CHAPTER 12 Introduction to Statistical Inference

Statistical Inference and Random Sampling
Summary and Conclusions

The usual goal of a statistical inference is a decision about which of two or more hypotheses a person will thereafter choose to believe and act upon. The strategy of such inference is to consider the behavior of a given universe in terms of the samples it is likely to produce, and if the observed sample is *not* a likely outcome of sampling from that universe, we then proceed as if the sample did not in fact come from that universe. (The previous sentence is a restatement in somewhat different form of the core of statistical analysis.)

Statistical inference and random sampling

Continuity and sameness is the fundamental concept in inference in general, as discussed in Chapter 11. Random sampling is the second great concept in inference, and it distinguishes probabilistic statistical inference from non-statistical inference as well as from non-probabilistic inference based on statistical data.

Let's begin the discussion with a simple though unrealistic situation. Your friend Arista a) looks into a cardboard carton, b) reaches in, c) pulls out her hand, and d) shows you a green ball. What might you reasonably infer?

You might at least be fairly sure that the green ball came from the carton, though you recognize that Arista might have had it concealed in her hand when she reached into the carton. But there is not much more you might reasonably conclude at this point except that there was at least one green ball in the carton to start with. There could be no more balls; there could be many green balls and no others; there could be a thousand red balls and just one green ball; and there could be one green ball, a hundred balls of different colors, and two pounds of

mud—given that she looked in first, it is not improbable that she picked out the only green ball among other material of different sorts.

There is not much you could say with confidence about the probability of yourself reaching into the same carton with your eyes closed and pulling out a single green ball. To use other language (which some philosophers might say is not appropriate here as the situation is too specific), there is little basis for induction about the contents of the box. Nor is the situation very different if your friend reaches in three times in a row and hands you a green ball each time.

So far we have put our question rather vaguely. Let us frame a more precise inquiry: What do we predict about the next item(s) we might draw from the carton? If we assume—based on who-knows-what information or notions—that another ball will emerge, we could simply use the principle of sameness and (until we see a ball of another color) predict that the next ball will be green, whether one or three or 100 balls is (are) drawn.

But now what about if Arista pulls out nine green balls and one red ball? The principle of sameness cannot be applied as simply as before. Based on the last previous ball, the next one will be red. But taking into account *all* the balls we have seen, the next will "probably" be green. We have no solid basis on which to go further. There cannot be any "solution" to the "problem" of reaching a general conclusion on the basis of these specific pieces of evidence.

Now consider what you might conclude if you were told that a single green ball had been drawn *with a random sampling procedure* from a box containing nothing but balls. Knowledge that the sample was drawn randomly from a given universe is grounds for belief that one knows much more than if a sample were not drawn randomly. First, you would be sure—if you had reasonable basis to believe that the sampling really was random, which is not easy to guarantee—that the ball came from the box. Second, you would guess that the proportion of green balls is not very small, because if there are only a few green balls and many other-colored balls, it would be unusual—that is, the event would have a low probability—to draw a green ball. Not impossible, but unlikely. And *we can compute the probability of drawing a green ball*—or any other combination of colors—*for different assumed compositions within the box*. So the knowledge that the sampling process is random greatly increases our ability—or our confidence in our ability—to infer the contents of the box.

Let us note well the strategy of the previous paragraph: *Ask about the probability that one or more various possible contents of the box (the "universe") will produce the observed sample*, on the assumption that the sample was drawn randomly. *This is the central strategy of all statistical inference*, though I do not find it so stated elsewhere. We shall come back to this idea shortly.

There are several kinds of questions one might ask about the contents of the box. One general category includes questions about our best guesses of the box's contents—that is, questions of *estimation*. Another category includes questions about our *surety* of that description, and our surety that the contents are similar or different from the contents of other boxes; the consideration of surety follows after estimates are made. The estimation questions can be subtle and unexpected (Savage, 1972, Chapter 15), but do not cause major controversy about the foundations of statistics. So we can quickly move on to questions about the extent of surety in our estimations.

