Chapter 1: Descriptive Statistics*

Instructor: Sergi Quintana Garcia
Notes from Manuel V. Montesinos (UAB and IDEA)

Statistics Brush-Up Course

Fall 2022

I. Introduction

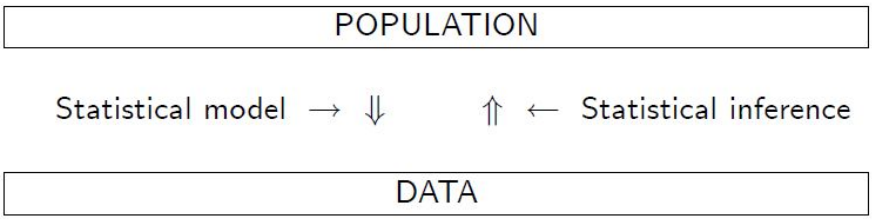
Statistics is the mathematical science pertaining to the collection, analysis, interpretation, and presentation of data to learn about the world around us. For example, what are the mean earnings of college graduates? Do mean earnings differ between men and women, and, if so, by how much? One way to answer these questions would be to perform an exhaustive survey of the population of workers, measuring the earnings of every worker and thus finding the population distribution of earnings. In practice, however, carrying out such a comprehensive survey would not be feasible. The process of designing the survey forms, managing and conducting the interviews, and compiling and analyzing the data would take years, apart from being extremely expensive. Therefore, a different, more practical approach is needed.

Statistics make our lives easier. Using statistical tools, we can learn about the characteristics of the full population by selecting a random sample from that population. Rather than surveying the entire population of a country, we might survey, say, 1,000 individuals selected at random, and analyze their data to reach tentative conclusions about the characteristics of the whole population.

A ***population*** is a set of individuals or objects (in our previous example, all the workers of a country), a ***sample*** is a subset of a population (the group of 1,000 workers we proposed to select at random), and a ***variable*** is a characteristic of a population which can take different values (workers' earnings). 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:

*These notes are partially based on James H. Stock and Mark W. Watson's textbook *Introduction to Econometrics*; Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer's textbook *Introduction to Econometrics with R*, and the Probability and Statistics courses taught by Joan Llull, Jordi Caballé and Anna Houštecká at the Universitat Autònoma de Barcelona and the Barcelona GSE. Typos, misprints, misconceptions and other errors are all mine.

Figure 1: The Statistical Method



- **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.

In this chapter, we will study descriptive statistics. Examples include numerical measures of the position or central tendency of the data (e.g. mean, median, mode) which give us a way to see where a certain data point or value falls in a sample (whether it is about the average, whether it is unusually high or low, etc.); measures of dispersion of the data (e.g. standard deviation, skewness, or kurtosis), the sample size size, or sample sizes of relevant subgroups.

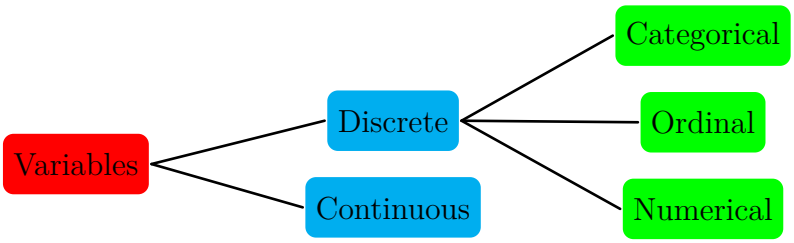
There are differences in the way we treat each type of data and variables. The data we analyze in Economics can be classified into three different types:

- **Cross-sectional data**: information of a sample of individuals at a given point in time (one observation per individual), such as the GPA of the students attending this class, or their earnings five years after graduating.
- **Time series data**: repeated observations of a given subject at different points in time (one observation per period), such as the GDP or the inflation rate in a country for the last ten years.
- **Panel data**: combination of multiple individuals with repeated observations at different points in time each, such as the earnings of the students attending this class, recorded year by year after graduating, or the GDP of a group of countries in the last ten years.

We typically distinguish between two types of variables: ***continuous***, which are the ones that can take values from an uncountable set of numbers (e.g. income), and

discrete or ***countable***, which are the ones that can take values from a countable set of numbers.[1] Discrete variables can be categorical, ordinal or numerical. A ***categorical*** or ***nominal*** variable is one that has two or more categories, but their values do not have a proper meaning, since there is no intrinsic ordering to the categories (e.g. a variable that equals 0 if the individual is a man, and 1 if she is a woman). An ***ordinal*** variable is similar to a categorical variable. The difference between the two is that there is a clear ordering of the variables (e.g. education level can have values such as elementary school graduate, high school graduate, some college and college graduate). Finally, a ***numerical*** or ***cardinal*** variable is similar to an ordinal variable, except that the intervals between the values are equally spaced (e.g. annual income divided in intervals of size $5,000.)

FIGURE 2: VARIABLE TYPES IN STATISTICS



II. Frequency Distributions

In this chapter, we build on an example to introduce the main notions we are after. Consider a dataset of 1,844 individuals with information on gross labor income in year 2008 for each of them. In Table 1 we describe the distribution of this variable in different ways. This variable is continuous, but we can ease its description by grouping the data in ***intervals*** (a.k.a. cells or groups) and deal with them as if we had a discrete variable.

The second column indicates the number of individuals in each category. This statistic is known as the ***absolute frequency***, or simply frequency. We denote it by n_g, where g indicates one of the G possible groups, and $\sum_{g=1}^{G} n_g = N$. The absolute frequency tells us how many individuals in the sample are in each income cell, but its values have limited information on the labor income distribution, unless they are compared to the frequencies in other cells.

