PROBABILITY

Probability is widely applied to measure risk associated with disease and drug therapy. *Risk factor* is a condition or behavior that increases the chance of developing disease in a healthy subject over a period of time. An example could be the risk of developing lung cancer over time with tobacco smoking. Another example of risk factor is the possibility of developing hearing loss after receiving an aminoglycoside antibiotic for a period of time.

The term *relative risk* (RR) or *risk ratio* is the most frequently used probability term to measure association of risk with exposure (to a drug or a behavior). Risk factors may be genetic, environmental, or behavioral. They have important implications in both pharmacokinetics and drug therapy (see). Risk factors may be casual or merely a marker that increases the probability of a disease.

$$RR = \frac{\text{risk}_{(\text{exposed})}}{\text{risk}_{(\text{unexposed})}}$$

$$RR = \text{relative risk} = \frac{\text{disease probability of exposed}}{\text{disease probability of unexposed}}$$

Often risk information is collected in a controlled manner over a period of time by survey, either from the past or forward in time. In a *prospective cohort study* (also known as a *cohort study, prospective, follow-up,* or *longitudinal study*), a cohort of healthy subjects exposed to different levels of a suspected risk is followed forward in time to determine the incidence of risk in each group. For example, in a hypothetical study, the risk of thrombophlebitis was studied in a group of randomly selected women: 500 women taking and 500 matched women not taking a birth control pill for 10 years. The RR of 10 for thrombophlebitis was calculated from a risk of thrombophlebitis in women exposed to the birth control pill versus women not exposed to the drug using a dichotomous 2 x 2 table as shown in , where A and B are the number of subjects who developed thrombophlebitis. The probability of *exposed* in this case is A/(A + B) and that of *not exposed* is C/(C + D). In this case, assume the exposed risk is 0.025 and the not-exposed risk is 0.0025. Then RR = 0.025/0.0025 = 10. Thus, the relative risk of thrombophlebitis in women on the birth control pill for 10 years is 10 for this group of women studied.

Table A.1 Tabulation from a Hypothetical Cohort Study of Female Subjects on the Pill—with and without Developing Thrombophlebitis for 10 Years

	Positive Outcome (+Thrombophlebitis)	**Negative Outcome (–Thrombophlebitis)**	**Subtotal**
Risk factor present (eg, taking the pill —exposed)	**A**	**B**	**A + B**
Risk factor absent (eg, not taking the pill—not exposed)	**C**	**D**	**C + D**
Subtotal	**A + C**	**B + D**	

A second method of studying risk is the *historical* or *retrospective cohort study,* which looks backward in time to determine the present risk. The cohort of exposed and unexposed subjects is retrieved from past records to determine risk outcome. In the next section relative risks are described in terms of *odds,* a similar concept that is also widely applied.

ODDS

The probability of drawing an ace from a deck of cards is 4/52, or 1/13 for a deck of cards containing 4 aces in 52 cards. The odds of drawing an ace is the number of times an ace will be drawn divided by the number of times it will not be drawn. The odds are

$$\text{ODDS} = \frac{4/52}{48/52} = \frac{4}{48} = \frac{1}{12}$$

This can be read as a 1:12 odds of drawing an ace. The absence of four aces in the denominator makes the difference in the odds outcome. Odds are numerically not equal to probability as defined.

The point may be illustrated by considering the opening of a standard deck of cards, one card at a time. For example, we may encounter 4 aces after 40 cards are opened, before the entire deck is open. We can see that the number of cards opened (we stopped at 40 cards) becomes a factor in the odds obtained. In this case, after opening the first 40 cards, 4 cards were aces and 36 cards were not aces. The odds are 4/36 = 1/9 instead of 1/12 as calculated for 52 cards. Using this analogy, it is inappropriate to pick sample sizes that do not reflect the natural risk course of the disease or drug treatment involved. If we decide to sample only 40 cards, stop sampling, and then calculate the odds (or RR in observing a disease), the results will be in error. In statistics, the sample size and how samples represent the population at large are important considerations for accurate

determination of the RR of a disease or drug treatment.

A common approach to studying risk outcome is the *case-control study* (also known as the *retrospective study*). In the case-control study, the exposure histories of two groups of subjects are selected on the basis of whether or not they develop a particular disease (eg, thrombophlebitis), in order to evaluate disease frequency resulting from drug (eg, the pill) exposure. The investigator selects the size of the subject population that has the disease or is disease free to determine exposure to the risk factor. The number of subjects who do and do not have the disease may not necessarily reflect the natural frequency of the disease. It is therefore improper to compute a "relative risk (RR)" from the odds ratio (OR) for a case-control study, because the investigator can manipulate the size of the relative risk.

$$\text{OR} = \frac{\text{odds of case exposure}}{\text{odds of case unexposed}}$$

If the disease is rare, then OR ≈ RR. When the sample size is large, the difference between OR and RR diminishes.

