METHODS OF PRESENTING DATA FROM EXPERIMENTS

Before we discuss the recording and presentation of data obtained from your experiments, there are a few terms that need be defined. When a body of knowledge becomes well developed it becomes possible to make predictions of certain outcomes. The beliefs on which the predictions are based are called axioms and the predictions are called hypotheses. Much of science advances by making predictions and testing them. This is called hypothesis testing. When predictions are not borne out, it usually indicates that axioms are not entirely correct or methods used were not appropriate. Thus, axioms are revised as predictions tested by properly designed experiments, fail to materialize. As an example, consider the problem of scientists who sought to determine if life existed in Martian soil.

You only test the null hypothesis. It can be accepted as true or it can be rejected as false. If accepted, then the prediction you have made as the hypothesis is untenable. In other words, your prediction is wrong. If the null hypothesis is rejected, then the hypothesis may be true but is not proven. Thus, you can be sure that your null hypothesis is wrong but your hypothesis is only supported, it is not established as a fact. The important distinction is that by rejecting the null hypothesis you show that your prediction is consistent with the axiom, not that the prediction is actually true.

Many of the exercises that we do in lab verify and confirm what you were taught in lecture. The function of lab exercises is to force you to critically evaluate the evidence you have on hand and see if the conclusions logically follow. Lab exercises are not intended to be an affirmation of established facts but a skeptical testing of dogma. In this light, you should establish the hypothesis that you are testing in the lab, use the lab exercise to test that hypothesis and discuss your results in light of rejecting or accepting the null hypothesis. As an example, consider the bacterial growth lab:

Axiom: Viable bacteria in the inoculum will grow rapidly and increase the turbidity of the medium.

Hypothesis: Absorbance of the culture medium will increase with time.

Null hypothesis: No change of absorbance will occur with time after an inoculum is introduced to the culture medium.

In some lab sections, the null hypothesis would have to be accepted because we observed that the absorbance either went down a little or stayed the same with respect to time. This indicates that something was not right with the axiom-- perhaps the bacteria were not viable, the culture medium was not appropriate or some other assumption about the method was violated. So you see that the lab does not have to work "correctly" for hypothesis-testing. There is enough information in the lab manual text for you to start with axioms, make some testable predictions and then use the given protocols to test your predictions.

The presentation of data (sing. datum) and their interpretation constitutes the core of any scientific investigation. There are many ways by which data can be presented. Each method is described in detail below.

Dr. Eby Bassiri **ebassiri@sas.upenn.edu**

Statements

The most common way of presentation of data is in the form of statements. This works best for simple observations, such as: "When viewed by light microscopy, all of the cells appeared dead." When data are more quantitative, such as- "7 out of 10 cells were dead", a table is the preferred form.

Tables

You should be familiar with the organization of information in tables from common experience. Here are some pointers:

1. The table should be identified by a number and have a title.

2. Experimental groups or treatments should be placed as rows in the table.

3. The first column should be labeled by identifying groups or treatments. Succeeding columns should contain measurements or observations on the groups.

4. The statistical analyses of data should be included in the table; i.e., means, a measure of deviation about the mean and sample size should be given

5. Units of measurement should be clearly stated for each column or row.

For example:

Table 1. Height of different letters on microscope slides as determined with the ocular micrometer.

Letter	Sample size	Mean (mm)	Standard deviation
I	10	0.11	0.05
E	9	0.09	0.03
K	10	0.13	0.04

Graphs

Graphs are commonly used scientific illustrations. There should be a good reason for using a graph rather than a table. Usually they are employed to show the functional relationship between dependent and independent continuous variables. An independent variable is one you can manipulate at will, such as the pH of a buffer or measurements during the time course of a reaction. This variable is plotted on the x-axis or abscissa. Dependent variables are the ones that are observed as the independent variable is changed, e. g., absorption, colony size, etc. The dependent variable is plotted on the y-axis or the ordinate. This convention allows the viewer to grasp the content of the graph easier because they intuitively view the x-axis and think to themselves- "at this level of treatment you get this response and at this higher level you see this much more effect". To invert the axes or plot discontinuous variables with the points connected with lines will confuse and mislead the reader.

