As we have noted, statistics is the science of learning from data. One important aspect of such learning is statistical inference, wherein we attempt to draw inferences about characteristics of a population based on a sample from that population. Statistical inference problems may take many forms. For the second week of our statistics module in this class, we will focus on the problem of inferring the value of a single numerical summary of the population. Such inferences should have two components: (a) an estimate of the numerical value about which we seek to learn, and (b) a precise statement quantifying our level of confidence (or, conversely, uncertainty) in that estimate.

Although we could conceive of many numerical summaries of a population, commonly people are interested in quantities like averages, proportions, quantiles, etc. That is, in the same types of quantities we use in descriptive statistics to summarize a dataset. Rather than trying to cover a broad range of such summaries in a short period of time, we will instead examine just one summary, but in significant depth and detail. We will concentrate on the estimation of proportions, and we will do so specifically within the context of opinion polls. However, the basic paradigm that we unveil and explore is representative of a much larger class of problems.

1 Background

1.1 Some Terminology

A population is the set of all subjects of interest to us. It is the total group about which we wish to draw some conclusions of some sort. The meaning of ‘subjects’ varies with context and the underlying issues of interest to us. For example, we may be interested in the prevalence of sexually transmitted diseases (STDs) among US adults in the 18-25 age group. Then \( \text{US adults in the 18-25 age group} \) is our population. If we wish to restrict our attention to those who also have a high school degree, driven perhaps by an interest in studying prevalence among those that are ‘more educated’, then the relevant population is \( \text{US adults in the 18-25 age group with a high school degree} \). Alternatively, if we are instead interested in understanding how universities in China are adapting to the quickly changing economy in that country, the ‘population’ is All Chinese universities.

A sample is a set of subjects belonging to a population for which we have data. The sample is therefore a subset of the population. A collection of young adults responding to a phone survey on sexual practices would be a sample of our population \( \text{US adults in the 18-25 age group} \). If at the start of the survey it is asked whether or not the respondent has a high school degree, and for those that respond ‘no’ the phone call is ended (with a ‘thank you for your time’) and the respondent is dropped from the survey, then the result is a sample from \( \text{US adults in the 18-25 age group with a high school degree} \). Similarly, if we use the Internet
and Google to collect data on the first 20 Chinese universities listed with English translation webpages, that is a sample from All Chinese universities.

A **parameter** is a numerical summary of the population. Such summaries can take on almost any form you might imagine. However, as mentioned previously, there are certain forms that are particularly common, such as means, medians, proportions, quantiles, etc. For example, parameters associated with *US adults in the 18-25 age group* and relevant to study of STD prevalence in that population might be the proportion subjects who have had gonorrhea or the average number of sexual partners among subjects in the past six months. Similarly, parameters of interest in the study of the adaptation of Chinese universities to a changing economy might be the change in average expenditure per student across universities over the past 10 years or the median change in international ranking over that period.

A **statistic** is a numerical summary of a sample taken from a population. Generally the choice of numerical summary defining a statistic mirrors the choice of parameter of interest to us. So a study interested in the average number of sexual partners in the past six months among young adults would naturally look at the same average in a sample of young adults. And similarly in our other examples above.

In statistical inference, we generally use a statistic as an **estimate** of its corresponding parameter – that is, as a “best guess” for the parameter, based on the sample. We also try to accompany that estimate by a corresponding statement of the level of (un)certainty associated with it. Typically this statement takes the form of a **confidence interval**, which is an interval of numbers within which the parameter is believed to fall.

Our focus in this class will be on one specific choice of numerical summary: proportions. That is, we will focus on the case where we wish to learn the proportion of subjects in a population that possess a certain characteristic. And we will do so by reporting the proportion of subjects in the sample that possess that same characteristic, as well as an accompanying interval.

Whether you have seen this type of statistical inference presented formally before or not, you have in fact encountered it frequently throughout your lives, in the form of the reporting of results from opinion polls.

### 1.2 Our Focus: Opinion Polls

We will concentrate on the estimation of proportions, and we will do so specifically within the context of opinion polls. Such polls are ubiquitous, and may be found in nearly all media, from mail to phones to TV to the Internet. They are used to ask opinions on nearly every topic under the sun (or so it seems!), but the most basic version ultimately reduces, in the abstract, to a binary response – yes/no, agree/disagree, democrat/republican, Pepsi/Coke, Facebook/MySpace.

Let’s look at a typical example of the type of statement produced as a result of such opinion polls.
Example 1. (Gauging People’s Sense of Financial Security) A February 22, 2009 poll commissioned by the Washington Post asked a random sample of 1001 adults across the country, “How financially secure do you feel? (Secure / Insecure)” Based on this poll, the Post reported that 57% of US adults feel financially secure, with a margin of error of “plus or minus three percentage points.” □

What precisely is this statement saying? We can begin to dissect this statement by identifying the various components that correspond to the terminology introduced above.

