Chapter 5

Inference for a Proportion

5.1 Introduction

A **dichotomous population** is a collection of units which can be divided into two nonoverlapping subcollections corresponding to the two possible values of a dichotomous variable, *e.g.* male or female, dead or alive, pass or fail. It is conventional to refer to one of the two possible values which dichotomize the population as "success" and the other as "failure." These generic labels are not meant to imply that success is good. Rather, we can think of choosing one of the two possible classifications and asking "does the unit belong to the subcollection of units with this classification?" with the two possibilities being yes (a success) and no (a failure). Thus, generically, we can refer to the two subcollections of units which comprise the dichotomous population as the **success group** and the **failure group**. When a unit is selected from the population and the unit is found to be a member of the success group we say that a **success** has occurred. Similarly, when a member of the failure group is selected we say that a **failure** has occurred.

The proportion of units in the population that belong to the success group is the **population success proportion**. This population success proportion is denoted by the lower case letter p. The population success proportion p is a parameter, since it is a numerical characteristic of the population. Notice that the **population failure proportion** $1 - p$ is also a parameter.

The **sample success proportion** or observed proportion of successes in a sample from a dichotomous population is denoted by $\hat{p}$ (read this as p hat). The observed proportion of successes in the sample $\hat{p}$ is a statistic, since it is a numerical characteristic of the sample.

We will consider two forms of inference about the population success proportion p of a dichotomous population. In Section 5.2 we will consider the use of the observed success proportion $\hat{p}$ to estimate the value of the population success proportion p. In Section 5.3 we will consider the use of the observed success proportion $\hat{p}$ to assess the support for conjectures about the value of the population success proportion p.

The approach to inference that we will use here and in other contexts in the sequel is based on the observed value of a statistic and the sampling distribution of the statistic. The **sampling distribution** of a statistic is the distribution of the possible values of the statistic that could be obtained from random samples. We can think of the sampling distribution of a statistic as a theoretical relative frequency distribution for the possible values of the statistic which describes the sample to sample variability in the statistic. The form of the sampling distribution of a statistic depends on the nature of the population the sample is taken from, the size of the sample, and the method used to select the sample.

The mean and the standard deviation of the sampling distribution are of particular interest. The mean of the sampling distribution indicates whether the statistic is biased as an estimator of the parameter of interest. If the mean of the sampling distribution is equal to the parameter of interest, then the statistic is said to be **unbiased** as an estimator of the parameter. Otherwise, the statistic is said to be **biased** as an estimator of the parameter. To say that a statistic is **unbiased** means that, even though the statistic will overestimate the parameter for some samples and will underestimate the parameter for other samples, it will do so in such a way that, in the long run, the values of the statistic will average to give the correct value of the parameter. When the statistic is **biased** the statistic will tend to consistently overestimate or consistently underestimate the parameter; therefore, in the long run, the values of a biased statistic will not average to give the correct value of the parameter. The standard deviation of the sampling distribution is known as the **standard error** of the statistic. The standard error of the statistic provides a measure of the sample to sample variability in the values of the statistic. The standard error of the statistic can be used to quantify how close we can expect the value of the statistic to be to the value of the parameter.

Note regarding formulae and calculations. Throughout this book selected formulae and intermediate calculations are provided to clarify ideas and definitions. Some readers may find it useful to reproduce these calculations; however, this is not necessary, since a modern statistical calculator or computer statistics program will perform these calculations and provide the desired answer.

5.2 Estimating a proportion

When sampling from a dichotomous population a primary goal is to estimate the population success proportion p, *i.e.*, to estimate the proportion of units in the population success group. The observed proportion of successes in the sample $\hat{p}$ is the obvious estimate of the corresponding population success proportion p.

Clearly there will be some variability from sample to sample in the computed values of the statistic $\hat{p}$. That is, if we took several random samples from the same dichotomous population, we would not expect the computed sample proportions, the $\hat{p}$'s, to be exactly the same. Two questions about $\hat{p}$ as an estimator of p that we might ask are: (1) Can we expect the sample success proportion $\hat{p}$ to be close to the population success proportion p? and (2) Can we quantify how close $\hat{p}$ will be to p? The sampling distribution of $\hat{p}$, which describes the sample to sample variability in $\hat{p}$, can be used to address these questions.

In the introduction to this chapter we mentioned that the sampling distribution of a statistic depends on the way in which the sample is selected, as well as the nature of the population being sampled. Therefore, before we continue with our discussion of the behavior of $\hat{p}$ as an estimator of p we need to describe a model for sampling from a dichotomous population. This model will be presented in terms of a sequence of n trials. In this context a **trial** is a process of observation or experimentation which results in one of two distinct outcomes (success or failure).

A sequence of n trials is said to constitute a sequence of **n Bernoulli trials with success probability p** if the following conditions are satisfied.

1. There is a common probability of success p for every trial. That is, on every trial the probability that the outcome of the trial will be a success is p.
2. The outcomes of the trials are independent of each other. That is, if we knew the outcome of a particular trial or trials this would provide no additional information about the probability of observing a success (or failure) on any other trial. For example, if we knew that a success (or failure) occurred in the first trial, this would not change the probability of success in any other trial.

The simple examples described below will help to clarify the definition of a sequence of n Bernoulli trials and the connection between sampling from a dichotomous population and Bernoulli trials.

Example. Tossing a fair die. Let a trial consist of tossing a fair (balanced) die and observing the number of dots on the upturned face. Define a success to be the occurrence of a 1, 2, 3, or 4. Since the die is fair, the probability of a success on a single trial is $p = 4/6 = 2/3$. Furthermore, if the die is always tossed in the same fashion, then the outcomes of the trials are independent. Therefore, with success defined as above, tossing the fair die n times yields a sequence of n Bernoulli trials with success probability $p = 2/3$.

Example. Drawing balls from a box. Consider a box containing balls (the population) of which 2/3 are red (successes) and 1/3 are green (failures). Suppose that a simple random sample of size n is selected with replacement from this box. That is, a ball is selected at random, its color is recorded, the ball is returned to the box, the balls in the box are mixed, and this process is repeated until n balls have been selected. Thinking of each selection of a ball as a trial we see that this procedure is abstractly the same as the die tossing procedure described above. That is, the outcomes of the draws are independent, and every time that a ball is drawn the probability of a success (drawing a red ball) is $p = 2/3$. Therefore, selecting a simple random sample of n balls with replacement from this collection of balls can be viewed as observing a sequence of n Bernoulli trials with success probability $p = 2/3$. In general, taking a simple random sample of size n selected with replacement from a population with success proportion p can be viewed as observing a sequence of n Bernoulli trials with success probability p.

Situations like the die tossing example above do not fit into the sample and population setup that we have been using. That is, in the die tossing example there is not a physical population of units from which a sample is obtained. In a situation like this we can think of the outcomes of the n Bernoulli trials (the collection of successes and failures that make up the sequence of outcomes of the n trials) as a sample of values of a variable. The probability model specifies that the probability of success on a single trial is p and the probability of failure is $1-p$. This model describes the population of possible values of the variable. Therefore, we can envision a dichotomous population of values (successes and failures) such that the population success proportion is p; and we can think of the outcome of a single trial as the selection of one value at random from this dichotomous population of values. With this idea in mind, we see that the success probability p of this probability model is a parameter and the observed proportion of successes in the n trials is a statistic.

Returning to our discussion of the sampling distribution of $\hat{p}$ we first present two important properties of this sampling distribution. The observed proportion of successes $\hat{p}$ in a sequence of n Bernoulli trials with success probability p (or equivalently the observed proportion of successes $\hat{p}$ in a simple random sample selected with replacement from a dichotomous population with population success proportion p) has a sampling distribution with the following properties.

1. The mean of the sampling distribution of $\hat{p}$ is the population success probability p. Therefore, $\hat{p}$ is unbiased as an estimator of p.
2. The **population standard error** of $\hat{p}$, denoted by S.E.($\hat{p}$), is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

The population standard error of $\hat{p}$ depends on n, which will be known, and p, which will be unknown. Notice that the population standard error gets smaller when n gets larger. That is, when sampling from a fixed dichotomous population, the variability in $\hat{p}$ as an estimator of p is smaller for a larger sample size than it is for a smaller sample size. This property reflects the fact that a larger sample provides more information than a smaller sample. The dependence of the population standard error of $\hat{p}$ on the population success probability p is more complicated. The quantity $p(1-p)$ attains its maximum value of $1/4$ when $p = 1/2$ and approaches zero as p approaches zero or one. Therefore, for a fixed sample size n, there will be more variability in $\hat{p}$ as an estimator of p when p is close to $1/2$ than there will be when p is close to zero or one. This behavior reflects the fact that a dichotomous population is most homogeneous when p is near the extremes $p = 0$ and $p = 1$, and is least homogeneous when p is close to $1/2$.

In many sampling situations the sample is not selected with replacement. For example, in an opinion poll we would not allow the same person to respond twice. We will now consider the effects of sampling without replacement.

Example. Drawing balls from a box (revisited). We will now consider how the ball drawing example from above changes when the simple random sample is selected without replacement. As before, let the box containing the balls (the population) be such that 2/3 are red (successes) and 1/3 are green (failures). However, suppose that the simple random sample of n balls is selected without replacement. That is, a ball is selected at random and its color is recorded and this process is repeated, without returning the ball to the box, until n balls have been selected. It is readily verified that the resulting simple random sample of size n selected without replacement cannot be viewed as a sequence of n Bernoulli trials. To see this suppose that the box contains 12 balls of which 8 are red and 4 are green. The probability of selecting a red ball on the first draw, denoted by $P(\text{red first})$, is $P(\text{red first}) = 8/12 = 2/3$. The probability that the second ball drawn is red clearly depends on the color of the first ball that was drawn. If the first ball drawn was red, then $P(\text{red second given red first}) = 7/11$. However, if the first ball drawn was green, then $P(\text{red second given green first}) = 8/11$. Notice that these probabilities are not the same and neither of them is equal to the population success proportion $p = 2/3$. Therefore, when the sample is selected without replacement, the sampling process is not the same as observing a sequence of Bernoulli trials, since the draws are not independent (the probability of drawing a red ball (success) depends on what happened in the earlier draws) and, as a consequence of this lack of independence, the probability of red (success) is not the same on each draw (trial).