In statistical analysis, it is important to guard against *selection error* or *bias.* Investigators may look harder for cancer, for example, in smokers than in a control group of healthy subjects. The resultant disparity is often called *surveillancebias.* In a case-control group there may also be *recall bias;* for example, medical history taken on surgical lung cancer (case) may be more likely to contain information on smoking than other type of surgical controls ().

EXPERIMENTAL DESIGN AND COLLECTION OF DATA

Statistics have important applications in scientific studies, whether in studies involving hypothesis testing or in finding ways to improve a product. Statistical design is widely used at the experimental planning stage. Later, when data are collected, statistical methods are applied for data analysis and to help draw conclusions from the studies.

Experimental design may be simple or may involve an elaborate model. The method may be applied to optimize a drug or drug product based on a set of criteria (). Experimental design may be used to optimize an analytical method to separate a drug from impurities, such as a HPLC method. In pharmacokinetics, experimental design is used to design better sampling time for drawing blood samples for drug analysis in pharmacokinetic parameter estimation. A common approach to optimizing a drug product is the *factorial design*. For example, we may be interested in determining whether 0, 0.25, 0.5, 0.75 or 1% of magnesium stearate should be formulated into a tablet granulation to allow adequate flow of the tablet granulation mixture during manufacturing. This problem may be viewed as a simple one-factorial-design experiment to determine the amount of lubricant needed to provide best powder flow (1 factor x 5 lubricant levels). The object of the experimental design is to try to pick the tablet lubricant level that will result in the optimal powder flow from the hopper to the die cavity during tablet compression.

In practice, tablet granulation flow is more complex. Moisture level, particle size, and tackiness (of the drug substance) are other factors that influence flow. We may decide to reduce the increment level and select the lubricant levels to 0, 0.5, and 1% only. This provides three lubricant levels—high, medium, and low—for each factor. We then can apply a factorial design of 4 factors at 3 levels, which involves $4^3 = 64$ trials in the above case to generate a geometric response space that represents all the factors involved. The full factorial design is tedious, so a reduced design, the *fractional factorial* design, is often used. How to reach the optimal point efficiently when several factors are involved is a problem for many optimization experiments.

The identical concept for optimizing flow can be applied to optimizing the media composition supporting antibiotic production by fermentation using *Streptomyces* or a new microbe engineered by recombinant DNA. The factorial design may be employed to find the optimum composition for the growth medium. Factorial design yields knowledge about the system but is tedious. An alternative approach to carry out the optimization is the *sequential simplex method,* which optimizes the factors through sequentially planned experiments. This method is often applied in parameter estimation from data using computerized iteration algorithms. In designing experiments, it is important to know the factors involved and the range of each factor. How much change in drug concentration is occurring in plasma samples over an hour or in a minute? How much difference in drug concentrations can the analytical method detect? Good design requires some knowledge of the system and a strategy to obtain data for analysis that saves time and resources.

The same principle applies to human clinical studies. In clinical trials, studies are often done with a limited number of subjects, due to either cost or the availability of subjects who meet the study requirements. Based on good clinical practices, the study subjects are selected according to exclusion and inclusion criteria that are written into the protocol. All subjects must give informed consent to be in the study. Since most studies are done over a period of time, it is important to ensure that both the treatment and control groups are balanced and to avoid any temporal influence. For example, in bioequivalency trials, adequate time (*wash-out time*) between study dosing periods is allowed for the drug to be eliminated from the subject and to avoid residual effects due to carryover of the drug from the first dosing period to the next dosing period.

All scientific studies must be designed properly to obtain valid conclusions that may be applied to the population intended. The experimental design of a study includes the following:

1. A clearly stated hypothesis for the study
2. Assurance that the samples have been randomly selected
3. Control of all experimental variables
4. Collection of adequate data to allow experimental testing of the hypothesis

For example, we may wish to test the hypothesis that the average weight of young males is greater than the average weight of females in the United States. First, we may decide that young male and female subjects aged 18 to 24 will be selected and other ages excluded. The subjects are randomly selected from a pool of subjects who are not interrelated in a way that might affect

their body weights. We may want to exclude subjects with certain diseases (exclusion criteria). A sufficient number of subjects must be selected (sampled) randomly so that the total number of subjects (sample size) represents the general population in the United States. The need for randomization is easily understood but often poorly met because of the difficulties in recruiting subjects, or in the methods used for recruiting. For example, if all the subjects in this example are randomly recruited from one health club in a given city, the samples will not be typical of the population intended even though the subjects are randomly selected. Are we too ambitious to include the population of the entire United States? Second, many of the subjects exercise at the health club to lose weight, and this may not be representative of the general population. A true sample is one selected randomly from the population of the entire country without connection by any variable that affects their weights.

The identification of all covariates in a study is generally difficult and requires thorough consideration. The subject of sample size, inclusion criteria, and exclusion criteria are major considerations in experimental design that will affect the statistical outcome. After careful consideration, we may realize that there are many variables to be considered and may wish to modify the scope to tailor the study objectives more efficiently.

