Statistical Experiments, Randomization, Hypotheses and Significance

Many times we need to make a simple decision between believing a claim or viewing it as unproved. One brand of food is substantially less expensive that another, yet the manufacturer of the cheaper brand claims that by taste it is indistinguishable from the more expensive one. Is this true? Someone claims that zinc lozenges help people recover from colds.[1] As a doctor, will you recommend zinc lozenges to your patients? A broker claims that he can predict the way the stock market will move with better than 80% accuracy. Will you hire him?

Before deciding whether to believe a claim you will ask for evidence. But what kind of evidence? How much? If someone in your family claims to be able to taste the difference between the two brands but you yourself cannot, how will you test the claim? What kind of evidence does a doctor need before making a recommendation that may improve peoples' health or may cause them needless expense? Will you trust the broker if he misses the direction of the market only once in the next 5 days?

In gathering and weighing evidence on which to base a judgment, the techniques of statistics come into play. We shall see how all of the material that we have looked at so far in this course comes into play when we set out to design the best way to answer questions such as those above—the concepts of variables, distributions, randomness, significance, probability models, and binomial coefficients.

Some Classics

Can we determine if a given individual is capable of telling the difference between the two brands? Note that determining if one particular person (Sam, say) can tell the difference and determining if people *in general* can tell the difference are different tasks. The problem of determining if Sam can distinguish is easier, and that's what we're thinking about.

The ideas that I am about to share with you are based on one of the greatest all-time classics in the whole literature on statistics, the second chapter of Ronald A. Fisher's book, *The Design of Experiments*, first published in 1935.[2] The chapter, titled "The Principles of Experimentation, Illustrated by a Psycho-Physical Experiment," begins:

> A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which the assertion can be tested. ...

The eminent biometrician, C. I. Bliss, began his 1969 textbook *Statistics in Biology*, by describing an experiment based on Fisher's discussion, which he had his Yale students conduct. They were to determine if they could discriminate between fresh and reconstituted milk. He describes the experiment as follows:

> Following the instructions of the manufacturer, we shall prepare a bottle of reconstituted skim milk and hold it overnight in a refrigerator next to a similar

[1] See: `http://www.med.umich.edu/pediatrics/ebm/cats/zinc.htm`, or `http://www.coldcure.com/`.

[2] `http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Fisher.html`

> bottle of fresh skim milk, in order to minimize any differences in temperature. At the time of the test, we shall letter eight 2-oz machine-made paper cups from a to h and arrange them in alphabetical order. Four cups are to be filled with fresh milk and the other four with reconstituted milk. The cups which are to receive each kind are to be decided at random. For this purpose, let us use a table of random numbers ..., selecting a starting point at random and writing under the letters a to h on our data sheet the digits 1 to 8 in the order in which each first appears in the table. Each letter is now associated with a digit. We next fill the cups corresponding to the digits 1 to 4 with fresh milk and those corresponding to the digits 5 to 8 with the reconstituted milk, keeping the cups in their original order, a to h. After checking that the cups are indistinguishable in th extent filled, in spacing, and in alignment, we shall conceal the code sheet and call the subject ... Although unaware of which cups contain each type of milk, the subject knows that there are four of each and, in consequence, will divide the cups into two groups of four. He tastes the milk in each cup in the order a to h, and separates the cups which he identifies as filled with fresh milk from those filled with reconstituted milk, tasting as often as may be necessary. In order to sharpen his taste, he may take a bite of cracker between cups. We then record his results ....

Testing Sam

The experimental design that Fisher and Bliss offered as a way of determining an individual's ability to distinguish between stimuli is not the only possible design, but in many ways it is a very intelligent, economical and really quite clever design. To appreciate why it is so good, we may consider some alternatives.

How might we test Sam to see if he can distinguish between Brand X and Brand Y. One way would be a side-by-side comparison—a cup of Brand X versus a cup of Brand Y. Sam sits at a table blindfolded, tastes a sample of each brand and tells us which he thinks is which. But in a single comparison, a correct discrimination would occur in half of all trials by pure chance. So success at this task would not be very convincing evidence of ability. Another severe disadvantage of this design is that it is impossible to treat the two stimuli the same, since one must be tasted first. This might well influence the outcome in a systematic ways that we do not know.

In an informal situation—around a dinner table, for example—we might ask Sam if he feels *sure*, and we might, in a friendly spirit, accept his estimation of his own ability. However, there are some people who are always sure of themselves, and even people who feel sure for no good reason whatsoever will still be right half the time by pure chance. If our goal is to gather evidence for or against the claim that Sam can tell the difference, then Sam's belief in his own ability does not have a clear and precise meaning. Since we have other ways of testing his ability that are clear and precise, we might as well ignore what he thinks of himself.

We could ask Sam to make several side-by-side comparisons. In 3 consecutive side-by-side comparisons, the probability of him being right 3 times by sheer luck rather than by the ability to distinguish is only 1 in 8. In 4 side-by-side comparisons the chance of a

perfect score by sheer luck is 1 in 16, and for 5 side-by-side comparisons the chances drop to 1 in 32. If Sam could call 5 comparisons correctly, we would be inclined to believe his ability to tell a difference.

There are other tasks that we could assign to Sam that are more efficient in the use of out time and equipment. Suppose we fill 4 cups with Brand X and and 4 more with Brand Y, im much the same way Bliss did for his Yale students. We label the cups, but keep the meaning of the labels carefully concealed. Sam will know that there are 4 cups of each brand, but he won't know which is which. We will be careful in presenting the cups to Sam that all are treated in nearly the same way, and that the differences in presentation which we cannot avoid are distributed randomly among the cups. Sam's task is to identify the four cups containing Brand X.

Suppose Sam correctly picks the 4 with Brand X. What does this signify? We know from our understanding of the binomial coefficients that here are $\binom{8}{4} = 70$ ways for him to pick 4 out of 8 cups if the brands are ignored. *If he has no ability at all to distinguish between the brands*, then he will be equally likely to make any one of the 70 choices. On the other hand, here is only one way to pick all four cups of Brand X, so the chances of him getting all four by sheer luck rather than by the ability is only 1 in 70. Consequently, success at this task is much better evidence of ability than success at 4, 5 or even 6 side-by-side comparisons.[3]

What should we conclude if Sam gets 3 right an 1 wrong? Sam must have mis-identified one of the X cups and he must have mis-identified one of the Y cups. Since there are 4 cups of each kind, there are $4 \times 4 = 16$ ways to have gotten 3 out of the 4 X's. This means there are 17 ways for him to have gotten 3 *or more* correct X's. Now, 17/70 is nearly 1/4. So someone with no skill at distinguishing whatsoever would be expected to get at least three right in nearly a quarter of all trials. A test that is passed by a quarter of all the incompetent people who take it is not very good.

[3] This task also requires 8 cups, just as 4 side-by-side comparisons, but it is more than 4 times more informative. You might wonder how it is that we can get so much more information without using more cups. The answer is that when we make 4 side-by-side comparisons, we reveal a lot to the subject that we do not reveal when we give him the sorting task. Imagine two true/false quizzes, each with 8 questions. On one quiz, you are told that there are 4 trues; on the other, the questions come in 4 pairs and you are told that in each pair one is true and one is false. Which is the easier quiz?