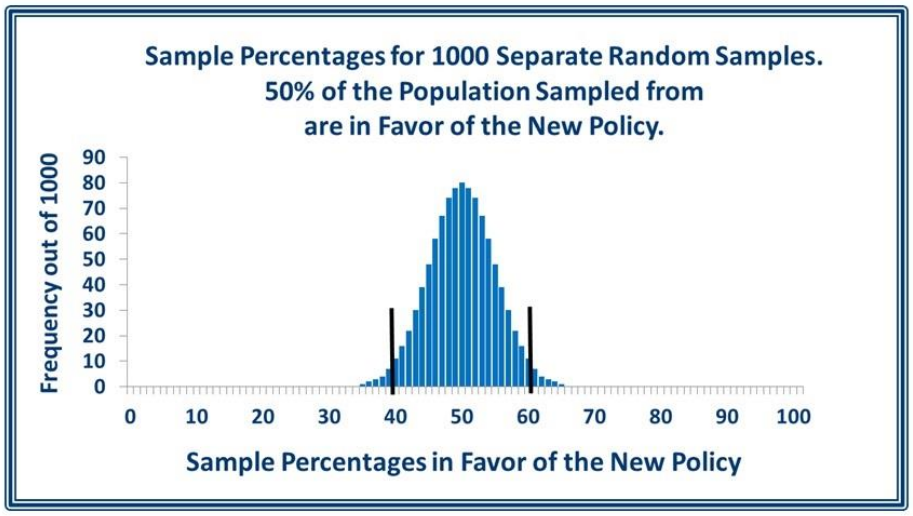
Statistical Analysis Illustrated



Foundations You Should Know

J. E. Kottemann

Statistical Analysis Illustrated
Foundations You Should Know

Introduction

Many people find statistical analysis puzzling. Why? Because it demands new ways of thinking. And how do you fathom these new ways of thinking? In order to fathom statistical analysis properly, you need to understand a number of its key foundations. Reading textbooks doesn't work because textbooks don't focus on foundations, and they cover so much ground in so much detail that readers are bound to lose sight of the forest for the trees. I remedy this state of affairs in a number of ways.

The first thing I do is focus on an application area that everyone should be familiar with—public opinion surveys. Second, I focus on the most straightforward type of survey questions and responses—Do you agree or disagree? Do you approve or disapprove?

The third thing I do is focus on foundations that are essential to understanding the nature of statistical analysis and the interpretation of results, be they public opinion survey results or pharmaceutical drug trial results. And fourth, I use a lot of illustrations.

Even if you've taken statistics courses in the past, and even if you know the mechanics of performing statistical analysis, you'll benefit from understanding these foundations.

The first group of essential foundations concerns the nature of sample statistics.

- Sampling distributions of statistic values are the key to everything. That's why there's a sampling distribution illustration on the front cover, as well as sampling distribution illustrations throughout the book.
- The lines we draw—literally and figuratively—define an interval that's used to decide whether or not a hypothesis should be rejected. You can see two interval lines on the cover too, superimposed on the sampling distribution.
- Two key facets—variance and sample size—determine a sampling distribution's shape and its interval's width. We'll see how these things interact.

With sampling distributions and confidence intervals we can conduct statistical analysis to our heart's content. But we do need to be careful when interpreting our results.

- Certain hypotheses can be *accepted*, but others can only be *not rejected*. There is a subtle but important difference between these two.
- The meaning of statistical significance is limited and it does not imply that there are meaningful real-world implications at hand.
- There are errors we will make that we will not know we've made, and probably never will know. But we can try to set limits on how often they might occur.
- Reducing the likelihood of one type of error increases the likelihood of the other. We can reduce the likelihood of both, but that costs extra.
- Strange things happen when we use a statistical analysis method without satisfying its assumptions.

This book, as noted, concentrates on survey/response data such as "Do you agree or disagree?" This allows us to concentrate on one type of sampling distribution: the normal distribution. But we'll look at more than that.

- In the second to last chapter, we'll look at other types of data and statistics. Through them, we will meet all of the "big four" distributions.

And this book concentrates on the most common type of statistical analysis: Frequentist analysis. But we won't overlook its worthy competitor: Bayesian analysis.

- In the last chapter, we'll see the one key difference between Frequentist and Bayesian statistical analysis that makes a world of difference. This chapter systematically introduces Bayesian analysis while pointing out where it differs from Frequentist analysis.

Each of these essential foundations are explored with many illustrations accompanied by thorough explanations. But even with this focused, streamlined, illustrated treatment, learning the foundations of statistical analysis is not easy. It does demand new ways of thinking. So please be patient and read carefully.

1. Sampling Distribution Basics

Statistics and Their Sampling Distributions

The town of Flowing Wells is a community of 80,000 residents. Imagine that you and I serve on the Town Council and that we are members of its public affairs committee. The council has recently drafted a new public health policy and we're tasked with assessing the public's opinion about it. The public opinion *statistic* we're interested in is the percentage of residents that are in favor of the proposed policy. We want to find out if a majority of residents are in favor of the policy.

You and I realize that asking all 80,000 residents their opinions is next to impossible, so we'll need to use statistical sampling and analysis. To start with, we decide to survey 100 residents. To avoid inadvertent bias when gathering sample opinions, everyone in the community must have an equal chance of being surveyed. So, we'll select 100 people *at random* from the town's list of residents, giving us a *random sample* of 100 peoples' opinion.

After contacting the 100 randomly selected residents we find that 55 of the 100 (55%) are in favor of the policy. We next ask ourselves whether the sample percentage of 55% is high enough for us to be confident that a majority of at least 40,001 (over 50%) of the town's population favors the policy. In other words, is the *sample* percentage of 55% high enough for us to be confident that the Flowing Wells *population* percentage exceeds 50%? We need to do some statistical analysis.

I volunteer to simulate the situation on the computer. Since we are interested in whether a majority are in favor, I focus on the majority dividing line of 50% in favor and assume that we're sampling from a population that is 50% in favor. I use the computer to simulate what occurs when we randomly sample 100 people from a population that is 50% in favor of something. Further, *I simulate this random sampling many, many times* to show us the range and frequency of sample percentage values to expect when we randomly sample 100 people from a 50%-in-favor population.

Figure 1.1 is a chart (a histogram) showing what the simulations reveal. It shows the frequency (out of 1,000) with which we should expect to get random samples that yield the various possible values for the percent-in-favor sample statistic.

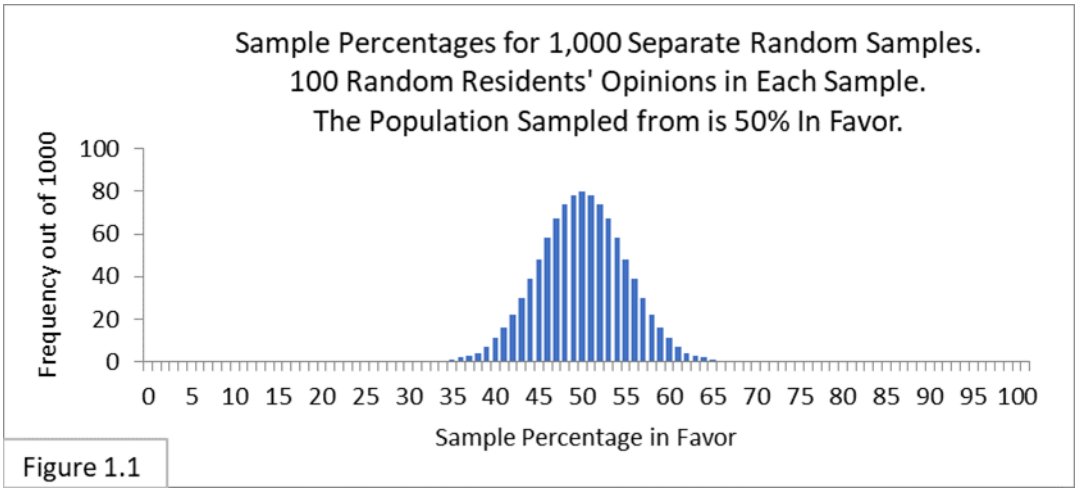


Figure 1.1

Figure 1.1 reflects the following theoretical scenario: 1) there is a large population that is 50% in favor of something; 2) 1,000 surveyors are hired and each one conducts a separate survey; 3) each of the 1,000 surveyors gathers a separate random sample of 100 people from the population to determine the percent in favor, and 4) the 1,000 separate sample percentages are used to make an aggregate chart, giving us Figure 1.1.

This notion of "what to expect when we randomly sample from a population many, many times" forms the basis of the most common approach to statistical analysis—*Frequentist* statistics: When we repeatedly sample from a given population, how frequently do we expect the various possible sample statistic values to arise? The chart in Figure 1.1 shows us what to expect. Charts like this illustrate what are called *sampling distributions*—the *distribution* of sample statistic values that occur with repeated random *sampling* from a population. The concept of sampling distributions is perhaps the most important concept in Frequentist statistics.

Looking at Figure 1.1, notice that the sampling distribution is centered on 50%, that most of the sample percentages we expect to get are relatively close to 50%, and that almost all of the sample percentages we expect to get are contained within the boundary lines of 35% and 65%. The sample percentage of 55% that we actually got is well within the 35%-to-65% boundary lines. What does that tell us? It tells us that getting a sample percentage value of 55% is not particularly unusual when

sampling 100 from a 50%-in-favor population. So, we can't rule out that we might indeed be sampling from a population that is 50% in favor. And since 50% is not a majority, we can't say that a majority of the Flowing Wells population favors the new policy. This type of logic, or *statistical inference*, is central to the Frequentist approach to statistical analysis.

Now let's suppose that our sample percentage is 70% instead of 55%. What would we infer then? Looking at Figure 1.1 (reproduced below) we can see that getting a sample percentage of 70% is highly unusual when randomly sampling 100 from a 50%-in-favor population, and that 70% is clearly to the right of the sampling distribution for 50%. Because of this, we'll infer that we are *not* sampling from a 50%-in-favor population, and that we are instead sampling from a population that is more than 50% in favor. With a sample percentage of 70% we'll say it seems likely that a majority of the Flowing Wells population favors the new policy.

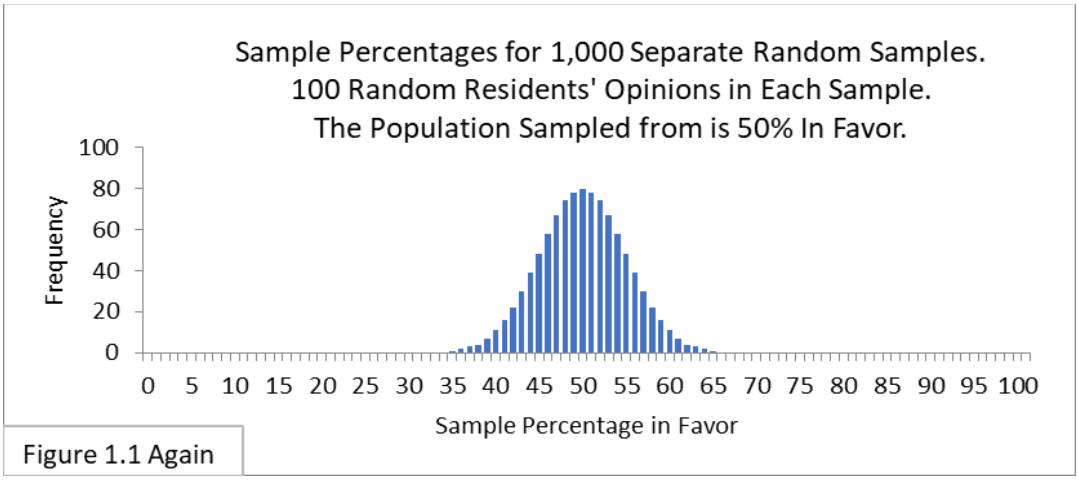


Figure 1.1 Again

Next, let's suppose that our sample percentage is 60% instead of 55% or 70%. What would we infer then? Looking at the sampling distribution in Figure 1.1, reproduced below, we can see that 60% is less clear cut. It does occur sometimes, but not very often. To make a judgment we'll need to "split hairs" using what are called *confidence intervals*.

Sampling Distributions and Their Confidence Intervals

Figure 1.2 shows what is called a *95% confidence interval*. It has the same sampling distribution as Figure 1.1, but now with two boundary lines added to indicate the 95% confidence interval spanning from 40% to 60%. That interval contains 95% (950/1,000) of the sample percentage values. That's what a 95% confidence means in this context: the interval containing 95% of the expected sample statistic values obtained by random sampling from a given population. 95% confidence intervals are widely used in statistical analysis. Some people find the term "confidence" in "confidence interval" to be confusing at first, so I'll often refer to it simply as the 95% interval.[i]

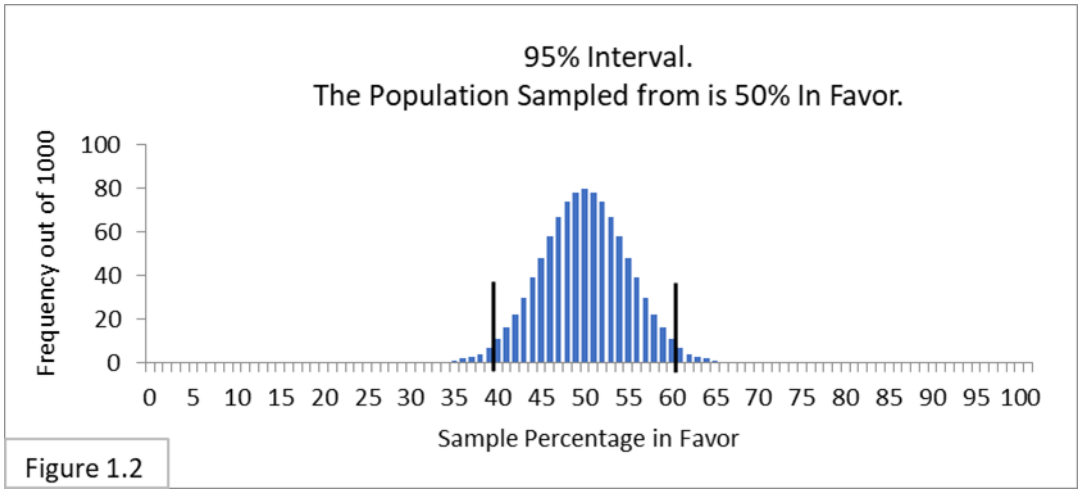


Figure 1.2

Recall that the survey sample percentage values we have considered so far are 55%, 60%, and 70%. Relative to the 40%-to-60% boundary lines, 55% and 60% are inside and 70% is outside. Using the 95% interval, we would say that the sample percentage values 55% and 60% *don't* allow us to infer that the (unknown) Flowing Wells population percentage value is different from 50%. After all, it is not very unusual to get those sample percentage values when randomly sampling 100 from a 50%-in-favor population.

On the other hand, the sample percentage value of 70% is outside the 40%-to-60% boundary lines, which *does* allow us to infer that the (unknown) Flowing Wells population percentage value is probably different from 50% and, further, is probably greater than 50%. That's because our sampling distribution shows that it's extremely unusual to get a sample percentage value of 70% when randomly sampling 100 people from a 50%-in-favor population, so we infer that we are not

sampling from a 50%-in-favor population. With a 70% sample percentage value we would say that we reject the hypothesis that the Flowing Wells population percentage is 50%.

What we are doing here is called *significance testing*. The difference between 50% and 70% is *statistically significant*. The differences between 50%, 55%, and 60% are not statistically significant.

It may seem like "hair splitting" to say that sample percentage values of 39%-or-less and 61%-or-more let us infer that a minority or a majority of the population is in favor of a policy, but that sample percentage values of 40%-to-60% do not. However, this is not the fault of statistical analysis itself, it's just that many real-world decisions force us to draw lines.

Coin Flipping Analogy I

To reinforce what we've covered so far, let's consider an analogy: Flipping a fair coin. A fair coin is perfectly balanced. When you flip it, there is a 50% chance it will come up heads and a 50% chance it will come up tails. This is analogous to randomly sampling from a population that is 50% in favor of and 50% not in favor of something: there is a 50% chance that a person randomly selected will be in favor and a 50% chance that a person randomly selected will not be in favor.

In this coin flipping experiment, we'll flip a fair coin 100 times in a trial, a trial being analogous to a sample. We'll perform 1,000 trials of 100 flips each. Figure 1.3 shows the results: the sampling distribution of the percentage of heads we expect to get with trial (sample) size of 100.

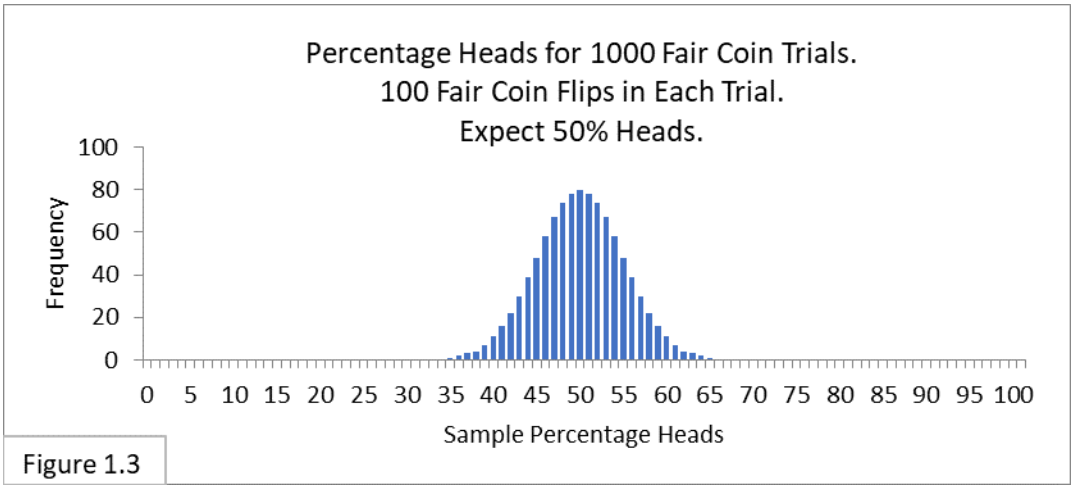


Figure 1.3

Notice that the sampling distribution in Figure 1.3 looks just like the sampling distribution in Figure 1.1 (reproduced below). They only differ in the text-labeling that's used to reflect the specific context (surveying vs. coin flipping).

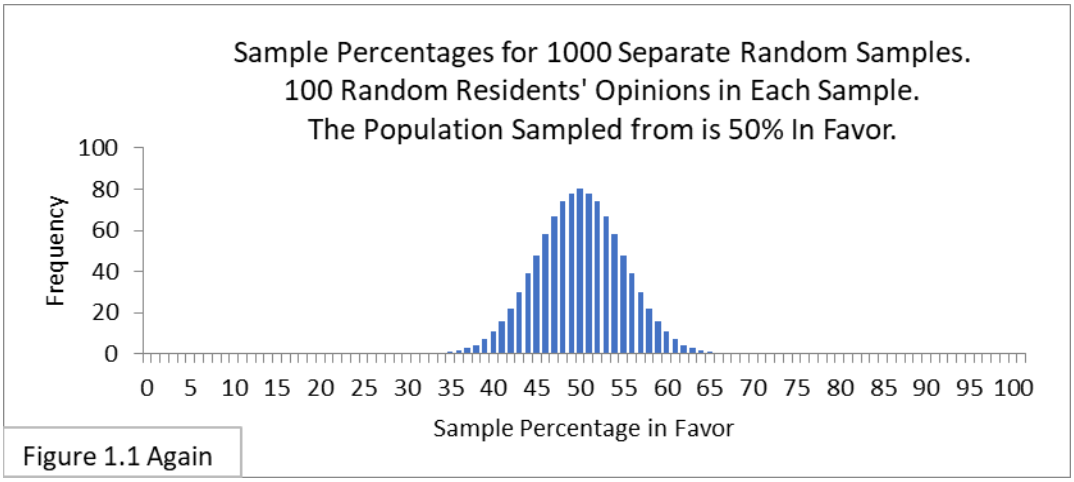


Figure 1.1 Again

Figure 1.4 shows the 95% interval of 40%-to-60% heads.

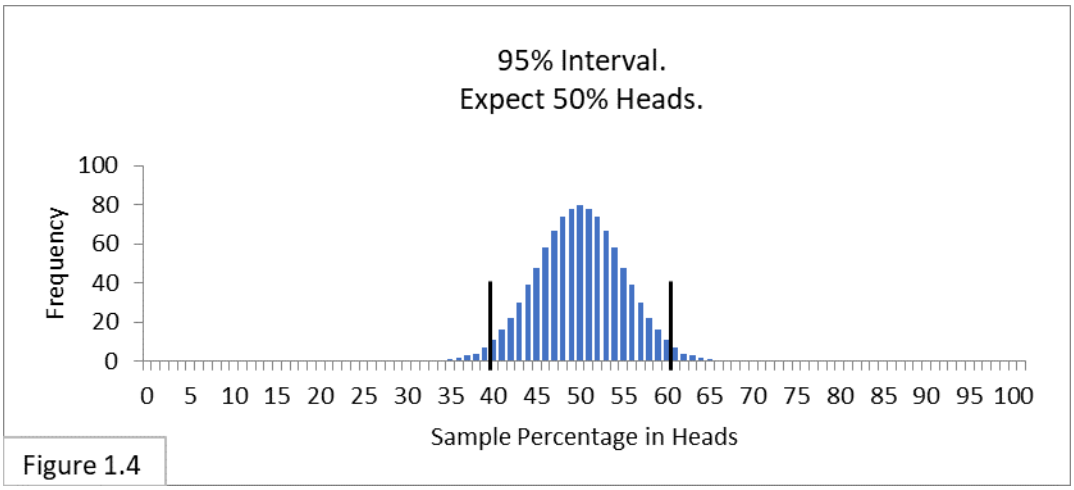


Figure 1.4

Notice that Figure 1.4 looks just like Figure 1.2 (reproduced below).

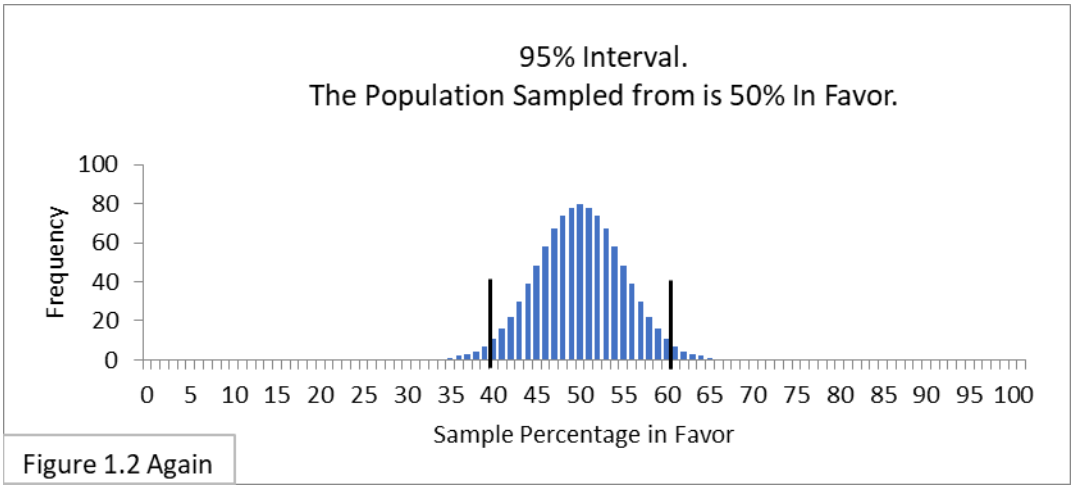


Figure 1.2 Again

If someone gave us a coin and asked us to judge whether it is a fair coin, we could flip it 100 times, and if we get fewer than 40 (40%) heads or more than 60 (60%) heads, we would say we didn't think the coin was fair. If we got anywhere between 40% and 60% heads, we would say we can't rule out that the coin is fair.

The reason these various Figures look alike is because they embody the same fundamental phenomenon. 1) They both involve randomness: random selection of respondents; randomness inherent in coin flips. Because of this, the items of interest—percent in favor and percent of heads—are called *random variables*. And 2) They both concern outcomes that have only two possible values: in favor vs. not in favor; heads vs. tails. Such things are called *binomial* random variables since there are only two possible values, and their sampling distributions are of a type called *binomial distributions*. Phenomena that share these two characteristics are analogous and can be analyzed in the same way: e.g., the percentage of people with a particular health condition, the percentage of defective items in a shipment, the percentage of students who passed an exam, the percentage of people who approve of the job the president is doing.

2. Sampling Distribution Dynamics

Sampling Distributions and Sample Sizes

We had decided to gather 100 opinions. Frankly, no deep thought went into that decision, 100 just seemed like a good sizable number. But, all things considered, the 40%-to-60% boundary lines are quite wide apart, hindering their usefulness: even with a sample percent-in-favor of 60% we can't infer that a majority of Flowing Wells residents are in favor of the new public health policy! Professional pollsters most often survey about 1000 people. Why is that? Let's see why.

Figure 2.1 shows the sampling distribution when randomly sampling 1000 people from a 50%-in-favor population. Notice how much narrower the sampling distribution is compared to our previous case of sampling 100 people. Why is that? Because as the size of our sample increases, the closer we expect the sample percentage values to be to the actual (but unknown) population percentage value. This is the *law of large numbers.*[ii] More evidence reduces our uncertainty and should get us closer to the truth. Later on, we'll see a simple formula that reflects the level of uncertainty due to sample size and its effect on the width of the sampling distribution.

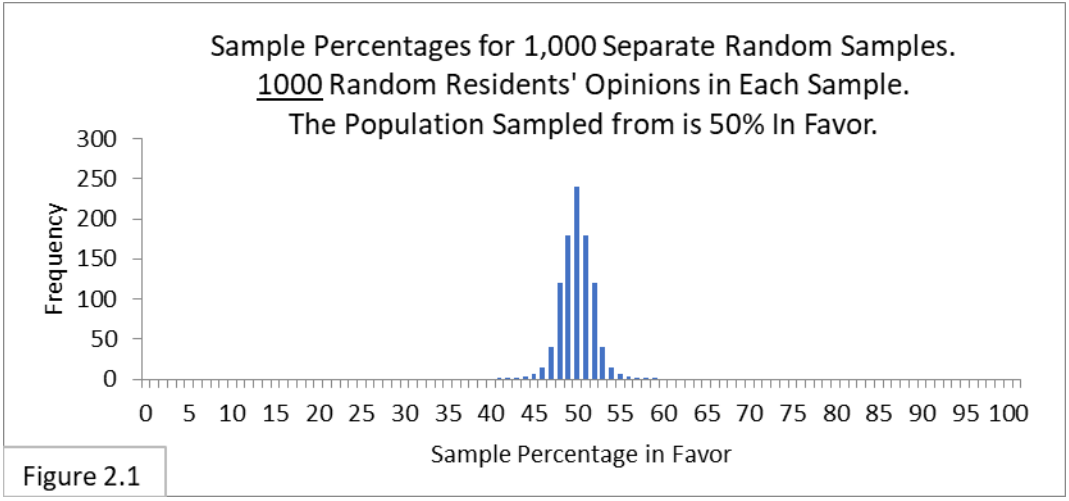


Figure 2.1

Now let's see how our 40%-to-60% boundary lines look when superimposed on the new, narrower sampling distribution. Figure 2.2 shows that.

