6.4 Inductive Statistical Reasoning

There's a 30% chance of rain tomorrow. The Republican candidate for governor is just 2 percentage points ahead, well within the margin of error. Las Vegas odds give Man o' War a 5 to 1 chance of winning the Kentucky Derby. Eating half a cup of blueberries everyday cuts your chance of getting cancer by 15%. Ebola carries a 70% mortality rate.

Our lives are awash in statistical claims as well as arguments that depend on them. It's important, therefore, for us as critical thinkers to have a sense of some of the principal errors people make in statistical reasoning. Statistics is an enormous field of its own, and the issues can become terribly complex, but nevertheless, a few central ideas can cut through a lot in our world that's not only difficult but also misleading.

Sampling: random and biased

Determining probabilities is often a deductive process (e.g., determining the probability of pulling two white marbles in a row from a bag of five, only three of which are white). The kinds of statistical reasoning that are, however, typically most interesting and most fraught for people are *inductive generalizations* about entire populations (the *target population*) on the basis of subsets (the *samples*) drawn from those populations or from related *sampled populations*. So, a study might draw inferences about a *target* population of all Scottish voters on the basis of a *sample* of surveys of just a few hundred people drawn from a *sample population* of those listed in the telephone directory. Perhaps the most important dimension of this kind of inquiry is the quality of the sample.

Qualitatively, samples ought as far as possible to *represent* the target populations about which they're to provide information. That is, in order to be *representative*, a sample ought to have the same relevant qualities as the target population. If 65% of the entire target population votes for the Labor Party, and voting patterns are the relevant property, then in order to be representative, 65% of the sample must also vote Labor.

When samples are not representative, they're said to be *biased*, and biased samples are perhaps responsible for most errors in statistical inferences. Researchers try not to acquire biased samples in a variety of ways. Perhaps the most common method is randomizing the sampling process in order to produce a *random sample* from the sampled population. But how does one "randomize"? There are computer programs that generate something approximating random numbers, which can be indexed to, say, entries in the phone book (though that method misses people not in the directories), or the numbers may be assigned to other sample set elements. But no computer program is really random (since program algorithms are composed of definite rules). And, therefore, the process of randomizing often attracts critical scrutiny. Researcher Shere Hite assembled a well-known report on human sexual conduct (*The Hite Report*; 1976), but some of her methods for assembling data drew criticism. For example, Hite had drawn data from survey forms placed in soft pornographic magazines, among other sources; but are the readers of that sort of publication representative of the general population? Are any data sets collected purely from voluntary participants average or normal, since the study then reports only on the kind of person who answers surveys? And is the average person who answers surveys about personal matters different from people on average?

Stratification

Even a truly random sample may not represent a population when the target population is unevenly distributed in some way. For example, suppose one wanted to study what percentage of drivers on a certain highway are on average under the influence of alcohol. Checking simply a random sample of drivers over a 24-hour period would not produce a representative sample, since more drivers are on the road at certain times of day – e.g., hypothetically, 75% of drivers during rush hours between 7–9am and 4:30–6:30pm. Similarly, if 90% of Conservative Party voters are densely clustered in a few postal codes in Scotland, it won't do to sample voters randomly and uniformly across all postal codes.

To deal with biases introduced by uneven distributions, researchers employ a strategy called *stratification*. Data sets are stratified so that the percentages of relevant subpopulations in the sample match that of the target population. So, in the study of drunk drivers, a researcher might be sure to take 75% of his or her otherwise randomized sample during the rush hours and then take the rest randomly from the remaining 20-hour period.

Data sets are often stratified in the social sciences by gender, by age, by geography, and by other factors that are thought to be relevant to studies. Note that assessing whether or not a sample is representative depends upon what characteristics are *relevant*. Sinistrality, or left-handedness, hair color, average weight of pets, and whether or not subjects' surnames have odd or even numbers of letters are unlikely to be relevant to an inquiry about voting preferences. Noting relevant qualities upon which to stratify and assess representation is, however, very important. A sample of only or mainly middle-aged American males will hardly yield results that tell us about all people across the globe or across history. Just this sort of error has characterized some medical research. Ruth Hubbard, a feminist philosopher of science, rightly criticized the important 1989 Physicians Health Study on the effects of daily aspirin on heart disease. Hubbard found that nearly all of the 22,000 subjects of the study were men. Other studies were found to be based upon similarly biased data sets. Unsurprisingly, subsequent research has found that heart disease occurs with different dynamics in women, and so therapies must be designed in light of sex as a relevant factor. Until the 1990s the FDA (the US Food and Drug Administration) routinely only tested new drug therapies on men, while research on women was largely restricted to reproductive issues. That bias may clearly have distorted scientific findings.

