

# The Theory of Polls

## Mathematical Topics

Students see that the probability distribution for large polls is approximately normal.

- Investigating the effect of changing the overall population on the theoretical distribution of poll results
- Investigating the effect of changing poll size on the theoretical distribution of poll results
- Seeing that increasing poll size leads to more accurate poll results
- Using combinatorial coefficients to find the theoretical distribution of poll results for polls of various sizes
- Seeing that as poll size increases, the distribution of sample proportions becomes approximately normal, and expressing this fact using the central limit theorem

## Outline of the Day

### In Class

1. Discuss *Homework 4: Graphs of the Theory*
  - Bring out that as the true proportion increases, the probability bar graph “shifts to the right”
  - Develop expressions for the probability of each outcome in terms of the true proportion
2. *The Theory of Polls*
  - Students find theoretical probability distributions for polls of several different sizes
3. Discuss *The Theory of Polls*
  - Review the use of combinatorial coefficients to get the probabilities
4. Introduce the central limit theorem
  - Discuss the probability of a “correct” poll prediction
  - Compare probability bar graphs for  $n = 5$  and  $n = 9$
  - Bring out that the graph is beginning to resemble the normal curve
  - Use the graph for the case  $n = 50$  to emphasize the resemblance
  - State the central limit theorem (for this context)

### At Home

*Homework 5: Civics in Action*

### Special Materials Needed

- Transparencies of probability bar graphs for results of 3-person polls from different overall populations (see Appendix C)
- A sequence of probability bar graphs using the same scale for the results of polls of different sizes (see Appendix C)

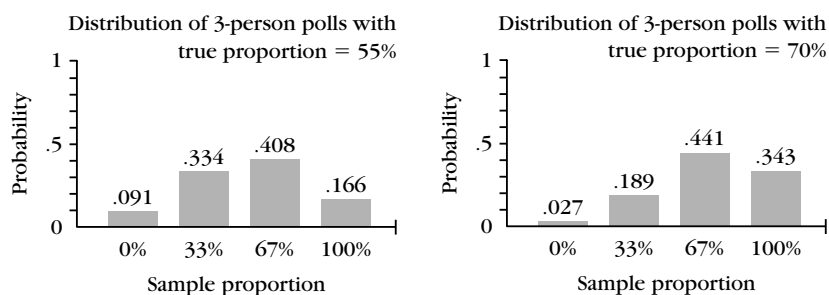
## 1. Discussion of Homework 4: *Graphs of the Theory*

Have groups discuss Question 3 while one group prepares an overhead of its graphs for Question 1 and another group does so for Question 2.

- *Questions 1 and 2*

Questions 1 and 2 are primarily for review and to bring out the idea that the probabilities change when the true proportion changes.

Probability bar graphs for the two situations look like this. (Both graphs are included in Appendix C. Note that because of rounding, the probabilities for the first graph do not total exactly 1.)



- *Comparing the results*

Students may describe the differences between the graphs in various ways, such as

“The graph for 70% has been shifted to the right, compared to the graph for 55%.”

or

“With a true proportion of 70%, you will be more likely to get two or three ‘yes’ votes and less likely to get one or zero ‘yes’ votes as compared to the graph for 55%.”

Have students compare these graphs with the case of a true proportion of 60% (from yesterday). Bring out that each of these graphs represents a case of the binomial distribution for  $n = 3$  and a particular value of  $p$ .

- *Question 3*

For Question 3, students should generalize Questions 1 and 2 to find the probability distribution for a 3-person poll with true proportion  $p$ . If a hint is

needed, ask specifically what the probability of a “no” vote is for a particular voter, to bring out that this probability is  $1 - p$ . Students should get these results:

- $P(0 \text{ “yes” votes}) = (1 - p)^3$
- $P(1 \text{ “yes” vote}) = 3p(1 - p)^2$
- $P(2 \text{ “yes” votes}) = 3p^2(1 - p)$
- $P(3 \text{ “yes” votes}) = p^3$

Point out that these probabilities represent the general version of the case  $n = 3$  of the binomial distribution. You might point out that these expressions have coefficients 1, 3, 3, and 1 (although the 1’s are implicit) and that these numbers are binomial (or combinatorial) coefficients. You might also mention that they form a row of Pascal’s triangle.

