2 Statistics of Immunological Testing

Albert D. Donnenberg

CONTENTS

2.1 OVERVIEW

Lord Kelvin* famously said, "Until you can measure something and express it in numbers, you have only the beginning of understanding." With few exceptions, interpreting immunological assays are all about interpreting numbers, and this can be aided (or not) by the application of statistical methods. This chapter is intended to help students of immunology, who have already had some exposure to statistics, navigate the most common statistical methods that are applied to tests of immunity. To quote the humorist Evan Esar, a statistician is one "who believes figures don't lie, but admits than under analysis some of them won't stand up either." By understanding a few basic statistical methods and avoiding the most common pitfalls of data analysis, the student of immunology increases the odds that his or her figures will, in fact, stand up.

Jim Deacon of the University of Edinburgh cites an unnamed Cambridge professor's take on statistics in Biology:

> In our Department we have a long corridor with a notice board at one end. I draw a histogram of my results, pin it to the notice board, then walk to the other end of the corridor. If I can still see a difference between the treatments then it's significant.

The don was advocating a rudimentary form of something that the great statistician and teacher John Tukey called exploratory data analysis (EDA) [1]. In our quest to understand our data, EDA is only a beginning, a method to discern trends and sniff out possible sources of variability *without making prior assumptions about the data*. Tukey drew a distinction between exploratory and confirmatory data analysis. The latter draws on a rich body of statistical methodology, which in the words of Tukey "[are] principles and procedures to look at a sample, and at what the sample has told us about the population from which it came, and assess the precision with which our inference from sample to population is made." Tukey advocated that exploratory and confirmatory data analysis proceed side by side.

2.2 EXPLORATORY DATA ANALYSIS IN THE COMPUTER AGE

The desktop computer did not exist when John Tukey wrote his seminal tract on EDA. Most of the methods that he invented or promoted were designed to be performed with a pencil, a ruler, and graph paper. Although this must have been somewhat cumbersome, EDA techniques are now implemented in most statistical packages. SYSTAT (Systat Software, Inc., San Jose, California), which I have used for the graphics and analyses in this chapter, has a powerful suite of EDA and graphic tools.

2.2.1 The Box Plot

Figure 2.1 is just what the don ordered. It is a dot plot of reciprocal antibody titers in nine healthy subjects before and 1 week after booster immunization with tetanus toxoid.

* William Thomson (1824–1907), First Baron Kelvin, who among other notable achievements gave us the concept of absolute zero, was fond of such pronouncements. He also said, "In science there is only physics, all the rest is only stamp collecting"; and "Large increases in cost with questionable increases in performance can only be tolerated in race horses and women."

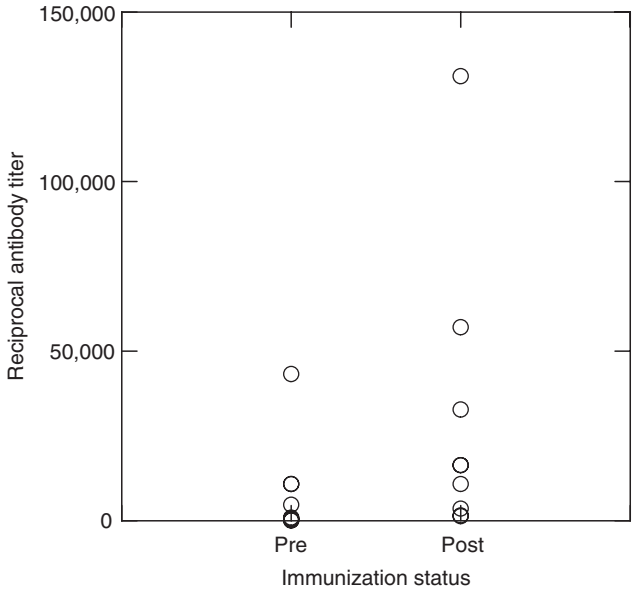

FIGURE 2.1 Dot plot of reciprocal antibody titers in nine subjects before and 1 week after immunization with tetanus toxoid vaccine.

It does not pass the don's test when I step to the back of my less -than -commodious office and look at the data on my computer screen. What I see is a lot of overlapping dots and a single dot in the *pre*group and several dots in the *post*group that are larger than the rest of the values. Did the immunization work? Perhaps not. Figure 2.2 shows the same data displayed using one of the most powerful tools of EDA, the box plot. Box plots display the sample range, median, and the spread of the central 50% of the data surrounding the median. Tukey saw the box as a *hinge* turned on its side. The outer bounds of the hinge give the location of the central 50% of the data (known as the hinge spread); where the two leaves of the hinge join (the *waist*) is the sample median (visible in the *post*group). The *whiskers* (another Tukey coinage) show the range (location of the smallest and largest values), exclusive of outliers, which are indicated by circles (explained in the figure legend). What one sees in the box plot is quite different from the raw dot plot. The median value of the *post*group is higher than that of the *pre*-group, which on this scale appears to be near the origin. We also notice that the data is not symmetrically distributed, but is quite skewed toward high values. In the *pre*group, the first and second quartile are smashed against the origin, and the *post*group has a short whisker below the second quartile, and a long whisker above the third. Both the *pre*- and *post*groups have one high outlier. I might be changing my opinion about whether the immunization worked, or at least formulating some confirmatory tests that might be used to shed light on the question.

2.2.2 Transformations in Exploratory Data Analysis

Tukey advocated transformation (the most common being logarithmic and square root) as the next logical step in EDA when one has *skewed* data. Figure 2.3 shows the

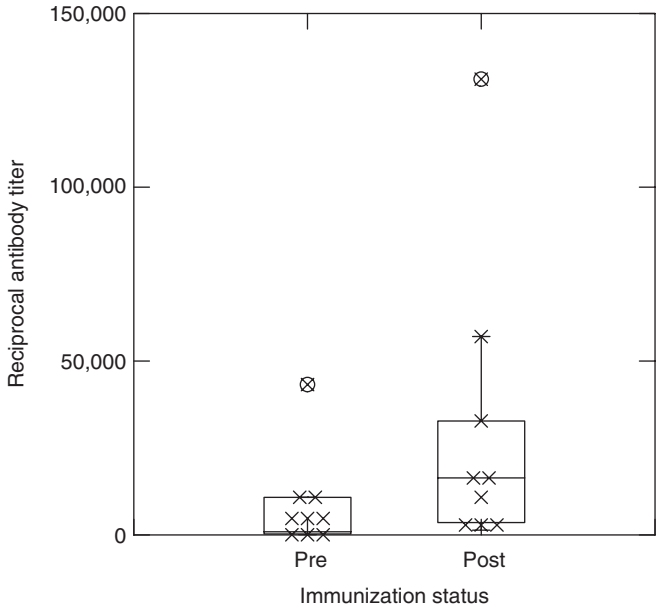

FIGURE 2.2 Classic Tukey box plots display nonparametric (distribution-free) descriptive statistics. Individual data points are marked by Xs. The waist indicates the group median and the hinges (upper and lower boundaries of the box) indicate interquartile distances. The whiskers (bars) give the ranges, exclusive of outliers. Outliers (more than 1.5 times the hinge spread from the median) are shown by asterisks (there are none in this illustration) and far outliers (more than three times the hinge spread from the median) are shown by circled Xs. The lower hinge cannot be seen in the pre group, because more than five of the nine data points lie close to the origin.

data of Figure 2.2, log transformed to the base 10. Now we can clearly see what was not apparent at all in Figure 2.1, and was only hinted in Figure 2.2. From the location of the median bars, we see that immunization led to a more than 10-fold increase in median antibody titer ($<10^3$ versus $>10^4$). Further, the data in both groups are more or less symmetrically distributed, one of the key requirements for the legitimate use of a variety of confirmatory statistical tests (explained in Section 2.3). The two far outlying values seen in Figure 2.2, now fit neatly into the distribution. AQ1

2.3 ESTIMATES OF DATA LOCATION AND DISPERSION

2.3.1 Data Location

The median (the midpoint of a series of ordered numbers, where half the values are above and half below) is the first data-location summary statistic to be evaluated in EDA because it does not require any assumptions about the distribution of the data. Additional summary statistics of data location are the arithmetic mean (the numerical average) and the mode (the most prevalent value within a distribution). Although the arithmetic mean is the most widely used statistic, we will see that

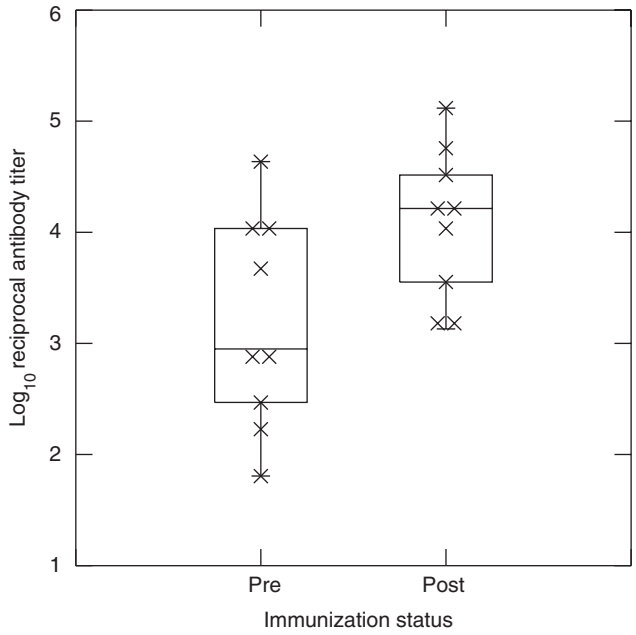

FIGURE 2.3 Box plot of log-transformed antibody data. The base 10 logarithm of the reciprocal titers shown in Figures 2.1 and 2.2 are plotted. Note that the distributions are fairly symmetrical about the waists (medians), and the two values that appeared as outliers in Figure 2.2 are now within 1.5 hinge spreads of the median.

this is not always to the best effect. When data are approximately symmetrically distributed, the median, mean, and mode are very close. As we saw in our EDA example of antibody titers, biological data, and especially immunological testing data, are often highly skewed. To continue with antibodies as an example, titers in a population may consist of individuals with no detectable antibodies, those with low titers, and a very small proportion with extremely high titers. If one were to accept the mode as a summary statistic to characterize the human immunodeficiency virus Enzyme-Linked ImmunoSorbent Assay (HIV ELISA) titers of university students, one would entirely discount the infrequent HIV positive subjects, because the most common result would be zero (or undetectable). The median would also be zero because more than half of the students are seronegative. Likewise, the arithmetic mean would barely register above zero, influenced as it is by the high proportion of subjects without detectable antibody. Although all these summary statistics are technically correct, they may not capture the biology that we seek to explain. Even if one confined the analysis to "positive" subjects, as is the case with tetanus toxoid because all individuals have been immunized some time in the past (Figure 2.4), the left-skewed distribution would result in mean and median values that differ markedly (the mean is higher than the median because the data distribution contains a tail of very high values). Thus, as in our EDA example, the choice of the most informative statistic to summarize data location is dependent on the data distribution. This will be discussed more below.