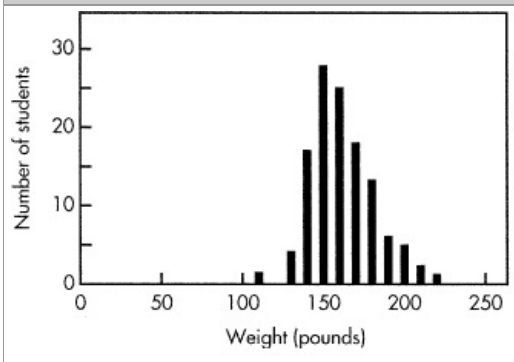
Age, gender, genetic background, and health of the subjects are important variables in clinical drug testing. The statistical design of a study is based on the study objectives. Each study should have clearly stated objectives and an appropriate study design indicating how the study is to be performed. Often, the population may be subdivided according to the objectives of the study. For example, a new drug for the treatment of Alzheimer's disease in the elderly may initially be tested in male subjects aged 55 and above. Later, clinical studies might test the drug in other patient populations. Many different statistical designs are possible. Some of these designs control experimental variables better than others. Specific statistical designs are given in other chapters and in standard statistical texts.

The quality of the data is very important and may be controlled by the researcher and the method of measurement. For example, if the weights of the young males and females in the example above are obtained using different scales, the investigator must ascertain that each scale weighs the subject accurately. *Accuracy* refers to the closeness of the observation (ie, observed weight) to the actual or true value. *Reproducibility* or *precision* refers to the closeness of repeated measurements.

ANALYSIS AND INTERPRETATION OF DATA

The objective of data analysis is to obtain as much information about the population as possible based on the sample data collected. A common method for analyzing data of a sample population is to classify the data and then plot the frequency of occurrence of all the samples. For example, the frequency of weight distribution of a class of students may be plotted in the form of a histogram that relates frequency to weight ().

Figure A-1.



Weight distribution of 120 students.

An important observation in this example is that the weight of most students lies in the middle of the weight distribution. There is a common tendency for most sample values to occur around the mean. This is described in the *central limit theorem,* which states that the frequency of the values of measurements drawn randomly from a population tends to approximate a *bell-shaped curve* or *normal distribution*. Extensive data collection is needed to determine the distributional nature of a sample population. Once the parameters of a distribution are determined, the probability of a given sample's occurrence in the population may be calculated.

DESCRIPTIVE TERMS

Descriptive terms are used in statistics to generalize the nature of the data and provide a measure of central tendency. The *mean* or *average* is the sum of the observations divided by the number, *n*, of observations (). The *median* is the middle value of the observation between the highest and lowest value. The *mode* is the most frequently occurring value. The term *range* is used to describe the *dispersion* of the observations and is the difference between the highest and lowest values. For data that are distributed as a normal distribution (discussed below), the mean, median, and mode have the same value.

Table A.2 Descriptive Statistics for a Set of Data[a]

Data		Descriptive Terms
21	63	Sum = 1274
25	67	Mean = 57.9
29	67	Mode = 67
35	67	Median = 62
37	67	*n* = 22
42	67	Range = 21–91
45	72	SD = 20.3
49	73	RSD = 35.1%
56	75	
57	88	
61	91	

[a] The data represent a set of measurements (observations) in study. The descriptive terms are often used to describe the data. Each term is defined in the text.

SD, standard deviation; RSD, relative standard deviation.

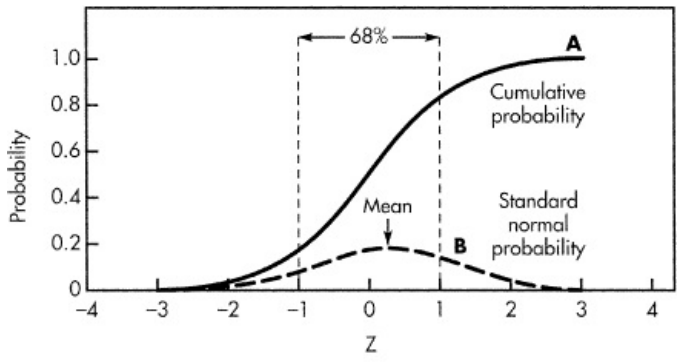
THE NORMAL DISTRIBUTION

If data are plotted according to the frequency of occurrence, a pattern for the distribution of the data is observed. Most data approximate a normal or *Gaussian distribution*. The *normal distribution* is a bell-shaped curve that is symmetric on both sides of the mean. Statistical tests that assume the data follow normal distribution patterns are known as *parametric tests. Nonparametric tests* do not make any assumption about the central tendency of the data and may be used to analyze data without assuming normal distribution. Nonparametric tests require no assumption of normality and are less powerful. Examples are the Wilcoxan test and the side test.

