Researchers Voice: The Effects of Sample on Polling

PUBLISHED BY MICHAEL LIEBERMAN ON JULY 20, 2020

Editor's Note: With the US presidential election just four months away, the polls are going to heat up. But, as we all know from 4 years ago, the results vary widely from one polling company to another and, in 2016, overall failed to accurately predict the outcome. Why – well it is all about the sample. This is true for any research conducted – be it political polling, brand, product, or employee research. With this in mind, we asked Michael Lieberman to take us back to the principles of sampling as a reminder of the different elements at play.

Sample Defined

Definition: Statistical Inference: the theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling.

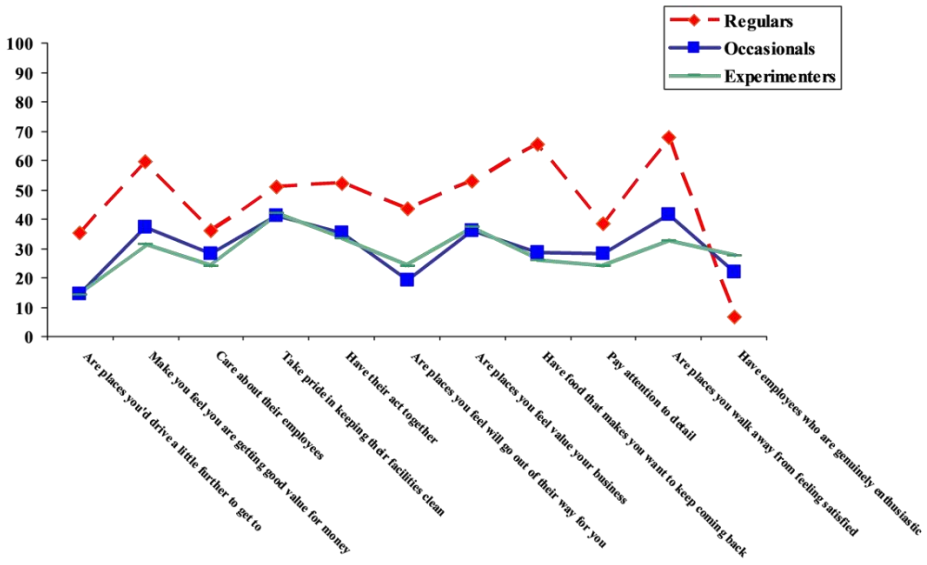
Let's put that another way though an example. A burger QSR (quick service restaurant) serves approximately 16 million people a day worldwide. They would like to know the key drivers to it overall sales among, say, 11 attributes. How much would it cost to ask all of their customers and then do the appropriate multivariate analysis? Well, let's figure it costs about $2 per answer, it would cost $32 million.

Or the burger QSR could take a representative sample of, say, 1,000 customers for three subgroups—a total of 3,000 surveys—and run the same analysis. The law of statistical inference suggests that the answers provided by the sample of 3,000 would be very much the same as the answers provided by all of their customers, within an error range of about 3%.

Below is an example chart where the burger QSR has asked 3,000 customers, one from each of the three groups represented on the chart, what they think about 11 attributes.

Characteristics Statements Closely Associated With Burger QSR

Image, Atmosphere, Service



The burger QSR would then be able to make decisions on marketing and messaging based on the sample Regulars, Occasionals, and Experimenters without having to ask all the people in these categories worldwide. With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. We use inferential statistics to try to infer from the sample data what the population might think. We use inferential statistics to make inferences from our data to more general conditions through the use of probability models.

Probability Models

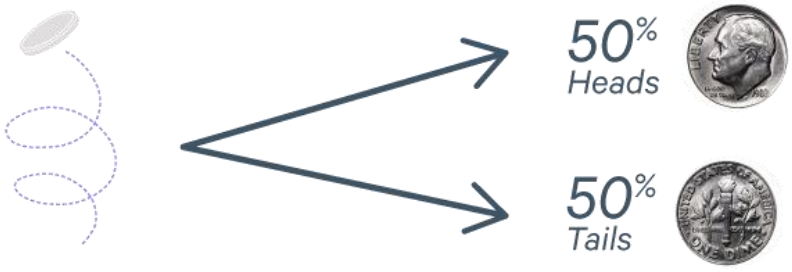
Survey research and political polling are heavily reliant on probability models. That is, the projection of the answers of a few onto an entire population. These projects are always quite accurate due to the concept of probability models.

A probability model is a mathematical representation of a random phenomenon. It is defined by its sample space, events within the sample space, and probabilities associated with each event. The sample space S for a probability model is the set of all possible outcomes. Probability is the number of times our outcome would prevail.

In other words, you have:

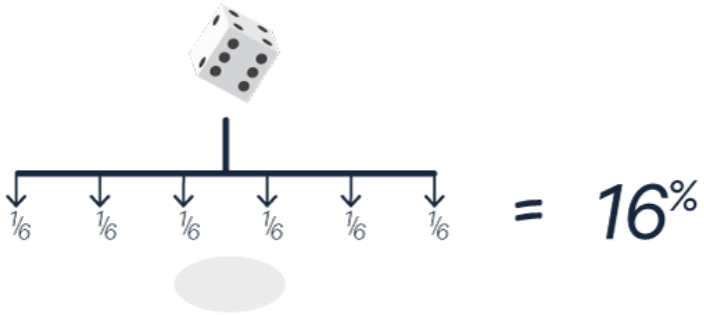
$$probability = \frac{your\ events}{\sum possible\ events}$$

For example, when you flip a coin, there are two possible outcomes, heads or tails.



