Probability and Statistics
Course Outline

By Joan Llull[*]

QEM Erasmus Mundus Master. Fall 2016

1. Descriptive Statistics

Frequency distributions. Summary statistics. Bivariate frequency distributions. Conditional sample means. Sample covariance and correlation.

2. Random Variables and Probability Distributions

Preliminaries: an introduction to set theory. Statistical inference, random experiments, and probabilities. Finite sample spaces and combinatorial analysis. Definition of random variable and cumulative density function. Continuous and discrete random variables. Commonly used univariate distributions. Transformations of random variables. Expectation and moments. Quantiles, the median, and the mode.

3. Multivariate Random Variables

Joint, marginal, and conditional distributions. Independence. Functions of random variables. Multivariate normal distribution. Covariance, correlation, and conditional expectation. Linear prediction.

4. Sample Theory and Sample Distributions

Random samples. Sample mean and variance. Sampling from a normal population: χ^2, t, and F distributions. Bivariate and multivariate random samples. Heterogeneous and correlated samples.

5. Estimation

Analogy principle. Desirable properties of an estimator. Moments and likelihood problems. Maximum likelihood estimation. The Cramer-Rao lower bound. Bayesian inference.

6. Regression

Classical regression model. Statistical results and interpretation. Nonparametric regression.

[*] 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.

7. Hypothesis Testing and Confidence Intervals

Hypothesis testing. Type I and type II errors. The power function. Likelihood ratio test. Confidence intervals. Hypothesis testing in a normal linear regression model.

8. Asymptotic Theory

The concept of stochastic convergence. Laws of large numbers and central limit theorems. Delta method. Consistency and asymptotic normality of ordinary least squares and maximum likelihood estimators. Asymptotic efficiency. Bootstrap.

References

Main references:

Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes (1974), *Introduction to the Theory of Statistics*, McGraw-Hill, Third Edition, 1974.

Lindgren, Bernard W. (1998), *Statistical Theory*, Fourth Edition, Chapman & Hall/CRC, 1993.

Goldberger, Arthur S. (1991), *A Course in Econometrics*, Harvard University Press, 1991.

Additional reference:

Greene, William E. (2003), *Econometric Analysis*, Prentice-Hall, Fifth Edition, 2003.

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

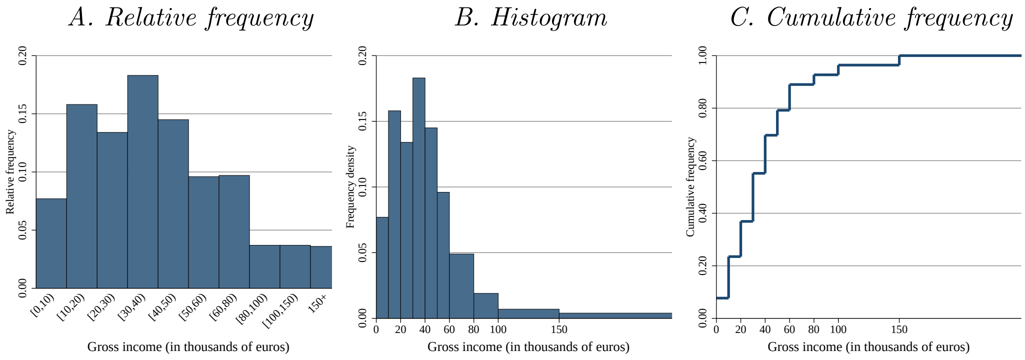
In this chapter, we build on a simple example to introduce the main notions that we are after. Consider a dataset of 2,442 households with information on household gross income in year 2010 for each of them. In Table 1 we describe the distribution of this variable in different ways. This variable is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



A. Relative frequency

B. Histogram

C. Cumulative frequency

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.

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:

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

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:

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

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 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.

[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

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

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

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

	Liquid assets (in euros):						
Gross Income (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

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:

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

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.

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

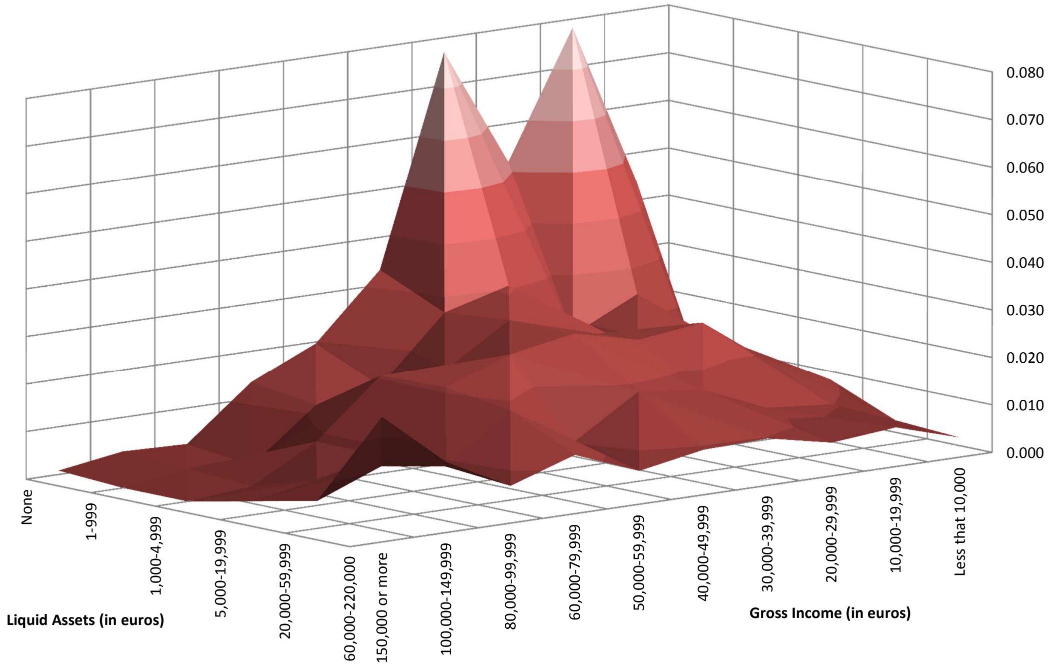


Table 4—Conditional Means of Income by Level of Assets (in Euros)

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

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.

Chapter 2: Random Variables and Probability Distributions

By Joan Llull[*]

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

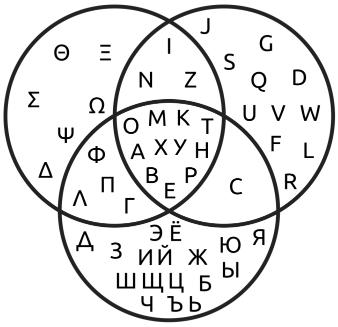
Main references:

— Mood: I.3.2-5; II; III.2.1, III.2.4, III.3.1-2; V.5.1; Appendix A.2.2, A.2.4

— Lindgren: 1.2; 3.1 to 3.3, 3.5; 4.1 to 4.6, 4.9, 4.10; 6.1, 6.5, 6.8; (3.2)

I. Preliminaries: An Introduction to Set Theory

We start this chapter with the introduction of some tools that we are going to use throughout this course (and you will use in subsequent courses). First, we introduce some definitions, and then describe some operators and properties of these operators. Consider a collection of objects, including all objects under consideration in a given discussion. Each object in our collection is an ***element*** or a point. The totality of all these elements is called the ***space***, also known as the universe, or the universal set, and is denoted by Ω. We denote an element of the set Ω by ω. For example, a set can be all the citizens of a country, or all the points in a plane (i.e. $\Omega = \mathbb{R}^2$, and $\omega = (x, y)$ for any pair of real numbers x and y). A partition of the space Ω is called a ***set***, which we denote by calligraphic capital Latin letters, with or without subscripts. When we opt for the second, we define the catalog of all possible incides as the ***index set***, which we denote by Λ (for example, if we consider the sets $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, then $\Lambda = \{1, 2, 3\}$). A ***Venn diagram*** is a diagram that shows all possible logical relations between a finite collection of sets. For example, we would represent the sets of capital letters in the Latin, Greek and Cyrillic scripts as:



[*] 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.

To express that an element ω is part of a set $\mathcal{A}$, we write $w \in \mathcal{A}$, and to state the opposite, we write $w \notin \mathcal{A}$. We can define sets by explicitly specifying all its elements (e.g. $\mathcal{A} = \{1, 2, 3, 4, 5, 6\}$), or implicitly, by specifying properties that describe its elements (e.g. $\mathcal{A} = \{(x, y) : x \in \mathrm{I\!R}, y \in \mathrm{I\!R}^+\}$). The set that includes no elements is called the ***empty set***, and is denoted by $\varnothing$.

Now we define a list of operators for sets:

- ***Subset***: when all elements of a set $\mathcal{A}$ are also elements of a set $\mathcal{B}$ we say that $\mathcal{A}$ is a subset of $\mathcal{B}$, denoted by $\mathcal{A} \subset \mathcal{B}$ ("$\mathcal{A}$ is contained in $\mathcal{B}$") or $\mathcal{B} \supset \mathcal{A}$ ("$\mathcal{B}$ contains $\mathcal{A}$").

- ***Equivalent set:*** two sets $\mathcal{A}$ and $\mathcal{B}$ are equivalent or equal, denoted $\mathcal{A} = \mathcal{B}$ if $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{A}$.

- ***Union***: the set that consists of all points that are either in $\mathcal{A}$, in $\mathcal{B}$, or in both $\mathcal{A}$ and $\mathcal{B}$ is defined to be the union between $\mathcal{A}$ and $\mathcal{B}$, and is denoted by $\mathcal{A} \cup \mathcal{B}$. More generally, let Λ be an index set, and $\{\mathcal{A}_\lambda\} \equiv \{A_\lambda : \lambda \in \Lambda\}$, a collection of subsets of Ω indexed by Λ. The set that consists of all points that belong to $\mathcal{A}_\lambda$ for at least one $\lambda \in \Lambda$ is called the union of the sets $\{\mathcal{A}_\lambda\}$, denoted by $\underset{\lambda \in \Lambda}{\cup} \mathcal{A}_\lambda$. If $\Lambda = \varnothing$, we define $\underset{\lambda \in \varnothing}{\cup} \mathcal{A}_\lambda \equiv \varnothing$.

- ***Intersection***: the set that consists of all points that are both in $\mathcal{A}$ and in $\mathcal{B}$ is defined to be the intersection between $\mathcal{A}$ and $\mathcal{B}$, and is written $\mathcal{A} \cap \mathcal{B}$ or $\mathcal{A}\mathcal{B}$. More generally, with the notation from the previous point, the set that consists of all points that belong to $\mathcal{A}_\lambda$ for every $\lambda \in \Lambda$ is called the intersection of the sets $\{\mathcal{A}_\lambda\}$, and is denoted by $\underset{\lambda \in \Lambda}{\cap} \mathcal{A}_\lambda$. If $\Lambda = \varnothing$, we define $\underset{\lambda \in \varnothing}{\cap} \mathcal{A}_\lambda \equiv \Omega$.

- ***Set difference***: the set that consists of all points in $\mathcal{A}$ that are not in $\mathcal{B}$ is defined to be the set difference between $\mathcal{A}$ and $\mathcal{B}$, and is denoted by $\mathcal{A} \setminus \mathcal{B}$ (or $\mathcal{A} - \mathcal{B}$ when the context is clear).

- ***Complement***: the complement of a set $\mathcal{A}$ with respect to the space Ω, denoted by $\mathcal{A}^c$ (or $\overline{\mathcal{A}}$) is the set that consists of all points that are in the space Ω and are not in $\mathcal{A}$, that is $\Omega \setminus \mathcal{A}$.

- ***Disjoint/mutually exclusive sets***: $\mathcal{A} \subset \Omega$ and $\mathcal{B} \subset \Omega$ are defined to be mutually exclusive or disjoint if $\mathcal{A} \cap \mathcal{B} = \varnothing$. Subsets $\{\mathcal{A}_\lambda\}$ are defined to be mutually exclusive is $\mathcal{A}_\lambda \cap \mathcal{A}_{\lambda'} = \varnothing$ for every λ and λ' such that $\lambda \neq \lambda'$.

- ***Cartesian product***: the set of all possible ordered pairs (a, b) where $a \in \mathcal{A}$

and $b \in \mathcal{B}$ is defined to be the Cartesian product of $\mathcal{A}$ and $\mathcal{B}$, and is denoted by $\mathcal{A} \times \mathcal{B}$.

- ***Power set***: the power set of a set $\mathcal{A}$, denoted by $2^{\mathcal{A}}$ (or $\mathcal{P}(\mathcal{A})$), is the set of all possible subsets of $\mathcal{A}$, including the empty set $\varnothing$, and $\mathcal{A}$ itself. For example, if $\mathcal{A} = \{x, y, z\}$, $2^{\mathcal{A}} = \{\varnothing, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$. If $\mathcal{A}$ includes n elements, $2^{\mathcal{A}}$ includes 2^n elements (hence its notation).

- ***Finite and countable sets***: a finite set is a set that has a finite number of elements (and an infinite set is a set with an infinite number of elements). A countable set is a set with the same number of elements as some subset of the set of natural numbers (can be finite or infinite).

- ***Sigma-algebra***: a sigma algebra (or σ-algebra), Σ, on a set $\mathcal{A}$ is a subset of the power set of $\mathcal{A}$, $\Sigma \subset 2^{\mathcal{A}}$, that satisfies three properties: i) it includes $\mathcal{A}$; ii) if the subset $\mathcal{B} \subset \mathcal{A}$ is included in Σ, $\mathcal{B}^c$ is also included; and iii) if a countable collection of subsets $\{\mathcal{A}_\lambda\}$ is included, its union $\underset{\lambda \in \Lambda}{\cup} \mathcal{A}_\lambda$ is also included.

Next we list some properties of the operators defined above. Some of the proofs will be done in class, others will be listed as exercises, and others are recommended to be done at own initiative. The properties are:

- ***Commutative laws***: $\mathcal{A} \cup \mathcal{B} = \mathcal{B} \cup \mathcal{A}$, and $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$.

- ***Associative laws***: $\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cup \mathcal{C}$, and $\mathcal{A} \cap (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cap \mathcal{C}$

- ***Distributive laws***: $\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C})$, and $\mathcal{A} \cup (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C})$.

- $\mathcal{A} \cap \Omega = \mathcal{A}$; $\mathcal{A} \cup \Omega = \Omega$; $\mathcal{A} \cap \varnothing = \varnothing$; $\mathcal{A} \cup \varnothing = \mathcal{A}$.

- $\mathcal{A} \cap \mathcal{A}^c = \varnothing$; $\mathcal{A} \cup \mathcal{A}^c = \Omega$; $\mathcal{A} \cap \mathcal{A} = \mathcal{A} \cup \mathcal{A} = \mathcal{A}$.

- $(\mathcal{A}^c)^c = \mathcal{A}$.

- ***DeMorgan's laws***: $(\mathcal{A} \cup \mathcal{B})^c = \mathcal{A}^c \cap \mathcal{B}^c$ and $(\mathcal{A} \cap \mathcal{B})^c = \mathcal{A}^c \cup \mathcal{B}^c$. Likewise, $\big(\underset{\lambda \in \Lambda}{\cup} \mathcal{A}_\lambda\big)^c = \underset{\lambda \in \Lambda}{\cap} \mathcal{A}_\lambda^c$, and $\big(\underset{\lambda \in \Lambda}{\cap} \mathcal{A}_\lambda\big)^c = \underset{\lambda \in \Lambda}{\cup} \mathcal{A}_\lambda^c$.

- $\mathcal{A} \setminus \mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$ and $(\mathcal{A} \setminus \mathcal{B})^c = \mathcal{A}^c \cup \mathcal{B}$.

- $(\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{B}^c) = \mathcal{A}$ and $(\mathcal{A} \cap \mathcal{B}) \cap (\mathcal{A} \cap \mathcal{B}^c) = \varnothing$.

- $\mathcal{A} \subset \mathcal{B} \Rightarrow \mathcal{A} \cap \mathcal{B} = \mathcal{A}$ and $\mathcal{A} \cup \mathcal{B} = \mathcal{B}$.

- $\mathcal{A} \times \mathcal{B} \neq \mathcal{B} \times \mathcal{A}$.

II. Statistical Inference, Random Experiments, and Probabilities

In Chapter 1 we started from a data sample, and we learned how to describe these data. Now we want to make general claims about the population the sample is meant to represent. These general claims are known as ***statistical inference***.

In order to do inference, we need a statistical model, this is, a data generating process. A ***random experiment*** or trial is a conceptual description of the process that generated the data that we observe. We call it random because, even though the process can be replicated under similar conditions, results are not known with certainty (there is more than one possible outcome). This is unlike a deterministic experiment, which has only one possible outcome. For example, a random experiment could be to roll a dice or toss a coin; the outcome of this experiment is random because if we repeat it (roll the dice again, or toss the coin again) we are not necessarily obtain the same result.

The ***probability space*** is a mathematical construct that formalizes a random experiment. The probability space consists of three parts:

- A ***sample space*** Ω, which is the set of all possible outcomes.
- A σ-***algebra***, $\mathcal{F} \subset 2^{\Omega}$, which is a set of events, $\mathcal{F} = \{\mathcal{A}_1, \mathcal{A}_2, ...\}$, where each event $\mathcal{A}_\lambda \subset \Omega$ is a subset of Ω that contains zero or more outcomes. An event $\mathcal{A}_\lambda$ is said to occur if the experiment at hand results in an outcome ω that belongs to $\mathcal{A}_\lambda$.
- A ***probability measure*** $P : \mathcal{F} \to [0, 1]$, which is a function on $\mathcal{F}$ that satisfies three axioms:

  1) $P(\mathcal{A}) \geq 0$ for every $\mathcal{A} \in \mathcal{F}$.
  2) $P(\Omega) = 1$ (Ω is sometimes called the sure event)
  3) If $\mathcal{A}_1, \mathcal{A}_2, ...$ is a sequence of mutually exclusive events in $\mathcal{F}$, then $P(\cup_{\lambda=1}^{\infty} \mathcal{A}_\lambda) = \sum_{\lambda=1}^{\infty} P(\mathcal{A}_\lambda)$.

  The third axiom has two implications. The first is that $P(\varnothing) = 0$. The second is that if we partition the space Ω in mutually exclusive events $\mathcal{A}_1, \mathcal{A}_2, ...$ the sum of the probabilities of all mutually exclusive events that form Ω equals 1.

For example, tossing a coin is a random experiment. The probability space of this experiment is as follows. The sample space is a set that includes heads and tails: $\Omega = \{head, tail\}$. The σ-algebra is the set of all possible combinations of outcomes, which is $\{\{\varnothing\}, \{head\}, \{tail\}, \{\Omega\}\}$. An event is an element of the σ-algebra, i.e. one of the four listed subsets. And the probability measure is the function that transforms this event into a probability, expressed between 0 and 1. The probability measure satisfies the three axioms: the probability of each of the four events is greater or equal than zero, it is sure that the outcome of the experiment will be either heads or tails, and if we consider the three mutually exclusive sets of the σ-algebra $\{\varnothing\}$, $\{heads\}$, and $\{tails\}$, the probability of the union of the three (which, since the union of the three is Ω, is equal to 1) is equal to the sum of the probabilities of the three events.

III. Finite Sample Spaces and Combinatorial Analysis

In the previous section, we generalized the concept of probability for any sample space. Now we focus on a particular type of sample spaces: those with a finite number of points, $\Omega = \{\omega_1, ..., \omega_N\}$. Let the operator $N(\mathcal{A})$ denote the number of elements of a finite set $\mathcal{A}$. We define $N \equiv N(\Omega)$ as the total number of possible outcomes of a random experiment with a finite number of outcomes.

Initially, consider the case of a finite sample space with ***equally likely*** points. In this case, the probability of each outcome is $1/N$. However, we can also implement the axiomatic definition of probability introduced in the previous section. Define a probability function $P(\cdot)$ over a finite sample space that satisfies two properties:

- $P(\{\omega_1\}) = P(\{\omega_2\}) = ... = P(\{\omega_N\})$.
- If $\mathcal{A} \subset \Omega$ includes $N(\mathcal{A})$ elements, then $P(\mathcal{A}) = N(\mathcal{A})/N$.

We shall call such function an equally likely probability function. It is trivial to check that an equally likely probability function satisfies the three axioms and hence is a probability function.

In this environment, the only problem left in determining the probability of a given event is a problem of counting: count the number of points in $\mathcal{A}$, $N(\mathcal{A})$ and the number of points in Ω, N. For example consider the experiment of tossing a coin twice. Let $\Omega = \{head, tail\} \times \{head, tail\} = \{(z_1, z_2) : z_1 \in \{head, tail\}, z_2 \in \{heads, tail\}\}$. There are $N = 2 \cdot 2 = 4$ sample points. It seems reasonable to attach a probability of $\frac{1}{4}$ to each point. Let $\mathcal{A} = \{\text{at least one head}\}$. Then, $\mathcal{A} = \{(head, tail), (tail, head), (head, head)\}$, and $P(\mathcal{A}) = \frac{3}{4}$.

In this example, counting was quite simple. However, in higher dimensional cases, we may need to count in a systematic way. For that purpose, it is useful to introduce an important tool: ***combinatorial analysis***. Specifically, let us introduce the following definitions:

- n **factorial**: a product of a positive integer n by all the positive integers smaller than it $[n! \equiv n(n-1)(n-2)...1 = \prod_{j=0}^{n-1}(n-j)]$. We define $0! \equiv 1$.
- $(n)_k$: a product of a positive integer n by the next $k-1$ smaller positive integers $[(n)_k \equiv n(n-1)...(n-k+1) = \prod_{j=0}^{k-1}(n-j) = \frac{n!}{(n-k)!}]$.
- **Combinatorial symbol (or n pick k)**: it is defined as:

$$\binom{n}{k} \equiv \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!}, \quad \text{with} \quad \binom{n}{k} \equiv 0 \text{ if } k < 0 \text{ or } k > n. \tag{1}$$

- **Binomial theorem**: the binomial theorem states that:

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}. \tag{2}$$

Consider an experiment such that each outcome can be represented as an n-tuple (as in our example before, where we expressed each outcome as a 2-tuple. Another example (that will be central in Chapter 4) is drawing a *sample* of size n from an *urn* with M *balls*, numbered from 1 to M. There are two basic ways of drawing a sample: *with replacement* and *without replacement.* In the case of sampling with replacement, the sample space is $\Omega = \{(z_1, ..., z_n) : z_1 \in \{1, ..., M\}, ..., z_n \in \{1, ..., M\}\}$, and in the case of sampling without replacement, it is $\Omega = \{(z_1, ..., z_n) : z_1 \in \{1, ..., M\}, z_2 \in \{1, ..., M\} \backslash \{z_1\} ..., z_n \in \{1, ..., M\} \backslash \{z_1, ..., z_{n-1}\}\}$.