AQ2

AQ3

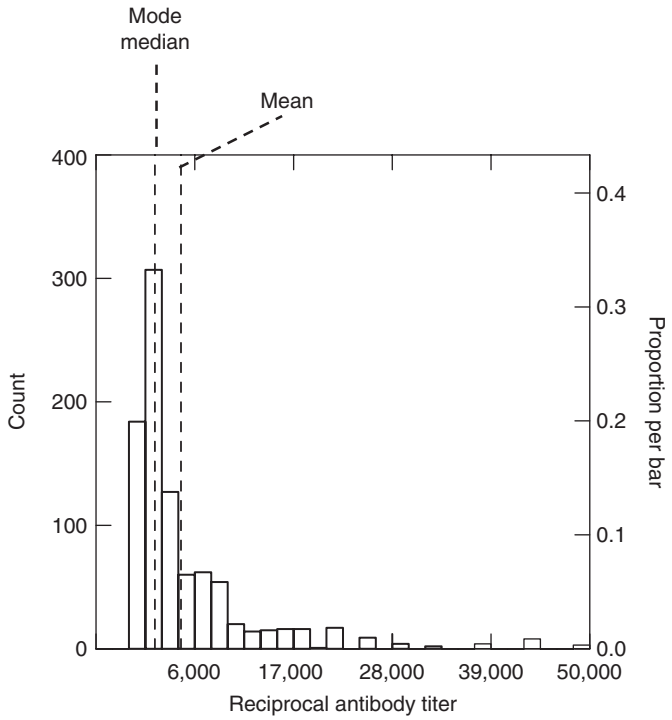

FIGURE 2.4 Distribution of antitetanus antibodies in 1000 subjects. Antibody titer was measured in serial dilutions of serum by ELISA and is expressed as a reciprocal titer (i.e., a value of 32 means that the serum the dilution curve crossed the threshold defining a positive result at a 1/32 dilution). The distributions of values is shown as a bar density histogram. Like many measures of immunity, antibody titer is not symmetrically distributed, but rather is left-skewed. Here, the median and modal values are seen to provide a better summary of the data, the mean being influenced by a few, very high values.

Technical Definition

Sample Standard Deviation. Experimental data sets are collected to make generalizations about the population from which test data are drawn. The standard deviation is a measure of average absolute difference in individual measurements (x_i), compared to the sample average (x bar).

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

The standard deviation tells us about the relative homogeneity or spread of a sample. It can be used, for example, to compute sample percentiles, or it can be normalized by dividing it by the sample mean to create a scale-independent measure of variability, the coefficient of variation.

Technical Definition

Normal Distribution. The normal distribution forms the basis of many parametric statistics, such as Student's *t* test. It is specified by the equation

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$$

This equation has only two parameters, the population mean, μ, and the population standard deviation, σ. The entire family of normal curves is generated by varying μ and σ. If you plug $\mu = 0$, $\sigma = 1$, and x values of -3 through 3 into this equation, you will get the standard normal curve (which we will see in a discussion of limits of detection) spanning three standard deviations above and below a mean of 0. This curve forms the basis of all standard statistical tables.

2.3.2 Data Dispersion

Although it is easy to find graphs in the scientific literature in which group means are compared without recourse to error bars, statistics of location alone do not provide an adequate summary of a set of observations, and certainly do not provide all the information required to compare results from different groups. In addition to location, we need an indication of the spread of individual observations within the group. In EDA, we chose the hinge spread (distance between the beginning of the second quartile and the end of the third) as a statistic that, along with the location of the median, provided useful information about the shape of the data distribution. In confirmatory data analysis, the sample standard deviation (STDEV) (see sidebar) is the most widely used and forms the basis for the many statistical tests based on the assumption of a normal or Gaussian distribution. In general, statistical tests, which assume that the data fit a particular distribution are called parametric tests.

Before we are ready to use parametric statistical tests (e.g., Student's *t* test*), there is something you must know about all tests relying on the normal distribution:

> The normal model uses only two parameters (mean and standard deviation) to describe your entire data set.

As discussed, these are measures of location and dispersion, respectively. The validity of tests that assume normality depends entirely on how well the data can be described by these two parameters. Figure 2.4 shows more antibody data, this time from a group of more than 1000 subjects. As we have begun to appreciate from Figures 2.1–2.3, antibody distributions are usually left-skewed, and the data in Figure 2.4 are no exception. In particular, notice how the high outlying values

* Student was the pseudonym of William Sealy Gosset, an industrial chemist for the Guinness Brewery in Dublin, Ireland. After discussions with the statistical pioneer Karl Pearson, he developed the *t* test to monitor the quality of batches of beer. Although he published the test in Pearson's statistical journal *Biometrika* [2], he was required by his employer to use a pseudonym, because they considered their monitoring process a trade secret.

cause the mean to be higher than the median or mode. Since the normal distribution is a family of bell-shaped curves that can be described by the mean and STDEV, we can create such a curve by plugging our sample mean and STDEV into an equation (see sidebar). Figure 2.5 shows what happens when we do this for the data in Figure 2.4. By definition, the resulting bell-shaped curve has its peak located at the sample mean, and its points of inflection at one STDEV below and above the mean. At this point, we have to question whether we want to entrust our data analysis to these two statistics. Considering the skewed nature of the distribution and the absence of zero or negative values, a logarithmic transformation looks promising. *In practice, a great deal of quantitative, continuous data from immunologic testing is lognormally distributed, so this is the first transformation that one would try.* Figure 2.6 shows the resulting distribution of the log-transformed data (bars) and the predicted normal distribution (curve). What an improvement—the mean, median, and mode are virtually identical (3.26, 3.31, and 3.30, respectively). The data themselves appear to be symmetrically distributed about the mean, and even the extreme values are fairly well modeled. However, this is real-world data and there is also a degree of positive Kurtosis (sharper peak with fatter tails, compared to the normal distribution). This is not a disease, but rather a way to quantify (along with skewness) a departure from the normal distribution. Practically, experience tells that with antibody data, this is just about as good as it gets, and the assumption of normality

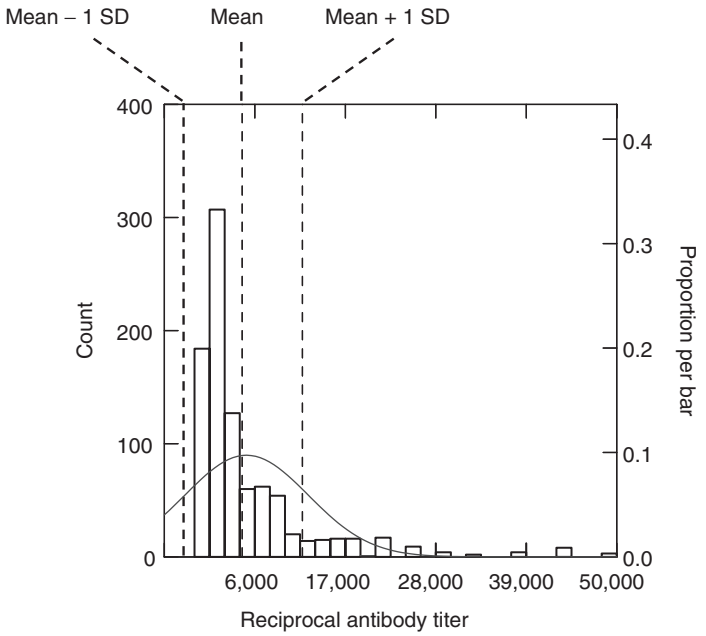

FIGURE 2.5 Application of the normal model to antibody data. The bars show the actual distribution of the data. The bell-shaped curve shows the normal distribution calculated from the sample mean and STDEV. One can see that the actual peak of the distribution is of a much lower value than the peak of the curve. It is also disturbing that the lower end of the distribution extrapolates to negative titers, and the upper end does not account for the highest values.

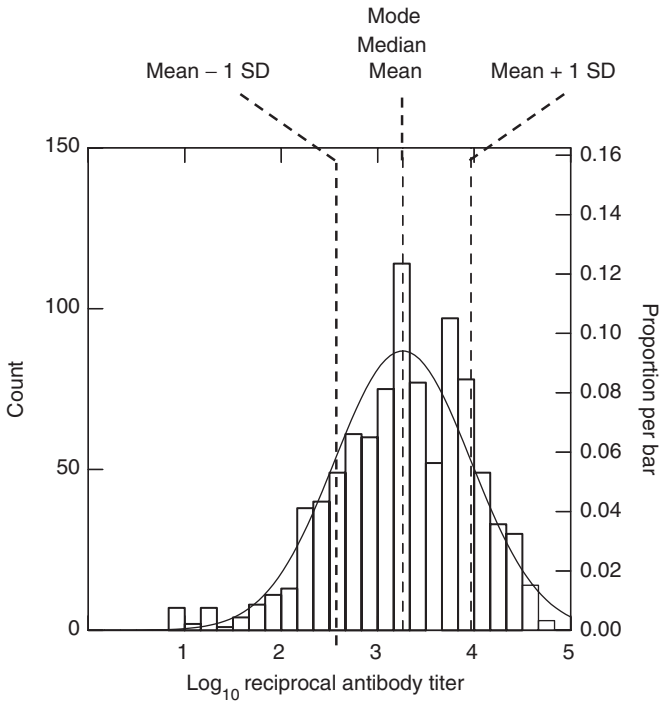

FIGURE 2.6 Distribution of log-transformed reciprocal titers. The mean, median, and mode are virtually identical and the superimposed normal distribution fits the data reasonably well. Although the data are in $\log_{10}$ units, you can get back to the original reciprocal titers by raising the numbers on the *x*-axis to the 10th power. Alternatively, one could simply plot the *x*-axis on a log scale and achieve the same effect.

is a reasonable one. We could then proceed to compare log-transformed titers. Going back to the data in Figures 2.1–2.3, which are naturally paired values (each subject was measured before and after immunization), Student's paired *t* test is the most appropriate confirmatory test (Figure 2.7). This test looks at the difference between the pre- and postimmunization values of the individual subjects. In our example, the average subject had a 7.9-fold rise in antibody titer following immunization (Student's two-tailed test, $p = 0.014$). This is as good a time as any for a discussion of one- and two-tailed *p* values. A *p* value can be put into words as the probability that a difference this great (i.e., 7.9-fold) would occur by chance alone (given the magnitude of the STDEV between individual differences and the number of paired observations). In the present case, a chance difference of 7.9-fold could be expected only 1 in 71 times ($p = 0.014$), well below the conventional probability traditionally considered statistically significant (1 in 20, or $p = 0.05$). Because this is a two-tailed value, we are considering the probability of a 7.9-fold increase *or* decrease in antibody titer. Some would favor a one-tailed *t* test (which would yield exactly half the probability, $p = 0.007$) arguing that we have a prior hypothesis that the only two expected outcomes of immunization are that either the antibody titer remains the same or AQ4 increases. I find such arguments to be an annoying example of *p*-value inflation. AQ5

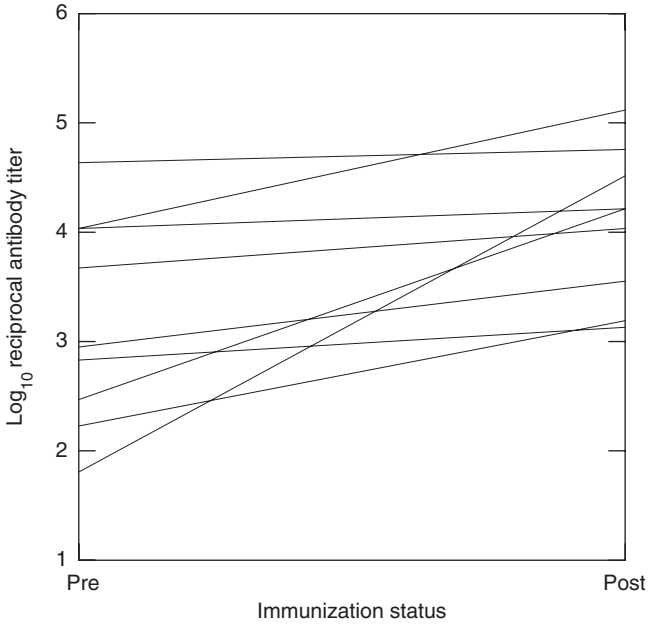

FIGURE 2.7 A line graph showing paired pre- and postimmunization log-transformed reciprocal titers. Student's paired *t* test calculates the difference between pre- and postvalues for each individual data pair. The mean difference is then tested to determine whether it is significantly different from zero (the famous null hypothesis). In this example, the mean difference was 0.896 ± 0.303 (SEM). Because these log units are not terribly intuitive, it is best to transform them back into something more intelligible. Taking the base 10 antilog of the mean difference (i.e., converting the number from the log domain, back into the linear domain), we get 100.896 = 7.87. This means that the subjects had, on average, a 7.9-fold increase in antibody titer. The *t* value associated with this value is 3.314, with a two-tailed *p* value of 0.014.