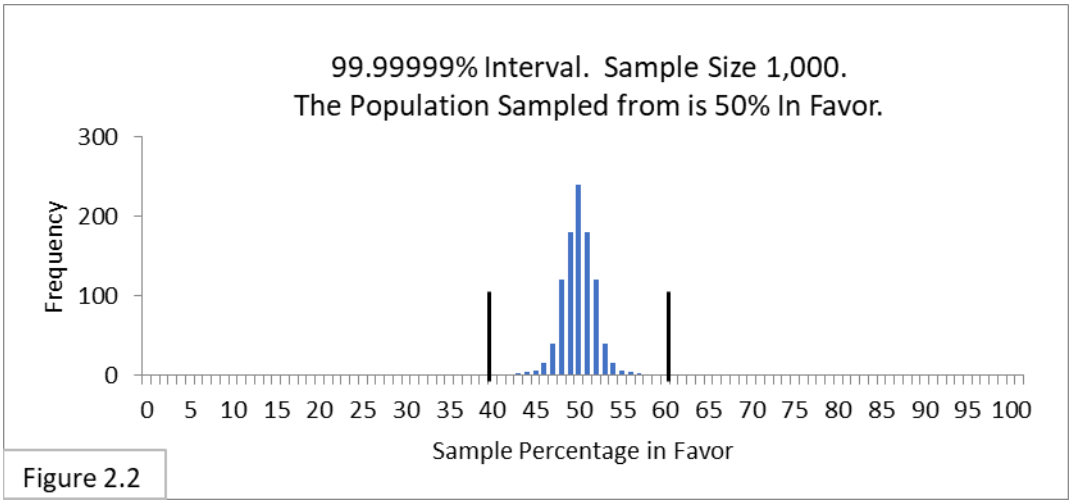


Figure 2.2

The 40%-to-60% boundary lines have been transformed from a 95% interval into a 99.99999% interval! 99.99999% of the sample percentage values we expect to get when sampling 1000 from a 50%-in-favor population are now within the 40%-to-60% boundary lines!

Let's fit a new 95% interval. Figure 2.3 shows the 95% interval superimposed on the new sampling distribution. Since the sampling distribution has become narrower, so has the 95% interval. With a sample size of 1000 the boundary lines are now 47%-to-53%. The 47%-to-53% boundary lines can also be expressed as 50%±3% (50% plus or minus 3%). Pollsters then refer to the 3% as the *margin of error* of the poll. (You may have noticed statements in the media such as "in the most recent poll the president's job approval rating is 49% with a margin of error of plus or minus 3 percentage points at the 95% confidence level.")

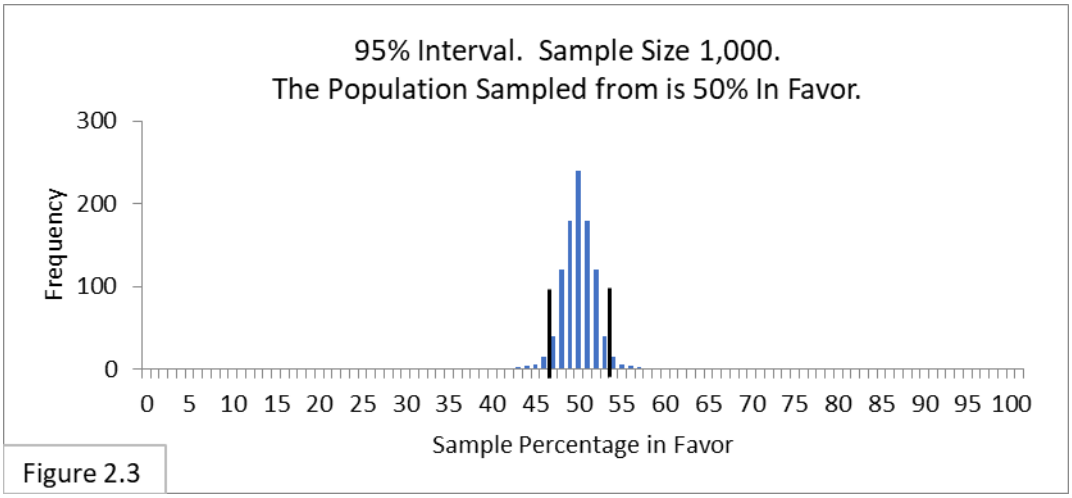


Figure 2.3

With the new 95% interval, the 55%, 60%, and 70% sample percentage values we've been considering are *all* outside the interval. All of them allow us to say it seems likely that a majority of the Flowing Wells population is in favor of the newly proposed public health policy. With sample size 100, we couldn't say 55% or 60% were significantly different, statistically, from 50%. With sample size 1000 we can. That's what increasing the sample size does for us. With the increased sample size, we've increased the *power* of our analysis. (We'll explore power in more detail later.)

This is why professional pollsters often survey 1000 people. It improves the power of the analysis over smaller sample sizes such as 100. So, you might ask, why not survey 10,000 to improve power even more? Because that would be more expensive. For most purposes, pollsters find sample sizes of about 1000 provide a good balance between power and expense. If more power is needed and the additional expense can be justified, larger sample sizes are used.

Coin Flipping Analogy II

In the next coin flipping experiment, we'll perform 1,000 trials (samples) of 1000 flips each (sample size). Figure 2.4 shows the sampling distribution for 1000 fair coin flips per trial.

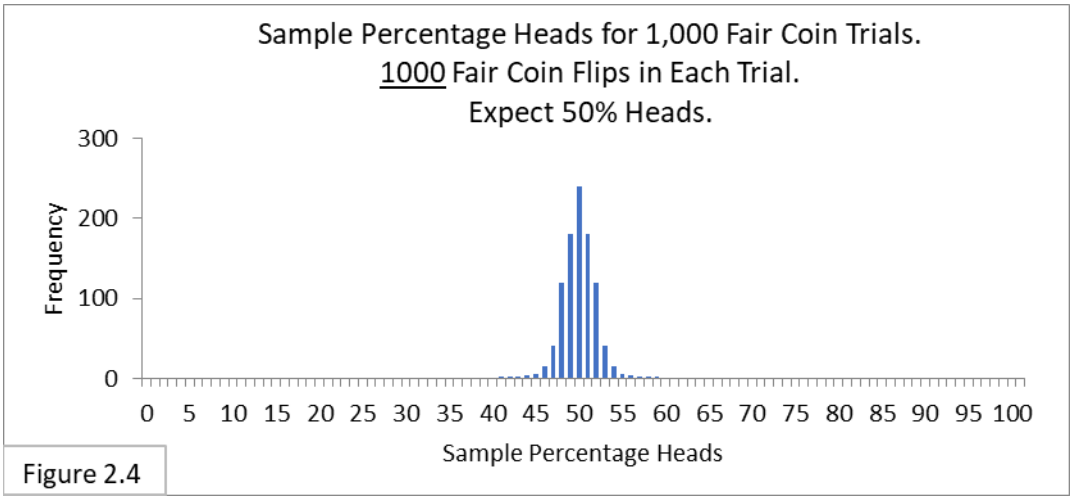


Figure 2.4

Notice that the sampling distribution in Figure 2.4 looks just like the sampling distribution in Figure 2.1 (reproduced below).

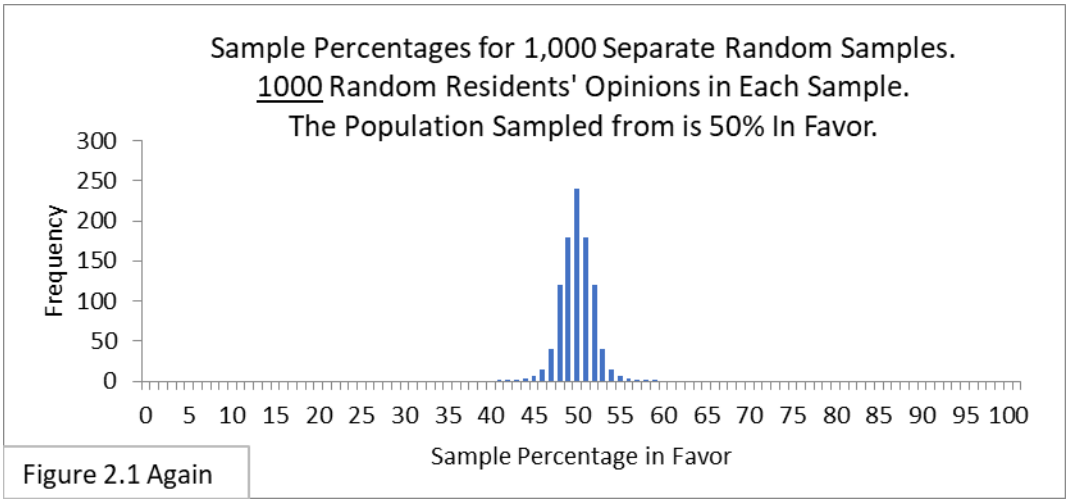


Figure 2.1 Again

Figure 2.5 shows the 95% interval of 47%-to-53% heads with 1000 flips per trial.

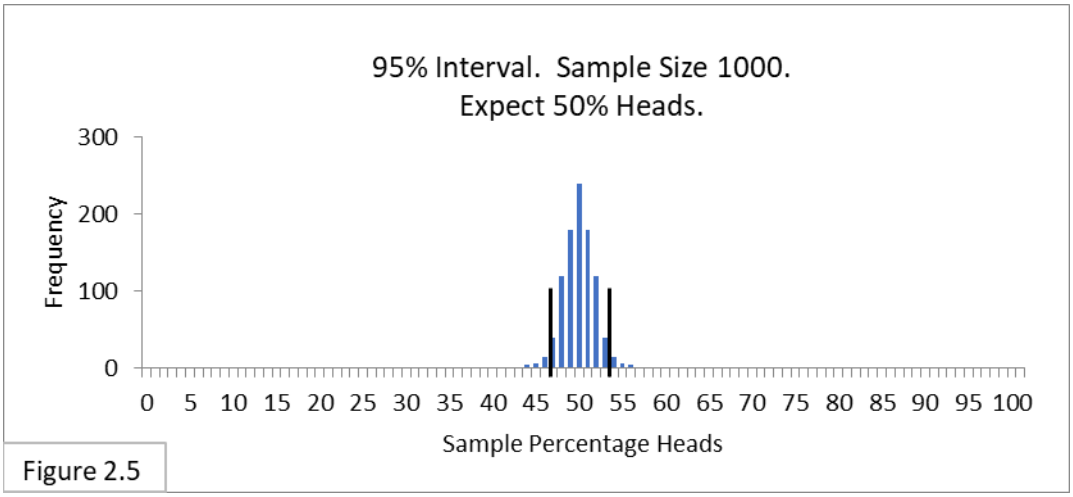


Figure 2.5

Notice that Figure 2.5 looks just like Figure 2.3 (reproduced below).

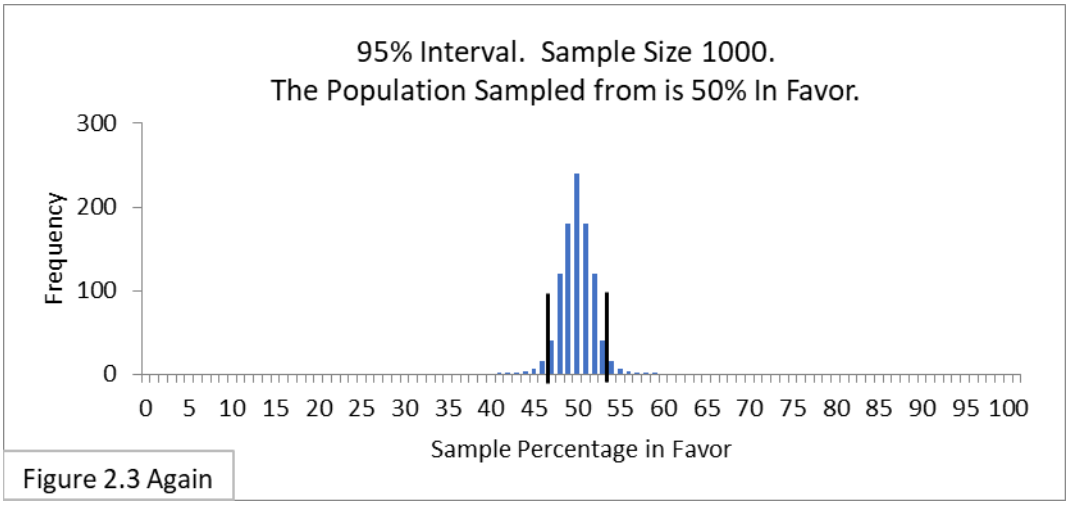


Figure 2.3 Again

If someone gave us a coin and asked us to judge whether it is a fair coin, we could flip it 1000 times, and, if we get fewer than 470 (47%) heads or more than 530 (53%) heads, we will say we think the coin is unfair. If we get anywhere between 47% and 53% heads, we will say we think the coin is fair. (Actually, as we'll delve into later, we'll be more circumspect and say that we can't rule out that the coin is fair rather than stick our neck out and say we think the coin is fair.)

Sampling Distributions of Other Populations

Let's suppose that the Flowing Wells population is only 30% in favor of the new public health policy. We'll experiment and randomly sample 100 residents, and again we'll amass 1,000 random samples. Figure 2.6 shows what to expect. The

sampling distribution, as expected, is centered on 30%. It is a "bell shaped" sampling distribution like we've been seeing all along (more on this later).

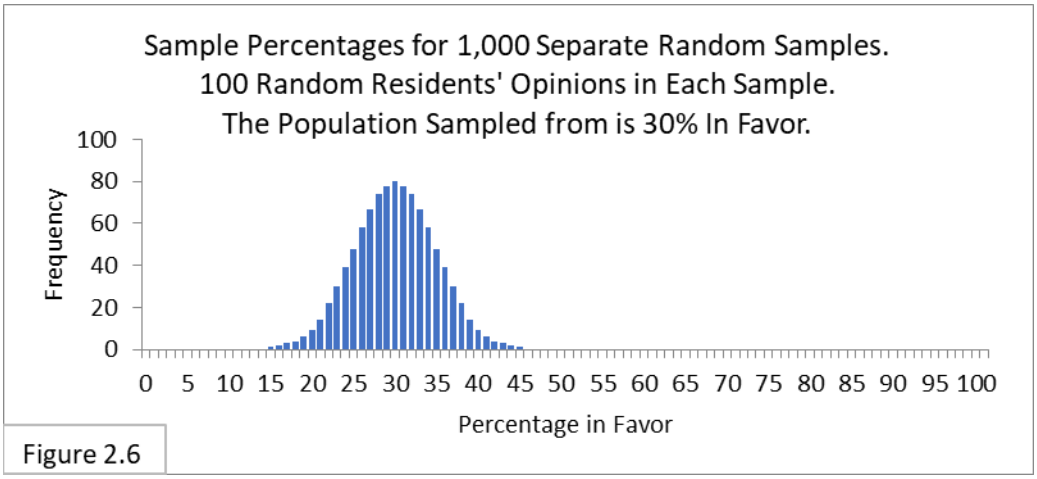


Figure 2.6

Next, let's look at the 95% confidence interval. That's shown in Figure 2.7.

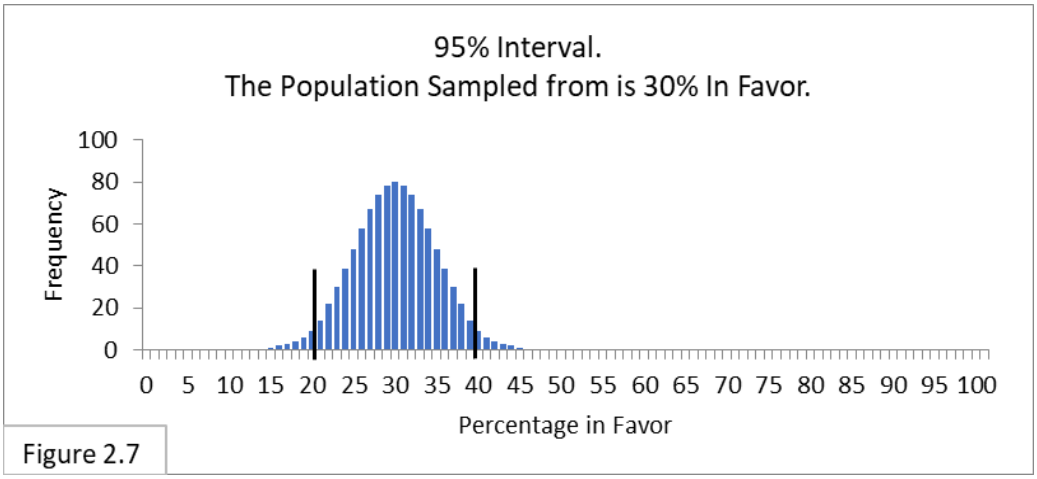


Figure 2.7

This 95% interval has boundary lines of 21% and 39%. When randomly sampling 100 from a population that is 30%-in-favor, we expect 95% of the sample percent-in-favor statistic values we get to be within this interval.

This 95% interval for a sample size of 100 and a population that is 30%-in-favor is 18 wide (39-21). Earlier (Figure 1.2, reproduced below) the 95% interval for a sample size of 100 and a population that is 50%-in-favor is 20 wide (60-40). Why is the interval in Figure 2.7 narrower than the interval in Figure 1.2?

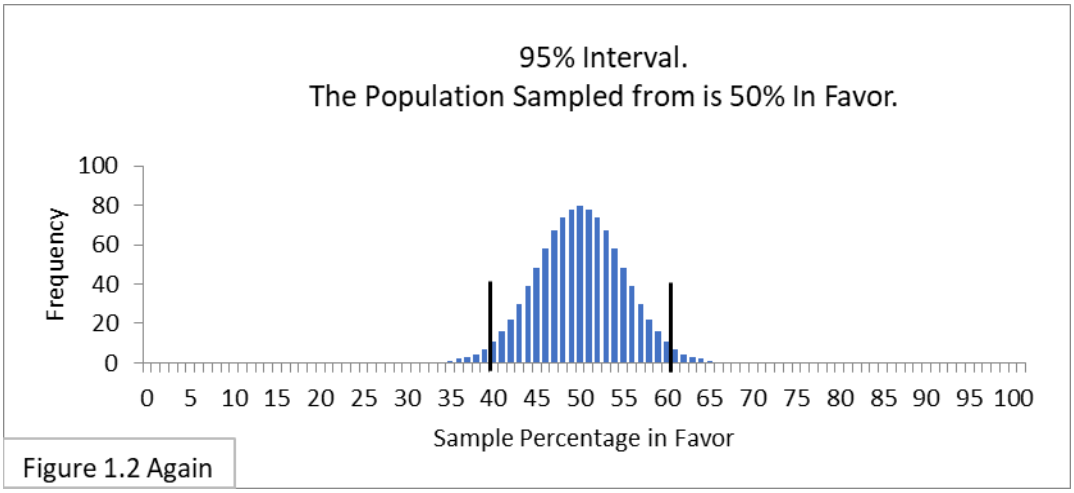


Figure 1.2 Again

It has to do with uncertainty. Imagine there are two contestants facing off in a judo match. Who's going to win? If they are evenly matched, each with a 50% chance of winning, you would be wholly uncertain who the winner will be. If the first contestant has a 70% chance of winning and the second has a 30% chance, you would be less uncertain who the winner will be. If the first contestant has a 100% chance of winning and the second has a 0% chance, you would not be uncertain at all. Likewise, if you were sampling from a population that is 0%-in-favor or 100%-in-favor, there is no uncertainty in what your sample percentage values will be—they'll all be 0% or 100%. Figure 2.8 shows the sampling distribution for a population that is 0%-in-favor. There's no uncertainty about sample percentage values when sampling from a population that is 0%-in-favor.

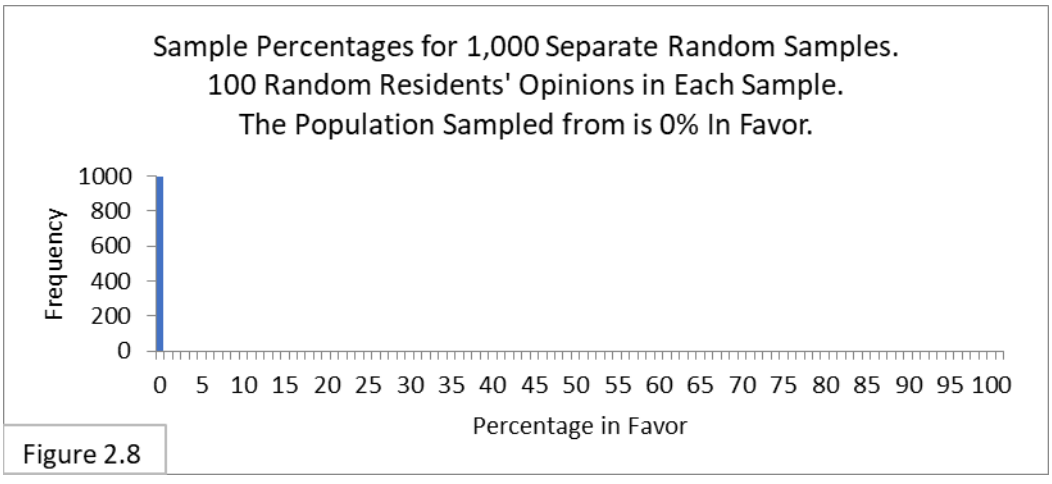


Figure 2.8

The width of the 95% interval reflects the level of uncertainty of our sample statistic values. Populations that are 50%-in-favor are the most uncertain. As the population percentage moves toward 100% or 0%, uncertainty decreases. As

uncertainty decreases, the width of the sampling distribution and its 95% interval decreases. Next, we'll see a simple formula that reflects the level of uncertainty due to the population percentage *and* sample size and their effects on the width of the sampling distribution.

3. Calculating Confidence Intervals

Time for Some Standardization

With random sampling of binomial values (in-favor vs. not-in-favor; heads vs. tails) we've seen that:

1) Sampling from populations with percent-in-favor close to 50% have wider sampling distributions than populations with percentages closer to 0% or 100%.

2) Larger sample sizes have narrower sampling distributions.

The various sampling distributions we've seen have different locations on the horizontal axis and they have different widths. It would be useful to convert them all to one standard scale. We'll need a common unit. And the rescaling to that unit must account for the effects of the population percent-in-favor value (number 1 above) and sample size (number 2 above).

The unit to be used is called *Standard Error.* It's labeled "Standard" because it serves as a standard unit. And it's labelled "Error" because we don't expect our sample statistic values to be exactly equal to the population statistic value; there will be some amount of error. The Standard Error formula, which I'll explain a piece at a time, is as follows: The square root of p times $(1\text{-}p)$ divided by n.

$$\sqrt{\frac{p * (1-p)}{n}}$$

The variable p is the *proportion*[iii] rather than percentage: .5 rather than 50% (and 0 rather than 0%; .01 rather than 1%; .1 rather than 10%; and 1 rather than 100%). The $p*(1\text{-}p)$ term in the numerator is called the proportion *variance*. Recall from the previous chapter that sampling from populations with percent-in-favor close to 50% have wider sampling distributions than populations with percentages closer to 0% or 100%.

The variance p*(1-p) reflects this dynamic:

```
0.0*(1-0) =   0.00
 .01*(1-.01) = .01
 .1*(1-.1) =   .09
 .3*(1-.3) =   .21
 .5*(1-.5) =   .25
 .7*(1-.7) =   .21
 .9*(1-.9) =   .09
 .99*(1-.99) = .01
1.0*(1-1) =   0.00
```

So, as p moves from .5 towards 0 or 1, variance decreases, and since variance is in the numerator, Standard Error decreases. Decreases in Standard Error correspond to narrowing of the sampling distribution. This reflects lower uncertainty. *Lower variance, lower uncertainty*.

Variance is itself a statistic and is very important in statistical analysis. We'll be seeing it in formulas from now on.

Now let's consider sample size, which is represented in the denominator of the formula by n. Recall from the previous chapter that larger sample sizes have narrower sampling distributions. Since n in the denominator of the Standard Error formula, as n increases Standard Error decreases. Again, decreases in Standard Error correspond to narrowing of the sampling distribution. Again, this reflects lower uncertainty. *Larger sample size, lower uncertainty*.

Now we can use the Standard Error scale to determine 95% intervals. First, an important fact: *The boundary lines of the 95% interval on the Standard Error scale are always -2 and +2* (they're actually -1.95996… and +1.95996…, but I'm rounding to -2 and +2 for the present purposes). Let's clarify all this by looking at several example calculations and illustrations.

Let's start with random sampling of 100 from a population that is 50% in favor of the new public health policy (Figure 1.2, reproduced below). Plugging in the numbers gives

$$\sqrt{\frac{.5 * (1 - .5)}{100}} = .05$$

Standard Error is .05 and two Standard Errors is .1 in proportions and 10% in percentages. Since we want to center the interval on the percentage p of 50%, we'll add and subtract 10% from 50%. This yields a calculated 95% interval of 50%±10% (50% minus 10% to 50% plus 10%) or 40%-to-60%. That's also what Figure 1.2 shows!

