1 Hitchhiker Guide to the Machine Learning Course

In this document we review some preliminary concepts of linear algebra, and statistics used in the course *Machine Leaning*, and how these tools can be used in Pyhton. This is not supposed to be an exhaustive document on these topics, for more information you should refer to:

- Petersen, K.B., Pedersen, M.S., "The matrix cookbook", 2008;
- Bishop, C.M., "Pattern recognition and machine learning", 2006, Springer;
- Montgomery, D.C., Runger, G.C., "Applied statistics and probability for engineers", 2010, John Wiley & Sons.

1.1 Linear Algebra

Let us recall some basics on the use of linear algebra about derivation and the definition of an eigenvector problem.

1.1.1 Matrix Derivatives

Consider a dataset composed of a set of N inputs $\mathbf{x}_i := (x_{i1}, \ldots, x_{iD})$, with $x_{ij} \in \mathbb{R}$, each of which has dimension D and a target $t_i \in \mathbb{R}$, corresponding to input $\mathbf{x}_i$. The goal of the least square method is to find a column vector $\mathbf{w} := (w_1, \ldots, w_D)^\top$ that minimize the following loss function:

$$L(\mathbf{w}) := \frac{1}{2} \sum_{i=1}^{N} \left(t_i - \sum_{j=1}^{D} x_{ij} w_j \right)^2 . \tag{1.1}$$

Equivalently, we can use a matrix notation to express the same goal. More specifically,

we define the input matrix as:

$$X = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{ND} \end{pmatrix},$$

and the target vector as $\mathbf{t} = (t_1, \dots t_N)^\top$. The loss one want to minimize in the least square problem becomes:

$$L(\mathbf{w}) := \frac{1}{2}||\mathbf{t} - X\mathbf{w}||_2^2 = \frac{1}{2}(\mathbf{t} - X\mathbf{w})^\top(\mathbf{t} - X\mathbf{w}). \tag{1.2}$$

Let us verify that Equation (1.2) is equivalent to Equation (1.1). By the definition of product between two matrices we have:

$$r_i = (\mathbf{t} - X\mathbf{w})_i := t_i - (x_{i1}, \dots, x_{iD}) \begin{pmatrix} w_1 \\ \vdots \\ w_D \end{pmatrix} = t_i - \sum_{j=1}^{D} x_{ij} w_j,$$

where $(\cdot)_i$ represents ith element of a vector and defining the residual vector $\mathbf{r} := (r_1, \dots, r_N)^\top$ we have:

$$L(\mathbf{w}) = \frac{1}{2}\mathbf{r}^\top\mathbf{r} = \frac{1}{2}\sum_{i=1}^{N}(r_i)^2 = \frac{1}{2}\sum_{i=1}^{N}\left(t_i - \sum_{j=1}^{D} x_{ij} w_j\right)^2,$$

which is the initial expression for the loss.

To minimize the loss we might compute the its derivatives w.r.t. each element of $\mathbf{w}$ and equal it to zero. To do that, we define the gradient as follows:

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} := \left(\frac{\partial L(\mathbf{w})}{\partial w_1}, \dots, \frac{\partial L(\mathbf{w})}{\partial w_D}\right).$$

Depending on the form of the loss we will be able to compute it in closed form or not. Usually, the matrix notation is less intuitive, but requires less computations to perform the same operations. Again, let us check the equivalence between the two forms. Consider an element of the gradient and the formulation of the loss in Equation (1.1). We have:

$$\begin{aligned} \left(\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}\right)_h &= \frac{\partial L(\mathbf{w})}{\partial w_h} = \frac{\partial}{\partial w_h}\left[\frac{1}{2}\sum_{i=1}^{N}\left(t_i - \sum_{j=1}^{D} x_{ij} w_j\right)^2\right] \\ &= \frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial w_h}\left[\left(t_i - \sum_{j=1}^{D} x_{ij} w_j\right)^2\right] = \sum_{i=1}^{N}\left[-x_{ih}\left(t_i - \sum_{j=1}^{D} x_{ij} w_j\right)\right]. \end{aligned}$$