The sampling distribution of the observed success proportion $\hat{p}$ is not the same when $\hat{p}$ is based on a sample selected without replacement as it is when $\hat{p}$ is based on a sample selected with replacement. In both sampling situations, the mean of the sampling distribution of $\hat{p}$ is the population success proportion p. Thus $\hat{p}$ is unbiased as an estimator of p whether the sample is selected with or without replacement. On the other hand, the standard error of $\hat{p}$ is not the same when $\hat{p}$ is based on a sample selected without replacement as it is when $\hat{p}$ is based on a sample selected with replacement. (The standard error of $\hat{p}$ is smaller when the sample is selected without replacement than it is when the sample is selected with replacement.) More specifically, unlike the formula for the standard error of $\hat{p}$ when the sample is selected with replacement which does not depend on the size of the population being sampled, when sampling without replacement the standard error of $\hat{p}$ depends on the size of the population. Fortunately, if the size of the population is very large relative to the size of the sample, then, for practical purposes, the probability of obtaining a success is essentially constant, the outcomes of the draws are essentially independent, and we can use the standard error formula based on the assumption of sampling with replacement even though the sample was selected without replacement.

Remark. *When $\hat{p}$ is computed from a simple random sample of size n selected without replacement from a dichotomous population of size N, the population standard error of $\hat{p}$, S.E.$(\hat{p}) = \sqrt{fp(1-p)/n}$, is smaller than the population standard error for a sample selected with replacement by a factor of $\sqrt{f}$, where $f = (N-n)/(N-1)$. The factor f is known as the finite population correction factor and its effect is most noticeable when N is small relative to n. If N is very large relative to n, then $f \approx 1$ and the two standard errors are essentially equal. Actually, if N is very large relative to n and the data correspond to a simple random sample, then the sampling distribution of $\hat{p}$ is essentially the same whether the sample is selected with or without replacement.*

The sampling distribution of $\hat{p}$ can be represented in tabular form as a probability distribution or in graphical form as a probability histogram. The **probability distribution** of $\hat{p}$ is a theoretical relative frequency distribution which indicates the probability or theoretical relative frequency with which each of the possible values of $\hat{p}$ will occur. The **probability histogram** of $\hat{p}$ is the theoretical relative frequency histogram corresponding to the probabilities (theoretical relative frequencies) in the probability distribution. It is possible to determine the exact sampling distribution of $\hat{p}$, in fact, it is even possible to find a formula which gives the probabilities of each of the possible values of $\hat{p}$. However, for our purposes it is more convenient to work with an approximation to the sampling distribution of $\hat{p}$. (The exact sampling distributions of $\hat{p}$ are discussed in Chapter 4a.) Before we discuss this approximation, which is based on the standard normal distribution, we need to briefly discuss the standard normal distribution.

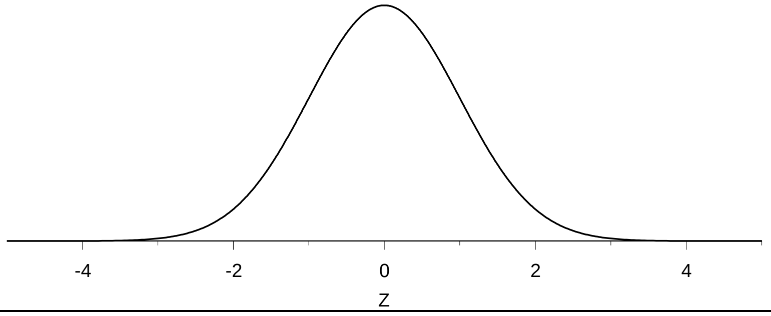
The normal distribution is widely used as a model for the probability distribution of a continuous variable. We will discuss normal distributions in general in more detail in Chapter 7. Here we will restrict our attention to the standard normal distribution and its use as an approximation to the sampling distribution of $\hat{p}$.

Before we discuss the standard normal distribution we first need to briefly consider the representation of a continuous probability model via a density curve. A **density curve** is a nonnegative curve for which the area under the curve (over the x–axis) is one. We can think of the density curve as a smooth version of a theoretical probability histogram with the rectangles of the histogram replaced by a smooth curve indicating where the tops of the rectangles would be. With a continuous variable it does not make sense to talk about the probability that the variable would take on a particular value, after all if we defined positive probabilities for the infinite collection (continuum) of possible values of the variable these probabilities could not add up to one. It does, however, make sense to talk about the probability that the variable will take on a value in a specified range. Given two constants $a < b$ the probability that the variable will take on a value in the interval from a to b is equal to the area under the density curve over the interval from a to b on the x–axis. In this fashion the density curve gives the probabilities which a single value of the variable, chosen at random from the infinite population of possible values, will satisfy.

When we use a continuous probability model to approximate the distribution of a discrete statistic, such as $\hat{p}$, we use the area under the density curve, over the appropriate interval on the number line, to approximate the area in the discrete probability histogram over the same interval. The idea here is that, if the density curve of the approximating continuous distribution matches the discrete probability histogram well, then the area under the density curve will provide a good approximation of the corresponding area in the histogram.

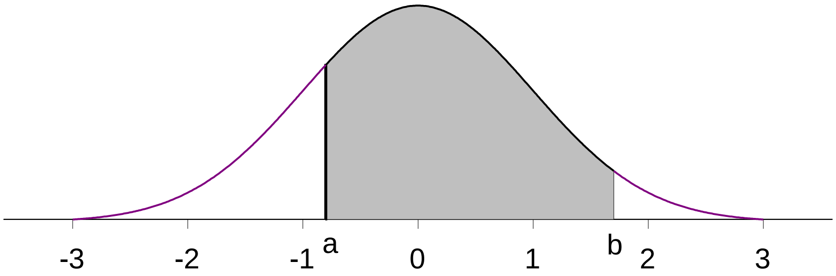
We will now discuss the standard normal distribution which we will use to approximate the sampling distribution of $\hat{p}$. The standard normal distribution can be characterized by its density curve which is the familiar bell shaped curve exhibited in Figure 1. The standard normal distribution and its density curve are symmetric around zero, *i.e.*, if we draw a vertical line through zero in Figure 1, then the two sides of the density curve are mirror images of each other. From Figure 1 it may appear that the standard normal density curve ends at -3 and 3; however, this density curve is actually positive (above the x–axis) for all possible values. The area under the standard normal density curve from -3 to 3 is .9974; thus, there is a small but positive area under the density curve outside of the interval from -3 to 3.

Figure 1. The standard normal density curve.



We will use the upper case letter Z to denote a variable which behaves in accordance with the standard normal distribution and we will refer to such a Z as a standard normal variable. The probability that the standard normal variable Z will take on a value between a and b, denoted by $P(a \leq Z \leq b)$ (read this as the probability that Z is between a and b), is the area under the standard normal density curve over the interval from a to b. A probability of the form $P(a \leq Z \leq b)$ is depicted, for particular values of a and b, as the area of the shaded region in Figure 2.

Figure 2. $P(a \leq Z \leq b)$**, drawn for** $a < 0$ **and** $b > 0$**.**



Computer programs and many calculators can be used to compute standard normal probabilities or equivalently to compute areas under the standard normal density curve. These probabilities can also be calculated using tables of standard normal distribution probabilities. We will not need to perform such calculations here.

The inferential methods we will consider are based on a large sample size normal approximation to the sampling distribution of $\hat{p}$. The normal approximation to the sampling distribution of $\hat{p}$, which is stated formally below, simply says that, for large values of n, the standardized value of $\hat{p}$ obtained by subtracting the population success proportion p from $\hat{p}$ and dividing this difference by the population standard error of $\hat{p}$, behaves in approximate accordance with the standard normal distribution. That is, for large values of n the quantity $(\hat{p}-p)/\text{S.E.}(\hat{p})$ behaves in approximate accordance with the standard normal distribution.

The normal approximation to the sampling distribution of $\hat{\mathbf{p}}$. *Let $\hat{p}$ denote the observed proportion of successes in a sequence of n Bernoulli trials with success probability p (or equivalently the observed proportion of successes in a simple random sample drawn with replacement from a dichotomous population with population success proportion p) and let $a < b$ be two given constants. If n is sufficiently large, then the probability that $(\hat{p}-p)/S.E.(\hat{p})$ is between a and b is approximately equal to the probability that a standard normal variable Z is between a and b. In symbols, using $\approx$ to denote approximate equality, the conclusion from above is that, for sufficiently large values of n,*

$$P\left(a \leq \frac{\hat{p}-p}{S.E.(\hat{p})} \leq b\right) \approx P(a \leq Z \leq b).$$

Remark. *If the population being sampled is very large relative to the size of the sample, then, for practical purposes, this normal approximation to the sampling distribution of $\hat{p}$ may also be applied when $\hat{p}$ is based on a simple random sample selected without replacement.*

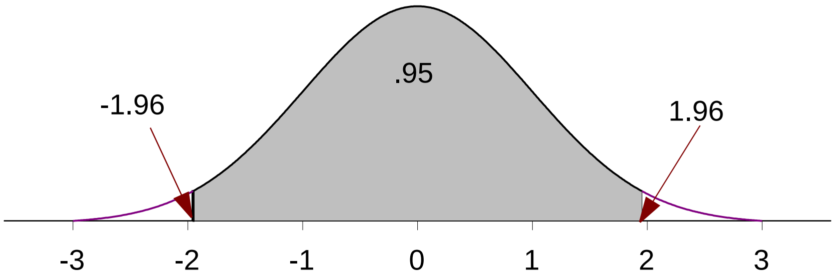
If we apply this approximation with $a = -k$ and $b = k$, then we find that the probability that $\hat{p}$ will take on a value within k population standard error units of p

(within kS.E.$(\hat{p})$ units of p) is approximately equal to the probability that a standard normal variable Z will take on a value between $-k$ and k, *i.e.*,

$$P\big(|\hat{p} - p| \le k\text{S.E.}(\hat{p})\big) = P\big(p - k\text{S.E.}(\hat{p}) \le \hat{p} \le p + k\text{S.E.}(\hat{p})\big) \approx P(-k \le Z \le k).$$

The most commonly used choice of the constant k in this probability statement is $k = 1.96$. The probability that the standard normal variable takes on a value between -1.96 and 1.96 is equal to .95, *i.e.*, $P(-1.96 \le Z \le 1.96) = .95$; therefore, the probability that $\hat{p}$ will take on a value within 1.96 population standard error units of p is approximately .95. This probability is indicated graphically as the shaded region of area .95 in Figure 3. Two other common choices of the constant k in this probability statement are $k = 1.645$ and $k = 2.576$, which give probabilities (areas) of .90 and .99, respectively.

Figure 3. $P(-1.96 \le Z \le 1.96) = .95$



We now return to our discussion of estimating the population success proportion. The following discussion is under the assumption that the data come from a simple random sample of size n drawn with replacement from a dichotomous population with population success proportion p or equivalently that the data correspond to a sequence of n Bernoulli trials with success probability p. For practical purposes, the confidence interval estimates described below are also applicable when the data come from a simple random sample of size n drawn without replacement, provided the population is very large.

