J. A. Tenreiro Machado
Carla M. A. Pinto

Probability and Statistics

— Selected Problems

L&H

J. A. Tenreiro Machado
Carla M. A. Pinto

Probability and Statistics

— Selected Problems

L&H

Authors

J. A. Tenreiro Machado, Carla M. A. Pinto
ISEP-Institute of Engineering, Polytechnic of Porto
Dept. of Electrical Engineering
Rua Dr. Antonio Bernardino de Almeida, 431
4200-072 Porto, Portugal
Email: jtm@isep.ipp.pt

ISBN 978-1-62155-006-8 e-ISBN 978-1-62155-007-5
DOI 10.5890/978-1-6215-007-5
L&H Scientific Publishing, Glen Carbon, USA

Library of Congress Control Number: 2013954046

Printed on acid-free paper

L&H Scientific Publishing, LLC (https://lhscientificpublishing.com/)
P.O. Box 99, Glen Carbon, IL62034, USA

Contents

Chapter 1
Descriptive statistics

1.1 Fundamentals

Definition 1.1. Consider a set of observations. We say that these observations are values from a given random variable (rv). We denote rv by capital letters, $X, Y, \cdots$ We will define random variable in a more precise way in the next chapter.

Remark 1.1. The random variables are of two types: discrete and continuous. If a rv takes values in a finite or an infinitely countable set then the rv is discrete. A continuous rv takes any value in a interval.

Definition 1.2. Let $x_1, x_2, \ldots, x_n$ be n observations of a rv. The (arithmetic) mean of a discrete and a continuous random variable X is, respectively:

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

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

where n is the number of observations of the random variable X, f_i is the absolute frequency of the observation x_i, and c_i is the mean point of the interval $[l_i, L_i]$, that contains x_i.

J. Tenreiro Machado, Carla M. Pinto, Probability and Statistics – Selected Problems
DOI: 10.5890/978-1-62155-007-3_1

Definition 1.3. Let $x_1, x_2, \ldots, x_n$ be n observations of a rv. The geometric mean is given by:

$$m_g = \sqrt[n]{x_1 x_2 \ldots x_n} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

Definition 1.4. Let $x_1, x_2, \ldots, x_n$ be n observations of a rv. The harmonic mean, m_h, is computed using the formula:

$$m_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

Definition 1.5. The median m_e of a rv X is the value of X such that 50% of the observations are to the left of that value.

In the case that X is a discrete rv, the median is computed as follows:

1. Odd number of observations n.
   The meadian is the central value after sorting the data in ascending order.
2. Even number of observations n.
   The median is the arithmetic mean of the two most central values, after sorting the data in ascending order.

Definition 1.6. The mode is the the most frequent value of a set of observations.

In the case that X is a continuous rv, the mode is computed by the formula:

$$m_o = l_i + a_i \frac{f_{i+1}}{f_{i+1} + f_{i-1}}$$

where i is the most frequent class (modal class), given by $[l_i, L_i[$, a_i is the amplitude of the class, and f_i is the absolute frequency of class i.

Definition 1.7. A quartile of order k, Q_k, is the value of the rv preceeded by $\frac{k}{4}n$ of the n observations $x_1, x_2, \ldots, x_n$.

The quartile Q_k is calculated as follows.

1. X is a discrete rv

$$Q_k = \begin{cases} x_{[kn/4]+1}, & \text{if } kn/4 \text{ is non integer} \\ \dfrac{x_{kn/4} + x_{kn/4+1}}{2}, & \text{if } kn/4 \text{ is an integer} \end{cases}$$

where $[a]$ is the characteristic of a.

2. X is a continuous rv

$$Q_k = l_i + a_i \frac{(k/4)n - a_{i-1}}{f_i}$$

where $k = 1,2,3$ and i is the class of cumulative frequency equal or immediatly greater than $\frac{kn}{4}$.

Definition 1.8. The variance of a sample, $x_i,\ i = 1,2,\ldots,n$, of a discrete random variable, is given by the expression:

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

where f_i is the absolute frequency of x_i. Note that the sample variance may also be computed using the formula

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2.$$

The variance of a sample, $x_i,\ i = 1,2,\ldots,n$, of a continuous rv is calculated as:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} f_i(c_i - \bar{x})^2$$

where f_i is the absolute frequency of class i, c_i is the mean value of class i, $\bar{x}$ is the arithmetic mean and n is the number of observations. Note that the sample variance may also be computed using the formula

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} f_i(c_i - \bar{x})^2.$$

Definition 1.9. The standard deviation of a sample, s, is the square root of the variance s^2.

Definition 1.10. Let $x_i, y_i,\ i = 1,2,\ldots n$ be observations of two rv X and Y. The covariance function is a number that measures the common variation of rv X and Y. It is defined as:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

Definition 1.11. The correlation coefficient between rv X and Y is defined by:

$$\rho_{XY} = \frac{cov(X,Y)}{s_x s_y}$$

where s_x, s_y are the standard deviations of samples x_i and y_i, $i = 1,2,\ldots,n$.