The target population is nominally the set of “US adults”. What does that mean however? To determine this we would have to obtain additional details about the sampling design and the manner in which the sampling was carried out. Is ‘adult’ well-defined, perhaps meaning 18 years and older? When the poll was administered, did those conducting the poll understand that this was to be the definition? In addition, was that criterion adequately applied when interviewing respondents (i.e., was everyone under 18 years old that responded excluded from the results)?

Note that none of these issues actually involve the response itself (i.e., secure/insecure). But they are critical precursors underlying the reliability of the inferences made about that response. Of course, professional polling agencies are aware of these issues and are experienced in working to ensure that they are addressed appropriately. But doing so is nevertheless nontrivial.

Moving on, we note that the parameter being inferred here is not the proportion of US adults that “feel secure”. Rather, it is the proportion of US adults (putting aside nuances about this definition) that, if asked, would respond to the question “How financially secure do you feel? (Secure / Insecure)” with the response ‘secure’. These are not necessarily the same thing! For example, some people might be uncomfortable about discussing their sense of financial security with a stranger, perhaps misunderstanding the purpose of the poll, and reply ‘secure’ even if they didn’t actually feel secure.

The definition of the sample here is clear (i.e., “1001 adults across the country”), but the term ‘random sample’ bears further scrutiny. Again, professional polling organizations are typically well-versed in how to conduct such polls, but the phrase ‘random sample’ is used in both technical and colloquial senses, and the colloquial sense (as used commonly in the media) tends to be a catch-all for many types of sampling methods.

The estimate here is the number 57%, and is the proportion of respondents that answered ‘secure’ when asked the question. The so-called margin-of-error is a standard but somewhat mysterious (to the casual observer) way of stating uncertainty. When combined with the estimate, by adding and subtracting it, we obtain the confidence interval. In the case of this poll, this interval takes the form (54%, 60%).

What is missing here – and this is not unusual – is some quantification of (un)certainty associated with this result. Sometimes there is fine print accompanying the basic polling statement. A standard phrase that is used is to say that the poll is accurate “19 times out of 20”. Note that this has the air of a probability statement, and indeed that is the case!

We see that polling statements are actually somewhat non-trivial. In the rest of our time on this topic, we will delve into the quantitative infrastructure underlying such statements, seeking to understand the quantitative reasoning behind them, building from the modeling assumptions upwards, and to discuss the resulting implications.
2 Taking a Closer Look at Polling Statements

In this section we look more formally at opinion polls and the statements emerging from them. We will state the basic sampling model, explain the probability statement deriving from it, and discuss its interpretation.

2.1 A Sampling Model: Simple Random Sampling

Polling statements are effectively a reporting of the observed data, accompanied by a probability statement. We have seen in our section on Probability that probabilities may come from many sources. Here they come from a model (i.e., as opposed to empirical data or subjective beliefs). In particular, they come from a model of the sampling; that is, from a model of the process by which the data in the sample were collected.

Ideally, our model describes precisely how the data were collected. More commonly, it’s an approximation to the reality of the sampling, attempting to balance fidelity to the reality with mathematical and computational tractability. Arguably the simplest, but also the most common, sampling model is the simple random sampling (SRS) model.

Suppose our population has \( N \) subjects and that we wish to sample a subset of \( n \) of those subjects (with, of course \( n \leq N \)). Simple random sampling is a method of sampling by which every possible sample of size \( n \) has the same chance of being selected. That is, it is a type of equally likely outcomes model.

Example 2. (SRS in a Toy Population) Suppose that there are \( N = 6 \) subjects in a population, and that we intend to sample \( n = 2 \) of these subjects by simple random sampling. An outcome in this setting is a specific subset of two subjects. The sample space is all possible subsets of two subjects. If we represent our population of six subjects as \( 1, 2, 3, 4, 5, 6 \), then the sample space is

\[
S = \{ \{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{1,6\}, \{2,3\}, \{2,4\}, \{2,5\}, \{2,6\}, \{3,4\}, \{3,5\}, \{3,6\}, \{4,5\}, \{4,6\}, \{5,6\} \} .
\]

Since there are 15 outcomes, under the equally likely outcomes model, each sample has a probability of \( 1/15 \) of being observed. \( \Box \)

More generally, for a population of size \( N \), and a sample of size \( n \), the sample space under SRS is the set of all subsets of size \( n \) of the integers 1 through \( N \). We would like to know the probability associated with samples in this general context. To do so, we need to calculate the size of the sample space.

