CMKL UNIVERSITY

AI + COMPUTER ENGINEERING

Probability and Statistics

Notes elaborated by

Dr. Irving Gómez Méndez

Fall 2025

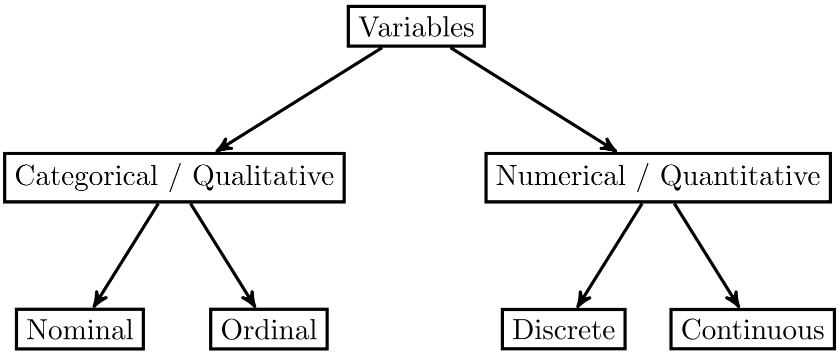
Contents

1 Descriptive Statistics

Descriptive analysis is a set of tools to summarize and describe data sets' main features.

Performing a descriptive analysis and an exploratory data analysis (EDA) are typically the first steps in a data analysis.

1.1 Types of variables



1.1.1 Categorical variables

Nominal variables

Variables that can be labeled or classified into mutually exclusive categories. These categories cannot be ordered in a meaningful way.

Examples of nominal variables:

- Gender: Female, male.
- Eyes color: Brown, blue, green.
- Species: Setosa, versicolor, virgnica.
- Country: Singapore, Thailand, Japan, Korea.

Ordinal variables

Variables that can be labeled or classified into natural, ordered categories. There is no appropriate way to take a distance between the categories.

Examples of ordinal variables:

- Education level: Uneducated, kindergarten, elementary school, high school, bachelor, postgraduate.

- Likert scale: Dislike, dislike somewhat, neutral, like somewhat, like.

Usually, it is necessary to code the categories as numerical values.

- Gender: 1, 2.
- Eyes color: 1, 2, 3.
- Species: 0, 1, 2.
- Country: 1, 2, 3, 4.

But we need to be careful, since the **scale doesn't mean anything**. For instance, if we code eyes color brown=1, blue=2, and green=3. It doesn't mean than two times brown color is equal to blue color.

1.1.2 Numerical variables

Discrete variables

Discrete variables can be obtained by counting, and the number of permitted values is either finite or countably infinite.

Examples of discrete variables:

- Number of kids: 0, 1, 2, ...
- Age in years: 0, 1, 2, ...
- Number of employees: 0, 1, 2, ...
- Number of girls in a group of 10 kids: 0, 1, 2, ..., 10

For numerical variables, including discrete variables, the scale is meaningful. For instance, having two kids is the double of having only one.

Continuous variables

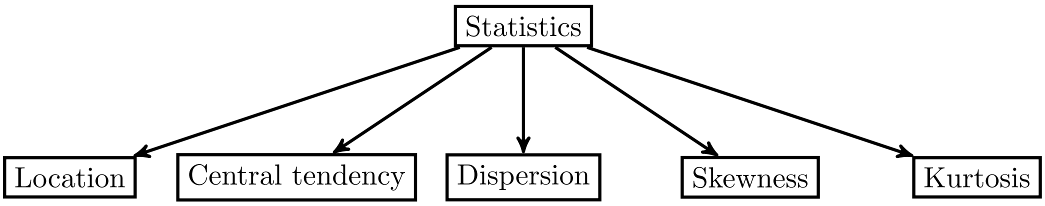
Continuous variables can take any value in an uncountable set.

Examples of continuous variables:

- Height: $(0,\infty)$
- Temperature: $(0,\infty)$
- Money in a bank account: $(-\infty,\infty)$

1.2 Statistics

A descriptive statistic is a summary that describes a features of a collection of information. Depending on the aspect that the statistics summarize, they can be divided in different categories.



1.2.1 Order statistics

Let be $\mathcal{D}_n = (x_1, x_2, \ldots, x_n)$ our observations, with $x_i \in \mathbb{R}$, $i = 1, \ldots, n$.

The kth order statistic of a statistical sample is equal to its kth-smallest value. It is usually denoted as $x_{(k)}$. That is, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

For example, assume that we have the sample:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1.8	1.4	1.2	1.0	0.8	0.1	4.6	2.4	1.3	1.3

The order statistics would be:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
0.1	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6

The first order statistic, $x_{(1)}$ is always the minimum of the sample, that is

$$x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}.$$

Similarly, for a sample of size n, the nth order statistic is the maximum, that is

$$x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}.$$

1.2.2 Empirical cumulative distribution function (ECDF)

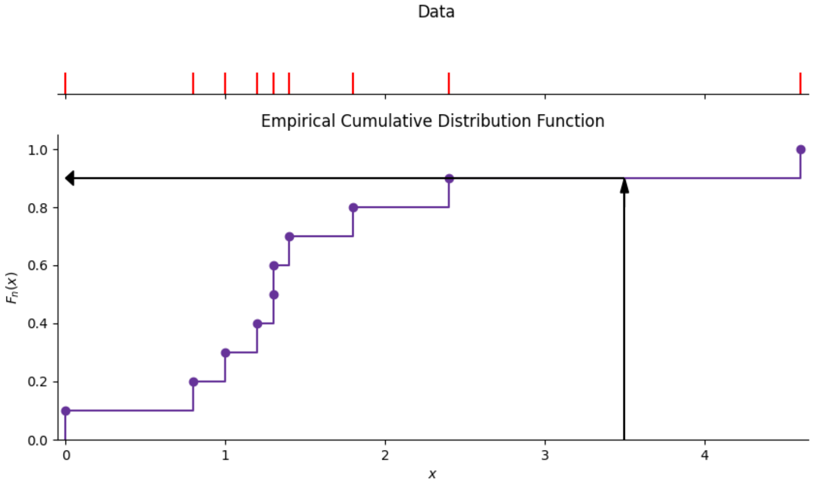
Using the observations $(x_1, x_2, \ldots, x_n)$, we define the empirical cumulative distribution function (ECDF) as:

$$F_n(x) = \frac{\text{number of elements in the sample} \leq x}{n}.$$

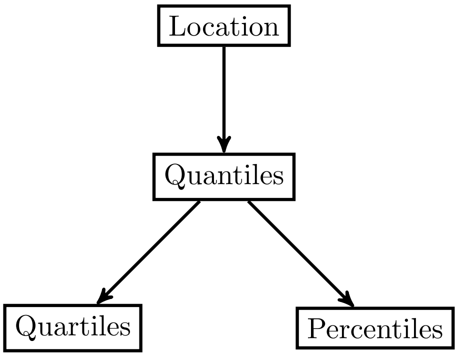
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1.8	1.4	1.2	1.0	0.8	0.1	4.6	2.4	1.3	1.3

	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
	0.1	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6
F_n	$1/10$	$2/10$	$3/10$	$4/10$	$5/10$	$6/10$	$7/10$	$8/10$	$9/10$	$10/10$

$$F_n(3.5) = \frac{\text{number of elements in the sample} \leq 3.5}{10} = \frac{9}{10}$$



1.2.3 Quantiles



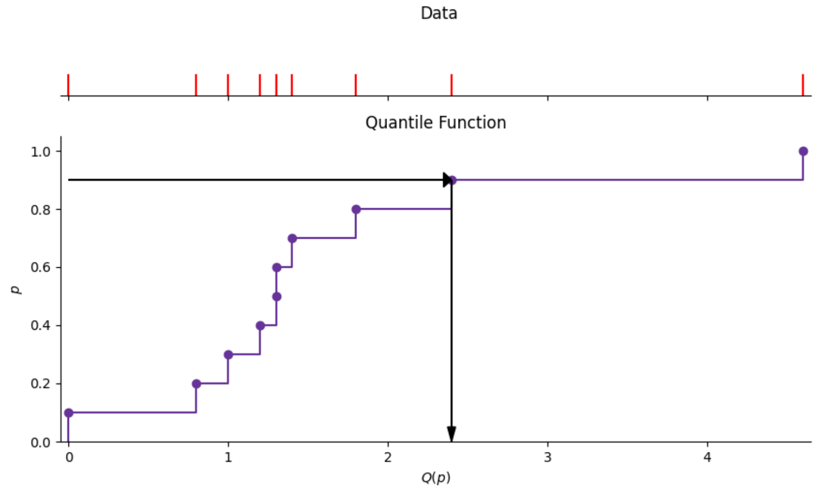
The (generalized) inverse function of the (empirical) cumulative distribution functions is called the (empirical) quantile function, also called the percent-point function, Q_p. It associates a point to a given probability.

$$Q_p : [0, 1] \mapsto x.$$

Q_p is called the quantile of probability p, and it is interpreted as the point where the variable has accumulated a probability equal to p.

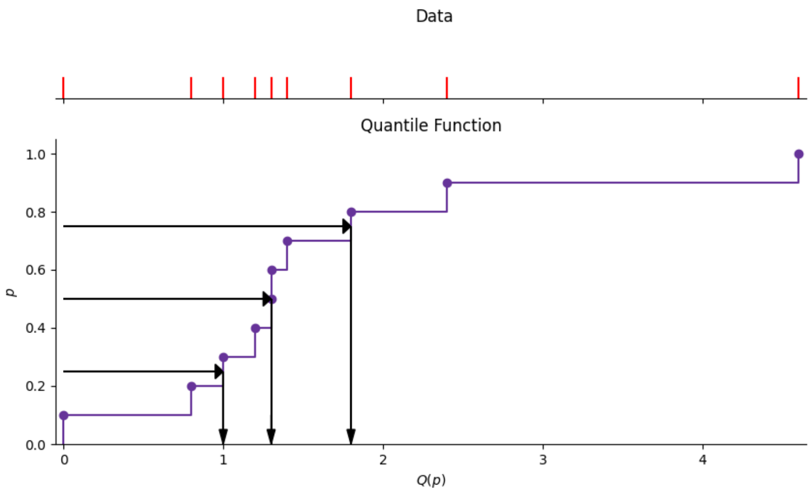
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1.8	1.4	1.2	1.0	0.8	0.1	4.6	2.4	1.3	1.3

	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
	0.1	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6
F_n	$1/10$	$2/10$	$3/10$	$4/10$	$5/10$	$6/10$	$7/10$	$8/10$	$9/10$	$10/10$



Quartiles

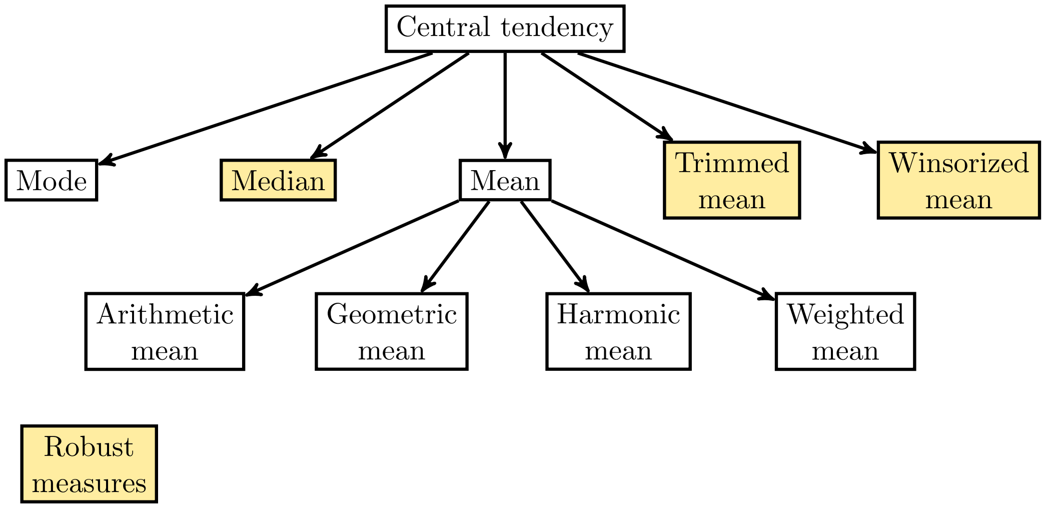
The quantiles of probability 0.25, 0.5, and 0.75 are of special interest. Such quantiles are called quartiles, and are usually denoted as Q_1, Q_2, and Q_3, respectively. In other words,25% of the observations are below or equal to Q_1, 50% of the observations are below or equal to Q_2, and 75% of the observations are below or equal to Q_3.



Percentiles

Similar to the quantiles, the percentiles correspond with quantiles of probability $0.01, 0.02, \ldots, 0.99$. That is, 1% of the observations are lower or equal to the first percentile, 2% are lower or equal to the second percentile, ..., 99% of the observations are lower or equal to the 99th percentile.

1.2.4 Central tendency



Mode

The mode is the value that appears more in the data. This statistic more suitable for categorical variables.

Median

The median, $\tilde{x}$, corresponds with the quantile of probability 0.5. That is, half of the observations are lower or equal to the median and half of the observations are greater than the median.

$$\operatorname{median}(\mathbf{x}) = \tilde{x} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd,} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{if } n \text{ is even.} \end{cases}$$

Arithmetic mean

Arithmetic mean, simply called mean or average:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Geometric mean

The numbers are restricted to be positive.

$$\bar{x}_{\text{GM}} = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log x_i\right\}.$$

The geometric mean is useful to calculate the mean interest rate.

Harmonic mean

The numbers are restricted to be positive.

$$\bar{x}_{\text{HM}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}.$$

The harmonic mean is useful to estimate the variance from a sample of variance estimators.

Weighted mean

To each observation in the sample, x_i, we associate a non-negative weight w_i.

$$\bar{x}_{\text{weighted}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

That is, the observations do not contribute equally to the final average.

Trimmed mean

Given a percentage $\alpha \in [0, 0.5)$, **we discard α of the lowest values and α of the largest values in the sample**. The α-trimmed mean, also called α-truncated mean, is then calculated with the remaining observations.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1.8	1.4	1.2	1.0	0.8	0.1	4.6	2.4	1.3	1.3

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
0.1	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6

Interquartile mean

The 25% trimmed mean is known as the interquartile mean.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1.8	1.4	1.2	1.0	0.8	0.1	4.6	2.4	1.3	1.3

In this example, we need to discard 2.5 observations from each side, which cannot be done. Thus, we can take the average between the 20% and the 30% trimmed means.

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
0.0	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6

20%-trimmed mean: $8/6 = 1.333$.

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
0.0	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6

30%-trimmed mean: $5.2/4 = 1.3$

$$\bar{x}_{\text{IQM}} = \frac{1.333 + 1.3}{2} = 1.317$$

Winsorizing

Winsorizing or winsorization is a transformation in which the extreme values of a sample data are replaced by a specified quantile of the data.

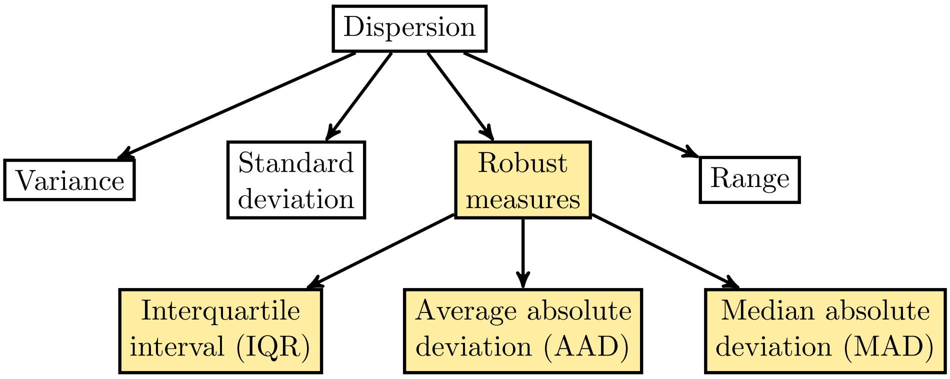
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1.8	1.4	1.2	1.0	0.8	0.1	4.6	2.4	1.3	1.3

Using the quantiles of probability 0.1 and 0.9 to winsorized, would replace the values of $x_{(1)}$ and $x_{(10)}$ for the values of $x_{(2)}$ and $x_{(9)}$, respectively.

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
0.1	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6
0.8	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	2.4

1.2.5 Dispersion

Dispersion statistics help to interpret the variability of data.


Dispersion
Variance
Standard deviation
Robust measures
Range
Interquartile interval (IQR)
Average absolute deviation (AAD)
Median absolute deviation (MAD)

Sample variance

The sample variance can be defined as two different expressions:

$$s^2_{\text{biased}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

If we know the real mean of the distribution, μ, we can redefine the sample variance as:

$$s^2_{\text{known } \mu} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

But, this is typically not the case.

Sample deviation

The standard deviation is defined as the squared root of the variance. Thus, the sample deviation can be defined as two different expressions:

$$s_0 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$s_1 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Range

$$R = x_{(n)} - x_{(1)}$$

Iterquartile interval (IQR)

Let be Q_1 and Q_3 the first and third quartiles, respectively. We define the interquartile interval, also called the interquartile range as

$$\mathrm{IQR} = Q_3 - Q_1$$

Median absolute deviation (MAD)

The median absolute deviation (MAD) is defined as

$$\mathrm{MAD} = \mathrm{median}(\mathbf{x} - \tilde{x})$$

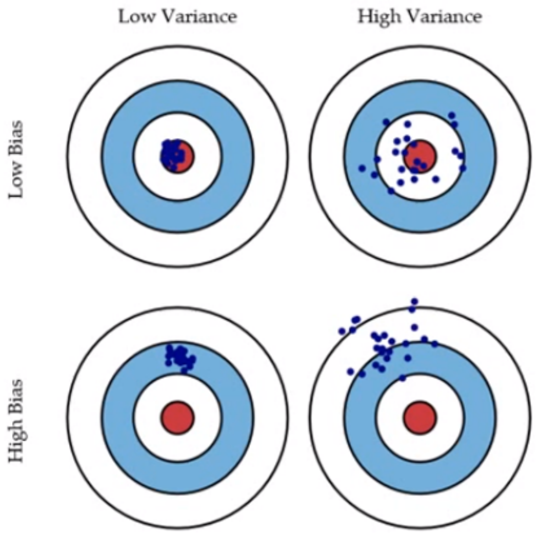
Average absolute deviation (AAD)

The average absolute deviation (AAD) is defined as

$$\mathrm{AAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Unbiased dispersion estimators

We say that an estimator is unbiased if its expected value corresponds with the population statistic.



Under normality assumptions, it can be shown that the next quantities are unbiased estimators for σ:

$$\sqrt{\frac{n}{n-1}}\frac{s_0}{c_4}, \quad \frac{s_1}{c_4}, \quad \frac{r}{d_2}, \quad \frac{\text{MAD}}{\Phi^{-1}(3/4)}, \quad \frac{\text{IQR}}{2\Phi^{-1}(3/4)}, \quad \text{AAD}\sqrt{\pi/2},$$

where

$$c_4 = \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma\left({}^{(n-1)}\!/_{2}\right)}, \quad d_2 = \int_{-\infty}^{\infty}\Big(1-(1-\Phi(x))^n-\Phi(x)^n\Big)dx,$$

and Φ is the cumulative distribution function of a standard normal variable.

1.2.6 Skewness

Empirical central moments

Let be

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^j, \quad j = 1, 2, \ldots$$

m_j is called the empirical (or sample) jth central moment. The theoretical jth central moment, μ_n, is defined as

$$\mu_j = \mathbb{E}(X-\mu)^j,$$

where $\mu = \mathbb{E}(X)$.

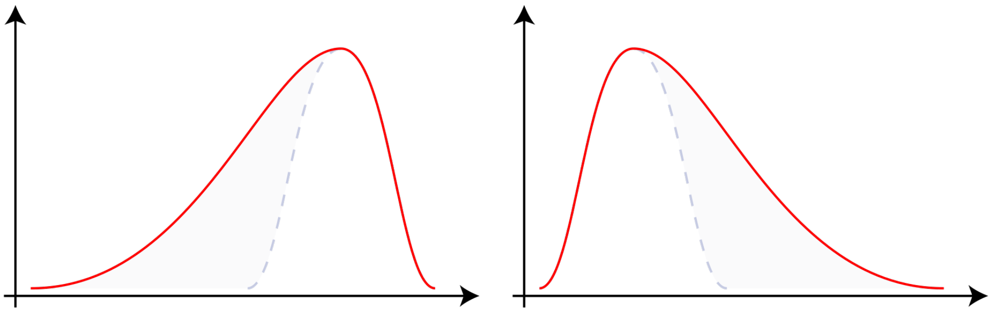
k-statistics

The nth k-statistic, k_n, is the minimum variance unbiased estimator of the nth cumulant, κ_n. The first four k-statistics are:

$$\begin{aligned} k_1 &= \bar{x} \\ k_2 &= \frac{n}{n-1}m_2 \equiv s^2_{\text{unbiased}} \\ k_3 &= \frac{n^2}{(n-1)(n-2)}m_3 \\ k_4 &= \frac{n^2[(n+1)m_4-3(n-1)m_2^2]}{(n-1)(n-2)(n-3)} \end{aligned}$$

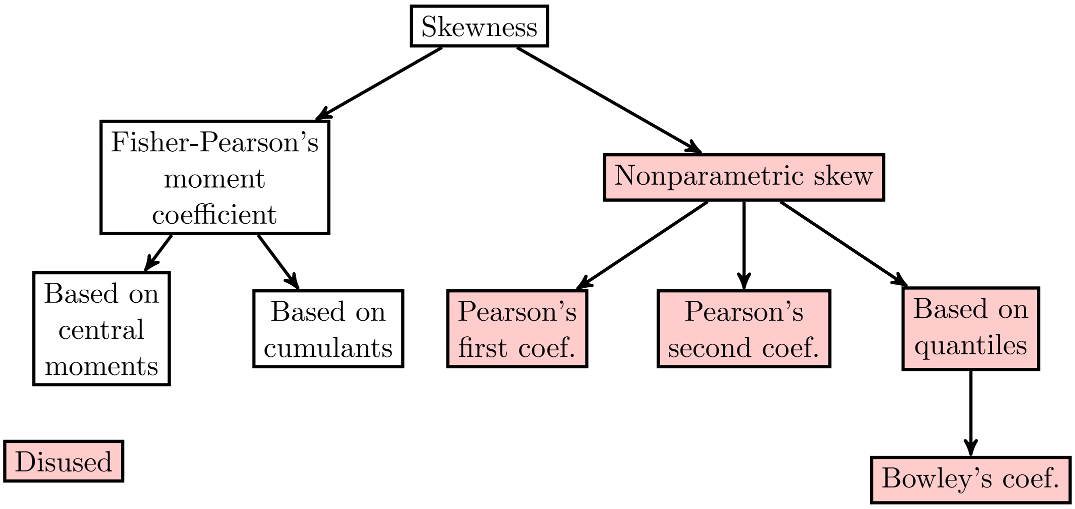
Skewness

Skewness is a measure of the asymmetry. The skewness value can be positive, zero, negative, or undefined.