Conversely, using the rule to derive matrix (which are similar to the one used to derive scalar functions), and using Equation (1.2) we have:

$$\begin{aligned}\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}}\left[\frac{1}{2}(\mathbf{t}-X\mathbf{w})^\top(\mathbf{t}-X\mathbf{w})\right] \\ &= -X^\top(\mathbf{t}-X\mathbf{w}).\end{aligned}$$

Let us compare this last derivatives with the previously derived one:

$$\begin{aligned}&\left(-X^\top(\mathbf{t}-X\mathbf{w})\right)_h = (-x_{1h},\ldots,-x_{Nh})(\mathbf{t}-X\mathbf{w}) \\ &= (-x_{1h},\ldots,-x_{Nh})\begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix} = \sum_{i=1}^{N} -x_{ih}\, r_i = \sum_{i=1}^{N}\left[-x_{ih}\left(t_i - \sum_{j=1}^{D} x_{ij}w_j\right)\right],\end{aligned}$$

which concludes our check.

Consequently, the minimum point of the previous loss function is computed looking at its stationary points:

$$\begin{aligned}X^\top(\mathbf{t}-X\mathbf{w}) &= 0 \\ X^\top X\mathbf{w} &= X^\top\mathbf{t}.\end{aligned}$$

1.1.2 Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{N\times N}$, its *eigenvalues* $\lambda_1,\ldots,\lambda_N$ and corresponding *eigenvectors* $\mathbf{v}_1,\ldots,\mathbf{v}_N$ are defined from the *eigenvector equations*:

$$A\mathbf{v}_i = \lambda_i\mathbf{v}_i,$$

for each $i \in \{1,\ldots,N\}$. Intuitively, an eigenvector $\mathbf{v}_i$ is a direction in the space that is only stretched when the transformation A is applied. The corresponding eigenvalue λ_i is the factor this vector is stretched. See Figure 1.1 for a simple example.

Using the matricial formulation, we have that to find the generic pair $(\lambda,\mathbf{v})$ it is possible to solve the following:

$$\begin{aligned}A\mathbf{v} &= \lambda\mathbf{v} \\ (A-\lambda I_N)\mathbf{v} &= 0,\end{aligned}$$

where I_N is the identity matrix of order N. The previous problem has a non-null solution only if the rank of the matrix $A-\lambda I_N$ is full, or, equivalently if its determinant is non-null. The equation corresponding to this condition is:

$$|A-\lambda I_N| = 0,$$

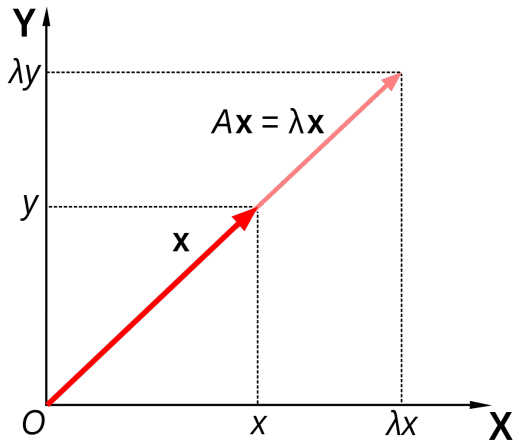


Figure 1.1: All the vectors in the direction of the eigenvector $\mathbf{x}$ are only stretched of a factor λ by the application of the linear transformation A. Vectors in other directions might also suffer from a change in terms of direction.

where we denoted the determinant of a matrix using $|\cdot|$. This is also known as the characteristic equation. Since it is a polynomial of order N in λ, it must have N solutions, which results in being the N eigenvalues of the matrix A.

Once we solved the previous equation, it is possible to show that the eigenvalues $\lambda_1, \ldots, \lambda_N$ satisfy the following properties:

- the rank of the matrix A is equal to the number of nonzero eigenvalues;
- the determinant of A is equal to the product of its eigenvalues:

$$|A| = \prod_{i=1}^{N} \lambda_i;$$

- the trace of A is equal to the sum of its eigenvalues:

$$Tr(A) = \sum_{i=1}^{N} \lambda_i.$$

Moreover, the positivity of the eigenvectors also determine some properties of the linear transformation. In the specific:

- a matrix A is said to be *positive definite*, if $\mathbf{x}^\top A\mathbf{x} > 0$ for all possible vectors $\mathbf{x} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$. Equivalently, a positive definite matrix has all positive eigenvalues, i.e., $\lambda_i > 0 \; \forall i$.