Remark. *The basic ideas underlying the inferential methods discussed in this chapter can be used to formulate confidence intervals and hypothesis tests when the data correspond to more complex types of random samples. However, the inferential methods discussed in this chapter are not appropriate for most national opinion polls and sample surveys which rely on complex stratified and/or cluster sampling.*

The observed proportion of successes in our sample $\hat{p}$ provides a single number estimate of the population success probability p. We can think of $\hat{p}$ as our “best guess” of the value of p. From the sampling distribution of $\hat{p}$ we know that $\hat{p}$ is unbiased as an estimator of p; therefore, on the average in the long run (under repeated sampling) we know that $\hat{p}$ provides a good estimate of the unknown parameter p. This unbiasedness, however, does not guarantee that the observed value of $\hat{p}$, based on a single sample, will be close to the true, unknown value of p.

Instead of reporting a single estimate of the unknown population success proportion p it would be more useful to report a range or interval of plausible values for p. In particular, given the data we would like to be able to say, with a reasonable level of confidence, that the true value of p is between two particular values. A **confidence interval estimate of** p consists of two parts. There is an interval of plausible values for p and a corresponding level of confidence. The **confidence level** indicates our confidence that the unknown p actually belongs to the corresponding interval. We will adopt the usual convention of using a confidence level of 95%. A 95% confidence interval estimate of p is an interval of plausible values for p constructed using a method which guarantees that 95% of such intervals will actually contain the unknown proportion p. That is, a 95% confidence interval is an interval constructed using a method of generating such intervals with the property that this method will work, in the sense of generating an interval that contains p, for 95% of all possible samples.

The starting point for using the normal approximation to the sampling distribution of $\hat{p}$ to construct a 95% confidence interval estimate of p is the approximate probability statement

$$P\big[|\hat{p}-p| \leq 1.96\text{S.E.}(\hat{p})\big] \approx .95.$$

This probability statement indicates that the probability that the statistic $\hat{p}$ is within 1.96S.E.($\hat{p}$) units of the parameter p is approximately .95. In other words, when we take a simple random sample and compute $\hat{p}$ the value we get will be within 1.96S.E.($\hat{p}$) of the true p approximately 95% of the time. This is equivalent to saying that the probability that the parameter p is within 1.96S.E.($\hat{p}$) units of the statistic $\hat{p}$ is approximately .95, which is exactly the type of statement we are looking for. Unfortunately, this interval of values is not computable, since it involves the population standard error S.E.($\hat{p}$) which depends on the unknown parameter p and is, therefore, also unknown.

We will consider two methods of forming a confidence interval for p. For ease of notation and greater generality we will let C denote the desired confidence level and k the corresponding cutoff point for the standard normal distribution, *i.e.*, C and k are chosen such that $P(-k \leq Z \leq k) = C$, where Z denotes a standard normal variable. (Some common choices of C and k are: $C = .95$ and $k = 1.96$ for 95% confidence, $C = .90$ and $k = 1.645$ for 90% confidence, and $C = .99$ and $k = 2.576$ for 99% confidence.) In terms of these symbols the starting point for using the normal approximation to the sampling distribution of $\hat{p}$ to construct a confidence interval estimate of p is the approximate probability statement

$$P\big[|\hat{p}-p| \leq k\text{S.E.}(\hat{p})\big] \approx C.$$

The first confidence interval estimate we consider is **the Wilson interval**. This interval estimate is obtained by re–expressing the basic inequality

$$|\hat{p} - p| \leq k\text{S.E.}(\hat{p})$$

as an interval of values for p. The **Wilson confidence interval** is given by

$$\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{M.E.}(\tilde{p}_k),$$

(read $\tilde{p}_k$ as p tilde sub k) where

$$\tilde{p}_k = \frac{n\hat{p} + \frac{k^2}{2}}{n + k^2}$$

determines the center of the interval, and the **margin of error of** $\tilde{p}_k$

$$\text{M.E.}(\tilde{p}_k) = \sqrt{\tilde{p}_k^2 - \frac{n\hat{p}^2}{n + k^2}}$$

determines the length of the interval.

If we use $k = 1.96$ in these expressions, then we can claim that we are 95% confident that the population success proportion p is between $\tilde{p}_k - \text{M.E.}(\tilde{p}_k)$ and $\tilde{p}_k + \text{M.E.}(\tilde{p}_k)$. There is some chance for confusion about what this statement actually means. The important thing to remember is that it is the statistic $\tilde{p}_k$ and the margin of error $\text{M.E.}(\tilde{p}_k)$ that vary from sample to sample. The population proportion p is a fixed, unknown parameter which does not vary. Therefore, the 95% confidence level applies to the method used to generate the confidence interval estimate. That is, the method (obtain a simple random sample and compute the numbers $\tilde{p} - \text{M.E.}(\tilde{p})$ and $\tilde{p} + \text{M.E.}(\tilde{p})$) used to generate the limits of the confidence interval is such that 95% of the time it will yield a pair of confidence interval limits which bracket the population success proportion p. Therefore, when we obtain a sample, compute the confidence interval, and say that we are 95% confident that this interval contains p what we mean is that we feel "pretty good" about claiming that p is in this interval, since the method works for 95% of all possible samples and so it probably worked for our sample.

Derivation of the Wilson interval. Since $|\hat{p} - p| \geq 0$ and $\text{S.E.}(\hat{p}) = \sqrt{p(1-p)/n}$, we can square each side of the basic inequality to get the equivalent inequality

$$(\hat{p} - p)^2 \leq \frac{k^2}{n}(p - p^2).$$

Straightforward algebra allows us to re–express this inequality as the following quadratic inequality in p

$$(n + k^2)p^2 - 2(n\hat{p} + \frac{k^2}{2})p + n\hat{p}^2 \leq 0.$$

Treating this inequality as an equality and solving for p gives the two values

$$\tilde{p}_k \pm \text{M.E.}(\tilde{p}_k),$$

$$\text{where} \quad \tilde{p}_k = \frac{n\hat{p} + \frac{k^2}{2}}{n + k^2} \quad \text{and} \quad \text{M.E.}(\tilde{p}_k) = \sqrt{\tilde{p}_k^2 - \frac{n\hat{p}^2}{n + k^2}}.$$

Thus the original probability statement

$$P\big[|\hat{p} - p| \leq k\text{S.E.}(\hat{p})\big] \approx C.$$

is equivalent to the probability statement

$$P\big[\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{M.E.}(\tilde{p}_k)\big] \approx C.$$

The endpoints of this interval, which are functions of $n, \hat{p}$, and k, are computable. Therefore, the Wilson confidence interval is given by

$$\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{M.E.}(\tilde{p}_k).$$

It is somewhat tedious to compute the Wilson interval by hand; but it is easy to program a calculator or computer to do the computations. An easy to compute approximation (the **Agresti–Coull interval**) to the Wilson 95% confidence interval is described after the following examples.

Example. Insects in an apple orchard. The manager of a large apple orchard is concerned with the presence of a particular insect pest in the apple trees in the orchard. An insecticide that controls this particular insect pest is available. However, application of this insecticide is rather expensive. It has been determined that the cost of applying the insecticide is not economically justifiable unless more than 20% of the apple trees in the orchard are infested. The manager has decided to assess the extent of infestation in the orchard by examining a simple random sample of 200 apple trees. In this example a unit an apple tree and the target population is all of the apple trees in this orchard. We will assume that the simple random sample is selected from all of the apple trees in the orchard so that the sampled population is the same as the target population. We will also assume that the 200 trees in the sample form a small proportion of all of the trees in the entire orchard so that we do not need to worry about whether the sample is chosen with or without replacement. An appropriate dichotomous variable is whether an apple tree is infested with possible values of yes (the tree is infested) and no (the tree is not infested). Since we are interested in the extent of the infestation we will view a tree that is infested as a success. Thus, the population success proportion p is the proportion of all of the apple trees in this orchard that are infested.

Two (related) questions of interest in this situation are:
(1) What proportion of all of the trees in this orchard are infested? (What is p?)
(2) Is there sufficient evidence to justify the application of the insecticide? (Is $p > .20$?)
We will consider four hypothetical outcomes for this scenario to demonstrate how a 95% confidence interval estimate can be used to address these questions.

Case 1. Suppose that 35 of the 200 apple trees in the sample are infested so that $\hat{p} = .175$. In this case we know that 17.5% of the 200 trees in the sample are infested and we can conjecture that a similar proportion of all of the trees in the entire orchard are infested. However, we need a confidence interval estimate to get a handle on which values of the population success proportion p are plausible when we observe 17.5% infested trees in a sample of size 200. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .1811$ and a 95% confidence interval ranging from .1286 to .2336. Thus we can conclude that we are 95% confident that between 12.86% and 23.36% of all of the trees in this orchard are infested. Notice that this confidence interval does not exclude the possibility that more than 20% of the trees in the entire orchard are infested, since the upper limit of the confidence interval 23.36% is greater than 20%. In other words, even though less than 20% of the trees in the sample were infested when we take sampling variability into account we find that it is possible that more than 20% (as high as 23.36%) of the trees in the entire orchard are infested.

Case 2. Suppose that 26 of the 200 apple trees in the sample are infested so that $\hat{p} = .13$. In this case we know that 13% of the 200 trees in the sample are infested. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .1370$ and a 95% confidence interval ranging from .0903 to .1837. Thus we can conclude that we are 95% confident that between 9.03% and 18.37% of all of the trees in this orchard are infested. In this case the entire confidence interval is below 20% excluding the possibility that more than 20% of the trees in the entire orchard are infested. Therefore, in this case we have sufficient evidence to conclude that less than 20% of the trees in the entire orchard are infested, *i.e.*, that $p < .20$.

Case 3. Suppose that 45 of the 200 apple trees in the sample are infested so that $\hat{p} = .225$. In this case we know that 22.5% of the 200 trees in the sample are infested. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .2302$ and a 95% confidence interval ranging from .1726 to .2877. Thus we can conclude that we are 95% confident that between 17.26% and 28.77% of all of the trees in this orchard are infested. Notice that this confidence interval does not exclude the possibility that less than 20% of the trees in the entire orchard are infested, since the lower limit of the confidence interval 17.26% is less than 20%. In other words, even though mores than 20% of the trees in the sample were infested when we take sampling variability into account we find that it is possible that less than 20% of the trees in the entire orchard are infested.

Case 4. Suppose that 54 of the 200 apple trees in the sample are infested so that $\hat{p} = .27$. In this case we know that 27% of the 200 trees in the sample are infested. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .2743$ and a 95% confidence interval ranging from .2132 to .3354. Thus we can conclude that we are 95% confident that between 21.32% and 33.54% of all of the trees in this orchard are infested. In this case the entire confidence interval is above 20% excluding the possibility that less than 20% of the trees in the entire orchard are infested. Therefore, in this case we have sufficient evidence to conclude that more than 20% of the trees in the entire orchard are infested, *i.e.*, that $p > .20$.

