Inference for a Proportion

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). When a unit is selected from the population and the unit is found to belong to the success subgroup we say that a **success** has occurred. Similarly, when a member of the failure subgroup is selected we say that a **failure** has occurred. The proportion of units in the population that belong to the success subgroup (the units classified as successes) 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. 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 success proportion $\hat{p}$ is a statistic, since it is a numerical characteristic of the sample.

We can use the selection of balls from a box of balls as a model for sampling from a dichotomous population. Consider a box containing balls of which some are red (successes) and the rest are green (failures). The population success proportion, p, is the proportion of red balls in the box. If we select a random sample of n balls from this box, then the sample success proportion, $\hat{p}$, is the proportion of red balls in the sample. Thus, in this model, a ball is a unit, the box of balls is the population, selecting a red ball is a success, the proportion of red balls in the box p is the parameter (the population success proportion), and the proportion of red balls in the sample $\hat{p}$ is the statistic (the sample success proportion).

The sampling distribution and the normal approximation

We will now discuss how the sample proportion $\hat{p}$ can be used to make inferences about the population proportion p. Assuming that the sample size n is reasonably large, it seems reasonable to view the sample proportion $\hat{p}$ as an estimate of the population 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 population, we would

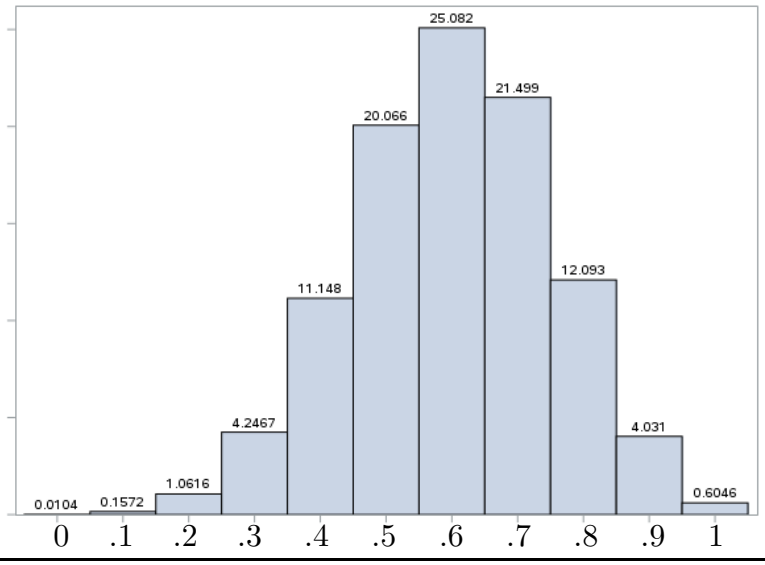
not expect the observed sample success proportions, the $\hat{p}$'s, to be exactly the same. In the box of balls example, if we took several samples from the box we would expect the proportion of red balls in the sample to vary from sample to sample.

Two questions we might ask about the sample proportion $\hat{p}$ as an estimator of the population proportion p are:
(1) Can we expect the sample proportion $\hat{p}$ to be close to the population proportion p?
(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 general, 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.

Figure 1. Sampling distribution of $\hat{p}$ when $n = 10$ and $p = .6$

$\hat{p}$	probability
0	.0001
.1	.0016
.2	.0106
.3	.0425
.4	.1115
.5	.2007
.6	.2508
.7	.2150
.8	.1209
.9	.0403
1	.0060



For example, suppose that we select a simple random sample of $n = 10$ balls, with replacement, from a box containing six red balls and four green balls. If we identify red as a success, then, in this example, the population success proportion is $p = .6$, since 60% of the balls in the box are red. The possible values of the sample success proportion $\hat{p}$ when $n = 10$ are: $0, .1, .2, \ldots, 1$. The sampling distribution of $\hat{p}$ summarized in Figure 1 gives the probabilities corresponding to these possible values. As you would expect, the

probabilities are highest for values near $p = .6$ and very small for values far away from $p = .6$. In particular, when we select a simple random sample of size $n = 10$ from this box, 25.08% of the time we would observe $\hat{p} = .6$, 20.07% of the time we would observe $\hat{p} = .5$, and 21.50% of the time we would observe $\hat{p} = .7$. On the other hand, it would be possible but very unlikely to observe $\hat{p} = .2$ (only 1.06% of the time) or $\hat{p} = .9$ (only 4.03% of the time).

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.

Returning to our discussion of the sampling distribution of $\hat{p}$ we first present two important properties of this sampling distribution. The observed proportion $\hat{p}$ in a simple random sample selected with replacement from a population with population proportion p has a sampling distribution with the following properties.

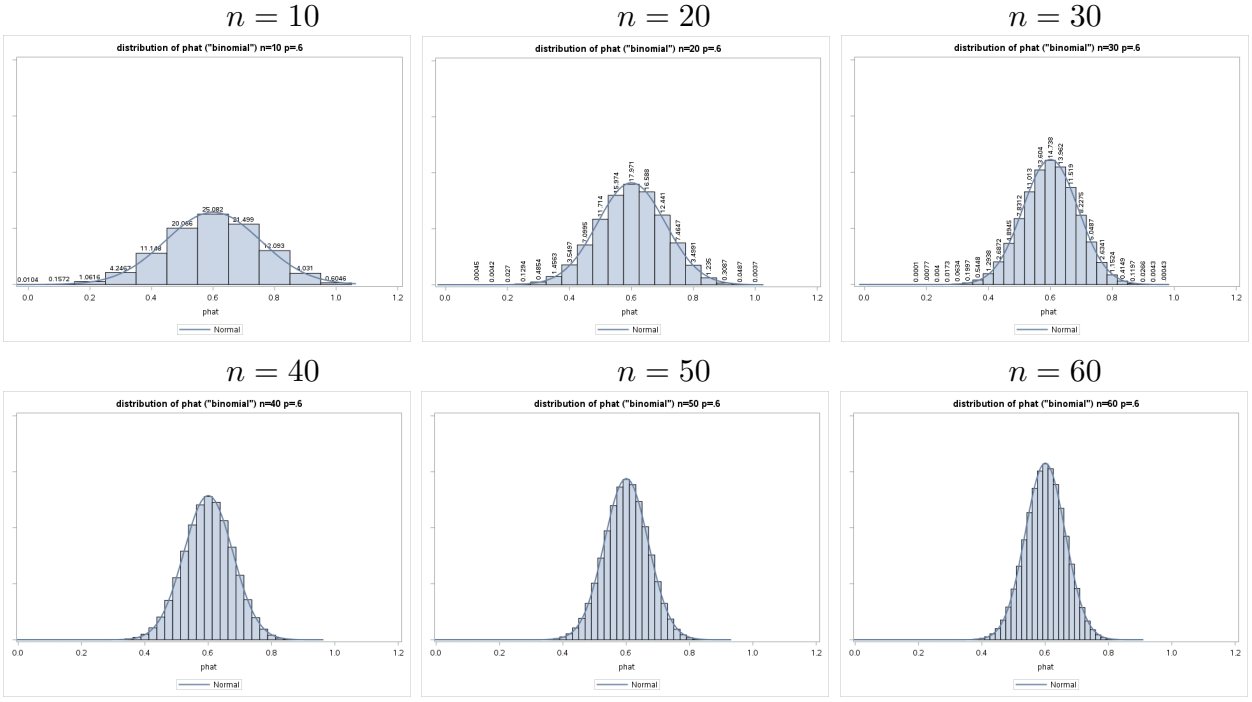
1. The mean of the sampling distribution of $\hat{p}$ is the population probability p. Therefore, $\hat{p}$ is unbiased as an estimator of p.
2. The population standard error of $\hat{p}$, denoted by $\text{SE}(\hat{p})$, is

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

The inferential methods we will consider are based on a large sample size normal approximation to the sampling distribution of $\hat{p}$. A detailed discussion of the exact form of the sampling distribution of $\hat{p}$, the normal distribution, and the normal approximation to the sampling distribution of $\hat{p}$ can be found in a separate document. Here we will provide an indication of some basic properties of the sampling distribution of $\hat{p}$ in terms of some graphical representations of the probability histogram of the distribution of $\hat{p}$ with superimposed fitted normal density curves.

First consider how the sampling distribution of $\hat{p}$ depends on the sample size n. From the expression given above for the population standard error of $\hat{p}$ we can see that, as you would expect, the variability in $\hat{p}$ as an estimator of p decreases as the sample size increases. This behavior of the sampling distribution is illustrated in Figure 2.