We can use the ***relative frequency*** as an alternative measure for this comparison. We denote it by f_g. This statistic gives the fraction of individuals in the sample

[1] Continuous variables can be treated as discrete if they are grouped in intervals. For instance, we could divide a sample of workers between those with earnings higher that the average and those with lower earnings.

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

that are in cell g, and is defined as

$$f_g = \frac{n_g}{N}. \tag{1}$$

The relative frequency of our example is presented in the third column of Table 1, and graphically in a bar chart in Figure 3. ***Bar graphs*** are composed of rectangular bars with proportional height to the values they represent. In our case, the height of the bars represents the relative frequencies, although they can also show the absolute frequencies.

Nevertheless, a misleading feature of the relative frequencies when we deal with continuous variables is that results are sensitive to the selection of bin widths. As it can be appreciated in Figure 3, the bars for intervals $[40,000 - 50,000)$ and $[60,000 - 80,000)$ have a similar height, but the intervals are differently sized. If we had grouped all observations in intervals of \$10,000, the last four bars of the figure would be shorter.

An alternative representation that avoids this problem is the ***histogram***. A histogram is a representation of frequencies shown as adjacent rectangles of area equal (or proportional) to the relative frequency. The height of the rectangles depicts the ***frequency density*** of the interval, which is the ratio of the relative frequency to the width of the interval. Histograms are sometimes normalized such that the total area displayed in the histogram equals 1. Figure 4 is a histogram of the data of our example. The height of the rectangles is normalized such that the frequency density of the intervals of the most common height (\$10,000) are relative frequencies.

The ***cumulative (relative) frequency***, denoted by F_g, indicates the fraction

FIGURE 3: RELATIVE FREQUENCY

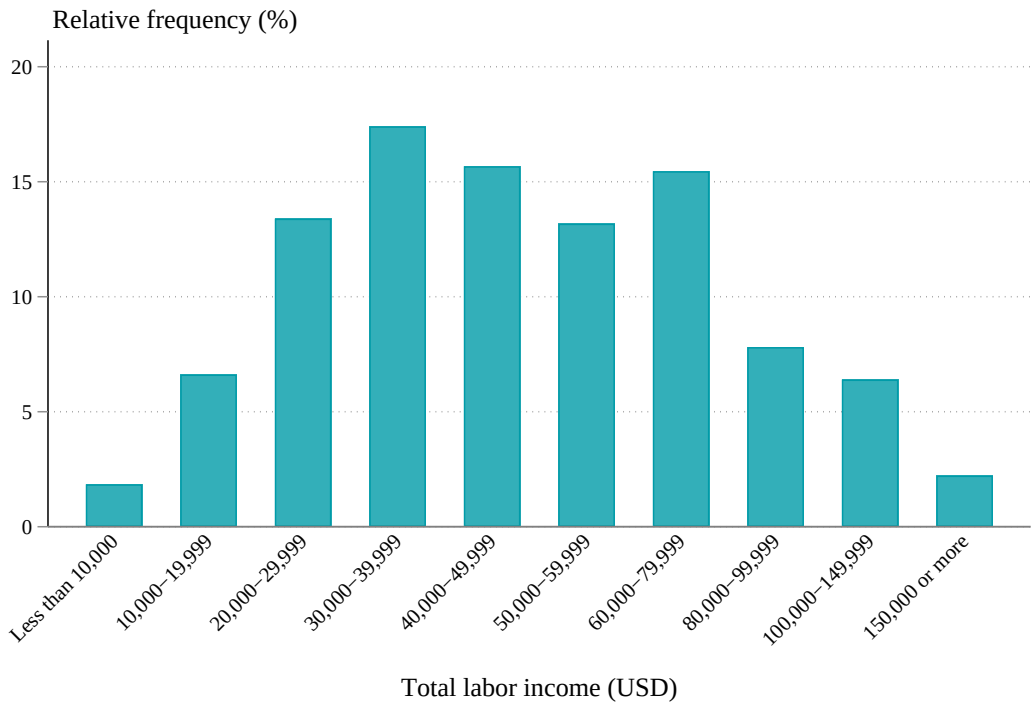
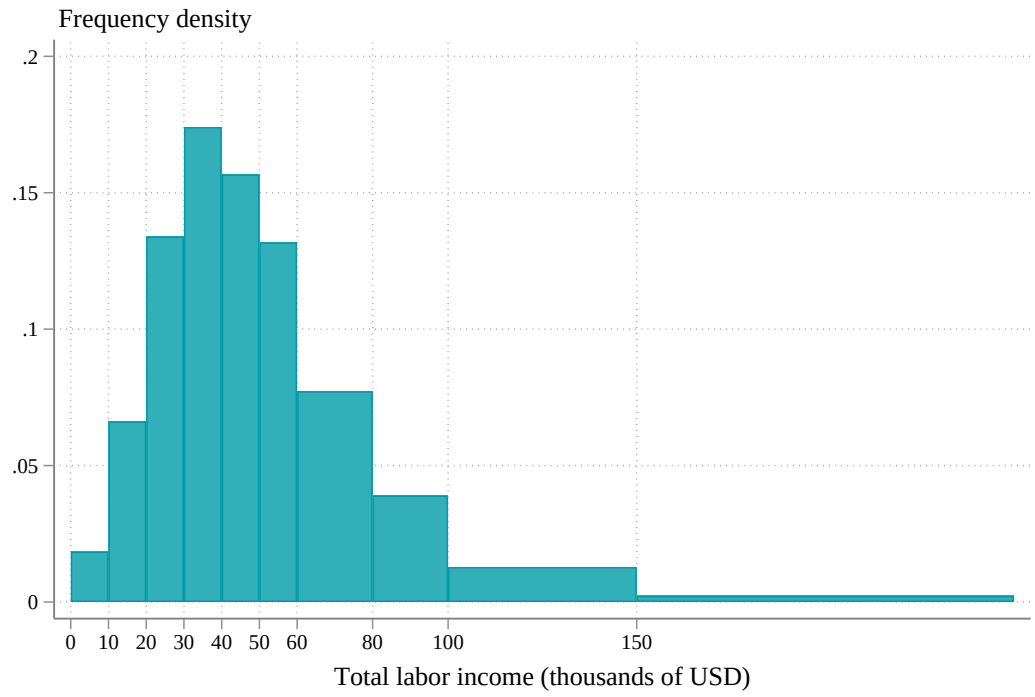