Consider your reaction if the sampling produces 10 green balls in a row, or 9 out of 10. If you had no other information (a very important assumption that we will leave aside for now), your best guess would be that the box contains all green balls, or a proportion of 9 of 10, in the two cases respectively. This estimation process seems natural enough.

You would be surprised if someone told you that instead of the box containing the proportion in the sample, it contained just *half* green balls. How surprised? Intuitively, the extent of your surprise would depend on the probability that a half-green "universe" would produce 10 or 9 green balls out of 10. This surprise is a key element in the logic of the hypothesis-testing branch of statistical inference.

We learn more about the likely contents of the box by asking about the probability that various *specific* populations of balls within the box would produce the particular sample that we received. That is, we can ask how likely a collection of 25 percent green balls is to produce (say) 9 of 10 green ones, and how likely collections of 50 percent, 75 percent, 90 percent (and any other collections of interest) are to produce the observed sample. That is, we ask about the *consistency* between any particular hypothesized collection within the box and the sample we observe. And it is reasonable to believe that those universes which have greater consistency with the observed sample—that is, those universes that are more likely to produce the observed sample—are more likely to be in the box than other universes. This (to repeat, as I shall repeat many times) is the

basic strategy of statistical investigation. If we observe 9 of 10 green balls, we then determine that universes with (say) 9/10 and 10/10 green balls are more consistent with the observed evidence than are universes of 0/10 and 1/10 green balls. So by this process of considering specific universes that the box *might* contain, we make possible more specific inferences about the box's *probable* contents based on the sample evidence than we could without this process.

Please notice the role of the assessment of probabilities here: By one technical means or another (either simulation or formulas), we assess the probabilities that a particular universe will produce the observed sample, and other samples as well.

It is of the highest importance to recognize that without additional knowledge (or assumption) one cannot make any statements about the probability of the sample having come from *any particular universe*, on the basis of the sample evidence. (Better read that last sentence again.) We can only speak about the probability that a particular universe *will produce* the observed sample, a very different matter. This issue will arise again very sharply in the context of confidence intervals.

Let us generalize the steps in statistical inference:

1. Frame the original question as: What is the chance of getting the observed sample x from population X? That is, what is probability of (If x then X)?

2. Proceed to this question: What kinds of samples does X produce, with which probability? That is, what is the probability of this particular x coming from X? That is, what is $p(x \mid X)$?

3. Actually investigate the behavior of X with respect to x and other samples. One can do this in two ways:

> a. Use the formulaic calculus of probability, perhaps resorting to Monte Carlo methods if an appropriate formula does not exist. Or,
>
> b. Use resampling (in the larger sense), the domain of which equals (all Monte Carlo experimentation) minus (the use of Monte Carlo methods for approximations, investigation of complex functions in statistics and other theoretical mathematics, and uses elsewhere in science). Resampling in its more restricted sense includes the bootstrap, permutation tests, and other non-parametric methods.

4. Interpretation of the probabilities that result from step 3 in terms of i) acceptance or rejection of hypotheses, ii) surety of conclusions, or iii) inputs to decision theory.

Here is a short definition of statistical inference: *The selection of a probabilistic model that might resemble the process you wish to investigate, the investigation of that model's behavior, and the interpretation of the results.*

We will get even more specific about the procedure when we discuss the canonical procedures for hypothesis testing and for the finding of confidence intervals in the chapters on those subjects.

The discussion so far has been in the spirit of what is known as *hypothesis testing*. The result of a hypothesis test is a decision about whether or not one believes that the sample is likely to have been drawn randomly from the "benchmark universe" X. The logic is that if the probability of such a sample coming from that universe is low, we will then choose to believe the alternative—to wit, that the sample came from the universe that resembles the sample. The underlying idea is that if an event would be very surprising if it really happened—as it would be very surprising if the dog had really eaten the homework (see Chapter 15)—we are inclined not to believe in that possibility. (This logic will be explored further in later chapters on hypothesis testing.)