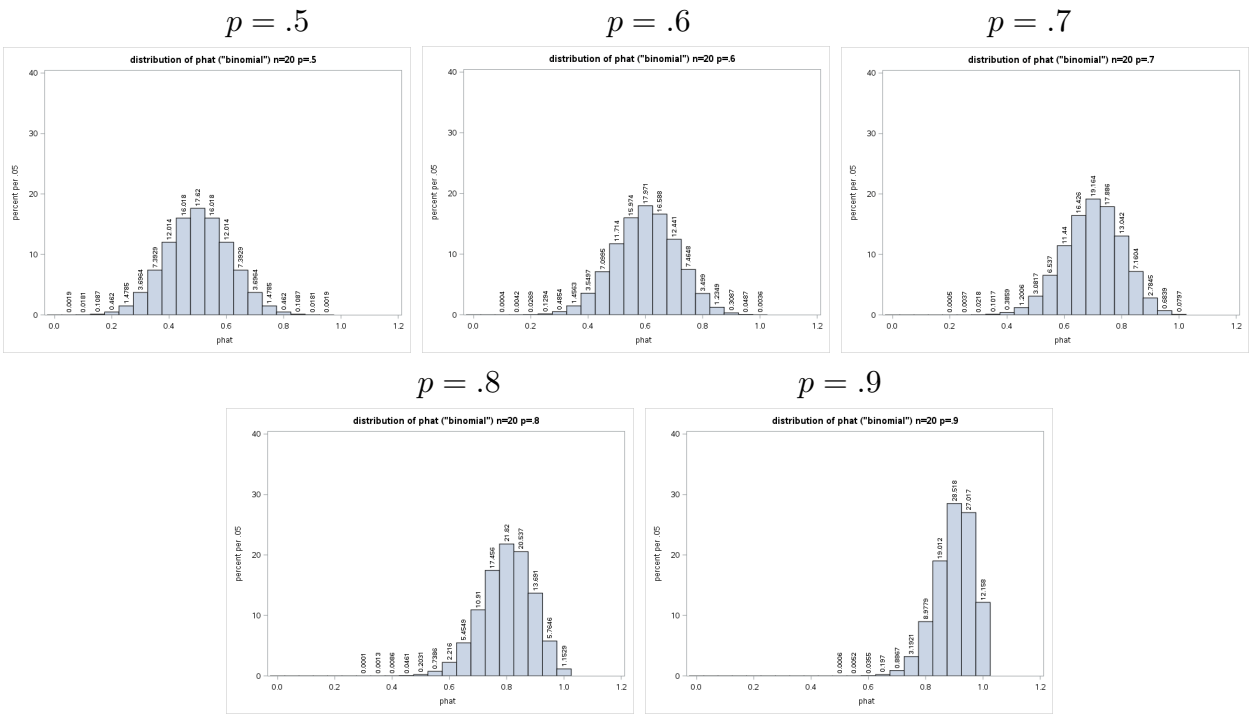
Figure 2. The sampling distribution of $\hat{p}$ with normal approximation for $p = .6$. Histograms are provided for $n = 10, 20, 30, 40, 50$, and 60. All of these distributions are centered at $p = .6$. Notice how the variability in the distribution decreases as the sample size n increases and how much better the normal density curve matches the histogram.



Next consider how the sampling distribution of $\hat{p}$ depends on the value of the population success proportion p. The sampling distribution of $\hat{p}$ is unimodal (single peaked) with its peak at p. If $p = .5$, then the sampling distribution of $\hat{p}$ is symmetric. If $p < .5$,

then sampling distribution of $\hat{p}$ is skewed right. If $p > .5$, then sampling distribution of $\hat{p}$ is skewed left. Since $\hat{p}$ is unbiased as an estimator of p, the mean of the distribution of $\hat{p}$ is p. Thus, the probability histogram of the distribution of $\hat{p}$ has its balance point at p. From the expression given above for the population standard error of $\hat{p}$ we can see that, for fixed n, the variability in $\hat{p}$ as an estimator of p is highest when $p = .5$ and decreases as p moves away from .5. These properties of the sampling distribution are illustrated in Figure 3.

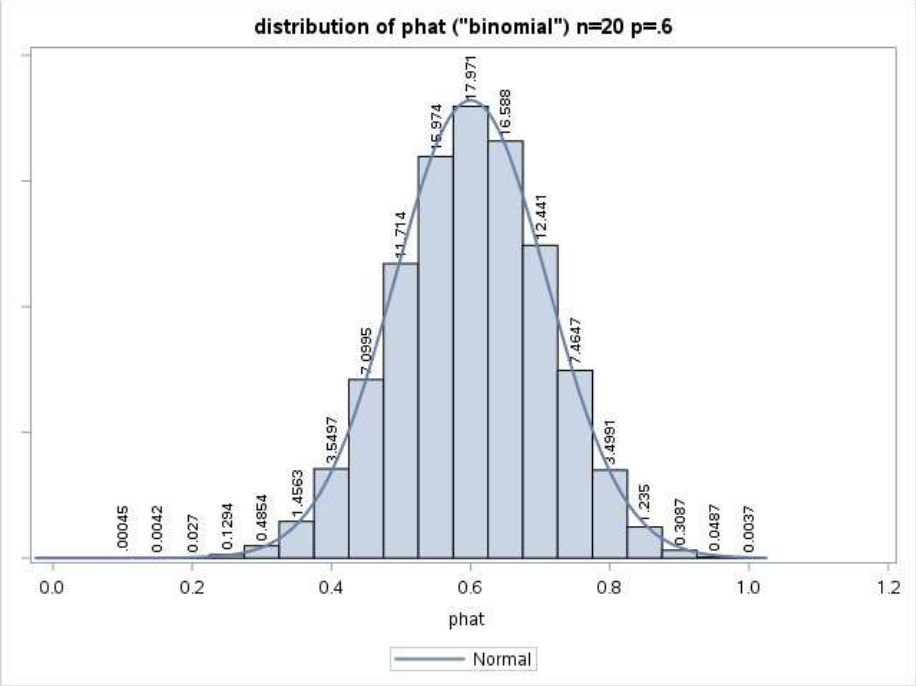
Figure 3. The sampling distribution of $\hat{p}$ with normal approximation for $n = 20$. Histograms are provided for $p = .5, .6, .7, .8$, and .9. Each distribution is centered at its p. Notice how the variability in the distribution decreases as p moves away from .5. Notice also how the distribution becomes more skewed left as p moves away from .5.



The normal approximation to the sampling distribution of $\hat{p}$ is illustrated graphically in Figure 4. In this figure, the probability histogram of the sampling distribution of $\hat{p}$ for the situation when a simple random sample of size $n = 20$ is selected with replacement from a dichotomous population with $p = .6$ is given along with the approximating normal density curve. The exact probabilities which correspond to the areas of the rectangles of the histogram are given as percentages, *e.g.*, the probability of observing $\hat{p} = .06$ is .17971 (17.97%). The normal approximation, explained more fully below, basically says that, over

a specified range of $\hat{p}$ values, the area under the normal density curve and the area in the rectangles of the histogram are similar. You can see that the curve matches the histogram reasonably well in Figure 4.

Figure 4. The sampling distribution of $\hat{p}$ with normal approximation for $n = 20$ and $p = .6$. This assumes a simple random sample selected with replacement. If the $\hat{p}$ values were converted to counts, $X = n\hat{p}$, then this would be a binomial distribution.



The normal approximation to the sampling distribution of $\hat{p}$ says that, for large values of n, the standardized value of $\hat{p}$ obtained by subtracting the population 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 $Z = (\hat{p} - p)/\mathrm{SE}(\hat{p})$ behaves like a standard normal variable. This means, as will be shown below, that we can use an area under the standard normal density curve (a probability in terms of Z) to approximate an area in the probability histogram of the sampling distribution of $\hat{p}$ (a probability in terms of $\hat{p}$). The relationship between $\hat{p}$ and p indicated by this expression for Z and the normal distribution itself allow us to use $\hat{p}$ to make formal, quantifiable inferences about p.

Estimation of p

As noted above, the sample proportion $\hat{p}$ is an unbiased estimator of the population proportion p. We can think of $\hat{p}$ as our "best guess" of the value of p. To allow for sampling variability 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 limiting values. We will now use the normal approximation to the sampling distribution of $\hat{p}$ to develop such an interval estimate of p.

The probability that a standard normal variable Z takes on a value between -1.96 and 1.96 is equal to .95, *i.e.*, $P(-1.96 \leq Z \leq 1.96) = .95$. Thus, when we observe the value of a standard normal variable Z, 95% of the time we will find that $-1.96 \leq Z \leq 1.96$. Graphically this means that the area under the standard normal density curve over the interval from -1.96 to 1.96 is .95. Thus, for sufficiently large values of n we have the approximation,

$$P\left[-1.96 \leq \frac{\hat{p} - p}{\text{SE}(\hat{p})} \leq 1.96\right] = .95$$

or equivalently

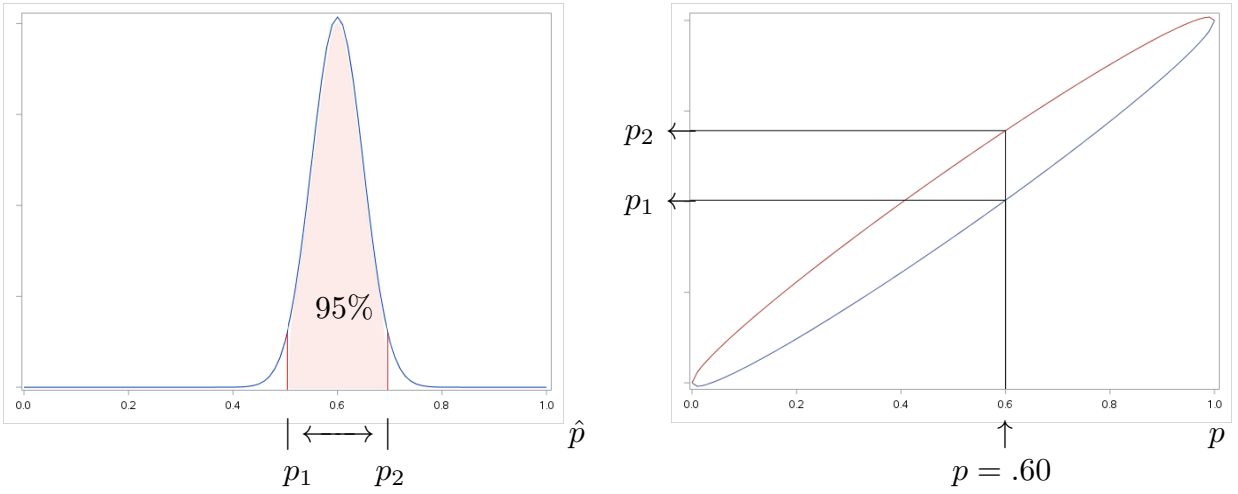
$$P\left[p - 1.96 \cdot \text{SE}(\hat{p}) \leq \hat{p} \leq p + 1.96 \cdot \text{SE}(\hat{p})\right] = .95.$$

