İST292 STATISTICS LESSON 1

1. INTRODUCTION TO STATISTICS AND SOME IMPORTANT DEFINITIONS

Statistics is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, presenting and interpreting numerical information.

Statistical methods are particularly useful for studying, analyzing and learning about ***populations*** of ***experimental units***.

Experimental (or Observational) Unit: An ***experimental (or observational) unit*** is an object (e.g., person, thing, transaction (işlem), or event) about which we collect data.

Population: A ***population*** is a set of units (usually people, objects, transactions, o events) that we are interested in studying. It is the data which we have not completely examined but to which our conclusions refer. The population size is usually indicated by a capital ***N***.

Examples:

- all unemployed workers in Turkey
- all registered voters in Ankara
- all the cars produced last year by a particular assembly line (montaj hattı/montaj üretim hattı)
-the set of all accidents occurring on a particular stretch (alan/kısım/saha/bölüm) of interstate highway (eyaleetlerarası otoban) during a holiday period.
- all students in Hacettepe University

In studying a population, we focus on one or more characteristics or properties of the units in the population. We call such characteristics ***variables***. For Example, we may be interested in the variables *age, gender and number of years of education* of the people currently unemployed in Turkey. Another example, in a research on the students in a university, we are interested in how many books (novel, poem, etc.) they are reading, how many hours they are spending for their education, sportive activities or social activities. Each reply for the questions in a public survey (questionnaire) is corresponded to a variable.

Variable: A ***variable*** is a characteristic or property of an individual experimental (or observational) unit in the population. A variable is a characteristic (e.g., temperature, age, race, growth, education level, etc.) that we would like to measure on individuals. The actual measurements recorded on individuals in the sample are called **data.**

In studying a particular variable, it is helpful to be able to obtain a numerical representation for it. Often, however, numerical representations are not readily avaliable, so meausrement plays an important supporting role in statistical studies.

Two Types: ***Quantitative variables*** have measurements (data) on a numerical scale. ***Categorical (Qualitative) variables*** have measurements (data) where the values simply indicate group membership.

Example: Which of the following variables are quantitative in nature? Which are categorical?

- age - ***Quantitative variables***
- advertising medium (reklam aracı) (radio/TV/internet)- ***Categorical variables***

- number of cigarettes smoked per day- ***Quantitative variables***
- smoking status (yes/no)- ***Categorical (Qualitative) variables***

Measurement: ***Measurement*** is the process we use to assign numbers to variables of individual population units. We might, for example, measure the performance of the president by asking a registered voter to rate it on a scale from 1 to 10. Or we might measure the age of the Turkey workforce simply by asking each worker, "How old are you?" In other cases, measurement involves the use of instruments such as stopwaches (kronometre), scales (ölçek) and calipers (kalınlık/çap ölçer). It can be used a scale from 0 to 10 or 0 to 100 for evaluating exam papers in a course.

Census (sayım, populasyon sayımı): If the population you wish to study is small, it is possible to measure a variable for every unit in the population. When we measure a variable for every unit of a population, it is called a ***census*** of the population. In Turkey, national cencus is made for every 5 years.

Typically, however, the populations of interest in most applications are much larger, involving perhaps many thousands, or even an infinite number, of units. An example of large population is all potential buyers of a new IPhone. For such large populations, condicting a census would be prohibitively (yanına yaklaşılmaz) time consuming or costly. A reasonable alternative would be to select and study a subset (or proportion) of the units in the population. The number of the units in the population is shown as capital N

Sample: A ***sample*** is a subset of the units of a population. That proportion of the population that is available, or to be made available, for analysis. A good sample is representative of the population. We will learn about probability samples and how they provide assurance that a sample is indeed representative. The sample size is shown as lower case ***n***.

Often data are selected from some larger set of data whose characteristics we wish to estimate. This selection process is called ***sampling***.

Example:
If your company manufactures one million laptops, they might take a sample of say, 500, of them to test quality. The population N = 1000000 and the sample n= 500.

Parameter: A characteristic of a population. The population mean, μ and the population standard deviation, σ, are two examples of population parameters. If you want to determine the population parameters, you have to take a census of the entire population. Taking a census is very costly.

Statistic: A ***statistic*** is a measure that is derived from the sample data. For example, the sample mean $(\overline{X})$ and the sample standard deviation (S) are statistics. They are used to estimate the population parameters.