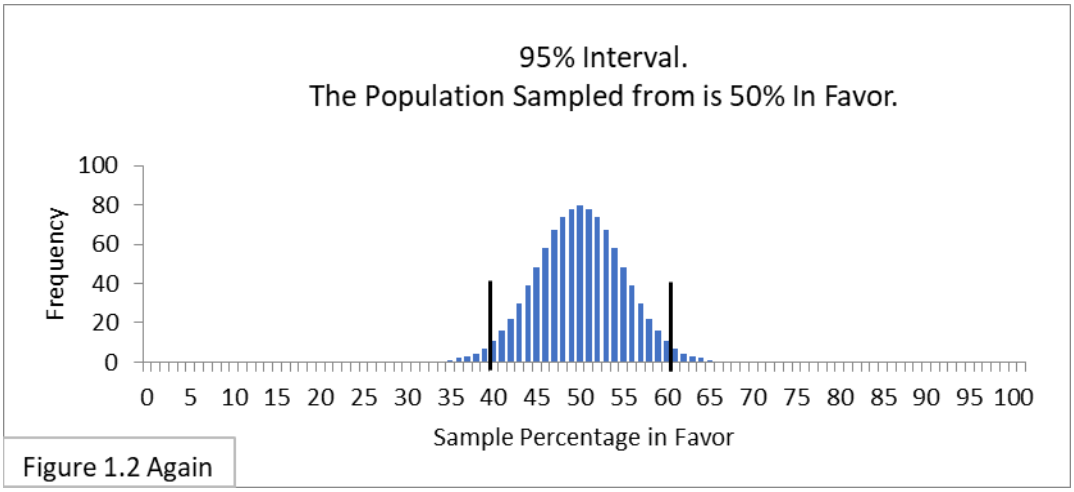


Figure 1.2 Again

Putting everything we just computed into a formula for calculating 95% intervals we get

$$p \pm 2 * \sqrt{\frac{p * (1-p)}{n}}$$

Next let's consider the 95% interval of random sampling of 100 from a population that is *30*% in favor of the new public health policy (Figure 2.7, reproduced below).

$$.3 \pm 2 * \sqrt{\frac{.3 * (1-.3)}{100}} = .3 \pm 2 * .045 = 30\% \pm 9\%$$

Standard Error is .045 and two Standard Errors is .09 in proportions and 9% in percentages. We want to center the interval on 30%, so we'll add and subtract 9% from 30%. This yields a 95% interval of 30%±9% (30% minus 9% to 30% plus 9%) or 21%-to-39%. That's also what Figure 2.7 shows!

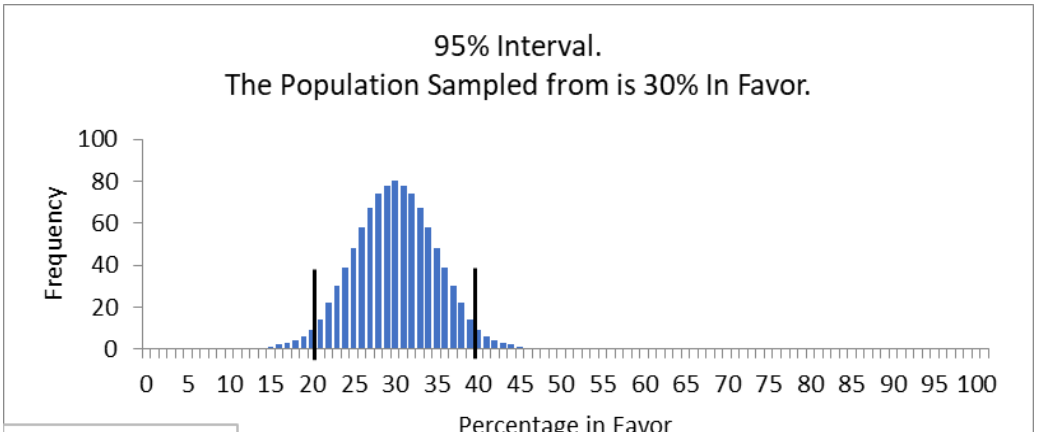


Figure 2.7 Again

Last let's consider the 95% interval of random sampling of *1000* from a population that is 50% in favor of the new public health policy (Figure 2.3, reproduced below).

$$.5 \pm 2 * \sqrt{\frac{.5 * (1 - .5)}{1000}} = .5 \pm 2 * .015 = 50\% \pm 3\%$$

Standard Error is .015 and two Standard Errors is .03 in proportions and 3% in percentages. We want to center the interval on 50%, so we'll add and subtract 3% from 50%. This yields a 95% interval of 50%±3% or 47%-to-53%. That's also what Figure 2.3 shows!

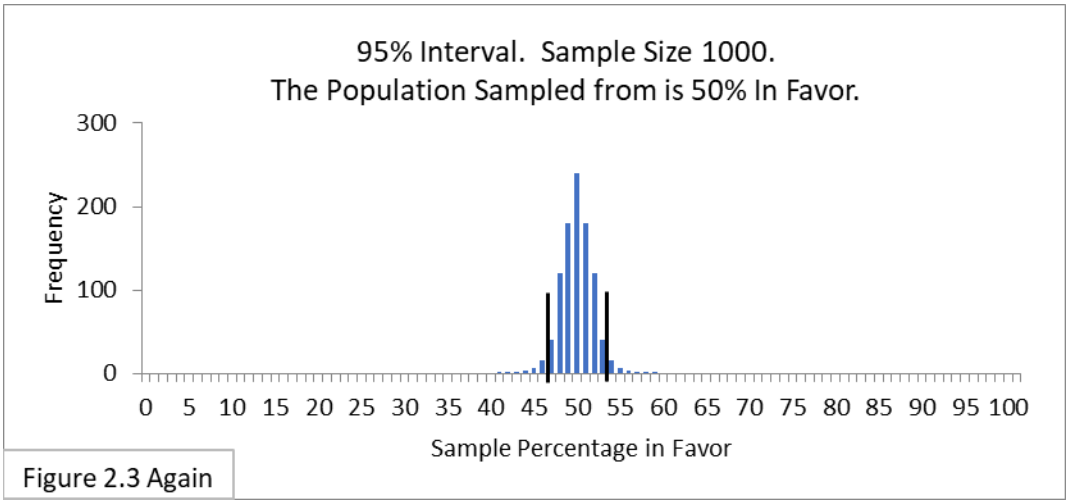


Figure 2.3 Again

The formula works!

The reason the formula works is because the sampling distributions are "bell shaped". More than that, they approximate the very special bell shape called the *Normal* distribution.[iv]

Let's go one step further and standardize an entire sampling distribution to get what's called *the Standard* Normal distribution. The Standard Normal Distribution is a normal distribution that uses Standard Error as its unit (rather than percentages or proportions). To illustrate, let's standardize Figure 1.1 (reproduced below).

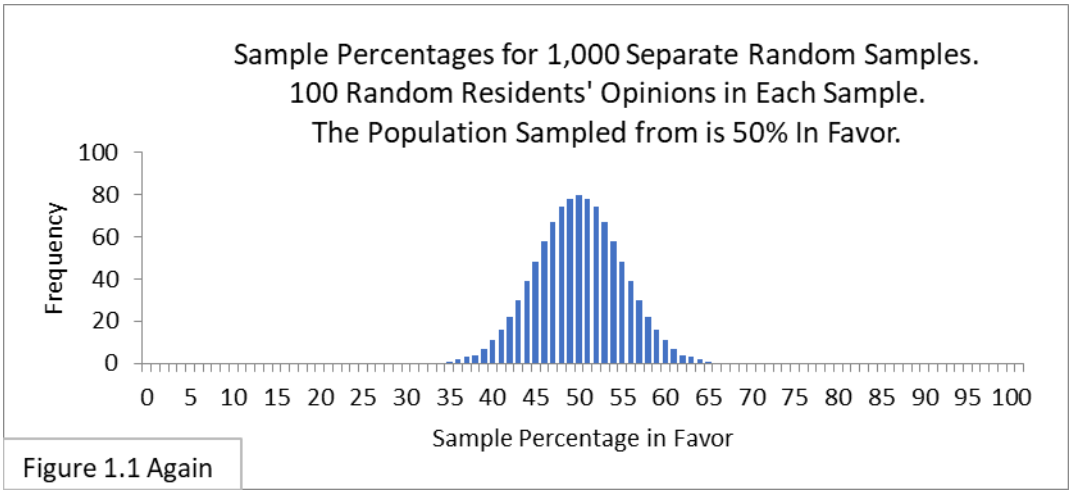


Figure 1.1 Again

Figure 3.1 is a standardized version of Figure 1.1.

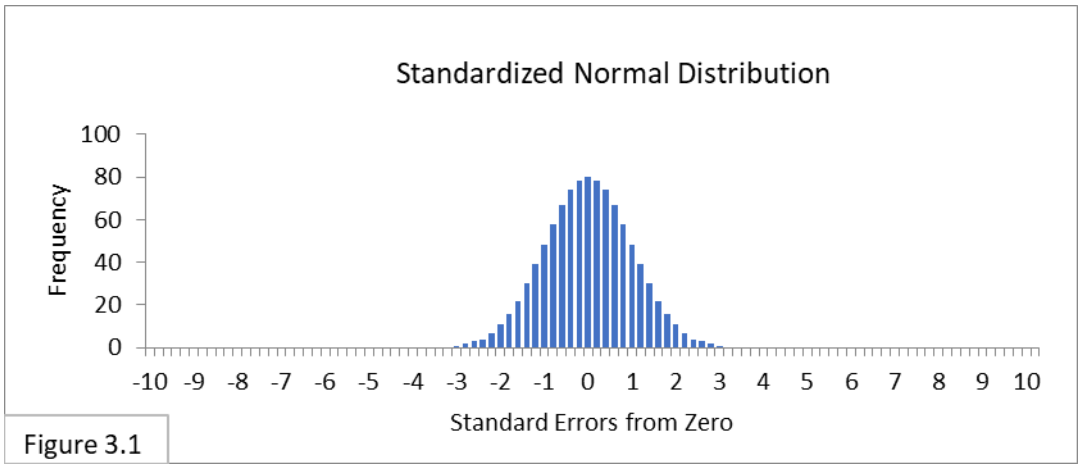


Figure 3.1

Notice that Standard Error is the unit used on the horizontal axis of Figure 3.1. This is done by rescaling the horizontal axis unit of Figure 1.1 to the Standard Error unit of Figure 3.1 using the below formula.

$$\frac{p - .5}{\sqrt{\frac{p * (1 - p)}{n}}}$$

This formula gives us how many Standard Errors a proportion, p, is from .5. First, we convert the percentages to proportions. Next, we recenter the axis: whereas Figure 1.1 is centered on the proportion value .5 (50%), Figure 3.1 is centered on zero Standard Errors; the numerator p-.5 centers the horizonal axis of Figure 3.1 onto zero. Finally, these differences are divided by the Standard Error to rescale the horizontal axis. Viola, Figure 1.1 has been standardized to the Standard Error scale of Figure 3.1.

Figure 3.2 shows its 95% interval below Figure 1.2 (reproduced below). Recall that the boundary lines of the 95% interval on the Standard Error scale are -2 and 2 (rounded). Plugging .4 (40%) and .6 (60%) from Figure 1.2 into the above formula gives us -2 and 2 Standard Errors as the 95% boundary lines in the Standard Error unit. As emphasized above: *The boundary lines of the 95% interval on the Standard Error scale are always -2 and +2 (rounded)*. If we standardized Figures 2.3 and 2.7, we'll again find the 95% interval boundary lines to be -2 and 2. (You can use the formula and do the arithmetic if you want to confirm this.)

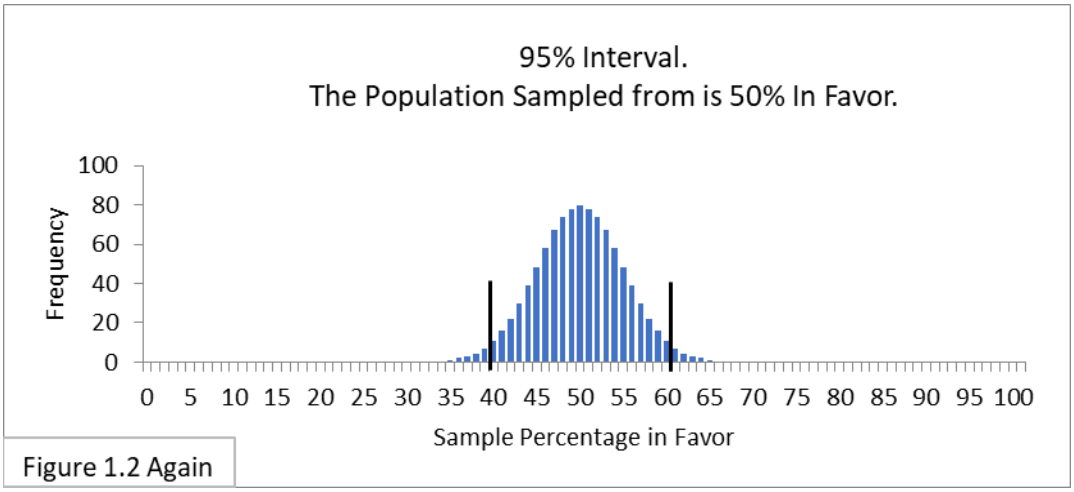


Figure 1.2 Again

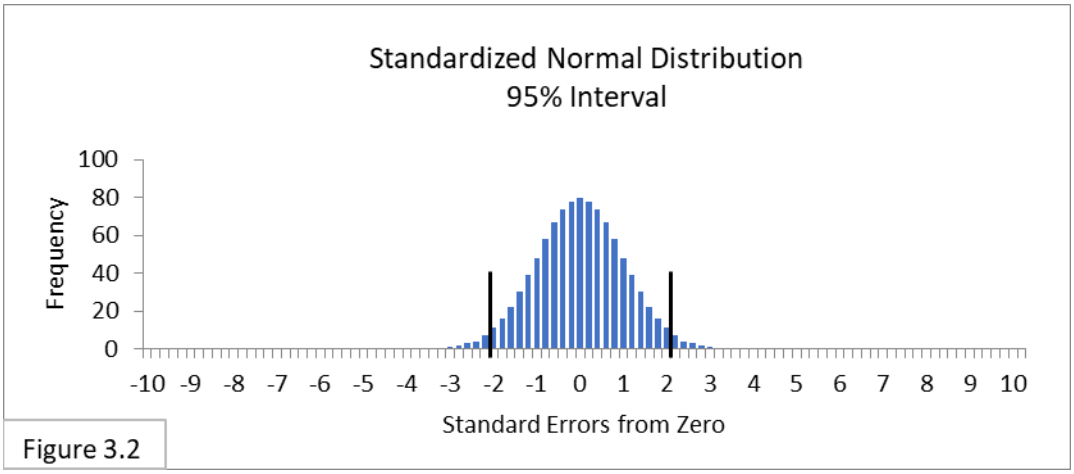


Figure 3.2

We can convert our units (e.g., percent-in-favor, percent-heads) into the Standard Error unit and vice versa by multiplying and dividing by Standard Error. That comes in very handy. All of the sampling distributions we've looked at so far can be standardized in this way. In practice, we don't convert entire sampling distributions to the standardized distribution; we use Standard Error in formulas as multipliers and divisors to calculate individual values, like we do to calculate the boundary lines for 95% intervals and to convert proportions to the Standard Error scale.

$p \pm 2 * \sqrt{\frac{p * (1-p)}{n}}$ used to calculate 95% intervals for proportions.

$\frac{p - .5}{\sqrt{\frac{p * (1-p)}{n}}}$ used to convert proportions to the Standard Error scale.

We'll further explore the Standard Normal distribution later on, but first let's put some of what we've covered so far into action, while also expanding our horizons.

1,000 Surveyors' Sample Statistics and Their 1,000 95% Confidence Intervals

In this section we're going to look at things from a different perspective. Surveyors won't be comparing their *sample* statistics of public opinion with what to expect when the *population* opinion statistic equals a particular value, like 50%. Instead, the surveyors want to determine, based on their sample statistic, what the value of the population statistic might be. For example, a surveyor who gets a sample statistic value of 34% will want to calculate a 95% interval surrounding 34% and explain what that interval might tell us about the overall population's opinions.

We are going to explore the subtleties involved by sending out 1,000 surveyors to survey *the same population* and see what they come up with and how they should interpret what they come up with. But first we'll need to set the stage by inventing a population that has certain characteristics that we know, but none of the surveyors know.

Our invented community, Artesian Wells, has about 70,000 residents. There is a new public health policy being debated and, since we are all-knowing, we know that 40% of the residents agree with the new policy. Only we know this. We want to know what to expect when many, many surveyors randomly sample 1000 people from this population. The survey respondents will be asked whether they "agree" or "disagree" with something, a binomial response. We'll use proportions rather than percentages, with the proportions rounded to two decimal places. Figure 3.3 shows us the sampling distribution of what to expect. (Don't get confused: There are 1,000 random samples, and each sample has a sample size of 1000.)

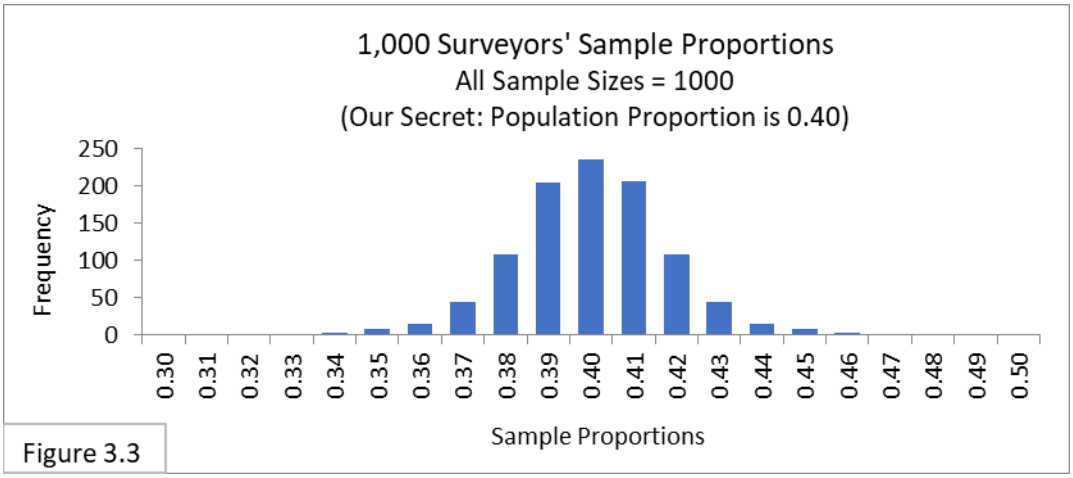


Figure 3.3

Based on visual inspection, notice that the great majority of the sample proportions are in the interval 0.37 to 0.43. Approximately 950 of the 1,000 sample proportions are contained within the interval 0.37 to 0.43, indicating that 0.37 to 0.43 is the 95% interval surrounding the population proportion of 0.40. The formula will give us the same boundary lines. (Feel free to double check.)

As always, we expect the 95% interval around the population proportion to contain 95% of all sample proportions obtained by random sampling.

Now, we hire 1,000 independent surveyors who converge on the town to do the "agree" or "disagree" survey. All 1,000 surveyors get their own random sample of

1000 residents and calculate their own sample proportion-agree statistic. How does each of the individual surveyors analyze their sample proportion?

First, let's look at the formula for calculating 95% confidence intervals for *sample* proportions. It looks much like the formula in the previous section. The variable $\hat{p}$ with a hat on denotes the sample proportion (as opposed to the population proportion). The square root term calculates the Standard Error for the sample proportion. Sample size is again represented by n. As for the constant 1.96, recall that earlier I rounded $\pm$1.95996... Standard Errors and used $\pm$2 Standard Errors; now I'm being more precise by using $\pm$1.96 Standard Errors, which is more common.

$$\hat{p} \pm 1.96 * \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

Each of the 1,000 surveyors calculates their individual interval using their sample proportion value, and *we expect that 95% of the surveyors' 95% confidence intervals will contain the population proportion* (0.4 in this example). You might want to reread that sentence a few times, keeping in mind that although we, as know-it-alls, know that the population proportion is 0.4, none of the surveyors have any idea what it is.

It's in this context that the term "confidence" in "confidence interval" came about: we are confident that 95% of all 95% confidence intervals for *sample* statistic values obtained via random sampling will contain the population statistic value. (But no individual surveyor will know whether their confidence interval contains the population statistic value or not!)

In a nutshell:
The 1,000 surveyors calculate their individual 95% confidence intervals.
About 950 of them will have an interval containing the population proportion.
About 50 of them won't.

Let's look at the 95% confidence intervals constructed via the formula by two surveyors: The first got a sample proportion of 0.38, and the second got a sample proportion of 0.34.

Surveyor #1 Result.
Using the formula with a sample proportion of 0.38 and a sample size of 1000, the 95% confidence interval is 0.35 to 0.41 (rounded).

Only we, being know-it-alls, know that this 95% confidence interval contains the population proportion of 0.4

Surveyor #2 Result.
Using the formula with a sample proportion of 0.34 and a sample size of 1000, the 95% confidence interval is 0.31 to 0.37.

Only we, being know-it-alls, know that this 95% confidence interval does *not* contain the population proportion of 0.4. Only we know that this surveyor is one of the 5% of unlucky surveyors who just happened to get a misleading random sample. This is called a Type I Error, which is covered in depth in the next chapter.

In summary,
1) We expect the 95% confidence interval around the population proportion to contain 95% of all sample proportions obtained by random sampling. We've been seeing this all along.
2) We expect 95% of all the 95% confidence intervals based on random sample proportions to contain the population proportion. We see this for the first time here; more detail is given next.

Table 3.1 is divided into three sections, left to right, and shows what the various surveyors will get. Overall, the Table shows the confidence intervals for surveyors with sample proportions of 0.3 through 0.5; sample proportions 0.30 through 0.36 are in the left section, 0.37 through 0.43 are in the middle section (shaded), and 0.44 through 0.50 are in the right section. Notice that the expected 950 surveyors in the middle section (shaded) with sample proportions within the interval of 0.37 to 0.43 also have 95% confidence intervals that contain the population proportion of 0.40. The expected 50 surveyors with sample proportions outside the interval of 0.37 to 0.43—the left and right sections of the Table—do not have 95% confidence intervals that contain the population proportion of 0.40.

Table 3.1

Surveyor's Sample Proportion	95% Interval Low	95% Interval High	Surveyor's Sample Proportion	95% Interval Low	95% Interval High	Surveyor's Sample Proportion	95% Interval Low	95% Interval High
0.30	0.27	0.33	0.37	0.34	0.40	0.44	0.41	0.47
0.31	0.28	0.34	0.38	0.35	0.41	0.45	0.42	0.48
0.32	0.29	0.35	0.39	0.36	0.42	0.46	0.43	0.49
0.33	0.30	0.36	0.40	0.37	0.43	0.47	0.44	0.50
0.34	0.31	0.37	0.41	0.38	0.44	0.48	0.45	0.51
0.35	0.32	0.38	0.42	0.39	0.45	0.49	0.46	0.52
0.36	0.33	0.39	0.43	0.40	0.46	0.50	0.47	0.53

Again, in summary, and for emphasis:
1) We expect the 95% confidence interval around the population proportion to contain 95% of all sample proportions obtained by random sampling.
2) We expect 95% of all the 95% confidence intervals based on random sample proportions to contain the population proportion.

Because of these two facts, we will reach the same conclusion whether we (1) check if a sample proportion is outside the 95% interval surrounding a hypothesized population proportion, or (2) check if the hypothesized population proportion is outside the 95% interval surrounding a sample proportion. The analysis can be done either way.

Here's a quick analogy: Suppose a stamping plant that makes coins was malfunctioning and produced unbalanced (i.e., unfair) coins. Unbeknownst to anyone, these unfair coins favored tails, and the chance of coming up heads is only 0.4. Now say 1,000 people flip these coins 1000 times each, while counting and then determining the proportion of heads. What would the 1,000 peoples' results be like? What would an analysis of a single coin and its 1000 flips be like? Answer: Just like the survey example above. Just replace the words "agree" and "disagree" with "heads" and "tails".

We expect 95% of the coin-flippers will get 95% confidence intervals that contain 0.4, and 5% of the coin-flippers will get 95% confidence interval that do *not* contain 0.4. In other words, we expect 95% of the results to be veridical and 5% of the results to be misleading. But no one knows whether their results are veridical or misleading.

The reason I use the word “veridical” is because it’s the perfect word: “Coinciding with reality.” I’m using it to mean the opposite of misleading.

4. Veridical vs. Misleading Results

In this chapter we'll start off using a sample size of 100 and its .4-to-.6 boundary lines to make a 95% confidence interval for testing coins. Any coin whose proportion of heads lies outside the interval we'll declare unfair. Only 5% of the time will a fair coin mislead us and lie outside the interval, leading us to erroneously declare it unfair. This is called *Type I Error*.

What about *unfair* coins that mislead us and lie *inside* the interval? That will lead us to erroneously declare them fair. This is called *Type II Error*. Let's explore these two types of potential errors.

Type I and II Error

Imagine that I present you with a basket full of coins. The basket has an unknown number of fair coins and an unknown number of unfair coins. Your task is to test two arbitrary coins by flipping each one 100 times. You're going to use the 95% interval to make your judgment: If the number of heads is within the .4-to-.6 interval then you'll judge the coin to be fair, and if the number of heads is outside the .4-to-.6 interval then you'll judge the coin to be unfair.

Figure 4.1 highlights the four things that can happen with fair/unfair coins that are within/outside the 95% interval for the hypothesis that the coin is fair. Table 4.1 presents the information in a tabular format. (Some readers prefer the Figure and others prefer the Table, so I'm including both.)

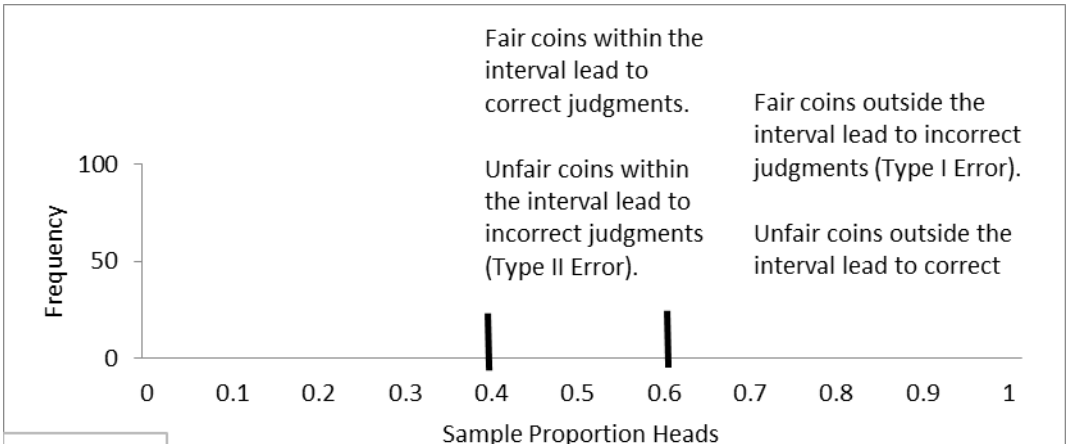


Figure 4.1

Table 4.1 Veridical and Misleading Results

Statistical Result / Unknown Truth	Result is Within the Interval (Judge the Coin to be Fair)	Result is Outside the Interval (Judge the Coin to be Unfair)
Coin is Fair	Veridical Result Correct Judgment	Misleading Result Type I Error
Coin is Unfair	Misleading Result Type II Error	Veridical Result Correct Judgement

Let's say you select a coin, flip it 100 times, and get a result of 55 heads. That's inside the interval, so you judge the coin to be fair. If in fact the coin is fair, then the result is veridical and leads you to make a correct judgment. If in fact the coin is unfair, then the result is misleading and leads you to make a Type II Error.

