Responsible Citizenship in a Technological Democracy

Note 10: Statistics and Probability

An observation often attributed to Mark Twain, but probably originally from Benjamin Disraeli, notes that there are three kinds of lies – lies, damn lies, and statistics.

Often public policy discussions involve statistical data, so to be a good consumer of, and participant in such discussions you need to know at least a little about such data, what they mean, and their limitations. There is a well known book, *How to Lie With Statistics* by Darrel Huff, that goes into a great deal more detail than we can here – check it out if you get interested.

Closely allied to statistics is probability, and we'll cover a bit of that too. But we're going to do them both lightly. Just what you need, no more.

Statistics

There are basically two kinds of statistics – descriptive and inferential. At their heart descriptive statistics are a "numerical summary" of a larger collection of numbers. The most common such summary, for example, is what we generally call the "average" – the sum of the collection divided by the number of numbers in it. The idea is that the average is intended to be sort of "typical", or "representative" of the collection.

Inferential statistics are used to draw inferences about a process or population being studied. The most common such statistic for our purposes is "correlation"; more about that later.

Let's look into a few descriptive statistics first – and begin by noting that there are three things that have some legitimate claim to be called the "average", or "typical", or "representative" of a collection of numbers:

> **Concept:** The *arithmetic mean* is the most familiar and what we described above – the sum of the collection divided by the number of numbers in it
>
> **Concept:** The *median* is that number for which half the numbers in the collection are larger and half are smaller.
>
> **Concept:** The *mode* is the most frequently occurring number in the collection.

Although the arithmetic mean is most familiar to you, I hope you can see that the median and mode also characterize something about the typical member of the collection. There are two things for you to remember. First, there is nothing sacred about formulas like that for the mean – there can be alternatives that might work as well of better. Second:

> **Concept:** The mean, median, and mode can be very different from one another. Be sure you are being given the one most relevant to the issue at hand, and when they are different, be suspicious!

To illustrate the last concept, consider the collection of numbers

> 1, 1, 3, 5, 990.

The mean is 200, the median is 3, and the mode is 1 (check it out). If these numbers were describing the reduction in taxes different people paid under a proposed new tax plan, for example, one might have quite different reaction to that plan based on which statistic is used. That's why you should be suspicious when the three statistics are quite different – there is something potentially odd about the collection of numbers that warrants closer scrutiny.

There are a zillion other descriptive statistics, but only one comes up often enough in policy discussions to warrant coverage here – the *standard deviation*. Consider two collections of numbers:

0, 0, 1, -1 , 2, -2, 3, -3
and
0, 0, 100, -100, 200, -200, 300, -300

I cooked these collections so that they both have the same mean, median, and mode, namely zero. But yet they feel quite different. The first one is more "compact" whereas the second is more "spread out." Enter the *standard deviation* to give a measure of just how spread out a collection is. The details of the calculation aren't important, but the idea is that the standard deviation is sort of an average of how far away each of the numbers in a collection is away from the arithmetic mean of the collection. So a larger standard deviation implies a collection whose elements are further apart.

> **Concept:** The *standard deviation* of a collection of numbers is a measure of how spread out the numbers are from the mean of the collection

At an intuitive level, a small standard deviation suggests that the mean, for example, is pretty representative of all the elements of the collection. Conversely, a large standard deviation may mean that it's not very representative.

There is one last idea to discuss concerning statistics – *sampling*. In many cases where, for policy reasons, we want a statistic on a *large* collection – say, the mean household income in the U.S. – there are something like a hundred million households in the U.S., so it's simply impractical to get all this data. Instead, a "representative sample" of a few thousands households is used. Done properly, this will generate an answer that is very close to the true one and is a lot less work.

Sampling works because of something called the "Law of Large Numbers":

> **Concept:** *The Law of Large Numbers* is that, given some large collection of numbers, if a large enough portion of that collection is chosen in a way that has no bias, the statistics of the portion will closely resemble that of the whole collection.

