Project: Proper Sampling

Bias (n): a. an inclination of temperament or outlook; especially: a personal and sometimes unreasoned judgment; prejudice b. deviation of the expected value of a statistical estimate from the quantity it estimates; systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others. (from *Webster's Online*)

Some of you might remember the graph at left. It’s from the New York Times...one of the many predictions of the chance of the presidency going to the two main candidates in 2016 (Hillary Clinton and Donald Trump). Throughout much of the election season, Clinton was given margins like the one at left: 3 to 1, 4 to 1. These percentages were estimated by pollsters – folks who make a living asking people their opinions and then estimating percentages based on the responses.

And these continued right up until election night:

And then, do you remember what happened once the votes started coming in?
How did this happen? How could ALL the pollsters\(^1\) have gotten it wrong?

Maybe you know your history, so you know the comparison being made in this tweet; if not, let’s learn a little:

In 1948, the US presidential race was (predominantly) between republican Thomas E. Dewey and democrat Harry Truman. The night before the election, most Americans believed that Dewey would win the presidency overwhelmingly, based on information from Gallup polls.

For those of you unfamiliar with the picture at right, that’s Truman holding the paper...laughing, because he just won. What happened? How could the papers (and most of America) have gotten it wrong? Simple: a biased sample.

The sample taken for the poll was based on a telephone survey. In 1948, telephones weren’t in every household. In fact, to have a telephone implied affluence. In 1948, most affluent people were republican. So, the poll was actually biased toward republicans, and was, therefore, not representative at all of the American people. When the votes were counted, Truman stood as president, and the Gallup organization had to rethink how they sampled.

How does this relate to the 2016 election? I’ll have you do some Googling later – but for now, let’s pick up some pieces that we need to fully understand this stuff.

---

**Types of data**

Let’s assume we’re using numerical data. First, we need to know how the data is measured. There are four main levels of measurement for numerical data\(^2\):

In **nominal** measurement, the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. For example, jersey numbers in basketball are measures at the nominal level. A player with number 30 on their jersey can’t be compared with a player with number 25 on their jersey *based only on that number*; there’s nothing significant about the numerical value. No statistics can be done on nominal data.

\(^{1}\) Well, not all: [http://www.realclearpolitics.com/articles/2016/10/14/why_pay_attention_to_the_la_times_poll.html](http://www.realclearpolitics.com/articles/2016/10/14/why_pay_attention_to_the_la_times_poll.html)

\(^{2}\) Thanks to [http://www.socialresearchmethods.net/kb/measlevl.htm](http://www.socialresearchmethods.net/kb/measlevl.htm) and [http://www.nizkor.org/features/fallacies/biased-sample.html](http://www.nizkor.org/features/fallacies/biased-sample.html)
In ordinal measurement, the attributes can be rank-ordered, but distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as 0=less than H.S.; 1=some H.S.; 2=H.S. degree; 3=some college; 4=college degree; 5=post college. This is called a Likert measure, or Likert scaling. In this measure, higher numbers mean more education. But is distance from 0 to 1 same as 3 to 4? Of course not. The interval between values is not interpretable in an ordinal measure. Ordinal data is tricky; some folks argue you can do meaningful statistics on ordinal data. I think it depends on what you’re studying, but that’s a discussion for a later time and place (i.e., in about 5 weeks).

In interval measurement, the distance between attributes does have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is the same as distance from 70-80. The interval between values makes sense; however, comparing two or more intervals does not make sense. For example, 80 degrees is not twice as hot as 40 degrees (although the value is twice as large). There is no zero starting point, so absolute value distances can’t be ascertained. Many statistical tests require that data be “at least interval”.

In ratio measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. Most discrete variables are ratio; for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say things like "...we had twice as many clients in the past six months as we did in the previous six months."

OK...now you know what kind of data you want. How to collect it?

Types of Sampling

Once you have decided you actually want to do a statistical study, you have to decide how to collect your data. Sample size is one point that must be considered, and we will discuss it often in class. For right now, all you need to know is that too small a sample is always bad. However, just having a large sample isn’t good enough; you must have a proper sample selection technique.

