Contents

3

Population

As is so often the case in statistics, some words have technical meanings that overlap with their common use but are not the same. Population is one such word. It is often difficult to decide which population should be sampled. For instance, if we wished to sample 500 listeners to an FM radio station specializing in music should the population be of listeners to that radio station in general, or of listeners to that stations classical music programme, or perhaps just the regular listeners, or any one of many other possible populations that you can construct for yourself? In practice the population is often chosen by finding one that is easy to sample from, and that may not be the population of first choice.

Sampling In medical trials (which are an important statistical application) the population may be those patients who arrive for treatment at the hospital carrying out the trial, and this may be very different from one hospital to another. If you look at any collection of official statistics (which are most important for state planning) you will be struck by the great attention that is given to defining the population that is surveyed or sampled, and to the definition of terms. For instance, it has proved difficult to get consensus on the meaning of unemployed in recent years but a statistician must be prepared to investigate the population of unemployed.

Think about this carefully. It should help you with your Sociology and Marketing and Market Research modules as well as this one:

Bias

In addition to the common-sense meaning of bias, there is also a more technical meaning for the word in statistics. This will be found both in Chapter 10 of this guide and in the work on estimators in Statistics 2. It seems natural enough to wish to avoid bias, but it is not helpful to be swayed by the value judgements inherited from the use of a word outside the limits of academic discussion.

Sampling

Explain the difference between sampling error and sampling bias. Explain briefly which is taking place in the following situations, if the population under study consists of all the pupils in a certain school:

i. Your sample is a list of all pupils at the school except those who arrived at the school in the last school year.

ii. You take a random sample of names from the school register of all pupils at the school.

Sampling Define quota sampling. In what circumstances would you use it? In what circumstances would you use stratified random sampling? Give two ways in which stratified random sampling differs from quota sampling.

Sampling

Although it seems sensible to sample a population to avoid the cost of a total enumeration (or census) of that population, it is possible to make a strong argument against the practice. One might well consider that sampling is fundamentally unfair because a sample will not accurately represent the whole population, and it allows the units selected for the sample to have more importance than those not selected. This might be thought undemocratic. Many countries continue to take a full census of their population, even though sampling might be cheaper. It is less obvious, but true, that sampling might well be more accurate, because more time can be spent verifying the information collected for a sample.

0.1 Syllabus

The concept of a random sample, the sampling distribution of the sample mean with applications to confidence intervals, hypothesis testing, and sample size determination, the sampling distribution of the sample proportion with applications to confidence intervals, hypothesis testing, and sample size determination, comparing two means, comparing two proportions, the chi-squared test of independence, Simpson's Paradox, simple linear regression, correlation, residuals.

On successful completion of this module, students should be able to:

1. Calculate probabilities based on the normal distribution.
2. Construct and use control charts based on individual measurements, subgroup means and subgroup ranges.
3. Interpret computer output from common statistical software packages for basic statistical inference procedures such as hypothesis testing and confidence intervals for: a mean, a proportion, difference between independent means, and differences between independent proportions.
4. Calculate the required sample size for tests of hypothesis and confidence intervals based on a single parameter.

5. Interpret computer output and diagnostic plots from common statistical software packages for simple linear regression and multiple linear regression.

6. Test the statistical significance of the difference between several conditional frequency distributions and outline the chi-squared formula used for the test.

0.2 Introduction to Statistics

What is Statistics

Statistics is a branch of mathematics in which groups of measurements or observations are studied. The subject is divided into two general categories ***descriptive statistics*** and ***inferential statistics***. In descriptive statistics one deals with methods used to collect, organize and analyze numerical facts. Its primary concern is to describe information gathered through observation in an understandable and usable manner.

Similarities and patterns among people, things and events in the world around us are emphasized. Inferential statistics takes data collected from relatively small groups of a population and uses inductive reasoning to make generalizations, inferences and predictions about a wider population. Throughout the study of statistics certain basic terms occur frequently. Some of the more commonly used terms are defined in the next sections.

Populations and Samples

A population is a complete set of items that is being studied. It includes all members of the set. The set may refer to people, objects or measurements that have a common characteristic. Examples of a population are all high school students, all cats, all scholastic aptitude test scores.

A relatively small group of items selected from a population is a sample. If every member of the population has an equal chance of being selected for the sample, it is called a random sample. Examples of a sample are all algebra students at Central High School, or all Siamese cats.

Types of Data

Data are numbers or measurements that are collected. Data may include numbers of individuals that make up the census of a city, ages of pupils in a certain class, temperatures in a town during a given period of time, sales made by a company, or test scores made by ninth graders on a standardized test.

Variables are characteristics or attributes that enable us to distinguish one individual from another. They take on different values when different individuals are observed. Some variables are height, weight, age and price. Variables are the opposite of constants whose values never change.