Negative skew

Positive skew



Fisher-Pearson's moment coefficient of skewness. The Fisher-Pearson's moment coefficient can be written in two equivalent expressions, using the central moments or the cumulants:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}.$$

Thus, two natural estimates are:

$$g_1 = \frac{m_3}{m_2^{3/2}}, \quad G_1 = \frac{k_3}{k_2^{3/2}}.$$

Nonparametric skew. And old measurement of skewness (**currently in disused**) is the nonparametric skew, defined as

$$S = \frac{\mu - \tilde{\mu}}{\sigma}.$$

The next quantities estimate the nonparametric skew:

- Pearson's first coefficient of skewness:

$$S_{k_1} = \frac{(\bar{x} - \text{mode}(\mathbf{x}))}{s_2}$$

- Pearson's second coefficient of skewness:

$$S_{k_2} = \frac{(\bar{x} - \tilde{x})}{s_2}$$

- Bowley's coefficient:

$$S_{q_{3/4}} = \frac{Q_3 + Q_1 - 2\tilde{x}}{\text{IQR}}$$

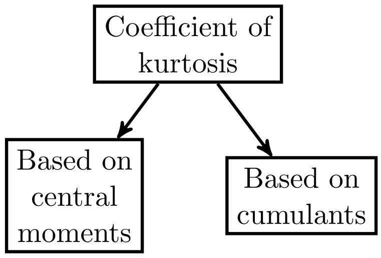
- Quantile coefficient of skewness: this is a generalization of Bowley's coefficient, given by

$$S_{q_p} = \frac{Q_p + Q_{1-p} - 2\tilde{x}}{Q_p - Q_{1-p}}$$

1.2.7 Kurtosis

Kurtosis

Kurtosis is a measure of the "tailedness". A higher kurtosis corresponds to greater extremity of deviations (or outliers).



The coefficient of kurtosis can be written in two equivalent expressions, using the central moments or the cumulants:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} = \frac{\kappa_4}{\kappa_2^2} + 3.$$

Thus, two natural estimates are:

$$g_2 = \frac{m_4}{m_2^2}, \quad G_2 = \frac{k_4}{k_2^2} + 3.$$

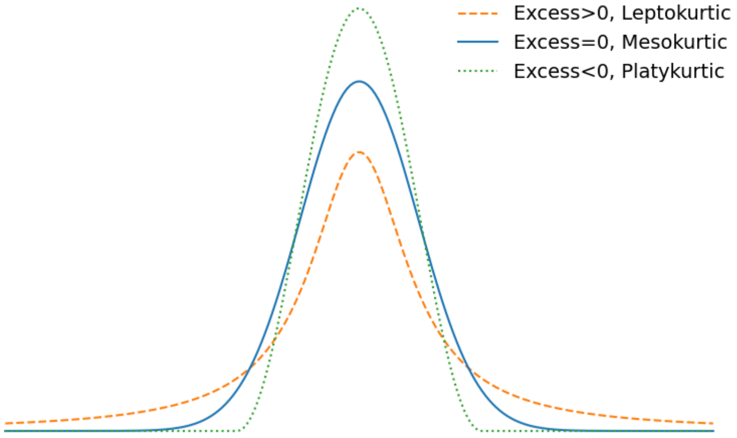
Excess of kurtosis

It is common to compare the tails with those of a normal distribution, whose kurtosis is 3. For this reason, it is common to work with the excess of kurtosis, rather than with the kurtosis itself, the excess of kurtosis is defined as

$$\gamma_{\text{Excess}} = \gamma_2 - 3.$$

There are 3 distinct regimes according to the value of the excess of kurtosis.

	Leptokurtic	Mesokurtic	Platykurtic
γ_{Excess}	> 0	0	< 0



2 Introduction to Probability

In nature, there are some phenomena that are say to be random, which means that they cannot be predicted without uncertainty. The randomness of the phenomena is not necessarily a property of the events by themselves, but due to the information of the observer.

For example, an eclipse is no longer a random event since we can predict them with complete certainty using gravitational mechanics. However, for someone that ignores this theory, an eclipse is essentially a random phenomenon.

The objective of the probability theory is to develop and study mathematical models for random phenomena which, by definition, cannot be predicted with complete certainty.

2.1 Brief history of probability theory

The theory of probability has a long history, which many authors date back at least to the XVII century when, at the request of their friend, the Chevalier de Meré, B. Pascal and P. de Fermat developed the mathematical formulation for gambling games, however it was during the XX century that the theory was notably raised.

One reason for this lack in the development of the area, compare to other fields of the mathematics, was the absence of an appropriate axiomatic system. In 1933, A. N. Kolmogorov proposed an axiomatic system through the ideas of Measure Theory, developed at the beginning of the century by H. L. Lebesgue. This axiomatic system models the random experiments using a probability space.

2.2 Probability space

A probability space is defined by a tuple $(\Omega, \mathcal{A}, \mathbb{P})$, where

- Ω is called the sample space which contains all the possible results of the experiment.
  The elements $\omega \in \Omega$ are called events.
- $\mathcal{A}$ is a system of subsets of Ω.
  $\mathcal{A}$ forms a $\sigma-$algebra of Ω.
- $\mathbb{P} : \mathcal{A} \to [0, 1]$ is the function that quantifies the uncertainty for each event $A \in \mathcal{A}$.
  $\mathbb{P}$ is called a probability measure.

Sample space

Every possible result of a random experiment is called an elementary event and the set of the elemental events is called the sample space. Usually, this set is denoted by Ω and the elemental events are denoted by ω.

1. In a fabric, one product is tested to determine if it is defective. In this case we can take $\Omega = \{0, 1\}$, where 0 means good quality and 1 means defective.

   If n products are tested, then we can take
   $$\Omega = \{(\epsilon_1, \epsilon_2, \ldots, \epsilon_n), \text{ s.t. } \epsilon_i = 1 \text{ or } 0, i = 1 \ldots, n\},$$
   where $\epsilon_i = 0$ means that the i-th product is fine and $\epsilon_i = 1$ means that it is defective.

2. In some point of a highway we count the number of cars that pass by, in some lapse of time. In this case, we can take $\Omega = \{0, 1, 2, \ldots\}$.

3. In a fabric of electronic components, we take one product at random, which is connected until it fails, observing its lifetime. We can take $\Omega = \{t, \text{ s.t. } t \in \mathbb{R}, t \geq 0\}$.

In practice, when an experiment is made, we are interested to know if some subset of Ω happened. These subsets are called events. For instance, if we roll a die, we might be interested in the subset $\{2, 4, 6\}$.

Thus, we are interested into a family of subsets of Ω, i.e., families $\mathcal{A}$ of events. These families are the second component of our probabilistic models, and must satisfy some conditions.

σ-algebra

The family of events $\mathcal{A}$ must satisfy:

1. $\Omega \in \mathcal{A}$, that is the result of the experiment must be an element of $\mathcal{A}$. Ω is called a certain event.

2. If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$, where $A^c = \Omega \setminus A = \{\omega, \text{ s.t. } \omega \in \Omega, \omega \notin A\}$. That is, if A is an event then "A does not happen" is also an event.

3. If $A_n \in \mathcal{A} \quad (n = 1, 2, \ldots)$ then $\cup_{n=1} A_n \in \mathcal{A}$. That is, the family $\mathcal{A}$ must satisfies that if $A_1, A_2, \ldots$ are events, "Some of the A_n" is also an event.

A family $\mathcal{A}$ that satisfies these conditions is called a σ-algebra of subsets of Ω.

Measure of probability

The third component of the model is a (measure of) probability. Let Ω be a sample space and $\mathcal{A}$ a σ-algebra of subsets of Ω. We want to assign to each event $A \in \mathcal{A}$ a real number $\mathbb{P}(A)$, which is called the probability of A, satisfying the conditions:

1. $\mathbb{P}(A) \geq 0$ for all $A \in \Omega$. The probability of any event is a non-negative real number.

2. $\mathbb{P}(\Omega) = 1$. A certain event has probability equals to one.

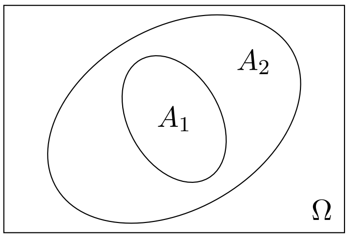
3. If $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$ are pairwise disjoint sets, i.e. $A_i \cap A_j = \varnothing$ if $i \neq j$, then
   $$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Some properties of the probability

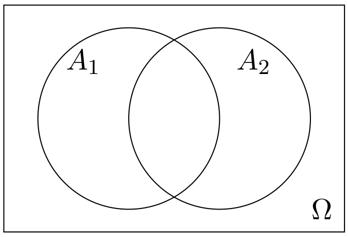
1. $\mathbb{P}(\varnothing) = 0$.

2. If $A_1 \cap A_2 = \varnothing$, then $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$.

3. If $A_1 \subset A_2$, then $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$.



4. $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$.



Exercises

1. In a certain residential suburb, 60% of all households get Internet service from the local cable company, 80% get television service from that company, and 50% get both services from that company. If a household is randomly selected, what is the probability that it gets at least one of these two services from the company, and what is the probability that it gets exactly one of these services from the company?

2. Suppose that 55% of all adults regularly consume coffee, 45% regularly consume carbonated soda, and 70% regularly consume at least one of these two products.

   (a) What is the probability that a randomly selected adult regularly consumes both coffee and soda?

   (b) What is the probability that a randomly selected adult doesn't regularly consume at least one of these two products?

2.3 Classical probability

2.3.1 Probability in finite spaces

Let be $\Omega = \{\omega_1, \dots, \omega_m\}$ a finite set and $\mathcal{A} = \mathcal{P}(\Omega)$ the family of all the subsets of Ω. Choose m real numbers p_i, $i = 1, 2, \dots, m$, such that

$$\begin{cases} p_i \geq 0, & \text{for all } i \\ \sum_{i=1}^{m} p_i = 1. \end{cases}$$

Set $\mathbb{P}(\omega_i) = p_i$ $(i = 1, 2, \dots, m)$, the probability of any event $A \in \mathcal{A}$ is defined as

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i$$

2.3.2 Classical probability

A particular case of interest is when $p_i = 1/m$ for all i. Thus, if A has n elements

$$\mathbb{P}(A) = \frac{n}{m}.$$

That is, if all the elemental events have the same probability, the probability of an event A is the ratio between the number of elements in A and the total number of elements of Ω.

The previous definition is known as *classical probability* and was proposed, among others, by P. S. Laplace. In this case, the problem of calculating the probability of an event is reduced to count how many results belong to the event of interest divided by how many results has the experiment.

Exercises

1. We choose three numbers at random between 1 and 10, one at a time and without replacement. What is the probability of getting 1, 2 and 3, in that order?
2. If the numbers of the previous example are chosen with replacement, what is the probability of getting 1, 2 and 3 in that order?
3. If we drop two dices, what is the probability that they sum 7?
4. If we toss a fair coin two consecutive times, what is the probability that at least one head appears?
5. If we toss a coin twice and one of the times we got tail, what is the probability that the other toss was head?

2.4 Non-equiprobable space

Consider now the case where Ω is an infinite numerable set:

$$\Omega = \{\omega_1, \omega_2, \ldots\}, \quad \mathcal{A} = \mathcal{P}(\Omega)$$

and

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i,$$

where

$$\begin{cases} p_i \geq 0, & \text{for all } i \\ \sum_{i=1}^{\infty} p_i = 1. \end{cases}$$

In this case, it is not possible that all the p_i being equal because they would not satisfy the previous conditions.

Example

We toss a coin until we get head for the first time.

- The possible results of this experiment are the natural numbers: $\Omega = \mathbb{N}$.
- The probability of getting head in the first toss is 1/2.
- The probability of getting tail and then head is $(1/2) \times (1/2) = 1/4$.
- The probability of getting tail two times and then head is 1/8 and so on.
- We see that the probability of getting head for the first time in the n-th toss is $p_n = 1/2^n$.

To see that this assignations defines a probability we have to check that

$$\sum_{n=1}^{\infty} p_n = 1$$

Remember that, if $|r| < 1$, then

$$1 + r + r^2 + r^3 + \cdots = \frac{1}{1-r}$$

multiplying both sides by r we have

$$r + r^2 + r^3 + r^4 + \cdots = \frac{r}{1-r}$$

Taking $r = 1/2$ shows that $\sum_{n=1}^{\infty} p_n = 1$.

Exercises

1. What is the probability of the event A :“the first tail is obtained in an even toss”?
2. Suppose that in a stock of $N > 0$ products n, $0 \leq n \leq N$, are defective. Consider that we sample with replacement three times, and observe if the selected product is defective or not.

   Determine the sample space, the probability of each elemental event, and verify that such probabilities define a measure of probability.
3. The probability that a loaded die shows the number k is proportional to k. Find the probability of the following events:
   (a) The result is an even number.
   (b) The result is less than 6.

2.5 Conditional probability

Consider a population of 20 people, of which:

- 14 are men.
- 6 are women.

Suppose that we select two people at random without replacement. That is, once we selected a person they cannot be selected again.

Consider the events:

- $\{S_1 = M\}$: “The first person is a man.”
- $\{S_1 = W\}$: “The first person is a woman.”
- $\{S_2 = M\}$: “The second person is a man.”
- $\{S_2 = W\}$: “The second person is a woman.”

2.5.1 Joint and marginal probabilities

The next table shows the number of simple events that correspond to the partition of Ω.

	$S_2 = M$	$S_2 = W$	
$S_1 = M$	14×13	14×6	14×19
$S_1 = W$	6×14	6×5	6×19
	14×19	6×19	20×19

Joint probability of S_1 and S_2, $\mathbb{P}(S_1, S_2)$

Using this table, it is easy to calculate probabilities like

$$\mathbb{P}(S_1 = M, S_2 = W) = \frac{14 \times 6}{20 \times 19}$$

	$S_2 = M$	$S_2 = W$	
$S_1 = M$	14×13	14×6	14×19
$S_1 = W$	6×14	6×5	6×19
	14×19	6×19	20×19

$$\mathbb{P}(S_1 = W, S_2 = W) = \frac{6 \times 5}{20 \times 19}$$

	$S_2 = M$	$S_2 = W$	
$S_1 = M$	14×13	14×6	14×19
$S_1 = W$	6×14	6×5	6×19
	14×19	6×19	20×19

Marginal probability of S_1, $\mathbb{P}(S_1)$

$$\mathbb{P}(S_1 = M) = \mathbb{P}(S_1 = M, S_2 = M) + \mathbb{P}(S_1 = M, S_2 = W) = \frac{14 \times 19}{20 \times 19}$$

	$S_2 = M$	$S_2 = W$	
$S_1 = M$	14×13	14×6	14×19
$S_1 = W$	6×14	6×5	6×19
	14×19	6×19	20×19

Marginal probability of S_2, $\mathbb{P}(S_2)$

$$\mathbb{P}(S_2 = W) = \mathbb{P}(S_1 = M, S_2 = W) + \mathbb{P}(S_1 = W, S_2 = W) = \frac{6 \times 19}{20 \times 19}$$

	$S_2 = M$	$S_2 = W$	
$S_1 = M$	14×13	14×6	14×19
$S_1 = W$	6×14	6×5	6×19
	14×19	6×19	20×19

2.5.2 Conditional probability

If we know that the first person is a man. What is the probability that the second person is also a man? In this case, we are asking for the probability of the event $S_2 = M$ *given that* we know that $S_1 = M$.

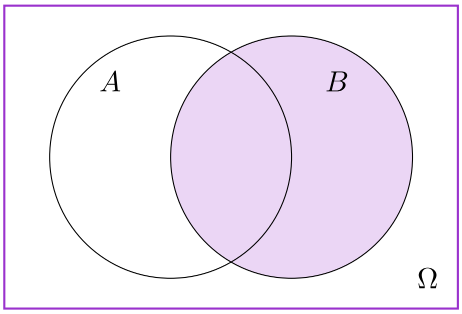
	$S_2 = M$	$S_2 = W$	
$S_1 = M$	14×13	14×6	14×19
$S_1 = W$	6×14	6×5	6×19
	14×19	6×19	20×19

The probability that we calculated is called *conditional probability* of $S_2 = M$ given $S_1 = M$.

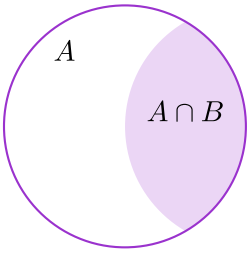
$$\mathbb{P}(S_2 = M | S_1 = M) = \frac{14 \times 13}{14 \times 19}$$

Note, that we can write it as

$$\mathbb{P}(S_2 = M | S_1 = M) = \frac{14 \times 13}{14 \times 19} = \frac{(14 \times 13)/(20 \times 19)}{(14 \times 19)/(20 \times 19)} = \frac{\mathbb{P}(S_1 = M, S_2 = M)}{\mathbb{P}(S_1 = M)}$$



$\mathbb{P}(B)$



$\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$

Exercises

1. We toss a die twice.
   (a) If the sum of the results is 8, what is the probability that the first number was k, $1 \leq k \leq 6$?
   (b) If the first number was 3, what is the probability that the second was k, $1 \leq k \leq 6$?
   (c) If the first number was 3, what is the probability that the sum of both is 7?
2. Two players toss two dice until the sum is 7 or 8. If the sum is 7 player A wins, if it is 8 player B wins. What is the probability that A wins?

2.6 Multiplication law or chain rule

For any finite collection of events $A_1, \dots, A_n$, we have

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1|A_2 \cap A_3 \cap \cdots \cap A_n)\mathbb{P}(A_2|A_3 \cap \cdots \cap A_n) \cdots \mathbb{P}(A_{n-1}|A_n)\mathbb{P}(A_n),$$

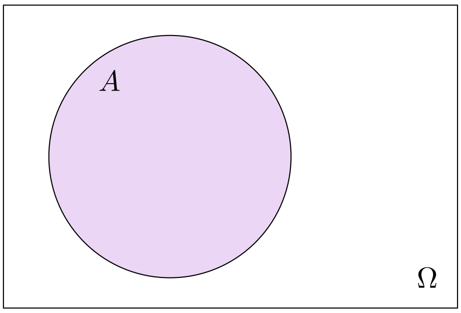
always that $\mathbb{P}(A_2 \cap A_3 \cap \cdots \cap A_n) > 0$.

2.7 Law of total probability

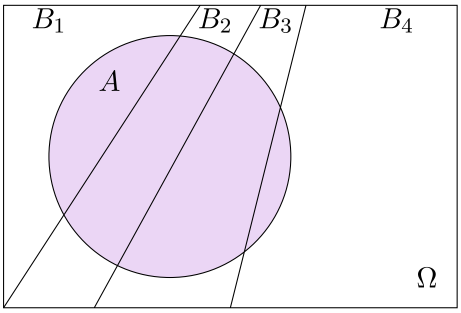
Le be $B_1, B_2, \dots$ a finite or numerable family of pairwise disjoint sets whose union is Ω. Then, for any event A,

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

where the sum is over all the indexes i such that $\mathbb{P}(B_i) > 0$.



$\mathbb{P}(A)$



$\mathbb{P}(A) = \sum_{i=1}^{4} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$

Exercises

1. A bag contains n cards, from which $m < n$ have female names, the rest have male names. If we select two cards, successively and without replacement, what is the probability that the second card has a female name?

2. What is the probability of getting 6 distinct numbers after throwing 6 dice?

3. We select two cards from a pack of 52 cards. Find the probability that the selected cards are an ace and a 10.

4. Craps: The game of *craps* has the next rules. The player throws two dice, if the result is 7 or 11, he/she wins. If it is 2, 3 or 12, he/she loses. If the sum is any other number, that number becomes his/her target and from that moment the player throws the dice until they sum his/her target, in such case the player wins, or when the sum is 7, in such case the player loses. What is the probability of winning this game?

2.8 Bayes' theorem

Let A and B be events, such that $\mathbb{P}(A) > 0$, then

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

Furthermore, let $B_1, B_2, \ldots$ be a finite or numerable partition of Ω and let A be any other event such that $\mathbb{P}(A) > 0$. Then

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\mathbb{P}(B_j)}$$

Exercises

1. From 100 patients in a hospital with certain disease, 10 are chosen for a treatment that augments the probability of heal from 0.5 to 0.75. Time after, a medic finds a healed patient, what is the probability that the patient received the treatment?

2. Three boxes contain two coins each one. In the first, B_1, both are coins of gold; in the second, B_2, both are coins of silver; and in the third, B_3, one coin is of gold and one coin is of silver.

   We pick one box at random and then one coin also at random. If the coin is of gold, what is the probability that it comes from the box with two gold coins?

3. Three mutually exclusive diseases A, B and C have the same symptoms H. Accordingly to a clinical study the probabilities of getting the diseases are 0.01; 0.005 and 0.02, respectively. Furthermore, the probability that a patient shows the symptoms H for each disease are 0.90; 0.95 and 0.75, respectively. If a sick person has the symptoms H, what is the probability that has the disease A?

4. A student answers a multiple-select question with four possible options. Assume that the probability that the student knows the answer to the question is 0.8 and the probability that he/she guesses is 0.2. If the student guess, the probability of selecting the correct answer is 0.25. If the student answered correctly, what is the probability that the student really knew the answer?

2.9 Independence

We say that two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

If A and B are independent then,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B).$$

That is, the occurrence of A does not give information on the occurrence of B.

Exercises

1. A stock of 10 objects has 4 defective products and 6 of good quality. Two objects are extracted successively without replacement. Consider the events:

   D_1 : "the first object is defective", and

   D_2 : "the second object is defective".

   Are these events independent? What happens if the objects are extracted with replacement?

2. What is the probability of getting three 6 when throwing 8 dice?

3 Probability Distributions

3.1 Random variables

Frequently, when we perform a random experiment we are interested mainly in some function of the outcome as opposed to the actual outcome itself. For instance, in throwing two dice, we are often interested in the sum of the two dice and are not really concerned about the separate values of each one. That is, we may be interested in knowing that the sum is 7 and may not be concerned over whether the actual outcome was $(1,6)$, $(2,5)$, $(3,4)$, $(4,3)$, $(5,2)$, or $(6,1)$.

These quantities of interest, or, more formally, these real-valued functions defined on the sample space, are known as random variables.

- **Random experiment:** Rolling two dice.
- **Random variable:** X = Sum of the two dice.

Event, ω	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	$\cdots$
Random variable, $X(\omega)$	2	3	4	5	6	7	$\cdots$

Examples

1. Suppose that our experiment consists of tossing 3 fair coins. If X denotes the number of heads that appear, then X is a random variable taking one value in $\{0,1,2,3\}$, with respective probabilities:

$$
\begin{aligned}
\mathbb{P}(X=0) &= \mathbb{P}\Big((T,T,T)\Big) = \frac{1}{8}\\
\mathbb{P}(X=1) &= \mathbb{P}\Big((T,T,H)\cup(T,H,T)\cup(H,T,T)\Big) = \frac{3}{8}\\
\mathbb{P}(X=2) &= \mathbb{P}\Big((T,H,H)\cup(H,T,H)\cup(H,H,T)\Big) = \frac{3}{8}\\
\mathbb{P}(X=3) &= \mathbb{P}\Big((H,H,H)\Big) = \frac{1}{8}
\end{aligned}
$$

2. Independent trials consisting of flipping a coin, having probability p of coming up heads, are continually performed until either a head occurs or a total of n flips are made. If we let X denote the number of times the coin is flipped, then X is a random variable taking one of the values $1,2,3,\cdots,n$ with respective probabilities:

$$\begin{aligned}
\mathbb{P}(X=1) &= \mathbb{P}((H)) = p \\
\mathbb{P}(X=2) &= \mathbb{P}((T,H)) = (1-p)p \\
\mathbb{P}(X=3) &= \mathbb{P}((T,T,H)) = (1-p)^2p \\
&\vdots \\
P(X=n-1) &= P((\underbrace{T,T,\cdots,T}_{n-2},H)) = (1-p)^{n-2}p \\
P(X=n) &= P((\underbrace{T,T,...,T}_{n-1},T)\cup(\underbrace{T,T,...,T}_{n-1},H)) = (1-p)^{n-1}
\end{aligned}$$

3. Three balls are randomly chosen from an urn containing 3 white, 3 red, and 5 black balls. Suppose that we win \$1 for each white ball selected and lose \$1 for each red ball selected. If we let X denote our total winnings from the experiment, then X is a random variable taking on the possible values 0, ±1, ±2, ±3 with respective probabilities:

$$\mathbb{P}(X=0) = \frac{\binom{5}{3}+\binom{3}{1}\binom{3}{1}\binom{5}{1}}{\binom{11}{3}} = \frac{55}{165}$$

$$\mathbb{P}(X=1) = \mathbb{P}(X=-1) = \frac{\binom{3}{1}\binom{5}{2}+\binom{3}{2}\binom{3}{1}}{\binom{11}{3}} = \frac{39}{165}$$

$$\mathbb{P}(X=2) = \mathbb{P}(X=-2) = \frac{\binom{3}{2}\binom{5}{1}}{\binom{11}{3}} = \frac{15}{165}$$

$$\mathbb{P}(X=3) = \mathbb{P}(X=-3) = \frac{\binom{3}{3}}{\binom{11}{3}} = \frac{1}{165}$$

3.2 Distribution of a random variable

If A is an arbitrary subset of $\mathbb{R}$, and we want to find the probability that the variable X takes values in A, we need to consider the set $\{\omega,\ \text{s.t. } X(\omega)\in A\}$. This relation defines a (measure of) probability induced by the variable X as follows:

$$\mathbb{P}_X(A) = \mathbb{P}(X\in A) = \mathbb{P}\Big(\{\omega\in\Omega,\ \text{s.t. } X(\omega)\in A\}\Big).$$

This (measure of) probability is known as the distribution of X and has all the probabilistic information about X.