- **Question 4**

The main focus of the discussion should be on Question 4, which leads into today’s activity, *The Theory of Polls*. Let students share ideas on the effect of changing the sample size.

This discussion should be viewed as preliminary. Through their work on *The Theory of Polls*, students should see that for larger polls, the graph should be “bunched” closer around the true proportion.

*“How ‘spread out’ would the bars be? Where would they cluster?”*

You can focus the class’s attention in this direction with general questions about how “spread out” the bars would be or where the bars would cluster. Students may have the intuitive sense that a larger poll is more likely to give a result close to the true proportion. However, if they don’t reach this conclusion on their own, tell them that they’ll see more about the effect of increasing sample size in the next activity.

*Note:* Students may make other observations about the effect on the graph of increasing sample size. For example, they may point out that there will be more “bars” in the graph and that each individual probability will be smaller (because there are more possible sample proportions). Do not neglect these ideas, but be sure to at least raise the issue of the “spread” of the data.

## 2. *The Theory of Polls*

With the homework discussion as background, tell students that their next activity will focus on what happens as poll size increases. Point out that for the sake of making clear comparisons among different sample sizes, the activity fixes the true proportion at 60%.

*“Why are we interested in these theoretical distributions?”*

Bring out that we are interested in the theoretical distribution of poll results in order to understand the reliability of polls. (You might point out that while an individual poll has many possible outcomes, some results are much more likely than others.) We want to know the likelihood of getting a sample proportion

“close to” the true proportion. Emphasize that “close to” is a vague term that will become more precise later in the unit. You can mention that this term is connected with the phrase *margin of error*.

One question of particular interest is the likelihood of correctly predicting the winner of the election. As students saw yesterday, if the candidate has the support of 60% of the overall population, there is about a 65% chance that a 3-person poll will show that candidate leading.

- *Hints as students work*

With this introduction, let students begin work on the activity in groups. As you observe them, you may decide that it’s worthwhile to bring the class together to go over a single case, such as the probability of getting two “yes” votes and three “no” votes (that is, a sample proportion of 40%). You can use the ideas below for the discussion of the activity as a guideline for this case.

You also may want to bring the class together when most groups are done with the case of 5-person polls, discuss this case, and review the use of combinatorial coefficients. Students can then return to their groups to work on the case of 9-person polls.

### 3. Discussion of *The Theory of Polls*

Today’s discussion of *The Theory of Polls* should lead to a strengthening of students’ understanding of how to find these probabilities and to the

observation that as sample size gets bigger, the probability bar graph begins to look more like the normal distribution.

- *Question 1: The five-person poll*

Begin by having club card students explain how to get each of the probabilities for Question 1a. For example, to find the probability of getting two “yes” votes and three “no” votes, they may begin with the fact that the probability of any *particular* sequence of two “yes” votes and three “no” votes (such as YNNYN) is  $.6^2 \cdot .4^3$ . (To clarify this, you may want to suggest that students view a sample of size 5 as a *sequence* of five individuals, like the sequence of games in a *Pennant Fever* problem, rather than as a set of five people chosen all at once.)

The presenter might then explain that there are ten such sequences (perhaps by making a list of cases) to see that the probability of getting two “yes” votes and three “no” votes is given by the expression  $10 \cdot .6^2 \cdot .4^3$ , which is .2304.

Ask students to explain what this number represents. They should be able to articulate that if the true proportion for the population is 60%, then about 23% of all 5-person polls will result in two “yes” votes and three “no” votes.

- *Using combinatorial coefficients*

Before moving on to study the probabilities more closely, we suggest that you review the use of combinatorial coefficients for expressing these values.

*“What symbol can be used to express the number of sequences with two ‘yes’ votes and three ‘no’ votes?”*

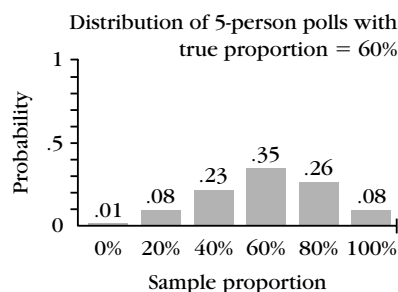
For instance, ask what symbol can be used to express the number of sequences with two “yes” votes and three “no” votes. Review that the number 10, in the expression  $10 \cdot .6^2 \cdot .4^3$ , is the combinatorial coefficient  ${}_5C_2$ . If needed, review the connection between the counting process here and that in the *Pennant Fever* scenario.