- a matrix A is said to be *semi-positive definite*, if $\mathbf{x}^\top A\mathbf{x} \geq 0$ for all possible vectors $\mathbf{x} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$. Equivalently, a semi-positive definite matrix has all non-negative eigenvalues, i.e., $\lambda_i \geq 0 \; \forall i$.

*** Exercise 1.1**

Show that the second derivatives of the previously considered quadratic form is:

$$\frac{\partial^2 L(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} = X^\top X.$$

1.2 Discrete Random Variables

A discrete random variable X is a variable with values in a discrete set E whose value is determined by a stochastic phenomenon, i.e., we are not able to predict its value even if we are given precise information about the phenomenon. For instance, consider a 20-faced dice: the event of throwing it can be modeled as a random variable X taking values in the finite set of events $E = \{1, \ldots, 20\}$, since we are not able to predict precisely which value might occur, even if we are given all the characteristics of the dice (e.g., dimensions, initial position, speed).

To properly model this phenomenon, we define a *probability function* $\mathbb{P} : E \to [0, 1]$ which tells you how often the event i belonging to a discrete set of events E occurs (e.g., probability that by trowing the dice you get 3) as:

$$\mathbb{P}(X = i) := \frac{|i|}{|E|},$$

where $|i|$ is the measure of the favorable events set and $|E|$ is the measure of set of the possible events. For instance, for a 20-faced dice we have:

$$\mathbb{P}(X = i) = \frac{1}{20},$$

since all the faces are equally likely to occur. In this case we have some properties that the probability function should have:

- $0 \leq \mathbb{P}(X = i) \leq 1$: an event can occur at least with zero probability and at most with probability one (the favorable cases are at most equal to the possible ones);
- $\sum_{i \in E} \mathbb{P}(X = i) = 1$: if we consider probability of the set of all the possible events E we should get one.

While this is the probability of a single events, we might be interested in the probability of multiple events. For instance, one might be interested in the probability that the dice

roll is small, say less than a specific value. This probability is captured by the *cumulative function* $F : E \to [0, 1]$, which specifies the probability that the random variable is lower than i:

$$F(i) := \mathbb{P}(X \leq i) = \sum_{h=1}^{i} \mathbb{P}(X = h) = \sum_{h \in E, h \leq i} \frac{|h|}{|E|}.$$

This newly defined function satisfies:

- $0 \leq F(i) \leq 1$: the sum of the probabilities should be between zero and one:
- $F(i) = 0, \ \forall i < \min_{h \in E} h$: if we consider a value small enough (smaller than the smaller element in the event space) the cumulative function should have value zero;
- $F(i) = 1, \ \forall i \geq \max_{h \in E} h$: if we consider a value large enough (larger or equal than the element in the event space) the cumulative function should have value one.

In the 20-faced dice case we have:

$$F(i) = \sum_{h=1}^{i} \frac{1}{20} = \frac{i}{20}.$$

There are two quantities which might be of major interest in the study of a random variable: the *expected value* and the *variance*. In the case of a dice, the former tells you what is value on average one could get from trowing the dice repeatedly, the latter gives us information about the spread in the single results we would have. Formally, for a generic random variable we have:

$$\mathbb{E}[X] := \sum_{i \in E} i \ P(X = i);$$
$$Var(X) := \sum_{i \in E} (\mathbb{E}[X] - i)^2 P(X = i).$$

In the case of the 20-faced dice we have:

$$\mathbb{E}[X] := \sum_{i=1}^{20} \frac{i}{20} = \frac{1}{20} \frac{20(20+1)}{2} = \frac{21}{2};$$
$$Var(X) := \sum_{i=1}^{20} \frac{\left(\frac{21}{2} - i\right)^2}{20} = \frac{57}{4}.$$

Sometimes the "spread" of a random variable is evaluated with the *standard deviation*, which is the square root of the variance $std(X) = \sqrt{Var(X)}$. This is due to the fact that if we are considering random variables expressed in some unit of measure, the variance is not compatible with the measurement itself, but its squared root does.

