Chapter 1: Descriptive Statistics

Joan Llull

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

joan.llull [at] movebarcelona [dot] eu

Introduction

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

Three types of data:

- Cross-sectional
- Time series
- Panel data

Two types of variables:

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

Frequency Distributions

Examples of kernels

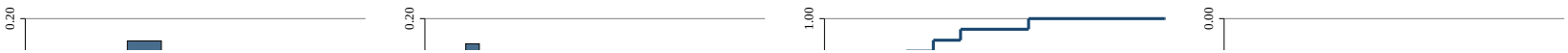
We build on a simple example: data for 2,442 **households** with information on **household gross income** in year 2010.

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

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

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

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



Kernel function

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

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

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

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

General conditions for kernels

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

- is symmetric,
- and integrates to 1.

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

Examples of kernels

Equivalent to what we did without the kernel:

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

Uniform kernel:

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

Gaussian kernel:

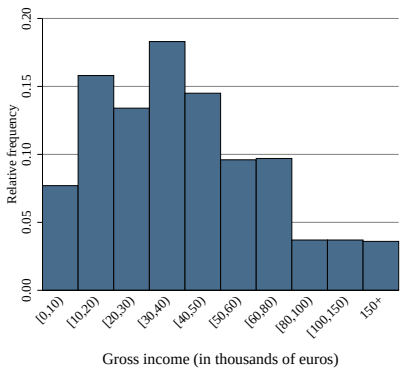
$$\kappa(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

Example

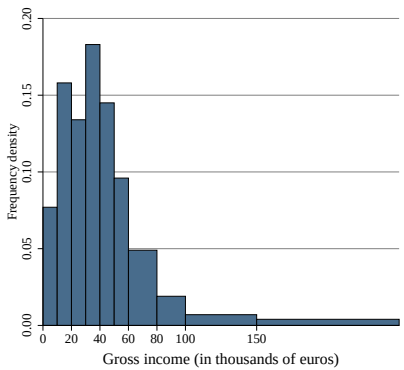
We build on a simple example: data for 2,442 **households** with information on **household gross income** in year 2010.

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

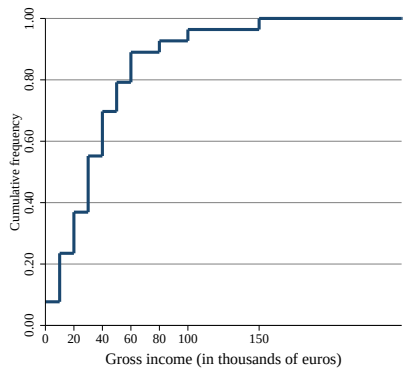
A. *Relative frequency*



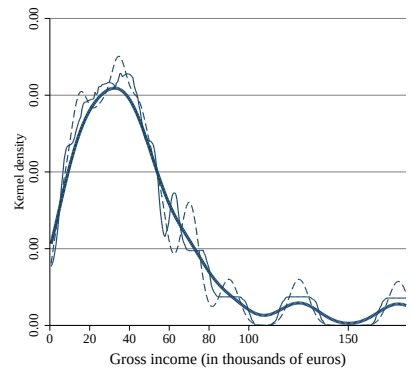
B. *Histogram*



C. *Cumul. frequency*



D. *Kernel density*



Summary Statistics

Arithmetic mean

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

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

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

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

Main problem: It is sensitive to extreme observations.

Median and mode

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

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

Main advantage: it is not sensitive to extreme values.

Main inconvenient: changes in the tails are not reflected.

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

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

Mean and median as loss minimizers

Loss function: is a function $L(\cdot)$ that satisfies $0 = L(0) \leq L(u) \leq L(v)$ and $0 = L(0) \leq L(-u) \leq L(-v)$ for any u and v such that $0 < u < v$.