*“How can you find numerical values for combinatorial coefficients on your calculators?”*

Ask students how they can find numerical values for combinatorial coefficients on their calculators. As needed, go over the mechanics of this process. (It is not necessary in this unit for students to know the formula for computing these combinatorial coefficients. Use your judgment about whether to take time to review this.)

- *The probability bar graph*

After all the probabilities have been found, have a group display its probability bar graph for the 5-person poll, or use a transparency of the graph shown here. (This graph is included in Appendix C.)



*Comment:* In this graph, the possible poll results—that is, the sample proportions given on the horizontal axis—are shown as percentages, but they could also be shown as fractions or decimals.

You may want to compare this graph to the graph for the case of 3-person polls, or you may prefer to wait until after discussing Question 2 (the case of the 9-person poll).

In any case, look at Question 1b, which asks for the likelihood that the poll will show the candidate leading. In the case of a 5-person poll, there is about a 68% chance of this happening. (If students use the rounded values shown in the graph, they will find the sum  $.08 + .26 + .35$ , which is  $.69$ , but the actual probability is closer to  $.68$ .)

Students should see that the chance of a “correct” poll is still fairly low but that it is up slightly from the figure of 65% for the 3-person poll. (*Note:* This result will be referred to later today, and again in the homework discussion tomorrow.)

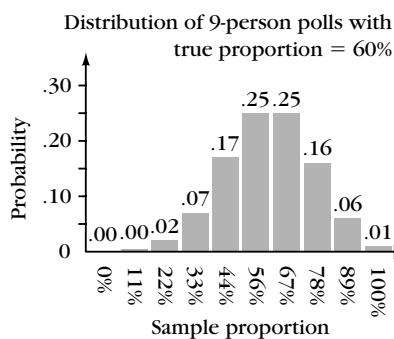
- **Question 2: The nine-person poll**

If students had trouble getting the probabilities for the 5-person poll, you may want to give groups more time to work on the case of the 9-person poll before proceeding with the discussion.

Let other club card students give probabilities for different possible results for a 9-person poll. Again, have students express the results using combinatorial coefficients.

At this stage in the unit, students should be fairly comfortable with the idea that in an  $n$ -person poll from a population with a true proportion  $p$ , the probability of getting exactly  $r$  “yes” votes is  ${}_n C_r \cdot p^r \cdot (1-p)^{n-r}$ . You can remind students that these probabilities represent the binomial distribution that they learned about in *Pennant Fever*.

Then have a group display its probability bar graph for the 9-person poll, or use a transparency of the graph shown here. (Again, the graph is included in Appendix C.) Notice that this graph has a different vertical scale from the graphs for  $n = 3$  and  $n = 5$ . Because there are more possible results, each result has a smaller probability than in the earlier cases. (You may want to emphasize that the probabilities for sample proportions of 0% and 11% are not actually 0, but that the probabilities shown in the graph are rounded to the nearest hundredth.)



*“What is the likelihood of a correct prediction?”*

Before comparing graphs, ask about the likelihood of a correct prediction (Question 2b). Students should see that this has now gone up to about 73% and that the chance for error is diminishing as the poll size grows. (This result, too, will be referred to again both today and tomorrow.)

- *Comparing the graphs*

*“What is happening to the graph of the theoretical distribution as the sample size gets larger?”*

Ask students to describe what is happening to the graph of the theoretical distribution as the sample size gets larger. To bring out the changes, you can use a set of graphs with a common scale and bars of a fixed width (see Appendix C). Try to elicit this conclusion, and post this principle.



