Chapter 1: Descriptive Statistics

By Joan Llull*

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

I. Introduction

Descriptive statistics is the discipline of qualitatively describing the main features of some data. It differs from inferential statistics in that the former aim to summarize a sample, whereas the latter uses the data to learn about the population that the sample is meant to represent. Examples include numerical measures of the position/central tendency of the data (e.g. mean, median, or mode), their dispersion (e.g. standard deviation, skewness, or kurtosis), the sample size, or sample sizes of relevant subgroups.

The data that we analyze in Economics can be classified in three different types:

- Cross-sectional data: information for a sample of individuals at a given point in time (one observation per individual).
- Time series data: repeated observations for a given subject at different points in time.
- Panel data: a sample that combines both types of information, i.e. multiple individuals with repeated observations at different points in time each.

We typically distinguish between two types of variables: continuous and discrete. Discrete variables can ordinal, cardinal, or categorical; in the latter case, their values do not have a proper meaning (e.g. a variable that equals 0 if the individual is a Male, and 1 if she is Female). There are differences in the way we treat each type of data and variables. However, continuous variables can be treated as discrete if they are grouped in intervals.

II. Frequency Distributions

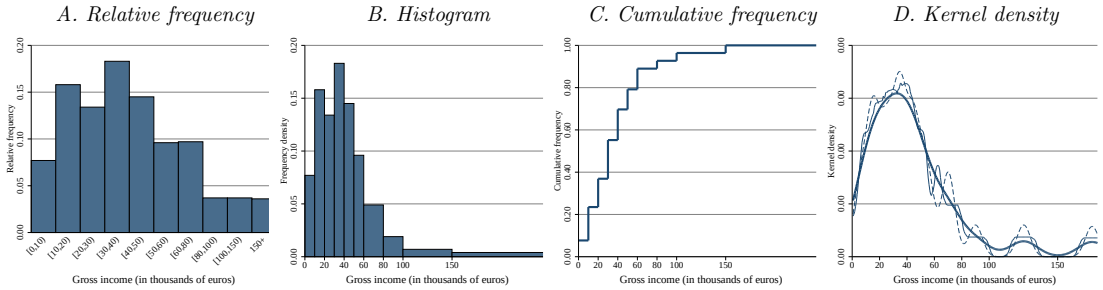
In this chapter, we build on a simple example to introduce the main notions that we are after. Consider a dataset of 2,442 households with information on household gross income in year 2010 for each of them. In Table 1 we describe the distribution of this variable in different ways. This variable is intrinsically continuous. In order to ease their description, the data are presented in intervals.

* Departament d'Economia i Història Econòmica. Universitat Autònoma de Barcelona. Facultat d'Economia, Edifici B, Campus de Bellaterra, 08193, Cerdanyola del Vallès, Barcelona (Spain). E-mail: joan.llull[at]movebarcelona[dot]eu. URL: http://pareto.uab.cat/jllull.

Table 1—Income Distribution (in Euros, 2,442 Households)

	Absolute frequency	Relative frequency	Cumul. frequency	Bandwidth	Frequency density	Central point
Less that 10,000	187	0.077	0.077	10,000	0.077	5,000
10,000-19,999	387	0.158	0.235	10,000	0.158	15,000
20,000-29,999	327	0.134	0.369	10,000	0.134	25,000
30,000-39,999	446	0.183	0.552	10,000	0.183	35,000
40,000-49,999	354	0.145	0.697	10,000	0.145	45,000
50,000-59,999	234	0.096	0.792	10,000	0.096	55,000
60,000-79,999	238	0.097	0.890	20,000	0.049	70,000
80,000-99,999	91	0.037	0.927	20,000	0.019	90,000
100,000-149,999	91	0.037	0.964	50,000	0.007	125,000
150,000 or more	87	0.036	1.000	100,000	0.004	200,000

Figure 1. Income Distribution (in Euros, 2,442 Households)



The first column indicates the number of households in each category. This statistic is known as ***absolute frequency***, or, simply, frequency. We denote it by N_a, where a indicates one of the A possible bins (a.k.a. cells or groups). The absolute frequency gives relevant information on how many households in the sample are in each income cell, but its values have limited information on the income distribution, unless they are compared to the frequencies in other cells.