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

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

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

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

Sample variance and standard deviation

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

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

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

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

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

Central moments

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

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

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

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

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

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

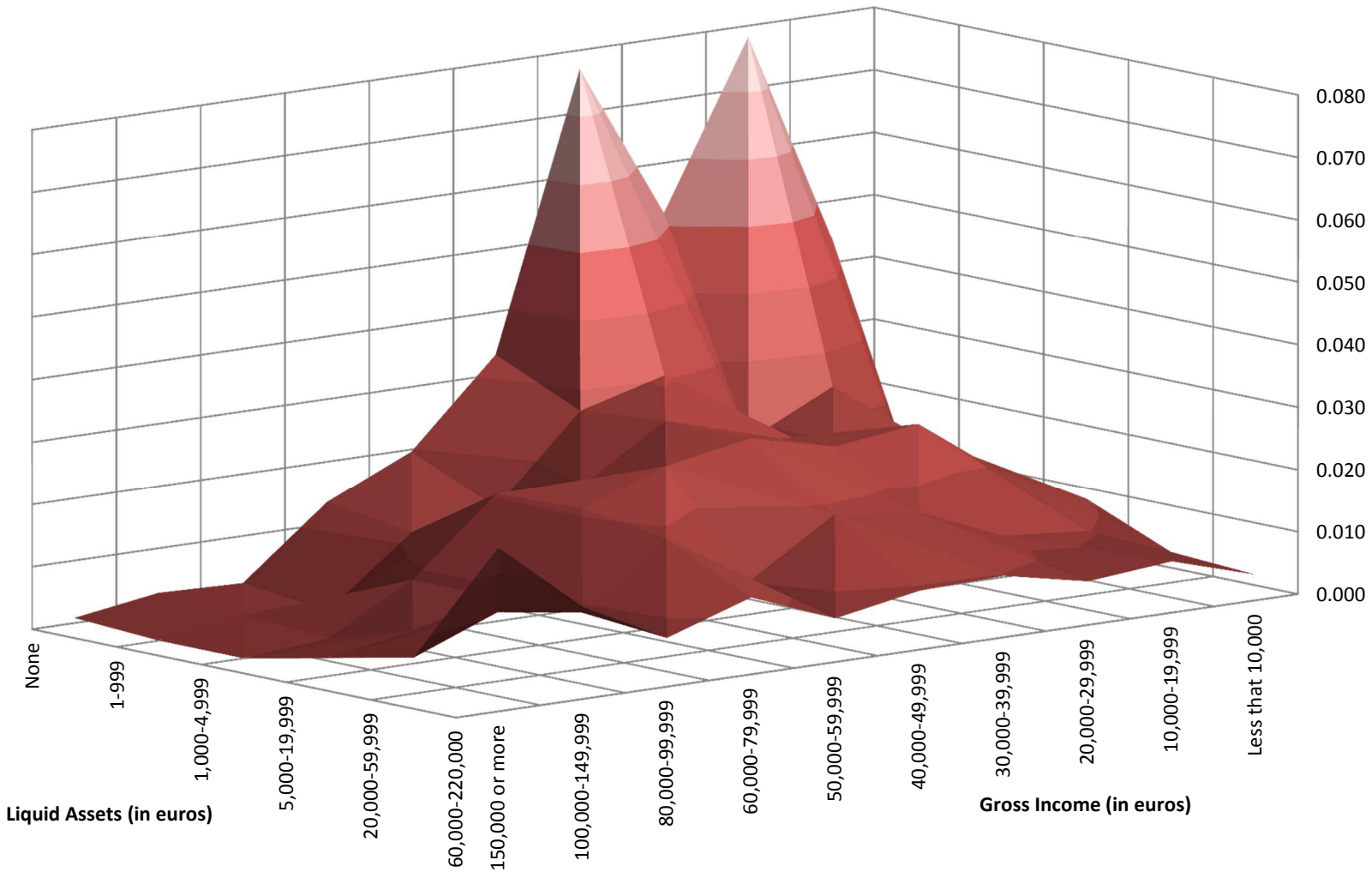
Fourth moment $\Rightarrow$ (excess) **kurtosis coefficient**:

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

Bivariate Frequency Distributions

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

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



Conditional relative frequencies

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

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

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

Conditional Sample Means

Conditional sample mean

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

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

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

In our example:

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

Kernel function

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

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

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

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

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

Sample Covariance and Correlation

Sample variance and correlation

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

The **sample covariance** is defined as:

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

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

The **correlation coefficient** is:

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

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

Chapter 7: Hypothesis Testing and Confidence Intervals

Joan Llull

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

joan.llull [at] movebarcelona [dot] eu

Hypothesis Testing

Hypothesis testing

Statistical hypothesis: a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables.

Null hypothesis: our main hypothesis of interest, denoted by H_0.

Alternative hypothesis: the set of possible alternative hypothetical data generating processes that would be feasible if the null hypothesis was not true, denoted by H_1.

Statistical hypothesis testing: 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.

Statistically significant comparison: the relationship between the observed and hypothetical data sets would be an unlikely realization of the null hypothesis according to a threshold probability.

Significance level: threshold of probability that allows to reject a hypothesis.

Hypothesis testing (cont'd)

Simple hypothesis: we completely specify f_X (up parameter values).

Composite hypothesis: any hypothesis that does not specify the distribution completely.

Test statistic: is a statistic that summarizes the comparison between the sample and the hypothetical sample obtained from the idealized model, denoted by $C(X)$.

Statistical test: a procedure to discern whether or not the test statistic *unlikely* have been generated by the model described by the null hypothesis.

Critical region or region of rejection: the set of values of the test statistic for which the null hypothesis is rejected, denoted by R_C.

Acceptance region: set of values of the test statistic for which we fail to reject the null hypothesis.

Critical value: threshold value of $C(X)$ delimiting the regions of acceptance and rejection.

Type I and Type II Errors

Type I and Type II errors

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

Size: $\alpha \equiv P_{H_0}(C(X) \in R_C)$.

Power: $(1 - \beta) \equiv P_{H_1}(C(X) \in R_C)$.

Power function: in a parametric test, the power under each possible value of θ, denoted by $\pi(\theta)$. If the parameter indicated by H_0 is θ_0, then $\pi(\theta_0) = \alpha$.

Significance level: upper bound imposed on the size of the test (i.e. maximum exposure to Type I error.

There is a **tension** between size and power.

(illustrate with one-sided normal example)

Likelihood Ratio Test

Likelihood ratio test

More on the **size vs power trade-off**: Why do we tend to consider critical regions at the tails of the distribution?

Consider the case in which **null and alternative hypotheses are simple**:

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

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

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

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

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

i.e., given the same size, the one that has more power is preferred.

Neyman-Pearson lemma if a size α critical region, R_α, and a constant $k > 0$ exist, such that:

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

then R_α is the most powerful critical region for any size α test of H_0 vs H_1.

(example with normal and two simple hypotheses)

Likelihood ratio test with composite hypotheses

Composite hypotheses are more useful:

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

A test with critical region R_α and power function $\pi(\theta)$ is **uniformly more powerful** for a size α if:

- $\max_{\theta \in \Theta_0} \pi(\theta) = \alpha$, that is, it is of size α.
- $\pi(\theta) \geq \pi'(\theta)$ for any $\theta \in \Theta$, and any test of size α and power function $\pi'(\cdot)$.

It is difficult that the second condition is **satisfied in general**.

Alternative: **generalized likelihood ratio test**:

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

Likelihood ratio test with composite hypotheses

Among other things, the generalized LR test is useful to test restrictions: $\lambda = \frac{\mathcal{L}(\hat{\theta}_r)}{\mathcal{L}(\hat{\theta}_u)}$.

Unbiased test: the power under H_0 is always smaller than the power under H_1.

Consistent test: the power under H_1 tends to 1 when $N \to \infty$.

(example with normal and two-tail test)

Confidence Intervals

Confidence intervals

So far we have been obtaining **point estimates**.