3.2.1 Cumulative distribution function

Let $(\Omega,\mathcal{A},\mathbb{P})$ be a probability space, and $X:\Omega\to\mathbb{R}$ a random variable.
We define the (cumulative) distribution function of the random variable X, denoted by F, as:

$$F(x) = \mathbb{P}\Big(\{\omega,\ \text{s.t. } X(\omega)\le x\}\Big) = \mathbb{P}(X\le x).$$

1. F is non-decreasing.

2. $\lim_{x\to-\infty} F(x) = 0,\ \lim_{x\to+\infty} F(x) = 1.$

3. F is right continuous.

Conversely, if a function $F:\mathbb{R}\to[0,1]$ satisfies the three previous properties, it can be proved that F is a distribution function of a random variable.

3.3 Discrete random variables

A random variable that can take at most a countable number of possible values is said to be discrete. For a discrete random variable X, we define the probability mass function $p(x)$ of X as

$$p(x) = \mathbb{P}(X = x).$$

The probability (mass) function $p(x)$ is positive for at most a countable number of values. That is, if X must assume one of the values $x_1, x_2, \ldots$, then

$$\begin{cases} p(x_i) \geq 0 & \text{for } i = 1, 2, \ldots \\ p(x) = 0 & \text{for all other values of } x. \end{cases}$$

Furthermore,

$$F(x) = \mathbb{P}(X \leq x) = \sum_{i : x_i \leq x} p(x_i).$$

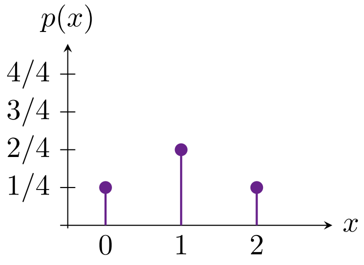
It is often instructive to present the probability (mass) function in a graphical format, by plotting $p(x_i)$ on the y-axis against x_i on the x-axis.

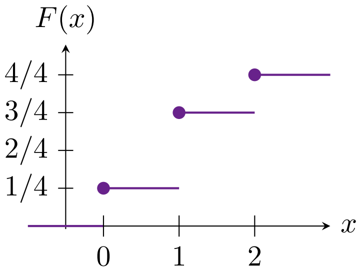
Examples

1. If the probability mass function of X is

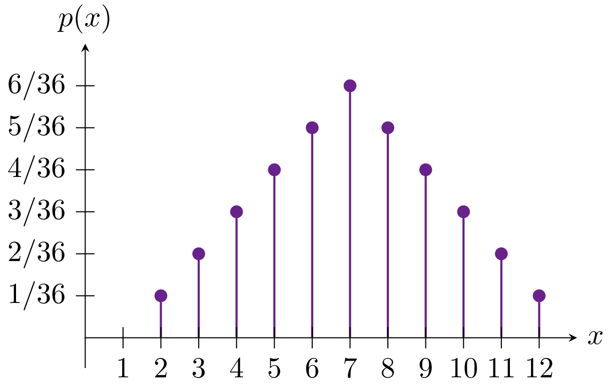
$$p(0) = \frac{1}{4},\, p(1) = \frac{1}{2},\, p(2) = \frac{1}{4},$$

we can represent this function graphically as

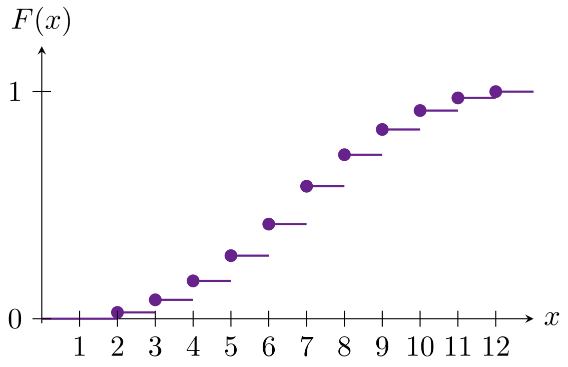




2. A graph of the probability function of the random variable, representing the sum when two dice are rolled looks like:



The distribution function of this random variables is:



3. A box has 6 cards numbered from 1 to 6. Two cards are taken with replacement and the maximum of the numbers is registered.

   (a) What is the probability function of this variable?

   (b) How is, if the sampling is done without replacement?

(a). First, consider the case with replacement. The sampling space for this experiment is the set or pairs (ω_1, ω_2) where $\omega_i \in \{1,2,3,4,5,6\}$ for $i = 1,2$.

The random variable $X : \Omega \to \mathbb{R}$ of interest is defined by

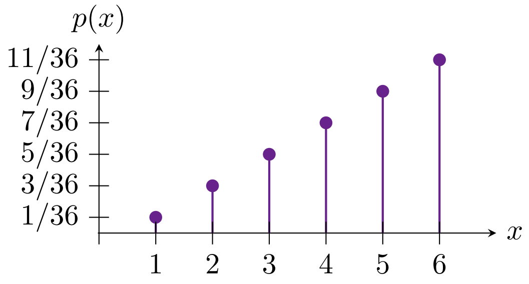
$$X(\omega_1, \omega_2) = \max\{\omega_1, \omega_2\}$$

which takes values in the set $\{1,2,3,4,5,6\}$.

If all the cards have the same probability to be selected then all the elementary events of the sample space have probability 1/36.It is easy to calculate the probability function of the random variable X:

x_i	1	2	3	4	5	6
$p(x_i)$	$^1/_{36}$	$^3/_{36}$	$^5/_{36}$	$^7/_{36}$	$^9/_{36}$	$^{11}/_{36}$

whose graphical representation is



(b). If the selection is done without replacement, then the sample space is the set

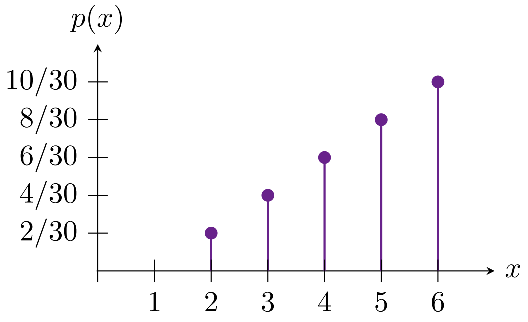
$$\{(\omega_1, \omega_2), \text{ s.t. } \omega_i \in \{1,2,3,4,5,6\}, \omega_1 \neq \omega_2\}.$$

The variable $X(\omega_1, \omega_2) = \max\{\omega_1, \omega_2\}$ now takes the values in the set $\{2,3,4,5,6\}$.

If the cards have the same probability to been selected, the elementary events have the same probability 1/30. The next table shows the probability function of the random variable X.

x_i	2	3	4	5	6
$p(x_i)$	$^2/_{30}$	$^4/_{30}$	$^6/_{30}$	$^8/_{30}$	$^{10}/_{30}$

whose graphical representation is



Exercises

1. A coin is toss repetitively, consider the first time that we observe "head." Find the probability function of this variable.

3.3.1 Bernoulli distribution

A Bernoulli distribution corresponds to a variable X that takes two values, 1 and 0, with probabilities p and $q = 1 - p$, respectively. That is

$$\mathbb{P}(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

In this case, we say that X has or follows a Bernoulli distribution with parameter p, and denote it as $X \sim \text{Bernoulli}(p)$.

Define the indicator function of a subset A of B, $\mathbb{1}_A(x) : B \to \{0, 1\}$ as

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

An equivalent notation for this function is $\mathbb{1}_{x \in A}$.

With this notation, the probability function of a Bernoulli random variable can be written as

$$p(x) = p^x (1 - p)^{1-x} \mathbb{1}_{x \in \{0,1\}}.$$

3.3.2 Binomial distribution

Consider the case of sampling with replacement, in which the variable of interest is the number of defective products X in a sample of n products. That is,

$$X = \sum_{i=1}^{n} e_i,$$

where $e_i = 1$ or 0 if the product is defective or not, respectively. The probability function of such random variable X is

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x} \mathbb{1}_{x \in \{0,1\ldots,n\}},$$

where p is the proportion of defective products.

In this case, we say that X has or follows a binomial distribution with parameters n, p, and denote it as $X \sim \text{Binomial}(n, p)$.

Note that, by the binomial theorem, the probabilities sum to 1. That is,

$$\sum_{x=0}^{n} \binom{n}{x} p^x (1 - p)^{n-x} = [p + (1 - p)]^n = 1.$$

Furthermore, X can be seen as the sum of n independent Bernoulli r.v. with parameter p.

Exercise

1. Five cards are selected with replacement from a deck of 52 cards. If X is the number of diamonds in the sample.

   (a) What is the probability that there are exactly two diamonds in the five cards?

   (b) What is the probability that there are at most 2 diamonds?

3.3.3 Uniform distribution

A random variable X with values in the set $\{x_1, x_2, \ldots, x_n\}$ has a uniform distribution if all the points x_i, $1 \leq i \leq n$ have the same probability.

Since there are n possible values this means that

$$p(x) = \frac{1}{n} \mathbb{1}_{x \in \{x_1, x_2, \ldots, x_n\}}.$$

In this case, we say that X has or follows a uniform distribution in $\{x_1, \ldots, x_n\}$, denoted as $X \sim \text{Uniform}\{x_1, \ldots, x_n\}$.

3.3.4 Poisson distribution

We say that a random variable X has or follows a Poisson distribution with parameter λ ($\lambda > 0$), denoted as $X \sim \text{Poisson}(\lambda)$, if its probability function is

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \mathbb{1}_{x \in \{0,1,\ldots\}}.$$

This relation effectively defines a probability function, using series of Taylor of the exponential function,

$$\sum_{x=0}^{\infty} p(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

Law of rare events

The Poisson distribution is also useful as an approximation to the binomial distribution for n large and p small.

If $X \sim \text{Binomial}(n, p)$, with $n \to \infty$ and $p \to 0$, such that $np = \lambda$, then

$$X \sim \text{Poisson}(\lambda).$$

Proof

Consider the binomial distribution when n increases and p tends to zero, but the product $np = \lambda$ remains fixed. The binomial distribution is

$$p_{n,k} = \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \frac{n(n-1)\cdots(n-k+1)}{k!} p^k (1-p)^{n-k}$$

$$= \frac{n(n-1)\cdots(n-k+1)}{k! n^k} (np)^k (1-p)^{n-k}$$

$$\begin{aligned}
p_{n,k} &= \frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{\lambda^k}{k!}(1-p)^{n-k} \\
&= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{\lambda^k}{k!}(1-p)^{n-k} \\
&= \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}{(1-p)^k}\frac{\lambda^k}{k!}(1-p)^{n}
\end{aligned}$$

On the other hand, note that

$$(1-p)^n = [(1-p)^{-1/p}]^{-np} = [(1-p)^{-1/p}]^{-\lambda},$$

from the definition of e we know that

$$\lim_{z\to 0}(1+z)^{1/z} = e.$$

So

$$\lim_{p\to 0}(1-p)^n = \lim_{p\to 0}[(1-p)^{-1/p}]^{-\lambda} = e^{-\lambda}.$$

Moreover,

$$\lim_{n\to\infty}\frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}{(1-p)^k} = 1$$

because we assumed that $p \to 0$ when $n \to \infty$ and $np = \lambda$ is constant.

Then we have

$$\lim_{n\to\infty} p_{n,k} = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Exercises

1. The number of calls per minute received by a call center follows a Poisson distribution with parameter $\lambda = 4$. If the center can manages at most 6 calls per minute, what is the probability that the center is insufficient for the calls received in one minute?

2. 400 fuses are sampled from a process that, on average, produces 1% of defectives. What is the probability that at most there are 5 defective fuses in the sample?

3.3.5 Hypergeometric distribution

Suppose a population of n elements:

- There are r elements of type I.
- There are $n-r$ elements of type II.

We extract a sample of k elements from this population without replacement, where every element has the same probability of being selected. Let X be the random variable that represents the number of elements of type I in the sample.

We want to find the probability function of X. That is, $\mathbb{P}(X = x)$, where $x \in \{\max\{0, k - (n-r)\}, \dots, \min\{k, r\}\}$. To find the probability, note that in the sample:

- There are x elements of type I.
- There are $k - x$ elements of type II.

Those of type I can be selected in $\binom{r}{x}$ different ways and those of type II in $\binom{n-r}{k-x}$ different ways. Since every selection of the x elements of type I can be combined with any selection of the $k-x$ elements of type II, we have that

$$\mathbb{P}(X=x)=\frac{\binom{r}{x}\binom{n-r}{k-x}}{\binom{n}{k}}\mathbb{1}_{x\in\{\max\{0,k-(n-r)\},\dots,\min\{k,r\}\}}.$$

In this case we say that X has or follows a hypergeometric distribution with parameters n, r, k, denoted as $X\sim\text{Hypergeometric}(n,r,k)$.

Binomial approximation to the hypergeometric distribution

If k individuals are randomly chosen without replacement from a population of n individuals, of which the fraction $p=r/n$ is of type I, then the number of individuals of type I selected is hypergeometric.

Now, when r and n are large in relation to k, it shouldn't make much difference whether the selection is being done with or without replacement.

No matter how many individuals have been selected previously, when r and n are large, each additional selection will be of type I with a probability approximately equal to p.

In other words, it seems intuitive that, **when r and n are large in relation to k, X should approximately be a binomial random variable with parameters k and p.**

Proof

To verify this intuition, note that if X is hypergeometric, then, for $x\leq k$,

$$\begin{aligned}
\mathbb{P}(X=x)&=\frac{\binom{r}{x}\binom{n-r}{k-x}}{\binom{n}{k}}\\
&=\frac{r!}{(r-x)!x!}\frac{(n-r)!}{(n-r-k+x)!(k-x)!}\frac{(n-k)!k!}{n!}\\
&=\binom{k}{x}\frac{r}{n}\frac{(r-1)}{(n-1)}\cdots\frac{(r-x+1)}{(n-x+1)}\frac{(n-r)}{(n-x)}\frac{(n-r-1)}{(n-x-1)}\\
&\qquad\cdots\frac{(n-r-(k-x-1))}{(n-x-(k-x-1))}\frac{\cancel{(n-r-(k-x))!}}{\cancel{(n-r-k+x)!}}\\
&\qquad\frac{\cancel{(n-k)!}}{\cancel{(n-x-(k-x))!}}\\
&\approx\binom{k}{x}p^x(1-p)^{k-x}
\end{aligned}$$

when $p=r/n$ and r and n are large in relation to k and x.

Example

Consider a population of 100 people, 10 of which have myopia. The probability that there are at most two people with myopia in a group of 10 chosen at random without replacement is:

$$\mathbb{P}(X\leq 2)=\sum_{x=0}^{2}\frac{\binom{10}{2}\binom{90}{8}}{\binom{100}{10}}\approx 0.94$$

An important property

Since all the individuals in the population have the same probability of been selected, then an important property of the hypergeometric distribution is that it assumes that **the proportion of individuals of type I in the selected sample** x/k **must be approximately equal to the proportion of individuals of type I in the population** r/n, that is

$$\frac{x}{k} \approx \frac{r}{n}.$$

In fact, if $X \sim$ Hypergeometric(n, r, k), its expected is

$$\mathbb{E}(X) = k\left(\frac{r}{n}\right).$$

We can derive useful estimations from this relation.

Case 1
If we know:

- The population size, n.
- The sample size, k.
- The number of elements of type I in the population, r.

Then, we can estimate the outcome of the random variable by

$$x \approx k\left(\frac{r}{n}\right).$$

Case 2
If we know:

- The population size, n.
- The sample size, k.
- The number of elements of type I in the sample, x.

Then, we can estimate the amount of elements of type I in the population:

$$r \approx n\left(\frac{x}{k}\right).$$

This is useful in quality control applications. To estimate the number of defective products produced in an industrial process, we can sample k products from a stock of n products. Assuming that defective products and good quality products are mixed, then the number of defective products X satisfies that $X \sim$ Hypergeometric(n, r, k).

Hence, the number of defective products in the stock can be estimated as

$$r \approx n\left(\frac{x}{k}\right).$$

Suppose that we take a sample of $k = 50$ screws from a stock of $n = 500$ pieces. If there is $x = 1$ defective screw in the sample, then we would estimate that there are approximately 10 defective pieces in the stock.

Case 3
If we know:

- The sample size, k.
- The number of elements of type I in the population, r.

- The number of elements of type I in the sample, x.

Then, we can estimate the population size:

$$n \approx r\left(\frac{k}{x}\right).$$

This is useful in ecology applications. To obtain some information about the size of a population, ecologists often perform the following experiment:

They first catch a number, say r, of these animals, mark them in some manner, and release them. After some time, allowing the marked animals to disperse throughout the region, a new catch of size k, is made.

Let X denote the number of marked animals in this second capture. If we assume that the population of animals in the region remained fixed between the two catches, then $X \sim$ Hypergeometric(n, r, k), and

$$n \approx r\left(\frac{k}{x}\right).$$

Suppose that the initial catch consisted of $r = 50$ animals, which are marked and then released. If a subsequent catch consists of $k = 40$ animals of which $x = 4$ are marked, then we would estimate that there are some 500 animals in the region.

3.3.6 Geometric distribution

Suppose that independent trials, each having a probability p, $0 < p < 1$, of being a success, are performed until a success occurs. If we let X be the number of trials required, then

$$\mathbb{P}(X = x) = (1-p)^{x-1} p \mathbb{1}_{\,x\in\{1,2,\ldots\}}$$

Any random variable X with this probability function is said to be a geometric random variable with parameter p, denoted as $X \sim$ Geometric(p).

Example

An urn contains w white and b black balls. Balls are randomly selected, one at a time, until a black one is obtained. If we assume that each ball selected is replaced before the next one is drawn, what is the probability that:

1. exactly n draws are needed?
2. at least k draws are needed?

If we let X denote the number of draws needed to select the first black ball, then $X \sim$ Geometric(p) with $p = b/(w + b)$. Then,

1. $$\mathbb{P}(X = n) = (1-p)^{n-1} p$$

2. $$\mathbb{P}(X \geq k) = p \sum_{x=k}^{\infty} (1-p)^{x-1}$$

$$= p\left(\frac{(1-p)^{k-1}}{p}\right)$$

$$= (1-p)^{k-1}$$

3.3.7 Negative binomial distribution

This distribution appears in the context of a succession of Bernoulli experiments with probability of success p. When we ask a similar question to the one of the geometric distribution, but instead of asking the number of experiments to obtain the first success, we ask the number of experiments to obtain the first k successes.

Let X be this variable, it takes the value of x if and only if the k-th success happens in the x-th experiment. That is, in the first $x-1$ experiments there are $k-1$ successes and the x-th experiment is a success. The probability of the latter is p, while the probability of $k-1$ successes in $x-1$ experiments is a binomial distribution:

$$\binom{x-1}{k-1} p^{k-1} q^{x-k}.$$

Since the experiments are independent, we have that $\mathbb{P}(X = x)$ is the product of these expressions, that is

$$\mathbb{P}(X = x) = \binom{x-1}{k-1} p^{k} q^{x-k}.$$

3.4 Continuous random variables

We have considered discrete random variables–that is, random variables whose set of possible values is either finite or countably infinite. However, there also exist random variables whose set of possible values is uncountable. Two examples are the time that a train arrives at a specified stop and the lifetime of a transistor.

In general, we say that a random variable X is continuous if its distribution function is continuous. Note that, since $\mathbb{P}(X = x) = F(x) - F(x^-)$. If F is continuous, then $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.

3.4.1 Density function

Let X be a random variable with distribution function F. We say that F has density (or is absolutely continuous), if exists a non-negative function f such that

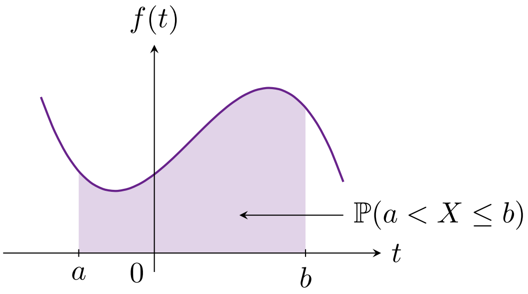
$$F(x) = \int_{-\infty}^{x} f(t)dt \quad \text{for all } x \in \mathbb{R}.$$

The function f is called the density of the distribution or the density of the random variable. Because $\lim\limits_{x\to\infty} F(x) = 1$, we have that $\displaystyle\int_{-\infty}^{\infty} f(t)dt = 1$.

We also have that

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a) = \int_{a}^{b} f(t)dt.$$

Thus, the probability that the random variable X belongs to the interval $(a, b]$ is the area between the graph of the function f, the x-axis and the verticals a and b.



In general, if B is any set of real numbers,

$$\mathbb{P}(X \in B) = \int_B f(t)dt.$$

Remember that we defined the density function f, of the random variable X with distribution function F, as the non-negative function such that

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Differentiating both sides of the equation yields

$$\frac{d}{dx}F(x) = f(x),$$

that is, the density is the derivative of the cumulative distribution function.

Examples

1. Suppose that X is a continuous random variable whose probability density function is given by

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

   (a) What is the value of C?

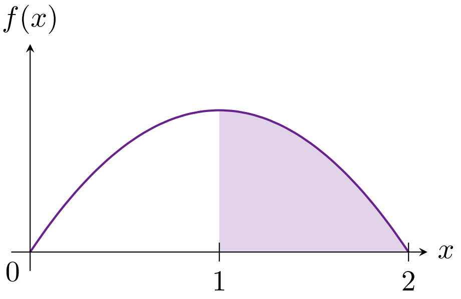
   (b) Find $\mathbb{P}(X > 1)$.

   (a). Since f is a probability density function, we must have

$$\begin{aligned} C\int_0^2 (4x - 2x^2)dx &= 1 \\ \Leftrightarrow C\left[2x^2 - \frac{2x^3}{3}\right]_{x=0}^{x=2} &= 1 \\ \Leftrightarrow C &= \frac{3}{8} \end{aligned}$$

   (b).

$$\mathbb{P}(X > 1) = \int_1^{\infty} f(x)dx = \frac{3}{8}\int_1^2 (4x - 2x^2)dx = \frac{1}{2}$$



2. The amount of time in hours that a computer works before breaking down is a continuous random variable with probability density function given by

$$f(x) = \begin{cases} \lambda e^{-x/100} & \text{for } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

   What is the probability that

(a) a computer will function between 50 and 150 hours before breaking down?

(b) it will function for fewer than 100 hours?

