Graphs of the Theory

Intent

Students investigate the effect of changing the overall population (the true proportion) on the theoretical distribution of poll results.

Mathematics

This activity continues the theme of *The Theory of Three-Person Polls*, with students making graphs using other true proportions. Students observe that as the true proportion increases, the probability bar graph shifts to the right. They also develop expressions for the probability of each outcome in terms of the true proportion.

Progression

Graphs of the Theory is essentially a continuation of *The Theory of Three-Person Polls*. Students now construct probability distributions for 3-person polls in two more populations, which have different true proportions. They are then asked to generalize their work for a true proportion of p and to speculate on the effect of increasing the sample size.

Approximate Time

30 minutes for activity (at home or in class)
10 minutes for discussion

Classroom Organization

Individuals, followed by whole-class discussion

Materials

Optional: Transparencies of *Graphs of the Theory* blackline masters

Doing the Activity

This activity requires no introduction.

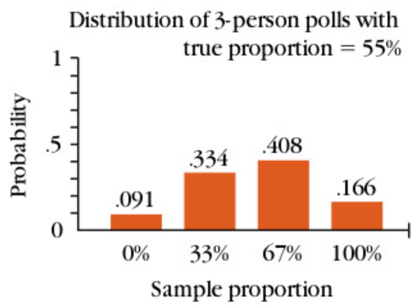
Discussing and Debriefing the Activity

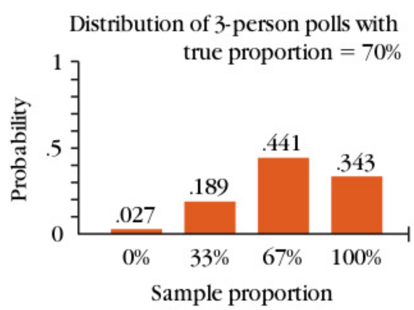
Have groups discuss Question 3 while one group prepares to present its graphs for Question 1 and another group does so for Question 2.

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 available as blackline masters. Note that because of rounding, the probabilities for the first graph do not total exactly 1.)





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

- "The graph for 70% has been shifted to the right, compared to the graph for 55%."
- "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 *The Theory of Three-Person Polls* and available as a blackline master).

Emphasize that each graph represents a case of the binomial distribution for $n = 3$ and a particular value of p.

Question 3

Students should generalize Questions 1 and 2 to find the probability distribution for a 3-person poll with true proportion p. If needed, ask specifically what the probability of a "no" vote is for a particular voter, and help students see that it is $1 - p$. Students should get these results:

- P(0 "yes" votes) = $(1 - p)^3$
- P(1 "yes" vote) = $3p(1 - p)^2$
- P(2 "yes" votes) = $3p^2(1 - p)$
- P(3 "yes" votes) = p^3

Explain 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 1s 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 *The Theory of Polls*. Let students share ideas on the effect of changing the sample size.

In *The Theory of Polls,* students will see that for larger polls, the graph will be "bunched" closer around the true proportion. Focus their attention in this direction with such general questions as, **How "spread out" would the bars be? Where would they cluster?** Students may intuitively sense that a larger poll is more likely to give a result close to the true proportion. (They'll see more about the effect of increasing sample size in the next activity.)

Students may make other observations about the effect of increasing sample size on the graph. For example, they may point out that there will be more bars 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.

Key Questions

How "spread out" would the bars be? Where would they cluster?

The Theory of Polls

Intent

This activity prepares students for the introduction of the central limit theorem.

Mathematics

Students use combinatorial coefficients to find the theoretical distribution of poll results for polls of various sizes. The discussion of the activity will strengthen their understanding of how to find these probabilities and lead to the observation that as sample size increases, the probability bar graph begins to look more like the normal distribution. The central limit theorem expresses this fact.

Progression

Students find theoretical probability distributions for polls of several sizes. In the follow-up discussion, the class reviews the use of combinatorial coefficients to get the probabilities and discusses the probability of a "correct" poll prediction. The class then compares the probability bar graphs for $n = 5$ and $n = 9$ and notes that the graph is beginning to resemble the normal curve. The graph for the case $n = 50$ is used to emphasize the resemblance to the normal curve and to state the central limit theorem for this context.

Approximate Time

40 minutes

Classroom Organization

Small groups, followed by whole-class discussion

Materials

Optional: Transparencies of *The Theory of Polls* blackline masters

Doing the Activity

Tell students they will now focus on what happens as the poll size increases. Point out that for the sake of making clear comparisons among different sample sizes, the true proportion is fixed at 60%.

Talk about the fact that we are interested in the theoretical distribution of poll results in order to understand the reliability of polls. 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.

One question of particular interest is the likelihood of correctly predicting the winner of the election. As students saw in *The Theory of Three-Person Polls*, 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.

Let students begin work on the activity in groups. As you observe, you may decide 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%) in Question 1a. Use the ideas below 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. Groups can then work on the case of 9-person polls.