Although we do not expect immunization to result in a decrease in titer, this is certainly an intelligible possibility, and as such, deserves to be factored into our analysis by the two-tailed test.

2.3.3 Data Sampling

Sampling has a special meaning in statistics. One of the tenets of the scientific method is that we can make inferences about an entire population by examining a sample drawn from that population. It is lost to history whether Student had a more convivial definition of sampling in mind when he invented his famous *t* test for the AQ6 Guinness brewery. To bring the subject back to immunology, the notion of sampling means that you do not have to measure the CD4 count of every living person to get a very good estimate of the distribution of CD4 counts in the population. Statisticians make a useful distinction between means and STDEVs calculated for entire populations, and those estimated from population samples. In educational testing, for example, it is possible to know the mean (location) and STDEV (dispersion) of the

test scores for the entire population of students taking the examination. In clinical testing, a population can only be sampled to infer normal ranges (i.e., the distribution of scores in subjects without known pathology). In experimental testing, replicate experiments are thought of as representing a sample from an infinite population of identical experiments that could have theoretically been performed. Although this sounds quite incredible, we will see how, by increasing the number of experimental replicates, our estimated mean and STDEV converge on the mean and STDEV of this fictitious population of all possible experiments.

2.4 EFFECT OF SAMPLE SIZE ON PARAMETRIC SUMMARY STATISTICS

The effect of sampling can easily be seen if we start with a data set in which, like our educational testing example discussed earlier, the results can be known for the entire population. Figure 2.8 shows data for white blood cell (WBC) counts measured on all patients who underwent peripheral stem cell mobilization at the University of Pittsburgh Medical Center between January 1999 and December 2006. To move hematopoietic stem cells from the bone marrow into the blood, these patients are treated with the growth factor G-CSF, and their WBCs are monitored for an increase that usually coincides with the peripheralization of hematopoietic stem and progenitor cells. Thus, the counts have much more variability than those of normal subjects. Results from the entire population (1224 determinations) were randomly sampled using increasingly large samples, until the entire population was included. Up to

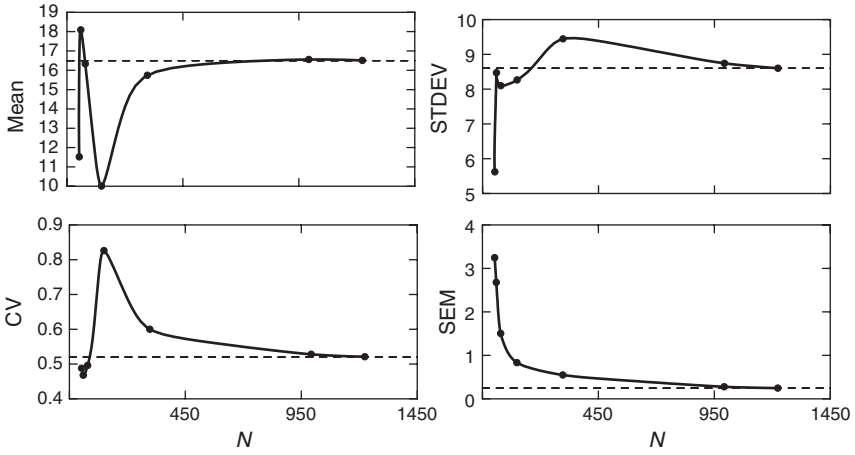

FIGURE 2.8 How the mean, sample STDEV, CV, and SEM change as a function of the number of observations (*N*). The data are peripheral WBCs (per microliter in 1000s) from patients receiving the hematopoietic growth factor G-CSF for mobilization of stem cells into their peripheral blood. Data were chosen at random from a set of 1224 observations. Mean, STDEV, CV, and SEM were calculated with an *n* of 3, 10, 30, 100, 300, 1000, and the full data set of 1224. The data points are connected by a spline fit. The statistics for the entire population are shown as a dashed line.

AQ10 and *n* of 300, the sample mean oscillates between 10.0 and 18.1, after which it stabi-
AQ11 lizes at about 16.5. The large STDEV indicates that there is a great deal of patient-to-patient variability, which when expressed relative to the mean (in the coefficient AQ12
AQ13 of variation [CV]), stabilizes at about 50%. Unlike the mean, STDEV, and CV, the estimates of which become more precise (not smaller) with an increasing number of observations, the standard error of the mean (SEM) is a measure of how precisely we have estimated the location of the population mean. Accordingly, it decreases as the number of observations grows, and theoretically, would be 0 when the entire population has been sampled.

2.5 STANDARD DEVIATIONS, STANDARD ERRORS, AND WHEN TO USE THEM

STDEVs and standard errors of the mean are the most commonly used statistics used to summarize data dispersion in the immunology literature. Now that we have seen how the sample mean, STDEV, and standard error change as the sample size increases, we can apply these statistics to our data with confidence. Bar graphs, line graphs, and dot plots are frequently shown with error bars to help us understand and compare data dispersion. Similarly, we often see numerical data summarized as 197 ± 62. Most commonly, these errors are expressed as STDEVs or standard errors. As seen, these two useful statistics have very different meanings. Despite this, they are sometimes (incorrectly) used interchangeably. Equally incorrectly, error bars or intervals are often used in technical publications without specifying which statistic they represent.

The *sample STDEV* tells you something about the spread of the data distribution. Use it as a summary statistic (e.g., as an error bar in a graph) when the point is to show the relative homogeneity or heterogeneity of the data. As already mentioned, when you use the sample STDEV to describe the spread of your data, you are implicitly endorsing the use of the normal distribution to model your data. You accept the premise that your data are well described by a family of symmetrical bell-shaped curves that are completely specified by the sample mean and STDEV. Your inferences about the location and spread of the population from which your sample was drawn will be correct only to the extent that the normal distribution fits this population. If the model does fit, you can use the STDEV to infer the upper and lower bounds, which encompass a given proportion of the population data. For example, 95th percentiles are calculated as the sample mean ± 1.96 STDEVs. Referring back to Figure 2.6, one can visualize how approximately two STDEVs above and below the mean include 95% of the normal distribution, with the remaining 5% lying out-
AQ14 side as *tails*. Below, we will provide an example where the STDEV of a measurement sampled from a healthy population is used to infer when an individual value from a patient is within normal limits. Remember that as the sample size (*n*) increases, the sample STDEV becomes a more precise estimate of the population STDEV. It does not become smaller. In this respect, it behaves something like the sample mean.

The *standard error of the sample mean* can be interpreted as a STDEV describing how means sampled from the entire population vary with respect to one another. Use the SEM (e.g., as an error bar in a graph) when you want to express the precision with which you know the location of the population mean or when comparing the means of

two or more or experimental treatments or groups (to determine the probability that they were drawn from the same population). As we saw in Figure 2.8, the greater the number of observations, the more precise is our estimate of the population mean.

2.6 ASSUMPTIONS OF CONFIRMATORY TESTS

Some of the most useful statistical tests rely on the normal distribution. These include Student's *t* test, analysis of variance (a generalization of Student's *t* test for more than two groups), linear regression analysis, and certain power calculations that are very helpful for experimental design. In addition to the assumption of normality, there are two other requirements that bear some explanation. The first is homogeneity of variance. For comparison of groups, both groups must have approximately equal variance (or equal STDEVs, the square root of the variance). The sample variances are then *pooled* to obtain the best estimate of the population variance. For linear regression, the STDEV of the dependent variable must not change as a function of the independent variable. In practice, a variant of Student's *t* test, the unequal variance *t* test, does not require this assumption, and can be used, provided the data are normally distributed [3]. For linear regression, a dependent variable with a variance that grows as a function of the magnitude of the independent variable will probably behave after transformation, as we will see in one of our case studies.

AQ15 The second assumption, known as continuity, requires that the universe of possible data values is continuous on the number line, that is, observations can be given to many decimal places arbitrarily. This is never actually the case in the real world where assay precision is limited by instrumentation and data are, in the strictest sense, discontinuous. To pick an extreme example, a hemagglutination assay may call for serum to be serially diluted in twofold increments. The titer is the lowest AQ16 dilution that causes the red blood cells to clump. Here, the only possible assay results (given as reciprocal titer) are 2, 4, 8, 16, etc. Given the assumptions of normality (hemagglutination titers are beautifully lognormally distributed) and homogeneity of variance, can we use Student's *t* test to compare data from two groups? Here is how John Tukey reconciled the "unrealism" of the assumption of continuity: "To the man with the data, such unrealism hardly looks helpful, but to the worker with a theory such an assumption turns out to make things simpler. Since all the man with the data can hope for is an approximation, and since an approximation to a few decimal places is often more than he deserves, he soon learns not to worry very often about this particular sort of unrealism." We may interpret this to mean that continuity is a formal requirement, and for us, the women and men with the data; it is sufficient that continuity of our measurement be intelligible, if not actual.

Why do I have to meet these requirements anyway?

- When you use a test like Student's *t* test to compare the data from groups A and B you are not really comparing the data itself.
- When you apply Student's *t* test to compare the results of group A and B, you are giving up the raw data in favor of a model that is *completely specified by only three numbers*: the mean of A, the mean of B, and a pooled STDEV that is supposed to accurately represent the spread of both groups.

- The test is invalid to the extent that the resulting normal distributions fail to accurately describe the experimental results.
- Remember that when the assumptions of normality or homogeneity of variance fail, transformations can often bring the data back into line.

2.6.1 A Word about the Null Hypothesis

In confirmatory statistics, the interpretation of a test is often related in terms of the null hypothesis. This elicits in most of us a mental cramp, since this is double negative, trying to refute the hypothesis of no difference, seems at a first glance the equivalent of trying to prove the hypothesis of a difference (what we students of science usually think we should be doing). This is not the case, as disproving the null hypothesis provides supporting evidence for, but does not prove an alternative hypothesis. Forcing ourselves to see our experimental problems in terms of the null hypothesis, and formulating statistical tests to reject the null hypothesis brings us into line, philosophically, with the falsifiability model of scientific discovery proposed by the twentieth century philosopher Karl Popper [4]. Popper believed that the closest science could come to truth is to provide theories with ample opportunity to be proven wrong by counterexamples. In other words, we gain confidence in a scientific hypothesis by attempting to prove it wrong and failing. Let us return to Figure 2.7 and our paired *t* test of antibody titers of sera obtained before and after immunization. Because this is a paired *t* test, the null hypothesis states that there is no difference between titers of individual subjects before and after immunization. To perform the statistical test, we subtract the log of the *pre* titer from the log of the *post*titer. When we test the null hypothesis, we determine the probability that the distribution of resulting *post–pre* differences could have been drawn from a (hypothetical) population of individuals for whom the mean difference between *pre*- and *post*values is zero, and the data dispersion, though unknown, can be estimated from the STDEV of our sample. Because we have sampled only nine pairs of data from this hypothetical population, the mean difference is very likely, by chance alone, to be a nonzero value. But how different from zero does it have to be before we can reject the idea that our data pairs were drawn from a population of subjects in whom there is no difference before and after immunization? As we have already seen, the mean difference of our sample was 0.896 (not 0) with a SEM of 0.303. A fast and intuitive way of thinking about this problem is to approximate the least significant difference (LSD) at $p = 0.05$. Just take the null hypothesis mean difference (0), and add two times the SEM (0.606). Is the sample mean difference (0.896) greater? Then we can reject the null hypothesis. Note that this does not formally prove that the immunization worked. There could have been other factors responsible for the rise in titers after immunization, or it could just be bad luck (in Figure 2.7, we found that a difference this large could be expected by chance alone, 1 in 71 times, if the nine data pairs were drawn from a population with a mean difference of 0). It does, however, provide evidence consistent with the notion that immunization caused a rise in titer. Further, unsuccessful attempts to falsify our hypothesis (can we replicate **AQ17** rejection of the null hypothesis in similar independent experiments, do we accept the null hypothesis when a sham immunization is given?) will give us greater confidence in our conclusion that the vaccine is efficacious.