There are two things to attend to in the description of the Law – “large enough” and “no bias”.

As for what is “large enough”, you can get a sense of that from the fact that when pollsters sample the public on things like their preference for political candidates they talk to a few thousand people. They then report their results by saying that there is a possible “sampling error” of some small percentage, maybe 3%. What they are saying is that it is highly probable that with the sample size they used, the odds are that the whole population would give the same response within that percentage of what they report. They don’t try for a 2% or 1% sampling error because they would have to talk to a *lot* more people

The issue of “no bias” is trickier. Indeed, many a scheme that was thought to be without bias was later discovered to have one – sometimes one that matters and sometimes not. For example, telephone pollsters generally use a scheme involving the generation of random phone numbers. That seems like it wouldn’t have a bias since your phone number isn’t related to your ethnicity, or gender, income, or just about anything else except rough geographical location. But of course it excludes homeless people and possibly some hearing impaired people. The exclusion of such groups may not matter for some questions, but for things like the household income survey, it might have a notable bias.

By the way, some WAGs like to talk about the “Law of Small Numbers” – the converse of the Law of Large Numbers. In effect it says that if you make the sample small, the statistics can be wildly different from those of the large collection. Beware of inferences from small samples!

Now let’s switch to inferential statistics, and the main concept here will be *correlation*. But to get to it we first need to discuss the notion of *independence*.

> **Concept:** Two things are independent if knowing something about one of them tells you *absolutely nothing* about the other!

I suspect that the color of shirt I choose each morning is independent of the number of passengers on the Queen Elizabeth II that day, for example. On the other hand, the color of shirt I choose is related to whether I am going in to the office that day – I tend to wear a white or blue “dress shirt” when I go to the office. But it’s not a 100% predictor since I may wear a sport shirt if I go in on weekends or holidays. I also wear a dress shirt when I go on business trips, attend conferences, and so on, where I will not be at the office. So, my shirt color is independent of the passengers on the QE II, but not of whether I am going to the office.

> **Concept:** Correlation is a measure of the deviation from independence.

Correlation is expressed as a number between -1.0 and +1.0, such that when things are independent their correlation is 0. When it’s +1.0 the two things are absolutely in lock step and when one goes up the other does too. When it’s -1.0, the two things are also in lock step, but when one goes up the other goes down. When it's a number in between -1.0 and +1.0, say -0.3, then some of the time when one goes up the other will go down, but not all the time. If it's a number such as -0.9 the "correlation is stronger" than -0.3 and the second will go down a lot more often when the first goes up. My shirt color is pretty highly correlated with whether I'm going to the office, but it's not 1.0!

The notion of correlation is extremely useful in a variety of policy contexts, but it is often misused. Please remember:

> **Concept:** Correlation is not necessarily causation!

Just because two things are highly correlated one cannot assume that one causes the other. The reason for the correlation may be quite indirect. There is an old saw that stock market rises are correlated with rising women's hemlines[1]. Although since discredited, this illustrates two points:

1. Nobody would argue that rising hemlines *cause* the stock market to increase, or the converse. They are at best correlated, not causal. Possibly, for example, both reflect a rising mood of confidence among the population caused by some completely unrelated event.
2. More data proved the correlation to be much weaker than initially thought. Beware of predictions based on correlations! The Law of Small Numbers applies here too!

While it's true that mere correlation does not prove causation, if A *does* cause B, then A and B will also be highly correlated. So, if two things are highly correlated it's at least worth asking the question whether there is a causal relation – you just can't use the fact of correlation in answering the question.

