CHAPTER 11

The Basic Ideas in Statistical Inference

Probabilistic statistical inference is a crucial part of the process of informing ourselves about the world around us. Statistics and statistical inference help us understand our world and make sound decisions about how to act.

More specifically, statistical inference is the process of drawing conclusions about populations or other collections of objects about which we have only partial knowledge from samples. Technically, inference may be defined as the selection of a probabilistic model to resemble the process you wish to investigate, investigation of that model's behavior, and interpretion of the results. Fuller understanding of the nature of statistical inference comes with practice in handling a variety of problems.

Until the 18th century, humanity's extensive knowledge of nature and technology was not based on formal probabilistic statistical inference. But now that we have already dealt with many of the big questions that are easy to answer without probabilistic statistics, and now that we live in a more ramified world than in earlier centuries, the methods of inferential statistics become ever more important.

Furthermore, statistical inference will surely become ever more important in the future as we voyage into realms that are increasingly difficult to comprehend. The development of an accurate chronometer to tell time on sea voyages became a crucial need when Europeans sought to travel to the New World. Similarly, probability and statistical inference become crucial as we voyage out into space and down into the depths of the ocean and the earth, as well as probe into the secrets of the microcosm and of the human mind and soul.

Where probabilistic statistical inference is employed, the inferential procedures may well not be the crucial element. For example, the wording of the questions asked in a public-opinion poll may be more critical than the statistical-inferential procedures used to discern the reliability of the poll results. Yet we dare not disregard the role of the statistical procedures.

Knowledge without probabilistic statistical inference

Let us distinguish two kinds of knowledge with which inference at large (that is, not just probabilistic statistical inference) is mainly concerned: a) one or more *absolute* measurements on one or more dimensions of a collection of one or more items—for example, your income, or the mean income of the people in your country; and b) *comparative* measurements and evaluations of two or more collections of items (especially whether they are equal or unequal)—for example, the mean income in Brazil compared to the mean income in Argentina. Types (a) and (b) both include asking whether there has been a *change* between one observation and another.

What is the conceptual basis for gathering these types of knowledge about the world? I believe that our rock bottom conceptual tool is the assumption of what we may call *sameness*, or *continuity*, or *constancy*, or *repetition*, or *equality*, or *persistence*; "constancy" and "continuity" will be the terms used most frequently here, and I shall use them interchangeably.

Continuity is a non-statistical concept. It is a best guess about the next point beyond the known observations, without any idea of the accuracy of the estimate. It is like testing the ground ahead when walking in a marsh. It is local rather than global. We'll talk a bit later about why continuity seems to be present in much of the world that we encounter.

The other great concept in statistical inference, and perhaps in all inference taken together, is *representative (usually random) sampling*, to be discussed in Chapter 12. Representative sampling—which depends upon the assumption of sameness (homogeneity) throughout the universe to be investigated—is quite different than continuity; representative sampling assumes that there is *no greater chance* of a connection between any two elements that might be drawn into the sample than between any other two elements; the order of drawing is immaterial. In contrast, continuity assumes that *there is a greater*

chance of connection between two contiguous elements than between either one of the elements and any of the many other elements that are not contiguous to either. Indeed, the process of randomizing is a device for doing away with continuity and autocorrelation within some bounded closed system—the sample "frame." It is an attempt to map (describe) the entire area ahead using the device of the systematic survey. Random representative sampling enables us to make probabilistic inferences about a population based on the evidence of a sample.

To return now to the concept of sameness: Examples of the principle are that we assume: a) our house will be in the same place tomorrow as today; b) a hammer will break an egg every time you hit the latter with the former (or even the former with the latter); c) if you observe that the first fifteen persons you see walking out of a door at the airport are male, the sixteenth probably will be male also; d) paths in the village stay much the same through a person's life; e) religious ritual changes little through the decades; f) your best guess about tomorrow's temperature or stock price is that will be the same as today's. This principle of constancy is related to David Hume's concept of *constant conjunction*.

When my children were young, I would point to a tree on our lawn and ask: "Do you think that tree will be there tomorrow?" And when they would answer "Yes," I'd ask, "Why doesn't the tree fall?" That's a tough question to answer.