We have done this type of argument early in the semester, although only for specific examples. To begin, we observe that there are \( N \) ways to choose the first subject. For each of these \( N \) ways, there are \( N - 1 \) ways to choose the second subject. Therefore, there are \( N(N - 1) \) ways to choose the first two subjects. Given these two subjects, there are \( N - 2 \) ways to choose the third subject, yielding for us that there are \( N(N - 1)(N - 2) \) ways to choose the first three subjects. Continuing in this fashion, we arrive at the conclusion that there are \( N(N - 1)(N - 2) \cdots (N - n + 1) \) ways to choose the \( n \) subjects.
But wait! This is the number of ways to choose an ordered subset of \( n \) subjects. For our purposes, however, any ordering of a given set of \( n \) subjects is the same sample. So we want to reduce this number by a factor reflecting the number of ways we can permute the order of \( n \) subjects. Arguing similarly to just above, given a subset of \( n \) subjects, there are \( n \) ways to place one of them 1st, \( n-1 \) ways to place another of them second after the first, \( n-2 \) ways to place another of them third after the first two, and so on until the last one is placed. Therefore, there are \( n(n-1)(n-2)\cdot \cdots \cdot 2\cdot 1 \) ways to permute the order of \( n \) subjects.

So the size of the sample space under SRS is

\[
\frac{N(N-1)(N-2)\cdots(N-n+1)}{n(n-1)(n-2)\cdots2\cdot1}.
\]

That is, this is the number of ways to choose \( n \) subjects for our sample from the \( N \) subjects in the population\(^1\). As \( N \) grows, this quantity can get large very quickly. For \( N = 6 \) and \( n = 2 \), we saw that it equals 15. If instead we have \( N = 60 \) and \( n = 20 \), it equals approximately \( 4.19 \times 10^{15} \). In other words, the probability of any particular subset of 20 people being chosen for the sample is on the order of \( 10^{-15} \).

What is important in the SRS model is the sense of fairness inherent in the sampling, in that there is no tendency towards any one particular subset of individuals being chosen over another. In that sense the resulting sample is expected to be a good reflection of the overall population. Not all sampling plans share this characteristic. We will discuss alternative sampling plans a bit in Section 5.

### 2.2 A Formal Inferential Statement

Recall our discussion during the probability section of this course regarding outcomes versus events. Outcomes are the actual realizations of a random process. Events are collections of possible outcomes (i.e., formally, events are subsets of the sample space), typically encoding higher-level descriptions of interest to us as to how outcomes might behave. Under SRS the outcome is a subset of \( n \) subjects selected randomly from the population of \( N \) subjects. But at the core of polling statements like that in Example\(^1\) is a probability statement phrased not in terms of the sample itself but in terms of a numerical summary of a characteristic of the subjects in the sample. That is, it is a probability statement pertaining to a certain event.

In order to be more precise, we need some additional mathematical notation. Let \( p \) be the proportion of interest to us in the population and, for a given sample, let \( \hat{p} \) be the corresponding proportion in the sample. Let \( ME \) denote the margin-of-error\(^2\). Then define the event \( A \) as

\[ A = \text{Event that } |p - \hat{p}| \leq ME . \]

Or, equivalently, we can write

\[ A = \text{Event that } \hat{p} - ME \leq p \leq \hat{p} + ME . \]

\(^1\)This number is in fact called ‘N-choose-n’, and formally expressed in the notation \( \binom{N}{n} \).

\(^2\)We will define this quantity in greater detail below, in Section 3.
The first formulation of our event, in (2), says that our sample proportion \( \hat{p} \) is within a distance \( ME \) of the population proportion \( p \). Or, rephrased using the terminology of estimation, it says that our estimate \( \hat{p} \) is within \( ME \) units of the parameter \( p \). The second (equivalent) formulation of our event, in (3), says that the interval \((\hat{p} - ME, \hat{p} + ME)\) contains the parameter \( p \). This interval is our confidence interval.

What accompanies the typical statement of polling results is the assertion that \( P(A) \approx 0.95 \). That is, the probability that our estimate \( \hat{p} \) is within a distance \( ME \) of the parameter \( p \) is approximately 95%. Or, alternatively, we can say that the probability that our confidence interval \((\hat{p} - ME, \hat{p} + ME)\) contains the parameter \( p \) is approximately 95%. This particular confidence interval is therefore referred to more precisely as a ‘95% confidence interval’. Since 19/20 = 0.95, we see that the colloquial way of citing confidence in a poll e.g., “this poll is expected to be accurate in 19 out of 20 times sampling”, is just an informal way of reporting the formal underlying probability statement.

Note that there must be something special about this so-called ‘margin-of-error’ to have it be the case that the probability \( P(A) \) is always approximately 95%. For example, intuitively we would expect that for larger sample sizes we get better accuracy, which would suggest that \( ME \) should decrease with increasing sample size. Such is indeed the case. Just precisely (i) how we set the margin-of-error, and (ii) why that setting yields 0.95 probability, is the focus of the next section.

### 3 A Look at the Machinery Behind Polling Statements

In the probability section of the course we saw that the calculation of probabilities under equally-likely-outcomes models boil down to counting. Under SRS we have an equally likely outcomes model and, in principle, we can employ such counting. But in reality this counting becomes highly non-trivial to do, and ultimately we resort to a mathematical approximation. In fact, we make use of one of the most celebrated approximations in probability and statistics. In this section we will motivate and introduce the approximation and, along the way, gain additional insight into the particular form of the margin-of-error.

