Introduction

1

Statistics

The science of creating, developing, and applying techniques by which the uncertainty of inductive inferences may be evaluated.

Statistics encompasses the processes of collection, summary, and drawing inferences from data. In the mid-1800s, statistics was not unanimously accepted as a science. This was best exemplified in the 1830s when a vote to form a statistics section as part of the British Association for the Advancement of Science was conducted. A committee, chaired by Thomas Malthus, was asked to report on the question: "Is statistics a branch of science?" The committee members were in agreement that the collection and orderly tabulation of data was science. However, they were resolutely split on the question, "Is the interpretation of results scientifically respectable?" The issue was hotly contested and prompted the statistics group to form its own separate association in 1834 called the Statistical Society of London, later renamed the Royal Statistical Society. Its purpose was "the procuring, arranging and publishing of facts calculated to illustrate the conditions and prospects of society." Its first rule of conduct was to exclude opinion; deductions were to be based on data and mathematical demonstration. The avoidance of bias was exemplified in the Statistical Society of London's choice of seal and motto, a sheaf of wheat, and motto, *Allis exterendum* which means "let others thrash it out" (Figure 1.1) (Cochran, 1976; Hilts, 1978). The science of statistics evolved through International Congresses, the founding of the journal Biometrika (1901), and through research and education in the early 1900s at University College (London, UK) and later at Iowa State (Ames), North Carolina State (Raleigh), and the Indian Statistical Institute (Kolkata, India). Although it had a fractious beginning, the science of statistics is now universally adopted and is entrenched in the approach to, and the analysis of, data.

> *Why employ the science of statistics? . . . To provide a degree of confidence to inductive inferences that are based on a limited number of observations and for which an objective approach is needed.*

Application of Statistics

Florence Nightingale (1820–1910), who was born in Italy, is known for her efforts to improve the nursing profession, sanitary conditions, and the administration of hospitals. What is less known is that in order for her to achieve these results, it was necessary to employ statistics: namely, data collection, arrangement, objective analysis, presentation of the results, and inferences drawn. Nightingale had introduced an improved collection and standardization of reporting of health statistics. Utilizing these data, she was able to demonstrate that British soldiers stationed in Britain had twice the mortality rate of the general population (Table 1.1). What was even more surprising

Applied Statistics in Biology: A Practical Guide Using SAS, R, and JMP, First Edition. Stephen R. Bowley, Edzard van Santen, Simon Riley, Diane K. Michelson and Ruth M. Hummel.

FIGURE 1.1 A sheaf of wheat is a prominent feature in the seal of the Royal Statistical Society, London. The British society was incorporated by Royal Charter in 1887.

Table 1.1 Relative mortality statistics of British soldiers and of British males of comparable age, compiled by Florence Nightingale.

	Annual deaths per 1000 (1839–1853)			Annual deaths per 1000 (1856)	
Years of age	Civilians	Soldiers stationed in Britain	Cause of death	Civilians	Soldiers stationed in Crimea
20–25	8.4	17.0	Infectious disease	0.2	18.7
25–30	9.2	18.3	Constitutional disease	0.4	0.3
30–35	10.2	18.4	Localized disease	0.3	0.9
35–40	11.6	19.2	Violent death	0.1	3.0
Mean	9.9	18.2	Total, all causes	1.0	22.9

Soldiers stationed in Britain had nearly twice the mortality rate of the civilian population. Further examination revealed that infectious disease was the primary cause of the difference in mortality.
Source: Royal Sanitary Commission report (1858)/George Edward Eyre and William Spottiswoode/Public domain.

was her finding that, for the month of May 1856, soldiers stationed in Britain had higher annual death rates (18.7 per 1000) compared to British soldiers who served on the front lines during the Crimean war (8.0 per 1000). One would have expected the reverse—that death rates would have been higher for soldiers serving on the battlefront compared to those held in reserve. Dissection of the data revealed that soldiers stationed in the Crimea were six times more likely to die from infectious diseases than from violent deaths. A Royal Commission, which was established on 5 May 1857 to review the issue, included her findings in its report. Because of the patriarchal nature of that era and the fact that Nightingale was a woman, her evidence had to be presented indirectly to the Commission in written form. Nonetheless, her documentation and presentation of sound statistical data confirmed that living conditions were deplorable for British soldiers. This provided the impetus for social reform, which led to improved health conditions for both soldiers and the general population. Nightingale was elected a fellow of the Royal Statistical Society in 1858 and later became an honorary member of the American Statistical Association.

Scientific Method

Scientific knowledge is knowledge generated by an individual or team using scientific methods. The scientific method is a broadly used term denoting the principles of research and experimentation and the philosophic basis of these principles. In complex matters, opinions, theories, or assertions should be arrived at through the application of the scientific method and not be based on speculation or unsupported personal opinion. Confirmation bias is when only ideas or evidence that supports existing beliefs is searched for and retained, while contrary evidence is discounted or ignored. This is how misinformation, fueled by confirmation bias, spreads through social media (Del Vicario et al., 2015). Neutral assessment of all relevant information is required to avoid flawed decisions due to confirmation bias.

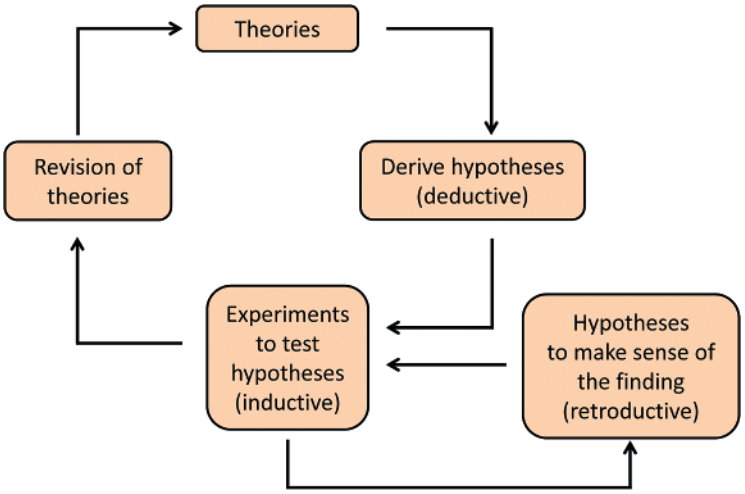


FIGURE 1.2 Outline of the scientific method illustrating the connections among the deductive, inductive, and retroductive processes. Hypotheses are developed based on existing knowledge, and experiments are used to challenge the hypotheses.

The scientific method, which is the way we refine, extend, and apply knowledge in all fields, is not a set of rigid steps. It involves the interplay of *inductive reasoning*, reasoning from specific observations and experiments; *deductive reasoning*, reasoning from theories to account for specific experimental results; and *retroductive reasoning*, to make sense of surprise findings. Figure 1.2 illustrates the connections among these three processes.

There is great flexibility in the combination and order of these paths. Discrete problems with a narrow focus may be solved using a single cycle, while complex or ill-structured problems may require skipping, back-tracking, overlapping, and other variations. Solutions that are identified should only be considered tentative. There are no certainties in science; it would be unreasonable to conclude that the problem was completely resolved. The solution will, at best, be a scientific theory, a logical explanation of observed events. Generation and testing of alternate hypotheses and in-depth challenge and test of scientific hypotheses are necessary before scientific theories can be converted to scientific laws: theories that have been widely tested and accepted as true.

A central process of the scientific method is the construction of a hypothesis, performing a series of experiments designed to test, not prove, the hypothesis, and, after examining the data from the experiments, drawing a conclusion. The conclusion may be one of three possibilities: (i) yes the data supported the hypothesis; (ii) no it did not; or (iii) the results are equivocal. The third possibility arises with underpowered experiments, which may provide visual or numerical indication of support, but the variability is too large to have confidence in the result. A key principle of the scientific method is that contrary evidence from well-designed and executed experiments must not be ignored. If the hypothesis was not supported, you must consider what may be wrong with it. This might lead to the formation of a hypothesis about a hypothesis.

Neyman and Pearson

The alternate hypothesis is one of the basic concepts in the Neyman–Pearson approach to scientific problems. Jerzy Neyman (1894–1981), whose parents were Polish, was born in Russia. During the boundary war between Poland and Russia, Neyman was arrested and was later relocated to Poland during a prisoner exchange. Fearful that he would again be interned as an enemy alien, he emigrated to England in 1934 to work at University College and four years later moved to the United States to work at the University of California. At Berkeley, Neyman's research was directed toward stochastic problems such as the study of carcinogens and the spread of epidemics. One reason he gave for transferring from University College, then the center of statistics, to California, which had no statistics unit, was that it was a place he could move his family that was as far away as possible from the developing war in Europe.

> Neyman's collaboration with Egon Pearson (1895–1980), much of which was through mail between Poland and England, provided the foundation for hypotheses tests. One of the concepts that they developed was that of the alternate hypothesis—one which might be valid if the hypothesis under test was not. There are two possible decisions regarding the null hypothesis: it is either rejected or accepted. Each of these decisions may be false, and the consequence of either mistake may have different importance. Neyman's research also led to the concept and use of confidence intervals. These ideas were widely adopted and are now commonly used by many researchers
>
> (Scott, 1997).

Statistical Null Hypothesis

The use of statistical tests for the analysis of experimental data enables an unbiased method to evaluate the evidence. It is a way of avoiding inferences such as "I believe these two means are different" or "I am pretty sure they are the same." A statistical test of significance replaces these biased statements with "Given this probability of uncertainty, these two means are different."

Do not confuse a science hypothesis with a statistical hypothesis. A science hypothesis is formulated based on all knowledge to date to explain how things work and is an active phrase such as "Given this scenario, this will happen." Experiments are then conducted to test this hypothesis. A statistical hypothesis is associated with a specific statistical test that is applied to evaluate the evidence from one or more experiments.

Statistical tests are based on a decision about the null hypothesis (H_0) that the statistic, such as a mean (μ_1), is some specified value (μ_0). This is usually expressed as H_0: $\mu_1 = \mu_0$, or that there is no difference between the two values. The hypothesis is either nullified or not nullified by the statistic test; this is why it is termed the null hypothesis. The alternate hypothesis (H_A) is one which is accepted if the null hypothesis is not.

