Chapter 3

Estimation of p

3.1 Point and Interval Estimates of p

Suppose that we have Bernoulli Trials (BT). So far, in every example I have told you the (numerical) value of p. In science, usually the value of p is unknown to the researcher. In such cases, scientists and statisticians use data from the BT to **estimate** the value of p. Note that the word *estimate* is a technical term that has a precise definition in this course. I don't particularly like the choice of the word *estimate* for what we do, but I am not the tsar of the Statistics world!

It will be very convenient for your learning if we distinguish between two creatures. First, is **Nature**, who knows everything and, in particular, knows the value of p. Second is the researcher who is ignorant of the value of p.

Here is the idea. A researcher plans to observe n BT, but does not know the value of p. After the BT have been observed the researcher will use the information obtained to make a statement about what p might be.

After observing the BT, the researcher counts the number of successes, x, in the n BT. We define $\hat{p} = x/n$, the proportion of successes in the sample, to be the **point estimate** of p.

For example, if I observe $n = 20$ BT and count $x = 13$ successes, then my point estimate of p is $\hat{p} = 13/20 = 0.65$.

It is trivially easy to calculate $\hat{p} = x/n$; thus, based on your experiences in previous math courses, you might expect that we will move along to the next topic. But we won't.

What we do in a Statistics course is *evaluate the behavior* of our procedure. What does this mean? Statisticians evaluate procedures by seeing how they perform *in the long run*.

We say that the point estimate $\hat{p}$ is **correct** if, and only if, $\hat{p} = p$. Obviously, any honest researcher wants the point estimate to be correct. Let's go back to the example of a researcher who observes 13 successes in 20 BT and calculates $\hat{p} = 13/20 = 0.65$.

The researcher schedules a press conference and the following exchange is recorded.

- Researcher: I know that all Americans are curious about the value of p. I am here today to announce that based on my incredible effort, wisdom and brilliance, I estimate p to be 0.65.

- Reporter: Great, but what is the actual value of p? Are you saying that $p = 0.65$?

- Researcher: Well, I don't actually know what p is, but I certainly hope that it equals 0.65. As I have stated many times, nobody is better than I at obtaining correct point estimates.
- Reporter: Granted, but is anybody worse than you at obtaining correct point estimates?
- Researcher: (Mumbling) Well, no. You see, the problem is that only Nature knows the actual value of p. No mere researcher will ever know it.
- Reporter: Then why are we here?

Before we follow the reporter's suggestion and give up, let's see what we can learn.

Let's bring Nature into the analysis. Suppose that Nature knows that $p = 0.75$. Well, Nature knows that the researcher in the above press conference has an incorrect point estimate. But let's proceed beyond that one example.

Consider a researcher who decides to observe $n = 20$ BT and use them to estimate p. What will happen?

Well, we don't know what will happen. The researcher *might* observe $x = 15$ successes, giving $\hat{p} = 15/20 = 0.75$ which would be a correct point estimate. Sadly, of course, the researcher would not know it is correct; only Nature would.

Given what we were doing in Chapters 1 and 2, it occurs to us to calculate a probability. After all, we use probabilities to quantify uncertainty.

So, before the researcher observes the 20 BT, Nature decides to calculate the probability that the point estimate will be correct. This probability is, of course,

$$P(X = 15) = \frac{20!}{15!5!}(0.75)^{15}(0.25)^{5},$$

which I find, with the help of the binomial website, to be 0.2023. There are two rather obvious undesirable features to this answer.

1. Only Nature knows whether the point estimate is correct; indeed, before the data are collected, only Nature can calculate the probability the point estimate will be correct.
2. The probability that the point estimate will be correct is disappointingly small.

(And note that for most values of p, it is impossible for the point estimate to be correct. For one of countless possible examples, suppose that $n = 20$ as in the current discussion and $p = 0.43$. It is impossible to obtain $\hat{p} = 0.43$.)

As we shall see repeatedly in this course, what often happens is that by collecting more data our procedure becomes 'better' in some way. Thus, suppose that the researcher plans to observe $n = 100$ BT, with p still equal to 0.75. The probability that the point estimate will be correct is,

$$P(X = 75) = \frac{100!}{75!25!}(0.75)^{75}(0.25)^{25},$$

which I find, with the help of the website, to be 0.0918. This is very upsetting! More data makes the probability of a correct point estimate smaller, not larger.

The difficulty lies in our desire to have $\hat{p}$ be *exactly* correct. Close is good too. In fact, statisticians like to say,

Close counts in horse shoes, hand grenades and estimation.