We have so far assumed that our only relevant knowledge is the sample. And though we almost never lack *some* additional information, this can be a sensible way to proceed when we wish to *suppress* any other information or speculation. This suppression is controversial; those known as Bayesians or subjectivists want us to take into account all the information we have. But even they would not dispute suppressing information in certain cases—such as a teacher who does not want to know students' IQ scores because s/he might want avoid the possibility of unconsciously being affected by that score, or an employer who wants not to know the potential employee's ethnic or racial background even though it might improve the hiring process, or a sports coach who refuses to pick the starting team each year until the players have competed for the positions. If the Bayesians will admit the reasonability of suppressing information in at least some situations, it will be a major step in accommodation and in bringing all views into greater harmony. (More about this topic in the appendix).

Now consider a variant on the green-ball situation discussed above. Assume now that you are told that samples of balls are

alternately drawn from one of two *specified* universes—two urns of balls, one with 50 percent green balls and the other with 80 percent green balls. Now you are shown a sample of nine green and one red balls drawn from one of those urns. On the basis of your sample you can then say how probable it is that the sample came *from one or the other universe*. You proceed by computing the probabilities (often called the *likelihoods* in this situation) that each of those two universes would individually produce the observed samples—probabilities that you could arrive at with resampling, with Pascal's Triangle, or with a table of binomial probabilities, or with the Normal approximation and the Z distribution, or with yet other devices. Those probabilities are .01 and .27, and the ratio of the two (0.1/.27) is a bit less than .04. That is, fair betting odds are about 1 to 27.

Let us consider a genetics problem on this model. Plant A produces 3/4 black seeds and 1/4 reds; plant B produces all reds. You get a red seed. Which plant would you guess produced it? You surely would guess plant B. Now, how about 9 reds and a black, from Plants A and C, the latter producing 50 percent reds on average?

To put the question more precisely: What betting odds would you give that the one red seed came from plant B? Let us reason this way: If you do this again and again, 4 of 5 of the red seeds you see will come from plant B. Therefore, reasonable (or "fair") odds are 4 to 1, because this is in accord with the ratios with which red seeds are produced by the two plants—4/4 to 1/4.

How about the sample of 9 reds and a black, and plants A and C? It would make sense that the appropriate odds would be derived from the probabilities of the two plants producing that particular sample, probabilities which we computed above.

Now let us move to a bit more complex problem: Consider two urns—urn G with 2 red and 1 black balls, and urn H with 100 red and 100 black balls. Someone flips a coin to decide which urn will be drawn from, reaches into that urn, and chooses two balls without replacing the first one before drawing the second. Both are red. What are the odds that the sample came from urn G? Clearly, the answer should derive from the probabilities that the two urns would produce the observed sample.

(Now just for fun, how about if the first ball drawn is thrown back after examining? What now are the appropriate odds?)

Let's restate the central issue. One can state the probability that a particular plant which produces *on average* 1 red and 3 black seeds will produce one red seed, or 5 reds among a sample of 10. But without further assumptions—such as the assumption above that the possibilities are limited to two specific universes—one cannot say how likely a given red seed is to have come from a given plant, even if we know that that plant produces only reds. (For example, it may have come from *other* plants producing only red seeds.)

When we limit the possibilities to two universes (or to a larger set of specified universes) we are able to put a probability on one hypothesis or another. But to repeat, in many or most cases, one cannot reasonably assume it is *only* one or the other. And then we cannot state any odds that the sample came from a particular universe. This is a very difficult point to grasp, experience shows, but a crucial one. (It is the sort of subtle issue that makes statistics so difficult.)

The additional assumptions necessary to talk about the probability that the red seed came from a given plant are the stuff of statistical inference. And they must be combined with such "objective" probabilistic assessments as the probability that a 1-red-3-black plant will produce one red, or 5 reds among 10 seeds.