Note that this indicates that 95% of the time when a simple random sample is selected and $\hat{p}$ is computed the observed value of $\hat{p}$ will be between $p - 1.96 \cdot \text{SE}(\hat{p})$ and $p + 1.96 \cdot \text{SE}(\hat{p})$, *i.e.*, $\hat{p}$ will be within 1.96 population standard error units of p. We will refer to the interval from $p - 1.96 \cdot \text{SE}(\hat{p})$ to $p + 1.96 \cdot \text{SE}(\hat{p})$ as the central 95% interval of the distribution of $\hat{p}$, since it is centered at p and it will contain the observed value of $\hat{p}$ 95% of the time.

The relationship between $\hat{p}$ and p is illustrated, for $n = 100$ and $p = .6$, in Figure 5. In this case, $p = .6$, $\text{SE}(\hat{p}) = \sqrt{.6(.4)/100} = .0490$, and $p_1 = .6 - 1.96 \cdot \text{SE}(\hat{p}) = .5040$ and $p_2 = .6 + 1.96 \cdot \text{SE}(\hat{p}) = .6960$ are the limits of the central 95% interval of the distribution of $\hat{p}$. The shaded region with area .95 in Figure 5 indicates that 95% of all samples will yield a sample proportion $\hat{p}$ which is between p_1 and p_2. That is, when $p = .6$ and $n = 100$, 95% of all samples will yield a sample proportion $\hat{p}$ which is between $p_1 = .5040$ and $p_2 = .6960$. In terms of a box of balls this means that, if exactly 60% of the balls in the box were red, then 95% of the time when we selected $n = 100$ balls from the box, at random and with replacement, we would find between 51 and 69 red balls (between 50.40 and 69.60) among the 100 balls in the sample.

The plot on the right in Figure 5 shows how the endpoints of the central 95% interval of the distribution of $\hat{p}$ depend on p. In this plot the sample size is $n = 100$. (The pattern is similar for other values of n.) The upper (red) curve gives values of $p + 1.96 \cdot \text{SE}(\hat{p}) = p + 1.96\sqrt{p(1-p)/n}$ as a function of p and the lower (blue) curve gives values of $p - 1.96 \cdot \text{SE}(\hat{p}) = p - 1.96\sqrt{p(1-p)/n}$ as a function of p. The intersections of a vertical line draw at a particular value of p with these curves are the endpoints of the central 95% interval of the distribution of $\hat{p}$ for that value of p. The lines drawn in the figure demonstrate this for the case $p = .6$. Note that, as mentioned above, $p_1 = .5040$ and $p_2 = .6960$, and these are the endpoints of the interval in the plot on the left in Figure 5.

Figure 5. The plot on the left shows the central 95% interval of the distribution of $\hat{p}$ for $n = 100$ and $p = .6$.
The curves in the plot on the right show the endpoints, $p \pm 1.96\sqrt{p(1-p)/n}$, of the central 95% interval of the distribution of $\hat{p}$ as a function of p for $n = 100$.
The endpoints for the case $p = .6$ are indicated by the lines marking the intersections at $p_1 = .5040$ and $p_2 = .6960$.



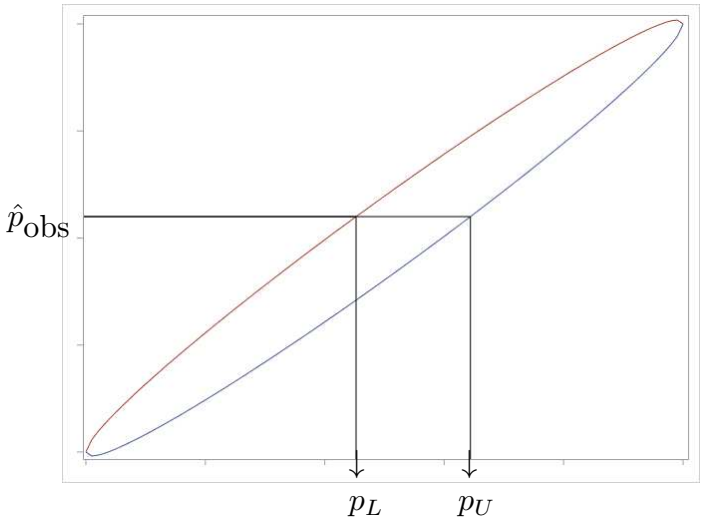
A confidence interval for p.

As noted above, for a particular value of p, 95% of all samples will yield a value of $\hat{p}$ within the corresponding central 95% interval from $p - 1.96\text{SE}(\hat{p})$ to $p + 1.96\text{SE}(\hat{p})$. Thus for each possible value of p we can find an interval of likely values for $\hat{p}$. But we need an interval of plausible values for p not for $\hat{p}$! We will now show how the central 95% intervals of the distribution of $\hat{p}$ can be used to form a 95% confidence interval estimate of p. We will provide a formal definition of a 95% confidence interval estimate later.

Suppose that a simple random sample has been selected and let $\hat{p}_{\text{obs}}$ denote the observed value of $\hat{p}$. Given this $\hat{p}_{\text{obs}}$, we want to know which values of p are plausible. More precisely, we want to know which values of p determine sampling distributions for $\hat{p}$ under which seeing $\hat{p} = \hat{p}_{\text{obs}}$ would not be surprising. We can formalize this goal by saying that we want to know which values of p determine a sampling distributions for $\hat{p}$ for which the central 95% interval of the distribution of $\hat{p}$ contains the observed value $\hat{p}_{\text{obs}}$.

The plot on the right in Figure 5 shows how the central 95% interval of the distribution of $\hat{p}$ depends on the value of p. We want to determine which values of p yield central 95% intervals which contain $\hat{p}_{\text{obs}}$. This requires using the graph of Figure 5 in the other direction. In Figure 6 a horizontal line is drawn at $\hat{p}_{\text{obs}}$ and its intersections, p_L and p_U, with the two curves are indicated. Notice that p_L is the smallest value of p for which the central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$ and p_U is the largest value of p for which the central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (Figure 6 is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. In this case, $p_L = .4524$ and $p_U = .6439$.) Thus, if we draw a vertical line, as in Figure 5, at any value of p between p_L and p_U, then the corresponding central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$.

Figure 6. The smallest and largest values of p, p_L and p_U, for which the central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$.
This example is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$.



In order to determine the value of p_L we simply set $p + 1.96\text{SE}(\hat{p})$ equal to $\hat{p}_{\text{obs}}$ and solve for p (draw the horizontal line as in Figure 6 and project down at the intersection with the upper (red) curve). To determine the value of p_U we set $p - 1.96\text{SE}(\hat{p})$ equal

to $\hat{p}_{\text{obs}}$ and solve for p (draw the horizontal line as in Figure 6 and project down at the intersection with the lower (blue) curve).

Table 1. Central 95% intervals $[p_1, p_2]$ of the distribution of $\hat{p}$ for selected values of p when $n = 100$. The intervals for $p = .46$ through $p = .64$ contain $\hat{p}_{\text{obs}} = .55$.

	central 95% interval	
p	p_1	p_2
.10	.041	.159
.20	.122	.278
.30	.210	.390
.40	.304	.496
.45	.352	.548 ***
.46	.362	.558
.47	.372	.568
.48	.382	.578
.49	.392	.588
.50	.402	.598
.51	.412	.608
.52	.422	.618
.53	.432	.628
.54	.442	.638
.55	.452	.648
.56	.463	.657
.57	.473	.667
.58	.483	.677
.59	.494	.686
.60	.504	.696
.61	.514	.706
.62	.525	.715
.63	.535	.725
.64	.546	.734
.65	.556	.743 ***
.70	.610	.790
.80	.722	.878
.90	.841	.959

A simple example will help to clarify this discussion. Consider a box containing a large number of balls. Suppose that some of the balls in the box are red. Let p denote the proportion of red balls in the box. Now suppose that a simple random sample of $n = 100$ balls has been selected and 55 of the 100 balls found to be red. In this example 55% of the balls in the sample are red, *i.e.*, $\hat{p}_{\text{obs}} = .55$. Table 1 contains central 95% intervals of

the distribution of $\hat{p}$ for $n = 100$ and selected values of p. Notice that when p is between .46 and .64 the intervals contain $\hat{p}_{\text{obs}} = .55$.