Now let's say you select another coin, flip it 100 times, and get a result of 65 heads. That's outside the interval, so you judge the coin to be unfair. If in fact the coin is unfair, then the result is veridical and leads you to make a correct judgment. If in fact the coin is fair, then the result is misleading and leads you to make a Type I Error.

At first, many people think (hope) that using a 95% confidence interval means that there is a 95% chance they're correct and a 5% chance they're incorrect. *But, unfortunately, that's not what it means*. Its meaning is much more limited. *Remember that our sampling distribution and its confidence interval portrays a very specific situation. In this case, it portrays the situation of flipping only fair coins. It does not portray flipping unfair coins.*

When we are in fact flipping fair coins, we do in fact expect that 95% of the outcomes will be within the .4-to-.6 boundary lines and 5% will be outside. *But* when we are in fact flipping unfair coins, this sampling distribution doesn't tell us what to expect.

What if *all* the coins in the basket are *fair*? Then you expect to be correct 95% of the time and to fall victim to Type I Error 5% of the time. Type II Error is irrelevant because it only applies to unfair coins.

What if *all* the coins in the basket are *unfair*? Then Type I Error is irrelevant because it only applies to fair coins. And, we have no idea, at this point, how many times we

should expect to be correct or how many times we should expect to fall victim to Type II Error. Our 95% interval has nothing definitive to say about unfair coins.

If we knew, for example, (1) that ½ of the coins in the basket are fair and ½ are unfair, and (2) that the unfair coins are all identical and favor heads ¾ versus tails ¼, then we could do some calculations to determine how likely it is that your judgments are correct. But we don't know those things.

We can make use of *estimates* of those things. In Frequentist statistics, we can, if we wish, estimate values for #2 in order to then estimate the likelihood of Type II Error. We'll explore that next. But we do *not* make nor use estimates for #1. (Estimates for #1 are used to determine what's called the *false discovery rate*, which is covered in an Addendum.)

Exploring Type II Error

Figure 4.2 shows the expected results of testing an unfair coin that comes up heads 30% (.3) of the time and tails 70% (.7) of the time. The person testing the coin has no knowledge of this. This person is testing whether the coin is fair and so is using a 95% interval for a fair coin that comes up heads 50% (.5) of the time and tails 50% (.5) of the time. Notice that the great majority of trials for the unfair coin will be outside the 95% interval for a fair coin and will lead to the correct judgment that the coin is unfair. However, about 2% of the time the unfair coin will be within the interval and be incorrectly judged to be fair. That is Type II Error.

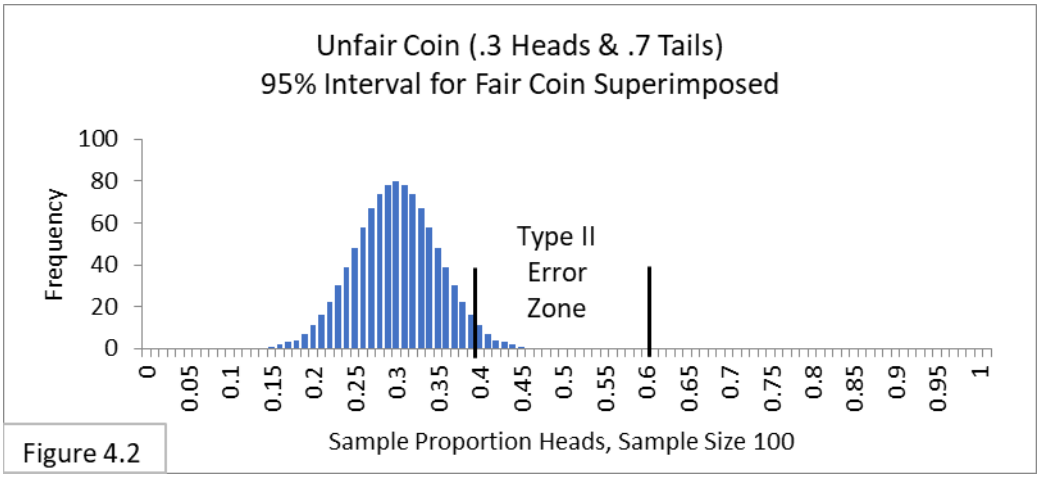


Figure 4.2

Using your imagination, you can envision that if the unfair coin is closer to fair—say it comes up heads 40% (.4) of the time—then the bell shape will be farther to the

right, so more of it will be between the boundary lines of .4-to-.6, and so there will be more frequent Type II Errors. Figure 4.3 illustrates what you might have imagined. The closer the unfair coin is to being fair, then the more Type II Errors we can expect. In Figure 4.3 about 50% of the trials are Type II Errors!

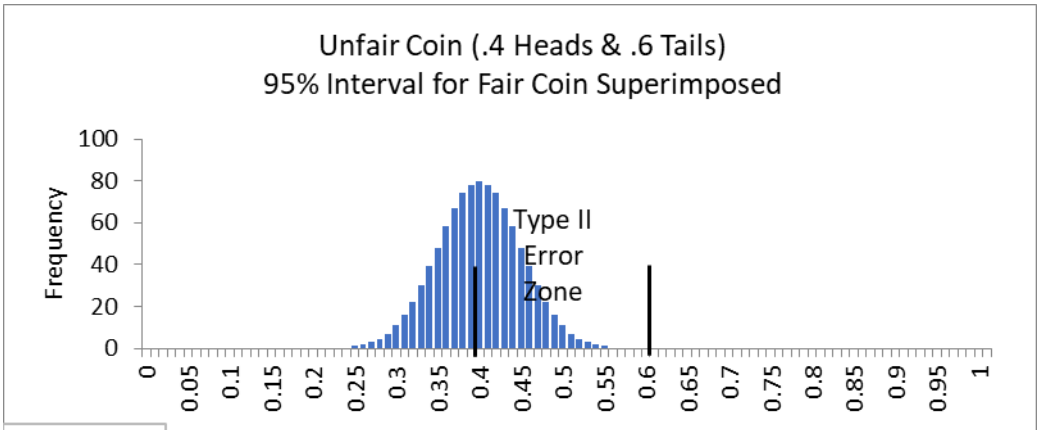


Figure 4.3

What if we increase our sample size? Recall that the sampling distribution and the 95% interval get narrower when we increase sample size (see Chapter 2 Section 1 *Sampling Distributions and Sample Sizes*). Using your imagination, narrow the bell shape and the 95% interval of Figure 4.2. That should lead to fewer Type II Errors, shouldn't it? Yes. Figure 4.4 illustrates what you might have imagined. With a larger sample size of 1000 (rather than 100) Type II Error seems extremely unlikely.

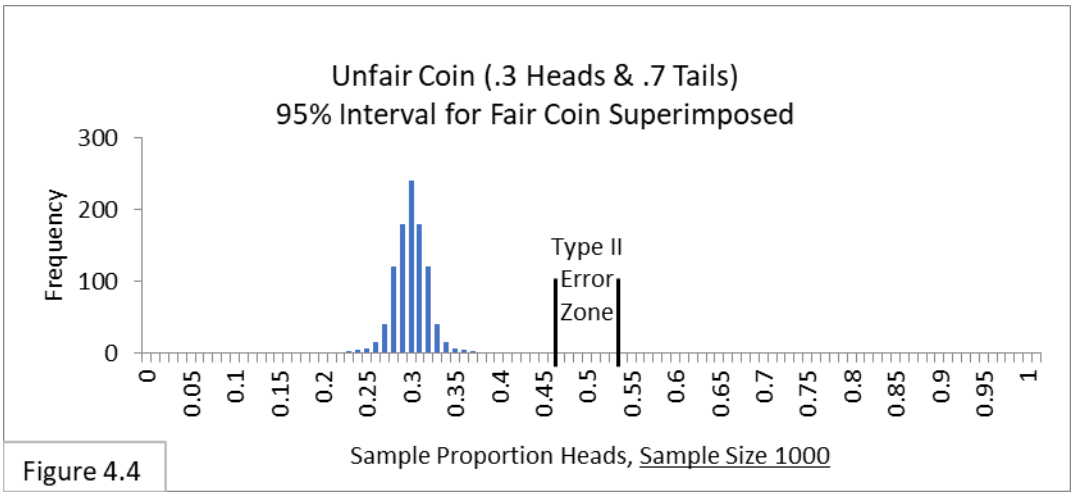


Figure 4.4

In summary, *the closer something is to what we're testing for, and the smaller the sample size, the more likely we are to suffer Type II Error*. There is another avenue to suffering higher Type II Error rates: making our Type I Error criterion stricter.

99% Intervals and Their Effects on Errors

While 95% intervals are the most commonly used (except in sciences such as particle physics), some researchers argue for stricter lines to be drawn: in particular, they argue for use of 99% intervals to better avoid Type I Errors (particle physicists use even stricter lines). Figure 4.5 shows the 99% interval for a fair coin and a sample size of 100.

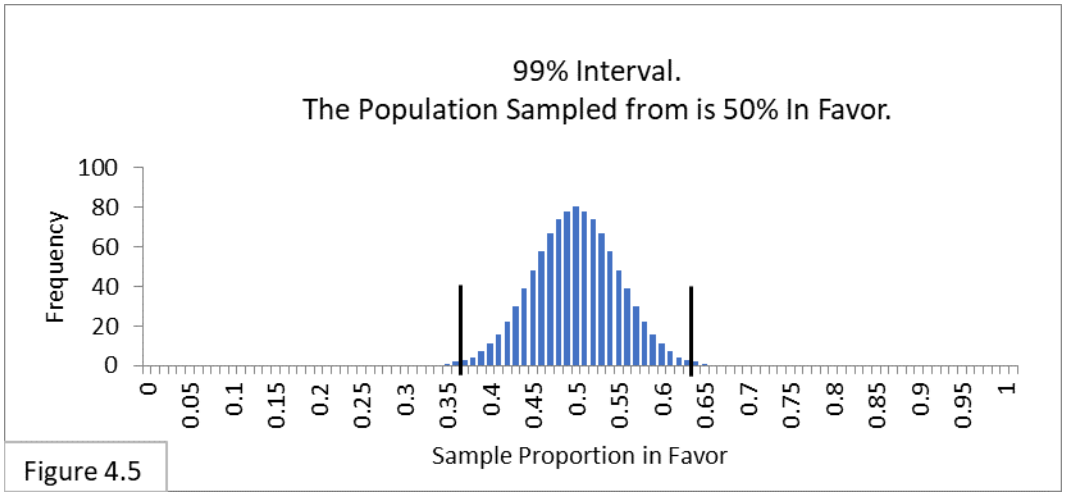


Figure 4.5

With a 99% interval, 99% of the expected sample proportion values with fair coins are contained within the interval. Since it contains more of the expected outcomes than the 95% interval, it is wider. In this case the boundary lines are .37-to-.63 (wider than the .40-to-.60 we have with our 95% interval). Now only 1% of the expected values are outside the boundary lines, ½% on each side.

With a wider interval, our sample statistic value needs to be more extreme in order to fall outside the interval. In that way, it is a stricter criterion with respect to Type I Error. We can use the stricter criterion to lessen the chances of Type I Error. But what happens with Type II Error? As you might expect, since you can't get something for nothing, Type II Error will become *more* likely. First, look at Figure 4.6 shown below.

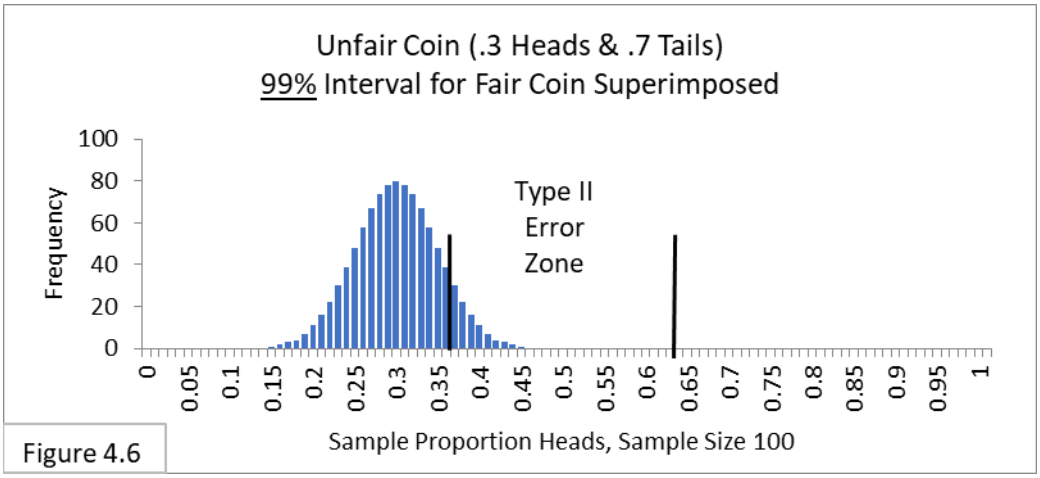


Figure 4.6

As you can see when comparing Figure 4.6 with Figure 4.2 (reproduced below), more outcomes of the same unfair coin are now within the 99% interval than were within the 95% interval.

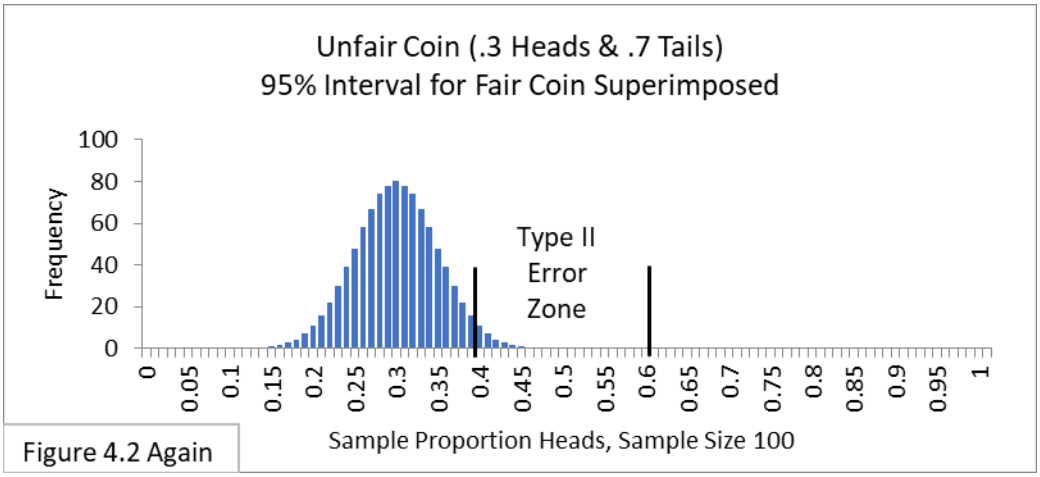


Figure 4.2 Again

This is not a surprise since the 99% interval is wider. Whereas about 10% of the unfair coins result in Type II Error with the 99% interval, only about 2% of the unfair coins result in Type II Error with the 95% interval. In short: When we make our Type I Error criterion stricter, we increase the likelihood of Type II Error.

In practice, we determine the acceptable likelihood of Type I Error by the confidence interval we choose to use—95% and 99% intervals are the most common in the social sciences, whereas 99.9999% is used in particle physics.

As we've just seen, the likelihood of Type II Error is determined by a number of factors, some within our control and some not. In summary, the likelihood of Type II Errors increases:

1) the closer something is to what we're testing for (.4 is closer to .5 than .3 is, and so .4 coins make Type II Errors more likely than .3 coins). We don't control this.

2) when we use smaller sample sizes (a sample size of 100 makes Type II Errors more likely than a sample size of 1000). We do control this, although larger samples cost more to gather.

3) when we use stricter Type I Error criteria (a 99% interval is stricter for Type I Error and makes Type II Error more likely than a 95% interval—but keep in mind that Type I Error becomes less likely; it's a trade-off). We do control this, but in practice the strictness of our Type I Error criteria is usually set by convention (social sciences use 95% and sometimes 99%; physical sciences use 95% through 99.9999% depending on the specific field).

Decreasing the likelihood of Type II Error increases the statistical *power* of our analysis (see Chapter 2 Section 1 *Sampling Distributions and Sample Sizes*, where we first met statistical power). In practice, increasing the sample size is a common way to reduce the likelihood of Type II Error. Moreover, there are formulas and software tools (including online calculators) available that help researchers estimate what their sample size should be in order to achieve their desired levels of Type I and Type II Error. With these tools you enter your estimated value for the population proportion, your expected sample proportion value, and your desired Type I and II Error rates. Keep in mind that, since you don't actually know what the population proportion is nor what your sample proportion will be, the sample size recommendations of these tools are "educated-guess" estimates.

In practice, Type I Error is feared more than Type II Error. In the social sciences the Type I Error limit is most often set at 5% via a 95% confidence interval. With surveys, for example, we want to limit how often we infer that a majority of the population is in favor of a new policy when in fact a majority is not in favor. What about Type II Error? That occurs if we infer that a majority of the population is not in favor of the new policy when in fact a majority is in favor. There is no well-established convention, but the most common guidance is to try and limit Type II Error to 20% (by having large enough sample sizes). These guidelines of 5% and

20% imply that we prefer to error on the conservative side, maintaining the status quo. Using our surveying context, this means that statistical survey evidence is more likely to erroneously undermine new policies than to erroneously provide support for them.[v]

While 95% and 99% intervals are common, much stricter Type I Error criteria are used in particle physics: 99.9999% (approximately). Using this interval with coin flipping, we would insist that a coin come up heads at least 75 times in 100 flips before declaring it unfair. Biologists conducting what are called genome-wide association studies use a similarly strict interval. Since these researchers test hundreds of thousands or even millions of separate genome locations, Type I Errors would be expected to occur far too frequently if they used a 95% or 99% interval. After all, 5% of 1,000,000 is 50,000 and 1% of 1,000,000 is 10,000.

5. A Series of Six Short Case Studies

In order to help reinforce the concepts introduced so far, the following six short case studies explore Type I Error and Type II Error under various circumstances. The odd numbered cases concentrate on Type I Error. These cases illustrate that the expected frequency of Type I Error does not change across the various circumstances. The even numbered cases concentrate on Type II Error and illustrate that the expected frequency of Type II Error does change across the various circumstances. Keep in mind that for all these cases we, being know-it-alls, know what the actual population percentages are, but the surveyors do not!

Case 1

Referring to Figure 5.1, suppose a population is 50% in favor of a new public health policy, and 1,000 surveyors survey the population using random sampling of sample size 100. All 1,000 surveyors hypothesize that the population is 50% in favor, and they use the appropriate 95% confidence interval spanning from 40% to 60%. Expect 95% (within the interval) to not reject the hypothesis. They don't know it, but they are in fact correct. Expect 5% (outside the interval) to reject the hypothesis and so suffer a Type I Error. They don't know it, but they are in fact incorrect. They just happened to get a misleading random sample. Type II is irrelevant because the hypothesis is in fact correct (unbeknownst to any of the surveyors).

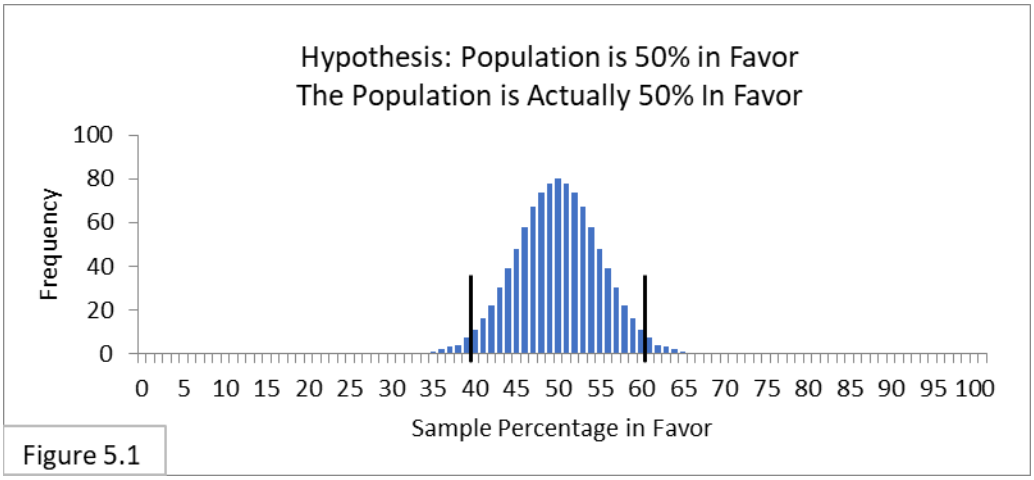


Figure 5.1

Now imagine that you are one of those surveyors. There's a 95% chance that you're one of the ones whose random sample led you to the correct conclusion, and a 5%

chance that you're one of the ones whose random sample led you to the incorrect conclusion (Type I Error).

Case 2

Referring to Figure 5.2, suppose a population is 55% in favor of a new public health policy, and 1,000 surveyors survey the population using random sampling of sample size 100. All 1,000 surveyors hypothesize that the population is 50% in favor and use the appropriate 95% confidence interval spanning from 40% to 60%. Expect about 85% (within the interval) to not reject the hypothesis and so suffer Type II Error. They don't know it, but they are in fact incorrect. Expect about 15% (outside the interval) to reject the hypothesis. They don't know it, but they are in fact correct. Type I Error is irrelevant because the hypothesis is in fact incorrect (unbeknownst to any of the surveyors). It may seem shocking, but because 55% is so close to 50% and because 100 is a somewhat small sample size, about 85% of the surveyors are expected to reach the wrong conclusion!

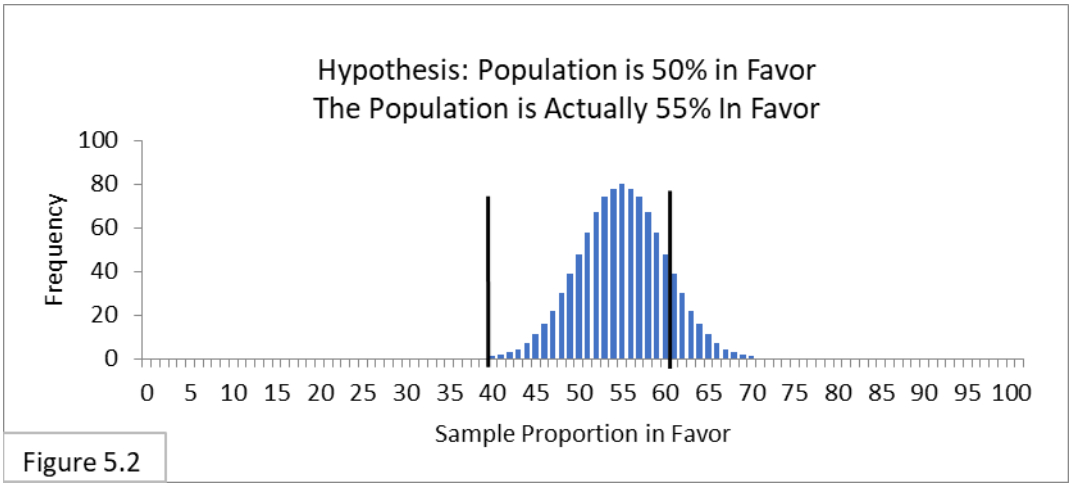


Figure 5.2

Now imagine that you are one of those surveyors. There's a 15% chance that you're one of the ones whose random sample led you to the correct conclusion, and an 85% chance that you're one of the ones whose random sample led you to the incorrect conclusion (Type II Error).

Case 3

Let's increase the sample size.

Referring to Figure 5.3, suppose a population is 50% in favor of a new public health policy, and 1,000 surveyors survey the population using random sampling of sample size *1000*. All 1,000 surveyors hypothesize that the population is 50% in favor. Nothing has changed regarding Type I Error because they're all still using a 95% interval, which now spans from 47% to 53%, and the hypothesis is true. We still expect 95% to be correct and 5% to be incorrect. Type II Error is irrelevant because the hypothesis is actually true.

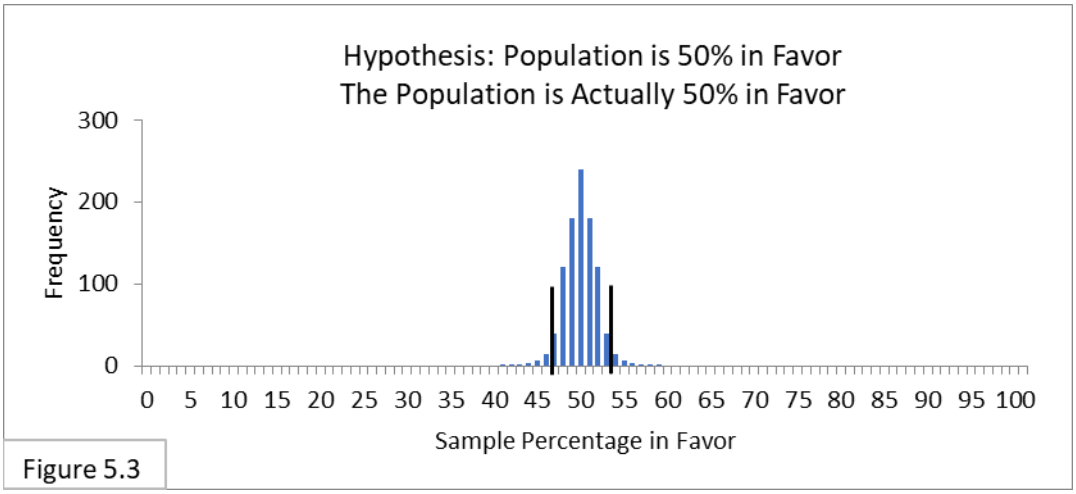


Figure 5.3

Now imagine that you are one of those surveyors. There's a 95% chance that you're one of the ones whose random sample led you to the correct conclusion, and a 5% chance that you're one of the ones whose random sample led you to the incorrect conclusion (Type I Error).