Remark 1. *Notice that these are the values obtained by knowing the random variable. If we only have some samples coming from the random variable, we are not able to compute the expected value and the variance, but we would be able to estimate their real values. We will see how in what follows.*

1.3 Continuous Random Variables

All the concept presented for a random variable X taking discrete values in a set E can be extended to the ones taking values in a continuous 1D set $\Omega \subseteq \mathbb{R}$. For instance, when we perform a length measurement, its values belongs to the interval $[0, +\infty)$. Similarly to what we did for discrete random variables with the probability function, we define the a *probability density function* (pdf) as follows:

$$f(x) := \lim_{\delta x \to 0} \frac{\mathbb{P}(x \leq X \leq x + \delta x)}{\delta x},$$

since, in this case, the probability of the event $X = x$ (having zero measure in a 1D space) is zero. In this case, the properties of the pdf are:

$$f(x) \geq 0 \ \ \forall x \in \Omega,$$
$$\int_{x \in \Omega} f(x)dx = 1.$$

A definition similar to what has been provided with the cumulative function for discrete variables can be provided also in the continuous case. When we want to evaluate the probabilities of intervals, we might resort to the *Cumulative Distribution Function* (CDF), defined as:

$$F(x) := \int_{s \in \Omega, s \leq x} f(s)ds,$$

having the following properties:

$$0 \leq F(x) \leq 1 \ \forall x \in \Omega,$$
$$F\left(\min_{x \in \Omega} x - \varepsilon\right) = 0,$$
$$F\left(\max_{x \in \Omega} x\right) = 1,$$

where $\varepsilon > 0$.

Similarly to the discrete case, the expected value and the variance are defined as:

$$\mu = \mathbb{E}[X] := \int_{x \in \Omega} xf(x)dx;$$
$$\sigma^2 = Var(X) := \int_{x \in \Omega} (\mathbb{E}[X] - x)^2 f(x)dx.$$

Among the most used continuous distributions we have the Gaussian one $X \sim \mathcal{N}(\mu, \sigma^2)$ defined over $\Omega = \mathbb{R}$ and having:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$
$$F(x; \mu, \sigma) = \int_{-\infty}^{x} f(t; \mu, \sigma) dt,$$

the uniform random variable $X \sim \mathcal{U}([0, 1])$ defined over $\Omega = [0; 1]$ and having:

$$f(x) = 1,$$
$$F(x) = x,$$

and the Beta random variable $X \sim Beta(\alpha, \beta)$ defined over $\Omega = [0; 1]$ and having:

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$ is a normalization term, and the CFD cannot be provided analytically.

1.4 Univariate Distributions in Python

In Python it is possible to define objects to handle different distributions in the package `scipy.stats`, among the others:

- `binom` Binomial distribution (discrete, `n` and `p` parameters);
- `norm` Gaussian distribution (continuous, `loc` and `scale` parameters);
- `unif` Uniform distribution (continuous, `loc` and `scale` parameters);
- `beta` Beta distribution (continuous, `a` and `b` parameters);

each of which will have specific parameters.

Remark 2. *Here the parameter `scale` for the Gaussian distribution is the standard deviation. In some other functions the second parameter of the Gaussian distribution is the variance. Check the documentation to know which one of the two is needed.*

If we want to get the pdf of a given point from a random variable with Gaussian distribution with mean 3 and standard deviation 4 we can write:

```
1 from scipy.stats import norm
2 norm.pdf(5, loc = 3, scale = 4) # pdf at x = 5
```

From the same distribution we can also get the value of the pdf function at a specific value, the value of the CDF function at a specific value, or the CDF function takes a specific value, in the following way:

```
norm.pdf(5, loc = 3, scale = 4) # pdf at x = 5
norm.cdf(3, loc = 3, scale = 4) # CDF at x = 3
norm.ppf(0.05, loc = 3, scale = 4) # x s.t. the CDF is 0.05
```