Alternatively, we can provide an interval approximation to the true parameter: **confidence interval**.

A confidence interval is defined by **a pair of values** $r_1(X)$ and $r_2(X)$ such that $P(r_1(X) < \theta_0 < r_2(X)) = 1 - \alpha$.

We can also construct **Bayesian confidence intervals**:

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

Hypothesis testing in a Normal Linear Regression Model

Tests for single coefficient hypotheses

We want to **test**:

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

If σ^2 is known, we define the following **statistic**:

$$Z_j \equiv \frac{\hat{\delta}_j - \delta_j}{\sigma\sqrt{(W'W)^{-1}_{jj}}} \quad Z_j \sim \mathcal{N}(0,1).$$

If σ^2 is unknown, we can define an alternative **statistic**:

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

Tests for multiple coefficients hypotheses

We want to **test**:

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

If σ^2 is known, we define the following **statistic**:

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

where $A \equiv (W'W)^{-1}$.

Chapter 6: Regression

Joan Llull

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

joan.llull [at] movebarcelona [dot] eu

Classical Regression Model

Introduction

We are interested in estimating $\mathbb{E}[Y|X]$ and/or $\mathbb{E}^*[Y|X]$.

Matrix version of $\mathbb{E}^*[Y|X]$:

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

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:

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

and if $\mathbb{E}^*[X_2|X_1] = \gamma + \delta X_1$, then:

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

Ordinary Least Squares (OLS)

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 $K \times 1$.

Analogy principle:

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

This estimator is called **Ordinary Least Squares**:

$$\hat{\beta} = \left[\sum_{\mathrm{i}=1}^{\mathrm{N}} (x_i - \bar{x}_N)(x_i - \bar{x}_N)' \right]^{-1} \sum_{\mathrm{i}=1}^{\mathrm{N}} (x_i - \bar{x}_N)(y_i - \bar{y}_N).$$

$$\hat{\alpha} = \bar{y}_N - \hat{\beta}' \bar{x}_N.$$

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

Albegraic properties of the OLS estimator

Let us introduce some compact notation:

- Parameter vector: $\delta \equiv (\alpha, \beta')'$.
- Vector of observations of Y: $y = (y_1, ..., y_N)'$.
- Matrix for other variables: $W = (w_1, ..., w_N)'$, where $w_i = (1, x_i')'$.

Then:

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

And the solution is:

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

or:

$$\hat{\delta} = (W'W)^{-1} W'y.$$

Note that we need $W'W$ to be full rank, such that it can be inverted (**absence of multicollinearity**).

Residuals and Fitted Values

From Chapter 3: **prediction error** $U \equiv y - \alpha - \beta' X = y - (1, X')\delta$.

Sample analog is called **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. $W'\hat{u} = 0$.
2. $\hat{y}'\hat{u} = 0$.
3. $y'\hat{y} = \hat{y}'\hat{y}$.
4. $\iota' y = \iota'\hat{y} = N\bar{y}$.

Var. decomp. and sample coeff. of determin.

As in the population, we can prove:

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

where $\widehat{\text{Var}}(z) \equiv N^{-1}\sum_{\text{i}=1}^{\text{N}}(z - \bar{z})^2$.

Given this, we write the **sample coefficient of determination** as:

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

Assumptions for the Classical Regression Model

In order to use the OLS estimator to obtain information about $\mathbb{E}[Y|X]$, we require additional assumptions $\Rightarrow$ **classical regression model**:

- **Assumption 1**: $\mathbb{E}[y|W] = W\delta \Leftrightarrow \mathbb{E}[y_i|x_1, ..., x_N] = \alpha + x_i'\beta \Leftrightarrow y \equiv W\delta + u, \mathbb{E}[u|W] = 0$. Two main implications:
  - **Linearity**: the optimal linear predictor and the conditional expectation function coincide.
  - **(Strict) Exogeneity**: $\mathbb{E}[y_i|x_1, ..., x_N] = \mathbb{E}[y_i|x_i]$. This implies $\text{Cov}(u_i, x_{kj}) = 0$ and $\mathbb{E}[u_i|W] = 0$ (satisfied by an i.i.d. sample).
- **Assumption 2**: $\text{Var}(y|W) = \sigma^2 I_N$ (**homoskedasticity**). Together with A1 it implies that $\text{Var}(y_i|x_1, ..., x_N) = \text{Var}(y_i|x_i) = \sigma^2$ and $\text{Cov}(y_i, y_j|x_1, ..., x_N) = 0$ for all $i \neq j$. It is satisfied by an i.i.d. sample.

Statistical Results and Interpretation

Unbiasedness and Efficiency

Given A1 and A2:

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

Gauss-Markov: under A1 and A2, OLS is a best (conditionally) linear estimator:

$$\forall\{\tilde{\delta} : \tilde{\delta} \equiv Cy, \mathbb{E}[\tilde{\delta}|W] = \delta\} \quad \Rightarrow \quad \text{Var}(\tilde{\delta}|W) \geq \sigma^2(W'W)^{-1}.$$

Note that this implies $\text{Var}(\tilde{\delta}) \geq \text{Var}(\hat{\delta})$!

Normal classical regression

- **Assumption 3**: $y|W \sim \mathcal{N}(W\delta, \sigma^2 I_N)$.

We can estimate δ and σ^2 by (conditional) **maximum likelihood** $\Rightarrow \hat{\delta}_{MLE} = \hat{\delta}_{OLS}$, and $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N}$.

$\Rightarrow$ OLS is **conditionally a BUE**.

For σ^2, there is **no conditional BUE**, as $\mathbb{E}[\hat{\sigma}^2|W] = \sigma^2 \frac{N-K}{N}$.

