PROBLEMS WITH STATISTICS

Athenaeum Society

Robert Y. Harper

12/7/2017

At this time last year, we were all recovering from a barrage of advertisements, news reports, and blogs about what people thought about various candidates, who might vote, who might not vote, how registered white male voters under 35 might vote, and who might actually be elected. The nattering nabobs bandied about a lot of numbers and statistics, which caused me to reconsider what I have learned about statistics. After some musing, I decided that you, Mr. President, Mr. Secretary-Treasurer, Mr. Fellow Presenter, and fellow members of the Athenaeum Society, will spend the next little bit listening to me drone on with some musings about statistics and their pitfalls. Make your exit now before you fall asleep.

Everyone in this room uses statistics: indeed, almost every human uses statistics or at least probability. Medical people look at drug or therapy studies that conclude that 90% of those tested had a better outcome than those given a placebo in a double blind study. Social scientists point out that 20.8% of Christian County residents live below the poverty line.[i] Manufacturers sample products to look at defect rates. Educators report that over 90% of school children graduate from public high schools in Christian County. Attorneys weigh their likelihood of success given the facts in a case and the conviction rates of particular judges in particular circumstances. If you are facing a hanging judge, you are probably much more likely to settle out of court. Politicians segment the population in a variety of ways, then look at polls to see how they are doing with these segments. They also look to polls to give them an indication of how the public feels about various issues. Marketers poll the public regularly to try to understand why people buy what they do. Insurers calculate the odds of you making particular claims; they may take age, gender, and certain behaviors into account in assessing the probabilities and then set your premium based on their estimates. Auditors using sampling

guidelines to examine your financial records; they do not have enough time to examine every single entry in your records, therefore they spot check. Even you, fellow Athenaeum members, calculate the likelihood of an enjoyable evening. Without any other information, you figure on two papers of about 20-25 minutes containing some recondite knowledge that will expand your intellectual horizons. If you know who the presenters are, you make a more informed calculation about the odds of an enjoyable evening: if William Turner is presenting, you figure that he will give an entertaining paper on history with a duration of up to 20 minutes; if I am presenting, you figure on a paper lasting at least 25 minutes, with mixed odds on the probability of enjoyment. Tonight the odds were not in your favor.

Let us first start off with a definition of statistics. While we are all educated men and know a statistic when we see it, here is the definition from Webster's New World Dictionary: "facts or data of a numerical kind, assembled, classified, and tabulated so as to present significant information about a given subject."[ii]

In addition, we should also review a few basic terms. You probably remember that a mean is number resulting from adding up all of the numerical answers and dividing by the number of answers. The median of a population is the number you find if you line up all of the responses from greatest to least and pick the one right in the middle. The mode is the answer given most frequently. This statistical measure is perhaps most useful when looking at responses given on a scale, say from 1-10.

Statistics are also related to odds or probability. Since most statistics rely on a sample of the population, statisticians usually speak of the probability that the result represents the truth about the population. When they say that with 95% certainty, the answer is x, then they mean

that the probability is that the true answer is within two standard deviations of x. Usually, they will give you the standard deviation as well.

As we think about probability a little, it is important to think about probabilities that are independent and probabilities that are dependent. An example of an independent probability is a coin toss. The odds of heads or tails is the same each time you toss the coin, regardless of what the previous outcome was. Although tossing the coin a large number of times should give you an equal number of heads and tails, it will not actually do so. In 1,000 tosses, 500 heads and 500 tails is one possible outcome out of 1,000, so statistically, while it is the expected outcome, it is unlikely to actually come to pass. Examples of dependent probabilities abound in actuarial work. The average life expectancy of an American male may be 76.3 years. However, once he reaches that age, we do not expect him to die immediately. In fact, he is expected to live another 10.2 years.[iii]

Another useful concept is that of the outlier. Sometimes a sample results in a data point or two that just do not fit the curve or the distribution. This may happen because of errors by the subject or the pollster, or it may happen because you have a subject who is completely unlike the rest of the population. For instance, if you time Athenaeum members on how long it takes them to run a 100 yard dash, you will find some kind of distribution, perhaps even a normal distribution. If we added Hussein Bolt or a participant on the television program "My 600 Pound Life" to our Society, their times would lie significantly outside the norm for the rest of the Society.

