3

Statistical Analyses

Some studies produce unambiguous results, in which case we do not need statistics. In most cases, however, we need some objective way to evaluate differences in our results. To provide a way to evaluate results with some degree of objectivity, we can use diverse statistical techniques, the subject of this chapter.

As mentioned in Chapter 2, the core statistical notion (provided by Sir Ronald A. Fisher) was that of seeing whether the effects of some variable of interest are likely to be larger than the effects of chance variation.[1] Statisticians have devised many procedures to do such comparisons and to establish relationships among variables.

Most statistical texts start, reasonably enough, by introducing the reader to the simpler ways by which to see how well we know the mean of a sample, and how sure we might be that it differs from the mean of a hypothetical population. Then they go on to tests that compare two sample means, and so on. I did not follow that pattern in this book, because this is not a book on statistics, but rather an introduction to principles (not to techniques) of doing science. I would have preferred to go right away to principles of design of scientific work, but that turned out to be difficult without some previous discussion of statistical concepts. Therefore, in this chapter I review a few statistical tests before going on to principles of experimental design in chapter 4, to provide readers with terms and strategies of data analysis. Some readers might want to read chapter 4 first and return to this chapter as needed. For the sake of reference, I do review the array from simpler to more complex tests in section 3.5.

Throughout, I refrain from entering into arithmetical details for each test, because these can be found in the many excellent statistics textbooks. Motulsky (1995) provides a lucid intuitive introduction to statistical analyses. Sokal and Rohlf (1995) give a thorough and authoritative review of the methods. Here we will emphasize concepts, but we will have to do a bit of algebra to sort out the concepts.

> [W]here measurement is noisy, uncertain, and difficult, it is only natural that statistics should flourish.
>
> *S. S. Stevens*

1. Chance or random variation is another way we refer to variation caused by additive contributions from many and unidentified variables. This is the "left-over" variation against which we want to compare the variation caused by the treatment we are studying.

This chapter therefore introduces the concepts underlying some selected kinds of statistical analyses, emphasizing the strategy of the tests, and what the tests are useful for. Out of the plethora of statistical methods available, I single out analysis *of variance,* regression, correlation, and *analysis* of frequencies. These provide the wherewithal to analyze data from most types of research discussed in chapter 1, are most frequently used in analyses that readers will encounter in the scientific literature, and provide the terms needed for chapter 4.

The chapter ends with a discussion of transformations of data. These are useful tools to better understand the nature of our data, and are also devices by which we can recast data so as to meet the assumptions of several of the statistical tests.

3.1 Analysis of Variance

Elements of ANOVA

The analysis of variance [a phrase usually shortened to ANOVA] was developed by the English statistical pioneer Sir Ronald A. Fisher. The ANOVA is fundamental to much of statistical analysis and to the design of experiments. It is a genera1 method by which we can compare differences *(as* variances) among means and assess whether the differences are larger than may be due to chance alone.

The ANOVA is applied widely in scientific literature. A survey of uses of ANOVA, however, showed that they were applied deficiently in 78% of the papers examined [Underwood 1981). The science community needs more critical application, reporting, and interpretation of this most useful statistical tool. Here we review only some basic principles.

Analysis of variance allows the separate calculation of estimates of variance attributable to treatments (or other components], by assuming that the various effects on a variable of interest are additive. The assumption of additivity is a core idea underlying the ANOVA, and leads to the notion that any value of a variable can be decomposed into components

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where $i = 1, \ldots, a$, and $j = 1, \ldots, n$. A given measurement of Y_{ij} is thus assumed to be made up of the sum of several terms. First, there is an effect due to being a Y, which is indicated as μ, the grand mean of all the values of Y. Then there is a term α_i that describes the effect of belonging to a subgroup of values of Y that we will call the treatment, and for which we ask the difference from the overall population. We answer that question by means of a third term, the *error,*[2] ε_{ij}. This third component of Y_{ij} represents the random variations in the jth individual value of Y from the ith group. The idea is that the random variation is the variability that is left after we have separated the effects of the grand mean and the groups (or treatments). For this ε_{ij} term to be truly random, the observations within

2. Statistical jargon uses the term *error* to refer to random variation, not to our more common use implying a blunder.

groups must have been taken at random from among the population of values. The mean of all the ε_{ij} has to be equal to zero; some of the deviations will be from values larger, and some smaller, than the mean of the distribution. The estimated variance of ε_{ij} is s^2.

These assumptions are another way to say that the observations must be independent of each other, and that the distribution of ε_{ij} must be normal. We also assume that variances are homogeneous, that is, that since s^2 calculated from different samples of observations estimates the same population σ^2, the s^2 values must be similar. As is the practice, Greek letters are used to indicate that we are referring to parameters, rather than statistical estimates.

The assumptions made for ANOVA, therefore, are *additivity* of components of variation, independence of the observations, homogeneity of variances, and *normality* of the observations. These assumptions are too often ignored in day-to-day analysis of scientific data. Too few of us actually carry out preliminary analyses to see if indeed our data do meet the assumptions. Although the various statistical procedures are fairly tolerant of violations of the assumptions, understanding of the assumptions is important because they have repercussions, as we will see in chapter 4, in the design of research as well as in the method of data analysis.

If our data violate the assumptions, there are two alternatives. The first option may be to apply a different suite of statistical tests that make no assumptions about distributions. Below we discuss nonparametric equivalents of parametric methods that can be applied to data that do not meet the assumptions of parametric tests. The second option is to transform the data into new scales that do meet the assumptions, and then carry out the appropriate ANOVA on the transformed data. Several transformations are available to solve different problems, as we also discuss below.

Examples of Types of ANOVA

Replicated One-way ANOVA

To make more real the concept of ANOVA, we examine first an example of one of the simplest versions: a one-way replicated ANOVA. This layout is applicable to test the effects of a variable or classification. Suppose we are interested in evaluating the firmness of sand along a series of stations on a beach. We use an instrument called a penetrometer to measure the resistance to displacement by sand; the smaller the number, the smaller the force need to penetrate the sand. We take five randomly located measurements at each of six stations along the beach (table 3.1).

Now, we could simply calculate standard errors for each of the means, and judge whether the means are likely to differ by seeing if the values for (mean ± se) for the different means overlap. That is a qualitative judgment; here we want a more quantitative assessment of the hypothesis that there are no differences among the means. We can see that there are differences among the stations (the statisticians want to have us refer to our stations as the groups). The issue is whether the variation amonggroups is larger than the *within-group* variation (the variation among replicates collected at each station, also called the error term).

Table 3.1. Measurements of Force Needed for Sand Penetration (Relative Units) Obtained from Five Replications at Each of Six Beach Stations.

	Station					
Replication	1	2	3	4	5	6
1	21	31	30	47	52	38
2	52	42	27	38	44	40
3	29	37	30	41	52	25
4	20	51	42	32	35	31
5	30	44	46	41	48	39
Totals	152	205	175	199	231	137
Means	25.3	41	35	39.8	46.2	27.4

Data from example used by Krumbein (1955)

To make this comparison, we first ascertain that the data meet the assumptions of ANOVA. It is easy to examine the data graphically to check on normality by means of a frequency histogram (fig. 3.1, left) and on homogeneity of variances by plotting variances versus means (right). The data are reasonably normal. The variances are similar, except for the one for station 1, which is about three times as large as the others. To decide whether the variances are homogeneous, we might try *Bartlett's test* (Sokal and Rohlf 1995, chap. 13) or the simpler *Hartley's* test. When we do these tests, we find that the variances in this data set do not differ sufficiently to invalidate the assumption. Variances have to differ more, as well as increase with the mean, to be a problem.

The data therefore are reasonably normal, and the variances do not change significantly in relation to the magnitudes of the means. To check for additivity we might calculate deviations from the overall mean, and see if the deviations are approximately similar for all groups. The other assumptions are likely to be less of a potential problem. In this case, we decide not to transform the data. Having checked the assumptions, we proceed to calculate variances; table 3.2 shows one way to organize the

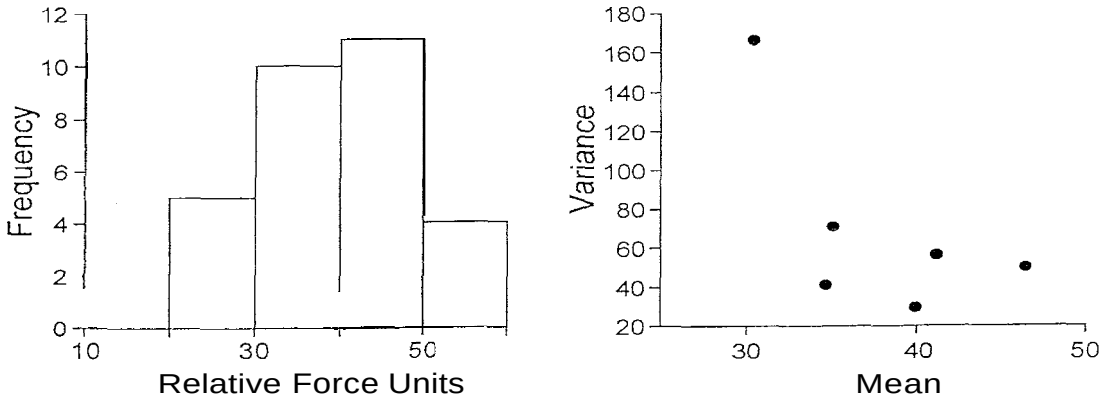


Fig. 3.1 Graphical examination of normality of frequency distribution (left) and homogeneity of variances (right) for the data of table 3.1.