Dr. Eby Bassiri **ebassiri@sas.upenn.edu**

It is important that you learn to read, interpret and make graphs because graphs are the best way to examine data in many instances. In biology, one usually needs only one quadrant from the standard Cartesian axes:

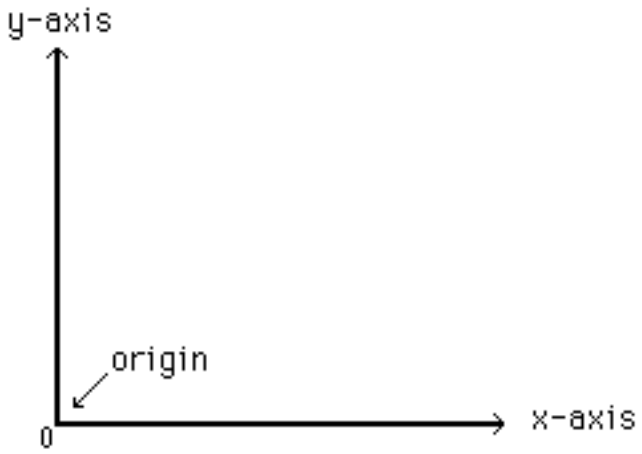

Fig. 1. Standard Cartesian axes.

The types of plots that one most often encounters in biology are (1) arithmetic and (2) semi-log. Arithmetic plots employ equal divisions between successive integers on both axes; semi-log plots use arithmetic scale on one axis and a logarithmic scale (equal divisions between successive powers of 10) on the other axis. (If you do not remember the meaning and use of $\log_{10}$ --and the ln function plus its relationship to $\log_{10}$ -- you should review them immediately because you will need them

In addition to the distinctions made above about the nature of variables, these points also pertain to graphs:

1. Give the figure a number and a label.

2. Use appropriate scales so that all of the graph is utilized.

3. Always label the axes and indicate the units.

4. Indicate the standard error in the measurement by vertical error bars that show the range, standard deviation or confidence interval (see Statistical Methods below).

5. Be cautious in the manner in which you connect data points. Obviously, you do not want to draw a line through a random scatter of points nor do you want to connect a zig-zagging line to connect these random points. Decide first if your data shows a non-random pattern. If the pattern is random omit the graph and state that in your text. If you think there is a pattern then use your judgment to fit a line or a curve through the data. This process should give you an appreciation of curve-fitting by statistical methods.

Dr. Eby Bassiri **ebassiri@sas.upenn.edu**

There are many errors that can be made in the making of graphs-- a number of the common ones are depicted in Fig. 2. Be sure that you understand the point of each illustration.

