Gush vs. Bore: A Look at the Statistics of Sampling

Open the Fathom file Random_Samples.ftm.

Imagine that in a nation somewhere nearby, a presidential election will soon be held with two candidates named Gush and Bore. In order to predict the outcome of the election, a news reporting agency randomly contacts 100 voters and asks who they will vote for. The results of the survey are shown in the table along the left side of your screen.

Next to the table of individual results is a smaller table showing how many of the people surveyed favored each candidate.

Looking at the results of the poll, one news agency reports, “62 percent of voters favor Gush.” (Your numbers may differ.) While that’s true of the voters surveyed, is this really the case for the population at large? Let’s see what happens if we take a survey of another 100 voters.

1. Right-click on the Poll table and select “Rerandomize.” This will take a survey of another 100 voters. What percentage of voters surveyed favored Gush this time?

2. Rerandomize the survey results five more times, each time recording the percentage of voters favoring Gush in the table below.
3. Why are there differences in your results? In theory, out of a survey of 100 people, what’s the largest number that might support Gush? What’s the smallest number?

4. If some all-knowing person told you that exactly 55 percent of the entire population favored Gush, would you say that any of the surveys you recorded were fairly inaccurate?

5. What was the average value of the measurements you took in Problem 2? How does it compare to the true percentage (55%) who support Gush?

As you most likely saw, different surveys will yield different conclusions as to how many voters favor Gush. So is it possible to trust the statistics we see on the news?

Let’s see what happens as we take more and more surveys from the same population. Right-click on the “Data from Polls” collection at the top of the screen and select “Inspect Collection” from the menu that appears. A window labeled “Inspect Data from Polls” will appear on your screen. In this window, check the “Animation on” box and enter “10” in the space next to “measures” a few lines below. The screen will look something like this:
Click on the button labeled “Collect More Measures.” Fathom shows little balls moving from the “Poll” collection to the “Data from Polls” collection at the top of the screen. Each ball represents another survey of 100 voters.

In the “Data from Polls” tables on the right side of your screen you now see the results of the 10 surveys we just took. In the first table, the “snapshot” column shows the percentage of voters which favored Gush in each survey. The table on the right shows the mean percentage of voter support for Gush reported over all 10 surveys, the minimum and maximum reported percentages among the 10 surveys, the number of surveys taken (10), and the standard deviation in the results.

6. Look at the first histogram shown below the tables. What percentages of voter support do the highest bars correspond to? Do any of those bars correspond to 55 percent?

7. Now look at the mean reported percentage of voter support as it’s shown in the table on the far right above the graph. How does it compare with the true level (55%) of voter support for Gush?

8. Neither the highest bar in the histogram nor the mean reported percentage has to agree with 55%, the level of support for Candidate Gush over the entire population, but let’s see what happens as we examine more surveys. In the “Inspect Data from Polls” box,
uncheck the “Animation on” box and change the number of measures being taken from 10 to 200. Click on “Collect More Measures.” After the computer is done with that, take another look at the mean percentage and standard deviation of voter support for Gush in the right-hand Data from Polls table. Is it closer to 55% than before? Also check the histogram. How is the graph different, and why do you think this is?

9. Look at the second histogram in the file. It shows the same boxes as before, but it also shows the graph of a function that we will talk about later in the worksheet. For now, just notice how well the rectangles fit the curve shown.

10. Let’s see how widely spread the survey results are. In the left “Data from Polls” table, right-click on the heading of the “snapshot” column. From the menu that pops up, click on “Sort Ascending.” This will reorder all the surveys based on the percentage reported of voters favoring Gush. We’d like to see between what values 95% of the reported percentages fall between. For that reason, we’re going to ignore the top 5 and bottom 5 responses (i.e. the 2.5% at each extreme).
   a) Record the following values of “snapshot”:

<table>
<thead>
<tr>
<th>6th value</th>
<th>195th value</th>
</tr>
</thead>
</table>

   b) What’s the average of your answers to part (a)?

c) How far apart is the average you found in part (b) from the two answers you found in part (a)?

d) Now fill in the blanks in the following statement: In 95% of the surveys conducted, the reported support for Gush lay in the interval of

   Answer to (b) ± Answer to (c)

Taking 200 surveys of 100 people each is a big task, and not one that many pollsters would attempt. Instead, a pollster usually conducts a single survey and focuses on talking to as many people as practical during that survey. Let’s see what happens to our reported survey results when we increase the sample size.

11. Write down the statistics currently shown in the “Data from Polls” table on the far right side of your page.
12. Right-click on the “Poll” table on the far left side of your screen. Select “New Cases...” from the menu that pops up, and in the “Add Cases to Poll” window that opens, enter 100 and click on OK. Now each survey conducted polls 200 people. If you click on the “Collect More Measures” button in the “Inspect Data from Polls” window, 200 new surveys will be taken, where this time each survey will poll 200 people. How do you expect that the reported percentages of support for Gush will change? Will the mean of the results change? Will the histogram?