The gambler's fallacy

Gamblers can often be heard saying things like, "A seven is due to come up on those dice"; "That roulette wheel has landed on red the last five spins, so the next spin is likely to be black." But that sort of thinking, as ought to be well known, is fallacious, partly an effect of magical or *wishful thinking*. Spins of the roulette wheel, throws of dice, and flips of coins are what statisticians call *independent events*, meaning that past instances of those events have no bearing upon future instances. The past flips of a coin do not somehow reach out from the past and influence present and future flips. Each flip happens on its own, and each flip of a normal coin carries a probability of roughly 50/50 heads or tails. Even if a coin has landed heads the last one hundred flips, on the next flip (assuming it truly is a normal coin) we still face under normal circumstances just a 50/50 chance of landing on tails. If that seems strange, think of it this way: sure, flipping 101 heads in a row is highly improbable, but having already flipped 100 heads, most of that extraordinary event has already occurred. The bit that's left is just a 50/50 event.

Now, of course, some difficulty may arise in figuring out whether or not events are *independent* or *dependent* in relation to one another. If the question involves bags of marbles, as it often does in statistical examples, what matters may be as simple as whether the marbles withdrawn are replaced. In a bag of 10 marbles, 5 of which are white and 5 black, there is a 50% or 5/10 chance of drawing a black marble. But if that black marble is not replaced, there remains then only a 4/9 chance of picking a black marble on the next draw.

With people, things get trickier and more psychological. If a team has lost five games in a row, those past losses may indeed be relevant to the team's next performance simply because athletes, unlike coins, remember and are affected by the past. The team may be demoralized after the string of losses and therefore play even less well than they have in the past. Or they may have become motivated by their recent defeats and may have gained extra determination to win the next game, thereby improving their chances. That's why blinding is so important in medical research, and why consumer expectations are so important in economics. People are historical beings, and they're influenced by their understanding of not only what is presently happening to them but also what has happened to them and to others in the past, as well as by what they believe will happen in the future.

Averages: mean, median, and mode

People are sometimes tempted to commit an error akin to the fallacy of division because of how some statistics are presented, especially averages. When the average American household is said to have "2.3 children," no one thinks, "Wow, I wonder what three-tenths of a child looks like?" But when someone says, "The average income of this neighborhood is $83,000/year," it is tempting to think that each household in that neighborhood makes about $83,000/year. This inference is fallacious. Consider the following four neighborhoods, each with seven households. Each has an average income of $83,000/year, but notice the vast differences in each household income and each neighborhood:

East Neighborhood	West Neighborhood	North Neighborhood	South Neighborhood
1. $83,000	1. $18,000	1. $50,000	1. $35,000
2. $82,000	2. $12,000	2. $60,000	2. $88,000
3. $84,000	3. $10,000	3. $70,000	3. $40,000
4. $82,000	4. $83,000	4. $60,000	4. $40,000
5. $84,000	5. $42,000	5. $95,000	5. $45,000
6. $82,000	6. $391,000	6. $83,000	6. $278,000
7. $84,000	7. $25,000	7. $163,000	7. $55,000

Average annual income of each neighborhood: $83,000

Different kinds of "average." One way to avoid this error – or detect it – is to keep vigilant about the difference between different kinds of averages. Principally, there are three kinds to consider when thinking critically about claims involving "averages." The *mean* is the result of the simple arithmetic process of adding up the relevant quantities and dividing by the number of them. It's sometimes called the *expected value*. Our example above of neighborhood income gives mean averages – each column of seven incomes is summed and divided, of course, by seven to reach an average of $83,000. A second kind of average is the *median*. You might think of the median as that element of the data that stands in the middle, so to speak, such that there are an equal number of data points above and below it. So, in our example above, the median income of East Neighborhood coincides with the mean at $83,000, whereas in West the median is just $25,000, less than one third of the mean. A third kind of average is the *mode*. The mode is just the most frequent or common element. So, in the North Neighborhood the mean is $83k; the median is $70k, and the mode is $60k. When drawn out as a curve, the top of the bump illustrates the mode (or multiple bumps in multi-modal distributions).