Statistics involves two different processes:
(1) describing sets of data and **(2)** drawing conclusions (making estimates, decisions, predictions, etc.) about the sets of data on the basis of sampling. So, the applications of statistics could be divided into two broad areas: ***descriptive statistics*** and ***statistical inference***.

Descriptive statistics utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present that information in a convenient form. Those statistics that summarize a sample of numerical data in terms of averages and other measures for the purpose of description, such as the mean and standard deviation. This includes the presentation of data in the form of graphs, charts, and tables.

Descriptive statistics, as opposed to inferential statistics, are not concerned with the theory and methodology for drawing inferences that extend beyond the particular set of data examined, in other words from the sample to the entire population. All that we care about are the summary measurements such as the average (mean). Thus, a teacher who gives a class, of say, 35 students, an exam is interested in the descriptive statistics. What was the class average, the median grade, the standard deviation, etc.? The teacher is not interested in making any inferences.

[For example, after grading an exam, a teacher may calculate the average grade to summarize the overall performance of the class. No inferences being made here.]

Statistical Inference is an estimate, prediction, or some other generalization about a population based on information contained in a sample. Why do we need statistical inference? Because we cannot get all information (data) about whole population.
The process of using sample statistics to draw conclusions about population parameters is known as ***statistical inference***. For example, using $(\overline{X})$ based on a sample of, say, n=1000) to draw conclusions about μ (population of, say, 300 million). This is a measure of performance in which the sample measurement is used to estimate the population parameter.

<u>Example 1:</u> Nielsen television ratings
The Nielsen ratings are based on a sample, not the population.
The sample consists of about 5,000 TV households
Population of more than 115,000,000 TV households
For example, if a show has a 10.0 rating, this means that 10% of the entire sample were watching that show. [Note: "Share of audience" is the percentage of those who have the TV on, i.e. of those actually watching TV.]

<u>Example 2:</u> market share of a product
Sample of supermarkets throughout the US to determine what percentage of people who buy a type of product (e.g., detergent (deterjan)) buy a specific brand (marka) (e.g., Tide (Ariel ya da Omo gibi marka)).

Both of these examples are of statistics that are used to make inferences about the population.

<u>Example for both of descriptive statistics and statistical inference</u>
For example, instead of polling (anket yapmak) all 145 million registered voters in the United States during a presidential election year, a pollster (anketör) might select and question a sample of 1500 votes. If he is interested in the variable "presidential pereference", he would record (measure) the preference of each vote sampled.

After the variables of interest for every unit in the sample (or population) are measured, the data are analyzed, either by descriptive or inferential statistical methods. The pollster, for example, may be interested only in describing the voting patterns of the sample of 1500 voters.

More likely, however, he will want to use the information in the sample to make inferences about the population of all 145 million voters. That is, we use the information conatined in the smaller sample to learn about the larger population. Thus, from the sample of 1500 voters, the pollster may estimate the percentage of all the voters who would vote for each presidential candidate if the election were held on the day the poll was conducted, or he might use the results to predict the outcome on election day. Note that pollsters do not call every adult who can vote for president. This would be very expensive. What pollsters do is call a representative sample of about 1500 people and use the sample statistics to estimate who is going to win the election.

2. TYPES OF SAMPLES

A) Nonprobability Samples – based on convenience (uygunluk, elverişlik) or judgment (yargı, hüküm). In these types of sampling methods sample of size n is selected from population of size N according to personal opinion. Hence, the problem with a nonprobability sample is that we do not know how representative our sample is of the population.
B) Probability Samples

Probability Sample: A sample collected in such a way that every element in the population has a known chance of being selected.

a) Simple Random Sample: A sample collected in such a way that every element in the population has an equal chance of being selected.

b) Systematic Random Sample: Choose the first element randomly, then every kth observation, where k = N/n

c) Stratified (Tabakalı) Random Sample: The population is sub-divided based on a characteristic and a simple random sample is conducted within each stratum (tabaka).

d) Cluster Random Sample: First take a random sample of clusters from the population of cluster. Then, a Stratified Random Sample (SRS) within each cluster. Example, election district (seçim bölgesi).