#### 3.1 The Sampling Distribution

The SRS model dictates that each sample of size \( n \) subjects from a population of \( N \) subjects has equal probability, the value of which is given by one over the quantity in (1). But, as we described above, polling statements involve a probability regarding an event \( A \) expressed in terms of the sample proportion \( \hat{p} \). In order to better understand the nature of this latter probability, we need to understand how the probabilities at the level of outcomes translate to probabilities at the level of our statistics, \( \hat{p} \). The key principle here is the notion of a so-called ‘sampling distribution’. We begin by illustrating this notion through an example.

**Example 3.** (Revisiting our Toy Population) Recall Example 2 in which we had just \( N = 6 \) subjects. Suppose that half of these subjects would, if asked today, declare that they intend to
vote for the republication in an upcoming election, and the other half, for the democrat. So, if our interest is in the proportion that say ‘republican’, in this population that proportion is \( p = 0.5 \).

Now let us consider the nature of \( \hat{p} \) in this setting. In Example 2, we found it convenient to label the 6 subjects simply by the numbers 1, 2, 3, 4, 5, and 6. Here, let us label them \( \{R_1, R_2, R_3, D_1, D_2, D_3\} \) (4) so as to emphasize their voting preference. If we sample \( n = 2 \) subjects, our sample space can now be expressed as follows.

\[
S = \{ \{R_1, R_2\}, \{R_1, R_3\}, \{R_1, D_1\}, \{R_1, D_2\}, \{R_1, D_3\}, \{R_2, R_3\}, \{R_2, D_1\}, \{R_2, D_2\}, \{R_2, D_3\}, \{R_3, D_1\}, \{R_3, D_2\}, \{R_3, D_3\}, \{D_1, D_2\}, \{D_1, D_3\}, \{D_2, D_3\} \}.
\]

For each of these 15 possible samples of two subjects, we can calculate a value for the estimate \( \hat{p} \). This value, of course, can change from sample to sample. But, in fact, for many samples the value of \( \hat{p} \) is the same. For example, with the samples \( \{D_1, D_3\} \) and \( \{D_2, D_3\} \) we get \( \hat{p} = 0 \) in both cases. In fact, note that \( \hat{p} \) can only take on the values 0, 0.5, and 1.0 in our particular population with samples of size \( n = 2 \).

It is useful to organize for ourselves how \( \hat{p} \) changes. Thinking back to our basic tools in descriptive statistics, we could in principle create a simple table of the 15 values. But more succinct would be a frequency table, as shown in Table 1. This table shows the distribution of values \( \hat{p} \) over the 15 possible different samples. We see, for example, that in 9 of the 15 samples (i.e., 60%) we will get \( \hat{p} = 0.5 \), while in 3 of the 15 samples (i.e., 20%) we will get \( \hat{p} = 0 \), and in the last 3 of 15 samples (i.e., again, 20%), we will get \( \hat{p} = 1.0 \). □

The frequency table in Table 1 showing the relative frequencies with which the different potential values of \( \hat{p} \) occur across all possible samples, is called the sampling distribution of the statistic \( \hat{p} \). This notion of a distribution for a statistic we intend to use to estimate a parameter is fundamental to inferential statistics. It is through these distributions that probability statements underlying things like confidence intervals (or the more informal polling statements!) are built. It is important to note, however, that sampling distributions change as the population size \( N \) and the sample size \( n \) change, as the following example illustrates.

**Example 4.** (Revisiting our Toy Population (cont)) Consider again the population from Example 3 with \( N = 6 \) and \( p = 0.5 \), but suppose now that we sample \( n = 3 \) subjects. What is the impact of our including one additional subject in the sample?

The sample space will consist of all subsets of size three chosen from the set in (4) above. Using the expression in (4), we can calculate that there are 20 possible samples of size \( n = 3 \)
from a population of size $N = 6$. Therefore, under the simple random sampling model, each such sample (e.g., $\{R_1, R_2, D_3\}$, or $\{R_2, D_2, D_2\}$, etc.) has probability $1/20$ of being chosen. Turning our attention to $\hat{p}$ now, clearly we can have $0, 1, 2$, or $3$ ‘republican’ responses from the subjects in the sample, and so $\hat{p}$ can take on only the values $0, 1/3, 2/3$, or $1$. And just as clearly, while there is only one way to get $\hat{p} = 0$ (i.e., $\{D_1, D_2, D_3\}$) and one way to get $\hat{p} = 1$ (i.e., $\{R_1, R_2, R_3\}$), there will be more than one sample corresponding to each of the values $\hat{p} = 1/3$ and $2/3$. We can either list all of the possible samples and count the number of ways in which these last two values of $\hat{p}$ may arise, or we may note that there must be a certain symmetry in the remaining samples, in that, for example, $\{R_1, R_2, D_1\}$ can be matched with $\{D_1, D_2, R_1\}$, and so on. In either case, we find that there are 9 ways each for $\hat{p}$ to take the values $1/3$ and $2/3$. Table 2 shows the resulting sampling distribution for $\hat{p}$. □

The sampling distribution for $\hat{p}$ for samples of size $n = 3$ is different from that for samples of size $n = 2$. But these two distributions also share certain characteristics. It is common to use numerical summary tools from descriptive statistics to quantify important aspects of sampling distributions. Two key quantities of this sort are (i) the mean of the sampling distribution, and (ii) the standard deviation of the sampling distribution, which measure the center/location and spread of the distribution, respectively.