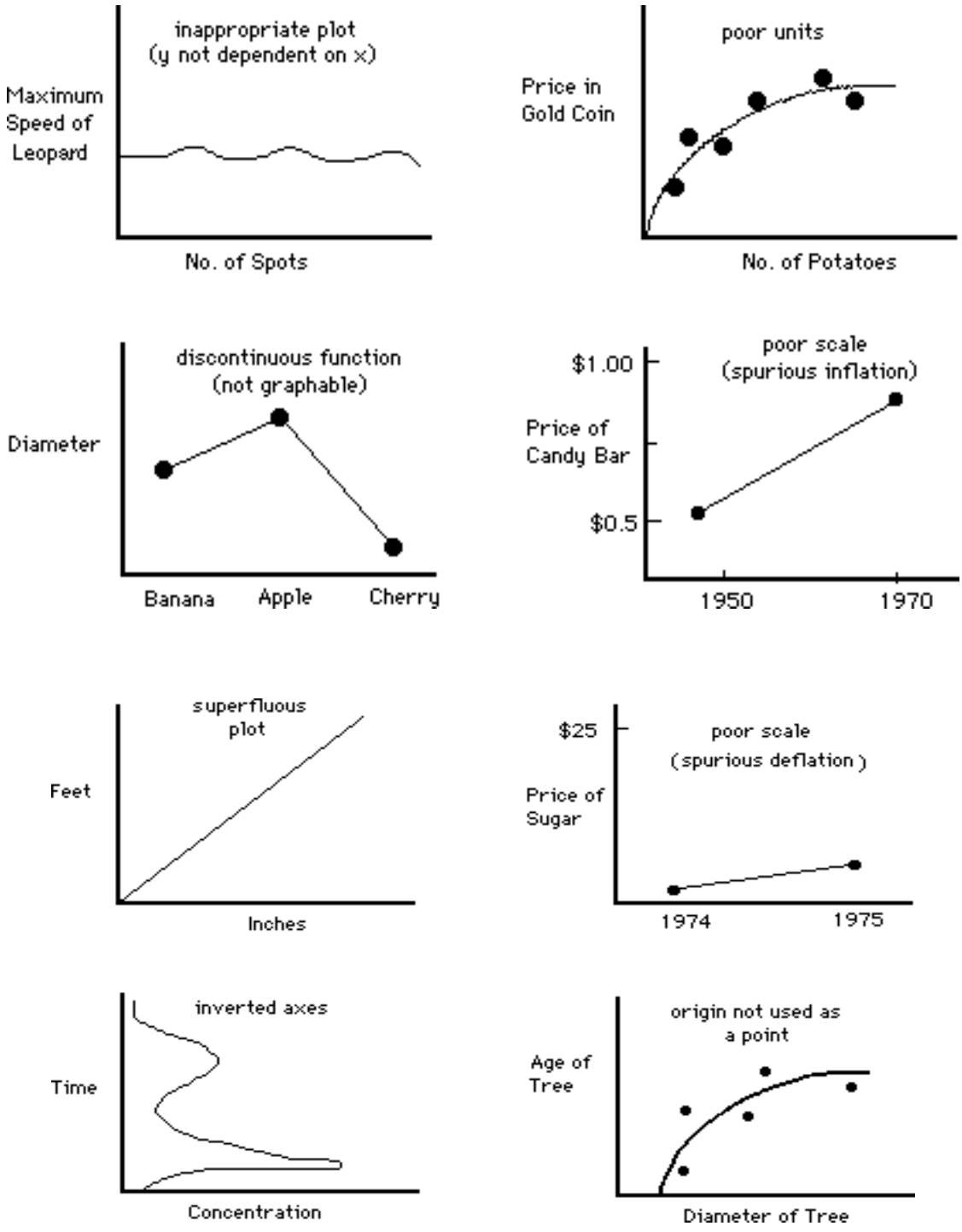

Fig. 2. Common errors in graphing.