In a Fisherian test of significance, there is no specified alternate hypothesis defined. The alternate is effectively an infinite range of possibilities other than the null value. The alternate hypothesis is therefore expressed as H_A: $\mu_1 \neq \mu_0$, or that there is a difference between the two values. With this approach, there is only one hypothesized model (μ_0).

In contrast, both the Neyman–Pearson and the Bayesian statistical tests involve a decision between two hypothesized models. In other words, the alternate hypothesis is a specific alternate value. There are two hypothesized models for the data: μ_0 and μ_1. The choice is between the null (μ_0) and the specific alternative (μ_1).

Application of a Bayesian statistical test requires knowledge of the prior distribution; such knowledge requires a lot of data. The prior probabilities are combined with the observed data to estimate the posterior probabilities of the null and the alternate hypotheses. The hypothesis with the greater posterior probability is accepted. A Bayesian test treats both the null and the alternate hypotheses on an equal footing. Neyman–Pearson tests are based on a likelihood ratio of the two alternate models; the likelihood of the null is divided by the likelihood of the alternate. The null is rejected when the ratio is below a critical threshold. Unlike a Bayesian statistical test, by using a Neyman–Pearson statistical test, it is possible to arrive at a different decision between the two competing models depending on which of the two is defined as the null.

Based on the evidence, a decision is made about the null hypothesis: Is it true or is it false? Unfortunately, this decision is not well defined; it is a balance of probabilities weighing the evidence for rejection of the null and the evidence toward not rejecting the null. Based on this balance, the decision is either the data support rejection or the evidence is insufficient to reject the null. Whichever the decision, there is a chance that the choice was wrong. If the null hypothesis is rejected, a mistake of the first kind (Type I error) may have been made. Conversely, if the null is not rejected, a mistake of the second kind (Type II error) may have been made.

Type I Error (α)

The Type I error, denoted by the symbol α, is the error rate of the test. It is the probability of rejecting H_0 when it is true.

To perform a statistical comparison, a particular probability of making a Type I error is established. The test statistic is compared to the critical value obtained from the probability distribution calculated under the assumption that H_0 is true. If the test statistic is equal to or lower than the critical value, then H_0 is not rejected. If the statistic is greater than the critical value, then the null hypothesis is rejected and H_A is accepted.

Computer programs use algorithms to compute an exact probability for the test statistic assuming the H_0 is true. This probability is termed the *P*-value. A *P*-value is the likelihood of observing a test statistic by random sampling that is as large as or larger than that obtained from the study. It is a measure of the strength of evidence against the null hypothesis: a very small *P*-value means that the test result would be very unlikely under H_0.

In practice, an investigator will specify a probability of a Type I error for which they are willing to reject the H_0 beyond reasonable doubt. If the *P*-value for a statistic is less than the designated α, then the null hypothesis is rejected. If the *P*-value for a test statistic is greater than or equal to the established error rate (α), then there is insufficient evidence to reject the null hypothesis. The Type I error benchmark of 0.05, although it is commonly used by many researchers, has no "magic" property other than it was the odds, 1 out of 20, that Fisher preferred (Cochran, 1976). Moreover, the traditional value of 0.05 (or 0.01) may not be optimal in all cases, and a risk assessment of the decision mistake needs to be applied (Carmer and Walker, 1988).

The level of α is established by you, the experimenter. Once you have set α, it must be consistently applied for all comparisons. For instance, if you have specified the Type I error to be 0.05, or whatever level is accepted by scientists in your discipline, then there is no place for conditional statements such as "highly significant." Either H_0 has been rejected or failed to be rejected; that is, either the difference was significant or was not. If you are not definitive in this declaration, you end up reintroducing personal biases that the science of statistics endeavors to avoid. In other words, you might be accused of applying the science only when it is convenient for your purpose. Reserve the use of adjectives for assessing the measurable outcome, the effect size, but not the significance of the statistics test.

For presentation purposes, do not round the *P*-value; present the value to four decimal places. For example, consider an analysis in which the Type I error rate was set at 0.05 and a statistic had a *P*-value of 0.0451. If the statistic's *P*-value is rounded to two decimal places (0.05), instead of rejecting the null hypothesis (0.0451 is less than the benchmark 0.0500), the null hypothesis would not be rejected. Furthermore, when results of comparisons are reported, ensure that the associated text and/or footnotes correspond to the decision made. When the *P*-value is less than the established benchmark (α), then the null hypothesis is rejected. Hence, sentences should be constructed as: . . . were different ($P < \alpha$) or . . .were significant ($P < \alpha$). Note that the sign is $<$ and not $\leq$. If the *P*-value is equal to or greater than the established benchmark (α), then the null hypothesis is not rejected. Hence, sentences should be constructed as: . . .did not differ ($P \geq \alpha$) or . . .were not significant ($P \geq \alpha$). It is also helpful to present the exact *P*-value, such as . . .were significant ($P = 0.0451$), which easily allows others to apply their own Type I error benchmark as they assess the results of your study.

Type II Error (β)

The probability of accepting H_0 when it is false is termed the Type II error. It is denoted by the symbol β. The Type II error relates to the power of the test ($1 - \beta$) which is the probability of rejecting H_0 when it is false.

The Type II error rate depends on the characteristics of the population, the precision of the experiment, and the specific hypothesis under test. Unlike the Type I error rate, one is not able to formally

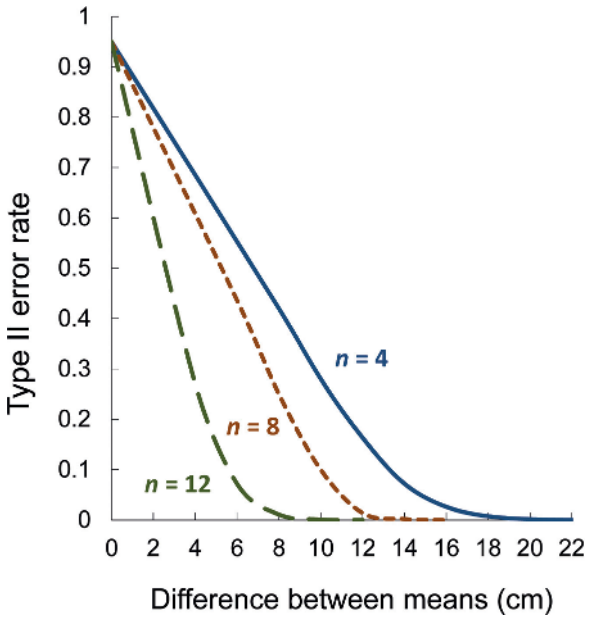


FIGURE 1.3 Effect of the sample size and the difference between means on the Type II error rate (β). The Type II error decreases as the sample size increases, assuming the errors remain constant as additional samples are measured. Values were computed for a population with $\mu = 127.8$ and $\sigma^2 = 68.9$ with sample sizes of 4, 8, and 12.

state a specific *a priori* Type II error rate for an experiment; it varies for each specific comparison being performed. Nonetheless, there are some general factors that affect the level of β:

1. The Type II error is affected by the magnitude of the difference between the two estimates; the closer they are, the higher will be the Type II error (Figure 1.3).
2. The size of β is also dependent upon the standard error of the estimate. This is determined by two elements: error variation and sample size. Through technical enhancements, improved experimental design, and appropriate statistical analyses, the error variance can be reduced. Decreasing experimental errors will decrease the probability of making a Type II error.
3. Ideally, the greater the sample size, the smaller will be the standard error. Thus, a greater sample size should reduce the Type II error rate (Figure 1.3). This assumes experimental errors remain unchanged, which is often not the case if the number of replications is sizably increased.

Power of the Test

The power of the test ($1 - \beta$) is the probability of rejecting H_0 when it is false. As the power of a statistical test increases, the probability of a Type II error decreases.

Consider a statistical comparison between two means μ_1 and μ_2. The null hypothesis of the comparison is that the two means are the same; the alternate is that the two means differ. Expressed in another way, the null hypothesis is that the difference between the two estimates is 0 (H_0: $\mu_1 - \mu_2 = 0$) and the alternate hypothesis is that there is a difference (H_A: $\mu_1 - \mu_2 \neq 0$). This notation is an example of a Fisherian test—the alternate mean is an infinite array of possibilities other than μ_1.

To determine the power of the test, four values are required: (i) the magnitude of the difference; (ii) the standard error of the estimates; (iii) the sample size; and, (iv) the Type I error rate of the test. Figure 1.4 illustrates how power can be determined. The curve on the left represents the expected distribution under the null hypothesis. For a mean, it follows a central *t*-distribution, which describes the distribution when H_0 is true. The mean (μ_0) coincides with the peak of the curve. When a Type I error rate (α) is set for the test, the critical value can be determined. This is indicated by the asterisk symbol. The critical value defines the boundary of the rejection region. If the second mean (μ_2) is to the right of this critical value, the null hypothesis is rejected. If the second mean is either at, or to the left of, the critical value, the null hypothesis, that they do not differ, is not rejected. The probability of

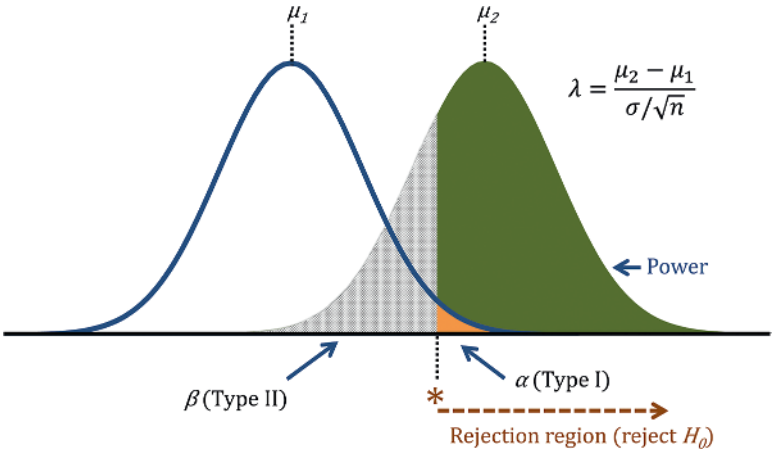


FIGURE 1.4 The curve on the left represents the central *t*-distribution under the null hypothesis, and the curve on the right is the expected distribution under the alternate—the non-central *t*-distribution. The null hypothesis is rejected if the mean is beyond the critical value (∗) for the test. The power of the test is the portion of the non-central *t*-distribution located within the rejection region of the test; the Type II error is the partially shaded area to the left. The larger the non-centrality parameter (λ), the greater is the power of the test.

a Type I error, rejection of the null hypothesis when it is true, is α. The critical value remains the same regardless of the true value of the mean. In this example, the null would be rejected; μ_2 is within the rejection region.