As a general rule, counting the number of elements of a set $\mathcal{A}$ composed of points that are n-tuples satisfying certain conditions consists of determining the number of points that may be used as each of the n components, say $N_1, ..., N_n$, and this way obtain $N(\mathcal{A}) = N_1 \cdot N_2 \cdot ... \cdot N_n$. Thus, in the case of sampling with replacement, M^n different samples could possibly be drawn. In the case without replacement, $N(\mathcal{A}) = M \cdot (M-1) \cdot ... \cdot (M-n+1) = (M)_n$.

To determine the size of the power set of a finite sample space with M elements, we can also use combinatorial analysis. For every subset of Ω that contains n elements, we can create $n!$ different (note that, for instance, $\{1, 2, 3\}$ is not different than $\{2, 3, 1\}$, since both contain the same three objects) combinations of n elements, drawing from the set without replacement. Denote the number of different sets of size n can be formed off Ω by x_n. Now, because we know from the previous paragraph that we can draw $(M)_n$ different size n samples from Ω, then $n!x_n = (M)_n$, which implies that $x_n = \frac{(M)_n}{n!} = \binom{M}{n}$. Therefore, the total number of sets that can be formed off Ω is $\sum_{n=0}^{M} \binom{M}{n}$. Thus, using the binomial theorem for $a = b = 1$, we see that $N(2^\Omega) = 2^M$.

We can also consider finite sample spaces ***without equally likely points***. In this case, we have to define our probability function in a different way. We can completely define values for $P(\mathcal{F})$ for each of the $2^{N(\Omega)}$ events by specifying a value of $P(\cdot)$ for each of the $N = N(\Omega)$ elementary elements. Let $\Omega = \{\omega_1, ..., \omega_N\}$, and define $p_j \equiv P(\{\omega_j\})$ for $j = 1, ..., N$. To satisfy the second and third axioms of the probability function, p_j for $j = 1, ..., N$ need to be such that $\sum_{\mathrm{j=1}}^{\mathrm{N}} p_j = 1$, since:

$$\sum_{\mathrm{j=1}}^{\mathrm{N}} p_j = \sum_{\mathrm{j=1}}^{\mathrm{N}} P(\{\omega_j\}) = P\left(\bigcup_{j=1}^{N} \{\omega_j\}\right) = P(\Omega) = 1. \tag{3}$$

For any event $\mathcal{A}$, define $P(\mathcal{A}) \equiv \sum_{\{j:\omega_j \in \mathcal{A}\}} p_j$. It is easy to prove that this function also satisfies the three axioms, and hence is a probability function.

IV. Definition of Random Variable and Cumulative Density Function

A ***random variable***, denoted by $X : \Omega \to \mathbb{R}$, is a function from Ω to the real line such that the set $\mathcal{A}_r$, defined by $\mathcal{A}_r \equiv \{\omega : X(\omega) \leq r\}$, belongs to $\mathcal{F}$ for every real number r. What is important from this definition is that a random variable is a transformation of an event into a numeric value.

In our example of tossing the coin, the number of heads obtained is a random variable because i) it is a transformation of elements of Ω into real numbers (e.g. $X(head) = 1$ and $X(tail) = 0$), and ii) the indicated condition is satisfied for any $r \in \mathbb{R}$: if $r < 0$, $\{\omega : X(\omega) \leq r\} = \varnothing$, if $r \geq 1$, $\{\omega : X(\omega) \leq r\} = \Omega$, and if $0 \leq r < 1$ $\{\omega : X(\omega) \leq r\} = \{tail\}$, and all three belong to $\mathcal{F}$.

A random variable is represented by its ***cumulative distribution function*** (cdf), denoted by F_X, which transforms real numbers into probabilities as follows:

$$F_X : \mathbb{R} \to [0, 1], \quad F_X(x) \equiv P(X \leq x). \tag{4}$$

This concept is analogous to the cumulative frequency that we defined in Chapter 1. In the coin tossing example, the cdf is as follows:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{2}, & \text{if } x \in [0, 1) \\ 1, & \text{if } x \geq 1 \end{cases}. \tag{5}$$

Notice that the main reason why we impose the mathematical condition $\mathcal{A}_r \subset \mathcal{F}$ for all $r \in \mathbb{R}$ is to ensure that the cdf of the random variable is well defined over the entire real line.

A cdf satisfies the following properties:

- In the limit, $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
- F_X is nondecreasing (because if $x_1 < x_2$ then $\{\omega : X(\omega) \leq x_1\} \subseteq \{\omega : X(\omega) \leq x_2\}$).
- F_x is continuous from the right (not necessarily from the left, as in the coin tossing example).

V. Continuous and Discrete Random Variables

We say that a random variable is ***discrete*** if its support includes a finite (or countably infinite) number of points of support. The cdf of a discrete random variable is a step function, with the discrete jumps occurring at the points of support. The cdf is fully characterized by the ***probability mass function*** (pmf), which is defined as $P(X = x_a)$, since:

$$F_X(x) \equiv \sum_{\{a : x_a \leq x\}} P(X = x_a). \tag{6}$$

Note that the concept of pmf is closely tied to the relative frequency defined in Chapter 1.

Analogously, we define a random variable as ***continuous*** if there exists a nonnegative function $f_X(\cdot)$ such that:

$$F_X(x) = \int_{-\infty}^{x} f_X(z)dz, \quad \forall x \in \mathbb{R}\,. \tag{7}$$

The function $f_X(\cdot)$ is known as ***probability density function*** (pdf). The pdf indicates the rate at which the probability is accumulated in the neighborhood of a point, which is also connected to the relative frequency explained in Chapter 1. This is easy to see using the definition of derivative (as, from Equation (7), if F_X is differentiable, the pdf is the derivative of the cdf at a given point):

$$f_X(x) = \lim_{h \to 0} \frac{F_X(x+h) - F_X(x)}{h}. \tag{8}$$

Continuous random variables (and their pdfs and cdfs) satisfy the following:

- $f_X(x) \geq 0$ in all the support where $F_X(x)$ is differentiable.
- $\int_{-\infty}^{\infty} f_X(z)dz = 1$, even though nothing prevents $f_X(x) > 1$ at some point x.
- F_X is continuous (from both sides).
- $P(X = x) = 0$ for all x in (and out of) the support of X.

- $P(x_1 < X < x_2) = \int_{x_1}^{x_2} f_X(z)dz$.
- $f_X(x) = \frac{d}{dx}F_X$ at all points where F_X is differentiable.

A random variable can also be ***mixed***: it is continuous in a part of its domain, but also has some points at which there is positive probability mass. More formally, a random variable is mixed if its cdf is of the form:

$$F_X(x) = pF_X^{(d)}(x) + (1-p)F_X^{(c)}(x), \quad 0 < p < 1, \tag{9}$$

where $F_X^{(d)}(\cdot)$ is the cdf of the discrete part, and $F^{(c)X(\cdot)}$ is the cdf of the continuous part. This type of cdf, formed as a convex combination of cdfs of continuous and discrete random variables is called a ***mixture***.

VI. Commonly Used Univariate Distributions

In this section we introduce a set of widely used discrete and continuous parametric families of distributions. In the problem sets throughout the course you may see additional distributions that are also commonly used. In any of the listed manuals, you can find a more extensive list of distributions.

The ***Bernoulli distribution*** is a discrete distribution with pmf given by:

$$f_X(x) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x \in \{0,1\} \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

where the parameter p satisfies $0 \leq p \leq 1$.

The ***binomial distribution*** is a discrete distribution function with pmf given by:

$$f_X(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & \text{for } x = 0,1,...,n \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

where $0 \leq p \leq 1$, and n ranges over the positive integers.

The ***Poisson distribution*** is a discrete distribution with pmf given by:

$$f_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{if } x = 0,1,2,... \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where the parameter λ satisfies $\lambda > 0$.

The ***uniform distribution*** is a continuous distribution (there is a discrete version of it) with pdf given by:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

where a and b are the inferior and superior limits of the support, and with cdf given by:

$$F_X(x) = \begin{cases} 0 & \text{if } x \in (-\infty, a) \\ \dfrac{x-a}{b-a} & \text{if } x \in [a, b] \\ 1 & \text{if } x \in (b, \infty) \end{cases} . \tag{14}$$

If X is uniformly distributed, we denote $X \sim \mathcal{U}(a, b)$.

The ***standard normal distribution*** is a continuous distribution with pdf given by:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \tag{15}$$

and cdf given by:

$$\Phi(x) = \int_{-\infty}^{x} \phi(z) dz. \tag{16}$$

Its pdf is symmetric around 0, its only maximum is at $x = 0$, and it has two inflection points at ±1. The indication that a random variable X is distributed as a standard normal is denoted as $X \sim \mathcal{N}(0, 1)$. The cdf of the normal distribution does not have a closed form solution, but its values are tabulated, and incorporated in most statistical softwares (even in spreadsheets!).

The standard normal distribution can be generalized by means of an affine transformation. This transformation is simply called the ***normal distribution***, and is denoted by $\mathcal{N}(\mu, \sigma^2)$. More specifically, let $Z \sim \mathcal{N}(0, 1)$, and let $X \equiv \mu + \sigma Z$, with $\sigma > 0$; then $X \sim \mathcal{N}(\mu, \sigma^2)$. The cdf of the normal distribution is given by:

$$F_X(x) \equiv P(X \leq x) = P(\mu + \sigma Z \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right), \tag{17}$$

and its pdf is equal to:

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right). \tag{18}$$

In this case, $f_X(\cdot)$ is symmetric with respect to μ, its only maximum is at $x = \mu$, and it has two inflection points at $\pm\sigma$.

VII. Transformations of Random Variables

In this section we want to learn what is the distribution of $Y \equiv g(X)$, given that we know that $X \sim F_X(\cdot)$. For example, suppose that we roll a dice once, and our random variable X is the number of points we obtain. Let $Y = X^2 - 7X + 10$. In this example:

X	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
Y	4	0	-2	-2	0	4

and thus:

Y	-2	0	4
$P(Y = y)$	1/3	1/3	1/3

More formally, what we have done is the following:

$$P(Y = y) = \sum_{\{i: g(x_i) = y\}} P(X = x_i), \tag{19}$$

this is, we summed the probability mass of all values of the support of X that generate the same value for $g(X)$.

When X is continuous, assuming that $g(\cdot)$ is invertible and differentiable, and that $g'(\cdot) \neq 0$, the cdf of Y is given by:

$$F_Y(y) \equiv P(Y \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)), \tag{20}$$

and the pdf is obtained differentiating:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = f_X\left(g^{-1}(y)\right) \left| \frac{1}{g'\left[g^{-1}(y)\right]} \right|. \tag{21}$$

In words, the probability in Y is accumulated the same way as it accumulates for X (first term) times the rate at which X is transformed —regardless of sign— into Y (second term). If $g(\cdot)$ is not invertible, it is still possible to follow a similar procedure if the function can be divided in invertible pieces.

VIII. Expectation and Moments

The mathematical ***expectation*** of a random variable X, denoted by $\mathbb{E}[X]$, is defined as follows:

- if X is discrete: $\mathbb{E}[X] \equiv \sum_a x_a P(X = x_a)$,
- and if X is continuous: $\mathbb{E}[X] \equiv \int_{-\infty}^{\infty} x f_X(x) dx$.

Note the analogy with the sample mean described in Chapter 1. This is not coincidental, as the expectation is the population equivalent to the sample mean. The two expressions above can be unified using the ***Riemann-Stieltjes integral***:

$$\mathbb{E}[X] \equiv \int_{-\infty}^{\infty} x dF_X(x). \tag{22}$$

The ***variance*** of a random variable X, denoted by $\text{Var}(X)$, is the expected quadratic deviation with respect to the mean $\mu_X \equiv \mathbb{E}[X]$:

- if X is discrete: $\text{Var}(X) \equiv \sum_a [(x_a - \mu_X)^2 P(X = x_a)]$,
- and if X is continuous: $\text{Var}(X) \equiv \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$.

In general, using the Riemann-Stieltjes integral:

$$\text{Var}(X) \equiv \int_{-\infty}^{\infty} (x - \mu_X)^2 dF_X(x). \tag{23}$$

The variance is a measure of dispersion of the probability mass (or density) of X around its mean, and it is the population counterpart of the sample variance. It is always nonnegative. And we can define the ***standard deviation*** as:

$$\sigma_X \equiv +\sqrt{\text{Var}(X)}, \tag{24}$$

where the positive sign indicates that it is given by the positive root only. More generally, the ***k^{th} central moment*** of the distribution of X is defined as:

$$\mathbb{E}[(x - \mu_X)^k] \equiv \int_{-\infty}^{\infty} (x - \mu_X)^k dF_X(x). \tag{25}$$

Its interpretation is analogous to the sample moments described in Chapter 1, and we can normalize the third and fourth moments in the way described there to obtain the coefficients of ***skewness*** and ***kurtosis***. We similarly define the ***k^{th} uncentered moment*** as $\mathbb{E}[X^k]$.

The expectation (and analogously any moment) of a transformation of X, $Y \equiv g(X)$, can be calculated directly with a transformation of Equation (22), without a need of obtaining the cdf of Y first:

$$\mathbb{E}[Y] \equiv \int_{-\infty}^{\infty} y dF_Y(y) = \int_{-\infty}^{\infty} g(x) dF_X(x). \tag{26}$$

This implication is very useful to establish a set of ***general properties*** of the expectation (and the variance). Let c be a constant, and let $g(X)$ and $h(X)$ denote two arbitrary functions of the random variable X. Then:

- $\mathbb{E}[c] = c$,
- $\mathbb{E}[cX] = c\mathbb{E}[X]$,

- $\mathbb{E}[g(X) + h(X)] = \mathbb{E}[g(X)] + \mathbb{E}[h(X)]$,
- $\mathbb{E}[g(X)] \geq \mathbb{E}[h(X)]$ if $g(X) \geq h(X)$ for every possible value of X,

and:

- $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
- $\text{Var}(c) = 0$,
- $\text{Var}(cX) = c^2 \,\text{Var}(X)$,
- $\text{Var}(c + X) = \text{Var}(X)$.

An additional interesting property of the expectations is known as the ***Jensen's inequality***, which is as follows. Let X denote a random variable, and let $g(\cdot)$ denote a continuous and convex function. Then:

$$\mathbb{E}[g(X)] \geq g\left(\mathbb{E}[X]\right). \tag{27}$$

The opposite is true if $g(\cdot)$ is concave. If the function is strictly convex (or concave), the inequality holds strictly. And if the function is linear, then Equation (27) is satisfied with equality. The proof is simple. If a function is (globally) convex, for every point x of its domain crosses a line $h(X) = a_x + b_x X$ that satisfies $h(X) \leq g(X)$ for all X (the tangent). Given that, we know (from the last property of the expectations listed above) that $\mathbb{E}[g(X)] \geq \mathbb{E}[h(X)]$. However, we know that $h(x) = g(x)$ by construction, and thus, at the point $x = \mathbb{E}[X]$, $h(\mathbb{E}[X]) = g(\mathbb{E}[X])$. Because $h(\mathbb{E}[X]) = \mathbb{E}[h(\mathbb{E}[X])]$, the result follows.

Another property is known as the ***Chebyshev's inequality***, which is satisfied by any distribution, and is given by:

$$P(|X - \mu_X| \geq c) \leq \frac{\sigma_X^2}{c^2} \quad \Leftrightarrow \quad P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}, \tag{28}$$

where X is a random variable, μ_X is its mean, σ_X^2 is its variance, c is an arbitrary positive constant, and $k \equiv \frac{c}{\sigma_X}$. Equation (28) states that not more than $\frac{1}{k^2}$ of the distribution's values can be more than k standard deviations away from the mean. For example, in the case of the normal distribution, $P(|X - \mu_X| \geq \sigma_X) \approx 1 - 0.6827 \leq 1$, $P(|X - \mu_X| \geq 2\sigma_X) \approx 1 - 0.9545 \leq 0.25$, and $P(|X - \mu_X| \geq 3\sigma_X) \approx 1 - 0.9973 \leq \frac{1}{9}$.

More generally, the ***Markov's inequality*** establishes that, for any positive constant c and nonnegative function $g(\cdot)$:

$$P(g(X) \geq c) \leq \frac{\mathbb{E}[g(X)]}{c}, \tag{29}$$

provided that the expectation $\mathbb{E}[g(X)]$ exists.

IX. Quantiles, the Median, and the Mode

The τ***th quantile*** of a distribution indicates the minimum value of X below which there is a fraction τ of the density of the distribution:

$$q_\tau \equiv \min\{x : F_X(x) \geq \tau\}, \tag{30}$$

for $\tau \in [0, 1]$. When $F_X(\cdot)$ is invertible, $q_\tau = F_X^{-1}(\tau)$. Thus, the quantiles also characterize the distribution of X, as so does the cdf. The ***median*** (which is the population equivalent to the sample median seen in Chapter 1) is the 0.5th quantile, $q_{0.5}$. The ***mode*** is the value of X that has the maximum density (or mass if X is discrete).

Chapter 3: Multivariate Random Variables

By Joan Llull[*]

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2015

Main references:

— Mood: IV: 1.2, I:3.6, IV:3, V:6, IV:4.1, IV:5, IV:4.2, IV:4.6, IV:4.3

— Lindgren: 3.1, 3.4, 2.7, 3.7, 12.1, 12.4, 4.3, 12.2, 12.6, 4.7, 4.8, 4.2, 4.5, 12.3

I. Joint and Marginal Distributions

In this chapter we will work with random vectors, which include a collection of (scalar) random variables. We call these vectors ***multivariate*** random variables. For them, we will define a ***joint*** cumulative density function. Let $X1, ..., X_K$ denote a collection of K random variables. The joint cdf is defined as:

$$F_{X_1 \dots X_K}(x_1, ..., x_K) \equiv P(X_1 \leq x_1, X_2 \leq x_2, ..., X_K \leq x_K). \tag{1}$$

When the random variables are discrete, we can define a joint probability mass function given by:

$$P(X_1 = x_1, X_2 = x_2, ..., X_K = x_K). \tag{2}$$

For example, consider the case of tossing two coins. Let $\Omega = \{head, tail\} \times \{head, tail\}$. Define the following two random variables:

$$X_1 = \begin{cases} 1 & \text{if } \omega = \{(head, head)\} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if } \omega = \{(x, y) : \{x\} = \{y\}\} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In words, X_1 equals one if we obtain two heads, zero otherwise, and X_2 equals one if we obtain the same outcome with the two coins (either both heads or both tails), zero otherwise. Note that, in this case, the pmf is:

$$P(X_1 = x_1, X_2 = x_2) = \begin{cases} \frac{2}{4} = \frac{1}{2} & \text{if } x_1 = 0, x_2 = 0 \\ \frac{1}{4} & \text{if } x_1 = 0, x_2 = 1 \\ \frac{1}{4} & \text{if } x_1 = 1, x_2 = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

[*] 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.

(note the connection with the joint relative frequency in Chapter 1) and the cdf is:

$$F_{X_1X_2}(x_1,x_2)=\begin{cases}0 & \text{if } x_1<0 \text{ or } x_2<0\\ \frac{1}{2} & \text{if } x_1\leq 0 \text{ and } 0\leq x_2<1\\ \frac{3}{4} & \text{if } 0\leq x_1<1, x_2\geq 1\\ 1 & \text{if } x_1\geq 1, x_2\geq 1.\end{cases} \tag{5}$$

In the case of continuous variables, we have a joint probability density function, $f_{X_1...X_K}(x_1,...,x_K)$, which is implicitly defined as:

$$F_{X_1...X_K}(x_1,...,x_K)\equiv\int_{-\infty}^{x_1}...\int_{-\infty}^{x_K}f_{X_1...X_K}(z_1,...,z_K)dz_1...dz_K. \tag{6}$$

A joint pdf satisfies the following properties:

- $f_{X_1...X_K}(x_1,...,x_K)\geq 0$ for all $x_1,...,x_K$.
- $F_{X_1...X_K}(\infty,...,\infty)=\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}f_{X_1...X_K}(z_1,...,z_K)dz_1...dz_K=1$.
- Probabilities:
  - $P(a_1\leq X_1\leq b_1,...,a_K\leq X_K\leq b_k)=\int_{a_1}^{b_1}...\int_{a_K}^{b_K}f_{X_1...X_K}(z_1,...,z_K)dz_1...dz_K$.
  - $P(X_1=a_1,...,X_K=a_K)=0$.
  - $P(X_1=a,a_2\leq X_2\leq b_2,...,a_K\leq X_K\leq b_K)=0$.
- $\dfrac{\partial^K}{\partial x_1...\partial x_K}F_{X_1...X_K}(\cdot)=f(\cdot)$.

For example, the following is a pdf of a bivariate continuous random variable:

$$f_{XY}(x,y)=\begin{cases}\frac{3}{11}(x^2+y) & \text{if } 0\leq x\leq 2, 0\leq y\leq 1\\ 0 & \text{otherwise.}\end{cases} \tag{7}$$

In this example, the cdf is:

$$\begin{aligned}F_{XY}(x,y)&=\begin{cases}\int_0^{\min\{2,x\}}\int_0^{\min\{1,y\}}\frac{3}{11}(x^2+y)dydx & \text{if } x\geq 0, y\geq 0\\ 0 & \text{otherwise}\end{cases}\\ &=\begin{cases}\frac{3}{11}\left[\frac{\min\{1,y\}\min\{8,x^3\}}{3}+\frac{\min\{1,y^2\}}{2}\min\{2,x\}\right] & \text{if } x\geq 0, y\geq 0\\ 0 & \text{otherwise,}\end{cases}\end{aligned} \tag{8}$$

which satisfies $F_{XY}(\infty,\infty)=\frac{3}{11}\left[\frac{8}{3}+\frac{1}{2}2\right]=1$.

Similarly, we can define the ***marginal*** cdf, which is given by:

$$\begin{aligned}F_i(x)\equiv P(X_i\leq x)&=P(X_1\leq\infty,...,X_i\leq x,...,X_K\leq\infty)\\ &=F_{X_1...X_K}(\infty,...,x,...,\infty),\end{aligned} \tag{9}$$

and, either the marginal pmf (discrete case), defined as:

$$P(X_i = x) \equiv \sum_{x_1} ... \sum_{x_K} P(X_1 = x_1, ...X_i = x, ..., X_K = x_K), \tag{10}$$

or the marginal pdf (continuous case), defined as:

$$f_i(x) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1...X_K}(z_1, ..., x, ..., z_K)dz_1...dx_{i-1}dx_{i+1}...dz_K. \tag{11}$$

Note that the marginal cdf can also be defined as:

$$F_i(x) = \int_{-\infty}^{x} f_i(z)dz. \tag{12}$$

In our discrete example from above, the marginal pmf for X_1 is:

$$P(X_1 = x) = \begin{cases} \frac{3}{4} & \text{if } x = 0 \\ \frac{1}{4} & \text{if } x = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Note that this is still a well defined probability function for the variable X_1, as it satisfies the three axioms of a probability function.