**The larger the poll size, the more the theoretical distribution of sample proportions is concentrated around the true proportion.**

If students don't see a clear pattern, show them the graph for the 50-person case. If possible, however, hold off showing this graph until the discussion of the normal curve in the next section (“The Normal Distribution

and the Central Limit Theorem”). If anyone suggests that the results are getting closer to a normal distribution, you can jump ahead to that section. But be sure to come back to the ideas in the next few paragraphs.

*“What happens to the percentage of ‘correct’ polls as the poll size increases?”*

Also, ask what happens to the percentage of “correct” polls (that is, those that show the true leader actually ahead) as the poll size increases. Review the values found so far.

- For 3-person polls: Approximately 65% of polls are “correct.”
- For 5-person polls: Approximately 68% of polls are “correct.”
- For 9-person polls: Approximately 73% of polls are “correct.”

This should lead students to this very reasonable principle.

**The larger the poll size, the more likely it is for the person who is actually leading in the race to be the winner in the poll.**

*“Why wouldn't a pollster simply use a large poll size?”*

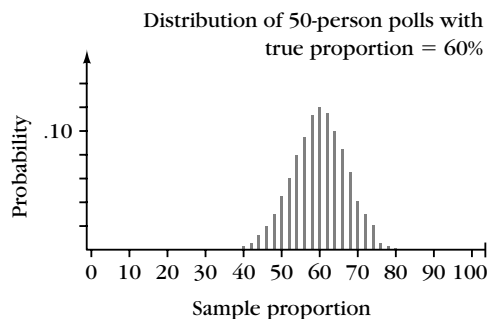
You might ask students, in light of this principle, why a pollster wouldn't simply use a large poll size to get a high probability of picking the winner correctly. Bring out that a larger poll requires greater resources, so pollsters need to balance cost against the desire for accurate results.

#### 4. The Normal Distribution and the Central Limit Theorem

*“Does the graph for 9-person polls suggest anything? Does its shape look familiar?”*

Have students look at the probability bar graph for 9-person polls, and ask if it suggests anything to them or if its shape looks familiar. Have them imagine what would happen as the poll size continued to grow. You may want to sketch a curve along the outline of the tops of the bars to suggest what happens as poll size increases.

If no one mentions the normal distribution, show students the graph below for the case of a 50-person poll. (This graph is included in Appendix C.) Point out that probabilities for sample proportions below 40% or above 80% were not zero, but they were so small that they don't show up.



You need not go over the computations for the probabilities, but you can review that they are calculated using the formula  ${}_{50}C_r \cdot .6^r \cdot .4^{50-r}$ .

Students may not remember many details about the normal distribution, but they should remember the general shape of the curve. (If this graph does not elicit recollection of the normal distribution, then mention the term yourself. Students can see examples of normal curves in *The Central Limit Theorem*—see Day 6.) Tell students that beginning tomorrow, they will have some activities and homework reviewing details about the normal distribution, including the role of standard deviation.

- *The central limit theorem*

Tell students that the connection between polls and the normal distribution is part of a profound principle in mathematics called the **central limit theorem**. (This unit treats only a special case of this theorem. Appendix B contains a more general statement of the theorem for your reference.)

Before stating the theorem, ask students to review the situation. That is, help them, as needed, to articulate that we are considering a population with a given overall proportion  $p$  in favor of the candidate. We take a poll of size  $n$  and find the proportion in favor of the candidate among the people polled. (You may want to review the terms *true proportion* and *sample proportion* and to remind students that we generally represent these by the symbols  $p$  and  $\hat{p}$ , respectively.)

Students have seen that for any given value of  $n$ , they can find the theoretical probability of obtaining each possible sample proportion. Therefore, for a given poll of size  $n$  and true proportion  $p$ , there is a theoretical probability distribution of sample proportions.

After reviewing the situation, post this statement of the central limit theorem.



**As the poll size gets larger, the probability distribution of sample proportions looks more and more like a normal distribution.**

*Note:* Presenting the central limit theorem this way assumes that the binomial distribution is a good model for finding the distribution of sample proportions, which means we are using sampling with replacement as our model for the polling process.

You may want to take this occasion to review this assumption (which is used throughout the unit) and to go over the fact that this assumption requires the overall population to be much larger than the sample size.