For example, from Tables 1 and 2 it is easy to calculate the mean. In both cases, the mean is 0.5. Note that 0.5 is the value of the parameter $p$ in the population that we seek to infer. This is not a coincidence! Instead, it is due to our use of simple random sampling. When the mean of the sampling distribution for an estimator is equal to the parameter that it estimates, we say that the estimator is unbiased. Otherwise, we say an estimator is biased. The unbiasedness of the estimator $\hat{p}$ under simple random sampling is true regardless of the values of $p$, $N$, and $n$, and derives directly from the fact that all samples have equal likelihood under this model.

Returning to the sampling distributions in Tables 1 and 2, we can similarly calculate the standard deviation for the values of $\hat{p}$. Formally, the standard deviation of an estimator is called its standard error. Recall that the standard deviation tells us something about the spread of the distribution about its mean value. The standard deviation for the sampling distribution with $n = 2$ samples is 0.327, while with $n = 3$ samples it is 0.229. So for $n = 3$ the sampling distribution is less spread out (i.e., more concentrated) about its mean than for $n = 2$. Intuitively, this means that in the case of $n = 3$ there is less uncertainty in our estimate $\hat{p}$. This change in standard deviations reflects the gain in accuracy we obtain from

<table>
<thead>
<tr>
<th>$\hat{p}$</th>
<th>0</th>
<th>0.333</th>
<th>0.667</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1/20</td>
<td>9/20</td>
<td>9/20</td>
<td>1/20</td>
</tr>
</tbody>
</table>

Table 2: Frequency table showing the sampling distribution of $\hat{p}$ for a population of size $N = 6$ and sample of size $n = 3$, when $p = 0.5$. 

\[\text{\footnotesize{\textsuperscript{3}In principle, to do this you expand the table back into a list of all values $\hat{p}$, across all possible samples (i.e., meaning that your list contains repeats), and then calculate the mean as you would with any set of data. For example, expanding Table 1 we obtain the list \{0,0,0,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,1,1,1\}.}}\]
increasing the sample size for \( n = 2 \) to 3.

How can we turn all of this into a probability statement? One way is to use the standard error of \( \hat{p} \) as a unit of accuracy. Mimicking the type of event defined in (2), let us ask, “What is the probability that our sampling yields an estimate \( \hat{p} \) within one standard error of the true value \( p \)?” For samples of size \( n = 2 \), the standard error was 0.327. The answer to our question therefore is given by how much probability is associated with values \( \hat{p} \) lying between 0.5 − 0.327 = 0.173 and 0.5 + 0.327 = 0.827. Since we see, consulting Table 1 that only the value \( \hat{p} = 0.5 \) falls in this range, the probability that our estimate falls within one standard error of \( p \) is equal to 0.60. For samples of size \( n = 3 \), the standard error was 0.229, and so the answer to our question now is given by how much probability is associated with values \( \hat{p} \) lying between 0.5 − 0.229 = 0.271 and 0.5 + 0.229 = 0.729. Consulting Table 2 since the values \( \hat{p} = 1/3 \) and 2/3 both fall in this range, but no other values of \( \hat{p} \) do, the probability that our estimate falls within one standard error of \( p \) is 2 \times \frac{9}{20} = \frac{18}{20} = 0.90.

This manner of calibrating a statement of accuracy using the standard error of an estimator is actually at the heart of the definition of the margin of error, in that the \( ME \) is closely associated with this standard deviation. In order to see how, we need to move beyond our toy examples.

### 3.2 The Central Limit Theorem

Typically our populations are of sizes far larger than \( N = 6 \), and our samples, far larger than \( n = 2 \) or 3. Working with these toy models is useful for gaining insight into the fundamentals of sampling distributions. But the size of the sample space under simple random sampling grows roughly like \( N^n \) i.e., exponentially fast! So it quickly becomes infeasible to do by hand the types of calculations we have been doing in order to obtain sampling distributions for \( \hat{p} \).

Fortunately, as \( N \) and \( n \) become even moderately large, these distributions all start sharing a certain characteristic shape. That this should indeed be the case can be shown in a mathematically formal manner, and is part of the celebrated Central Limit Theorem (CLT). While a proof of this theorem is beyond the scope of this class, it is nevertheless easy to demonstrate the CLT in action and to characterize its behavior.