For example, when $p = .50$, 95% of all samples will yield a $\hat{p}$ value between .402 and .598. In a case like this, where the central interval contains .55, it would not be surprising to see $\hat{p}_{\text{obs}} = .55$ and the correspond value of p would be deemed plausible. That is, if exactly 50% of the balls in the box were red, then 95% of the time when we selected a random sample of $n = 100$ balls we would find that between 40.2% and 59.8% of the 100 balls in the sample were red. Since 55% belongs to this range it is plausible that exactly 50% of all the balls in the box are red ($p = .50$).

On the other hand, if $p = .30$, 95% of all samples will yield a $\hat{p}$ value between .210 and .390. Thus, if exactly 30% of the balls in the box were red, then 95% of the time when we selected a random sample of $n = 100$ balls we would find that between 21.0% and 39.0% of the 100 balls in the sample were red. Since this entire interval is less than 55%, it would be surprising to see $\hat{p}$ as large as .55 when p was .30. Therefore, when we see $\hat{p}_{\text{obs}} = .55$ it is reasonable to conclude that the percentage of red balls in the box is not 30% ($p \neq .30$).

Similarly, if $p = .70$, 95% of all samples will yield a $\hat{p}$ value between .610 and .790. Thus, if exactly 70% of the balls in the box were red, then 95% of the time when we selected a random sample of $n = 100$ balls we would find that between 61.0% and 79.0% of the 100 balls in the sample were red. Since this entire interval is greater than 55%, it would be surprising to see $\hat{p}$ as small as .55 when p was .70. Therefore, when we see $\hat{p}_{\text{obs}} = .55$ it is reasonable to conclude that the percentage of red balls in the box is not 70% ($p \neq .70$).

In cases like the last two, where the entire interval is less than .55 or the entire interval is greater than .55, it would be surprising to see $\hat{p}_{\text{obs}} = .55$ and the correspond value of p would not be deemed plausible. From the table we can see that $p_L \approx .45$, since the upper endpoint of the central 95% interval p_2 is approximately .55 when $p = .45$, and $p_U \approx .64$, since the lower endpoint p_1 of the central 95% interval is approximately .55 when $p = .64$. More precise computation indicates that $p_L = .4524$ and $p_U = .6439$. Using this reasoning, we see that when $n = 100$ and we observe $\hat{p}_{\text{obs}} = .55$ we can argue that values of p between $p_L = .4524$ and $p_U = .6439$ are plausible. In other words, if we selected a simple random sample of $n = 100$ balls from the box and found that 55% of the balls in the sample were red, then we would conclude that somewhere between 45.24% and 64.39% of all the balls in the box are red.

The interval of plausible values for p with endpoints p_L and p_U discussed above is a 95% confidence interval estimate of p. In the preceding example, the more formal way of stating the conclusion is as follows. If we selected a simple random sample of $n = 100$ balls from the box and found that 55% of the balls in the sample were red, then we would conclude that we were 95% confident that the percentage of red balls in the box was somewhere between 45.24% and 64.39% (in terms of p, $.4524 \leq p \leq .6439$).

A confidence interval for p – more formally.

We will now give a more formal definition of a confidence interval estimate of p. The purpose of a confidence interval estimate is to provide 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 (95% in the example above), that the true value of p is between two particular values (p_L and p_U in the example above). 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. We will adopt the usual convention of using a confidence level of 95%. The confidence level indicates our confidence that the unknown p actually belongs to the corresponding interval.

There is some chance for confusion about what it means to say we are 95% confident that p is between p_L and p_U. The important thing to remember is that it is the endpoints of the interval p_L and p_U 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 p_L and p_U forming 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 used to construct the interval works for 95% of all possible samples and so it probably worked for our sample.

A 95% confidence interval for p – summary and computation.

As noted above, a 95% confidence interval estimate of p is an interval of plausible values for p 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 Wilson (score) 95% confidence interval for p.

The 95% confidence interval estimate of the population proportion p discussed above is usually called the Wilson interval or the score interval. The graphical derivation above is readily modified for a confidence level other that 95%. The requisite modification is to change the 95% confidence level multiplier 1.96 to the value appropriate for the desired confidence level, *e.g.*, the multiplier 1.645 leads to a 90% confidence interval.

Computation of the Wilson (score) 95% confidence interval for p.

Computations of the Wilson 95% confidence interval estimate of p for the box of balls example with $n = 100$ and 55 red balls are illustrated in Figures 7–10. Recall that in this example, we selected a simple random sample of $n = 100$ balls from the box and found that 55% of the balls in the sample were red, then we concluded that we were 95% confident that the percentage of red balls in the box was somewhere between 45.24% and 64.39% (in terms of p, $.4524 \leq p \leq .6439$).

Figure 7. SAS output giving the Wilson 95% confidence interval for the box of balls example with $n = 100$ and $\hat{p}_{\text{obs}} = .55$ with SAS command file.

box of balls example

The FREQ Procedure

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
red	55	55.00	55	55.00
green	45	45.00	100	100.00

Binomial Proportion	
outcome = red	
Proportion	0.5500
ASE	0.0497

Confidence Limits for the Binomial Proportion		
Proportion = 0.5500		
Type	95% Confidence Limits	
Wilson	0.4524	0.6439

SAS command file: boxofballs_onepropexample.sas

```
/* box of balls example */
data boxofballs;
   input outcome $ count;
/* the coding outcome $ indicates that outcome is character valued */
cards;
red 55
green 45
;
proc freq data=boxofballs order=data;
   tables outcome / alpha=.05 binomial(cl=wilson level='red');
   weight count;
title 'box of balls example';
/* tables option
   -- alpha=.05 requests 95% confidence level
        this is the default Use alpha=.10 to
        get a 90% confidence interval and use
        the appropriate endpoint of the 90% CI
        as a 95% lower/upper confidence bound */
/* binomial options
   -- cl=wilson Wilson confidence interval
   -- level='red' define "success" default is
         the first level which is "red" in this example */
run;
/* ---------------------------------------------------------*/
```

Figure 8. R commands and output giving the Wilson 95% confidence interval for the box of balls example with $n = 100$ and $\hat{p}_{\text{obs}} = .55$.

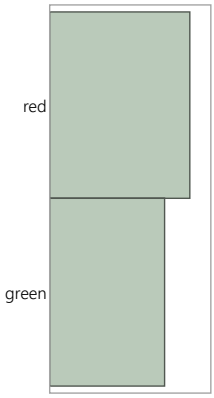
R commands and output for the box of balls example

```
R Commands:

# box of balls example
prop.test(55,100,correct=0)

R Output:

1-sample proportions test without continuity correction

data:  55 out of 100, null probability 0.5
X-squared = 1, df = 1, p-value = 0.3173
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4524460 0.6438546
sample estimates:
   p
0.55
```

Figure 9. JMP commands and output giving the Wilson 95% confidence interval for the box of balls example with $n = 100$ and $\hat{p}_{\text{obs}} = .55$.

Distributions

color



Frequencies

Level	Count	Prob
green	45	0.45000
red	55	0.55000
Total	100	1.00000

N Missing 0

2 Levels

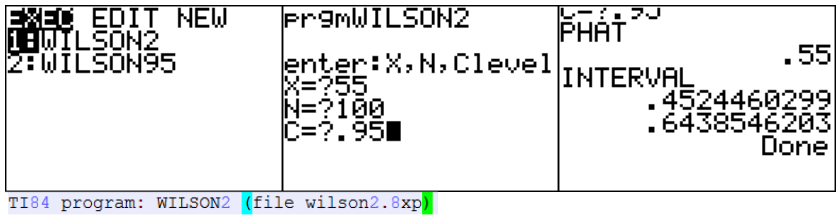
Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
green	45	0.45000	0.356145	0.547554	0.950
red	55	0.55000	0.452446	0.643855	0.950
Total	100				

Note: Computed using score confidence intervals.

```
JMP directions for the box of balls example

step 1: create a data table: file --> new --> data table
input in columns 1 and 2 as:

color count
red    55
green  45

step 2: analyze --> distribution
assign variables as:

y, columns : color
freq : count

under the "color tab" select options
display options -->  horizontal layout
histogram options -->  vertical
confidence interval -->  .95
```

Figure 10. TI84 program "WILSON2" screen captures followed by command listing. This output shows how to get the Wilson 95% confidence interval for the box of balls example with $n = 100$ and $\hat{p}_{\text{obs}} = .55$.




```
TI84 program: WILSON2 (file wilson2.8xp)

Disp "enter:X,N,Clevel"
Prompt X,N,C
X/N→R
C+(1-C)/2→P
invNorm(P)→K
(X+K^2/2)/(N+K^2)→Y
(Y^2-X^2/(N*(N+K^2)))^.5→W
Y-W→L
Y+W→U
Disp "PHAT",R
Disp "INTERVAL",L,U
```

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 is 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 the entire 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. The SAS output for these examples is provided in Figures 11a and 11b.

Figure 11a. SAS output for the apple orchard example – confidence intervals

apple orchard example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	35	17.50	35	17.50
notinfested	165	82.50	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.1750
ASE	0.0269

Confidence Limits for the Binomial Proportion		
Proportion = 0.1750		
Type	95% Confidence Limits	
Wilson	0.1286	0.2336

apple orchard example

The FREQ Procedure

case=2

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	26	13.00	26	13.00
notinfested	174	87.00	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.1300
ASE	0.0238

Confidence Limits for the Binomial Proportion		
Proportion = 0.1300		
Type	95% Confidence Limits	
Wilson	0.0903	0.1837

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 we get a 95% confidence interval ranging from .1286 to .2336 (see Figure 11a). 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. Of

course the interval also indicates that it is possible that less than 20% (as low as 12.86%) 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 we get a 95% confidence interval ranging from .0903 to .1837 (see Figure 11a). 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 $p < .20$.