Usually, statistics are an attempt to convey information about a population. We often assume that a population is normally distributed. As a reminder for those of you whose last statistics class was more than 5 years ago, a normal population is one in which the mean answer

has the greatest likelihood of occurring and in which responses greater or lesser than the mean occur with equal and symmetric likelihood. When graphed, the probability gives a bell curve. Sometimes the curve is fairly flat, sometimes quite peaked, but in all cases symmetric. We use the concept of standard deviation to describe the width of the bell. By definition, 65.26% of a population fits within one standard deviation of the mean and 95.44% fits within two standard deviations of the mean.[iv]

One use of standard deviation is to determine statistical significance. Two data points can differ, but in statistical terms not be different at the same time. If the two points are within one, or sometimes two, standard deviations of each other, it is said that the difference is not statistically significant. An example might be that Christian County students increased their average ACT score by 1.0 points. While this looks good, if the standard deviation for a test score is 1.0 point, then statistically speaking, there was no improvement. You may ask how there can be standard deviation on a test? Well, a person taking the same test might score differently on different days, depending on factors such as sleep the night before, breakfast that morning, whether or not his girlfriend announced she was pregnant the night before, or even the temperature of the test room. Furthermore, there are several variations of the test given each year; none of them will yield exactly the same score from the same student. Therefore, standardized tests have standard deviations.

It is important to note that not all populations are normally distributed. For instance, there are binary populations: except in very rare cases, you have either male or female genitalia. Sexuality is probably not normally distributed. Most people will self-identify as straight – maybe a zero or one on the Kinsey scale. A few will identify in the bi-sexual range of 2, 3, 4 or

5. Then there will be small kick up at the tale amongst those who self-identify as strictly homosexual, a 6 on the Kinsey scale.[v]

As an aside, the Kinsey scale is an interesting example of the misuse of statistics. Most people understand the Kinsey report to say that 10% of the American population is homosexual. What the Kinseys really reported was that 10% of the population at the time of the study was engaged in a same-sex relationship. They did not say that this portion of the population was exclusively homosexual for a lifetime.[vi] Many of you can probably think of someone, usually a woman, who engaged in an exclusive relationship with a member of the same sex, but later returned to heterosexual relationships. Although I won't, I could poll the room and create a statistic. While interesting in itself, the whole concept of this poll raises other interesting questions about statistics, in this case related to recall. Some of you, when polled, would say no, you do not know anyone who briefly engaged in an exclusive same-sex relationship. You would not be lying, but it could be that, at the time of polling, you merely did not recollect someone from your misspent youth who fit the category. Later on, with the aid of some peer, you might be brought to recall the individual. Others among you might honestly give an answer of no because the truth was hidden from you – the population of your acquaintance might hold someone with such a relationship who has carefully kept the knowledge of said relationship from you. Assuming that you have all tried to answer truthfully, we might still not achieve an accurate representation of the true character of the population just because of problems related to recall.

There are numerous other distributions including exponential and Chi squared. Exponential distributions look like a curved downward sloping line, similar to a demand curve. An example might be the probability that a truck of grain will be unloaded in a certain amount of

time. Usually the Mill can unload a trailer of grain in 30 minutes, including the time to sample and test the load. However, there will be occasion when it make take 45 minutes; there will be just a few occasions when it takes 60 minutes, and maybe even an outlier taking 90 minutes. You can count yourself fortunate that my fascination with statistics does NOT extend far enough to subject you to Chi squared distributions and their various degrees of freedom.

Almost every statistic we see has been derived from a sample. Those creating the statistic then attempt to project the answers onto the entire population. Sometimes, the population is the entire universe, such as the population of the world. Usually the population is more limited, such as the population of the United States, American voters, white male Republican voters, or Athenaeum members. Good pollsters will always reveal the population they have tried to poll, however, those reporting and repeating the results, whether on television, in the newspaper, or in blogs, are not always careful about repeating this information. We need to ask who the population was, especially if the statistic seems in any way odd. For instance, if you hear that 3 out of 5 Americans prefer Pepsodent, you might want to ask which population was polled. While Pepsodent is a perfectly nice brand of toothpaste, it is not the best seller by a long margin.