Discussing and Debriefing the Activity

Question 1: The 5-Person Poll

Have several students explain how to get each probability 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 they view a sample of size 5 as a *sequence* of five individuals, like the sequence of games in a *Pennant Fever* problem (see *"Pennant Fever" Reflection*), 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), so 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

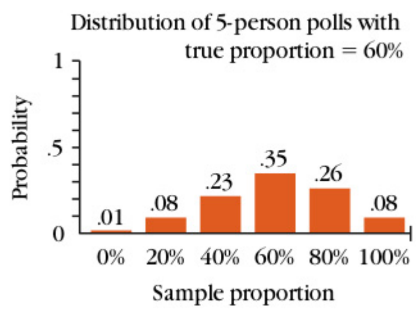
Before moving on to study the probabilities more closely, review the use of combinatorial coefficients for expressing these values.

Ask, **What symbol can we use 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 numeric values for combinatorial coefficients on your calculator? As needed, go over the mechanics of this process. (It is not necessary for students to know the formula for computing these combinatorial coefficients in this unit. Use your judgment about whether 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. (The possible poll results—that is, the sample proportions given on the horizontal axis—are shown here as percentages, but they could also be shown as fractions or decimals.)



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

In the case of 5-person polls, as in Question 1c, there is about a 68% chance. (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 3-person polls.

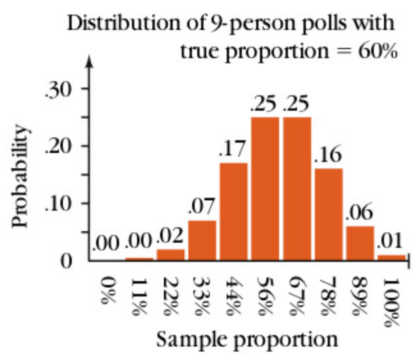
Question 2: The 9-Person Poll

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

Let other students give probabilities for different possible results for a 9-person poll, expressing the results using combinatorial coefficients.

At this stage, 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 ${}_nC_r \cdot p^r \cdot (1 - p)^{n-r}$.

Then have a group display its probability bar graph for the 9-person poll, or use a transparency of the graph shown here. 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. This rounding is also the reason that the sum of the probabilities is not exactly 1 for either the 5-person or the 9-person graph.)



Before comparing graphs, ask, **What is the likelihood of a correct prediction?** Students should see that this has now risen to about 73% and that the chance for error is diminishing as the poll size grows.

Comparing the Graphs

Ask, **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. 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 the results are getting closer to a normal distribution, you can jump ahead to that material, but be sure to come back to the ideas in the next few paragraphs.

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:

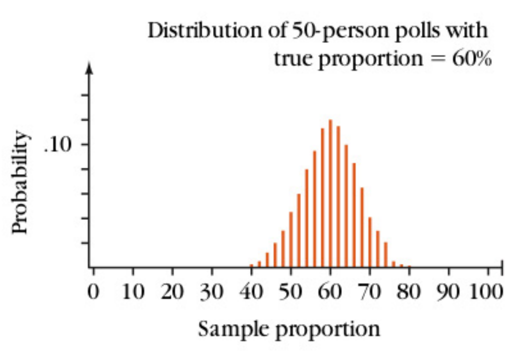
> **The larger the poll size, the more likely it is for the person who is actually leading in the race to be the winner in the poll.**

You might also ask, in light of this principle, **Why wouldn't a pollster simply use a large poll size to get a high probability of picking the winner correctly?** A larger poll requires greater resources, and pollsters need to balance cost with the desire for accurate results.

The Normal Distribution and the Central Limit Theorem

Have students look at the probability bar graph for 9-person polls, and ask, **Does this graph suggest anything to you? Does its shape look 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 also included on a blackline master.) Point out that probabilities for sample proportions below 40% or above 80% are not zero, but are so small they don't show up.



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, mention the term yourself. (Students can see examples of normal curves in *The Central Limit Theorem*.)

Explain 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. *The Central Limit Theorem: An Overview for Teachers* contains a more general statement of the theorem for your reference.)

Before stating the theorem, ask students to review the situation. Help them 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 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 probability distribution of sample proportions.

Then, 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.**

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.

Tell students 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).

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 during the discussion of *The Pollster's Dilemma*.

How Does the Central Limit Theorem Fit Into the Unit?

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

Also explain 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.

Key Questions

What symbol can we use to express the number of sequences with two "yes" votes and three "no" votes?

How can you find numeric values for combinatorial coefficients on your calculator?

What is the likelihood of a correct prediction?

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

What happens to the percentage of "correct" polls (that is, those that show the true leader actually ahead) as the poll size increases?

Why wouldn't a pollster simply use a large poll size to get a high probability of picking the winner correctly?

Does the graph for 9-person polls suggest anything to you? Does its shape look familiar?

How do you think you might use the central limit theorem in the unit?