Another useful functionality available in Python allows to draw samples from a specific distribution. For instance, if we want to sample 100 realizations of the Gaussian variable X we could do:

```
norm.rvs(loc = 3, scale = 4, size = 100)
```

1.5 Multivariate Distributions in Python

Up to now we described only distributions whose domain was a subset of $\mathbb{R}$. There exists also some distributions which takes values in $\Omega \subseteq \mathbb{R}^n$, with $n \in \mathbb{N}$, $n > 1$. They are usually called *multivariate distributions*. If we want to resort to such random variable in Python we should use different tools. For instance, for the multivariate Gaussian we have:

```
from scipy.stats import multivariate_normal
multivariate_normal.pdf([0, 0], mean = [0.5, -0.2], cov = [[2.0, 0.3], [0.3,
    0.5]])
multivariate_normal.rvs(mean = [0.5, -0.2], cov = [[2.0, 0.3], [0.3, 0.5]],
    size = 100)
```

where $x = (0, 0)$ is the point considered, `mean` is the mean vector, `cov` is the covariance matrix, and `size` is the number of instances to be sampled.

For instance, if we want to sample 100 points from a 5-variate normal distribution with mean $[1\ 1\ 1\ 1\ 1]^T$ with identity covariance matrix and plot these points, we can use:

```
import numpy as np
multivariate_normal.rvs(mean = [1 1 1 1 1], cov = np.eye(5), size = 100)
```

1.6 Central Limit Theorem

In the previous sections we assumed that the distribution was known, i.e., that the parameters characterizing the distribution were known. Otherwise we built some *estimators* to infer them from data coming from the specified distribution. For instance, given a set of N independent samples $\{x_1, \ldots, x_n\}$ coming from the same distribution

(or formally, i.i.d. samples), the consistent estimators for the expected value and for the (sample) variance are:

$$\bar{X} := \frac{\sum_{i=1}^{n} x_i}{n},$$
$$s^2 := \frac{\sum_{i=1}^{n} (\bar{X} - x_i)^2}{n-1},$$

respectively.[1]

Let us focus on the empirical expected value $\bar{X}$. We recall the *Central Limit Theorem* (CLT):

Theorem 1 (Central Limit Theorem)**.** *Assume $\{X_1, \dots X_n\}$ is a sequence of independent and identically distributed (i.i.d.) random variables, with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, then:*

$$\sqrt{n}\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right) \to \mathcal{N}(0; \sigma^2),$$

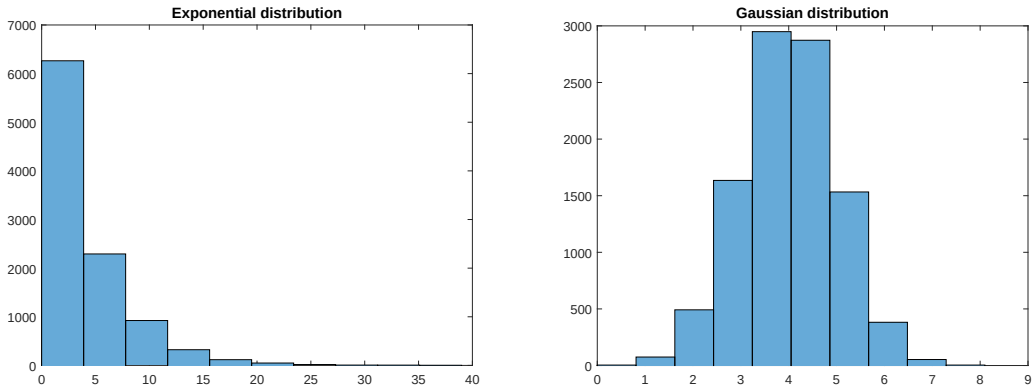
where the convergence holds in distribution.