Under the alternate hypothesis, the expected distribution follows a non-central *t*-distribution defined by the non-centrality parameter (λ). This describes the expected distribution when H_0 is false. The non-centrality parameter is based on the distance between the alternate and the null ($\mu_2 - \mu_1$), the standard deviation (σ), and the sample size (n). The non-centrality parameter becomes larger as the distance between the means increases, as the sample size gets larger, and/or as the variance decreases.

The power, the probability of rejecting the null hypothesis when it is false, is the portion of the non-central *t*-distribution which is located within the rejection region of the test. This area is the power of the test. In Figure 1.4, this is the shaded region of the distribution located to the right of the critical value.

The area of the non-central *t*-distribution to the left of this critical boundary is the Type II error; in Figure 1.4, it is depicted by the partially shaded region.

The power of the test is affected by four factors: distance, variance, size of the sample, and the Type I error. The further apart the means, the greater will be the power. The closer they are, the smaller will be the area within the rejection region and the lower will be the power of the test. The effects of variance and sample size are reflected in the standard error of the estimates—the lower the variance and/or the greater the sample size, the lower will be the standard error. As the standard error of the estimates decreases, the non-central *t*-distribution will be narrower; it will have less overlap with the central-*t* distribution, and the test will have greater power.

The power can be altered only through two of these four factors: namely, the variance and the sample size. Researchers have no control on the distance between the two means—this distance is fixed by nature. Although it is possible to specify any Type I error rate to use, the α rate used is one that is commonly accepted by scientists in their respective disciplines. Nonetheless, the shape and overlap of the non-central *t*-distribution can be altered by decreasing the errors associated with the estimates, which decreases the denominator of the non-centrality parameter. This can be achieved through use of improved techniques and experimental design, increased sample size, and by application of appropriate statistical analyses. Although no specific benchmark has been established as an optimal power value, most use the level 0.80, a level of $\beta = 0.20$ or less, set by Jacob Cohen (1988) as adequate to reject a false null hypothesis. Cohen viewed this level as an optimal balance between the Type I and II risks, reasoning that most researchers would view a Type I error as being four times more serious than a Type II error.

A retrospective, or post-hoc, power analysis uses the sample size and error variances from the experiment to determine what the power was for each statistical comparison. There is not a single power value for a study *per se*; the power value will vary with each specific statistical comparison performed in the analysis. Before performing a research study, an *a priori* power analysis should be conducted in order to predict if the sample size is sufficient to achieve the desired power and

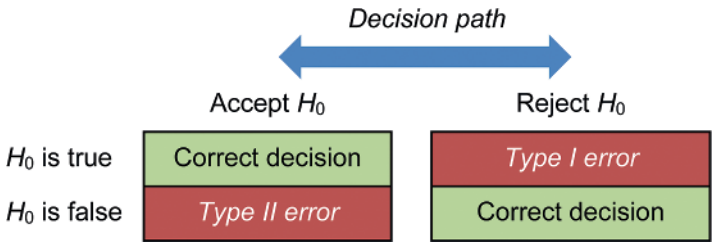


FIGURE 1.5 Relationship between decisions and mistakes associated with the acceptance or rejection of a given null hypothesis.

to compare the predicted power of alternate experimental designs. An *a priori* power analysis requires prediction of the means and error variation for the proposed study. Methods to conduct both retrospective and *a priori* power analyses are outlined in Chapter 5.

Figure 1.5 summarizes the possible decisions and mistakes that can be made for a given hypothesis. The Type I error rate is established by the experimenter, while the Type II error rate is a function of the sample size, errors, and the population under study. You only need to be concerned about making one of the two types of mistakes: which one depends on the decision you have made about the null hypothesis. If you reject the null hypothesis, you may be making a Type I error; if you do not reject the null hypothesis, you may be making a Type II error.

Researchers should always be cognizant of the likelihood, and impact, of making a Type II error whenever the null hypothesis is not rejected. Only through assessment of the study design and the power of the test is the chance of making a Type II error minimized.

We strive to reduce the chance of making either mistake through the experimental design and methodologies applied. Since it is not possible to reduce the chance to 0, one might wonder: which of the two mistakes is more serious to make? The decision to either accept or reject a hypothesis has consequences. This is termed the *inductive risk*, the risk of error in the decision about the null hypothesis. The inductive risk depends on the situation and involves value judgement. To explain, let us consider these two mistakes in the context of a screening test for a disease. If there is no disease present but one declares that there is disease, then a Type I error has been made. The null hypothesis has been incorrectly rejected. This is also termed a *false positive*. A Type II error is made when one incorrectly fails to reject the null hypothesis and declares that the subject is disease-free. The test has incorrectly provided evidence against the alternate hypothesis. In the context of a screening test, a Type II error can also be termed a *false negative*.

By thinking of the Type I and Type II errors in terms of false positives and false negatives, we are better able to consider the relative consequences of the two possible mistakes. If the comparison is related to a medical diagnosis, then a false positive would result in a patient needlessly undergoing a course of treatment and likely additional testing. Through additional testing of the patient, the mistake would likely be detected, the course of treatment discontinued, and the consequence of the false-positive reading would be minimized. On the other hand, a Type II error would give the patient the incorrect assurance that they did not have the disease. This misdiagnosis would result in a delay of treatment for the condition, and this delay may have serious consequences for the patient. From the viewpoint of the patient, a Type II error would be of greater concern than a Type I. The relative impact of either mistake is subjective and will vary depending on the hypothesis under consideration.

P-Value Misuse

Too often, researchers focus on applying a *P*-value as a benchmark without placing the findings in context. This is often referred to as the "*P*-Hunt" for experiments in which the null is rejected. By itself, a *P*-value does not provide a good measure of evidence regarding a model or hypothesis. A *P*-value can indicate how incompatible the data are with a specific model, but it does not measure the probability that the studied hypothesis is true. Rejecting the null does not "prove" the alternate is true. Similarly, failing to reject it also does not mean that H_0 is true. What the study provides is contribution toward our understanding of the true situation. If H_0 is indeed true, it would require the

combined evidence of many well-designed and executed studies, each with high power of the test, in order to confidently conclude that this was the most likely situation.

Scientific conclusions should not be based solely on whether a *P*-value passes a specific threshold. Moreover, *P*-values should not be numerically compared—they are single values with unknown distribution. In other words, if one comparison had a *P*-value of 0.0345 and another had a value of 0.0123, you are not able to conclude that the latter comparison is "more significant" than the first. Indeed, there is wide variation for *P*-values among repetitions of an experiment—a variation termed the dance of the *P*-values. Decisions based on effect size and confidence intervals show less test-to-test variability than the *P*-value. Furthermore, the *P*-value, or statistical significance, does not measure either the size of the effect or the importance of the result.

Effect Size

A common hypothesis that is tested in the analysis of experimental data is whether differences exist among the treatments. Some view this as an exercise in futility (see Chew, 1977). Given a properly designed experiment and sufficient replication, the statistical null hypothesis of equal treatment effects will always be rejected. The hypothesis of greater concern is whether the difference is of practical value or biologically meaningful. In other words, is the difference of importance or is it much ado about naught? The *P*-value can reveal whether a difference exists, but it does not convey the magnitude of the effect. Indeed, with very large sample sizes, significant *P*-values can be obtained when there is little measurable outcome. Inferences based on a decision to reject the null, declaring a difference, should be formulated with regard to the effect size. The effect size is an index which quantifies the magnitude of a treatment difference or the strength of an association: Is it huge or is it negligible? In the case of binary variables, the effect size is computed as an *odds ratio*. For quantitative variables, it can be computed either as a ratio of the means ($\text{mean}_1/\text{mean}_2$), abbreviated as *RoM*, or the standardized difference between means, $(\text{mean}_1 - \text{mean}_2)/s$, termed *Cohen's d*. The latter expresses the difference in terms of standard deviations. Other variations of the mean difference index include *Glass'* Δ, which uses the standard deviation of mean_2 as a divisor, and *Hedge's g,* which incorporates a correction factor for small sample bias. For a measure of association such as Pearson's linear correlation coefficient, it is based on the degree of the linear association. Table 1.2 indicates relative benchmarks for three common effect size indices; these are based on the relative amount of variation explained by the effect. For odds ratios in which the numerator is less than the denominator, the benchmarks would be the inverse of those listed in Table 1.2. The thresholds are a general guide only; they need to be adjusted for the impact of the change within the context of the study.

A comparison may be declared statistically significant, but it may not represent a biological, or practical, meaningful difference. Inferences should be based on both the statistical significance and the magnitude of the change.

Table 1.2 Relative thresholds for three effect size indices.

Relative effect size	Odds ratio	Cohen's *d*	Correlation
Small	1.5	0.2	±0.2
Medium	2.0	0.5	±0.5
Large	3.0	0.8	±0.8

Diagnostic Tests

Diagnostic tests, such as antibody tests, are often used to assess the health of a subject or detection of a compound. Typically, these tests have a binary response, producing either a positive or negative response. Diagnostic assays are never perfect, and knowledge of their reliability is important in their application. Consider a screening test for a particular disease in a nursery. Assume that absence of the disease is the null hypothesis. If there is no disease present, but the diagnostic test indicates that there is infection, then the null hypothesis has been incorrectly rejected by the assay (a Type I error). Conversely, if the disease is present and the diagnostic test result indicates otherwise, then the null hypothesis has been incorrectly accepted (a Type II error). As indicated earlier, these mistakes with diagnostic tests are termed false negatives and false positives, respectively.