Example. Opinions about a change in tax law. Consider a public opinion poll conducted to assess the support for a proposed change in state tax law among the taxpayers in a particular metropolitan area. The target population is the group of approximately 200,000 taxpayers in the particular metropolitan area. Suppose that the opinion poll is conducted as follows: first a simple random sample of 100 taxpayers in the metropolitan area is obtained, restricting the sample to taxpayers who have telephones, then these 100 taxpayers are contacted by telephone and each person is asked to respond to the question "Do you favor or oppose the proposed change in state tax law?" In this example we will define a unit to be an individual taxpayer in this metropolitan area. (Note that, technically, a unit is a household, since more than one taxpayer may share the same telephone number.) The variable is the response of the taxpayer to the indicated question with possible values of: "I favor the change," "I oppose the change," and "I do not have an opinion regarding the change." We will dichotomize this variable (and the population) by recording the responses as either "I favor the change" or "I do not favor the change." Notice that in this example the target and sampled populations are not the same. Since there might well be a relationship between having a telephone and opinion about the proposed tax law change, we will restrict our attention to the sampled population of all taxpayers in this metropolitan area who have telephones. The parameter of interest is the proportion p of all taxpayers in this metropolitan area who have a telephone who favor the proposed tax law change at the time of the survey. In this example the random sample would be selected without replacement. However, since the size of the population, approximately 200,000, is much larger than the sample size $n = 100$, we can use the confidence interval estimation procedure as described above.

Suppose that the poll was conducted and 55 of the 100 taxpayers in the sample responded that they favor the tax law change. The observed proportion who favor the change is thus $\hat{p} = .55$, *i.e.*, 55% of the 100 taxpayers in the sample favored the change. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .5482$ and a 95% confidence interval ranging from .4524 to .6438. Therefore, we are 95% confident that the actual proportion of taxpayers in this metropolitan area (who have telephones) who favored the proposed change in state tax law at the time of the survey is between 45.24% and 64.38%. Notice that this confidence interval contains values for p that are both less than .5 and greater than .5. Therefore, based on this outcome of the opinion poll there is not sufficient evidence to conclude that more than half of the taxpayers in this metropolitan area (who have telephones) favored the proposed tax law change at the time of this poll.

Now suppose that the poll was conducted and 64 of the 100 taxpayers in the sample responded that they favor the tax law change. The observed proportion who favor the change is thus $\hat{p} = .64$, *i.e.*, 64% of the 100 taxpayers in the sample favored the change. Using the Wilson method with $k = 1.96$ we get $\tilde{p}_k = .6348$ and a 95% confidence interval ranging from .5424 to .7273. Therefore, we are 95% confident that the actual proportion of taxpayers in this metropolitan area (who have telephones) who favored the proposed change in state tax law at the time of the survey is between 54.24% and 72.73%. In this case all of the values for p in the confidence interval are greater than .5. Therefore, based on this outcome of the opinion poll there is sufficient evidence to conclude that more than half of the taxpayers in this metropolitan area (who have telephones) favored the proposed tax law change at the time of this poll. However, we would also note that, based on this confidence interval, the actual percentage in favor of the change might be as small as 54.24%.

In the preceding analysis of this opinion poll example we dichotomized the responses to the question by combining the "oppose" and "no opinion" responses as "do not favor". As an alternative we might prefer to restrict our attention to only those people who are willing to express a definite opinion by restricting our inference to the subpopulation of taxpayers who would have been willing to respond "I favor the change" or "I oppose the change" at the time of the survey. Thus we redefine the population success proportion p as the proportion of all taxpayers in this metropolitan area (who have a telephone and have reached an opinion at the time of the survey) who favor the proposed tax law change at the time of the survey. To implement this approach we simply ignore the part of the sample for which the respondents did not express an opinion, redefine the sample size as n^* the number who responded "favor" or "oppose", and compute the confidence interval conditional on the reduced sample size n^*. For example, if $n = 100$ and if 64 taxpayers favor the change, 26 taxpayers oppose the change, and 10 taxpayers have no opinion, then we restrict our attention to the $n* = 64 + 26 = 90$ taxpayers who expressed an opinion. For this sample the observed proportion who favor the change is $\hat{p} = 64/90 = .7111$. Using the Wilson method with $k = 1.96$ (and $n* = 90$) we get $\tilde{p}_k = .7025$ and a 95% confidence interval ranging from .6104 to .7946. Therefore, we are 95% confident that the actual proportion of taxpayers in this metropolitan area (who have telephones and have reached an opinion) who favored the proposed tax law change at the time of this poll is between 61.04% and 79.46%.

Remark. *When a confidence interval for a proportion p is based on a simple random sample selected with replacement or a simple random sample selected without replacement from a much larger population the precision of the confidence interval as an estimate of p depends on the absolute size of the sample not the size of the sample relative to the size of the population. For example, if a simple random sample of size $n = 200$ yields $\hat{p} = .65$ (and $k = 1.96$), then $\tilde{p}_k = .6472$, M.E.$(\tilde{p}_k) = .0655$, and we are 95% confident that p is between $.6472 - .0655 = .5817$ and $.6472 + .0655 = .7127$. Any sample of size $n = 200$ for which $\hat{p} = .65$ yields this confidence interval; which has length $2(.0655) = .1310$. Thus if we were sampling from a population of 200,000 or a population of 2,000,000 and if we obtained $\hat{p} = .65$, we would get the same confidence interval. Hence the precision of the confidence interval, as measured by its length, depends on the sample size but does not depend on what fraction of the population was sampled.*

We will now consider a simpler method for computing a confidence interval for p. This confidence interval estimate, known as the **Wald interval**, is in widespread use and many calculators and computer programs will compute it. Unfortunately, this confidence interval estimate has some undesirable properties and we **do not recommend** its use.

As we noted above (for $C = .95$ and $k = 1.96$), the probability statement

$$P\big[|\hat{p} - p| \le k\text{S.E.}(\hat{p})\big] \approx C$$

is equivalent to the probability statement

$$P\big[\hat{p} - k\text{S.E.}(\hat{p}) \le p \le \hat{p} + k\text{S.E.}(\hat{p})\big] \approx C,$$

but this interval of values is not computable, since the population standard error $\text{S.E.}(\hat{p}) = \sqrt{p(1-p)/n}$ depends on the unknown parameter p. The **Wald interval** is obtained by replacing the unknown population standard error by an estimated standard error. The **estimated standard error of** $\hat{p}$

$$\widehat{\text{S.E.}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is obtained by replacing the unknown parameter p in the population standard error by the observable statistic $\hat{p}$. The **Wald confidence interval** is given by

$$\hat{p} - \text{M.E.}(\hat{p}) \le p \le \hat{p} + \text{M.E.}(\hat{p}), \quad \text{where} \quad \text{M.E.}(\hat{p}) = k\widehat{\text{S.E.}}(\hat{p})$$

is the **margin of error of** $\hat{p}$.

Notice that the Wald interval is centered at $\hat{p}$ and its length is determined by the margin of error of $\hat{p}$,

$$\text{M.E.}(\hat{p}) = k\widehat{\text{S.E.}}(\hat{p}).$$

As with the Wilson interval, if we use $k = 1.96$ in these expressions, then we can claim that we are 95% confident that the population success proportion p is between $\hat{p} - \text{M.E.}(\hat{p})$ and $\hat{p} + \text{M.E.}(\hat{p})$. With the same interpretation of "95% confident" as before.

We will now discuss the "undesirable properties" of this interval estimate and the reason we do not recommend it. Even though $\hat{p}$ performs well as a single number estimate of p the Wald confidence interval estimate, based on $\hat{p}$ and $\widehat{\text{S.E.}}(\hat{p})$, does not perform well. When we say that we are 95% confident that the population success proportion p is between $\hat{p} - \text{M.E.}(\hat{p})$ and $\hat{p} + \text{M.E.}(\hat{p})$ we realize that our indicated 95% confidence level is actually an approximation to the true confidence level. For this confidence interval estimate the indicated 95% confidence level differs from the actual confidence level because of the two approximations used to construct this interval, *i.e.*, because of our use of the normal approximation and our use of the estimated standard error. We would hope, at least for reasonably large values of n, that the difference between the indicated 95% confidence level of our interval estimate and its actual confidence level would be small. Unfortunately, this is not necessarily the case and the actual confidence level of this confidence interval estimate may be quite different from the indicated 95%. In particular, the actual confidence level of this 95% confidence interval estimate may be much smaller than 95%. Furthermore, this discrepancy between the indicated 95% confidence level and the actual confidence level is not necessarily negligible even when the sample size n is quite large.

On the other hand, the Wilson confidence interval estimate, based on $\tilde{p}_k$ and $\text{M.E.}(\tilde{p}_k)$, only requires one approximation (the normal approximation) and for this reason it performs better than the Wald confidence interval.

We will now describe the easy to compute approximation (the **Agresti–Coull interval**) to the Wilson 95% confidence interval. If we add 4 artificial observations to the data, 2 success and 2 failures, and then compute the Wald 95% confidence interval, it turns out that we obtain a reasonably accurate approximation of the Wilson 95% confidence interval. More formally, the **Agresti–Coull interval** is obtained by replacing the estimator $\hat{p}$ and its margin of error in the Wald interval by the alternate estimator $\tilde{p}$ (read this as p tilde) and its margin of error. The estimator $\tilde{p}$ is obtained by adding 2 successes and 2 failures to the data, *i.e.*,

$$\tilde{p} = \frac{\text{the number of successes plus 2}}{\text{the number of observations plus 4}} = \frac{n\hat{p} + 2}{n + 4}.$$

The corresponding 95% **margin of error of $\tilde{p}$** is

$$\text{M.E.}(\tilde{p}) = 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}},$$

which is analogous to the margin of error of $\hat{p}$ with $\tilde{p}$ in place of $\hat{p}$ and $n+4$ in place of the actual sample size n. The **Agresti–Coull 95% confidence interval** estimate of p, is given by

$$\tilde{p} - \text{M.E.}(\tilde{p}) \le p \le \tilde{p} + \text{M.E.}(\tilde{p}),$$

where $\tilde{p}$ and M.E.($\tilde{p}$) are as defined above. The reason that this works is that the value of k for a 95% confidence interval is $k = 1.96$ which implies that $k^2/2 \approx 2$ and $k^2 \approx 4$ so that $\tilde{p} \approx \tilde{p}_k$ (for $k = 1.96$) and M.E.($\tilde{p}$) $\approx$ M.E.($\tilde{p}_k$) (for $k = 1.96$). If you have a calculator or computer program which computes the Wald interval, then you can use this "add 2 successes / add 4 observations" trick to approximate the Wilson 95% confidence interval.

