ECO220 Lecture Notes

Tianyu Du

May 17, 2018

Contents

1 Lecture 1 May. 8 2018

Content Chapter 1-4,

- Statistics
- Data
- Population

- Sample

1.1 Statistics

What is statistics Quantitive methods.

1.1.1 Example 1

Question This summer, 120 students enrolled in ECO220. Find out the number of courses that students are taking, the average number of courses they take, and the % of student taking 1 or 2 courses.

Population 120 students in ECO220. Noted as $N = 120$

Analyze:

1. Number of courses they take.
2. Average number of courses they take.
3. Percent of students taking 1 or 2 courses.

Data information collected from the whole *population* (all individuals). Use data to answer questions above.

number of courses	number of students	percent
1	40	0.33
2	30	0.25
3	30	0.25
4	15	0.14
5	5	0.03
Total	120	1.00

Parameters Parameters are fixed numbers. They can be calculated once we measure everyone in population.
<u>Examples of parameters from population</u>

- **Average** $\mu = 2.29$

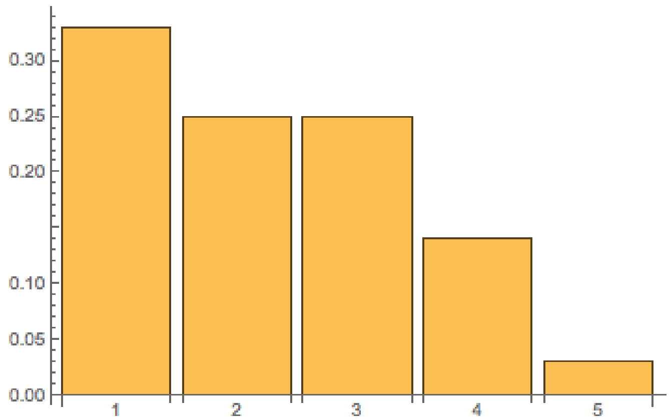

Figure 1: Frequency

1.1.2 Example 2

Question Find out the percentage of people in Ontario who are in favour of government policy.

Population People in Ontario.

In favour of policy	# of people in Ontario	%
Very much in favour	*	*
In favour	*	*
neutral	*	*
not in favour	*	*
strongly against	*	*
Total	N = Population of Ontario	1.00

Sample Since N is too large to handle, we select a sample, which is a subset of population, denoted as n, and then analyze the sample.

In favour of policy	# of people in Ontario	%
Very much in favour		
In favour		
neutral		
not in favour		
strongly against		
Total	n = Size of sample	1.00

The above chart based on sample data to *estimate* the chart using population data.

Let p be the % of people in Ontario(population) who are "very in favour" or "in favour"

Let $\hat{p}$ be the % of people in sample who are "very in favour" or "in favour", can be calculated based on the sample data.

The parameter p has an unknown value. The value of $\hat{p}$ can be calculated from sample data, $\hat{p}$ is an **estimate** for p.

Note p is a fixed value, but $\hat{p}$ will change from sample to sample. We call $\hat{p}$ an **estimator** (or **sample statistic**). The value of sample statistic will change from sample to sample, we call $\hat{p}$ a *random value.*

Parameters on population

- μ: Average
- p: Percentage

Sample Statistic on sample

- $\overline{x}$: Average
- $\hat{p}$: Percentage

Statistics

$$\text{Statistics}\begin{cases}\text{Descriptive statistics}\begin{cases}\text{Graph}\\ \text{Numerical measures}\end{cases}\\ \text{Inferential statistics: } \textit{Draw conclusions on a population based on sample data.}\end{cases}$$

2 Lecture 2 May. 9 2018

What is statistics? **Population** with size denoted with N and **sample** with its size denoted as n. Analyze the population from data from sample.