There are two reasonable bases for predicting that the tree will be standing tomorrow. First and most compelling for most of us is that almost all trees continue standing from day to day, and this particular one has never fallen; hence, what has been in the past is likely to continue. This assessment requires no scientific knowledge of trees, yet it is a very functional way to approach most questions concerning the trees—such as whether to hang a clothesline from it, or whether to worry that it will fall on the house tonight. That is, we can predict the outcome in this case with very high likelihood of being correct even though we do not utilize anything that would be called either science or statistical inference. (But what do you reply when your child says: "Why should I wear a seat belt? I've never been in an accident"?)

A second possible basis for prediction that the tree will be standing is scientific analysis of the tree's roots—how the tree's weight is distributed, its sickness or health, and so on. Let's put aside this sort of scientific-engineering analysis for now.

The first basis for predicting that the tree will be standing tomorrow—sameness—is the most important heuristic device in all of knowledge-gathering. It is often a weak heuristic; certainly the prediction about the tree would be better grounded (!) after a skilled forester examines the tree. But persistence alone might be a better heuristic in a particular case than an engineering-scientific analysis alone.

This heuristic appears more obvious if the child—or the adult—were to respond to the question about the tree with another question: Why should I expect it to *fall*? In the absence of some reason to expect change, it is quite reasonable to expect no change. And the child's new question does not duck the central question we have asked about the tree, any more than one ducks a probability estimate by estimating the complementary probability (that is, unity minus the probability sought); indeed, this is a very sound strategy in many situations.

Constancy can refer to location, time, relationship to another variable, or yet another dimension. Constancy may also be cyclical. Some cyclical changes can be charted or mapped with relative certainty—for example the life-cycles of persons, plants, and animals; the diurnal cycle of dark and light; and the yearly cycle of seasons. The courses of some diseases can also be charted. Hence these kinds of knowledge have long been well known.

Consider driving along a road. One can predict that the price of the next gasoline station will be within a few cents of the gasoline station that you just passed. But as you drive further and further, the dispersion increases as you cross state lines and taxes differ. This illustrates continuity.

The attention to constancy can focus on a single event, such as leaves of similar shape appearing on the same plant. Or attention can focus on single sequences of "production," as in the process by which a seed produces a tree. For example, let's say you see two puppies—one that looks like a low-slung dachshund, and the other a huge mastiff. You also see two grown male dogs, also apparently dachshund and mastiff. If asked about the parentage of the small ones, you are likely—using the principle of sameness—to point—quickly and with surety—to the adult dogs of the same breed. (Here it is important to notice that this answer implicitly assumes that the fathers of the puppies are among these dogs. But the fathers might be somewhere else entirely; it is in these ways that the principle of sameness can lead you astray.)

When applying the concept of sameness, the object of interest may be collections of data, as in Semmelweiss's data on the consistent differences in rates of maternal deaths from childbed fever in two clinics with different conditions (see Table 11-1), or the similarities in sex ratios from year to year in Graunt's data on London births (Table 11-2), or the stark effect in John Snow's data on the numbers of cholera cases associated with two London wells (Table 11-3), or the reduction in beriberi among Japanese sailors as a result of a change in diet (Table 11-4). These data seem so overwhelmingly clear cut that our naive statistical sense makes the relationships seem deterministic, and the conclusions seems straightforward. (But the same statistical sense frequently misleads us when considering sports and stock market data.)

Table 11-1
Deaths of Mothers

	First Clinic			Second Clinic		
	Births	**Deaths**	**Rate**	**Births**	**Deaths**	**Rate**
1841	3,036	237	7.8	2,442	86	3.5
1842	3,287	518	15.8	2,659	202	7.6
1843	3,060	274	8.9	2,739	164	6.0
1844	3,157	260	8.2	2,956	68	2.3
1845	3,492	241	6.9	3,241	66	2.0
1845	4,010	459	11.4	3,754	105	2.8
Total	20,042	1,989		17,791	691	
Avg.			9.9			3.9

Source: Semmelweis, Ignaz, *The Etiology, Concept, and Prophylaxis of Childbed Fever*, Translated and edited by K. Codell Carter (Madison, Wisconsin: Univ. of Wisconsin Press, 1983), p. 64.