Now let us move one step further. Instead of stating as a fact under our control that there is a .5 chance of the sample being drawn from each of the two urns in the problem above, let us assume that we do not *know* the probability of each urn being picked, but instead we *estimate* a probability of .5 for each urn, based on a variety of other information that all is uncertain. But though the facts are now different, the most reasonable estimate of the odds that the observed sample was drawn from one or the other urn will not be different than before—because in both situations we were working with a "prior probability" of .5. (The term "prior probability" is the language of the Bayesian approach to statistics.) And when we view the situation this way, the Neyman-Pearson model may be seen perfectly well in a Bayesian framework.

Now let us go a step further by allowing the universes from which the sample may have come to have different assumed probabilities as well as different compositions. That is, we now consider prior probabilities other than .5.

How do we decide which universe(s) to investigate for the probability of producing the observed sample, and of producing samples that are even less likely, in the sense of being more

surprising? That judgment depends upon the purpose of your analysis, upon your point of view of how statistics ought to be done, and upon some other factors.

It should be noted that the logic described so far applies in exactly the same fashion whether we do our work estimating probabilities with the resampling method or with conventional methods. We can figure the probability of nine or more green chips from a universe of (say) p = .7 with either approach.

So far we have discussed the *comparison* of various hypotheses and possible universes. We must also consider where the consideration of the *reliability* of estimates comes in. This leads to the concept of *confidence limits*, which will be discussed in Chapters 20 and 21.

Samples Whose Observations May Have More Than Two Values

So far we have discussed samples and universes that we can characterize as proportions of elements which can have only one of two characteristics—green or other, in this case, which is equivalent to "1" or "0." This expositional choice has been solely for clarity. All the ideas discussed above pertain just as well to samples whose observations may have more than two values, and which may be either discrete or continuous.

Summary and conclusions

A statistical question asks about the probabilities of a sample having arisen from various source universes in light of the evidence of a sample. In every case, the statistical answer comes from considering the behavior of particular specified universes in relation to the sample evidence and to the behavior of other possible universes. That is, a statistical problem is an exercise in postulating universes of interest and interpreting the probabilistic distributions of results of those universes. The preceding sentence is the key operational idea in statistical inference.

Different sorts of realistic contexts call for different ways of framing the inquiry. For each of the established models there are types of problems which fit that model better than other models, and other types of problems for which the model is quite inappropriate.

Fundamental wisdom in statistics, as in all other contexts, is to employ a large tool kit rather than just applying only a hammer, screwdriver, or wrench no matter what the problem is at hand. (Philosopher Abraham Kaplan once stated Kaplan's Law of scientific method: Give a small boy a hammer and there is nothing that he will encounter that does not require pounding.) Studying the text of a poem statistically to infer whether Shakespeare or Bacon was the more likely author is quite different than inferring whether bioengineer Smythe can produce an increase in the proportion of calves, and both are different from decisions about whether to remove a basketball player from the game or to produce a new product.

Some key points: 1) In statistical inference as in all sound thinking, one's *purpose is central*. All judgments should be made relative to that purpose, and in light of costs and benefits. (This is the spirit of the Neyman-Pearson approach). 2) One cannot avoid making judgments; the process of statistical inference cannot ever be perfectly routinized or objectified. Even in science, fitting a model to experience requires judgment. 3) The best ways to infer are different in different situations—economics, psychology, history, business, medicine, engineering, physics, and so on. 4) Different tools must be used when the situations call for them—sequential vs. fixed sampling, Neyman-Pearson vs. Fisher, and so on. 5) In statistical inference it is wise not to argue about the proper conclusion when the data and procedures are ambiguous. Instead, whenever possible, one should go back and get more data, hence lessening the importance of the efficiency of statistical tests. In some cases one cannot easily get more data, or even conduct an experiment, as in biostatistics with cancer patients. And with respect to the past one cannot produce more historical data. But one can gather more and different kinds of data, e.g. the history of research on smoking and lung cancer.

Endnotes

1. Hence I shall merely mention that the method of moments and the method of maximum likelihood serve most of our needs, and often agree in their conclusions; furthermore, we often know when the former may be inappropriate.