The shape of the normal distribution is determined by only two parameters, the population mean and the variance, both of which may be estimated from the samples. The *variance* is a measure of the spread or variability of the sample. Many biologic and physical random variables are described by the normal distribution (or may be transformed to a normal distribution). These may include the weight and height of humans and animal species, the elimination half-lives of many drugs in a population of patients, the duration of a telephone call, and other variables. In statistics, the item investigated is termed the *random variable*. For convenience, the standardized normal distribution is introduced to allow easy probability calculation when the standard deviation is known (). The probability of a sample value occurring from 1 *standard deviation* (SD) above to 1 SD below the mean is 68% (*z* of −1 to +1). This value is calculated by finding the probability corresponding to *z* = −1 and z = 1 from curve B in as follows:

- Probability between *z* of −4 to −1 is 0.16.
- Probability between *z* of −4 to +1 is 0.84.
- Therefore, the probability between *z* of −1 to +1 is 0.84 − 0.16 = 0.68 or 68%.

Figure A-2.



Source: Shargel S, Wu-Pong S, Yu ABC: *Applied Biopharmaceutics & Pharmacokinetics*, 5th Edition: http://www.accesspharmacy.com

Probability of a normal distribution. (A = cumulative probability, B = standard normal probability.)

The area representing probability between any two points on the normal distribution is calculated from this graph. In practice, a cumulative standardized normal distribution table is used to allow better accuracy.

Standard deviation measures the variability of a group of data. SD for *n* number of measurements is calculated according to the following equation:

$$SD = \sqrt{\frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1}} \qquad (A.1)$$

where $\bar{X}$ is the mean, x_i is the observed value, and n is the number of observations (data). The standard deviation is often calculated by computer or calculator and gives an indication of the spread of data (). A larger standard deviation indicates that the spread of data about the mean is larger compared to data with the same mean but with a smaller standard deviation.

Relative standard deviation or *coefficient of variation* allows comparison of the variance of measurements. The standard deviation is divided by the mean to give the relative standard deviation (RSD) or coefficient of variation (CV):

$$RSD = \frac{SD}{\bar{x}} \qquad (A.2)$$

The RSD may be expressed as a percent or %CV by multiplying the RSD by 100. This is commonly known as *percent standard deviation* or *percent variation*.

The difference between the mean, $\bar{X}$, and each observed datum, x_i , is the deviation from the mean. Because the deviation from the mean can be either negative or positive, the deviations are squared and summed to give an estimation of the total spread or deviation of the data from the mean. The term $\Sigma^n{}_i (x_i - \bar{X})^2$ is the *sum of the squares.* This term incorporates measurement error as well as inherent variance of the samples. If a single sample is measured several times, the sum of the squares should be very small if the method of measurement is reproducible. The concept of least squares for minimizing error due to model fitting is fundamental in many statistical methods.

CONFIDENCE LIMIT

If normal distribution of the data is assumed, the probability of a random variable in the population can be calculated. For example, data that falls within 1 SD above and below the mean ($\bar{X}$ ± 1 SD) represents approximately 68% of the data, whereas data that falls within 2 SD above and below the mean ($\bar{X}$ ± 2 SD) represents approximately 95% of the data. In the examples below, the random variable in which we are interested is the diameters of drug particles measured from a powdered drug sample lot.

Example

The particle size of a powdered drug sample was measured. The average (mean) particle size was 130 µm with a standard deviation of 20 µm.

1. Determine the range of particle sizes that represents the middle 68% of the powdered drug.

Solution

From a normal distribution table or , 68% of the middle particles represent 34% above and below the mean, corresponding to the mean ±1 SD.

Small particle size = mean – SD = 130 – 20 = 110 µm

Large particle size = mean + SD = 130 + 20 = 150 µm

Therefore, 68% of the particles will have a particle size ranging from 110 to 150 µm.

2. Determine the range of particle sizes that represents the middle 95% of the powdered drug.

Solution

95% ÷ 2 or 47.5% on each side of the mean, corresponding to ±2 SD ()

Smallest particle size = 130 – (2 x 20) = 90 µm

Largest particle size = 130 + (2 x 20) = 170 µm

Therefore, 95% of the particles will have a particle size ranging from 90 to 170 µm.

In the above example, the calculation shows that most of the particles lie around the mean. To be 95% certain, simply extend from the mean ±2 SD. This approach estimates the 95% confidence limit. A 95% confidence limit implies that if an experiment is performed 100 times, 95% of the data will be in this range above and below the mean.

This example shows how to reconstruct a population based on the two parameters, mean and variance (approximated by the SD). A more common application is the estimation of experimental data such as assay measurements. Such 95% confidence limits are often calculated from the standard deviation to estimate the reliability of the assay measurement.