An **unbiased estimator** is $s^2 \equiv \frac{\hat{u}'\hat{u}}{N-K}$, and an unbiased estimator of the variance of $\hat{\delta}$ is $\widehat{\text{Var}}(\hat{\delta}) = s^2(W'W)^{-1}$.

Chapter 5: Estimation

Joan Llull

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

joan.llull [at] movebarcelona [dot] eu

Analogy Principle

Analogy Principle

Estimation problem: obtaining an approximation to a population characteristic θ_0 combining the information provided in a given sample.

Estimator: is a rule for calculating this approximation to the given quantity based on observed data, $\hat{\theta}(X_1, ..., X_N)$.

Estimate: the value obtained from implementing the estimation rule to the provided sample $\hat{\theta}(x_1, ..., x_N)$.

With some **abuse of notation**, we often denote by $\hat{\theta}$ both a given estimator and the corresponding estimate.

Analogy Principle: define in the sample a statistic that satisfies similar properties to those satisfied by the **true parameter** in the population.

Desirable Properties of an Estimator

Desirable properties 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.

Mean squared error (MSE):

$$MSE(\hat{\theta}) \equiv \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta)^2.$$

Bias: $\mathbb{E}[\hat{\theta}] - \theta \Rightarrow$ an estimator with no bias is an **unbiased estimator**.

Efficiency: among unbiased estimators, if one has lower variance than the other, we say it is more efficient.

Best Unbiased Estimator (BUE): it is the most efficient estimator of all possible estimators.

Best Linear Unbiased Estimator (BLUE): like the BUE, but on the class of linear estimators.

Moments and Likelihood Problems

Moments problem

Two equivalent conditions for the parameter of interest define a ***moments problem***:

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

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

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

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

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.

Likelihood problem

We assume that the population distribution is a known function, except for the parameters of interest, which are unknown:

- **Density** of X: $\{f(X;\theta) : \theta \in \Theta\}$, known.
- **Parameter space**: Θ (set of possible parameters values), known.
- **True parameter value**: $\theta = \theta_0$, the only unknown element.

Likelihood problem: the true parameter satisfies:

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

Score: $z(X;\theta_0) \equiv \frac{\partial \ln f(X;\theta_0)}{\partial \theta'}$.

Zero expected score condition: $\mathbb{E}[z(X;\theta_0)] = 0$.

The likelihood problem can also be seen as a **moments problem**.

Maximum Likelihood Estimation

The maximum likelihood estimator (MLE)

Likelihood function: $L_N(\theta) = \prod_{i=1}^{N} f(X_i; \theta)$.

Log-likelihood function:$\mathcal{L}_{\mathrm{N}}(\theta) \equiv \ln L_N(\theta) = \sum_{i=1}^{N} \ln f(X_i; \theta)$.

Maximum Likelihood Estimator (MLE):

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

This estimator is the **sample analog** of the condition that the true parameter θ_0 satisfies in the population.

Likelihood principle: approximate θ_0 by the value of θ that maximizes the likelihood of obtaining the sample that we observe.

First order conditions (sample analogs of zero expected score rule):

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

The Cramer-Rao Lower Bound

The Cramer-Rao lower bound

Fisher information: $I(\theta_0) \equiv \text{Var}(z(X;\theta_0))$ (intuition).

Information equality:

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

Cramer-Rao inequality: any unbiased estimator $\tilde{\theta}$ satisfies

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

$\Rightarrow$ an unbiased estimator with $\text{Var}(\tilde{\theta}) = I(\theta_0)^{-1}$ is the **BUE**.

When the sample size tends to infinity, the variance of the MLE tends to the Cramer-Rao lower bound $\Rightarrow$ **MLE is the BUE** if $N \to \infty$.

If there exists a BUE, it is the MLE.

Illustration with the **normal distribution**.

Bayesian Inference

Bayes' theorem

Recall the **Bayes theorem** in Chapter 3:

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

More generally, if we **partition** the sample space in N mutually exclusive sets that cover the entire sample space, $\mathcal{A}_1, ..., \mathcal{A}_N$:

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

Bayesian inference

Subjective probability: the probability function that describes our beliefs about the true probabilities of the different outcomes.

Inputs:

- **Likelihood** of the sample: $f_N(X|\theta)$.
- **A priori distribution**: $g(\theta)$.

Output: **a posteriori distribution** of the parameter given the information in the sample (Bayes theorem):

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

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. (**frequentist inference** vs **Bayesian inference**).

Probabilities associated to θ are in **Bayesian interpretation**: a quantity assigned to representing a state of knowledge or belief.

Bayesian estimation

In Bayesian estimation, we are primarily interested in obtaining a **posterior distribution** $h(\theta|X)$.

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:

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

Likewise:

- the **median** of the posterior distribution minimizes the expected absolute loss
- the **mode** maximizes the posterior density

Chapter 4: Sample Theory and Sample Distributions

Joan Llull

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

joan.llull [at] movebarcelona [dot] eu

Random Samples

Simple random samples

Our **population** is described by a **probabilistic model**.

The **data** are a set of realizations from the probabilistic model.

The process of obtaining the data is called **sampling**.
(e.g. what we did in Chapter 2 for finite sets)

Simple random sampling: a collection of random variables $(X_1, ..., X_N)$ is a simple random sample from F_X if:

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

and thus:

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

Sample Mean and Variance

Sample Mean

Statistic: single measure of some attribute of a sample.

Chapter 1, descriptive statistics. Now, we are using them to **infer** some characteristic of the population.

A statistic is a **random variable** $\Rightarrow$ **sample distribution**.

Sample mean: $\bar{X}_N \equiv \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} X_i$.

Some properties (regardless of the functional form of F_X):

- $\mathbb{E}[\bar{X}_N] = \mathbb{E}[X]$.
- $\mathrm{Var}(\bar{X}_N) = \frac{\mathrm{Var}(X)}{N}$ (**precision**).