As already mentioned, statistics are generally trying to convey information about a population. Sometimes the population is the entire universe, but usually it is more limited. If we really wanted to know something about the entire population, we would have to conduct a time-consuming expensive survey of everyone in the population. The only statistics that come close to this level of inclusion are the decennial U.S. population surveys, which do try to obtain basic information about the entire population of the United States. Even this survey is not fully inclusive, for a variety of reasons, so it also relies on projections in an attempt to capture under-

represented populations. For instance, people on the welfare are often reluctant to participate out of concern that the information given may somehow be used against them to reduce their benefit, so they hide from a pollster or give incomplete information. Therefore, U.S. Census Bureau statisticians make estimates to try to capture the citizens not captured by the surveys. Some geographic areas are also harder to capture, such as rural Alaska, so different projections would be made in that area.

There are statisticians whose job is to create representative samples. They look at the population to be studied and determine the size and composition of the sample necessary for pollsters to declare that their results, with 95% certainty, represent the actual traits of the population. For instance, a long time ago, some statistician created a sampling plan for grain trucks. It is not feasible for elevator operators to sample every kernel of grain in the truck, therefore we probe a truck in a few spots determined by some statistician, probably in Kansas.

Most surveys do not try to sample the entire population; they rely on sampling a subset of the population. Again, as a reminder, the population itself might be limited, such as polls of registered white males under 35 who intend to vote. If you are going to sample a population, how do you do it? You want your sample to be random, but this is difficult to achieve. Historically, phone surveys using land lines were felt to be fairly good because most people had phones, most people answered the phone, and most people would participate in a survey. However, the poor were less likely to have telephones, and a pollster could wind up with underrepresentation if he relied exclusively on telephonic polling. He might project his numbers up a little or might have to make some extra phone calls to try to obtain enough representation in that subcategory if he needed to capture it. Many of you have probably had the experience of being asked if you were willing to participate in a poll, being asked two or three demographic

questions about age, gender, race, and income and then being thanked for your time. The pollsters were screening you in order to obtain statistically valid representative samples of their population. If they are polling 1,000 people to obtain information about all Kentuckians, they cannot have all of their responses come from white professional males.

In today's world, telephone sampling is even less reliable. Fewer people have land lines, fewer people answer phone calls, and fewer people are willing to participate in surveys, especially ones run by a group with which they are not familiar. As a digression, if a pollster has to rely primarily on participants who are familiar with the group employing him, then his sample is already biased, unless his population is exclusively people familiar with the organization running the survey. So, since people have a tendency to participate only in surveys given by groups they like, you should be wary of the statistics given by organizations, especially politically focused ones. Now, back to the issue of sampling using telephones. Cell phones, while callable, are even less statistically reliable than land lines used to be. While a large percentage of the population does have cell phones these days, the usage is underrepresented among the poor and the elderly. There are also certain geographic areas with poor coverage where usage is below average. And, usage is above average among young, non-poor people. In addition to usage issues that impact the reliability of statistics attained through cell phone surveys, there is also the issue of who answers. Since cell phones show the number and sometimes name of incoming callers , many, if not most, of us will not answer calls from numbers we do not recognize, especially from an out of area number. This reluctance to participate further degrades the reliability of the statistic obtained.

There are, of course, alternative ways to sample populations. You can sample through the mail. However, then you have the problem of persuading participants to return the surveys.

Generally, with mailed surveys, only people really interested in the issue will return the survey. If you are asking political questions, the people with the strongest opinions, both positive and negative will participate, but not those whose feelings are ambivalent or mild. Your results will tend to skew to the edges, especially if you are using 1-5 type scales and you may wind up with a U-shaped distribution rather than one looking more like a normal curve because you missed all of the population in the middle. Your sample has ceased to be random and becomes statistically invalid, or at least suspect.

