Motivation for Math 325 Probability Theory: Statistical Inference

by Professor T. Fiore

We are interested in probability theory because it forms the mathematical basis for statistical inference. Of course, there are many other applications of probability theory, but we highlight statistical inference now because Math 425 on the mathematics of statistical inference is Part II of this course. We are working towards statistical inference this semester, although we will never mention it.

Statistical Inference has two goals:

(1) To infer something about a "population" when only given information about a "sample" from that population,

(2) quantify how much confidence we have in that inference about the population.

Example of Situation for Statistical Inference: 2018 Ballot Initiative to Stop Transporting Oil in the Pipeline under the Straits of Mackinac

In November 2018, Michigan voters might decide to halt the ongoing transportation of oil under the Straits of Mackinac. Ten months before the vote, an oil company would like to determine what proportion of Michigan voters would vote to stop the oil. This proportion is of vital interest to the oil company, because stopping the oil would affect their business. If the proportion of voters in favor of stopping the oil is high ten months before the vote, then the oil company could invest in an advertising campaign to highlight the benefits of keeping the pipeline in use, and possibly influence the vote in its favor.

Of course it is impossible for the oil company to ask every single voter how he/she would vote, so instead the oil company hires a polling company to call some voters and ask.

Population=all the voters in Michigan

Sample=only the voters in Michigan called by the polling company

The polling company wants to keep the sample size as small as possible (i.e. as cheap as possible) while still making an inference about the population with a certain confidence level.

Thus, this is a real-life situation where information about the population is sought, but is impossible to obtain. Instead, information is obtained only about a sample, and the polling company makes an inference about the whole population.

Example of Situation for Statistical Inference: Determining the Average Time of Recurrence of Kidney Cancer

A medical researcher would like to know the average time until the recurrence of kidney cancer for patients who have had a tumor removed. She does a retrospective study in which she obtains the medical records of 3 cancer treatment centers for the past 45 years, for patients with kidney cancer who have had a tumor removed. Then she computes the average recurrence time for the patients in those records.

Population=all human beings with kidney cancer who have had a tumor removed, in the past 45 years, the present, and the near future.

Sample=only the kidney cancer patients who have had a tumor removed at the 3 centers in the past 45 years

Of course it is impossible to know the average recurrence time for the population of interest, since it includes people in the future. The researcher would like to make an inference about the average time of recurrence of kidney cancer for the population, based on information only from the sample (the patients in the retrospective study).

Example of Using Probability to Make an Inference: Determining when a Coin is Biased

Suppose you flip a coin 100 times and each time it is heads. Would you think the coin is an ordinary fair coin?

No, of course not. You would not think it is an ordinary fair coin because you know the probability of 100 consecutive heads is very low, if the coin were fair.

The observed data has very low probability, assuming the coin is fair. Thus, the assumption the coin is fair is very probably false. I would actually believe the coin has heads on both sides!

Example of Using Probability to Make an Inference: Determining if an Advertising Campaign to Increase Seat Belt Usage Succeeded

Suppose the proportion of drivers who use a seatbelt in a certain state is known to be .22. An advertising campaign to increase seatbelt usage is done. Two months after the campaign, 590 random drivers in the state are observed, and 158 of them were wearing seat belts, the other 432 were not. Did the campaign increase seat belt usage?

Let's find the probability of an outcome *as extreme or more extreme than the actually observed outcome, assuming the proportion of seatbelt users remains at .22.* Let Y be the random variable the counts the number of people who are wearing a seatbelt when we do 590 consecutive random checks, Y is a binomial random variable with success probability .22 and 590 trials. Then the probability of an outcome *as extreme or more extreme than the actually observed outcome, assuming the proportion of seatbelt users remains at .22,* is

$$P(Y \geq 158) = 1 - P(Y \leq 157) = 1 - \texttt{pbinom(157,size=590,prob=.22)}$$
$$= 0.003487\ldots$$

This probability is very small, so we infer that the advertising campaign increased seat belt usage.

Compare the thought process here with the coin example!

Example of Using Probability **NOT** to Make an Inference: Effectiveness of a New Drug

This example is by Moore, McCabe, and Craig in their textbook Introduction to the Practice of Statistics, 6th Edition, page 355.

"Researchers want to know if a new drug is more effective than a placebo. Twenty patients receive the new drug, and 20 receive a placebo. Twelve (60%) of those taking the drug show improvement versus only 8 (40%) of the placebo patients. Our unaided judgment would suggest that the new drug is better. However, probability calculations tell us that a difference this large or larger between the results in the two groups would occur about one time in five simply because of chance variation. Since this probability is not very small, it is better to conclude that the observed difference is due to chance rather than a real difference between the two treatments."