FIGURE 4: HISTOGRAM

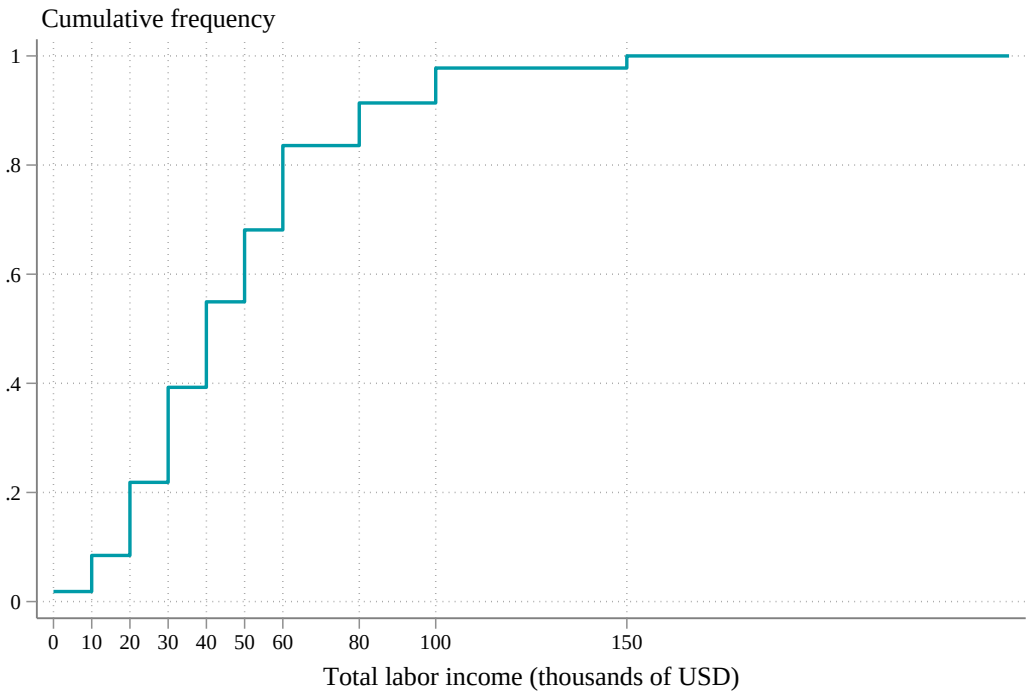


of observations in a given cell g or in the cells below.[2] Formally, the cumulative frequency is defined as:

$$F_g = \sum_{h=1}^{g} f_h. \tag{2}$$

In our example, the cumulative frequency is presented in the fourth column of Table 1 and in Figure 5.

FIGURE 5: CUMULATIVE FREQUENCY



All the exposition so far is on computing frequency distributions for discrete variables. When variables are continuous, we can use ***kernels*** to compute these distributions. In this case, we compute the frequency density as

$$d(g) = \frac{1}{N} \sum_{i=1}^{N} \kappa \left(\frac{x_i - x_g}{\gamma} \right), \tag{3}$$

where $\kappa(\cdot)$ is a ***kernel function***. In general, a kernel is a non-negative, real-valued, integrable function that is symmetric and integrates to 1. It assings a weight to each observation x_i based on the distance between this observation and the value we are conditioning on, x_g. An extreme example, which matches with the way we computed

[2] Analogously, the *cumulative absolute frequency* N_g is the number of observations in a given cell or in the cells below: $N_g = \sum_{h=1}^{g} n_h$.

relative frequencies with Equation (1) is

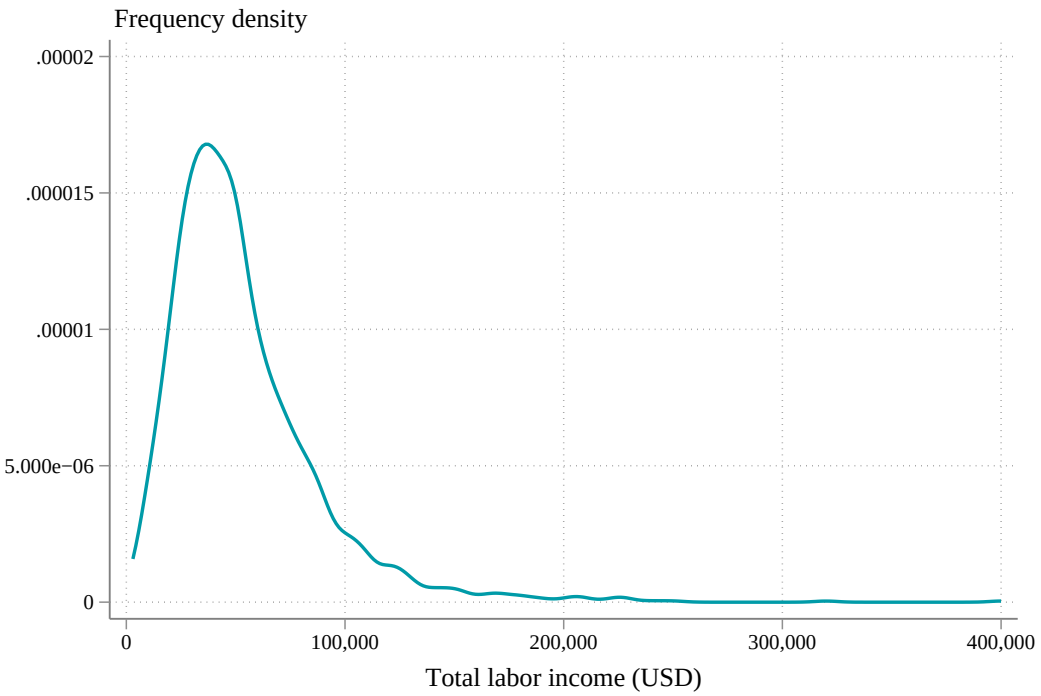
$$\kappa(u) = \begin{cases} 1, & \text{if } u = 0 \\ 0, & \text{if } u \neq 0, \end{cases} \tag{4}$$