In developing a diagnostic assay, one strives to construct one that reports the true situation: all true positives are identified to be positive, and all true negatives are identified to be negative. The probability that a diagnostic test of an infected plant produces a positive result is referred to as the *sensitivity* of the test. Conversely, the probability that a test of a non-infected plant produces a negative result is referred to as the *specificity* of the test. Exact confidence intervals for these estimates are based on binomial probabilities. Unfortunately, there is usually a trade-off between sensitivity and specificity; as one is increased the other decreases. Which of the two is more important will depend on the objectives of the assessment. If a low frequency of false positives is the objective, then high specificity is desired. Conversely, if a low frequency of false negatives is the objective, then high sensitivity is desired. However, we often wish to simultaneously achieve both goals, which require a balance between sensitivity and specificity.

The pretest probability is the probability that an individual has the disease prior to any diagnostic test; it is the ratio of true positives divided by the total number. The predictive value of a positive test, termed the *likelihood ratio of a positive test* (LR^+), is the ratio of true positives divided by the sum of (true + false positives). This can also be computed as the sensitivity divided by 1 − specificity. The predictive power of a negative test, termed the *likelihood ratio of a negative test* (LR^-), is the ratio of true negatives divided by the sum of (true + false negatives), which equals 1 − sensitivity divided by the specificity. The LR^+ ratio indicates how much a positive test result will increase the posttest probability of the disease being present in the subject compared to the pretest probability. Likewise, the LR^- ratio indicates how much a negative test result will lower the posttest probability the subject has the disease. LR ratios can be used to assess the potential utility of the diagnostic test. The higher the LR^+ value, or the lower the LR^- value, the greater the utility (Table 1.3).

The *diagnostic odds ratio* can be used to compare the discriminative power of two or more diagnostic procedures. It is computed as the ratio of LR^+/LR^-; the test with the higher diagnostic odds ratio, with high sensitivity and specificity, has greater discriminative power.

The sensitivity, specificity, and LR^+/LR^- ratios of a diagnostic test can be determined using samples that are known to contain, and samples known not contain, the constituent to be detected. In many cases, such as detection of a disease organism, the true disease status may not be known. In these situations, the comparisons will be to a *gold standard* test—the best available diagnostic test for determining whether the disease or compound is present. Gold standard tests are often expensive, invasive, or risky and are typically reserved for retesting subjects that test positive using a lower cost, timelier, diagnostic test.

Table 1.3 Relative effects on the posttest probability of a disease.

Relative effect size	LR^+ value	LR^- value
None	1	1
Slight	2–5	0.2–0.5
Moderate	5–10	0.1–0.2
Large	>10	<0.1

The confidence interval (CI) is a measure of uncertainty for proportion estimates such as sensitivity, specificity, and other ratios. There are a number of types of intervals that can be computed. One type that is commonly described in statistic textbooks is the *Wald interval*. It is based on the properties of the normal distribution (see Chapter 3) and is computed as

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

where $\hat{p}$ is the estimated sample proportion, $Z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution, and N is the sample size.

When the sample size is small, or when the proportion estimate is low (closer to 0) or high (closer to 1), the Wald interval not a suitable measure of uncertainty. In view of its limitations, it is recommended that a different type of confidence interval be employed. A commonly used alternative is the *Clopper–Pearson interval,* which is based on the cumulative probability of the binomial distribution. Although preferable to the Wald interval, some view the Clopper–Pearson interval as being too conservative. See Brown et al. (2001) for further details and comparison of confidence intervals for binomial proportions.

EXAMPLE 1.1 DIAGNOSTIC TEST ANALYSIS

Sudden oak death, caused by *Phytophthora ramorum* Werres et al., causes high tree mortality and has a wide range of host species. Because of quarantine regulations and a need for efficient identification of infected plants, various diagnostic techniques were compared on 428 plant samples (Vettraino et al., 2010). The results of the nested polymerase chain reaction (PCR) assay in that study will be used for illustration (Figure 1.6A). For this evaluation, the gold standard was a combination of an agar culture of plant tissue and a PCR test for the causal agent.

A total of 323 samples were diseased (Figure 1.6A). Of these, 302 tested positive and 21 tested negative using the PCR assay. The sensitivity of this test was 302/323 = 93.5%, and the false-negative rate was 1 − 0.935 = 6.5% (Figure 1.6B). There were a total of 105 samples that were disease-free.

A Diagnostic test results

	Test positive	Test negative
Gold standard, positive	302	21
Gold standard, negative	14	91

B Sensitivity and specificity

		Clopper-Pearson 95% confidence limits	
	Estimate	Lower	Upper
Sensitivity	0.935	0.902	0.959
Specificity	0.867	0.786	0.925
Positive predictive value	0.956	0.927	0.976
Negative predictive value	0.813	0.728	0.880

C Likelihood ratio estimates

	Estimate	Lower 95%	Upper 95%
LR+	7.01	4.30	11.43
LR-	0.08	0.05	0.11
Diagnostic odds	93.5	45.7	191.2

D Probability with disease

Pre-test	Positive post-test	Negative post-test
0.755	0.956	0.187

FIGURE 1.6 Analysis summary of the nested PCR diagnostic test for *Phytophthora ramorum*. Diagnostic test results (A), sensitivity and specificity estimates (B), likelihood ratio estimates (C), and pre- and post-test probability with disease (D). LR+ = likelihood ratio of a positive test, LR− = likelihood ratio of a negative test.

The specificity of the PCR test was 91/105 = 86.7%, and the false-positive rate was 1 − 0.867 = 13.3%. The positive likelihood ratio (LR^+) was 7.01, and the negative likelihood ratio (LR^-) was 0.08 (Figure 1.6C).

There were false readings for both the disease-free as well as the diseased samples compared to the gold standard method; the proportion incorrectly classified, the *error rate*, was (14 + 21)/428 = 8.2%. The LR values indicated the test offered a moderate change in posttest probability. The pretest probability of disease was 75.5% (Figure 1.6D). For a positive diagnostic test, this probability, also termed *positive predictive value*, increased to 95.6%, and a negative diagnostic test reduced this probability to 18.7%. Alterations in the methodology, or repeated testing over time to reduce false negatives, could be ways to improve the utility of this diagnostic test.

The default binomial confidence intervals generated in SAS are Wald intervals, and in R, they are Clopper–Pearson intervals. The Wald interval is based on the normal approximation, whereas the Clopper–Pearson interval is based on the cumulative probability of the binomial distribution. The procedures in both systems include options by which you can request other types of intervals.

Bias

Bias is the tendency to over- or underestimate the true value. In experiments, biased estimates can arise due to the study design and methodology such as unrepresentative sampling. Biased estimates can also be encountered when statistical tests are applied to complex models. For the latter, the bias is predictable, or systematic, and can be adjusted for. Diagnostic assays also have systematic bias. This is because the result indicated by the test is the number of samples that test positive, not the true number that are positive. To illustrate, consider a population which had 5% of the plants infected by *Phytophthora ramorum*. For the PCR assay (Example 1.1), its sensitivity was 93.5% and its specificity was 86.7%. If 100 plants were to be randomly selected, on average, 5 would be true positives and 95 true negative. The PCR test would only detect 93.5% of the five diseased plants as positive (4.675). However, due to the false-positive rate (1 − specificity = 13.3%), the diagnostic test would also report 12.635 of the 95 disease-free plants as being positive. The apparent incidence based on the test would be 4.675 + 12.635 = 17.31% infected—a much higher level (>3 times) than the true situation. Conversely, when the true incidence of the disease is very high, the diagnostic test will indicate a lower incidence than the true value due to the number of false negatives.

Figure 1.7 illustrates graphically the bias due to false positives and false negatives if this diagnostic test was used to monitor the disease prevalence in a population. The dashed line indicates the true frequency of the disease. The blue area is the frequency of true positives that are detected by the assay, and the gray area is the frequency of false positives. When the disease is at a low frequency in the population, the test would indicate, due to the false positives, a higher prevalence than is the true situation. When the disease is at a high frequency in the population, the test would indicate, due to the false negatives, a lower prevalence than that which exists. The degree of bias is a function of the test sensitivity, specificity, and disease prevalence. As sensitivity and/or specificity increases, the degree of bias is reduced. Provided the sensitivity and specificity are well-characterized under the conditions that the diagnostic test is employed, the bias in the estimate of disease frequency in a population can be corrected for (see Campbell et al., 2008).

The utility of a particular diagnostic test to monitor the incidence of a disease in a population will depend upon its prevalence. For Example 1.1, there is little bias in the range of 65–75% disease prevalence since this is the region where the number of false negatives and positives balances each other

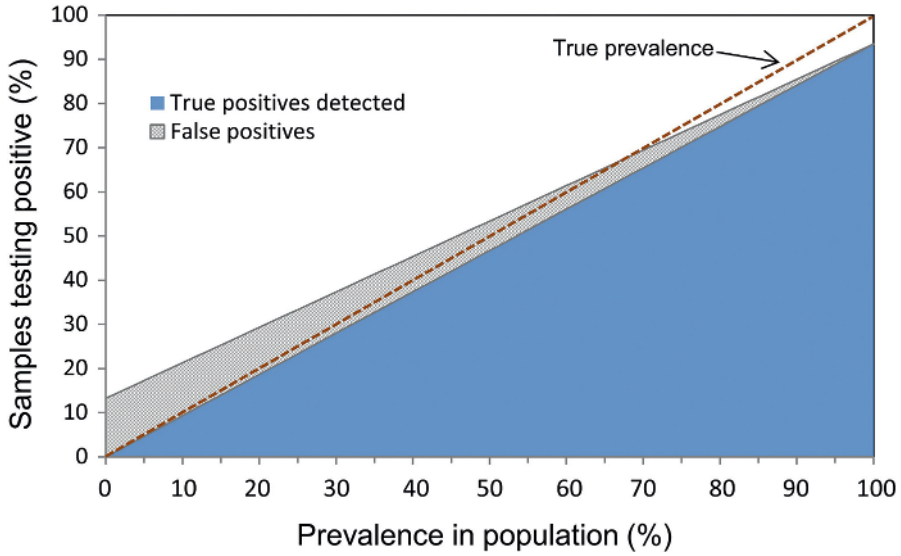


FIGURE 1.7 Effect of disease prevalence on the bias when a diagnostic test is used to estimate the frequency of a disease in a population. The dashed line indicates the true prevalence, the blue region indicates the number correctly detected, and the gray region the number of false positives for a diagnostic test with a specificity of 93.5% and sensitivity of 86.7%. When the prevalence of the disease is low, the diagnostic test overestimates the frequency (due to false positives); when the prevalence of the disease is high, the test underestimates the frequency (due to false negatives).