You can also poll people based on physical proximity. This means you are standing somewhere and polling people who come by. This technique will work if you are interested in the population that is in the area at the time. For instance, if you are interested in statistics about eclipse-viewers, then standing at Ninth and Main to poll people in Hopkinsville on August 21 works. If you are interested in the population of Hopkinsville, then polling people at Ninth and Main on August 21 does not work. First of all, a significant percentage of those at the corner were from out-of-town. Secondly, many citizens were working that day and unable to be at the corner. And, thirdly, it would be easily conceivable that the demographics on that corner that day did not conform to those of the community at large – i.e. blacks and Latinos were underrepresented. Man-in-the-street polls lead to entertaining reporting, such as when Jay Leno asks people basic questions about history and most cannot name the father of our country. The result may give an indication of a topic warranting further investigation, but the result is statistically invalid.

Internet polls are becoming increasingly popular, but the validity of their results is almost nil. They have all of the drawbacks of mail surveys, namely that only the truly committed will participate and the ambivalent or mildly interested will generally pass. Added to these

drawbacks, people have to find the survey in the first place in order to participate. How many liberals are searching conservative web-sites or blogs to participate in polls or vice versa? Again, you have the problem in spades of the population mostly participating in surveys for groups with which they are already familiar, if not in agreement. And you have the problem of underrepresentation of certain segments of the population, such as the elderly and the poor. You may even have the problem of underrepresentation of young people who spend all of their time on social apps and do not see traditional www websites.

Every sampling technique has its drawbacks, so there are statisticians whose job it is to try to make corrections for sampling error. The first thing he will do to try to make the results statistically valid is to break down the population into subgroups in an attempt to reduce sampling error. I have already mentioned the phone surveyors who hang up after two or three questions. The pollsters are trying to make sure that they are capturing the entire population desired. Sometimes, in spite of the best efforts to obtain a valid sample, there are gaps and, as already discussed in the U.S. census, a statistician will overweigh certain responses to try make the sample more representative of the population.

So, assuming that a pollster has done his best to obtain a statistically valid sample, the results should be valid and convincing, no? No. The results are only as good as the answers the participants gave. Usually, the pollster has no ability to determine if the participant told the truth or if the participant lied. This is why you have the famous old quote by Leonard H. Courtney popularized by Mark Twain: "There are three kinds of lies: lies, damn lies, and statistics".[vii]

Some statistics are not as subject to this damning observation. Statistics based on non-human samples, such as yield per acre, error rates in manufacturing, or test grades are directly observable. Admittedly humans are making the observations and therefore error can creep in,

but they are not so much matters of opinion. Likewise, observations about humans which are directly observable, such as hair color, weight, skin color, or gender, are not as subject to error. It is certainly possible to make an error of judgment, but these are accounted for in estimates of statistical accuracy, which I already discussed.

Your real problems come when you survey people to find out what they think or feel. If you are not already skeptical of statistics, once I finish this section on the problem of questions, you certainly will be. One of the first problems is not obviously a problem. Most people want to answer a question correctly. We all spent at least twelve years in school and correct answers always received positive reinforcement. Sometimes there is a definite correct answer, such as to the question of who is the father of our country. Sometimes, we are uncertain what the correct answer may be – there may not even be a correct answer, such as when we are asked about our feelings about a particular candidate or issue. In this instance, the correct answer is the one that correctly states our feelings. However, people also have a desire to please. We want to please others, including pollsters whom we do not know and have never seen. Researchers have found that respondents strive to please pollsters with their answers. One example was the survey taken at the height of the Monica Lewinski scandal that asked whether Bill Clinton or a prostitute was better suited to be president. Most people taking the survey understood that this question was humorous and new what the "correct" answer was; they intuitively understood which answer would please the pollster by giving and entertaining result. Therefore, it was no surprise that something like 80% answered the prostitute.

Another problem is how the question is asked. Leading questions indicate to the participant what the answer should be. Therefore, they are inadmissible in court and statistically suspect in surveys. In court, you should not ask a witness if he heard a customer threaten to

shoot the bartender. The question contains the desired answer, making it too easy for a witness to answer yes in order to please the attorney. You could ask the same question in a survey, and the results would be invalid for the same reason. Another example is the old question: "Hey ref, have you stopped beating your wife yet?" It is a close cousin of the leading question, in that a "yes" answer implies that the ref formerly beat his wife but has now stopped and a "no" answer implies that he continues to beat his wife. Especially in a survey, these kinds of questions are difficult for participants to answer clearly and should be avoided.