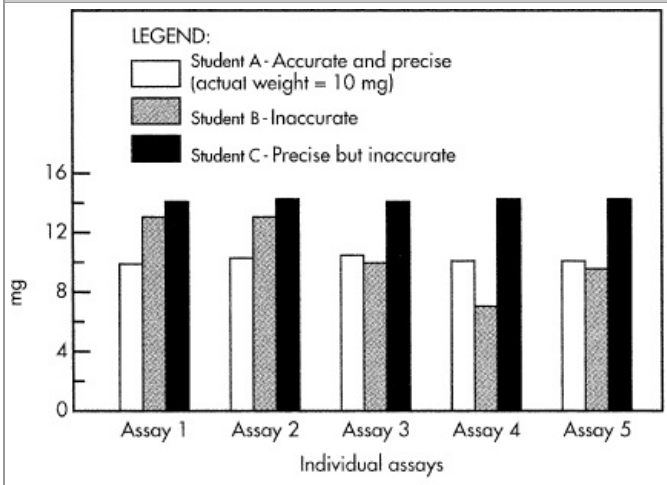
In the example above, the mean for the particle size was 130 µm and the SD was 20 mm. Therefore, from equation A.2, the RSD = 20/130 or 0.15. The RSD may be expressed as a percent or %CV by multiplying the RSD by 100.

As mentioned, *accuracy* refers to the agreement with the observed value or mea-surement in a group of data and the actual or true value of the population. Unfortunately, the true value is unknown in many studies. The term *precision* refers to the reproducibility of the data or the variation within a set of measurements. Data that are less precise will demonstrate a larger variance or a larger relative standard deviation, whereas more precise data will have a smaller variance.

Example

A lot of 10-mg tablets was assayed five times by three students (). Which student assayed the tablets most accurately?

Figure A-3.



Source: Shargel S, Wu-Pong S, Yu ABC: *Applied Biopharmaceutics & Pharmacokinetics*, 5th Edition: http://www.accesspharmacy.com

Comparison of assays (5 each) of 10-mg tablet by three students.

Solution

The mean and SD of assays by each student were determined. Because the same lot was assayed, the difference in SD among the students is attributed to assay variations. Student A is closest to the target—that is, the labeled claimed dose (LCD) of 10 mg. Student C is most precise (with the smallest %SD), but is consistently off target.

The data obtained by Student C is considered biased because all the observed data are above 10 mg. Data are also considered biased if all the observed data are below the true value of 10 mg.

Bias refers to a systematic error when the measurement is consistently not on target. Repeated measurements may be very reproducible (precise) but miss the target. In the example above, Student C was most precise, but Student A was most accurate. In determining accuracy and precision, a standard (known sample) is usually prepared and assayed several times to determine the variation due to assay errors. In the example above, we assumed that the students used known 10-mg standard tablets. If the tablets were unknown samples, it would not be possible to conclude which student was more accurate, because the true value would be unknown. In practice, assay methods are validated for precision and accuracy based on known standards before unknown samples (eg, plasma samples) are assayed.

In the analysis and interpretation of data, statistics makes inferences about a population using experimental data gathered in a sample. After analysis of data, the statistician calculates the likelihood or probability that a given result would happen. *Probability* (*P*) is the fraction of the population indicating that a given result or event would occur by random sampling or chance. For example, if $P < 0.05$, then the likelihood that a result occurs by random sampling is 5/100, 1/20, or 5%. By convention, if the statistical inference produces a *P* value of 0.05 or less, it is considered atypical or uncommon of the population. As shown in , the probability of finding a student weighing above 250 lb is small ($P < 0.05$), and we may conclude (somewhat erroneously) that a student who weighs 250 lb is significantly different from the rest. This concept for determining the probability of how typical a given sample value occurs in a population may be extended to hypothesis testing. Hypothesis testing estimates the probability of whether a given value is typical of the control group or of the treated group.

STATISTICAL DISTRIBUTIONS AND APPLICATION

The frequency distribution of some data does not appear to be symmetrically shaped. The term *skewness* relates to the asymmetry of the data. The data distribution may be skewed to the left or right of the mean. In many pharmacologic studies, the sample size in the study is small, and the investigator cannot always be certain that the data obtained from the study are normally distributed. An incorrect assumption of a normal distribution may lead to a biased conclusion. In such cases, a nonparametric test may be used, because it does not make any assumption about the underlying test except that it may be continuous.

During data analysis, a value or observation may be observed that is several standard deviations above or below the mean; such a value is called an *outlier*. A value that is an outlier is difficult to use statistically. An observed value that deviates far from the majority of the data may indicate non-normal distribution, an error in measurement, or an error in data entry. If an error is found during checking, it should be corrected. In general, outlier values should not be excluded from the statistical analysis. Some investigators use log transformation of the data to make the distribution of the data appear to be more normally distributed. A geometric mean is obtained after log transformation of the data. In some cases, with sufficient data collection, a bimodal distribution may be observed. For example, the acetylation of isoniazid in humans follows a bimodal distribution, indicating two populations consisting of fast acetylators and slow acetylators. In this study, if the data were obtained from subjects of whom all but one were fast acetylators, then the single datum for the slow acetylator might be considered an outlier (and possibly be discarded from the data analysis).