In the continuous example above, the marginal pdf for X is:

$$f_X(x) = \begin{cases} \int_0^1 \frac{3}{11}(x^2 + y)dy = \frac{3}{11}\left(x^2 + \frac{1}{2}\right) & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

which is a well defined pdf, as it integrates to 1, and the marginal cdf is:

$$F_X = \begin{cases} 1 & \text{if } x \geq 2 \\ \int_0^x \frac{3}{11}\left(x^2 + \frac{1}{2}\right) = \frac{3}{11}\left(\frac{x^3}{3} + \frac{x}{2}\right) & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

II. Conditional Distributions and Independence

A. Conditional probability

Let us first introduce the concept of ***conditional probability***. In probability theory, a conditional probability measures the probability of an event given that another event has occurred. Let $\mathcal{A}$ and $\mathcal{B}$ be two events included in the σ-algebra of the sample space. The probability that $\mathcal{A}$ occurs given that $\mathcal{B}$ occurred, denoted by $P(\mathcal{A} \,|\, \mathcal{B})$ is formally defined as:

$$P(\mathcal{A} \,|\, \mathcal{B}) \equiv \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{B})}. \tag{16}$$

To illustrate it, consider the example of tossing two coins. We want to know that is the probability of obtaining two heads, conditional on the fact that the first coin already delivered a head. In this case, the sample space would be $\Omega = \{\{head, head\}, \{head, tail\}, \{tail, head\}, \{tail, tail\}\}$, the set $\mathcal{A}$ would be $\mathcal{A} = \{head, head\}$, the set $\mathcal{B}$ would be $\mathcal{B} = \{\{head, tail\}, \{head, head\}\}$, and the intersection of the two would be $\mathcal{A} \cap \mathcal{B} = \{head, head\}$. Thus, $P(\mathcal{A} \cap \mathcal{B})$, assuming coins are regular and, hence, events are equally likely, would be equal to $\frac{1}{4}$. Likewise, $P(\mathcal{B})$ would be $\frac{2}{4}$. Hence, $P(\mathcal{A} \,|\, \mathcal{B}) = \frac{1}{2}$.

This definition can be reversed to obtain the probability of $\mathcal{B}$ given that $\mathcal{A}$ occur, as they are both connected by $P(\mathcal{A} \cap \mathcal{B})$:

$$P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B}) = P(\mathcal{B} \,|\, \mathcal{A})P(\mathcal{A}) \Rightarrow P(\mathcal{B} \,|\, \mathcal{A}) = \frac{P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B})}{P(\mathcal{A})}. \quad (17)$$

This identity is called the ***Bayes' rule*** (a.k.a. Bayes' law, or Bayes' theorem).

The conditional probability allows us to talk about the ***independence*** of two events. We say that events $\mathcal{A}$ and $\mathcal{B}$ are independent if the conditional and marginal probabilities coincide. That is:

- $P(\mathcal{A} \,|\, \mathcal{B}) = P(\mathcal{A})$
- $P(\mathcal{B} \,|\, \mathcal{A}) = P(\mathcal{B})$
- $P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A})P(\mathcal{B})$.

Notice that these three conditions are equivalent, so we only need to check whether one of them holds.

B. Conditional distribution

Let X be a random variable, and let $\mathcal{A}$ be an event, with $P(\mathcal{A}) \neq 0$. The ***conditional*** cdf of X given $\mathcal{A}$ occurred is:

$$F_{X|\mathcal{A}}(x) \equiv P(X \leq x|\, \mathcal{A}) = \frac{P(X \leq x \cap \mathcal{A})}{P(\mathcal{A})}. \quad (18)$$

Very often, the event we are conditioning on is represented by a random variable(s), so that both X (which itself could also be a scalar or a random vector) and this random variable(s) form a random vector. In general, let X_1 denote the partition of the random vector that is our outcome of interest, and X_2 be the partition that includes the random variables we are conditioning on. The cdf of X_1 conditional on $X_2 = x_2$ is defined as:

$$F_{X_1|X_2}(x|x_2) \equiv \begin{cases} P(X_1 \leq x|X_2 = x_2) & \text{if } X_2 \text{ is discrete} \\ \lim_{h\to 0} P(X_1 \leq x|x_2 + h \geq X_2 \geq x_2) & \text{if } X_2 \text{ is continuous.} \end{cases} \quad (19)$$

The distinction between continuous and discrete is because we require that the marginal probability of the condition is not equal to zero for it to be defined. In the discrete case, the pmf is:

$$P(X_1 = x|X_2 = x_2) = \frac{P(X_1 = x, X_2 = x_2)}{P(X_2 = x_2)}. \tag{20}$$

Similarly, we can develop an analogous definition for the case of a continuous random vector. The conditional pdf of X_1 conditional on X_2 is:

$$f_{X_1|X_2}(x|x_2) \equiv \frac{f_{X_1X_2}(x, x_2)}{f_{X_2}(x_2)}, \tag{21}$$

where $f_{X_1|X_2}$ denotes the conditional pdf, $f_{X_1X_2}$ is the joint pdf, and f_{X_2} is the marginal pdf for X_2. Note that we can use this expression to ***factorize*** the joint pdf, in a Bayes' rule fashion, as follows:

$$f_{X_1X_2}(x_1, x_2) = f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2) = f_{X_2|X_1}(x_2|x_1)f_{X_1}(x_1). \tag{22}$$

We use these factorizations very often in econometrics.

We can also use Equation (20) to reformulate the conditional cdf for a continuous random vector as:

$$F_{X_1|X_2}(x|x_2) = \int_{-\infty}^{x} f_{X_1|X_2}(z|x_2)dz \tag{23}$$

Note that the conditional pdf is a well defined pdf:

- $f_{X_1|X_2}(x_1|x_2) \geq 0$.
- $\int_{-\infty}^{\infty} f_{X_1|X_2}(x, x_2)dx = 1$.
- $\frac{\partial}{\partial x}F_{X_1|X_2}(x|x_2) = f_{X_1|X_2}(x|x_2)$.

Also note that, if X_1 is a random vector of size K_1, the above integrals and differentials are K_1-variate.

To illustrate all this, consider the example used in previous section (Equation (7)). The conditional pdf of Y given X is:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{\frac{3}{11}(x^2 + y)}{\frac{3}{11}\left(x^2 + \frac{1}{2}\right)} = \frac{x^2 + y}{x^2 + \frac{1}{2}}, \tag{24}$$

for the relevant interval, and zero otherwise. Now we can use this expression to easily compute, for example:

$$P\left(0 \leq y \leq 1/2|x = 1\right) = \int_0^{1/2} \frac{1 + y}{1 + \frac{1}{2}}dy = \frac{2}{3}\left(y + \frac{y^2}{2}\right)\bigg]_0^{1/2} = \frac{5}{12}, \tag{25}$$

where we are abusing of notation by stating $x = 1$ instead of the limit $\lim_{h\to 0} x \in [1, 1+h]$. Similarly, we can compute:

$$P\left(0 \le y \le \frac{1}{2} \middle| \frac{1}{2} \le x \le \frac{3}{2}\right) = \frac{\displaystyle\int_{\frac{1}{2}}^{\frac{3}{2}} \int_0^{\frac{1}{2}} \frac{3}{11}\left(x^2 + y\right) dy dx}{\displaystyle\int_{\frac{1}{2}}^{\frac{3}{2}} \frac{3}{11}\left(x^2 + \frac{1}{2}\right) dx}. \tag{26}$$

C. *Independence*

We say that two random variables X_1 and X_2 are ***independent*** if and only if:

- The conditional distributions $f(X_1|X_2)$ and $f(X_2|X_1)$ do not depend on the conditioning variable, X_2 and X_1 respectively.
- $F_{X_1X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$ for all X_1 and X_2.
- $P(x_1 \in \mathcal{A}_{X_1} \cap x_2 \in \mathcal{A}_{X_2}) = P(x_1 \in \mathcal{A}_{X_1})P(x_2 \in \mathcal{A}_{X_2})$.

The three conditions are equivalent. Note, for example, that the third condition implies that we can formulate the conditional probability as:

$$\begin{aligned} P(x_1 \in \mathcal{A}_{X_1} | x_2 \in \mathcal{A}_{X_2}) &= \frac{P(x_1 \in \mathcal{A}_{X_1} \cap x_2 \in \mathcal{A}_{X_2})}{P(x_2 \in \mathcal{A}_{X_2})} \\ &= \frac{P(x_1 \in \mathcal{A}_{X_1})P(x_2 \in \mathcal{A}_{X_2})}{P(x_2 \in \mathcal{A}_{X_2})} \\ &= P(x_1 \in \mathcal{A}_{X_1}), \end{aligned} \tag{27}$$

which does not depend on X_2 (as the first condition indicates. Likewise, the second condition implies:

$$f_{X_1X_2}(x_1, x_2) = \frac{\partial^2 F_{X_1X_2}(x_1, x_2)}{\partial X_1 \partial X_2} = \frac{\partial F_{X_1}(x_1)}{\partial X_1}\frac{\partial F_{X_2}(x_2)}{\partial X_2} = f_{X_1}(x_1)f_{X_2}(x_2). \tag{28}$$

Thus, similarly to what we obtained in Equation (27), $f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1)$ for any x_1 and x_2. As a corollary, we can state that $(X_1, ..., X_K)$ are independent if and only if $F_{X_1...X_K}(x_1, ..., x_K) = \prod_{i=1}^K F_i(x_i)$.

III. Transformations of Random Variables

Let $(X_1, ..., X_K)'$ be a size K vector of independent random variables, and $g_1(\cdot), ..., g_K(\cdot)$ be K functions such that $\{Y_i = g_i(X_i) : j = 1, ..., K\}$ are random variables, then $(Y_1, ..., Y_K)'$ is also a vector of independent random variables.

To see it, note that the cdf of $(Y_1, ..., Y_K)'$ is:

$$\begin{aligned} F_{Y_1 \ldots Y_K}(y_1, ..., y_K) &= P(Y_1 \leq y_1, ..., Y_K \leq y_K) \\ &= P(X_1 \leq g_1^{-1}(y_1), ..., X_K \leq g_K^{-1}(y_K)) \\ &= F_{X_1 \ldots X_K}(g_1^{-1}(y_1), ..., g_K^{-1}(y_K)) \\ &= \prod_{i=1}^{K} F_{X_i}(g_1^{-1}(y_i)), \end{aligned} \tag{29}$$

where the last equality results from the fact that $X_1, ..., X_K$ are independent, and that $g_i(X_i)$ only takes X_i as an argument, and not X_j for $j \neq i$.

Finally, let X be a size K vector of continuous random variables with pdf $f_X(x)$, and let K-dimensional function $Y = g(X)$ with a unique inverse $X = g^{-1}(Y)$, and:

$$\det\left(\frac{\partial g^{-1}(Y)}{\partial Y'}\right) \neq 0. \tag{30}$$

Then, the joint pdf of $Y = g(X)$ is:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \det\left(\frac{\partial g^{-1}(Y)}{\partial Y'}\right) \right|. \tag{31}$$

IV. Multivariate Normal Distribution

The ***multivariate normal distribution*** is defined over a random vector $X = (X_1, ..., X_K)'$ by the following pdf:

$$f_X(x) = (2\pi)^{-\frac{K}{2}} \det(\Sigma_X)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu_X)'\Sigma_X^{-1}(x - \mu_X)\right), \tag{32}$$

where Σ_X is a $K \times K$ positive definite and symmetric matrix of parameters, and μ_X is a size $K \times 1$ vector of parameters. The fact that a random vector follows a multivariate normal is expressed as $X \sim \mathcal{N}_K(\mu_X, \Sigma_X)$. Thus, the normal distribution is completely characterized by $K + (\frac{1}{2}K(K+1))$ (the second term comes from the fact that Σ_X is symmetric).

The multivariate normal distribution is obtained as a linear transformation of a vector of independent random variables that are (standard) normally distributed. Formally, let $Z = (Z_1, ..., Z_K)'$ be a vector of independent variables such that $\{Z_i \sim \mathcal{N}(0,1) : j = 1, ..., K\}$. Define the random vector X as $X \equiv \mu_X + \Sigma_X^{\frac{1}{2}} Z$, where μ_X is a size K vector, and $\Sigma_X^{\frac{1}{2}}$ is a $K \times K$ nonsingular matrix that satisfies $\left(\Sigma_X^{\frac{1}{2}}\right)\left(\Sigma_X^{\frac{1}{2}}\right)' = \Sigma_X$. Then, implementing the result in Equation (31), we can see that $X \sim \mathcal{N}(\mu_X, \Sigma_X)$:

$$\left.\begin{array}{l} \phi_k(z) = \prod_{i=1}^{K} \phi(z_i) = (2\pi)^{-\frac{K}{2}} \exp\left(-\frac{1}{2}z'z\right) \\ X = \mu_X + \Sigma_X^{\frac{1}{2}} Z \Leftrightarrow Z = \Sigma_X^{-\frac{1}{2}}(x-\mu_X) \\ \left|\det\left(\Sigma_X^{-\frac{1}{2}}\right)\right| = \det(\Sigma_X)^{-\frac{1}{2}} \end{array}\right\} \Rightarrow \begin{array}{l} f_X(x) = (2\pi)^{-\frac{K}{2}} \det(\Sigma_X)^{-\frac{1}{2}} \times \\ \quad \exp\left(-\frac{1}{2}(x-\mu_X)'\Sigma_X^{-1}(x-\mu_X)\right). \end{array} \tag{33}$$

Using a similar derivation we can prove that $Y = a + BX \sim \mathcal{N}(a + B\mu_X, B\Sigma_X B')$.

V. Covariance, Correlation, and Conditional Expectation

A. Covariance and Correlation between Two Random Variables

Let $(X_1, X_2)'$ be two random variables with expectations $\mu_{X_1} \equiv \mathbb{E}[X_1]$ and $\mu_{X_2} \equiv \mathbb{E}[X_2]$. The ***covariance*** between X_1 and X_2 is defined as:

$$\operatorname{Cov}(X_1, X_2) \equiv \mathbb{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]. \tag{34}$$

Note that the variance is a special case of covariance: the one of a variable X with itself. Some properties of the covariance are:

- $\operatorname{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2]$.
- $\operatorname{Cov}(X_1, X_2) = \operatorname{Cov}(X_2, X_1)$.
- $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$.
- $\operatorname{Cov}(c, X) = 0$.
- $\operatorname{Cov}(aX_1, bX_2) = ab\operatorname{Cov}(X_1, X_2)$.
- $\operatorname{Cov}(X_1 + X_2, X_3) = \operatorname{Cov}(X_1, X_3) + \operatorname{Cov}(X_2, X_3)$.
- $\operatorname{Var}(X_1 + X_2) = \operatorname{Var}(X_1) + \operatorname{Var}(X_2) + 2\operatorname{Cov}(X_1, X_2)$.

The magnitude of the covariance depends on the units of measure, the same way than the descriptive statistic counterpart seen in Chapter 1 did. That is why we define the ***correlation coefficient*** as:

$$\rho_{X_1 X_2} \equiv \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}}. \tag{35}$$

The ***Cauchy-Schwarz inequality*** implies that $\rho^2_{X_1 X_2} \leq 1$, or, in other words, that coefficient ranges between -1 and 1. To prove it, define $U \equiv X_1 - \mu_{X_1}$ and $V \equiv X_2 - \mu_{X_2}$, and the function $w(t) = \mathbb{E}[(U - tV)^2] \geq 0$. Because $w(t)$ is a quadratic function in t, $(\mathbb{E}[V^2])t^2 - (2\,\mathbb{E}[UV])t + \mathbb{E}[U^2]$, it either has no roots or one root. This implies the discriminant (for $at^2 + bt + c$, the discriminant is $b^2 - 4ac$) has to be either zero (one root for $w(t)$) or negative (no roots).

Thus, $(2\,\mathbb{E}[UV])^2 - 4\,\mathbb{E}[V^2]\,\mathbb{E}[U^2] \leq 0$, which implies $\mathbb{E}[UV]^2 \leq \mathbb{E}[V^2]\,\mathbb{E}[U^2]$, from which the result follows trivially. In the case when $\rho^2_{X_1X_2} = 1$, there exists a value for t^* such that $w(t^*) = 0$, which is equivalent to say that $U = t^*V$, i.e. $X_1 - \mu_{X_1} = t^*(X_2 - \mu_{X_2})$, or, in words, X_1 is a linear transformation of X_2.

B. Expectation and Variance of Random Vectors

Let $X = (X_1, ..., X_K)'$ be a size K vector of random variables. The ***expectation*** of the random vector X is defined as:

$$\mathbb{E}[X] \equiv \int_{-\infty}^{\infty} x dF_{X_1 \dots X_K}(x_1, ..., x_K) = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_K] \end{pmatrix}, \tag{36}$$

where we make use of the Rienman-Stiljes integral. Similarly, we can define the expectation of a transformation of X as:

$$\mathbb{E}[g(X)] \equiv \int_{-\infty}^{\infty} g(x) dF_{X_1 \dots X_K}(x_1, ..., x_K). \tag{37}$$

A corollary of this is that, since $g(X) = (c_1, ..., c_K)X = c_1X_1 + ... + c_KX_K$, we can see that $\mathbb{E}[c_1X_1 + ... + c_KX_K] = c_1\,\mathbb{E}[X_1] + ... + c_K\,\mathbb{E}[X_K]$. Moreover, even though we cannot derive a general result for the expectation of the product of random variables, in the special case where the random variables are independent, we can establish that:

$$\begin{aligned} \mathbb{E}[X_1X_2] &= \int_{-\infty}^{\infty} x_1x_2 dF_{X_1X_2}(x_1, x_2) \qquad (38) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f_{X_1X_2}(x_1, x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1) dx_1 \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2) dx_2 \\ &= \mathbb{E}(X_1)\,\mathbb{E}(X_2). \end{aligned}$$

Thus, note that the fact that two variables are independent imply that the expectation of the product is the product of the expectations. However, the reverse implication is not true.

The ***variance-covariance matrix*** is defined as:

$$\begin{aligned}\mathrm{Var}(X) \equiv \mathbb{E}[(X-\mu_X)(X-\mu_X)'] = \\ = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \dots & \mathrm{Cov}(X_1,X_K) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2,X_K) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_K,X_1) & \mathrm{Cov}(X_K,X_2) & \dots & \mathrm{Var}(X_K) \end{pmatrix},\end{aligned} \tag{39}$$

where $\mu_X \equiv \mathbb{E}[X]$. This matrix is symmetric (because $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$), and positive-semidefinite (which is the equivalent to say that the variance is non-negative). To prove the last, recall that a matrix M is positive-semidefinite if, for all non-zero vectors $c \in \mathbb{R}^K$, $c'Mc \geq 0$. In the case of the variance-covariance matrix:

$$c' \mathrm{Var}(X)c = c' \mathbb{E}[(X-\mu_X)(X-\mu_X)']c = \mathbb{E}[c'(X-\mu_X)(X-\mu_X)'c] = \mathbb{E}[Y^2] \geq 0, \tag{40}$$

where we make use of the fact that the linear combination $c'(X - \mu_X)$ delivers a scalar random variable.

Retaking the example of the multivariate normal distribution, $\mathbb{E}[X] = \mu_X$, and $\mathrm{Var}(X) = \Sigma_X$. We can write Σ_X as:

$$\Sigma_X = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1K}\sigma_1\sigma_K \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots & \rho_{2K}\sigma_2\sigma_K \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K}\sigma_1\sigma_K & \rho_{2K}\sigma_2\sigma_K & \dots & \sigma_K^2 \end{pmatrix}. \tag{41}$$

It can be proved that $\det \Sigma > 0 \Leftrightarrow -1 < \rho < 1$. Thus, if two variables are perfectly correlated, we cannot write its joint normal density (which makes sense, given that one variable is a linear transformation of the other).

C. Conditional Expectation

The ***conditional expectation*** of a continuous random variable X_1 given X_2 is defined as:

$$\mathbb{E}[X_1|X_2 = x_2] \equiv \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2)dx_1. \tag{42}$$

If X_1 is discrete, it is analogously defined as:

$$\mathbb{E}[X_1|X_2 = x_2] \equiv \sum_{x_1 \in (-\infty,\infty)} x_1 P(X_1 = x_1|X_2 = x_2). \tag{43}$$

In general, using the Rienman-Stiltjes integral, we can write:

$$\mathbb{E}[X_1|X_2 = x_2] \equiv \int_{-\infty}^{\infty} x_1 dF_{X_1|X_2}(x_1|x_2). \tag{44}$$

The ***conditional variance*** is defined as:

$$\text{Var}[X_1|X_2 = x_2] \equiv \int_{-\infty}^{\infty} (x_1 - \mathbb{E}[X_1|X_2])^2 dF_{X_1|X_2}(x_1|x_2), \tag{45}$$

which can be expressed as $\mathbb{E}[X_1^2|X_2 = x_2] - \mathbb{E}[X_1|X_2 = x_2]^2$ (the derivation is analogous to the one for the unconditional variance).

Since we can compute the conditional expectation for every possible value that X_2 can take, we can simply denote $\mathbb{E}[X_1|X_2]$, which is a function of X_2. Let $h(X_2) \equiv \mathbb{E}[X_1|X_2]$ denote this function. We can compute $\mathbb{E}[h(X_2)]$ (i.e. $\mathbb{E}[\mathbb{E}[X_1|X_2]]$) integrating over the marginal distribution of X_2:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X_1|X_2]] &= \int_{-\infty}^{\infty} h(X_2) f_{X_2}(x_2) dx_2 && (46)\\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2) dx_1 \right) f_{X_2}(x_2) dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2) dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f_{X_1 X_2}(x_1, x_2) dx_1 dx_2 \\
&= E[X_1]. && (47)
\end{aligned}$$

This result is known as the ***law of iterated expectations***, and, even though we derived it for continuous variables, it also applies to discrete ones.

D. Revisiting Independence

In Section II.C above we defined the concept of independence, based on the distributions of the variables. Now, given the inputs in this section, we can revise the notion of independence, and define alternative degrees of independence. In particular, from strongest to weakest notion of independence (i.e. the first implies the second, which implies the third, but the reverse is not true), we define:

- ***Independence***: $F_{X_1|X_2}(x_1|x_2) = F_{X_1}(x_1)$ (or any of the equivalent specifications defined in Section II.C).
- ***Mean independence***: X_1 is mean independent of X_2 if $\mathbb{E}[X_1|X_2] = \mathbb{E}[X_1]$ for all values of X_2. Unlike the other two, this relation is not symmetric, as X_2 being mean independent of X_1 does not necessarily imply that X_1 is mean independent of X_2.