2.1 Inferential statistics

Involves *uncertainty*, to deal with the uncertainty, we need **probability**

2.2 Data

Two types of data

1. Quantitive data
   (a) Discrete
   (b) Continuous
2. Qualitative(Categorical) data

Note Some categorical data might be sensitive (e.g. income, age), to handle this, we could **categorize** the answers to handle this while collecting data.

2.3 Descriptive Statistics: Graphs

Example 1 Incomes in Toronto.

Example 2 Market shares of computers.

Example 3 Home price in Toronto.

Example 4 Age and income

Note There is no unique (or, correct) way of drawing graphs. A good graph is a picture that tells the audience a true picture of a population or sample.

2.4 Descriptive Statistic: Numerical Measures

2.4.1 Measures of centre (location)

Mean also called average and expected value, let $x_1, x_2, \ldots x_n$ be the measurements for the population of size N. The population mean is denoted by μ and defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Let $x_1, x_2, \ldots, x_n$ be measurements for the sample of size n, then the sample mean is denoted by $\overline{x}$ and defined as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note μ is population mean, therefore a *parameter*. That's μ has a *fixed value* if all units in population is measured. $\overline{x}$ is sample mean, and therefore a *sample statistic (estimator)* and $\overline{x}$ does not have a fixed value. The values of $\overline{x}$ change from sample to sample.

Note The mean is a good measure of centre, but it is sensitive to extreme values.

Median is the value in the middle when all data are sorted in order of magnitude.

Note For the data set with event numbers of observations, we defined the median as the mean of values of two observations in the middle.

Note 50% of data are less than the median.

Mode the value(s) that occurs most often.

Note there could be multiple modes in a dataset. (if there are tow modes, the data is called **bimoded**). Also it is possible for a dataset to have **no mode** (e.g. values of all observations are unique).

Percentile In general the k^{th} percentile is a number such that $k\%$ of data fall below this number.

Terminology

- 25^{th} percentile, also called 1^{st} quartile, denoted as $Q1$.
- 50^{th} percentile, also called 2^{nd} quartile, denoted as $Q2$. *Notice that $Q2$ is always the same as median.*
- 75^{th} percentile, also called 3^{rd} quartile, denoted as $Q3$.
- Interquartile is defined as $Q3 - Q1$.

3 Lecture 4 May. 17 2018

Notations

Variable	Population	Sample
size	N	n
mean	μ	$\overline{x}$
variance	σ^2	s^2
std dev	σ	s

Definition Coefficient of variation of a set of data is defined as $cv = \frac{std}{mean}$. Therefore, CV in population is defined as

$$CV_{population} = \frac{\sigma}{\mu}$$

And CV in sample is defined as

$$CV_{sample} = \frac{s}{\overline{x}}$$

3.1 Chapter 6. Covariance and Correlation

Data(Population) consider two sets(population) of data, $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_N\}$ with size N. And let μ_X and μ_Y denote the means of population X and Y, let σ_X and σ_Y denote the standard deviation of two sets of data.

Definition The covariance of two data sets, X and Y is defined as

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

Definition The correlation coefficient between X and Y is defined as

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Data(Sample) Consider two samples from data sets. $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ with size n.

Definition The covariance on sample is defined as

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

Definition The sample correlation coefficient is defined as

$$r = \frac{Cov(X,Y)}{s_X s_Y}$$

3.2 Interpretation

Example Consider two sets of data with $Cov(X,Y) = -25.3$.

1. The negative **sign** means X and Y have a negative relationship (*linear relationship*).
2. The **magnitude** has no meaning.

To have a measure that both the sign and magnitude of it have meaning, consider the correlation coefficient. If $r = -0.94$

1. The negative **sign** implies X and Y have a negative relationship.
2. The **magnitude** 0.94 means the relationship between X and Y is strong.

General Interpretation By definition of correlation coefficients on population and sample, we have

$$\rho \in [-1,1]$$

and

$$r \in [-1,1]$$

The sign suggests the direction of correlation, and the magnitude(absolute value) of coefficient shows the strength of correlation.