To better understand this concept, let us sample from an *exponential* distribution with $\mu = 4$ and a *Gaussian* distribution with $\mu = 4$ and $\sigma = 1$. The histograms of the sampled distributions (we considered $n = 10000$ samples) are shown in Figure 1.2. Since the distributions have the same expected value, the empiric mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ will concentrate around this value. If we repeatedly sample from these distributions and consider the empiric mean obtained at each repetition we will have the results shown in Figure 1.3, which confirms that the distribution of the empirical mean is approximated by a Gaussian, as stated in the CLT. Moreover, we can see how the first one is more spread since the variance of the exponential distribution is 4 times the one of the Gaussian we considered.

1.7 Confidence Intervals

Once we have some estimates of the distribution parameters we would like to understand how much we should rely on them. For instance, if we used few data to estimate the true expected value $\mathbb{E}[X]$ it is likely that the true value might be far from the estimated one, while if we used N large enough we are more certain about the true value. Another factor determining how much we can rely on the estimator is the variance of the phenomenon itself: with high variance we will have samples which are more

[1] The term *consistent* means that if we have an infinite number of samples we would converge (in probability) to the true value of the parameter.



(a) Exponential (b) Gaussian

Figure 1.2: Histograms of the samples coming from two different distributions.

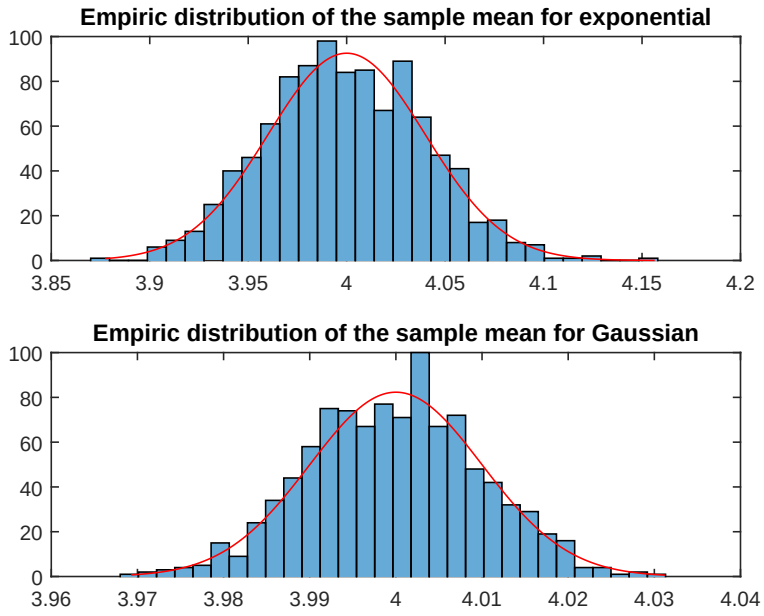


Figure 1.3: Histograms of the estimated means $\bar{X}$ coming from two different distributions.

likely to be far from each other and, thus, we are less certain about the value of the expected value (as it was shown in the previous section). Since we are in a stochastic environment, we need to set a level identifying that our estimator is "good enough". The probability that $\mathbb{E}[X]$ is exactly $\bar{X}$ is zero since the expected realization is a continuous random variable itself. Thus, we need to built some intervals, where we have high

confidence that the true mean $\mathbb{E}[X]$ is in.

Since we have the characterization of the distribution of the empirical mean, we can built intervals in which the expected value $\mathbb{E}[X]$ is with a specific confidence α. For instance, if we want to consider a confidence interval for the empirical mean, in the case we know the value of the standard deviation σ, with confidence at least $1-\alpha$, we have:

$$\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \tag{1.3}$$

where, $z_{\alpha/2}$ is the $1-\alpha/2$ quantile (i.e., the point where the CDF takes value $1-\alpha/2$ or $F^{-1}(1-\alpha/2)$) for a Normal distribution $\mathcal{N}(0,1)$, or equivalently:

$$\mathbb{P}\left(|\bar{X} - \mu| \geq \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) \leq \alpha.$$

Remark 3. *The previous inequality is true only under Gaussian assumption. Since we are provided a finite number of samples, the confidence intervals are only an approximations of the real ones if we are considering random variables which are not Gaussian distributed.*