2.6.2 A Word about Statistical Significance

The word *significant* does two jobs in the language of biomedical science. When we are speaking of a *statistically significant* difference, we usually mean that we have rejected the null hypothesis because, for example, the probability of two groups of data drawn from the same population is less than 1 in 20 ($p < 0.05$). This is an arbitrary cutoff, and we may wish to set a more rigorous criterion for statistical significance (say, $p < 0.01$) if our problem demands that avoiding type I error (false positives, see sidebar) is more important than avoiding type II error (false negatives). Strictly speaking, this concept of statistical significance applies to comparisons that are performed in isolation. If one were to perform 100 comparisons, then 5 (1 in 20) should turn up spuriously significant at a p value of 0.05, even when no differences exist. In such cases, the Bonferroni correction for multiple comparisons is used to adjust the required p value [5]. Finally, we have the semantic problem of the word significance, and the double job that it does in the language of science. Statistical significance in no way implies significance in the ordinary sense of meaningfulness or importance. We have seen earlier how the SEM decreases with sample size. Thus, with increasing sample size, even small and biologically unimportant differences may grow to be statistically significant. In words attributed to Gertrude Stein, author and Johns Hopkins School of Medicine dropout, "A difference, to be a difference, must make a difference."

Technical Definition

A type I error (or alpha error) is a false positive. In other words, it is rejecting the null hypothesis when it is actually true. The *p value* is the probability of alpha error.

A type II error (or beta error) is a false negative. It means failing to observe a difference when in fact a difference exists (or failing to reject a false null hypothesis).

2.6.3 Sensitivity and Specificity: Positive and Negative Predictive Values

When we are dealing with a test that provides dichotomous outcomes, such as a test to detect a viral infection (positive or negative), we can evaluate it in terms of sensitivity and specificity, which are defined in comparison to a reference or *gold standard*. The gold standard may be an existing and accepted test, or it may be an outcome that will eventually become known with time, such as the onset of clinical symptoms (e.g., HIV infection). Table 2.1 shows the relationship between a hypothetical-test result and the gold standard. We can see that sensitivity, the ability to identify true positives, is defined as the ratio of true positives to the sum of true positives plus false negatives. Similarly, specificity, the ability to detect true negatives is the ratio of true negatives to true negative plus false positives. Table 2.1 can also be used to conceptualize two other important parameters that depend not only on the test, but also on the prevalence of true positives in the population. These are the

TABLE 2.1
Sensitivity and Specificity are Defined in Relation to a Gold Standard

	Gold Standard	
Test Result	**Positive**	**Negative**
Positive	TP	FP
Negative	FN	TN

Note: About the test—sensitivity = TP/(TP + FN), ability to identify TP; specificity = TN/(FP + TN), ability to identify TN. About the population—positive predictive value = TP/(TP + FP), will be low when prevalence of TP is low; negative predictive value = TN/(FN + TN), will be low if prevalence of TP is very high.

positive and negative predictive values of the test. For a test of a given sensitivity and specificity, the positive predictive value will be low when the prevalence of true positives is low. Similarly, the negative predictive value is low when the prevalence of true positives is high. This can be very important in the interpretation of test results. For example, in a blood donor pool that has been prescreened to exclude subjects with risk behavior, the prevalence of HIV infection may be very low (say 1/10,000). On screening a million blood products from this donor group, we would expect to find 100 positive samples. Using an assay with 99.9% sensitivity and specificity, we would correctly identify virtually all the 100 positive samples. Of the 999,900 products that are not infected, the test would also correctly identify 999,000 as negative, but would incorrectly find 1000 samples to be HIV positive. Despite the excellent sensitivity and specificity, the positive predictive value is only 9%.

AQ18 AQ19

We also sometimes use the term sensitivity in the context of quantitative testing, for example, nephelometry is more sensitive than radial immunodiffusion for the measurement of serum immunoglobulin. Here, the term is used to compare the lower limit of detection of the assays. We mean that nephelometry can quantify immunoglobulins down to a lower concentration than immunodiffusion. This is usually related to the signal-to-noise ratio, a term borrowed from electrical engineering. It is based on the assumption that a measurement is an admixture of signal (the parameter that we wish to measure) and noise (a random or irrelevant contribution that contaminates the signal). Sensitivity depends on the relative strengths of signal and noise. In the context of assays of immunity, noise is the intensity of the background or negative control measurement (e.g., optical density of the negative control in an ELISA assay and fluorescence intensity of an isotype control in flow cytometry) and the signal is inferred to be what remains when the background is subtracted from the test result. Sensitivity in this sense can be increased both by strengthening the signal (in flow cytometry using a fluorescent dye with a higher quantum efficiency) or reducing the noise (blocking nonspecific antibody binding). From a statistical perspective, the lower limit of detection can be defined as an upper limit (e.g., 95th percentile) of an appropriate negative control value. For example, the lower limit of detection of an

AQ20

MHC class I-tetramer assay was defined by determining the mean and STDEV of the percent of log tetramer positive cells in an MHC mismatched setting (permitting only nonspecific binding). The upper 95th percentile of this negative control, determined to be about 1 in 10,000 cells, was considered to be the lower limit of detection [6]. AQ21

2.6.4 Precision and Accuracy

Precision and accuracy are two additional terms that have a specialized meaning in the context of assays of immunity. Precision is closely allied to reproducibility, and reflects the closeness of replicate determinations. The CV (Figure 2.8) among replicate determinations provides an estimate of precision. Like sensitivity and specificity, accuracy relies on comparison to some gold standard. An intuitive example is a pipetting device that may aliquot the same amount of liquid each time within a very close tolerance, but is out of calibration. This device is precise but not accurate. It is entirely possible to be accurate without being precise, and vice versa (Figure 2.9).

An example in which the precision of a new assay was determined is shown in Figure 2.10, where a nuclear stain was used to quantify leukocytes contaminating filtered platelet products [7]. Each assay was performed in triplicate and the CV was determined and plotted as a function of the leukocyte count. The assay was very precise at high leukocyte concentrations, but the precision deteriorated rapidly at counts less than 100 cells/mL. Accuracy of the same assay was determined by carefully diluting leukocytes in a sample of known concentration (Figure 2.11). The cell count calculated from the dilution (x-axis) was taken as a gold standard, and the observed results (y-axis) were compared by linear regression analysis. The unexpected nonlinearity at the lower region of the curve may have come from the dilution error, but the model assumes that it is from counting error. Despite this, an intercept near 0 and a slope near 1, and small differences between observed and expected values indicate the accuracy of this assay.

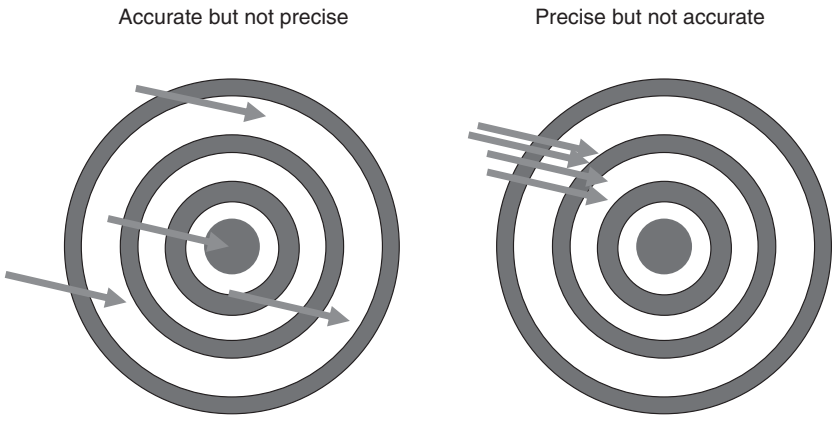

FIGURE 2.9 Accuracy and precision are independent attributes. The target on the left-hand side illustrates accuracy without precision (the position of the arrows average to a perfect bull's-eye), whereas, the arrows on the right-hand side form a tight cluster (high precision) but fail to hit the mark.

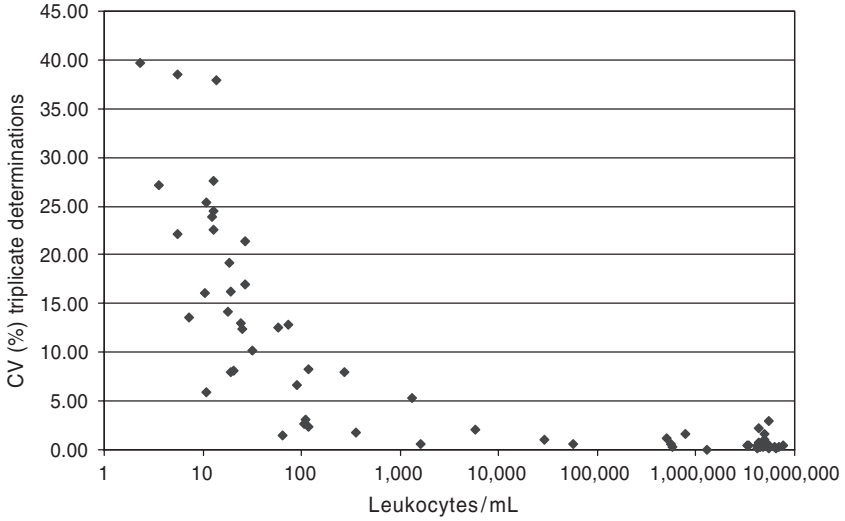

FIGURE 2.10 The precision of an assay that measures the leukocyte content of leukoreduced platelet products decreases at lower cell concentrations, as determined by the CV of triplicate determinations.

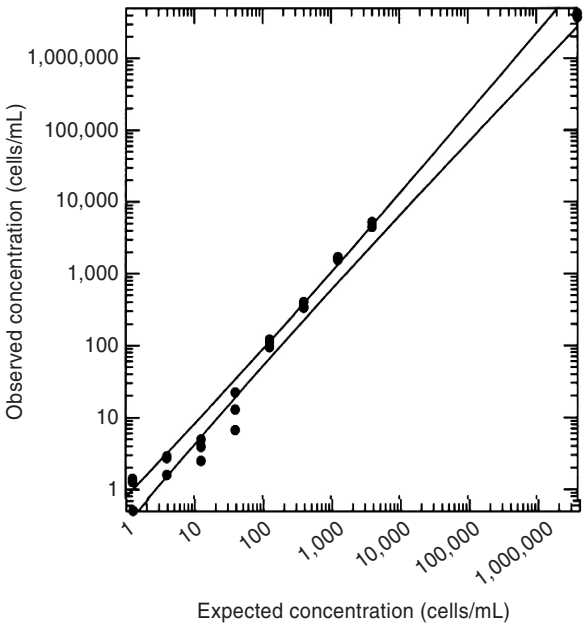

FIGURE 2.11 Careful dilution of a sample of known concentration can be used to create a gold standard (expected concentration) to judge the accuracy of an assay designed to count cells at very low concentration. The sample was diluted in half-$\log_{10}$ increments and assayed in triplicate. There is unexplained nonlinearity at some of the lower concentrations. The lines represent the 95% confidence intervals about the linear regression line of best fit.