Table 3.2. Analysis of Variance Procedure

source of Variation	Sum of Squares (SS)	Degrees of Freedom (*df*)	Mean Square (MS)	Estimate of of Variance	F Test
Among groups	$SS_a = \sum(C_i^2/n) - CT$	$k - 1$	$MS_a = SS_a/(k-1)$	$\sigma^2 + \sigma g^2$	MS_a/MS_w
Within groups	$SS_w = SS_t - SS_a$	$k(n-1)$	$MS_w = SS_w/k(n-1)$	σ^2	
Totals	$\sum(X_i)^2 - CT$	$kn - 1$			

55, and SS_w refer to sums of squares among and within groups. The value X_i represents an observation. C, is the total for each column in table 3.1; $\sum$ indicates the process of summation across rows or down columns in table 3.1. The "correction term" CT is G^2/kn, where G is the grand total. Degrees of freedom (*df*) are arrived at by the number of observations we made ($k = 6$ stations, $n = 5$ replicates), minus 1. We then divide SS_a and SS_w by *df* to get the mean squares (MS_a and MS_w). The mean squares, in turn, are our estimates of among-group and within-group variances. The value to be used in the *F* test is obtained by dividing by the within-group estimate of variation, and separates out the variation due to among-group variation. If $F = 1$, the variation among groups is the same as the variation within groups, and there is no group effect.

procedure. [I have added tables such as this and others for those readers desiring an explicit account.) Having done these calculations, we can now put together the ANOVA table for the beach firmness data (table *3.3).*

The ANOVA allows us to test whether differences among groups are significant relative to random variation estimated by the within-group terms. These tests are carried out using the F distribution, so named in honor of Fisher. The ratios of the estimated variances of a treatment relative to random variation are compared to F values that vary depending on the degrees of freedom associated with the two estimates of variances being tested.

So, the F value we get in table 3.3 is *2.28.* We look up the range of values for the *F* distribution in tables provided in most statistics texts, and find that, for 5 and 24 df, an F value has to be larger than 2.62 to be significant at the 5% probability level. The value in table *3.3* does not exceed the 5% cutoff, and we report this finding by adding "NS" after the F value, for "not significant." Incidentally, the convention is that if the F value is significant at the 5% or 1% probability level (i.e., if the calculated F is greater than 2.62 for $\alpha = 0.05$ or the corresponding value for $\alpha = 0.01$), the F value is followed by one or two asterisks, respectively.

In any case, by comparison with the table of *F* values, we conclude that the null hypothesis cannot be rejected: firmness of sand over the beach in question is homogeneous over the distances sampled. The mean firmness of *37.8,* calculated from all measurements, and the within-group variance of 69.17 can be taken as estimates of the population mean and variance.

Table 3.3. Analysis of Variance Table for Data of Table 3.1

Source of Variation	SS	Degrees of Freedom (*df*)	MS	F
Among groups	788	5	157.60	2.28NS
Within groups	1660	24	69.17	—
Totals	2448	29	—	—

SS = sum of squares; MS = mean squares. NS = not significant.

Analysis of variance might tell us that there are significant differences among the groups or treatments, but if we were testing different kinds of insect repellent or airplane wing design, we woulcl want to know *which* of the treatments differed. To do this sort of comparison, people have applied *t* tests or other techniques for comparisons of means.

Differences between two specific means are often tested with the *t* test, which is a special case of the more general ANOVA. Application of the *t* test to multiple means is problematic, although commonly done. If we have five means, we have at least 10 possible *t* tests, if the means are ordered by size. In this context, degrees of freedom tell us how many comparisons are possible. With five means we have $(n - 1)$ degrees of freedom, or four comparisons possible (one *df* is taken up when we estimate the overall mean of the values). Thus, multiple *t* tests, if they are done at all, need to be limited to four comparisons, and the comparisons have to be selected before we see the results. The problematic issue of multiple tests is a general difficulty; as I have already mentioned.

In addition to the matter of using degrees of freedom that we do not really have, multiple tests often run the risk of committing Type II errors. As mentioned in chapter 2, whether we make 20 or 100 comparisons among a set of means, at the 5% probability level by chance alone we expect 5% (1 or 5 tests, respectively) to be declared significant, *even if* the difference is not truly significantly different. Indiscriminate application of multiple tests is not a desirable practice, because we are courting Type II errors.

There are many kinds of *multiple comparison* tests developed to examine differences among sets of means in rather specific situations. Statisticians do not agree about the use of such tests. Some suggest cautious use (Sokal and Rohlf 1995), but others think that "multiple comparison methods have no place at all in the interpretation of data" (O'Neill and Wetherill 1971). Mead (1988) recommends strongly that multiple comparison methods be avoided and that critical graphical scrutiny be done instead.

At this point we have to note that there are two different types of ANOVA. In Model I ANOVA the tseatments are fixed. Treatments could be fixed by the researcher, as in testing the effects of different drugs on patients or of different dosages of fertilizer on a crop. Treatments may also be classifications that are inherently fixed, such as age of subjects, color, or sex. For example, we could test whether weights of Italian, Chinese, and U.S. women differ by collecting data in the three sites. Note that in some Model I situations the researcher knows the mechanism behind the presumed effects, but in other cases, such as the women's weight question, we deal with a complex set of unidentified mechanisms that determine the variable.

In Model II ANOVAS, treatments are not fixed by nature or by the experimenter, but are chosen randomly. Examples of this may be a study of concentration of mercury in 30 crabs that were collected in each of three sites, and the sites were chosen randomly. We do not know what might be the meaning of differences among sites. The question this design allows us to ask is whether among-group (sites) variation is larger than within-group variation. If the F test is significant, the inference from a Model II ANOVA is that there was a significant added variance component

associated with the treatment, while the inference from a Model I analysis is that there was a significant treatment effect.

It is not always easy to differentiate between the two kinds of ANOVAS. For example, if the sites selected in the beach firmness study were chosen at random from among many beaches, the study would be Model II. On the other hand, if we selected specific positions along the elevation of the beach, to correspond to specific locations, or geological features (beach face, berm, crest, etc.), the study would fit a Model I ANOVA. The identity of the model to be used matters because, as we saw above, the inferences differ somewhat, and the calculations for the two types of ANOVA differ to some extent (see Sokal and Rohlf 1995, chap. 8). In the end, the differences in conclusions reached via a Model I or II analysis are a matter of nuances meaningful to the statistically versed. The larger benefit of considering whether we apply a Model I or Model II analysis is that it fosters critical thinking about how we do science.

Multiway ANOVA

So far we have concentrated on ANOVAS in which the data are classified in one way. One of the reasons why the ANOVA has been an attractive way to scrutinize data is that it is applicable to much more complicated data sets. For example, in our examination of the weights of women from Italy, China, and the United States, we might be concerned with the matter of age, so we might want to do the analysis separating groups of females of different ages. In this case, we have a data set with two-way classification: country and age. We might further be interested in asking whether women from urban or rural settings respond differently; in this case we have a three-way ANOVA. Such multiway classifications can be rather powerful analytic tools, allowing us to inquire about important and subtle issues such as the possible interactions among the treatment classifications. These studies permit asking of questions such as, "Do the age-related differences remain constant in rural settings, regardless of country of residence?" Of course, the offsetting feature is that actually carrying out such studies and doing their analysis becomes progressively more demanding as the variables multiply. ANOVA layouts are diverse, and can be used to investigate many levels of several variables. Here we limit discussion to two types that introduce the essential concepts.

Unreplicated Two-Way ANOVA. We can run an experiment in which we have two treatments that are applied to experimental units (table 3.4). For simplicity and generality, we can use *Columns* and *Rows* as the names of the two treatments. If we have fixed groups (Model I), we take it that the observations are randomly distributed around a group mean (x_{ij}); if we have random groups (Model 11), the observations are randomly distributed around an overall mean for the groups (x). We can set out the procedural concepts as in table 3.5, a slightly more complicated ANOVA table than table 3.2. If we are dealing with Model I ANOVA, we test the row and column effects by dividing their mean squares (MS) by the error MS; the divisions sort out the effects of both treatments (rows and columns) from random error. If we have a Model II ANOVA, we have to calculate the

Table 3.4. Layout of Unreplicated Two-way ANOVA.

	Columns				
Rows	1	2	j	c	Row Totals
1	X_{11}	X_{12}	X_{1j}	X_{1c}	R_1
2	X_{21}	X_{22}	X_{2j}	X_{2c}	R_2
i	X_{i1}	X_{i2}	X_{ij}	X_{ic}	R_i
r	X_{r1}	X_{r2}	X_{rj}	X_{rc}	R_r
Column totals	C_1	C_2	C_j	C_c	G

The X_{ij} in the cells are the observations, and R_i, C_j, and G are the row, column, and grand totals.

components of variation from the last column in the table. For example, for the row variance, the residual MS is subtracted from the row MS, and the difference is divided by the number of columns.

Replicated *Two-Way ANOVA.* The unreplicated two-way layout is seldom used in research, but it is a template for many elaborations of experimental design. Depending on the questions we ask, and the material available, we can add replicates at each row-by-column cell, we can split cells, we can run an experiment with only partial columns or rows, we can make the groups be levels of a factor, or we can use one of the variables to isolate uninteresting variation so that the effects of the treatment of interest are better evaluated. Some of these strategies of treatment design are dealt with in chapter 4. Mead (1988) is an excellent reference for all these designs.