Case 4

Referring to Figure 5.4, suppose a population is 55% in favor of a new public health policy, and 1,000 surveyors survey the population using random sampling of sample size 1000. Their hypothesis is that the population is 50% in favor of the new policy, and they use the corresponding 95% interval of 47%-to-53%. Because of the increased sample size—increased power—we now expect about 90% of the surveyors to be correct (rather than 15% in Case 2). And we expect about 10% to be incorrect and suffer Type II Error (rather than 85% in Case 2). That's much better. Type I Error is irrelevant because the hypothesis is actually false.

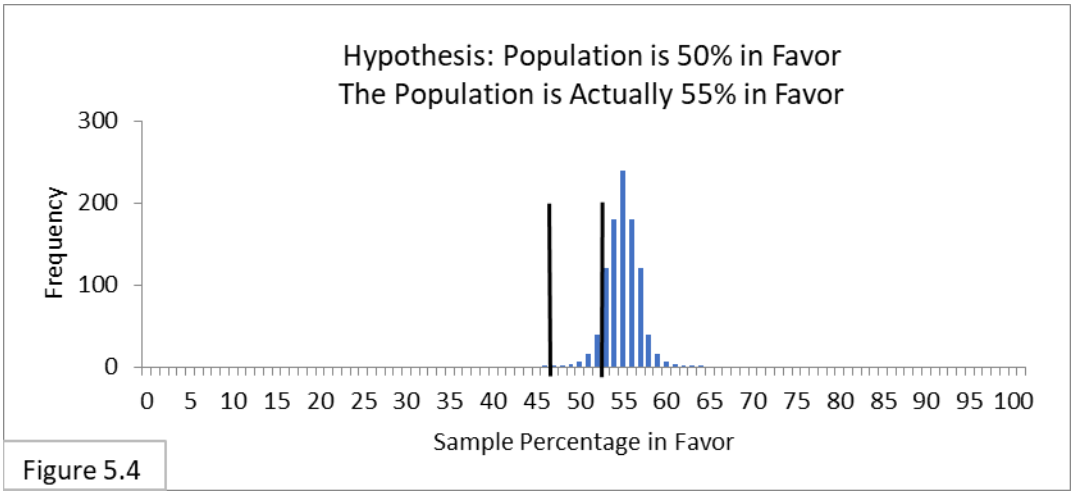


Figure 5.4

Now imagine that you are one of those surveyors. There's a 90% chance that you're one of the ones whose random sample led you to the correct conclusion, and a 10% chance that you're one of the ones whose random sample led you to the incorrect conclusion (Type II Error). Much better odds than in Case 2!

Case 5

Let's make our Type I Error criterion stricter.

As noted earlier, Type I Error is feared more than Type II Error, and since we've managed, in Case 4, to lower the expected Type II Error rate to about 10%, let's take advantage of that and make our Type I Error criterion stricter, knowing full well that that will increase the likelihood of Type II Error.

Referring to Figure 5.5, suppose a population is 50% in favor of a new public health policy, and 1,000 surveyors survey the population using random sampling of sample size 1000. Their hypothesis is that the population is 50% in favor of the new policy. Now they are going to use 99% intervals, which span from 46% to 54%. Now we expect 99% of the surveyors to be correct and 1% to be incorrect. Type II Error is irrelevant.

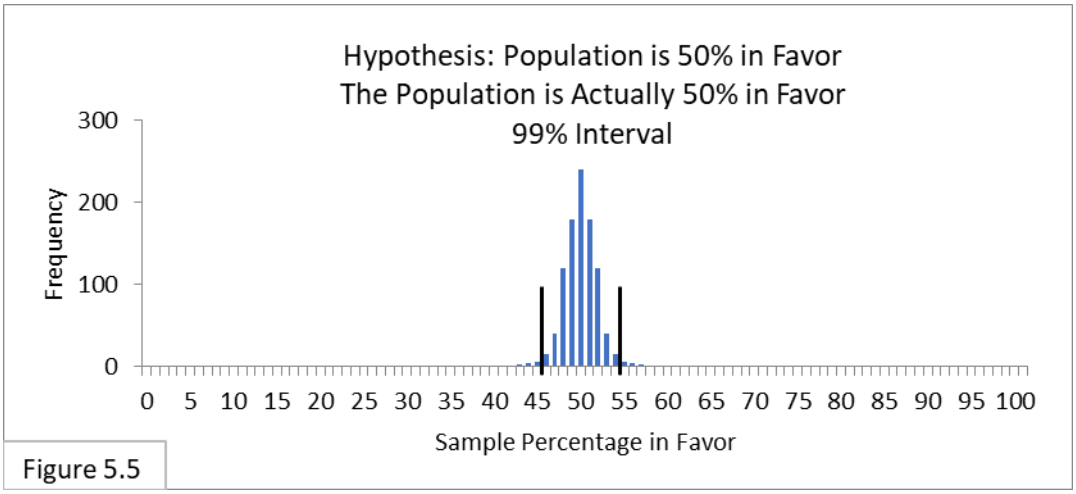


Figure 5.5

Now imagine that you are one of those surveyors. There's a 99% chance that you're one of the ones whose random sample led you to the correct conclusion, and a 1% chance that you're one of the ones whose random sample led you to the incorrect conclusion (Type I Error).

Case 6

Referring to Figure 5.6, suppose a population is 55% in favor of a new public health policy, and 1,000 surveyors survey the population using random sampling of sample size 1000. Their hypothesis is that the population is 50% in favor of the new policy, and they use the corresponding 99% interval of 46%-to-54%. Because of the stricter Type I Error criterion and its wider 99% interval, we now expect about 70% of the surveyors to be correct (rather than 90% in Case 4). And we expect about 30% to be incorrect and suffer Type II Error (rather than 10% in Case 4). Type I Error is irrelevant.

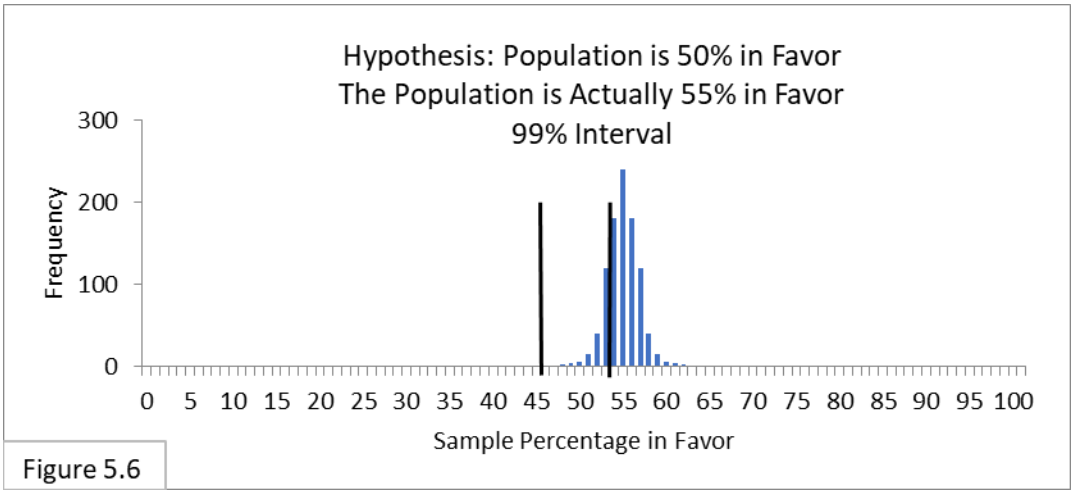


Figure 5.6

Now imagine that you are one of those surveyors. There's a 70% chance that you're one of the ones whose random sample led you to the correct conclusion, and a 30% chance that you're one of the ones whose random sample led you to the incorrect conclusion (Type II Error).

With sample size 1000, which would you choose, the Type I and II Errors rates of 5% and 10% as in Cases 3 and 4, or 1% and 30% as in Cases 5 and 6? Well, that depends on how costly making each kind of error is. And that depends on context. If the repercussions of Type I Error are much worse than those of Type II Error, then you'd pick 1% and 30%. If not, you'd pick 5% and 10%. Or, you could pay to have even larger samples gathered and try for 1% and 10% to get the best of both! (But remember, we've only considered Type II Error in cases where the population is 55% in favor. If we considered 51% to 54% we would see more Type II Error and if we considered 56% or more we would see less Type II Error.)

The Bottom Line: Sample size is an important way to manage error rates. Larger sample sizes allow you to make your Type I Error level stricter, if desired, while also making sure your Type II Error level remains reasonable.

Terminology and Notation: Mathematically, the probability of Type I Error is denoted by the lower-case Greek letter alpha, α. The percentage level for confidence—which we've been referring to a lot—is (1- α)*100%. So, an alpha-level of 0.05 is equivalent to a confidence level of 95%. The probability of Type II Error is denoted by the lower-case Greek letter beta, β. Power is 1-β (*not* made into a percentage). So, for example, a beta level of 0.20 is equivalent to a power level of 0.80.

6. Formalizing Hypotheses

When we looked at Type I & II Error, I referenced various *hypotheses*. And in Chapter 5 *A Series of Six Short Case Studies,* I casually posited the surveyors' hypotheses that "the population is 50% in favor of the new policy". It's time to get more formal.

The central hypothesis is called the *Null Hypothesis*. It's central, not because you think or hope it's true. *Usually, it's the exact opposite*. For example, a pharmaceutical company testing a new drug obviously wants it to be shown effective via drug trial data, but their Null Hypothesis is that the drug is *not* effective. Then, they hope to reject the Null Hypothesis. The Null Hypothesis is central because it's the hypothesis you are testing. We can frame a statistical analysis simply by whether or not we reject the Null Hypothesis. But if desired, we can also formally state an Alternate Hypothesis that is to be accepted if the Null Hypothesis is rejected. As we'll see, though, the Null Hypothesis is never accepted; it's either rejected or not rejected—a subtle but fundamental difference.

The Null Hypothesis is a statement that may be ruled out by evidence (the sample data). Typically, the Null Hypothesis is an equality (and the optional Alternate Hypothesis is an opposing inequality). For example

Null: The population percentage in favor *is equal to* 50%
Alternate (optional): The population percentage in favor *is not equal to* 50%

With these terms, we can define Type I & II Errors more formally.

Type I Error: Rejecting the Null Hypothesis when it is actually true.
Ex. The population is 50% in favor, but your sample leads you to reject that it's 50% in favor.
Type II Error: Not rejecting the Null Hypothesis when it is actually false.
Ex. The population is not 50% in favor, but your sample leads you to not reject that it is 50% in favor.

Let's consider why the Null Hypothesis may be rejected or not rejected but never accepted. We'll look at it a few different ways, starting with a classic analogy: There is a hypothesis that a certain animal native to Tasmania is extinct. That is the Null Hypothesis. A group of scientists goes out looking for it. If they find one, can they

reject the Null Hypothesis that the animal is extinct? Yes, of course. But if they don't find one, can they accept the Null Hypothesis that the animal is extinct? Of course not. The absence of evidence is not evidence of absence.

Now let's look at some cases closer to home: Surveyor #1's Null Hypothesis is that the Flowing Wells population is 50% in favor of a new public health policy. Her frame of reference is Figure 5.1 (reproduced below).

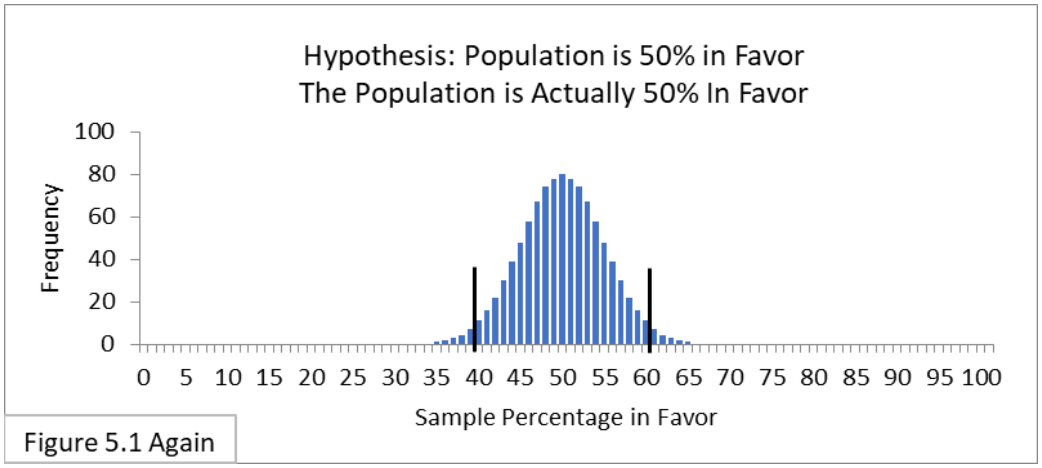


Figure 5.1 Again

Surveyor #2's Null Hypothesis is that the Flowing Wells population is 55% in favor of the same new public health policy. His frame of reference is Figure 6.1.

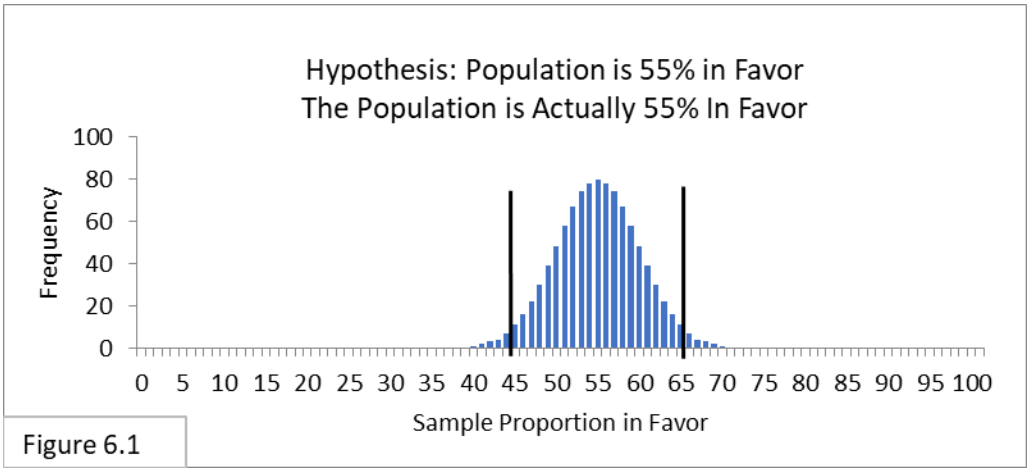


Figure 6.1

Both surveyors use the same survey sample of 100 random people, and the sample proportion turns out to be 53%. 53% is within both of the surveyors' 95% intervals. If we allowed each of them to accept their Null Hypothesis, then she would accept the Null Hypothesis that the population proportion is 50%, and he would accept the

Null Hypothesis that the population proportion is 55%. They can't both be right! No, they can't both be right, and so we can't let them accept their Null Hypothesis. Each can only not reject their Null Hypothesis.

Next, let's look at a variation on that theme. We'll use Figure 5.2, reproduced below. The Null Hypothesis is that the Flowing Wells' population proportion is equal to 50%, but let's say that we, as know-it-alls, know that it's actually equal to 55%.

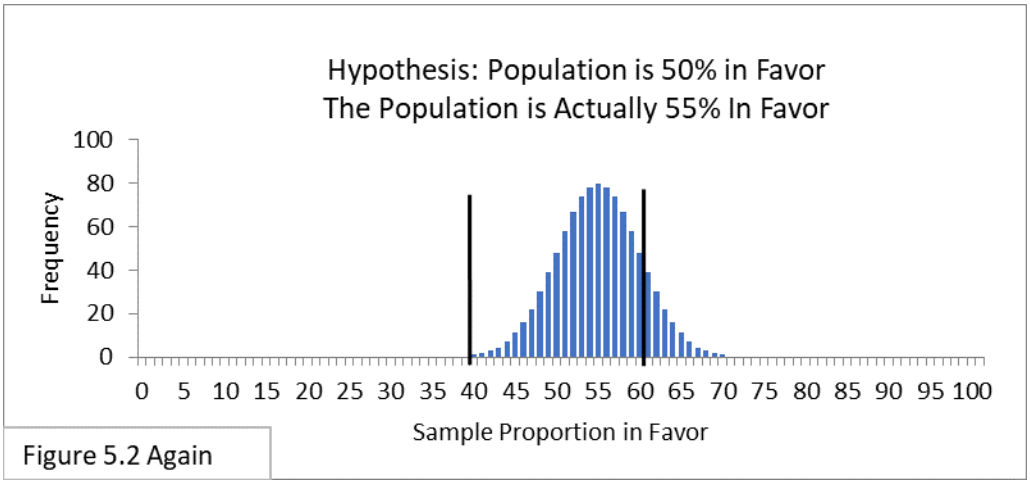


Figure 5.2 Again

Here, a population that is actually 55% in favor has a sampling distribution with the majority of sample percentages within the 95% interval for 50%. Should all the surveyors who get sample proportions within the 40%-to-60% interval accept their Null Hypothesis and say they are confident that the population percentage is 50%? No. All the surveyors whose percentages are outside the 40%-to-60% interval will reject the Null Hypothesis that the population percentage is 50%, and all the surveyors whose proportions are within the 40%-to-60% interval will not reject the Null Hypothesis (and unknowingly suffer Type II Error).

And finally, in many cases it would actually be very unlikely for the Null Hypothesis to be true anyway. Flowing Wells has 80,000 residents. For the Null Hypothesis that 50% of the Flowing Wells' population is in favor of the new public health policy to be true, 40,000 residents must be in favor of the policy. Not 39,999. Not 40,001. It must be exactly 40,000. That's another reason why we never accept the Null Hypothesis as true. It's often too exacting. *But always remember, the Null Hypothesis is something we want to see if we can reject; it's not something we want to see if we can accept.*[vi]

In a nutshell: We may infer that the Null Hypothesis is false, but we never infer that it's true. If the evidence is strong enough, we reject the Null Hypothesis. And when we reject the Null Hypothesis, we call the result *statistically significant.* If the evidence is not strong enough, we don't reject the Null Hypothesis. We never accept the Null Hypothesis.

7. The Limited Meaning of Statistical Significance

When our sample statistic is outside of our 95% confidence interval, we reject the Null Hypothesis and call the result statistically significant. What does "statistically significant" mean? What does it tell us? Recall from Chapter 4 *Veridical vs. Misleading Results* that, at first, many people think that a statistically significant result using a 95% confidence interval tells them that there is a 95% chance they're correct and a 5% chance they're incorrect. But, unfortunately, that's not what it means. Its meaning is much more limited. It only tells us that there is a 95% chance we're correct and a 5% chance we're incorrect *when the Null Hypothesis is true* (see cases 1 and 3 in Chapter 5 *A Series of Six Short Case Studies*). It's on this basis that we reject the Null Hypothesis.

The meaning of statistical significance is limited in another way as well. Many people think, at first, that statistical significance tells them that the results must have meaningful real-world implications, that the results are *practically significant*. But, unfortunately, that's not what it means either. The term *significant* is qualified with the term *statistically.* It doesn't mean generally significant or practically significant or meaningfully significant. To illustrate, let's look at three examples, the first two of which we've seen before. All involve calculating the 95% interval surrounding .5. They involve sample sizes of 100, 1000, and 10,000.

$$.5 \pm 2 * \sqrt{\frac{.5 * (1 - .5)}{100}} = .5 \pm 0.10 = 40\% \text{ to } 60\%$$

$$.5 \pm 2 * \sqrt{\frac{.5 * (1 - .5)}{1000}} == .5 \pm 0.03 = 47\% \text{ to } 53\%$$

$$.5 \pm 2 * \sqrt{\frac{.5 * (1 - .5)}{10{,}000}} == .5 \pm 0.01 = 49\% \text{ to } 51\%$$

As you know, the 95% interval narrows as sample size increases. At some point the 95% interval will narrow to effectively nothing. In the case here of a sample size of 10,000 the interval is very narrow. The margin of error is only 0.01, or 1%. While a sample size of ten thousand may seem ridiculously large, in the modern age of digitized "big data" it actually isn't. Regardless, the main point remains that with large enough sample sizes we can make nearly any result statistically significant.

So, if we had a sample of ten thousand and our sample percentage was 51.1% what would we infer? Well, we would reject the Null Hypothesis that the population is 50% in favor, and we would call the result statistically significant. *Because it is*. *But* what about practical significance? Is 51.1% meaningfully different from 50% in terms of its practical implications? Maybe, maybe not.

Statistical significance *is* important when assessing the results of statistical analysis, but you also need to look at the actual statistic values involved and decide whether they are practically significant, with meaningful real-world implications.[vii]

And while we do want to have sample sizes large enough to avoid undue risk of Type II Error, we also have to be wary when using sample sizes so large that negligible results become statistically significant.

Here's an example of that: A study found that a certain dietary supplement lowered the risk of getting a certain minor ailment from 2 in a 1000 (0.2%) down to 1 in a 1000 (0.1%). The sample size of the study was 30,000, so the difference between 0.2% and 0.1% is statistically significant (at 95% confidence). That gives a *relative risk* difference of 50% ((0.2%-0.1%)/.2%) but an *absolute risk* difference of only 0.1% (0.2%-0.1%). Advertisements for the supplement highlighted the facts that the supplement's positive effect was statistically significant and that the supplement reduced the risk of getting the ailment by 50%, but the advertisements did not mention that the absolute risk reduction was only 0.1%. Many people would find that misleading. And many people would consider an absolute risk difference of 0.1% to be negligible and practically insignificant. Bottom line: You definitely want to know both the relative and absolute differences in order to better assess practical significance.

8. Approximating Binomial Distributions: The z-distribution

Next, we need to finalize the Standard Normal distribution we met earlier (in Chapter 3 Section 1 *Time for Some Standardization*). Here, we'll go further and formulate something called the *z-distribution.* The z-distribution is important because it is the ultimate source of the formulas we've been using; it is where ±1.95996... comes from.

The sampling distributions for binomial variables we've been looking at are *discrete* distributions (discrete values such as 0, .01, .02, ..., 1 on the horizontal axis and discrete frequency counts on the vertical axis), but the equation we'll see in a moment defines a *continuous* distribution (continuous values on the real number line for both the horizontal and vertical axes). Let's transform the discrete binomial distribution of Figure 8.1 to the Standard Error scale, as we did earlier, and then to the continuous Standard Normal distribution, which is most often called the z-distribution.

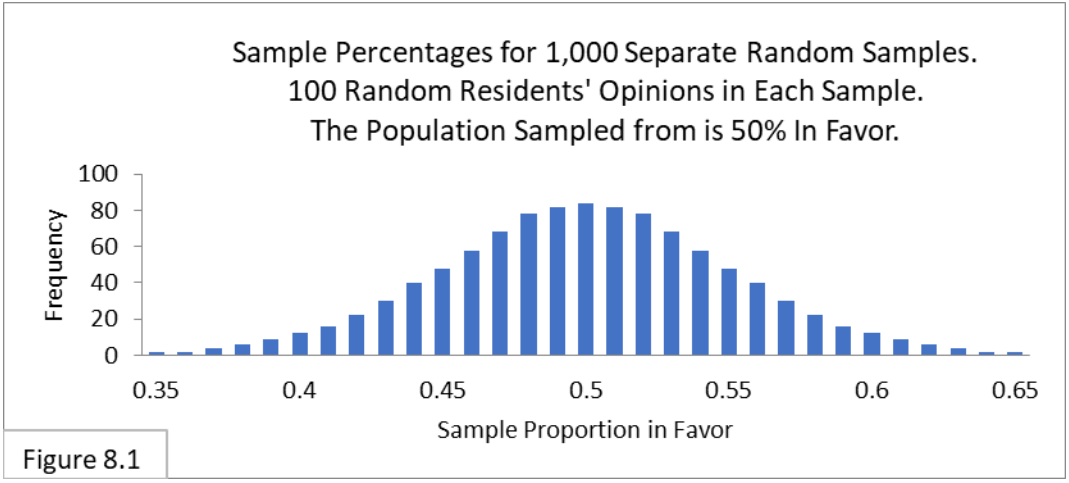


Figure 8.1

First, on the horizontal axis we simply subtract .5 from each value (to center the distribution on zero) and divide each by the Standard Error determined using the below formula, as we saw earlier.

$$\frac{p - .5}{\sqrt{\frac{p * (1 - p)}{n}}}$$

Viola, we get Figure 8.2.

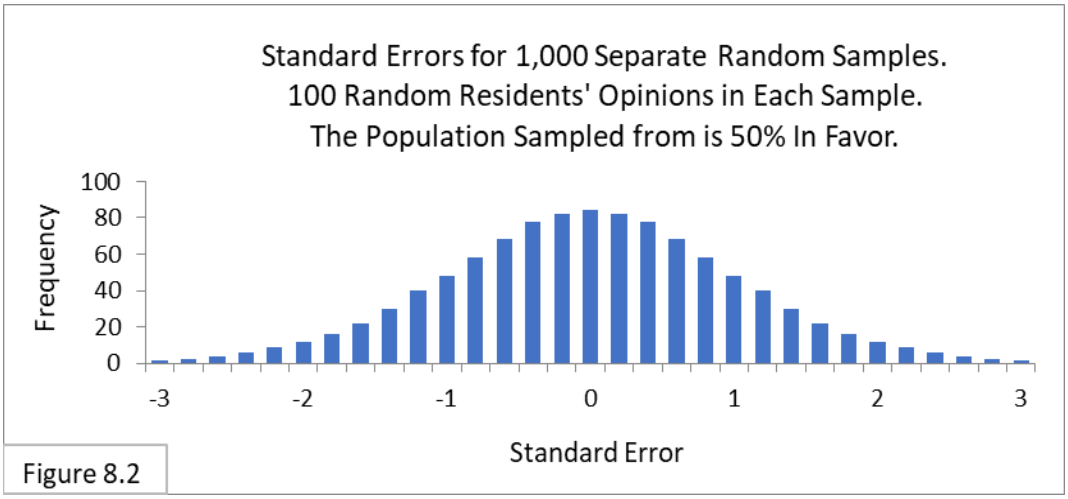


Figure 8.2

Next, we connect the tops of all the bars to make a continuous distribution, and we replace the frequency scale on the vertical axis with a probability scale (more on this below). Viola, we get Figure 8.3, the famous z-distribution. The letter z denotes the Standard Error scale.

Figure 8.3

The scale on the vertical axis in Figure 8.3 is now probability. Probability is a continuous scale going from 0 to 1. A probability of 1 means something will always happen and a probability of 0 means something will never happen. A probability of 0.5 means something will happen half the time.[viii]

The z-distribution is a probability distribution that is normal, standardized, and continuous. It is one of the most important *standardized probability distributions* in statistics. Since it's a probability distribution, the entire area under the curve is 1. And, since it is a continuous function, probabilities for specific values—such the probability of z exactly equaling 1.96—are zero. We always need to refer to the

probability of ranges—such as the probability that z is less than -1.96 or greater than 1.96.