2.6.5 Statistics and Experimental Design

The only complaint that statisticians have about experimentalists (and especially immunologists) is that we think about statistics only after getting the results of our experiments. Modern statistical packages contain tools to assist with experimental design, the most important of which are for power analysis. For a two-way comparison, power analysis forces one to think about one's experiment in terms of the smallest difference that one would care to detect between groups (remember Gertrude Stein?), and the degree of data dispersion expected within these groups. In practice, it is like backing into a test of significance. For a real-world example, a graduate student came to me with a data set comparing the expression of an immunophenotypic marker in two groups. After doing a preliminary experiment with three subjects, each in two groups, he noticed a difference in the hypothesized direction, but on performing a *t* test, failed to reject the null hypothesis. He decided that the minimal difference in marker expression that he wished to be powered to detect was 25% (a smaller difference seemed to him unlikely to be biologically important). In the example shown in Figure 2.12, his hypothesized results were a mean of 10% in group 1 versus 35% in group 2. From his preliminary data, he estimated the STDEV

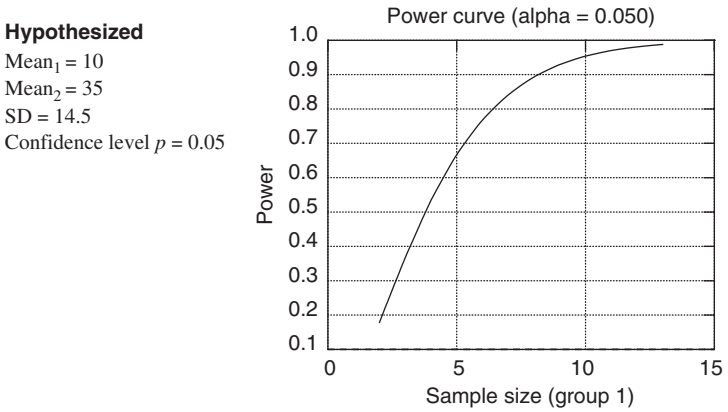

FIGURE 2.12 This is a true example from the doctoral dissertation work of Dr. Aki Hoji, then an immunology graduate student at the University of Pittsburgh Graduate School of Public Health. He was comparing phenotypic markers on two T-cell subsets, therefore the data is expressed as percent positive for a given marker. He did a preliminary experiment with three subjects per group. At that time the data looked promising, but had not reached statistical significance. He decided that the minimal difference that he wished to be powered to detect was 25%. That is, a difference of less than 25% expression for a given marker was not deemed to be of biological importance. In the example shown, the hypothesized percent positive were 10 and 35% for the two subsets being compared. This is arbitrary and the same conclusion would have been reached for hypothesized values of 0% versus 25% and 50% versus 75%. The STDEV (14.5%) was calculated from his preliminary data. The results were encouraging. They indicated that at his present group size ($n = 3$), he had only a 37% probability of detecting a 25% difference between the groups (at $p = 0.05$), even if such a difference actually existed. Further, the power calculation indicated that increasing the group size to $n = 7$ would give him an 84% probability of detecting such a difference.

at 14.5%, and he chose a conventional alpha value (significance level) of 0.05. The results of the power curve, which examines the effect of sample size on power (1—beta error), showed that his preliminary data ($n = 3$) had only a 37% probability of detecting a 25% difference between the groups, if any existed, at a p value of 0.05. Increasing the sample size to $n = 7$ increased this probability to 84%. Knowing this, he went back to the bench. To be kosher, the power calculation should have been made before the experiment was initiated, or should have been made using data from an independent set of experiments. Performing interim analyses (taking peeks at the data) influences the meaning of a p value in much the same way that multiple comparisons do.

How many mice/patients/samples do you need to complete your experiment? All you need to know is

- The assay results conform to the requirements of the test.
- The smallest between-group difference that you would call biologically meaningful.
- An estimate of the STDEV of the assay itself.
- The alpha level. The p value you will call significant (conventionally at $p = 0.05$ or less).
- How much power you need (usually 0.8, i.e., in 80% of such experiments you would detect such a difference, if it exists, at the stated confidence level)? AQ22

2.6.6 A Word about Nonparametric Statistics

Nonparametric statistical tests, such as the Wilcoxon signed rank test, the Mann–Whitney U test, the Kruskal Wallace test, and the Spearmen rank coefficient are confirmatory tests based on the rank order of data points rather than their magnitude. They, therefore, do not require the data to conform to a particular distribution, and are not sensitive to the magnitude of outlying values. They can also accommodate data points that are outside the dynamic range of the test (e.g., undetectable and too numerous to count). These tests can be very useful when comparing data that are drawn from a complex distribution, such as one with two or more peaks, or small data sets, where it is difficult to determine the distribution from which the data are drawn. As robust as nonparametric tests are, experience teaches that they are rarely necessary for the vast majority of immunological testing data. Given a data set that can legitimately be analyzed, raw, or in transformation, by tests that rely on normality, an equivalent nonparametric test will often have less statistical power (sensitivity).

2.6.7 Graphic Representation of Data

2.6.7.1 Words of Wisdom Concerning Data Graphics

A picture may be worth a thousand words. Here are a few hundred words about pictures.

> Pictures based on exploration of data should force their messages upon us. Pictures that emphasize what we already know—security blankets to reassure us—are frequently not worth the space they take. Pictures that have to be gone over with a reading glass to

see the main point are wasteful of time and inadequate of effect. The greatest value of a picture is when it forces us to notice what we never expected to see.

John Tukey

Less is more.

Ludwig Mies Van der Rohe

Pie charts are among the most abused graphics icons.

We cannot think of a single instance in which a perspective bar graph should be used for any application.

A nasty relative of the perspective bar chart is the pseudo perspective bar chart. Illustrators frequently feel the need to make two dimensional bars look like blocks or sky scrapers.

Leland Wilkinson [8]

Forego chartjunk, including moiré vibration, the grid, and the duck*.

The task of the designer is to give visual access to the subtle and the difficult—that is, the revelation of the complex.

Edward Tufte [9]

2.6.7.2 Practical Advice on Data Presentation

- Watch your proportions. In the olden days, presentation data were shown using 35-mm slides that have an aspect ratio (width/height) of 1.55. Today, most of our graphics end up in PowerPoint, which has an aspect ratio of 1.33. Therefore, graphs, tables, and word charts should be about 30% wider than tall. A significant departure means wasted space and a graphic that will read smaller when projected on a screen. AQ23
- "Wider-than-tall shapes usually make it easier for the eye to follow from left to right ... smoothly-changing curves can stand being taller than wide, but a wiggly curve needs to be wider than tall" (John Tukey).
- Choose backgrounds carefully. White is fine for print, but creates a terrible glare when projected in a large auditorium. I personally prefer a light gray, but deep blue with white letters can also be effective in a retro kind of way, in imitation of diazo print photographic slides.
- For publication graphics, anticipate the final size of the printed graphic. Choose the type size and the line thickness accordingly. Letters should not

* Tufte defines chartjunk as ink that does not tell the viewer anything new. This includes all forms of spurious or superfluous decoration and especially computer-generated grids and crosshatching patterns that cause the Moiré effect: "designs [that] interact with the physiological tremor of the eye to produce the distracting appearance of vibration and movement." To know more about the duck, see Ref. 8.

clot when they are reduced (e.g., as when the upper half of the lower case "e" fills in). Likewise, lines should not become so thin that they fade.
- Choose fonts carefully. Sans serif fonts (e.g., Ariel and Helvetica) tend to look cleaner, but words are more easily recognized by sight when seriffed fonts (e.g., Times Roman and Georgia) are used.

2.7 THREE CASE STUDIES

2.7.1 Case I: Thymidine Uptake Data

Uptake of radiolabeled thymidine is a tried and true method for measuring the proliferation of lymphocytes in response to mitogens or specific antigens [10]. Cells are cultured for a sufficient duration to enter the log phase of proliferation, after which they are given a short pulse (usually 4 h) with tritiated thymidine, which is taken up specifically by proliferating cells. The cells are trapped on a filter, washed, and the incorporated thymidine is measured by scintillation spectrometry. The results involve a comparison of signal (label incorporation in cultures exposed to a mitogenic stimulus) and noise (background incorporation of label in the absence of stimulation). In the present example, the data are from an antigen-specific lymphoproliferation test performed on peripheral blood from a series of 76 patients. The results are expressed as net counts per minute (net CPM), that is, the CPM incorporated in the presence of antigen minus the CPM incorporated in a no antigen control from the same sample (Figure 2.13). The net CPM were very skewed, with a median of 1,750 and a mean of 11,125, the mean being strongly influenced by a tail of large values. The data were, therefore, log-transformed and evaluated. Zero or negative values, resulting from background counts equal to or greater than the experimental counts ($n = 6$) were assigned an arbitrary small value of 1 prior to transformation, because logarithms are undefined for these values. In this case, negative values represented less than 10% of the data. A preponderance of negative values would have steered us away from log transformations and assumptions of normality, and toward a nonparametric test. Log transformation went a long way to make our data fit the normal distribution, with a mean of 3.023 (1054 when back-transformed to a geometric mean by taking the antilog, and much closer to the median of 3.243). This can be appreciated in the horizontal notched box plots shown in Figure 2.13. Figure 2.13 also introduces the normal probability plot, another useful and commonly available tool for evaluating data distributions. Here, the ordered data (x-axis) are plotted against the value predicted by the normal distribution (normalized to a mean of 0 and a STDEV of 1). A linear relationship indicates that the data are well described by the normal distribution and is evidence of the appropriateness of using tests requiring normality. The EDA tools applied in Figure 2.13 point to the same conclusion. Although the log-transformed data are not perfect (there is a hump on the left resulting from negative values and a resulting flattening of the curve [negative Kurtosis]), the transformed data are decidedly better described by the normal distribution than are the raw data. This is about all one can expect from real-world data.

As mentioned earlier, the use of t tests, analysis of variance (a generalization of the t test for multiple groups), linear regression analysis, and several other statistical

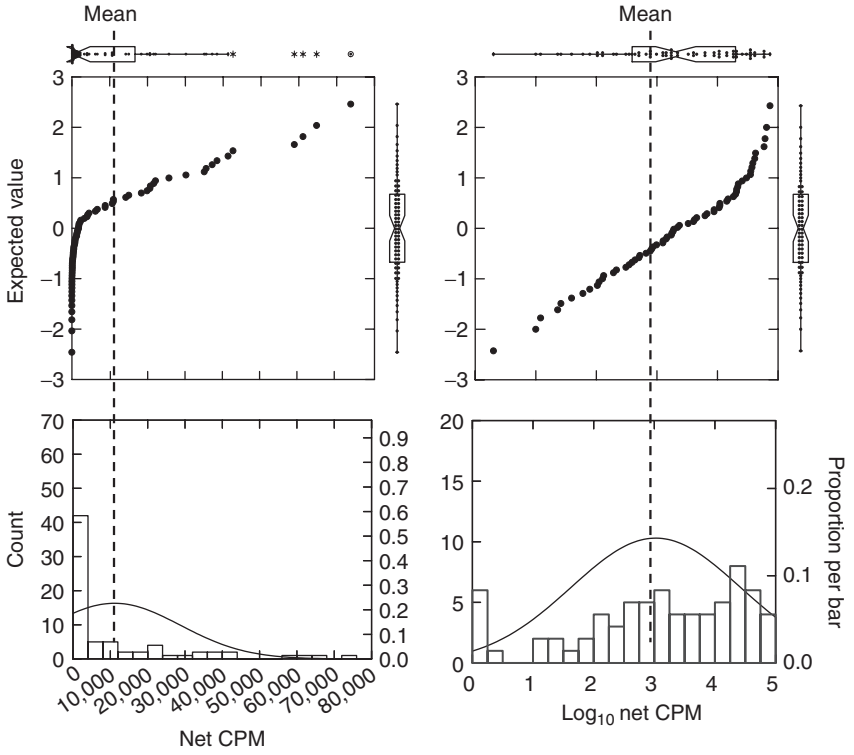

FIGURE 2.13 Use of notched box plots, probability plots, and histograms to assess the conformance of a data set to the normal distribution. In the top panels, net tritiated thymidine uptake (net CPM) and the log transform of net CPM are displayed as normal probability plots, where the *x*-axis shows ordered observed results, and the *y*-axis shows results predicted by the normal distribution. The linearity of the resulting relationship indicates the normality of the data. The probability plots are flanked by notched box plots, a variation of the box plot in which the notches at the waist show the 95% confidence interval about the median. The lower panels show histograms with superimposed normal distributions calculated from the data means and standard deviations, comparable to those shown in Figures 2.5 and 2.6. The CV of the raw data is 158% compared to a CV of 46% for the log-transformed data.

tests depend on the assumption that the data are normally distributed. Often immunological data are not. When data are not normally distributed, the mean and STDEV are poor summary statistics. The greater the departure from normality, the less sensitive the test is to detect real differences in the data. Many types of immunological data are log-normally distributed. Examples include radionuclide uptake, antibody titers, absolute blood counts, MHC tetramer binding frequencies, precursor frequencies, fluorescence intensity of cell surface immunophenotypic markers, and frequencies of rare populations detected by flow cytometry. When such data have been log transformed, the mean and STDEV better represent the data, tests depending on these parameters are more sensitive, and summary statistics are more appropriate.