5.3 Testing for a proportion

The hypothesis testing procedures discussed in this section are based on the normal approximation to the sampling distribution of $\hat{p}$. Hence we will continue to assume that the data form a simple random sample of size n, selected with replacement, from a dichotomous population with population success proportion p, or equivalently, that the data correspond to the outcomes of a sequence of n Bernoulli trials with success probability p. As before if the population is very large, then these methods can also be used when the data form a simple random sample of size n, selected without replacement.

A **hypothesis** (statistical hypothesis) is a conjecture about the nature of the population. When the population is dichotomous, a hypothesis is a conjecture about the value of the population success proportion p.

A **hypothesis test** (test of significance) is a formal procedure for deciding between two complementary hypotheses. These hypotheses are known as the null hypothesis (H_0 for short) and the research (or alternative) hypothesis (H_1 for short). The research hypothesis is the hypothesis of primary interest, since the testing procedure is designed to address the question: "Do the data support the research hypothesis?" The null hypothesis is defined as the negation of the research hypothesis. The test begins by tentatively assuming that the null hypothesis is true (the research hypothesis is false). The data are then examined to determine whether the null hypothesis can be rejected in favor of the research hypothesis. The probability of observing data as unusual (surprising) or more unusual as that actually observed under the tentative assumption that the null hypothesis is true is computed. This probability is known as the P–value of the test. (The P in P–value indicates that it is a probability it does not refer to the population success proportion p.) A small P–value indicates that the observed data would be unusual (surprising) if the null hypothesis was actually true. Thus if the P–value is small enough, then the null hypothesis is judged untenable and the test rejects the null hypothesis in favor of the research (alternative) hypothesis. On the other hand, a large (not small) P–value indicates that the observed data would not be unusual (not surprising) if the null hypothesis was actually true. Thus if the P–value is large (not small enough), then the null hypothesis is judged tenable and the test fails to reject the null hypothesis.

There is a strong similarity between the reasoning used for a hypothesis test and the reasoning used in the trial of a defendant in a court of law. In a trial the defendant is presumed innocent (tentatively assumed to be innocent) and this tentative assumption is not rejected unless sufficient evidence is provided to make this tentative assumption untenable. In this situation the research hypothesis states that the defendant is guilty and the null hypothesis states that the defendant is not guilty (is innocent). The P–value of a hypothesis test is analogous to a quantification of the weight of the evidence that the defendant is guilty with small values indicating that the evidence is unlikely under the assumption that the defendant is innocent.

Example. Insects in an apple orchard (revisited). Recall that the manager of a large apple orchard examined a simple random sample of 200 apple trees to gauge the extent of insect infestation in the orchard. The manager has determined that applying the insecticide is not economically justifiable unless more than 20% of the apple trees in the orchard are infested. Since the manager does not want to apply the insecticide unless there is evidence that it is needed, the question of interest here is: "Is there sufficient evidence to justify application of the insecticide?" In terms of the population success proportion p (the proportion of all of the apple trees in this orchard that are infested) the research hypothesis is $H_1 : p > .20$ (more than 20% of all the trees in the orchard are infested); and the null hypothesis is $H_0 : p \leq .20$ (no more than 20% of all the trees in the orchard are infested). A test of the null hypothesis $H_0 : p \leq .20$ versus the research hypothesis $H_1 : p > .20$ begins by tentatively assuming that no more than 20% of all the trees in the orchard are infested. Under this tentative assumption it would be surprising to observe a proportion of infested trees in the sample $\hat{p}$ that was much larger than .20. Thus the test should reject $H_0 : p \leq .20$ in favor of $H_1 : p > .20$ if the observed value of $\hat{p}$ is sufficiently large relative to .20.

Case 1. Suppose that 52 of the 200 apple trees in the sample are infested so that $\hat{p} = .26$. In this case we know that 26% of the 200 trees in the sample are infested and we need to decide whether this suggests that the proportion of all the trees in the orchard that are infested p exceeds .20. More specifically, we need to determine whether observing 52 or more infested trees in a simple random sample of 200 trees would be surprising if in fact no more than 20% of all the trees in the orchard were infested. Assuming that exactly 20% of all the trees in the orchard are infested, we find that the probability of observing 52 or more infested trees in a sample of 200 trees (seeing $\hat{p} \geq .26$), is .0169 (this is the P–value of the test). In other words, if no more than 20% of all the trees in the orchard were infested, then a simple random sample of 200 trees would give $\hat{p} \geq .26$ about 1.69% of the time. Therefore, observing 52 infested trees in a sample of 200 would be very surprising if no more than 20% of all the trees in the orchard were infested and we have sufficient evidence to reject the null hypothesis $H_0 : p \leq .20$ in favor of the research hypothesis $H_1 : p > .20$. In the case we would conclude that there is sufficient evidence to contend that more than 20% of all the trees in the orchard are infested and, in this sense, application of the insecticide is justifiable.

Case 2. Next suppose that 45 of the 200 apple trees in the sample are infested so that $\hat{p} = .225$. Assuming that exactly 20% of all the trees in the orchard are infested, we find that the probability of observing 45 or more infested trees in a sample of 200 trees (seeing $\hat{p} \geq .225$), is .1884 (this is the P–value of the test). In other words, if no more than 20% of all the trees in the orchard were infested, then a simple random sample of 200 trees would give $\hat{p} \geq .225$ about 18.84% of the time. Therefore, observing 45 infested trees in a sample of 200 would not be very surprising if no more than 20% of all the trees in the orchard were infested and we do not have sufficient evidence to reject the null hypothesis $H_0 : p \leq .20$ in favor of the research hypothesis $H_1 : p > .20$. In the case we would conclude that there is not sufficient evidence to contend that more than 20% of all the trees in the orchard are infested and, in this sense, application of the insecticide not is justifiable.

The research hypothesis in the apple orchard example is a directional hypothesis of the form $H_1 : p > p_0$, where $p_0 = .20$. We will now discuss the details of a hypothesis test for a directional research hypothesis of this form. For the test procedure to be valid the specified value p_0 and the direction of the research hypothesis must be motivated from subject matter knowledge before looking at the data that are to be used to perform the test.

Let p_0 denote the hypothesized value (with $0 < p_0 < 1$) which we wish to compare with p. The research hypothesis states that p is greater than p_0 ($H_1 : p > p_0$). The null hypothesis is the negation of $H_1 : p > p_0$ which states that p is no greater than p_0 ($H_0 : p \leq p_0$). The research hypothesis $H_1 : p > p_0$ specifies that the population is one of the dichotomous populations for which the population success proportion p is greater than p_0. The null hypothesis $H_0 : p \leq p_0$ specifies that the population is one of the dichotomous populations for which the population success proportion p is no greater than p_0. Notice that this competing pair of hypotheses provides a decomposition of all possible dichotomous populations into the collection of dichotomous populations where $p > p_0$ and the research hypothesis is true and the collection of dichotomous populations where $p \leq p_0$ and the null hypothesis is true. Our goal is to use the data to decide which of these two collections of dichotomous populations contains the actual population we are sampling from.

Since a hypothesis test begins by tentatively assuming that the null hypothesis is true, we need to decide what constitutes evidence against the null hypothesis $H_0 : p \leq p_0$ and in favor of the research hypothesis $H_1 : p > p_0$. The relationship between the observed proportion of successes in the sample $\hat{p}$ and the hypothesized value p_0 will be used to assess the strength of the evidence in favor of the research hypothesis. Generally, we would expect to observe larger values of $\hat{p}$ more often when the research hypothesis $H_1 : p > p_0$ is true than when the null hypothesis $H_0 : p \leq p_0$ is true. In particular, we can view the observation of a value of $\hat{p}$ that is sufficiently large relative to p_0 as constituting evidence against the null hypothesis $H_0 : p \leq p_0$ and in favor of the research hypothesis $H_1 : p > p_0$. Deciding whether the observed value of $\hat{p}$ is "sufficiently large relative to p_0" is based on the corresponding P–value, which is defined below.

The P–value for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ is the probability of observing a value of $\hat{p}$ as large or larger than the value of $\hat{p}$ that we actually do observe. The P–value quantifies the consistency of the observed data with the null hypothesis and may be interpreted as a, somewhat indirect, measure of the strength of the evidence in the data in favor of the research hypothesis and against the null hypothesis. Because the P–value is computed under the assumption that the null hypothesis is true (and the research hypothesis is false), the smaller the P–value is, the less consistent the observed data are with the null hypothesis. Therefore, since one of the hypotheses must be true, when we observe a small P–value we can conclude that the research hypothesis is more consistent with the observed data than is the null hypothesis.

The P–value is computed under the assumption that the research hypothesis $H_1 : p > p_0$ is false and the null hypothesis $H_0 : p \leq p_0$ is true. Because the null hypothesis only specifies that $p \leq p_0$, we need to choose a particular value of p (that is no larger than p_0) in order to compute the P–value. It is most appropriate to use $p = p_0$ for this computation. (Recall that in the apple orchard example we used $p_0 = .20$ to compute the P–value.) Using $p = p_0$, which defines the boundary between $p \leq p_0$, where the null hypothesis is true, and $p > p_0$, where the research hypothesis is true, provides some protection against incorrectly rejecting $H_0 : p \leq p_0$.

To compute the P–value we need to know how much variability there is in the sampling distribution of $\hat{p}$ when $p = p_0$. When $p = p_0$ the standard error of $\hat{p}$, which provides a suitable measure of the variability in $\hat{p}$, is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{p_0(1-p_0)}{n}}.$$

To use the normal approximation to the sampling distribution of $\hat{p}$ to compute the P–value we first need to determine the calculated Z statistic or Z score corresponding to the observed value of $\hat{p}$. This calculated Z statistic, denoted by Z_{calc}, is

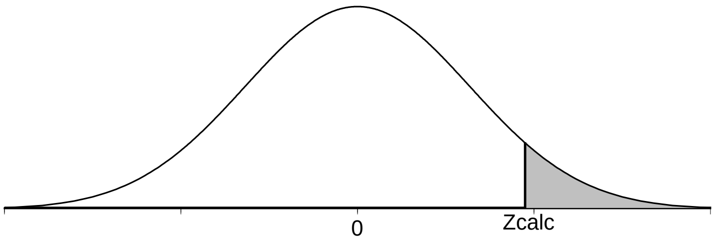
$$Z_{calc} = \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})},$$

where the standard error S.E.$(\hat{p})$ is as defined above. Recall that the P–value for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ is the probability of observing a value of $\hat{p}$ as large or larger than the value of $\hat{p}$ that we actually do observe, computed assuming that $p = p_0$. Using the normal approximation, this P–value is equal to the probability that a standard normal variable takes on a value at least as large as Z_{calc}. This P–value is

$$P\text{–value} = P(Z \geq Z_{calc}),$$

where Z denotes a standard normal variable, *i.e.*, this P–value is the area under the standard normal density curve to the right of Z_{calc}, as shown in Figure 4. Notice that the P value (the area to the right of Z_{calc}) is small when Z_{calc} is far to the right of zero which is equivalent to $\hat{p}$ being far to the right of p_0.