- ***Absence of correlation***: $\text{Cov}(X_1, X_2) = 0 = \rho_{X_1X_2}$.

To illustrate all this, consider a simple example. Let X_1 and X_2 be two discrete random variables, with pmf defined by the following table:

$X_1 \backslash X_2$	1	2	3
2	1/12	1/6	1/12
3	1/6	0	1/6
4	0	1/3	0

In this example:

$$F(X_1 = x_1|X_2 = 1) = \begin{cases} \frac{1/12}{1/12+1/6} = \frac{1}{3} & \text{if } x_1 = 2 \\ \frac{1/6}{1/12+1/6} = \frac{2}{3} & \text{if } x_1 = 3 \\ 0 & \text{if } x_1 = 4 \end{cases} \tag{48}$$

$$F(X_1 = x|X_2 = 2) = \begin{cases} \frac{1/6}{1/6+1/3} = \frac{1}{3} & \text{if } x_1 = 2 \\ 0 & \text{if } x_1 = 3 \\ \frac{1/3}{1/6+1/3} = \frac{2}{3} & \text{if } x_1 = 4 \end{cases}$$

$$F(X_1 = x|X_2 = 3) = \begin{cases} \frac{1}{3} & \text{if } x_1 = 2 \\ \frac{2}{3} & \text{if } x_1 = 3 \\ 0 & \text{if } x_1 = 4, \end{cases}$$

so clearly there is no independence, as $F(X_1|X_2)$ depends on X_2. Now, to check whether there is mean independence, we have to compute the conditional expectations:

$$\mathbb{E}[X_1|X_2 = x_2] = \begin{cases} 2 * \frac{1}{3} + 3 * \frac{2}{3} = \frac{8}{3} & \text{if } x_2 = 1 \\ 2 * \frac{1}{3} + 4 * \frac{2}{3} = \frac{10}{3} & \text{if } x_2 = 2 \\ 2 * \frac{1}{3} + 3 * \frac{2}{3} = \frac{8}{3} & \text{if } x_2 = 3 \end{cases} \tag{49}$$

$$\mathbb{E}[X_2|X_1 = x_1] = \begin{cases} 1 * \frac{1}{4} + 2 * \frac{1}{2} + 3 * \frac{1}{4} = 2 & \text{if } x_1 = 2 \\ 1 * \frac{1}{2} + 3 * \frac{1}{2} = 2 & \text{if } x_1 = 3 \\ 2 & \text{if } x_1 = 4, \end{cases}$$

so X_2 is mean independent of X_1, but X_1 is not mean independent of X_2. This implies that $\text{Cov}(X_1, X_2) = 0$. It is easy to show. Applying the law of iterated expectations, one can trivially see that $\text{Cov}(X_1, X_2) = \text{Cov}(X_1, \mathbb{E}[X_2|X_1]) = \text{Cov}(X_1, \mathbb{E}[X_2]) = 0$.

Another example is the multivariate normal distribution. In this case, independence, mean independence, and absence of correlation are equivalent, and all three occur (for the relation between X_i and X_j) if and only if $\rho_{ij} = 0$. To illustrate it, we can think of the bivariate normal, but it is trivially extended generally for the multivariate normal of any dimension. Independence is proved by checking that, when ρ is equal to zero, the joint density can be factorized as the product of the two marginals. Additionally, we can prove that the conditional distribution of X_1 given X_2 is:

$$X_1|X_2 \sim \mathcal{N}\left(\mu_1 + \rho_{12}\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2), \sigma_1^2(1 - \rho_{12}^2)\right), \tag{50}$$

which implies mean independence if and only if $\rho_{12} = 0$ (provided that X_1 is not a degenerate random variable, and thus $\sigma_1 \neq 0$). Finally, we prove that absence of correlation is equivalent to $\rho_{12} = 0$ from the definition of ρ_{12}.

VI. Linear Prediction

A. Expectations and Prediction

The conditional expectation (and the expectation in general) can be written as the result of a minimization process. We already pointed out something similar for the sample mean in Chapter 1. The expectation is the ***optimal predictor*** in the sense that it minimizes the expected quadratic loss. More formally, let $h(X)$ denote a prediction of a variable Y based on the information in X, and define $U \equiv Y - h(X)$ the prediction error. The conditional expectation satisfies:

$$\mathbb{E}[Y|X] = \arg\min_{h(X)} \mathbb{E}[(Y - h(X))^2] = \arg\min_{h(X)} \mathbb{E}[U^2]. \tag{51}$$

This property is trivial to prove by checking that, for any other function $m(X)$, $\mathbb{E}[(Y - m(X))^2] = \mathbb{E}\left[\{(y - \mathbb{E}[y|X]) + (\mathbb{E}[Y|X] - m(x))\}^2\right] \geq \mathbb{E}[(y - \mathbb{E}[y|X])^2]$. By extension, $\mathbb{E}[Y]$ is the ***optimal constant predictor***.

This notion of prediction is interesting because it allows us to separate the variation in Y that can be explained by X from the one that cannot. A way of quantifying to what extent a variable Y is explained by X compared to other factors is through the variance. The variance of Y can be decomposed as:

$$\text{Var}(Y) = \text{Var}(\mathbb{E}[Y|X]) + \text{Var}(U) + 2\,\text{Cov}(\mathbb{E}[Y|X], U), \tag{52}$$

where we used the fact that $h(X) = \mathbb{E}[Y|X]$ is the optimal predictor, and, thus, $Y \equiv \mathbb{E}[Y|X] + U$. To compute the $\text{Var}(U)$, we need $\mathbb{E}(U^2)$ and $\mathbb{E}[U]^2$. For the second term, we can use the law of iterated expectations:

$$\mathbb{E}[U|X] = \mathbb{E}[Y - \mathbb{E}[Y|X]|X] = \mathbb{E}[Y|X] - \mathbb{E}[Y|X] = 0 \quad \Rightarrow \quad \mathbb{E}[U] = 0. \tag{53}$$

Thus:

$$\text{Var}(U) = \mathbb{E}[U^2] = \mathbb{E}[\mathbb{E}[U^2|X]] = \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]] = \mathbb{E}[\text{Var}(Y|X)]. \tag{54}$$

Finally, using the result in (53), the last term in Equation (52) is:

$$\text{Cov}(\mathbb{E}[Y|X], U) = \mathbb{E}(\mathbb{E}[Y|X]U) = \mathbb{E}(\mathbb{E}[Y|X]\,\mathbb{E}[U|X]) = 0. \tag{55}$$

Hence, we can write:

$$\text{Var}(Y) = \text{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\text{Var}(Y|X)]. \tag{56}$$

The first term of the right-hand-side of the above expression is the variation of Y that is explained by X. The second term gives the expected variation of Y for a given value of X. Hence, we can introduce the ***population*** R^2:

$$R^2 \equiv \frac{\text{Var}(\mathbb{E}[Y|X])}{\text{Var}(Y)} = 1 - \frac{\mathbb{E}[\text{Var}(Y|X)]}{\text{Var}(Y)}. \tag{57}$$

This coefficient ranges between 0 and 1, and its interpretation is the fraction of the variation in Y that is explained by the variation in the prediction of Y given X. Thus, it is a measure of the ***goodness of fit*** of the model.

B. Optimal Linear Predictor

Now we focus on the ***optimal linear predictor***. Given a random vector (Y, X), the optimal linear predictor of Y given X is the function $\mathbb{E}^*[Y|X] \equiv \alpha + \beta X$ that satisfies:

$$(\alpha, \beta) = \arg\min_{(a,b)} \mathbb{E}[(Y - a - bX)^2]. \tag{58}$$

Solving for the first order conditions:

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \qquad \alpha = \mathbb{E}[Y] - \beta\,\mathbb{E}[X], \tag{59}$$

and, hence, it is equal to:

$$\mathbb{E}^*[Y|X] = \mathbb{E}[Y] + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - \mathbb{E}[X]), \tag{60}$$

The optimal linear predictor is optimal (in the sense of minimal quadratic error) in the class of linear predictors. Thus, when the conditional expectation function is linear, the optimal linear predictor is equal to the conditional expectation.

Using the optimal linear predictor we can compute a goodness of fit statistic that is analogous to the population R^2, defined in Equation (57):

$$\rho^2_{XY} \equiv \frac{\text{Var}(\mathbb{E}^*[Y|X])}{\text{Var}(Y)} = \beta^2 \frac{\text{Var}(X)}{\text{Var}(Y)} = \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)\,\text{Var}(Y)}, \tag{61}$$

and, hence, the notation ρ^2_{XY}. Notice that $0 \leq \rho^2_{XY} \leq R^2$. Also note that, if X is itself a random vector, then:

$$\beta = \mathrm{Var}(X)^{-1}\,\mathrm{Cov}(X,Y), \qquad \alpha = \mathbb{E}[Y] - \beta'\,\mathbb{E}[X]. \tag{62}$$

Let us introduce some properties of the optimal linear predictor:

- $\mathbb{E}^*[c|X] = c$.
- $\mathbb{E}^*[cX|X] = cX$.
- $\mathbb{E}^*[Y + Z|X] = \mathbb{E}^*[Y|X] + \mathbb{E}^*[Z|X]$.
- $\mathbb{E}^*[Y|X_1] = \mathbb{E}^*[\mathbb{E}^*[Y|X_1, X_2]|X_1]$.

An interesting case in which the optimal predictor is linear (i.e. the conditional expectation function of Y given X is linear in X) is the multivariate normal distribution (e.g. see the bivariate case in Equation (50)).

Chapter 4: Sample Theory and Sample Distributions

By Joan Llull[*]

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

Main references:

— Mood: IV:1, IV:2.1, IV:2.2-2, IV:3.1, IV:4

— Lindgren: 7.1, 7.3, 7.4, 7.5

I. Random samples

The objective of this chapter is to make inference about some characteristics of a population from a set of observations in the data. The population is described by a probabilistic model like those seen in previous chapters. The observations in the data are considered as realizations from the probabilistic model. Recall that one of the main features of a random experiment is that it can be replicated under the same conditions.

The process through which we obtain our data is called ***sampling***. There are several ways of sampling. In Chapter 2 we introduced sampling in finite sets as an example to illustrate the use of combinatorial analysis to compute probabilities.

Simple random sampling is the easiest way of selecting a sample. It is not always the best way we can do it in Economics, but its simplicity puts it as the starting point for all others. A collection of random variables (or random vectors) $(X_1, ..., X_N)$ is a (simple) ***random sample*** from F_X if $(X_1, ..., X_N)$ are independent and identically distributed (i.i.d) with cdf F_X. We can use the word sample to refer both to this random vector $(X_1, ..., X_N)$, and to the realization of it $(x_1, ..., x_N)$. Each of the elements of this vector is known as an ***observation***.

Given that the observations are i.i.d., the cdf of the sample is:

$$F_{X_1 \ldots X_N}(x_1, ..., x_N) = \prod_{i=1}^{N} F_X(x_i), \tag{1}$$

and, thus:

$$f_{X_1 \ldots X_N}(x_1, ..., x_N) = \prod_{i=1}^{N} f_X(x_i), \tag{2}$$

[*] 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.

where f_X is the pmf of the sample if X is discrete, and the corresponding pdf if X is continuous.

For example, consider a Bernoulli random variable with pmf equal to:

$$f_X(x) = \begin{cases} \frac{2}{3} & \text{if } x = 0 \\ \frac{1}{3} & \text{if } x = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Now consider a random sample of three observations obtained from this population. As we discussed in Chapter 2, there are $2^3 = 8$ possible permutations, and the pmf is given by:

Sample:	1	2	3	4	5	6	7	8
x_1:	0	0	0	0	1	1	1	1
x_2:	0	0	1	1	0	0	1	1
x_3:	0	1	0	1	0	1	0	1
$f_{X1...X_3}(x_1, x_2, x_3)$:	$\frac{8}{27}$	$\frac{4}{27}$	$\frac{4}{27}$	$\frac{2}{27}$	$\frac{4}{27}$	$\frac{2}{27}$	$\frac{2}{27}$	$\frac{1}{27}$

II. Sample mean and variance

A ***statistic*** is a single measure of some attribute of a sample. It is calculated by applying a function to the values of the items of the sample. Any of the synthetic measures that we computed in Chapter 1 were statistics. In that chapter we were using them to summarize the data. Now, we are going to use them to infer some properties of the probability model that generated the data.

As a transformation of random variables, a statistic is a random variable. As such, it has a probability distribution. This probability distribution is called ***sample distribution***.

The first of the statistics that we introduced in Chapter 1 is the ***sample mean***. In a simple random sample, the weights used to compute the sample mean are all equal, and thus equal to $\frac{1}{N}$. Therefore, here we define the sample mean as:

$$\bar{X}_N \equiv \frac{1}{N} \sum_{\mathrm{i=1}}^{\mathrm{N}} X_i. \tag{4}$$

In the example before, for each of the possible samples we would obtain a different sample mean:

Sample:	1	2	3	4	5	6	7	8
$\bar{x}$:	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	1

Note that all combinations of the same inputs give the same sample mean, so we could alternatively count the number of combinations instead of permutations.

Importantly, note that our statistic (the sample mean) has a sample distribution:

$$f_{\bar{X}_N}(\bar{x}) = \begin{cases} \frac{8}{27} & \text{if } \bar{x} = 0 \\ \frac{12}{27} & \text{if } \bar{x} = \frac{1}{3} \\ \frac{6}{27} & \text{if } \bar{x} = \frac{2}{3} \\ \frac{1}{27} & \text{if } \bar{x} = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Given Equation (5), we could compute $\mathbb{E}[\bar{X}_N]$ and $\text{Var}(\bar{X}_N)$:

$$\mathbb{E}[\bar{X}_N] = \frac{8}{27} \cdot 0 + \frac{12}{27} \cdot \frac{1}{3} + \frac{6}{27} \cdot \frac{2}{3} + \frac{1}{27} \cdot 1 = \frac{1}{3}, \tag{6}$$

and:

$$\begin{aligned} \text{Var}(\bar{X}_N) &= \mathbb{E}[\bar{X}_N^2] - \mathbb{E}[\bar{X}_N]^2 \\ &= \left[\frac{8}{27} \cdot 0^2 + \frac{12}{27} \cdot \left(\frac{1}{3}\right)^2 + \frac{6}{27} \cdot \left(\frac{2}{3}\right)^2 + \frac{1}{27} \cdot 1^2\right] - \left(\frac{1}{3}\right)^2 \\ &= \frac{2}{27} = \frac{2/9}{3}. \end{aligned} \tag{7}$$

Note that, for this variable, $\mathbb{E}[X] = p = 1/3$, and $\text{Var}(X) = p(1-p) = 2/9$. Therefore, at least in this example, $\mathbb{E}[\bar{X}_N] = \mathbb{E}[X]$ and $\text{Var}(\bar{X}_N) = \text{Var}(X)/N$. This result is general, as discussed in the following paragraph.

Let $(X_1, ..., X_N)$ be a random sample from a population described by the cdf F_X which has mean $\mathbb{E}[X] = \mu$ and variance $\text{Var}(X) = \sigma^2$. Let $\bar{X}_N$ denote the sample mean of this sample. Then, $\mathbb{E}[\bar{X}_N] = \mu$, and $\text{Var}(\bar{X}_N) = \sigma^2/N$. Let us check that:

$$\mathbb{E}[\bar{X}_N] = \mathbb{E}\left[\frac{1}{N}\sum_{\text{i=1}}^{\text{N}} X_i\right] = \frac{1}{N}\sum_{\text{i=1}}^{\text{N}} \mathbb{E}[X_i] = \frac{1}{N}\sum_{\text{i=1}}^{\text{N}} \mu = \frac{1}{N} N\mu = \mu. \tag{8}$$

And for the variance:

$$\begin{aligned} \text{Var}(\bar{X}_N) &= \mathbb{E}[(\bar{X}_N - \mu)^2] = \mathbb{E}\left[\left(\frac{1}{N}\sum_{\text{i=1}}^{\text{N}}(X_i - \mu)\right)^2\right] \\ &= \mathbb{E}\left[\frac{1}{N^2}\left(\sum_{\text{i=1}}^{\text{N}}(X_i - \mu)^2 + \sum_{\text{i=1}}^{\text{N}}\sum_{\text{j=1}}^{\text{N}}{}_{j \neq i} (X_i - \mu)(X_j - \mu)\right)\right] \\ &= \frac{1}{N^2}\sum_{\text{i=1}}^{\text{N}} \mathbb{E}\left[(X_i - \mu)^2\right] + \frac{1}{N^2}\sum_{\text{i=1}}^{\text{N}}\sum_{\text{j=1}}^{\text{N}}{}_{j \neq i} \mathbb{E}\left[(X_i - \mu)(X_j - \mu)\right] \\ &= \left[\frac{1}{N^2}\sum_{\text{i=1}}^{\text{N}} \mathbb{E}\left((X_i - \mu)^2\right)\right] = \frac{1}{N} N\sigma^2 = \frac{\sigma^2}{N}, \end{aligned} \tag{9}$$

where, from the third to the fourth line, we used the fact that, given that the observations are i.i.d., the covariance between X_i and X_j is equal to zero.

There are three main conclusions to extract from this general result. The first one is that $\mathbb{E}[\bar{X}_N]$, and $\mathrm{Var}(\bar{X}_N)$ do not depend of the form of F_X, they only depend on its first two moments. The second one is that $\bar{X}_N$ is "centered" around the population mean μ. And the third one is that the dispersion of $\bar{X}_N$ is reduced when we increase N, tending to zero when $N \to \infty$. We care about the variance of $\bar{X}_N$ as an indicator of the (inverse of the) ***precision*** of $\bar{X}_N$ as a proxy for μ: the smaller $\mathrm{Var}(\bar{X}_N)$, the more likely is that $\bar{X}_N$ is "close" to μ. Thus, the larger the sample, the more "accurate" is $\bar{X}_N$ as an approximation to μ. We will discuss extensively all this in the following chapters.

A similar analysis can be performed with respect to another of the statistics that we introduced in Chapter 1: the ***sample variance***. Again, given the observations are obtained from a random sample, the weight we give to each observation is equal for all of them, and equal to $\frac{1}{N}$. Thus, the sample variance, which we denote as $\hat{\sigma}^2_N$, is defined as:

$$\hat{\sigma}^2_N \equiv \frac{1}{N} \sum_{\mathrm{i=1}}^{\mathrm{N}} (X_i - \bar{X}_N)^2. \tag{10}$$

Let us first compute the expectation:

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2_N] &= \mathbb{E}\left[\frac{1}{N} \sum_{\mathrm{i=1}}^{\mathrm{N}} (X_i - \bar{X}_N)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N} \sum_{\mathrm{i=1}}^{\mathrm{N}} (X_i - \mu - (\bar{X}_N - \mu))^2\right] \\
&= \frac{1}{N} \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathbb{E}\left[(X_i - \mu)^2\right] - \mathbb{E}\left[(\bar{X}_N - \mu))^2\right] \\
&= \mathrm{Var}(X) - \mathrm{Var}(\bar{X}_N) = \sigma^2 - \frac{\sigma^2}{N} = \frac{(N-1)}{N}\,\sigma^2.
\end{aligned} \tag{11}$$

Thus, with the sample variance we expect to obtain less dispersion that the dispersion in the population, except when $N \to \infty$.

We often propose an alternative statistic to measure dispersion in the sample, the ***corrected sample variance***, which is defined as:

$$s^2_N \equiv \frac{N}{N-1}\,\hat{\sigma}^2_N = \frac{1}{N-1} \sum_{\mathrm{i=1}}^{\mathrm{N}} (X_i - \bar{X}_N)^2. \tag{12}$$

Easily we can check that $\mathbb{E}[s^2_N] = \sigma^2$. Therefore, unlike the sample variance, the corrected sample variance is centered around the population dispersion of the data. This is a desirable property when we want to make inference about the population, and, thus, s_N^2 is commonly used instead of $\hat{\sigma}_N^2$.

Even though we are not going to prove it (it is recommended as an exercise), it is messy but easy to show that:

$$\text{Var}(s_N^2) = \frac{2\sigma^4}{N-1} + \frac{\mu_4 - 3\sigma^4}{N}. \tag{13}$$

There exists an alternative measure of dispersion that has lower variance (i.e. that is more precise) than s_N^2:

$$\tilde{\sigma}_N^2 \equiv \frac{1}{N}\sum_{\text{i=1}}^{\text{N}}(X_i - \mu)^2. \tag{14}$$

Trivially, we can check that $\mathbb{E}[\tilde{\sigma}_N^2] = \sigma^2$. To compute the variance of $\tilde{\sigma}_N^2$, we only need to compute $\mathbb{E}[(\tilde{\sigma}_N^2)^2]$. To do so, define $Z_i \equiv X_i - mu$, so that notation is less messy:

$$\begin{aligned}\mathbb{E}[(\tilde{\sigma}_N^2)^2] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{\text{i=1}}^{\text{N}} Z_i^2\right)^2\right] = \frac{1}{N^2}\mathbb{E}\left[\sum_{\text{i=1}}^{\text{N}} Z_i^4 + \sum_{\text{i=1}}^{\text{N}}\sum_{\text{j=1}}^{\text{N}}{}_{i\neq j} Z_i^2 Z_j^2\right] \\ &= \frac{1}{N^2}[N\mu_4 + (N^2 - N)\sigma^4] = \frac{1}{N}[\mu_4 - (N-1)\sigma^4]. \end{aligned} \tag{15}$$

And, hence:

$$\text{Var}(\tilde{\sigma}_N^2) = \frac{1}{N}[\mu_4 - (N-1)\sigma^4] - \sigma^4 = \frac{1}{N}[\mu_4 - \sigma^4] < \text{Var}(s_N^2). \tag{16}$$

Therefore, this statistic would be preferred to the previous two to make inference about the variance of the distribution of X because it is centered at σ^2, like s_N^2, but it is more precise. However, this is an ***unfeasible estimator***, which means that, in general, we cannot compute it, because we do not know μ.

III. Sampling form a normal population: χ^2, t, and F distributions

Let $(X_1, ..., X_N)$ be a random sample from the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. From previous chapters we know that, as a linear combination of normal random variables, the sample mean is also normally distributed. And, from previous section, we know the parameters of this normal distribution:

$$\bar{X}_N \sim \mathcal{N}(\mu, \sigma^2/N). \tag{17}$$

Also using the materials from previous chapters, we also know the distribution of the following transformation:

$$Z \equiv \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1). \tag{18}$$

This result will allow us to make inference about μ based on $\bar{X}_N$ in future chapters, provided that σ^2 is known, because, given σ, the distribution of this statistic is known (standard normal).