Recall that the boundary lines for the 95% interval on the Standard Normal Distribution are always -1.96 and 1.96, as shown in Figure 8.4. The area under the curve within the boundary lines is 0.95 and the total area outside is 0.05, with 0.025 on each side.

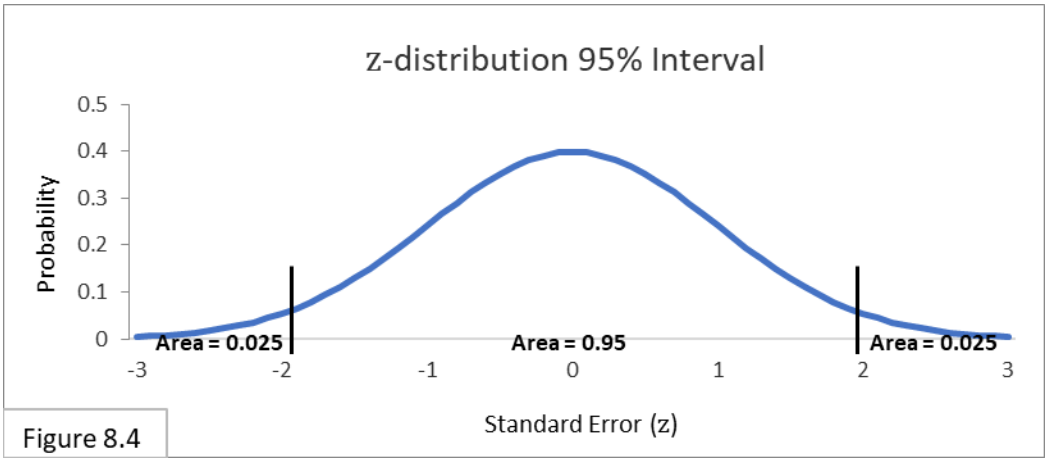


Figure 8.4

In light of the z-distribution and its Standard Error scale, let's revisit the two primary formulas we've been using. In the below formula, multiplying ±1.96 times the Standard Error gives us the 95% interval expressed in proportions.

$$p \pm 1.96 * \sqrt{\frac{p * (1-p)}{n}}$$ the 95% interval scaled in proportions.

And in the below formula, dividing the difference between a proportion and a fixed proportion value (such as .5) by Standard Error gives us the difference expressed in Standard Errors.

$$\frac{p - .5}{\sqrt{\frac{p * (1-p)}{n}}}$$ proportion scaled to the z-distribution.

Many statistical formulas are, or involve, such *scale conversions*. Let's go through complementary examples.

As examples of using each of the two formulas, with illustrations, let's say we took a random sample of 100 and got a sample proportion of 0.62. Our hypothesis is that the population proportion is 0.50. The 95% interval calculation gives us

$$.5 \pm 1.96 * \sqrt{\frac{.5 * (1 - .5)}{100}} = 0.4 \text{ to } 0.6$$

Figure 8.1, annotated below, shows the sampling distribution along with the confidence interval of 0.4 to 0.6 and the sample proportion value of 0.62 superimposed on it.

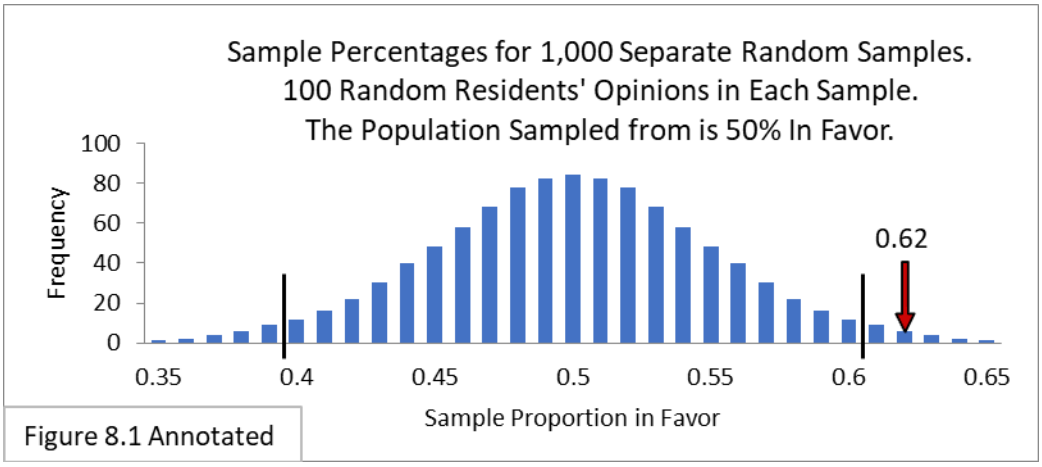


Figure 8.1 Annotated

Using the second formula we get

$$\frac{.62 - .5}{\sqrt{\frac{.5 * (1 - .5)}{100}}} = 2.40$$

Figure 8.4, annotated below, shows the z-distribution along with its standard confidence interval of -1.96 to 1.96 and the Standard Error value of 2.40 superimposed on it.

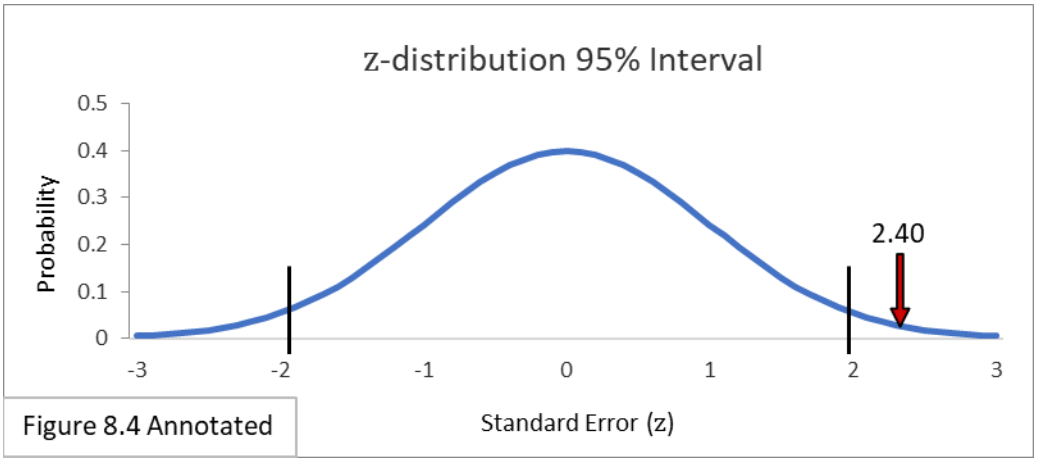


Figure 8.4 Annotated

This illustrates the equivalence between the two: the statistical result is outside the 95% interval in the same relative location. You get the same result whether you 1) calculate the confidence interval boundaries in terms of proportions using the z-distribution's ±1.96 multiplier, and then see if your sample proportion value is outside the interval, or 2) calculate the Standard Error for your sample proportion, and then see if the Standard Error value is outside the z-distribution's standard ±1.96 interval. From here on, we'll be using both types of calculation and illustration.

The z-distribution *approximates* the binomial distribution. It is not an *exact* match, because the binomial is a discrete distribution, but, as long as the sample size is large enough, it is a very useful approximation. (We'll look at exceptions in Chapter 9 *Addressing Assumptions*.)

Note: The equation for the continuous z-distribution curve itself is shown below. There is no practical need for you to know this equation. I put it here so you could see that there is indeed an equation that defines the z-distribution curve.

$$\frac{1}{\sqrt{2\pi}} e^{\left(-\frac{x^2}{2}\right)}$$

where, x is Standard Error (the horizontal axis). For those of you acquainted with *calculus*, you know that you can find the area under specific regions of the curve using *integration* to determine the probability of having values within those regions.[ix]

9. Addressing Assumptions

Nearly all statistical tests make specific assumptions about the data being analyzed. If the assumptions are not met, then the statistical test will yield questionable results and shouldn't be used. The assumptions are often related to sample sizes and the nature of the distributions of the data values themselves. This chapter overviews these types of assumptions for binomials to give you a feel for what they are and why they matter. In practice, the assumptions for any given statistical analysis method must be carefully assessed prior to its use. When a given method's use cannot be justified, then a less restrictive method must be chosen—one that makes fewer or different assumptions—or the data must be transparently "massaged" in a professionally appropriate way to make it better adhere to a method's assumptions. A professional will report everything that was done as part of any analysis.

For binomial data the two most important assumptions have to do with sample size and extreme proportion values. In §*8. Approximating Binomial Distributions: The z-Distribution,* we saw that binomial sampling distributions can be approximated with the z-distribution—Figure 8.3, reproduced below—but only when sample sizes are large enough. In addition, the proportion values cannot be too close to zero or one. These stipulations are critical because the formulas we've been using assume that the z-distribution is a proper approximation. Let's look at these issues in more detail.

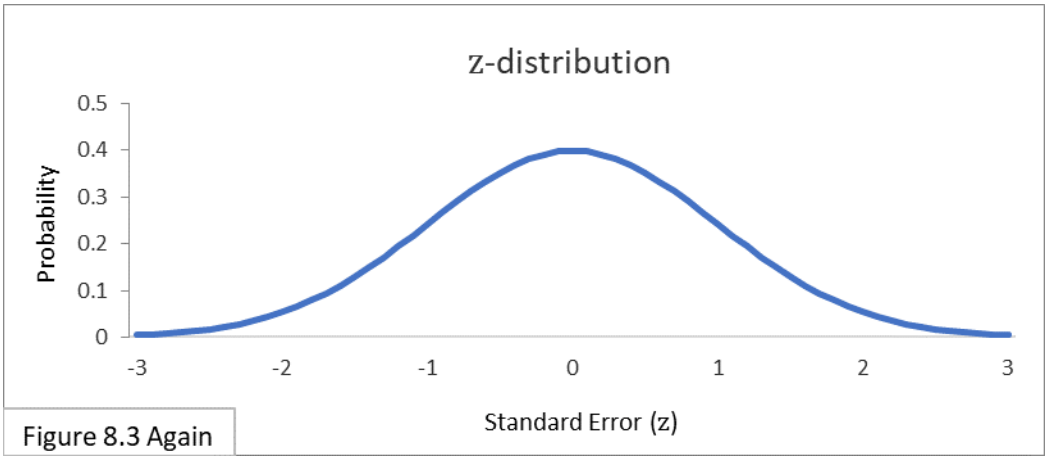


Figure 8.3 Again

If the sample size is too small, the z-distribution is a poor approximation to the binomial distribution. For example, if our sample size is only 2, then the z-

distribution is obviously a poor approximation as shown by the sampling distribution in Figure 9.1.

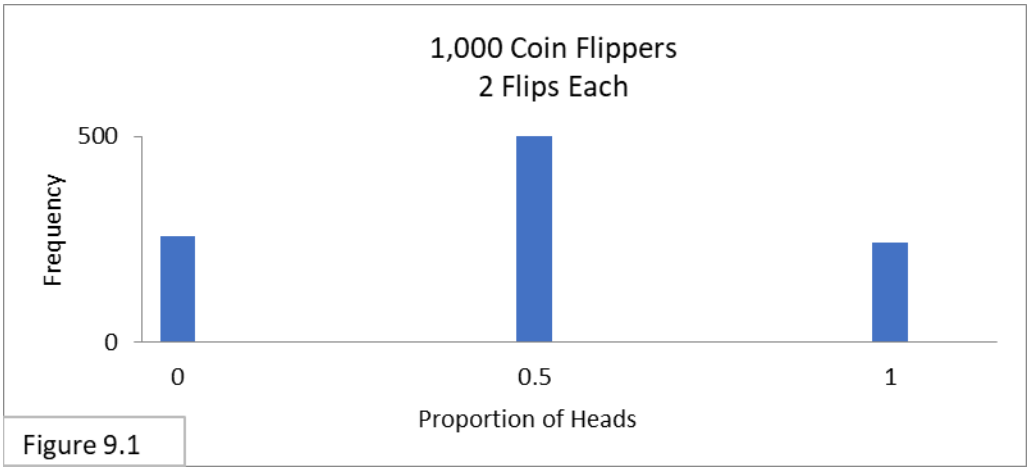


Figure 9.1

In the case of Figure 9.1, the approximation is so poor that the 95% confidence interval calculated via the formula, shown below, goes *out-of-bounds* on both sides with proportion boundary lines of -.2 and 1.2. These are both infeasible values for proportions!

$$.5 \pm 1.96 * \sqrt{\frac{.5 * (1 - .5)}{2}} = .5 \pm .69 = \text{-.2 to 1.2 (rounded)}$$

When p is too close to zero or one, then the binomial distribution will distort from the normal bell shape. In such cases the z-distribution will also be a poor approximation. Figure 9.2 shows that the sampling distribution for p of 0.9 with sample size of 30 is distorted from the normal bell shape.

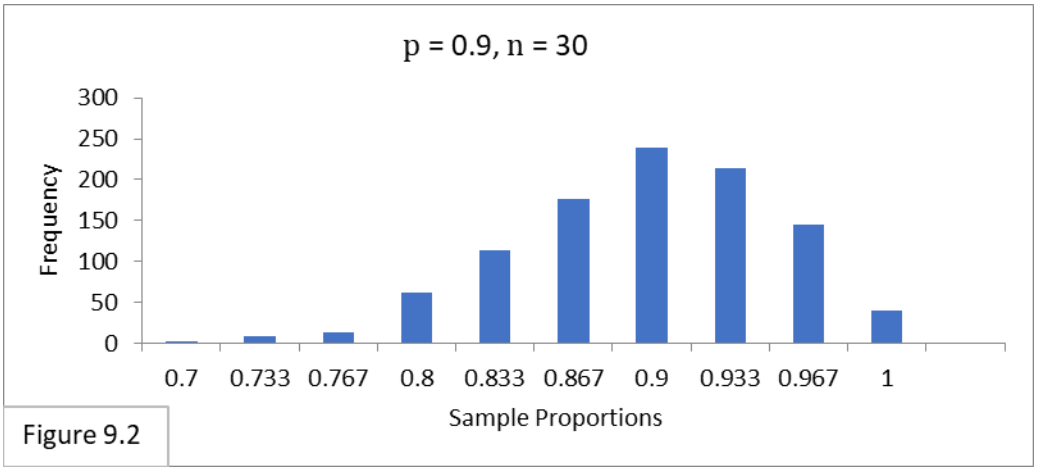


Figure 9.2

In the case of Figure 9.2, the 95% confidence interval calculated via the formula, shown below, has the extreme proportion boundary line value of 1 on the right-hand side (prior to rounding, it actually calculates to slightly greater than 1).

$$.9 \pm 1.96 * \sqrt{\frac{.9 * (1 - .9)}{30}} = .9 \pm .1 = .8 \text{ to } 1.0 \text{ (rounded)}$$

How can we determine whether the z-distribution will be a good approximation for a given binomial sampling distribution? Recall that both p and n influence the nature of the sampling distribution (Chapter 2 *Sampling Distribution Dynamics*). So, the values of both n and p must be considered together. A commonly used rule-of-thumb for the minimum sample size, n, needed for any given proportion, p, is to make sure n is large enough so that

$$n * p \geq 10 \quad \text{and} \quad n * (1 - p) \geq 10$$

For Figure 9.1, by using this rule-of-thumb for p of 0.5 we can determine that n of 2 is too small because 2*.5 only equals 1.

On the other hand, with n of 20 we get 20*(.5) which equals 10. Figure 9.3 illustrates that with a sample size of 20, the sampling distribution has filled in and narrowed enough to attain a normal shape.

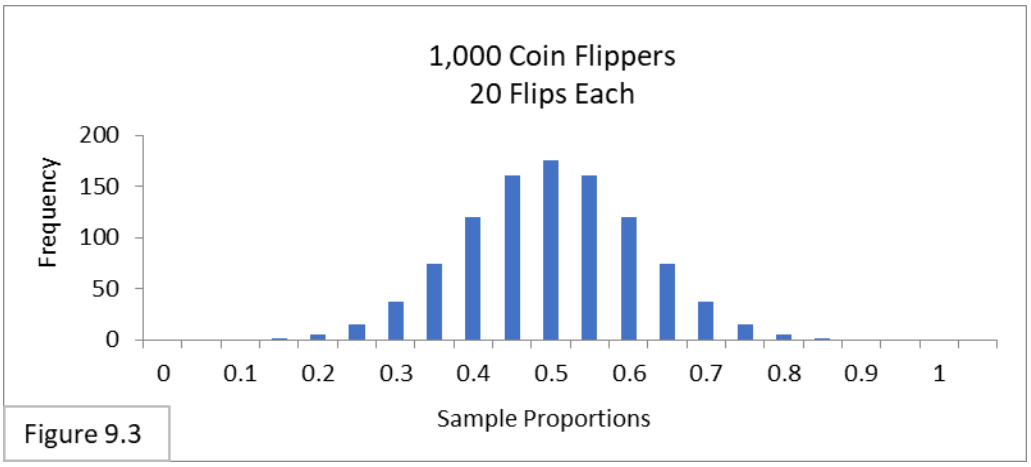


Figure 9.3

And the 95% confidence interval calculated with the formula, shown below, appears to be accurate in light of Figure 9.3.

$$.5 \pm 1.96 * \sqrt{\frac{.5 * (1 - .5)}{20}} = .5 \pm .2 = .3 \text{ to } .7 \text{ (rounded)}$$

For Figure 9.2, by using this rule-of-thumb for p of 0.9, we can determine that n of 30 is too small because 30*(1-0.9) only equals 3. On the other hand, with n of 100 we get 100*(1-0.9) which equals 10. Figure 9.4 illustrates that with a sample size of 100, the sampling distribution has narrowed enough to attain a normal shape.

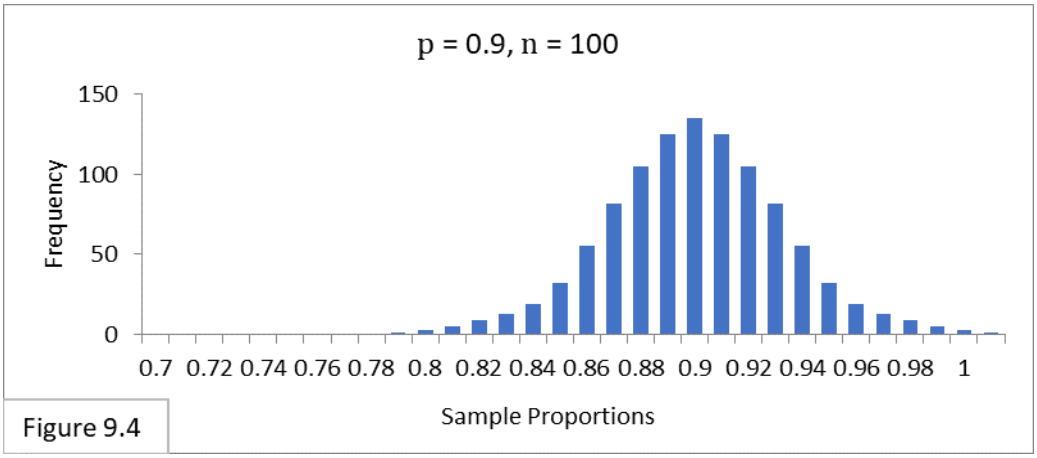


Figure 9.4

And the 95% confidence interval calculated with the formula, shown below, now appears to be accurate in light of Figure 9.4.

$$.9 \pm 1.96 * \sqrt{\frac{.9 * (1 - .9)}{100}} = .9 \pm .06 = .84 \text{ to } .96 \text{ (rounded)}$$

So, in summary, the z-distribution *approximates* the binomial distribution, and sample sizes must be adequate for the formulas to work correctly, and p cannot be too close to zero or one. With sample sizes that are too small and p that are too close to zero or one, alternatives called *exact methods* can be used.[x]

10. Analyzing the Difference Between Two Groups Using Binomial Proportions

Now that we've covered many of the essential foundations laid out in the introduction, let's put them into action while looking at another common type of analysis involving binomial proportions. We'll keep surveying the population in Flowing Well, but now we'll compare them to the population in a neighboring town, Artesian Wells. The reason for our surveys is to assess how residents feel about a newly proposed state government initiative.

The state government is considering state tax increases to fund some new government programs. We're going to survey Flowing Wells residents as we have been doing, but we're also going to survey residents in the nearby town of Artesian Wells. Historically, the town of Flowing Wells is known to lean socialist and probably favors the new taxes and programs, but the town of Artesian Wells is thought to lean libertarian and probably doesn't. The survey is going to be conducted in each of the two towns to see if the proportions of residents that are in favor of the new taxes and programs are different in the two towns. The corresponding Null Hypothesis is that they are *not* different:

> There is *no difference* between the Flowing Wells and the Artesian Wells communities' population proportions. That is, the *difference* between the Flowing Wells and the Artesian Wells population proportions *equals zero*.

The relevant sample statistic is the Flowing Wells' sample proportion minus the Artesian Wells' sample proportion. (Using the Artesian Wells' sample proportion minus the Flowing Wells' sample proportion will give us fundamentally the same results.)

Unfortunately, the not-for-profit organization that is conducting the survey only has resources to survey 100 random residents in each town.

First, let's get a "bird's eye view" of the situation. The Null Hypothesis is that there's no difference between the Flowing Wells and the Artesian Wells communities' population proportions, so let's look at simulation results that assume that is true. We'll assume that both Flowing Wells and Artesian Wells have an overall community population opinion of 50% (0.50) agree, such that the difference in the population proportions equals zero. If 100 people are randomly surveyed in each of the two

communities, how likely is it that various sample proportion *differences* could arise by chance, due to the randomness inherent in the sampling? The simulation takes two random samples of size 100 from populations that are 50%-in-favor and subtracts one of the sample proportions from the other. It does this over and over and over.

Figure 10.1 illustrates the simulation results, which is the sampling distribution of what to expect when the Null Hypothesis is true.

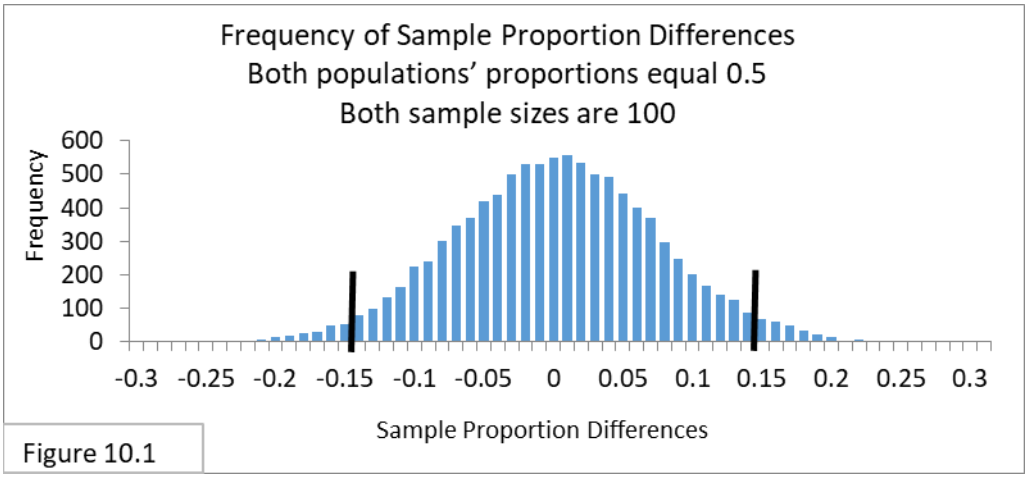


Figure 10.1

You can see that the sampling distribution for the *difference* between two sample proportions is a *normal* distribution. And it appears the 95% interval for this situation has boundary lines of -.14 and .14. So, if the difference between two sample proportions is within the interval -.14 to .14, then we won't reject the Null Hypothesis. If it's outside the interval, then we will reject the Null Hypothesis and say that the difference between the two communities is statistically significant. The limited resources that have constrained the sample sizes to 100 gives a fairly wide interval, where the two sample proportions have to be at least 0.15 apart to reject the Null Hypothesis. The small samples and wide confidence interval raise concerns about Type II Error.

The Standard Error formula for the *difference* between two population proportions is

$$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

Notice that the terms within the square root follow the same "variance divided by sample size" structure we saw with the Standard Error for a single proportion. Now there is a term for each of the two populations.

Both samples are size 100, and we are assuming both have population p equal to .5, making their difference equal to zero. Using the 95% confidence interval formula we get

$$0.0 \pm 1.96 * \sqrt{\frac{.5(1-.5)}{100}+\frac{.5(1-.5)}{100}} = 0.0 \pm .14 \text{ (rounded)}$$

That agrees with Figure 10.1; the sampling distribution is approximated nicely by the z-distribution that's embodied in the formula.

For an example analysis, let's suppose the sample proportion for the survey in Flowing Wells is .52 and for Artesian Wells is .44. The difference is .08, which is within the 95% interval. Conclusion: Do not reject the Null Hypothesis. (Remember, we never accept the Null Hypothesis, we just don't reject it.)

Now that we've looked at the bird's eye view, let's get back to conducting the surveys. The first survey is conducted in March. We are going to analyze the survey data using the other primary formula type we've been utilizing. We'll convert a sample proportion *difference* to the Standard Error scale. Below is the formula for the Standard Error of the difference between two sample proportions.

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Let's suppose the survey data yields a sample proportion for Flowing Wells of .52 and for Artesian Wells of .44. Are these two sample proportions far enough apart that we can say the difference is statistically significant? First, we'll calculate the Standard Error of the sample proportion difference.

$$\frac{.52 - .44}{\sqrt{\frac{.52(1-.52)}{100} + \frac{.44(1-.44)}{100}}} = 1.136 \text{ (rounded)}$$

The proportion difference of .08 (.52-.44) equates to 1.136 Standard Errors. Looking now at the Standard Normal Distribution (z-distribution) in Figure 10.2 we can see that the result is inside the 95% interval and so we do not reject the Null Hypothesis that Flowing Wells and Artesian Wells have equal population proportions. The .08 sample proportion difference is not statistically significant.