(a). Since

$$1 = \int_0^\infty \lambda e^{-x/100} dx = \lambda \int_0^\infty e^{-x/100} dx$$

we obtain

$$1 = -\lambda(100) e^{x/\lambda}|_0^\infty = 100\lambda \Leftrightarrow \lambda = \frac{1}{100}.$$

Hence, the probability that a computer will function between 50 and 150 hours before breaking down is given by

$$\mathbb{P}(50 < X < 150) = \int_{50}^{150} \frac{1}{100} e^{-x/100} dx = -e^{-x/100}\Big|_{50}^{150}$$

$$= e^{-1/2} - e^{-3/2} \approx 0.384$$

(b). Similarly,

$$\mathbb{P}(X < 100) = \int_0^{100} \frac{1}{100} e^{-x/100} dx = -e^{-x/100}\Big|_0^{100} = 1 - e^{-1} \approx 0.633$$

3. The lifetime in hours of a certain kind of radio tube is a random variable having a probability density function given by

$$f(x) = \begin{cases} 0 & x \leq 100, \\ \frac{100}{x^2} & x > 100. \end{cases}$$

What is the probability that exactly 2 of 5 such tubes in a radio set will have to be replaced within the first 150 hours of operation? Assume that the events E_i, $i = 1, 2, 3, 4, 5$, that the ith such tube will have to be replaced within this time, are independent.

From the statement of the problem, we have

$$\mathbb{P}(E_i) = \int_0^{150} f(x) dx = 100 \int_{100}^{150} \frac{1}{x^2} dx = \frac{1}{3}.$$

Hence, from the independence of the events E_i , it follows that the desired probability is

$$\binom{5}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3 = \frac{80}{243}.$$

3.4.2 Uniform distribution

A random variable X has uniform distribution in the interval $[a, b]$ if for any interval I contained in $[a, b]$, $\mathbb{P}(X \in I)$ is proportional to the length of I.

We denote it as $X \sim \text{Uniform}[a, b]$.

We can calculate the distribution function of X,

$$F(x) = \mathbb{P}(X \in [a, x]) = K(x - a)$$

where K is the constant of proportionality.

Because $F(b) = \mathbb{P}(X \in [a, b]) = 1$ we have

$$K(b - a) = 1 \Leftrightarrow K = \frac{1}{b - a}.$$

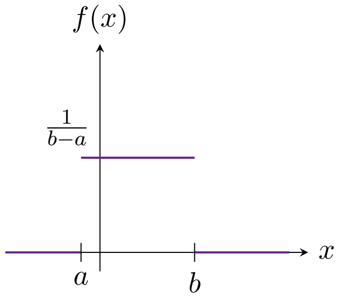
Therefore, the distribution function of X is

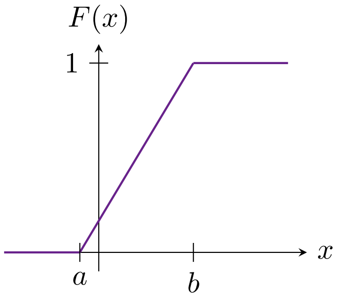
$$F(x) = \frac{x-a}{b-a}\mathbb{1}_{a\leq x\leq b} + \mathbb{1}_{x>b}$$

The density of this random variable is

$$f(x) = \frac{1}{b-a}\mathbb{1}_{a\leq x\leq b}$$

since we can verify that $F(x) = \int_{-\infty}^{x} f(t)dt$ for all $x \in \mathbb{R}$.





3.4.3 Exponential distribution

We say that a random variable X has exponential distribution if

$$\mathbb{P}(X > x) = e^{-x/\lambda}, \quad x \geq 0$$

where $\lambda > 0$.

Thus, its distribution function is

$$F(x) = \mathbb{P}(X \leq x) = 1 - \mathbb{P}(X > x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - e^{-x/\lambda} & \text{if } x \geq 0. \end{cases}$$

And the density function of this distribution is

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{\lambda}e^{-x/\lambda} & \text{if } x \geq 0. \end{cases}$$

We use the notation $X \sim \text{Exponential}(\lambda)$.

Memoryless property

An important property of the exponential distribution is that, for a and b non-negative

$$\mathbb{P}(X > a + b) = \mathbb{P}(X > a)\mathbb{P}(X > b).$$

We can verify this property from the definition of the distribution.

As a consequence, we have that

$$\mathbb{P}(X > a + b|X > a) = \mathbb{P}(X > b), \quad a \geq 0, \quad b \geq b,$$

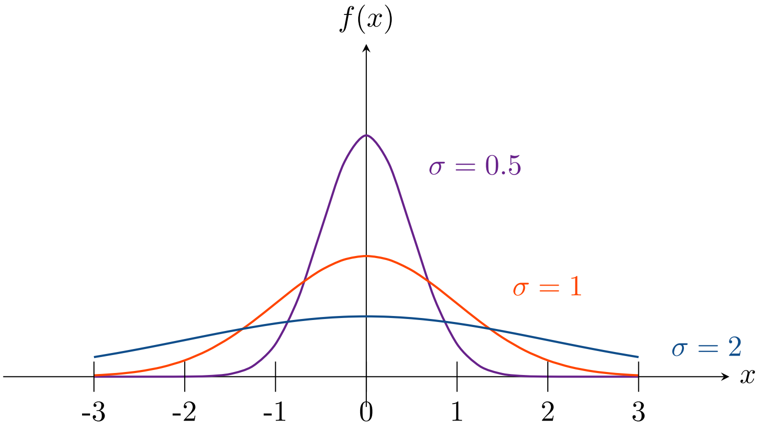
which is known as the "memoryless" property of the exponential distribution or the property of a lifetime product that never gets old.

3.4.4 Normal distribution

A random variable X is said to have or follow a normal distribution with parameters μ and σ^2 if its density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathbb{1}_{x\in\mathbb{R}}.$$

If X has a normal distribution with parameters μ and σ^2, we denote it as $X \sim \mathcal{N}(\mu, \sigma^2)$.



Exercises

1. If X is a normal random variable with parameters $\mu = 3$ and $\sigma^2 = 9$, find
   (a) $\mathbb{P}(2 < X < 5)$.
   (b) $\mathbb{P}(X > 0)$.

Standard normal distribution

If $X \sim \mathcal{N}(0,1)$, it is said that X has a standard normal distribution, whose density is usually denoted as ϕ, so

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \mathbb{1}_{x\in\mathbb{R}}.$$

The distribution function corresponding to the density ϕ is usually denoted as Φ.

De Moivre - Laplace theorem

Let be $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, and $q = 1 - p$. Then, for n sufficiently large

$$\frac{1}{n}\sum_{i=1}^{n} X_i \dot{\sim} \mathcal{N}(p, pq)$$

This result was proved originally for the special case of $p = 1/2$ by A. de Moivre in 1733 and was then extended to general p by P. S. Laplace in 1812. It represents the first version of the central limit theorem, proved rigorously by A. M. Lyapunov in the period 1901–1902.

Examples

1. Let X be the number of times that a fair coin lands on heads, when it is flipped 40 times. Find the probability that $X = 20$. Use the normal approximation and compare it with the exact solution.

Because the binomial is a discrete integer-valued random variable whereas the normal is a continuous random variable, it is best to write $\mathbb{P}(X = i)$ as $\mathbb{P}(i - 1/2 < X < i + 1/2)$ before applying the normal approximation (this is called the continuity correction).

Therefore,

$$\begin{aligned}
\mathbb{P}(X = 20) &= \mathbb{P}(19.5 < X < 20.5) \\
&= \mathbb{P}\left(\frac{19.5 - 20}{\sqrt{10}} < \frac{X - 20}{\sqrt{10}} < \frac{20.5 - 20}{\sqrt{10}}\right) \\
&\approx \mathbb{P}(-0.158 < Z < 0.158) \\
&\approx \Phi(0.158) - \Phi(-0.158) \approx 0.1255.
\end{aligned}$$

The exact result is

$$\mathbb{P}(X = 20) = \binom{40}{20}\left(\frac{1}{2}\right)^{40} \approx 0.1254.$$

2. To determine the effectiveness of a certain diet in reducing the amount of cholesterol in the bloodstream, 100 people are put on the diet. After they have been on the diet for a sufficient length of time, their cholesterol count will be taken.

   The nutritionist running this experiment has decided to endorse the diet if at least 65 percent of the people have a lower cholesterol count after going on the diet. What is the probability that the nutritionist endorses the new diet if, in fact, it has no effect on the cholesterol level?

   Let us assume that if the diet has no effect on the cholesterol count, then, strictly by chance, each person's count will be lower than it was before the diet with probability $1/2$.

   Hence, if X is the number of people whose count is lowered, the probability that the nutritionist will endorse the diet when it actually has no effect on the cholesterol count is

$$\begin{aligned}
\sum_{i=65}^{100}\binom{100}{i}\left(\frac{1}{2}\right)^{100} &= \mathbb{P}(X \geq 65) \\
&= \mathbb{P}\left(\frac{X - 50}{\sqrt{25}} \geq \frac{65 - 50}{\sqrt{25}}\right) \\
&\approx \mathbb{P}(Z \geq 3) \\
&1 - \Phi(3) \approx 0.0014
\end{aligned}$$

3. Fifty-two percent of the residents of Bangkok are in favor of outlawing cigarette smoking in publicly owned areas. Approximate the probability that more than 50 percent of a random sample of n people from Bangkok are in favor of this prohibition when (a) $n = 11$, (b) $n = 101$, (c) $n = 1001$.

   How large would n have to be to make this probability exceed 0.95?

Let N denote the number of residents of Bangkok.

To answer the preceding question, we must first understand that a random sample of size n is a sample such that the n people were chosen in such a manner that each of the $\binom{N}{n}$ subsets of n people had the same chance of being the chosen subset.

Consequently, S_n, the number of people in the sample who are in favor of the smoking prohibition, is a hypergeometric random variable. That is,

$$S_n \sim \text{Hypergeometric}(N, r, n),$$

where $r = 0.52N$.

But because N and $0.52N$ are both large in comparison with the sample size n, it follows from the binomial approximation to the hypergeometric that the distribution of S_n is closely approximated by a binomial distribution with parameters n and $p = 0.52$. That is,

$$S_n \dot{\sim} \text{Binomial}(n, 0.52)$$

The normal approximation to the binomial distribution then shows that

$$\mathbb{P}(S_n > 0.5n) = \mathbb{P}\left(\frac{S_n - 0.52n}{\sqrt{n(0.52)(0.48)}} > \frac{0.5n - 0.52n}{\sqrt{n(0.52)(0.48)}}\right)$$

$$= \mathbb{P}(Z > -0.04\sqrt{n})$$

$$\approx \Phi(0.04\sqrt{n}).$$

Thus,

$$\mathbb{P}(S_n > 0.5n) \approx \begin{cases} \Phi(0.1327) \approx 0.5528 & \text{if } n = 11, \\ \Phi(0.4020) \approx 0.6562 & \text{if } n = 101, \\ \Phi(1.2655) \approx 0.8972 & \text{if } n = 1001. \end{cases}$$

In order for this probability to be at least 0.95, we would need

$$\Phi(0.04\sqrt{n}) > 0.95.$$

Because $\Phi(x)$ is an increasing function and $\Phi(1.645) = 0.95$, this means that

$$0.04n > 1.645 \Leftrightarrow n > 1691.266.$$

That is, the sample size would have to be at least 1692.

3.4.5 Laplace distribution

A random variable X that can take either positive or negative values, and whose absolute value is exponentially distributed with parameter λ, $\lambda > 0$ is said to have a Laplace distribution, its density and distribution function are given by

$$f(x) = \frac{1}{2}e^{-\lambda|x|}\mathbb{1}_{x\in\mathbb{R}}$$

and

$$F(x) = \frac{\lambda}{2}e^{-\lambda x}\mathbb{1}_{x<0} + \left[1 - \frac{1}{2}e^{-\lambda x}\right]\mathbb{1}_{x\geq 0},$$

respectively.

3.4.6 Gamma distribution

A random variable is said to have a gamma distribution with parameters α, β, $\alpha > 0, \beta > 0$, if its density function is given by

$$f(t) = \frac{\beta^\alpha}{\Gamma(\alpha) t^{\alpha-1}} e^{-\beta t} \mathbb{1}_{t \geq 0},$$

where $\Gamma(\alpha)$, called the gamma function, is defined as

$$\Gamma(\alpha) = \int_0^\alpha e^{-y} y^{\alpha-1} dy,$$

β is called the rate parameter and α the shape parameter.

It is often to see the density parametrized in terms of α and $\theta = 1/\beta$, θ is called the scale parameter of the distribution. This distribution includes the exponential as a special case ($\alpha = 1$).

3.4.7 χ^2 distribution

The gamma distribution with $\alpha = n/2$ and $\beta = 1/2$, n a positive integer, is called the χ^2 distribution with n degrees of freedom, denoted as χ^2_n.

If a vector of random variables follows a multivariate standard normal distribution, $\mathbf{X} \sim \mathcal{N}_p(0, I)$, then

$$X_1^2 + \cdots + X_p^2 \sim \chi^2_p.$$

3.4.8 Lognormal distribution

The lognormal distribution has been used as a model in diverse applications in engineering, medicine, and other areas.

A random variable T is said to be log-normally distributed if $X = \log T$ is normally distributed, say with mean μ, variance σ^2. Its density function is given by

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left\{ -\frac{1}{2} \left(\frac{\log t - \mu}{\sigma} \right)^2 \right\} \mathbb{1}_{t > 0}$$

3.4.9 Weibull distribution

The Weibull distribution is perhaps the most widely used lifetime distribution model.

Applications to the lifetimes or durability of manufactured items is common, and it is used as a model with diverse types of items, such as ball bearings, automobile components, and electrical insulation. It is also used in biological and medical applications, for example, in studies on the time to the occurrence of tumors in human populations or in laboratory animals.

The probability density function of a Weibull random variable is

$$f(t) = \frac{\rho}{\lambda} \left(\frac{t}{\lambda} \right)^{\rho - 1} \exp\left\{ -({}^{t}\!/_{\lambda})^\rho \right\} \mathbb{1}_{t \geq 0}.$$

where $\rho > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter.

3.5 Mixed random variables

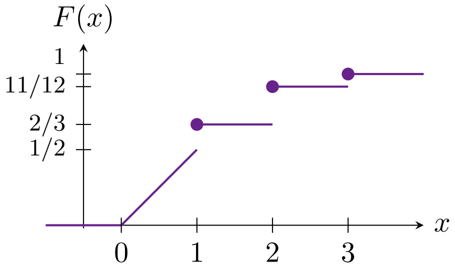
These are random variables that are neither discrete nor continuous, but are a mixture of both. In particular, a mixed random variable has a continuous part and a discrete part.

Examples

1. If the distribution function of the random variable X is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/2 & \text{if } 0 \leq x < 1, \\ 2/3 & \text{if } 1 \leq x < 2, \\ 11/12 & \text{if } 2 \leq x < 3, \\ 1 & \text{if } 3 \leq x. \end{cases}$$

A graph of $F(x)$ is



2. Let X be a continuous random variable with the following density: $f(x) = 2x\mathbb{1}_{x\in[0,1]}$. Let

$$Y = g(X) = \begin{cases} X & \text{if } 0 \leq X \leq \frac{1}{2}, \\ \frac{1}{2} & \text{if } X > \frac{1}{2}. \end{cases}$$

Find the distribution function of Y.

Note that the support of X is $[0, 1]$. For $x \in [0, 1]$, $0 \leq g(x) \leq \frac{1}{2}$. Thus, the support of Y is $[0, \frac{1}{2}]$, so

$$F_Y(y) = 0, \quad \text{for } y < 0,$$
$$F_Y(y) = 1, \quad \text{for } y > \tfrac{1}{2}.$$

Now, note that

$$\mathbb{P}\left(Y = \frac{1}{2}\right) = \mathbb{P}\left(X > \frac{1}{2}\right) = \int_{\frac{1}{2}}^{1} 2xdx = \frac{3}{4}.$$

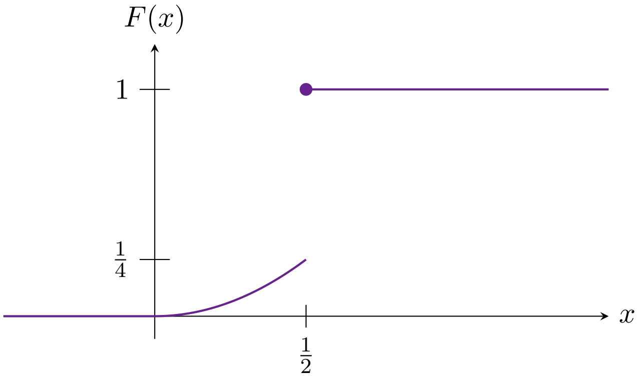
Also, for $0 < y < \frac{1}{2}$,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq y) = \int_0^y 2xdx = y^2.$$

Thus, the distribution function of Y is given by

$$F_Y(y) = y^2\mathbb{1}_{y\in(0,\frac{1}{2})} + \mathbb{1}_{y\geq\frac{1}{2}},$$

whose graph is



3.6 Joint distributions and independence

3.6.1 Joint distribution

If we have a pair of random variables (X, Y) defined in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, their joint distribution function is given by

$$F_{XY}(x, y) = F(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

If both distributions are discrete and take values x_i, $i \geq 1$ and y_j, $j \geq 1$ respectively, their joint probability function is

$$p_{XY}(x_i, y_j) = \mathbb{P}(X = x_i, Y = y_j), \quad i \geq 1, j \geq 1.$$

A joint distribution function has (joint) density if exists a function f_{XY} of two variables such that

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(s, t) dt ds, \quad \text{for all } x, y.$$

3.6.2 Marginal distribution

If both random variables are discrete, the marginal probability functions are given by

$$p_X(x_i) = \sum_j p_{XY}(x_i, y_j) \quad \text{and} \quad p_Y(y_j) = \sum_i p_{XY}(x_i, y_j).$$

If F has a joint density f, the marginal densities of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx.$$

3.6.3 Independence

If for any x and y, $F(x, y) = F_X(x) F_Y(y)$, then X and Y are independent.

If the variables are discrete with joint probability p_{XY}, they are independent if and only if

$$p_{XY}(x, y) = p_X(x) p_Y(y).$$

Similarly, if the variables are continuous with joint density $f_{XY}(x, y)$, they are independent if and only if

$$f_{XY}(x, y) = f_X(x) f_Y(y).$$

We denote that X is independent of Y as $X \perp\!\!\!\perp Y$.

3.6.4 Conditional distribution

Let X, Y be discrete random variables. The conditional probability function $p_{X|Y}(x|y)$ of X given $Y = y$ is defined by

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}, \quad \text{if } p_Y(y) > 0.$$

Note that if $X \perp\!\!\!\perp Y$, then

$$p_{X|Y}(x|y) = p_X(x)$$

Let X, Y be random variables with joint density $f_{XY}(x, y)$. We define the conditional density as

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}, \quad \text{if } f_Y(y) > 0.$$

Note that if $X \perp\!\!\!\perp Y$, then

$$f_{X|Y}(x|y) = f_X(x)$$

3.7 Statistics of random variables

3.7.1 Expected value

Moments of a distribution

If X is a random variable, the moment of order n of X is given by

$$\mathbb{E}[X^n] = \begin{cases} \sum_i x_i^n \mathbb{P}(X = x_i), & \text{if } X \text{ is a discrete random variable,} \\ \int_{-\infty}^{\infty} x^n f(x) dx, & \text{if } X \text{ is a continuous random variable,} \end{cases}$$

always that the series or integral converge absolutely. If they diverge, we say that the moment does not exists.

Expected value

The first moment, corresponding to $n = 1$, is known as the mean or expected value of X, denoted by μ. That is,

$$\mu \equiv \mathbb{E}(X)$$

Expected value of a function of a r.v.

If X is a random variable and g is a function, then $g(X)$ is also a random variable, then the expected value of $g(X)$ is given by

$$\mathbb{E}[g(X)] = \begin{cases} \sum_j g(x_i) \mathbb{P}(X = x_i), & \text{if } X \text{ is a discrete random variable,} \\ \int_{-\infty}^{\infty} g(x) f(x) dx, & \text{if } X \text{ is a continuous random variable,} \end{cases}$$

always that the series or integral converge absolutely.

Conditional expectation

Let X, Y be random variables, and g be a function such that $\mathbb{E}[g(X)] < \infty$. We define the conditional expected value of $g(X)$ given $Y = y$ as

$$\mathbb{E}[g(X)|Y = y] = \begin{cases} \sum_x g(x) p_{X|Y}(x|y), & \text{if } X, Y \text{ are a discrete random variables, and } p_Y(y) > 0, \\ \int g(x) f_{X|Y}(x|y) dx, & \text{if } X, Y \text{ are a continuous random variables, and } f_Y(y) > 0. \end{cases}$$

3.7.2 Median and mode

Median

The median of a random variable X is any value $\tilde{\mu}$ such that

$$\mathbb{P}(X \geq \tilde{\mu}) \geq \frac{1}{2}, \quad \mathbb{P}(X < \tilde{\mu}) \geq \frac{1}{2}.$$

Mode

If X is a discrete random variable, the mode is the value x at which the probability mass function takes its maximum value.

When the density function of a continuous distribution has multiple local maxima it is common to refer to all of the local maxima as modes of the distribution.

3.7.3 Variance and standard deviation

Central moments

The central moment of order n is the moment of order n of the variable $(X - \mu)$, always that μ exists. The nth central moment is denoted as μ_n. That is

$$\mu_n = \mathbb{E}(X - \mu)^n$$

Variance

The second central moment is called the variance of X, denoted as $\mathbb{V}[X]$. It can be shown that

$$\mathbb{V}[X] = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2.$$

Standard deviation

The squared root of the variance is called the standard deviation of X, denoted by $\mathrm{sd}(X)$.

It is common to denote the variance of X by σ^2, and the standard deviation by σ.

3.7.4 Skewness and kurtosis

The coefficient of skewness of a random variable X is defined by

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}.$$

The kurtosis of a random variable X is defined as

$$\gamma_2 = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}.$$

3.7.5 Quantile

Let $F : \mathbb{R} \to [0, 1]$ be the distribution function of a random variable X, we define the quantile function $Q : (0, 1) \to \mathbb{R}$ as the generalized inverse of F. The quantile of probability p of X, Q_p is any value such that

$$F(Q_p) = p$$

3.7.6 Covariance and correlation

Let X and Y be random variables with joint distribution, means μ_x and μ_Y, and finite variances σ_X^2 and σ_Y^2.

The covariance of X and Y, denoted as σ_{XY} or $\mathrm{Cov}(X, Y)$ is defined as

$$\sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y.$$

The correlation of X and Y, denoted as ρ_{XY} or $\mathrm{Corr}(X, Y)$ is defined as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

We say that X and Y are uncorrelated if $\rho_{XY} = 0$.

Independent variables with finite variance are uncorrelated, but there are uncorrelated variables that are not independent.

4

Statistical Inference

4.1 Estimation

4.1.1 Law of large numbers

Let $X_1, \dots, X_n$ be i.i.d. r.v. with the same expected value $\mu = \mathbb{E}(X_i)$, $i = 1, \dots, n$.

Weak law of large numbers

$\bar{X}_n \xrightarrow[n\to\infty]{P} \mu$

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{X}_n - \mu| < \varepsilon\right) = 1$$

Strong law of large numbers

$\bar{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \mu$

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$$

Due to the law of large numbers, we can estimate the expected value of a function of a random variable, using its average. That is, we can estimate $\mathbb{E}[g(X)]$ by

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

4.1.2 Method of moments

To estimate the parameters of a distribution, we can choose the expected value of several functions (as much as needed) and equalize them with their empirical versions.