One alternative is to replace σ^2 by s_N^2, but s_N^2 is itself a random variable, and, hence, the distribution above is altered. To see how, we first need to derive the distribution of s_N^2, and, to do that, we have to introduce some intermediate results:

1) Let $\tilde{Z} \equiv (\tilde{Z}_1, ..., \tilde{Z}_K)'$ be a vector of K i.i.d. random variables, with $\tilde{Z}_i \sim \mathcal{N}(0,1)$. Then, we say that $\tilde{W} = \tilde{Z}_1^2 + ... + \tilde{Z}_K^2 = \tilde{Z}'\tilde{Z}$ is distributed as a ***chi-squared*** with K ***degrees of freedom***: $\tilde{W} \sim \chi_K^2$. The degrees of freedom are the number of independent squared standard normal distributions that are adding. The support of this distribution is $\mathbb{R}^+$. Interesting results for this distribution are that $\mathbb{E}[\tilde{W}] = K$ and $\text{Var}(\tilde{W}) = 2K$ (you are strongly encouraged to prove them).

2) Let $\tilde{X} \sim \mathcal{N}_N(0, \Sigma)$. Then, $\tilde{X}'\Sigma^{-1}\tilde{X} \sim \chi_N^2$. To see it, decompose $\Sigma = \Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}$ as in previous chapter. Thus, $\tilde{X}'\Sigma^{-1}\tilde{X} = (\tilde{X}'\Sigma^{-\frac{1}{2}})(\Sigma^{-\frac{1}{2}}\tilde{X}) = \tilde{Z}'\tilde{Z}$, where $\tilde{Z} \sim \mathcal{N}_N(0, I)$, which is equivalent to say that all its elements are independently distributed as a standard normal. Given the definition of the chi-squared distribution in the previous bullet, we therefore know that $\tilde{Z}'\tilde{Z} \sim \chi_N^2$, completing the proof.

3) Let M be a size $K \times K$ idempotent (satisfies $MM = M$) and symmetric (satisfies $M' = M$) matrix, with $\text{rank}(M) = R \leq K$. Because it is idempotent, M is singular (with the only exception of $M = I$), it is also diagonalizable, and its eigenvalues are either 0 or 1. In particular, it can always be diagonalized as $M = C'\Lambda C$ such that $C'C = I$, and Λ is a matrix that include ones in the first R elements of the diagonal and zeros elsewhere. As a result, the trace of M (the sum of its diagonal elements) is equal to its rank (and thus always a natural number).

4) Let $\tilde{Z} \sim \mathcal{N}_K(0, I)$, and M be a size $K \times K$ idempotent and symmetric matrix with $\text{rank}(M) = R \leq K$. Then $\tilde{Z}'M\tilde{Z} \sim \chi_R^2$. To prove it, consider the diagonalization above: $\tilde{Z}'C'\Lambda C\tilde{Z}$. If we let C be the equivalent to $\Sigma^{\frac{1}{2}}$ above, $\tilde{Z}C \sim \mathcal{N}_K(0, C'C) = \mathcal{N}_K(0, I)$. Therefore, $\tilde{Z}'M\tilde{Z}$ is a sum of

R independent squared standard normals(given that Λ has R elements in the diagonal that are equal to one, and the rest are equal to zero), and thus $\tilde{Z}'M\tilde{Z} \sim \chi^2_R$.

5) Let $\tilde{Z} \sim \mathcal{N}_K(0, I)$, and M be a size $K \times K$ idempotent and symmetric matrix with $\text{rank}(M) = R \leq K$. Also let P be a $Q \times N$ matrix such that $PM = 0$. Then $\tilde{Z}'M\tilde{Z}$ and $P\tilde{Z}$ are independent. To prove it, note that, as linear combinations of a standard normal vector, both $M\tilde{Z}$ and $P\tilde{Z}$ are normal (thus, independence and absence of correlation are equivalent, as we saw in Chapter 3). Additionally:

$$\text{Cov}(P\tilde{Z}, M\tilde{Z}) = P\,\text{Cov}(\tilde{Z}, \tilde{Z})M = P\,\text{Var}(\tilde{Z})M = PIM = PM = 0. \quad (19)$$

(last step by assumption). Because M is idempotent and symmetric, $\tilde{Z}'M\tilde{Z} = \tilde{Z}'M'M\tilde{Z} = (M\tilde{Z})'M\tilde{Z}$. Thus, $\tilde{Z}'M\tilde{Z}$ is a function of $M\tilde{Z}$ so, since $M\tilde{Z}$ and $P\tilde{Z}$ are independent, $\tilde{Z}'M\tilde{Z}$ and $P\tilde{Z}$ are independent.

We now can use these intermediate results to derive the distribution of s^2_N. Define $\iota \equiv (1, ..., 1)'$ a size N vector of ones. Clearly, $\iota'\iota = N$, and, thus:

$$\bar{X} = \frac{1}{N}\sum_{\text{i=1}}^{\text{N}} X_i = (\iota'\iota)^{-1}\iota'X \equiv PX. \quad (20)$$

Similarly:

$$X - \iota\bar{X} = \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_N - \bar{X} \end{pmatrix} = (I - \iota(\iota'\iota)^{-1}\iota')X \equiv MX. \quad (21)$$

Trivially, the matrix M is symmetric and idempotent. Thus, we can write:

$$\sum_{\text{i=1}}^{\text{N}} (X_i - \bar{X})^2 = X'M'MX = X'MX. \quad (22)$$

This result implies that:

$$\frac{(N-1)s^2_N}{\sigma^2} = \frac{\sum_{\text{i=1}}^{\text{N}}(X_i - \bar{X})^2}{\sigma^2} = \frac{1}{\sigma^2}X'MX = \tilde{Z}'M\tilde{Z}. \quad (23)$$

where $\tilde{Z}_i \equiv \frac{X_i - \mu}{\sigma}$. The last equality is obtained by noting that $\bar{\tilde{Z}} \equiv N^{-1}\sum_{\text{i=1}}^{\text{N}} \tilde{Z}_i = \frac{1}{\sigma}\bar{X} - \frac{\mu}{\sigma}$, and thus:

$$M\tilde{Z} = \tilde{Z}_i - \bar{\tilde{Z}} = \tilde{Z}_i + \frac{\mu}{\sigma} - \left(\bar{\tilde{Z}} + \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}MX. \quad (24)$$

Therefore, since $\tilde{Z} \sim \mathcal{N}_N(0, I)$ and $\text{rank}(M) = \text{tr}(M) = N-1$, we conclude that:[1]

$$W \equiv \frac{(N-1)s_N^2}{\sigma^2} \sim \chi^2_{N-1}. \tag{25}$$

Finally, we introduce a new distribution: the ***Student-t***. Let $Z \sim \mathcal{N}(0, 1)$ and $W \sim \chi^2_K$, with Z and W being independent. Then:

$$t \equiv \frac{Z}{\sqrt{\frac{W}{K}}} \sim t_K, \tag{26}$$

which we read t follows a Student-t distribution with K degrees of freedom. The pdf of this distribution is symmetric with respect to zero, and its support is the real line. Also, $\mathbb{E}[t] = 0$ and $\text{Var}(t) = \frac{K}{K-2}$ for $K > 2$ (with $K \leq 2$ then the variance does not converge). When $K \to \infty$, the distribution is very similar to a normal distribution.

The choice of Z and W as a notation for this definition are not coincidental. The Z and W respectively defined in Equations (18) and (25) satisfy $Z \sim \mathcal{N}(0, 1)$ and $W \sim \chi^2_K$, as we proved above. Thus, we only need to prove that they are independent to be able to use the ***t-statistic*** from Equation (26) for our Z and W. To do so, we start by checking that $PM = 0$:

$$PM = \iota(\iota'\iota)^{-1}\iota'(I - \iota(\iota'\iota)^{-1}\iota') = \iota(\iota'\iota)^{-1}\iota' - \iota(\iota'\iota)^{-1}\iota' = 0. \tag{27}$$

Also, we note that $Z = \sqrt{N}P\tilde{Z}$, with $\tilde{Z} \sim \mathcal{N}_N(0, I)$. Thus, given that $W = \tilde{Z}'M\tilde{Z}$, and using the intermediate result number 5 above, we conclude that $P\tilde{Z}$ and W are independent, as so are Z and W. Therefore:

$$\frac{Z}{\sqrt{\frac{W}{N-1}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{N}}}{\sqrt{\frac{(N-1)s_N^2/\sigma^2}{N-1}}} = \frac{(\bar{X}-\mu)}{s/\sqrt{N}} \sim t_{N-1}. \tag{28}$$

Hence, with this statistic, we can make inference about μ without knowing σ^2.

There is another distribution that is useful to make inference about the variance. Even though we will not enter into the details of it, let us define it. Let W_1 and W_2 be two independent random variables such that $W_1 \sim \chi^2_K$ and $W_2 \sim \chi^2_Q$. Then:

$$F \equiv \frac{W_1/K}{W_2/Q} \sim F_{K,Q}, \tag{29}$$

or, in words, the statistic F follows a ***F-distribution*** with K and Q degrees of freedom. This distribution satisfies that $\mathbb{E}[F] = \frac{Q}{Q-2}$ (for $Q > 2$, otherwise the

[1] To prove that $\text{tr}(M) = N - 1$, note that $\text{tr}(I_N) = N$, and $\text{tr}(P) = \text{tr}(\iota(\iota'\iota)^{-1}\iota') = \text{tr}(\iota'\iota(\iota'\iota)^{-1}) = \text{tr}(1) = 1$ (since $\text{tr}(AB) = \text{tr}(BA)$), and, thus, $\text{tr}(M) = \text{tr}(I_N) - \text{tr}(P) = N - 1$.

integral does not converge). Also, $(t_K)^2 \sim F_{1,K}$, since the numerator is one squared normal (i.e. a chi-squared with one degree of freedom), and the denominator is a chi-squared with K degrees of freedom divided by K.

IV. Bivariate and Multivariate Sampling

So far we have analyzed the case in which we sample from a univariate distribution. However, we can also sample from a multivariate distribution. Let X be a size K random variable with joint pdf equal to $f_X(x)$. Now, we extract a random sample $(X_1, ..., X_N)$ where X_i for $i = 1..., N$ are random vectors. Given the random sampling, the joint pdf is given by:

$$f_{X_1 \dots X_N}(x_1, ..., x_k) = \prod_{i=1}^{N} f_{X_i}(x_i). \tag{30}$$

Thus, we can define the following "joint" statistics:

- Sample mean: $\mathbb{E}[X]$.
- Sample variance-covariance matrix: $\frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}}(X_i - \bar{X})(X_i - \bar{X})'$.

The sample variance-covariance matrix includes variances and covariances. We showed above that the expectation of the sample variance was not equal to the population variance, and thus we created a corrected variance. Should we do the same thing with the covariance? The answer is yes. The proof is analogous to the univariate case discussed above (one only needs to know that the expectation of the matrix is the matrix of the expectations, as we also discussed, and then operate individually).

Chapter 5: Estimation

By Joan Llull[*]

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

Main references:

— Goldberger: 11.1-11.13, 12.1, 12.2

— Lindgren: 8.1, 8.2, 7.7, 7.10, 2.8, 2.3, 8.7

I. Analogy Principle

The ***estimation problem*** consists of obtaining an approximation to a population characteristic θ_0 combining the information provided in a random sample. An ***estimator*** is a rule for calculating an ***estimate*** of a given quantity based on observed data. Hence, an estimator is a function from the sample $(X_1, ..., X_N)$, which we denote by $\hat{\theta}(X_1, ..., X_N)$, and an estimate is the result of implementing the estimator to the given sample, denoted by $\hat{\theta}(x_1, ..., x_N)$. In general, when the context is clear, we typically abuse of notation and simply use $\hat{\theta}$ to denote both the estimator and the estimate. The estimator is a statistic, and the estimate is a particular realization of this statistic.

A general rule to decide which estimator to implement is to define in the sample a statistic that satisfies similar properties to those satisfied by the ***true parameter*** in the population. This is called the ***analogy principle***. For example, to estimate the population mean, we often compute the sample mean; to estimate the variance, we often compute the sample variance; to estimate the median, we compute the median in the sample.

II. Desirable Properties of an Estimator

We define now certain criteria to determine the "quality" of an estimator. An estimator is good if it is a good approximation to the true parameter no matter which is the true value of the parameter. For example, consider a population with mean θ. If we are interested in estimating θ, we could propose as an estimator $\hat{\theta} = 3$ (regardless of what information I have in my sample). This estimator will be very good only if $\theta = 3$, but, in general, it is going to be bad. Instead, if we propose $\bar{X}_N$, the estimator will be generally good, as we expected $\bar{X}_N$ to be centered around θ, no matter what is the real value of θ.

[*] 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.

A measure of how good is an estimator is the ***mean squared error*** (MSE):

$$MSE(\hat{\theta}) \equiv \mathbb{E}[(\hat{\theta}-\theta)^2]. \tag{1}$$

The MSE can be decomposed as follows:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta}-\theta)^2] = \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta)^2] \\ &= \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2] + \mathbb{E}[(\mathbb{E}[\hat{\theta}]-\theta)^2] + 2\,\mathbb{E}[\hat{\theta}-\mathbb{E}[\hat{\theta}]](\mathbb{E}[\hat{\theta}]-\theta) \\ &= \text{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}]-\theta)^2. \end{aligned} \tag{2}$$

The difference $\mathbb{E}[\hat{\theta}]-\theta$ is called ***bias***. Thus, a good estimator, which is one that has small MSE, is an estimator that have small variance (precision), and small (or no) bias. An ***unbiased*** estimator is an estimator that has zero bias, that is, an estimator that satisfies $\mathbb{E}[\hat{\theta}]=\theta$.

For example, we saw in Chapter 4 that $\text{Var}(\bar{X}_N) = \frac{\text{Var}(X)}{N}$, and $\mathbb{E}[\bar{X}_N] = \mathbb{E}[X]$. Thus, $MSE(\bar{X}_N) = \frac{\text{Var}(X)}{N}$. We could do the same for s_N^2 or for $\hat{\sigma}_N^2$. Indeed, in an exercise in this chapter, you will compare the MSE of these two estimators of $\text{Var}(X)$. There, you will see that $\hat{\sigma}_N^2$, even though it is a biased estimator of $\text{Var}(X)$, has a lower MSE than the unbiased estimator s^2, no matter what the value of σ^2 or the sample size are. There are other examples where this is not true in general, but only for some values of the true parameter or for certain sample sizes.

Then the question is: what is more important, absence of a bias or lower MSE? There is obviously a trade-off, and no clear answer. What is clear, though, is that among the estimators that are unbiased, we prefer those that have less variance. We say that an estimator is more ***efficient*** than another if, being both unbiased, it has lower variance. Among all the estimators that are unbiased, the on that has the minimum possible variance is called ***best unbiased estimator*** (BUE), the minimum variance unbiased estimator, or simply the most efficient estimator. A more restrictive criterion is to search for the ***best linear unbiased estimator*** (BLUE), which is the best unbiased estimator of the class of estimators that are linear combinations of the data.

III. Moments and Likelihood Problems

A ***moments problem*** is defined by two equivalent conditions on the parameter of interest:

- It optimizes an expectation function. E.g.:

$$\mu = \arg\min_c \mathbb{E}[(Y-c)^2] \quad \text{or} \quad (\alpha,\beta) = \arg\min_{(a,b)} \mathbb{E}[(Y-a-bX)^2]. \tag{3}$$

- It solves a moment condition. E.g.:

$$\mathbb{E}[(Y-\mu)] = 0 \quad \text{or} \quad \mathbb{E}\left[(Y-\alpha-\beta X)\begin{pmatrix}1\\X\end{pmatrix}\right] = \begin{pmatrix}0\\0\end{pmatrix}. \tag{4}$$

Notice that any of these two requirements makes any assumption on the population distribution beyond the relevant moments. A method of moments estimator would use sample analogs to these conditions, and would obtain $\hat{\mu}$ or $(\hat{\alpha}, \hat{\beta})$ that satisfy them.

The ***likelihood problem*** is defined by a more completely specified environment. We assume that the population distribution is a known function, except for the parameters of interest, which are unknown. For instance, we know that the population pdf is $\{f(X;\theta) : \theta \in \Theta\}$, where Θ is the space of parameters for which the function is defined, but we do not know the value $\theta = \theta_0$, which is the true parameter value.

The fact that we know more information about the population of interest allows us to obtain better estimators from a given sample, provided that is extra information (the functional form of the population distribution) is correct.

As we prove below, the true parameter satisfies:

$$\theta_0 = \arg\max_{\theta\in\Theta} \mathbb{E}[\ln f(X;\theta)]. \tag{5}$$

To prove it, consider the first order condition of the above optimization problem:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial \ln f(X;\theta)}{\partial \theta'}\right] \equiv \mathbb{E}[z(X;\theta)] = 0 \\
\Leftrightarrow &\int_{-\infty}^{\infty} z(X;\theta) f(X;\theta_0) dX = 0 \\
\Leftrightarrow &\int_{-\infty}^{\infty} \frac{\partial \ln f(X;\theta)}{\partial \theta'} f(X;\theta_0) dX = 0 \\
\Leftrightarrow &\int_{-\infty}^{\infty} \frac{1}{f(X;\theta)} \frac{\partial f(X;\theta)}{\partial \theta'} f(X;\theta_0) dX = 0
\end{aligned} \tag{6}$$

Now, note that, because $f(X;\theta_0)$ is a pdf, it must integrate to 1:

$$\begin{aligned}
&\int_{-\infty}^{\infty} f(X;\theta_0) dX = 1 \\
\Leftrightarrow &\frac{\partial}{\partial \theta'} \int_{-\infty}^{\infty} f(X;\theta_0) dX = 0 \\
\Leftrightarrow &\int_{-\infty}^{\infty} \frac{\partial f(X;\theta_0)}{\partial \theta'} dX = 0.
\end{aligned} \tag{7}$$

(note that we are assuming that the range of integration does not depend on θ_0). Hence, replacing $\theta = \theta_0$ in Equation (6), we obtain the same expression as in Equation (7), and thus we conclude that θ_0 is a solution of the problem in Equation (5), because it satisfies the first order condition. We call $z(X; \theta_0)$ the ***score***, and the fact that $\mathbb{E}[z(X; \theta_0)] = 0$ is called ***zero expected score*** condition. Finally, note that the likelihood problem can also be seen as a moments problem.

IV. Maximum Likelihood Estimation

Consider a sample of size X, $(X_1, ..., X_N)$. In a likelihood problem, the pdf of this sample is known (up to parameter values), as we assume that $\{f(X; \theta) : \theta \in \Theta\}$ is known. The pdf of the sample written as a function of θ is known as the ***likelihood function***, and is equal to:

$$L_N(\theta) = \prod_{i=1}^{N} f(X_i; \theta). \tag{8}$$

$L_N(\theta)$ can be seen both as a function of the data for a given parameter θ (the pdf of the sample if $\theta = \theta_0$), or a function $\{L_N(\theta) : \theta \in \Theta\}$ of the parameter for a fixed sample (the likelihood function). The log-likelihood function is defined as:

$$\mathcal{L}_{\mathrm{N}}(\theta) \equiv \ln L_N(\theta) = \sum_{i=1}^{N} \ln f(X_i; \theta). \tag{9}$$

Given a sample with log-likelihood $\mathcal{L}_{\mathrm{N}}(\theta)$, we define the ***Maximum Likelihood Estimator*** (MLE) as:

$$\hat{\theta}_{MLE} \equiv \arg\max_{\theta \in \Theta} \mathcal{L}_{\mathrm{N}}(\theta) = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \ln f(X_i; \theta). \tag{10}$$

Note that this estimator is the sample analog of the condition that the true parameter θ_0 satisfies in the population, as described by Equation (5). The idea of the MLE is to approximate θ_0 by the value of θ that maximizes the likelihood (probability in the discrete case, density in the continuous case) of obtaining the sample that we observe. This is called the ***likelihood principle***.

The MLE satisfies the first order conditions of the optimization problem in Equation (10):

$$\frac{1}{N} \sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{\partial \ln f(X_i; \hat{\theta}_{MLE})}{\partial \theta'} = 0. \tag{11}$$

These conditions are sample analogs of the zero expected score rule.

V. The Cramer-Rao Lower Bound

The variance of the score is called the ***Fisher information*** (or simply information), $I(\theta_0) \equiv \text{Var}(z(X;\theta_0))$. The name information is motivated by the fact that $I(\theta_0)$ is a way of measuring the amount of information that a random variable X contains about an unknown parameter θ_0 for the distribution that models X (variation in the slope of the likelihood function). The ***information equality*** is an interesting property satisfied by the variance of the score:

$$\text{Var}(z(X;\theta_0)) = \mathbb{E}[z(X;\theta_0)z(X;\theta_0)'] = -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta\partial\theta'}\right]. \tag{12}$$

To prove it, note that, if the expected score is equal to zero, the derivative of the expected score should also be equal to zero:

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta'}f(X;\theta_0)dX = 0\\
&\Leftrightarrow \int_{-\infty}^{\infty}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta\partial\theta'}f(X;\theta_0)dX + \int_{-\infty}^{\infty}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta}\frac{\partial f(X;\theta_0)}{\partial\theta'}dX = 0\\
&\Leftrightarrow \mathbb{E}\left[\frac{\partial^2\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta\partial\theta'}\right] + \int_{-\infty}^{\infty}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta}\frac{\partial f(X;\theta_0)}{\partial\theta'}\frac{1}{f(X;\theta_0)}f(X;\theta_0)dX = 0\\
&\Leftrightarrow \mathbb{E}\left[\frac{\partial^2\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta\partial\theta'}\right] + \mathbb{E}\left[\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta}\frac{\partial\,\mathcal{L}_{\text{N}}(\theta_0)}{\partial\theta'}\right] = 0,
\end{aligned} \tag{13}$$

which, after rearranging the terms, delivers the result.

The ***Cramer-Rao inequality*** states that any unbiased estimator $\tilde{\theta}$ satisfies:

$$\text{Var}(\tilde{\theta}) \geq I(\theta_0)^{-1}. \tag{14}$$

Thus, this inequality indicates that the inverse of the information matrix is the lower bound for the variance for any unbiased estimator. Therefore, an unbiased estimator that has variance equal to the Cramer-Rao lower bound is the BUE. As we will be able to discuss after we introduce some concepts in Chapter 8, when the sample size tends to infinity, the variance of the MLE tends to the Cramer-Rao lower bound, and, hence, in that case, it is the BUE. Moreover, if a BUE exists, it is the MLE (we are not going to prove it).

