# Introduction to Statistical Inference

## I. Knowledge and Uncertainty

Imagine the following game. Unseen by you, the Dealer places 19 beads of one color in 19 boxes and one bead of a different color in a 20th box. The boxes, which all look identical from the outside, are closed and placed in a random arrangement on the table. You get to open one box and observe the color of the bead it contains. At this point, if you wish, you can make a wager about the dominant color. To do so, you put $20 on the table and announce your guess. The Dealer opens all the remaining boxes to reveal their contents. If you have guessed correctly, you take back your $20 and the Dealer pays you a dollar as a prize. If you are wrong, the Dealer takes your $20.

Problems

1. Will you play this game?
2. Would you play if instead of risking $20, you had to put a different amount down? For what amounts would you play?
3. This is a philosophical question. What do you *know* about the contents of the boxes after observing the contents of one box? Clearly, you know the rules of the game and you know the color of the bead in the box you opened. Do you know more than this?

Solutions

1. Will you play this game? *You should always bet that the dominant color is the color you've seen, since any other way of guessing will have a smaller chance of success. In the long run, in 19 out of 20 plays you will be right, resulting in a gain of $19. But in one out of 20 plays, you'll lose $20. So, on average in each 20 plays, you'll lose a dollar. If you are smart, you will not play.*
2. Would you play if instead of risking $20, you had to put a different amount down? For what amounts would you play? *By the same reasoning as in the previous answer, if you put down $19 to play, then you will break even in the long run. If the amount you need to put down is any less than $19, the game is in you favor.*
3. What do you *know* about the contents of the boxes after observing the contents of one box? *What you know is very similar to what you know about random experiment (with a chance of success of 0.95) before the experiment in performed. You know enough to be able to make bets on the outcome in such a manner that the advantage will be yours in the long run.*

This simple example illustrates, in the simplest form I can think of, the pattern of reasoning that occurs in statistical inference. There are certain features of this example that we will see in more complex and sophisticated forms in the justification for many standard techniques of statistical inference. First, there is an objective situation about which I have some specific—but incomplete—knowledge, and there is a feature of the situation that I do not know but that I wish to know. Second, I obtain some additional information about the unknown feature through a process that involves chance or randomness. (From the outside, I can obtain no knowledge any differences among the boxes, so my choice of which one to open is random.) Because of the randomness, the additional information does not permit me to make inferences with certainty, yet nonetheless it does give me a basis for a hypothesis. Third, I am able to quantify my uncertainty in a manner that permits me to predict how often my hypothesis will be correct if I repeat the same pattern of inference over and over again.

## II. A Second Example: Estimating Population Proportions

Suppose we have a very large but homogeneous population, some members of which have a particular property and some members of which do not. We wish to determine the proportion of members with the property, but we are not able to examine all the members individually. Therefore, we plan to select a sample, examine the members of the sample and make an inference about the unknown population proportion from the observed sample proportion. This kind of procedure arises in numerous "real-life" scenarios. For example, we may wish to determine the proportion of a voting population that has a particular opinion by polling a random sample. Or, we may wish to determine the fraction of manufactured items coming off an assembly line that are defective by examining a sample of items taken at random times.

Now, there are a number of assumptions that must be satisfied before we can go any further. We shall assume that the population is so big that taking a sample does not affect the probability that the remaining items will have—or not have—the property. We shall also assume that we can draw a sample that is truly random: every member of the population should have an equal chance of being in the sample. We can state our assumptions with more precision as follows. We assume that:

1) we can select members of the population at random,
2) any randomly selected member has the property with and the same probability,
3) we have a sample consisting of $n$ randomly selected members,
4) the presence or absence of the property in each of these $n$ members of the sample is independent of how the property occurs in the rest of the sample.

The probability referred to in 2) is unknown to us. For convenience, let us call it $p$. We do not know the numerical value of $p$. It is somewhat analogous to the unknown "dominant color" in the example in §I, except that while the unknown color was a categorical term, $p$ is quantitative. We wish to infer something about $p$ from the sample, which gives us some information about $p$ but also contains some randomness. Our knowledge of the sample is analogous to the knowledge we obtained when we opened on box and viewed a single bead; it is informative, but it does not permit an absolute conclusion.