Look at Figure 1 where we use bar plots to visualize four sampling distributions for the estimator \( \hat{p} \). On the left are the sampling distributions we saw in Examples 3 and 4 where \( N = 6 \) and \( n = 2 \) and 3, respectively. On the right are the sampling distributions for the cases where \( N = 60 \) and \( n = 20 \) and 30, respectively. That is we have simply scaled up both \( N \) and \( n \) by a factor of 10. The parameter \( p \) is equal to 0.5 in all cases.

We see that for our two larger populations and sample sizes, the sampling distribution has filled out into the well-known ‘bell curve’ that we all knew and dreaded in high school! These distributions can be seen to be unimodal (i.e., has only one mode), symmetric, centered on a mean value 0.5, and fairly concentrated about their mean (i.e., their ‘tails’ decay quickly to the left and right). Note too that, having now seen these distributions for the larger population and sample size, side by side with those for the smaller, it is clear that there are hints of this bell curve even at very small populations and sample sizes.

The ‘bell curve’ describing these sampling distributions formally is called the normal distribution. The CLT in this context states that under simple random sampling from a
Figure 1: Sample distributions for the estimator $\hat{p}$ under the simple random sampling model, for $N = 6$ (left) and $N = 60$, and $n/N = 1/3$ (top) and $n/N = 1/2$ (bottom), respectively.

Population with proportion $p$ of interest, for sufficiently large population size $N$ and sample size $n$,

1. the sampling distribution of the estimate $\hat{p}$ will be well-approximated by a normal distribution, 

2. the mean of this normal distribution is $p$, and
3. the standard deviation of this normal distribution is
\[ \sqrt{\frac{p(1-p)}{n}}. \] (5)

In Figure 2 you can see the normal approximation as it applies to the two right-hand plots in Figure 1. The normal distribution is defined by a curve whose formula can be written down in an algebraic expression. Probabilities, in analogy to the bar plots we are approximating, can be calculated according to area under the curve. For example, the probability that \( \hat{p} \) is within one standard error of its mean is simply equal to the area under the curve between plus and minus standard deviation distance from the mean \( p \). These areas cannot be written down using simple algebraic expressions, but can be calculated using a computer or by consulting tables created for the purpose.

### 3.3 Explaining the Margin-of-Error

Recall at the end of Section 3.1 that we suggested using the standard error as a unit of accuracy in making probability statements about our estimate \( \hat{p} \) of \( p \). The margin of error cited in the standard reporting of polling results derives from a simple relationship between the normal distribution and its standard deviation.

Look at Figure 3. A schematic representation of the normal distribution is shown in black. Also indicated, using vertical lines, are distances of one, two, and three standard deviations (shown in blue, red, and yellow, respectively) from the mean (shown in gray). A nice property of the normal distribution is that, regardless of the actual values of the mean and standard deviation, it has 68% of its area between plus and minus one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations. And this fact holds regardless of the actual values of the mean and standard deviation.

So let’s put together everything we’ve learned now. We use the sample proportion \( \hat{p} \) as an estimate of the population proportion \( p \). We want to accompany that estimate by a statement of accuracy. The statement we wish to make is that

\[ P(\text{Event that } |p - \hat{p}| \leq M) \approx 0.95. \]

4This formula is for the case where \( N \) is much bigger than \( n \). It is in turn an approximation to a slightly more precise formula, which states that the standard deviation has the form

\[ \sqrt{\frac{p(1-p)}{n}} \left(1 - \frac{n}{N}\right). \]

This formula was used, for example, in creating the plots in Figure 2 since \( n = 20 \) or \( 30 \) is quite sizable compared to \( N = 60 \).

5This curve can be expressed as a function \( f(x) \) that depends on two parameters, \( m \) and \( s \). When evaluated at a given point \( x \) on the \( x \)-axis, \( f \) takes the value

\[ f(x) = \frac{1}{\sqrt{2\pi s^2}} \exp\left\{-\frac{(x-m)^2}{2s^2}\right\}. \]

The parameters \( m \) and \( s \) can be shown to be equal to the mean and standard deviation of this distribution.
We know that the sampling distribution for $\hat{p}$ is centered on $p$, and so to set the margin of error $ME$, we need to figure out how many units we go to the left and right of $p$ until we have accumulated an area in our bar chart equal to 0.95. Since this sampling distribution may be unwieldy to work with, we instead approximate it by a normal distribution, with a mean of $p$ and a standard deviation of

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

Finally, we know that for a normal distribution, 95% of the area under the curve is between plus and minus two standard deviations from the mean.

As a result of the above reasoning, we are led to the conclusion that we should set

$$ME = 2\sqrt{\frac{p(1-p)}{n}}. \quad (6)$$

This is the explicit formula for the so-called ‘margin of error’. Note, however, that the expression in equation (6) is a function of $p$. But it is $p$ that we are trying to estimate … presumably because we do not know its value! So how can we actually calculate the margin-of-error in practice?