Policy Relevance

As suggested in the opening paragraphs, statistics are often misused in policy discussions – sometimes intentionally, but often unintentionally. There are just a few things you ought to keep in mind:

1. There are three notions that can claim some right to being called the "average" of a collection of numbers: the mean, the median, and the mode. They can be quite different from one another, and which is the "best" characterization of the collection may depend on the use as well as the data in the collection.
2. The degree to whether any of the "averages" is representative of numbers in the collection depends in part on how "spread out" the numbers are. When someone uses "the average" in a policy argument, it's a good idea to ask about the standard deviation!
3. Correlation, even a strong one, is not necessarily causation!

The most common misuse of correlation is to suggest a causal relation when one doesn't exist. Just because every inmate of a U.S prison eats potatoes does not mean that potatoes cause crime! But it goes the other way too. For example, in the current discussion of global climate change

[1] I checked with my wife to see if this example is PC. She, at least, is not offended by it.

scientists have shown an indisputable correlation between the concentration of CO_2 in the atmosphere and a rise in temperature since the beginning of the industrial revolution. But, say the critics, that's only a correlation and doesn't prove causation.

In a sense they are right, but as noted earlier, a causal relation also creates a strong correlation. So a strong correlation suggests that we should at least ask about whether there is a causal relation. In this case there is other evidence concerning the mechanism of the "green house effect" that has convinced the vast majority of climatologists that the relation is causal.

Probability

Statistics characterize, or summarize, existing information. Probability is about predicting the likelihood of some extant or future event. Hurricane Katrina happened. OK, the two questions are – how likely was that to happen (the extant event), and how likely is it to happen again (the future event)? Statistics and probability are related in that statistical data from the past is often used to suggest the probability of events in the future.

> **Concept:** Probability is expressed as a number between 0 and 1 (or sometimes between 0 and 100%, or as a pair of numbers such as 60/40 that sum to 100). Zero means it hasn't or won't happen, no how, no way! One means it has or will happen without any question or doubt.

The concept of the probability of a single event is pretty intuitive. It's pretty clear that the probability of a coin toss coming us heads is 0.5 (or 50%, or 50/50 if you prefer), or the probability of rolling a one on a six-sided dice is 1/6th. The thing you need to know to be a responsible citizen relates to combining the probabilities of two or more distinct events – the probability of an anthrax attack *and* a simultaneous assault on the 911 emergency phone system, for example.

There are four cases that you need to know about. I am about to give you some formulas for these four cases, but I don't expect you to memorize them. What I do expect is that you will remember that there are different ways to combine probabilities and when you see it being done, to then to dig out this note and see if it's being done properly. I expect you may be surprised how often it isn't!

Forgive the notation, but consider two events, A and B. Also consider P(A) and P(B) as the probability, likelihood, of those events happening. Then,

1. The question might be "what the likelihood of <u>both</u> happening?". The answer is $P(A \text{ and } B) = P(A) \times P(B)$. So, if the question is what's the likelihood of tossing two coins in a row to be heads, the answer is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

2. The question might be "what is the likelihood of one <u>or</u> the other happening"—say rolling a one <u>or</u> a two on a six-sided die. The answer depends on whether the results are "mutually exclusive".

a. In cases like the toss of a die, the result has to be a one, a two or something between 3 and 6. No ambiguity – the outcomes are "mutually exclusive" – only one number can be on top. And in that case the probability is just P(A or B) = P(A) + P(B). The probability of rolling a one or a two is just twice that of rolling just one of them; that is, 1/6+1/6 = 1/3.

b. Alas, in other cases, the outcomes may not be mutually exclusive. You might win the lottery <u>and</u> get the girl, for example. In that case, the probability is P(A) + P(B) – P(A and B).

3. When A and B are not independent, then another kind of question arises – namely what is the probability of A happening given that B has already happened. What is the probability that I will go to the office given that I am wearing a white shirt, for example. This is called the "conditional probability of A given B" and it written P(A│B). The formula for it is P(A│B) = P(A and B) ⁄ P(B).