Remind students that there are many normal distributions. Also point out that the normal distribution that approximates a given poll depends on the true proportion for the overall population and on the poll size. Review the earlier observation that as the poll size increases, the distribution becomes more concentrated around the true proportion. (Students will see later in the unit that the mean for the “limiting” normal distribution is equal to the true proportion for the overall population.)

Tell students that there is no easy rule of thumb about how big a poll should be to have the theoretical distribution look “close enough” to normal. This depends on the true proportion for the overall population (and on what “close enough” means).

One common guideline is that if the proportion of “yes” votes in the whole population is  $p$ , then  $n$ , the sample size, should be chosen so that both  $np$  and  $n(1 - p)$  are at least 5. This is fairly easy to achieve if the population isn’t wildly unbalanced. For instance,

if  $p$  is between .2 and .8, then the guideline is satisfied as long as  $n$  is at least 25. Bring out that the probability bar graph students have for  $n = 50$  and  $p = .6$  looks quite a lot like the normal curve.

For the rest of the unit, we will assume that the sample sizes given in problems and those selected in student projects are large enough that the normal distribution is a good approximation. You can add this to the list of assumptions posted on Day 1.

- *How does the central limit theorem fit into the unit?*

*“How might you use the central limit theorem in the unit?”*

Ask students how they think they might use the central limit theorem in the unit. Let them share ideas (this will probably be fairly speculative at this point). Then tell them that they will be learning more about this theorem and about the normal distribution, and will see how these ideas can be used to understand the unit problem.

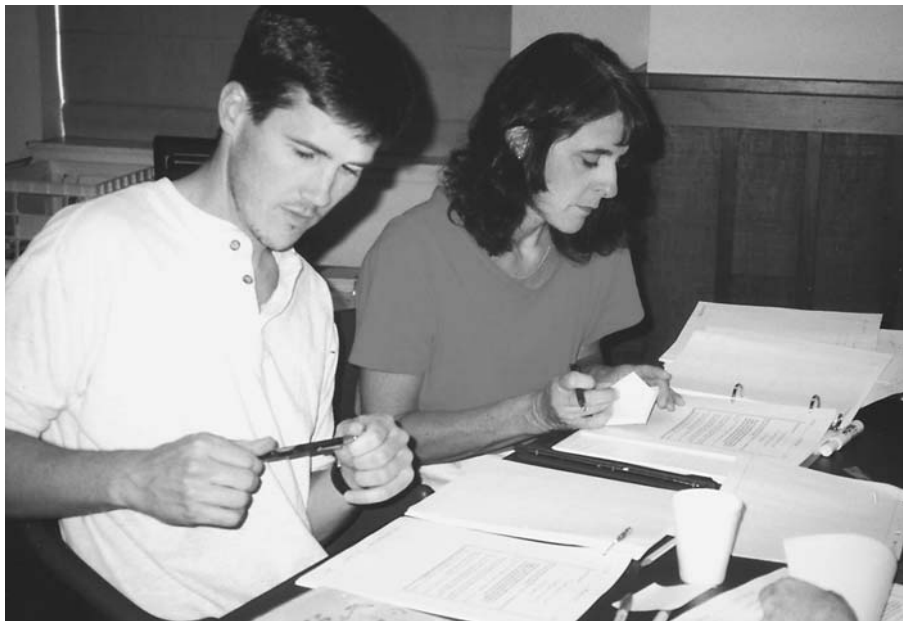
Also tell students that the central limit theorem applies to many other situations involving averaging. (You might briefly indicate how the sample proportion for a poll is a kind of average.) Emphasize that this theorem helps to explain why the normal distribution is so important in the study of statistics. The theorem shows that normal distributions pop up of their own accord every time we do averaging.

The supplemental problem *Another View of the Central Limit Theorem* gives students a chance to examine the principle in a slightly more general setting. However, that activity involves ideas about mean and

standard deviation that are discussed on Day 11 and developed further in *The Search Is On!* (Days 12 and 13), so we suggest that you not assign that activity until at least Day 13.

## Homework 5: Civics in Action

This assignment involves another look at the probabilities for correctly picking the winner of an election using a small poll.



**IMP teachers Kevin Drinkard and Claire Meranda prepare questions they can ask in class to help prevent their students from losing sight of the big picture.**