Two solutions typically are employed. One is to substitute the value of $\hat{p}$ in (6). The other is to note that the term $p(1-p)$ in equation (6) is never bigger than 1/4, and is equal to this value only when $p = 1/2$. Substituting 1/4 for $p(1-p)$ in (6), we find that the

---

6Formally, we would say we are ‘bounding’ $p(1-p)$ above by 1/4.
margin of error is never bigger than $1/\sqrt{n}$. This last value i.e.,

$$ME \approx \frac{1}{\sqrt{n}}, \quad (7)$$

is typically what is reported in the media. So, for example, for the Washington Post poll
in Example 1, the margin of error being quoted i.e., “plus or minus 3 percentage points”, is
obtained as $1/\sqrt{1001} \approx 0.0316$.

4  An Application: The General Social Survey

The General Social Survey (GSS), is a (currently) every-other-year survey of the American
public, conducted by the National Opinion Research Center at the University of Chicago.
Going back to 1972, this survey seeks to gather data useful in painting a picture of the
demographics and attitudes of U.S. residents. As such, the data from this survey are of
particular interest to researchers in the social sciences.

Using the Survey Documentation and Analysis (SDA) web application\footnote{Available at \url{http://sda.berkeley.edu/GSS/}}, we can easily
explore the results of this survey, summarizing the responses to various questions. In doing
so, even though the software calculates all of the relevant output for us, it is critical that we

---

\(\text{Figure 3: Schematic representation of the normal distribution (black), with units of one, two, and three standard deviations indicated in blue, red, and yellow, respectively.}\)
read this output with an understanding of the material in the preceding sections in order to be best positioned to effectively both interpret and judge what we see.

**Example 5. (Mental Health)** Recall the example on mental health from the course notes on descriptive statistics. In 2006, GSS survey participants were read a description of a hypothetical person that had been “feeling really down” for “[t]he last several weeks.” Symptoms were said to include a “sad mood and heavy feeling,” a lack of enjoyment and pleasure, difficulty concentrating, reduced energy, weight-loss, and an inability to sleep. Participants were asked whether or not they would recommend that this person pursue various steps, including, “Go a doctor.” Possible answers were simply ‘yes’ or ‘no’, making the response a categorical variable with two categories.

The summary and analysis of the data for this survey question can be generated in GSS by selecting the variable ‘MHDODOC’, and putting it in the ‘ROW’ entry. There are options on what sort of analysis to use. If we check off the box next to ‘Confidence Intervals’, leaving the default ‘95 percent’ choice, and click the radio button next to ‘SRS’ for choice of ‘Sample Design’, we get the analysis summarized in Figure 4. Each cell of the output contains four entries: the percent, the confidence interval, the standard error, and the number of cases. We can use these entries to extract information for the types of opinion poll statements we studied in the previous sections.

The GSS polls US adults. Suppose our focus is on the proportion of the adults in the US population that would respond ‘Yes’ to the question of whether the hypothetical person described should go to a doctor. Call this proportion \( p \). We see that of a total of 1407 respondents, 878 of them, or 62.4%, responded ‘Yes’. So \( \hat{p} = 0.624 \) is our estimate of \( p \). In order to provide an accompanying notion of accuracy for this estimate, we need to compute the margin of error. The standard error is computed here using the formula in (5), with \( \hat{p} \) substituted for \( p \), which yields for us

\[
\sqrt{\frac{0.624 \times 0.376}{1407}} = 0.0129.
\]

Therefore, following the formula in (7), the margin of error is \( ME = 2 \times 0.0129 = 0.0258 \). Or, if we wish to use the approximate margin of error typically quoted in the media, per the formula in (8), we get that \( ME \approx 1/\sqrt{1407} = 0.0267 \). Note that because \( \hat{p} = 0.624 \) is close to 0.5, and the approximation in formula (8) is derived from the formula in (7) by plugging in 0.5 for \( p \) and \( 1-p \), it is no surprise that these two values for the margin of error are so close.

Alternatively, we might prefer to provide our results in the form of a 95% confidence interval. Recall that the form of this interval is given by \( (\hat{p} - ME, \hat{p} + ME) \). Plugging in the numbers above, we obtain the interval \((0.5982, 0.6498)\).

These various numbers are summarized in the first cell of the table in Figure 4. To get the numbers in the second cell, we simply switch our focus to concentrate on the proportion of US adults that would respond ‘No’ to our question, and we now call this proportion \( \hat{q} \). The calculations in this case are completely analogous to those above. □
5 Departures from Our Basic Model

Let’s step back a moment now from all of the various details we have developed above. The basic statement of opinion poll results, such as that given in Example 1, is based on a model. In particular, it is based on a sampling model i.e., a model of how the data were collected. If the assumption of this model is incorrect, it is reasonable to wonder how far we can trust the results and statements that rest upon it.

While such questions can (and should!) be asked for any of the types of models we have seen throughout this class, within the world of statistics there is a very nice expression that is relevant here. Attributed to Sir David Cox, a eminent statistician of the 1900s, it says simply, “All models are wrong; some are useful.”