Dr. Eby Bassiri **ebassiri@sas.upenn.edu**

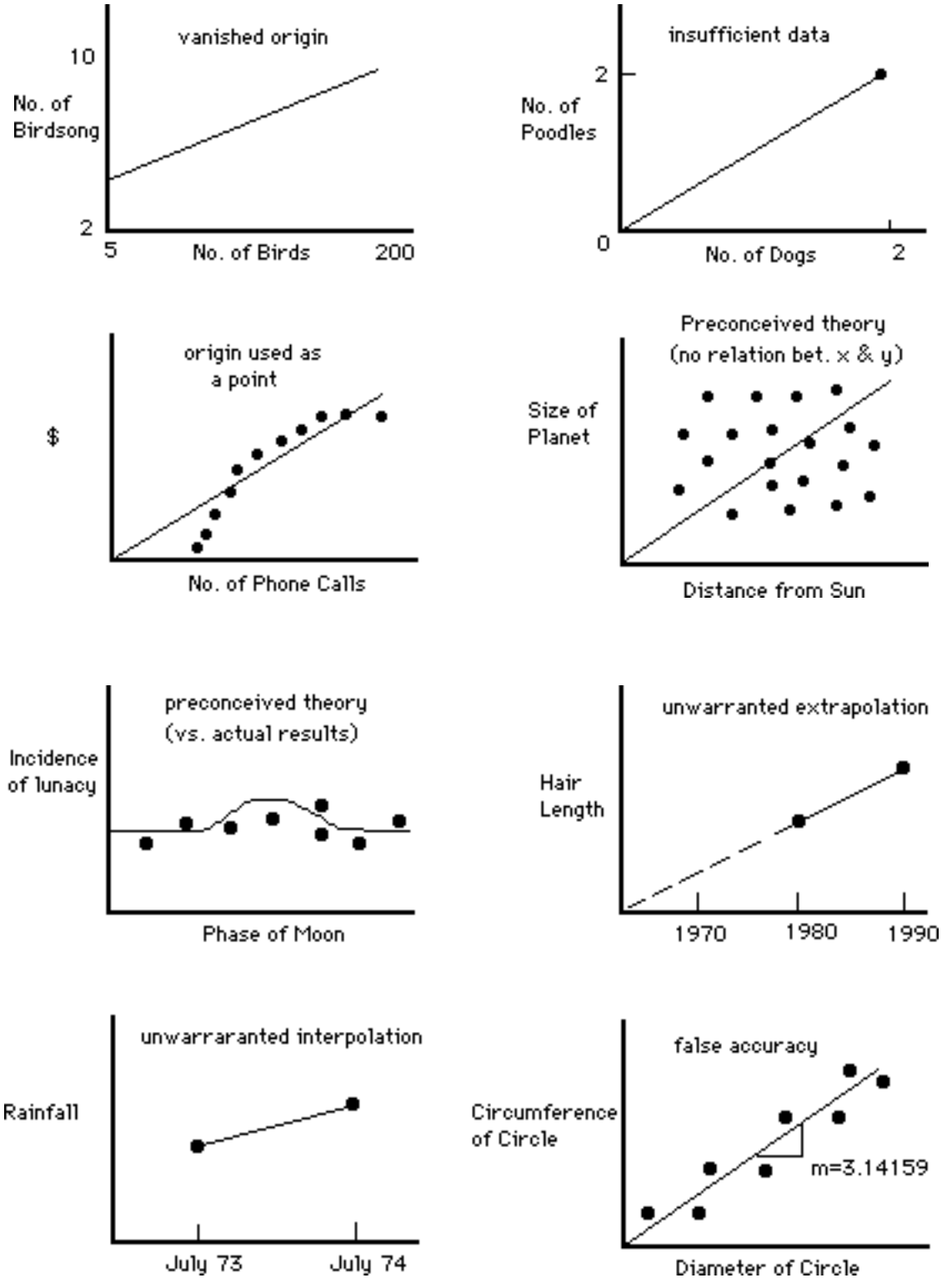

Fig. 2. Common errors in graphing (cont'd).

Dr. Eby Bassiri ebassiri@sas.upenn.edu

Making graphs can be quite easy if you keep a few basic rules in mind:

1. The independent variable (the one which the experimenter varies as he wishes e.g., time) is placed on the x-axis.

2. The dependent variable (the one that varies as a result of the variation in the experimental variable, e.g., number of cells present/ml) is placed on the y-axis.

3. All axes must be labeled with meaning and units.

4. Units should be appropriate and in meaningful proportion on the two axes.

5. The graph should fill as much of the sheets as possible -- no miniatures to be read with magnifying lenses allowed.

6. You may be asked to measure the slope of a line (m) you have drawn. To refresh your memory, use the x and y coordinates of two points on the line in this equation:

$$m = \frac{y_2 - y_1}{x_2 - x_1} \qquad (1)$$

Histograms

In some cases you have measured continuous variables in response to several discrete treatments. For example, if you measure bacterial populations in pond water in the winter, spring, summer and fall you should not graph the data as a graph with lines connecting the four seasons. The lines will mislead the reader into thinking you measured populations continually throughout the year. But you want to show the seasonal trend that is apparent. The solution is to use a histogram that shows the bacterial populations as columns representing the four seasons. The presentation preserves the logical order of the sampling times, permits graphing the number of bacteria as a continuous variable and allows seasonal trends to be readily grasped. The same pointers mentioned for graphs apply to histograms, with the exception of line fitting.

Statistical Methods

The inherent variability of biological material, coupled with the degree of error normally encountered in most measuring systems, demands the use of some form of data evaluation before valid conclusions and inferences can be drawn. It is particularly important that students of the biological sciences become familiar with statistical methods of handling data. The following section briefly outlines some of the basic methods used in this form of analysis.

The procedures fall into two groups. The first is concerned with those statistics that define the nature and distribution of the data. The second outlines procedures that may be used to compare two or more sets of data. This introduction to statistical methods is not to be regarded as complete, nor is it expected to substitute for formal training in this area.

Defining the data

In order to define a sample of data, one must have some knowledge of the central tendencies and degree of dispersion of the data. The statistics usually used for this purpose are the arithmetic mean, the standard deviation, and the confidence interval of the mean.