(Figure 1.7). The lower the prevalence, the greater will be the upward bias to the test result, and the higher the prevalence, the greater will be the downward bias to the test result. When prevalence is low, a diagnostic test with high specificity is desired to reduce bias. Such a test would minimize the level of false positives (rejecting the null when it is true). However, when the incidence of the disease in a population is high, a diagnostic test with high sensitivity is desired in order to minimize the false negatives. The degree of bias and the mid-point location where the false negatives and false positives will balance each other depend upon the particular sensitivity and specificity of the diagnostic assay.

Summary

Experiments are conducted, and data are generated in order to challenge scientific hypotheses. Statistical tests are a way to draw unbiased inferences about the evidence collected. Does the data support the hypothesis or does it not? The decision is not clear-cut as it is accompanied by a degree of uncertainty. By conducting well-planned experiments, you can lessen, but not eliminate, the degree of doubt associated with the inferences drawn. The decision to either reject or failing to reject a statistical null hypothesis is always tied to a possible mistake: rejecting the null when it is true or failing to reject it when it is false. Designing and conducting studies with high power will help lessen the degree of uncertainty, but you must always be cognizant of its presence. Regardless of the decision, you are never able to prove which hypothesis, H_0 or H_A, is true, but the evidence can contribute toward our understanding of which is most likely the true situation.

Drawing inferences from a study is only a small part of your contribution toward our collective knowledge. You need to be efficient in this process, which requires proper planning, organization, and execution. For other researchers to draw upon the experiment for future studies, they will need to know details: what was the evidence, how was it collected, and how was it analyzed? This provides a mechanism for others to replicate your study and/or the analysis techniques applied. Your data can be incorporated with new information or techniques in the future to better assess the question at hand. Moreover, by having access to the evidence in its raw form, new hypotheses can be evaluated—many of which you may not envision today. The next chapter describes processes and techniques that can be used to achieve these outcomes.

SAS Code

Data Files

The following provides coding for the data analysis of the example provided in this chapter. The code has been presented in the default color scheme used in a SAS editor window. Comments are presented in green, procedures in dark blue, the procedure statements in light blue, and data areas, if present, are highlighted in yellow. For this example, the data are arranged in three columns: the gold standard classification (variable *gstandard*), the test result (variable *test*), and the number of observations in each of the four groups (variable *count*). For Example 1.1 the data have been including within the coding. Note that each SAS statement is ended with a semicolon.

The **DATA pcr;** statement indicates that a data set is to be created, and it will be called **pcr**. Some analyses will involve a number of datasets, so the PROC statements will explicitly indicate which dataset to use. Following the data statement is an *infile* statement indicating how the columns of values in the data are delimited (in this example, it is a comma). The *input* statement then names each of the columns of values. If the name is followed by a $ sign, this indicates it is an alphanumeric variable (both letters and numbers). If there is no $ symbol after the name, then the range of values are converted to numbers (any characters are ignored). For this example, the variables *gstandard* and *test* are alphanumeric and the variable *count* is numeric.

Within a data step, calculations can be performed and other variables created. These can call upon a wide range of SAS functions and can include *if then else* decisions in the calculation. In this example, two additional numeric variables have been created, *result* and *gtest*, and have the values 0 or 1 depending on the situation. Following these statements is a *datalines* statement indicating the following rows are to be interpreted based on the input statement. Subsequent lines will be considered data until a blank line with the semicolon is encountered. Thereafter, the SAS program will once again treat the region as SAS program steps.

CSV file import. For the coding in later chapters, the data will be imported from an external spreadsheet file (csv format). The first line of the spreadsheet contains each of the variable names. There are many ways to import data into the SAS system. For those that use the online version called SAS STUDIO, this version uses virtual directories. Create a subdirectory in your SAS STUDIO session and upload the csv file(s) to it. We have called our subdirectory: csv_files. The SAS procedure PROC IMPORT can then be used to import the data from the spreadsheet into a SAS session. Here is an example for SAS STUDIO:

```
PROC IMPORT datafile='~/csv_files/pcr.csv' out=pcr replace;
    datarow=2;
    delimiter=',';
    getnames=yes;
    guessingrows=max;
RUN;
```

The *datafile* option indicates the path to the file name containing the data (pcr.csv). Note that you need to include the .csv extension to the file name. The tilde ~ symbol is a shortcut path to the SAS STUDIO directory. The *out*= option indicates the name of the SAS dataset to create (in this example, it is pcr), and the *replace* option indicates that if there is already a dataset called pcr, then it is to be replaced with this set of data.

For other versions of the SAS system, the location of the file would be based on its location in the system (drive and path). In the case of UNIX-based systems, the forward slashes (/) in the directory path would be replaced with backward slashes (\).

The variable names are assumed to be in the first row of the csv file. The *datarow* statement indicates what line in the spreadsheet the data start (it is row 2 in all the example files). The *delimiter* statement indicates what separates each field (a comma). The *getnames=yes* statement then uses the names in the first row as the names of the variables created.

For files that are delimited, by default the program will scan the first 20 rows to determine each column's data type (numeric or alphanumeric) and, if alphanumeric, the length in characters. The *guessingrows=max* statement overrides the default scan of 20 rows. For many datasets with alphanumeric variables, the values may get truncated if longer names are located lower in the spreadsheet. For instance, if values such as Exp01 to Exp20 are in the first 20 rows, these are five characters in length. Without overriding this default of 20 rows, a value such as Exp200, which may be located further down, would be truncated to five characters, resulting in the erroneous value of Exp20.

After importing the data, additional manipulations can be performed to the dataset. The following would be the commands required to create two additional variables necessary for Example 1.1. The *DATA* statement indicates that the new data set is to be called pcr (we could have given it another name), and the *set* statement indicates that the program is to begin with the existing dataset called pcr. The first if-then-else statements create a new variable called *result* that has the value 0 if the alphanumeric variable *test* exactly matches the character string "negative." Otherwise, the variable *result* is assigned the value 1.

```
DATA pcr;
  set pcr;
  if test='negative' then result=0;
  else result=1;
  if gstandard='absent' then gtest=0;
  else gtest=1;
RUN;
```

Upper and Lower Case Letters

Most commands and operations in the SAS system are case-insensitive. Thus, factor names and statements can be written as lower case, upper case, or a mixture thereof. For instance, a factor such as length can be written as length, Length, and LENGTH and a procedure/statement written as proc import, Proc Import, or PROC IMPORT, and the command will be recognized as equivalent. In the coding for all examples, we have listed all the procedure calls in capital letters. This was not necessary but was done to assist your understanding as to the steps within the data analysis.

However, what are case-sensitive are the values of alphanumeric variables. Levels of the variable that have different case styles are treated as separate things. For instance, consider a dataset involving information collected on a series of cultivars. Here is a subset of data involving the same cultivar:

eunit	cultivar	Location	yield_Mgha
24	Exp01	New_York	3.9
131	exp01	new_york	4.2
400	EXP01	NEW YORK	4.7

Note that the letters in the cultivar designation in each row differ in their case style: the first has an initial capital, the second is all lower case, and the third is all upper case. SAS will treat each of these forms to be different cultivars. Thus, for an if-then-else statement such as:

```
if cultivar='exp01' then result=0;
```

only the second instance in the dataset matches the request computation for the *result* variable. The other two versions of the cultivar designation are not a perfect match and would be ignored and not included by the IF command.

Differences in case style often arise when data are compiled over years, from different research groups or different software systems. The SAS system has a function that can be used to convert all levels of an alphanumeric variable to have the same case style. The function *propcase* will convert all levels to initial capitals, *lowcase* will convert all to lower case letters, and *upcase* will convert all to upper case letters.

For example, if you had a dataset (called *study2023*) containing an alphanumeric variable called *cultivar*, the following code could be used to convert all levels of the *cultivar* variable to initial capitals:

```
DATA study2023;
   set study2023;
   cultivar=propcase(cultivar);
RUN;
```

All instances of exp01, EXP01, and Exp01 would then be converted to Exp01.

Propcase will capitalize the first instance of a letter and every instance of a letter that follows a blank, hyphen, or tab. In order to also include a capital after a letter that follows an underscore, as in new_york, you also have to include that delimiter as an option:

```
location=propcase(location, "_");
```

If the function *lowcase* was substituted for *propcase* in the above code, all versions would be converted to exp01, and if the function *upcase* was used, then all versions would be converted to EXP01.

Spaces are also treated as a character; thus, the versions Exp01 and Exp 01 would not be recognized as the same cultivar. If the dataset involves differences such as this, then the function *compress* can be used to remove all spaces between characters. A related function, *compbl*, can be used to remove extra spaces between characters—instances of multiple spaces are converted to a single space, such as Exp 01 to Exp 01. It is also possible to combine a series of processes such as these within a DATA step.

```
DATA study2023;
```

```
    set study2023;
    cultivar=lowcase(cultivar);
    cultivar=compress(cultivar);
RUN;
```

This example coding would change the style to all lower case and concurrently remove all blanks within the designation. Thus, versions such as EXP01 and Exp 01 would all be converted to exp01.

The function *tranwrd* can be used to replace a character. For instance, to replace the underscore in the location designation new_york with a blank, this could be achieved by the command:

```
location=tranwrd(location, "_", " ");
```

or the reverse, blanks replaced by an underscore:

```
location=tranwrd(location, " ", "_");
```

Analysis of Diagnostic Tests

SAS does not have a specific procedure for analysis of diagnostic tests. A two-step approach to the analysis is required. The first step uses PROC FREQ to obtain the sensitivity and specific measures. This procedure requires the variable levels to be in the numeric form of 0 and 1 rather than in words such as negative and positive. The variables *gtest* and *result* were created to match the requirement for this phase of the analysis. The second step uses PROC IML to calculate the odds ratios and 95% confidence limits of the estimates. The PROC IML coding presented requires the data to be in the specific order we have listed in the DATA step.