where we only add the values if $x_i = x_g$ (or $u = x_i - x_g = 0$). The problem of this kernel function is that it is not smooth. A commonly used smooth alternative is the ***Gaussian kernel***, given by the probability density function of the normal distribution:

$$\kappa(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}. \tag{5}$$

Figure 6 depicts the result of using this kernel on our data.

FIGURE 6: GAUSSIAN KERNEL



Meanwhile, the parameter γ, used in the argument of the kernel, is known as the ***bandwidth***. Its role is to penalize observations that are far from the conditioning points, so that we can decide how much weight to give to observations with a value that is very different from the reference point g without having to change the function $\kappa(\cdot)$. The larger the γ, the lower the penalty to deviations, and hence the larger the window of relevant observations used in the computation.

III. Summary Statistics

Summary statistics are used to summarize a set of observations from the data, helping to communicate the largest amount of information as simply as possible. Typical summary statistics include measures of location or central tendency (e.g. mean, median, mode) and statistical dispersion (e.g. standard deviation, skewness, kurtosis).

A. Location statistics

Location statistics indicate a central or typical value in the data. The most commonly used one in the ***sample mean*** (a.k.a. average, arithmetic mean, or, when the context is clear, simply the mean). This statistic is defined as the weighted sum of the numerical values of our variable of interest for each and every observation. Formally, the sample mean is defined as

$$\bar{x} = \sum_{i=1}^{N} w_i x_i, \tag{6}$$

where x_i is the value of variable x for observation i, N is the total number of observations, and w_i is the weight of the observation, such that $\sum_{i=1}^{N} w_i = 1$. When all observations have the same weight, $w_i = 1/N$, and the sample mean is simply the sum across observations of all values of x_i divided by the number of observations. In that case, the sample mean can be written as

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{7}$$

Sometimes we are interested in giving a different weight to each observation. For instance, consider the sample mean of the labor income variable in our example. Giving to each interval g a value equal to the central point x_g (listed in the last column of Table 1) and computing a sample mean of the ten bins without using weights would not be appropriate, because each bin includes a different set of individuals. In this case it would be more accurate to compute the sample mean using the relative frequencies as weights:[3]

$$\bar{x}_i = \frac{\sum_{g=1}^{G} x_g n_g}{N} = \sum_{g=1}^{10} x_g f_g. \tag{8}$$

[3]Note that the relative frequencies are valid weights, as they sum to 1.

Properties of the mean

1. $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$

   Proof.
   $$\sum_{i=1}^{N}(x_i - \bar{x}) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \bar{x} = N\bar{x} - N\bar{x} = 0.$$

   □

2. $\overline{kx} = k\bar{x}$, where k is a constant (or scalar).

   Proof. Note that the values taken by the variable kx for the N individuals are $kx_1, kx_2, ..., kx_N$. Therefore,
   $$\overline{kx} = \frac{\sum_{i=1}^{N} kx_i}{N} = k \cdot \left(\frac{\sum_{i=1}^{N} x_i}{N}\right) = k\bar{x}.$$

   □

3. Let X_1 be $x_{11}, x_{12}, ..., x_{1N}$, X_2 be $x_{21}, x_{22}, ..., x_{2N}$, and consider the scalars α_1 and α_2, such that the values taken by the linear combination $Z = \alpha_1 X_1 + \alpha_2 X_2$ are $Z : \alpha_1 x_{11} + \alpha_2 x_{21}, \alpha_1 x_{12} + \alpha_2 x_{22}, ..., \alpha_1 x_{1N} + \alpha_2 x_{2N}$. The mean of Z is $\bar{Z} = \alpha_1 \bar{X}_1 + \alpha_2 \bar{X}_2$.

   Proof.
   $$\begin{aligned}\bar{Z} &= \frac{\sum_{i=1}^{N} z_i}{N} = \frac{\sum_{i=1}^{N}(\alpha_1 x_{1i} + \alpha_2 x_{2i})}{N} = \frac{\sum_{i=1}^{N} \alpha_1 x_{1i} + \sum_{i=1}^{N} \alpha_2 x_{2i}}{N} \\ &= \frac{\alpha_1 \sum_{i=1}^{N} x_{1i} + \alpha_2 \sum_{i=1}^{N} x_{2i}}{N} = \frac{\alpha_1 \sum_{i=1}^{N} x_{1i}}{N} + \frac{\alpha_2 \sum_{i=1}^{N} x_{2i}}{N} \\ &= \alpha_1 \bar{X}_1 + \alpha_2 \bar{X}_2.\end{aligned}$$

   □

The main problem of the sample mean as a location statistic is that it is very sensitive to extreme values. A very extreme observation can deviate its value substantially. An alternative measure that is not sensitive to extreme values is the ***median***. The median is the value of the observation that separates the upper half

of the distribution from the lower half:

$$\text{med}(x) = \min\left\{x_g : F_g \geq \frac{1}{2}\right\}. \tag{9}$$

This means that, if we sort all the observations as $x_1 \leq x_2 \leq ... \leq x_{N-1} \leq x_N$, the median is the value of the variable for the observation that leaves the same number of observations above and below her. Then, the median of x is

$$\text{med}(x) = \begin{cases} x_{\frac{N}{2}+\frac{1}{2}} \text{ if } N \text{ is odd,} \\ \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} \text{ if } N \text{ is even.} \end{cases} \tag{10}$$

Examples

1. $N = 7$ and the ordered values are $1, 1, 3, 6, 8, 8, 12$. The median is $x_4 = 6$.
2. $N = 8$ and the ordered values are $1, 1, 3, 6, 8, 8, 12, 3472$. The median is

$$\frac{x_4 + x_5}{2} = \frac{6 + 8}{2} = 7.$$

As noted above, the main advantage of the median is that it is not sensitive to extreme values. However, its main inconvenience is that changes in the tails of the distribution are not reflected, because the median only takes into account the frequencies of these values, but not the values themselves.