Table 11-2
Ratio of Number of Males to Number of Females

	London
Period	**Christenings**
1629-1636	1,072
1637-1640	1,073
1641-1648	1,063
1649-1656	1,095
1657-1660	1,069

Source: Graunt, John, *Natural and Political Observations Mentioned in a Following Index and Made Upon the Bills of Mortality* (Reprint Edition) (New York; Arno Press, 1662/1975).

Table 11-3
John Snow's Data on Cholera Rates for Three Wells

Southwark and Vauxhall Supply	71 deaths per 10,000 houses
Lambeth Supply	5 deaths per 10,000 houses
Rest of London	9 deaths per 10,000 houses

Source: Winslow, Charles-Edward Amory, *The Conquest of Epidemic Disease* (Madison, Wisconsin: Univ. of Wisconsin Press, 1980), p. 276.

Table 11-4
Takaki's Japanese Naval Records of Deaths from Beriberi

Year	Diet	Total Navy Personnel	Deaths from Beriberi
1880	Rice diet	4,956	1,725
1881	Rice diet	4,641	1,165
1882	Rice diet	4,769	1,929
1883	Rice Diet	5,346	1,236
1884	Change to new diet	5,638	718
1885	New diet	6,918	41
1886	New diet	8,475	3
1887	New diet	9,106	0
1888	New diet	9,184	0

Source: K. Takaki, in Kornberg, 1989, p. 9

Constancy and sameness can be seen in macro structures; consider, for example, the constant location of your house. Constancy can also be seen in micro aggregations—for example, the raindrops and rain that account for the predictably fluctuating height of the Nile, or the ratio of boys to girls born in London, cases in which we can *average* to see the "statistical" sameness. The total sum of the raindrops produces the level of a reservoir or a river from year to year, and the sum of the behaviors of collections of persons causes the birth rates in the various years.

Statistical inference is only needed when a person thinks that s/he *might* have found a pattern but the pattern is not completely obvious to all. Probabilistic inference works to test—either to confirm or discount—the belief in the pattern's existence. We will see such cases in the following chapter.

People have always been forced to think about and act in situations that have not been constant—that is, situations where the amount of variability in the phenomenon makes it impossible to draw clear cut, sensible conclusions. For example, the appearance of game animals in given places and at given times has always been uncertain to hunters, and therefore it has always been difficult to know which target to hunt in which place at what time. And of course variability of the weather has always made it a very uncertain element. The behavior of one's enemies and friends has always been uncertain, too, though uncertain in a manner different from the behavior of wild animals; there often is a gaming element in interactions with other humans. But in earlier times, data and techniques did not exist to enable us to bring statistical inference to bear.

The treatment of uncertainty

The purpose of *statistical* inference is to help us peer through the veil of variability when it obscures the main thrust of the data, so as to improve the decisions we make. Statistical inference (or in most cases, simply probabilistic estimation) can help a) a gambler deciding on the appropriate odds in a betting game when there seems to be little or no difference between two or more outcomes; b) an astronomer deciding upon one or another value as the central estimate for the location of a star when there is considerable variation in the observations s/he has made of the star; c) a basketball coach pondering whether to remove from the game her best shooter who has heretofore done poorly tonight; d) an oil-drilling firm debating whether to follow up a test-well drilling with a full-bore drilling when the probability of success is not overwhelming but the payoff to a gusher could be large.

Returning to the tree near the Simon house: Let's change the facts. Assume now that one major part of the tree is mostly dead, and we expect a big winter storm tonight. What is the danger that the tree will fall on the house? Should we spend $1500 to have the mostly-dead third of it cut down? We know that last year a good many trees fell on houses in the neighborhood during such a storm.

We can gather some data on the proportion of old trees this size that fell on houses—about 5 in 100, so far as we can tell. Now it is no longer an open-and-shut case about whether the tree will be standing tomorrow, and we are using statistical inference to help us with our thinking. We proceed to find a

set of trees *that we consider similar to this one*, and study the variation in the outcomes of such trees. So far we have estimated that the *average* for this group of trees—the mean (proportion) that fell in the last big storm—is 5 percent. Averages are much more "stable"—that is, more similar to each other—than are individual cases.

Notice how we use the crucial concept of sameness: We assume that our tree is like the others we observed, or at least that it is not systematically different from most of them and it is more-or-less average.

How would our thinking be different if our data were that one tree in 10 had fallen instead of 5 in 100? This is a question in statistical inference.