Sample variance

Sample variance:

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

$\Rightarrow \mathbb{E}[\hat{\sigma}_N^2] = \frac{N-1}{N}\sigma^2$: expect less dispersion than in population.

Corrected sample variance:

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

- $\mathbb{E}[s_N^2] = \sigma^2$.
- $\mathrm{Var}(s_N^2) = \frac{2\sigma^4}{N-1} + \frac{\mu_4 - 3\sigma^4}{N}$.

Sample variance

"Ideal" sample variance:

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

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

$\Rightarrow$ This statistic cannot be computed without knowing μ.

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

Distribution of the sample mean

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $\bar{X}_N \sim \mathcal{N}(\mu, \sigma^2/N)$, and:

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

This would help in making inference about μ if we knew σ^2...

... but we don't know σ^2.

Use s_2^N instead $\Rightarrow$ s_N^2 **is a random variable**: we need to derive the distribution of the transformed random variable.

Some **intermediate steps** first.

Intermediate steps I

1. **Chi-squared**: 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:

$$\tilde{W} = \tilde{Z}_1^2 + ... + \tilde{Z}_K^2 = \tilde{Z}'\tilde{Z} \sim \chi_K^2.$$

   - The degrees of freedom (K): number of independent squared standard normals included.
   - The support of this distribution is $\mathbb{R}^+$.
   - $\mathbb{E}[\tilde{W}] = K$ and $\text{Var}(\tilde{W}) = 2K$.

2. Let $\tilde{X} \sim \mathcal{N}_N(0, \Sigma)$. Then:

$$\tilde{X}'\Sigma^{-1}\tilde{X} \sim \chi_N^2$$

Intermediate steps II

3. Let M be a size $K \times K$ matrix that:
   - is idempotent (satisfies $MM = M$),
   - symmetric (satisfies $M' = M$),
   - and has $\text{rank}(M) = R \leq K$.

   Then:
   - M is singular (with the only exception of $M = I$).
   - M is diagonalizable, and its eigenvalues are either 0 or 1.
   - 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.

   $\Rightarrow \text{tr}(M) = \text{rank}(M)$ (and thus always a natural number).

Intermediate steps III

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:

$$\tilde{Z}' M \tilde{Z} \sim \chi^2_R.$$

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 $Q \times K$ matrix such that $PM = 0$. Then $\tilde{Z}' M \tilde{Z}$ and $P\tilde{Z}$ are independent.

Student-t

Using these steps:

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

Student-t: Let $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_K$, with Z and W being independent. Then:

$$t \equiv \frac{Z}{\sqrt{\frac{W}{K}}} \sim t_K.$$

Some characteristics:

- $\mathbb{E}[t] = 0$.
- $\text{Var}(t) = \frac{K}{K-2}$ for $K > 2$.
- Symmetric with respect to zero, support is $\mathbb{R}$.
- When $K \to \infty$ it is similar to a normal.

Thus we can make **inference** on μ without knowing σ:

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

F distribution

F-distribution: 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:

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

Some characteristics:

- $\mathbb{E}[F] = \frac{Q}{Q-2}$ for $Q > 2$.
- $(t_K)^2 \sim F_{1,K}$.

Used to make **inference** about σ^2.

Bivariate and Multivariate Sampling

Bivariate and Multivariate Sampling

In a multivariate random sample, $(X_1, ..., X_N)$ are N realizations of a **random vector**.

All the results above apply.

We can also construct "corrected covariances".

Chapter 3:
Multivariate Random Variables

Joan Llull

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

joan.llull [at] movebarcelona [dot] eu

Joint and Marginal Distributions

Joint and marginal cdfs

Multivariate random variable: a vector that includes several (scalar) random variables:

$$X = (X_1, \ldots, X_K)'.$$

Joint cdf:

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

Marginal cdf:

$$F_i(x) \equiv P(X_i \leq x) = P(X_1 \leq \infty, ..., X_i \leq x, ..., X_K \leq \infty)$$

$$= F(\infty, ..., x, ..., \infty).$$

Joint pmfs and pdfs

Joint pmf (discrete var.): $P(X_1 = x_1, X_2 = x_2, ..., X_K = x_K)$.

Joint pdf (continuous var.): a joint pdf $f_{X_1...X_K}(z_1, ..., z_K)$ satisfies:

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

Properties of joint pdfs:

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

(examples with discrete and continuous variables)

Marginal pmfs and pdfs

Marginal pmf (discrete var.):

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

Marginal pdf (continuous var.):

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

or equivalently:

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

(examples with discrete and continuous variables)

Conditional Distributions and Independence

Conditional probability and independence

Let $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$. The probability that $\mathcal{A}$ occurs **given that** $\mathcal{B}$ occurred, denoted by $P(\mathcal{A} \,|\, \mathcal{B})$ is formally defined as:

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

Bayes' rule:

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

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

$\mathcal{A}$ and $\mathcal{B}$ are **independent** if (the three below are equivalent):

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

Conditional cdfs

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

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

If $\mathcal{A}$ is represented as a random vector X_2:

$$F_{X_1|X_2}(x|x_2) \equiv P(X_1 \leq x|X_2 = x_2)$$

if discrete, and:

$$F_{X_1|X_2}(x|x_2) \equiv \lim_{h \to 0} P(X_1 \leq x|x_2 + h \geq X_2 \geq x_2)$$

if continuous.

Conditional pmfs and pdfs

Conditional pmf (discrete variable):

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

Conditional pdf (continuous variable):

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

or implicitly through the cdf:

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

(check it is a well-defined pdf)

Factorization (application of the Bayes' rule):

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

Independence

Two random variables X_1 and X_2 are **independent** if and only if:

- $f(X_1|X_2)$ and $f(X_2|X_1)$ do not depend on 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**.

Corollary: $(X_1, ..., X_K)$ independent if and only if $F_{X_1...X_K}(x_1, ..., x_K)$ $\prod_{i=1}^{K} F_i(x_i)$.