When a specific distribution is not known, it may be possible to use the bootstrap/jackknife method. A *bootstrap* is a paradoxical means of getting started on something when you need some of that something in order to get started. The concept derives from the phrase "to pull oneself up by one's bootstraps." Rather than attributing an assumption to the distribution, the actual data collected will be used to extrapolate further about the larger population it represents. Instead of collecting more samples by actual experiment, one simply puts all the data into a "virtual bag" and randomly samples (not actually taking the original data out) from the bag until the desired number of samples is collected to yield a glimpse of what the population will be like. The method was a great innovation by Bradley Efron and is easily adapted to many applications with computer technology. The basic bootstrap method is essentially the same. The method has worked well and has been modified for many practical applications where assumption of specific distribution failed (). It should be appreciated that the nature of the data from an original study is new and unknown. The greatest difficulty facing the investigator is characterizing the distribution and applying it to a proper model. The bootstrap method avoids the problem of assuming the wrong distribution. In , statistical distribution is discussed with further application to pharmacokinetics.

The normal, binomial, Poisson, chi-square distributions are frequently applied in statistics and engineering. Some distributions are related mathematically. References for further information are listed at the end of this appendix.

Statistical analysis is referred to as a statistical *test* when the data are formulated for hypothesis testing. The most common test involves the *Student's t-test,* which is a test of means of two groups assuming the same variance. If the variances are different, the Student's *t*-test will be a *t*-test with unequal variance. When the sample size increases to very large, the *t*-distribution approaches the normal distribution. The *t*-test becomes more powerful if the subjects meet the criteria for a paired *t*-test. The *F*-test is a simple test of variance of two groups. When more groups are involved, analysis of variance may be applied.

Statistics are broadly applied in pharmacokinetics. Special statistical methods have been developed to provide a platform for rational argument or decision making in accepting or rejecting a theory. The information generated in many controlled clinical studies is often sparse due to limited sample size (eg, few subjects) and limited sampling (eg, few blood samples). For this reason, all drugs under development in clinical studies are monitored for side effects and rare events. After FDA approval and marketing, pharmacovigilance is needed to assure safety in the larger population of patients who will be using the drug.

In testing the effect of a drug across different groups, the term *interaction* is often heard. An *effect* is a difference in treatment response, and an *interaction* is a difference in differences. Examples might be treatment-by-center interactions, or treatment-by-gender interactions. When large interactions are noted, it is not appropriate to combine groups and make an overall assessment of the treatment effect statistically. Clinical trials often involve pivotal and supporting studies. An analysis that combines multiple studies to obtain an overall result, such as an overall estimate of the size of the treatment effect of a drug, is termed *meta-analysis*.

HYPOTHESIS TESTING

Hypothesis testing is an objective way of analyzing data and determining whether the data support or reject the hypothesis postulated. For example, we might want to test the hypothesis that a given steroid causes a weight increase. We want to test this hypothesis using two groups of healthy volunteers, one group (treated) that took the given steroid and another (control) that took no drug. The two hypotheses generated are as follows:

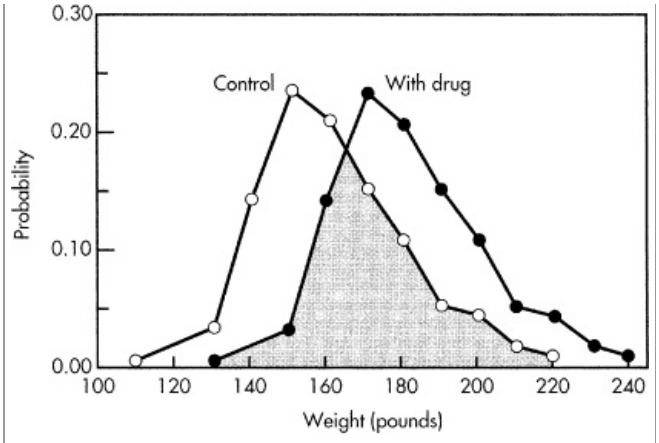
H_0: There is no difference in weight between the treated and control groups (null hypothesis).

H_1: There is a difference in weight between the treated and control groups (alternative hypothesis).

The null hypothesis H_0, states that there is no difference between the treatments. The experimental data will either reject or fail to reject (accept) the null hypothesis. If the null hypothesis is rejected, then the alternative hypothesis is accepted, since there are only two possibilities. A simple hypothesis testing data from two groups is the two-sample Student's *t*-test involving a control group and a treated group (the *t*-distribution approximates the normal distribution and is commonly used).

The data for the study (simulated) is shown in using 120 students in the control and treated groups. The mean weight of the treated group is about 175 lb, whereas the mean weight for the control group is about 155 lb. There is a shift to the right in the weight distribution of the treated group. However, a considerable overlapping (shaded area) in weights is observed, making it difficult to reject the null hypothesis. In practice, H_0 is rejected at a known level of uncertainty called the *level of significance*. A level of 5% ($\alpha = 0.05$) is considered statistically significant, and a level of 1% ($\alpha = 0.01$) is considered highly significant. Commonly, this level of significance is reported as $P < 0.05$ or $P < 0.01$, respectively, to indicate the different levels of significance. Because uncertainty is involved, whenever the null hypothesis is rejected or accepted, the level of significance is stated. A significance level of 25% ($P < 0.25$) in the above example suggests that there is a 25% probability that the weight change is not due to drug treatment. A 25% probability is a level far too large to reject the H_0 with certainty. The level of significance is therefore related to the probability of incorrectly rejecting H_0 when it should have been accepted. This level of error is called *Type I error*. Whenever a decision is made, there is the possibility of making the wrong decision. Four possible decisions for a statistical test may be made (). A *Type II error* is committed when H_0 is accepted as being true when it should have been rejected. In contrast, a Type I error is committed when H_0 is rejected when it should have been accepted.

