Chapter 1: Descriptive Statistics

Joan Llull

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

joan.llull [at] movebarcelona [dot] eu

Introduction

Descriptive statistics is the discipline of qualitatively describing the main features of some data.

Three types of data:

- Cross-sectional
- Time series
- Panel data

Two types of variables:

- Discrete (ordinal, cardinal, or categorical)
- Continuous (can be treated as discrete if grouped in intervals)

Frequency Distributions

Examples of kernels

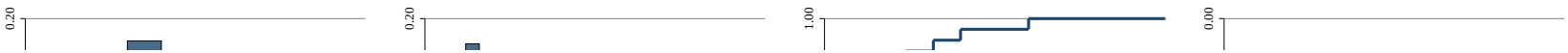
We build on a simple example: data for 2,442 **households** with information on **household gross income** in year 2010.

Table : Income Distribution (in Euros, 2,442 Households)

	Absolute frequency	Relative frequency	Cumul. frequency	Band-width	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 : Income Distribution (in Euros, 2,442 Households)

A. *Relative frequency* B. *Histogram* C. *Cumul. frequency* D. *Kernel density*



Kernel function

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

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

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

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.

General conditions for kernels

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

- is symmetric,
- and 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:

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

Uniform kernel:

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

Gaussian kernel:

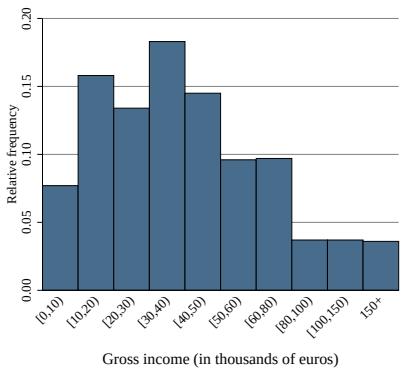
$$\kappa(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

Example

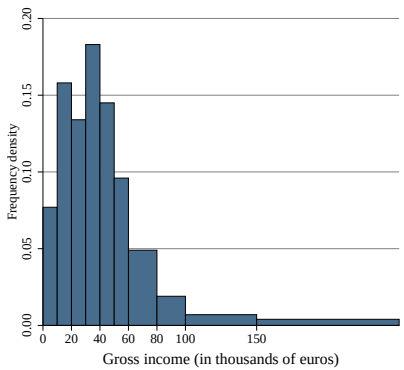
We build on a simple example: data for 2,442 **households** with information on **household gross income** in year 2010.

Figure : Income Distribution (in Euros, 2,442 Households)

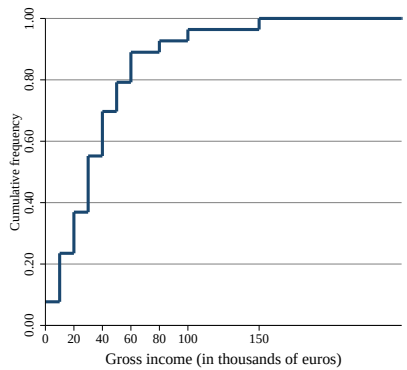
A. *Relative frequency*



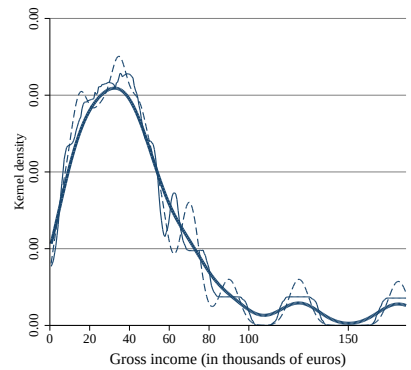
B. *Histogram*



C. *Cumul. frequency*



D. *Kernel density*



Summary Statistics

Arithmetic mean

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.

The **arithmetic mean**, also known as average, sample mean, or, when the context is clear, simply the mean, is defined as:

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

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

Main problem: It is sensitive to extreme observations.

Median and mode

The **median** is value for the observation that separates the higher half of the data from the lower half:

$$\text{med}(x) \equiv \min \left\{ a : c_a \geq \frac{1}{2} \right\}$$

Main advantage: it is not sensitive to extreme values.

Main inconvenient: changes in the tails are not reflected.

The **mode** is the value with the highest frequency:

$$\text{mode}(x) \equiv \left\{ a : f_a \geq \max_{j \neq a} f_j \right\}$$

Mean and median as loss minimizers

Loss function: is a function $L(\cdot)$ that 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 $0 < u < v$.

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

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

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

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

Sample variance and standard deviation

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:

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

The **standard deviation** is $s \equiv \sqrt{s^2}$.

The variance and the standard deviation are not easy to interpret. $\Rightarrow$ **coefficient of variation**:

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

Central moments

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

The (sample) **central moment** of order k, denoted by m_k, is defined as:

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

The **0th to 2nd moments** are: $m_0 = 1$, $m_1 = 0$, and $m_2 = s^2$.

