Point and Interval Estimates

Suppose we want to estimate a parameter, such as \( p \) or \( \mu \), based on a finite sample of data. There are two main methods:

1. **Point estimate:** Summarize the sample by a single number that is an estimate of the population parameter;

2. **Interval estimate:** A range of values within which, we believe, the true parameter lies with high probability.

*Example.* Suppose I wanted to estimate the mean height of all female students at UNC. I took a sample in this class and the sample mean was \( \bar{x} = 65.5 \) (inches). So the obvious thing to do is to take that as an estimate for the population mean. But I didn’t have to use the sample mean. I could have taken the sample median (65) or the sample mode (63). It makes sense to ask which is better.
What properties make a good point estimator?

1. It’s desirable that the sampling distribution be centered around the true population parameter. An estimator with this property is called *unbiased*.

2. It’s desirable that our chosen estimator have a small standard error in comparison with other estimators we might have chosen.

The sample mean is exactly unbiased (whereas the sample median may not be), and also, if the true population is normal, the sample mean has a smaller standard error than the sample median. Both of these would indicate that the sample mean is preferable to the sample median as an estimator of the population mean. However there are other properties that could nevertheless make the median preferable (e.g. it’s more resistant to outliers).
In the case of a binomial proportion, the obvious point estimator is the sample proportion. For example, consider our example about President Obama’s popularity rating (class posted 03/05/09 — Chapter 6 material).

In this example, 68% of respondents gave Obama a positive rating after he had been in office for one month (the answer could be different if we repeated the poll now). The most natural interpretation of this is that 68% or 0.68 is a statistic which serves as an estimator of the true but unknown proportion of people who would have approved of Obama if the whole population had been surveyed. It seems obvious that we would use the sample proportion as an estimator of the population proportion, but we don’t have to.
Now let’s turn to interval estimates. The simplest way to introduce this is through an example.

*Example.* In a college of 25,000 students, the administration would like to know for what proportion of students both parents had completed college. A sample of 350 students was drawn at random and in that sample, 276 of the students said that both their parents had completed college.

The sample proportion is \( \frac{276}{350} = 0.789 \) (or 78.9%), so by the same logic as in the last example, it makes sense to use that number also as an estimator of the population proportion (in this case the “population” is all 25,000 students at this college). But, how accurate is that?
The idea: Consider the interval

Sample Proportion $\pm 1.96 \times \text{Standard Error}$  

Why would this work?
Sample Proportion $\pm 1.96 \times$ Standard Error \hfill (*)

First let’s calculate $\Pr\{-1.96 < X < 1.96\}$ when $X$ is a standard normal random variable (mean 0, standard deviation 1).

From the normal table, for $z = 1.96$ we have left-tail probability 0.9750. For $z = -1.96$ we have left-tail probability 0.0250. The difference is $0.975-0.025=0.95$.

But given that the sample proportion has an approximately normal distribution, this means that the probability that the sample proportion lies within 1.96 standard errors of the true mean is also 0.95. Or in other words, the probability that the interval (*) includes the true mean is 0.95. This is what we mean by saying that the interval we calculated is a 95% confidence interval.
A side comment. Earlier in the course, we said that there is a 95% chance that a normal random variable lies within two standard deviations of the mean (this is part of the empirical rule first discussed in Chapter 2, because although at that time we didn’t use the words normal distribution, that’s actually what the empirical rule refers to). So why have we now replaced the number 2 with 1.96? Actually, 1.96 is more accurate. If we repeat the above normal probability calculations with $z = \pm 2$ instead of $z = \pm 1.96$, the probability becomes $0.9772 - 0.0228 = 0.9544$. That’s still quite close to 95%, and 1.96 is quite close to 2, so in practice, it doesn’t make much difference whether we use 2 or 1.96 standard deviations. But at this stage of the course, we’re trying to be more precise about things than we were earlier on, hence the change.
Confidence interval for a population proportion

To construct a confidence interval to measure the proportion $p$ of a population that has a particular characteristic (e.g. supporters of President Obama):

Step 1: Take a sample of size $n$, calculate $\hat{p}$ (pronounced $p$-hat) as the sample proportion of people who have that characteristic (e.g. saying they support Obama in a survey)

Step 2: Calculate the standard error $SE = \sqrt{\hat{p}(1-\hat{p})/n}$.

Note: The formula should really be $\sqrt{p(1-p)/n}$, but we don’t know $p$, so we use $\hat{p}$ instead.