Figure 4. P–value for $H_0 : p \leq p_0$ versus $H_1 : p > p_0$.



Once the P–value has been computed we need to decide whether the P–value is small enough to justify rejecting the null hypothesis in favor of the research hypothesis. In the apple orchard example we argued that observing 52 infested trees in a sample of 200 would be very surprising if no more than 20% of all the trees in the orchard were infested, since the corresponding P–value of .0169 was very small. We also argued that observing 45 infested trees in a sample of 200 would not be very surprising if no more than 20% of all the trees in the orchard were infested, since the corresponding P–value of .1884 is fairly large. Deciding whether a P–value is small enough to reject a null hypothesis requires a subjective judgment by the investigator in the context of the problem at hand.

The following general remarks regarding the use of P–values to assess the evidence against a null hypothesis and in favor of a research hypothesis apply to hypothesis tests in general, not just hypothesis tests for a proportion.

One approach to hypothesis testing is to use a fixed cutoff value to decide whether the P–value is "large" or "small". The most common application of this approach is to conclude that there is sufficient evidence to reject the null hypothesis in favor of the research hypothesis only when the P–value is less than .05. When a fixed cutoff value like .05 (5%) is used to decide whether to reject the null hypothesis in favor of the research hypothesis this cutoff value is known as the **significance level** of the test. Hence, if we adopt the rule of rejecting the null hypothesis in favor of the research hypothesis only when the P–value is less than .05, then we are performing a hypothesis test at the 5% level of significance. In accordance with this terminology, the P–value is also known as the **observed significance level** of the test and if the P–value is less than the prescribed significance level, then the results are said to be **statistically significant**.

To perform a hypothesis test at the 5% level of significance we compute the appropriate P–value and compare it to the fixed significance level .05. If the P–value is less than .05, then we conclude that there is sufficient evidence, at the 5% level of significance, to reject the null hypothesis H_0 in favor of the research hypothesis H_1, *i.e.*, if the P–value **is less than** .05, then the data **do** support H_1. If the P–value is not less than .05, then we conclude that there is not sufficient evidence, at the 5% level of significance, to reject the null hypothesis H_0 in favor of the research hypothesis H_1, *i.e.*, if the P–value **is not less than** .05, then the data **do not** support H_1.

Instead of, or in addition to, using a fixed significance level like 5% we can use the P–value as a measure of the evidence (in the data) against the null hypothesis H_0 and in favor of the research hypothesis H_1. Some guidelines for deciding how strong the evidence is in favor of the research hypothesis H_1 are given below.

Guidelines for interpreting a P–value:

1. If the P–value is greater than .10, there is no evidence in favor of H_1.
2. If the P–value is between .05 and .10, there is suggestive but very weak evidence in favor of H_1.
3. If the P–value is between .04 and .05, there is weak evidence in favor of H_1.
4. If the P–value is between .02 and .04, there is moderately strong evidence in favor of H_1.
5. If the P–value is between .01 and .02, there is strong evidence in favor of H_1.
6. If the P–value is less than .01, there is very strong evidence in favor of H_1.

Whether you choose to use a fixed significance level or the preceding guidelines based on the P–value you should always report the P–value since this allows someone else to interpret the evidence in favor of H_1 using their personal preferences regarding the size of a P–value.

In the U.S. legal system there is a similar set of guidelines for assessing the level of proof or weight of the evidence against the null hypothesis of innocence and in favor of the research hypothesis of guilt. The weakest level of proof is "the preponderance of the evidence" (this is similar to a reasonably small P–value), the next level of proof is "clear and convincing evidence" (this is similar to a small P–value), and the highest level of proof is "beyond a reasonable doubt" (this is similar to a very small P–value).

We now return to our discussion for the particular research hypothesis $H_1 : p > p_0$. The steps for performing a hypothesis test for

$$H_0 : p \le p_0 \quad \text{versus} \quad H_1 : p > p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \ge Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = (\hat{p} - p_0)/\text{S.E.}(\hat{p})$, and $\text{S.E.}(\hat{p}) = \sqrt{p_0(1 - p_0)/n}$. This P–value is the area under the standard normal density curve to the right of Z_{calc}, as shown in Figure 4.

2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p > p_0$ over $H_0 : p \le p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is greater than p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p > p_0$ over $H_0 : p \le p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is greater than p_0.

The procedure for testing the null hypothesis $H_0 : p \le p_0$ versus the research hypothesis $H_1 : p > p_0$ given above is readily modified for testing the null hypothesis $H_0 : p \ge p_0$ versus the research hypothesis $H_1 : p < p_0$. The essential modification is to change the direction of the inequality in the definition of the P–value. Consider a situation where the research hypothesis specifies that the population success proportion p is less than the particular, hypothesized value p_0, *i.e.*, consider a situation where the research hypothesis is $H_1 : p < p_0$ and the null hypothesis is $H_0 : p \ge p_0$. For these hypotheses values of the observed success proportion $\hat{p}$ that are sufficiently small relative to p_0 provide evidence in favor of the research hypothesis $H_1 : p < p_0$ and against the null hypothesis $H_0 : p \ge p_0$. Therefore, the P–value for testing $H_0 : p \ge p_0$ versus $H_1 : p < p_0$ is the probability of observing a value of $\hat{p}$ as small or smaller than the value actually observed. As before, the P–value is computed under the assumption that $p = p_0$. The calculated Z statistic Z_{calc} is defined as before; however, in this situation the P–value is the area under the standard normal density curve to the left of Z_{calc}, since values of $\hat{p}$ that are small relative to p_0 constitute evidence in favor of the research hypothesis.

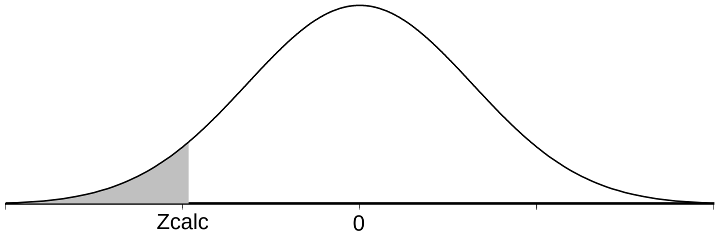
The steps for performing a hypothesis test for

$$H_0 : p \geq p_0 \quad \text{versus} \quad H_1 : p < p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \leq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = (\hat{p} - p_0)/\text{S.E.}(\hat{p})$, and $\text{S.E.}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. This P–value is the area under the standard normal density curve to the left of Z_{calc} as shown in Figure 5.

Figure 5. P–value for $H_0 : p \geq p_0$ versus $H_1 : p < p_0$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p < p_0$ over $H_0 : p \geq p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is less than p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p < p_0$ over $H_0 : p \geq p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is less than p_0.

Example. Acceptance sampling for electronic devices. A large retailer receives a shipment of 10,000 electronic devices from a supplier. The supplier guarantees that no more than 6% of these devices are defective. In fact, if more than 6% of the devices in the shipment are defective, then the supplier will allow the retailer to return the entire shipment, provided this is done with 10 days of receiving the shipment. Therefore, the retailer needs to decide between accepting the shipment and returning the shipment to the supplier. This decision will be based on the information provided by examining a simple random sample of electronic devices selected from the shipment.

In this example one of these electronic devices is a unit and the collection of 10,000 units constituting the shipment is the population. Notice that, in this example, the target population and the sampled population are the same (each is the shipment of 10,000 devices). A suitable variable for the indicated objective is whether an electronic device is defective with the two possible values: yes (it is defective) and no (it is not defective). A relevant parameter is the proportion p of defective devices in the shipment of 10,000 devices. The corresponding statistic $\hat{p}$ is the proportion of defective devices in the sample of devices that is examined.

The boundary between the null and research hypotheses is clearly $p_0 = .06$, since we need to decide whether the population proportion of defective devices p exceeds .06. Assuming that the supplier is trustworthy, it would seem to be a reasonable business practice to accept the shipment of electronic devices unless we find sufficient evidence, by examining the sample of devices, to conclude that more than 6% of the devices in the shipment are defective. Hence, we will use a hypothesis test to determine whether there is sufficient evidence to conclude that the population defective proportion p exceeds .06. More formally, our research hypothesis is $H_1 : p > .06$ and our null hypothesis is $H_0 : p \leq .06$.

To continue with this example we need to know the sample size n and the results of the examination of the sample of electronic devices. Suppose that the simple random sample contains $n = 200$ electronic devices. For a sample of size $n = 200$ the standard error of $\hat{p}$ for testing a hypothesis with $p_0 = .06$ is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{(.06)(.94)}{200}} = .0168.$$

Case 1. Suppose that 16 of the 200 devices in the sample are defective so that $\hat{p} = .08$. In this case we know that 8% of the 200 devices in the sample are defective and we need to decide whether this suggests that more than 6% of all the devices in the shipment are defective. The calculated Z statistic is

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})} = \frac{.08 - .06}{.0168} = 1.1910$$

and the P–value is

$$P\text{–value} = P(Z \geq Z_{calc}) = P(Z \geq 1.1910) = .1168.$$

Since this P–value is large there is not sufficient evidence to reject the null hypothesis $p \leq .06$ in favor of the research hypothesis $p > .06$. Therefore, if we observe 16 defective devices in a random sample of $n = 200$ devices, then we should accept the shipment of devices, since there is not sufficient evidence to conclude that more than 6% of the shipment of 10,000 devices is defective.

Case 2. Now suppose that 20 of the 200 devices in the sample are defective so that $\hat{p} = .10$. In this case

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})} = \frac{.10 - .06}{.0168} = 2.3820$$

and the P–value is

$$P\text{–value} = P(Z \geq Z_{calc}) = P(Z \geq 2.3820) = .0086.$$

This P–value is very small indicating that we have strong evidence against the null hypothesis $p \leq .06$ and in favor of the research hypothesis $p > .06$. Therefore, if we observe 20 defective devices in a random sample of $n = 200$ devices, then we are justified in returning the shipment of devices, since there is strong evidence that more than 6% of the shipment of 10,000 devices is defective.

In both of the cases described above, in addition to the conclusion of the hypothesis test the retailer might also wonder exactly what proportion of devices in the shipment of 10,000 devices are defective. We can use a 95% confidence interval estimate of p to answer this question.

In the first case there are 16 defective devices in the sample of $n = 200$ giving an observed proportion of defective devices of $\hat{p} = .08$. The confidence interval estimate is based on $\tilde{p}_k = .0879$ and the 95% margin of error M.E.$(\tilde{p}_k) = .0381$. Therefore, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is between $.0879 - .0381 = .0498$ and $.0879 + .0381 = .1260$. As expected, since we did not reject the tentative assumption that $p \leq .06$, we see that this confidence interval includes proportions that are both less than .06 and greater than .06.