0.2.1 Types of Data

Data are the facts and figures collected, analyzed and summarized for presentation and interpretation.

- All the data collected in a study is the **data set** for the study.
- There are several types of data and identifying the type of data is vital in determining the statistical method used to describe it.
- Most statistical analysis are specific to a certain data type.

- Data can be classified as either ***qualitative*** or ***quantitative***.
- Categorical Data / Nominal Data
- Ordinal Data
- Interval Data
- Ratio Data

0.3 Samples, Statistics and Parameters

What is a Sample? A sample is a relatively small subset of people, objects, groups, or events, that is selected from the population. Instead of surveying every recent college graduate in the United States, which would cost a great deal of time and money, we could instead select a sample of recent graduates, which would then be used to generalize the findings to the larger population.

- A sample is a subset of a population.
- Since it is usually impractical to test every member of a population, a sample from the population is typically the best approach available.

Statistic

- This is a numerical characteristic of the sample; a value known when the sample is taken but that can change from sample to sample.

- In the clinical trial example, the statistic is the number of male out the 30 individuals that respond well to the drug.

Parameter

- This is a numerical characteristic of the population; it is a fixed number with an unknown value.

- In the clinical trial example, the parameter could be the total number of adult male that respond well to the drug.

- Inferential statistics generally require that sampling be **random** although some types of sampling (such as those used in voter polling) seek to make the sample as representative of the population as possible by choosing the sample to resemble the population on the most important characteristics.

The major use of statistics is to use information from a ***sample*** to infer something about a ***population***.

- A ***population*** is a collection of data whose properties are analyzed. The population is the complete collection to be studied, it contains all subjects of interest.
- A ***sample*** is a part of the population of interest, a sub-collection selected from a population.
- A ***parameter*** is a numerical measurement that describes a characteristic of a population, while a ***sample statistic*** is a numerical measurement that describes a characteristic of a sample.
- In general, we will use a statistic to infer something about a parameter.

Populations and Samples

The collection of everyone or everything that is to be analyzed in a study is called a **population**. As we have seen in the examples above, the population could be enormous in size. There could be millions or even billions of individuals in the population. But we must not think that the population has to be large. If our group being studied is fourth graders in a particular school, then the population consists only of these students. Depending on the school size, this could be less than a hundred students in our population.

To make our study less expensive in terms of time and resources, we only study a subset of the population. This subset is called a **sample**. Samples can be quite large or quite small. In theory one individual from a population constitutes a sample. Many applications of statistics require that a sample have at least 30 individuals.

Parameters and Statistics

The main objective of Statistics as a science is to estimate a population parameter by use of sample statistics.

What we are typically after in a study is the **parameter**. A parameter is a numerical value that states something about the entire population being studied. For example, we may want to know the mean wingspan of the American bald eagle. This is a parameter, because it is describing all of the population.

Parameters are difficult if not impossible to obtain exactly. On the other hand, each parameter has a corresponding **statistic** that can be measured exactly. A statistic is a numerical value that states something about a sample. To extend the example above, we could catch 100 bald eagles and then measure the wingspan of each of these. The mean wingspan of the 100 eagles that we caught is a statistic.

The value of a parameter is a fixed number. In contrast to this, since a statistic depends upon a sample, the value of a statistic can vary from sample to sample. Suppose our population parameter has a value, unknown to us, of 10. One sample of size 50 has corresponding statistic with value 9.5. Another sample of size 50 from the same population has corresponding statistic with value 11.1.

Examples of Parameters and Statistics

Below are some more example of parameters and statistics:

Suppose we study the population of dogs in Kansas City. A parameter of this population would be the mean height of all dogs in the city. A statistic would be the mean height of 50 of these dogs. We will consider a study of high school seniors in the United States. A parameter of this population is the standard deviation of grade point averages of all high school seniors. A statistic is the standard deviation of the grade point averages of a sample of 1000 high school seniors.

Mnemonic Device

There is a simple and straightforward way to remember what a parameter and statistic are measuring. All that we must do is look at the first letter of each word. A parameter measures something in a population, and a statistic measures something in a sample.

What Is the Difference Between a Parameter and a Statistic?

In several disciplines the goal is to study a large group of individuals. These groups could be as varied as a species of bird, college freshmen in the U.S. or cars driven around the world. Statistics is used in all of these studies when it is infeasible or even impossible to study each and every member of the group of interest. Rather

than measuring the wingspan of every bird of a species, asking survey questions to every college freshman, or measuring the fuel economy of every car in the world, we instead study and measure a subset of the group.