Figure 11b. SAS output for the apple orchard example – confidence intervals

apple orchard example

The FREQ Procedure

case=3

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	45	22.50	45	22.50
notinfested	155	77.50	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.2250
ASE	0.0295

Confidence Limits for the Binomial Proportion		
Proportion = 0.2250		
Type	95% Confidence Limits	
Wilson	0.1726	0.2877

apple orchard example

The FREQ Procedure

case=4

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	54	27.00	54	27.00
notinfested	146	73.00	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.2700
ASE	0.0314

Confidence Limits for the Binomial Proportion		
Proportion = 0.2700		
Type	95% Confidence Limits	
Wilson	0.2132	0.3354

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 we get a 95% confidence interval ranging from .1726 to .2877 (see Figure 11b). 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 more 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% (as low as 17.26%) of the trees in the entire orchard are infested. Of course the interval also indicates that it is possible that more than 20% (as high as 28.77%) of the trees in the entire orchard are infested.

Case 4. Finally, 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 we get a 95% confidence interval ranging from .2132 to .3354 (see Figure 11b). 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 $p > .20$.

A formula for the Wilson (score) 95% confidence interval for p.

Some readers may find an algebraic expression for the Wilson confidence interval useful. Refer to the computer code and output of Figures 7–10 for an illustration of the computations of the Wilson 95% confidence interval for a simple example. For greater generality we will provide these expressions in terms of k, where k is the multiplier for the desired confidence level. For a 95% confidence level $k = 1.96$ and for a 90% confidence level $k = 1.645$. The Wilson confidence interval estimate of p is given by

$$\tilde{p}_k - \text{ME}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{ME}(\tilde{p}_k),$$

(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** $\tilde{p}_k$

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

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 $\tilde{p}_k - \text{ME}(\tilde{p}_k)$ and $\tilde{p}_k + \text{ME}(\tilde{p}_k)$. As noted above, 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 $\text{ME}(\tilde{p}_k)$ 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 $\tilde{p} - \text{ME}(\tilde{p})$ and $\tilde{p} + \text{ME}(\tilde{p})$) 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.

Aside – derivation of the Wilson interval formula.

This aside contains an algebraic derivation of the Wilson confidence interval for p. The starting point for this derivation is the interval $p - k\text{SE}(\hat{p}) \leq \hat{p} \leq p + k\text{SE}(\hat{p})$ for $\hat{p}$. Notice that we can re–express this relationship as $|\hat{p} - p| \leq k\text{SE}(\hat{p})$. Since $|\hat{p} - p|$, k, and $\text{SE}(\hat{p}) = \sqrt{p(1-p)/n}$ are positive, we can square each side of the inequality

$$|\hat{p} - p| \leq k\text{SE}(\hat{p})$$

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

$$\tilde{p}_k \pm \text{ME}(\tilde{p}_k),$$

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

Thus, letting C denote the desired confidence level, the probability statement

$$P\big[|\hat{p} - p| \leq k\text{SE}(\hat{p})\big] = C.$$

is equivalent to the probability statement

$$P\big[\tilde{p}_k - \mathrm{ME}(\tilde{p}_k) \le p \le \tilde{p}_k + \mathrm{ME}(\tilde{p}_k)\big] = C.$$

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

$$\tilde{p}_k - \mathrm{ME}(\tilde{p}_k) \le p \le \tilde{p}_k + \mathrm{ME}(\tilde{p}_k).$$

Testing a hypothesis about p

In the apple orchard example we used a confidence interval estimate of p to decide whether the data supported the contention that more than 20% of the apple trees in the entire orchard were infested. We will now develop some formal methodology for assessing the evidence in favor of such a contention. We will start with some terminology.

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.

We will introduce hypothesis testing in the context of the apple orchard example. Details and formalization will follow the example.

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. For example, when $n = 200$ and $p = .2$, the central 95% interval for $\hat{p}$ ranges from .1446 to.2554 and the central 99% interval for $\hat{p}$ ranges from .1272 to.2728. Therefore, if exactly 20% of all the trees in the orchard were infested, then it would be surprising to see much more than about 25% infested trees in the sample (above the central 95% interval) and it would be very surprising to see more than about 27% infested trees in the sample (above the central 99%

interval). On the other hand, if exactly 30% of all the trees in the orchard were infested, then the central 95% interval for the sample percentage ranges from 23.65% to 36.35% and the central 99% interval for the sample percentage ranges from 21.65% to 38.35%, so that sample percentages around 25% to 28% would not be surprising. 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. That is, if $\hat{p}$ is large enough to make us doubt our tentative assumption that p itself is not larger than .20.

Figure 12a. SAS output for the apple orchard example – tests and confidence intervals. The commands to produce this output are provided in Figure 12b.

apple orchard example

The FREQ Procedure

case=5

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	52	26.00	52	26.00
notinfested	148	74.00	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.2600
ASE	0.0310

Confidence Limits for the Binomial Proportion		
Proportion = 0.2600		
Type	95% Confidence Limits	
Wilson	0.2041	0.3249

Test of H0: Proportion = 0.2	
ASE under H0	0.0283
Z	2.1213
One-sided Pr > Z	0.0169
Two-sided Pr > |Z|	0.0339

Sample Size = 200

apple orchard example

The FREQ Procedure

case=6

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	45	22.50	45	22.50
notinfested	155	77.50	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.2250
ASE	0.0295

Confidence Limits for the Binomial Proportion		
Proportion = 0.2250		
Type	95% Confidence Limits	
Wilson	0.1726	0.2877

Test of H0: Proportion = 0.2	
ASE under H0	0.0283
Z	0.8839
One-sided Pr > Z	0.1884
Two-sided Pr > |Z|	0.3768

Sample Size = 200

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 the SAS output of Figure 12a (case 5) this P–value

is labeled "one–sided Pr $> Z$". 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. Referring to the SAS output, we see that we can also conclude with 95% confidence that between 20.41% and 32.49% of all the trees in the entire orchard are infested. As expected, we find that the entire 95% confidence interval is above 20%. Note, however, that values as low as 20.41%, which is not much above 20%, are deemed plausible here.

The SAS commands which produced the output in Figure 12a are provided in Figure 12b.

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 the SAS output of Figure 12a (case 6) this P–value is labeled "one–sided Pr $> Z$". 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. Referring to the SAS output, we see that we can also conclude with 95% confidence that between 17.26% and 28.77% of all the trees in the entire orchard are infested. As expected, we see that the 95% confidence interval contains values which are both above and below 20%.

Figure 12b. SAS commands for the apple orchard example – tests and confidence intervals

SAS command file: Apple orchard examples cases 5 and 6

```
/* apple orchard example */
data orchard;
   input case outcome : $ 11. count;
/* the coding outcome : $ 11. indicates that we want
                       to allocate 11 characters for outcome */
cards;
5 infested 52
5 notinfested 148
6 infested 45
6 notinfested 155
;
proc freq data=orchard order=data;
   tables outcome / alpha=.05 binomial(cl=wilson p=.2 level='infested');
   weight count;
   by case;
title 'apple orchard example';
/* binomial options
   -- cl=wilson Wilson confidence interval
   -- p=.2 use p_0=.2 in the hypothesis
   -- level='infested' define "success" default is
         the first level which is infested in this example */
run;
```

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.

Testing a directional hypothesis of the form $p > p_0$

Research question. Is there sufficient evidence to conclude that the population proportion p is greater than the hypothesized value p_0?

Research hypothesis. $H_1 : p > p_0$, The population proportion p is greater than the hypothesized value p_0.

Tentative assumption – null hypothesis. $H_0 : p \leq p_0$, We tentatively assume that the population proportion p is not greater than the hypothesized value p_0.

Evidence in favor of the research hypothesis. 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 $H_1 : p > p_0$.

Assessment of the strength of the evidence – the P–value of the test. Deciding whether the observed value of $\hat{p}$ is "sufficiently large relative to p_0" is based on the P–value of the test. 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.

Computation of the P–value. The P–value of the test 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 $H_0 : p \leq p_0$.

The derivation below is meant to clarify the procedure. We can use a suitable calculator or computer program to perform these computations. We will use the normal approximation to the sampling distribution of $\hat{p}$ to compute the P–value. As noted above we will use the hypothesized value p_0 in our computation of the P–value. Thus we will use the population standard deviation of $\hat{p}$

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

in our computation of the Z–score. The calculated Z statistic or Z score corresponding to the observed value of $\hat{p}$, denoted by Z_{calc}, is

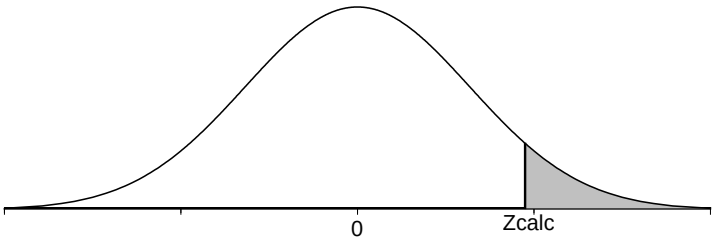
$$Z_{calc} = \frac{\hat{p} - p_0}{\text{SE}(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

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 13. 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 13. 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. Some guidelines for interpreting a P–value are provided below.