1. Arithmetic Mean

The mean ($\bar{X}$) is computed by summing (Σ) the individual sample measurements (X_i) and dividing by the total number of measurements (N):

$$\bar{X} = \frac{\sum Xi}{N} \qquad (2)$$

2. Standard Deviation

The mean of a group of data gives little information concerning the distribution of the data about the mean. Obviously, different numerical values can give the same mean value. For example:

Set 1: 32, 32, 36, 40, 40 $\quad \bar{X}_1 = 36$
Set 2: 2, 18, 24, 36, 100 $\quad \bar{X}_2 = 36$

The means for both sets of figures are identical. Yet the variation from the mean in Set 2 is so great as to make the average meaningless. The standard deviation is a measure of data dispersion about the mean. The range covered by the mean plus or minus one standard deviation includes about 68% of the data on the basis of which the mean was calculated. The range covered by the mean ± two standard deviations will include approximately 95% of the data. The calculation of the standard deviation is summarized below:

- Compute the arithmetical mean ($\bar{X}$) by summing (Σ) the individual measurements (X_i) and dividing by the total number of measurements (N):

$$\bar{X} = \frac{\sum X_i}{N}$$

- Calculate the deviation from the mean for each measurement: (X_i - $\bar{X}$).
- Square each of the individual deviations from the mean: (X_i - $\bar{X}$)^2. This allows one to deal with positive values.
- Determine the sum of the squared deviations: $\Sigma(X_i - \bar{X})^2$.
- Calculate the standard deviation(s) of the sample using the formula:

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N-1}} \qquad (3)$$

It would be a good exercise for you to see if you can derive the following from formula (3):

$$s = \sqrt{\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{N}}{N-1}} \qquad (4)$$

Dr. Eby Bassiri **ebassiri@sas.upenn.edu**

Formula (4) would be a short-cut, especially if the number of data points is large. An example is given in Table 2 where:

Number of individual measurements (N) = 10
Arithmetical means ($\overline{X}$) = 220/10 = 22
Degrees of freedom (N - 1) = 9

Using formula (3), we have:

$$s = \sqrt{\frac{\sum(X_i - \overline{X})^2}{N-1}} = \sqrt{\frac{42}{9}} = 2.16$$

And we come up with the same answer by using formula (4):

$$s = \sqrt{\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{N}}{N-1}} = \sqrt{\frac{4882 - \frac{220^2}{10}}{9}} = \sqrt{\frac{42}{9}} = 2.16$$

Table 2. Sample calculation of the standard deviation.

Observation number	Measurement (X_i) of stem length (mm)	X_i^2	Deviation from mean ($X_i - \overline{X}$)	$(X_i - \overline{X})^2$
X_1	20	400	-2	4
X_2	24	576	+2	4
X_3	22	484	0	0
X_4	19	361	-3	9
X_5	26	676	+4	16
X_6	22	484	0	0
X_7	24	576	+2	4
X_8	20	400	-2	4
X_9	22	484	0	0
X_{10}	21	441	-1	1
Total	220	4882	0	42
Stat. equivalent	ΣX_i	ΣX_i^2	$\Sigma(X_i - \overline{X})$	$\Sigma(X_i - \overline{X})^2$

3. Confidence Interval for a Sample Mean

The confidence interval (C. I.) for a sample mean equals the standard deviation of the sample (s) divided by the square root of the sample number (N) and multiplied by a factor (t) that is determined by the probability level desired and the value of the sample number:

$$C.I. = \pm t[\frac{s}{\sqrt{N}}] \qquad (5)$$

It is highly improbable that a sample mean, based on a relatively small series of data, will correspond exactly to the true mean calculated from an infinitely large sample of the population. It is necessary, therefore, to define a range within which the true mean might be

Dr. Eby Bassiri **ebassiri@sas.upenn.edu**

expected to lie. To define this range, the standard error of the mean (S.E. $\bar{X}$) must be known. The standard error equals the standard deviation divided by the square root of the number in the sample:

$$S.E.\bar{X} = \frac{s}{\sqrt{N}} \qquad (6)$$

The confidence interval is then calculated by multiplying the S.E. $\bar{X}$. by t, whose value depends on the number in the sample N, and the level of probability selected. Normally, a probability or significance level of 0.05 is accepted in biological studies. This implies that in only 5% of the samples taken separately from a given population would the parameters defined by the sample fail to have significance.

