CHAPTER 16

The Concept of Statistical Significance in Testing Hypotheses

This chapter offers an interpretation of the meaning of the concept of statistical significance and the term "significant" in connection with the logic of significance tests. It also discusses the concept of "level of significance."

The logic of hypothesis tests

Let's address the logic of hypothesis tests by considering a variety of examples in everyday thinking:

> Consider the nine-year-old who tells the teacher that the dog ate the homework. Why does the teacher not accept the child's excuse? Clearly it is because the event would be too "unusual." But why do we think that way?

Let's speculate that you survey a million adults, and only three report that they have ever heard of a real case where a dog ate somebody's homework. You are a teacher, and a student comes in without homework and says that a dog ate the homework. It could have happened—your survey reports that it really has happened in three lifetimes out of a million. But the event happens *only very infrequently*.

Therefore, you probably conclude that because the event is so unlikely, something else must have happened—and the likeliest alternative is that the student did not do the homework. The logic is that if an event seems very unlikely, it would therefore surprise us greatly if it were to actually happen, and therefore we assume that there must be a better explanation. This is why we look askance at unlikely coincidences when they are to someone's benefit.

The same line of reasoning was the logic of John Arbuthnot's hypothesis test about the ratio of births by sex in the first published hypothesis test, though his extension of his logic to God's design as an alternative hypothesis goes beyond the standard modern framework. It is also the implicit logic in the research on puerperal fever, cholera, and beri-beri, the data for which were shown in Chapter 11, though no explicit mention was made of probability in those cases.

> Two students sat next to each other at an ACT college-entrance examination in Kentucky in 1987. Out of 219 questions, 211 of the answers were identical, including many that were wrong. Student A was a high school athlete in Kentucky who had failed two previous SAT exams, and Student B thought he saw Student A copying from him. Should one believe that Student A cheated? (*The Washington Post*, April 19, 1992, p. D2.)

You say to yourself: It would be most unlikely that the two test-takers would answer that many questions identically by chance—and we can compute how unlikely that event would be. Because that event is so unlikely, we therefore conclude that one or both cheated. And indeed, the testing service invalidated the athlete's exam. On the other hand, if all the questions that were answered identically were *correct*, the result might not be unreasonable. If we knew in how many cases they made the *same mistakes*, the inquiry would have been clearer, but the newspaper did not contain those details.

> The court is hearing a murder case. There is no eye-witness, and the evidence consists of such facts as the height and weight and age of the person charged, and other circumstantial evidence. Only one person in 50 million has such characteristics, and you find such a person. Will you convict the person, or will you believe that the evidence was just a coincidence? Of course the evidence *might* have occurred by bad luck, but the probability is very, very small (1 in 50 million). Will you therefore conclude that because the chance is so small, it is reasonable to assume that the person charged committed the crime?

Sometimes the unusual really happens—the court errs by judging that the wrong person did it, and that person goes to prison or even is executed. The best we can do is to make the criterion strict: "Beyond a reasonable doubt." (People ask: What

probability does that criterion represent? But the court will not provide a numerical answer.)

> Somebody says to you: I am going to deal out five cards and it will be a royal flush—ten, jack, queen, king, and ace of the same suit. The person deals the cards and lo and behold! the royal flush appears. Do you think the occurrence happened just by chance? No, you are likely to be very dubious that it happened by chance. Therefore, you believe there must be some other explanation—that the person fixed the cards, for example.

Note: You don't attach the same meaning to any *other* permutation (say 3, 6, 7, 7, and king of various suits), even though that permutation is just as rare— unless the person announced exactly that permutation in advance.

Indeed, even if the person says *nothing*, you will be surprised at a royal flush, because this hand has *meaning*, whereas another given set of five cards do not have any special meaning.

> You see six Volvos in one home's driveway, and you conclude that it is a Volvo club meeting, or a Volvo salesperson's meeting. Why? Because it is unlikely that six people not connected formally by Volvo ownership would be friends of the same person.

Two important points complicate the concept of statistical significance:

1. With a large enough sample, *every* treatment or variable will seem different from every other. Two faces of even a good die (say, "1" and "2") will produce different results in the very very long run.

2. Statistical significance does not imply economic or social significance. Two faces of a die may be statistically different in a huge sample of throws, but a 1/10,000 difference between them is too small to make an economic difference in betting. Statistical significance is only a *filter*. If it appears, one should then proceed to decide whether there is *substantive* significance.

Interpreting statistical significance is sometimes complex, especially when the interpretation depends heavily upon your prior expectations—as it often does. For example, how should a basketball coach decide whether or not to bench a player for poor performance after a series of missed shots at the basket?