Transformations of Random Variables

Transformations of random variables

Let:

- $(X_1, ..., X_K)'$: vector of independent random variables.
- $\{Y_i = g_i(X_i) : j = 1, ..., K\}$: transformed random variables.

Then $(Y_1, ..., Y_K)'$ are also **independent**.

Let:

- $(X_1, ..., X_K)'$: vector of continuous random variables with pdf $f_X(x)$.
- $Y = g(X)$: K-dimensional function with a unique inverse, $X = g^{-1}(Y)$, and $\det\left(\frac{\partial g^{-1}(Y)}{\partial Y'}\right) \neq 0$.

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

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

Multivariate Normal Distribution

Multivariate normal distribution

Multivariate normal: $X \sim \mathcal{N}_K(\mu_X, \Sigma_X)$. Pdf is given by:

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

where μ_K is a size K vector of parameters, and Σ_X is $K \times K$ positive definite and symmetric matrix of parameters.

It is a **transformation** of a vector of independent standard normals $Z = (Z_1, ..., Z_K)'$ with $\{Z_i \sim \mathcal{N}(0,1) : j = 1, ..., K\}$:

$$X = \mu_X + \Sigma_X^{\frac{1}{2}} Z.$$

Using **similar derivation**: $Y = a + BX \sim \mathcal{N}(a + B\mu_X, B\Sigma_X B')$.

Covariance, Correlation, and Conditional Expectation

Covariance and correlation: two random vars.

Covariance: $\mathrm{Cov}(X_1, X_2) \equiv \mathbb{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]$.

Properties:

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

Correlation coefficient:

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

Cauchy-Schwarz inequality: $\Rightarrow \rho^2_{X_1 X_2} \leq 1$.

Expectation of random vectors

Expectation of a random vector:

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

Similarly:

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

X_1 and X_2 are **independent** $\Rightarrow \mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\,\mathbb{E}[X_2]$.

Variance-covariance matrix

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

$$\text{Var}(X) \equiv \mathbb{E}[(X - \mu_X)(X - \mu_X)'] =$$

$$= \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_K) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_K) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_K, X_1) & \text{Cov}(X_K, X_2) & \dots & \text{Var}(X_K) \end{pmatrix}.$$

This matrix is **symmetric** and **positive-semidefinite**.

Retaking the **multivariate normal**: if $X \sim \mathcal{N}_K(\mu_X, \Sigma_X)$,

$$\mathbb{E}[X] = \mu_X, \quad \text{Var}(X) = \Sigma_X = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1K}\sigma_1\sigma_K \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots & \rho_{2K}\sigma_2\sigma_K \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K}\sigma_1\sigma_K & \rho_{2K}\sigma_2\sigma_K & \dots & \sigma_K^2 \end{pmatrix}.$$

Conditional expectation

Conditional expectation X_1 given X_2 (continuous):

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

(discrete):

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

In general, using the **Rienman-Stiltjes integral**:

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

Conditional variance is defined as:

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

Law of iterated expectations:

$$\mathbb{E}[\mathbb{E}[X_1|X_2]] = E[X_1].$$

Independence (revisited)

Three concepts:

1. ***Independence***: $F_{X_1|X_2}(x_1|x_2) = F_{X_1}(x_1)$ (or any of the equivalent specifications defined above).

2. ***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. (this relation is not symmetric).

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

From strongest to weakest: 1. $\Rightarrow$ 2. $\Rightarrow$ 3.

Application: bivariate normal distribution

In the multivariate normal case, independence, mean independence, and absence of correlation are **equivalent**.

They occur (for X_i and X_j) **if and only if** $\rho_{ij} = 0$.

Conditional distribution:

$$X = (X_1, X_2)' \sim \mathcal{N}_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

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

Linear Prediction

Expectations and prediction

$h(X)$: **prediction** of a variable Y based on the information in X.

$U \equiv Y - h(X)$: **prediction error**.

The expectation is the **optimal predictor**:

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

By extension, $\mathbb{E}[Y]$ is the **optimal constant predictor**.

Nice **decomposition**:

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

Population R^2:

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

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:

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

Solving for the first order conditions $\beta = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$ and $\alpha = \mathbb{E}[Y] - \beta\,\mathbb{E}[X]$, and, hence:

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

Properties:

- If expectation is linear (e.g. bivariate normal), $\mathbb{E}^*[Y|X] = \mathbb{E}[Y|X]$.
- $\mathbb{E}^*[c|X] = c$.
- $\mathbb{E}^*[cX|X] = cX$.
- $\mathbb{E}^*[Y + Z|X] = \mathbb{E}^*[Y|X] + \mathbb{E}^*[Z|X]$.
- $\mathbb{E}^*[Y|X_1] = \mathbb{E}^*[\mathbb{E}^*[Y|X_1, X_2]|X_1]$.

Optimal linear predictor (cont'd)

Goodness of fit statistic:

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

Notice that $0 \leq \rho^2_{XY} \leq R^2$.

Chapter 2: Random Variables and Probability Distributions

Joan Llull

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2016

joan.llull [at] movebarcelona [dot] eu

Preliminaries: An Introduction to Set Theory

Definitions

Consider a collection of objects, including all objects in consideration in a given discussion.

We start introducing some definitions:

- **Element** (or point): each object in our collection (ω).
- **Space** (universe, or universal set): the totality of all elements (Ω).
- **Set**: a partition of the space ($\mathcal{A}$, $\mathcal{B}$,... or $\mathcal{A}_1$, $\mathcal{A}_2$,...).
- **Index set**: for the second type of notation for sets, the collection of all possible indexes (Λ).
- **Venn diagram**: a diagram that shows all possible logical relations between a finite collection of sets

We denote that an element ω is part of a set $\mathcal{A}$ by $\omega \in \mathcal{A}$. To express the opposite, we use $\omega \notin \mathcal{A}$.