Table 3. Significance limits of student's *t* distribution

Degrees of freedom	Probability levels		
	0.10	0.05	0.01
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845
21	1.721	2.080	2.831
22	1.717	2.074	2.819
23	1.714	2.069	2.807
24	1.711	2.064	2.797
25	1.708	2.060	2.787
26	1.706	2.056	2.779
27	1.703	2.052	2.771
28	1.701	2.048	2.763
29	1.699	2.045	2.756
30	1.697	2.042	2.750
40	1.684	2.021	2.704
60	1.671	2.000	2.660
120	1.658	1.980	2.617
>120	1.645	1.960	2.576

Dr. Eby Bassiri **ebassiri@sas.upenn.edu**

The t values may be obtained from the t table (Table 3). These values are listed in columns for 0.10, 0.05, and 0.01 probability levels. Note that it is necessary to have a value for the number of degrees of freedom (D.F.) of the sample. In this case, the D.F. for the sample equals the number (N) in the sample minus one (N - 1).

Comparison of data

1. Standard Error of the Difference of Means

The standard error of the difference of means is computed by using the formula:

$$S.E.(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1}{N_1}.\frac{s_2}{N_2}} \qquad (7)$$

where s_1 and s_2 represent the standard deviations of two different groups, N_1 and N_2 represent the number of individuals in each group (preferably at least 20), and $\overline{X}_1$ and $\overline{X}_2$ are the respective means. Using this formula, if the difference between the two means is larger than two times the standard error of the difference, it can be concluded that the difference between the groups is not due to chance alone, but is due to the treatment given. It can be further concluded that similar plants or animals under similar treatment would be expected to respond in a similar manner.

2. Student's t Test

The student's t test is used to determine whether, within a selected degree of probability, two groups of data represent samples taken from the same or different populations of data. In other words, it is used to determine if two groups of data are significantly different. This test uses both the means and standard deviations of the two samples. It is calculated as:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)\sqrt{\frac{N_1 N_2}{(N_1 + N_2)}}}{\sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}} \qquad (8)$$

where s_1 and s_2 represent standard deviations of two different groups, N_1 and N_2 represent the number of individuals in each group, and $\overline{X}_1$ and $\overline{X}_2$ are the respective means.

The calculated t value is then compared to the value in the table (Table 3) at the probability level chosen and at the combined degrees of freedom of the two samples ($N_1 + N_2$ - 2). If the value for t is less than that found in the table, then the two groups of data are not considered significantly different at the chosen level of probability. If the value for t exceeds that in the table, then the two groups of data may be considered significantly different.

3. Chi-square Test

The statistical test most frequently used to determine whether data obtained experimentally provides a good fit, or approximation, to the expected or theoretical data, is relatively simple to carry out. Basically, this test can be used to determine if any deviations from the expected values are due to chance alone or to factors or circumstances other than chance.

The formula for chi square (χ^2) is:

$$\chi^2 = \sum[\frac{(O-E)^2}{E}] \quad (9)$$

where O = the observed number of individuals
E = the expected number of individuals and
$\sum$ = sum of all values of $(O - E)^2/E$ for the various categories of phenotypes.

The following example shows how this type of analysis can be applied to genetics data. It can be used in many other analyses as long as numerical data are used and not percentages or ratios.

In a cross of tall maize (corn) plants to dwarf plants, the F_1 generation consisted entirely of tall plants. The F_2 generation had 84 talls and 26 dwarfs. The question we want to answer is whether this F_2 data fits the expected 3:1 monohybrid ratio. Using the data given in Table 4, chi square was calculated.

$$\chi^2 = \sum[\frac{(O-E)^2}{E}] = [0.027 + 0.082] = 0.109$$

Table 4. Summary of the calculations of chi squares for the hypothetical cross.

Phenotype	Genotype	O	E	O - E	$(O - E)^2$	Chi-Square
Tall	T-	84	82.5	1.5	2.25	0.027
Dwarf	tt	26	27.5	-1.5	2.25	0.082
Total		110	110	0	4.50	0.109

What does this chi-square value of 0.109 mean? Consider that if the observed values were exactly equal to the expected values (i.e., O = E) then we would have a perfect fit, and χ^2 would equal zero. Thus, if you obtain a small value of χ^2, this would indicate a close agreement of the observed and expected ratios, whereas a large value for χ^2 would indicate marked deviation from the expected ratio. However, deviations from the expected values are always bound to occur due to chance alone. The question is: "Are the observed deviations within the limits expected by chance alone?" Statisticians have generally agreed, for these types of studies, on the arbitrary limits of 1 chance in 20 (probability = 0.05 = 5%) for making the distinction between acceptance or rejection of the data as fitting the expected ratio.

