ECON 57 - Lecture 2

Descriptive Statistics

Augusto Gonzalez-Bonorino

Pomona College

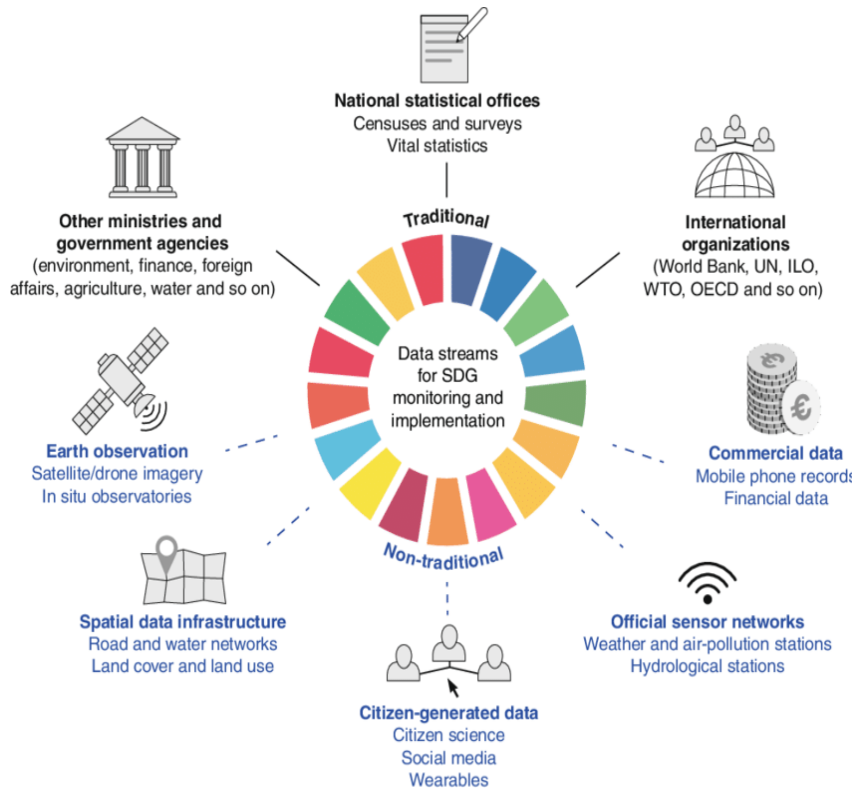
Fall 2023

Table of Contents

Data, Information, & Statistics

Introduction

- Data is measurable (can be encoded) and objective (raw data should not have human biases).
  - If and how the data of interest can be encoded determines the appropriate tools to extract their information.
  - Objectiveness will depend on the data generation process (if collected from Nature or simulated artificially) or data source (primary or secondary)



- Information is generated through measurement (i.e., the collection of observations).
- Measurement reduces our uncertainty or ignorance of a system relative to a unit of measurement.
  - Naturally, the importance or relevance of reducing our ignorance of a given subject will determine the level of interest and investment available.
- Statistical methods guide us on how to perform measurement (data collection and sampling methods) and reduce our ignorance of a subject relative to some context.

Data, Information, & Statistics

- **Data:** Observations collected from the real world.
  - Primary Data: Collected directly from participants, often through surveys, experiments, or observations. For example, conducting a survey to gather households' income data.
  - Secondary Data: Collected from already published sources, such as government reports, academic journals, or databases. For instance, using historical GDP data from a national economic database.

Data, Information, & Statistics

- **Data:** Observations collected from the real world.
- **Information:** Data that has been processed to be meaningful.
  - It is what is needed to make decisions and derive conclusions. For instance, turning raw data on consumer spending into insights about spending patterns and trends.
  - Analysis of data provides information. Economic indicators like unemployment rates or inflation rates are calculated from raw data and provide valuable information about the state of the economy.

Data, Information, & Statistics

- **Data:** Observations collected from the real world.
- **Information:** Data that has been processed to be meaningful.
- **Statistics:** The science of collecting, analyzing and interpreting data.
  - Descriptive Statistics: Describe, show, or summarize data. Measures like mean, median, and standard deviation help us understand the central tendencies and variability of economic data.
  - Inferential Statistics: Making inferences about populations using sample data. Techniques like hypothesis testing and regression analysis allow economists to draw conclusions about larger populations based on a sample.