A third statistic that is often used to describe location is the ***mode***. The mode is the value that appears the largest number of times, i.e., the one with the highest absolute (or relative) frequency:

$$\text{mode}(x) = \left\{x_g : n_g \geq \max_{h \neq g} n_h\right\}. \tag{11}$$

While the mean and the median are measures of the centrality of the data in the strictest sense, the mode gives the most typical value.[4] If we work with grouped data, as we are doing with labor income in our example, we can use the midpoint values of the intervals, leading to an approximation error.

As central statistics, both the sample mean and the median can be computed by minimizing the distance between the data points in the sample and the statistic. The function that describes the distance between the data and a parameter or statistic

[4]Note that some variables can have more than one mode.

of interest θ is called the ***loss function***, denoted by $L(\cdot)$. For any values u and v such that $0 < u < v$, the loss function needs to satisfy $0 = L(0) \leq L(u) \leq L(v)$ and $0 = L(0) \leq L(-u) \leq L(-v)$. It can be proved that the sample mean minimizes the sum of squared deviations (quadratic loss), and the median minimizes the sum of absolute deviations (absolute loss):

$$\bar{x} = \min_{\theta} \sum_{i=1}^{N} w_i (x_i - \theta)^2 \tag{12}$$

$$\text{med}(x) = \min_{\theta} \sum_{i=1}^{N} w_i |x_i - \theta|. \tag{13}$$

B. Dispersion statistics

Dispersion statistics indicate how the values of a variable differ from each other. They summarize the deviations with respect to a location measure, typically the sample mean.

The ***sample variance*** or, when the context is clear, simply the variance, is given by the average squared deviation with respect to the sample mean:

$$s^2 = \sum_{i=1}^{N} w_i (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} = \sum_{g=1}^{G} (x_g - \bar{x})^2 f_g, \tag{14}$$

where the last two equalities apply when all the observations are weighted equally by $1/N$.

Properties of the variance

1. $s^2 = \overline{x^2} - \bar{x}^2$.

   Proof.

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} = \frac{1}{N} \sum_{i=1}^{N} \left[x_i^2 - 2 x_i \bar{x} + \bar{x}^2 \right] \\ &= \frac{1}{N} \left[\sum_{i=1}^{N} x_i^2 - 2\bar{x} \sum_{i=1}^{N} x_i + N\bar{x}^2 \right] = \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\bar{x} \left(\frac{\sum_{i=1}^{N} x_i}{N} \right) + \bar{x}^2 \\ &= \overline{x^2} - 2\bar{x}\bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2. \end{aligned}$$

   □

2. As a consequence of the previous property,

$$s_x^2 = \overline{x^2} - \bar{x}^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 = \frac{\sum_{g=1}^{G} x_g^2 n_g}{N} - \bar{x}^2 = \sum_{g=1}^{G} x_g^2 f_g - \bar{x}^2.$$

3. $s(kx)^2 = k^2 s(x)^2$, where k is a scalar.

   Proof.

$$\begin{aligned} s(kx)^2 &= \frac{\sum_{i=1}^{N}(kx_i - \overline{kx})^2}{N} = \frac{\sum_{i=1}^{N}(kx_i - k\bar{x})^2}{N} = \frac{k^2 \sum_{i=1}^{N}(x_i - \bar{x})^2}{N} \\ &= k^2 s(x)^2. \end{aligned}$$

□

4. Let X_1: $x_{11}, x_{12}, ..., x_{1N}$, X_2: $x_{21}, x_{22}, ..., x_{2N}$, and consider the scalars α_1 and α_2, such that the values taken by the linear combination $Z = \alpha_1 X_1 + \alpha_2 X_2$ are Z: $\alpha_1 x_{11} + \alpha_2 x_{21}, \alpha_1 x_{12} + \alpha_2 x_{22}, ..., \alpha_1 x_{1N} + \alpha_2 x_{2N}$. The variance of Z is $\bar{Z} = \alpha_1^2 s_{X_1}^2 + \alpha_2^2 s_{X_2}^2 + 2\alpha_1\alpha_2 \cdot s_{X_1,X_2}$.

   Proof.

$$\begin{aligned} s_z^2 &= \frac{\sum_{i=1}^{N}(z_i - \bar{Z})^2}{N} = \frac{\sum_{i=1}^{N}\left[\alpha_1 x_{1i} + \alpha_2 x_{2i} - \left(\alpha_1 \bar{X}_1 + \alpha_2 \bar{X}_2\right)\right]^2}{N} \\ &= \frac{\sum_{i=1}^{N}\left[\alpha_1\left(x_{1i} - \bar{X}_1\right) + \alpha_2\left(x_{2i} - \bar{X}_2\right)\right]^2}{N} \\ &= \alpha_1^2\left[\frac{\sum_{i=1}^{N}(x_{1i} - \bar{X}_1)^2}{N}\right] + \alpha_2^2\left[\frac{\sum_{i=1}^{N}\left(x_{2i} - \bar{X}_2\right)^2}{N}\right] \\ &+ 2\alpha_1\alpha_2\left[\frac{\sum_{i=1}^{N}\left(x_{1i} - \bar{X}_1\right)\left(x_{2i} - \bar{X}_2\right)}{N}\right] \\ &= \alpha_1^2 s_{X_1}^2 + \alpha_2^2 s_{X_2}^2 + 2\alpha_1\alpha_2 \cdot s_{X_1,X_2}. \end{aligned}$$