But what do I mean by close? Well, for an example to move us along, suppose we decide that if $\hat{p}$ is within 0.05 of p then it is *close enough* for us to be happy. Revisiting the two computations above, we see that for $n = 20$ and $p = 0.75$, close enough means $(14 \leq X \leq 16)$. The probability of this happening, again with the help of the website, is 0.5606. For $n = 100$ close enough means $(70 \leq X \leq 80)$. The probability of this happening is 0.7967. As a final example, for $n = 1000$, close enough means $(700 \leq X \leq 800)$. The probability of this happening is 0.9998, a virtual certainty to a statistician.

Here is another way to view my 'close enough' argument above. Instead of estimating p by the single number (point) $\hat{p}$ we use an **interval estimate**, in this example the closed interval is $\hat{p} \pm 0.05$. As you may have learned in a math class, a closed interval is an interval the includes its endpoints. In this class, all interval estimates are closed intervals. Analogous to our earlier definition, we say that the interval estimate is correct if, and only if, the interval contains p. Thus, saying that $\hat{p}$ is within 0.05 of p (my working definition of close enough in the example above) is equivalent to saying that p is in the interval estimate; i.e. the interval estimate is correct.

Henceforth, I will not talk about $\hat{p}$ being close enough to p; I will talk about whether an interval estimate is correct. Let's look at the example above again with this new perspective.

For the value $p = 0.75$ I studied the performance of the interval estimate $\hat{p} \pm 0.05$ for three possible values of n: 20, 100 and 1000. I found that as n becomes larger, the probability that the researcher would obtain a correct interval estimate also becomes larger.

My example above—the interval estimate $\hat{p} \pm 0.05$—is called a fixed-width interval estimate because the researcher decides in advance to have an interval estimate with a width (the distance between the upper and lower boundaries of the estimate) of 0.10 (or, as statisticians prefer to say, a half-width of 0.05). Fixed-width interval estimates are *not very popular* because of the following feature.

Let us return to the example of $n = 100$ BT with a fixed-width interval estimate of $\hat{p} \pm 0.05$. As I stated above, if $p = 0.75$ then the probability that the interval estimate will be correct is 0.7967. As we shall now see, this probability changes if p changes.

In particular, suppose that $p = 0.95$. Then, the interval estimate $\hat{p} \pm 0.05$ will be correct if, and only if, $(90 \leq X \leq 100)$. Using the website, the probability of this event is 0.9885. If, however, $p = 0.50$, then the interval estimate will be correct, if, and only if, $(45 \leq X \leq 55)$. The probability of this event is 0.7288.

I will summarize these computations in the following table:

p :	0.50	0.75	0.95
Prob. of being correct:	0.7288	0.7967	0.9885

This is a very unsatisfactory result! With a sample of size $n = 100$ the researcher has very little idea as to the probability the interval will be correct because he/she does not know the value of p. There are some general properties (not quite theorems) of fixed-width interval estimates, some of which are hinted at in our table above, namely:

1. The probability of being correct is symmetric in p around 0.50; i.e. the probability of being correct is the same for success rate $1 - p$ as it is for success rate p. Thus, for example, if

$p = 0.05$ then the probability the interval will be correct is 0.9885, the same as it is for $p = 0.95$.

2. Viewed as a function of p, the probability of being correct is minimized at or near $p = 0.50$ and it generally grows larger as we move away from 0.50, towards either 0 or 1.

B/c statisticians are disappointed with fixed-width interval estimates, we will turn our attention to the idea of 'fixed probability of being correct.' The method is described below.

In Chapter 2, we saw pictures of probability histograms that suggest approximating binomial probabilities by using a normal curve. I did examples and you did homework that revealed that in many instances these approximate answers are quite good. In fact, the method works very well provided that p is not too close to 0 and 1 and that n is pretty large. At this time, we will use these admittedly extremely vague expressions 'not too close' and 'pretty large.' We will eventually deal with this issue, but not now.

First, it is bothersome to keep saying 'p is not too close to either 0 or 1.' So we avoid this, as follows.

I will assume that the researcher is a good enough scientist to distinguish between situations in which p is very close to 0 (say 0.01 or smaller) and very close to 1 (say 0.99 or larger). I really cannot imagine that a researcher would be sufficiently ignorant of the subject of study to not be able to do this!