Third moment $\Rightarrow$ **skewness coefficient**:

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

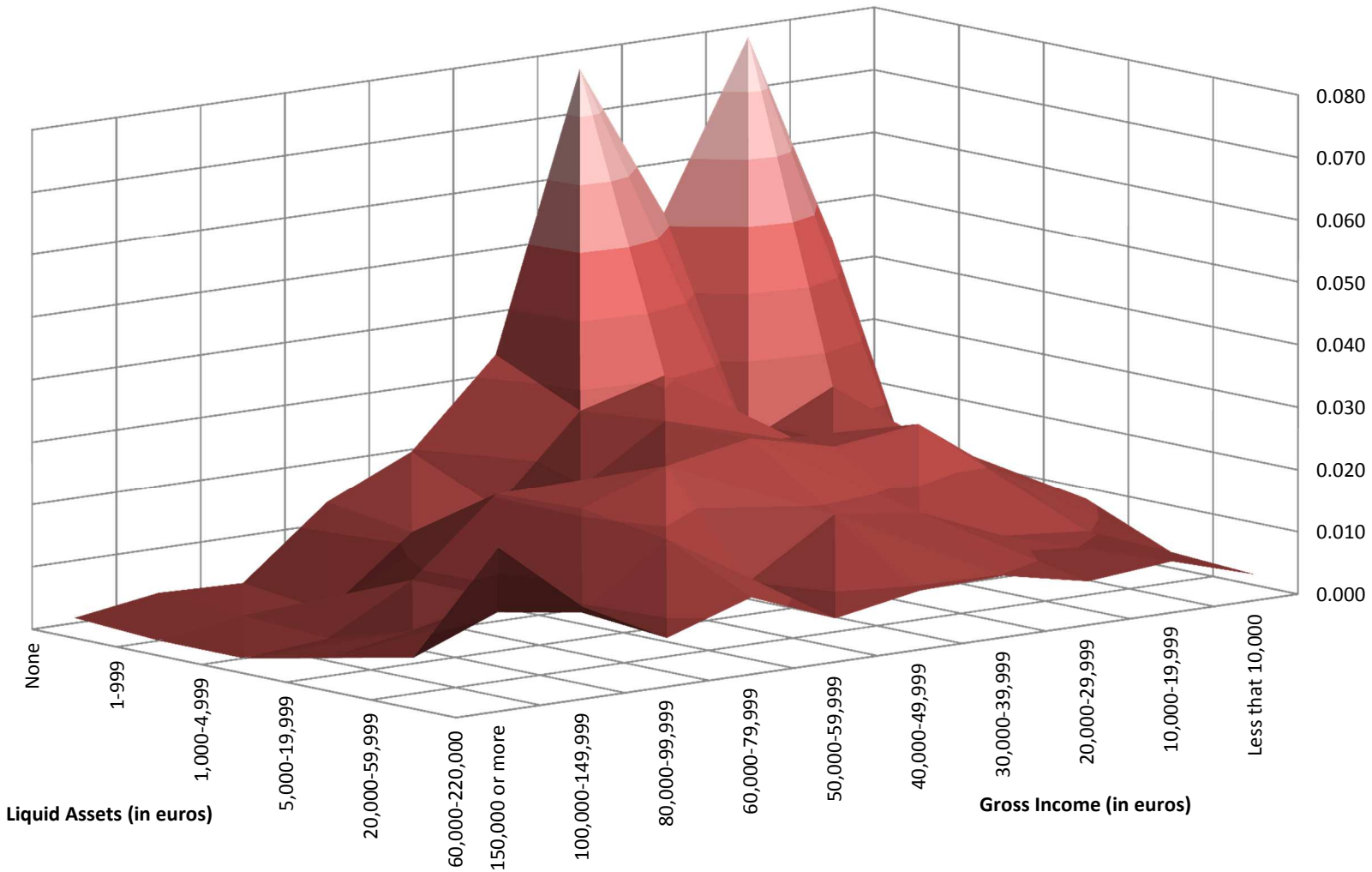
Fourth moment $\Rightarrow$ (excess) **kurtosis coefficient**:

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

Bivariate Frequency Distributions

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

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



Conditional relative frequencies

On top of absolute and relative **joint** frequencies, we can be interested in computing **conditional** relative frequencies.

The **conditional relative frequency** is computed as:

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

Conditional Sample Means

Conditional sample mean

The **conditional sample mean** is given by:

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

where $\mathbb{1}\{\cdot\}$ is the indicator function that equals one if the argument is true, and zero otherwise.

In our example:

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

Kernel function

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

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

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

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

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.

Sample Covariance and Correlation

Sample variance and correlation

Finally, we introduce two measures that provide information on the (linear) **co-movements** of two variables.

The **sample covariance** is defined as:

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

Signs contain information, but **magnitudes** are hard to interpret.

The **correlation coefficient** is:

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

and it ranges between -1 and 1, and the magnitude is interpretable. A value of 0 indicates that the two variables are (linearly) uncorrelated.