For instance, let μ and σ^2 be the mean and variance of a distribution, we can estimate such quantities solving the system of equations

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Distribution	$\mathbb{E}(X)$, μ	$\mathbb{V}(X)$, σ^2	**Estimator**
Bernoulli(p)	p	$p(1-p)$	$\hat{p} = \bar{X}_n$
Binomial(n, p)	np	$np(1-p)$	$\hat{p} = X/n$
Poisson(λ)	λ	λ	$\hat{\lambda} = \bar{X}_n$
Geometric(p)	$1/p$	$(1-p)/p^2$	$\hat{p} = 1/\bar{X}_n$
Exponential(λ)	λ	λ^2	$\hat{\lambda} = \bar{X}_n$
$\mathcal{N}(\mu, \sigma^2)$	μ	σ^2	$\hat{\mu} = \bar{X}_n$ $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

Inference for the gamma distribution

If $X \sim \text{Gamma}(\alpha, \beta)$, then

$$\mathbb{E}(X) = \frac{\alpha}{\beta}, \quad \mathbb{V}(X) = \frac{\alpha}{\beta^2}.$$

Thus, using the method of moments, we can estimate α and β solving the next equations:

$$\bar{X} = \frac{\hat{\alpha}}{\hat{\beta}}, \quad s^2 = \frac{\hat{\alpha}}{\hat{\beta}^2 0}.$$

Getting,

$$\hat{\alpha} = \frac{\bar{X}^2}{s^2}, \quad \hat{\beta} = \frac{\bar{X}}{s^2}.$$

Inference for the Weibull distribution

Similarly to the method of moments, we can estimate parameters equalizing theoretical quantiles with their empirical counterparts. For instance if $X \sim \text{Weibull}(\rho, \lambda)$, then its quantile function is given by

$$Q(p) = \lambda(-\log(1-p))^{1/\rho}.$$

If $p = 1 - e^{-1}$, then

$$Q(1 - e^{-1}) = \lambda.$$

Thus, we can estimate λ through

$$\hat{\lambda} = Q(1 - e^{-1}).$$

To estimate ρ, we can another quantile of an arbitrary probability. For instance, we can choose $p = 1/2$, getting

$$\hat{\rho} = \frac{\log\log 2}{\log\left(\tilde{X}/\hat{\lambda}\right)}.$$

4.1.3 Central limit theorem

Central limit theorem

Let $X_1, \ldots, X_n$ be i.i.d. r.v. with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2 < \infty$, then

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

Thus, if $\hat{\sigma}$ is a consistent estimator of σ,

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\hat{\sigma}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

Hence, for n sufficiently large

$$\bar{X}_n \mathrel{\dot\sim} \mathcal{N}\left(\mu, \hat{\sigma}^2/n\right)$$

Central limit theorem for the median

Let $X_1, \ldots, X_n$ be i.i.d. r.v. with density function f. For n sufficiently large,

$$\tilde{X}_n \dot{\sim} \mathcal{N}\left(\tilde{\mu}, \frac{1}{4nf^2(\tilde{\mu})}\right),$$

where $\tilde{X}_n$ is the median of the sample, and $\tilde{\mu}$ is the median of the distribution.

In particular, if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\tilde{\mu} = \mu$ and

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Thus,

$$f(\tilde{\mu}) = \frac{1}{\sqrt{2\pi}\sigma},$$

and

$$\tilde{X}_n \dot{\sim} \mathcal{N}\left(\mu, \frac{\pi\sigma^2}{2n}\right)$$

4.1.4 Confidence interval

Let θ be an unknown parameter, or a function of unknown parameters, of a distribution. Fix $\alpha \in (0,1)$, where α is a small positive number, known as the *level of significance* (a widely used value for α is 0.05). An interval CI $= (L, U)$ such that

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha$$

is called a *confidence interval of probability* $(1-\alpha) \times 100\%$ for θ.

Confidence interval for μ

Therefore,

$$\left(\mathcal{N}\left(\bar{X}_n, \hat{\sigma}^2/n\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\bar{X}_n, \hat{\sigma}^2/n\right)_{1-\alpha/2}\right)$$

is a confidence interval of approx. $(1-\alpha) \times 100\%$ of probability for μ.

Confidence interval for $\tilde{\mu}$, for the normal distribution

A confidence interval of approx. $(1-\alpha) \times 100\%$ of probability for μ is given by

$$\left(\mathcal{N}\left(\tilde{X}_n, {}^{\pi\hat{\sigma}^2}/{}_{2n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\tilde{X}_n, {}^{\pi\hat{\sigma}^2}/{}_{2n}\right)_{1-\alpha/2}\right),$$

where, $\hat{\sigma}$ is a consistent estimator of σ.

4.1.5 Prediction interval

A common misinterpretation of confidence intervals, especially for the mean μ, is that they predict, with some level of confidence, an interval where a new observation might arise.

That is, a confidence interval CI $= (L, U)$, such that

$$\mathbb{P}(L < \mu \leq U) = 1 - \alpha$$

is sometimes **misinterpreted as "there is a probability of** $(1-\alpha) \times 100\%$ **that a new observation will be in the interval** (L, U)**"**.

Prediction interval

A confidence interval indicates, with some confidence level, where might be the unknown parameter, but does not indicate where might be a new observation, X_{new}.

A prediction interval, PI $= (L, U)$, does indicate where a new observation might be, with some confidence level. That is, a prediction interval PI $= (L, U)$ with a confidence level of $(1-\alpha) \times 100\%$ is such that

$$\mathbb{P}(L \leq X_{\text{new}} \leq U) = 1 - \alpha.$$

Prediction interval for the normal distribution, using the mean

Let be $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, and $X_{\text{new}} \sim \mathcal{N}(\mu, \sigma^2)$ a new observation that we want to predict, such that X_{new} is independent from our sample. That is,

$$X_{\text{new}} \perp\!\!\!\perp \mathbf{X} \Rightarrow X_{\text{new}} \perp\!\!\!\perp \bar{X}_n.$$

Then,

$$X_{\text{new}} - \bar{X}_n \sim \mathcal{N}\left(0, \sigma^2 + {}^{\sigma^2}\!/_{n}\right)$$

$$\Rightarrow X_{\text{new}} \sim \mathcal{N}\left(\bar{X}_n, \sigma^2 + {}^{\sigma^2}\!/_{n}\right).$$

Hence,

$$\left(\mathcal{N}\left(\bar{X}_n, \sigma^2 + {}^{\sigma^2}\!/_{n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\bar{X}_n, \sigma^2 + {}^{\sigma^2}\!/_{n}\right)_{1-\alpha/2}\right)$$

is a prediction interval of probability $(1-\alpha) \times 100\%$.

Usually, σ is unknown, so we need to substitute σ by a consistent estimator $\hat{\sigma}$. In the particular case that we estimate σ by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2},$$

then

$$\left(t_{n-1}\left(\bar{X}_n, s^2 + {}^{s^2}\!/_{n}\right)_{\alpha/2} \quad ; \quad t_{n-1}\left(\bar{X}_n, s^2 + {}^{s^2}\!/_{n}\right)_{1-\alpha/2}\right)$$

would be a prediction interval of probability $(1-\alpha) \times 100\%$.

Prediction interval for the normal distribution, using the median

Under the same assumptions as before, we would have that

$$X_{\text{new}} - \tilde{X}_n \overset{\cdot}{\sim} \mathcal{N}\left(0, \sigma^2 + {}^{\pi\sigma^2}\!/_{2n}\right)$$

$$\Rightarrow X_{\text{new}} \overset{\cdot}{\sim} \mathcal{N}\left(\tilde{X}_n, \sigma^2 + {}^{\pi\sigma^2}\!/_{2n}\right).$$

Hence,

$$\left(\mathcal{N}\left(\tilde{X}_n, \sigma^2 + {}^{\pi\sigma^2}\!/_{2n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\tilde{X}_n, \sigma^2 + {}^{\pi\sigma^2}\!/_{2n}\right)_{1-\alpha/2}\right)$$

is a prediction interval of approx. probability $(1-\alpha) \times 100\%$..

Prediction interval for the normal distribution, using Q_1 and Q_3

Consider the same assumptions as before. Let $\alpha \in (0, 0.5]$ be the significance level, and

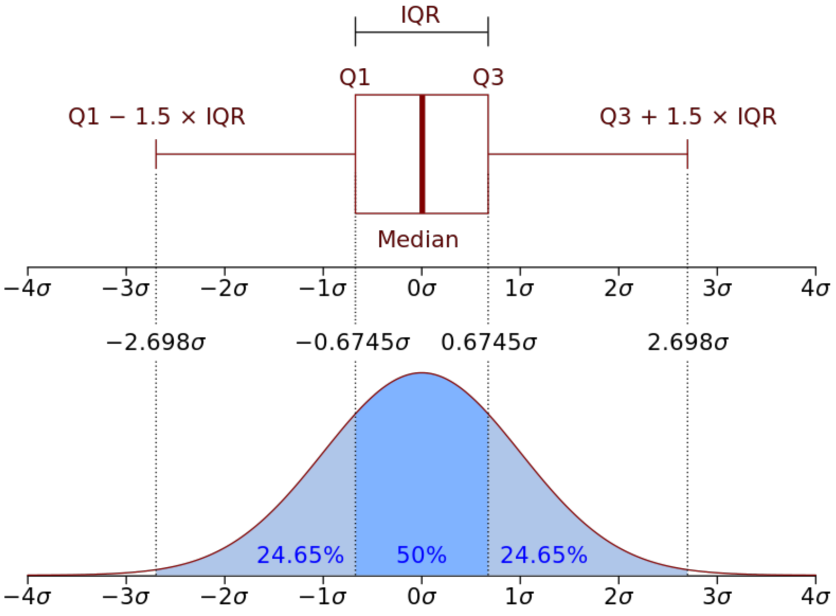
$$\delta = \frac{1}{2}\left(\frac{\Phi^{-1}(1-\alpha/2)}{\Phi^{-1}(3/4)} - 1\right).$$

Then,

$$\Big(Q_1 - \delta\,\mathrm{IQR} \quad ; \quad Q_3 + \delta\,\mathrm{IQR}\Big)$$

is a prediction interval of probability $(1-\alpha) \times 100\%$.

In particular, $\delta = 1.5 \Leftrightarrow \alpha \approx 0.007$.



4.1.6 Determining the sample size

Let $X_1, \ldots, X_n$ be i.i.d. r.v. with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2 < \infty$. Then, according to the central limit theorem

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

Thus,

$$\mathbb{P}\left(-\Phi^{-1}(1-\alpha/2) \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq \Phi^{-1}(1-\alpha/2)\right) \approx 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(\sqrt{n}\frac{|\bar{X}_n - \mu|}{\sigma} \leq \Phi^{-1}(1-\alpha/2)\right) \approx 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(|\bar{X}_n - \mu| \leq \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha/2)\right) \approx 1-\alpha$$

Let be

$$n = \left(\frac{\Phi^{-1}(1-\alpha/2)\sigma}{\varepsilon}\right)^2,$$

for some value $\varepsilon > 0$, fixed before hand. Then,

$$\mathbb{P}\left(|\bar{X}_n - \mu| \leq \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha/2)\right) \approx 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(|\bar{X}_n - \mu| \leq \varepsilon\right) \approx 1-\alpha.$$

That is, if we take a sample of size

$$n = \left(\frac{\Phi^{-1}(1-\alpha/2)\sigma}{\varepsilon}\right)^2,$$

we can be $(1-\alpha) \times 100\%$ confident that the error $|\bar{X}_n - \mu|$ will not exceed ε.

4.1.7 Summary

Punctual estimation for the normal distribution

Let be $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then,

- Usual estimators:

$$\hat{\mu} = \bar{X}_n; \quad \hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2; \quad \hat{\sigma} = s$$

- Robust estimators:

$$\hat{\mu} = \tilde{X}_n; \quad \hat{\sigma} = \frac{\text{MAD}}{\Phi^{-1}(3/4)}; \quad \hat{\sigma} = \frac{\text{IQR}}{2\Phi^{-1}(3/4)}; \quad \hat{\sigma} = \text{AAD}\sqrt{\pi/2}$$

Confidence interval for μ using the average

- CI for μ based on $\bar{X}_n$ and σ:

$$\left(\mathcal{N}\left(\bar{X}_n, {}^{\sigma^2}\!/_{n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\bar{X}_n, {}^{\sigma^2}\!/_{n}\right)_{1-\alpha/2}\right)$$

- CI for μ based on $\bar{X}_n$ and s:

$$\left(t_{n-1}\left(\bar{X}_n, {}^{s^2}\!/_{n}\right)_{\alpha/2} \quad ; \quad t_{n-1}\left(\bar{X}_n, {}^{s^2}\!/_{n}\right)_{1-\alpha/2}\right)$$

- CI for μ based on $\bar{X}_n$ and a consistent estimator $\hat{\sigma}$:

$$\left(\mathcal{N}\left(\bar{X}_n, {}^{\hat{\sigma}^2}\!/_{n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\bar{X}_n, {}^{\hat{\sigma}^2}\!/_{n}\right)_{1-\alpha/2}\right)$$

Confidence interval for μ using the median

- CI for μ based on $\tilde{X}_n$ and σ:

$$\left(\mathcal{N}\left(\tilde{X}_n, {}^{\pi\sigma^2}\!/_{2n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\tilde{X}_n, {}^{\pi\sigma^2}\!/_{2n}\right)_{1-\alpha/2}\right)$$

- CI for μ based on $\tilde{X}_n$ and a consistent estimator $\hat{\sigma}$:

$$\left(\mathcal{N}\left(\tilde{X}_n, {}^{\pi\hat{\sigma}^2}\!/_{2n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\tilde{X}_n, {}^{\pi\hat{\sigma}^2}\!/_{2n}\right)_{1-\alpha/2}\right)$$

Confidence interval for σ^2 and σ

- CI for σ^2 based on s^2:

$$\left(\frac{s^2(n-1)}{\chi^2_{n-1,1-\alpha/2}} \quad ; \quad \frac{s^2(n-1)}{\chi^2_{n-1,\alpha/2}}\right)$$

- CI for σ based on s:

$$\left(\frac{s\sqrt{n-1}}{\sqrt{\chi^2_{n-1,1-\alpha/2}}} \quad ; \quad \frac{s\sqrt{n-1}}{\sqrt{\chi^2_{n-1,\alpha/2}}}\right)$$

Prediction interval using the average

- PI based on $\bar{X}_n$ and σ:

$$\left(\mathcal{N}\left(\bar{X}_n, \sigma^2 + {\sigma^2}/{n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\bar{X}_n, \sigma^2 + {\sigma^2}/{n}\right)_{1-\alpha/2}\right)$$

- PI based on $\bar{X}_n$ and s:

$$\left(t_{n-1}\left(\bar{X}_n, s^2 + {s^2}/{n}\right)_{\alpha/2} \quad ; \quad t_{n-1}\left(\bar{X}_n, s^2 + {s^2}/{n}\right)_{1-\alpha/2}\right)$$

- PI based on $\bar{X}_n$ and a consistent estimator $\hat{\sigma}$:

$$\left(\mathcal{N}\left(\bar{X}_n, \hat{\sigma}^2 + {\hat{\sigma}^2}/{n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\bar{X}_n, \hat{\sigma}^2 + {\hat{\sigma}^2}/{n}\right)_{1-\alpha/2}\right)$$

Prediction interval using the median

- PI based on $\tilde{X}_n$ and σ:

$$\left(\mathcal{N}\left(\tilde{X}_n, \sigma^2 + {\pi\sigma^2}/{2n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\tilde{X}_n, \sigma^2 + {\pi\sigma^2}/{2n}\right)_{1-\alpha/2}\right)$$

- PI based on $\tilde{X}_n$ and a consistent estimator $\hat{\sigma}$:

$$\left(\mathcal{N}\left(\tilde{X}_n, \hat{\sigma}^2 + {\pi\hat{\sigma}^2}/{2n}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(\tilde{X}_n, \hat{\sigma}^2 + {\pi\hat{\sigma}^2}/{2n}\right)_{1-\alpha/2}\right)$$

Prediction interval using Q_1 and Q_3

Let be $\alpha \in (0, 0.5]$, and

$$\delta = \frac{1}{2}\left(\frac{\Phi^{-1}(1-\alpha/2)}{\Phi^{-1}(3/4)} - 1\right),$$

$$\left(Q_1 - \delta\,\mathrm{IQR} \quad ; \quad Q_3 + \delta\,\mathrm{IQR}\right)$$

4.2 Maximum likelihood

4.2.1 Framework

Likelihood function

Let $X_1, \ldots, X_n$ be i.i.d. discrete random variables with joint probability function

$$p(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} p(X_i; \theta),$$

We define the likelihood function as a function of the parameters θ that is proportional to the joint probability function. That is,

$$L(\theta) = Cp(X_1, \ldots, X_n; \theta) = C \prod_{i=1}^{n} p(X_i; \theta),$$

where $C > 0$ is a constant that might depend on the sample $\mathbf{X}$ but not on the parameters.

Similarly, if $X_1, \ldots, X_n$ are i.i.d. continuous random variables with joint density function

$$f(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta),$$

We define the likelihood function as a function of the parameters θ that is proportional to the joint density function. That is,

$$L(\theta) = Cf(X_1, \ldots, X_n; \theta) = C \prod_{i=1}^{n} f(X_i; \theta),$$

where $C > 0$ is a constant that might depend on the sample $\mathbf{X}$ but not on the parameters.

Maximum likelihood estimator (MLE)

We define the maximum likelihood estimator (MLE), $\hat{\theta}$, for the parameters θ, as the values where the likelihood is maximized. That is,

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta)$$

In other words, $\hat{\theta}_{\text{MLE}}$ are the values of the parameters that maximize the probability of the observed sample.

Log-likelihood

We define the log-likelihood as

$$\ell(\theta) = \log L(\theta).$$

If $\hat{\theta}_{\text{MLE}}$ is the value that maximizes the likelihood, it can be proved that it also maximizes the log-likelihood.

That is,

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell(\theta)$$

Remember that if $X_1, \ldots, X_n$ are i.i.d. r.v., then

$$L(\theta) = C \prod_{i=1}^{n} f(X_i; \theta).$$

Hence,

$$\ell(\theta) = \log C + \sum_{i=1}^{n} \log f(X_i; \theta).$$

Entropy

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(X_i;\theta)$$

$$\Leftrightarrow$$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i;\theta)$$

$$\Leftrightarrow$$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} - \log f(X_i;\theta)$$

In information theory, the entropy of a random variable X with density function f is defined as

$$H(X) = \mathbb{E}[-\log f(X)],$$

Thus, note that

$$\frac{1}{n} \sum_{i=1}^{n} - \log f(X_i;\theta)$$

is just the empirical version of the entropy. And the MLE is such that minimizes the empirical entropy of the model.

Score function and likelihood equations

We define the empirical score function as

$$sc_n(\theta) = \frac{\partial \ell(\theta)}{\partial \theta},$$

when such derivatives exist.

Because

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \ell(\theta).$$

Then, it must satisfies that

$$\frac{\partial \ell(\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} = 0.$$

That is, the MLE are the values where the empirical score function equals zero,

$$sc_n(\hat{\theta}_{\mathrm{MLE}}) = 0.$$

Thus, we can find the MLE solving this system of equations known as *likelihood equations.*

Relative likelihood

We define the relative likelihood as

$$R(\theta) = \frac{L(\theta)}{L\left(\hat{\theta}_{\mathrm{MLE}}\right)}.$$

That is, R is the likelihood L, but it is normalized to take values in $(0,1)$, and $R(\hat{\theta}_{\mathrm{MLE}}) = 1$.

Relative log-likelihood

We define the relative log-likelihood as

$$r(\theta) = \log R(\theta) = \ell(\theta) - \ell(\hat{\theta}_{\text{MLE}}).$$

It can be shown that, under very general conditions known as regularity conditions,

$$-2r(\theta_0) \dot{\sim} \chi^2_d,$$

where $d = \dim(\theta)$, and θ_0 is the real value of θ.

Likelihood region

A likelihood region of level $c \in (0, 1)$ is defined as the values of θ such that $R(\theta) \geq c$,

$$\{\theta : R(\theta) \geq c\}.$$

Confidence region

It can be shown that, under very general conditions, known as regularity conditions, a likelihood region of level

$$c = \exp\left\{-\frac{1}{2}\chi^2_{d,1-\alpha}\right\}, \quad \text{where } d = \dim(\theta),$$

corresponds (approx.) with a confidence region of $(1 - \alpha) \times 100\%$ of probability.

4.2.2 Inference for the exponential distribution

Consider that we have a sample $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{Exponential}(\lambda)$, whose density function is given by

$$f(X|\lambda) = \frac{1}{\lambda}e^{-X/\lambda}.$$

Likelihood, log-likelihood and score functions

The likelihood, log-likelihood and score functions are given by

$$L(\lambda) = \frac{1}{\lambda^n}e^{-\frac{1}{\lambda}\sum X_i},$$

$$\ell(\lambda) = -n\log\lambda - \frac{n}{\lambda}\bar{X},$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{-n}{\lambda} + \frac{n\bar{X}}{\lambda^2}.$$

Maximum likelihood estimator

The maximum likelihood estimator must satisfy

$$\left.\frac{\partial \ell}{\partial \lambda}\right|_{\lambda=\hat{\lambda}} = 0 \Rightarrow \hat{\lambda} = \bar{X}.$$

Confidence interval based on the relative likelihood

The relative log-likelihood is then given by

$$
\begin{aligned}
r(\lambda) &= \ell(\lambda) - \ell(\hat{\lambda}) \\
&= -n \log \lambda - \frac{n}{\lambda}\bar{X} + n \log \bar{X} + n \\
&= n\left(\log \frac{\bar{X}}{\lambda} - \frac{\bar{X}}{\lambda} + 1\right)
\end{aligned}
$$

The relative likelihood can be calculated as

$$R(\lambda) = \exp\left\{r(\lambda)\right\}.$$

A likelihood interval of level $c = \exp\left\{-\frac{1}{2}\chi^2_{1,1-\alpha}\right\}$ corresponds with a confidence interval of probability $1-\alpha$.

Confidence interval based on the central limit theorem

On the other hand, $\mathbb{E}(X|\lambda) = \lambda$ and $\mathbb{V}(X|\lambda) = \lambda^2$. Thus, using the central limit theorem, a confidence interval of (approx.) probability $1-\alpha$ is given by

$$\left(\mathcal{N}(\bar{X}, {}^{\lambda^2}\!/_{n})_{\alpha/2} \quad ; \quad \mathcal{N}(\bar{X}, {}^{\lambda^2}\!/_{n})_{1-\alpha/2}\right),$$

estimating $\hat{\lambda} = \bar{X}$:

$$\left(\mathcal{N}(\bar{X}, {}^{\bar{X}^2}\!/_{n})_{\alpha/2} \quad ; \quad \mathcal{N}(\bar{X}, {}^{\bar{X}^2}\!/_{n})_{1-\alpha/2}\right)$$

Predictive distribution

It can be proved that the predictive distribution is given by

$$\mathsf{Lomax}\left(\alpha = n+1, \beta = n\bar{X}\right),$$

where α is the shape parameter, and β is the scale parameter. Thus, prediction intervals can be calculated from this distribution.