How about if we investigate further and find that 4 of 40 *elms* fell, but only one of 60 *oaks*, and ours is an oak tree. Should we consider that oaks and elms have different chances of falling? Proceeding a bit further, we can think of the question as: Should we or should we not consider oaks and elms as different? This is the type of statistical inference called "hypothesis testing": We apply statistical procedures to help us decide whether to treat the two classes of trees as the same or different. If we should consider them the same, our worries about the tree falling are greater than if we consider them different with respect to the chance of damage.

Notice that statistical inference was not necessary for accurate prediction when I asked the kids about the likelihood of a live tree falling on a day when there would be no storm. So it is with most situations we encounter. But when the assumption of constancy becomes shaky for one reason or another, as with the sick tree falling in a storm, we need a more refined form of thinking. We collect data on a large number of instances, inquire into whether the instances in which we are interested (our tree and the chance of it falling) are representative—that is, whether it resembles what we would get if we drew a sample randomly—and we then investigate the behavior of this large class of instances to see what light it throws on the instances(s) in which we are interested.

The procedure in this case—which we shall discuss in greater detail later on—is to ask: If oaks and elms are *not* different, how likely is it that only one of 60 oaks would fall whereas 4 of 40 elms would fall? Again, notice the assumption that our tree is "representative" of the other trees about which we have information—that it is not systematically different from most of them, but rather that it is more-or-less average. Our tree cer-

tainly was not chosen randomly from the set of trees we are considering. But for purposes of our analysis, we proceed *as if* it had been chosen randomly—because we deem it "representative."

This is the first of two roles that the concept of randomness plays in statistical thinking. Here is an example of the second use of the concept of randomness: We conduct an experiment—plant elm and oak trees at *randomly-selected* locations on a plot of land, and then try to blow them down with a wind-making machine. (The random selection of planting spots is important because some locations on a plot of ground have different growing characteristics than do others.) Some purists object that *only* this sort of experimental sampling is a valid subject of statistical inference; it can never be appropriate, they say, to simply *assume* on the basis of other knowledge that the tree is representative. I regard that purist view as a helpful discipline on our thinking. But accepting its conclusion—that one should not apply statistical inference except to randomly-drawn or randomly-constituted samples—would take from us a tool that has proven useful in a variety of activities.

As discussed earlier in this chapter, the data in some (probably most) scientific situations are so overwhelming that one can proceed without probabilistic inference. Historical examples include those shown above of Semmelweiss and puerperal fever, and John Snow and cholera. But where there was lack of overwhelming evidence, the causation of many diseases long remained unclear for lack of statistical procedures. This led to superstitious beliefs and counter-productive behavior, such as quarantines against plague often were. Some effective practices also arose despite the lack of sound theory, however—the waxed costumes of doctors, and the burning of mattresses, despite the wrong theory about the causation of plague; see Cipolla, 1981)

So far I have spoken only of *predictability* and not of other elements of statistical knowledge such as *understanding* and *control*. This is simply because statistical *correlation* is the bed rock of most scientific understanding, and predictability. Later we will expand the discussion beyond predictability; it holds no sacred place here.

---

[1] It is because hypothesis testing focuses on this most basic of inferential processes—deciding "same" or "different"—that I believe it to be a more basic technique than estimating confidence intervals, which focus on the accuracy of estimates.

Where statistical inference becomes crucial

There was little role for statistical inference until about three centuries ago because there existed very few scientific data. When scientific data began to appear, the need emerged for statistical inference to improve the interpretation of the data. As we saw, statistical inference is not needed when the evidence is overwhelming. A thousand cholera cases at one well and zero at another obviously does not require a statistical test. Neither would 999 cases to one, or even 700 cases to 300, because our inbred and learned statistical senses can detect that the two situations are different. But probabilistic inference is needed when the number of cases is relatively small or where for other reasons the data are somewhat ambiguous.

For example, when working with the 17th century data on births and deaths, John Graunt—great statistician though he was—drew wrong conclusions about some matters because he lacked modern knowledge of statistical inference. For example, he found that in the rural parish of Romsey "there were born 15 Females for 16 Males, whereas in London there were 13 for 14, which shows, that London is somewhat more apt to produce Males, then the country" (p. 71). He suggests that the "curious" inquire into the causes of this phenomenon, apparently not recognizing—and at that time he had no way to test—that the difference might be due solely to chance. He also notices (p. 94) that the variations in deaths among years in Romsey were greater than in London, and he attempted to explain this apparent fact (which is just a statistical artifact) rather than understanding that this is almost inevitable because Romsey is so much smaller than London. Because we have available to us the modern understanding of variability, we can now reach sound conclusions on these matters.