Univariate Frequency Distributions/tables

A frequency distribution is a table used to organize data. It contains two columns:

- **Values:** Stores all possible values or responses of the variable being studied.
- **Frequency:** Stores all the frequencies (i.e., number of observations) for each class in **values**.

Univariate Frequency Distributions/tables

1. The absolute frequency of a value is the number of times it occurs in a dataset. $Fr(x = 1) = 4$
2. The relative frequency of a value is the number of time it occurs relative to the frequency of the remaining values in a dataset of size n. $p(x = 1) = \frac{Fr(x=1)}{n}$.
3. A frequency distribution is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.
4. A relative frequency distribution measures how relative frequencies are distributed over the values of the variables in the dataset.
5. Tables and charts are used to study frequencies.

Discrete Type Frequency Distribution

The frequency distribution of a discrete variable X is a list of each possible values for X along with the frequency with which each value occurs.

$$f(x) = \frac{\text{number of times x occurs}}{\text{total number of observations}}$$

X = Economic Class	**f(x) = Frequency of class**	**Relative Frequency**
poverty	40	40/4673
middle	1255	1255/4673
upper	598	598/4673
working	2780	2780/4673

Table: Economic Class Frequency Table with Relative Frequencies

Continuous Type Frequency Distribution

The same as for discrete, but instead of exact values we have ranges (bins).

X = Income Range ($)	**f(x) = Frequency of bin**	**Relative Frequenc**
$0 - $20,000	35	35/4085
$20,001 - $40,000	1530	1530/4085
$40,001 - $60,000	1250	1250/4085
$60,001 - $80,000	1255	1255/4085
$80,001 - $100,000	15	15/4085

Table: Income Distribution Frequency Table

Formal Representation

Let $x :=$ variable being analyzed, $n(x_i) :=$absolute frequency associated with x_i where $i : 1, 2, ..., k$, and $\frac{n(x_i)}{n} :=$ relative frequency of x_i.

X	$n(x_i)$	$\frac{n(x_i)}{n}$
x_1	$n(x_1)$	$\frac{n(x_1)}{n}$
x_2	$n(x_2)$	$\frac{n(x_2)}{n}$
x_3	$n(x_3)$	$\frac{n(x_3)}{n}$
...	...	...
x_{k-1}	$n(x_{k-1})$	$\frac{n(x_{k-1})}{n}$
x_k	$n(x_k)$	$\frac{n(x_k)}{n}$

Table: Generic Frequency Distribution/Table

Note that $\sum_i^k n(x_i) = n$ (the sample size)

Measures of centrality

Measures of central tendency are numbers that describe what is average or typical within a distribution of data. The choice of the measure of central tendency depends on the type of data and the context of the analysis.

- **Mean or Average:** The mean is the (arithmetic) average of the dataset $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Median:** The median is the middle value in the dataset when sorted in ascending order. It separates the lowest 50% from the highest 50% of values.
  - **Odd-numbered data:** Compute the middle point $\frac{n+1}{2}$ to get the median.
  - **Even-numbered data:** Compute the two middle points $\frac{n}{2}$ and $\frac{n}{2}+1$, and take their average $\frac{2n+2}{4}$.
- **Mode:** The mode is the most frequent value in the dataset.

Measures of Centrality

Consider the following dataset of luxury cars which could represent inventory, sales, or orders.

Luxury Car	Frequency
Mercedes-Benz	25
BMW	40
Audi	15
Jaguar	30
Porsche	20

Statistic	Value
Average	26
Median	25
Mode	BMW

Measures of Dispersion

Measures of dispersion quantify the spread of the data points.

- **Range:** The difference between the maximum and minimum values in the dataset.
- **Variance:** A measure of how far a set of numbers is spread out from their average value. It is the second central moment of a distribution. $\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$.
- **Standard deviation:** A measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set. $\sigma = \sqrt{\sigma^2}$
- **Coefficient of Variation:** It represents the ratio of the standard deviation to the mean and gives the relative measure of dispersion $CV = \frac{\sigma}{\text{mean(x)}} \times 100\%$

Measures of Dispersion

Example

Consider the following dataset representing the monthly incomes (in thousands of dollars) of 10 individuals:

$$x = \{30, 35, 40, 38, 42, 45, 50, 48, 55, 60\}$$

Now, let's calculate the variance, standard deviation, and coefficient of variation (CV) for this dataset.