We are going to prove the Cramer-Rao inequality for the simple case in which θ_0 is a scalar (and so is $I(\theta_0)$), but the results are directly generalizable to the case in which θ_0 is a vector (and thus $I(\theta_0)$ is a matrix). To do so, we are going to use the Schwartz inequality that we proved in Chapter 3. The Schwartz inequality states that $\text{Cov}(X,Y)^2/(\text{Var}(X)\,\text{Var}(Y)) \leq 1$. We are going to define two random variables: $\tilde{\theta}$ and $z(X;\theta_0)$. To compute $\text{Cov}(\tilde{\theta}, z(X;\theta_0))$ we start from the fact that $\tilde{\theta}$ is an unbiased estimator, and hence satisfies $\mathbb{E}[\tilde{\theta}] = \theta_0$:

$$\begin{aligned}
&\mathbb{E}[\tilde{\theta}] = \theta_0 \\
\Leftrightarrow & \frac{\partial}{\partial\theta_0'}\mathbb{E}[\tilde{\theta}] = \frac{\partial}{\partial\theta_0'}\int_{-\infty}^{\infty}\tilde{\theta}f(X;\theta_0)dX = \int_{-\infty}^{\infty}\tilde{\theta}\frac{\partial f(X;\theta_0)}{\partial\theta_0'}dX = 1 \\
\Leftrightarrow & \int_{-\infty}^{\infty}\tilde{\theta}\frac{\partial \ln f(X;\theta_0)}{\partial\theta_0'}f(X;\theta_0)dX = \mathbb{E}[\tilde{\theta}z(X;\theta_0)] = 1.
\end{aligned} \tag{15}$$

Because $\mathbb{E}[x(X;\theta_0)] = 0$, this implies that $\text{Cov}(\tilde{\theta}, z(X;\theta_0)) = 1$. Thus, the Schwartz inequality reads as:

$$\frac{1}{\text{Var}(\tilde{\theta})\,\text{Var}(z(X;\theta_0))} \leq 1 \quad \Leftrightarrow \quad \text{Var}(\tilde{\theta}) \geq \frac{1}{\text{Var}(z(X;\theta_0))} = \frac{1}{I(\theta_0))}, \tag{16}$$

proving the result. The same logic with a bit more tedious algebra applies to the multivariate case.

Let us illustrate all this with the normal distribution. Consider a random variable $X \sim \mathcal{N}(\mu_0, \sigma_0^2)$, for which we have an i.i.d. sample $(X_1, ..., X_N)$. We are interested in estimating the parameters $\theta = (\mu, \sigma^2)'$. The likelihood function for the sample is:

$$L_N(\mu,\sigma^2) = \prod_{i=1}^{N} f(x_i) = (2\pi)^{-\frac{N}{2}}(\sigma^2)^{-\frac{N}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{\text{i=1}}^{\text{N}}(x_i-\mu)^2\right), \tag{17}$$

and the log-likelihood is:

$$\mathcal{L}_{\text{N}}(\mu,\sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{\text{i=1}}^{\text{N}}(x_i-\mu)^2. \tag{18}$$

The score is:

$$z(X;\theta) = \frac{\partial\,\mathcal{L}_{\text{N}}(\mu,\sigma^2)}{\partial(\mu,\sigma^2)} = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{\text{i=1}}^{\text{N}}(x_i-\mu) \\ -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{\text{i=1}}^{\text{N}}(x_i-\mu)^2 \end{pmatrix}. \tag{19}$$

Evaluated at $\theta = (\mu_0, \sigma_0^2)$, it is easy to check that the expected score is equal to zero. The MLE picks $\theta = (\hat{\mu}, \hat{\sigma}^2)'$ such that $z(X;\hat{\theta}) = 0$ (i.e., the first order condition is satisfied), which, with simple algebra, delivers:

$$\hat{\mu} = \frac{1}{N}\sum_{\text{i=1}}^{\text{N}} X_i = \bar{X}_N \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{\text{i=1}}^{\text{N}}(X_i - \bar{X}_N)^2 = \hat{\sigma}_N^2. \tag{20}$$

To compute the Cramer-Rao lower bound, we can use the information matrix equality and compute $I(\theta_0) = -\mathbb{E}\left[\frac{\partial^2 z(X;\theta_0)}{\partial\theta\partial\theta'}\right]$:

$$\mathbb{E}\left[\begin{pmatrix} N\dfrac{1}{\sigma_0^2} & \dfrac{1}{\sigma_0^4}\displaystyle\sum_{\mathrm{i=1}}^{\mathrm{N}}(x_i-\mu_0) \\ \dfrac{1}{\sigma_0^4}\displaystyle\sum_{\mathrm{i=1}}^{\mathrm{N}}(x_i-\mu_0) & -\dfrac{N}{2}\dfrac{1}{\sigma_0^4}+\dfrac{1}{\sigma_0^6}\displaystyle\sum_{\mathrm{i=1}}^{\mathrm{N}}(x_i-\mu_0)^2 \end{pmatrix}\right] = \begin{pmatrix} \dfrac{N}{\sigma_0^2} & 0 \\ 0 & \dfrac{N}{2\sigma_0^4} \end{pmatrix}. \tag{21}$$

Therefore, the Cramer-Rao lower bound is:

$$I(\theta_0)^{-1} = \begin{pmatrix} \dfrac{\sigma_0^2}{N} & 0 \\ 0 & \dfrac{2\sigma_0^4}{N} \end{pmatrix}, \tag{22}$$

which allows us to extract three conclusions:

1) $\hat{\mu} = \bar{X}_N$ is the best unbiased estimator (BUE) of μ_0: it is an unbiased estimator (we saw in Chapter 4 that $\mathbb{E}[\bar{X}_N] = \mu_0$), and the variance of the sample mean (derived in Chapter 4) is equal to the Cramer-Rao lower bound.

2) $\hat{\sigma}^2 = \hat{\sigma}_N^2$ is a biased estimator of σ^2, as we noted in Chapter 4, and, hence, the Cramer-Rao result is not applicable. On the other hand, s_N^2, which is an unbiased estimator of σ_0^2, is not the BUE, because its variance is $\mathrm{Var}(s_N^2) = \frac{2\sigma^4}{N-1} + \frac{\mu_4 - 3\sigma^4}{N}$ (as noted in Chapter 4), larger than the Cramer-Rao bound. The latter does not surprise us, because we knew that, if there exist a BUE, it is the MLE.

3) If we knew μ_0, and we were to estimate only σ_0^2, the estimator that we would obtain would be $\hat{\sigma} = \tilde{\sigma}_N^2$, which is unbiased, and whose variance is $\mathrm{Var}(\tilde{\sigma}_N^2) = \frac{\mu_4 - \sigma_0^4}{N}$, where μ_4 is the fourth central moment. It turns out that, for the normal distribution, $\mu_4 = 3\sigma^4$, which would imply that the "ideal" estimator of the variance is indeed a BUE, because its variance is equal to the Cramer-Rao bound.

As a final note, it is left as an exercise to check that the equality of the information, which we have used in Equation (21), holds.

VI. Bayesian Inference

Recall the Bayes theorem in Chapter 3:

$$P(\mathcal{A}\,|\,\mathcal{B}) = \frac{P(\mathcal{B}\,|\,\mathcal{A})P(\mathcal{A})}{P(\mathcal{B})} = \frac{P(\mathcal{B}\,|\,\mathcal{A})P(\mathcal{A})}{P(\mathcal{B}\,|\,\mathcal{A})P(\mathcal{A}) + P(\mathcal{B}\,|\,\mathcal{A}^c)P(\mathcal{A}^c)}. \tag{23}$$

The second equality is new: we have used the fact that $P(\mathcal{B}) = P(\mathcal{B} \cap \mathcal{A}) + P(\mathcal{B} \cap \mathcal{A}^c)$. If instead of $\mathcal{A}$ and $\mathcal{A}^c$ we partition the sample space in N mutually exclusive sets that cover the entire sample space, $\mathcal{A}_1, ..., \mathcal{A}_N$, we can write:

$$P(\mathcal{A}_i \,|\, \mathcal{B}) = \frac{P(\mathcal{B} \,|\, \mathcal{A}_i)P(\mathcal{A}_i)}{P(\mathcal{B} \,|\, \mathcal{A}_1)P(\mathcal{A}1) + ... + P(\mathcal{B} \,|\, \mathcal{A}_N)P(\mathcal{A}_N)}. \tag{24}$$

We often have *a priori* beliefs about the probability that an event occurs. Imagine we want to estimate whether a coin is fair or not. It is natural to start from the belief that the probabilities assigned to heads and tails are 0.5 each. If we are only able to toss the coin once, it makes sense to give a lot of weight to our belief, and little to the sample, it looks a better strategy than assigning an estimate of $\hat{P}(\text{heads}) \equiv \hat{p} = 1$ for the outcome we obtained. If we can keep tossing the coin, we will update our beliefs with each toss. If after 10 times we have obtained 8 heads and only 2 tails, we will start to think that maybe the coin is not fair. If, after tossing it 1,000 times, we obtained 800 heads and 200 tails, then we will probably conclude that the coin is not fair, and our estimate will be $\hat{p} = \frac{8}{10}$. This is the intuition of the type of estimators that we are introducing in this section.

More formally, define ***subjective probability*** as the probability function that describes our beliefs about the true probabilities of the different outcomes. We can define subjective probabilities even when we do not have any *a priori* information or belief. We assume that we know the likelihood of the sample $f_N(X|\theta)$ up to the parameter θ, the ***a priori distribution*** $g(\theta)$ that describes our beliefs about the true parameter before observing the sample. The ***Bayesian inference*** is based on the ***a posteriori distribution*** of the parameter given the information in the sample, obtained from the application of the Bayes theorem:

$$h(\theta|X) = \frac{f(X|\theta)g(\theta)}{\int_{-\infty}^{\infty} f(X|c)g(c)dc} \propto f(X|\theta)g(\theta). \tag{25}$$

Importantly, note that we are treating θ (the "parameter") as a random variable now, not as a given (but unknown) value as we have been doing so far. The approaches used so far form the basis of the ***frequentist inference***, which is opposed to the Bayesian inference. In a frequentist approach, the unknown parameters of the model are not capable of being treated as random variables in any way. In contrast, a Bayesian approach to inference does allow probabilities to be associated with unknown parameters. Normally, these associated probabilities are probabilities in its Bayesian interpretation, which is a quantity that we assign for the purpose of representing a state of knowledge or belief.

Therefore, in Bayesian estimation, we are primarily interested in obtaining a posterior distribution $h(\theta|X)$. However, we can also obtain point estimates using the posterior distribution $h(\theta|X)$. For example, the mean of the posterior distribution minimizes the expected quadratic loss:

$$\mathbb{E}_h[\theta|X] = \arg\min_c \int_{-\infty}^{\infty} (c-\theta)^2 h(\theta|X) d\theta. \tag{26}$$

Likewise, the median of the posterior distribution minimizes the expected absolute loss, and the mode maximizes the posterior density.

Let us illustrate all this by retaking the example of tossing a coin. Let X be a random variable that takes the value of 1 if the outcome is a head, and let p denote the probability that $X = 1$ ($1-p$ is the probability of $X = 0$, and 0 is the probability for any other value of X). The likelihood of a sample of size N for X is:

$$f(X|p) = p^r(1-p)^{N-r}, \tag{27}$$

where $r \equiv \sum_{\mathrm{i=1}}^{\mathrm{N}} X_i$. Now we need a prior. Consider the beta distribution with given values for parameters α and β (such that $\alpha, \beta > 0$) as our prior:

$$g(p) = \begin{cases} p^{\alpha-1}(1-p)^{\beta-1}\frac{1}{B(\alpha,\beta)} & \text{if } p \in [0,1] \\ 0 & \text{otherwise,} \end{cases} \tag{28}$$

where the last term $B(\alpha, \beta) \equiv \int_0^1 z^{\alpha-1}(1-z)^{\beta-1}$ is a constant that guarantees the distribution to integrate to 1. In this case, the posterior distribution of p given the data is:

$$h(p|X) \propto p^{r+\alpha-1}(1-p)^{N-r+\beta-1}. \tag{29}$$

A point estimate for p is the mean of the posterior distribution:

$$\tilde{p} = \mathbb{E}_h[p|X] = \frac{\alpha+r}{\alpha+r+\beta+N-r} = \frac{\alpha+r}{\alpha+\beta+N}. \tag{30}$$

Note that we can rewrite $\tilde{p}$ as follows:

$$\tilde{p} = \frac{N}{\alpha+\beta+N}\frac{r}{N} + \frac{\alpha+\beta}{\alpha+\beta+N}\frac{\alpha}{\alpha+\beta} \equiv w(N)\hat{p} + (1-w(N))\,\mathbb{E}_g[p], \tag{31}$$

with $w(N) \in (0,1)$, and $\partial w(N)/\partial N = (\alpha+\beta)/(\alpha+\beta+N)^2 > 0$. Therefore, our estimate is a convex combination of the sample mean and the mean of our prior, with weights that are such that the weight on the former increases (and that on the latter decreases as a result) when the sample size N increases.

Chapter 6: Regression

By Joan Llull[*]

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

Main references:

— Goldberger: 13.1, 14.1-14.5, 16.4, 25.1-25.4, (13.5), 15.2-15.5, 16.1, 19.1

I. Classical Regression Model

A. Introduction

In this chapter, we are interested in estimating the conditional expectation function $\mathbb{E}[Y|X]$ and/or the optimal linear predictor $\mathbb{E}^*[Y|X]$ (recall that they coincide in the case where the conditional expectation function is linear). The generalization of the result in Chapter 3 about the optimal linear predictor for the case in which Y is a scalar and X is a vector is:

$$\mathbb{E}^*[Y|X] = \alpha + \beta' X \quad \Rightarrow \quad \begin{array}{c} \beta = [\mathrm{Var}(X)]^{-1}\,\mathrm{Cov}(X,Y) \\ \alpha = \mathbb{E}[Y] - \beta'\,\mathbb{E}[X]. \end{array} \tag{1}$$

Consider the bivariate case, where $X = (X_1, X_2)'$. It is interesting to compare $\mathbb{E}^*[Y|X_1]$ and $\mathbb{E}^*[Y|X_1,X_2]$. Let $\mathbb{E}^*[Y|X_1] = \alpha^* + \beta^* X_1$ and $\mathbb{E}^*[Y|X_1,X_2] = \alpha + \beta_1 X_1 + \beta_2 X_2$. Thus:

$$\mathbb{E}^*[Y|X_1] = \mathbb{E}^*[\mathbb{E}^*[Y|X_1,X_2]|X_1] = \alpha + \beta_1 X_1 + \beta_2\,\mathbb{E}^*[X_2|X_1]. \tag{2}$$

Let $\mathbb{E}^*[X_2|X_1] = \gamma + \delta X_1$. Then:

$$\mathbb{E}^*[Y|X_1] = \alpha + \beta_1 X_1 + \beta_2(\gamma + \delta X_1) \quad \Rightarrow \quad \begin{array}{c} \beta^* = \beta_1 + \delta\beta_2 \\ \alpha^* = \alpha + \gamma\beta_2. \end{array} \tag{3}$$

This result tells us that the effect of changing variable X_1 on Y is given by a direct effect (β_1) and an indirect effect through the effect of X_1 on X_2 and X_2 on Y. For example, consider the case in which Y is wages, X_1 is age, and X_2 is education, with $\beta_1, \beta_2 > 0$. If we do not include education in our model, then we could obtain a β_1^* that is negative, as older individuals may have lower education.

B. Ordinary Least Squares

Consider a set of observations $\{(y_i, x_i) : i = 1, ..., N\}$ where y_i are a scalars, and x_i are vectors of size $K \times 1$. Using the analogy principle, we can propose a natural

[*] 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.

estimator for α and β:[1]

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(a,b)} \frac{1}{N} \sum_{\text{i=1}}^{\text{N}} (y_i - a - b'x_i)^2. \tag{4}$$

This estimator is called ***Ordinary Least Squares***. The solution to the above problem is:

$$\begin{aligned} \hat{\beta} &= \left[\sum_{\text{i=1}}^{\text{N}} (x_i - \bar{x}_N)(x_i - \bar{x}_N)' \right]^{-1} \sum_{\text{i=1}}^{\text{N}} (x_i - \bar{x}_N)(y_i - \bar{y}_N), \\ \hat{\alpha} &= \bar{y}_N - \hat{\beta}'\bar{x}_N. \end{aligned} \tag{5}$$

Note that the first term of $\hat{\beta}$ is a $K \times K$ matrix, while the second is a $K \times 1$ vector.

C. Algebraic Properties of the OLS Estimator

Let us introduce some compact notation. Let $\delta \equiv (\alpha, \beta')'$ be the parameter vector, let $y = (y_1, ..., y_N)'$ be the vector of observations of Y, and let $W = (w_1, ..., w_N)'$ such that $w_i = (1, x_i')'$ be the matrix (here we are using capital letters to denote a matrix, not a random variable) of observations for the remaining variables. Then:

$$\hat{\delta} = \arg\min_d \sum_{\text{i=1}}^{\text{N}} (y_i - w_i'd)^2 = \arg\min_d (y - Wd)'(y - Wd). \tag{6}$$

And the solution is:

$$\hat{\delta} = \left(\sum_{\text{i=1}}^{\text{N}} w_i w_i' \right)^{-1} \sum_{\text{i=1}}^{\text{N}} w_i y_i = (W'W)^{-1}W'y. \tag{7}$$

Let us do the matrix part in detail. First note:

$$\begin{aligned} (y - Wd)'(y - Wd) &= y'y - y'Wd - d'W'y + d'W'Wd \\ &= y'y - 2d'W'y + d'W'Wd. \end{aligned} \tag{8}$$

The last equality is obtained by observing that all elements in the sum are scalars. The first order condition is:

$$\begin{gathered} -2W'y + 2(W'W)\hat{\delta} = 0, \\ W'y = (W'W)\hat{\delta}, \\ \hat{\delta} = (W'W)^{-1}W'y. \end{gathered} \tag{9}$$

[1] To avoid complications with the notation below, in this chapter we follow the convention of writing the estimators as a function of realizations (y_i, x_i) instead of doing it as functions of the random variables (Y_i, X_i).

Note that we need $W'W$ to be full rank, such that it can be inverted. This is to say, we require ***absence of multicollinearity***.

D. Residuals and Fitted Values

Recall from Chapter 3 the prediction error $U \equiv y - \alpha - \beta' X = y - (1, X')\delta$. In the sample, we can define an analogous concept, which is called the ***residual***: $\hat{u} = y - W\hat{\delta}$. Similarly, we can define the vector of ***fitted values*** as $\hat{y} = W\hat{\delta}$. Clearly, $\hat{u} = y - \hat{y}$. Some of their properties are useful:

1) $W'\hat{u} = 0$. This equality comes trivially from the derivation in (9): $W'\hat{u} = W'(y - W\hat{\delta}) = W'y - (W'W)\hat{\delta} = 0$. Looking at these matrix multiplications as sums, we can observe that they imply $\sum_{\text{i=1}}^{\text{N}} \hat{u}_i = 0$, and $\sum_{\text{i=1}}^{\text{N}} x_i \hat{u}_i = 0$. Interestingly, these are sample analogs of the population moment conditions satisfied by U.
2) $\hat{y}'\hat{u} = 0$ because $\hat{y}'\hat{u} = \hat{\delta} W'\hat{u} = \hat{\delta} \cdot 0 = 0$.
3) $y'\hat{y} = \hat{y}'\hat{y}$ because $y'\hat{y} = (\hat{y} + \hat{u})'\hat{y} = \hat{y}'\hat{y} + \hat{u}'\hat{y} = \hat{y}'\hat{y} + 0 = \hat{y}'\hat{y}$.
4) $\iota' y = \iota'\hat{y} = N\bar{y}$, where ι is a vector of ones, because $\iota'\hat{u} = \sum_{\text{i=1}}^{\text{N}} \hat{u}_i = 0$, and $\iota' y = \iota'\hat{y} + \iota'\hat{u}$.

E. Variance Decomposition and Sample Coefficient of Determination

Following exactly the analogous arguments as in the proof of the variance decomposition for the linear prediction model in Chapter 3 we can prove that:

$$y'y = \hat{y}'\hat{y} + \hat{u}'\hat{u} \quad \text{and} \quad \widehat{\text{Var}}(y) = \widehat{\text{Var}}(\hat{y}) + \widehat{\text{Var}}(\hat{u}), \tag{10}$$

where $\widehat{\text{Var}}(z) \equiv N^{-1} \sum_{\text{i=1}}^{\text{N}} (z - \bar{z})^2$ To prove the first, we simply need basic algebra:

$$\hat{u}'\hat{u} = (y - \hat{y})'(y - \hat{y}) = y'y - \hat{y}'y - y'\hat{y} + \hat{y}'\hat{y} = y'y - \hat{y}'\hat{y}. \tag{11}$$

The last equality is obtained following the result $y'\hat{y} = \hat{y}'\hat{y}$ obtained in item 3) from the list above. To prove the second equality in (10), we need to recall from Chapter 4 that we can write $\sum_{\text{i=1}}^{\text{N}} (y - \bar{y})^2 = (y - \iota\bar{y})'(y - \iota\bar{y})$. And now, we can operate:

$$(y - \iota\bar{y})'(y - \iota\bar{y}) = y'y - \bar{y}\iota' y - y'\iota\bar{(y)} + \bar{y}^2 \iota'\iota = y'y - N\bar{y}^2. \tag{12}$$

Given the result in item 4) above, we can conclude that $(\hat{y} - \iota\bar{y})'(\hat{y} - \iota\bar{y}) = \hat{y}'\hat{y} - N\bar{y}^2$. Thus:

$$N\widehat{\text{Var}}(\hat{u}) = \hat{u}'\hat{u} = y'y - \hat{y}'\hat{y} = y'y - N\bar{y}^2 - (\hat{y}'\hat{y} - N\bar{y}^2) = N\widehat{\text{Var}}(y) - N\widehat{\text{Var}}(\hat{y}), \tag{13}$$

completing the proof.

Similar to the population case described in Chapter 3, this result allows us to write the ***sample coefficient of determination*** as:

$$R^2 \equiv 1 - \frac{\sum_{\mathrm{i=1}}^{\mathrm{N}} u_i^2}{\sum_{\mathrm{i=1}}^{\mathrm{N}} (y_i - \bar{y})^2} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{N}} (\hat{y}_i - \bar{y})_i^2}{\sum_{\mathrm{i=1}}^{\mathrm{N}} (y_i - \bar{y})^2} = \frac{\widehat{\mathrm{Var}}(\hat{y})}{\widehat{\mathrm{Var}}(y)} = \frac{[\widehat{\mathrm{Cov}}(y, \hat{y})]^2}{\widehat{\mathrm{Var}}(\hat{y})\widehat{\mathrm{Var}}(y)} = \rho_{y,\hat{y}}^2. \tag{14}$$

The last equality is obtained by multiplying and dividing by $\hat{y}'\hat{y}$, and using that $\hat{y}'\hat{y} = y'\hat{y}$ as shown above.