Returning to our discussion for the directional research hypothesis $H_1 : p > p_0$. The final steps for performing a hypothesis test for

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

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \geq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = (\hat{p} - p_0)/\text{SE}(\hat{p})$, and $\text{SE}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. This P–value is the area under the standard normal density curve to the right of Z_{calc}, as shown in Figure 13. This P–value is the "one–sided Pr > Z" of the SAS output.

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 $H_0 : p \leq p_0$. 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 $H_0 : p \leq p_0$. 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.

Remarks and guidelines about P–values.

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).

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.

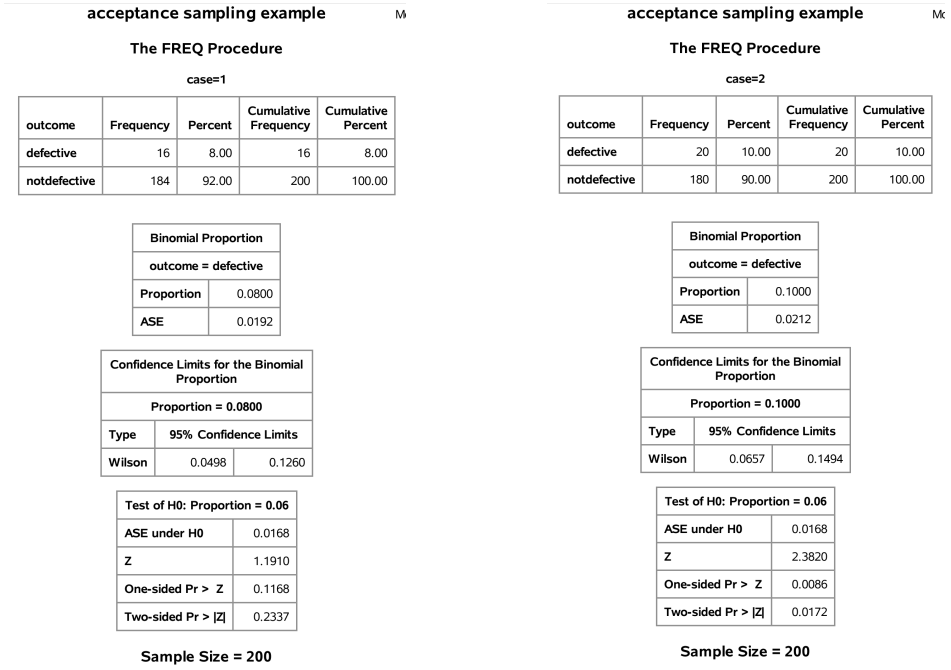
In this example one of these electronic devices is a unit and the collection of 10,000 units constituting the shipment is the population. Notice that, in this example, the target population and the sampled population are the same (each is the shipment of 10,000 devices). A suitable variable for the indicated 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

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

Figure 14. SAS output for the acceptance sampling example



acceptance sampling example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
defective	16	8.00	16	8.00
notdefective	184	92.00	200	100.00

Binomial Proportion	
outcome = defective	
Proportion	0.0800
ASE	0.0192

Confidence Limits for the Binomial Proportion		
Proportion = 0.0800		
Type	95% Confidence Limits	
Wilson	0.0498	0.1260

Test of H0: Proportion = 0.06	
ASE under H0	0.0168
Z	1.1910
One-sided Pr > Z	0.1168
Two-sided Pr > |Z|	0.2337

Sample Size = 200

acceptance sampling example

The FREQ Procedure

case=2

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
defective	20	10.00	20	10.00
notdefective	180	90.00	200	100.00

Binomial Proportion	
outcome = defective	
Proportion	0.1000
ASE	0.0212

Confidence Limits for the Binomial Proportion		
Proportion = 0.1000		
Type	95% Confidence Limits	
Wilson	0.0657	0.1494

Test of H0: Proportion = 0.06	
ASE under H0	0.0168
Z	2.3820
One-sided Pr > Z	0.0086
Two-sided Pr > |Z|	0.0172

Sample Size = 200

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

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

and the P–value is

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

In the SAS output of Figure 14 this P–value is labeled "one–sided Pr $> Z$. 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

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

and the P–value is

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

In the SAS output of Figure 14 this P–value is labeled "one–sided Pr $> Z$. 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 $\hat{p} = .08$. From the SAS output in Figure 14, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is between .0498 and .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 $\hat{p} = .10$. From the SAS output in Figure 14, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is between .0656 and .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).

Testing a directional hypothesis of the form $p < p_0$

The procedure for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ given above is readily modified for testing the null hypothesis $H_0 : p \geq p_0$ 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 $H_0 : p \geq p_0$. 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 $H_0 : p \geq p_0$. Therefore, the P–value for testing $H_0 : p \geq p_0$ 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

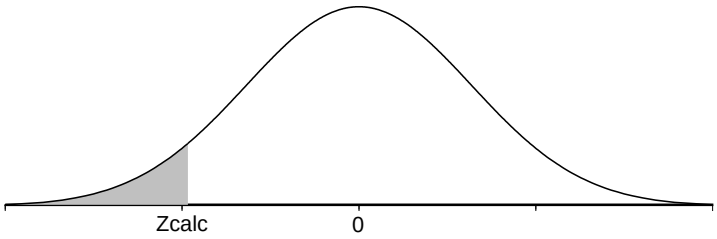
$$H_0 : p \geq p_0 \quad \text{versus} \quad H_1 : p < p_0$$

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, $Z_{calc} = (\hat{p} - p_0)/\text{SE}(\hat{p})$, and $\text{SE}(\hat{p}) =$

$\sqrt{p_0(1-p_0)/n}$. This P–value is the area under the standard normal density curve to the left of Z_{calc} as shown in Figure 15. This P–value is the "one–sided Pr < Z" of the SAS output.

Figure 15. 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.

Bernoulli trials

Some applications of these inferential methods for a proportion, such as the following machine parts example, correspond to a sequence of n trials. In this context a dichotomous 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 example 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 $p = 2/3$.

Note that this process of tossing a die is abstractly the same as the process of selecting a ball at random from a box containing six balls with the balls numbered from 1 to 6. Thus tossing the die n times is equivalent to selecting a simple random sample of size n with replacement from this box containing six balls.

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

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

The calculated Z statistic is

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

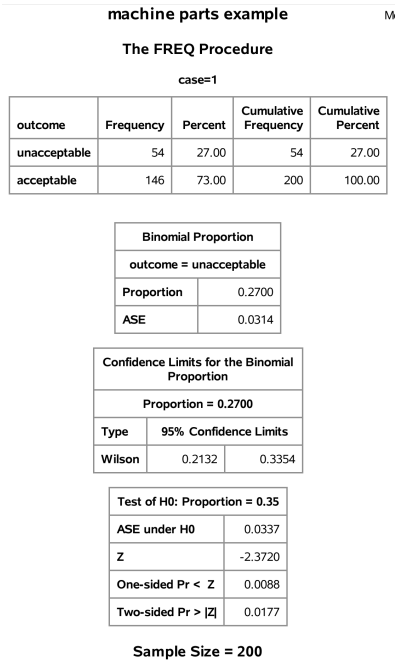
and the P–value is

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

You can find these values in the SAS output of Figure 16. 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 $\hat{p} = .27$ and a 95% confidence interval ranging from .2132 to .3354. Therefore, we are 95% confident that the probability of a part produced using the new production process being unacceptable is between .2132 and .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.

Figure 16. SAS output for the machine parts example



machine parts example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
unacceptable	54	27.00	54	27.00
acceptable	146	73.00	200	100.00

Binomial Proportion	
outcome = unacceptable	
Proportion	0.2700
ASE	0.0314

Confidence Limits for the Binomial Proportion		
Proportion = 0.2700		
Type	95% Confidence Limits	
Wilson	0.2132	0.3354

Test of H0: Proportion = 0.35	
ASE under H0	0.0337
Z	-2.3720
One-sided Pr < Z	0.0088
Two-sided Pr > |Z|	0.0177

Sample Size = 200

Testing a nondirectional research hypothesis

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.

Testing a nondirectional research (alternative) hypothesis of the form $p \neq p_0$.

Research question. Is there sufficient evidence to conclude that the population proportion p is different from the hypothesized value p_0?

Research hypothesis. $H_1 : p \neq p_0$, The population proportion p is not equal to the hypothesized value p_0.

Tentative assumption – null hypothesis. $H_0 : p = p_0$, We tentatively assume that the population proportion p is exactly equal to the hypothesized value p_0.

Evidence in favor of the research hypothesis. As with the directional hypothesis cases, 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 values of $\hat{p}$ farther away from p_0 more often when the research hypothesis $H_1 : p \neq p_0$ is true than when the null hypothesis $H_0 : p = p_0$ is true. In particular, we can view the observation of a value of $\hat{p}$ that is sufficiently far away from p_0, in either direction, as constituting evidence against the null hypothesis $H_0 : p = p_0$ and in favor of the research hypothesis $H_1 : p \neq p_0$.

Assessment of the strength of the evidence – the P–value of the test. Deciding whether the observed value of $\hat{p}$ is "sufficiently far away from p_0 in either direction" is based on the P–value of the test. The P–value for testing the null hypothesis $H_0 : p = p_0$ versus the research hypothesis $H_1 : p \neq p_0$ is the probability of observing a value of $\hat{p}$ for which the distance $|\hat{p} - p_0|$ (the absolute value of the difference between $\hat{p}$ and p_0) is as large or larger than the value of this distance that we actually do observe.