If you are looking for the probability of tossing heads, the answer is ½, or 50%.

When you roll a dice, the chances of rolling a 4 is:



When you roll a dice, the chances of rolling a 4 is 1/6, or 16%.

The Accuracy of Sample Data

This is a broad simplification of a complex definition but explains why our survey sample, though representative of the entire burger QSR, is correctly within a certain probability. Are the answers from 1000 able to project onto a total sample size of 100,000. The confidence interval is 3%, so we are 97% sure that our answers from the representative sample is correct.

Let's move on to the accuracy of survey research to gauge brand equity, consumer sentiment, political polls, etc. In these cases, we use probability samples to get our most accurate picture.

Probability Sampling vs. Non-Probability Sample

A probability sampling method is any method of sampling that utilizes some form of random selection. In order to have a random selection method, you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen. That is, if a study is being done on grocery store shoppers, and your sample is grocery store shoppers, randomly choosing 1000 people to survey is an example of a probability sample. In general researchers prefer probability sampling.

The difference between nonprobability and probability sampling is that nonprobability sampling does not involve random selection and probability sampling does.

Examples of nonprobability sampling include:

- Convenience, haphazard or accidental sampling – members of the population are chosen based on their relative ease of access. To sample friends, co-workers, or shoppers at a single mall, are all examples of convenience sampling.
- Judgmental sampling or purposive sampling – The researcher chooses the sample based on who they think would be appropriate for the study. This is used primarily when there is a limited number of people that have expertise in the area being researched. For example, surveying CEOs on the chances of their firm purchasing a private jet.
- Case study – The research is limited to one group, often with a similar characteristic or of small size.
- Panel sampling: No matter how thorough panel data might be, it cannot be considered a random sample because respondents who may qualify for the survey and who are not contained in the panel will not be surveyed.
- Quotas – A quota is established (e.g. 65% women) and researchers are free to choose any respondent they wish as long as the quota is met.

Non-probability and quota samples are very common in survey research. A good example is the phenomena of over-sampling. A luxury car brand would like to gauge its corporate brand love story. They take a random sample of Drivers 18+, asking questions of brand image. Then they oversample the key demographic, owners of the automobile brand between the ages of 34-55—these folks have more communication value than the population at large. The luxury car brand owners have an average age of 45 in 2019. They oversample young, affluent customers because that is the market they would like to penetrate. This should give them the insights they are looking for.

Weighting Sample

Weighting adjusts the poll data in an attempt to ensure that the sample more accurately reflects the characteristics of the population from which it was drawn and to which an inference will be made. Weighting amplifying answers of people underrepresented or lower voice of those over-represented.

Weighting is used to adjust the relative contribution of the respondents, but it does not involve any changes to the actual answers to survey questions. A good way to view case weights is to show a group is [Percentage in Sample Population]/[Sample in Actual Sample].

The Table, below, shows how to create demographics weights for a study. In more sophisticated weighting schemes weight from all three dimensions can be balanced.

	Sampled	Population	Case Weight
Gender			
Men	45%	50%	1.11
Women	55%	50%	0.91
Age			
18-24	34%	30%	0.88
25-34	21%	20%	0.95
35-44	18%	15%	0.83
45-54	19%	15%	0.79
55+	8%	20%	2.50
Ethnicity			
White	55%	60%	1.09
African-American	25%	23%	0.92
Hispanic	12%	10%	0.83
Other Ethnicity	8%	7%	0.88

To display a simple use of weights, below is a table of weighted average, the grading system of a statistics class. The percentage of grade is, in fact, a case weight. The final grade is a weighted average.

	Percentage of Grade	Grade
First Exam	25%	80%
Second Exam	25%	72%
Excel Project	10%	80%
Final Exam	40%	83%
	Final Grade	79%

Non-Response Bias

Finally we will touch on the subject of non-response bias. Nonresponse bias occurs when some respondents included in the sample do not respond. The key difference here is that the error comes from an absence of respondents instead of the collection of erroneous data. Most often, this form of bias is created by refusals to participate or the inability to reach some respondents.

Nonresponse is a problem for survey quality because it almost always introduces systematic bias into the data. This results in poorer data quality and can significantly bias any estimates derived from the data. There are several techniques researchers can use to minimize nonresponse and to offset the bias it introduces into data. During the data collection period, researchers can use:

- Callbacks
- Incentives
- Oversampling
- Weighting Up known Non-Response Groups.

When designing a study, statistical inference, probability modeling and weighting vital for methodological and ethical reasons, as well as for reasons financial resource. The understanding of the terms and effects of the subjects covered in this brief summary can build an understanding and foundation in the field of empirical survey research.

Michael Lieberman is founder and president of Multivariate Solutions, a New York consulting firm offering comprehensive statistical consulting. He can be reached at 646-257-3794 or at michael@mvsolution.com.