Chapter 1: Introduction to Statistics and Data Analysis

Department of Engineering Sciences
Izmir Katip Celebi University

Week 1
2014-2015 Spring

Overview: Statistical Inference, Samples, Populations, and the Role of Probability

Use of Scientific Data

The use of statistical methods in manufacturing, development of food products, computer software, energy sources, and many other areas involves the gathering of information or **scientific data**. For over a thousand years, data have been collected, summarized, reported, and stored for perusal. However, there is a profound distinction between collection of scientific information and **inferential statistics**.

The offspring of inferential statistics has been a large toolbox of statistical methods employed by statistical practitioners. These statistical methods are designed to contribute to the process of making scientific judgments in the face of **uncertainty** and **variation**. The product density of a particular material from a manufacturing process will not always be the same. Indeed, if the process involved is a batch process rather than continuous, there will be not only variation in material density among the batches that come off the line (batch-to-batch variation), but also within-batch variation.

Statistical methods are used to analyze data from a process in order to gain more sense of where in the process changes may be made to improve the **quality** of the process. In this process, quality may well be defined in relation to closeness to a target density value in harmony with what portion of the time this closeness criterion is met.

An engineer may be concerned with a specific instrument that is used to measure sulfur monoxide in the air during pollution studies. If the engineer has doubts about the effectiveness of the instrument, there are two **sources of variation** that must be dealt with.

The first is the variation in sulfur monoxide values that are found at the same locale on the same day. The second is the variation between values observed and the true amount of sulfur monoxide that is in the air at the time. If either of these two sources of variation is exceedingly large (according to some standard set by the engineer), the instrument may need to be replaced.

Variability in Scientific Data

In the problem discussed above the statistical methods used involve dealing with variability, and the variability to be studied is that encountered in scientific data. If the observed product density in the process were always the same and were always on target, there would be no need for statistical methods. If the device for measuring sulfur monoxide always gives the same value and the value is accurate (i.e., it is correct), no statistical analysis is needed.

Statistics researchers have produced an enormous number of analytical methods that allow for analysis of data from systems like the one described above. This reflects the true nature of the science that we call inferential statistics, namely, using techniques that allow us to go beyond merely reporting data to drawing conclusions (or inferences) about the scientific system.

Statisticians make use of fundamental laws of probability and statistical inference to draw conclusions about scientific systems. Information is gathered in the form of **samples**, or collections of **observations**.

Samples are collected from **populations**, which are collections of all individuals or individual items of a particular type. At times a population signifies a scientific system. For example, a manufacturer of computer boards may wish to eliminate defects. A sampling process may involve collecting information on 50 computer boards sampled randomly from the process. Here, the population is all computer boards manufactured by the firm over a specific period of time. If an improvement is made in the computer board process and a second sample of boards is collected, any conclusions drawn regarding the effectiveness of the change in process should extend to the entire population of computer boards produced under the improved process.

Often, it is very important to collect scientific data in a systematic way, with planning being high on the agenda. At times the planning is, by necessity, quite limited. We often focus only on certain properties or characteristics of the items or objects in the population. Each characteristic has particular engineering importance to the customer, the scientist or engineer who seeks to learn about the population.

For example, in the illustration above the quality of the process had to do with the product density of the output of a process. An engineer may need to study the effect of process conditions, temperature, humidity, amount of a particular ingredient, and so on. He or she can systematically move these **factors** to whatever levels are suggested according to whatever prescription or **experimental design** is desired. However, a forest scientist who is interested in a study of factors that influence wood density in a certain kind of tree cannot necessarily design an experiment. This case may require an **observational study** in which data are collected in the field but factor levels can not be preselected. Both of these types of studies lend themselves to methods of statistical inference.

In the former, the quality of the inferences will depend on proper planning of the experiment. In the latter, the scientist is at the mercy of what can be gathered.

The importance of statistical thinking by managers and the use of statistical inference by scientific personnel is widely acknowledged. Research scientists gain much from scientific data. Data provide understanding of scientific phenomena. Product and process engineers learn a great deal in their off-line efforts to improve the process. They also gain valuable insight by gathering production data (online monitoring) on a regular basis. This allows them to determine necessary modifications in order to keep the process at a desired level of quality.

There are times when a scientific practitioner wishes only to gain some sort of summary of a set of data represented in the sample. In other words, inferential statistics is not required. Rather, a set of single-number statistics or **descriptive statistics** is helpful. These numbers give a sense of center of the location of the data, variability in the data, and the general nature of the distribution of observations in the sample.

Sometimes, descriptive statistics are accompanied by graphics. Modern statistical software packages allow for computation of **means**, **medians**, **standard deviations**, and other single number statistics as well as production of graphs that show a footprint of the nature of the sample. Definitions and illustrations of the single-number statistics and graphs, including histograms, stem-and-leaf plots, scatter plots, dot plots, and box plots, will be given in sections that follow.

The Role of Probability

In this course, we will deal with fundamental notions of probability between weeks 3 and 10. A thorough grounding in these concepts allows us to have a better understanding of statistical inference Without some formalism of probability theory, we cannot appreciate the true interpretation from data analysis through modern statistical methods. It is quite natural to study probability prior to studying statistical inference.

Elements of probability allow us to quantify the strength or confidence in our conclusions. In this sense, concepts in probability form a major component that supplements statistical methods and helps us measure the strength of the statistical inference. The discipline of probability, then, provides the transition between descriptive statistics and inferential methods. Elements of probability allow the conclusion to be put into the language that the science or engineering practitioners require.