Figure A-4.



Hypothesis testing example.

Table A.3 Decisions Based on a Statistical Test

	Accept Null Hypothesis (H_0)	**Reject Null Hypothesis (H_0)**
H_0 true	Correct decision	Type I error
H_1 true	Type II error	Correct decision

The probability of committing a Type I error is defined as the significance level of the statistical test, and is denoted as P or α (alpha). The probability of a Type II error, denoted as β (beta), can also be computed.

The *power test* determines the probability that the statistical test results in the rejection of the null hypothesis if H_0 is really false. The larger the power, the more sensitive are the tests. Because power is defined as $1 - \beta$, the larger the β-error (Type II error), the weaker is the power. The power of the test would equal $1 - \beta$ for a particular power.

To reduce Type I or Type II errors, the sample sizes need to be increased or the assay method improved. Because time, expense, and ethical concerns for performing a study are important issues, the investigator generally tries to keep the sample size (usually the number of subjects in a clinical study) to a minimum. The variability within the samples, number of samples (sample size), and desired level of significance will affect the power of the statistical test. Usually, the greater the variability within the samples, the larger will be the sample size needed to obtain sufficient power.

ANALYSIS OF VARIANCE (ONE-WAY)

When more than two data sets are compared, analysis of variance (ANOVA) is used to determine the probability of the data sets being identical or different among groups. *One-way* analysis of variance is a method for testing the differences between the population means of k treatment groups, where each group i ($i = 1, 2, \ldots, k$) consists of n_i observations X_{ij} ($j = 1, 2, \ldots, n_i$).

For example, we may want to test whether there is a difference in the peak plasma level of a drug resulting from the administration of three different dosage forms—solution, capsule, and tablet. If we decide to have three groups of 20 patients per group, $i = 1, 2, 3$, $j = 1, 2, \ldots, 20$, and the observation in each group will be X_{ij}. The formulas for calculation are as follows:

$$\text{Sum of observation in group} = \sum_{j}^{n_i} x_{ij}$$

$$\text{Total sum of squares SS} = \sum_{i}^{k} \sum_{j}^{n_i} x_{ij}^2 - \frac{\left(\sum_{i}^{k} \sum_{j}^{n} x_{ij}^2 \right)}{\sum_{i}^{k} n_i}$$

$$\text{Total sum of squares TSS} = \sum_{i}^{k} \sum_{j}^{n_i} x_{ij}^2 - \frac{\left(\sum_{i}^{k} \sum_{j}^{n_i} x_{ij}^2 \right)}{\sum_{i}^{k} n_i}$$

$$\text{Error sum squares (ESS)} = \text{SS} - \text{TSS} = \text{total sum of squares} - \text{treatment sum of squares}$$

The value of the *F*-statistic is

$$F = \frac{DF_2 \times TSS}{DF_1 \times ESS} = \frac{TSS/DF_1}{ESS/DF_2}$$

Alteratively, *F* is expressed as the ratio of treatment mean squares (TMS)/error mean squares (EMS):

$$F = \frac{TSS/DF_1}{ESS/DF_2} = \frac{TMS}{EMS}$$

$$\text{Error degrees of freedom } DF_1 = k - 1$$

$$\text{Error degrees of freedom } DF_2 = \sum_{j=1}^{k} n_i - k$$

The ANOVA employed depends on the objectives and design of the study. ANOVA methods can estimate the variance among different subjects (intersubject variability), groups, or treatments. Using ANOVA, the statistician determines whether to accept or reject the null hypothesis (H_0), deciding whether there is no significant difference (accept H_0) or there is a significant difference (reject H_0) between the data groups.

ANOVA is a test for a difference in means between two or more groups. It is very similar to the *t*-test, which tests for a difference between two groups. It is applicable when several groups are involved. An ANOVA on two groups is analogous to as a two-sample *t*-test. ANOVA assumes that the groups have equal variances and that the outcome is normally distributed within each group. These assumptions are less important when one has large samples and equal sample sizes for each group (as with the *t*-test). If one rejects the null hypothesis in the ANOVA, one can conclude that the groups are not all equal, but the test does not yield any information on each pair of comparisons.