□

The ***standard deviation*** is defined as the square root of the variance: $s = \sqrt{s^2}$. The standard deviation is easier to interpret than the variance, since its value is in the same units as the variable of interest. An alternative measure that does not

depend on the units in which the outcome of interest is measured is the ***coefficient of variation***, which is a standardized measure of dispersion computed as the ratio between the standard deviation and the sample mean:

$$cv = \frac{s}{\bar{x}} \tag{15}$$

The coefficient of variation can be interpreted as the percentage deviation with respect to the average value of the variable. We say that it is immune to the units of measurement because for any scalar $k > 0$

$$cv(kx) = \frac{s(kx)}{|\overline{kx}|} = \frac{ks(x)}{|k\bar{x}|} = \frac{ks(x)}{k|\bar{x}|} = \frac{s(x)}{|\bar{x}|} = cv(x). \tag{16}$$

The variance belongs to a more general class of statistics known as ***central moments***. The (sample) central moment of order k, denoted by m_k, is defined as

$$m_k = \sum_{i=1}^{N} w_i (x_i - \bar{x})^k = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^k}{N} = \sum_{g=1}^{G} (x_g - \bar{x})^k f_g, \tag{17}$$

where the last two equalities hold if all observations are weighted by $1/N$. The central moment of order 0, m_0, is equal to 1, as $m_0 = \sum_{i=1}^{N} w_i = 1$. From the definition of the sample mean, it also follows that $m_1 = 0$. The second order central moment m_2 is the variance.

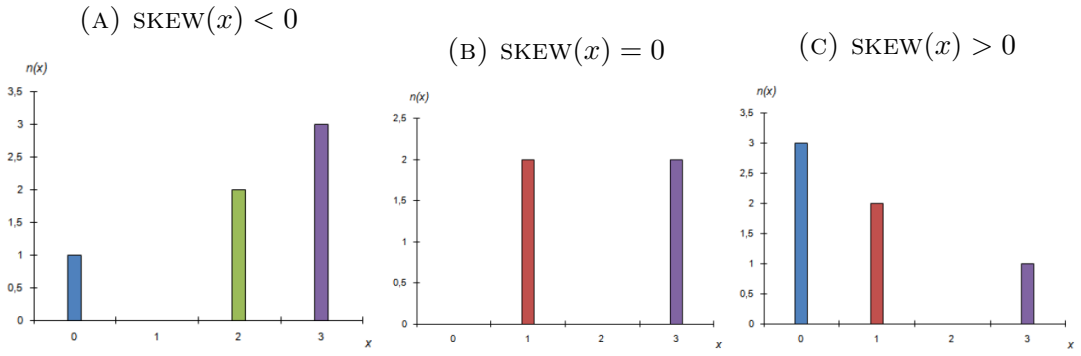
Other two central moments that are popular are the third and fourth order moments. The third moment is used to compute the ***skewness coefficient*** (a.k.a. coefficient of asymmetry), which is defined as

$$\text{skew}(x) = \frac{m_3}{s^3} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N s^3} = \frac{\sum_{g=1}^{G} (x_g - \bar{x})^3 f_g}{s^3}. \tag{18}$$

If the distribution is symmetric, $m_3 = 0$, because the right cubic deviations from the mean compensate exactly with the left ones (as the sample mean is the value that makes left and right deviations from it to compensate exactly). A positive skewness indicates that the distribution is skewed to the right, and a negative value implies the opposite. In a distribution that is skewed to the right, the mean is above the median, and the opposite is true if the distribution is skewed to the left. Figure 7 represents the three cases.

An analogous statistic computed from the fourth central moment is the ***(excess)***

FIGURE 7: EXAMPLE: SKEWNESS

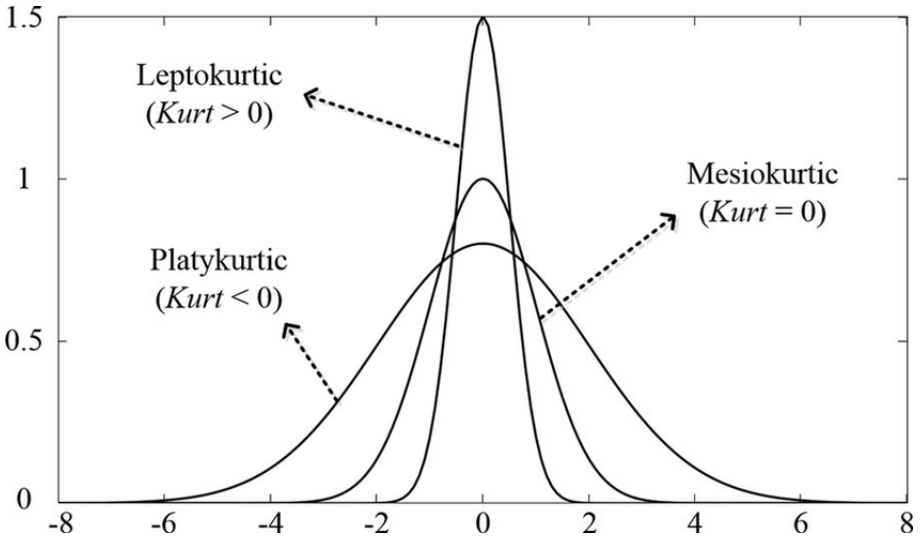


kurtosis coefficient, defined as

$$K = \frac{m_4}{s^4} - 3 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{Ns^4} - 3 = \frac{\sum_{g=1}^{G}(x - \bar{x})^4 f_g}{s^4} - 3. \tag{19}$$

This statistic measures the thickness of the tails of the distribution. For a normal distribution, $K = 0$ (that is why we normalize it by subtracting 3 from it).[5] Negative values indicate a *platykurtic* distribution (fatter tails than the normal distribution), whereas positive values indicate a *leptokurtic* distribution (thinner tails than the normal distribution). Figure 8 compares these types of distribution.

