An Overview of Statistics

What *is* Statistics?

Statistics, quite simply, is about learning from sample data.[1] You face a group of individuals – perhaps people, but maybe cans of tomatoes, or automobiles, or fish in a lake, or even something as nebulous as calendar weeks. This group is the **population** of interest to you. There is something you would like to know about this population: How likely are the people to try a new product you are thinking of bringing to the market? Are the cans properly sealed? What determines the cost of keeping the cars in working condition? How many fish are there? What will demand for your product be in the weeks to come? The answer to your question will guide you in making a decision.

If you could simply collect data from all the members of your population, you would know what you need to know. However, there can be many reasons why this might not be possible. It might be too expensive: If the potential purchasers of your product are all the adult consumers in the United States, the sheer size of the population makes contacting every individual prohibitively costly. It may be that collecting data does direct damage: If you open all the cans of tomatoes to test the contents, you have nothing left to sell. More subtly, the population is often somewhat ill-defined. If you manage a fleet of automobiles, you might consider the population of interest to be cars actually in your fleet in recent months, together with cars potentially in your fleet in the near future. In this case, some members of the population are not directly accessible to you.

For any of these reasons, you might find yourself unable to examine all members of the population directly. So, you content yourself with collecting data from a **sample** of individuals drawn from the population. Your hope is that the sample is representative of the population as a whole, and therefore anything learned from the sample will give you information concerning the entire population, and will consequently help you make your decisions.

How Does Statistics Work?

All statistical studies are carried out by following some **statistical procedure**, and every statistical procedure has three elements: You must specify *how* the sample data will be collected and *how much* data will be collected, and *what* you'll do with the data once it's in hand.

As a simple example, consider the following *estimation procedure*:

- An individual will be selected at random from the population, with every member of the population having an equal chance to be chosen. Relevant information will be obtained from the selected individual, who then will be returned to the population. (This method of selecting individuals is known as *simple random sampling with replacement*.)

- The process described above will be repeated 20 times.
- Assume, for purposes of this example, that the individuals are people, and that we obtain from each sampled individual his or her gross income over the previous twelve months. We will then average the twenty observations, and use this average (the *sample mean*) as an estimate of the average across all members of the population (the *population mean*).

If you were facing a decision problem in which the "best" decision depended on the population mean income, you might now use your estimate to guide your decision.

What is the Principle Focus of Statistics?

In the example above, the estimate we obtain might – if we are incredibly lucky – be precisely equal to the population mean. However, it will probably be the case that the sample is not perfectly representative of the population, and our sample mean is somewhat different from the true population mean. The possibility that the sample might fail to be perfectly representative is called **exposure to sampling error**. How far off is our estimate from the truth? Using only the data at hand, we can't say. (If we could, we'd simply correct the estimate!) Instead, we focus our attention on the procedure used to make the estimate, and we determine how exposed to sampling error we were in using that procedure.

The Language of Estimation

When we make an estimate, we summarize our exposure to sampling error by using a standard "language" to report our result. We say:

"I conducted a study to estimate {something} about {some population}. My estimate is {some value}. The way I went about making this estimate, I had {a large chance} of ending up with an estimate within {some small amount} of the truth."

For example, "I conducted a study to estimate the mean gross income over the past year of current subscribers to our monthly magazine. My estimate is $65,230. The way I went about making this estimate, I had a 95% chance of ending up with an estimate within $1,500 of the truth."

Notice how parsimonious this language is. We don't bore the listener with unnecessary detail concerning the data-collection procedure. Instead, we cut directly to the important issue: How exposed to sampling error were we when we carried out the procedure? How much can our procedure (and, perforce, the estimate we derived using it) be trusted?

The Language of Hypothesis Testing

Sometimes we wish to examine statistical evidence, and determine whether it supports or contradicts a claim that has been made (or that we might wish to make) concerning the population. This is done in a somewhat asymmetric fashion, analogous to the approach taken in the British system of criminal justice (adopted throughout most of the modern world): We take a statement, presume it to be "innocent," i.e., true, and ask how strongly the evidence contradicts our initial assumption.

Typically, we only do this if *some* evidence weighs against the statement, and our statistical analysis determines how strongly the evidence contradicts the original statement. This is done by calculating the probability that the procedure we carried out would – in a world where the statement really *is* true – provide such contradictory evidence purely due to sampling error.

This probability, called the *significance level* of the sample data with respect to the statement, is then interpreted. If it is relatively large, then we conclude that the evidence against the statement is weak, since we must acknowledge that, in a presumed world in which the statement is true, our studies would frequently still provide such evidence purely due to our exposure to sampling error. However, if this probability is small, we conclude that the evidence at hand is quite different from that we would expect to see if the statement were true, i.e., we conclude that the evidence strongly argues against the statement's truth, and we lean towards finding the statement "guilty."

Just as in a criminal trial, we never conclude that the statement is "innocent" – at most, we find it "not guilty." In other words, our analysis leaves us in one of two camps: We have strong evidence that the original statement is false, or we do not have such evidence. Therefore, if we wish to make an affirmative case for a claim, we are forced to take the opposite of that claim as the statement we put on trial. Only in this way might we conclude, at the end, that the data – if strong evidence against the claim on trial – serves to support the original claim.

Studying Relationships

With the two languages of statistics in place, one is ready to use the most important (to managers) of statistical "tools," *regression analysis*. This does what its name implies: It "goes back" from something of primary interest – for example, the market value of a piece of residential real estate – to a set of determining factors – such as lot size, interior living space, number of bedrooms, and (of course) location. It untangles the effects of the various factors, providing a clear view of the role of each in the overall relationship.

One reason for performing a regression analysis is to be able to make predictions in individual cases. A county tax assessor, or a real estate agency, may wish to appraise the market value of a home that has not changed possession for many years.

Another reason is to estimate the impact of some particular factor in the overall relationship. For example, one might wonder what the addition of an exercise room, or a living-room fireplace, typically adds to the resale value of one's property.[2]

In either of these applications, the language of estimation comes into play, being used to say how much a prediction, or an estimate of the impact of some factor, can be trusted. As well, the language of hypothesis testing is used when investigating whether there is evidence that some factor truly *has* an effect – for example, that after differences in experience and work assignments are taken into account, there is still evidence that a firm's compensation scheme rewards one demographic group differently from another.

[1] Actually, this is the focus of what is sometimes called "inferential statistics." In contrast, "descriptive statistics" refers to useful ways of summarizing data at hand without direct reference to an underlying population.

[2] Regression analysis cannot, by itself, demonstrate that an added feature *causes* an increase in resale value. However, it can estimate the typical difference in resale value of comparable homes that differ only in the presence or absence of the feature. We could then use our judgment and experience to assert that the resale price difference is due to the feature.