Run and Quit Statements

At the end of each procedure step, a **RUN;** statement is required. This instructs the program to execute the statements listed for that procedure. In the case of PROC IML and PROC REG, a **QUIT;** statement is required to terminate those procedures. If you are changing titles within a series of steps, make sure a **RUN;** statement is placed prior to the title change. Otherwise, only the last version of the title will be the one used for the preceding phases of an analysis.

Title Statements

Title statements can be included in the coding. These are listed on the first line of every page of output to provide context to the various analysis tables and/or figures generated. Subtitles can be included, the *Title2* statement places the text on the second line, and *Title3* on the third line. Titles are changed whenever one is placed after a RUN; statement. If the Title line is blank, then that deletes further listing of the title on subsequent phases of an analysis. In the example coding, you will note the titles are deleted at the end of the sequence of analyses so that they do not carry over into subsequent SAS procedures you may submit.

Procedure and Statement Options

There are many options available for modifying the computation methods and the display of results. This applies to both the PROC statement and to each of the command lines within a procedure. Use the help option to determine what options are available along with details of its usage and a description and/or reference to the statistical methodology employed.

For example, in the PROC FREQ example coding, the specificity and sensitivity estimates are obtained using the *senspec* option to the *tables* statement. The default computation method for these confidence intervals is the Wald interval.

The tables statement contains the option to obtain a different type of confidence interval by specifying the type desired. The following statement would be used to obtain adjusted Wald intervals:

```
tables result*gtest /senspec binomial(CL= wald(correct));
```

The following statement would be used to obtain Clopper–Pearson intervals:

```
tables result*gtest /senspec binomial(CL= clopperpearson);
```

There are a number of other intervals available—see the SAS documentation for the *tables* statement in this procedure.

> **Note**
> *At the time of writing, the CL= option does not alter the CI computation method used by the senspec option—it is limited to computing Wald intervals. Example work-around coding has been included within the Example 1.1 code.*

EXAMPLE 1.1 Diagnostic Test Analysis. Data: pcr.csv

```
/*****************************************************************************************/

/*                                                                                       */
/* Dataset: pcr.csv  Phytothopthora detection via PCR.                                   */
/* gstandard=gold standard; test=test result                                             */
/*                                                                                       */
/*****************************************************************************************/


/* PROC FREQ senspec option requires the class values to be 0 or 1 */
/* result & gtest variables were created for this purpose          */

DATA pcr;
     infile datalines delimiter=',';
     input gstandard$ test$ gtest result count;
     datalines;
     present, positive, 0, 0, 302
     present, negative, 0, 1, 21
     absent, positive, 1, 0, 14
     absent,  negative, 1, 1, 91
     ;
RUN;

TITLE 'Sensitivity and Specificity estimates';

PROC FREQ data=pcr order=data;
     weight count;
     tables result*gtest /senspec;
     exact binomial;
RUN;

TITLE 'LR+, LR-, Odds ratio estimates';

PROC IML;
     use pcr;
     read all into y;
     tp=y[1,3];
     fn=y[2,3];
     fp=y[3,3];
     tn=y[4,3];
     /* Listing of dataset */
     data=j(2,2);
     data[1,1]=tp;
     data[1,2]=fn;
     data[2,1]=fp;
     data[2,2]=tn;
     print data[r={"Gold standard positive" "Gold standard negative"} c={"Test positive" "Test
     negative"} l={"Diagnostic test results"}];
     /* Calculation of estimates */
     m1=tp+fn;
     m2=fp+tn;
     sen=tp/m1;
     spe=tn/m2;
```

```
     t=1.96;
     ans=j(3,3);
     ans[1,1]=sen/(1-spe);
     ans[2,1]=(1-sen)/spe;
     ans[3,1]=ans[1,1]/ans[2,1];
     sePos=sqrt((1/tp)-(1/m1)+(1/fp)-(1/m2));
     seNeg=sqrt((1/tn)-(1/m1)+(1/fn)-(1/m2));
     seodds=sqrt((1/y[,3])[+,1]);
     ans[1,2]=exp((log(ans[1,1])-(t*sePos)));
     ans[1,3]=exp((log(ans[1,1])+(t*sePos)));
     ans[2,2]=exp((log(ans[2,1])-(t*seNeg)));
     ans[2,3]=exp((log(ans[2,1])+(t*seNeg)));
     ans[3,2]=exp((log(ans[3,1])-(t*seodds)));
     ans[3,3]=exp((log(ans[3,1])+(t*seodds)));
     print ans[r={"LR+" "LR-" "Diagnostic odds"} c={"Estimate" "Lower 95%" "Upper 95%"}
     l="Likelihood Ratio Estimates" format=8.2];
     */Pre- and Post-test probabilities/;
     prob=j(1,3);
     prob[1,1]=m1/sum(y[,3]);
     prob[1,2]=ans[1,1]*prob[1,1]/(1-prob[1,1]);
     prob[1,2]=prob[1,2]/(1+prob[1,2]);
     prob[1,3]=ans[2,1]*prob[1,1]/(1-prob[1,1]);
     prob[1,3]=prob[1,3]/(1+prob[1,3]);
     print prob[c={"Pre-test" "Positive Post-test" "Negative Post-test"} l="Probability with
     disease" format=8.3];
QUIT;

/* At the time of writing there was no senspec option to obtain confidence intervals
other than Wald intervals. Here is an example workaround to obtain alternate confidence
intervals for the sensitivity and specificity estimates.

With one-way tables the ratio of the first cell divided by the total of all cells is used as the
default for computing the proportion estimate. This is why the counts for the specificity
had to be reversed compared to the order they were listed in the original pcr dataset. */

DATA senspec;
     infile datalines delimiter=',';
     length statistic$ 15;
     input statistic$ class count;
     datalines;
     +ve Pred value, 1, 302
     +ve Pred value, 2, 14
     -ve Pred value, 1, 91
     -ve Pred value, 2, 21
     sensitivity, 1, 302
     sensitivity, 2, 21
     specificity, 1, 91
     specificity, 2, 14
     ;
RUN;
/* Four intervals are computed using the following example coding (See SAS documentation for other
types):
Wald, Clopper-Pearson, Continuity-corrected Wald, and Continuity-corrected Wilson Score */

PROC FREQ data=senspec order=data;
     weight count;
     tables class / binomial(CL=wald);
     tables class / binomial(CL=clopperpearson);
     tables class / binomial(CL=wald(correct));
     tables class / binomial(CL=wilson(correct));
     by statistic;
RUN;

/* Clearing the title line */
TITLE;
RUN;
```

R Code

Getting Started

All of the data sets used in these examples, as well as some functions, are found in the companion R package `ASB`. To download and install that package, copy and paste the following lines of code into the R console:

```
# First, make sure devtools package is available and install if needed
if (!'devtools' %in% rownames(installed.packages())){
  install.packages('devtools')
}
# Minimal installation (fast)
devtools::install_github(repo = "IFAS-SCU/ASB-R",
                         INSTALL_opts = c("-no-multiarch"))
```

These examples also make use of functions contained in a large number of other packages. To download and install all of those other packages as well, you can run the following code, but be warned that doing so can be very slow. If you encounter any issues with the installation, refer to https://www.github.com/IFAS-SCU/ASB-R/ for help troubleshooting.

```
# First, make sure devtools package is available and install if needed
if (!'devtools' %in% rownames(installed.packages())){
  install.packages('devtools')
}
# Complete installation (slow)
devtools::install_github(repo = "IFAS-SCU/ASB-R",
                         dependencies = TRUE,
                         INSTALL_opts = c("-no-multiarch"))
```

The examples to follow have been structured so that the code is fully self-contained at the level of each individual example (as opposed to chapter or the book itself). This was done to make the book more useful as a reference, at the price of introducing considerable redundancy. The text between blocks of code provides additional details that will be useful for those adapting the R code for the analysis of their own data. For the sake of brevity, however, these additional details generally appear only in the first example for which they are relevant. Likewise, only the minimal of output (which appears following "#>") has been included. To view the complete output, simply copy or type the code into an R script and run it. The code itself and the R output are shown in `consolas font`, with the code being colored using "syntax highlighting" and each line of output (other than figures and tables) proceeded by #>. Whenever R packages, functions, and function arguments are mentioned in the text, they are indicated by the following formatting conventions: `package`, `function()`, and `argument =`, respectively.

Finally, while many of the basics of the R language itself are reviewed in the following examples, those who have never used R before or who need assistance installing and running the R or RStudio software may wish to consult one of the many excellent introductions to R programming which are available for free online, such as the hands-on programming with R book by Garrett Grolemund.