Policy Relevance

Back in Note 7 on risk, we implicitly used the notion of probabilities when we said that risk was often defined as the likelihood of an event times the consequence of that event happening. We also asserted that there are "risk analysis" techniques that can be used to determine these two quantities. Well, we're still not going to go into detail on those techniques, but in essence they involve thinking deeply about all the bad things that can happen in a given situation, assigning a probability to each such bad thing and then combining those individual probabilities using the techniques described above.

Given that managing risk, and ensuring safety and reliability in particular, are a primary function of government, the notion of probabilities sneaks into a lot of policy discussions. The misuse of combining them is rampant – forgetting to distinguish between events that are mutually exclusive or not, for example.

All through the 1980's there was a campaign by several advocacy groups to force auto-makers to install airbags, and finally in 1991 President Bush signed a bill to do that. Then, in the late 90's a hugh-and-cry went up from pretty much the same advocacy groups about the fact that passenger-side airbags were killing small people – especially children and small-framed women. So now you can turn off the passenger-side airbag. But ironically, it was perfectly predictable in the 80's that small people would be killed – in fact it *was* predicted! Some arm waving "simplification for the public" by the advocacy groups with an improper combination of probabilities allowed this to be ignored. Airbags are a good thing, a really good thing! But be careful of argument for a good thing that ignore its downside. Be especially cautious of statistical arguments that seem to give a lopsided picture. Nothing is a pure good or a pure evil!

Note one final thing about P(A and B) vs. P(A or B). Since probabilities are always less than one, their product is always *smaller* than either the individual ones. Since probabilities are always positive numbers, their sum is always *larger* than either of the individual ones. As a result, if you use the wrong formula the answer can get really out of whack.

Again, you don’t need to remember these formulas, only that they exist and that it *really* matters that the right one gets used!

Randomness

Your intuition is probably pretty good about the notion of randomness. However, it’s a word that gets tossed around rather loosely sometimes, so there are a couple of points I’d like you to be aware of.

> **Concept:** Given a collection of things, a *random sample* of it is obtained if each element of the collection has an equal chance of being selected.

Note that randomness is a property of the *process* for selection, not of the selected items. So, a perfectly random selection from the collection 1, 2, 3 could be 1, 1, 1. In fact, 1, 1, 1 is just as likely as any other specific selection! Keep this in mind in policy discussions. The word “random” doesn’t mean that a selection will be “fair” or “egalitarian”, or that there will be “equal representation” in the selected items. A random selection is one of the best approximations to “fair” that we have, especially if you draw a selection large enough for the Law of Large Numbers to apply, but “random” and “fair” aren’t the same thing.

> **Concept:** The notion of randomness is about a process of selection, not about the properties of the selected items.

One of the ways that people sometimes delude themselves, or try to persuade others, is to forget about independence. Suppose you’ve flipped a coin 50 times and they all came up heads. Something inside us screams that the odds on the next toss must be *much* higher for tails. *Much, much* higher! But, nope. Each toss is independent of those before it, and so there is only a 50-50 chance that the next toss will be tails.

Changing the subject a bit, earlier I gave an example of the sample selection technique used by pollsters that involved generating random phone numbers. Well, I fudged a bit there. Those “random” phone numbers were generated by a computer, and computers can’t generate really random numbers. Computers are “deterministic” – which is fancy way of saying given the same input they will always do the same thing, and therefore produce the same output. Usually that’s a good thing. But, “deterministic” is the antithesis of “random” – in the former, given an input you always get the same output, in the later you have no idea what will come out.

Computer Scientists have devised techniques that produce what are called “pseudo random” numbers. If you look at a sequence of numbers produced by one of these techniques, it will look like it could have been produced by a random process. But, if you run it again with the same input, you’ll get the same output! The trick, of course, is to start the process with different input, and that’s not hard to do. But, you have to know to do it. Wouldn’t it hilarious, or maybe deeply misleading, if those pollsters didn’t know to start with different input and wound up calling the same “random” phone numbers for each poll they take?