The simple random sampling model captures the situation where no particular sample of size $n$ has more of a chance than any other of being chosen. If this assumption is not too unreasonable, then everything that follows is reasonable as well, since we have used (or made reference to) only mathematical arguments and proofs to build up the standard polling statement from the SRS model.

If the SRS model is *not* reasonable, then in order to try to answer the question of how trustworthy are the results based upon it, we must have some idea of how the assumptions
of SRS fail. With this information in hand, we can then try to do two things: (i) quantify the impact of this failure on our stated results; (ii) produce alternative sampling models that better match reality, and repeat the process of deriving appropriate accuracy statements.

Actually, there are many sampling models in use. Often they were developed as conscious alternatives to SRS, driven by necessity and/or convenience. Simple random sampling is sometimes simply unrealistic to implement. Two common alternatives are **systematic sampling** and **stratified random sampling**. In systematic sampling, we pick a random starting point in the population and then sample a subset in a systematic manner. For example, given a town with houses in a traditional block system, we might start polling at a randomly selected house and then continue polling every, say, 10th house on every other block or some such thing. In stratified random sampling, we first group subjects into separate sub-populations, we randomly sample a subset of the sub-populations, and then randomly sample subjects within the chosen sub-populations. For example, we might randomly sample city blocks, and then within those city blocks, we could randomly sample some small number of houses.

Both of these sampling plans can be much more amenable to faithful implementation (e.g., by survey workers) than simple random sampling. And in both cases, accuracy statements may be derived in a manner similar to those we did for SRS. The notion of a sampling distribution for \( \hat{p} \) is still critical. In fact, a form of the central limit theorem still holds for approximating the sampling distribution under these sampling plans. And, if set up appropriately, the estimate \( \hat{p} \) will still be unbiased. The primary difference then, between these plans and SRS, is found in the calculation of the standard errors, and hence in the margin of error reported. A detailed study of these points, however, is beyond the scope of this class.

Models can also fail because certain sources of ‘bias’ enter the study. Two common types of bias are **non-response bias** and **misleading response bias**. Examples of non-response bias are when subjects fail to return mail-in surveys, fail to click ‘Yes, I will take the survey’ on the web, or hang up when called by a survey worker. It is easy to see how such actions might bias a study, using our toy model in Example 3. Suppose that the third democrat, \( D_3 \), unknown to us, refuses to participate. For simplicity, suppose that this subject is not even on our lists (e.g., perhaps he/she did not register to vote). Then our simple random sampling is really from the population \( \{R_1, R_2, R_3, D_1, D_2\} \), not \( \{R_1, R_2, R_3, D_1, D_2, D_3\} \), and \( p \) is equal to 0.6 in this population, not 0.5. Our estimate \( \hat{p} \), being based still on simple random sampling, will indeed be an unbiased estimate of \( p \), but for \( p = 0.6 \). In other words, the impact of this non-response bias is to leave us with an estimate \( \hat{p} \) that will be biased towards values that on average overshoot the proportion 0.5 in the actual population that is of interest to us.

Misleading response bias is when subjects are recorded with values other than actually pertain to them. This may be due to anything from errors on the part of the survey workers to simple deception on the part of subjects. Note that not all subjects are deceptive to be malicious. A particularly interesting example of this phenomenon is the so-called ‘Bradley Effect’, where the cause is social and psychological in nature. The Bradley effect has its roots in an empirical finding during the 1982 gubernatorial race in California, in which Tom Bradley, an African-American, was ahead of his opponent (who was white) in the polls preceding the election, but inexplicably came out behind in the actual voting. The theory
behind the Bradley effect is that voters may claim (and perhaps even believe) that they will vote in a politically correct manner where sensitive matters or concerns are present (e.g., such as the race of the candidates), but once in the polling booth, they may change their mind. The Bradley effect was much discussed, and watched for, in the past presidential election that saw Barack Obama succeed in being elected President.

6 What to Take With You

There are many different problems in estimation, each with their own unique flavor in terms of the exact nature of the parameter, estimate, and margin-of-error. Yet there are broad classes of problems whose flavor is basically the same as what you have seen here for the particular case of opinion polls. So while it may seem like we have only studied one very specific problem, in fact we have encountered quite closely the very foundation upon which a broad spectrum of statistical results and statements are based.

Things to take away with you include the following.

1. Concepts
   - The distinction between population and sample, and between parameter and statistic.
   - The nature and interpretation of the probability statement that underlies standard polling statements.
   - The notion of a sampling distribution for an estimate of a population parameter, and the accompanying notions of (un)biased estimators and the standard error of an estimator.

2. Skills
   - Decompose a standard polling statement into its constituent parts, identifying the population, the sample, the parameter, the estimate, and the margin of error.
   - Given summary data, such as in the SDA output from the GSS, construct the estimate \( \hat{p} \), the margin of error, and a 95% confidence interval for \( p \).
   - Calculate the sampling distribution of \( \hat{p} \) under simple sampling plans (e.g., SRS) in a toy population, and evaluate related quantities such as the mean and standard deviation.