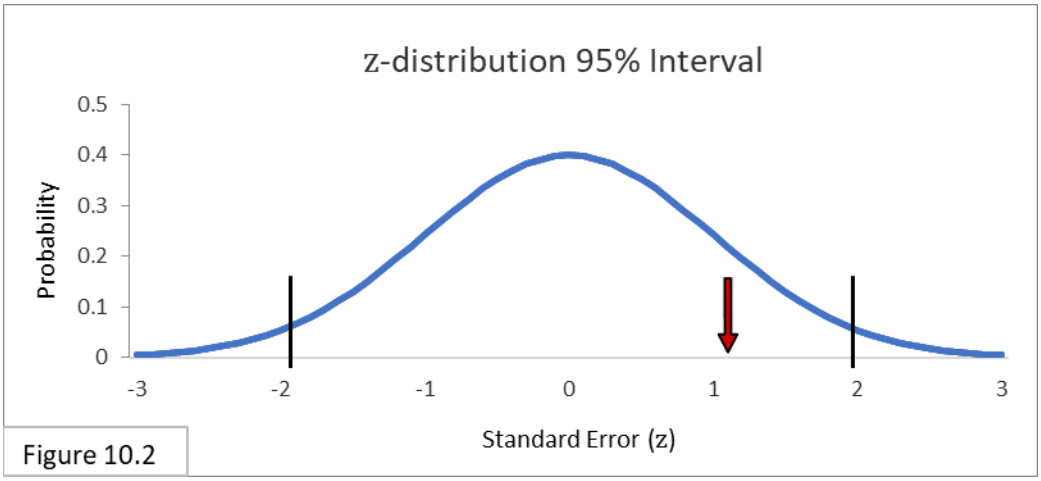


Figure 10.2

Someone from the not-for-profit organization raises the issue of practical significance and suggests that 52% and 44% are different enough to be considered politically meaningful… *Wait!* you say. *Without statistical significance we really shouldn't even be thinking about that! The results do not allow us to talk as if the population proportions are different at all!* Nonetheless, you add helpfully, with the small sample sizes, Type II Error does seem like a real possibility.

Suppose the survey is conducted again a month later, in April, and the two sample proportions are farther apart: .52 and .34.

First off, since .34 is closer to zero than most of the proportions we've encountered so far, let's check the assumption rule-of-thumb we saw in the previous chapter:

$$n * p \geq 10 \quad \text{and} \quad n * (1 - p) \geq 10$$

With a sample size of 100 and proportion of .34 we get

$$100 * .34 = 34 \geq 10 \quad \text{and} \quad 100 * (1 - .34) = 66 \geq 10$$

This assumption is met comfortably, so we'll calculate the Standard Errors of the difference between .52 and .34 and then check it with the z-distribution.

$$\frac{.52 - .34}{\sqrt{\frac{.52(1 - .52)}{100} + \frac{.34(1 - .34)}{100}}} = 2.614 \text{ (rounded)}$$

Looking at Figure 10.3 we see that 2.614 Standard Errors is outside the 95% interval. The .18 difference between .52 and .34 is statistically significant, and we do reject the Null Hypothesis that Flowing Wells and Artesian Wells have equal population proportions.

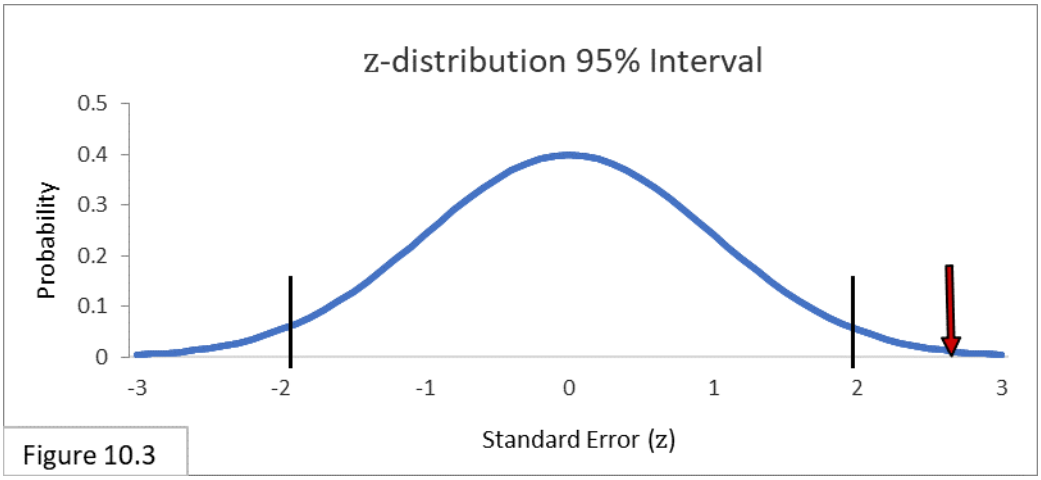


Figure 10.3

Plus, it does seem that 52% and 34% are different enough to claim practical significance—that the difference has meaningful political implications.

Next, let's say that for the following month's survey, in May, additional resources are committed so that larger samples can be gathered. The sample statistic values are the same as two months ago, in March—sample proportions of .52 and .44—but now the sample sizes are 1000.

$$\frac{.52 - .44}{\sqrt{\frac{.52(1-.52)}{1000} + \frac{.44(1-.44)}{1000}}} = 3.592 \text{ (rounded)}$$

Looking at Figure 10.4 we see that 3.592 Standard Errors is outside the 95% interval. (It's off the chart shown here, but the z-distribution itself actually goes from negative infinity to positive infinity.) With sample sizes of 1000, the difference of .08 between .52 and .44 is statistically significant, and we do reject the Null Hypothesis that Flowing Wells and Artesian Wells have equal population proportions. While the .08 difference was not statistically significant in March with sample sizes of 100, it is statistically significant in May with sample sizes of 1000. The additional statistical power due to the larger sample size does that. And while it is always possible that a Type I Error has occurred *whenever* we reject a Null Hypothesis, our experience over the last few months convinces us that this month's difference is most likely real.

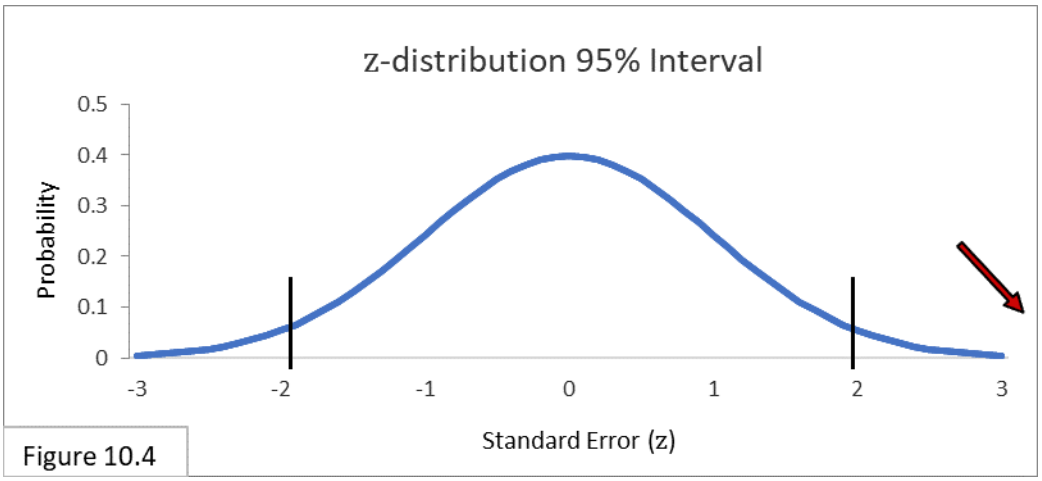


Figure 10.4

It is unclear, however, whether 52% and 44% are different enough to have serious political implications. Perhaps we need to ask some political scientists whether they think these statistically significant survey results have any practical significance.[xi]

11. The Rest of the (Frequentist) Iceberg

At first blush, it seems that we have only explored the tip of the iceberg. But really, you can learn many of the important characteristics of an iceberg by studying the tip of the iceberg: for one, you can learn all about its general composition. Likewise, by studying the analysis of binomials, you can learn the general composition of statistical analysis. No matter what types of statistics you're going to analyze, they are all going to have sampling distributions. And those sampling distributions are all going to have corresponding standardized probability distributions. And when using those distributions, you will always reject the null hypothesis when the statistic value is located far enough out in the tail of the distribution. There will always be statistical and practical significance assessments to perform, and there will always be Type I and Type II Errors to worry about. And there will always be assumptions that must be met.

In addition to the z-distribution, there are three other types of standardized probability distributions that form the "big four" of Frequentist statistics: the t-distributions, the χ^2-distributions, and the F-distributions. Each of these three distributions is really a family of distributions, with each of the three families comprised of many instances of the same type of distribution. All four are highlighted below. To streamline the highlights, I've relegated details to a number of endnotes. (I hope to produce a sequel—tentatively titled *Statistical Analysis Illustrated: Variance Everywhere*—to cover these topics in depth.)

For sample statistics related to *binomials and ranks*[xii] we use the z-distribution, illustrated in Figure 11.1.

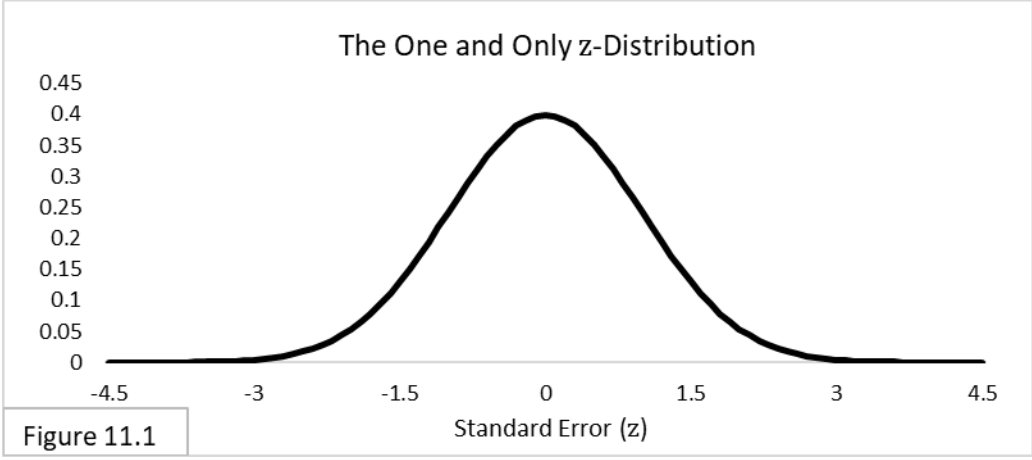


Figure 11.1

For sample statistics related to *averages*—such as the average height of a sample of people, or the difference in average height between two samples of people—we use t-distributions, illustrated in Figure 11.2. The t-distribution incorporates additional uncertainty into the equation. When there is no additional uncertainty, it is the same as the z-distribution. That's the solid-line curve. When there is additional uncertainty, the t-distribution is more spread out, reflecting the additional uncertainty. The more additional uncertainty there is, the more spread out the t-distribution is and the wider the 95% confidence interval will be. In Figure 11.2, the dotted-line curve is the t-distribution with the most additional uncertainty of the three.[xiii]

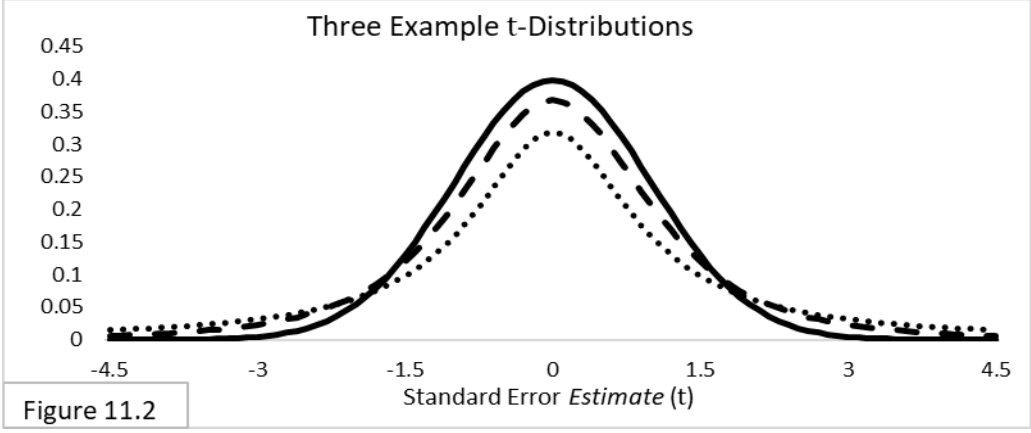


Figure 11.2

For sample statistics related to *variances*—such as the variance (variety) of heights within a sample of people—we use χ^2-distributions, called Chi-squared distributions and illustrated in Figure 11.3. As you can see by the horizontal axis, χ^2-statistic values are always greater than or equal to zero.[xiv]

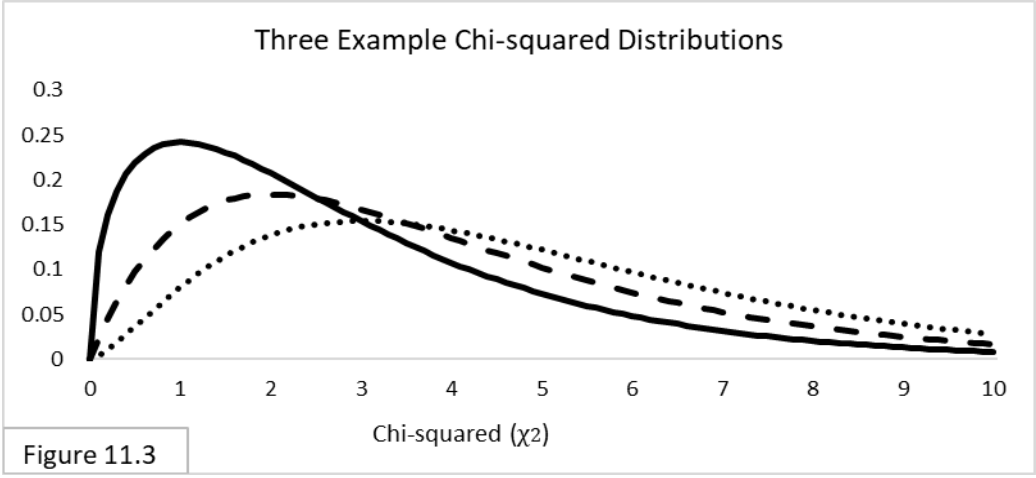


Figure 11.3

For *comparing two sample variances*—such as comparing the variety of heights between two samples of people—we use their ratio rather than their arithmetic difference, and we use F-distributions, illustrated in Figure 11.4. F-statistic values are also always greater than or equal to zero.[xv]

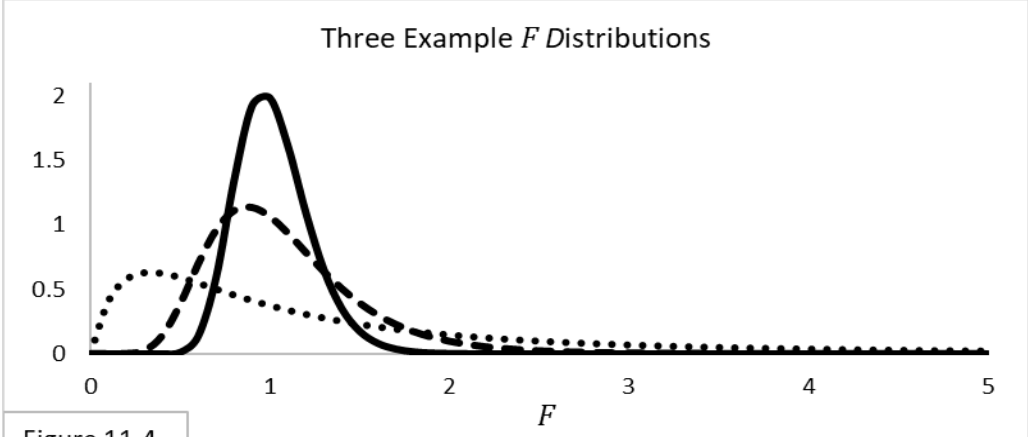


Figure 11.4

These four distributions are by far the most widely used standardized probability distributions in Frequentist statistics. For all the types of data you are likely to analyze, there are statistics that can be used to summarize that data, and those statistics can be analyzed using a standardized probability distribution.[xvi] When using any of these distributions, you'll typically reject the null hypothesis when the statistic value is located far enough out in the tail. And in all cases, assessments need to be made regarding assumptions, Type I and Type II Error probabilities, and statistical and practical significance.

Frequentist statistics, which we've focused on, has been the dominant statistical analysis methodology for over a century. However, a different methodology, with a long history itself, continues to challenge that dominance: Bayesian statistics.

12. Bayesian Analysis

Two major approaches to statistical analysis are the Frequentist approach and the Bayesian approach. This book focuses on the Frequentist approach, but I couldn't leave without introducing you to the Bayesian way. *Bayesian statistics* provides a special method for calculating probability estimates, for choosing between hypotheses, and for learning about population statistic values. To explore its basic workings, we'll start with a statistical scenario involving medical testing and diagnosis.

Basics of Bayesian Analysis

Table 12.1 shows the various possible conditions and outcomes of a diagnostic test for a fictional disease, Krobze. The rows show the unknown truth of not having or having Krobze, and the columns show the known negative or positive test results. The four shaded "intersection" cells show the familiar breakdown into veridical results (true negative test results and true positive test results) and misleading results (Type I Error and Type II Error, which are false positive test results and false negative test results). As we'll see, and unlike Frequentist statistics, Bayesian statistics makes central use of assumptions about population statistic values in order to calculate probability estimates for the truth of hypotheses (something that's considered *unthinkable* in Frequentist statistics!).

Table 12.1 Medical Diagnosis Scenario Structure

Test Result / Unknown Truth	Test Negative	Test Positive
Don't have Krobze	True Negative	False Positive (Type I Error)
Do have Krobze	False Negative (Type II Error)	True Positive

In particular, Bayesian analysis can be used to estimate *conditional* (if…then…) *probabilities* like:

1) If you test negative, then what is the probability that you really don't have Krobze?

2) If you test positive, then what is the probability that you really do have Krobze?

The first involves the Test Negative column: we need to calculate the probability of a true negative divided by the sum of the probabilities of true and false negatives. The second involves the Test Positive column: we need to calculate the probability of a true positive divided by the sum of the probabilities of true and false positives.

In order to calculate these, we need to know, or estimate, the probability of having the disease in general–that is, how prevalent Krobze is in the population. Notice that this is really saying that we need to know, or estimate, the population proportion, and that we will incorporate it into the heart of the analysis (something that is not done in Frequentist statistics). We also need to know how reliable the diagnostic test is.

Listed below is this required information, which is then used to fill in Table 12.2.

1 out of 100 have Krobze, for a probability of **0.01** of having Krobze and **0.99** of not.

For the diagnostic test reliability, we are given the following information:

For people who don't have Krobze, 90% (**0.9**) test negative (true negative) and 10% (**0.1**) test positive (false positive)

For people who do have Krobze, 80% (**0.8**) test positive (true positive) and 20% (**0.2**) test negative (false negative)

Notice that in addition to using the above information to fill in Table 12.2—shown in bold type—another row has been added to the bottom of the Table for column sums.

Table 12.2 Probabilities after a Single Test

Test Result / Unknown Truth	Test Negative	Test Positive
Don't have Krobze (0.99 of the population)	True Negative **0.99*0.9**=0.891	False Positive **0.99*0.1**=0.099
Do have Krobze (0.01 of the population)	False Negative **0.01*0.2**=0.002	True Positive **0.01*0.8**=0.008
Column Probability Sums	0.893 (89.3% will test negative)	0.107 (10.7% will test positive)

Let's use Table 12.2 to answer the questions posed above.

1) If you *test negative*, what is the probability that you *don't* have Krobze?
This asks for the true negative rate. In jargon it's called the *specificity* of the test.
We use the Test Negative column.
True negative/true and false negatives=0.891/0.893=0.99776 which is nearly 100%.

2) If you *test positive*, what is the probability that you *do* have Krobze?
This asks for the true positive rate. In jargon it's called the *sensitivity* of the test.
We use the Test Positive column.
True positive/true and false positives=0.008/0.107=0.074766 which is between 7% and 8%. It's this low because Krobze is fairly rare, with only a 1% prevalence in the population, so false positives dominate the true positives.

More key Bayesian terminology:
The two Don't & Do have Krobze cells show *prior probabilities*.
The four shaded True False Negative Positive cells show *joint probabilities*.
The two Column Probability Sums cells show *marginal probabilities*.
The two calculated solutions 0.99776 and 0.074766 are *posterior probabilities*.

Many people feel that all of this makes better sense when we use frequencies rather than probabilities, so let's do it that way too.

Consider 1,000 random people. Below are the frequencies we expect, which are then used to fill in Table 12.3.

990 won't have Krobze (99 out of 100 don't have it) and of those,
    891 (90% of 990) test negative (true negative)
    99 (10% of 990) test positive (false positive)
10 will have Krobze (1 out of 100 have it), and of those
    8 (80% of 10) test positive (true positive)
    2 (20% of 10) test negative (false negative)

Table 12.3 Using Frequencies Instead of Probabilities

Test Result / Unknown Truth	Test Negative	Test Positive
Don't have Krobze (990 out of 1,000)	True Negative 990*0.9=891	False Positive 990*0.1=99
Do have Krobze (10 out of 1,000)	False Negative 10*0.2=2	True Positive 10*0.8=8
Column Frequency Sums	893 (893 will test negative)	107 (107 will test positive)

1) If you *test negative*, how likely is it that you *don't* have Krobze?

891/893=0.99776; same as above.

2) If you *test positive*, how likely is it that you *do* have Krobze?

8/107=0.074766; same as above.

The Bayesian method is especially useful because it can be used successively to update probability estimates. Let's say you tested positive the first time and want to have another type of test (with the same diagnostic reliability) performed. Table 12.4 shows the conditions and outcomes for the second test. The posterior probability of 0.074766 based on your first positive test now becomes the prior probability for the second test.

Table 12.4 Second Test Following the First Positive Test Result

Second Test Result / Unknown Truth (Given first positive test)	Test Negative	Test Positive
Don't have Krobze 1-0.074766=0.925234	True Negative 0.925234*0.9=0.832711	False Positive 0.925234*0.1=0.092523
Do have Krobze 0.074766	False Negative 0.074766*0.2=0.014953	True Positive 0.074766*0.8=0.059813
Column Probability Sums	0.847664	0.152336

Say that you test negative the second time. The probability that you *don't* have Krobze given that the second test is negative is:

0.832711/0.847664=0.98236; about 98%

Instead, say that you again test positive. The probability that you *do* have Krobze following the second positive test result is:

0.059813/0.152336=0.392637; about 39%

Keep in mind that Krobze is fairly rare (1 in a 100). So, false positive test results tend to dominate the Test Positive column: that's largely why we got 7% after the first positive test and 39% after the second positive test. Test reliability matters too, as we'll see next.

Let's say that instead, the doctor orders a much more reliable (and much more expensive) test the second time. It detects both true negatives and true positives 99% of the time. Table 12.5 covers this situation.

Table 12.5 Second Super-Test Following the First Positive Test Result

Second Super Test Result / Unknown Truth (Given first positive test)	Test Negative	Test Positive
Don't have Krobze 1-0.074766=0.925234	True Negative 0.925234*0.99=0.915982	False Positive 0.925234*0.01=0.009252
Do have Krobze 0.074766	False Negative 0.074766*0.01=0.000748	True Positive 0.074766*0.99=0.074018
Column Probability Sums	0.916729	0.083271

Say that you test negative the second time with the super-test. The probability that you *don't* have Krobze given the first positive test result and the second negative super-test result is:

0.915982/0.916729=0.999184; nearly 100%

Finally, say that you test positive the second time with the super-test. The probability that you *do* have Krobze given the first positive test and the second positive super-test is:

0.074018/0.083271=0.888888; about 89%

Bayes' Formula

So far, I've used a table format to illustrate the application of Bayes' Formula because I have found that people have an easier time following along than when I use the official formula. That said, I would be remiss if I didn't show you the formula. There are several renditions of Bayes' formula to choose from. This rendition aligns best with the table format we've been using.

$$\frac{\mathrm{P(A)} * \mathrm{P(E|A)}}{\mathrm{P(A)} * \mathrm{P(E|A)} \ + \ \mathrm{P(B)} * \mathrm{P(E|B)}} = \mathrm{P(A|E)}$$

P(A) is the prior; the prevalence of Krobze in the population (.01)
P(B) is the probability of not having Krobze in the population (1-.01=.99)
E is the evidence of a positive test result (corresponding to the positive test column in the table)
| is the symbol for "given that"
P(E|A) is the probability of testing positive given that you do have Krobze (0.8)
P(E|B) is the probability of testing positive given that you don't have Krobze (0.1)
P(A|E) is what you want to know: the probability you have Krobze given that you test positive

Using again the first Krobze example when you test positive the first time, we get

$$\frac{.01 * .8}{(.01 * .8) \ + (.99 * .1)} = .074766$$

This is the same posterior probability that we got earlier using the table format.

A Note on Priors

With the above analyses we were extremely fortunate to know that, in general, 1 in 100 people have Krobze. That gave us our initial prior probabilities (0.01 and 0.99) which we needed to get started. What if we don't know these initial prior probabilities? One alternative is to use expert opinion, which is somewhat subjective and may be quite incorrect. Another is to use what are called *noninformative priors*.

For example, if we have no idea what the initial prior probability of Krobze is, we could use the noninformative priors of equal probabilities for Don't Have Krobze (0.5) and Have Krobze (0.5). As you can imagine, we'll get quite different results using priors of 0.5 and 0.5 instead of 0.99 and 0.01. Determining valid priors is important, and it can be tricky. But luckily, the more data we have to update the prior with, the less important the prior becomes.

Frequentist statistics does not incorporate prior probabilities into the analysis and so it doesn't need to make assumptions about priors.

A Note on Testing Competing Hypotheses

Let's say we are planning to survey Flowing Wells regarding a new public health policy and that we have two competing hypotheses. For a change of pace, we'll use rational numbers (fractions) rather than decimal numbers. Our two competing hypotheses are H_1 that the population is 1/3 in favor and H_2 that the population is 2/3 in favor. For the prior probabilities we'll assume equal probabilities of 1/2 that each of the hypotheses is true. Next, suppose we have a random sample of 10 survey respondents and that 4 are in favor (4/10). Based on this sample data, the Bayesian method updates the 1/2 prior probability for H_1 to the posterior probability of 4/5 and updates the 1/2 prior probability for H_2 to the posterior probability of 1/5. Given these results we'd opt for H_1 over H_2.

Since Frequentist statistics does not incorporate prior probabilities into its analysis, it never derives a probability that a given hypothesis is true (see Chapter 4 *Veridical vs. Misleading Results*).