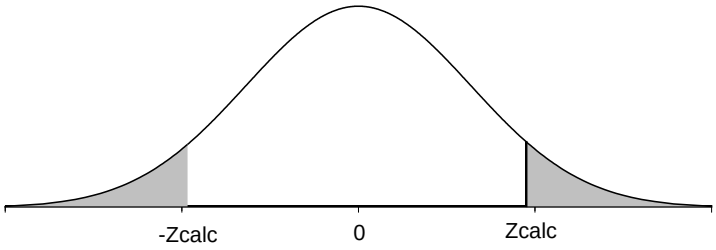
Computation of the P–value. The P–value of the test is computed under the assumption that the research hypothesis $H_1 : p \neq p_0$ is false and the null hypothesis $H_0 : p = p_0$ is true. In this situation the calculated Z statistic Z_{calc} is the absolute value of the Z statistic that would be used for testing a directional hypothesis. That is, the calculated Z statistic is

$$Z_{calc} = \left| \frac{\hat{p} - p_0}{\mathrm{SE}(\hat{p})} \right| = \left| \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right|.$$

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}, as shown in Figure 17. 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 $p = p_0$. Notice that this P value (the area to the left of $-Z_{calc}$ plus the area to the right of Z_calc) is small when Z_{calc} is far away from zero in one direction or the other which is equivalent to $\hat{p}$ being far away from p_0 in one direction or the other.

Figure 17. P–value for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$.



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, $Z_{calc} = |(\hat{p} - p_0)/\text{SE}(\hat{p})|$, and $\text{SE}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. 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 17. This P–value is the "two–sided Pr $> |Z|$" of the SAS output.

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

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

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

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

You can find these values in the SAS output of Figure 18. 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.

Figure 18. SAS output for the Mendel pea flower color example

Mendel pea flower color example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
red	705	75.89	705	75.89
white	224	24.11	929	100.00

Binomial Proportion	
outcome = red	
Proportion	0.7589
ASE	0.0140

Confidence Limits for the Binomial Proportion		
Proportion = 0.7589		
Type	95% Confidence Limits	
Wilson	0.7303	0.7853

Test of H0: Proportion = 0.75	
ASE under H0	0.0142
Z	0.6251
One-sided Pr > Z	0.2660
Two-sided Pr > |Z|	0.5319

Sample Size = 929

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 "nondirectional" confidence intervals of the form $p_L \leq p \leq p_U$. A 95% confidence interval of this form, consisting of a lower bound p_L for p and an upper bound p_U for p, gives 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.

A lower confidence bound on p allows us to estimate the smallest value of p which is plausible in light of the observed value of $\hat{p}$. Similarly, an upper confidence bound on p allows us to estimate the largest value of p which is plausible in light of the observed value of $\hat{p}$. For example, in the acceptance sampling example we might argue that we are less concerned with a limit on how large p might be than with a limit on 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.

The reasoning which led to the "nondirectional" Wilson interval (a range of values for p) can be adapted to yield a directional interval (a lower or upper confidence bound for p). Recall that our development of the Wilson 95% confidence interval estimate of p was based on the observation that, for each possible value of p, we can view the central 95% interval from $p - 1.96\text{SE}(\hat{p})$ to $p + 1.96\text{SE}(\hat{p})$ as an interval which is likely to contain $\hat{p}$. This starting point led us to a confidence interval estimate of p which provided a range of values between a lower limit and an upper limit. If we use an upper 95% interval for $\hat{p}$ or a lower 95% interval for $\hat{p}$ instead, we will end up with a directional confidence interval, *i.e.*, we will end up with a 95% confidence bound (lower or upper) for p. In practice, 95% confidence bounds are usually computed by selecting the appropriate endpoint of a 90% confidence interval. That is, if $p_L \leq p \leq p_U$ is a 90% confidence interval estimate of p, then P_L is a 95% lower confidence bound for p and p_U is a 95% upper confidence bound for p.

We will illustrate some applications of confidence bounds in the context of some of the examples we discussed earlier. Detailed derivations of 95% lower and upper confidence bounds are provided after the examples.

Example. Acceptance sampling for electronic devices (revisited). Recall that in this example the retailer had received a shipment of 10,000 electronic devices from a supplier with a guarantee that no more than 6% of these devices were defective. The retailer is interested in p, the value of the proportion of defective devices in the shipment of 10,000 devices. In particular, the retailer is concerned that this proportion might be too large (greater than .06). Thus we proposed a test of the null hypothesis $H_0 : p \leq .06$ versus the research hypothesis $H_1 : p > .06$. We considered two cases to illustrate the

corresponding hypothesis testing procedure. We will now show how a directional confidence bound can supplement the formal hypothesis test.

As noted above, in this example, we are less concerned with a limit on how large p might be than with a limit on 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. Note that a lower confidence bound will always include values greater than .06 (6%). The important consideration is whether it also includes values less than .06.

Figure 19. SAS output for the acceptance sampling example

acceptance sampling example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
defective	16	8.00	16	8.00
notdefective	184	92.00	200	100.00

Binomial Proportion	
outcome = defective	
Proportion	0.0800
ASE	0.0192

Confidence Limits for the Binomial Proportion		
Proportion = 0.0800		
Type	90% Confidence Limits	
Wilson	0.0538	0.1174

acceptance sampling example

The FREQ Procedure

case=2

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
defective	20	10.00	20	10.00
notdefective	180	90.00	200	100.00

Binomial Proportion	
outcome = defective	
Proportion	0.1000
ASE	0.0212

Confidence Limits for the Binomial Proportion		
Proportion = 0.1000		
Type	90% Confidence Limits	
Wilson	0.0703	0.1404

Case 1 If there are 16 defective devices in a sample of $n = 200$, then the observed proportion of defective devices is $\hat{p} = .08$. The lower endpoint .0538 of the 90% confidence interval in the SAS output for case 1 in Figure 19 is the 95% confidence lower bound for p. Hence, we can conclude that we are 95% confident that the actual percentage of defective devices in the shipment of 10,000 is at least 5.38%. More importantly, with 95% confidence we can say that the percentage could be as low as 5.38%. Since this lower bound is less than 6% we do not have sufficient evidence to claim that more than 6% of the 10,00 devices are defective.

Recall that, in this case, the P–value for testing $H_0 : p \leq .06$ versus $H_1 : p > .06$ was .1168 and we concluded that there is not sufficient evidence to claim that more than

6% of the 10,000 devices are defective. Thus the lower confidence bound and the formal hypothesis test lead to the same conclusion.

Case 2 If there are 20 defective devices in a sample of $n = 200$, then the observed proportion of defective devices is $\hat{p} = .10$. The lower endpoint .0703 of the 90% confidence interval in the SAS output for case 2 in Figure 19 is the 95% confidence lower bound for p. In this case we can conclude that we are 95% confident that the actual percentage of defective devices in the shipment of 10,000 is at least 7.03%. That is, with 95% confidence we can say that the percentage is not less than 7.03%. Since this lower bound is greater than 6% we have sufficient evidence to claim that more than 6% of the 10,000 devices are defective. Note also that the lower bound 7.03% indicates that the percentage of defective devices among the 10,000 is at least 1.03 percentage points higher than the 6% cutoff value.

In this case, the P–value for testing $H_0 : p \leq .06$ versus $H_1 : p > .06$ was .0086 and we concluded that there is strong evidence that more than 6% of the 10,000 devices are defective. Again, the lower confidence bound and the test lead to the same conclusion.

Example. Machine parts (revisited). Recall that, in this example, 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. In this example p denotes the probability that a part produced using the new production process will be unacceptable and our goal is to decide whether this probability is less than .35. A sample of 200 parts was produced using the new production process and 54 of these parts were found to be unacceptable. The SAS output for this example is provided in Figure 20. In this example, the P–value for testing $H_0 : p \geq .35$ versus $H_1 : p < .35$ is .0088 and we concluded that 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. Using the upper endpoint of the 90% confidence interval in the SAS output of Figure 20, we can conclude that we are 95% confident that the probability that a part produced using the new production process will be unacceptable is no larger than .3245. Since this value is less than .35, the 95% confidence upper bound leads to the conclusion that the new production process is superior to the old process and indicates the extent of the improvement.

