Lecture 7. Fundamental Sampling Distributions and Data Descriptions

Vu Nguyen Son Tung

Faculty of information technology
HANU

Statistics is a scientific subject that studies the rules of random phenomena of large numbers based on collecting and processing statistical data of observation results about random phenomena.

Statistics is a scientific subject that studies the rules of random phenomena of large numbers based on collecting and processing statistical data of observation results about random phenomena.

If we collect all the data related to the object to be researched, we can know more about that object. However, in reality that is impossible, because the scale of the research object is too large or during the research process the research object is destroyed. So we need to take samples for research. From that sample, we need to draw conclusions about the subject to be researched. That is the main task of this statistics section.

Populations and Samples

1. **A population** consists of the totality of the observations with which we are concerned. The number of observations in the population is defined to be the size of the population.
   Each observation in a population is a value of a random variable X having some probability distribution $f(x)$.
2. **A sample** is a subset of a population.
3. To eliminate any possibility of bias in the sampling procedure, it is desirable to choose **a random sample** in the sense that the observations are made independently and at random.

Data

Variable is a sign we are interested in studying as a whole.
Variables can be quantitative variables or qualitative variables.

1. A variable is called **quantitative variable** if it can be measured on each individual and has a numerical value. We call that the value of the variable.
2. The set of values of quantitative variables over the entire population gives us **quantitative data**.
3. A variable is called a **qualitative variable** if the value of that variable on each individual is the assignment to that individual of an attribute or assigns it to a certain category or level.
4. The set of all values of the qualitative variable over the entire population is **qualitative data**.

Data is the result, the observed value of the variables.
For example. To research students in universities, people are often interested in variables such as gender, age, ethnicity, major,....

Data collection

There are 2 data sources:

- Secondary: data from an available source, published or not.
- Primary: is data that researchers collect themselves to serve a specific content and research. There are 2 methods of collecting primary data:
  - Conducting experiments.
  - Conducting observations, investigations, surveys: by phone, written survey, and in-person survey, ...

Types of statistics

- **Descriptive statistics.** Provides methods to organize, describe and present collected data so that readers can best understand these data. Use charts and graphs to present data.
- **Inferential statistics.** After arranging the data appropriately, we need to analyze the data to understand the content hidden in the data. The task is to build methods that allow us to draw conclusions and make predictions with some accuracy on the whole based on a sample of data.

Presentation of sample data

1. Frequency distribution table. Suppose in a data sample of size n the values of the variable X have m different values $x_1 < x_2 < \ldots < x_m$. Suppose the value x_i has a repetition count of r_i. Then r_i is called the frequency of x_i. The following table is called the frequency distribution table:

X	x_1	x_2	$\ldots$	x_m
frequence	r_1	r_2	$\ldots$	r_m

2. Experimental distribution table. To compare the results when the sample size changes, we consider **the relative frequency** of the sample values. The quantity $f_i = r_i/n$, where n is the sample size, is called the relative frequency of x_i. The following table is called the experimental distribution table of X

X	x_1	x_2	$\ldots$	x_n	Sum
Frequency	r_1	r_2	$\ldots$	r_n	$\sum r_i = n$
Rrelative frequency	f_1	f_2	$\ldots$	f_n	$\sum f_i = 1$

3. Group distribution table. In the case of an investigation with a large sample size or when the variable has many different but close values, we often determine a number of intervals C_1, C_2, ..., C_m so that each value of the variable belongs to one and only one interval. The division is arbitrary. The widths of the intervals are not necessarily equal.

4. Some other ways of presentation. In addition to the above methods, we also use graphs and charts to represent them: stick frequency chart, stick frequency chart, frequency polygon chart, histogram, ...

Example 1. To evaluate the income of workers in an industrial park, 50 workers were randomly surveyed. The income results of each person (unit is million VND) by month are as follows:
4; 4.5; 4.2; 3.8; 4; 4.8; 5; 4.5; 4.2; 3.5; 5; 4.8; 4.5; 3.8; 4; 3.8; 3.5; 5; 4.2; 4.5; 3.8; 3.5; 5.2; 5; 4.5; 5.5; 5; 4.8; 4.5; 3.8; 4; 4.2; 4.5; 5; 5.2; 4.8; 4.8; 5; 3.5; 3.8; 3.5; 4.5; 4.2; 5; 4.5; 4.8; 5; 3.8; 3.5; 4.

1. Make a frequency distribution table.
2. Make a table of the experimental distribution.
3. Make a group distribution table.

Location Measures of a Sample

There are two important groups of sample characteristics:

1. Characteristics for images about the central position of the sample: Sample mean, sample median, mode.
2. Characteristics that show the degree of dispersion and variability of the data: Sample amplitude, interquartile range, sample mean deviation, variance and sample standard deviation.

Consider a random variable X with $\mu = EX, \sigma^2 = DX$.
Let $(x_1, x_2, \ldots, x_n)$ be a random sample with size n taken from X.

Sample mean

1. **The sample mean** is the average of the observed values, denoted by $\overline{x}$ and is determined by

$$\overline{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{m} r_i x_i}{\sum_{i=1}^{m} r_i}.$$

2. **Property:** $E\overline{x} = \mu, D\overline{x} = \dfrac{\sigma^2}{n}$.

In case the obtained sample has an unusual value, the sample mean does not give us an accurate conclusion, then we use the **sample median**.

Sample median

1. **The sample median** m is a number that satisfies the property: the number of values less than or equal to m is equal to the number of sample values greater than or equal to m.
2. To find m we arrange the values of the sample in ascending order, say $x_1 \leq x_2 \leq \ldots \leq x_n$. Then,
   - If n is odd then $m = x_{\frac{n+1}{2}}$.
   - If n is even then $m = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$.

Mode

The sample mode is the value of the sample that occurs most often.

Example. A shoe shop statistics the results of selling 200 pairs of shoes in the following table:

Selling price (thousand VND)	Number of pairs sold
$[30; 40)$	12
$[40; 50)$	27
$[50; 55)$	22
$[55; 60)$	35
$[60; 65)$	37
$[65; 70)$	16
$[70; 80)$	10
$[80; 90)$	21
$[90; 100)$	20
Sum	200

Find the mode range, median range, mean value of sales.

Solution. Price range $[60; 65)$ has the highest frequency 37 so that is the mode range.
Also, $12 + 27 + 22 + 35 < 100 < 12 + 27 + 22 + 35 + 37$ so $[60; 65)$ is the median.
The sample mean is $\overline{x} = 63.15$.

Measures of dispersion

1. The difference between the maximum value and the minimum value of the sample is called the sample range.
2. The sample mean deviation M_d is the average value of the distance from each value to the sample mean.

$$M_d = \frac{|x_1 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n}.$$

3. Sample variance ís the average of the squares of the deviations of the observations from their mean:

$$s^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

4. Sample standard deviation: $s = \sqrt{s^2}$.

Theorem

If s^2 is the variance of a random sample of size n, we may write

$$s^2 = \frac{\sum r_i (x_i - \overline{x})^2}{n-1} = \frac{1}{n(n-1)} \left(n. \sum r_i x_i^2 - \left(\sum r_i x_i \right)^2 \right)$$

THANK YOU FOR YOUR ATTENTION!