Remark 4. *In the derivation of the confidence interval we assumed that the standard deviation σ was known. In the case we resort to its estimates s^2, there exist different confidence intervals for the mean of the Gaussian distribution (see the Montgomery book for more details).*

If we are considering unbounded domains (e.g., $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^+$), we also might resort to the following inequality to design confidence bounds:

Theorem 2 (Chebichev Inequality)**.** *Suppose X is random variables with $\mathbb{E}[X] = \mu < \infty$ and $Var[X] = \sigma^2 < \infty$, then:*

$$\mathbb{P}\left(|\mu - X| \geq \frac{\sigma}{\sqrt{\alpha}}\right) \leq \alpha.$$

which, if we consider the empirical mean $\bar{X}$ and a symmetric bound, leads to:

$$\bar{X} - \frac{\sigma}{\sqrt{n}\sqrt{\alpha}} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}\sqrt{\alpha}} \tag{1.4}$$

Remark 5. *The Chebichev inequality is more general w.r.t. the one derived from CLT, since it can be applied to generic random variables. In fact, CLT holds asymptotically for every random variable and exactly for Gaussian ones.*

Finally, if we also assume that the random variables have have finite support (e.g., $\Omega = [a,b]$ or $\Omega = \{0,1\}$), we might rely on:

Theorem 3 (Chernoff-Hoeffding bound)**.** *Assume to have a sequence $\{X_1, X_2, \ldots, X_n\}$ of n i.i.d. random variables with support in $[a, b]$ and $\mathbb{E}[X_i] = \mu, \forall i$, then for each $\varepsilon > 0$ we have:*

$$\mathbb{P}\left(|\bar{X} - \mu| \geq \varepsilon\right) \leq 2\exp\left\{-\frac{2n\varepsilon^2}{(b-a)^2}\right\} = \alpha.$$

This statistical bound leads to the following confidence intervals with confidence at least $1 - \alpha$ confidence:

$$\bar{X} - (b-a)\sqrt{\frac{-\log(\alpha/2)}{2n}} \leq \mu \leq \bar{X} + (b-a)\sqrt{\frac{-\log(\alpha/2)}{2n}}. \tag{1.5}$$

1.7.1 Hypothesis Testing

Sometimes we would like to make statement like:

The estimated parameter $\bar{X}$ is equal to μ.
The estimated parameter $\bar{X}'$ is different from another estimated parameter $\bar{X}''$.

In the case stochastic quantities are involved, the answer to such questions is the *test of hypotheses*. More specifically, we need to specify a null hypothesis H_0 and an alternative hypothesis H_1, e.g.,:

$$H_0 \ : \ \mu = \mu_0 \quad vs. \quad H_1 \ : \ \mu \neq \mu_0.$$

Given the data $\{x_1, \ldots, x_n\}$, we need that some of them support either the null H_0 or the alternative H_1 hypothesis. If we are able to compute an estimates for μ and we know its distribution, we are also able to say how likely is that the estimates has been drawn by the distribution. For instance, by considering the CLT we might say that:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow t = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

and, thus, that there is α probability that the test statistic t is lower than the quantile of order $\alpha/2$ of the standard Gaussian distribution or that it is larger than the quantile of order $1 - \alpha/2$. More formally:

$$\mathbb{P}(t < z_{\alpha/2} \vee t > z_{1-\alpha/2}) = \alpha.$$

If the test statistic has absolute value greater than $z_{1-\alpha/2}$ or smaller than $z_{\alpha/2}$ there is small (α) probability that the data are coming from a distribution where H_0 holds. Nonetheless, if we sample data from distribution in the case H_0 holds repeatedly we have that α of the times the test will say that the data are not coming from H_0. One

		Decision	
		Fail to reject H_0	reject H_0
True	H_0	Correct	Type I error
	H_1	Type II error	Correct

Table 1.1: Possible situations for a hypothesis test.