Summary statistics—such as the simple mean—are devices for reducing a large mass of data (inevitably confusing unless they are absolutely clear cut) to something one can manage to understand. And probabilistic inference is a device for determining whether patterns should be considered as facts or artifacts.

Here is another example that illustrates the state of early quantitative research in medicine:

> Exploring the effect of a common medicinal substance, Boecker examined the effect of sasparilla on the nitrogenous and other constituents of the urine. An individual receiving a controlled diet was given

> a decoction of sasparilla for a period of twelve days, and the volume of urine passed daily was carefully measured. For a further twelve days that same individual, on the same diet, was given only distilled water, and the daily quantity of urine was again determined. The first series of researches gave the following figures (in cubic centimeters): 1,467, 1,744, 1,665, 1,220, 1,161, 1,369, 1,675, 2,199, 887, 1,634, 943, and 2,093 (mean = 1,499); the second series: 1,263, 1,740, 1,538, 1,526, 1,387, 1,422, 1,754, 1,320, 1,809, 2,139, 1,574, and 1,114 (mean = 1,549). Much uncertainty surrounded the exactitude of these measurements, but this played little role in the ensuing discussion. The fundamental issue was not the quality of the experimental data but how inferences were drawn from those data (Coleman in Kruger, 1987, p. 207).

The experimenter Boecker had no reliable way of judging whether the data for the two groups were or were not meaningfully different, and therefore he arrived at the unsound conclusion that there was indeed a difference. (Gustav Radicke used this example as the basis for early work on statistical significance.)

Another example: Joseph Lister convinced the scientific world of the germ theory of infection, and the possibility of preventing death with a disinfectant, with these data: Prior to the use of antiseptics—16 post-operative deaths in 35 amputations; subsequent to the use of antiseptics—6 deaths in 40 amputations (Winslow, 1943, p. 303). But how sure could one be that a difference of that size might not occur just by chance? No one then could say, nor did anyone inquire, apparently.

Here's another example of great scientists falling into error because of a too-primitive approach to data (Feller, 3rd ed, 1968, pp. 69-70): Charles Darwin wanted to compare two sets of measured data, each containing 16 observations. At Darwin's request, Francis Galton compared the two sets of data by ranking each, and then comparing them pairwise. The a's were ahead 13 times. Without knowledge of the actual probabilities Galton concluded that the treatment was effective. But, assuming perfect randomness, the probability that the a's beat [the others] 13 times or more equals 3/16. This means that in three out of sixteen cases a perfectly ineffectual treatment would appear as good or better than the treatment classified as effective by Galton.

That is, Galton and Darwin reached an unsound conclusion. As Feller says, "This shows that a quantitative analysis may be a valuable supplement to our rather shaky intuition" (p. 70).

Looking ahead, the key tool in situations like Graunt's and Boecker's and Lister's is creating *ceteris paribus*—making "everything else the same"—with random selection in experiments, or at least with statistical controls in non-experimental situations.

Conclusions

In all knowledge-seeking and decision-making, our aim is to peer into the unknown and reduce our uncertainty a bit. The two main concepts that we use—the two great concepts in all of scientific knowledge-seeking, and perhaps in all practical thinking and decision-making—are a) continuity (or non-randomness) and the extent to which it applies in given situation, and b) random sampling, and the extent to which we can assume that our observations are indeed chosen by a random process.

Endnotes

1. These are cases of David Hume's "constant conjunction."

2. I benefited from the discussion of this matter by Hald, 1990, p. 93ff.

3. A peculiar perverseness associated with the new knowledge of statistical inference is that very strong findings, which require little or no formal inference to demonstrate and which are so powerful that they can be shown with a simple graph or table, are very hard to publish in social science literature because they do not meet the tests of "rigor," and "elegance." Editors view them as detracting from the "technical level" of their journals. A good many of the greatest discoveries of the past would nowadays fall in this category of being difficult or impossible to publish.