TABLE 2.2
Student's t Test Performed on CPM Data from Two Groups

	Mean CPM (Groups 1 and 2)			*P* Value (Two Tailed)	
N	Arithmetic	Log	Geometric	Raw Data	Logged
3	2960, 6293	3.465, 3.703	2951, 5045	0.221	0.369
10	1890, 3180	3.239, 3.445	1734, 2784	0.063	0.040
30	1980, 4675	3.262, 3.546	1827, 3512	0.001	7×10^{-5}
175	1949, 4826	3.248, 3.550	1772, 3547	4×10^{-14}	6×10^{-23}

Note: Samples of increasing size were drawn at random from a dataset containing 175 observations per group.

Before we leave the effects of log transformation on radioactive thymidine uptake data, let us consider a comparison of the lymphoproliferative responses of two groups of 175 subjects each (Table 2.2). Samples of increasing size (n) were drawn from this data set at random. The sampled data were compared by Student's t test using the raw data (arithmetic means) or geometric means. The geometric mean is obtained by log transforming the data, taking the average, and then back transforming the average to the linear domain by taking the antilog. At $n = 3$ (three subjects per group), group 2 appears to have a higher mean than group 1, but the comparison is not statistically significant. An independent sample of 10 subjects yields statistical significance using both raw and log-transformed data, but the p value is somewhat smaller using logs. In this data set, the superiority of log transformation is most evident when n is large. This case study demonstrates both the practical advantages of transforming the data to better toconform to normality, and the incredible robustness of Student's t test, which even when the assumption of normality is flagrantly violated, still manages to provide the correct answer, albeit with some loss of sensitivity. It is worth noting here that despite our insistence on the importance of transformation, the central limit theorem states that no matter what the shape of the population distribution, the sampling distribution of the mean approaches normality as the sample size increases. However, a very large sample size is required before a highly skewed distribution converges to normality, and we rarely have this luxury with experimental data.

2.7.2 Case II: T-Cell Receptor Heterogeneity, as Measured by the v-Beta Distribution

The antigen receptors on T-cell clones acquire their incredible diversity by a process involving gene rearrangement and mutation at splice sites (see details in Chapter XXX). One of the elements contributing to receptor diversity is the 57-member v-beta gene family that codes for the variable portion of the T-cell receptor beta chain [11]. Only one v-beta specificity is expressed on a given T cell or its clonal descendants. Since 24 of the 57 members account for more than 75% of all TCR v-beta usage, a panel of 24 antibodies can be used to classify the majority of human T cells. v-beta usage within these 24 families follows a characteristic distribution,

AQ24

with some members representing a larger proportion of the population than others. At most, a given specificity will represent 10% of the total T cells, but most are much less prevalent. This is illustrated in Figure 2.14, which shows the v-beta distribution on $CD4^+$ T cells from 21 healthy subjects. Applying a Gaussian model to our data (Figure 2.15), we can determine normal ranges for each v-beta specificity, arbitrarily defining individual v-beta deletions and expansions as outside the bounds of the 2.5th and 97.5th percentiles, respectively, of the normal distribution calculated from the means and STDEVs of our data. For each v-beta specificity, we expect, by definition, a frequency of 2.5% expansions and 2.5% deletions in our healthy population. We can check this by applying the calculated cutpoints to our control data set. Given that we have measured 24 v-beta specificities in 21 subjects, we would expect to find, on average 12.6 deletions and 12.6 expansions ($24 \times 21 \times 0.025 = 12.6$). However, when we apply our cutpoints, we get 25 expansions and only three deletions. What went wrong? To find out, let us examine our assumptions about normality. Lest you begin to think that all immunological data require transformation, let us

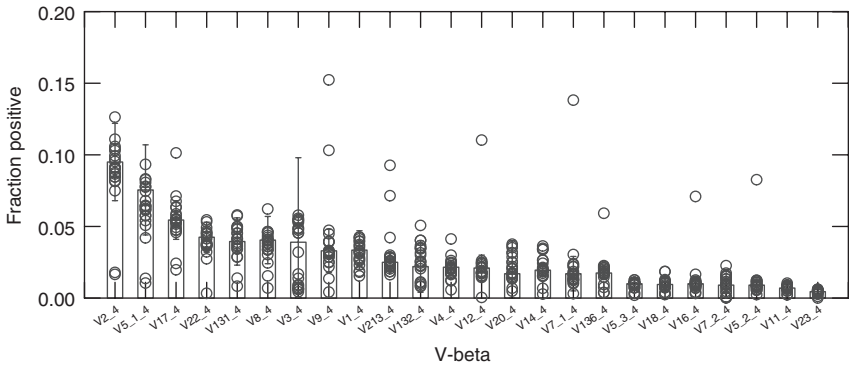

FIGURE 2.14 Distribution of v-beta specificities in peripheral blood $CD4^+$ T cells from 21 healthy subjects. Individual v-beta specificities are ordered by prevalence. The circles show individual data points. The bars show the arithmetic mean values, and the error bars represent 1.96 standard deviations about the mean.

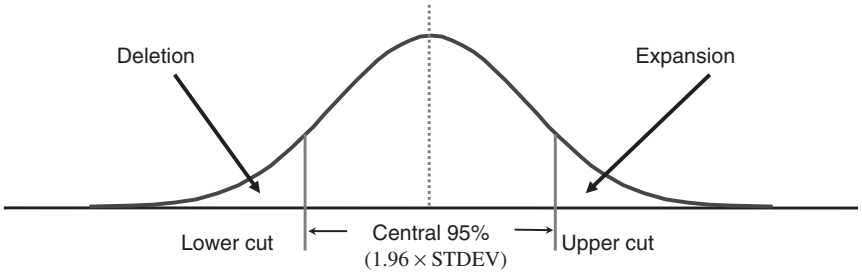

FIGURE 2.15 The Gaussian model can be used to define normal ranges of usage for each v-beta specificity. Deletions and expansions were defined as lying ouside the central 95% of the distribution.

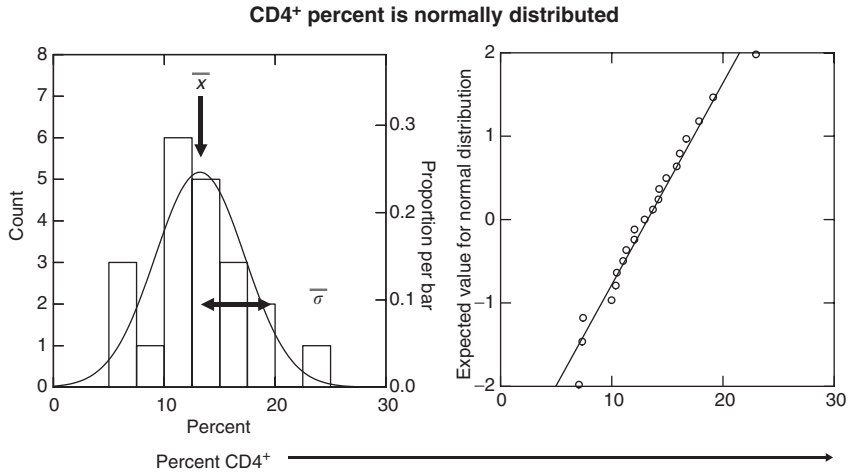

FIGURE 2.16 Distribution of $CD4^+$ cells among lymphocytes. The histogram shows the distribution of the actual data; the superimposed normal curve is calculated from the sample mean and standard deviation. The probability plot confirms the normality of the data.

begin by applying our EDA tools to the percent of $CD4^+$ cells in our 21 samples (Figure 2.16). Visual inspection of a bar histogram of the data, the superimposed normal distribution calculated from the sample mean and STDEV, and a probability plot all confirm that CD4 expression is normally distributed among lymphocytes of healthy subjects. In Figure 2.17, I have applied probability plots to both raw and log-transformed frequencies of usage for two v-beta specificities, v-beta 1 and v-beta 7.2, demonstrating quite conclusively that v-beta usage is lognormally distributed. In general, this appears to be the case for all measurements of rare events. Figure 2.18 is a reworking of Figure 2.14, using a logarithmic scale on the *y*-axis to reflect the lognormal distribution of TCR v-beta. Defining deletions and expansions as lying outside the central 95% of the log-transformed control data set, as we did previously on the raw data, we now calculate 12 expansions and 14 deletions among our controls. This is remarkably close to the 12.6 deletions and expansions predicted on the basis of the normal distribution. Unlike the thymidine uptake example, where failure to log transform merely reduced our power to detect a difference, failure to transform TCR v-beta usage data gave us completely erroneous normal ranges. Had we blindly applied the normal range cutpoints calculated from the raw control data to a set of patient data, we would have greatly overestimated the frequency of expansions and underestimated the frequency of deletions, and therefore seriously misread the underlying immunobiology.

The following happens if you fail to transform your data when you should:

- The sample mean is greater than the sample median (lognormal data is asymmetrical and has a tail of high values).
- The STDEV is larger than it would be for log-transformed values. The example in Figure 2.13 shows that the CV was more than three times greater in the untransformed data.