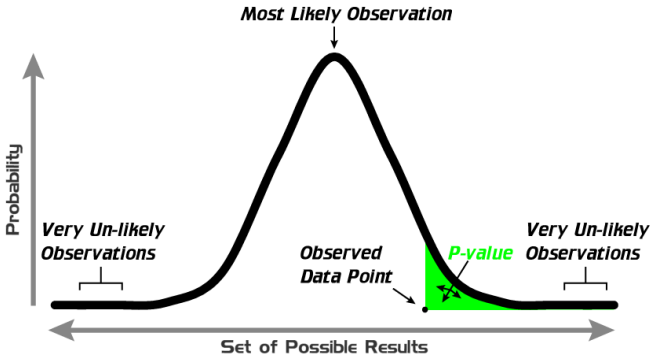


Figure 1.4: P-value in a right tailed test.

might think to use values of $\alpha \approx 0$ to reduce this probability and solve the problem. This would decrease the so called "type I error", but increase the "type II error", that even if the data are coming from a distribution for which H_1 holds we are not able to state that. Table 1.1 summarize all the possible situations.

The same reasoning used for the (two-tail) test could be used to develop test for hypothesis of the kind:

$$H_0 : \mu \leq \mu_0 \quad vs. \quad H_1 : \mu > \mu_0,$$
$$H_0 : \mu \geq \mu_0 \quad vs. \quad H_1 : \mu < \mu_0,$$

by considering only quantiles of order $1 - \alpha$ and α, respectively.

If we want to avoid to define a specific confidence α and let the data tell us how much we might be confident about their correspondence to a specific hypothesis, we could compute the p-value. The p-value is defined as the smallest confidence $\bar{\alpha}$ s.t. we are still able to reject the null hypothesis H_0. If we want to visualize the p-value in a right tailed hypothesis test (see Figure 1.4), it is the area under the distribution pdf and above the test statistics we computed

1.8 Frequentist vs. Bayesian Approach

The bounds we derived allow one to consider the cases in which we have only information about the bound of the variable which is considered. They do not allow to incorporate in a straightforward way information about the data distribution. In the case we have further information we might resort to a Bayesian approach for the parameter estimation. Indeed, by adopting the Bayesian framework, the value of the expected value of the random variable μ is a random variable itself. This is particularly interesting if we have information coming from the domain or from previously observed data.

For instance, let us say that we are considering a Bernoulli variable and we have some information coming from the past that tells us that a previously analysed phenomenon, **similar** to the one in analysis, had 3 successes over 10 trials. It would be wrong to consider these samples as drawn from the considered variable (i.e., using them to compute the empirical mean).

Consider the Bayes formula:

$$\begin{aligned}\mathbb{P}(\mu|x_1,\dots,x_t) &= \frac{\mathbb{P}(x_1,\dots,x_{t-1},x_t|\mu)\mathbb{P}(\mu)}{\mathbb{P}(x_1,\dots,x_t)}\\ &\propto \mathbb{P}(x_t|\mu)\mathbb{P}(x_1,\dots,x_{t-1}|\mu)\mathbb{P}(\mu)\\ &= \mathbb{P}(x_t|\mu)\mathbb{P}(x_{t-1}|\mu)\mathbb{P}(x_1,\dots,x_{t-2}|\mu)\mathbb{P}(\mu)\\ &= \mathbb{P}(\mu)\prod_{h=1}^{t}\mathbb{P}(x_h|\mu),\end{aligned}$$

where we assumed conditional independence of x_t from all the other data.[2] This way we are able to incorporate information starting from a prior distribution $\mathbb{P}(\mu)$ incrementally.

In the case we consider Bernoulli realizations, if we consider a Beta distribution as prior for the expected value μ ($\mu \sim \text{Beta}(3,7)$ in the example), we have that the posterior is still a Beta (i.e., Bernoulli and Beta are conjugate prior-posterior), thus allowing us to have an update rule for the next datum x_t.

Remark 6. *In the case we incorporate meaningful information the Bayesian learning process could be faster than the frequentist one. Though, if the information provided by the prior are misleading, the process could slow down and in some case prevent the estimation process to converge to the real value (e.g., if the prior assigns zero probability to the real value of the parameter).*

[2] We do not report the denominators since they do not change depending on μ, therefore they constitute only a normalization term for the posterior probability.