A Note on Estimating Population Statistic Values

More sophisticated Bayesian analysis involves *entire distributions*. For example, let's say we are trying to estimate a population proportion for an agree-or-disagree opinion survey question. Figure 12.1 shows 1) the prior distribution for the population proportion estimate proposed by an expert, 2) the proportion sampling distribution based on random sample data from the population, and 3) the resulting posterior distribution for the population proportion estimate derived via Bayesian updating of #1 using #2.

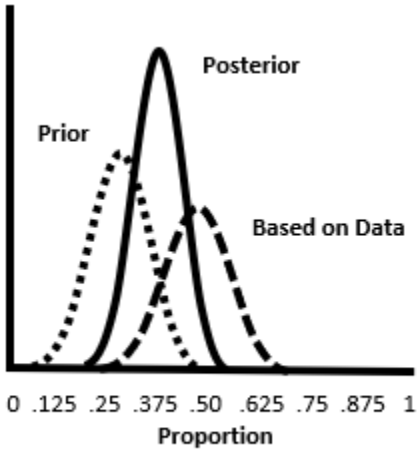


Figure 12.1

From the posterior, we can see that the most likely estimate for the population proportion is .375. We can also determine the interval containing 95% of the posterior's area: from .25 to .50. This is called a Bayesian *credible interval*, which is subtly different from a Frequentist confidence interval. Credible Interval: Given our priors and our sample data, there is a 95% chance that the population proportion falls within the 95% credible interval; valid priors and distributional assumptions need to be incorporated into the analysis. Confidence Interval: 95% of the 95% confidence intervals derived from sample data will contain the population proportion; no priors are needed and the statistical methods themselves embody distributional assumptions (such as assuming that the z-distribution is an appropriate approximation to the binomial distribution).

You can see that Bayesian analysis leads to stronger declarations than Frequentist analysis does, but that the legitimacy of those declarations rests, in part, on the validity of the prior probabilities. Keep in mind, though, that the prior probabilities become less important as more data is used to update the probabilities. Given truly realistic priors *or* noninformative priors plus adequate quantities of data, Bayesian analysis becomes—for those so inclined—an attractive alternative to Frequentist statistics. Bayesian analysis is quite flexible, and it can get extraordinarily complicated. Here, we've taken a peek at the tip of the Bayesian iceberg.

Addendum. The False Discovery Rate

Terminology and Notation (reproduced from the end of Chapter 5): Mathematically, the probability of Type I Error is denoted by the lower-case Greek letter alpha, α. The percentage level for confidence—which we've been referring to a lot—is (1-α)*100%. So, an alpha-level of 0.05 is equivalent to a confidence level of 95%. The probability of Type II Error is denoted by the lower-case Greek letter beta, β. Power is 1-β (*not* made into a percentage). So, for example, a beta level of 0.20 is equivalent to a power level of 0.80.

Imagine that there are 1000 *hypotheses* to be tested and that 100 of the Null Hypotheses are actually false and 900 of the Null Hypotheses are actually true. Let's suppose that these are 1000 separate studies to determine whether certain foods and dietary supplements affect peoples' health, and so each Null Hypothesis states that a certain food or dietary supplement has no effect on health. Assume an alpha-level of 0.05 is used (95% confidence). Also assume that the probability for Type II Error, beta, is 0.20 and so statistical power is 0.80. These are fairly realistic levels, although 0.80 power is probably higher (better) than many studies would have.

With a 0.05 alpha probability for Type I Error, expect 900*0.05=**45** Type I Errors, and expect 900*0.95=**855** correct non-rejections of the Null Hypothesis.

With a 0.20 beta probability for Type II Error, expect 100*0.20=**20** Type II Errors, and expect 100*0.80=**80** correct rejections of the Null Hypothesis.

Therefore, we expect a total of 45+80=**125** rejected Null Hypotheses.

What percentage of the rejected Null Hypotheses do we expect to be erroneously rejected? This is called the *false discovery rate*. Per the above, we expect 45 to be erroneously rejected, and we expect 80 to be correctly rejected. Therefore, we expect 45/(45+80)=**36%** of the rejected Null Hypotheses to be erroneously rejected. Given that rejected Null Hypotheses tend to get all the publicity, you should find this eye-opening. (Headlines for rejected null hypotheses might be statements like "This supplement significantly improves health!" and "This food is a significant health risk!")

We can also lay this out in a Table (A.1) of probabilities. Using the rightmost column to calculate the proportion of studies rejecting the Null Hypothesis that are Type I Errors gives us 0.045/0.125=0.360=36%.

Table A.1 Correct and Incorrect Research Study Findings

	Studies Not Rejecting Null	Studies Rejecting Null
Null Hypothesis true 90% of the time	0.9*0.95 = **0.855** Correct Findings	0.9*0.05 = **0.045** Type I Error
Null Hypothesis false 10% of the time	0.1*0.2 = **0.020** Type II Error	0.1*0.8 = **0.080** Correct Findings
Column Sums	**0.875** 87.5% of studies won't reject Null	**0.125** 12.5% of studies will reject Null

The 36% can be improved upon, as you know, by increasing the sample sizes in the 1000 studies. This will increase power. If we increase power to 0.9, for example, we'll increase the correct rejections from 80 to 90. Ideally, if we can increase sample size sufficiently, we can set our alpha down to 0.01, thus decreasing the 45 to 9, while also maintaining or even increasing power. With alpha of 0.01, beta of 0.10 and thus power of 0.90, we would reduce the false discovery rate to 9/(9+90)=9%. This is much better, but it's also much more expensive to gather the large amounts of additional data that are required.

Lastly, keep in mind that the 36% false discovery rate we came up with is contingent on our assumption that only 100 of the 1000 Null Hypotheses are actually false. If researchers have good theories to guide their choice of hypotheses, then the proportion of Null Hypotheses that are truly false should be higher and the false discovery rate should be lower.

About the Author

Jeffrey E. Kottemann is Professor Emeritus of Information and Decision Science at Salisbury University in the University System of Maryland, USA. He was previously on the faculty of the University of Hawaii, Manoa and the University of Michigan, Ann Arbor. Jeff received his Ph.D. in Systems and Quantitative Methods from The University of Arizona, Tucson. He has published research articles in a variety of disciplines including computer science, decision science, economics, engineering, information science, and psychology, as well as a prior book on statistical analysis entitled *Illuminating Statistical Analysis Using Scenarios and Simulations* published by John Wiley and Sons.

---

[i] Some people find the word “confidence” in “confidence interval” to be somewhat counterintuitive, but that’s been the term since the 1930’s. Some people think they should be called something like "uncertainty intervals." Referring to Figure 1.2, if you get a sample percentage in the 40%-to-60% boundary lines, then you're too uncertain to be able to say that the population percentage is anything other than 50%. Another suggestion is to call them “expectation intervals” since they concern the sample statistic values we expect to get with random sampling. In fairness, the term “confidence interval” does enjoy more intuitive appeal in some other contexts, as we’ll see later on.

[ii] Consider two extreme sample size examples on opposite sides of the continuum: surveying only two people and surveying everyone. 1) When surveying two random people in a 50%-in-favor population, ¼ of the time we’ll get the first and second who are both in favor, ¼ of the time we’ll get the first person in favor and the second not in favor, ¼ of the time we’ll get the first person not in favor and the second in favor, and ¼ of the time we’ll get both people not in favor. So, ¼ of the time we’ll get the extremely far-off sample percentage value of 100% and ¼ of the time we’ll get the extremely far-off sample percentage value of 0%. 2) If we increased our sample size to the cover the entire population of 80,000, all our “sample” percentage values will be exactly equal to the population percentage value. None are off even a small amount. In general, the larger our sample size the closer we expect our sample percentages to be to the population percentage. This is the *law of large numbers.*

[iii] Officially, the Greek lowercase letter pi, π, is used. Sometimes the Roman uppercase letter P is used. I'll use the Roman lowercase letter p.

[iv] The mathematical proof of this is the de Moivre–Laplace Theorem, which is a special case of The Central Limit Theorem.

[v] Likewise, in testing a new drug against an old drug (or a placebo), the working-hypothesis is that the new drug is *not* better than the old drug. The difference between the efficacy of the new drug and the old drug must be outside the 95% confidence interval for the new drug to be deemed more efficacious. Being more conservative in this case means that we'll be less likely to approve an ineffective drug (Type I Error) but more likely to not approve an effective drug (Type II Error). It is in the company's interest to have large samples in order to decrease the Type II Error rate.

[vi] So, if the results are close to the confidence interval borderline, a researcher would justifiably wonder about the possibility of Type II Error and whether the research should be conducted again with a larger sample size.

[vii] There is a class of statistics called *effect size* statistics that have been developed to help address the issue of practical significance. (The term "effect size" reflects the terminology used in experiments where researchers analyze the size of the effects of various treatments.) For a binomial proportion the effect size statistic, denoted g, is simply the sample proportion minus the hypothesized proportion—in other words g is simply how far apart the sample and hypothesized proportion values are. For example, if our actual sample proportion is 0.65 and our hypothesized value is 0.50, then g equals 0.15. Rules-of-thumb Tables—such as the below, intended primarily for the behavioral sciences—are then used to categorize and label the practical significance of a statistical result. Given this Table the practical significance of our g of 0.15 is "medium".

Effect Size, g	Strength of Effect
Near 0.05	small
Near 0.15	medium
Near or above 0.25	large

Keep in mind that *effect sizes don't take real-world context into account*. So, even though g of 0.15 is categorized as "medium" by the generic Rules-of-thumb, in a

particular situation this survey result may well have “large” (or “small”) practical implications.

[viii] Probability is similar to *relative* frequency, although relative frequency is a discrete scale not a continuous scale. For example, if something happens 500 out of a thousand times, the relative frequency is 500/1000=0.5. That corresponds to a probability of 0.5. But relative frequency is a *discrete* scale (1/1000, 2/1000, …, 1000/1000) and probability is a *continuous* scale from 0 to 1.

[ix] Technical research reports often refer to what are called *alpha-levels* and *p-values*. Together they comprise an alternate way of specifying a confidence interval and whether a statistical result is outside the confidence interval. Alpha-levels and p-values both specify probabilities—areas under specific regions of a probability distribution curve such as the z-distribution.

Alpha-levels specify the area under the curve outside the confidence interval. So, using a 0.05 alpha-level is equivalent to using a 95% confidence interval, and using a 0.01 alpha-level is equivalent to using a 99% confidence interval. Since the probabilities in these cases involve the areas outside the confidence interval on both sides, they are called *two-tail* probabilities (each side of the probability distribution is thought of as a tail).

P-values also refer to areas under a probability distribution curve and they indicate how far out a statistical result is on the tail of a probability distribution. Figure 8.5 illustrates. When a statistical result has a two-tail p-value less than the alpha-level of .05, that means it lies outside the 95% interval, and the result is said to be *statistically significant* at the 0.05 alpha-level (or, equivalently, statistically significant at the 95% confidence level).

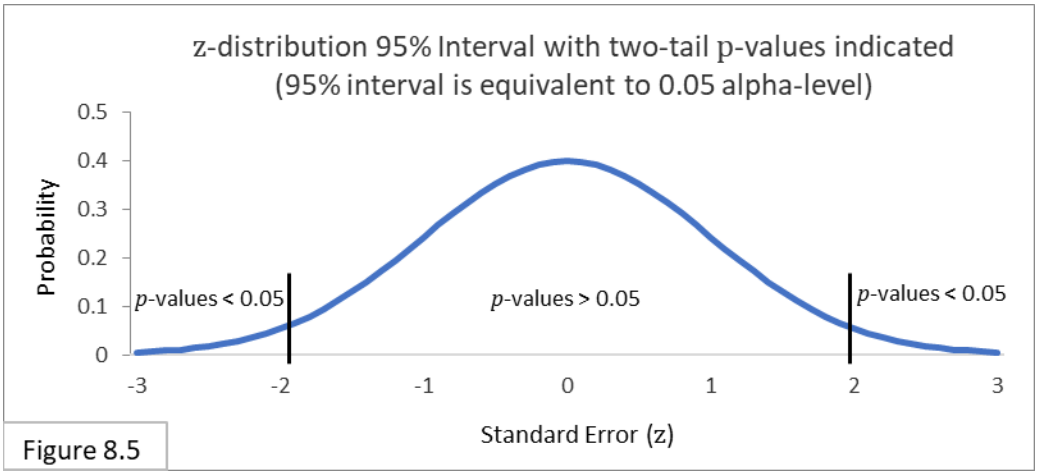


Figure 8.5

The farther out a statistical result is on the tail, the smaller the p-value area becomes: When a statistical result has a p-value less than the alpha-level of .01, that means it lies outside the 99% interval, and the result is said to be statistically significant at the 0.01 alpha-level (or, equivalently, statistically significant at the 99% confidence level).

The use of p-values has come under substantial criticism because they can be misleading and are frequently misinterpreted. The use of confidence intervals is often suggested as a better alternative. That's why we are using confidence intervals and not p-values. Nonetheless, I thought I should at least describe what they are since they are still commonly used in reporting research results. *Simply remember that p-value<0.05 means a statistical result is outside the 95% confidence interval, and p-value<0.01 means a statistical result is outside the 99% confidence interval.*

[x] With binomials, the various possible combinations of outcomes can be enumerated mathematically. Let's use coin flipping for an example, and we'll flip the coin 10 times. Since each coin flip has two possible outcomes and we are considering ten separate outcomes together, there are a total of 2^{10}=1024 unique possible patterns (called permutations) of heads and tails with 10 flips of a coin. Of these, there is only one with 0 heads and only one with 10 heads. These are the least likely outcomes.

TTTTTTTTTT HHHHHHHHHH

So, no heads will occur 1/1024 of the time, as will all heads.

There are ten combinations with 1 head, and ten combinations with 9 heads:

HTTTTTTTTT	THHHHHHHHH
THTTTTTTTT	HTHHHHHHHH
TTHTTTTTTT	HHTHHHHHHH
TTTHTTTTTT	HHHTHHHHHH
TTTTHTTTTT	HHHHTHHHHH
TTTTTHTTTT	HHHHHTHHHH
TTTTTTHTTT	HHHHHHTHHH
TTTTTTTHTT	HHHHHHHTHH
TTTTTTTTHT	HHHHHHHHTH
TTTTTTTTTH	HHHHHHHHHT

So, one head will occur 10/1024 of the time, as will nine heads.

The formula for the number of combinations is n!/[h!*(n-h)!] where n is the number of flips, h is the number of heads you're interested in, and ! is the factorial operation (for example: 5!=5*4*3*2*1=120).

Going further using the formula, there are 45 combinations with 2 or 8 heads, so they will each occur 45/1024 of the time. There are 120 combinations with 3 or 7 heads, so they will each occur 120/1024 of the time. There are 210 combinations with 4 or 6 heads, so each will occur 210/1024 of the time. Finally, there are 252 combinations with 5 heads, which is the most likely outcome at 252/1024 and therefore the most frequently expected outcome. With this mathematical approach we can calculate exact probabilities for all the possible outcomes, hence the term *exact methods*. With large sample sizes this method becomes computationally intensive.

Also, using simulation to create sampling distributions is always a viable alternative. Figure 9.1 (reproduced below), for example, was generated using simulation and can be used to analyze results with sample size of two. You can tell that one-quarter of the expected results are 0, one-half are 0.5 and one-quarter are 1. We can see that there is no way to make a 95% confidence interval. Figure 9.2 (reproduced below) was also generated using simulation and shows that the sampling distribution is asymmetric (not normal). Nonetheless, we can make an approximate (and asymmetrical) 95% confidence interval from 0.8 to 0.967.

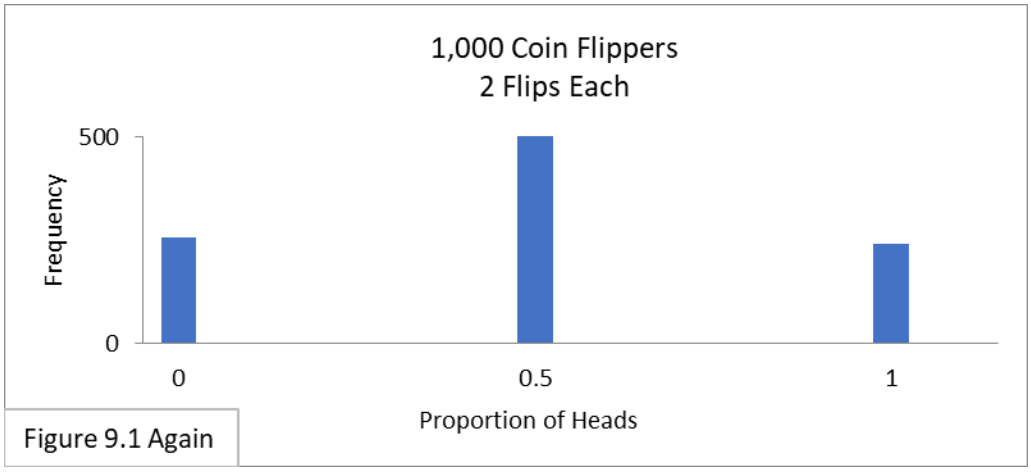


Figure 9.1 Again

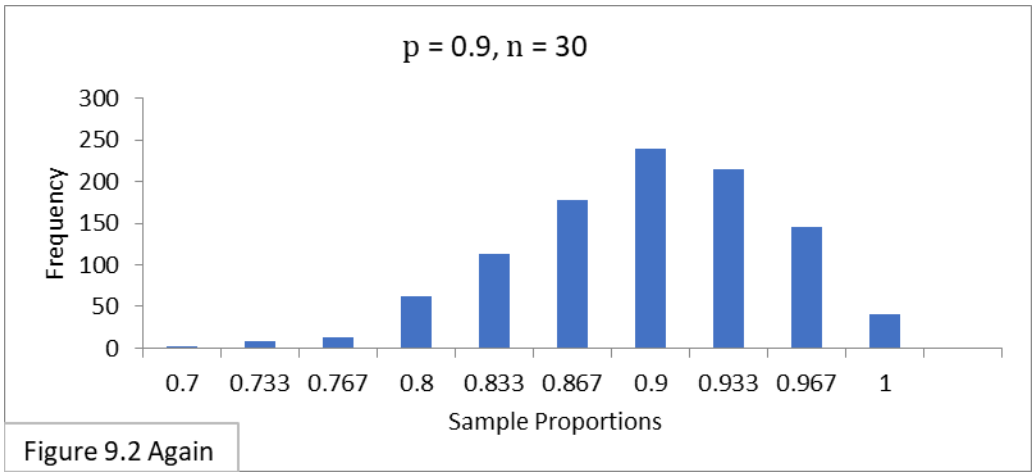


Figure 9.2 Again

[xi] The reason I say that we need to consult some political scientists is because specific domain expertise is often needed to assess practical significance. Nonetheless, as noted in a Chapter 7 endnote, general methods for quantifying practical significance have been developed and go by the name of *effect sizes*. (The term "effect size" reflects the terminology used in experiments where researchers analyze the size of the effects of various treatments.) Below I'll calculate and interpret the effect size, denoted h, for the sample proportion difference between 0.52 and 0.44.

First, each of the two sample proportions is rescaled using the below formula. (The rescaling adjusts the effect sizes as sample proportions approach 0 or 1.)

$$2 * arcsine\left(\sqrt{p}\right)$$

For 0.52, this gives 1.61, and for 0.44 this gives 1.45. Then we take the difference, which gives 0.52-0.44=0.16. Finally, the below Table shows Rules-of-thumb

(intended primarily for the behavioral sciences) to categorize and label the strength of the effect size, h. Our h of 0.16 is categorized as a "small" effect size.

Effect Size, h	Strength of Effect
Near 0.2	small
Near 0.5	medium
Near or above 0.8	large

Keep in mind that *effect sizes don't take real-world context into account*. So, even though the quantified effect size of 0.52 versus 0.44 is categorized as "small" using the generic Rules-of-thumb, a political scientist might well tell us that this statistically significant difference in this particular context actually represents a meaningful difference of opinion with "large" political implications.

[xii] We've looked at many analyses involving binomial data and statistics, but none with rank data and statistics. As an example, students' class rank is rank data, with the valedictorian = 1, the salutatorian = 2, etc. We can also assign rank numbers ourselves and use them to perform statistical analysis. For example, let's say we have a sample of 100 people, 50 males and 50 females, and their heights. We can sort the 100 by height and assign rank numbers 1 to 100 based on the sorted order. After that, we can separately add up the rank numbers for the males and for the females. Based on those rank sums we can see whether males or females rank higher with respect to height. In fact, we can use a statistical method called the Mann-Whitney test that uses the z-distribution to assess whether rank sums are statistically significant.

[xiii] For example, one common type of "average" is the *arithmetic mean* or *mean* for short. Below is the formula for the *sample mean*. Its symbol is an x with a bar on top. I imagine everyone has calculated a mean before: sum (Σ) all the numbers (x) and divide by the number of numbers (n).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

If we have, for example, a sample of 100 people and their heights, we can calculate the sample mean for heights. The additional uncertainty comes in when we try to calculate the Standard Error of the sample mean. There is no way to calculate the true Standard Error unless we know the true value of the *population* variance, which we almost never do. So, we're stuck using the sample variance (see next endnote)

as an estimate for the population variance. This makes our calculated Standard Error an estimate too. This is the source of the additional uncertainty. Notice that the horizontal axis for the t-distribution of Figure 11.2 is labelled Standard Error *estimate*.

The amount of additional uncertainty is a function of sample size. Since we expect larger sample sizes to improve our sample variance estimates (law of large numbers), they also improve our Standard Error estimates, thereby lessening the additional uncertainty. Thus, the selection of which specific t-distribution to use is a function of sample size. (Recall that with binomial variables, which only have two possible values, we can calculate the true variance via p*(1-p) and use it to calculate the true Standard Error. Notice that the horizontal axis for the z-distribution in Figure 11.1 is labelled Standard Error, without the *estimate* qualifier. This is also true when calculating the variance and standard error for rank sums (see prior endnote).)

[xiv] For example, below is the formula for the *sample variance*. Its symbol is s^2. It sums (Σ) squared differences as a measure of how spread out the data is around the sample mean (see previous endnote). In other words, it's a measure of how much the data values *vary* from the mean. Because of the squaring, variances are always greater than or equal to zero.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$ where $\bar{x}$ is the sample mean

If we have, for example, a sample of 100 people and their heights, we can calculate the sample mean for heights and then calculate the sample variance for heights.

For binomials, which only have two possible values, the variance formula can be simplified to

$$p * (1-p)$$

which is also always positive.

One widely used application of Chi-squared distributions is to analyze multinomial variables. Whereas a binomial variable has only two possible values, a multinomial variable has two or more values (e.g., a political affiliation variable may have the three possible values of Democrat, Republican, and Independent). The formula for

the Chi-squared statistic shown below is a variance formula because it calculates the squared differences between the counts we observed in our sample and the counts we expect under our Null Hypothesis. In other words, it's a measure of how much the observed counts *vary* from the hypothesized counts. And since it is a variance formula, we use the Chi-squared distribution to assess it. The farther apart the observed and expected values are, the larger the Chi-squared statistic value becomes, and the farther out in the tail of the distribution it is located.

$$\sum \frac{(observed - expected)^2}{expected}$$ for all the categories =the χ^2 (Chi-squared) statistic

[xv] Let's say we want to know whether there is a difference in the *variance* of opinions between Flowing Wells and Artesian Wells. The Null Hypothesis is that there is no difference. We are going to sample 100 from each community, calculate the sample variances of each, and then divide one sample variance by the other to get the F value.

$$F = \frac{s_1^2}{s_2^2}$$

The F-distribution used for sample sizes of 100 and 100 is shown below in Figure 11.5. The 95% confidence for the F value is from 0.67 to 1.5. If our F value is outside this interval, we will reject the Null Hypothesis.

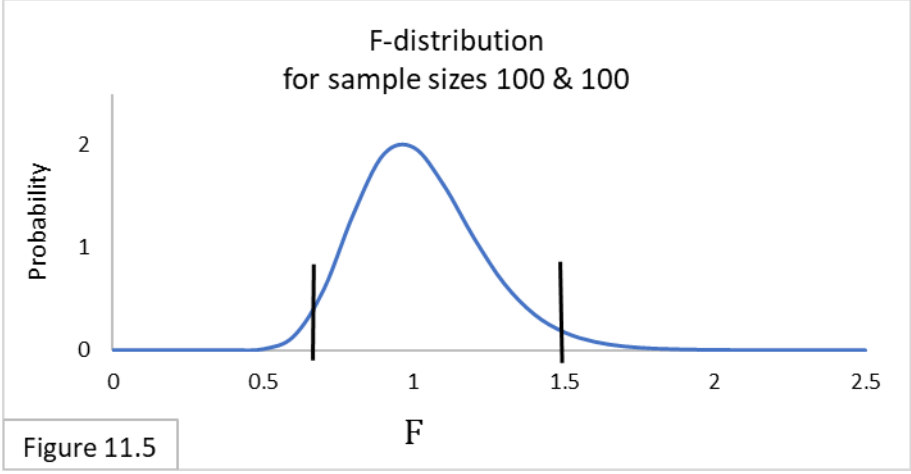


Figure 11.5

One widely used application of F-distributions—one that might initially seem counterintuitive—is to compare multiple sample *means* simultaneously via a method called Analysis of Variance (ANOVA). Let's explore how it works—hang on

to your hats. Imagine a survey that asks for respondents' political affiliation (Democrat, Republican, or Independent) as well as their opinion on a 1-to-7 scale. We want to know if the sample mean opinion for each of the three political affiliations indicate that there are differences of opinion across political affiliations in the population. The Null Hypothesis is that all three groups have the same population mean.

First, we administer the survey to Flowing Wells residents and calculate the sample mean and sample variance for each of the three political affiliations. Let's say the sample means are 3, 4, and 5, and the sample variances are 0.9, 1.0, and 1.1. ANOVA first calculates an estimate for the overall population variance by averaging the sample variances 0.9, 1.0, and 1.1. That gives 1.0.

Next, recall that the larger a population variance is the wider the sampling distribution becomes (Chapter 2 *Sampling Distribution Dynamics*). So, ANOVA determines how large the population variance would need to be for all the sample means—3, 4, and 5—to be contained within one and the same sampling distribution. Let's say that works out to be 30.0.

Now we have the required population variance of 30 and the estimate for the actual population variance of 1. We look up the F value of 30/1 on the F-distribution and find that it is way, way out in the tail. That suggests that these sample means didn't come from the same population. So, we reject the Null Hypothesis that Democrats, Republicans, and Independents in Flowing Wells have the same average opinion.

[xvi] Examples: binomials (z), rank sums (z), means (t or F), correlations(t), variances (χ^2), multinomials (χ^2), regression coefficients (t), regression models (F).