4.2.3 Inference for the normal distribution

Consider that we have a sample $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, whose density function is given by

$$f(X|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(X-\mu)^2}{2\sigma^2}\right\}.$$

Likelihood, log-likelihood functions

The likelihood, and log-likelihood functions are given by

$$L(\mu,\sigma) = \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\}$$

$$\ell(\mu,\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

Score functions and likelihood equations

The likelihood equations are given by

$$\left.\frac{\partial \ell}{\partial \mu}\right|_{(\mu,\sigma)=(\hat{\mu},\hat{\sigma})} = 0, \quad \left.\frac{\partial \ell}{\partial \sigma}\right|_{(\mu,\sigma)=(\hat{\mu},\hat{\sigma})} = 0$$

where,

$$\begin{cases} \frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) \\ \\ \frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(X_i-\mu)^3}{\sigma^3} \end{cases}$$

Maximum likelihood estimator

Thus,

$$\sum_{i=1}^{n} X_i - n\hat{\mu} = 0 \Rightarrow \hat{\mu} = \bar{X},$$

and

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(X_i-\hat{\mu})^2 = n \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

That is, the maximum likelihood estimators for μ and σ are given by

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})}$$

Relative likelihood and likelihood region

We now calculate the relative log-likelihood:

$$\begin{aligned} r(\mu,\sigma) &= \ell(\mu,\sigma) - \ell(\hat{\mu},\hat{\sigma}) \\ &= -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 + n\log\hat{\sigma} + \frac{n}{2} \\ &= n\log\frac{\hat{\sigma}}{\sigma} + \frac{n}{2} - \frac{1}{2\sigma^2}\left(n\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2n\mu\bar{X} + n\mu^2\right) \\ &= n\log\frac{\hat{\sigma}}{\sigma} + \frac{n}{2}\left(1 - \frac{1}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\mu\bar{X} + \mu^2\right)\right) \\ r(\mu,\sigma) &= n\left[\log\frac{\hat{\sigma}}{\sigma} + \frac{1}{2}\left(1 - \frac{1}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\mu\bar{X} + \mu^2\right)\right)\right]. \end{aligned}$$

The relative likelihood is then obtained doing

$$R(\mu,\sigma) = \exp\left\{r(\mu,\sigma)\right\}.$$

A likelihood region of level $c = \exp\left\{-\frac{1}{2}\chi^2_{2,1-\alpha}\right\}$ corresponds with a confidence region of probability $(1-\alpha)\times 100\%$.

4.3 Hypothesis testing

4.3.1 Framework

Null hypothesis

The hypothesis testing enables us to assess the extent to which a data set $\mathcal{D}_n$ is consistent with a particular hypothesis, known as the *working* or *null hypothesis*.

The null hypothesis generally represents a simplified view of the data-generating process, such as:

- $\mu = \mu_0$.
- $\tilde{\mu} = \tilde{\mu}_0$.
- Two variables are independent.
- Etc.

The null hypothesis is then the hypothesis that will be adopted, and subsequently acted upon, unless the data indicate that it is untenable.

Test statistic

The first step is to formulate a test statistic that **measures the extent to which the observed data depart from the null hypothesis**.

It is so constructed that the larger its value, the greater the departure from the null hypothesis.

P-value

Once the value of the test statistic has been obtained from the observed data, we calculate the **probability of obtaining a value as extreme or more extreme than the observed value**, when the null hypothesis is true.

This quantity **summarizes the strength of the evidence** in the sample data **against the null hypothesis** and is known as the *probability value* or *P-value*.

If the P-value is large, we would conclude that it is quite likely that the observed data would have been obtained when the null hypothesis was true, and that there is no evidence to reject the null hypothesis. On the other hand, if the P-value is small, this would be interpreted as evidence against the null hypothesis; the smaller the P-value, the stronger the evidence.

Null distribution

In order to obtain the P-value for a hypothesis test, the statistic must have a probability distribution that is known, or at least approx. known, when the null hypothesis is true. This probability distribution is referred to as the *null distribution* of the test statistic.

Level of significance

If a P-value is smaller than some value α, we say that the hypothesis is rejected at an $\alpha \times 100\%$ level of significance. Instead of reporting that a null hypothesis is rejected or not rejected at some specified significance level, a more satisfactory policy is to report the actual P-value.

P-value	Interpretation
$P > 0.1$	No evidence to reject the null hypothesis
$0.05 < P \leq 0.1$	Slight evidence against the null hypothesis
$0.01 < P \leq 0.05$	Moderate evidence against the null hypothesis
$0.001 < P \leq 0.01$	Strong evidence against the null hypothesis
$P \leq 0.001$	Overwhelming evidence against the null hypothesis

These guidelines are not hard-and-fast rules and should not be interpreted rigidly. For example, there is no practical difference between a P-value of 0.046 and 0.056.

In deciding on a course of action, the statistical evidence summarized in the P-value will be just one ingredient of the decision-making process. **In addition to the statistical evidence, there will also be scientific evidence to consider**.

4.3.2 One sample Student's t-test

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, if the hypothesis

$$H : \mu = \mu_0$$

were true, then

$$\sqrt{n}\frac{\bar{X}_n - \mu_0}{s}\Big| H \sim t_{n-1}.$$

Thus, we can take $T = \sqrt{n}\frac{\bar{X}_n - \mu_0}{s}$ as our test statistic.

Note that if T were large positive that would indicate that $\bar{X}_n$ is much greater than μ_0, and would correspond with a departure from the null hypothesis.

Similarly, large negative values would indicate that $\bar{X}_n$ is much smaller than μ_0, and would correspond with a departure from the null hypothesis as well. Thus, if t is the observed value of T, the appropriate P-value is

$$\mathbb{P}(T \leq -|t|) + \mathbb{P}(T \geq |t|).$$

In view of the symmetry of the t distribution,

$$P\text{-value} = 2\mathbb{P}\left(T \geq |t| \Big| H\right) = 2\mathbb{P}\left(t_{n-1} \geq |t|\right)$$

4.3.3 Two samples Student's t-test

Suppose that $X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)}$ and $X_1^{(2)}, X_2^{(2)}, \ldots, X_{n_2}^{(2)}$ are i.i.d. samples from two populations, each with a normal distribution. Assume that we want to test the hypothesis of equal means,

$$H : \mu_1 = \mu_2$$

Same sample size, and same variance

If $n_1 = n_2$, and $\sigma_1^2 = \sigma_2^2$. Then,

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{2}{n}}}\Big| H \sim t_{2n-2},$$

where

$$s_p = \sqrt{\frac{s_1^2 + s_2^2}{2}},$$

and s_1^2 and s_2^2 are the unbiased estimators of the population variance.

Same sample size, but different variance

If $n_1 \neq n_2$, but $\sigma_1^2 = \sigma_2^2$. Then,

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\Big| H \sim t_{n_1+n_2-2},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

Different sample size (Welch's t-test)

If $\sigma_1^2 \neq \sigma_2^2$. Then,

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \Big| H \dot{\sim} t_{df},$$

where

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

4.3.4 Levene's and Bartlett's tests

Suppose that $X_1^{(1)}, X_2^{(1)}, \ldots, X_{n_1}^{(1)}$ and $X_1^{(2)}, X_2^{(2)}, \ldots, X_{n_2}^{(2)}$ are i.i.d. samples from two populations, each with a normal distribution. Assume that we want to test the hypothesis of equal variances,

$$H : \sigma_1^2 = \sigma_2^2$$

F-test

Then,

$$F = \frac{s_1^2}{s_2^2} \Big| H \sim F_{n_1, n_2},$$

where F_{n_1,n_2} is an F distribution with n_1 and n_2 degrees of freedom.

Levene's and Bartlett's tests

However, **the F-test is known to be extremely sensitive to non-normality**, so Levene's and Bartlett's tests are better for testing the equality of two variances.

4.3.5 One-way ANOVA

Suppose that we have samples from J populations, each one with a normal distribution.And we want to test if their means can be considered equal:

$$H : \mu_1 = \mu_2 = \cdots = \mu_J$$

Imagine that we take a sample from each population, being n_j the population size for the jth population. Suppose, that the observations, y_{ij} are independent between them and from other populations. Define

- $\bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{j=1}^{J} y_{ij}$
- $n_{\text{tot}} = \sum_{j=1}^{J} n_j$
- $\bar{y}_{\cdot\cdot} = \frac{1}{n_{\text{tot}}} \sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij}$
- Sum of squares between groups:

$$SS_B = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$$

- Sum of squares within groups:

$$SS_W = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot j})^2$$

- Total sum of squares:

$$SS_T = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{..})^2$$

We define the ANOVA table as

	df	SS	MS	F
Between groups	$df_B = J - 1$	SS_B	$MS_B = SS_B/df_B$	$F = MS_B/MS_W$
Within groups	$df_W = n_{\text{tot}} - J$	SS_W	$MS_W = SS_W/df_W$	
Total	$df_T = n_{\text{tot}} - 1$	SS_T		

Under the hypothesis

$$H : \mu_1 = \mu_2 = \cdots = \mu_J,$$

$F = MS_B/MS_W$ follows an F distribution with df_B and df_W degrees of freedom. Hence, the appropriate P-value for this hypothesis test is

$$\mathbb{P}(F_{df_B,df_W} \geq f),$$

where f is the observed value of the test statistic F.

4.4 Linear regression

4.4.1 Framework

In regression analysis we consider a pair (X, Y), where X is $\mathbb{R}^p$-valued $(X \in \mathcal{X} \subseteq \mathbb{R}^p)$ and Y is $\mathbb{R}$-valued $(Y \in \mathcal{Y} \subset \mathbb{R})$.

The aim is to find a function $f : \mathcal{X} \to \mathcal{Y}$, such that $f(X)$ is a good approximation of Y.

Linear regression

We can propose f in different ways. For instance, we can use a linear expression:

$$f(X) = \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)}.$$

This is the linear regression model.

Because f just approximates Y. We need to add some *errors* to the model. Thus,

$$\begin{aligned} Y &= f(X) + \varepsilon \\ &= \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)} + \varepsilon \\ &= \beta^T X + \varepsilon \end{aligned}$$

The error ε is the part of Y that cannot be explained by our model f.

To complete our model, we need to specify a distribution of probability for the error. It is usual to assume that the errors follow a normal distribution with mean zero and variance $\sigma^2 < \infty$. So,

$$Y = \beta^T X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

That is,

$$Y \sim \mathcal{N}(\beta^T X, \sigma^2).$$

4.4.2 Estimation of the parameters

The parameters of the model, $\beta_1, \ldots, \beta_p$, σ^2, are unknown and need to be estimated.

Suppose that we have n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let be

$$\mathbf{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

$\mathbf{X}$ is known as the design matrix, and $\mathbf{Y}$ is known as the response vector.

Estimation of β

Let be $\hat{\beta}$ the maximum likelihood estimator (MLE) of β. It can be proved that $\hat{\beta}$ can be found solving the system of equations:

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0.$$

These equations are called the *normal equations*, whose solution is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Once we have estimated β, we can predict Y through

$$\widehat{Y} = \hat{\beta}_1 X^{(1)} + \hat{\beta}_2 X^{(2)} + \hat{\beta}_p X^{(p)}$$

Estimation of σ^2

Remember that our model can be written as

$$Y = \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)} + \varepsilon,$$

and $\mathbb{V}(\varepsilon) = \sigma^2$.

Thus, the error in the observation i can be estimated by

$$\hat{\varepsilon}_i = Y_i - \widehat{Y}_i,$$

this quantities are known as *residuals.*

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2,$$

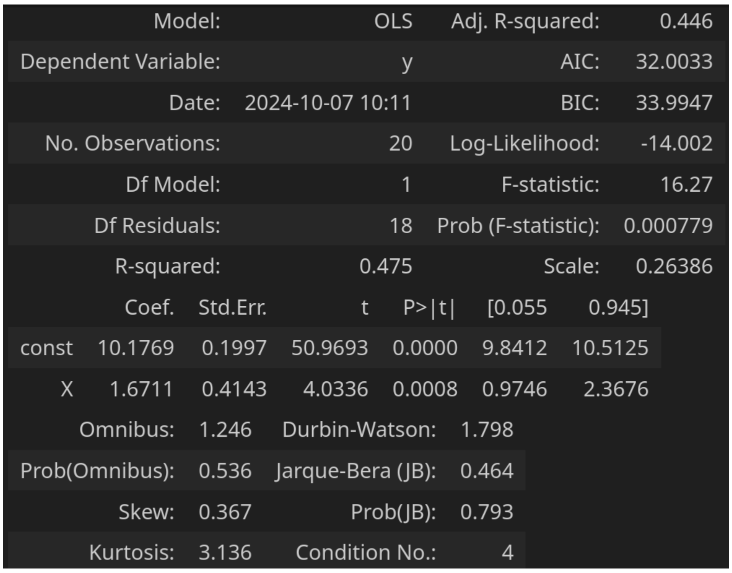
known as the *mean squared error (MSE).*

4.4.3 Linear regression in practice

Assume that we want to fit a regression model of the form

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Here is a possible summary of the trained model:

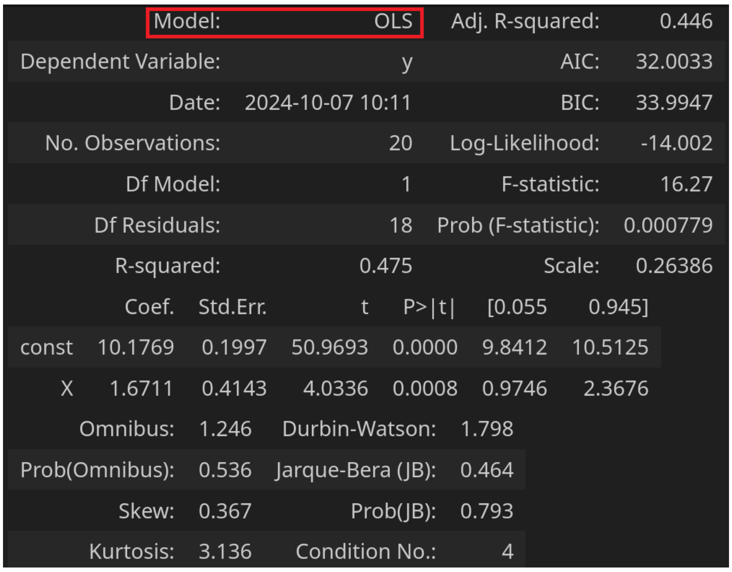


Model:	OLS	Adj. R-squared:	0.446
Dependent Variable:	y	AIC:	32.0033
Date:	2024-10-07 10:11	BIC:	33.9947
No. Observations:	20	Log-Likelihood:	-14.002
Df Model:	1	F-statistic:	16.27
Df Residuals:	18	Prob (F-statistic):	0.000779
R-squared:	0.475	Scale:	0.26386

	Coef.	Std.Err.	t	P>|t|	[0.055	0.945]
const	10.1769	0.1997	50.9693	0.0000	9.8412	10.5125
X	1.6711	0.4143	4.0336	0.0008	0.9746	2.3676

Omnibus:	1.246	Durbin-Watson:	1.798
Prob(Omnibus):	0.536	Jarque-Bera (JB):	0.464
Skew:	0.367	Prob(JB):	0.793
Kurtosis:	3.136	Condition No.:	4

OLS



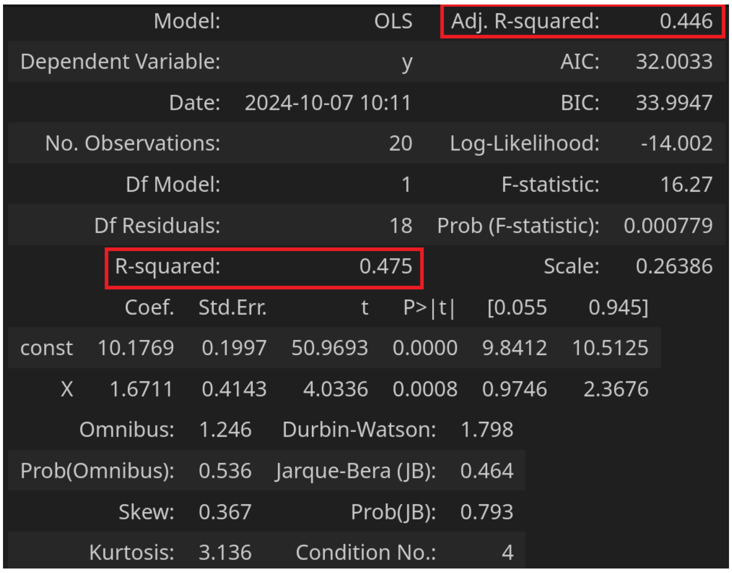
```
                Model:               OLS      Adj. R-squared:     0.446
   Dependent Variable:                 y                 AIC:   32.0033
                 Date:  2024-10-07 10:11                 BIC:   33.9947
     No. Observations:                20      Log-Likelihood:   -14.002
             Df Model:                 1         F-statistic:     16.27
         Df Residuals:                18  Prob (F-statistic):  0.000779
            R-squared:             0.475               Scale:   0.26386
           Coef.  Std.Err.        t   P>|t|   [0.055   0.945]
const    10.1769    0.1997  50.9693  0.0000   9.8412  10.5125
    X     1.6711    0.4143   4.0336  0.0008   0.9746   2.3676
       Omnibus:  1.246    Durbin-Watson:  1.798
Prob(Omnibus):  0.536  Jarque-Bera (JB):  0.464
          Skew:  0.367          Prob(JB):  0.793
      Kurtosis:  3.136     Condition No.:      4
```

OLS refers to Ordinary Least Squares, which means that β is estimated minimizing the loss function:

$$\sum_{i=1}^{n} (\beta^T X_i - Y_i)^2.$$

It can be proved that these estimators correspond with the MLE. Other loss functions can be considered to solve specific issues, such as heteroskedasticity, or to get robust estimators.

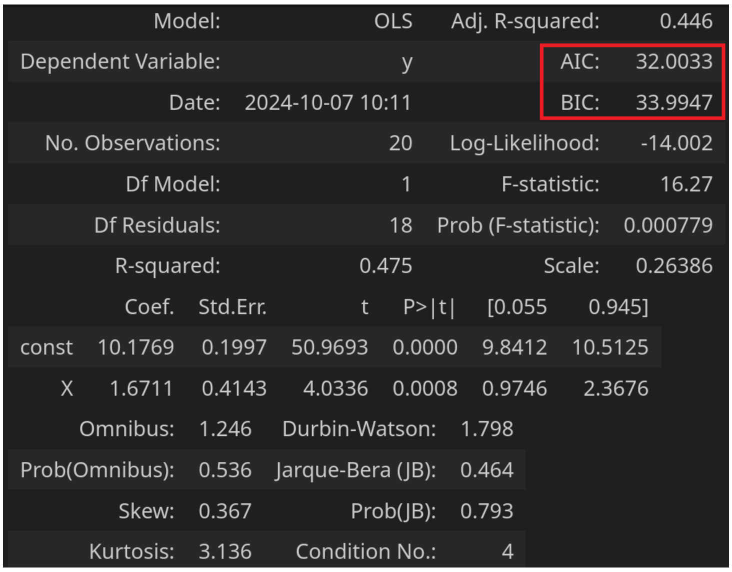
R^2 and adjusted R^2



```
                Model:               OLS      Adj. R-squared:     0.446
   Dependent Variable:                 y                 AIC:   32.0033
                 Date:  2024-10-07 10:11                 BIC:   33.9947
     No. Observations:                20      Log-Likelihood:   -14.002
             Df Model:                 1         F-statistic:     16.27
         Df Residuals:                18  Prob (F-statistic):  0.000779
            R-squared:             0.475               Scale:   0.26386
           Coef.  Std.Err.        t   P>|t|   [0.055   0.945]
const    10.1769    0.1997  50.9693  0.0000   9.8412  10.5125
    X     1.6711    0.4143   4.0336  0.0008   0.9746   2.3676
       Omnibus:  1.246    Durbin-Watson:  1.798
Prob(Omnibus):  0.536  Jarque-Bera (JB):  0.464
          Skew:  0.367          Prob(JB):  0.793
      Kurtosis:  3.136     Condition No.:      4
```

The statistic R^2 measures the proportion of the variance of $\mathbf{Y}$ that is explained by our model. It is between zero and one. The closer it is to one, the better the model can predict Y. However, it can be inflated artificially adding more variables, even if they are not related with Y. The adjusted R^2 takes this into account, penalizing by the number of variables that are in the model.

AIC, and BIC

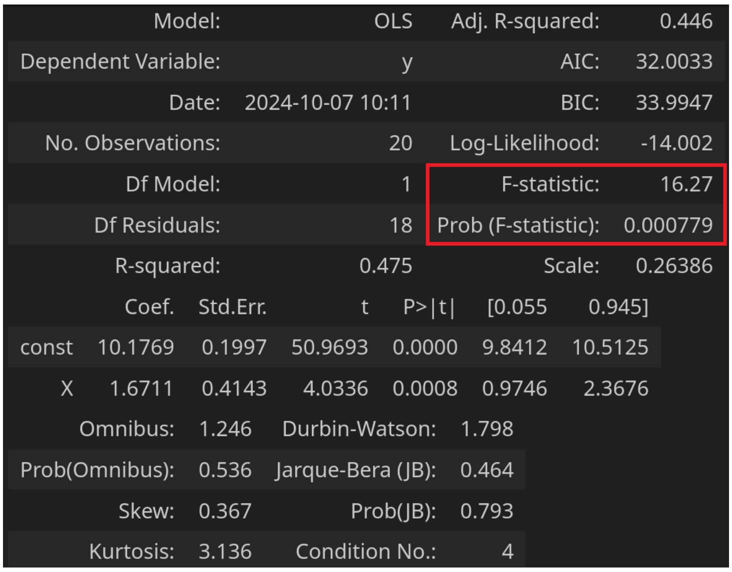


```
              Model:              OLS    Adj. R-squared:     0.446
Dependent Variable:                 y               AIC:   32.0033
               Date:  2024-10-07 10:11              BIC:   33.9947
   No. Observations:               20     Log-Likelihood:  -14.002
           Df Model:                1        F-statistic:    16.27
       Df Residuals:               18  Prob (F-statistic): 0.000779
          R-squared:            0.475              Scale:  0.26386
          Coef.   Std.Err.       t    P>|t|   [0.055    0.945]
const   10.1769   0.1997   50.9693   0.0000   9.8412   10.5125
    X    1.6711   0.4143    4.0336   0.0008   0.9746    2.3676
       Omnibus:  1.246    Durbin-Watson:  1.798
Prob(Omnibus):   0.536  Jarque-Bera (JB):  0.464
          Skew:  0.367          Prob(JB):  0.793
      Kurtosis:  3.136     Condition No.:      4
```

AIC, and BIC state for Akaike Information Criterion and Bayesian Information Criterion, respectively.

Information criteria are used to compare different models, taking into account how intricate a model is, and how well it can explain the observations. According to the principle of parsimony, the simplest model capable to explain our phenomenon should be the preferred one. Thus, we should select the model with the lowest AIC (or BIC) when we are comparing different models.

F-test



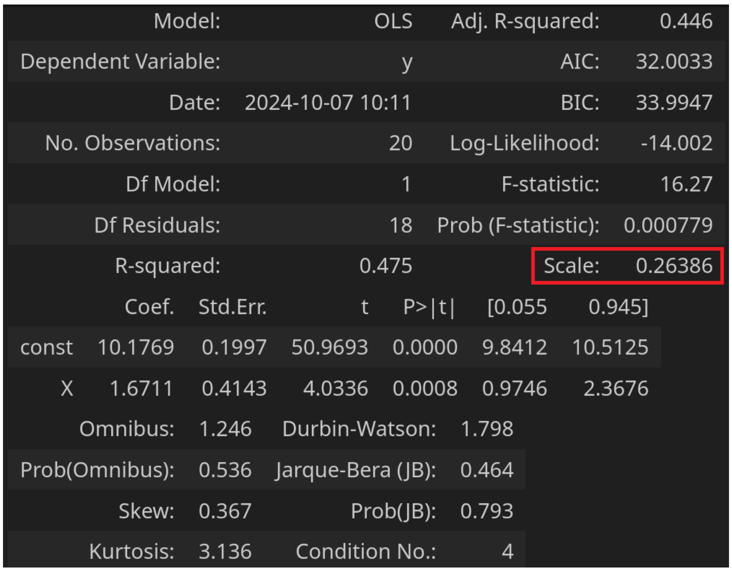
```
              Model:              OLS    Adj. R-squared:     0.446
Dependent Variable:                 y               AIC:   32.0033
               Date:  2024-10-07 10:11              BIC:   33.9947
   No. Observations:               20     Log-Likelihood:  -14.002
           Df Model:                1        F-statistic:    16.27
       Df Residuals:               18  Prob (F-statistic): 0.000779
          R-squared:            0.475              Scale:  0.26386
          Coef.   Std.Err.       t    P>|t|   [0.055    0.945]
const   10.1769   0.1997   50.9693   0.0000   9.8412   10.5125
    X    1.6711   0.4143    4.0336   0.0008   0.9746    2.3676
       Omnibus:  1.246    Durbin-Watson:  1.798
Prob(Omnibus):   0.536  Jarque-Bera (JB):  0.464
          Skew:  0.367          Prob(JB):  0.793
      Kurtosis:  3.136     Condition No.:      4
```

The F statistic is used to test the hypothesis that specific linear combinations of the parameters are equal to zero. By default the linear combinations are

$$\begin{aligned} \beta_1 &= 0 \\ \beta_2 &= 0 \\ &\vdots \\ \beta_p &= 0 \end{aligned}$$

Thus, it tests if all the parameters are equal to zero at the same time.