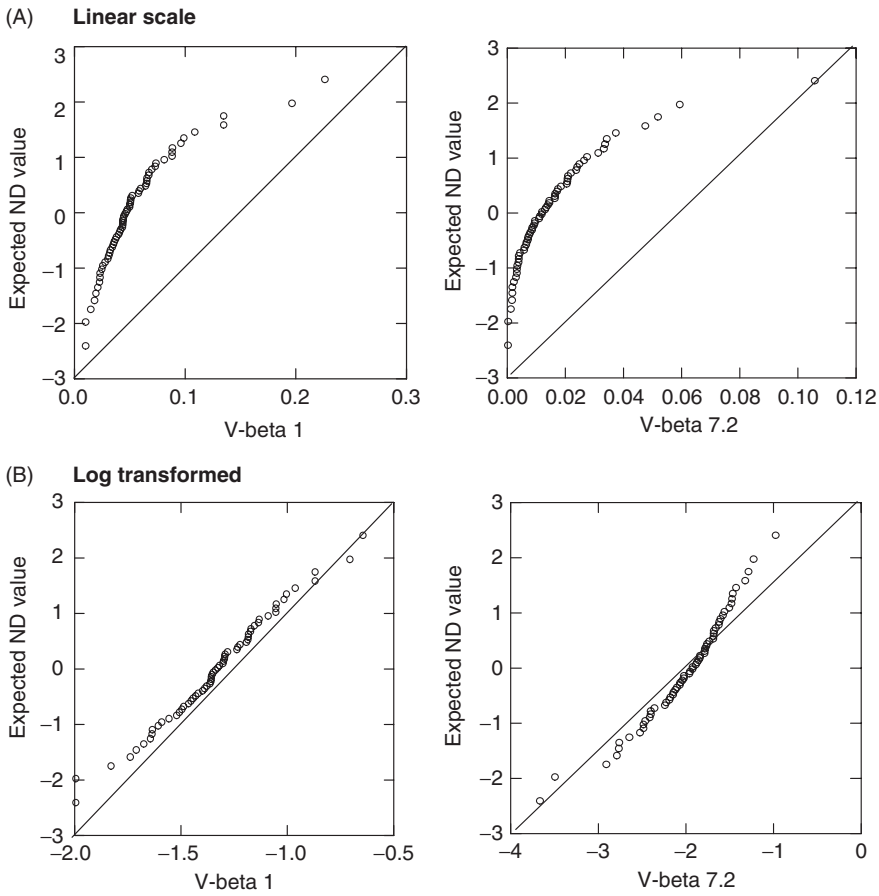

FIGURE 2.17 V-beta usage is lognormally distributed. The distribution of two v-beta specificities, v-beta 1 and v-beta 7.2, are shown for both raw (A) and log-transformed (B) results from a larger dataset. The probability plots clearly demonstrate that v-beta usage is more closely modeled by the normal distribution after log transformation.

- The STDEV does not represent the distance from the mean to the point of inflection of the sample distribution (how could it since the distribution is asymmetrical?). This greatly throws off percentile (probability) calculations based on the number of STDEVs above and below the mean. AQ25
- It takes a greater difference between the means of two groups to attain statistical significance, becausee the sample STDEVs are larger.

2.7.3 Case III: Linear Regression and Correlation—Peripheral Blood CD34 Counts Predict the CD34 Dose Harvested by Leukapheresis

Simple linear regression analysis is designed to test a particular hypothesis concerning the relationship between two variables: The dependent variable increases in proportion to the independent variable. The distinction between the independent

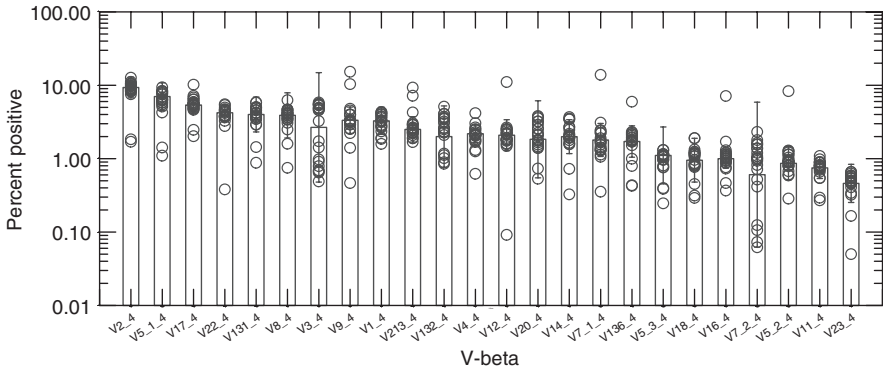

FIGURE 2.18 Distribution of log-transformed v-beta specificities in peripheral blood $CD4^+$ T cells from 21 healthy subjects. Individual v-beta specificities are ordered by prevalence. The circles show individual data points. The bars show the geometric mean values, and the error bars represent 1.96 standard deviations about the geometric mean.

variable, conventionally plotted on the *x*-axis, and the dependent variable, plotted on the *y*-axis, is an important one. The independent variable is like a gold standard; it is assumed to have no variance, whereas, the dependent variable is a measurement that entails inherent variability. An example might be the relationship between the measurement cell number (dependent variable) and time in culture (independent variable), or a response predicted by theory (independent variable), and the actual measured response (dependent variable). There is often the implicit assumption that the independent variable causes changes in the dependent variable. Linear regression assumes that the relationship is a straight line, which can be stated as

$$y = a + bx$$

where

y = predicted value of the dependent variable
a = y intercept of the line
b = slope of the line
x = value of the independent variable (Figure 2.19).

You may have learned the same equation in algebra class as $y = mx + b$. Simple linear regression analysis uses the method of least squares to determine the line of best fit through a series of data points (pairs of observations) by minimizing the difference between the observed values and those predicted by the linear equation (residuals). Unless the line of best fit passes exactly through each pair of expected and observed values, the intercept (a) and the slope (b) will have variance. We can express this variance in terms of the STDEVs of the intercept and slope, or by calculating confidence intervals about these parameters. We can also determine the coefficient of correlation (r^2), a parameter that is estimated as 1—the sum of the squared differences between predicted and observed y values (regression sum

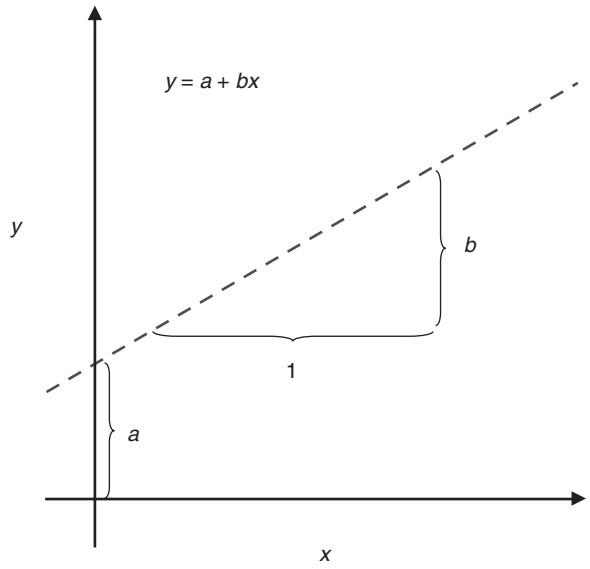

FIGURE 2.19 A linear relationship between the dependent variable (y-axis) and the independent variable (x-axis) is modeled by an intercept parameter (a) and a slope parameter (b).

of squares), divided by the sum of the squared differences between the predicted y values and the mean of the observed y values (total sum of squares). In other words, it is the fraction of the total variance in the variable y that is explained by the regression of y against the variable x. r^2 is a fraction between 0 and 1, where 0 indicates no correlation and 1 indicates that all the variance in y can be explained by the value of x (i.e., a perfect fit). Unlike regression analysis, there is no assumption that one variable is dependent and the other independent. The correlation coefficient r^2 can be computed between any two measurements, even if both have variance.

It is important to remember that a "line of best fit" with its intercept, slope, and r^2 can be calculated for any set of paired values. Even a slope that is significantly different from 0 does not guarantee that a linear relationship exists between x and y. Figure 2.20 shows the line of best fit through a hyperbola, defined by the equation $y = 1/x$. Despite the high coefficient of correlation ($r^2 = 0.915$) and a slope significantly different from 0 (-0.301, $p < 0.0005$), a straight line does not make a very compelling fit. The residual plot (a plot of the difference between observed and predicted values as a function of the estimated value) provides a very powerful EDA tool with which to judge the appropriateness of the linear model. In Figure 2.20, we can see that at increasing values of x, the linear model first underestimates, then overestimates, then underestimates the observed values.

Linear regression also assumes that data are normally distributed, continuous, and homoskedastic (here, this means that the variance of y does not change as a function of x). The residual plot can help you assess the homoskedasticity of the data, because the residuals should be evenly distributed across the estimated values in both sign and magnitude.

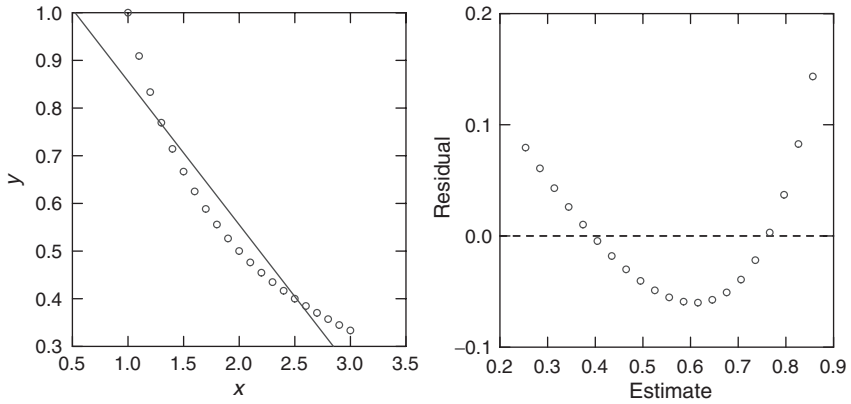

FIGURE 2.20 Fitting a linear model to hyperbolic data. The data points plotted on the left were generated from the hyperbolic function $y = 1/x$. Despite the obvious departure of these data from linearity, the coefficient of correlation was quite high (0.915). The nonrandom distribution of residuals (right-hand side), clearly show the shortcomings of the linear model.

Autologous or matched allogeneic hematopoietic stem cells can be used to rescue the bone marrow of patients treated with high-dose chemotherapy for a variety of hematologic malignancies. The hematopoietic stem cell graft is obtained by administering hematopoietic growth factors such as G-CSF [12]. This causes hematopoietic stem cells to mobilize from the bone marrow into the peripheral circulation. G-CSF is administered for several days after which leukapheresis, a technique that captures WBCs, while returning red cells, platelets, and plasma to the donor, is performed. The duration of treatment required to mobilize stem cells into the blood is variable, especially in patients who have had extensive previous chemotherapy. If you leukapherese the patient too soon, few stem cells are obtained necessitating further leukapheresis. Leukapherese too late and the patient will have to be rested and remobilized at a later time delaying the critical therapy. A biomarker predicting the results of leukapheresis would be most helpful.

To determine whether the peripheral CD34 count predicts the cell dose obtained on the same day by leukapheresis, we performed linear regression analysis of the $CD34^+$ cell dose obtained by leukapheresis (measured in $CD34^+$ cells/kilogram of patient body weight) as a function of the peripheral CD34 count (measured in $CD34^+$ cells/μL). Our first order of business would be to determine whether these variables are approximately normally distributed. As we have seen, there are many tools for this, probability plots, box plots, and density histograms among them. For the purpose of this exercise, we will blast ahead with the analysis at our peril, without EDA. Figure 2.21 shows the poor results. The data are crammed against the *y*-axis. A few very high values are dragging the regression downward (this is called leverage, and is not a good thing). The residual plot tells the complete story. The leftward skewing of the data is even more apparent (departure from normality), and the residuals are small for low values and get larger for high values (heteroskedasticity). The residuals seem pretty evenly distributed around 0 for low values, but are all negative and large for the four highest values (leverage). Had we started with probability plots, we

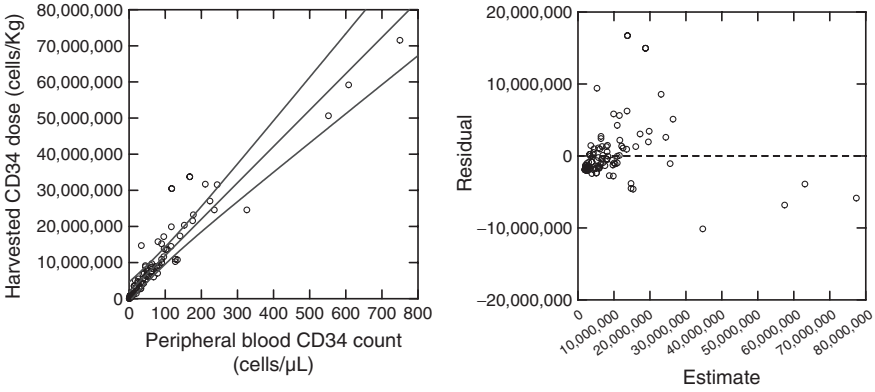

FIGURE 2.21 Linear regression of CD34 harvest dose as a function of peripheral CD34 count. The individual data points are shown as circles, the least squares line of best fit and its 95% confidence bounds are shown on the left-hand side. A plot of the residuals of regression is shown on the right-hand side.

