Chapter 1: Descriptive Statistics

Manuel V. Montesinos

Statistics Brush-Up Course
Competition and EPP Master Programs

Fall 2020

Introduction

Statistics is the mathematical science pertaining to the collection, analysis, interpretation, and presentation of data to learn about the world around us.

Using statistical tools, we can learn about the characteristics of a population by selecting a random sample:

- **Population**: set of individuals or objects.
- **Sample**: subset of a population.
- **Variable**: characteristic of a population which can take different values.

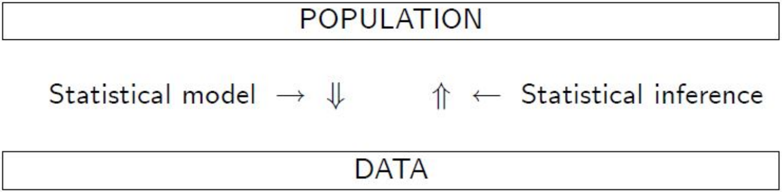
Introduction

Depending on what we would like to know about a population, a sample, or the relationship between these two, we will need to use a different item in the **statistician's toolkit**:

- **Probability theory** explains how data are generated from a population by means of statistical (or probability) models.
- **Statistical inference** uses the data to learn about the population that the sample is meant to represent. This is achieved by "inverting" the statistical model.
- **Descriptive statistics** aim to summarize a sample to provide a qualitative description of its main features.

Introduction

Figure 1: The Statistical Method



Introduction

In this chapter we will focus on descriptive statistics.

Data can be classified into **three types**:

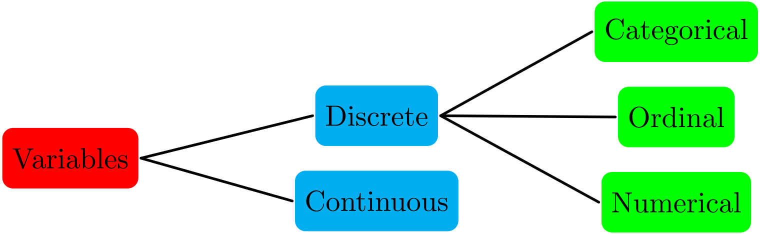
- Cross-sectional
- Time series
- Panel data

Types of variables:

- Discrete: categorical, ordinal or numerical.
- Continuous: can be treated as discrete if grouped in intervals.

Introduction

Figure 2: Types of Variables in Statistics



Frequency Distributions

Frequency Distributions

We build on an example: data for 1,844 individuals with information on **gross labor income** in year 2008.

Table 1: Labor Income Distribution (in USD, 1,844 Individuals)

	Absolute frequency	Relative frequency	Cumul. frequency	Bandwidth	Frequency density	Central point
Less than 10,000	34	0.018	0.018	10,000	0.018	5,000
10,000-19,999	122	0.066	0.085	10,000	0.066	15,000
20,000-29,999	247	0.134	0.219	10,000	0.134	25,000
30,000-39,999	321	0.174	0.393	10,000	0.174	35,000
40,000-49,999	289	0.157	0.549	10,000	0.157	45,000
50,000-59,999	243	0.132	0.681	10,000	0.132	55,000
60,000-79,999	285	0.155	0.836	20,000	0.077	70,000
80,000-99,999	144	0.078	0.914	20,000	0.039	90,000
100,000-149,999	118	0.064	0.978	50,000	0.013	125,000
150,000 or more	41	0.022	1	100,000	0.002	200,000

Frequency Distributions

The second column in Table 1 indicates the **absolute frequency**, which is the number of individuals in each category $g \in G$:

$$\sum_{g=1}^{G} n_g = N.$$

An alternative measure to compare how many individuals are in each income cell is given by the **relative frequency**:

$$f_g = \frac{n_g}{N}.$$

Relative (or absolute) frequencies can be represented by **bar graphs**. However, these can be misleading when we deal with continuous variables, since the results are sensitive to the selection of the **bandwidth**.

Frequency Distributions

Figure 3: Relative Frequency

Frequency Distributions

Histograms represent the **frequency density** of each interval, which is the ratio of the relative frequency to the width.