Multiway replicated layouts are most useful to study the simultaneous effects of two or more independent variables on the dependent variable. This joint influence is referred to as the interaction of the independent variables and is a powerful concept made available only by this type of analysis. The multilevel layout makes possible the investigation of joint effects of variables, something that no amount of study of the separate factors can reveal. We have to note, however, that with an unreplicated design the joint effect of the two variables is not separable from the random, residual variation. This separation becomes possible only when we have replicates within cells affected by both

Table 3.5. Analysis of Variance Table for Layout of Table 3.4.

Source of Variation	SS	df	MS	Estimate of Variation	F Test
Rows	$\sum(R_i^2/c) - CT$*	$r - 1$	$SS_R/(r-1)$	$\sigma^2 + c\sigma_R^2$	MS_R/MS_e
Columns	$\sum(C_j^2/r) - CT$	$c - 1$	$SS_C/(c-1)$	$\sigma^2 + r\sigma_C^2$	MS_C/MS_e
Residual variation (or error:	$SS_G - (SS_R + SS_C)$	$(r-1)(c-1)$	$SS_e/(r-1)(c-1)$	σ^2	
Total	$SS(X_{ij}^2) - CT$	$rc - 1$			

*CT = "correction term," a shoe-hand way to refer to remainder variation.

r and c are total number of rows, and total number of cells within a row, respectively. SS_e and MS_e are error sum of squares and error mean square, respectively.
For other definitions of terms, refer to tables 3.2 and 3.4.

independent variables. This is the major reason for replicated multiway ANOVAS.

Suppose that instead of the X_{ij} observations in cells of the unreplicated two-way layout above, we set out n replicates, so we have X_{ijn} observations. Since it is awkward in this situation to refer to rows and columns, we discuss this design as involving two factors, A and B, both of which are applied to or affect n replicates. The layout (table 3.6) is called *cross-*classified if each level of one factor is present at each level of the second factor. In this kind of analysis, it is advantageous if that equal replication be present in all cells; missing replicates or unbalanced designs require much additional computational effort.

The model for such an analysis, where there are two factors, A and B, and cells hold n replicates, is

$$X_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk}.$$

In this equation, X_{ijk} represents the kth replicate ($k = 1, \ldots, n$) in the treatment combination of the ith level of factor A and the jth level of factor B. A, and B_j are the effects at the ith and jth levels of factors A and G. We will test the hypothesis that neither the A, B, nor AB effects are significant by the tests implicit in table 3.7.

The models of expected MS differ when A and B are random or fixed (table 3.8). It is not always obvious which MS should be in the numerator and which in the denominator of Ftests with multiway ANOVA designs of this level of complexity or greater. The distinction between random and fixed models becomes more important with more complex layouts, because, as in table 3.8, the model determines which MS we divide by to examine the significance of the effects of factor and interaction terms. Mead (1988) gives rules by which we can select the appropriate MSs to use in Ftests. Table 3.8 is no doubt daunting; it is included here as a signpost to warn the reader that at this level, the statistical analyses may be powerful but increasingly complicated.

If you have gotten to this stage on your own, you will find it a good idea to consult a statistician about these analyses before going on with your work. In fact, experience teaches that it is wise to consult with someone with statistical expertise *before* starting research that demands experimental designs described in this section; otherwise, much time and effort may be lost.

Table 3.6. Layout of a Replicated, Cross-Classified Two-way ANOVA.

	Variable A			
Variable B	Subgroup 1		Subgroup 2	
Subgroup 1	X_{111}	X_{112}	X_{211}	X_{212}
Subgroup 2	X_{121}	X_{122}	X_{221}	X_{222}

In this case, only two replicate assertions are included. "Subgroups" could refer to a classification (e g , males and females) *or* a level (e.g., doses X and 3X of a given chemical treatment).

Table 3.7. Analysis of Variance Formulas for Data of Table 3.6

Source of Variation	Sum of Squares	Degrees of Freedom
Factor *A*	$\frac{\sum^{a}\left(\sum^{b}\sum^{n} X_{ijk}\right)^2}{bn} - K$	$(a-1)$
Factor B	$\frac{\sum^{b}\left(\sum^{a}\sum^{n} X_{ijk}\right)^2}{an} - K$	$(b-1)$
$A \times B$	$\frac{\sum^{a}\sum^{b}\left(\sum^{n} X_{ijk}\right)^2}{n} - K - SS_A - SS_B$	$(a-1)(b-1)$
Within cells	$\sum^{a}\sum^{b}\left[\sum^{n} X_{ijk}^2 - \frac{\left(\sum^{n} X_{ijk}\right)^2}{n}\right]$	$ab(n-1)$
Total	$\sum^{a}\sum^{b}\sum^{n} X_{ijk}^2 - K$	abn – 1

The "correction term" in this case is $K = \left(\sum^{a}\sum^{b}\sum^{n} X_{ijk}\right)^2 / abn$.

Nonparametric Alternatives to ANOVA

If transformations do not manage to recast data so that assumptions of ANOVA are met, we can opt for nonparametric alternatives. These are procedures that are distribution-free, in contrast to ANOVA, which makes assumptions as to parametric distributions underlying the test. For single samples, groups, or classifications, the *Kruskal–Wallis* test is available. For tests comparing two samples, the *Mann–Whitney U* or the Wilcoxon two-sample tests are recommended; both these nonparametric tests are based on rankings of observations, and calculations of likelihood of deviations from chance. The *Kolmogorov–Smirnov* two-sample test assays differences between two distributions.

Where we need nonparametric alternatives to parametric Model I two-way ANOVA, the *Friedman's two-way* test is appropriate. Where data are paired, *Wilcoxon's* signed *ranks* test is available. Both of these methods

Table 3.8. Estimates of Mean Squares for Replicated Two-way ANOVAS of Different Model Types

Layout in which		Mean Squares Estimate the Following			
A is	B is	Within Cells $ab[df = (n-1)]$	$A \times B$ $[df = (a-1)(b-1)]$	B $[df = (b-1)]$	A $[df = (a-1)]$
Fixed	Fixed	σ_e^2	$\sigma_e^2 + n\sigma_{AB}^2$	$\sigma_e^2 + anK_B^2$	$\sigma_e^2 + bnK_A^2$
Fixed	Random	σ_e^2	$\sigma_e^2 + n\sigma_{AB}^2$	$\sigma_e^2 + anK_B^2$	$\sigma_e^2 + n\sigma_{AB}^2 + bnK_A^2$
Random	Fixed	σ_e^2	$\sigma_e^2 + n\sigma_{AB}^2$	$\sigma_e^2 + n\sigma_{AB}^2 + anK_B^2$	$\sigma_e^2 + bnK_A^2$
Random	Random	σ_e^2	$\sigma_e^2 + n\sigma_{AB}^2$	$\sigma_e^2 + n\sigma_{AB}^2 + anK_B^2$	$\sigma_e^2 + n\sigma_{AB}^2 + bnK_A^2$

From Underwood (1981).

The "correction terms" in these cases are $K_B^2 = \sum^{b}(B_j - \bar{B})^2/(b-1)$, $K_A^2 = \sum(A_i - \bar{A})^2/(a-1)$

depend on analyses of ranked data. A much simpler test is the sign *test,* which merely counts the number of positive and negative differences among pairs of data and then ascertains whether the frequencies of + and – are in equal proportions.

3.2 Regression

Elements of Regression

In the ANOVA we have in reality been considering the effects of a variety of treatments on one dependent variable. That is, we had categories that we called treatments, and we *measured* values of a dependent variable in the experimental units. Regression addresses the more general case of measurements of two variables.

In regression, we express the relationship of one variable to another by an equation that describes one as a function (linear in the simplest case) of the other variable. The regression can be $Y = \alpha + \beta X$ and $dY/dX = \beta$, where Y is the *dependent variable,* α is the *intercept,* X is the *independent variable,* and β, the slope of the line, is called the *regression coefficient.*

Regression merely establishes the form of the function that links X and Y. Regression cannot by itself establish a causal link between the two variables. To ascertain whether changes in the independent variable X lead to changes in the dependent variable Y, we need to apply manipulative experimental approaches discussed above.

In any data set, we expect that the points lie in a scatter around a regression line whose intercept is α and slope is β. The line merely represents the position of the expected values, if three assumptions are met. This model of regression requires the following:

"Regression"?

"Regression" sounds odd to us today, since in lay use this word has a fairly negative connotation. It was used in a rather different way by Sir Francis Galton in a paper published in 1885, to describe the relationship between adult height5 of children and of their parents. He first used the term "reversion" in a lecture, but he finally titled the paper "Regression towards mediocrity in hereditary stature." Our reaction to his use of words is a reflection of changes in usage; we must not think his intention was to suggest a degrading descent to undesirable (but inherited) height, which is what the title might mean to us today. In any case, statisticians have retained the term to describe the relationship between variables.

The paper is also notable because it contains one of the earliest bivariate plots (see frontispiece for this chapter). Curiously, the data, and Calton's treatment of them, are more of a correlation than a regression as we might consider it today. The plot also includes a derived variable version of the data, because the data are reported as differences for each observation from 68.25 inches (presumably the average height). The numbers in the body of Galton's graph represent the number of individuals in that particular "cell," so the format is a two-dimensional frequency distribution, with lines added to show the orientation of axes. This may be an early effort, but shows sophisticated graphical representation.