Related to leading questions is the issue of how a question is worded, which can influence how a respondent perceives the issue and thus answers the question. For instance, if a pollster asks whether potential voters prefer Hillary Clinton, who is married to her long-time husband Bill, or Donald Trump, pussy grabber currently married to his third wife Melania, he is likely to receive a different answer than he would if the question were if the voter preferred Clinton, former attorney involved in Whitewater and former Secretary of State involved in Bengazi or Trump, real estate mogul and billionaire. Admittedly, these are pointed examples, but you see the point. Even the order of choices can make a difference. People have a slight inclination to pick the first response, whether in a survey or in the polling booth. Furthermore, we have short attention spans, so if there are more than three responses in a verbal survey, we have a hard time remembering all of our choices and tend to pick something more recently mentioned since we are embarrassed to admit we cannot remember all of our choices.

Not only is word order and choice order important in understanding how valid a statistic is, so is the order of the questions. Research shows that once a respondent is asked to state a position or preference, he will defend that position or preference under further questioning, even if he had no position or preference prior to the first question. I have noticed this dynamic in

negotiations. If I ask someone what their preference is, it can be quite difficult to move them off of that articulated preference. If I do not ask the other party to state their preference, it is often much easier to sway him in the direction I prefer. This dynamic is closely akin to anchoring.

An informed citizenry must also be aware of input source influences when evaluating statistics. These come into play in particular when people are asked to make estimates or project into the future. For instance, there is the concept of anchoring. This input source influence works thusly: the pollster asks a question, gives a potential answer, and asks the respondent to react to it. Just by putting out a response, the pollster anchors the respondent's perception of where the correct answer would be. For instance, if a pollster asks how large a deficit we should allow the United States government to run, we might each have an answer. Many of us would probably say $0. However, if the pollster says is $10 trillion too large or too small, our perception of what the deficit could or should be is immediately anchored and while we might say that $10 trillion is much too large, we might now settle for $1 trillion or $100 billion. However, if the pollster asks if $100 billion is too large, that anchors our perception of the correct answer in a different place, and we would now say that $1 trillion is much too large, whereas before, we might have been tempted to accept a $1 trillion deficit.

Haloing is another input source influence of which to be aware. In this case, we like some attribute about the subject, whether a person or an item, and transfer that good impression to other attributes of the person or item. For instance, we like Oreo cookies, therefore we are prepared to believe that other Nabisco products will be tasty and safe. Or, we find someone physically attractive and are then easily persuaded that the person is also smart and good. In politics, we tend to believe that success in business will transfer to successful public service. Sometimes it does, sometimes it does not.

A third important input source influence is contrast. We evaluate potential answers based on the contrast with other choices or with past occurrences. For instance, if you receive a $3,000 raise, that may seem like a little if your last raise was $10,000, or it may seem like a lot if your last raise was $1,000. You will also evaluate the size of your raise compared to the size of your salary. If you earn $25,000 a year, a $3,000 raise looks pretty good. If you earn $500,000 a year, a $3,000 raise looks pretty paltry. Now in survey terms, let's go back to our example of the federal deficit. If we are asked if a $10 trillion deficit is acceptable, our answer will vary depending on whether last year's deficit was $100 trillion or $100 million and whether the overall budget is $1 trillion or $100 trillion.

We often hear statistical statements made along the lines of 9 out of 10 doctors who smoke prefer Chesterfields. When we hear such statements, we should applaud the advertiser for at least making it clear what the population is. However, the advertiser has not clearly stated how they arrived at the statistic. Is the answer from unaided or aided recall? For unaided recall, the pollster might simply ask: "What brand of cigarette do you prefer?" The respondent must name off whatever his brand preference is. For aided recall, after asking "What brand of cigarette do you prefer?", the pollster will list off several brands to aid the respondent in recalling the choices. Particularly in a crowded field like cigarette brands, the pollster is unlikely to include all of the brands available. An unethical pollster might even omit one or more of the favorite brands in an effort to skew the results towards the brand employing him.