The **cumulative absolute frequency** is the number of observations in a given cell g or in the cells below:

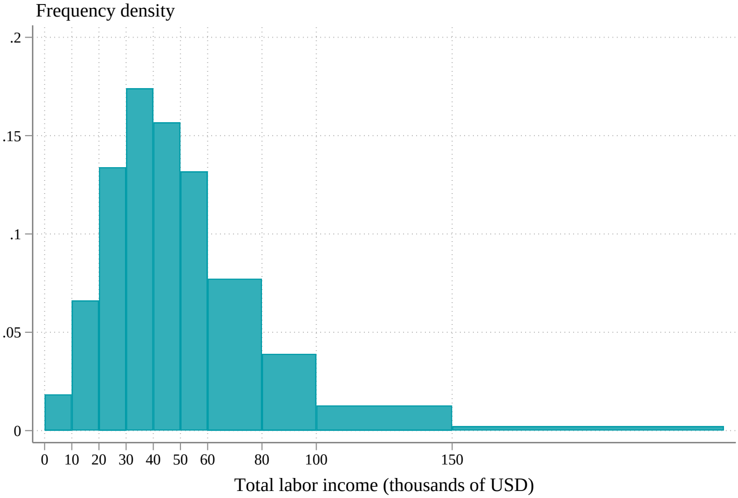
$$N_g = \sum_{h=1}^{g} n_h.$$

Analogously, the **cumulative relative frequency** is the fraction of observations in cell g or in the cells below:

$$F_g = \sum_{h=1}^{g} f_h.$$

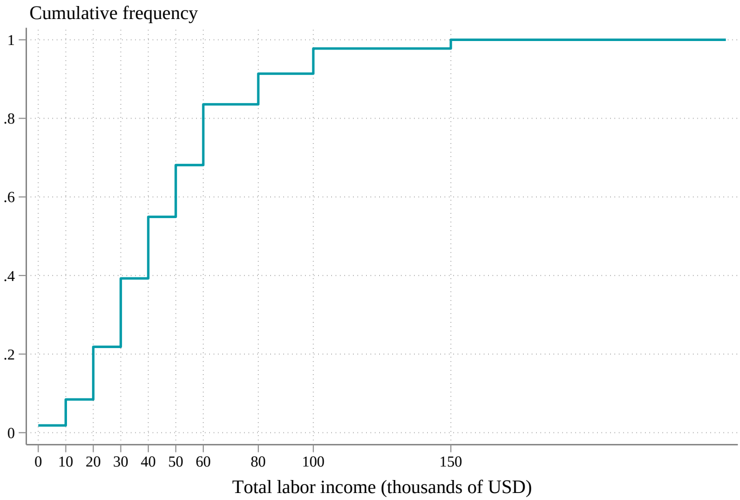
Frequency Distributions

Figure 4: Histogram



Frequency Distributions

Figure 5: Cumulative Frequency



Frequency Distributions

Discretizing continuous data in **intervals** may be misleading (relevant variation is gone vs. curse of dimensionality)

To compute the frequency density of x without discretizing it, we can use a **kernel function**:

$$d(g) = \frac{1}{N} \sum_{i=1}^{N} \kappa \left(\frac{x_i - x_g}{\gamma} \right),$$

where we use $\kappa \left(\frac{x_i - x_g}{\gamma} \right)$ as a weight, and the ratio outside of the sum is a normalization, such that the weights add up to one.

Frequency Distributions

In general, a **kernel** is a non-negative, real-valued, integrable function that:

- is symmetric,
- integrates to 1.

The parameter γ, used in the argument of the kernel, is known as the **bandwidth**, and its role is to penalize observations that are far from the conditioning point.

Examples of kernels:

- Equivalent to what we did without the kernel:

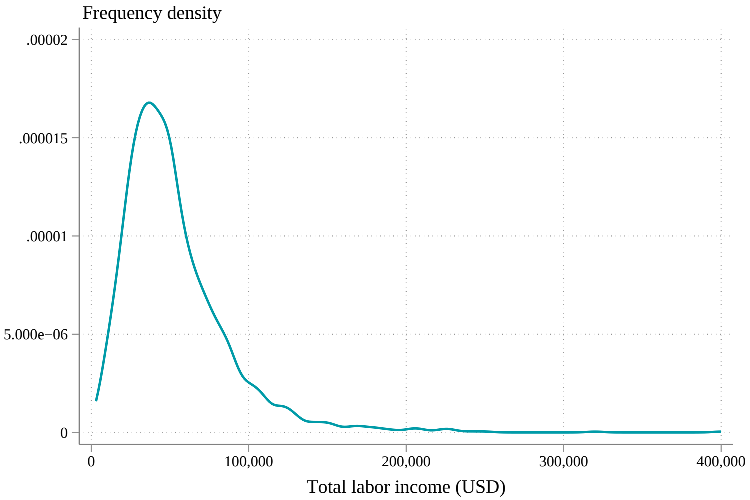
$$\kappa(u) = \begin{cases} 1 \text{ if } u = 0 \\ 0 \text{ if } u \neq 0. \end{cases}$$