An alternative measure that eases this comparison is the ***relative frequency***, denoted by $f(x = a)$, or simply f_a. The relative frequency gives the fraction of households in a sample that are in a given cell a, and is defined as:

$$f_a \equiv \frac{N_a}{N}, \tag{1}$$

where N_a is the number of observations in cell $a \in \{1, ..., A\}$, and $N \equiv \sum_{a=1}^{A} N_a$. In our example, the second column of Table 1 gives the relative frequencies. Graphically, the relative frequency is plotted in a bar chart in Figure 1A. A bar graph is a chart with rectangular bars with proportional height to the values they represent. In this case, the height of the bars represent the relative frequencies.

A misleading feature of the relative frequencies to represent continuous variables, as it can be appreciated in Figure 1A, is that results are sensitive to the selection

of bin widths. For example, the last three bars have a similar height, but they correspond to differently sized intervals. If we had grouped all observations in intervals of 10,000 euros, the bars at the right 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. Thus, the height of the rectangles depicts the ***frequency density*** of the interval, which is the ratio of the relative frequency and the width of the interval. Sometimes, histograms are normalized such that the total area displayed in the histogram equals 1.

Figure 1B is a histogram of the data presented in Table 1. The height of the rectangles is normalized such that the frequency density of the intervals of the most common height (10,000 euros) are relative frequencies (fifth column of Table 1).

The ***cumulative frequency***, $c(x = a)$ or simply c_a, calculated in the third column of Table 1, indicates the fraction of observations in a given cell a, or in the cells below. More formally, the cumulative frequency is defined as:

$$c_a \equiv \sum_{j=1}^{a} f_j. \tag{2}$$

In our example, the cumulative frequency is depicted in Figure 1C.

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

$$f(a) = \frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} \kappa \left(\frac{x_i - a}{\gamma} \right), \tag{3}$$

where $\kappa(\cdot)$ is a ***kernel function***. In general, a kernel function is a non-negative real-valued integrable function that is symmetric and integrates to 1.

The kernel function gives weight to observations based on the distance between x_i and the value we are conditioning on, a. An extreme example, which matches exactly 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 $x_i = a$ (or $u = x_i - a = 0$), exactly as before.

If we had the raw disaggregated data, and we wanted to use (equal size) intervals,

we could use the following kernel:

$$\kappa(u) = \begin{cases} 1, & \text{if } |u| \leq \tilde{u} \\ 0, & \text{if } |u| > \tilde{u} \end{cases}. \tag{5}$$

In this case, we are constructing intervals of size $2\tilde{u}$ centered at a. The slight difference between this case and what we did before with the intervals is that in this case we have a completely defined function for the conditional mean of y given x for all values of x. This function is constant for a while and then jumps every time a new observation comes in or steps out.

The problem of these two kernel functions is that they are not smooth. A commonly used smooth alternative is a ***Gaussian kernel***, which is given by the density of the normal distribution:

$$\kappa(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}. \tag{6}$$

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, so that we can decide how much weight to give to observations with x_i very different from a 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.

Following with our example, Figure 1D plots three different kernel frequencies. Thick and dashed lines use a Gaussian kernel, the former with the optimal bandwidth, and the latter with a bandwidth of 10,000. The thin solid line depicts a rectangular kernel (Equation (5)) with bandwidth equal to 10,000.

III. 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. Typical summary statistics include measures of location or central tendency (e.g. mean, median, mode) and statistical dispersion (e.g. standard deviation, skewness, kurtosis).

Location statistics indicate a central or typical value in the data. The most commonly used one is the ***arithmetic mean***, also known as average, sample 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. More formally, the sample mean is defined as:

$$\bar{x} \equiv \sum_{\mathrm{i=1}}^{\mathrm{N}} w_i x_i, \tag{7}$$

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 observation, such that $\sum_{\mathrm{i=1}}^{\mathrm{N}} w_i = 1$. When all observations have the same weight, $w_i = \frac{1}{N}$, and the sample mean is simply the sum across observations of all values of x_i divided by the number of observations.