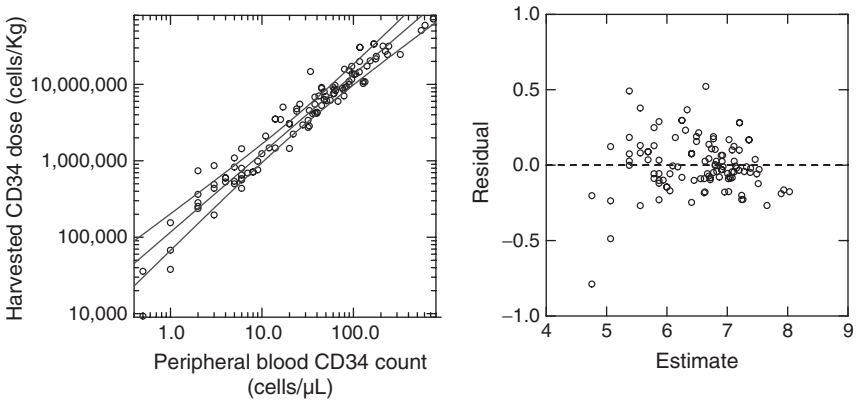

FIGURE 2.22 Linear regression of log CD34 harvest dose as a function of log peripheral CD34 count.

would have immediately recognized that both variables are log normal. Figure 2.22 shows what happens when we log transform the data and repeat the regression analysis. The data are fairly evenly spread, and the residuals are much better behaved, although now, at low values some negative residuals with high leverage have popped up (these are after all data from the real world). Most importantly, we have a regression equation of the form as follows:

$$y = 10^{5.1+1.03\times x}, \quad r^2 = 0.94$$

where y is the predicted CD34 harvest dose in cells per kilogram. This can be used to predict the cell dose that can be expected for a patient with a given peripheral CD34 count. In practice, we were able to greatly improve on this regression equation, using a

technique called multiple linear regression, by factoring in other variables that carried independent information. These included how long the patient was on the leukapheresis machine, how many times the patient had been previously leukapheresed, and the sex of the patient. This has provided a practical tool that is used everyday in the clinic.

2.8 CAVEATS AND HANDY FACTS

- *Look at the data first.* Data analysis should always begin with EDA. You can tell a lot about your data just by plotting them. This includes the shape of the distributions, the presence of extreme values, and the nature of the relationships between variables. EDA tools such as probability plots can help you choose your models and your transformations.
- *The mean of differences.* It is not the same thing as the difference of two means. If data are naturally paired, you usually want to take the mean of individual differences. For example, if you have the data from which you subtract the "background," the average of these differences is not mathematically equivalent to taking the mean of the data and subtracting the mean background.
- *Beware of ratios.* Ratios are sometimes essential to the analysis of immunological data but they are rarely well behaved from a statistical standpoint. Also, remember that, as mentioned earlier, the average of ratios is not computationally equivalent to the ratio of averages.
- *Outliers.* Victor McKusick, a pioneer of medical genetics, exhorted his students "cherish your exceptions." Outliers often provide a key to normal biology. In contrast, if you know that technical problems encountered during an experiment explain the outlying values, it is a disservice to leave them in the analysis where they are a known source of noise. Let your conscience be your guide.

2.9 SOME HANDY PROPERTIES OF LOGS, ADAPTED FROM TUKEY [1]

- Logs in different bases differ only by multiplicative constants. You can convert a logarithm of base x to a logarithm of base y by dividing it by the base x log of y. For example, to convert a natural logarithm to the base 10, divide it by the natural log of 10. Because one base is as good as another, serial fourfold dilutions of serum, for example, can be treated as arbitrary whole number units: 0, 1, 2 for 1/8, 1/32, 1/128, respectively. After averaging these arbitrary units, the mean can be "back-converted" to a reciprocal titer by the formula: $4^x \times 8$, where 4 is the fold-dilution, x the titer in arbitrary fourfold units, and 8 is the starting dilution. In the traditional serological literature, this is known as a geometric mean. Another way to get to the same place is to take the logs of the reciprocal titers, average them, and then take the antilog of the average.
- Engineers and mathematicians often express logarithms to the base e because e (2.71828183) crops up naturally in a variety of situations involving

derivatives and limits and simplifies calculations. Base *e* logarithms are known as natural logarithms (ln). In the biological sciences, base 10 logarithms are more common (in fact, they are known as common logarithms) because it is easier for most of us to think in powers of 10 rather than in powers of 2.7: the $\log_{10}$ of 10 is 1, of 100 is 2, of 1000 is 3, etc.

- The log of 1 is 0 (any base).
- The log of 0 is undefined. This sometimes poses a problem in the case of data such as CPM above background that can take on 0 or negative values. Most data realists agree that it is acceptable to add an arbitrary small number to data values before taking the log. Thus, radioisotopic data might be transformed as $\log_{10}$ (net CPM + 1).
- The log of a product is the sum of the logs of the factors: $\log (x \times y) = \log x + \log y$.
- The difference of log values is a ratio. For example, in base 10 logs, subtracting 3 from 6 is equal to 3. In the linear domain, this translates to $10^6/10^3 = 10^3$, or 1,000,000/1,000 = 1,000.
- The geometric mean is obtained by logging the values, taking their average, and then taking the antilog of the average.

2.10 PRESENTING DATA THAT HAVE BEEN LOG TRANSFORMED

Although working in base 10 logarithms gets to be intuitive very quickly, it is always a good idea to plot or tabulate data in the linear domain. This is accomplished by taking the antilog of the logged values. Be aware of the properties of logarithms mentioned. For example, if you need to calculate error bars, the lower limit of your error bar should be $10^{(\text{mean}-\text{SEM})}$ and the upper limit $10^{(\text{mean}+\text{SEM})}$, not $10^{\text{mean}} \pm 10^{\text{SEM}}$. You will know that your error bars are correct if they are symmetrical about the mean when plotted on a log scale, but asymmetrical (with the lower bar being shorter) when plotted on a linear scale.

AQ26 2.11 CLOSING REMARKS

This chapter discusses how a handful of exploratory and confirmatory statistical techniques can improve the design and interpretation of immunological experiments. Although immunology has always been a quantitative science, improvements in instrumentation and the direct interface of instruments to computers have allowed us to design more complex experiments in which more parameters are measured with higher resolution and greater dynamic range. Although this is certainly a blessing, it is a mixed one, because the interpretation of experimental results is commensurately more complex and, as we saw in the TCR v-beta example, the chance that rote application of statistical methods will lead to erroneous interpretations is magnified. The key to good data analysis then is to begin with techniques that allow one to visualize the lay of the data before one chooses confirmatory tests that require assumptions about it. One cannot overemphasize the importance of data transformation during the exploratory phase of data analysis. Finally, the end of scientific investigation is to communicate the findings as effectively as possible. In this regard, good graphic

design is almost as critical as good data analysis. On recognizing the power and beauty of EDA data transformation, confirmatory data analysis, and graphical data display, we may find ourselves agreeing with George Bernard Shaw who said that, "It is the mark of a truly intelligent person to be moved by statistics."

ACKNOWLEDGMENTS

This chapter is written from the perspective of a biologist, with data to analyze and present, not a professional statistician. I am deeply indebted to my heroes of data analysis, John Tukey, Leland Wilkinson, and Edward Tufte, who are all statisticians.

REFERENCES

1. Tukey, JW. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, MA, 1977.
AQ27 2. Student. The probable error of a mean. *Biometrika* 6: 1–25, 1908.
3. Ruxton, GD. The unequal variance t-test is an underused alternative to student's t-test and the Mann–Whitney U test. *Behav. Ecol*. 17(4): 688–690, 2006.
4. Popper, K. *The Logic of Scientific Discovery*. Routledge Classic, London, 2002.
5. Shaffer, JP. Multiple hypothesis testing. *Annu. Rev. Psychol*. 46: 561–584, 1995.
6. Hoffmann, TK, Donnenberg, AD, Finkelstein, SD, Donnenberg, VS, Friebe-Hoffmann, U, Myers, EN, Appella, E, DeLeo, AB, Whiteside, TL. Frequencies of tetramer + T cells specific for the wild-type sequence p53(264–272) peptide in the circulation of patients with head and neck cancer. *Cancer Res*. 5;62(12): 3521–3529, 2002.
7. Triulzi, DJ, Meyer, EM, Donnenberg, AD. Leukocyte subset analysis of leukoreduced platelet components. *Transfusion* 40: 771–780, 2000.
8. Wilkinson, L. *SYSTAT, Version 9*. SPSS Inc., Chicago, IL, 1999.
9. Tufte, ER. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983, 117p.
10. Schellekens, P, Eijsvoogel, VP. Lymphocyte transformation *in vitro*: Tissue culture conditions and quantitative measurements. *Clin. Exp. Immunol*. 3: 571–584, 1968.
11. Robinson, MA. The human T cell receptor beta-chain gene complex contains at least 57 variable gene segments. Identification of six V beta genes in four new gene families. *J. Immunol*. 46(12): 4392–4397, 1991.
12. To, LB, Haylock, DN, Simmons, PJ, Juttner, CA. The biology and clinical uses of blood stem cells. *Blood* 89: 2233–2258, 1997.

QUERY FORM

BOOK TITLE:	**1984**
CHAPTER NO:	**CH002**

Queries and / or remarks

Query No.	Query / remark	Response
AQ1	“Below” has been changed to “Section 2.3.” Please confirm if it is okay.	
AQ2	“HIV ELISA” has been expanded as “human immunodeficiency virus Enzyme-Linked ImmunoSorbent Assay.” Please confirm.	
AQ3	Please provide the Section number or specify exactly where instead of “below.”	
RQ4	significance	
RQ5	since the “p-value actually deflates	
RQ6	reference?	
RQ7	reference?	
AQ8	Is the acronym "G-CSF" well known? If not, please provide the expansion at the first occurrence.	
RQ9	reference?	
AQ10	”Up to and *n* of 300” is not clear. Please clarify.	
RQ11	which is?	
RQ12	which is?	
RQ13	show equation	
AQ14	Please provide the section number instead of “below” or the specify exactly where instead of “below.”	
AQ15	“The third assumption” has been changed to “The second assumption.” Please confirm.	
RQ16	isn’t it the highest?	
AQ17	The sentence “can we ... is given? is not clear. Please clarify.	
AQ18	Please confirm if the deletion of “99.9%” in the sentence “Using an ... positive samples.” is okay.	
RQ19	do you want to show calculation for ppv and npv?	
AQ20	Please confirm if the change made in the sentence “Sometimes the ... serum immunoglobulin.” is okay.	
AQ21	Is the acronym “MHC” well known? If not, please provide the expansion at the first occurrence.	
RQ22	is there a calculation here that a novice would use to determine n?	
AQ23	Should this be “1.55:1 and 1.33:1?” Please check.	
AQ24	Please specify the Chapter number.	
AQ25	The sentence “how ... assymetrical?” is not clear. Please clarify	
AQ26	Can the section head “closing remarks” be replaced with “Conclusion.” Please confirm.	
AQ27	Should the name “W.S. Gosset” be inserted in parentheses after “Student.” Please advice.	