- Gaussian kernel:

$$\kappa(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

Frequency Distributions

Figure 6: Gaussian Kernel



Summary Statistics

Summary Statistics

Summary statistics are used to summarize a set of observations from the data in order to communicate the largest amount of information as simply as possible.

Location statistics indicate a central or typical value in the data. The most commonly used one is the **sample mean**:

$$\bar{x} = \sum_{i=1}^{N} w_i x_i \underbrace{= \frac{\sum_{i=1}^{N} x_i}{N}}_{\text{if } w_i = 1/N \ \forall i},$$

where x_i is the value of x for observation i, N is the total number of observations, and w_i is the weight of the observations, such that $\sum_{i=1}^{N} w_i = 1$.

Main problem: it is sensitive to extreme values.

Summary Statistics

The **median** is the value of the observation that separates the upper half of the distribution from the lower half:

$$\text{med}(x) = \min\left\{x_g : F_g \geq \frac{1}{2}\right\}.$$

In other words, it leaves the same number of observations above and below her:

$$\text{med}(x) = \begin{cases} x_{\frac{N}{2}} + \frac{1}{2} \text{ if } N \text{ is odd,} \\ \frac{x_{\frac{N}{2}} + x_{\frac{N+1}{2}}}{2} \text{ if } N \text{ is even.} \end{cases}$$

Main advantage: it is not sensitive to extreme values.

Main inconvenient: changes in the tails of the distribution are not reflected.

Summary Statistics

The **mode** is the value with the highest absolute (or relative) frequency:

$$\operatorname{mode}(x) = \left\{ x_g : n_g \geq \max_{h \neq g} n_h \right\}.$$

A **loss function** $L(\cdot)$ describes the distance between the data and θ. For any u and v such that $0 < u < v$, it satisfies $0 = L(0) \leq L(u) \leq L(v)$, and $0 = L(0) \leq L(-u) \leq L(-v)$.

The **sample mean** is the minimizer of the *quadratic loss*:

$$\bar{x} = \min_{\theta} \sum_{i=1}^{N} w_i (x_i - \theta)^2.$$

The **median** is the minimizer of the *absolute loss*:

$$\operatorname{med}(x) = \min_{\theta} \sum_{i=1}^{N} w_i |x_i - \theta|.$$

Summary Statistics

Dispersion statistics indicate how the values of a variable differ from each other.

The **sample variance** is given by the average squared deviation with respect to the sample mean:

$$s^2 = \sum_{i=1}^{N} w_i (x_i - \bar{x})^2 \underbrace{= \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} = \sum_{g=1}^{G} (x_g - \bar{x})^2 f_g,}_{\text{if } w_i = 1/N \ \forall i}$$

The **standard deviation** is $s = \sqrt{s^2}$. It is in the same units as x.

The **coefficient of variation** does not depend on the units of x:

$$cv = \frac{s}{\bar{x}}.$$

Central moments

The variance belongs to a more general class of statistics known as **central moments**.

The (sample) central moment of order k:

$$m_k = \sum_{i=1}^{N} w_i (x_i - \bar{x})^k \underbrace{= \frac{\sum_{i=1}^{N} (x_i - \bar{x})^k}{N} = \sum_{g=1}^{G} (x_g - \bar{x})^k f_g}_{\text{if } w_i = 1/N \ \forall i} .$$

Some central moments: $m_0 = 1$, $m_1 = 0$, $m_2 = s^2$.