For dichotomous trials the labels of success and failure are arbitrary. In my experience it seems to be human nature to called the preferred outcome, if there is one, the success. For example, if I am shooting free throws, I call a made shot a success and a miss a failure. We will follow this practice *unless* we believe that one of the outcomes is unlikely; that is, either p or q is close to 0. For reasons that will become apparent later, we greatly prefer to have p near 0 than to have p near 1. As a result, henceforth we will obey the following rule:

> **For BT, if one of the possible outcomes has probability of occurring that is believed to be close to 0, we will designate that outcome as the success.**

I have talked about Nature knowing the value of p and the researcher not knowing it. As a mathematician, I think about p having a continuum of possible values between 0 and 1. (Exclusive; remember we are not interested in BT that always give successes or always give failures.) But scientifically, unless p is very close to 0, I am happy with knowing p to, say, three digits of precision. I will give two examples.

Recall that one of the most important applications of BT is when a researcher selects a random sample, with replacement, from a finite population. Consider the 2008 presidential election in Wisconsin. Barack Obama received 1,677,211 votes and John McCain received 1,262,393 votes. In this example, I will ignore votes cast for any other candidates. The population size is $N =$ 1,677,211 + 1,262,393 = 2,939,604. I will designate a vote for Obama as a success, giving $p = 0.571$ and $q = 0.429$.

Notice I say that $p = 0.571$. I conjecture that this imprecision did not bother you. In particular, you did not jump up (figuratively or literally) and say,

No! The value of p is the rational number 1,677,211 divided by 2,939,604, which as a decimal is 0.570556782.... And I apologize for not writing this decimal until it repeats, but this is the size of the display on my calculator and I have other work I must do.

Personally, and this is clearly a value judgment that you don't need to agree with, 0.571 is precise enough for me: Obama received 57.1% of the votes. If I am feeling particularly casual, I would be happy with 0.57. I would never be happy, in an election, to round to one digit, in this case 0.6, because for so many elections rounding to one digit will give 0.5 for each candidate, which is not very helpful! (Of course, sometimes we must focus on total votes, not proportions. For example, in the 2008 Minnesota election for U.S. Senator, Franken beat Coleman by a small number of votes. The last number I heard was that Franken had 312 more votes out of nearly 3 million cast. So yes, to three digits, each man received 50.0% of the votes.)

For p close to 0 (remember, we don't let it be close to 1), usually we want much more precision than simply the nearest 0.001. At the time of this writing, there is a great deal of concern about the severity with which the H1N1 virus will hit the world during 2009–10. Let p be the proportion of, say, Americans who die from it. Now, if p equals one in 3 million, about 100 Americans will die, but if it equals one in 3,000, about 100,000 Americans will die. To the nearest 0.001, both of these p's is 0.000. Clearly, more precision than the nearest 0.001 is needed if p is close to 0.

3.2 The Approximate 95% Confidence Interval for p

In this section we learn about a particular kind of interval estimate of p which is called the **confidence interval** (CI) estimate.

I will first give you the confidence interval formula and then derive it for you. Remember, first and foremost, a confidence interval is a closed *interval*. An interval is determined by its two endpoints, which we will denote by l for lower (smaller) endpoint and u for upper (larger) endpoint. Thus, I need to give you the formulas for l and u. They are:

$$l = \hat{p} - 1.96\sqrt{\hat{p}\hat{q}/n} \text{ and } u = \hat{p} + 1.96\sqrt{\hat{p}\hat{q}/n}.$$

If you note the similarity of these equations and recall the prevalence of laziness in math, you won't be surprised to learn that we usually combine these into one expression for the 95% confidence interval for p:

$$\hat{p} \pm 1.96\sqrt{\hat{p}\hat{q}/n}.$$

We often write this as

$$\hat{p} \pm h,$$

with

$$h = 1.96\sqrt{\hat{p}\hat{q}/n},$$

called the half-width of the 95% CI for p.

I will now provide a brief mathematical justification of our formula.

As discussed in Chapter 2, if $X \sim$ Bin(n, p) then probabilities for Z,

$$Z = \frac{X - np}{\sqrt{npq}},$$

can be well approximated by the standard normal curve (snc), provided n is reasonably large and p is not too close to 0. It turns out that for the goal of interval estimation, the unknown p (and $q = 1 - p$) in the denominator of Z creates a major difficulty. Thanks, however, to an important result of Eugen Slutsky (1925) (called *Slutsky's Theorem*) probabilities for Z',

$$Z' = (X - np)/\sqrt{n\hat{p}\hat{q}},$$

can be well approximated by the snc, provided n is reasonably large, p is not too close to 0 and $0 < \hat{p} < 1$ (we don't want to divide by 0!). Note that Z' is obtained by replacing the unknown p and q in the denominator of Z with the values $\hat{p}$ and $\hat{q}$ which will be known once the data are collected.