Distributions

In what's called a *normal distribution* the mean, median, and mode all coincide – it's the perfect "bell" curve. Curves where the mode is less than the mean will tilt toward the left; where the mode is greater than the mean, they'll tilt rightward. Note, however, that curves can be more or less broad, depending upon how broadly the data are distributed. So, the East Neighborhood will present a high graphical curve, without much spread, as all its data points are clustered around $83,000. East Neighborhood, however, presents a bimodal distribution, where there are three households at $82k and $84k, while only one draws $83k. That means the curve for East Neighborhood will not only be a high, narrow curve; it will also have two bumps of equal height (three households each) with a dip in between them (illustrating the one household that corresponds to the mean and median). South Neighborhood, in contrast, will be much flatter. It will rise quickly to its mode at $40k and then slope more gradually downward stretching way, way out to that one outlying household at $278k.

The *standard deviation* and *variance* describe the distribution or spread of a data set by giving metrics for distance of data from the mean. (Variance of the sample is mathematically defined in several ways depending upon the different kinds of variable involved, but a good general definition is the *mean average of the squared distance of each data point from the sample mean*. Standard deviation is the square root of the variance.) Roughly, the larger the standard deviation or variance, the broader the distribution. If all the data are the same, variance and standard deviation will be zero. In a normal distribution, one standard deviation captures about two-thirds of the data, two standard deviations about 95% of the data, and three standard deviations 99.7% of the data. (Non-normal distributions don't show this pattern.) Obviously, a wide distribution of data will exhibit a larger standard deviation or variance.

All this suggests a few important lessons. When scrutinizing statistical claims about averages, strong critical thinkers will consider how the data might present different averages – that is, different means, medians, and modes. Strong critical thinkers will also consider how broadly or narrowly the data behind any statistical claim are distributed. Highly unequal economic orders might share the same mean wealth with those that are relatively egalitarian, but the median and variance in wealth will be substantially different.

Exercises and study questions

1. Determine the mean, median, and mode for each of the data sets in the neighborhood income table in this section. Draw a graph or curve illustrating each data set. In a few paragraphs, compare and contrast the data and reflect upon what social and political implications one might draw from them.
2. How might one gather a representative sample of illicit drug use or private sexual conduct? How might doing so be different from gathering data about bird behavior or human residential patterns? Write a short essay to explain.
3. "This just in: 65% of Americans approve of the president's performance. (This data was gathered from a survey of 900 registered voters, by calls made between 7pm and 9pm, Monday through Friday.)"

SEE ALSO

Chapter 5: Tools for Detecting Informal Fallacies
6.1 Inductive vs. Deductive Arguments Again
6.3 Fallacies about Causation

READING

Sherri L. Jackson, *Research Methods and Statistics: A Critical Thinking Approach*, 5th edn (2015)
Charles Whelan, *Naked Statistics* (2013)
Timothy C. Urdan, *Statistics in Plain English*, 3rd edn (2011)
David J. Hand, *Statistics: A Very Short Introduction* (2008)
Joel Best, *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists* (2001)
Darrell Huff, *How to Lie With Statistics* (1993)

6.5 Base Rate Fallacy

In the base rate fallacy, someone draws an unjustified conclusion about an event, often in causal terms, because he or she ignores (intentionally or unintentionally) the rate at which that event normally occurs. For example, someone might say: "When I exercise strenuously, I get sore. But when I drink water immediately after strenuous exercise, my soreness goes away in a couple of days. So, if you're afraid you'll be sore, you should try drinking water." The speaker here is inferring a causal relationship between *drinking water immediately after working out* and *the waning of soreness*. But the careful thinker (and experienced athlete) will recognize that, whether you drink water or not, soreness will wane in a couple of days. Even if every time you drink water soreness wanes, since soreness still wanes 100% of the time *regardless of whether you drink water*, this successive relationship (*A*, then *B*) is not sufficient for concluding that *A* is the cause of *B*.

Similarly, physicians will often say, "You have a cold. I can prescribe a medication, and it will last about two weeks. Or we can just leave it alone, and it will last about 14 days." This is because physicians understand the base rate duration of a cold, and that, since colds are viral, medical treatments have no causal effect on that rate.

Other examples are more interesting. Consider someone who takes a medical test that is 99% reliable and she is told that she has tested positive. What is the likelihood that she actually has the disease? It is tempting to think 99%, but this is misleading. The probability that she actually has the disease depends on *both* how reliable the test is *and* on how likely it is that anyone ever gets the disease. Let's say you also learn that only 1 person in 100,000,000 ever gets this disease (that's about 3 people in the entire United States). Even if your *test* is really good (99% reliable), the likelihood that you actually have the disease is low (just over 0.0001% or about one in a million).[1] Ignoring the base rate of the disease can lead to gross miscalculations, and therefore, unjustified beliefs.