Sometimes we are interested in giving different weight to each observation. For example, consider the sample average of the income variable presented in bins in Table 1 above. Giving to each bin a value equal to the central point of the interval (listed in the last column of the table) and computing a sample mean of the 10 bins without using weights would not provide the desired result because each bin includes a different set of individuals. Thus, in that case, it would be more appropriate to compute the sample average using the relative frequencies as weights:

$$\overline{inc}_i = \sum_{i=1}^{10} f_i \times inc_i. \tag{8}$$

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

The main problem of the sample mean as a location statistic is that it is very sensitive to extreme values. A single but 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 data from the lower half. Informally, the median is the value of the variable for the individual that, if we sort all observations, leaves the same number of observations above and below her. More formally, it is defined as:

$$\mathrm{med}(x) \equiv \min \left\{ a : c_a \geq \frac{1}{2} \right\}, \tag{9}$$

that is, the minimum value for which the cumulative frequency is above one half. The main advantage of the median, as noted above, is that it is not sensitive to extreme values. However, its main inconvenience is that changes in the tails 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 with the highest frequency. More formally:

$$\mathrm{mode}(x) \equiv \left\{ a : f_a \geq \max_{j \neq a} f_j \right\}. \tag{10}$$

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. Note that, in some instances, we can have more than one mode.

As central statistics, both the sample mean and the median can be computed minimizing the distance between the different data points in the sample and the statistic. The function that describes the distance between the data and a parameter or statistic of interest is called the ***loss function***, denoted by $L(\cdot)$. The loss function satisfies $0 = L(0) \leq L(u) \leq L(v)$ and $0 = L(0) \leq L(-u) \leq L(-v)$ for any u and v such that $0 < u < v$. With trivial algebra, it can be proved that the sample mean minimizes the sum of squared deviations (quadratic loss):

$$\bar{x} = \min_{\theta} \sum_{\mathrm{i}=1}^{\mathrm{N}} w_i (x_i - \theta)^2. \tag{11}$$

Similarly (though slightly more difficult to prove), the median minimizes the sum of absolute deviations (absolute loss):

$$\mathrm{med}(x) = \min_{\theta} \sum_{\mathrm{i}=1}^{\mathrm{N}} w_i |x_i - \theta|. \tag{12}$$

Dispersion statistics indicate how the values of a variable across different observations differ from each other. More specifically, 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:

$$s^2 \equiv \sum_{\mathrm{i}=1}^{\mathrm{N}} w_i (x_i - \bar{x})^2. \tag{13}$$

The ***standard deviation*** is defined as $s \equiv \sqrt{s^2}$. The interest in the standard deviation is because it is easy to interpret, as 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:

$$cv \equiv \frac{s}{\bar{x}}. \tag{14}$$

The coefficient of variation can be interpreted as a percentage deviation with respect to the average value of the variable.

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:

$$m_k \equiv \sum_{\mathrm{i}=1}^{\mathrm{N}} w_i (x_i - \bar{x})^k. \tag{15}$$

The central moment of order 0, m_0, is equal to one, as $m_0 = \sum_{\mathrm{i}=1}^{\mathrm{N}} w_i = 1$. From the definition of the sample mean $\bar{x}$, it also follows that $m_1 = 0$. The second order central moment m_2 is the sample variance. Other two central moments that are popular are the third and fourth order moments. The third order moment is used to compute the ***skewness coefficient***, which we denote by sk, and is defined as:

$$sk \equiv \frac{m_3}{s^3}. \tag{16}$$

If the distribution is symmetric, $m_3 = 0$, because the right cubic deviations from the mean exactly compensate with the left ones (as the sample mean is the value that makes left and right deviations from it to exactly compensate, since $m_1 = 0$ by construction). A positive sign for sk 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.

An analogous statistic computed from the fourth central moment is called the (excess) ***kurtosis coefficient***, and is defined as:[1]

$$K \equiv \frac{m_4}{s^4} - 3. \tag{17}$$

This statistic indicates how "fat" are the tails of the distribution. For a normal distribution, $K = 0$ (that is why we normalize it by subtracting 3 from it). Negative values indicate a platykurtic distribution (fatter tails than the normal distribution), whereas positive values indicate a leptokurtic distribution (thiner tails than the normal distribution).

Following with the example from Table 1, we compute all these descriptive statistics, 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 46,253 euros, way above the median, which is 25,000 euros (i.e. the 20,000-29,999 euro interval). The most frequent interval is 30,000-39,999 euros (whose central point is 35,000 euros). The variance is hard to interpret, but the

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

Table 2—Summary Statistics

Statistic:	Value
Sample mean ($\bar{x}$)	46,253
Median (med)	25,000
Mode	35,000
Variance (s^2)	1,575,784,440
Std. deviation (s)	39,696
Coef. variation (cv)	0.858
Skewness (sk)	2.24
Kurtosis (K)	5.82

standard deviation, which is 39,696 is quite high. The coefficient of variation is 0.858, which indicates that the standard deviation is 85.8% of the the mean. The skewness coefficient is 2.24, which indicates a positively skewed distribution (and indeed the sample mean is larger than the median), and the kurtosis is quite high.