1. The independent variable X is measured without error (again we are using "error" here in the statistical sense of an estimate of variation, not in the sense of a mistake]. In this sense, the X values are fixed by the researcher (as in the case of Model I ANOVA), but the Y values are free to vary randomly.
2. The linear equation $\mu_Y = \alpha + \beta X$ describes the expected mean value of Y for a given X.
3. For a given value X_i, the corresponding values of Y are distributed independently and normally, so that $Y_i = \alpha + \beta X_i + \varepsilon_i$. The error terms ε_i are assumed to be distributed normally with a mean of zero. There may be more than one value of Y for given values of X.

Uses of Regression

Definition of the Empirical Relationship of Y and X

[I]n Sicily, thigh bones and shoulder hones have been found of so immense a size, that from thence of necessity by the certain rules of [regression], we conclude that the men to whom they belonged were giants, as big as huge steeples.

Miguel de Cervantes, The History of Don Quixote de la Mancha

The most common use of regression is to decide if indeed there is a significant empirical relationship between dependent and independent variables, and to define the relationship quantitatively. We may be interested in ascertaining whether, given the scatter of the data, fish yields significantly increase as temperature increases, and if so, what are the slope, intercept, and variation associated with the relationship. The regression establishes the empirical relationship, even if we have no knowledge of exactly how temperature of seawater leads to larger fish yields.

We can also use regression to quantify a relationship that has a causal origin. If we experimentally manipulated the independent variable, we can justifiably add the idea of causality to interpretation of the regression between X and Y. We have already discussed the idea of causal relationships above; the regression merely allows us to define the quantitative nature of the relationship.

Estimation of Y from X

If we have an equation that relates Y and X, an obvious use is to make predictions about unknown values of Y from the equation and known values of X. We might have data on seawater temperature and fish harvest from the same areas, and it might be of interest to calculate fish yield for any given seawater temperature. This is readily done by use of the linear regression equation fitted to the data.

Comparison of Regressions

Regression can also be used to ascertain whether the relationship between Y and X is the same in bivariate samples taken from more than one population. For example, we might be interested in testing whether the relationship of feldspar to quartz content in samples of igneous rocks taken from the northeast of Brazil is similar from that in samples collected near the Gulf of Guinea in Africa.

Analysis of Covariance

An additional use of regression that merits mention is that of analysis of covariance (ANCOVA). If we have data from several groups, say, nitrogen (N) content of leaves of different species of grasses, and we plot N content of soil as the X variate, and N content of leaves of grasses as the Y variate, we might find that the N content in each species depends on the soil N content. We are also likely to find that although grasses as a group respond in similar fashion (the slopes of the regressions are similar) to soil nitrogen, the regression lines are offset, that is to say, the intercepts (the α in the regression equation) along the Y axis differ. ANCOVA is designed for just such cases; it examines the regressions of each grass species, assuming that they are similar and so can be pooled, then uses the pooled regression to "correct" for the effect of the X variate (soil N in our example) and applies F tests to determine whether the intercepts on the Y axis differ.

Analysis of covariance is probably the most restrictive of the analyses we have discussed. Use of ANCOVA in tests of hypotheses requires meeting all the assumptions of ANOVA and of regression. and assumes that the regressions used to eliminate the effect of the covariate are similar.

Significance Tests in Regression

Establishing the significance of regressions is done by means of tests of significance much like the ones used in ANOVA (table 3.9). If there is a change in the X variable, $X_1 - \bar{X}$, there will be a concomitant change in Y (fig. 3.2). Part of the change in Y, $\hat{Y}_1 - \bar{Y}$, is due to the regression relationship.[3] The remainder, $Y - \hat{Y}_1$, can be thought of as the residual variation attributable to random effects of many unidentified variables or chance.

To do tests of significance in regressions, we therefore partition the overall variation in the data set into a component that measures the effect of the regression (the effect of variation of the independent variable) on the dependent variable. We also estimate the remaining variation (the departure in position of individual points away from the regression line) and treat that term as an estimate of the error due to random variation.

Table 3.9. Sources of Variation, Sum of Squares, and Mean Squares That Estimate the Model in a Regression Analysis.

Source	*df*	SS	M S	M S Estimates
Explained by regression (differences between estimated Y and mean of Y)	1	$\sum (\hat{Y} - \bar{Y})^2$	$s_{\hat{Y}}^2$	$\sigma_{YX}^2 + \beta^2 \sum (X - \bar{X})^2$
Unexplained variation (differences between measured Y and estimated Y)	$n-2$	$\sum (Y - \hat{Y})^2$	s_{YX}^2	σ_{YX}^2
Total (differences between measured Y and mean of Y)	$n-1$	$\sum (Y - \bar{Y})^2$	s_Y^2	

3. Y is the observed value of the dependent variable; $\hat{Y}$ is the estimate of such a value obtained using the regression relationship; $\bar{X}$ and $\bar{Y}$ are the mean estimates of the independent and dependent variables, respectively.

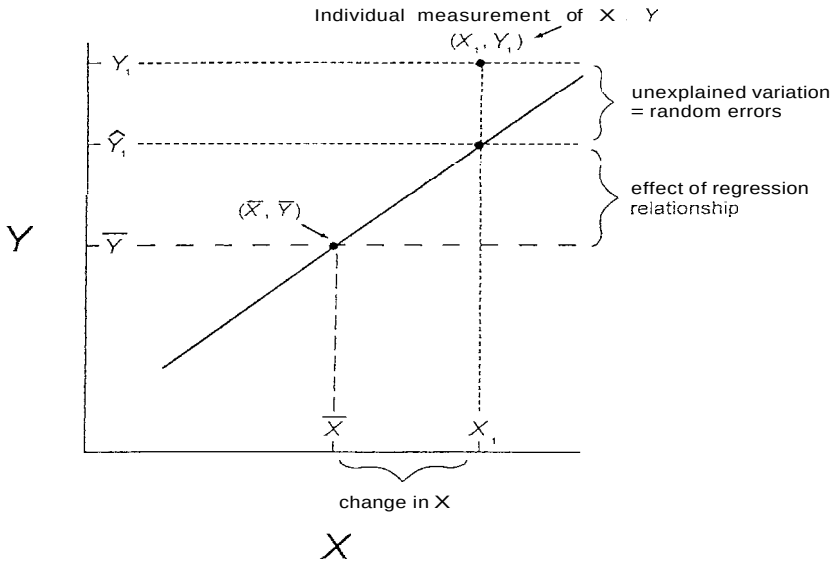


Fig. 3.2 Diagram to illustrate the partition of variation in the dependent variable Y into variation due to the regression relationship, and variation due to unexplained or random variation. X, Y show specific values of variables, X, $\bar{Y}$ show means of all values of X and Y. $\hat{Y}$ shows estimate of mean of Y.

We then compare the significance of the regression term by comparisons using an F ratio, as in the case of ANOVA.

In the case of table 3.9 the F test is the quotient of the regression and residual MS. The regression MS is based on one degree of freedom. The total MS has $(n-1)$ df, so only $(n-2)$ are left for the residual MS.

The *coefficient of determination* (r^2) is a useful additional statistic that can be obtained from regression tables such as table 3.9. Values of r^2 are obtained by estimates of the total change in Y created by the change in X, carried out in the calculations of table 3.9. If we further divide $s^2_{\hat{Y}}$ by s^2_Y, and multiply by 100, we estimate r^2, which is the percentage of the variation in Y that is explained by variation in X. The r^2 is used rather frequently and too freely (Prairie *1996)*. We will discuss its properties, utility, and drawbacks after correlations are introduced in the following section.

We have dealt with Model I regression, in xvhich the Xs are fixed. Model II regression applies to circumstances in which both variables are subject to error. Model II regression is a more complicated subject, with several different cases, whose properties are still not well understood, and in which tests of significance are less straightforward than those of table 3.9. Model II regressions require somewhat different calculations and tests. One way to do unbiased Model II calculations is to use the geometric mean approach (see Sokal and Rohlf *1995*, chap. 14, which reviews several different Model II cases and provides the formulas needed). Clear discussions of applications of Model II regression in marine biology and fisheries sciences are provided by Laws and Archie (1981) and Ricker *(1973)*. When scatter around regression lines is relatively large, use of Model I and Model II calculations yields different results, so with such data it is more important to apply the most appropriate model. Distinc-

tion of the two models is less important in cases in which the scatter of the data around the regression line is relatively modest, because there the two models lead to similar results.

Of course, not all relationships are linear, nor are we interested only in two-variable relationships. For such applications (multiple and curvilinear regression), consult Sokal and Rohlf (1995, chap. 16). These topics are also treated well by Draper and Smith (1981), who provide a clear account of methods, but demand understanding of matrix algebra. Fortunately, the complicated calculations for nonlinear regression are done for us by most software packages, so we need not be deterred from their use.

If transformations fail to make data meet the requisite assumptions for regression analyses, we can apply nonparametric methods. These tests ascertain only whether the Y increases or decreases as X changes. *Kendall's* rank correlation is one option for a nonparametric alternative to regression.