Example

Suppose that an engineer encounters data from a manufacturing process in which 100 items are sampled and 10 are found to be defective. It is expected and anticipated that occasionally there will be defective items. Obviously these 100 items represent the sample. However, it has been determined that in the long run, the company can only tolerate 5% defective in the process. Now, the elements of probability allow the engineer to determine how conclusive the sample information is regarding the nature of the process. In this case, the population conceptually represents all possible items from the process. Suppose we learn that if the process is acceptable, that is, if it does produce items no more than 5% of which are defective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process. This small probability suggests that the process does, indeed, have a long-run rate of defective items that exceeds 5%. In other words, under the condition of an acceptable process, the sample information obtained would rarely occur. However, it did occur! Clearly, though, it would occur with a much higher probability if the process defective rate exceeded 5% by a significant amount.

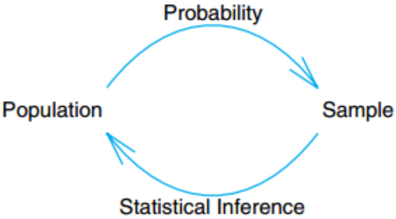
From this example it becomes clear that the elements of probability aid in the translation of sample information into something conclusive or inconclusive about the scientific system. In fact, what was learned likely is alarming information to the engineer or manager. Statistical methods, which we will actually detail in weeks 13 and 14, produced a *P*-value of 0.0282. The result suggests that the process **very likely is not acceptable**.

How Do Probability and Statistical Inference Work Together?

It is important for us to understand the clear distinction between the discipline of probability, a science in its own right, and the discipline of inferential statistics. As we have already indicated, the use or application of concepts in probability allows real-life interpretation of the results of statistical inference. As a result, it can be said that statistical inference makes use of concepts in probability.

One can see from the previous example that the sample information is made available to the analyst and, with the aid of statistical methods and elements of probability, conclusions are drawn about some feature of the population (the process does not appear to be acceptable in Example 1). Thus for a statistical problem, **the sample along with inferential statistics allows us to draw conclusions about the population, with inferential statistics making clear use of elements of probability.** This reasoning is inductive in nature.

In the next 9 weeks, unlike what we do in the previous example, we will not focus on solving statistical problems. Many examples will be given in which no sample is involved. There will be a population clearly described with all features of the population known. Then questions of importance will focus on the nature of data that might hypothetically be drawn from the population. Thus, one can say that **elements in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population, based on known features of the population.**



Fundamental relationship between probability and inferential statistics

So, which one is more important? The field of probability or the field of statistics?

They are both very important and clearly are complementary. The only certainty concerning the pedagogy of the two disciplines lies in the fact that if statistics is to be taught at more than merely a cookbook level, then the discipline of probability must be taught first. This rule comes from the fact that nothing can be learned about a population from a sample until the analyst learns the rudiments of uncertainty in that sample.

For example, consider the previous example. The question centers around whether or not the population, defined by the process, is no more than 5% defective. In other words, the assumption is that on the average 5 out of 100 items are defective. Now, the sample contains 100 items and 10 are defective. Does this support the assumption or refute it? On the surface it would appear to be a refutation of the assumption because 10 out of 100 seem to be "a bit much". But without elements of probability, how do we know?

Only through the study of material in future chapters will we learn the conditions under which the process is acceptable (5% defective). The probability of obtaining 10 or more defective items in a sample of 100 is 0.0282.

Sampling Procedures; Collection of Data

In the previous section, we discussed very briefly the notion of sampling and the sampling process. While sampling appears to be a simple concept, the complexity of the questions that must be answered about the population may be very complex at times. We will discuss the notion of sampling in a technical way in week 11, so here we just give some common-sense notions of sampling. This is a natural transition to a discussion of the concept of variability.

Simple Random Sampling

The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked. Let us assume that only a single population exists in the problem. **Simple random sampling** implies that any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size. The term **sample size** simply means the number of elements in the sample. Obviously, a table of random numbers can be utilized in sample selection in many instances. The virtue of simple random sampling is that it aids in the elimination of the problem of having the sample reflect a different population than the one about which inferences need to be made.

For example, a sample is to be chosen to answer certain questions regarding political preferences in a certain state in the United States. The sample involves the choice of, say, 1000 families, and a survey is to be conducted. Now, suppose it turns out that random sampling is not used. Rather, all or nearly all of the 1000 families chosen live in an urban setting. It is believed that political preferences in rural areas differ from those in urban areas. In other words, the sample drawn actually restricted the population and thus the inferences need to be restricted to the "limited population", and in this case restriction may be undesirable. If, indeed, the inferences need to be made about the state as a whole, the sample of 1000 individuals here is often referred to as a **biased sample**.

Hence, **simple random sampling** is a procedure used to select a sample of *n* objects from a population in such a way that

- each member of the population is chosen strictly by chance,
- the selection of one member does not influence the selection of any other member,
- each member of the population is equally likely to be chosen, and
- every possible sample of a given size, *n*, has the same chance of selection.

The resulting sample is called a **(simple) random sample**.

As we hinted earlier, simple random sampling is not always appropriate. Which alternative approach is used depends on the complexity of the problem. Often, for example, the sampling units are not homogeneous and naturally divide themselves into nonoverlapping groups that are homogeneous. These groups are called strata, and a procedure called stratified random sampling involves random selection of a sample within each stratum. The purpose is to be sure that each of the strata is neither over- nor underrepresented.

For example, suppose a sample survey is conducted in order to gather preliminary opinions regarding a bond referendum that is being considered in a certain city. The city is subdivided into several ethnic groups which represent natural strata. In order not to disregard or overrepresent any group, separate random samples of families could be chosen from each group.