This mistake is common when people try to evaluate the effects of a change after the change has already been made. For example, if your throat is sore, you might start looking at it in the mirror and thinking it looks pretty bad. But if you don't often look at your throat when you're well, how could you tell? What does your throat *normally* look like?

Similarly, if we pass a law to decrease the rate at which a social problem occurs, and the rate does, in fact, decrease, it might seem that the law was successful. But that conclusion follows only if the reduction was not already occurring, that is, if the base rate of the reduction was changed by the law. For example, in 1974, a federal speed limit of 55 miles per hour was set on interstate highways in order to reduce the number of automobile deaths. From 1973 to 1980, the rate of interstate deaths dropped 17 percent. It might seem that the law was successful. But as it turns out, in the seven years prior to the law (1966–1973), the rate dropped 26 percent. And the greatest drop occurred between 1934 and 1949, when there were virtually no safety regulations on interstate travel at all. When you calculate the rate at which interstate deaths decreased from 1934 to 1980, the 55 mile per hour speed law does not seem to have played a causal role in that decrease. In part as a recognition of this, in 1987 Congress began allowing individual states the option to increase speed limits on rural interstate highways.

Legitimate reasons to ignore base rates?

Base rates are not relevant to all probability calculations – for instance, determining the reliability of a test (false positive and false negative rates) or determining the probability of drawing an ace of spades from a deck of 52 cards. In addition, in some cases, we have good reasons to prefer other information to base rates. Very low probability events do occur, and base rates do not always help determine when they have.

Consider the case of miracles (supernatural suspensions of natural operations). Some have argued that we have good reason to disbelieve that any miracles occur because this is not "the way the world works." In other words, the base rate of the occurrence of miracles is practically zero (even if there were one or two in the history of the world, the base rate is extremely low). On the other hand, weird things do happen in nature at a calculable rate. For example, spontaneous remission of many types of cancer occurs at a base rate of just under 1%. That makes remission unlikely, but possible. But since we don't have anything like this for miracles, it may seem that remission is always more likely than that a miracle has occurred.

But now consider, alternatively, the evolution of consciousness (of the sort we humans seem to have). Assuming that it's in some sense unique, consciousness of a human sort has occurred only once in all of evolutionary history: in other words, the base rate at which evolution has produced human consciousness independently of humans is zero. Should we then believe that evolution is not responsible for consciousness? Note that this is the same reasoning strategy that was applied to miracles. (Be careful not to fall prey to the *ad ignorantiam* fallacy (5.7) here, though; not having reasons or evidence for disbelief is not the same as having reasons or evidence for belief.)

Someone might object that, unlike miracles, we know evolution produces unique, new things, such as photosynthetic processes and echolocation. But notice this reply introduces new information, something else we know about evolution. The miracle-defender can make a similar move: *if* a divine being exists, there is no obstacle to its suspending the laws of nature that it created. In both cases, other information about the nature of the putative cause (whether natural selection produces new things; whether God exists) is more important than base rates in evaluating whether the cause has been accurately identified. What these examples suggest is not that we should ignore base rates, but that, in some cases, we need to know much more than base rates in order to form a rational belief about the cause of an event.

Examples of the base rate fallacy

1. "Ever since I bought this bear repellent, I haven't seen a bear. It must really work."
2. "My cold goes away after I take echinacea. I'm telling you, it really works."
3. "Every time I wash my car, it rains. Why me!?"
4. "I read my horoscope every morning, and nothing too bad ever happens to me. Therefore, if you want your life to go well, you should start reading your horoscope."

SEE ALSO

6.3 Fallacies about Causation
6.4 Inductive Statistical Reasoning
8.10 Evidence: Weak and Strong

READING

Rolf Dobelli, *The Art of Thinking Clearly* (2014)
Lawrence Shapiro, "A drop in the sea: What are the odds that Jesus rose or that Moses parted the waves? Even with the best witnesses, vanishingly small," *Aeon Magazine* (2013)
Daniel Kahneman, *Thinking, Fast and Slow* (2011)

6.6 Slippery Slope and *Reductio ad Absurdum*

In the slippery slope fallacy, an arguer attempts to refute a proposed claim or behavior on the grounds that believing the claim or performing the act initiates a causal chain leading to dire, unacceptable, or unwanted consequences. In doing so, however, the arguer fails to justify one or more of the causal links. The slippery slope fallacy is related to the *post hoc* fallacy in that it attributes causal relationships where there may