Figure 20. SAS output for the machine parts example

machine parts example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
unacceptable	54	27.00	54	27.00
acceptable	146	73.00	200	100.00

Binomial Proportion	
outcome = unacceptable	
Proportion	0.2700
ASE	0.0314

Confidence Limits for the Binomial Proportion		
Proportion = 0.2700		
Type	90% Confidence Limits	
Wilson	0.2217	0.3245

Test of H0: Proportion = 0.35	
ASE under H0	0.0337
Z	-2.3720
One-sided Pr < Z	0.0088
Two-sided Pr > |Z|	0.0177

For completeness, we will now provide detailed derivations of 95% lower and upper confidence bounds. The probability that a standard normal variable Z takes on a value less than 1.645 is equal to .95, $P(Z \leq 1.645) = .95$. That is, when we observe the value of a standard normal variable Z, 95% of the time we will find that $Z \leq 1.645$. Graphically this means that the area under the standard normal density curve over the interval from $-\infty$ to 1.645 is .95. Thus, for sufficiently large values of n we have the approximation,

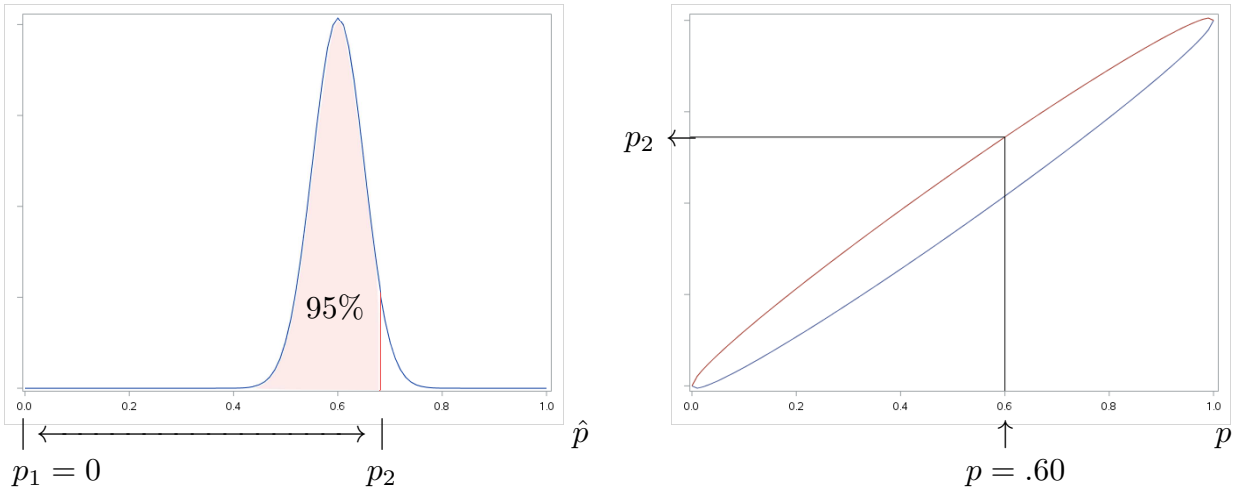
$$P\left[\frac{\hat{p} - p}{\text{SE}(\hat{p})} \leq 1.645\right] = .95.$$

Note that this indicates that 95% of the time when a simple random sample is selected and $\hat{p}$ is computed the observed value of $\hat{p}$ will be between zero and $p + 1.645\text{SE}(\hat{p})$, *i.e.*, $\hat{p}$ will be no more than 1.645 population standard error units above p. We will refer to the interval from zero to $p + 1.645\text{SE}(\hat{p})$ as the lower 95% interval of the distribution of $\hat{p}$, since it will contain the observed value of $\hat{p}$ 95% of the time.

The plots in Figure 21 show how the lower 95% interval of the distribution of $\hat{p}$ depends on the value of p. On the left we have a representation of the lower 95% interval, the interval from $p_1 = 0$ to p_2. The plot on the right shows how p_2 depends on p. We want to determine the values of p which yield lower 95% intervals which contain $\hat{p}_{\text{obs}}$. To

do this we need to use the graph on the right of Figure 21 in the other direction, as shown in Figure 22. In Figure 22 a horizontal line is drawn at $\hat{p}_{\text{obs}}$ and its intersection, p_L, with the upper curve is indicated. Notice that p_L is the smallest value of p for which the lower 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (Figure 22 is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. In this case, $p_L = .4679$.) Thus, if we draw a vertical line, as in Figure 21, at any value of p between p_L and $p_U = 1$, then the corresponding lower 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. Note that p_L is the 95% confidence lower bound for p.

Figure 21. The plot on the left shows the lower 95% interval of the distribution of $\hat{p}$ for $n = 100$ and $p = .6$.
The upper (red) curve in the plot on the right shows the upper endpoint, $p + 1.645\sqrt{p(1-p)/n}$, of the lower 95% interval of the distribution of $\hat{p}$ as a function of p for $n = 100$. The upper endpoint for the case $p = .6$ is indicated by the line marking the intersections at $p_2 = .6806$. (The lower endpoint is zero.)



In order to determine the value of p_L we simply set $p + 1.645\text{SE}(\hat{p})$ equal to $\hat{p}_{\text{obs}}$ and solve for p (draw the horizontal line as in Figure 22 and project down at the intersection with the upper (red) curve).

An analogous argument starting with the upper 95% interval of the distribution of $\hat{p}$, *i.e.*, the interval from $p - 1.645\text{SE}(\hat{p})$ to one, leads to a 95% confidence upper bound for p. In this case, as shown in Figure 23, we start with the upper 95% interval of the distribution of $\hat{p}$, the interval from $p_1 = p - 1.645\text{SE}(\hat{p})$ to one. The graph on the right in Figure 23 shows how the upper 95% interval of the distribution of $\hat{p}$ depends on value of p.

Figure 22. The smallest value of p, p_L, for which the lower 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (The interval goes from p_L to $p_U = 1$.)
This example is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. Here $p_L = .4679$.

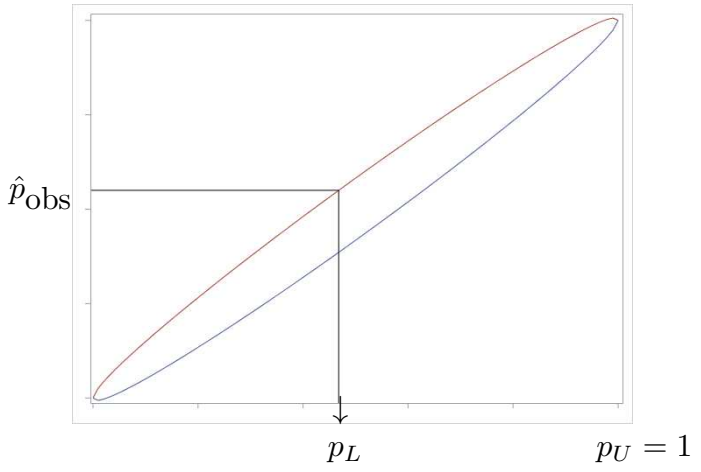
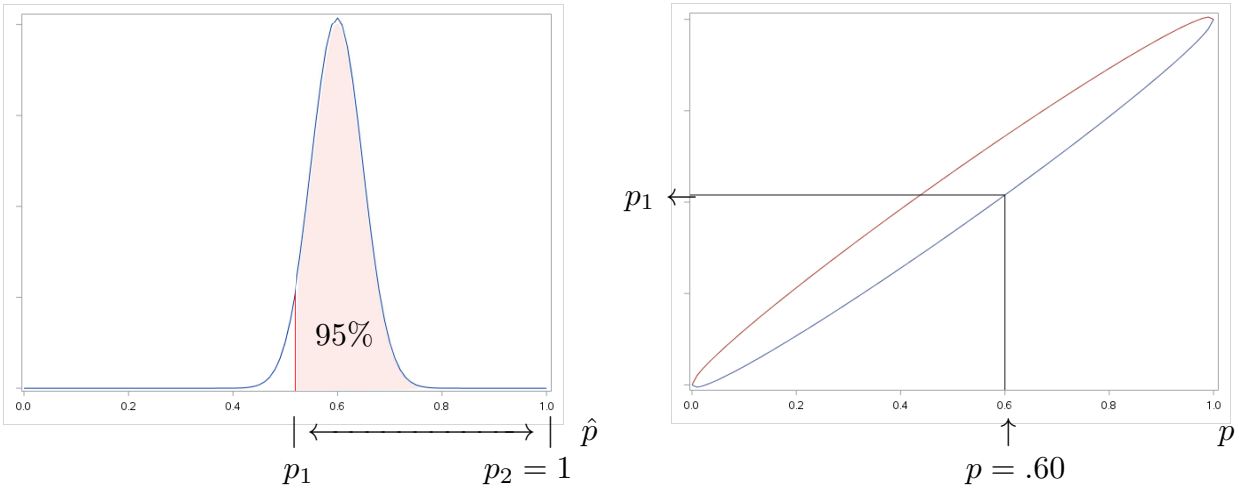


Figure 23. The plot on the left shows the upper 95% interval of the distribution of $\hat{p}$ for $n = 100$ and $p = .6$.
The lower (blue) curve in the plot on the right shows the lower endpoint, $p - 1.645\sqrt{p(1-p)/n}$, of the upper 95% interval of the distribution of $\hat{p}$ as a function of p for $n = 100$. The lower endpoint for the case $p = .6$ is indicated by the line marking the intersections at $p_1 = .6806$. (The upper endpoint is one.)



In Figure 24 a horizontal line is drawn at $\hat{p}_{\text{obs}}$ and its intersection, p_U, with the lower curve is indicated. Notice that p_U is the largest value of p for which the upper central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (Figure 24 is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. In this case, $p_U = .6806$.) Thus, if we draw a vertical line, as in Figure 23, at

any value of p between $p_L = 0$ and p_U, then the corresponding upper 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. Note that p_U is the 95% confidence upper bound for p.

In order to determine the value of p_U we simply set $p - 1.645\text{SE}(\hat{p})$ equal to $\hat{p}_{\text{obs}}$ and solve for p (draw the horizontal line as in Figure 24 and project down at the intersection with the lower (blue) curve).

Figure 24. The largest value of p, p_U, for which the upper 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (The interval goes from $p_L = 0$ to p_U.)
This example is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. Here $p_U = .6294$.