Regression Analyses with Multiple Variables

In general, it seems reasonable to think that more than one independent variable may affect values of a dependent variable. Often we can measure responses of a dependent variable to the influence of several independent variables, and subject the data to examination by methods such as multiple regression or the related *path* analyses, techniques that are well described in Sokal and Rohlf (1995). These methods are not a panacea. First, the analyses require all the assumptions of regression analysis. Second, if there are correlations among the independent variables whose effects are to be evaluated (a phenomenon called collinearity), it is not feasible to unambiguously estimate the effects of each variable. Methods to test whether there are collinearities among variables thought to be independent are given by Myers (1990).

The inappropriate use of multiple-variable analyses is common. For example, Petraitis et al. (1996) found moderate to serious collinearity in 65% of examples of use of path analysis in evolutionary biology. Moreover, these analyses should not be interpreted as showing causality, but co-relationships (see section 3.3). Results coming from these sorts of analyses are, in the terms of chapter 1, more characteristic of the initial descriptive phase of scientific work, creating interesting observations whose causes need study by manipulative methods.

3.3 Correlation

Correlation is a measure of the degree to which two variables vary together; this is noi the same as regression, which expresses one variable as a function of the other. Correlation and regression are related in that both treat relationships between two variables and in that the formulas used in calculations are similar. It is therefore not surprising that they are often confused. Table 3.10 summarizes the applications of regression and correlation.

Table 3.10. Situations Where Regression and Correlation Are Applicable.

	Nature of the Two Variables	
Purpose	Y Random, X Fixed	Y_1, Y_2 Both Random
Describe relationship of one variable to another, or predict one from the other	Model I regression	Model II regression[a]
Establish relationship between variables	Meaningless,[b] but can use r^2 as estimate of % of variation in Y associated with variation in X	Correlation coefficient r

Adapted from Sokal and Rohlf (1995)

[a]Model I is generally inappropriate, except in the common Berkson case, where values of X are subject to error, but the levels of X are controlled by the experimenter. Since it is unlikely that the errors introduced by the experimenter and the random errors are correlated, Model I applies.
[b]Meaningless because correlation is not definable if we fix one of the two variables.

If we wish to establish and estimate the dependence of Y on X, or describe the relationship of Y and X, we can use Model I regression if Y is random and X is fixed. If the two variables are random (let us call them now Y_1 and Y_2, since they are random, and we have been using Y for random variables), we can use Model II regression, except in a few cases listed by Sokal and Rohlf (1995, chap. 15). If both Y_1 and Y_2 are random (and normally distributed), we can calculate the *correlation coefficient*, r, and carry out significance tests.

If we wish to establish the association or interdependence between the two variables, Y being random and X fixed, we can stretch the interpretation of r by calculating r^2 as an estimate of the proportion of the variation in Y that is explained by variation in X. This r^2 has been defined in section 3.2, where we called it the *coefficient of determination*. It is possible to calculate r even if we have data best suited to a regression analysis. In these cases the r calculated can be taken only as a numerical value, not as an estimate of the parametric correlation between the two variables.

Correlation coefficients and coefficients of determination are among the most frequently used statistical tools. When we use these statistics, however, we must be aware of certain pitfalls, as noted by Berthouex and Brown (1994) and many others.

First, as discussed in chapter 1, correlations cannot be taken to mean that changes in X *cause* changes in Y. As another example, consider figure 3.3, where the data yield an $r = 0.864$, which can be shown to be significant. Yet the values plotted on the x axis of figure 3.3 are the first six digits of π, versus the first six nonzero Fibonacci numbers on the y axis.[4] There is no reason to think that there is any link — causal or otherwise — between these numbers, and in fact the relationship is not even predictive — the line fitted to the data does not predict the next Fibonacci number (13).

4. Fibonacci numbers are the sequence of numbers formed by adding the two prior numbers (0, 1, 1, 2, 3, 5, 8, 13, 21, . . .), named after the mathematician Leonardo Fibonacci (c. 1170) of Pisa.

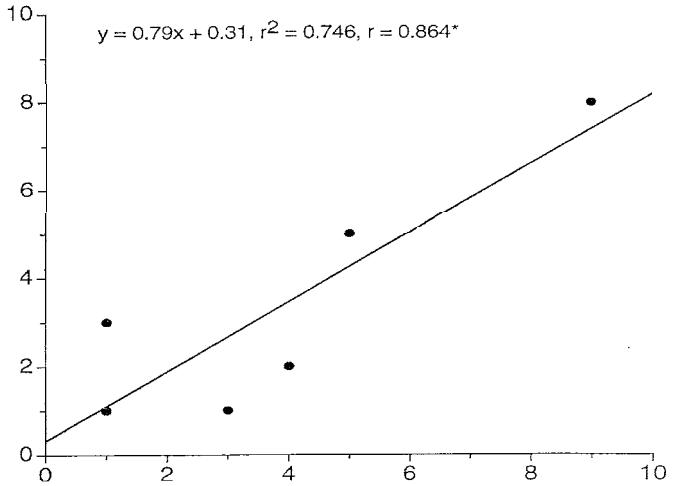


Fig. 3.3 Another example of why care is needed in interpretation of correlations: values of X are the first six digits of π (314159);values of Y are the first six nonzero Fibonacci numbers (112358). Modified from Berthouex and Brown (1994).

Second, data with quite different features might yield the same r. Figure 9.4 shows that remarkably different data sets can yield an r of 0.82. This example reminds us that it is always desirable to plot data graphically before proceeding to statistical analyses.

Third, the likelihood of finding significant r or r^2 increases as the number of observations increases, even if there is no relationship between Y_1 and Y_2. I-Iahn (1973)calculated values of r^2 between unrelated Xs and Ys that would be required to find a level of significance. With just three observations, for instance, r^2 would have to be 0.9938 before it could be declared significant at the 0.05 level. With 100 observations of Y_1 and Y_2, a significant relationship would be declared even with an r^2 of 0.04. We should therefore be wary of correlations or regressions computed from low numbers of observations, because in such circumstances it is hard to show that possibly important differences are significant. We should also be wary of correlations or regressions clone with many, many observations, because in these cases it is too easy to show that minor, perhaps uninteresting, differences are statistically significant (cf.fig. 10.2 bottom). Regardless of the number of observations, the ability to predict Y from X using a regression is limited by the scatter of the points. Regressions with $r^2 < 0.65$ have low predictive power, and should be interpreted accordingly (Prairie 1996).

Fourth, the estimates of r or r^2 depend on the range of values (and number of observations and their spacing) of the Y_2 variable. This is evident in figure 3.4, where different r^2 values result from the use of different subsets of the full data set shown at the top. Data sets in the top and second panels provide a fair assessment of the relationship between Y_1 and Y_2. The narrow range in Y_2 in the third panel makes it impossible to discern the relationship (note nonsignificant r).Although the fourth panel yields a good assessment of the correlation, in the absence of further data a skeptical reader would not be convinced that the linear relationship

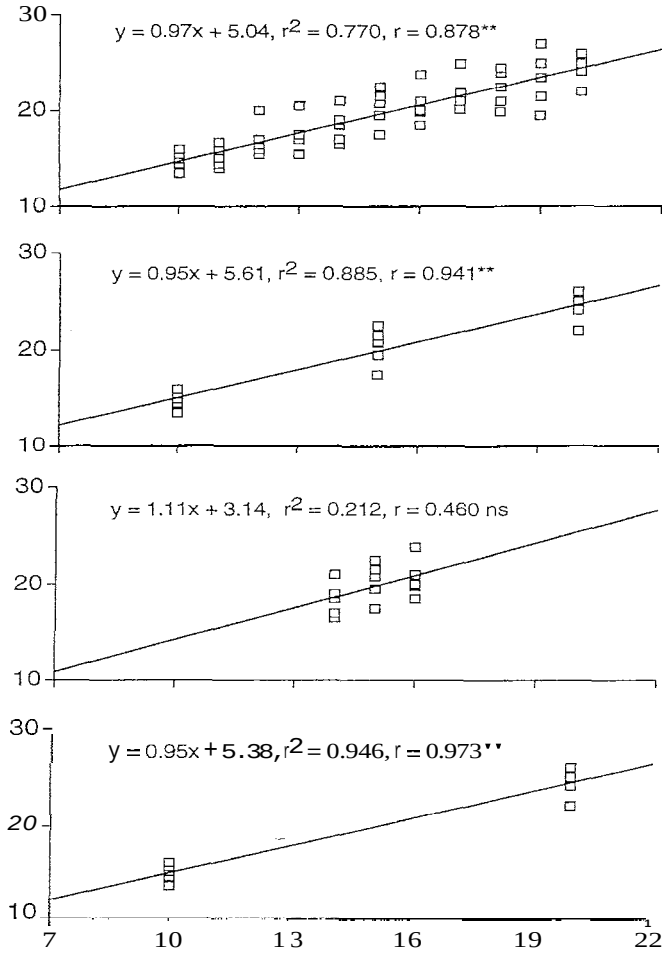


Fig. 3.4 Correlation statistics calculated for a data set (top) and for subsets of the same data (below). Modified from Berthouex and Brown (1994).

exists. The spacing of the observations in the X variable and the range thus need to be carefully planned in designing studies, as discussed in more detail in section 4.2.

Correlation and its relationship to regression, its near cousin, and to other statistics have always been confusing. For some definitions and identifications of the different ways r has been understood, see Rodgers and Nicewander (1988). To clarify some of the issues, table 3.10 summarizes the purposes of regression and correlation, and conditions for the application of these two related kinds of analysis.