Here is the derivation. Suppose that we want to calculate $P(-1.96 \leq Z' \leq 1.96)$. Because of Slutsky's result, we can approximate this by the area under the snc between -1.96 and 1.96. Using the website, you can verify that this area equals 0.95. Next, dividing the numerator and denominator of Z' by n gives

$$Z' = \frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}}.$$

Thus,

$$-1.96 \leq Z' \leq 1.96 \text{ becomes } -1.96 \leq \frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}} \leq 1.96;$$

rearranging terms, this last inequality becomes

$$\hat{p} - 1.96\sqrt{\hat{p}\hat{q}/n} \leq p \leq \hat{p} + 1.96\sqrt{\hat{p}\hat{q}/n}.$$

Examine this last expression. In terms of my definitions at the beginning of this section, it is

$$l \leq p \leq u.$$

Thus, we have shown that, before we collect data, the probability that we will obtain a correct confidence interval estimate is (approximately) 95% and that this is true for all values of p!

This is a great result. The only concern is whether the approximation is good. I will do a few examples to investigate this question.

Suppose that a researcher decides to observe $n = 200$ BT and plans to compute the above 95% confidence interval for p. Is the approximation any good? Well, to answer this question we must bring Nature into the argument. To investigate the quality of the approximation we need not only to specify n, which I have done, but also p. So suppose that $p = 0.40$.

We note that the interval will be correct, if, and only if, it contains $p = 0.40$. That is,

$$\hat{p} - 1.96\sqrt{\hat{p}\hat{q}/200} \leq 0.40 \leq \hat{p} + 1.96\sqrt{\hat{p}\hat{q}/200}.$$

After some algebra, it follows that $l \leq 0.400$ corresponds to $\hat{p} \leq 0.470$ and $u \geq 0.400$ corresponds to $\hat{p} \geq 0.340$. Remembering that $\hat{p} = x/200$, we conclude that the confidence interval will be correct if, and only if, $68 \leq X \leq 94$, where probabilities for X are given by the Bin(200,0.40). With the help of the binomial website, this probability is found to be 0.9466. Not ideal—I would prefer 0.9500—but a reasonably good approximation.

I will repeat the above example for the same $n = 200$, but for a p that is closer to 0, say $p = 0.10$. In this case, by algebra, the confidence interval is correct if, and only if, $15 \leq X \leq 30$. The probability of this event is 0.8976, which is not very close to the desired 95%.

For one last example, suppose that $n = 200$ and $p = 0.01$. The interval is correct if, and only if, $1 \leq X \leq 8$. The probability of this event is 0.8658, which is a really bad approximation to 0.9500.

We have seen that for $n = 200$, if p is close to 0 the 95% in the 95% confidence interval is *not* a very good approximation to the exact probability that the interval will be correct. We will deal with that issue soon, but first I want to generalize the above result from 95% to other confidence levels.

3.2.1 Other Confidence Levels and One-sided Intervals

The 95% confidence level is very popular with statisticians and scientists, but it is not the only possibility. You could choose any level you want, provided that it is above 0% and below 100%. There are six levels that are most popular and we will restrict attention to those in this class. They are: 80%, 90%, 95%, 98%, 99% and 99.73%. Consider again our derivation of the 95% confidence interval. The choice of 95% for level led to 1.96 appearing in the formula, but otherwise had absolutely no impact on the algebra or probability theory used.

Thus, for any other level, we just need to determine what number to use in place of 1.96. For example, for 90% we need to find a positive number, let's call it z, so that the area under the snc between $-z$ and $+z$ is 90%. It can be shown that $z = 1.645$ is the answer. Thus, to summarize: The 90% confidence interval for p is

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Extending these ideas we get the following result. The **(two-sided) confidence interval for** p is given by:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

In this formula, the number z is determined by the desired confidence level, as given in the following table.

Confidence Level	80%	90%	95%	98%	99%	99.73%
z:	1.282	1.645	1.960	2.326	2.576	3.000

Thus, for example,

$$\hat{p} \pm 2.576\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

is the 99% two-sided confidence interval for p. Also,

$$\hat{p} \pm 3\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

is the 99.73% CI for p. We also recognize this as the *pretty certain interval* of Chapter 1. Thus, the pretty certain interval of Chapter 1 was simply the 99.73% CI for r. (Remember, in Chapter 1 the probability of interest was denoted by r.)

You have no doubt noticed that I have added the modifier *two-sided* to the technical term confidence interval. We call our answer the two-sided CI because it has both upper and lower bounds. Sometimes in science we want a one-sided bound on the value of p. This is especially true when p is close to 0.

Below are the two results.