The 3rd central moment is the **skewness coefficient**:

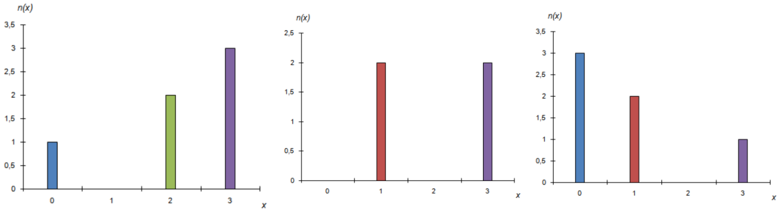
$$\text{skew}(x) = \frac{m_3}{s^3} \underbrace{= \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N s^3} = \frac{\sum_{g=1}^{G} (x_g - \bar{x})^3 f_g}{s^3}}_{\text{if } w_i = 1/N \ \forall i} .$$

Central moments

If $\text{skew}(x) > 0$, the distribution is skewed to the right (mean above the median). If $\text{skew}(x) < 0$, the distribution is skewed to the left (mean below the median).

Figure 7: Examples of Skewness

(a) $\text{skew}(x) < 0$ **(b)** $\text{skew}(x) = 0$ **(c)** $\text{skew}(x) > 0$



Central moments

The 4th central moment is the **kurtosis coefficient**:

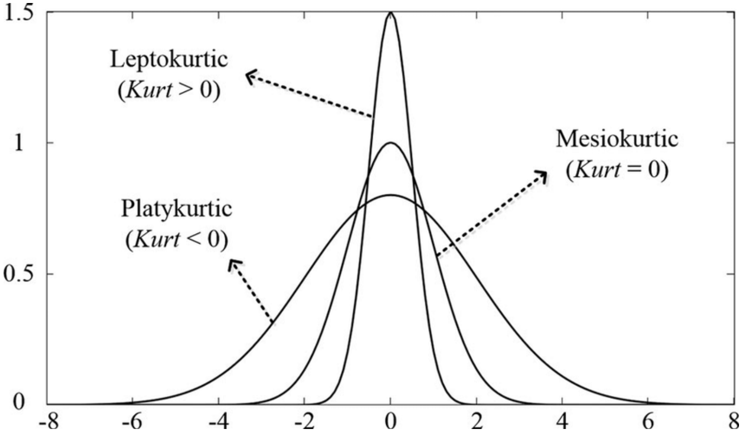
$$K = \frac{m_4}{s^4} - 3 \underbrace{= \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{Ns^4} - 3 = \frac{\sum_{g=1}^{G}(x - \bar{x})^4 f_g}{s^4} - 3}_{\text{if } w_i = 1/N \ \forall i}.$$

It measures the **thickness** of the tails of the distribution:

- $K < 0 \Rightarrow$ thicker tails than normal distribution.
- $K = 0 \Rightarrow$ normal distribution.
- $K > 0 \Rightarrow$ thinner tails than normal distribution.

Central moments

Figure 9: Kurtosis



Example

Table 2: Summary Statistics from Table 1

Statistic	Value
Sample mean ($\bar{x}$)	55,115
Median (med)	45,000
Mode	35,000
Variance (s^2)	1,263,061,746.57
Std. deviation (s)	35,539.58
Coef. variation (cv)	0.645
Skewness (skew)	1.8
Kurtosis (K)	4.377

Bivariate Frequency Distributions

Bivariate Frequency Distributions

Table 3: Joint Distribution of Income and Wealth (1,844 Individuals). Absolute Frequencies

	Wealth (in USD):						
Labor income (in USD):	Less than 1,000	1,000 -4,999	5,000 -19,999	20,000 -59,999	60,000 -199,999	200,000 or more	Total
A. Absolute Frequencies							
Less than 10,000	3	8	9	4	7	3	34
10,000-19,999	22	18	30	16	32	4	122
20,000-29,999	18	42	73	62	47	5	247
30,000-39,999	14	34	59	79	124	11	321
40,000-49,999	8	21	58	66	114	22	289
50,000-59,999	0	12	25	82	109	15	243
60,000-79,999	3	10	34	72	133	33	285
80,000-99,999	3	2	12	31	77	19	144
100,000-149,999	1	2	6	21	64	24	118
150,000 or more	0	1	1	6	25	8	41
Total	72	150	307	439	732	144	1,844

Bivariate Frequency Distributions

Table 4: Joint Distribution of Income and Wealth (1,844 Individuals). Relative Frequencies