Consider Coach John Thompson who, after Charles Smith missed 10 of 12 shots in the 1989 Georgetown-Notre Dame NCAA game, took Smith out of the game for a time (*The Washington Post*, March 20, 1989, p. C1). The scientific or decision problem is: Should the coach consider that Smith is not now a 47 percent shooter as he normally is, and therefore the coach should bench him? The statistical question is: How likely is a shooter with a 47 percent average to produce 10 of 12 misses? The key issue in the statistical question concerns the total number of shot attempts we should consider.

Would Coach Thompson take Smith out of the game after he missed *one* shot? Clearly not. Why not? Because one "expects" Smith to miss a shot half the time, and missing one shot therefore does not seem unusual.

How about after Smith misses two shots in a row? For the same reason the coach still would not bench him, because this event happens "often"—more specifically, about once in every sequence of four shots.

How about after 9 misses out of ten shots? Notice the difference between this case and 9 females among ten calves. In the case of the calves, we expected half females because the experiment is a single isolated trial. The event *considered by itself* has a small enough probability that it seems *unexpected* rather than expected. ("Unexpected" seems to be closely related to "happens seldom" or "unusual" in our psychology.) And an event that happens seldom seems to call for explanation, and also seems to promise that it will yield itself to explanation by some unusual concatenation of forces. That is, unusual events lead us to think that they have unusual causes; that is the nub of the matter. (But on the other hand, one can sometimes benefit by paying attention to unusual events, as scientists know when they investigate outliers.)

In basketball shooting, we expect 47 percent of Smith's *individual* shots to be successful, and we also expect that average for each set of shots. But we also expect *some* sets of shots to be far from that average because we observe many sets; such variation is inevitable. So when we see a single set of 9 misses in ten shots, we are not very surprised.

But how about 29 misses in 30 shots? At some point, one must start to pay attention. (And of course we would pay more attention if beforehand, and never at any other time, the player said, "I can't see the basket today. My eyes are dim.")

So, how should one proceed? Perhaps proceed the same way as with a coin that keeps coming down heads a very large proportion of the throws, over a long series of tosses: At *some* point you examine it to see if it has two heads. But if your investigation is negative, in the absence of an indication *other than the behavior in question*, you continue to believe that there is no explanation and you assume that the event is "chance" and *should not be acted upon*. In the same way, a coach might *ask* a player if there is an explanation for the many misses. But if the player answers "no," the coach should not bench him. (There are difficulties here with truth-telling, of course, but let that go for now.)

The key point for the basketball case and other repetitive situations is not to judge that there is an unusual explanation from the behavior of a *single sample alone*, just as with a short sequence of stock-price changes.

We all need to learn that "irregular" (a good word here) sequences are less unusual than they seem to the naked intuition. A streak of 10 out of 12 misses for a 47 percent shooter occurs about 3 percent of the time. That is, about every 33 shots Smith takes, he will begin a sequence of 12 shots that will end with 3 or fewer baskets—perhaps once in every couple of games. This does not seem "very" unusual, perhaps. And if the coach *treats* each such case as unusual, he will be losing some of the services of a better player than he replaces him with.

In brief, how hard one should search for an explanation should depend on the probability of the event. But one should (almost) assume the absence of an explanation unless one actually finds it.

Bayesian analysis could be brought to bear upon the matter, bringing in your prior probabilities based on the knowledge of research that has shown that there is no such thing as a "hot hand" in basketball (see Chapter 9), together with some sort of cost-benefit error-loss calculation comparing Smith and the next best available player.

The concept of statistical significance

"Significance level" is a common term in probability statistics. It corresponds roughly to the probability that the assumed benchmark universe could give rise to a sample as extreme as the observed sample by chance. The results of Example 16-1 would be phrased as follows: The hypothesis that the radiation treatment affects the sex of the fruit fly offspring is accepted as true at the probability level of .16 (sometimes stated as the 16 percent level of significance). (A more common way of expressing this idea would be to say that the hypothesis is *not rejected* at the .16 probability level or the 16 percent level of significance. But "not rejected" and "accepted" really do mean much the same thing, despite some arguments to the contrary.) This kind of statistical work is called hypothesis testing.

The question of *which* significance level should be considered "significant" is difficult. How great must a coincidence be before you refuse to believe that it is only a coincidence? It has been conventional in social science to say that if the probability that something happens by chance is less than 5 percent, it is significant. But sometimes the stiffer standard of 1 percent is used. Actually, *any* fixed cut-off significance level is arbitrary. (And even the whole notion of saying that a hypothesis "is true" or "is not true" is sometimes not useful.) Whether a one-tailed or two-tailed test is used will influence your significance level, and this is why care must be taken in making that choice.