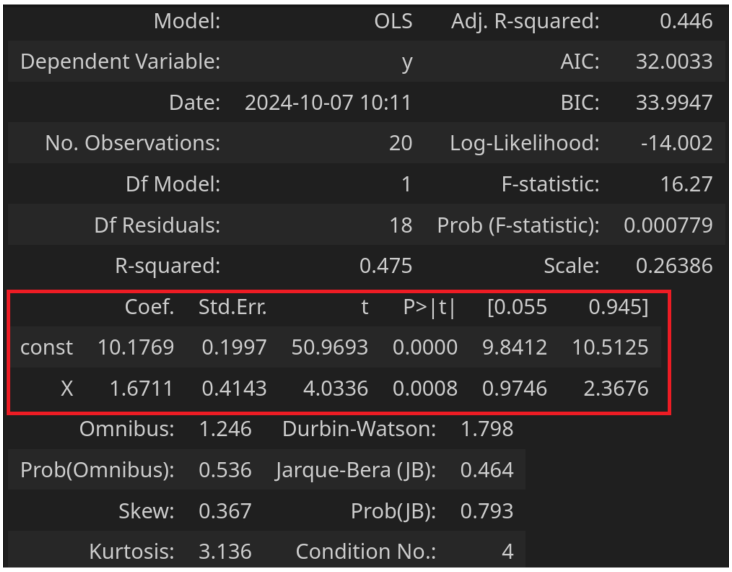
Scale



```
                Model:              OLS   Adj. R-squared:    0.446
   Dependent Variable:                y              AIC:  32.0033
                 Date: 2024-10-07 10:11              BIC:  33.9947
     No. Observations:               20   Log-Likelihood:  -14.002
             Df Model:                1      F-statistic:    16.27
         Df Residuals:               18 Prob (F-statistic): 0.000779
            R-squared:            0.475            Scale:  0.26386
          Coef.  Std.Err.       t    P>|t|   [0.055   0.945]
const   10.1769   0.1997  50.9693  0.0000   9.8412  10.5125
    X    1.6711   0.4143   4.0336  0.0008   0.9746   2.3676
       Omnibus: 1.246   Durbin-Watson: 1.798
 Prob(Omnibus): 0.536  Jarque-Bera (JB): 0.464
          Skew: 0.367        Prob(JB): 0.793
      Kurtosis: 3.136    Condition No.:     4
```

Scale factor for the covariance matrix of the errors. The default value is the MSE.

Estimates of β

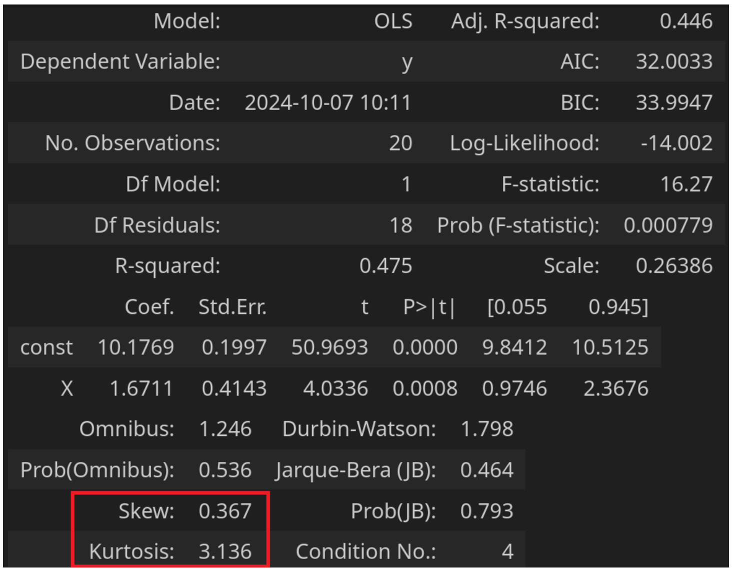


```
                Model:              OLS   Adj. R-squared:    0.446
   Dependent Variable:                y              AIC:  32.0033
                 Date: 2024-10-07 10:11              BIC:  33.9947
     No. Observations:               20   Log-Likelihood:  -14.002
             Df Model:                1      F-statistic:    16.27
         Df Residuals:               18 Prob (F-statistic): 0.000779
            R-squared:            0.475            Scale:  0.26386
          Coef.  Std.Err.       t    P>|t|   [0.055   0.945]
const   10.1769   0.1997  50.9693  0.0000   9.8412  10.5125
    X    1.6711   0.4143   4.0336  0.0008   0.9746   2.3676
       Omnibus: 1.246   Durbin-Watson: 1.798
 Prob(Omnibus): 0.536  Jarque-Bera (JB): 0.464
          Skew: 0.367        Prob(JB): 0.793
      Kurtosis: 3.136    Condition No.:     4
```

This part reports:

- The value of the estimators $\hat{\beta}_1, \ldots, \hat{\beta}_p$.
- The standard error of the estimates. That is, the standard deviation of the estimators.
- The t statistic, used to test the hypothesis $H : \beta_j = 0$ **given that all the other variables are included in the model**.
- The P-value of the hypothesis $H : \beta_j = 0$ **given that all the other variables are included in the model**.
- A confidence interval for each parameter.

Skew and kurtosis



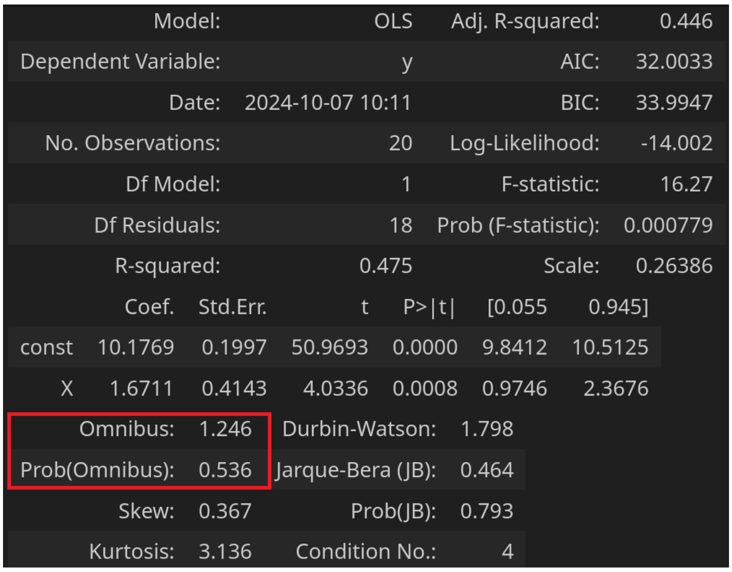
Model:	OLS	Adj. R-squared:	0.446
Dependent Variable:	y	AIC:	32.0033
Date:	2024-10-07 10:11	BIC:	33.9947
No. Observations:	20	Log-Likelihood:	-14.002
Df Model:	1	F-statistic:	16.27
Df Residuals:	18	Prob (F-statistic):	0.000779
R-squared:	0.475	Scale:	0.26386

	Coef.	Std.Err.	t	P>|t|	[0.055	0.945]
const	10.1769	0.1997	50.9693	0.0000	9.8412	10.5125
X	1.6711	0.4143	4.0336	0.0008	0.9746	2.3676

Omnibus:	1.246	Durbin-Watson:	1.798
Prob(Omnibus):	0.536	Jarque-Bera (JB):	0.464
Skew:	0.367	Prob(JB):	0.793
Kurtosis:	3.136	Condition No.:	4

Skewness and kurtosis of the residuals. If the errors do follow a normal distribution the skewness should be zero, and the kurtosis should be 3.

Omnibus test



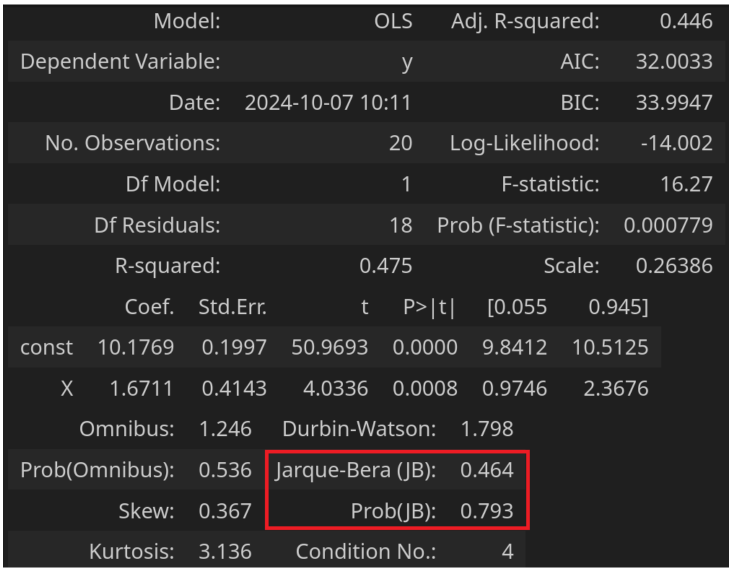
Model:	OLS	Adj. R-squared:	0.446
Dependent Variable:	y	AIC:	32.0033
Date:	2024-10-07 10:11	BIC:	33.9947
No. Observations:	20	Log-Likelihood:	-14.002
Df Model:	1	F-statistic:	16.27
Df Residuals:	18	Prob (F-statistic):	0.000779
R-squared:	0.475	Scale:	0.26386

	Coef.	Std.Err.	t	P>|t|	[0.055	0.945]
const	10.1769	0.1997	50.9693	0.0000	9.8412	10.5125
X	1.6711	0.4143	4.0336	0.0008	0.9746	2.3676

Omnibus:	1.246	Durbin-Watson:	1.798
Prob(Omnibus):	0.536	Jarque-Bera (JB):	0.464
Skew:	0.367	Prob(JB):	0.793
Kurtosis:	3.136	Condition No.:	4

The omnibus statistic is used to test if the errors follow a normal distribution based on the skewness and the kurtosis of the residuals. A value close to zero indicates a normal distribution for the errors.

Jarque-Bera test



Model:	OLS	Adj. R-squared:	0.446
Dependent Variable:	y	AIC:	32.0033
Date:	2024-10-07 10:11	BIC:	33.9947
No. Observations:	20	Log-Likelihood:	-14.002
Df Model:	1	F-statistic:	16.27
Df Residuals:	18	Prob (F-statistic):	0.000779
R-squared:	0.475	Scale:	0.26386

	Coef.	Std.Err.	t	P>|t|	[0.055	0.945]
const	10.1769	0.1997	50.9693	0.0000	9.8412	10.5125
X	1.6711	0.4143	4.0336	0.0008	0.9746	2.3676

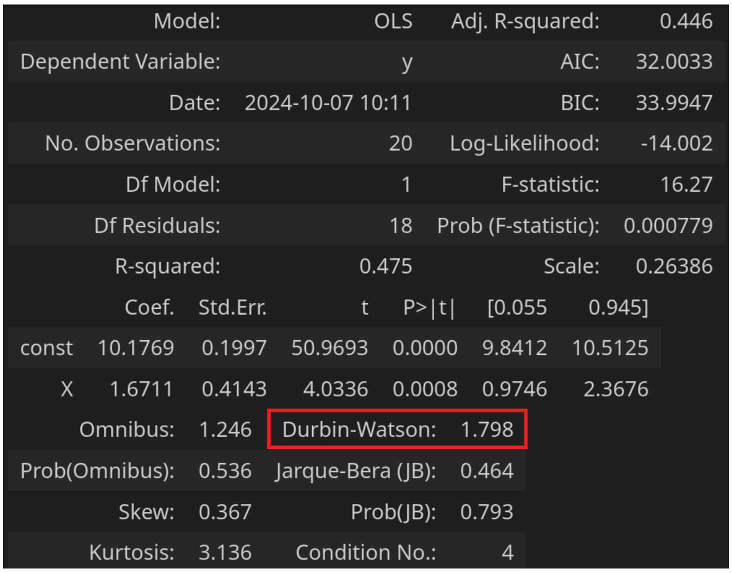
Omnibus:	1.246	Durbin-Watson:	1.798
Prob(Omnibus):	0.536	Jarque-Bera (JB):	0.464
Skew:	0.367	Prob(JB):	0.793
Kurtosis:	3.136	Condition No.:	4

The Jarque-Bera statistic is used to test if the errors follow a normal distribution. It is defined as

$$\frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right),$$

where S is the skewness, and K is the kurtosis. If the errors are normal, the statistic should be zero.

Durbin-Watson statistic



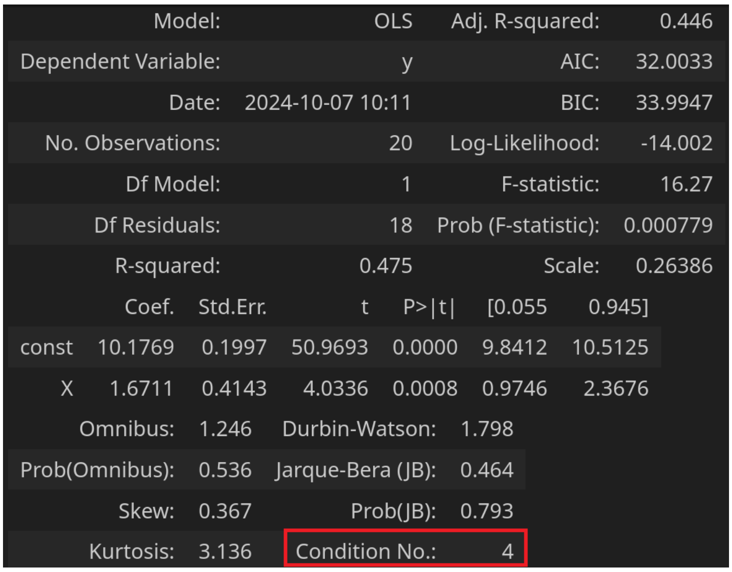
Model:	OLS	Adj. R-squared:	0.446
Dependent Variable:	y	AIC:	32.0033
Date:	2024-10-07 10:11	BIC:	33.9947
No. Observations:	20	Log-Likelihood:	-14.002
Df Model:	1	F-statistic:	16.27
Df Residuals:	18	Prob (F-statistic):	0.000779
R-squared:	0.475	Scale:	0.26386

	Coef.	Std.Err.	t	P>|t|	[0.055	0.945]
const	10.1769	0.1997	50.9693	0.0000	9.8412	10.5125
X	1.6711	0.4143	4.0336	0.0008	0.9746	2.3676

Omnibus:	1.246	Durbin-Watson:	1.798
Prob(Omnibus):	0.536	Jarque-Bera (JB):	0.464
Skew:	0.367	Prob(JB):	0.793
Kurtosis:	3.136	Condition No.:	4

The Durbin-Watson statistic helps to determine if the residuals are independent. It should be around two. Otherwise, it might be an indicative of dependence between the errors.

Condition number



Model:	OLS	Adj. R-squared:	0.446
Dependent Variable:	y	AIC:	32.0033
Date:	2024-10-07 10:11	BIC:	33.9947
No. Observations:	20	Log-Likelihood:	-14.002
Df Model:	1	F-statistic:	16.27
Df Residuals:	18	Prob (F-statistic):	0.000779
R-squared:	0.475	Scale:	0.26386

	Coef.	Std.Err.	t	P>|t|	[0.055	0.945]
const	10.1769	0.1997	50.9693	0.0000	9.8412	10.5125
X	1.6711	0.4143	4.0336	0.0008	0.9746	2.3676

Omnibus:	1.246	Durbin-Watson:	1.798
Prob(Omnibus):	0.536	Jarque-Bera (JB):	0.464
Skew:	0.367	Prob(JB):	0.793
Kurtosis:	3.136	Condition No.:	4

Condition number of the matrix **X**. A large value is interpreted as **X** being numerically unstable, compromising the results. A large value might also indicate a problem of collinearity. A variance decomposition analysis should be conducted to determine if this is the case.

A

MAD, IQR, and Boxplot

A.1 Median absolute deviation (MAD)

Assume that $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then,

$$\mathbb{P}(|X-\mu| \leq \mathrm{MAD}) \xrightarrow[n\to\infty]{} 1/2.$$

On the other hand,

$$\mathbb{P}(|X-\mu| \leq \mathrm{MAD}) = \mathbb{P}\left(|Z| \leq \frac{\mathrm{MAD}}{\sigma}\right) = \mathbb{P}\left(-\frac{\mathrm{MAD}}{\sigma} \leq Z \leq \frac{\mathrm{MAD}}{\sigma}\right),$$

where $Z = (X-\mu)/\sigma$. That is,

$$\begin{aligned}\mathbb{P}(|X-\mu| \leq \mathrm{MAD}) &= \Phi\left(\frac{\mathrm{MAD}}{\sigma}\right) - \Phi\left(-\frac{\mathrm{MAD}}{\sigma}\right) \\ &= \Phi\left(\frac{\mathrm{MAD}}{\sigma}\right) - \left[1 - \Phi\left(\frac{\mathrm{MAD}}{\sigma}\right)\right] \\ &= 2\Phi\left(\frac{\mathrm{MAD}}{\sigma}\right) - 1.\end{aligned}$$

Therefore we conclude that,

$$\begin{aligned} 2\Phi\left(\frac{\mathrm{MAD}}{\sigma}\right) - 1 &\xrightarrow[n\to\infty]{} \frac{1}{2} \\ \Rightarrow \frac{\mathrm{MAD}}{\sigma} &\xrightarrow[n\to\infty]{} \Phi^{-1}(3/4) \\ \Rightarrow \frac{\mathrm{MAD}}{\Phi^{-1}(3/4)} &\xrightarrow[n\to\infty]{} \sigma. \end{aligned}$$

Thus,

$$\hat{\sigma} = \frac{\mathrm{MAD}}{\Phi^{-1}(3/4)}$$

is a consistent estimator for σ.

A.2 Interquartile range (IQR)

Let be $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\Phi(x) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq x\right)$$

$$= \mathbb{P}(X \leq \mu + \sigma x)$$

$$= F_X(\mu + \sigma x)$$

Making $x = \Phi^{-1}(p)$,

$$p = F_X(\mu + \sigma \Phi^{-1}(p)),$$

and

$$q_X(p) = \mu + \sigma \Phi^{-1}(p).$$

Thus,

$$q_1 = q_X(1/4) = \mu + \sigma \Phi^{-1}(1/4),$$

but $\Phi^{-1}(1/4) = -\Phi^{-1}(3/4)$. Hence,

$$q_1 = \mu - \sigma \Phi^{-1}(3/4), \quad \text{and} \quad q_3 = \mu + \sigma \Phi^{-1}(3/4).$$

Therefore,

$$\text{IQR} = q_3 - q_1 = 2\sigma \Phi^{-1}(3/4) \Rightarrow \sigma = \frac{\text{IQR}}{2\Phi^{-1}(3/4)}.$$

Finally, we conclude that

$$\hat{\sigma} = \frac{\text{IQR}}{2\Phi^{-1}(3/4)}$$

is a consistent estimator for σ.

A.3 Boxplot

Let be $\alpha \in (0, 1/2]$. Then,

$$1 - \alpha/2 \geq 1 - 1/4$$

$$\Rightarrow \Phi^{-1}(1-\alpha/2) \geq \Phi^{-1}(3/4)$$

$$\Rightarrow \frac{1}{2}\left(\frac{\Phi^{-1}(1-\alpha/2)}{\Phi^{-1}(3/4)} - 1\right) \geq 0$$

That is,

$$\delta = \frac{1}{2}\left(\frac{\Phi^{-1}(1-\alpha/2)}{\Phi^{-1}(3/4)} - 1\right) \geq 0$$

On the other hand, let be $X \sim \mathcal{N}(\mu, \sigma^2)$. It follows that

$$\begin{aligned}
Q_1 - \delta\,\mathrm{IQR} &= \mu - \sigma\Phi^{-1}(3/4) - 2\delta\sigma\Phi^{-1}(3/4) \\
&= \mu - \sigma\Phi^{-1}(3/4)\,(1 + 2\delta) \\
&= \mu - \sigma\Phi^{-1}(3/4)\left(\frac{\Phi^{-1}(1-\alpha/2)}{\Phi^{-1}(3/4)}\right) \\
&= \mu - \sigma\Phi^{-1}(1-\alpha/2).
\end{aligned}$$

Similarly,

$$Q_3 + \delta\,\mathrm{IQR} = \mu + \sigma\Phi^{-1}(1-\alpha/2).$$

Hence,

$$\begin{aligned}
&\mathbb{P}\Big(Q_1 - \delta\,\mathrm{IQR} \le X \le Q_3 + \delta\,\mathrm{IQR}\Big) \\
=&\mathbb{P}\Big(\mu - \sigma\Phi^{-1}(1-\alpha/2) \le X \le \mu + \sigma\Phi^{-1}(1-\alpha/2)\Big) \\
=&1 - \alpha
\end{aligned}$$

In particular, $\delta = 1.5 \Leftrightarrow \alpha \approx 0.007$.

B
Measuring robustness

B.1 Influence function

To measure robustness, statisticians have developed a tool called the influence function (IF). The IF at a point x of a functional (a statistic) T for a probability distribution F is defined as the Gateaux derivative:

$$\mathrm{IF}(x; T, F) := \lim_{t \searrow 0} \frac{T(t\delta_x + (1-t)F) - T(F)}{t}$$

To understand the IF, replace F by the empirical distribution F_{n-1}, and take $t = 1/n$, so

$$\mathrm{IF}(x; T, F_{n-1}) = \lim_{n\to\infty} n \left[T\left(\frac{1}{n}\delta_x + \left(1 - \frac{1}{n}\right) F_{n-1} \right) - T\left(F_{n-1}\right) \right]$$

$$= \lim_{n\to\infty} n \Big[T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1}) \Big]$$

That is, the IF measures approximately n times the change of T caused by an additional observation in x when T is applied to a large sample of size $n-1$.