$$\text{Mean(x)} = \frac{30 + 35 + 40 + 38 + 42 + 45 + 50 + 48 + 55 + 60}{10} = 44$$

$$\text{Variance(x)} = \frac{(30 - 44)^2 + (35 - 44)^2 + \ldots + (60 - 44)^2}{10} \approx 105.6$$

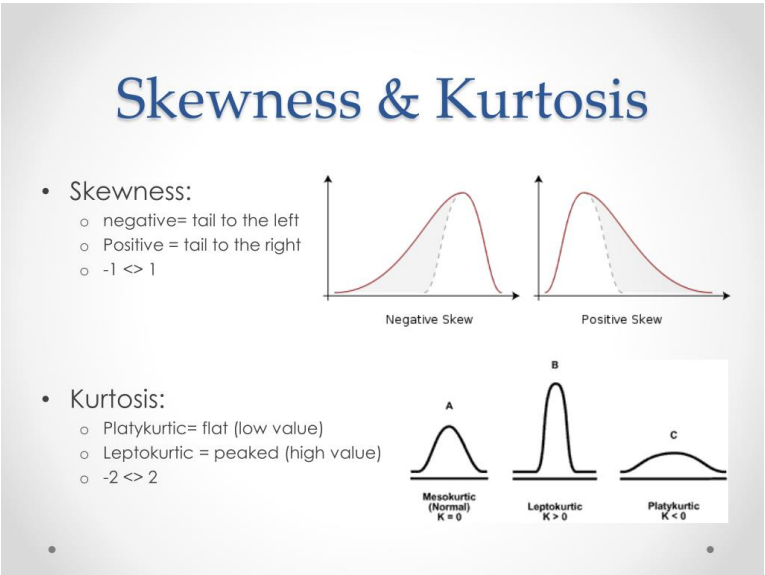
$$\text{Standard Deviation(x)} = \sqrt{\text{Variance(x)}} \approx 10.27$$

$$\text{Coefficient of Variation (CV)} = \frac{\text{sd(x)}}{\text{Mean(x)}} \times 100\% \approx \frac{10.27}{44} \times 100\% \approx 23.3\%$$

Skewness and Kurtosis

Skewness and kurtosis are measures of the shape of a distribution.

- **Skewness:** It measures the asymmetry of the distribution.
- **Kurtosis:** It measures the thickness of the tail of the distribution.
  - Positive skewness indicates a longer tail on the right, while negative skewness indicates a longer tail on the left. High kurtosis indicates heavier tails and more outliers.



These measures help us understand the departure of a distribution from normality.

Conceptual Exercises

1. Describe the differences between data, information, and statistics. Provide examples for each to illustrate their distinctions.
2. Consider two datasets: Dataset A with the following values: 10, 15, 20, 25, 30 and Dataset B with the following values: 50, 55, 60, 65, 70. Compare the measures of central tendency (mean, median, and mode) for both datasets and discuss how they differ and what insights they provide about the data.
3. For a dataset with positive skewness, explain how the mean, median, and mode relate to each other. Use an economic example to illustrate this relationship.
4. What is the main difference between discrete and continuous data? Illustrate your reasoning with examples of your daily life.

Mathemetical Exercises

1. Calculate the mean, median, and mode of the following dataset: 15, 20, 23, 30, 35, 40, 45, 45, 50, 55, 60, 65, 70. Interpret the results and explain which measure of central tendency best represents the dataset.
2. The variance of a dataset is given by $\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$. Consider the dataset $\{10, 15, 20, 25, 30\}$. Calculate the variance of this dataset and discuss what the variance value represents in the context of the data.
3. A dataset of test scores is given by $\{78, 85, 92, 88, 95\}$. Calculate the coefficient of variation (CV) for this dataset. Discuss what the CV value tells us about the relative variability of the test scores.

Economic Applications

Descriptive statistics are extensively used in economics to gain insights from data and make informed decisions. Here are some economic applications:

- **Income Distribution:** Using measures of central tendency and dispersion, we can analyze the income distribution in a country to assess income inequality.
- **Price Indexes:** Descriptive statistics are crucial in constructing price indexes, such as the Consumer Price Index (CPI), to measure inflation and track changes in the cost of living.
- **Market Research:** Companies use descriptive statistics to analyze consumer preferences, estimate demand, and make pricing decisions.
- **Unemployment Rate:** Measures of central tendency are used to calculate the unemployment rate, which is a vital indicator of an economy's health.