0.4 Preliminaries

0.4.1 Analyzing the Data

- Statistical data analysis divides the methods for analyzing data into two categories: exploratory methods and confirmatory methods.
- Exploratory methods are used to discover what the data seems to be saying by using simple arithmetic and easy-to-draw pictures to summarize data.
- The objectives of Exploratory Data Analysis are to:
  - Suggest hypotheses about the causes of observed phenomena,
  - Assess assumptions on which statistical inference will be based,
  - Support the selection of appropriate statistical tools and techniques,
  - Provide a basis for further data collection through surveys or experiments.

0.5 Types of Data

Data are the facts and figures collected, analyzed and summarized for presentation and interpretation.

- All the data collected in a study is the **data set** for the study.
- There are several types of data and identifying the type of data is vital in determining the statistical method used to describe it.
- Most statistical analysis are specific to a certain data type.
- Data can be classified as either ***qualitative*** or ***quantitative***.
- Categorical Data / Nominal Data
- Ordinal Data
- Interval Data
- Ratio Data

0.6 Variables

Variables

The key terms used in data collection can be dened as follows:

- A variable is the phenomenon being measured in the experiment or observational study. item A continuous variable takes any value on a range of real numbers. item A discrete variable takes only distinct values, usually often integers (analogous to counting)

0.6.1 What is a variable

In statistics, a variable has two defining characteristics:

- A variable is an attribute that describes a person, place, thing, or idea.
- The value of the variable can "vary" from one entity to another.

For example, a person's hair color is a potential variable, which could have the value of "blonde" for one person and "brunette" for another.

0.6.2 Categorical vs. Quantitative Variables

Variables can be classified as categorical (or ***qualitative***) or numerical (or ***quantitative***).

- **Categorical**. Categorical variables take on values that are names or labels. The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, terrier) would be examples of categorical variables.
- **Quantitative**. Quantitative variables are numerical. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable.

In algebraic equations, quantitative variables are represented by symbols (e.g., x, y, or z).

Qualitative data

- Qualitative data includes labels or names used to identify an attribute of each element.
- Qualitative data may be numeric (e.g. area codes), but usually it is non-numeric.
- Examples: gender, region, colour, socio-economic status.

Quantitative data

- Quantitative data require numeric values that indicate how much or how many.
- Quantitative data is always numeric.
- Examples: height, weight, age, expenditure.

0.6.3 Discrete vs. Continuous Variables

Quantitative variables can be further classified as ***discrete*** or ***continuous***. If a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise, it is called a discrete variable. Discrete data has distinct whole number values with no intermediate points.

- Discrete variables are often used as "counting" variables. For example, the number of employees in a company is discrete data.
- Continuous variables are often used as "measurement variables

Some examples will clarify the difference between discrete and continouous variables.

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
- Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and infinity. However, it could not be any number between 0 and infinity.
  We could not, for example, get 2.3 heads. Therefore, the number of heads must be a discrete variable.

0.6.4 Discrete vs. Continuous Variables

Quantitative variables can be further classified as ***discrete*** or ***continuous***. If a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise, it is called a discrete variable. Discrete data has distinct whole number values with no intermediate points.

- Discrete variables are often used as "counting" variables. For example, the number of employees in a company is discrete data.
- Continuous variables are often used as "measurement variables

Some examples will clarify the difference between discrete and continouous variables.

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
- Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and infinity. However, it could not be any number between 0 and infinity.
  We could not, for example, get 2.3 heads. Therefore, the number of heads must be a discrete variable.

- A pharmaceutical firm might be interested in conducting an experiment (i.e. a clinical trial) to learn about how a new drug affects blood pressure in adult males.
- To obtain data about the effect of the new drug, researchers select a sample of 30 individuals from a list of volunteers.

- A pharmaceutical firm might be interested in conducting an experiment (i.e. a clinical trial) to learn about how a new drug affects blood pressure in adult males.
- To obtain data about the effect of the new drug, researchers select a sample of 30 individuals from a list of volunteers.

For the clinical trial example

Population : all adult males.

Unit : any adult male.

Sample : the 30 individuals.

Sampling frame : the list of volunteers.

Variable : the blood pressure.

Bivariate Data

- Univariate statistics describes statistics related to one variables.
- Bivariate statistics describes statistics related to two variables X and Y.
- Multivariate statistics describes statistics related to multiple variables (not part of course).

Covariance Covariance is a strength of the measure of the linear relationship between two variables.

$$cov(x, Y) =$$

Variables in Regression Analysis

- The X variable is called the independent (or predictor) variable.
- The Y variable is called the dependent (or response) variable.
- Using the scatter plot we can state the strength and type (linear/non-linear) of the relationship.

Correlation and cause-effect

- Note that a strong relationship between two variables does not imply a cause-effect relationship.
- For example, there is a strong negative correlation between the sales of ice cream and the number of flu infections.
- This does not mean that ice cream protects against flu.
- This relationship results from a latent variable (a variable that has not been observed).
- Such a latent variable in this case is the weather. Low temperatures and wet weather result in a high number of flu infections and low ice cream sales.
- Hot, sunny weather leads to the opposite.