The significance values of correlation coefficients are tested by t tests. These significance tests ascertain whether the association between the two variables is greater than expected by chance alone. Recall the discussion in section 1.1 about interpretation of the mechanisms that give rise to correlations.

Correlational statistics have proliferated in many fields in which experimental approaches are not readily available. There are also many ways to extend the correlations to more than two variables, including such methods as principal components and factor analysis. These are usually quite demanding computationally and ambiguous in interpretation. Much more accessible are nonparametric tests, including *Kendall's* or Spear*man's* rank correlation, which measure the magnitude of correlation. The breathtakingly simple *Olmstead* and Tukey's corner test (Sokal and Rohlf 1995, chap. 15) is useful, but discerns only the presence or absence of correlation.

3.4 Analysis of Frequencies

So far we have addressed analyses of continuous measurement data. Recall, however, that there are other kinds of data that are not continuous. Moreover, we can easily convert continuous data into noncontinuous data by binning, discussed above. Noncontinuous data are also often shown or obtained as frequencies. These kinds of data require different methods of analysis. Here we first discuss goodness-of-fit tests in one-sample and multiple-sample situations, then go on to tests of independence.

Goodness-of-Fit Tests

If we collect a set of data that can be expressed as frequencies, we often want to know whether the frequencies in our sample match those expected on the basis of a theory or some previous knowledge. Such situations are common; for example, in genetics, expected frequencies of offspring can be calculated based on accepted rules, and compared to measured frequencies. Most people who have had any training in science at all have been exposed to the chi-square (χ^2) test that has been traditional for such purposes. Sokal and Rohlf (1995, chap. 17) suggest that the χ^2 test be replaced by something called the G test, for theoretical reasons, plus the G test involves easier computation. G statistics are distributed approximately as χ^2 statistics. G tests of goodness of fit to an expected frequency can be readily done for a single data set, to be compared to an expected frequency. G tests are also possible for frequency data that are tabulated in more than one way, for example, in the case where number of young per nest is recorded for n individual parent birds, or where the question refers to the frequency of sex of the young in the nests being studied.

The Kolmogorov–Smirnov test is another nonparametric procedure useful for continuous frequency data. This test is more powerful than the G or χ^2 tests, particularly when dealing with small sample sizes.

Tests of Independence

There are circumstances in which it is more interesting to ask whether two variables or properties interact with each other, rather than to ascertain the exact frequency of occurrence. We have already encountered the

notion of interaction between variables in discussing multiway ANOVA analyses.

One example of a question that addresses interactions with frequency data might be whether moths with light or dark color gain differential protection from predators. An experiment to test the question would involve exposing 100 moths of each color to predators in the field and recording survivors after a certain interval of time. We then count the frequencies of light and dark survivors, and of light and dark moths presumably eaten by predatory birds. If the properties of color and survival do not interact, we would expect that frequencies of the four classes should equal the product of the proportion in each color that were exposed (0.5 in this experiment), multiplied by the proportion of moths eaten in the overall sample.

Such two-way (as well as multiway) data sets are shown as contingency tables. Data of such structure can be evaluated by application of G tests of independence (Sokal and Rohlf 1995, chap. 17), which make use of the proportions of marginal totals (the sums of rows and columns) to calculate departure from expected frequencies. These tests are similar to the χ^2 contingency tests also used for the same purposes. The G tests of independence are applicable to data for which either the marginal totals are not fixed or one property is fixed.

One advantage of the contingency χ^2 or the G tests is that the frequencies are additive. This permits testing of *any* selected specific comparisons among cells, rows, or columns in a contingency table. We could compare, for example, the significance of color only within survivors in our moth experiment. This flexibility provides a way to extract much information from frequency data.

In some selected circumstances, which we will refer to as repeated measures, there may be interest in changes in a property measured in the same individual or set of experimental units. The *McNemar test* and Cochran's Q *test* are two nonparametric statistics available to assess the degree of correlated proportions in such special circumstances.

3.5 Summary of Statistical Analyses

I have mentioned a number of statistical analyses in the preceding sections. Table 3.11 summarizes these analyses and links them to the different types of data discussed in chapter 2. Table 3.11 is by no means exhaustive; instead, it lists representative ways to scrutinize a variety of data and situations that arise commonly in doing science. Sokal and Rohlf (1995) discuss other options.

Regarding statistical tests in general,

- use tests after you are well acquainted with the data (let the data speak first);
- apply tests that are appropriate (test assumptions, note the type of data and the nature of the question);
- subject test results to skeptical scrutiny (graph the data first, know what results of tests mean);
- avoid using tests that you do not understand, or whose assumptions you may not have tested; and

Table 3.11 Type of Data, and Comparisons Provided by Various Statistical Analyses

Type of Question	Nature of Samples or Groups	Type of Data: Measurement (parametric)	Ordinal (nonparametric)	Nominal (nonparametric)
One-sample goodness of fit to randomness		G or χ^2	Kolmogorov–Smirnov	G or χ^2
Difference between two samples or groups	Independent	Unpaired t test	Mann–Whitney U test	G or χ^2
Difference between two samples or groups	Related	Paired t test	Wilcoxon	McNemar
Differences among more than two samples or groups	Independent	One-way ANOVA	Kruskal–Wallis one-way	G or χ^2 goodness of fit
Differences among more than two samples or groups	Related	Two-way ANOVA	Friedman's two-way	Cochran's Q test
Relationship of a variable to another	Y random, X fixed	Model I regression		
Relationship of a variable to another	Y_1 and Y_2 random	Model II regression		
Relationship of a variable to others	Y random, $X_1, \ldots X_n$ fixed	Model 1 multiple regression		
Covariation of two variables	Y_1 and Y_2 random	Correlation	Kendall or Spearman rank correlation	Contingency G or χ^2 test
Covariation among more than two variables	$Y_1, \ldots, Y_n$ random	Multiple correlation, principal axes, factor analysis		

- avoid the temptation to apply a test just because it is available in your software programs.

Often, a well-drawn figure, with measures of variation and a clear visual message (see chapter 9), is a far better way to examine, show, and understand your data than complex calculations done by a software package and presented in a fancy though perhaps indiscernible graphic.

3.6 Transformations of Data

In chapter 2, I mentioned that transformations were convenient ways to recast data so as to convert frequencies of data to normal distributions, a basic assumption of many statistical analyses. In this chapter I have introduced further assumptions associated with ANOVA and, most particularly, with regression analyses.

I should add that despite the space I give to assumptions and transformations, these are issues that are readily resolved and are not usually a problem. Fortunately, it is often the case that one transformation helps solve more than one violation of assumptions of a particular test. In addition, both ANOVA and regression analyses are fairly tolerant of violations

Frequentist vs. Bayesian Statistics

As we end the twentieth century, there are revisionist scientists who prefer to replace the conventional "frequentist" approach to data analyses with a Bayesian approach. Frequentist refers to the approach based on how frequently one would expect to obtain a given result if an experiment were repeated and analyzed many times. Bayesian statistics derive fi·om a theorem formulated in a paper published in 1763 by the Reverend Thomas Bayes, an English amateur mathematician. Bayesian analyses allow the user to start with what is already known, or supposed, and to see how new information changes that prior knowledge, hunches, or beliefs.

Bayesians find the frequentist assumption of a fixed expected mean value for given variables unacceptable; even if such fixed values existed, they argue, such values would not readily defined by random sampling, in view of the pervasive variation that characterizes nature. Frequentists are limited to making statistical claims about "significant differences" based on probability disti·ibutions, variously expressed as "confidence intervals." Everyone admits that such intervals are ambiguous in definition, interpretation, and use. For example, we can consult a statistical table of r values, where a irequentist might Find that, with 50 observations (not an unusually large number of observations), a relationship can be said to be statistically significant at the 0.05 level, even though the correlation coefficient might "account" for only 7% of the variation among the observations. While statistically significant, would such a conclusion be scientifically significant?

Bayesians openly admit that science is subjective and argue that explicitly admitting the use of prior insights — informed hunches — to search for scientific explanations is a more rational approach, rather than make statistical claims based on unattainable objectivity. Frequentists respond that no human endeavor is perfect and that their approach provides a way to reduce possible biases; they fear that use of prior probabilities allows biases to enter the field of science and at worst intimates that science is just another socially constructed belief. To many frequentists, this is an alarming concept, as discussed more extensively in the last chapter of this book.

In any event, the dispute is not just about statistical methods, but about ways of thinking about science, and the arguments will no doubt continue and will likely invigorate how we do science in coming decades. More details on the issue appear in Science (1999) 286:1460–1464, American Statistician (1997) 51:241–274, and in *Ecological Applications* (1996) 6:1034–1123.

of assumptions. Nonetheless, I devote space to these matters because they force scrutiny of data and heighten our awareness of their nature. These are issues that we tend to rush through in our anxiety to get the answer to the question, "Are the differences significant or not?"

I end this chapter with a discussion of derived variables. These are the result of yet another common class of transformation, carried out to express relationships such as rates or percentages, or for removing effects of a second variable by an arithmetical operation.