Well, a random sample is perfect. In a random sample, every member of a population has an equal chance of being chosen. In a simple random sample, we first fix a sample size, call it n. Then, we design a method so that every possible sample of size n has the same chance of being chosen.

But...how do you actually do a random sample? Suppose I wanted to conduct a simple random sample of 100 current COCC students. I need to be random. Humans aren’t very random, as you might recall from MTH 243. So how can I do this? One way is to generate random strings of digits, 6 digits long. These 6 digits stand for the last 6 digits of your COCC ID number (we all have an 820 at the start of our numbers).

How would you do that? Random digits are fairly easy to come by in the present day. Before the technological explosion, we had to rely on tables of random numbers, which could be a pain in the rear. At right is a little bit of one that I found online. To use it, you start by selecting the first 6 digits (396346). The student with 396346 in their last 6 digits is our first member of the sample. We then move onto the next six digits (234974), find that student, and so on.
With more computing technology available, we can also generate random numbers with technology. Excel’s the bomb – you can use the “randbetween” command to generate a number of 6–digit random numbers. Here’s what you type in the formula bar:

```
=RANDBETWEEN(0,999999)
```

Then, you can refresh the formula (by pressing F9 as many times as you need) or copy the formula to other cells:

```
E1 =RANDBETWEEN(0,999999)
```

Of course, you might run into a problem with this simulation. Look at the random strings I generated. Suppose there are no students with 072737 (or 388639, or 407752) as their ID number suffixes. How would you proceed? Well, you could just skip those numbers, but, with a school as small as COCC (and with 100,000 possible arrangements of these 6–digit numbers), maybe we should try something different...

One way would be to simply arrange all COCC students in alphabetical order. Then, number them. The first student whose last name starts with “A” is number 1, the next is number 2, and so on. This would give you a list of, say, 17000 students, each paired with an integer between 1 and 17000. You can then use the randbetween command again....randbetween(1,17000), and then copied to 100 cells:

```
J10 =RANDBETWEEN(1,17000)
```

Then, use your master list to find those 100 numbers!

---

3 Current today (8.28.15)
Random sampling is wonderful, and is the best method for obtaining a great sample. Unfortunately, it can also be very, very tedious. You basically have to create a list that accounts for every member of your population, and (even with technology), that can be daunting. To see why, suppose your population is “all adults in the US.” The Census Bureau needs to deal with this parameter every 10 years. Hundreds of millions of numbers to match with people, with people being born or dying every few seconds...a simple random sample’s impossible. So how to proceed? Well, let's look at some other types of sampling...

- **Systematic Sampling** is where your data is arranged in some sort of array; you pick a random starting point, then pick an interval, and then “skip” to your sample elements. For example, you may start with the 5th element in your data, then select every 7th member. In the Roman Army, systematic sample was used to instill fear in the troops. Any soldiers guilty of mutiny, cowardice or desertion were arranged in a line, and then every 10th soldier was executed.

- **Stratified Random Sampling** is a very efficient method of sampling. In this method, your population is divided into two or more subsets that each have the same characteristic (called “strata”, the plural of “stratum” which means “to spread out”). Then, you randomly sample within each of the subsets. Of course, if your subsets are not the same size, you may have to adjust your weights accordingly. For example, if you wanted to sample engineers from each of the mid-Atlantic states, you might randomly select a few from Delaware, a few from New Jersey, and few from Pennsylvania, and a few from Virginia. Of course, there might be more engineers in a large state like Pennsylvania, and fewer engineers in a small state like Delaware, so results might have to be weighted (remember that from MTH 243?).

- **Cluster Sampling** is accomplished by first dividing the population into “clusters”, or non-overlapping subgroups, and then randomly selecting some of those clusters and using all members in each cluster. It is somewhat similar to stratified random sampling.

- Last (and definitely least) is **Convenience Sampling**. In this form of sampling, you collect data that’s simple to get. For example, if I were interested in the proportion of people in Oregon that played guitar, I could lean out of my office and ask my co-workers, “Yo! You play the guitar?” It shouldn’t take you too long to realize that’s probably nowhere near a viable sample.