Scatter-plots Subsequent Slides

- Relatively strong positive relationship (as height increases weight on average increases), reasonably linear.
- No relationship/weak negative relationship
- Negative, very strong, non-linear relationship.
- Non-linear relationship.

0.7 What is Bivariate data?

- A dataset with two variables contains what is called bivariate data
- For example, the heights and weights of people (i.e. for the purposes of determining the extent to which taller people weigh more)

0.8 Univariate and Bivariate Data

Statistical data is often classified according to the number of variables being studied.

- **Univariate data**. When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of high school students. Since we are only working with one variable (weight), we would be working with univariate data.
- **Bivariate data**. When we conduct a study that examines the relationship between two variables, we are working with bivariate data. Suppose we conducted a study to see if there were a relationship between the height and weight of high school students. Since we are working with two variables (height and weight), we would be working with bivariate data.

0.9 Observational studies and experiments

- In industrial and agricultural applications of statistics it is possible to control the levels of the important factors that affect the results. Bias from the factors that cannot be controlled is dealt with by randomization. These investigations are **designed experiments**.
- In medical, economic and other social science applications of statistics one usually just observes a sample of the population available, without control of any of the factors that may influence the measures observed. These studies are **observational studies**.
- In industrial and agricultural applications of statistics it is possible to control the levels of the important factors that affect the results.
- Bias from the factors that cannot be controlled is dealt with by randomization.
- These investigations are designed experiments.

In medical, economic and other social science applications of statistics one usually just observes a sample of the population available, without control of any of the factors that may influence the measures observed. These studies are observational studies.

0.10 Observational studies and experiments

- In industrial and agricultural applications of statistics it is possible to control the levels of the important factors that affect the results. Bias from the factors that cannot be controlled is dealt with by randomization. These investigations are **designed experiments**.

- In medical, economic and other social science applications of statistics one usually just observes a sample of the population available, without control of any of the factors that may influence the measures observed. These studies are **observational studies**.
- In industrial and agricultural applications of statistics it is possible to control the levels of the important factors that affect the results.
- Bias from the factors that cannot be controlled is dealt with by randomization.
- These investigations are designed experiments.

In medical, economic and other social science applications of statistics one usually just observes a sample of the population available, without control of any of the factors that may influence the measures observed. These studies are observational studies.

0.11 Descriptive Statistics and Inferential Statistics

Statistical procedures can be divided into two major categories: descriptive statistics and inferential statistics.
Descriptive Statistics

- Descriptive statistics includes statistical procedures that we use to describe the population we are studying. The data could be collected from either a sample or a population, but the results help us organize and describe data. Descriptive statistics can only be used to describe the group that is being studying. That is, the results cannot be generalized to any larger group.
- Descriptive statistics are useful and serviceable if you do not need to extend your results to any larger group. However, much of social sciences tend to include studies that give us universal truths about segments of the population, such as all parents, all women, all victims, etc.
- Frequency distributions, measures of central tendency (mean, median, and mode), and graphs like pie charts and bar charts that describe the data are all examples of descriptive statistics.

Inferential Statistics

- Inferential statistics is concerned with making predictions or inferences about a population from observations and analyses of a sample. That is, we can take the results of an analysis using a sample and can generalize it to the larger population that the sample represents. In order to do this, however, it is imperative that the sample is representative of the group to which it is being generalized.
- To address this issue of generalization, we have tests of significance. A Chi-square or T-test, for example, can tell us the probability that the results of our analysis on the sample are representative of the population that the sample represents.
- In other words, these tests of significance tell us the probability that the results of the analysis could have occurred by chance when there is no relationship at all between the variables we studied in the population we studied.
- Examples of inferential statistics include linear regression analyses, logistic regression analyses, ANOVA, correlation analyses, structural equation modeling, and survival analysis, to name a few.

0.12 Various Theory Components

- Distinguish between a bimodal distribution and a unimodal distribution
- Compare and contrast interval and ordinal data.

Example Given that $p_1 = 1/4, p_2 = 1/8, p_3 = 1/8, p_4 = 1/3, p_5 = 1/6$ find:

- $\sum_{i=1}^{i=n} p_i \times x_i$
- $\sum_{i=1}^{i=n} p_1 \times x_i^2$

Joint probability tables

- A joint probability table is a table in which all possible events (or outcomes) for one variable are listed as row headings, all possible events for a second variable are listed as column headings, and the value entered in each cell of the table is the probability of each joint occurrence.
- Often, the probabilities in such a table are based on observed frequencies of occurrence for the various joint events.
- The table of joint-occurrence frequencies which can serve as the basis for constructing a joint probability table is called a contingency table.