	Wealth (in USD):						
Labor income (in USD):	Less than 1,000	1,000 -4,999	5,000 -19,999	20,000 -59,999	60,000 -199,999	200,000 or more	Total
B. Relative Frequencies (%)							
Less than 10,000	0.163	0.434	0.488	0.217	0.380	0.163	1.844
10,000-19,999	1.193	0.976	1.627	0.868	1.735	0.217	6.616
20,000-29,999	0.976	2.278	3.959	3.362	2.549	0.271	13.395
30,000-39,999	0.759	1.844	3.200	4.284	6.725	0.597	17.408
40,000-49,999	0.434	1.139	3.145	3.579	6.182	1.193	15.672
50,000-59,999	0.000	0.651	1.356	4.447	5.911	0.813	13.178
60,000-79,999	0.163	0.542	1.844	3.905	7.213	1.790	15.456
80,000-99,999	0.163	0.108	0.651	1.681	4.176	1.030	7.809
100,000-149,999	0.054	0.108	0.325	1.139	3.471	1.302	6.399
150,000 or more	0.000	0.054	0.054	0.325	1.356	0.434	2.223
Total	3.905	8.134	16.649	23.807	39.696	7.809	100.000

Bivariate Frequency Distributions

Tables 3 and 4 are **contingency tables**. They present the absolute and relative **joint frequencies** of labor income and wealth:

- Each value of Table 3 is the absolute frequency n_{gh} for the cell with $g \in \{1, ..., G\}$ labor income and $h \in \{1, ..., H\}$ wealth.
- The values in Table 4 are computed as

$$f_{gh} = \frac{n_{gh}}{N}.$$

To obtain the relative frequencies of one of the variables, i.e. the **marginal frequencies**, we sum over one of the dimensions:

$$f_g = \sum_{h=1}^{H} f_{gh} = \frac{\sum_{h=1}^{H} n_{gh}}{N} = \frac{n_g}{N}.$$

Bivariate Frequency Distributions

We can also be interested in computing **conditional relative frequencies**, i.e. the relative frequency of $y_i = h$ for the subsample with $x_i = g$:

$$f(y = h|x = g) = \frac{n_{gh}}{n_g} = \frac{\frac{n_{gh}}{N}}{\frac{n_g}{N}} = \frac{f_{gh}}{f_g}.$$

Conditional Sample Mean

Conditional Sample Mean

Restricting the sample to observations with $y_i = y$, we can calculate the conditional version of all the summary statistics introduced before.

The **conditional sample mean** is given by

$$\bar{x}_{|y=y_h} = \sum_{g=1}^{G} f(x_g|y = y_h) \times x_g.$$

Table 5: Conditional Means of Labor Income by Level of Wealth

Wealth	Mean labor income
Less than $1,000$	31,250
$1,000 - 4,999$	36,566.67
$5,000 - 19,999$	41,628.66
$20,000 - 59,999$	54,009.11
$60,000 - 199,999$	63,381.15
$200,000$ or more	76,527.78

Conditional Sample Mean

All previous discussion is for the case in which we **condition** on a **discrete** variable.

To compute the conditional mean of x given y without discretizing y, we can use a **kernel function**:

$$\bar{x}_{|y=y_h} = \frac{1}{\sum_{i=1}^{N} \kappa\left(\frac{y_i - y_h}{\gamma}\right)} \sum_{i=1}^{N} x_i \times \kappa\left(\frac{y_i - y_h}{\gamma}\right),$$

where we use $\kappa\left(\frac{y_i - y_h}{\gamma}\right)$ as a weight, and the ratio outside of the sum is a normalization, such that the weights add up to one.

Sample Covariance and Correlation

Sample Covariance and Correlation

We now review two measures that provide information on the (linear) **co-movements** of two variables.

The **sample covariance** is defined as

$$
\begin{aligned}
s_{x,y} &= \sum_{i=1}^{N} w_i (x_i - \bar{x})(y_i - \bar{y}) \\
&\underbrace{= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N} = \sum_{g=1}^{G} \sum_{h=1}^{H} (x_g - \bar{x})(y_h - \bar{y}) f_{gh} \,.}_{\text{if } w_i = 1/N \ \forall i}
\end{aligned}
$$

If $s_{x,y}$ is positive (negative), it is more common to have individuals with deviations of x and y of the same (opposite) sign.

Sample Covariance and Correlation

The problem with the covariance is that **magnitudes** are hard to interpret.

The **correlation coefficient** indicates the strength of the linear relation:

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y}.$$

It ranges between -1 and 1. A value of 0 implies that the two variables are (linearly) uncorrelated.

One way to graphically illustrate the relationship between two variables is to use a **scatter plot**.

Sample Covariance and Correlation

Figure 10: Scatter Plot (Wealth vs. Labor Income)