FIGURE 8: KURTOSIS



Source: Yang Wu.

Following with the example from Table 1, we compute all these descriptive statis-

[5]The kurtosis coefficient that is normalized by substracting 3 is often known as the *excess kurtosis coefficient*. In that terminology, the kurtosis coefficient would be defined as m_4/s^4.

tics, using the central point of the intervals as values for the variable, and the relative frequencies as weights. Table 2 presents the results. The sample mean is \$55,115, way above the median, which is \$45,000 (i.e. the 40,000-49,000 interval). The most frequent interval is 30,000-40,000, whose central point is 35,000. The variance is hard to interpret, but the standard deviation, which is 35,539.58 is quite high. The coefficient of variation is 0.645, which indicates that the standard deviation is 64.5% of the mean. The skewness coefficient is 1.8, which indicates a positively skewed distribution (indeed, the sample mean is larger than the median), and the kurtosis is quite high, so the distribution of total labor income is leptokurtic.

Table 2: Summary Statistics

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

IV. Bivariate Frequency Distributions

In this section, we extend the concepts of Section II and introduce some new ideas to describe the co-movements of two variables. In our example, we could be interested in comparing the distribution of labor income for individuals with different levels of wealth. Table 3 presents the absolute and relative ***joint frequencies*** of the same variable as in the example above (labor income) and wealth. This type of tables are known as ***contingency tables***. Note that the totals in the last column coincide with the absolute and relative frequencies in Table 1, but the table includes additional information. Each value of Panel A of the table is the absolute frequency n_{gh} of the cell with $g \in \{1, ..., G\}$ labor income and $h \in \{1, ..., H\}$ wealth. The relative frequencies in Panel B, denoted by $f(x = g, y = h)$ or simply f_{gh}, are computed analogously to Equation (1):

$$f_{gh} = \frac{n_{gh}}{N}. \tag{20}$$

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

	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
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

In order to obtain the relative frequencies of one of the variables (i.e., the last column or the last row of Panel B in Table 3), which are known in this context as ***marginal frequencies***, we have to sum over one of the dimensions. Thus, the totals in the last column of Panel B are obtained as

$$f_g = \sum_{h=1}^{H} f_{gh} = \frac{\sum_{h=1}^{H} n_{gh}}{N} = \frac{n_g}{N}, \tag{21}$$

and analogously for the totals in the last row, f_h.

We can also be interested in computing ***conditional relative frequencies***, that is, the relative frequency of wealth $y_i = h$ for the subsample of observations that have income $x_i = g$, denoted by $f(y = h|x = g)$:

$$f(y = h|x = g) = \frac{n_{gh}}{n_g} = \frac{\frac{n_{gh}}{N}}{\frac{n_g}{N}} = \frac{f_{gh}}{f_g}. \tag{22}$$

V. Conditional Sample Mean

Restricting the sample to observations with $y_i = y$, we can calculate the conditional version of all the descriptive statistics introduced in Section III. As they are all analogous, we focus on the conditional mean, which is

$$\bar{x}_{|y=y_h} = \sum_{g=1}^{G} f(x_g|y = y_h) \times x_g. \tag{23}$$

Table 4 shows the conditional mean of labor income for each level of wealth in our example.

Table 4: Conditional Means of Labor Income by Level of Wealth (in USD)

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

So far, we have assumed that the data is either discrete or grouped in discrete intervals. However, grouping data in intervals for a continuous variable can be

problematic. If intervals are too wide, we might lose relevant variation, but if they are two narrow, we might have very few observations to compute our statistics, and we can even have empty cells (we would suffer from the so called *curse of dimensionality*). Therefore, sometimes we might be interested in analyzing the data without grouping them in intervals.

We can compute the conditional mean of x given y without discretizing y by using a kernel function. The idea is to compute the mean of x for the observations with $y_i = y$, but also for other observations that have y_i close to y, giving a lower weight to those, based on how far they are. We can write this conditional mean as

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

where we use the kernel function $\kappa(\frac{y_i - y_h}{\gamma})$ as a weight, and the ratio outside of the sum is a normalization so that the weights sum to one. Using the kernel function defined in Equation (4), the conditional mean would match the result of applying Equation (23).

VI. Sample Covariance and Correlation

The final set of descriptive statistics presented in this chapter includes two measures that inform about the co-movements of two variables. These two measures speak about the existence of linear relations between two variables, but they can fail at detecting a nonlinear relation between them.

The first statistic is the ***sample covariance*** or, when the context is clear, simply the covariance, which is the average of the product of deviations of each variable with respect to its sample mean. Formally, the covariance between x and y is defined as

$$s_{x,y} = \sum_{i=1}^{N} w_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N} = \sum_{g=1}^{G} \sum_{h=1}^{H} (x_g - \bar{x})(y_h - \bar{y}) f_{gh}, \quad (24)$$

where the last two equalities hold if all observations are weighted equally by $1/N$.

The covariance can be positive, negative or equal to zero. A positive covariance indicates that it is more common to have individuals with deviations of x and y of the same sign, whereas a negative covariance indicates that deviations of the opposite sign are more common.

Properties of the covariance

1. $s_{x,x} = s_x^2$.

2. $s_{x,y} = \overline{x \cdot y} - \bar{x} \cdot \bar{y}$.