In the second case there are 20 defective devices in the sample of $n = 200$ giving $\hat{p} = .10$, $\tilde{p}_k = .1075$, and the 95% margin of error M.E.$(\tilde{p}_k) = .0419$. Therefore, in this case we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is between $.1075 - .0419 = .0656$ and $.1075 + .0419 = .1494$. As expected, since we did reject the tentative assumption that $p \leq .06$, we see that all of the values in this confidence interval are greater than .06. Notice that in this case the P–value .0086 is quite small indicating that there is very strong evidence that the proportion of defective devices in the shipment is larger than .06. However, from the 95% confidence interval estimate of p we find that this proportion of defective devices might actually be as small as .0656, which is not much larger than .06. Thus, the small P–value indicates strong evidence that p is greater than .06 but it does not necessarily indicate that p is a lot larger than .06. Of course the 95% confidence interval estimate also indicates that p may be as large as .1494 which is a good bit larger than .06.

The scenario in the acceptance sampling example where there is strong evidence that $p > .06$ (P–value .0086) but the lower limit of the 95% confidence interval .0656 is not much larger than .06 highlights the need for a confidence interval to estimate the value of p in addition to a hypothesis test to clarify the practical importance of the result of the test. Bear in mind that a hypothesis test addresses a very formal distinction between two complementary hypotheses and that in some situations the results may be statistically significant (in the sense that the P–value is small) but of little practical significance (in the sense that p is not very different from p_0).

Example. Machine parts. The current production process used to manufacture a particular machine part is known (from past experience) to produce parts which are unacceptable, in the sense that they require further machining, 35% of the time. A new production process has been developed with the hope that it will reduce the chance of producing unacceptable parts. Suppose that 200 parts are produced using the new production process and that 54 of these parts are found to be unacceptable.

In this example we have a sequence of 200 dichotomous trials, where a trial consists of producing a part with the new production process and determining whether it is unacceptable. In this example p denotes the probability that a part produced using the new production process will be unacceptable. We will model these 200 trials as a sequence of $n = 200$ Bernoulli trials with population success probability p. This assumption is reasonable provided: (1) the probability that a part is unacceptable is essentially constant from part to part; and, (2) whether a specific part is unacceptable or not has no effect on the probability that any other part is unacceptable.

In this example the boundary between the null and research hypotheses is clearly $p_0 = .35$. Since these data were collected to determine if the new production process is better than the old process, we want to know whether there is sufficient evidence to conclude that less than 35% of the parts produced using the new production process would be unacceptable. Thus our research hypothesis is $H_1 : p < .35$ and our null hypothesis is $H_0 : p \geq .35$. Since 54 of the 200 parts in our sample are unacceptable we know that $\hat{p} = .27$ and we need to determine whether this is small enough to suggest that the corresponding population probability p is also less than .35. For a sample of size $n = 200$ the standard error of $\hat{p}$ for testing a hypothesis with $p_0 = .35$ is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{(.35)(.65)}{200}} = .0337.$$

The calculated Z statistic is

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})} = \frac{.27 - .35}{.0337} = -2.3739$$

and the P–value is

$$P\text{–value} = P(Z \leq Z_{calc}) = P(Z \leq -2.3739) = .0088.$$

Since this P–value is very small, there is sufficient evidence to reject the null hypothesis $p \geq .35$ in favor of the research hypothesis $p < .35$. Hence, based on this sample of 200 parts there is very strong evidence that the new production process is superior in the sense that the probability of producing an unacceptable part is less than .35.

Clearly this conclusion should be accompanied by an estimate of how much smaller this probability is likely to be. Observing 54 unacceptable parts in the sample of $n = 200$ gives $\hat{p} = .27$, $\tilde{p}_k = .2743$, and the 95% margin of error $\text{M.E.}(\tilde{p}_k) = .0611$. Therefore, we are 95% confident that the probability of a part produced using the new production process being unacceptable is between $.2743 - .0611 = .2132$ and $.2743 + .0611 = .3354$. As expected, since we did reject the tentative assumption that $p \geq .35$, we see that all of the values in this confidence interval are less than .35. The P–value .0088 is quite small indicating that there is very strong evidence that the probability of producing an unacceptable part is less than .35. However, from the 95% confidence interval estimate of p we find that this probability might actually be as large as .3354 which is not much smaller than .35. Of course the 95% confidence interval estimate also indicates that p may be as small as .2132 which is a good bit smaller than .35.

The hypothesis tests we have discussed thus far are only appropriate when we have enough *a priori* information, *i.e.*, information that does not depend on the data to be used for the hypothesis test, to postulate that the population success proportion p is on one side of a particular value p_0. That is, we have only considered situations where the research hypothesis is directional in the sense of specifying either that $p > p_0$ or that $p < p_0$. In some situations we will not have enough *a priori* information to allow us to choose the appropriate directional research hypothesis. Instead, we might only conjecture that the population success proportion p is different from some particular value p_0. In a situation like this our research hypothesis specifies that the population success proportion p is different from p_0, *i.e.*, $H_1 : p \neq p_0$ and the corresponding null hypothesis specifies that p is exactly equal to p_0, *i.e.*, $H_0 : p = p_0$. As we will see in the inheritance model considered below, when testing to see whether p is equal to a specified value p_0 the null hypothesis $H_0 : p = p_0$ often corresponds to the validity of a particular theory or model and the research hypothesis or alternative hypothesis specifies that the theory is invalid.

In order to decide between the null hypothesis $H_0 : p = p_0$ and the research hypothesis $H_1 : p \neq p_0$, we need to decide whether the observed success proportion $\hat{p}$ supports the null hypothesis by being "close to p_0", or supports the research hypothesis by being "far away from p_0". In this situation the P–value is the probability that the observed success proportion $\hat{p}$ would be as far or farther away from p_0 in either direction as is the value that we actually observe. In other words, the P–value corresponds to large values of the distance $|\hat{p} - p_0|$ (the absolute value of the difference between $\hat{p}$ and p_0). The P–value is computed under the assumption that $p = p_0$ so that the null hypothesis is true. In this situation the calculated Z statistic Z_{calc} is the absolute value of the Z statistic that would be used for testing a directional hypothesis. That is, the calculated Z statistic is

$$Z_{calc} = \left| \frac{\hat{p} - p_0}{\text{S.E.}(\hat{p})} \right| .$$

In terms of this Z statistic the P–value is the probability that the absolute value of a standard normal variable Z would take on a value as large or larger than Z_{calc} assuming that $p = p_0$. This probability is the sum of the area under the standard normal density curve to the left of $-Z_{calc}$ and the area under the standard normal density curve to the right of Z_{calc}. We need to add these two areas (probabilities) since we are finding the probability that the observed success proportion $\hat{p}$ would be as far or farther away from p_0 in either direction as is the value that we actually observe, when $p = p_0$.

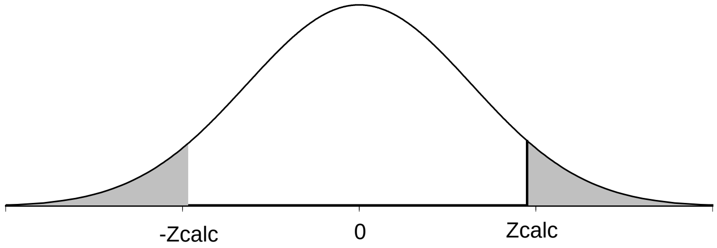
The steps for performing a hypothesis test for

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(|Z| \geq Z_{calc}) = P(Z \leq -Z_{calc}) + P(Z \geq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = |(\hat{p} - p_0)/\text{S.E.}(\hat{p})|$, and $\text{S.E.}(\hat{p}) = \sqrt{p_0(1 - p_0)/n}$. This P–value is the sum of the area under the standard normal density curve to the left of $-Z_{calc}$ and the area under the standard normal density curve to the right of Z_{calc} as shown in Figure 6.

Figure 6. P–value for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p \neq p_0$ over $H_0 : p = p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is different from p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p \neq p_0$ over $H_0 : p = p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is different from p_0.

Example. Inheritance in peas (flower color). In his investigations, during the years 1856 to 1868, of the chromosomal theory of inheritance Gregor Mendel performed a series of experiments on ordinary garden peas. One characteristic of garden peas that Mendel studied was the color of the flowers (red or white). When Mendel crossed a plant with red flowers with a plant with white flowers, the resulting offspring all had red flowers. But when he crossed two of these first generation plants, he observed plants with white as well as red flowers. We will use the results of one of Mendel's experiments to test a simple model for inheritance of flower color. Mendel observed 929 pea plants arising from a cross of two of these first generation plants. Of these 929 plants he found 705 plants with red flowers and 224 plants with white flowers.

The gene which determines the color of the flower occurs in two forms (alleles). Let R denote the allele for red flowers (which is dominant) and r denote the allele for white flowers (which is recessive). When two plants are crossed the offspring receives one allele from each parent, thus there are four possible genotypes (ordered combinations) RR, Rr, rR, and rr. The three genotypes RR, Rr, and rR, which include the dominant R allele, will yield red flowers while the fourth genotype rr will yield white flowers. If a red flowered RR genotype parent is crossed with a white flowered rr genotype parent, then all of the offspring will have genotype Rr and will produce red flowers. If two of these first generation Rr genotype plants are crossed, each of the four possible genotypes RR, Rr, rR, and rr is equally likely and plants with white as well as red flowers will occur. Under this simple model for inheritance, with each of the four genotypes having the same probability of occurring (and with each plant possessing only one genotype), the probability that a plant will have red flowers is $p = 3/4$ and the probability that a plant will have white flowers is $1 - p = 1/4$. In other words, this model for inheritance of flower color says that we would expect to see red flowers 3/4 of the time and white flowers 1/4 of the time.

We can test the validity of this model by testing the null hypothesis $H_0 : p = 3/4$ versus the alternative hypothesis $H_1 : p \neq 3/4$. Notice that the model is valid under the null hypothesis and the model is not valid under the alternative hypothesis. Mendel observed 705 plants with red flowers out of the $n = 929$ plants giving an observed proportion of plants with red flowers of $\hat{p} = 705/929 = .7589$. The standard error of $\hat{p}$, computed under the assumption that $p = p_0 = 3/4$, is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{(.75)(.25)}{929}} = .0142$$

and the calculated Z statistic is $Z_{calc} = .6251$ giving a P–value of

$$P\text{–value} = P(|Z| \geq Z_{calc}) = P(|Z| \geq .6251) = .5319.$$

This P–value is quite large and we are not able to reject the null hypothesis; therefore, we conclude that the observed data are consistent with Mendel's model. Technically, we should say that the data are not inconsistent with the model in the sense that we cannot reject the hypothesis that $p = 3/4$. In this example, the 95% confidence interval estimate of p ranges from .7303 to .7853.