Logarithmic Transformations

The logarithmic transformation is useful in a variety of ways. We have already seen in chapter 2 how it ensures normality. In regressions (fig. 3.5, top three panels), log transformations linearize relationships.[5] A log

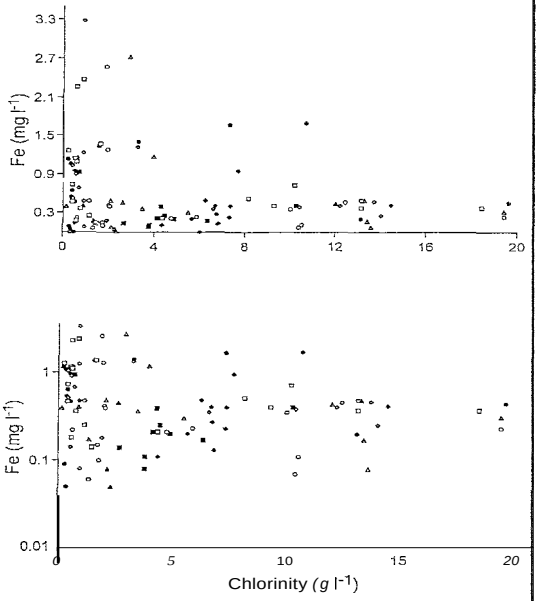
5. I should note here that use of linearized regression can give rise to serious errors in estimates of slopes and intercepts (Motulsky 1095, Berthouex and Brown 1994). Software available for desktop computers allows painless calculation of nonlinear relationships where needed.

Transformations and Graphical Analysis of Data

The preparation of data for statistical analysis is only one reason why we should be aware of data transformations. A second, and probably more general and important reason is that our understanding of the nature of data and data presentation is furthered by knowing how transformations reveal different aspects of data.

Consider, for example, the two graphs in this box. The different symbols may be disregarded for present purposes. A quick glance at the top graph might lead us to conclude that variability and central tendency of concentrations of iron decrease as chlorinity increases. Similarly, cursory examination of the bottom graph might suggest that variability in iron concentration decreases, but that, contrary to what was concluded from the top graph, central tendency remains about the same. In fact, both graphs show exactly the same data; the only difference is that the *Y* axis is in an arithmetical scale in the top graph and as a logarithmic scale in the bottom. There is no sleight of hand intended in either case; it is simply that use of different scales leads us to see different features of the data. In the case of the arithmetic scale, the data display makes us focus on the higher values; the log scale expands the lower value range and lets us see more of the structure of the data there. Both representations are "true"; it is just that choice of scale changes the depiction in set ways.

This example makes clear that (a) routine examination of axis scales (and of units) should precede interpretation of any graph, and (b) transforming data in different ways makes apparent different aspects of the data. Both of these features are eminently useful in practicing science.



Different scales reveal different features of data: scatter plot of concentrations of dissolved iron and chloride in waters of the Ebro Delta Lagoons, collected by my colleague, Francisco Comin. Data are shown plotted in arithmetical (top) and logarithmic *(bottom)* scales.

transformation also assures additivity even if components of variation are multiplicative, for example, $Y_{ij} = \mu\, \alpha_i\, \varepsilon_{ij}$. The log transformation will convert the equation to an additive form that meets the assumption of additivity: $\log Y_{ij} = \log \mu + \log \alpha_i + \log \varepsilon_{ij}$. Log transformations also are useful where, as in the third panel of figure 3.5, variances increase as means increase; in such instances, logarithmic transformations make the variance independent of the mean and improve homogeneity of the variances. Log transformations therefore sustain assumptions of normality, linearity, additivity, and homogeneity, and make a linear regression analysis possible.

Of course, we can transform the Y, the *X*, or both variables before regression analysis [fig. 3.5, top three panels). The choice depends on the nature of the data. Transformation of Y values (fig. 3.5, top) is appropriate where percentage changes in Y vary linearly with changes in X.

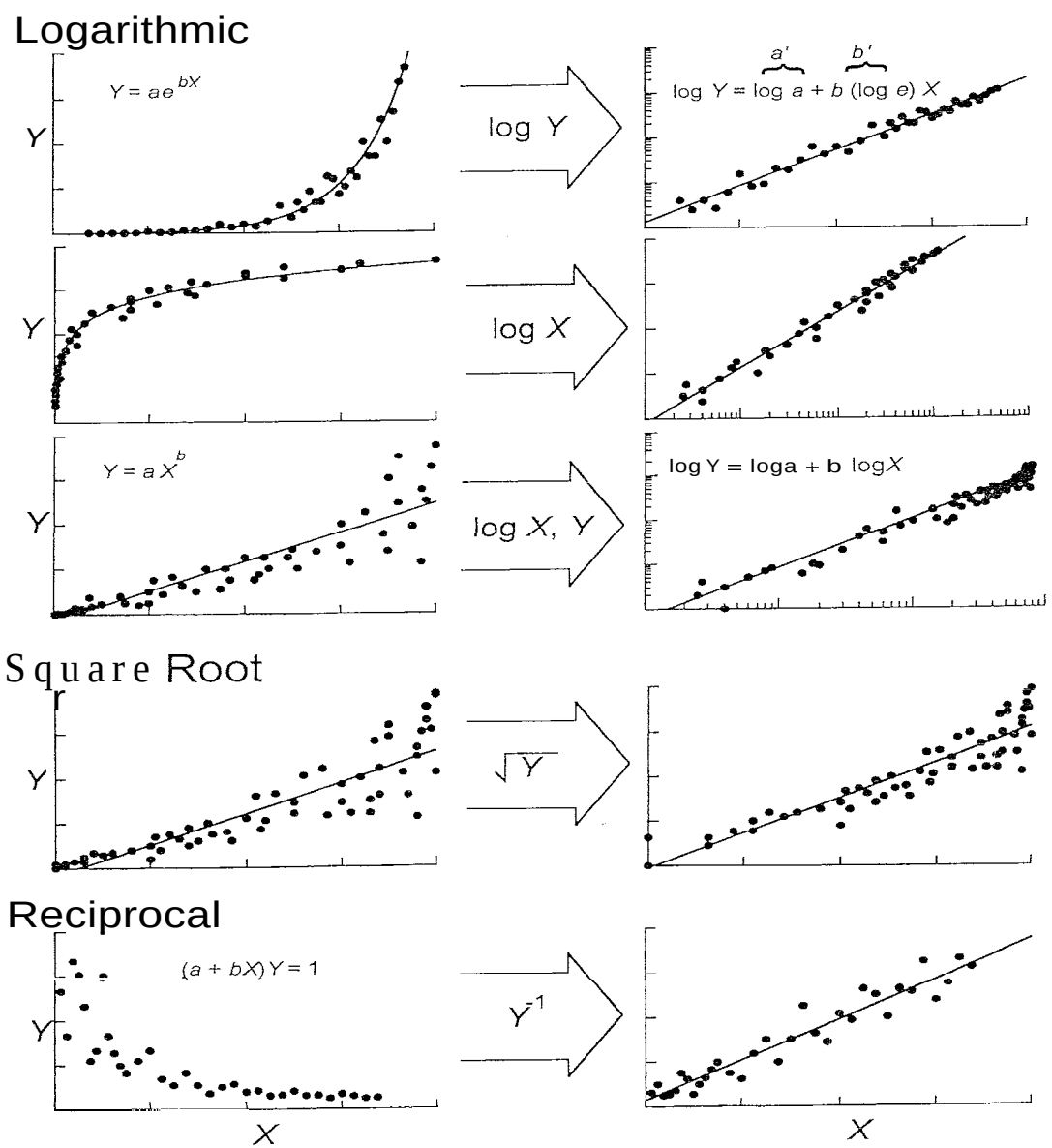


Fig. 3.5 Transformations in regression. For each data type, the arithmetical version is on the left and the transformed variables are on the right. Top three panels: Different forms of logarithmic transformation. *Fourth panel*: Square root transformation. *Bottom* panel: Reciprocal transformation.

Logarithmic transformation of the X values is reasonable where percentage changes in X are related to linear changes in Y (fig. 3.5, second panel). Logarithmic transformations of Y and X are useful in data where there is a much larger increase in one of the variables relative to increases in the other when data are plotted in arithmetical scales (fig. 3.5, third panel).

Scientific data are commonly analyzed after log transformations. This comes from the expectation that variability of data will be proportional to the magnitude of the observations. We want to evaluate differences among means in a way that expresses variation relative to magnitude of the values. Since log transformations do exactly this, they are a natural and convenient scale in which to examine scientific data. Mead (1988) therefore suggests that rather than ask, "When should data be transformed logarithmically?" we should ask, "When is it reasonable to analyze data in other than a logged scale?"

Square Root Transformations

In chapter 2, 1 noted that square root'transformations make count data appear normally distributed and assure independence of mean and variance. Square root transformations of Y values also add linearity, as well as homogenize variances (fig. 3.5, fourth panel), helping meet the assumptions of regression. Note that the square root transformation has an effect similar to, but less powerful than, that of log transformation.

Reciprocal Transformations

Reciprocal transformations are of the form $1/Y$ (fig. 3.5, bottom). This transformation is important to allow regression studies of data such as found in figure 3.5, bottom left. The reciprocal transformation linearizes the relationship of Y to X in data sets that originally have a hyperbolic relationship. One example of a hyperbolic relationship is a dilution series, commonly used in microbiology, in which a fluid containing microorganisms is serially diluted by the transfer of a unit volume from one dilution to the next.

Linearization of data may lead, however, to biased estimates of intercepts, slopes, and r. In the reciprocal transformation, for example, the values at the large x (small $1/x$) end of the range will be squeezed together, and values at the other end of the range will appear to vary greatly. This distortion biases the position of the line of fit. This transformation should therefore be used with caution. Refer to the review by Berthouex and Brown (1994) for further details before using linearization transformations.