13. Click on “Collect More Measures.” When the computer is finished, record the data from the right-hand “Data from Polls” table below:

<table>
<thead>
<tr>
<th>mean</th>
<th>min</th>
<th>max</th>
<th>count</th>
<th>stdDev</th>
</tr>
</thead>
</table>

Also look at the first histogram below the tables. As the sample size changed, what else changed? What stayed the same? Did this agree with your predictions?

14. Let’s check on how widely spread the new survey results are. As you did in Problem 10, sort the values in the “Data from Polls” table from largest to smallest. Once again, we’re going to ignore the top and bottom 2.5% of responses—for 200 surveys conducted, this means throwing away the top and bottom 5 values.

   e) Record the following values of “snapshot”:

<table>
<thead>
<tr>
<th>6th value</th>
<th>195th value</th>
</tr>
</thead>
</table>

f) What’s the average of your answers to part (a)?

g) How far apart is the average you found in part (b) from the two answers you found in part (a)?

h) Now fill in the blanks in the following statement: In 95% of the surveys conducted, the reported support for Gush lay in the interval of
15. As you did in Step 9, add some more cases—this time, **add 300 more cases**, so that each survey contacts 500 people for their opinion. Then **click on “Collect More Measures”** in the “Inspect Data from Polls” window, and once again record the data from the right-hand “Data from Polls” window. How are these data shown in the first histogram shown below the tables? Are the patterns you observed in the last problem upheld?

<table>
<thead>
<tr>
<th>mean</th>
<th>min</th>
<th>max</th>
<th>count</th>
<th>stdDev</th>
</tr>
</thead>
</table>

16. As you did in Problems 10 and 14, complete the statement: In 95% of the surveys conducted, the reported support for Gush lay in the interval of

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>

Clearly, the more surveys we take, the better the likelihood we have of being able to accurately report the true percentage of the population that favors Gush in the election. However, time and resources do not often allow pollsters to take more than one survey at a time. Instead, pollsters will often try to predict how likely it is that a single survey they’ve conducted reflects the true mean. How do they do this?

17. As you have answered the worksheet questions so far, what have you noticed about the shapes of the histograms that have been graphed?

You may have noticed that the heights of the bars in the histograms have seemed to follow what is popularly known as the bell or normal curve. This is no accident. Statisticians have proven that when a population is surveyed several times, as our population of voters has been, then the results of each survey tend to be normally distributed about the mean. In a nutshell, you can think of the bell curve as showing how likely it is that a survey will report a given percentage as favoring Gush. The bell curve reaches its highest point at the true mean of the population, and as you may have noticed in the last few problems, larger sample sizes produce narrower curves.

18. Look at the lower of the two histograms. Like the upper histogram, it now shows the results of polling 500 people 200 different times. You’ll notice, however, that the blue curve no longer fits the data that well. That’s because the standard deviation of the
normal curve modeling the distribution of the sample means depends on the sample size. **Double click on the formula displayed below the bottom histogram.** In the “Expression for function” window that pops up, carefully replace the “100” in the formula by “500,” as shown.

![](image)

Click on OK, and examine the new curve that you have graphed. How well does it fit the boxes drawn in the histogram? Can you make an educated guess what the standard deviation is for samples of size n?

Let’s now talk about how this all fits together. As we’ve said, results of surveys tend to cluster in a normal distribution (a bell curve) around the true population-wide value. And the more samples you take, the more those set of samples look like a normal distribution with the same mean and with standard deviation as above. But if all you know is the result of one survey, and certainly not the population-wide value, then what can you do?

In Problems 10, 14, and 16 above, you calculated a range of values that tells you two numbers between which about 95% of all possible surveys will report their values. We will soon be able to calculate the length of this interval on paper (rather than experimentally) by learning a bit more about normal distributions, which we’ll do later on in our class. We can then use that length like this: If we know that 95% of all polls will return a voter support level for Gush that is within $P$ percentage points (let’s say) of the true population support level, then we can turn this information around. If 95% of the time the result of a random survey will lie within $P$ points of the population-wide value, then that means that the population-wide value will lie within $P$ points of our sample’s result (assuming our survey was one of the lucky 95%). That allows us to
specify a range about our survey’s result, called the *confidence interval*, in which we can be reasonably certain that the population-wide value lies.

19. Take a random poll from your last experiment and look at the interval centered on the poll average with width equal to the 95% confidence interval you calculated in problem 16. Is the true mean 55% contained in this interval?

**Going Further:**

1. If you wanted a poll to have a 95% confidence interval that was half as large as another poll (e.g. +/- 2% instead of +/- 4%), what would you have to do to the size of the poll taken? (Look back at Problem 18.)

2. It seems like political polls still vary a lot more than the confidence interval tells us they should. Why might this be?

3. How would you word the Obama-McCain article title first cited?