   Proof.

$$
\begin{aligned}
s_{x,y} &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i \cdot y_i - y_i \cdot \bar{x} - x_i \cdot \bar{y} + \bar{x} \cdot \bar{y}) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} x_i \cdot y_i - \bar{x}\sum_{i=1}^{N} y_i - \bar{y}\sum_{i=1}^{N} x_i + N \cdot \bar{x} \cdot \bar{y}\right) \\
&= \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N} - \bar{x}\left(\frac{\sum_{i=1}^{N} y_i}{N}\right) - \bar{y}\left(\frac{\sum_{i=1}^{N} x_i}{N}\right) + \bar{x} \cdot \bar{y} \\
&= \overline{x \cdot y} - \bar{x} \cdot \bar{y} - \bar{y} \cdot \bar{x} + \bar{x} \cdot \bar{y} = \overline{x \cdot y} - \bar{x} \cdot \bar{y}.
\end{aligned}
$$

   □

3. As a consequence of the previous property,

$$
\begin{aligned}
s_{x,y} &= \overline{x \cdot y} - \bar{x} \cdot \bar{y} = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y} = \sum_{g=1}^{G}\sum_{h=1}^{H} \frac{x_g \cdot y_h \cdot n_{gh}}{N} - \bar{x} \cdot \bar{y} \\
&= \sum_{g=1}^{G}\sum_{h=1}^{H} x_g \cdot y_h \cdot f_{gh} - \bar{x} \cdot \bar{y}.
\end{aligned}
$$

4. $s_{\alpha x,\beta y} = \alpha\beta s_{x,y}$, where α and β are scalars.

   Proof.

$$
\begin{aligned}
s_{\alpha x,\beta y} &= \frac{\sum_{i=1}^{N}\left(\alpha x_i - \overline{\alpha x}\right)\left(\beta y_i - \overline{\beta y}\right)}{N} = \frac{\sum_{i=1}^{N}\left(\alpha x_i - \alpha\bar{x}\right)\left(\beta y_i - \beta\bar{y}\right)}{N} \\
&= \frac{\alpha\beta\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{N} = \alpha\beta s_{x,y}
\end{aligned}
$$

   □

5. $s_{x,y} = s_{y,x}$.

The problem with the covariance is that its magnitude is not easy to interpret. Alternatively, the ***correlation coefficient*** indicates the strength of the linear re-

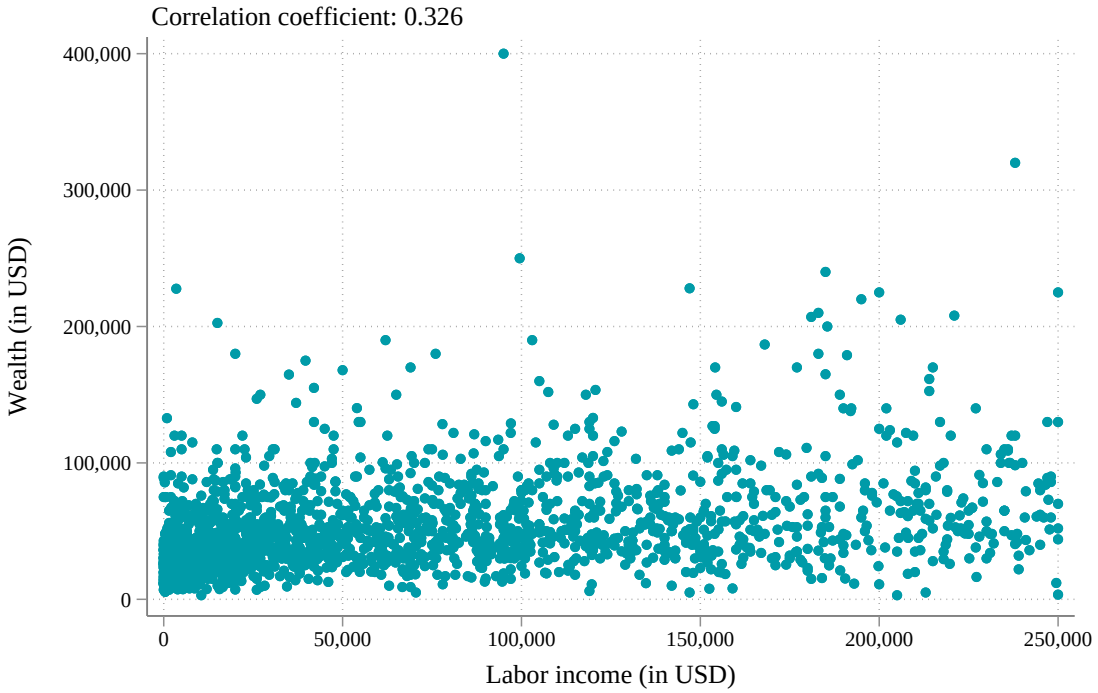
lation. It is defined as

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y}, \tag{25}$$

and it ranges between -1 and 1, with the former indicating perfect negative correlation, and the latter indicating perfect positive correlation. A value of 0 indicates that the two variables are linearly uncorrelated. Note that $r_{\alpha x,\beta y} = r_{x,y}$ if $\alpha > 0$ and $\beta > 0$, which implies that the correlation coefficient is immune to the units of measurement.

In our example, one way to illustrate the relationship between labor income and wealth is by means of a ***scatter plot***, like the one in Figure 9. In this type of diagram, the data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis (labor income, in our case) and the value of the other variable determining the position on the vertical axis (wealth). The figure also includes the value of the correlation coefficient, which indicates that there is a positive correlation between the two variables, but it is weak and likely unimportant.[6]

FIGURE 9: WEALTH VS. LABOR INCOME



[6]Usually, correlations are not considered strong until the correlation coefficient surpasses at least an absolute value of 0.8.