F. Assumptions for the Classical Regression Model

So far we have just described algebraic properties of the OLS estimator as an estimator of the parameters of the linear prediction of Y given X. In order to use the OLS estimator to obtain information about $\mathbb{E}[Y|X]$, we require additional assumptions. This extra set of assumptions constitute what is known as the ***classical regression model***. These assumptions are:

- **Assumption 1 (linearity+strict exogeneity)**: $\mathbb{E}[y|W] = W\delta$, which is equivalent to say $\mathbb{E}[y_i|x_1, ..., x_N] = \alpha + x_i'\beta$, or to define $y \equiv W\delta + u$ where $\mathbb{E}[u|W] = 0$. There are two main conditions embedded in this assumption. The first one is ***linearity***, which implies that the optimal linear predictor and the conditional expectation function coincide. The second one is that $\mathbb{E}[y_i|x_1, ..., x_N] = \mathbb{E}[y_i|x_i]$, which is called ***(strict) exogeneity***. Exogeneity implies that $\mathrm{Cov}(u_i, x_{kj}) = 0$ and $\mathbb{E}[u_i|W] = 0$. To prove it, note that $\mathbb{E}[u_i] = \mathbb{E}[\mathbb{E}[u_i|W]] = \mathbb{E}[\mathbb{E}[y_i - \alpha - x_i'\beta|W]] = \mathbb{E}[\mathbb{E}[y_i|W] - \alpha - x_i'\beta] = 0$, and, hence, $\mathrm{Cov}(u_i, x_{kj}) = \mathbb{E}[u_i x_{kj}] = \mathbb{E}[x_{kj}\,\mathbb{E}[u_i|W]] = 0$. This assumption is satisfied by an i.i.d. random sample:

$$\begin{aligned} f(y_i|x_1, ..., x_N) = \frac{f(y_i, x_1, ..., x_N)}{f(x_1, ..., x_N)} &= \frac{f(y_i, x_i) f(x_1)...f(x_{i-1}) f(x_{i+1})...f(x_N)}{f(x_1)...f(x_N)} \\ = \frac{f(y_i, x_i)}{f(x_i)} = f(y_i|x_i), & \end{aligned} \tag{15}$$

  which implies that $\mathbb{E}[y_i|x_1, ..., x_N] = \mathbb{E}[y_i|x_i]$. This is not satisfied, for example, by time series data: if $x_i = y_{i-1}$ (that is, a regressor is the lag of the dependent variable), as $\mathbb{E}[y_i|x_1, ..., x_N] = \mathbb{E}[y_i|x_i, x_{i+1} = y_i] = y_i \neq \mathbb{E}[y_i|x_i]$.

- **Assumption 2 (homoskedasticity)**: $\mathrm{Var}(y|W) = \sigma^2 I_N$. This assumption implies (along with the previous one) that $\mathrm{Var}(y_i|x_1, ..., x_N) = \mathrm{Var}(y_i|x_i) =$

σ^2 and $\text{Cov}(y_i, y_j|x_1, ..., x_N) = 0$ for all $i \neq j$:

$$\begin{aligned}\text{Var}(y_i|x_i) &= \text{Var}(\mathbb{E}[y_i|x_1, ..., x_N]|x_i) + \mathbb{E}[\text{Var}(y_i|x_1, ..., x_N)|x_i] \\ &= \text{Var}(\mathbb{E}[y_i|x_i]|x_i) + \mathbb{E}[\sigma^2|x_i] = 0 + \sigma^2 = \sigma^2. \qquad (16)\end{aligned}$$

We could also check as before that an i.i.d. random sample would satisfy this condition.

II. Statistical Results and Interpretation

A. Unbiasedness and Efficiency

In the classical regression model, $\mathbb{E}[\hat{\delta}] = \delta$:

$$\mathbb{E}[\hat{\delta}] = \mathbb{E}[\mathbb{E}[\hat{\delta}|W]] = \mathbb{E}[(W'W)^{-1}W'\,\mathbb{E}[y|W]] = \mathbb{E}[\delta] = \delta, \qquad (17)$$

where we crucially used the Assumption 1 above. Similarly, $\text{Var}(\hat{\delta}|W) = \sigma^2(W'W)^{-1}$:

$$\text{Var}(\hat{\delta}|W) = (W'W)^{-1}W'\,\text{Var}(y|W)W(W'W)^{-1} = \sigma^2(W'W)^{-1}, \qquad (18)$$

where we used Assumption 2. Note that $\text{Var}(\hat{\delta}) = \sigma^2\,\mathbb{E}[(W'W)^{-1}]$:

$$\text{Var}(\hat{\delta}) = \text{Var}(\mathbb{E}[\hat{\delta}|W]) + \mathbb{E}[\text{Var}(\hat{\delta}|W)] = 0 + \sigma^2\,\mathbb{E}[(W'W)^{-1}]. \qquad (19)$$

The first result that we obtained indicates that OLS gives an unbiased estimator of δ under the classical assumptions. Now we need to check how good is it in terms of efficiency. The ***Gauss-Markov Theorem*** establishes that OLS is a BLUE (best linear unbiased estimator). More specifically, the theorem states that in the class of estimators that are conditionally unbiased and linear in y, $\hat{\delta}$ is the estimator with the minimum variance.

To prove it, consider an alternative linear estimator $\tilde{\delta} \equiv Cy$, where C is a function of the data W. We can define, without loss of generality, $C \equiv (W'W)^{-1}W' + D$, where D is a function of W. Assume that $\tilde{\delta}$ satisfies $\mathbb{E}[\tilde{\delta}|W] = \delta$ (hence, $\tilde{\delta}$ is another linear unbiased estimator). We first check that $\mathbb{E}[\tilde{\delta}|W] = \delta$ is equivalent to $DW = 0$:

$$\begin{aligned}&\mathbb{E}[\tilde{\delta}|W] = \mathbb{E}[\delta + (W'W)^{-1}W'u + DW\delta + Du|W] = (I + DW)\delta \\ &\qquad (I + DW)\delta = \delta \Leftrightarrow DW = 0, \qquad (20)\end{aligned}$$

given that $\mathbb{E}[Du|W] = D\,\mathbb{E}[u|W] = 0$. An implication of this is that $\tilde{\delta} = \delta + Cu$, since $DW\delta = 0$. Hence:

$$\begin{aligned}\text{Var}(\tilde{\delta}|W) &= \mathbb{E}[(\tilde{\delta} - \delta)(\tilde{\delta} - \delta)'|W] = \mathbb{E}[Cuu'C'|W] = C\,\mathbb{E}[uu'|W]C' = \sigma^2 CC' \\ &= (W'W)^{-1}\sigma^2 + \sigma^2 DD' = \text{Var}(\hat{\delta}|W) + \sigma^2 DD' \geq \text{Var}(\hat{\delta}|W). \qquad (21)\end{aligned}$$

Therefore, $\mathrm{Var}(\hat{\delta}|W)$ is the minimum conditional variance of linear unbiased estimators. Finally, to prove that $\mathrm{Var}(\hat{\delta})$ is the minimum as a result we use the variance decomposition and the fact that the estimator is conditionally unbiased, which implies $\mathrm{Var}(\mathbb{E}[\tilde{\delta}|W]) = 0$. Using that, we obtain $\mathrm{Var}(\tilde{\delta}) = \mathbb{E}[\mathrm{Var}(\tilde{\delta}|W)]$. Hence, proving whether $\mathrm{Var}(\tilde{\delta}) - \mathrm{Var}(\hat{\delta}) \geq 0$, which is what we need to prove to establish that $\mathrm{Var}(\hat{\delta})$ is the minimum for this class of estimators, is the same as proving $\mathbb{E}[\mathrm{Var}(\tilde{\delta}|W) - \mathrm{Var}(\hat{\delta}|W)] \geq 0$. Note that, given a random matrix A, because $Z'\,\mathbb{E}[A]Z = \mathbb{E}[Z'AZ]$ if A is positive semidefinite, $\mathbb{E}[A]$ is also positive semidefinite. Therefore, since we proved that $\mathrm{Var}(\tilde{\delta}|W) - \mathrm{Var}(\hat{\delta}|W) \geq 0$, that is, it is positive semidefinite, then its expectation should be positive semidefinite, which completes the prove.

B. Normal classical regression model

Let us now add an extra assumption:

- **Assumption 3 (normality)**: $y|W \sim \mathcal{N}(W\delta, \sigma^2 I_N)$, that is, we added the normality assumption to Assumptions 1 and 2.

In this case, we can propose to estimate δ by ML (which we know provides the BUE). The conditional likelihood function is:

$$L_N(\delta, \sigma^2) = f(y|W) = (2\pi)^{-\frac{N}{2}} \left(\sigma^{2N}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - W\delta)'(y - W\delta)\right), \quad (22)$$

and the conditional log-likelihood is:

$$\mathcal{L}_{\mathrm{N}}(\delta, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - W\delta)'(y - W\delta). \quad (23)$$

The first order conditions are:

$$\frac{\partial\,\mathcal{L}_{\mathrm{N}}}{\partial\delta} = \frac{1}{\sigma^2}W'(y - W\delta) = 0 \quad (24)$$

$$\frac{\partial\,\mathcal{L}_{\mathrm{N}}}{\partial\sigma^2} = \frac{1}{2\sigma^2}\left(\frac{(y - W\delta)'(y - W\delta)}{\sigma^2} - N\right) = 0, \quad (25)$$

which easily delivers that the maximum likelihood estimator of δ is the OLS estimator, and $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N}$. Therefore, we can conclude that, under the normality assumption, the OLS estimator is conditionally a BUE. We could prove, indeed, that $\sigma^2(W'W)^{-1}$ is (conditionally) the Cramer-Rao lower bound. Even though we are not going to prove it (it is not a trivial proof), unconditionally, there is no BUE. To do it, we would need to use the unconditional likelihood $f(y|W)f(W)$ instead of $f(y|W)$ alone.

Regarding $\hat{\sigma}^2$, similarly to what happened with the variance of a random variable, the MLE is biased:

$$\hat{u} = y - W\hat{\delta} = y - W(W'W)^{-1}W'y = (I - W(W'W)^{-1}W')y = My. \tag{26}$$

Similar to what happened in Chapter 5 (check the arguments there to do the proofs), M, which is called the residual maker, is idempotent and symmetric, its rank is equal to its trace, and equal to $N-K$, where K is the dimension of δ (because $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, and hence $\operatorname{tr}(W(W'W)^{-1}W') = \operatorname{tr}(I_K)$), and $MW = 0$. Therefore, $\hat{u} = My = M(W\delta + u) = Mu$. Hence:

$$\hat{u}'\hat{u} = (Mu)'Mu = u'M'Mu = u'Mu = \operatorname{tr}(u'Mu) = \operatorname{tr}(uu'M) = \operatorname{tr}(Muu'), \tag{27}$$

where we used the fact that $u'Mu$ is a scalar (and hence equal to its trace), and some of the tricks about traces used above. Now:

$$\begin{aligned}\mathbb{E}[\hat{u}'\hat{u}|W] &= \mathbb{E}[\operatorname{tr}(Muu')|W] = \operatorname{tr}(\mathbb{E}[Muu'|W]) = \operatorname{tr}(M\,\mathbb{E}[uu'|W]) \\ &= \operatorname{tr}(M\sigma^2 I_N) = \sigma^2 \operatorname{tr}(M) = \sigma^2(N-K). \end{aligned} \tag{28}$$

Hence, an unbiased estimator is $s^2 \equiv \frac{\hat{u}'\hat{u}}{N-K}$, and, as a result (easy to prove using the law of iterated expectations) an unbiased estimator of the variance of $\hat{\delta}$ is $\widehat{\operatorname{Var}}(\hat{\delta}) = s^2(W'W)^{-1}$.

Chapter 7: Hypothesis Testing and Confidence Intervals

By Joan Llull[*]

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

Main references:

— Mood: IX: 1, 2.1, 2.2, 3.1, 3.2, 3.3; VIIII: 1, 2.1-2.3

— Lindgren: 9.1-9.4, 9.8-9.12, 8.8

I. Hypothesis Testing

A ***statistical hypothesis*** is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. Consider a sample $(X_1, ..., X_N)$. What we want is to use this sample to see whether, with high enough chances of success, we can reject this hypothesis about the population that generated that sample.

The hypothesis that we are interested in testing is called the ***null hypothesis***, and is denoted by H_0. The null hypothesis describes a hypothetical data generating process. We call the ***alternative hypothesis***, denoted by H_1, or sometimes H_a, to the set of possible alternative hypothetical data generating processes that would be feasible if the null hypothesis was not true.

Statistical hypothesis testing is a method of statistical inference that compares our sample to a hypothetical sample obtained from an idealized model. The null hypothesis describes a specific statistical relationship between the two data sets. The comparison is deemed ***statistically significant*** if the relationship between the observed and hypothetical data sets would be an unlikely realization of the null hypothesis according to a threshold probability: the ***significance level***.

For example:

$$\begin{aligned} H_0: &\quad X_i \sim \mathcal{N}(0,1), \\ H_1: &\quad X_i \sim \mathcal{N}(\mu,1). \end{aligned} \tag{1}$$

In this example, we assume that the distribution of X_i is $\mathcal{N}(\cdot, 1)$, but we want to test whether the mean of the data generating process is equal to μ or equal to 0.

In this example, our hypothesis is called a ***simple hypothesis***, because we completely specified f_X (up parameter values). Alternatively, a ***composite hypothesis*** is any hypothesis that does not specify the distribution completely. The

[*] 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.

key difference if we knew that a simple hypothesis is satisfied, we would know the entire distribution of X_i, because one and only one distribution satisfies it; instead, infinite distributions satisfy a composite hypothesis.

A ***test statistic*** $C(X)$ is a statistic that summarizes the comparison between the sample and the hypothetical sample obtained from the idealized model. A ***statistical test*** is a procedure to discern whether or not the test statistic *unlikely have been generated by the model described by the null hypothesis.* The ***critical region*** or region of rejection, denoted by R_C, is the set of values of the test statistic for which the null hypothesis is rejected. The set of values of the test statistic for which we fail to reject the null hypothesis is often called the ***acceptance region***. The ***critical value*** is the threshold value of $C(X)$ delimiting the regions of acceptance and rejection.

II. Type I and Type II Errors

As a combination of random variables, a test statistic is a random variable. As such, with certain probability it can lead us to take wrong decisions. The following table summarizes the possible situations:

$H_0 \backslash C(X)$	$C(X) \in R_C$	$C(X) \in R_C^c$
true	Type I error	Ok
false	Ok	Type II error

Therefore, we define as ***Type I error*** the situation in which we reject a true null hypothesis, and ***Type II error*** is the situation in which we do not reject the null hypothesis despite even though it was false. The probability that Types I and II errors occur are relevant to judge how *good* is the test. We define the ***size*** of a test as $\alpha \equiv P_{H_0}(C(X) \in R_C)$, the probability of rejecting a correct hypothesis, i.e. the false positive rate.[1] The ***power*** of a test is $(1-\beta) \equiv P_{H_1}(C(X) \in R_C)$, the probability of correctly rejecting the null hypothesis, i.e. the complement of the false negative rate, β. In the above expressions, P_{H_i} indicates that the probabilities are computed using the cdf described by the hypothesis H_i. In a parametric test, we can define $\pi(\theta)$ as the function that gives the power of the test for each possible value of θ. This function is called the ***power function***. If θ_0 is the parameter indicated in H_0, then $\pi(\theta_0) = \alpha$. Finally, the ***significance level*** of a test is the upper bound imposed on the size of the test, that is, the value

[1] For composite hypothesis, the size is the supremum of the probability of rejecting the null hypothesis over all cases covered by the null hypothesis.

chosen by the statistician that determines the maximum exposure to erroneously rejecting H_0 he/she is willing to accept.

Note that there exists a tension between size and power. In the classical method of hypothesis testing, also known as Neyman-Pearson method, we give most of the importance to minimize the size of the test. But note that if we pick a critical value that takes α to zero, then the power of the test would be also zero.

For example, consider the following (one-sided) tests for the mean of a normal distribution. Assume that $X_i \sim \mathcal{N}(\mu, \sigma^2)$. The hypothesis we want to test is:

$$\begin{aligned} H_0: &\quad \mu = \mu_0, \\ H_1: &\quad \mu > \mu_0. \end{aligned} \tag{2}$$

We consider two situations, depending on whether σ^2 is known or not. If σ^2 is known, we know from Chapter 4 that:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1). \tag{3}$$

Hence, define the following statistic:

$$C \equiv \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}. \tag{4}$$

Under H_0, $\mu = \mu_0$, and, hence, $C \underset{H_0}{\sim} \mathcal{N}(0, 1)$, but under H_1, $C \sim \mathcal{N}(\theta, 1)$, where $\theta \equiv \frac{\sqrt{N}(\mu - \mu_0)}{\sigma}$, since C is a linear transformation of the statistic defined in Equation (7).

Now we consider a critical region: $R_\alpha = \{C > C_\alpha\}$. As we know that $\mu \geq \mu_0$ (these are all the possible cases included in the null and alternative hypotheses), the critical region is defined by the set of values that are so large that are unlikely to be obtained if $C \sim \mathcal{N}(0, 1)$, and hence constitute evidence against the null hypothesis, and in favor of the alternative hypothesis that we defined.

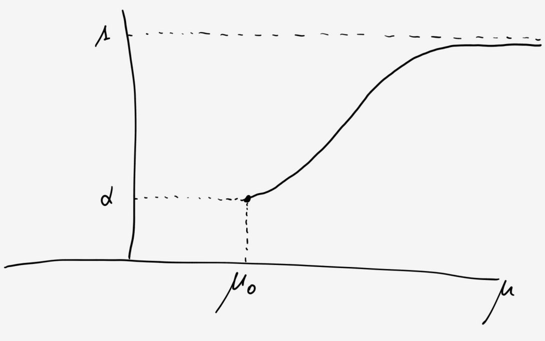
The critical value is the value C_α that satisfies:

$$P_{\mu_0}(C > C_\alpha) = \alpha = 1 - \Phi(C_\alpha) \quad \Rightarrow C_\alpha = \Phi^{-1}(1 - \alpha). \tag{5}$$

The power function is:

$$\pi(\mu) = P_\mu(C > C_\alpha) = 1 - \Phi(C_\alpha - \theta) = \Phi(\theta - C_\alpha). \tag{6}$$

Hence, the power function has the following shape:



If σ^2 is unknown, we also know from Chapter 4 that:

$$\frac{\bar{X}-\mu}{s/\sqrt{N}} \sim t_{N-1}. \tag{7}$$

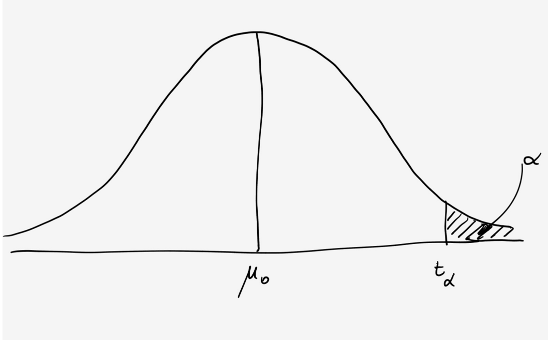
Hence, define the following statistic:

$$t \equiv \frac{\bar{X}-\mu_0}{s/\sqrt{N}}. \tag{8}$$

Under H_0, $\mu = \mu_0$, and, hence, $t \underset{H_0}{\sim} t_{N-1}$, independently of the value of σ^2. Now the critical value is the value t_α such that:

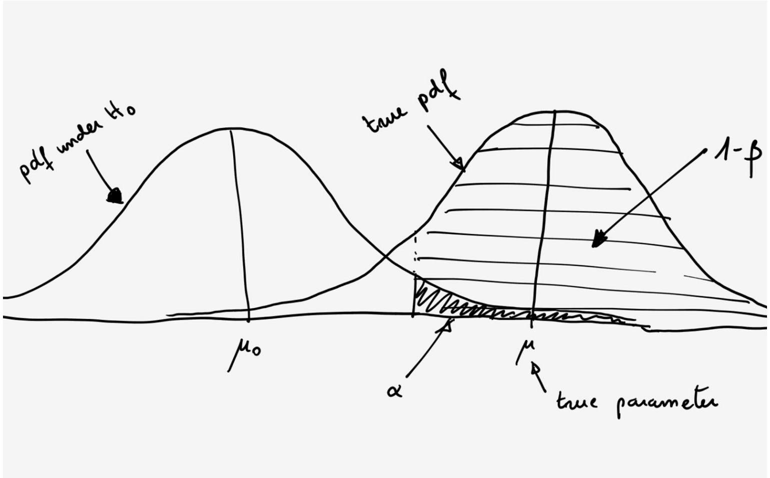
$$P_{\mu_0}(t > t_\alpha) = \alpha = 1 - F_t(t_\alpha). \tag{9}$$

Hence, we need the distribution tables for the Student-t distribution. Graphically, if H_0 is true:



III. Likelihood Ratio Test

Note that, in the previous example, we could define infinite different regions of size α (any region that contains 5% of the area below the pdf if H_0 is true). Then, the question is: why do we tend to consider the critical region in the tail(s)? The answer has to do with maximizing the power of the test. Intuitively, consider the following graphical representation:



By choosing the critical region at the tails, we are increasing the power of the test. In this Section, we formally prove that, by choosing the critical region at the tails (one tail for one-sided alternative hypotheses, the two tails for two-sided hypotheses, e.g. $\mu \neq \mu_0$) we maximize the power function). More generally, we analyze what is the best possible test in different situations.