We can define sets **explicitly** ($\mathcal{A} = \{1, 2, 3, 4, 5, 6\}$) or **implicitly** ($\mathcal{A} = \{(x, y) : x \in \mathbb{R}, y \in \mathbb{R}^+\}$).

The set that includes no elements is the **empty set**, $\varnothing$.

Operators for sets

We define a list of operators for sets:

- **Subset**: when all elements of $\mathcal{A}$ are also in $\mathcal{B}$ ($\mathcal{A} \subset \mathcal{B}$ or $\mathcal{B} \supset \mathcal{A}$).
- **Equivalent set**: if $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{A}$ ($\mathcal{A} = \mathcal{B}$).
- **Union**: the set that consists of all points that are either in $\mathcal{A}$ or in $\mathcal{B}$ or in both ($\mathcal{A} \cup \mathcal{B}$ or $\bigcup_{\lambda \in \Lambda} \mathcal{A}_\lambda$, and we define $\bigcup_{\lambda \in \varnothing} \mathcal{A}_\lambda \equiv \varnothing$).
- **Intersection**: the set that consists of all points that are both in $\mathcal{A}$ and $\mathcal{B}$ ($\mathcal{A} \cap \mathcal{B}$ or $\bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$, and we define $\bigcap_{\lambda \in \varnothing} \mathcal{A}_\lambda \equiv \Omega$).
- **Set difference**: the set that consists of all points in $\mathcal{A}$ that are not in $\mathcal{B}$ ($\mathcal{A} \setminus \mathcal{B}$ or, when the context is clear, $\mathcal{A} - \mathcal{B}$).
- **Complement**: the complement of $\mathcal{A}$ with respect to the space Ω is the set that consists of all points in Ω that are not in $\mathcal{A}$, that is $\Omega \setminus \mathcal{A}$ ($\mathcal{A}^c$ or $\overline{\mathcal{A}}$).
- **Disjoint/mutually exclusive sets**: $\mathcal{A} \subset \Omega$ and $\mathcal{B} \subset \Omega$ are mutually exclusive if $\mathcal{A} \cap \mathcal{B} = \varnothing$. Subsets $\{\mathcal{A}_\lambda\}$ are mutually exclusive if $\mathcal{A}_\lambda \cap \mathcal{A}_{\lambda'} = \varnothing$ for every λ and λ' such that $\lambda \neq \lambda'$.

Operators for sets

- **Cartesian product**: the set of all ordered pairs (a, b) where $a \in \mathcal{A}$ and $b \in \mathcal{B}$ ($\mathcal{A} \times \mathcal{B}$).
- **Power set**: the power set of $\mathcal{A}$ is the set of all possible subsets of $\mathcal{A}$, including $\varnothing$ and $\mathcal{A}$ itself ($2^{\mathcal{A}}$ or $\mathcal{P}(\mathcal{A})$). If $\mathcal{A}$ includes n elements, $2^{\mathcal{A}}$ includes 2^n elements.
- **Finite and countable sets**: a finite set is a set that has a finite 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 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:
  - it includes $\mathcal{A}$.
  - if the subset $\mathcal{B} \subset \mathcal{A}$ is included in Σ, $\mathcal{B}^c$ is also included.
  - if a countable collection of subsets $\{\mathcal{A}_\lambda\}$ is included, its union $\bigcup_{\lambda \in \Lambda} \mathcal{A}_\lambda$ is also included.

Some properties of these operators

- **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, $\left(\bigcup\limits_{\lambda \in \Lambda} \mathcal{A}_\lambda \right)^c = \bigcap\limits_{\lambda \in \Lambda} \mathcal{A}_\lambda^c$, and $\left(\bigcap\limits_{\lambda \in \Lambda} \mathcal{A}_\lambda \right)^c = \bigcup\limits_{\lambda \in \Lambda} \mathcal{A}_\lambda^c$.
- $\mathcal{A} \setminus \mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$ 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 $\mathcal{A} \cup \mathcal{B} = \mathcal{B}$.
- $\mathcal{A} \times \mathcal{B} \neq \mathcal{B} \times \mathcal{A}$.

Statistical Inference, Random Experiments, and Probabilities

Statistical inference and random experiment

Statistical inference: the process of deducing properties or claims from a population of interest (as opposed to descriptive statistics).

Random experiment (or trial): conceptual description of the process that generated the data that we observed.

We call it **random** because though the process can be replicated under similar conditions, results are not known with certainty (there is more than one possible outcome).

$\Rightarrow$ e.g. tossing a coin

The definition of probability

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 $P(\cup_{\lambda=1}^{\infty} \mathcal{A}_\lambda) = \sum_{\lambda=1}^{\infty} P(\mathcal{A}_\lambda)$.

$\Rightarrow$ e.g. tossing a coin

Finite Sample Spaces and Combinatorial Analysis

Finite sample spaces

We focus now on **finite sample spaces**: $\Omega = \{\omega_1, ..., \omega_N\}$.

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.

We analyze **two cases**:

- All elements of the sample space are **equally likely**.
- The elements of the sample space are **not equally likely**.

Finite sample spaces w/ equally likely elements

In this case, the **probability** of each outcome is $\frac{1}{N}$,

Axiomatic definition: 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 $P(\mathcal{A}) = N(\mathcal{A})/N$.

Such function satisfies the **three axioms** and hence is a probability function.

Combinatorial analysis

The only part left is a **counting** exercise, to obtain $N(\mathcal{A})$ and N (e.g. tossing a coin twice).

To do it systematically, we will use **combinatorial analysis**. Define:

- $n! \equiv n(n-1)(n-2)...1 = \prod_{j=0}^{n-1}(n-j)$. We define $0! \equiv 1$.
- $(n)_k \equiv n(n-1)...(n-k+1) = \prod_{j=0}^{k-1}(n-j) = \frac{n!}{(n-k)!}$.
- **Combinatorial symbol (or n pick k)**: it is defined as:

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

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

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

Application: random sampling

Example: drawing a **sample** of size n from an **urn** with M **balls**, numbered from 1 to M.