Derived Variables

Types of Derived Variables

Scientists use a remarkable variety of variables that are created by arithmetical transformations, such as division of two original variables. Such

manipulations lead to definitions of rates, percentages, and ratios, all of which are core features of doing science. Another common data manipulation is to remove the effect of a second variable (implicitly assuming additivity of effects) by subtracting the effect of the second variable from that of a first variable.

It is not widely appreciated, however, that such data manipulations may give rise to artifacts that need to be kept clearly in mind to prevent confusing artifacts and actual effects. First, consider the 1,000 random values of Y (restricted to numbers between 1100 and 1220) and X (any three-digit number) plotted in figure 3.6 (top left): these are random numbers, so there is no correlation at all among values. If, on the other hand,

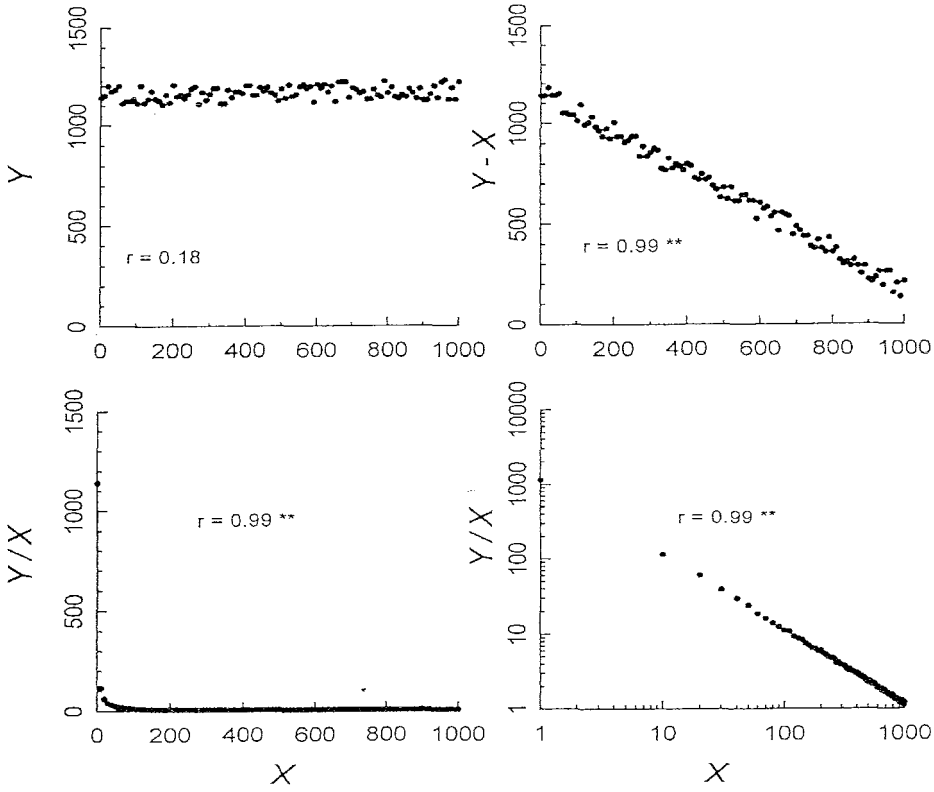


Fig 3 6 Spurious correlations created by use of derived variables Data are series of random numbers fn(X), random numbers between 1100 and 1220 fn(Y) (Kenney 1982) *Top left* Value of Y plotted versus values of X *Top right* Same data, plotted as Y – X on y axis, versus X on axis *Bottom left* Same data, plotted as Y/X versus X, with arithmetical scales on the axes *Bottom right* Values plotted as Y/X versus X, this time with axes bearing log scales Reprinted with permission from Kenney, B. C. 1982. Beware of spurious self-correlations! *Water Res. Bull.* 18:1041–1048, copyright of American Water Resources Association.

we plot Y/X versus X, a marked relationship appears (fig. 3.6, bottom left), merely because of the presence of X in both axes. Quite often we show such data manipulations in log scales (bottom right), which enhance the artifacts. Similarly, $(Y - X)$, a derived variable that is often used, when plotted versus X (top right) shows a "relationship." The degree of spurious correlation increases as the variation of the common term (X in our examples) increases, relative to variation in Y. Correlations of derived variables with common terms are best avoided; if it is essential to use such variables, Atchley et al. (1976) and Kenney (1982) suggest procedures to see if spurious relationships are a problem.

Error Propagation Techniques

It is often necessary to make comparisons among derived variables, but we are likely to have estimates of variation only for the original variables. To estimate variation that is associated with the derived variables, there are two approaches available: errorpropagation *techniques*, and the newer resampling *methods*.

To calculate the error of a derived variable, we weight the contribution of each component of the derived variable to variation of the derived variable. Note that this can apply to a simple ratio, to a difference, or to a complex equation (called a model in chapter 1) with different components. The essential assumption needed is that the terms of the derived variable are independent, because if the terms are correlated, their contribution to variation of the derived variable is undefined. [Formulas to calculate propagated errors for different arithmetical operations are given in the accompanying box.)

Formulas for Calculating Propagated Errors in Different Arithmetical Operations

$$\left.\begin{array}{l} z = x_y \\ z = x/y \end{array}\right\} \quad \frac{s_z}{\bar{Z}} = \sqrt{\left(\frac{s_x}{\bar{X}}\right)^2 + \left(\frac{s_y}{\bar{Y}}\right)^2}$$

$$z = x \pm y \quad \} \quad s_z = \sqrt{s_x^2 + s_y^2}$$

$$z = x^m y^n \quad \} \quad \frac{s_z}{\bar{Z}} = \sqrt{m^2\left(\frac{s_x}{\bar{X}}\right)^2 + m^2\left(\frac{s_y}{\bar{Y}}\right)^2}$$

$$z = kx \quad \} \quad s_z = k s_x$$

These equations define how we might calculate propagated standard deviations (s_z) in cases where the terms that contribute to the propagated standard deviation are multiplied, divided, summed or subtracted, raised to powers, or subject to a constant multiplier. In all cases, the z refers to the propagated term derived from the independent terms, x and y (modifiedfrom Meyer 1975).

A newer way to assess the variation associated with a derived variable is to make use of resampling methods. One such procedure makes use of what is called the bootstrap technique (Diaconis and Efron 1983, Manly 1991). This method assumes that the frequency distribution of the population is closely approximated by the frequency distribution of a sample. Using this supposition, the sample frequency distribution of the variable (or for derived variables, the result of the operation being studied) is itself repeatedly resampled n times, by randomly selecting subsets of the sampled values. Then the bootstrap mean is calculated from the repeated samplings. This procedure is repeated many times, until the mean of the derived variable does not change with further repetition. The measures of variation, such as the bootstrap standard deviation, can be calculated from the sets of subsamples.

SOURCES AND FURTHER READING

Atchley, W. R., C. T. Gaskins, and D. Anderson. 1976. Statistical properties of ratios. 1. Empirical results. Syst. Zool. 25:137–148.

Berthouex, P. M., and L. C. Brown. 1094. Statistics *for Environmental* Engineers. Lewis.

Diaconis, P., and B. Efron. 1983. Computer-intensive methods in statistics. Sci. Am. 248:116–130.

Draper, N. R., and H. Smith. 1981. Applied Regression Analysis, 2nd ed. Wiley.

Hahn, G. J., 1973. The coefficient of determination exposed! *Chemtech* October, 609–611.

Kenney, B. C. 1982. Beware of spurious self-correlations! Water Res. Bull. 18:1041–1048.

Krumbein, W. C. 1955. Experimental design in the earth sciences. Trans. Am. Geophys. Union 36:1–11.

Laws, E. A., and J. W. Archie. 1981. Appropriate use of regression analysis in marine biology. Mar. Biol. 65:13–16.

Manly, B. F. J. 1991. The Design and Analysis of Research Studies. Cambridge University Press.

Mead, R. 1988. The Design of Experiments. Cambridge University Press.

Meyer, S. L. 1975. Data Analysis *for* Scientists and Engineers. Wiley.

Motulsky, H. 1995. Intuitive Statistics. Oxford University Press.

Myers, R. H. 1990. Clussical and Modern Regression with Applications, 2nd ed. P. W. S. Kent.

O'Neill, R., and G. P. Wetherill. 1971. The present state of multiple comparison methods. J. Statist. *Soc.* B 33:218–241.

Petraitis, P. S., A. E. Dunham, and P. H. Niewianowski. 1996. Inferring multiple causality: The limitations of path analysis. *Funct.* Ecol. 10:421–431.

Prairie, Y. T. 1996. Evaluating the power of regression models. Can. J. Fish. Aquat. *Sci.* 53:490–492.

Ricker, W. E. 1973. Linear regressions in fishery research. *J.* Fish. Res. Board Can. 30:409–434.

Rodgers, J. L., and W. A. Nicewander. 1988. Thirteen ways to look at the correlation coefficient. Am. Stat. 42:59–66.

Sokal, R. R., and F. J. Rohlf. 1995. *Biometry*, 3rd ed. Freeman.
Tukey, J. W. 1977. *Exploratory Data Analysis*. Addison-Wesley.
Underwood, A. J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. *Oceanogr. Mar. Biol.* 19:513–605.