The chi-square value for a two-term ratio (i.e., 3:1) that corresponds to a 0.05 or 1 in 20 probability is 3.841 (see Table 5). Therefore, you would expect to obtain this value, due to chance deviations only, in only 5% of similar trials if the hypothesis is true. When χ^2 for this two-term ratio is larger than 3.841, then the probability that the variation is due to chance alone is less than 5% or 1 in 20. You would therefore reject the hypothesis that the observed and expected ratios are in close agreement.

Dr. Eby Bassiri **ebassiri@sas.upenn.edu**

In our example, χ^2 is 0.109, which is considerably less than 3.841. Thus we can say that the variation between the observed and expected values is due to chance alone and accept the data as fitting the 3:1 ratio.

Where did we obtain the value 3.841? Mathematicians have developed a variety of statistical tables. Table 5 is an example of a table of chi-square values. The table is set up so that probability (P) values extend across the top, and "degrees of freedom" values are down the left margin. The number of degrees of freedom in test of genetic ratios is generally always one less than the number of classes in the ratio being analyzed. Thus in tests of such ratios as 1:1 or 3:1 there is one degree of freedom, a test of a 1:2:1 ratio would have two degrees of freedom, and a test of a 1:2:1:2:4:2:1:2:1 would have eight degrees of freedom.

Table 5. Distribution of $\chi 2$

n	Probability (P)									
	0.99	**0.95**	**0.90**	**0.80**	**0.70**	**0.50**	**0.10**	**0.05**	**0.01**	**0.001**
1	0.0002	0.00393	0.0158	0.0642	0.148	0.455	2.706	3.841	6.635	10.827
2	0.0201	0.103	0.211	0.446	0.713	1.386	4.605	5.991	9.210	13.815
3	0.115	0.352	0.584	1.005	1.424	2.366	6.251	7.815	11.345	16.268
4	0.297	0.711	1.064	1.649	2.195	3.357	7.779	9.488	13.277	18.465
5	0.554	1.145	1.610	2.343	3.000	4.351	9.236	11.070	15.086	20.517
6	0.872	1.635	2.204	3.070	3.828	5.348	10.645	12.592	16.812	22.457
7	1.239	2.167	2.833	3.822	4.671	6.346	12.017	14.067	18.475	24.322
8	1.646	2.733	3.490	4.594	5.527	7.344	13.362	15.507	20.090	26.125
9	2.088	3.325	4.168	5.380	6.393	8.343	14.684	16.919	21.666	27.877
10	2.558	3.940	4.865	6.179	7.267	9.342	15.987	18.307	23.209	29.588

The general idea of degrees of freedom can be exemplified by the situation encountered by a small boy when he is putting on his shoes. He has two shoes, but only one degree of freedom. Once one shoe is filled by a foot, right or wrong, the other shoe is automatically committed to being right or wrong too. Similarly, in a two-place table, one value can be filled arbitrarily, but the other is then fixed by the fact that the total must add up to the precise number of observations made in the experiment, and the deviations in the two classes must compensate for each other. When there are four classes, any three are usually free, but the fourth is fixed. Thus, when there are four classes, there are usually three degrees of freedom.

In our example, we have two classes in the ratio (i.e., 3:1) and therefore have one degree of freedom when we interpret the chi-square table. Look at the one-degree-of-freedom row under the .05 probability column, and you will find the value 3.841. This number represents the maximum value for chi-square that you should be willing to accept and yet consider the deviations observed as due to chance alone. If you were willing to accept a P value representing 1 chance in 10, what value of chi-square would you accept as maximum?

In our example, chi-square was calculated to be 0.109. Looking across the one-degree-of-freedom line in Table 5, we find that this value falls between the 0.70 (χ^2 = .148) and the 0.80 (χ^2 = .0642) columns. This says that the probability that the deviations we obtained from the expected values could be attributed to chance alone is 70-80%. That is, if we were to repeat the study 100 times, we would obtain deviations as large as those observed about 70% of the time (i.e., 7 out of every 10 experiments). We can thus reasonably regard this deviation as simply a sampling, or chance, error.