Two ways of **drawing** a sample:

- with replacement $\Rightarrow \Omega = \{(z_1, ..., z_n) : z_1 \in \{1, ..., M\}, ..., z_n \in \{1, ..., M\}\}$.
- without replacement $\Rightarrow \Omega = \{(z_1, ..., z_n) : z_1 \in \{1, ..., M\}, z_2 \in \{1, ..., M\}\backslash\{z_1\} ..., z_n \in \{1, ..., M\}\backslash\{z_1, ..., z_{n-1}\}\}$.

Counting the number of elements in a set $\mathcal{A}$ composed of points that are n-tuples satisfying certain conditions: $N(\mathcal{A}) = N_1 \cdot N_2 \cdot ... \cdot N_n$.

- with replacement $\Rightarrow M^n$ different samples could possibly be drawn.
- without replacement $\Rightarrow M \cdot (M-1) \cdot ... \cdot (M-n+1) = (M)_n$ different samples.

Application: size of the power set

Counting elements of the power set:

- Every subset of Ω with n elements $\Rightarrow n!$ different 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.
- Since we know we can get $(M)_n$ different size-n samples without replacement we know $n!x_n = (M)_n$
- Thus $x_n = \dfrac{(M)_n}{n!} = \binom{M}{n}$, and the number of sets is $\displaystyle\sum_{n=1}^{M} x_n$.

$\Rightarrow$ Using the binomial theorem $\displaystyle\sum_{n=1}^{M} x_n = \sum_{n=1}^{M} \binom{M}{n} = 2^M$

Finite sample spaces w/o equally likely elements

We should define our **probability function** in a different way.

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:

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

For any **event** $\mathcal{A}$, define $P(\mathcal{A}) \equiv \sum_{\{j:\omega_j \in \mathcal{A}\}} p_j$.

This function also **satisfies the three axioms**, and hence is a probability function.

Definition of Random Variable and Cumulative Density Function

Random variable

Random variable:

a function $X : \Omega \to \mathbb{R}$

such that $\mathcal{A}_r \equiv \{\omega : X(\omega) \leq r\}$

satisfies $\mathcal{A}_r \subset \mathcal{F}$ for every real number r.

$\Rightarrow$ e.g. tossing a coin

Cumulative distribution function

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

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

Properties of a cdf:

- 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).

$\Rightarrow$ e.g. tossing a coin

Continuous and Discrete Random Variables

Discrete random variables

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:

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

(connection with Chapter 1)

Continuous random variables

We define a random variable as **continuous** if there exists a non-negative function $f_X(\cdot)$ such that:

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

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 x:

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

Continuous random variables

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.

Mixed random variables

Mixed random variable: continuous in a part of its domain, but with some points at which there is positive probability mass.

More formally, a random variable is mixed if its **cdf** is of the form:

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

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**.

Commonly Used Univariate Distributions

Discrete distributions

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

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

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

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

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

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

Uniform distribution

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

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

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

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

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

Standard normal distribution

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

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

and cdf given by:

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

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!).

Normal distribution

The standard normal distribution can be generalized by means of an **affine transformation**: if $Z \sim \mathcal{N}(0,1)$, then $X \equiv \mu + \sigma Z$.

This transformation is simply called the **normal distribution**, and is denoted by $\mathcal{N}(\mu, \sigma^2)$.

The cdf of the normal distribution is given by:

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

and its pdf is equal to:

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

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 $\pm\sigma$.

Transformations of Random Variables

Transformations of random variables

Let $Y \equiv g(X)$, with $X \sim F_X(\cdot)$. Then, $Y \sim ?$

If the support of Y is **discrete**:

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

If the support of Y is **continuous**, $g(\cdot)$ is invertible and differentiable, and $g'(\cdot) \neq 0$:

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

and:

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

Expectation and Moments

Expectation

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

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

Analogy with the **sample mean** described in Chapter 1 $\Rightarrow$ the expectation is the population equivalent to the sample mean.

The two expressions above can be unified using the **Riemann-Stieltjes integral**:

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

For $Y \equiv g(X)$:

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

Variance and central moments

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: $\text{Var}(X) \equiv \sum_a \left[(x_a - \mu_X)^2 P(X = x_a)\right]$,
- 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:

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

The **standard deviation** is defined as:

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

where the positive sign indicates that it is given by the positive root.

The $\mathbf{k}^{th}$ **central moment** of the distribution of X is defined as:

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

General properties of expectations

Let c be a constant, and $g(X)$ and $h(X)$ denote two arbitrary functions of a 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 value of X,

and:

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

Jensen's inequality

Jensen's inequality: Let X denote a random variable, and let $g(\cdot)$ denote a continuous and convex function.

Then:

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

The opposite is true if $g(\cdot)$ is concave.

If the function is strictly convex, the inequality holds strictly.

If the function is linear, then it is satisfied with equality.

Chebyshev's and Markov's inequalities

Chebyshev's inequality: any distribution satisfies:

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

where X is a random variable, μ_X is its mean, σ_X^2 is its variance, c is an arbitrary positive constant, and $k \equiv \frac{c}{\sigma_X}$.

It 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 $P(|X-\mu_X| \geq 3\sigma_X) \approx 1-0.9973 \leq \frac{1}{9}$.

Markov's inequality: for any positive constant c and nonnegative function $g(\cdot)$ (such that the expectation $\mathbb{E}[g(X)]$ exists):

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

Quantiles, the Median, and the Mode

Quantiles, the median, and the mode

τth quantile: minimum value of X below which there is a fraction τ of the density of the distribution:

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

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.

Median: the 0.5th quantile, $q_{0.5}$.

Mode: value of X that has the maximum density (or mass if X is discrete).