IV. Bivariate Frequency Distributions

In this section, we extend the concepts in Section II (and introduce new ideas) to describe the co-movements of two variables. Table 3 presents the absolute and relative ***joint frequencies*** of the same variable as in the example above (gross income) and liquid assets. This type of tables are also know as ***contingency tables***. Note that the totals in the last column coincide with the absolute and relative frequencies presented in Table 1. However, the table includes additional information. Each value of the top panel of the table N_{ij} is the absolute frequency for the cell with $a \in \{1, ..., A\}$ income, and $b \in \{1, ..., B\}$ assets. The relative frequencies in the bottom panel, denoted by $f(x = a, y = b)$ or simply f_{ab}, are computed analogously to Equation (1):

$$f_{ab} = \frac{N_{ab}}{N}. \tag{18}$$

The relative frequencies are also presented in Figure 2.

To obtain the relative frequencies of one of the variables (i.e., the last column or last row of the bottom panel of Table 3), which are known in this context as ***marginal frequencies***, we sum over the other dimension:

$$f_a = \sum_{b=1}^{B} f_{ab} = \frac{\sum_{b=1}^{B} N_{ab}}{N} = \frac{N_a}{N}, \tag{19}$$

and analogously for f_b.

We can also be interested in computing ***conditional relative frequencies***, that is, the relative frequency of $y_i = b$ for the subsample of observations that

Table 3—Joint Distribution of Income and Liquid Assets (2,442 Households)

Gross Income (in euros):	Liquid assets (in euros): None	1-999	1,000-4,999	5,000-19,999	20,000-59,999	60,000-220,000	Total
A. Absolute Frequencies							
Less that 10,000	107	16	16	26	12	10	187
10,000-19,999	191	61	49	41	25	20	387
20,000-29,999	127	45	45	65	28	17	327
30,000-39,999	188	75	56	61	42	24	446
40,000-49,999	81	66	69	69	46	23	354
50,000-59,999	48	33	48	63	25	17	234
60,000-79,999	33	28	50	51	46	30	238
80,000-99,999	6	2	21	21	22	19	91
100,000-149,999	7	5	3	13	27	36	91
150,000 or more	2	0	0	7	14	64	87
Total	790	331	357	417	287	260	2,442
B. Relative Frequencies							
10,000-19,999	0.078	0.025	0.020	0.017	0.010	0.008	0.158
20,000-29,999	0.052	0.018	0.018	0.027	0.011	0.007	0.134
30,000-39,999	0.077	0.031	0.023	0.025	0.017	0.010	0.183
40,000-49,999	0.033	0.027	0.028	0.028	0.019	0.009	0.145
50,000-59,999	0.020	0.014	0.020	0.026	0.010	0.007	0.096
60,000-79,999	0.014	0.011	0.020	0.021	0.019	0.012	0.097
80,000-99,999	0.002	0.001	0.009	0.009	0.009	0.008	0.037
100,000-149,999	0.003	0.002	0.001	0.005	0.011	0.015	0.037
150,000 or more	0.001	0.000	0.000	0.003	0.006	0.026	0.036
Total	0.324	0.136	0.146	0.171	0.118	0.106	1.000

have $x_i = a$, which is denoted by $f(y = b|x = a)$:

$$f(y = b|x = a) \equiv \frac{N_{ab}}{N_a} = \frac{\frac{N_{ab}}{N}}{\frac{N_a}{N}} = \frac{f_{ab}}{f_a}. \tag{20}$$

In our example, we could be interested in comparing the distribution of income for individuals with no assets to the distribution of income for individuals with more than 60,000 euros in liquid assets.

V. Conditional Sample Means

Restricting the sample to observations with $x_i = x$, 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 is:

$$\bar{y}_{|x=a} \equiv \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}\{x_i = a\} \times f(y_i|x_i = a) \times y_i, \tag{21}$$

Figure 2. Joint Distribution of Income and Liquid Assets (2,442 Households)