EXAMPLE 1.1 Diagnostic Test Analysis

In this example, as in subsequent ones, the first task is to load all required packages and data sets. In this case, only the `ASB` package and the pcr data set are required. Since the data set consists of only a 2 × 2 contingency table, it can be printed in its entirety; in subsequent examples, additional methods will be introduced, which are appropriate for exploring larger and more complex data sets. Note that the `data()` function only works to load data sets which are part of a package: to read in data contained in an external file (such as .csv file), a different function, such as `read.csv()`, would be required.

```
library(ASB)    # Load the ASB package

data("pcr")     # Load the pcr data set
print(pcr)
#>               Gold Standard Test
#> Alternate Test   +  -
#>              + 302 14
```

```
#>              -  21 91
```

The following code will print row and column totals, respectively. Note that R ignores everything after a "#" symbol; these notes are said to be "commented out":

```
margin.table(pcr, margin = 1) # Margin 1 is rows
#> Alternate Test
#>   +   -
#> 316 112

margin.table(pcr, margin = 2) # Margin 2 is columns
#> Gold Standard Test
#>   +   -
#> 323 105
```

While prop.table() will print cell, row or column sums as proportions:

```
prop.table(pcr)             # Cell proportions
#>               Gold Standard Test
#> Alternate Test          +          -
#>              + 0.70560748 0.03271028
#>              - 0.04906542 0.21261682

prop.table(pcr, margin = 1) # Row proportions
#>               Gold Standard Test
#> Alternate Test         +         -
#>              + 0.9556962 0.0443038
#>              - 0.1875000 0.8125000

prop.table(pcr, margin = 2) # Column proportions
#>               Gold Standard Test
#> Alternate Test          +         -
#>              + 0.93498452 0.1333333
#>              - 0.06501548 0.8666667
```

Running the full diagnostic summary requires specifying which values in the table represent "true positives" (i.e. the alternate test agrees with the gold standard that the result is positive), "false negatives" (i.e. the PCR test returned a negative result while the gold standard showed it was positive), etc. One way to reference individual values in a table is via indexing: brackets containing comma separated values are used to indicate row and column values (e.g. `table_name[row_number, column_number]`):

```
diagnostic_summary(TP = pcr[1, 1], # First row, first column
                   FN = pcr[2, 1], # Second row, first column
                   FP = pcr[1, 2], # First row, second column
                   TN = pcr[2, 2]) # Second row, second column
```

	Estimate	Lower CI (95%)	Upper CI (95%)
Sensitivity	0.9350	0.9023	0.9593
Specificity	0.8667	0.7864	0.9251
Pos.Pred.Val.	0.9557	0.9268	0.9756
Neg.Pred.Val.	0.8125	0.7278	0.8800
LR+	7.0124	4.3024	11.4293
LR-	0.0750	0.0493	0.1142
Odds ratio	93.4762	45.6951	191.2198
Youden index	0.8017	0.7313	0.8720
Accuracy	0.9182	0.8881	0.9424
Error rate	0.0818	0.0576	0.1119

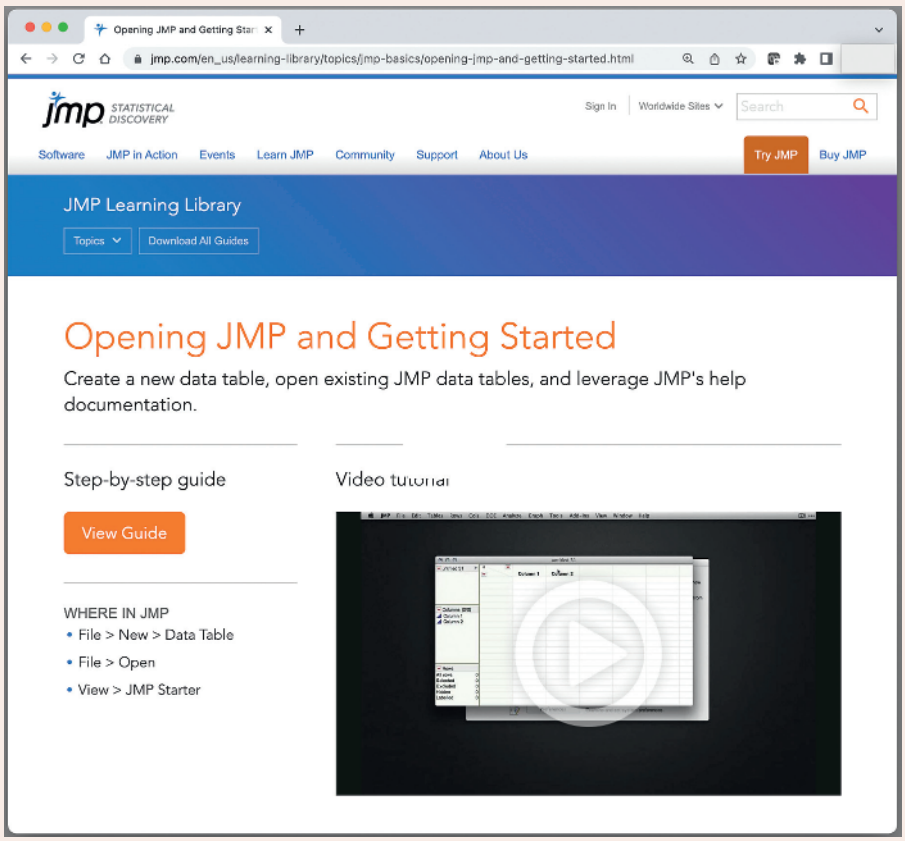
Note that the confidence intervals provided here are calculated following the "Clopper–Pearson" procedure, which differs slightly from the Wald-type confidence intervals which SAS employs by default.

JMP Method

Getting Started

JMP Basics and Tips

Go to jmp.com/learn and select the "JMP Basics" category to find topics including "Opening JMP and Getting Started," "Navigating JMP (in Windows or Mac)," "JMP Tables Menu," "Creating Formulas in JMP," and "JMP Data Tables" (for joining and splitting, etc.).



Time and Date

Dates in JMP are internally coded as "the number of seconds since midnight on January 1, 1904." JMP has built-in functions to transform date/time data into different formats or to extract the day, month, year, quarter, day-of-week, calculate time gaps, etc. Learn more in the JMP Help by searching "date time functions."

Data Formulas and Transformations

Data manipulation through transformations and formulas can usually be done very efficiently in JMP. For more information on the depth of transformation tools, see https://community.jmp.com/t5/Mastering-JMP-Videos-and-Files/Using-Formulas-to-Get-the-Most-from-Your-Data/ta-p/286131 and https://community.jmp.com/t5/JMP-On-Air/Formulas-in-JMP/ta-p/258755.

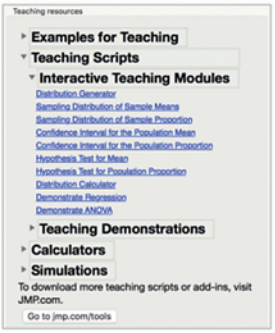
Go to jmp.com/learn for instructions on basic (and advanced) statistical concepts, such as confidence intervals and hypothesis tests. For learning tools about *P*-values and sampling distributions and other fundamental concepts, see the guide called "Interactive Tools for Teaching and Learning."

Statistical Concepts

jmp.com/learn Ver 04/2021

Interactive Tools for Teaching and Learning

A variety of tools for teaching statistical concepts are available within JMP. These tools allow you to explore fundamental concepts covered in introductory statistics courses, including sampling distributions, confidence intervals, hypothesis testing, ANOVA, regression, and more.



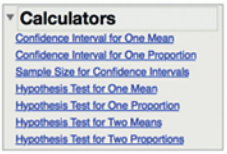
Interactive Teaching Modules

JMP provides built-in interactive modules for teaching core statistical concepts: sampling distributions, confidence intervals, hypothesis testing for means and proportions, plus probability distributions, regression, and t-Test and ANOVA.

1. You can run these modules directly from JMP Student Subscription: **Student > Applets**.
2. If you are using JMP or JMP Pro, access these tools from **Help > Sample Data > Teaching Resources > Teaching scripts > Interactive Teaching Modules**.
3. Click the Help button in any teaching module to learn more.
4. These modules can also be downloaded and installed as a JMP menu item as a free add-in from jmp.com/tools.

Calculators

JMP provides calculators for confidence intervals, hypothesis tests, and sample size for means and proportions, using either summary statistics or raw data. These are ideal for exploring how the width of confidence intervals change, how test results change, or how the calculated sample size changes as different input values are used.



1. In JMP Student Subscription: **Student > Calculators**.
2. In JMP or JMP Pro: **Help > Sample Data > Teaching Resources > Calculators**.
3. These calculators, along with additional calculators (for differences in two means, two proportions, and two variances), are available in an add-in bundle at jmp.com/tools.

Note: Additional information on these and other tools for exploring statistical concepts can be found at jmp.com/tools or in our user community at community.jmp.com. Data sets and additional teaching scripts, including a variety of teaching demonstrations and simulators, are available in the JMP Sample Data Directory under *Help > Sample Data*. Visit www.jmp.com/studentsubscription to learn more about JMP Student Subscription.

EXAMPLE 1.1 Diagnostic Test Analysis

Arrange the data in a data table of three columns, as shown in Figure 1.C.1, or open the Phytophthora ramorum.jmp data table. (To create a new data table in JMP, go to File > New > New Data Table and click in the column headers and cells to enter new column names and cell entries.)

Go to Analyze > Fit Y by X and enter the **Result** and **Truth** columns as *Y* and *X* and enter **Count** as the *Freq*, as shown in Figure 1.C.2.

Phytophthora ramorum

3/0 Cols / 4/0	Result	Truth	Count
1	Positive	Diseased	302
2	Negative	Diseased	21
3	Positive	Healthy	14
4	Negative	Healthy	91

FIGURE 1.C.1 Phytophthora ramorum.jmp data table.

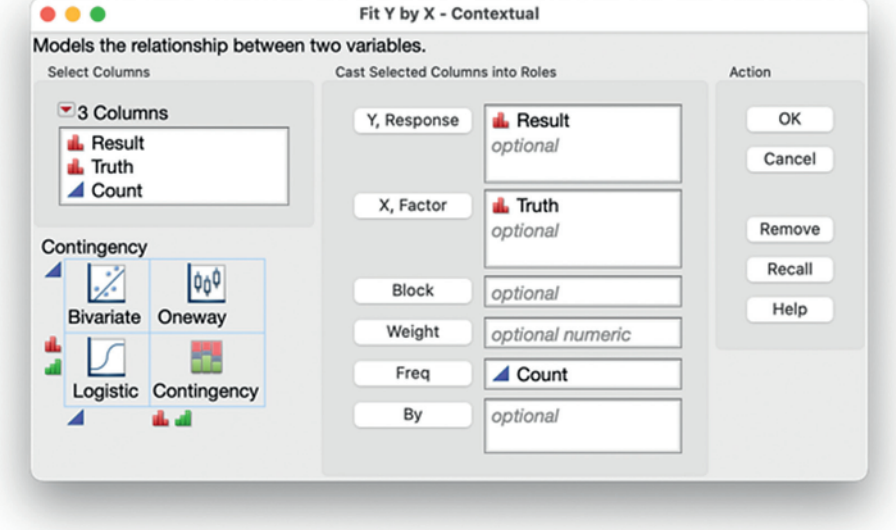


FIGURE 1.C.2 Completed dialog box for Fit Y by X.