While most statistics report on individual efforts or experiences, they can be subject to group influence. This problem will apply particularly to polls taken of individuals who have participated in a group experience. In the group, all of the input source influences come into play, along with others such as first mover advantage, in which the first person to move, i.e.

speak, anchors the rest of the discussion. The natural tendency of humans to please others and avoid conflict can lead people in groups to move from disparate views to a unified view. When polled after the group session, whether in a focus group or a jury, the subject will have different responses than he would have before the group experience. For instance, you may have had the experience of eating at a restaurant and finding it excellent. You then return with a group of friends and have an experience that your friends feel is average at best. You may modify your opinion of the restaurant towards the group mean, even if overall you were pleased with your second experience.

Some statistics, while expressed as numbers, are not subject to the full array of mathematical manipulations. If a pollster asks you to respond using a scale, he will wind up reporting that a certain number of people answered at each level of the scale. Some scales are 1-10 or 1-5, but these numbers are really labels, not numbers. They are more similar to scales like never, usually no, sometimes, usually yes, and always. Really, they are labels for your opinion and they give some indication of the strength of your opinion or the frequency with which you do something. Because they are labels, a scrupulous pollster will not calculate a mean or a median answer. He can give the mode, which you will remember is the most popular answer. He can also sort of add up responses and say that a certain number of people answer 1 or 2, or that a certain number of people answered always or usually. For instance, a pollster might ask how well you thought our current president was doing on a scale of 1 to 10, with 10 being superbly and 1 being horribly. While the different levels give the respondent an opportunity to express the strength of their opinion, they are not addable or averageable. We cannot take all of the answers and say that Americans felt that Donald Trump was performing at an average level. The best a pollster can do is to say that 45% of Americans felt that Donald Trump was

performing at a 6 or better on a 10 point scale. Another example would be if a pollster asks if you always vote Democrat, usually vote Democrat, sometimes vote Democrat and sometimes Republican, usually Republican, or always Republican. If, for the moment, we assume that the American population splits roughly evenly, we still cannot compute a mean. It is hard to say what the mean would even be in this instance. Would it be that we vote independent? Or, would it be that we vote 50% of the time for Republicans and 50% of the time for Democrats? Now you see why all the pollster can report is that some percentage of Americans always or usually vote for Democrats and some percentage always or usually vote for Republicans.

So, as you have now heard, statistics present many challenges of collection and interpretation. There are issues related to defining the population, creating a valid sample, crafting the questions themselves, and even ordering the questions. There are also response issues related to recall, the desire to please, the wish to be correct, and sometimes group dynamics. I hope that this exceptionally brief exploration of the problems will remind you of the issues and help you be skeptical of any statistic you see or hear, even when provided by reputable sources. If it sounds unlikely, it probably is.

ENDNOTES

---

[i] U.S. Census Bureau, 2010 survey
[ii] Webster's New World Dictionary
[iii] Social Security Administration actuarial life tables
[iv] Anderson, Sweeney, Williams: Statistics for Business and Economics, pp. 200-202
[v] Kinsey Scale, from Wikipedia entry: hhtp://en.wikipedia.org/wiki/Kinsey-scale
[vi] Ibid.
[vii] http://www.twainquotes.com/Statistics.html

BIBLIOGRAPHY

Anderson, David, Dennis, J. Sweeney, and Thomas A. Williams. Statistics for Business and Economics. St. Paul. MN: West Publishing Company, 1990

http://www.twainquotes.com/Statistics.html

Kinsey Scale, from Wikipedia entry: hhtp://en.wikipedia.org/wiki/Kinsey-scale

Northcraft, Gregory B. and Margaret A. Neale. Organizational Behavior: A Management Challenge. Chicago: The Dryden Press, 1990

Social Security Administration actuarial life tables

United States Census Bureau, 2010 survey

Webster's New World Dictionary

Zikmund, William G.. Exploring Marketing Research. Fort Worth: The Dryden Press, 1991