The upper confidence bound for p is given by:

$$\hat{p} + z_1\sqrt{\frac{\hat{p}\hat{q}}{n}},$$

and the lower confidence bound for p is given by:

$$\hat{p} - z_1\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

In these formulas, the number z_1 is determined by the desired confidence level, as given in the following table.

Confidence Level	90%	95%	97.5%	99%	99.5%	99.86%
z_1:	1.282	1.645	1.960	2.326	2.576	3.000

For example, suppose that $n = 200$ and $\hat{p} = 0.250$. The 95% upper confidence bound for p is given by:

$$0.250 + 1.645\sqrt{\frac{0.25(0.75)}{200}} = 0.250 + 1.645(0.0306) = 0.250 + 0.050 = 0.300.$$

In words, I am 95% confident that p is 0.300 or smaller.

3.3 Exact Confidence Intervals

In an earlier example we saw that if $n = 200$ and p is close to 0, our above method, based on the snc approximation, is not very good. It is not very good because the actual true probability that the 95% confidence interval will be correct is substantially smaller than 95%.

There is an exact method available for obtaining a confidence interval for p. It can be obtained by using the website:

http://statpages.org/confint.html

There is a link to this website on our course webpage.

I will illustrate the use of this site.

Suppose that we have $n = 200$ BT and we observe a total of only $x = 10$ successes. This gives $\hat{p} = 0.05$. We do not know the value of p (only Nature does) but such a small value of $\hat{p}$ suggests strongly that p is close to 0, and, hence, that the approximate CI might not be very good.

In fact, the 95%, two-sided snc CI is:

$$0.050 \pm 1.96\sqrt{\frac{0.05(0.95)}{200}} = 0.050 \pm 0.030 = [0.020, 0.080].$$

Next, let's see what we get if we use the website. (It might help you if you go to the website and mimic what I am doing.)

In the section named 'Binomial Confidence Intervals' type in '10' for Numerator (x) and '200' for Denominator (N). (Aside: I don't know why they use N; every sensible person uses n :-].) Hit the compute button and the program produces the exact confidence interval, in this case:

$$0.0242 \text{ to } 0.0900.$$

The exact site can also be used for one-sided CI's. I will illustrate this technique for our data above, $n = 200$ and $x = 10$, and the one-sided 95% upper confidence bound.

- Scroll down to the section titled 'Setting Confidence Levels.'
- I want the 95% upper confidence bound for p, so I type 95 in the 'Confidence Level' box. (Be careful not to type 0.95.)
- I type 5 in the '% Area in Upper Tail' box and 0 in the '% Area in Lower Tail' box.
- Scroll back up and type in x and n as for the two-sided case.
- Remaining in the upper section, click 'Compute.' The answer I get is

$$0.0000 \text{ to } 0.0833.$$

Thus, I am 95% confident that p is 0.0833 or smaller.

I now turn to two technical questions.

First, why is this called *exact*? Well, because there is no approximation involved. Let me explain. Well, only a little bit. The website uses the binomial distribution, not the approximating snc, to obtain its answers. Sadly, the technique is beyond the scope of this course. (It really is quite messy and no fun at all, not even for a statistician!) The key idea is the following. For ease of exposition, let's focus on the two-sided 95% CI. The website method has the following property.

> For every value of p between 0 and 1, the website answer has the property that the probability that the CI will be a correct interval is 95% or larger.

If you want, say, 90%, instead of 95%, the above is true with the number 90%. It is also true for any one-sided CI (upper bound or lower bound).

To make sure this is clear: If you want, say, 95% confidence:

> You can use the website which guarantees we have a 95% chance or more of getting a correct interval regardless of the value of p

OR

> You can use the snc which guarantees nothing; it gives approximately 95% and sometimes the approximation is bad.

The second question now is rather obvious: Why would one *ever* use the snc approximation?

One advantage of the snc is that it actually makes sense in that we can see how it relates to the shape of the binomial distribution. By contrast, the website answer is totally mysterious.

Next, for really large studies the two methods give about the same answer. For example, if $n = 2000$ and $x = 1000$ it can be shown that the exact 95% CI is 0.4778 to 0.5222, while the snc answer is 0.4781 to 0.5219. If we round these answers to the third digit after the decimal, we get $[0.478, 0.522]$ for both.

The exact answer involves some pretty serious computations, but the snc approximate answer can be obtained easily with a hand calculator.

Finally, as a statistician, I feel that I understand pretty well the strengths and weaknesses of using the snc method. I don't know who wrote the program for the website that gives exact CI's and although it seems ok to me, I don't really *know* that. I do not recommend you believe everything you find on a website. (Nor should you automatically believe everything anyone tells you, including me.)