Results are shown in Figure 1.C.3. Click the red triangle at the top left to select and display the relative risk output. Here, we see the contingency table with counts (302, 21, 14, and 91) and row and column percents, notably 95.57% for the probability of being diseased after seeing a positive test result, 93.50% for the probability of getting a positive result for a truly diseased specimen, 81.25% for the probability of not being diseased after seeing a negative result, and 86.67% for the probability of seeing a negative test result for a truly healthy specimen. The relative risk table shows the LR^+ of 7.01 for the ratio of the probability of a positive test result for a diseased specimen compared to the probability of a positive result for a truly healthy specimen and the LR^- of 0.08 for the ratio of the negative test result probabilities.

For the standard errors on the contingency table probabilities, go to Help > Sample Data. In the *Teaching Resources* in the bottom right, expand the *Calculators* sections and choose the *Confidence Interval for Two Proportions*. For *Choose input method* select *Summary Statistics* and enter 302, 323, 91, and 105. Click *Update Results*. Results are shown in Figure 1.C.4.

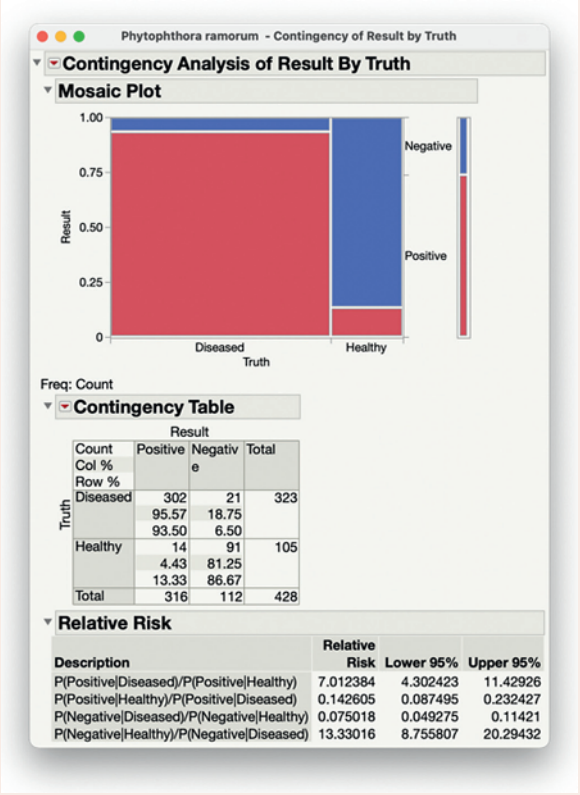


FIGURE 1.C.3 Output for Fit Y by X.

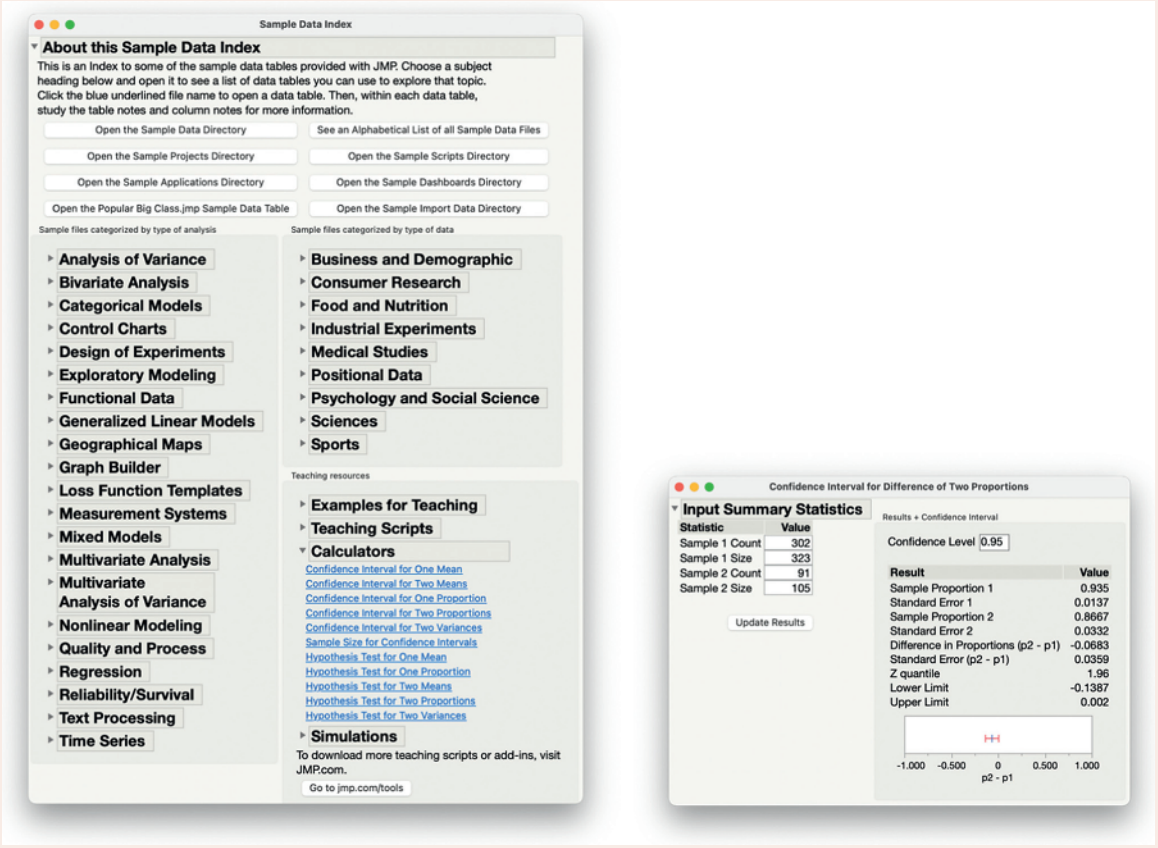


FIGURE 1.C.4 Standard errors for the contingency table probabilities.

REFERENCES

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, 101–133.

Campbell, H., Biloglav, Z., & Rudan, I. (2008). Reducing bias from test misclassification in burden of disease studies: use of test to actual positive ratio—a new test parameter. *Croatian Medical Journal*, 49, 402–414.

Carmer, S. G. & Walker, W. M. (1988). Significance from a statistician's viewpoint. *Journal of Production Agriculture*, 1, 27–33.

Chew, V. (1977). Statistical hypothesis testing: an academic exercise in futility. *Proceedings of the Florida State Horticultural Society*, 90, 214–215.

Cochran, W. G. (1976). Early development of techniques in comparative experimentation. In Owen, D. B., Minton, P. D., & Pratt, J. W. (Eds.), *On the history of statistics and probability*. (pp. 3–25). Marcel Dekker, Inc., NY.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Lawrence Erlbaum Associates, NJ.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2015). The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, 113(3), 554–559.

Hilts, V. L. (1978). *Allis exterendum*, or, the origins of the Statistical Society of London. *ISIS*, 69, 21–43.

Royal Sanitary Commission. (1858). *Report I of the Commissioners appointed to inquire into the regulations affecting the sanitary conditions of the Army, the organization of military hospitals, and the treatment of the sick and wounded; with evidence and appendix* (pp 368, 369, 555). Great Britain Parliament, House of Commons Sessional Papers, 1857–1858. Volume XVIII. London.

Scott, E. L. (1997). Jerzy Neyman. In Johnson, N. L. & Kotz, S. (Eds.) *Leading personalities in statistical sciences* (pp 137–145). John Wiley & Sons, NY.

Vettraino, A. M., Sukno, S., Vannini, A., & Garbelotto, M. (2010). Diagnostic sensitivity and specificity of different methods used by two laboratories for the detection of *Phytophthora ramorum* on multiple natural hosts. *Plant Pathology*, 59, 289–300.

ADDITIONAL READING

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *American Statistician*, 59(2), 121–126.

Cohen, I. B. (1984). Florence Nightingale. *Scientific American*, 250(3), 128–137.

Cohen, J. (1994). The earth is round ($p<0.05$). *American Psychologist*, 49, 997–1003.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, 67, 559–579.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and interpretation of research results*. Cambridge University Press, NY.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology-Research and Practice*, 40, 532–538.

Fowler, N. (1990). The 10 most common statistical errors. *The Bulletin of the Ecological Society of America*, 71(3), 161–164.

Hayden, S. R. & Brown, M. D. (1999). Likelihood ratio: A powerful tool for incorporating the results of a diagnostic test into clinical decision making. *Annals of Emergency Medicine*, 33(5), 575–580.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, 67, 160–167.

Kanji, G. K. (1993). *100 statistical tests*. Sage Publications Limited, London.

Kendall, D. G., Bartlett, M. S., & Page, T. L. (1982). Jerzy Neyman, 16 April 1894 - 5 August 1981. *Biographical Memoirs of Fellows of the Royal Society*, 28, 378–412.

Mitroff, I. I. & Featheringham, T. R. (1974). On systemic problem solving and the error of the third kind. *Behavioral Sciences*, 19, 383–393.

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *The Annals of Mathematical Statistics*, 19, 58–65.

Neave, H. R. (1976). The teaching of hypothesis testing. *Journal of Applied Statistics*, 3, 55–63.

Read, C. B. (1997). Florence Nightingale. In Johnson, N. L. & Kotz, S. (Eds.) *Leading personalities in statistical sciences* (pp 311–315). John Wiley & Sons, NY.

Rhode, C. A. (2014). *Introductory Statistical inference with the likelihood function*. Springer, NY.

Sokal, F. J. & Rohlf, R. R. (1981). *Biometry* (2nd ed.). W. H. Freeman and Company, NY.

Steel, R. G. D., Torrie, J. H., & Dickey, D. A. (1997). *Principles and procedures of statistics: a biometrical approach* (3rd ed.). McGraw-Hill Co. Inc., NY.

Sullivan, G. M. & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, 4, 279–282.

Weimer, W. B. (1979). *Notes on the methodology of scientific research*. John Wiley & Sons, NY.

Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. John Wiley and Sons.