Step 3: The 95% confidence interval $(\hat{p} - 1.96 \times SE, \hat{p} + 1.96 \times SE)$. 
Example from text. In one question of the GSS in 2000, 1154 people were asked whether they would be willing to pay higher prices to support the environment. 518 said yes.

Find a 95% confidence interval for $p$, the true proportion in the whole population who would be willing to pay higher prices to support the environment.
Step 1: \( \hat{p} = \frac{518}{1154} = 0.449 \) to 3 decimal places.

Step 2: The standard error is \( \sqrt{\frac{0.449 \times 0.551}{1154}} = 0.0146 \).

Step 3: \( 1.96 \times 0.0146 = 0.029 \) to 3 decimal places.

The 95\% confidence interval is (0.420, 0.478).

In practice, we wouldn’t usually express this to three decimal places and simply say that we believe the true proportion of people who support the proposition (i.e. who would be willing to pay higher prices to protect the environment) is between 42\% and 48\%. 
A side comment. In another GSS survey people were asked whether they would support legislation to force industry to adopt more environment-friendly policies. This time close to 80% answered yes. Yet it seems likely that tighter regulation on industry will result in higher prices for consumers (this will certainly be true if you ask the industry representatives). Another case where the wording of a question arguably influences the answer to a much greater extent than standard error calculations indicate.
Sample size condition

For these calculations to be valid (standard error formula including \( \hat{p} \) in place of \( p \), normal approximation for the distribution of \( \hat{p} \)) we require the sample size to be “reasonably large”. In practice, it is sufficient that \( n\hat{p} \geq 15 \) and \( n(1 - \hat{p}) \geq 15 \).
Confidence intervals with other confidence coefficients

So far we have worked with 95% confidence intervals, signifying that there is supposed to be a 95% probability that the interval includes the true population parameter. However, there’s nothing special here about the probability 95% — we could equally well work with 90%, or 99%, or any other probability we care to specify.

If we want a 99% confidence interval, we make the same calculation but replace 1.96 by 2.58. If we want a 90% confidence interval, we make the same calculation but replace 1.96 by 1.645. See Table 7.2, page 328.
Example. In an earlier example, we considered a sample of 350 students from a college (with a total student population of 25,000) and asked for what fraction of students it was true that both their parents had been to college. In that case, the sample proportion was .789. Suppose we wanted a 90% confidence interval.

The standard error is \( \sqrt{\frac{.789 \times (1-.789)}{350}} = .0218 \) — multiply by 1.645, the margin of error is 0.036 to three decimal places. Thus the 95% confidence interval is (.753, .825).
Interpretation of a confidence interval

Continue the previous example: Does the answer mean there is a 90% chance that the true proportion is between .753 and .825?

Strictly speaking, such a statement doesn’t make sense — we’re talking about a finite population of parents; either they went to college or they didn’t; what does it mean to talk of a 90% chance?

What the 90% confidence statement really means is that in many repetitions of the procedure, the interval will cover the true value 90% of the time.
Example: Let's suppose the true proportion is 80%. I ran 10 simulations experiments where I generated a binomial random variable with $n = 350, p = 0.8$. The results were: 272 285 266 289 287 278 269 277 285 271.

I now constructed a 90% confidence interval for each of the ten hypothetical samples. The results were:

<table>
<thead>
<tr>
<th>$X$</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>$X$</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>272</td>
<td>0.741</td>
<td>0.814</td>
<td>278</td>
<td>0.759</td>
<td>0.830</td>
</tr>
<tr>
<td>285</td>
<td>0.780</td>
<td>0.848</td>
<td>269</td>
<td>0.731</td>
<td>0.806</td>
</tr>
<tr>
<td>266</td>
<td>0.722</td>
<td>0.798</td>
<td>277</td>
<td>0.756</td>
<td>0.827</td>
</tr>
<tr>
<td>289</td>
<td>0.792</td>
<td>0.859</td>
<td>285</td>
<td>0.780</td>
<td>0.848</td>
</tr>
<tr>
<td>287</td>
<td>0.786</td>
<td>0.854</td>
<td>271</td>
<td>0.738</td>
<td>0.811</td>
</tr>
</tbody>
</table>

The interval is slightly different each time, and in fact, in 9 out of the 10 cases the interval covered the true value 0.8.
**Conclusion:** The *interval* is random. The confidence coefficient (in this case, 90%) represents the long-run probability that the interval would cover the true value in many repetitions of the sampling procedure.