After completion of the study and statistical analysis of the data, the investigator must decide whether any statistically significant differences in the data groups have clinical relevance. For example, it may be possible to demonstrate that a new antihypertensive agent lowers the systolic blood pressure in patients by 10 mm Hg and that this effect is statistically significant ($P < 0.05$) using the appropriate statistical test. From these results and statistical treatment of the data, the principal investigator must decide whether the study is clinically relevant and whether the drug will be efficacious for its intended use. An example of ANOVA appears in , .

In clinical tests involving a comparison of a drug to a control, it is important to establish some criteria in advance with the clinicians. What is the size of the difference (Δ) that is considered clinically meaningful? How large should the sample size be in order to have adequate statistical power? If the variance is large, a large sample may be needed. On the other hand, "Over-powering" may occur if the study sample size is so large that an extremely small treatment effect could be found statistically significant but not clinically useful. When two similar studies testing the treatment effect of a drug found different results and conclusions, as is sometimes reported in the literature, it is important to look for study flaws and biases. If an estimate is biased, it means that it is not accurately estimating the true value (true mean, true median, true standard deviation, etc). It may be due to a flaw in the study design, conduct, or analysis. Any of these can shift the results in favor of either the new drug or the control. Is the drug tested in one of the studies using an old standard of care? A new drug believed by most physicians to be superior can influence the result. Bias includes not blinding the study or having many dropouts, especially if the dropouts are due to the study drug. The consequence of the study bias may lead to a drug that may be found statistically better than the control but not clinically important. Clinical judgment should be used to evaluate the magnitude of the treatment and eliminate over-powering. This is especially important for superiority trials of new drugs. Most clinical trials for regulatory purposes are quality checked or audited to assure accuracy. On the other hand, the literature or information published, although useful, may not be adequately reviewed or does not always meet statistical design requirements. It is important for the pharmacist to check the source of information and determine whether test methods are validated.

POWER TEST

A Type I error may be observed when the result of an ANOVA rejects the null hypothesis when it should have been accepted. The power of a statistical hypothesis test provides a high level of certainty that the correct decision was made—-that is, to reject the null hypothesis when it is actually false. The power of a hypothesis test is the probability of not committing a Type II error. It is calculated by subtracting the probability of a Type II error from 1, usually expressed as

$$\text{Power} = 1 - \text{P (Type II error)} = 1 - \beta$$

The values for the power test range from 0 to 1. Ideally, the values for the power test should have a high power or value close to 1. To calculate the power of a given test, it is necessary to specify α (the probability that the test will lead to the rejection of the hypothesis tested when that hypothesis is true) and to specify a specific alternative hypothesis. Usually, α is set at 0.05. The power test is influenced by sample size and by intrasubject variability. For drugs whose bioavailability demonstrates high intrasubject variability (>30% CV), a larger number of subjects (larger sample size) is required to obtain a high power (>0.95).

BIOEQUIVALENCE STUDIES

Statistics have wide application in bioequivalence studies for the comparison of drug bioavailability for two or more drug products (see). The FDA has published two Guidance for Industry for the statistical determination of bio-equivalence () that describe the comparison between a test (T) and reference (R) drug product. These trials are needed for approval of new or generic drugs. If the drug formulation changes, bioequivalence studies may be needed to compare the new drug formulation to the previous drug formulation. For new drugs, several investigational formulations may be used at various stages, or one formulation with several strengths must show equivalency by extent and rate (eg, 2 x 250-mg tablet versus 1 x 500-mg tablet, suspension versus

capsule, immediate-release versus extended-release product). The blood levels of the drug are mea-sured for both the new and the reference formulation. The derived pharmacokinetic parameters, such as maximum concentration (C_{max}) and area under the curve (AUC), must meet accepted statistical criteria for the two drugs to be considered bioequivalent. In bioequivalence trials, a 90% confidence interval of the ratio of the mean of the new formulation to the mean of the old formulation (Test/Reference) is calculated. That confidence interval needs to be completely within 0.80 to 1.25 for the drugs to be considered bioequivalent. Adequate power should be built into the design and validated methods used for analysis of the samples. Typically, both the rate (reflected by C_{max}) and extent (AUC) are tested. The ANOVA may also reveal any sequence effects, period effects, treatment effects, or inter- and intrasubject variability. Because of the small subject population usually employed in bioequivalence studies, the ANOVA uses log-transformed data to make an inference about the difference of the two groups.

PHARMACOKINETIC MODELS

In data analysis involving a model, the number of data points should exceed the number of parameters in the model with a sufficient degree of freedom. Otherwise, the model is unconstrained and the parameters estimated are not valid.

REFERENCES

Deming SN, Morgan SL: *Experimental Design: A Chemometric Approach.* Elsevier, 1987

Efron, Bradley, Tibshirani RJ: *An Introduction to the Bootstrap.* New York, London, Chapman & Hall, 1993

FDA Guidance for Industry—Statistical Approaches to Establishing Bioequivalence. www.fda.gov/cder/guidance, 2001

FDA Guidance for Industry—Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design. www.fda.gov/cder/guidance, 1992

Knapp RG, Miller MC III: *Clinical Epidemiology and Biostatistics*. Lippincott Williams & Wilkins, 1992