B.1.1 Sensitivity curve

The expression in the previous limit is called the sensitivity curve (SC) of T at the point x.

$$\mathrm{SC}_n(x) = n\Big[T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})\Big]$$

Example

Let be T the average, so

$$\mathrm{SC}_n(x) = n\Big[T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})\Big]$$

$$= n \left[\frac{1}{n} \left(\sum_{i=1}^{n-1} x_i + x \right) - \frac{1}{n-1} \sum_{i=1}^{n-1} x_i \right]$$

$$= x + \sum_{i=1}^{n-1} x_i - \frac{n}{n-1} \sum_{i=1}^{n-1} x_i$$

$$\begin{aligned}
\mathrm{SC}_n(x) &= x + \sum_{i=1}^{n-1} x_i - \frac{n}{n-1} \sum_{i=1}^{n-1} x_i \\
&= x - \frac{1}{n-1} \sum_{i=1}^{n-1} x_i \\
&= x - \bar{x}_{n-1}.
\end{aligned}$$

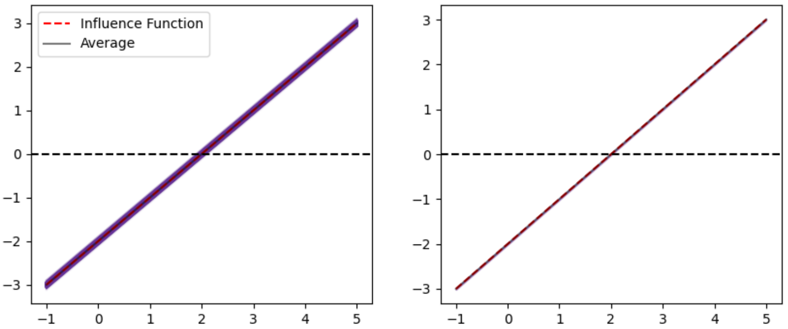
By law of large numbers,

$$SC_n(x) \xrightarrow[n\to\infty]{a.s.} x - \mu$$

B.1.2 Influence function of some statistics

Average

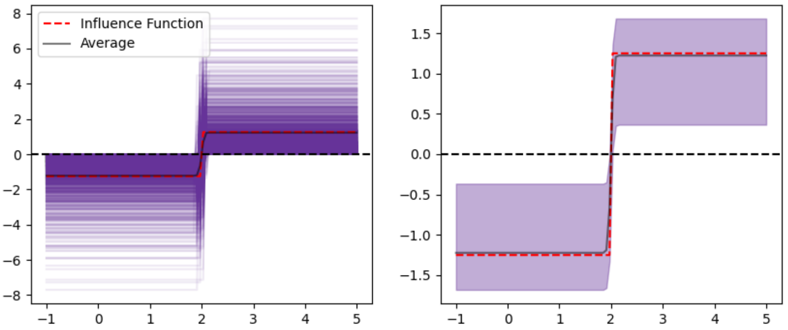
$$\mathrm{IF}(x; T, F) = x - \mu$$



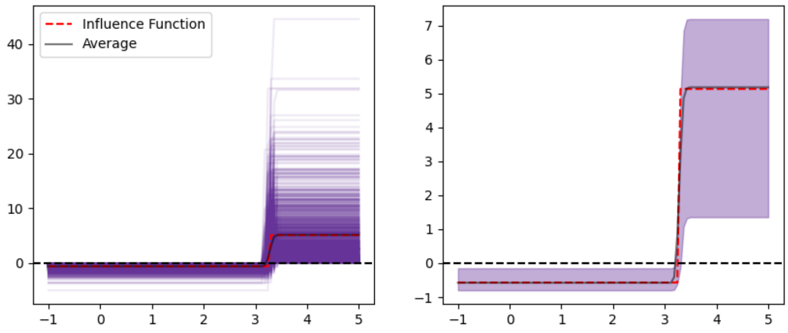
Quantile of probability p, Q_p

$$\mathrm{IF}(x; T, F) = \frac{p - \mathbb{1}\,(x \leq Q_p)}{dF(Q_p)}$$

$p = 0.5$

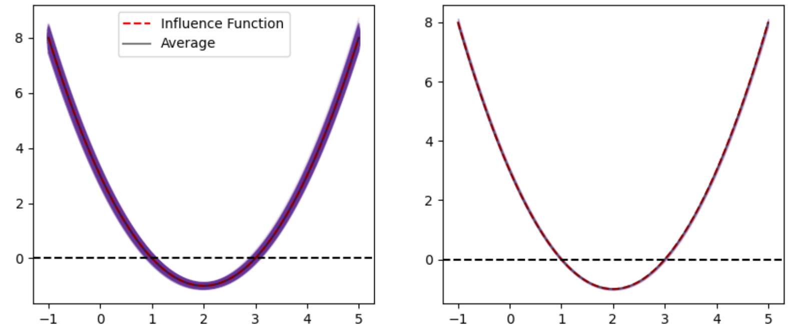


$p = 0.9$



Variance

$$\mathrm{IF}(x; T, F) = (x - \mu)^2 - \sigma^2$$



B.1.3 Robustness measures

Gross-error sensitivity

The gross-error sensitivity measures the worst influence which a small amount of contamination of fixed size can have on the value of the estimator.

$$\gamma^* = \sup_x |\mathrm{IF}(x; T, F)|$$

It is a desirable feature that γ^* be finite, in which case we say that T is B-robust at F.

Local-shift sensitivity

The local-shift sensitivity has to do with small fluctuations in the observations.

$$\lambda^* = \sup_{x \neq y} \frac{|\mathrm{IF}(y; T, F) - \mathrm{IF}(x; T, F)|}{|y - x|}$$

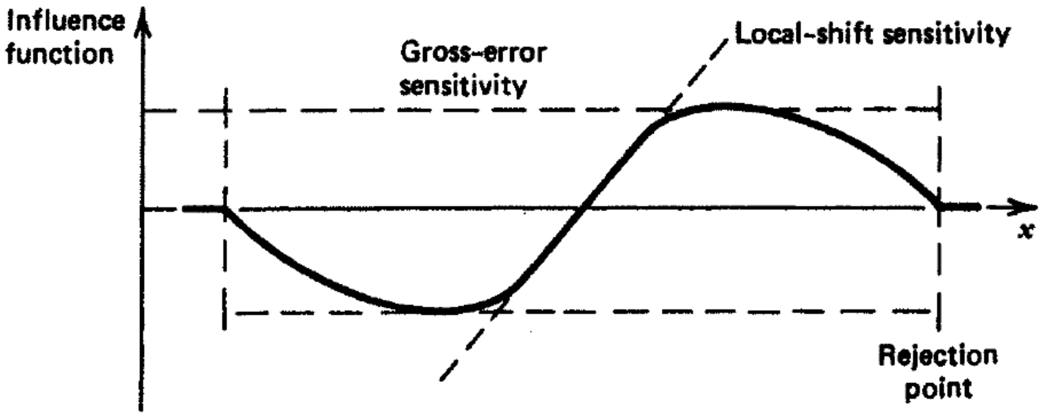
It measures the worst effect of "wiggling" an observation. Note that even an infinite value of λ^* may refer to a very limited actual change, because of the standardization.

Rejection point

It is an old robustness idea to reject extreme outliers entirely. If the IF is identically zero in some region, then contamination in those points does not have any influence at all.

$$\rho^* = \inf\{r > 0; \mathrm{IF}(x; T, F) = 0 \text{ when } |x| > r\}$$

Therefore, it is a desirable feature that ρ^* is finite.



Breakdown point

The breakdown point, ε^*, is the largest amount of contamination (proportion of atypical points) such that the estimator still gives some relevant information.

$$\varepsilon^* = \sup_t \left\{ \sup_x |T(t\delta_x + (1-t)F) - T(F)| < \infty \right\}$$

In some cases, the breakdown point is more important than the efficiency of any corresponding estimator.

Finite-sample breakdown point

The finite-sample breakdown point of $T_n(x_1, \ldots, x_n)$ is the largest proportion of data points that can be arbitrarily replaced by outliers without $T_n(x_1, \ldots, x_n)$ leaving a bounded set.

$$\varepsilon_n^*(T_n; x_1, \ldots, x_n) = \frac{1}{n} \max_m \left\{ \max_{i_1, \ldots, i_m} \sup_{y_1, \ldots, y_m} |T_n(z_1, \ldots, z_n)| \in \Theta \right\},$$

where Θ is bounded, and the sample $(z_1, \ldots, z_n)$ is obtained by replacing m data points $x_{i_1}, \ldots, x_{i_m}$ by arbitrary values $y_1, \ldots, y_m$.

Statistic	γ^*	λ^*	ρ^*	ε^*
Arithmetic mean	∞	1	∞	0
α-trimmed mean	$< \infty$	$< \infty$	∞	α
Median	$\sqrt{\pi/2}$	∞	∞	0.5
Quantile q	$< \infty$	$< \infty$	∞	$\min\{q, 1-q\}$

C

Proof that the density of a normal distribution integrates one

A random variable X is said to have or follow a normal distribution with parameters μ and σ^2 if its density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathbb{1}_{x\in\mathbb{R}}.$$

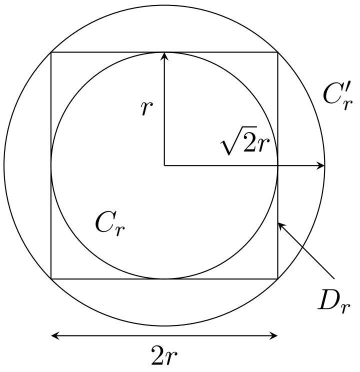
Since this is a non-negative function, to see that it is effectively a density of probability we have to prove that

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Consider the change of variable $z = (x-\mu)/\sqrt{2}\sigma$, then

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2}\sigma\sqrt{2}dz = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-z^2}dz$$

One way to calculate this last integral is as follows. Let be C_r the disc with center in the origin and radius r, and C'_r the disc with the same center and radius $\sqrt{2}r$. Let D_r be the square with center in the origin and side $2r$.



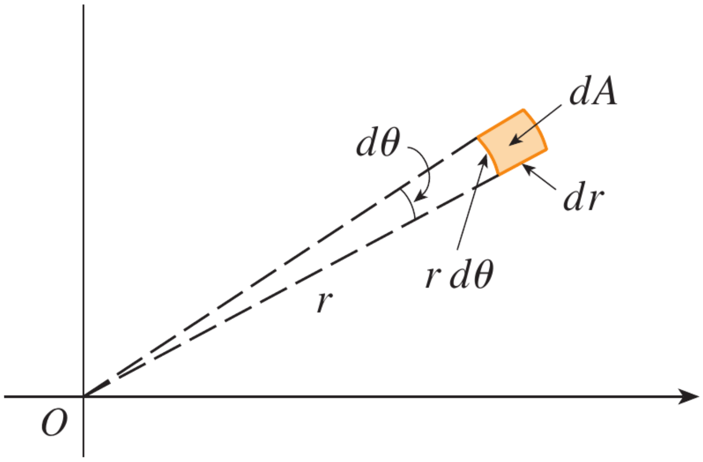
Given that the common integrand of the following integrals is non-negative, we have

$$\int\int_{C_r} e^{-(x^2+y^2)}dxdy \leq \int\int_{D_r} e^{-(x^2+y^2)}dxdy \leq \int\int_{C'_r} e^{-(x^2+y^2)}dxdy,$$

also

$$\int\int_{D_r} e^{-(x^2+y^2)}dxdy = \int_{-r}^{r} e^{-x^2}dx \int_{-r}^{r} e^{-y^2}dy = \left(\int_{-r}^{r} e^{-x^2}dx\right)^2.$$

Consider now the first integral of the inequalities. Changing to polar coordinates ρ, θ through the transformation $x = \rho\cos\theta$ and $y = \rho\sin\theta$, we have



$$\int\int_{C_r} e^{-(x^2+y^2)}dxdy = \int_0^{2\pi} d\theta \int_0^r e^{-\rho^2}\rho d\rho$$

$$= 2\pi \left[-\frac{1}{2}e^{-\rho^2}\right]_0^r$$

$$= 2\pi \left[\frac{1}{2}\left(1-e^{-r^2}\right)\right]$$

$$= \pi\left(1-e^{-r^2}\right).$$

Analogously, changing r by $\sqrt{2}r$, we have

$$\int\int_{C_r'} e^{-(x^2+y^2)}dxdy = \pi\left(1-e^{-2r^2}\right).$$

Replacing these quantities in the previous inequalities,

$$\pi\left(1-e^{-r^2}\right) \leq \left(\int_{-r}^{r} e^{-x^2}dx\right)^2 \leq \pi\left(1-e^{-2r^2}\right),$$

letting $r \to \infty$

$$\pi \leq \left(\int_{-\infty}^{\infty} e^{-x^2}dx\right)^2 \leq \pi,$$

thus

$$\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}.$$

Finally, we have

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-z^2}dz = 1.$$

$\mathcal{D}$ Sample distributions

If $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then,

1. $$\bar{X}_n \sim \mathcal{N}\left(\mu, {}^{\sigma^2}\!/_{n}\right) \Rightarrow \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

2. Let $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, $$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

3. $\bar{X}_n \perp\!\!\!\perp s^2$, so $$\sqrt{n}\frac{\bar{X}_n - \mu}{s} \sim t_{n-1}$$

D.1 Confidence interval for μ

If $X \sim t_\nu$, then
$$\mu + \sigma X$$
is said to follow a location-scale-t distribution with ν degrees of freedom, parameter of location μ, and parameter of scale σ, which is denoted as

$$\mu + \sigma X \sim t_\nu(\mu, \sigma^2)$$

Thus, from expression (3),
$$\mu \sim t_{n-1}\left(\bar{X}_n, s^2/n\right),$$
and the interval
$$\left(t_{n-1}\left(\bar{X}_n, s^2/n\right)_{\alpha/2} \quad ; \quad t_{n-1}\left(\bar{X}_n, s^2/n\right)_{1-\alpha/2}\right)$$
corresponds with a CI of probability $(1-\alpha) \times 100\%$ for μ.

D.2 Confidence interval for σ^2

From expression (2):

$$\mathbb{P}\left(\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-n-1,\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} \leq \frac{\sigma^2}{(n-1)s^2} \leq \frac{1}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha$$

That is,

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \quad ; \quad \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right)$$

is a confidence interval of probability $(1-\alpha) \times 100\%$ for σ^2.

D.3 Confidence interval for σ

It is typically more useful to work with the standard deviation σ rather than the variance σ^2. A confidence interval of probability $(1-\alpha) \times 100\%$ for σ is given by

$$\left(\frac{s\sqrt{n-1}}{\sqrt{\chi^2_{n-1,1-\alpha/2}}} \quad ; \quad \frac{s\sqrt{n-1}}{\sqrt{\chi^2_{n-1,\alpha/2}}}\right)$$

E

Maximum likelihood theory

E.1 Fisher information

E.1.1 Expected Fisher information per sample unit

Assume that X is a continuous r.v. with density function f (if X is a discrete random variable, f would be its mass function). We define the expected Fisher information per sample unit as

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 \middle| \theta\right]$$

Under very general conditions, known as regularity conditions, it can be proved that

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) \middle| \theta\right]$$

E.1.2 Observed Fisher information of the sample

We define the observed Fisher information of the sample $\mathbf{X}$ as

$$\begin{aligned} I &= -\frac{\partial^2}{\partial\theta^2}\ell(\theta)\Big|_{\theta=\hat{\theta}_{\text{MLE}}} \\ &= -\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log f(X_i;\theta)\Big|_{\theta=\hat{\theta}_{\text{MLE}}} \end{aligned}$$

E.2 Asymptotic distribution of the MLE

Under regularity conditions, it can be proved that $\hat{\theta}_{\text{MLE}}$ is a consistent estimator of θ. Thus, if we denote by θ_0 the real value of θ, then

$$\hat{\theta}_{\text{MLE}} \xrightarrow[n\to\infty]{P} \theta_0.$$

Hence, by law of large numbers,

$$\frac{1}{n}I \xrightarrow[n\to\infty]{P} \mathcal{I}(\theta_0).$$

On the other hand, it can be proved that

$$\sqrt{n}\left(\hat{\theta}_{\text{MLE}} - \theta_0\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \mathcal{I}^{-1}(\theta_0)\right).$$

Using Slutsky's theorem, we conclude that

$$\left(\hat{\theta}_{\text{MLE}} - \theta_0\right)\sqrt{I} \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, 1\right).$$

E.3 Wilk's theorem

Remember that $r(\theta) = \ell(\theta) - \ell(\hat{\theta}_{\text{MLE}})$. Thus,

$$r(\hat{\theta}_{\text{MLE}}) = 0,$$

$r'(\theta) = sc_n(\theta)$, so

$$r'(\hat{\theta}_{\text{MLE}}) = 0,$$

and $r''(\theta) = \frac{\partial^2}{\partial\theta^2}\ell(\theta)$, so

$$r''(\hat{\theta}_{\text{MLE}}) = -I$$

The Taylor's series of r around $\hat{\theta}_{\text{MLE}}$ would be

$$r(\theta) = \cancelto{0}{r(\hat{\theta}_{\text{MLE}})} + \cancelto{0}{r'(\hat{\theta}_{\text{MLE}})}(\theta - \hat{\theta}_{\text{MLE}}) + \frac{1}{2}\cancelto{-I}{r''(\hat{\theta}_{\text{MLE}})}(\theta - \hat{\theta}_{\text{MLE}})^2 + \cdots .$$

Hence,

$$r(\theta_0) \approx -\frac{1}{2}I(\theta_0 - \hat{\theta}_{\text{MLE}})^2$$

$$\Rightarrow -2r(\theta_0) \approx (\hat{\theta}_{\text{MLE}} - \theta_0)^2 I.$$

But,

$$\left(\hat{\theta}_{\text{MLE}} - \theta_0\right)\sqrt{I} \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, 1\right).$$

Therefore,

$$-2r(\theta_0) \mathrel{\dot\sim} \chi^2_d,$$

where $d = \dim(\theta)$.

Properly said, we only proved the case where $d = 1$, but the general case is analogous, using the Taylor series of a multivariate function.

E.4 Confidence region

Thus,

$$\mathbb{P}\left(-2r(\theta_0) \leq \chi^2_{d,1-\alpha}\right) \approx 1 - \alpha$$

$$\mathbb{P}\left(r(\theta_0) \geq -\frac{1}{2}\chi^2_{d,1-\alpha}\right) \approx 1 - \alpha$$

$$\mathbb{P}\left(R(\theta_0) \geq e^{-\frac{1}{2}\chi^2_{d,1-\alpha}}\right) \approx 1 - \alpha$$

That is, there is a probability of approx. $(1 - \alpha) \times 100\%$ that the real value of θ, θ_0, is in the likelihood region of level

$$c = e^{-\frac{1}{2}\chi^2_{d,1-\alpha}}.$$

In other words, a likelihood region of level

$$c = e^{-\frac{1}{2}\chi^2_{d,1-\alpha}}$$

corresponds, approx., with a confidence region of $(1 - \alpha) \times 100\%$ of probability.

$\mathcal{F}$

Mathematical statistics

F.1 Empirical cumulative distribution function (ECDF)

F.1.1 Glivenko-Cantelli theorem

Let $X_1, \dots, X_n$ be i.i.d. random variables. We define the empirical cumulative distribution function (ECDF) as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \le x}.$$

The Glivenko-Cantelli theorem states that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{\text{a.s.}} 0$$

This theorem is so relevant that it is sometimes referred as the Fundamental Theorem of Statistics.

F.1.2 Dvoretsky-Kiefer-Wolfowitz inequality

In 1956 Dvoretsky, Kiefer, and Wolfowitz proved that

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \le Ce^{-2n\varepsilon^2}, \quad \text{for every } \varepsilon > 0.$$

In 1990, Massart proved that $C = 2$. Thus,

$$\mathbb{P}\Big(|F_n(x) - F(x)| \le \varepsilon\Big) \ge 1 - 2e^{-2n\varepsilon^2}$$

F.1.3 Confidence intervals for the ECDF

Let be

$$\varepsilon = \sqrt{\frac{1}{2n}\log\left({}^{2}\!/_{\alpha}\right)}.$$

Then,

$$\mathbb{P}\Big(|F_n(x) - F(x)| \le \varepsilon\Big) \ge 1 - \alpha.$$

That is,

$$(F_n(x) - \varepsilon \quad ; \quad F_n(x) + \varepsilon)$$

is approximately a confidence interval of probability $(1-\alpha) \times 100\%$ for $F(x)$.

F.1.4 Asymptotic distribution of the ECDF

Let $X_1, \dots, X_n$ be i.i.d. r.v. with cumulative distribution F. Note that,

$$\mathbb{1}_{X_i \leq x} = \begin{cases} 1 & \text{if } X_i \leq x, \\ 0 & \text{if } X_i > x. \end{cases}$$

That is, $\mathbb{1}_{X_i \leq x}$ is a Bernoulli r.v. with parameter $p = \mathbb{P}(X_i \leq x) = F(x)$.

$$\mathbb{1}_{X_i \leq x} \sim \text{Bernoulli}\Big(F(x)\Big)$$

Hence, by the De Moivre - Laplace limit theorem:

$$\sqrt{n}\Big(F_n(x) - F(x)\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0, F(x)(1-F(x))\Big)$$

F.2 Slutsky's theorem

Let X_n and Y_n be sequences of random variables, such that X_n converges in distribution to a random variable X and Y_n converges in probability to a constant c.

That is, if $X_n \xrightarrow[n\to\infty]{d} X$ and $Y_n \xrightarrow[n\to\infty]{P} c$. Then,

- $X_n + Y_n \xrightarrow[n\to\infty]{d} X + c$,
- $X_n Y_n \xrightarrow[n\to\infty]{d} Xc$,
- $X_n / Y_n \xrightarrow[n\to\infty]{d} X/c$, always that $c \neq 0$.

F.3 Delta method

Let X_n be a sequence of random variables such that

$$\sqrt{n}(X_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2),$$

then

$$\sqrt{n}\Big(g(X_n) - g(\theta)\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \sigma^2[g'(\theta)]^2\right),$$

for any function g such that $g'(\theta)$ exists and is non-zero valued.

F.4 Asymptotic distribution for the quantiles

Remember that,

$$\sqrt{n}\Big(F_n(x) - F(x)\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0, F(x)(1-F(x))\Big).$$

On the other hand, let $g(t) = F^{-1}(t)$, and note that, $(g(t)) = t$. By the chain rule,

$$f(g(t))\frac{dg}{dt} = 1 \quad \Rightarrow \quad \frac{dg}{dt} = \frac{1}{f(g(t))}.$$

Using Delta method:

$$\sqrt{n}\Big(g(F_n(x)) - g(F(x))\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{F(x)(1-F(x))}{f^2(g(F(x)))}\right)$$

But $g(t) = F^{-1}(t)$, so

$$\sqrt{n}\Big(F^{-1}(F_n(x)) - x\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{F(x)(1-F(x))}{f^2(x)}\right)$$

Now, let be $x = F^{-1}(p) \equiv q_p$, for some fixed $p \in (0,1)$. Hence,

$$\sqrt{n}\Big(F^{-1}(F_n(q_p)) - q_p\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(q_p)}\right).$$

By Glivenko-Cantelli theorem, for n sufficiently large, we can interchange F by F_n and vice-versa,

$$\sqrt{n}\Big(F_n^{-1}(F(q_p)) - q_p\Big) \dot{\sim} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(q_p)}\right).$$

That is,

$$\sqrt{n}\Big(F_n^{-1}(p) - q_p\Big) \dot{\sim} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(q_p)}\right).$$

Denote by $Q_p = F_n^{-1}(p)$, the empirical quantile of probability p.

Then, for n sufficiently large,

$$\sqrt{n}\Big(Q_p - q_p\Big) \dot{\sim}\mathcal{N}\left(0, \frac{p(1-p)}{f^2(q_p)}\right)$$

$$\Rightarrow Q_p \dot{\sim}\mathcal{N}\left(q_p, \frac{p(1-p)}{nf^2(q_p)}\right).$$

F.4.1 Confidence interval for the quantile q_p

A confidence interval of approx. $(1-\alpha) \times 100\%$ of probability for the quantile of probability p, q_p, is given by

$$\left(\mathcal{N}\left(Q_p, \frac{p(1-p)}{nf^2(q_p)}\right)_{\alpha/2} \quad ; \quad \mathcal{N}\left(Q_p, \frac{p(1-p)}{nf^2(q_p)}\right)_{1-\alpha/2}\right)$$