**Final Thoughts**

1. Of course, no matter how good your sample, it’s still possible that it’s not exactly representative of the population. This is called **sampling error**. It happens; you can use a quality, unbiased sample and still have unexpected results. We’ll learn techniques to offset that as much as possible in class. However, **nonsampling error** results from the way your data is processed. You could take a potentially valid sample and ruin it. You want to try to avoid this. For example, suppose you’re performing a study about the efficiency of email communications over paper methods. You properly randomly sample from your population, and ask, “Which, in your opinion, is more effective: lightning – fast email, or sluggish and often misrouted interoffice memoranda?” Even if you don’t know what “interoffice memoranda” are, chances are, you’re steering your respondents away from that choice (and towards email) simply because they don’t want to appear silly for choosing the “sluggish” option. Careful with your wording!

2. With increased use of handheld technology, social media, and 24 – hour “news”, self – selected sampling is on the rise (“text your answer in now!”). As a result, more and more of the “statistical” results you see on many outlets are meaningless.

And last, but certainly not least...

---

4 Hence, the origin of the word “decimate”, which literally means “thin by a tenth.”
3. Observational studies (where behaviors are monitored by objective researchers) are much more reliable than self-reported studies. Of course, they’re also harder (and usually, more expensive) to implement. Consider this study: you’re interested in whether or not there’s a difference in heights of men in two different areas. What will give you better results: an anonymous observational study that measures the men, or simply asking the men to self-report their heights? Yup.

For those of you moving on to research fields, sampling will be a large part of your career. In your further studies, you will learn more and more methods of sampling, even though they’re just variations of the ones above. You will also learn about good questioning techniques and interview/data collection skills, such as double-blind studies. For some of you, this is just the beginning!
1. (2 points each) Determine which of the four levels of measurement is most appropriate.

   a. High temperatures for Bend this week  
      nominal  ordinal  interval  ratio
   b. OSU/Cascades student ID numbers  
      nominal  ordinal  interval  ratio
   c. Rating this class 1 = great, 2 = so – so, 3 = terrible  
      nominal  ordinal  interval  ratio
   d. Distances traveled by students commuting to COCC  
      nominal  ordinal  interval  ratio

2. (4 points) Recently (2015), my home state of Delaware was voted the “3rd most bicycle friendly state in America” (behind Washington and Minnesota, with Oregon coming in 6th). Check out the rankings here, if you’d like. I was immediately skeptical – I vividly remember getting run off the road, having things thrown at me, and even being hit while biking. Granted, it was a long time again, but…it was still hard to believe that things had gotten so much better. And, as it turns out, they may not have: Google “bike friendly state ranking” and look for the League of American Bicyclists link that explains their sampling methodology. Anything potentially fishy with the way the data were gathered (hint: who did they ask)?

3. (4 points) Please read this article: http://www.bendbulletin.com/localstate/4230600-151/survey-1-in-10-cocc-students-has-felt. Among the numbers in that article that surprised me when I first read it were “16 percent of students reported having observed or been made aware of conduct directed toward others that created an offensive or hostile learning environment” (not to mention the headline! Yikes!). I mean, heck! There are usually about 35 – 40 students in my classes…3 – 4 of them feel disrespected? 6 feel that they’re in a hostile learning environment?!? But, once I realized the statistical blunder that was made during the data collection, I felt much better…from the article, find that blunder and tell me what it is. And no – it’s not that only 9% of the students were surveyed (that’s actually an OK sample size; assuming around 10000 credit students, that’s 900, which would give us a MOE of about 3.1%. I can live with that).

4. (6 points….3 for each) Do some online research and give at least two potential explanations for how so very many pollsters, right up until election night, underestimated Trump’s chances of winning the presidency (and overestimated Clinton’s). Make sure that the reasons you find aren’t of the “people are fed up of traditional politicians” kind; if that were the case, the polls would have borne that out. It had to have been with the sampling they used. Find two ways that could have explained the discrepancies.

5. For the remaining (3 points), please go to the site here (if you need the hyperlink, it’s http://coccweb.cocc.edu/srule/MTH244/projects/sampling.html). The activity outlined there (and summarized in class) will help you see why randomness is so hard to achieve by humans.