3. DATA

3.1. Types of Data

Statistics is the science of data and that data are obtained by measuring the values of one or more variables on the units in the sample (or population). All data (and hence the variables we measure) can be classified as one of two general types: ***quantitative (nicel/sayısal) data and qualitative (nitel) data***.

A) Qualitative (or Categorical) Data: Qualitative data are measurements that cannot be measured on a natural numerical scale; they can only be classified into one of a group of categories.

<u>Qualitative Data Examples:</u>
1. Sex □Male □Female (Nominal). Nanionality: German, French, British, Turkish, Arabian, etc.

2. The political party affiliation (üyeliği) (Democrat, Rebuplican, or Independent) in a sample of 50 voters. (Nominal)
3. The defective status (defective or not) of each of 100 computer chips manufactured by Intel
4. A taste tester's ranking (best, worst, etc.) of four brands of barbecue sauce for a panel of 10 testers (Ordinal)

Often, we assign arbitrary (keyfi) numerical values to qualitative data for ease of computer entry and analysis. But these assigned numerical values are simply codes: They can't be meaningfully added, subtracted, multipilied or divided. For example, we might code Democrat=1, Rebuplican=2, and Independent=3. Smilarly, a taste tester might rank the barbecue sauces from 1 (best) to 4 (worst). These are simply arbitrarily (keyfi olarak) selected numerical codes for the categories and have no utility beyond that.

B) Quantitative Data: ***Quantitative data*** are measurements that are recorded on a naturally occurring numerical scale.

Quantitative Data Examples:
1. The temperature (in degrees Celsius) at which each piece in a sample of 20 pieces of heat-resistant plastic begins to melt (Interval)
2. The current unemployment rate (measured as a percentage) in each of the 81 cities (Ratio)
3. The number of convicted (mahkum edilmiş) murderers (katil) who receive the death penalty each year over a 10-year period (Ratio)

Quantitative data result in numerical responses, and may be ***discrete*** or ***continuous***.

a) Discrete data arise from a counting process.
Example:
How many courses have you taken at this College? _________

b) **Continuous data** arise from a measuring process.
Example:
How much do you weigh? ________

One way to determine whether data is continuous, is to ask yourself whether you can add several decimal places to the answer. You may weigh 150 pounds but in actuality may weigh 150.23568924567 pounds. On the other hand, if you have 2 children, you do not have 2.3217638 children.

3.2. Levels of Data

Qualitative (Categorical) data can be subclassified as either ***nominal data*** or ***ordinal data.*** The categories of an ***ordinal data*** set can be ranked or meaningfully ordered, but the categories of a ***nominal data*** set can't be ordered.

A)Nominal Data
Classification – categories

When objects are measured on a nominal scale, all we can say is that one is different from the other

Examples: gender (sex) (cinsiyet), male, female
occupation (iş/meslek), doctor, teacher, researcher, astronuat, engineer
ethnicity (etnik yapı),
marital status (medeni hal), single, maried, divorced
etc.

B)Ordinal Data

Ranking, but the intervals between the points are not equal. We can say that one object has more or less of the characteristic than another object when we rate them on an ordinal scale. Thus, a category 5 hurricane is worse than a category 4 hurricane which is worse than a category 3 hurricane, etc.

Examples: hardness of minerals scale, income as categories, class standing (kaçıncı sınıf), rankings of football teams, military rank (general, colonel (albay), major (binbaşı), lieutenant (yüzbaşı/üsteğmen), sergeant (astsubay/çavuş), etc.), hurricane rankings (category 1, 2, …, category 5)

Quantitative data can be subclassified as either ***interval data*** or ***ratio data***. For ***ratio data***, the origin (i.e. the value 0) is a meaningful number. But the origin has no meaning with interval data. Consequently, we can add and subtract interval data, but we can't multiply and divide them.

C)Interval Data

Equal intervals, but no "true" zero.

Examples: IQ, temperature

Since there is no true zero – the complete absence (eksiklik) of the characteristic you are measuring – you can't speak about ratios.

Example:
Suppose
New York temperature = 40 degrees
Buffalo temperature = 20 degrees
Does that mean it is twice as cold in Buffalo as in NY? No.

D)Ratio Data

Equal intervals and a "true" zero.

Examples: height, weight, length

All scales, whether they measure weight in kilograms or pounds, start at 0. The 0 means nothing and is not arbitrary.

100 kilograms is double 50 kilograms
$100 is half as much as $200