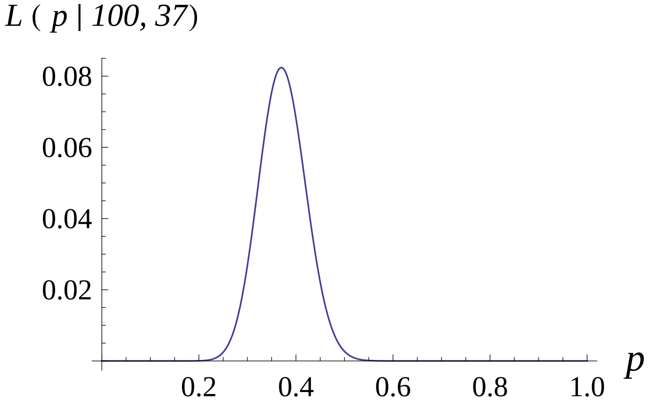
So now let us suppose that we have taken a sample and determined the number of members of the sample that have the property of interest. Let us call the proportion of the sample that has the property $\overline{x}$. This number is known to us, since we know our sample. We want to infer something about $p$ from $\overline{x}$. How can we do this?

Because of our assumptions, we know with certainty that if the sampling were repeated over and over again, we would see a pattern that was dependent on $p$. To be able to talk about this without ambiguity, let us introduce a new symbol—the random variable $\overline{X}$—to describe the sample proportion that we might see in one of a vast imaginary collection consisting of numerous samples. Because each sample consists of $n$ items, each having the property with probability $p$, $\overline{X}$ has a binomial $(n, p)$ distribution:

$$P(\overline{X} = k/n \mid p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

In other words, the probability that a random sample of size $n$ contains $k$ items with the property is given as a function of $p$ by the right hand side of the equation above. We actually have a family of random variables, each determined by a specification of $n$ and $p$. When we wish to indicate which variable we are talking about, we write $\overline{X}_{n,p}$.

In the equation, $n$, $k$ and $p$ are all quantities to which we might assign values. Typically, we would decide on the sample size $n$ at the outset. The number of items with the property in sample that we have actually taken gives us a value for $k$. If we fix $n$ and $k$ at the values in the actual sample, then the equation above tells us the likelihood of obtaining a sample with the same composition as ours, as a function of $p$. We call this $L\left(p \mid n, k\right)$. (See page 290, Definition 6.3.1.) For some values of $p$, our sample is more likely, and for other values it is less. The following figure shows the likelihood of a sample of size $n = 100$ having $k = 37$ members with the property:



Not surprisingly, the most likely value of $p$ is 0.37. Yet we see that even when $p = 0.37$, the probability of getting a sample with $k = 37$ is rather small—about 0.082. The *most likely* value of $p$ is *not bloody likely*!

The probability mass function of $\overline{X}$ is concentrated around $p$, but not *at* $p$. For every possible value of $p$, we can determine a range of values of $k$ that are more probable than others. If we set $n = 100$ and proceed to make the calculations (I will spare you the details), then we find that no matter what $p$ is, there is at least a 95% probability that a sample proportion will be within 0.10 of $p$:

$$\text{For all } p \in [0,1],\ P(|p - \overline{X}_{100,p}| \leq 0.10) > 0.95$$

This means that is we bet that $p \in [\overline{x} - 0.1, \overline{x} + 0.1]$, we will win least 19 out of 20 times. This is independent of what $p$ happens to be.

To tie this back to the first example, suppose we play the following game. The Dealer creates populations of beads with with proportion $p$ red. He chooses $p$ as he pleases, possibly always the same value, possibly different, possibly with greater frequency in one range, possibly not. He may choose any pattern or rule that he wishes. After he creates a population, he allows us to draw a random sample of 100 beads. We count the number of reds in it and then we have the opportunity to bet about the value of $p$. What shall we do? If we abide by the following rule, we will come out ahead in the long run: If we count $k$ beads in our sample of 100, then we bet \$19 against the Dealer's dollar that $p \in [(k - 10/100, (k + 10)/100]$. We will win on average, at least 19 times out of 20 in the long run.

**End Note**

As a matter of fact, the samples are most variable when $p = .5$. So, if the Dealer wanted to place someone following the rule described above at a maximum disadvantage, he would always set $p = .5$. Nonetheless, even in this case if you follow the rule, you will lose in less than 4% of the games, since

$$\sum_{k=40}^{60} \binom{100}{k} 2^{-100} > 0.964.$$

James J. Madden, LSU
August 26, 2010