First, we consider the case in which null and alternative hypotheses are simple:

$$\begin{aligned} H_0 :& \quad C(X) \sim F_0(\cdot), \\ H_1 :& \quad C(X) \sim F_1(\cdot). \end{aligned} \tag{10}$$

Let R_α and R'_α be two critical regions of size α:

$$P_{H_0}(C \in R_\alpha) = P_{H_0}(C \in R'_\alpha) = \alpha. \tag{11}$$

We say that R_α is ***preferred*** to R'_α for the alternative H_1 if:

$$P_{H_1}(C \in R_\alpha) > P_{H_1}(C \in R'_\alpha). \tag{12}$$

Therefore, among two tests with the same size, the one that has more power is preferred. More formally, the ***Neyman-Pearson lemma*** states that in the test of $F_0(\cdot)$ vs $F_1(\cdot)$ (or, equivalently, $f_0(\cdot)$ vs $f_1(\cdot)$), if a size α critical region, R_α, and a constant $k > 0$ exist, such that:

$$R_\alpha = \left\{ X : \lambda(X) = \frac{f_0(X)}{f_1(X)} < k \right\}. \tag{13}$$

then R_α is the most powerful critical region for any size α test of H_0 vs H_1. This lemma is very strong, even though it is also quite restrictive, as it requires both hypotheses to be simple (e.g. it is not applicable to $\mu = \mu_0$ vs $\mu > \mu_0$). $\lambda(X)$ is know as the ***likelihood ratio***. Small values of $\lambda(X)$ indicate small likelihood of H_0 and large likelihood of H_1, and, hence, we want those cases in the critical region. The critical region R_α defined in Equation (13) is known as the ***size*** α ***critical region of likelihood ratio***.

For example, consider a random sample obtained from a normal distribution with known variance σ^2. Consider the following hypotheses:

$$\begin{aligned} H_0 :& \quad \mu = \mu_0, \\ H_1 :& \quad \mu = \mu_1, \end{aligned} \tag{14}$$

with $\mu_1 > \mu_0$. This is the example drawn in the figure above. These hypotheses are simple, because the distributions under the null and alternative hypotheses are completely specified. The likelihood ratio is:

$$\lambda(X) = \frac{\prod_{i=1}^N \frac{1}{\sigma}\phi\left(\frac{X_i-\mu_0}{\sigma}\right)}{\prod_{i=1}^N \frac{1}{\sigma}\phi\left(\frac{X_i-\mu_1}{\sigma}\right)} = \exp\left(\frac{N}{2\sigma^2}\left[\mu_1^2 - \mu_0^2 - 2\bar{X}(\mu_1-\mu_0)\right]\right). \tag{15}$$

Now we need to find critical regions of size α for this test statistic. Since we do not directly know the distribution of $\lambda(X)$, we can transform it so that we have an expression in terms of something for which we can compute probabilities:

$$\begin{aligned} \lambda(X) < k \Leftrightarrow & \ln\lambda(X) < \ln k \\ & \Leftrightarrow -\frac{N}{\sigma^2}(\mu_1-\mu_0)\bar{X} < \ln k - \frac{N}{2\sigma^2}(\mu_1^2-\mu_0^2) \\ & \Leftrightarrow \bar{X} > \frac{\mu_1+\mu_0}{2} - \frac{\sigma^2 \ln k}{N(\mu_1-\mu_0)}, \end{aligned} \tag{16}$$

where in the last step we used the fact that $\mu_1^2-\mu_0^2 = (\mu_1-\mu_0)(\mu_1+\mu_0)$. Therefore, $\lambda(X) < k$ is equivalent to $C > C_\alpha$, which is what we did above, where, in this case, since we know σ^2, C is defined by Equation (8). This explains why we pick the critical region in the tail of the distribution: the statistic C does not depend on μ_1. Therefore, the critical region will be the same regardless of the alternative hypothesis.

The case of composite hypothesis is much more useful in practice. The simple hypotheses case, nonetheless, is useful because it allows us to implement the Neyman-Pearson lemma. It will also serve as a key ingredient in the implementation of the Neyman-Pearson lemma to the composite case.

Let H_0 and H_1 be composite hypotheses (the case in which only one of them is composite is a special case):

$$\begin{aligned} H_0 :& \quad \theta \in \Theta_0, \\ H_1 :& \quad \theta \in \Theta_0^c = \Theta \backslash \Theta_0, \end{aligned} \tag{17}$$

where Θ is the set of all possible parameters. A test with critical region R_α and power function $\pi(\theta)$ is ***uniformly more powerful*** for a size α if:

1) $\max_{\theta \in \Theta_0} \pi(\theta) = \alpha$, that is, it is of size α.

2) $\pi(\theta) \geq \pi'(\theta)$ for any $\theta \in \Theta$, and any test of size α and power function $\pi'(\cdot)$.

In general, uniformly more powerful tests do not exist because it is difficult that the second condition is satisfied in general. Therefore, there is no equivalent of the Neyman-Pearson lemma for the composite case. However, we proceed with an alternative: the ***generalized likelihood ratio test***, which defines the likelihood ratio as:

$$\lambda(X) = \frac{\max_{\theta \in \Theta_0} L(X;\theta)}{\max_{\theta \in \Theta} L(X;\theta)} = \frac{L(X;\hat{\theta}_0)}{L(X;\hat{\theta}_1)}. \tag{18}$$

This test statistic is very useful to test equality restrictions on the parameters. In this case, $\lambda = \frac{L(\hat{\theta}_r)}{L(\hat{\theta}_u)}$, where $\hat{\theta}_r$ and $\hat{\theta}_u$ indicate, respectively, the estimated coefficients for the restricted and unrestricted models.

To build the test, we need to know the distribution of $\lambda(X)$ or of a transformation of it. Interestingly, if the samples are large (see Chapter 8 for a reference) the distribution of $-2\ln\lambda$ is approximately χ^2.

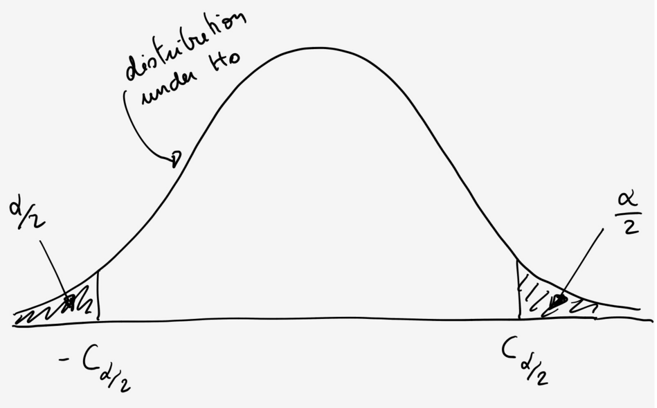
Finally, we say that a test is ***unbiased*** if the power under H_0 is always smaller that the power under H_1, and we say that the test if ***consistent*** if the power under H_1 tends to 1 when $N \rightarrow \infty$.

To illustrate all this, we retake some of the examples above, and we introduce some new examples. The first example is the one-sided test of the normal mean with known σ^2 described above. In this case, as the null hypothesis is simple, we can simply apply the Neyman-Pearson lemma for all the possible values of the alternative (i.e., the test $\mu = \mu_0$ vs $\mu > \mu_0$ is an infinite sequence of tests of the form $\mu = \mu_0$ vs $\mu = \mu_1$ for all $\mu_1 > \mu_0$). Therefore, $C = \frac{\sqrt{N}(\bar{X}-\mu_0)}{\sigma^2} > C_\alpha$ describes a test that is uniformly more powerful for a size α. The case of unknown σ^2 will be analogous, except that the test will be defined by $t > t_\alpha$, which will be distributed as a Student-t instead of a normal.

Consider, as a second example, the two tail test for the mean of a normal distribution with known variance, that is:

$$\begin{aligned} H_0 &: \quad \mu = \mu_0, \\ H_1 &: \quad \mu \neq \mu_0. \end{aligned} \tag{19}$$

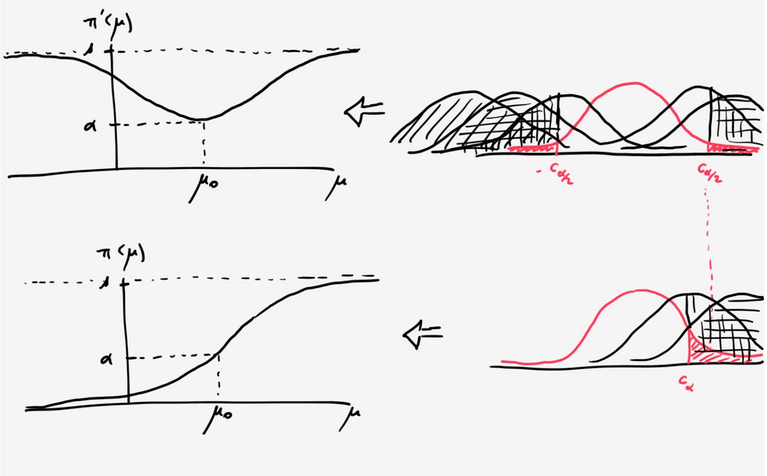
The critical region is still defined by the statistic C defined above, but now the critical region is defined by $|C| > C'_\alpha = C_{\alpha/2}$:



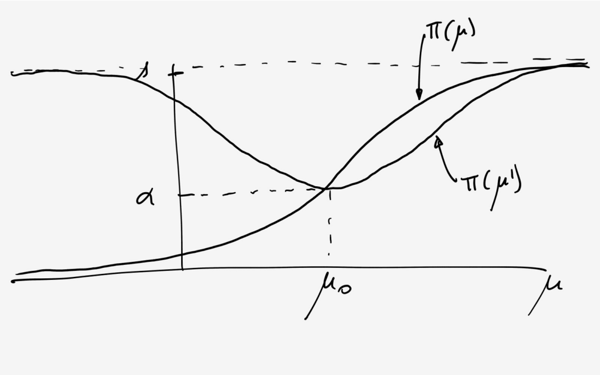
Note that, as before, the distribution under the alternative is given by $\mathcal{N}(\theta, 1)$ with $\theta \equiv \frac{\sqrt{N}}{\sigma}(\mu - \mu_0)$. Therefore, unlike in the previous case, the power function is:

$$\pi'(\mu) = P_\mu(|C| > C_{\alpha/2}) = 1 - [\Phi(C_{\alpha/2} - \theta) - \Phi(-C_{\alpha/2} - \theta)]. \tag{20}$$

To illustrate that a uniformly most powerful test does not exist, let us compare the $\pi'(\mu)$ with the power function $\pi(\mu)$ defined in Equation (6):



It is illustrative to put the two power functions in the same graph to compare:



Clearly, the one-sided test is preferred for the alternatives that imply $\mu > \mu_0$, and the two tail is preferred for $\mu < \mu_0$ (a one sided test of the type $\mu = \mu_0$ vs $\mu < \mu_0$ would be the most powerful one when $\mu < \mu_0$). The key here is that the one tail test is very good for the alternatives with $\mu > \mu_0$, but very bad for the alternatives with $\mu < \mu_0$. We would need to prove that $|C| > C_{\alpha/2}$ is the Neyman-Pearson critical region for the two-tail test, but we will not do it in class (it is strongly recommended as an exercise).

IV. Confidence Intervals

In Chapters 5 and 6, we provided very specific approximations to the parameter value of interest: ***point estimates***. However, it is useful to provide an approximation in the form of an interval. In this section, we provide an interval approximation to the true parameter value that we call ***confidence interval***. It is natural to study here, as they are closely related to hypothesis testing.

A confidence interval is defined by a pair of values $r_1(X)$ and $r_2(X)$ (or $r_1(\hat{\theta})$ and $r_2(\hat{\theta})$) such that $P(r_1(X) < \theta_0 < r_2(X)) = 1 - \alpha$, where α indicates the significance level as in the previous sections. In words, the confidence interval is a range of possible values for θ that, given the sample obtained, we infer contains the true parameter with probability $1 - \alpha$. Importantly, the functions that define the confidence intervals do not depend on the true parameter value.

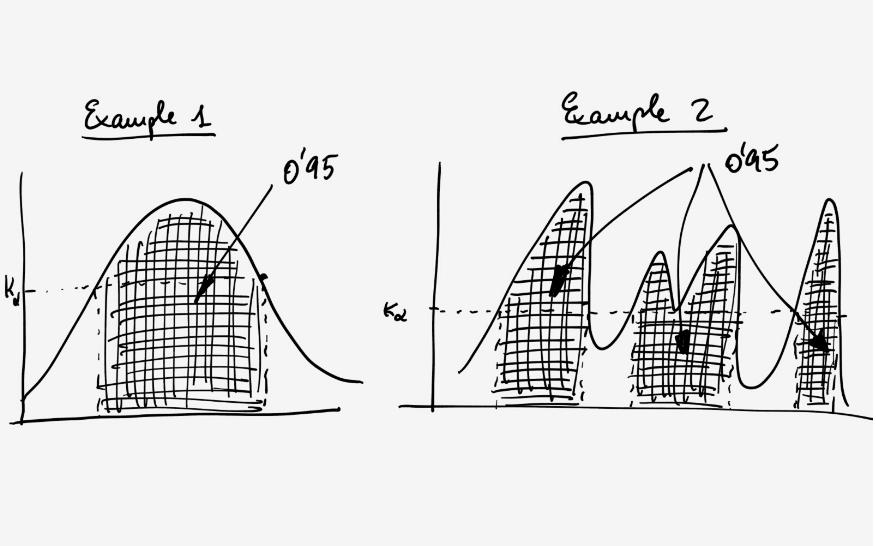
The confidence intervals are constructed in the exact same way that we find the critical value for a two-tail hypothesis test. If the distribution of $r(\hat{\theta})$ is symmetric and unimodal, the confidence intervals will typically be symmetric. One could also build one-sided confidence intervals (i.e., either $r_1(X) = -\infty$ or $r_2(X) = \infty$, but this practice is rather rare.

For example, in our example of the mean, if $C_{0.025} \approx 1.96$ (obtained from the tables of the normal distribution), the confidence interval for the mean would be $[\bar{X} - 1.96 * \frac{\sigma}{\sqrt{N}}, \bar{X} + 1.96 * \frac{\sigma}{\sqrt{N}}]$.

In Bayesian inference, we construct confidence intervals based on the posterior distribution. In that case, we define ***Bayesian confidence intervals*** grouping the regions of the posterior distribution that accumulate more density so that we accumulate density up to $1 - \alpha$. Thus:

$$R_{\theta,\alpha} \equiv \{\theta : h(\theta|X) > k_\alpha\}, \tag{21}$$

so that $P_h(\theta \in R_{\theta,\alpha} \geq 1 - \alpha)$. The interval does not need to be a contiguous set. In the following figure, you can see two examples, one in which the area is contiguous (Example 1), and one in which it is not (Example 2):



V. Hypothesis Testing in a Normal Linear Regression Model

A. Tests for Single Coefficient Hypotheses

Consider the normal linear regression model defined by Assumptions 1 through 3 in Chapter 6. Recall that:

$$\begin{aligned} &Y = W\delta + U, \\ &\hat{\delta} = (W'W)^{-1}W'y = \delta + (W'W)^{-1}W'U, \\ &Y|W \sim \mathcal{N}(W\delta, \sigma^2 I_N), \end{aligned} \tag{22}$$

which implies that:

$$(\hat{\delta} - \delta)|W \sim \mathcal{N}(0, \sigma(W'W)^{-1}). \tag{23}$$

For this model, we want to test hypotheses of the form:

$$\begin{matrix} H_0: & \delta_j = \delta_{j0}, \\ H_1: & \delta_j \neq \delta_{j0}, \end{matrix} \quad \text{or} \quad \begin{matrix} H_0: & \delta_j = \delta_{j0}, \\ H_1: & \delta_j > \delta_{j0}. \end{matrix} \tag{24}$$

Define the following statistic:

$$Z_j \equiv \frac{\hat{\delta}_j - \delta_j}{\sigma\sqrt{(W'W)^{-1}_{jj}}}, \tag{25}$$

where $(W'W)^{-1}_{jj}$ indicates the jjth element of the matrix $(W'W)^{-1}$. Cearly:

$$Z_j|W \sim \mathcal{N}(0, 1). \tag{26}$$

To derive the unconditional distribution of the statistic, we note that $f(Z_j|W)$ does not depend on W, and, thus, $f(Z_j|W) = f(Z_j)$. Hence, we can conclude that:

$$Z_j \sim \mathcal{N}(0,1). \tag{27}$$

If σ^2 is unknown, we follow the analogy for the sample mean derived in Chapter 4, and derive a t statistic:

$$t \equiv \frac{\hat{\delta}_j - \delta_j}{\widehat{s.e.}(\hat{\delta}_j)} \sim t_{N-K}, \tag{28}$$

where K is the size of the vector δ and $\widehat{s.e.}(\hat{\delta}_j) = s\sqrt{(W'W)^{-1}_{jj}}$ is the estimated standard error of the coefficient. To prove that $t \sim t_{N-K}$, we proceed as in Chapter 4. Dividing the numerator by the standard error, we obtain Z_j, which is distributed as a standard normal. More specifically:

$$Z_j = \frac{\hat{\delta}_j - \delta_j}{\sigma\sqrt{(W'W)^{-1}_{jj}}} = \frac{1}{\sqrt{(W'W)^{-1}_{jj}}}(W'W)^{-1}W'\tilde{U} \equiv P\tilde{U} \sim \mathcal{N}(0,1), \tag{29}$$

where $\tilde{U} \sim \mathcal{N}(0, I_N)$. Therefore, we can rewrite the t statistic as:

$$t = \frac{Z_j}{\sqrt{V/(N-K)}}, \tag{30}$$

where $V \equiv \frac{(N-K)s^2}{\sigma^2}$. Now:

$$V \equiv \frac{(N-K)s^2}{\sigma^2} = \frac{\hat{U}'\hat{U}}{\sigma^2} = \frac{U'MU}{\sigma^2} = \tilde{U}'M\tilde{U} \sim \chi^2_{N-K}, \tag{31}$$

where $M = (I_N - W(W'W)^{-1}W')$, which is symmetric, idempotent, and its rank is $N-K$ (the proof is exactly as in Chapter 4). Now we only need to prove that Z_j and V are independent, which we do by showing that $PM = 0$:

$$\begin{aligned} PM &= \frac{1}{\sqrt{(W'W)^{-1}_{jj}}}(W'W)^{-1}W'[I_N - W(W'W)^{-1}W'] \\ &= \frac{1}{\sqrt{(W'W)^{-1}_{jj}}}[(W'W)^{-1}W' - (W'W)^{-1}W'W(W'W)^{-1}W'] \\ &= \frac{1}{\sqrt{(W'W)^{-1}_{jj}}}[(W'W)^{-1}W' - (W'W)^{-1}W'] \\ &= 0. \end{aligned} \tag{32}$$

This completes the proof of $t|W \sim t_{N-K}$. As a final step, to derive the unconditional distribution, we note again that $f(t|W)$ does not depend on W, and, thus,

$f(t) = f(t|W)$, hence concluding that:

$$t \sim t_{N-K}. \tag{33}$$

Given all these test statistics and their distributions, we then proceed with inference in the same way as described in previous sections.

B. Tests for Multiple Coefficients Hypotheses

Consider the following test of linear restrictions:

$$\begin{aligned} H_0 : \quad & R\delta = R\delta_0, \\ H_1 : \quad & R\delta \neq R\delta_0. \end{aligned} \tag{34}$$

where R is a matrix of size $Q \times K$, where $Q \leq K$, and $\text{rank}(R) = Q$. This is general enough to test any linear combination of regressors. For notation compactness, we define $A \equiv (W'W)^{-1}$. We can write the following statistic:

$$F \equiv \frac{(\hat{\delta} - \delta)' R' [RAR']^{-1} R(\hat{\delta} - \delta)/Q}{s^2} \sim F_{Q,N-K}. \tag{35}$$

Note that the rank condition for R is necessary for RAR' to be invertible. The detailed proof is left as an exercise, but intuitively:

$$\begin{aligned} & \hat{\delta}|W \sim \mathcal{N}(\delta, \sigma^2 (W'W)^{-1}) \\ \Rightarrow \quad & R\hat{\delta}|W \sim \mathcal{N}(R\delta, \sigma^2 R(W'W)^{-1}R') \\ \Rightarrow \quad & R(\hat{\delta} - \delta)|W \sim \mathcal{N}(0, \sigma^2 R(W'W)^{-1}R'). \end{aligned} \tag{36}$$

Now note that the numerator is a combination of Q squared standard normals (provided that we divide by σ), that we should prove that are independent (in which case is a χ^2_Q), divided by the degrees if freedom Q. The denominator (once divided by σ) is a χ^2_{N-K} divided by the degrees of freedom $N - K$, as we have been doing for the t-test several times. Thus, proving that the χ^2 variables in the numerator and denominator are independent, we would complete the proof, given the definition of the F-distribution in Chapter 4.

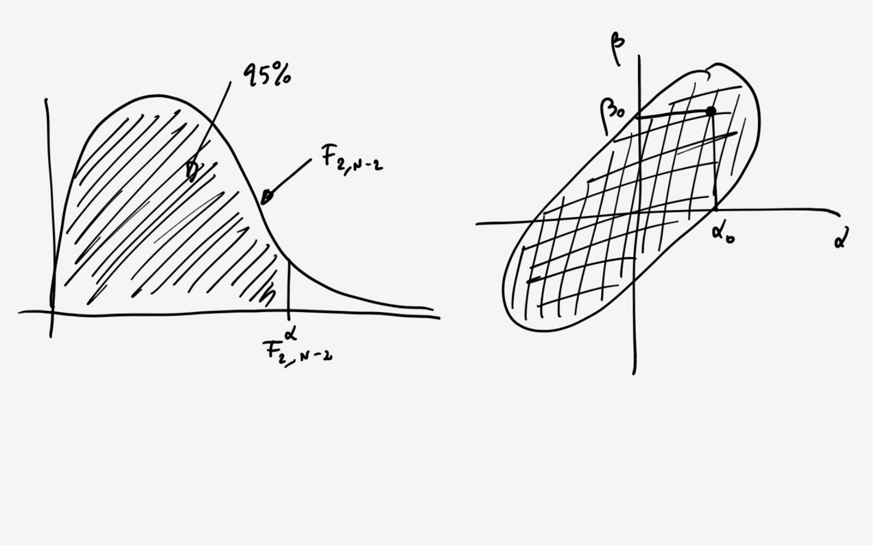
One particular application of this test is testing for the values of several coefficients simultaneously. For example, consider the following case for $\delta = (\alpha, \beta)'$:

$$\begin{aligned} H_0 : \quad & \alpha = \alpha_0 \text{ and } \beta = \beta_0, \\ H_1 : \quad & \alpha \neq \alpha_0 \text{ or } \beta \neq \beta_0. \end{aligned} \tag{37}$$

In this case, $R = I_2$. And, hence, the statistic boils down to:

$$F = \frac{(\hat{\delta} - \delta)' W'W (\hat{\delta}' - \delta)/2}{s^2} \sim F_{2,N-2}. \tag{38}$$

The following figure determines how we construct the confidence interval (given by $P[F < F^{\alpha}_{2,N-2}] = 1 - \alpha$):


95%
$F_{2,N-2}$
$F^{\alpha}_{2,N-2}$
β
β_0
α_0
α