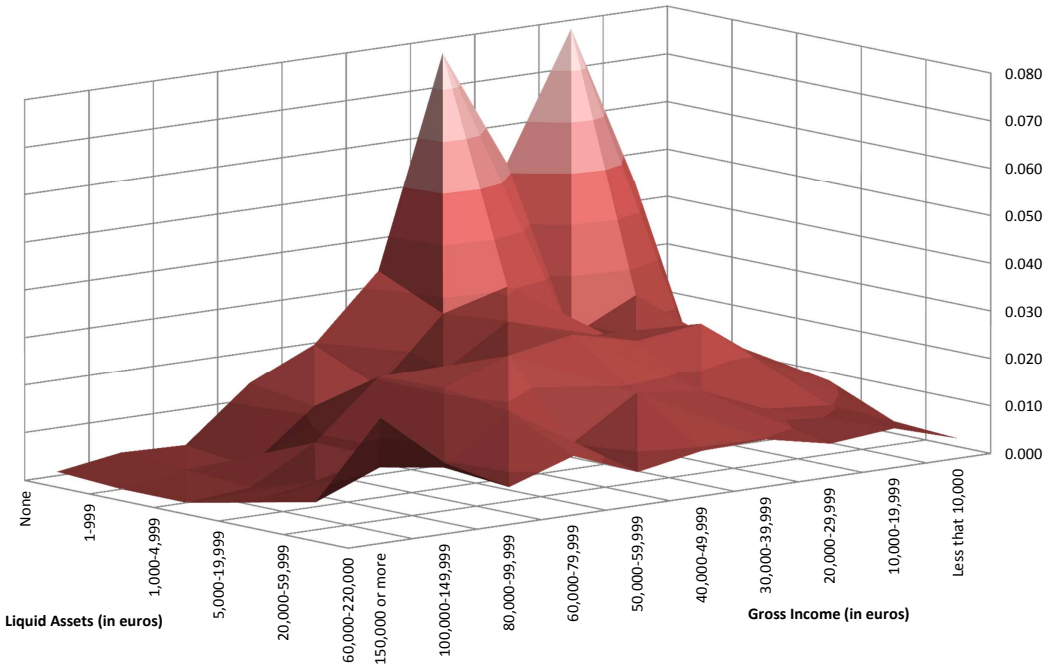


Table 4—Conditional Means of Income by Level of Assets (in Euros)

Liquid assets:	Mean gross income:
None	29,829
1-999	37,145
1,000-4,999	43,165
5,000-19,999	46,906
20,000-59,999	60,714
60,000-220,000	94,981
Unconditional	46,253

where $\mathbb{1}\{\cdot\}$ is the indicator function that equals one if the argument is true, and zero otherwise. Table 4 shows the conditional means of gross income for each level of liquid assets in our example.

All this assumes that the data is either discrete, or grouped in discrete intervals. However, grouping data for a continuous variable in intervals can be problematic. If intervals are too wide, we might be loosing relevant variation, but if they are two thin, we will be computing our statistics with very few observations, and we can even have empty cells (course of dimensionality). Thus, we might be interested in analyzing the data without grouping them in intervals.

To compute the conditional mean of y given x without discretizing x we can

use a kernel function. The intuition is that we compute the mean of y_i for the observations with $x_i = x$, but also for other observations that have x_i that are close to x, giving to those a lower weight, based on how far they are. More formally, we can write the conditional mean as:

$$\bar{y}_{|x=a} = \frac{1}{\sum_{\mathrm{i=1}}^{\mathrm{N}} \kappa\left(\frac{x_i - a}{\gamma}\right)} \sum_{\mathrm{i=1}}^{\mathrm{N}} y_i \times \kappa\left(\frac{x_i - a}{\gamma}\right), \tag{22}$$

where we use $\kappa\left(\frac{x_i - a}{\gamma}\right)$ as a weight, and the ratio outside of the sum is a normalization such that the weights sum to one. Using the kernel function defined in Equation (4), the resulting conditional mean would match Equation (21) exactly.

VI. Sample Covariance and Correlation

The final set of descriptive statistics presented in this chapter includes two measures that provide information on the co-movements of two variables. Importantly, 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 covariance, which is the average of the product of deviations of each variable with respect to its sample mean. More formally, the covariance is defined as:

$$s_{xy} \equiv \sum_{\mathrm{i=1}}^{\mathrm{N}} w_i (x_i - \bar{x})(y_i - \bar{y}). \tag{23}$$

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 correlation indicates that deviations are more commonly of opposite sign.

One of the main problems of the covariance is that its magnitude is not easy to interpret. Alternatively, the ***correlation coefficient*** is a statistic whose magnitude indicates the strength of the linear relation. The correlation coefficient is defined as:

$$r_{xy} \equiv \frac{s_{xy}}{s_y s_x}, \tag{24}$$

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.