5.4 Directional confidence bounds

In our discussion of hypothesis testing we considered directional research hypotheses of the form $p > p_0$ and $p < p_0$ as well as nondirectional research hypotheses of the form $p \neq p_0$. However, in our discussion of 95% confidence intervals for p we only considered confidence intervals of the form

$$\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{M.E.}(\tilde{p}_k).$$

A 95% confidence interval of this form consists of a lower bound $\tilde{p}_k - \text{M.E.}(\tilde{p}_k)$ for p and an upper bound $\tilde{p}_k + \text{M.E.}(\tilde{p}_k)$ for p, thereby giving a range of plausible values for p. In a situation where we have enough *a priori* information to justify a directional research hypothesis we might argue that it would be more appropriate to determine a 95% confidence bound (a lower bound or an upper bound) for p instead of a range of values.

For example, in the acceptance sampling example we might argue that we are less concerned with how large p might be than with how small it might be. Therefore, we might be satisfied with an estimate of the smallest value of p which would be consistent with the data, *i.e.*, we might only need a 95% confidence lower bound for p.

We will now show how a 90% confidence interval for p can be used to provide a 95% confidence lower (or upper) bound for p. The cutoff point k for the margin of error for a 90% confidence interval for p is $k = 1.645$. Three relevant probabilities associated with the 90% confidence interval with lower limit L and upper limit U are:

$$P[L \leq p \leq U] = .90, \quad P[p < L] = .05, \quad \text{and} \quad P[U < p] = .05.$$

Combining the probability that p is between L and U and the probability that p is greater than U we see that

$$P[p > L] = .90 + .05 = .95.$$

In other words, 95% of the time the computed value of the lower limit L of a 90% confidence interval for p will be less than p. Therefore, the lower limit L of a 90% confidence interval for p can be used as a 95% confidence lower bound for p. An analogous argument shows that the upper limit U of a 90% confidence interval for p can be used as a 95% confidence upper bound for p.

Example. Acceptance sampling for electronic devices (revisited). If there are 16 defective devices in a sample of $n = 200$, then the observed proportion of defective devices is $\hat{p} = .08$ and taking $k = 1.645$ gives $\tilde{p}_k = .0856$ and a 90% margin of error of $\text{M.E.}(\tilde{p}_k) = .0318$. Therefore, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is at least $.0856 - .0318 = .0538$, which allows for the possibility that $p < .06$.

If there are 20 defective devices in a sample of $n = 200$, then the observed proportion of defective devices is $\hat{p} = .10$ and taking $k = 1.645$ gives $\tilde{p}_k = .1053$ and a 90% margin of error of $\text{M.E.}(\tilde{p}_k) = .0351$. Therefore, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is at least $.1053 - .0351 = .0702$, which supports the conclusion that $p > .06$.

5.5 Summary

Basic inferential methods (confidence intervals and hypothesis tests) were introduced in this chapter in the context of making inferences about a population proportion p. These inferential methods are based on the sampling distribution of a statistic (the sample proportion $\hat{p}$ in this chapter) which describes the sample to sample variability in the statistic as an estimator of the corresponding parameter.

The specific inferential methods introduced in this chapter involve the use of the observed proportion of successes $\hat{p}$ in a random sample to make inferences about the corresponding population success proportion p. In particular, we discussed confidence interval estimates of p and formal tests of hypotheses about p. These inferences about p are based on a normal approximation to the sampling distribution of $\hat{p}$ and require certain assumptions about the random sample. Strictly speaking, the inferential methods discussed in this chapter are not appropriate unless these assumptions are valid. The requisite assumptions are that the sample is a simple random sample selected with replacement or equivalently that the sample corresponds to a sequence of Bernoulli trials. We also noted that this approximation works well for a simple random sample selected without replacement provided the population being sampled is very large. The sampling distribution of $\hat{p}$ is the theoretical probability distribution of $\hat{p}$ which indicates how $\hat{p}$ behaves as an estimator of p. Under the assumptions described above, the sampling distribution of $\hat{p}$ indicates that $\hat{p}$ is unbiased as an estimator of p ($\hat{p}$ neither consistently overestimates p nor consistently underestimates p) and provides a measure of the variability in $\hat{p}$ as an estimator of p (the population standard error of $\hat{p}$, $\text{S.E.}(\hat{p}) = \sqrt{p(1-p)/n}$). The normal approximation allows us to compute probabilities concerning $\hat{p}$ by re–expressing these probabilities in terms of the standardized variable $Z = (\hat{p} - p)/\text{S.E.}(\hat{p})$ and using the standard normal distribution to compute the probabilities.

A 95% confidence interval estimate of p is an interval of plausible values for p constructed using a method which guarantees that 95% of such intervals will actually contain the unknown proportion p. That is, a 95% confidence interval is an interval constructed using a method of generating such intervals with the property that this method will work, in the sense of generating an interval that contains p, for 95% of all possible samples. We recommended the Wilson interval as a confidence interval estimate of p. For a confidence level C (usually .95) and the corresponding standard normal cutoff point k ($k = 1.96$ when $C = .95$) the Wilson interval is of the form

$$\tilde{p}_k - \text{M.E.}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{M.E.}(\tilde{p}_k),$$

where

$$\tilde{p}_k = \frac{n\hat{p} + \frac{k^2}{2}}{n + k^2}$$

determines the center of the interval, and the **margin of error of** $\tilde{p}_k$

$$\text{M.E.}(\tilde{p}_k) = \sqrt{\tilde{p}_k^2 - \frac{n\hat{p}^2}{n + k^2}}$$

determines the length of the interval. We also discussed a simple but less accurate confidence interval for p (the Wald interval) and a simple approximation to the 95% Wilson interval (the Agresti–Coull interval). The 95% Agresti–Coull interval is of the form

$$\tilde{p} - \text{M.E.}(\tilde{p}) \leq p \leq \tilde{p} + \text{M.E.}(\tilde{p}),$$

where

$$\tilde{p} = \frac{\text{the number of successes plus 2}}{\text{the number of observations plus 4}} = \frac{n\hat{p} + 2}{n + 4}$$

and

$$\text{M.E.}(\tilde{p}) = 1.96\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}.$$

The “add 2 successes / add 4 observations” trick used in the Agresti–Coull interval can be used in a computer program or calculator implementation of the Wald interval to approximate the 95% Wilson interval.

A hypothesis test is used to compare two competing, complementary hypotheses (the null hypothesis H_0 and the research or alternative hypothesis H_1) about p by tentatively assuming that H_0 is true and examining the evidence, which is quantified by the appropriate P–value, against H_0 and in favor of H_1. Since the P–value quantifies evidence against H_0 and in favor of H_1, a small P–value constitutes evidence in favor of H_1. Guidelines for interpreting a P–value are given on page 99.

If there is sufficient *a priori* information to specify a directional hypothesis of the form $H_1 : p > p_0$ or $H_1 : p < p_0$, then we can perform a hypothesis test to address the respective questions "Is there sufficient evidence to conclude that $p > p_0$?" or "Is there sufficient evidence to conclude that $p < p_0$?" The null hypotheses for these research hypotheses are their negations $H_0 : p \leq p_0$ and $H_0 : p \geq p_0$, respectively. The hypothesis test proceeds by tentatively assuming that the null hypothesis H_0 is true and checking to see if there is sufficient evidence (a small enough P–value) to reject this tentative assumption in favor of the research hypothesis H_1. The P–values for these directional hypothesis tests are based on the observed value of the Z–statistic $Z_{calc} = (\hat{p} - p_0)/\text{S.E.}(\hat{p})$, where $\text{S.E.}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$ is the standard error for testing. For $H_1 : p > p_0$ large values of $\hat{p}$, relative to p_0, favor H_1 over H_0 and the P–value is the probability that $Z \geq Z_{calc}$. For $H_1 : p < p_0$ we look for small values of $\hat{p}$, relative to p_0, and the P–value is the probability that $Z \leq Z_{calc}$.

For situations where there is not enough *a priori* information to specify a directional hypothesis we considered a hypothesis test for the null hypothesis $H_0 : p = p_0$ versus the alternative hypothesis $H_1 : p \neq p_0$. Again we tentatively assume that H_0 is true and check to see if there is sufficient evidence (a small enough P–value) to reject this tentative assumption in favor of H_1. In this situation the hypothesis test addresses the question "Are the data consistent with $p = p_0$ or is there sufficient evidence to conclude that $p \neq p_0$?" For this non–directional hypothesis test we take the absolute value when computing the observed value of the Z–statistic $Z_{calc} = |(\hat{p} - p_0)|/\text{S.E.}(\hat{p})$, since values of $\hat{p}$ which are far away from p_0 in either direction support $p \neq p_0$ over $p = p_0$. Thus the P–value for this hypothesis test is the probability that $|Z| \geq Z_{calc}$.

For all of these hypothesis tests, the P–value is computed under the assumption that H_0 is true, and the P–value is the probability of observing a value of $\hat{p}$ that is as extreme or more extreme, relative to p_0, than the value we actually observed, under the assumption that H_0 is true (in particular $p = p_0$). In this statement the definition of extreme (large, small, or far from in either direction) depends on the form of H_1.

5.6 Exercises

For each of the following examples:

a) Define the relevant population success proportion or probability. Be sure to indicate the corresponding population.

b) Using the information provided, formulate an appropriate research hypothesis about the population success proportion and briefly explain why your hypothesis is appropriate.

c) Perform a hypothesis test to determine whether the data support your research hypothesis. Provide the P–value and briefly summarize your conclusion in the context of the example.

d) Construct a 95% confidence interval for the success proportion and interpret it in context of the example.

1. A company which provides telephone based support for its products has found that 20% of the users of this service file complaints about the quality of the service they receive. Recently this company retrained its support personnel with the hope of reducing the percentage of users who file complaints. A random sample of 150 customers who used the telephone support after the support personnel had been retrained revealed that 20 customers were not satisfied with the quality of support they received.

2. A manufacturer has found that 15% of the items produced at its old manufacturing facility fail to pass final inspection and must be remanufactured before they can be sold. This manufacturer has recently opened a new manufacturing facility and wants to determine whether the items produced at the new facility are more or less likely to fail inspection and require remanufacture. A random sample of 200 items is selected from a large batch of items produced at the new facility and of these 42 fail inspection and require remanufacturing.

3. A supplier of vegetable seeds has a large number of bean seeds left over from last season and is trying to decide whether these seeds are suitable for sale for the current season. This supplier normally advertises that more than 85% of its bean seeds will germinate. A random sample of 200 of the leftover beans seeds was selected and of these 200 seeds 181 germinated.