Remark 1.2. The properties of the correlation coefficient ρ_{XY} are:

- $-1 \leq \rho_{XY} \leq 1$.
- if $\rho_{XY} = 0$ then X and Y are said to be uncorrelated.
- If $\rho_{XY} < 0$ then X and Y are negatively correlated.
- If $\rho_{XY} > 0$ then X and Y are positively correlated.

Definition 1.12. Let $x_1, x_2, \ldots, x_n$ be n observations of a rv X. The skew may be computed by the following formula:

$$s^3 = \frac{\frac{\sum_i (x_i - \bar{x})^3}{n}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}}$$

The skewness is a measure of the symmetry in the distribution of the observations.

Remark 1.3.

- If $s^3 < 0$, then the distribution has a negative skew.
- If $s^3 > 0$, then the distribution has a positive skew.
- If $s^3 = 0$, then the distribution is symmetrical.
- The more different s^3 is from 0, the greater the skew in the distribution.

Definition 1.13. Let $x_1, x_2, \ldots, x_n$ be n observations of a rv X. The measure of kurtosis is given by:

$$s^4 = \frac{1}{n} \sum_i \left(\frac{x_i - \bar{x}}{\sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}} \right)^4$$

Kurtosis measures the spread of the observations from a normal (Gaussian) distribution (a gaussian distribution is symmetric around the mean, please see Chapter 4, for more information).

Remark 1.4.

- When the distribution is normally distributed (see Chapter 4), its kurtosis equals 3 and it is said to be mesokurtic.
- When the distribution is less spread than the normal distribution, its kurtosis is greater than 3 and it is said to be leptokurtic.
- When the distribution is more spread than the normal distribution, its kurtosis is less than 3 and it is said to be platykurtic.

1.2 Worked Examples

Problem 1.1
The following set of observations came from a statistical variable X

$$1,2,3,1,2,2,4,2,4,3,5,5,4$$

Compute:

a) the geometric mean of the observations.
b) the harmonic mean of the observations.

Resolution

a) The geometric mean is computed as follows.

$$m_g = \sqrt[13]{1\times 2\times 3\times 1\times 2\times 2\times 4\times 2\times 4\times 3\times 5\times 5\times 4} \simeq 2.58522$$

Thus, the geometric mean is $m_g = 2.5852$.

b) The harmonic mean is computed as follows.

$$m_h = \frac{13}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{1}+\frac{1}{2}+\frac{1}{2}+\frac{1}{4}+\frac{1}{2}+\frac{1}{4}+\frac{1}{3}+\frac{1}{5}+\frac{1}{5}+\frac{1}{4}} \simeq 2.23495$$

Thus, the harmonic mean is given by $m_h = 2.2349$.

Problem 1.2
The results of the observation of a continuous statistical variable X are written in the table below, where I_i represents the range of class $i = 1,2,\ldots,6$ and f_i is the corresponding absolute frequency.

a) the arithmetic mean, m_a, of the statistical variable.

I_i	$[0,3[$	$[3,5[$	$[5,6[$	$[6,7[$	$[7,10[$
f_i	30	80	80	60	60

b) the geometric mean, m_g, of the statistical variable.
c) the harmonic mean, m_h, of the statistical variable.
d) the standard deviation, s, of the statistical variable.
e) the cumulative distribution function.
f) the mode, m_o, of the statistical variable.
g) the median, m_e, of the statistical variable.

Resolution

a) In the computation of m_a, we use the mean value of each class. The arithmetic mean is thus given by:

$$\mu = \frac{1}{310}[1.5(30)+4(80)+5.5(80)+6.5(80)+8.5(60)] = 5.50$$

b)

$$m_g = \sqrt[310]{1.5^{30}\cdots 4^{80}\cdots 5.5^{80}\cdots 6.5^{60}+8.5^{60}} = 5.1099$$

Thus, $m_g = 5.11$.

c)

$$m_h = \frac{310}{\frac{30}{1.5}+\frac{80}{4}+\frac{80}{5.5}+\frac{60}{6.5}+\frac{60}{8.5}} = 4.3764$$

Thus $m_h = 4.38$.

d) To compute s, we first compute the variance

$$s^2 = \frac{1}{310}\sum_{i=1}^{310} f_i(c_i-\bar{x})^2$$

After some algebra, we obtain $s^2 = 4.0645$.

$$s = \sqrt{s^2} = \sqrt{4.0645} = 2.0161$$

We obtain $s = 2.02$.

e) The cumulative distribution curve $C(x)$ is obtained by summing up the absolute frequencies f_i of the values of x in all intervals I_j such that $I_j \leq I_i$. We obtain:

$$C(x) = \begin{cases} 30, & 0 \leq x < 3 \\ 110, & 3 \leq x < 5 \\ 190, & 5 \leq x < 6 \\ 250, & 6 \leq x < 7 \\ 310, & 7 \leq x \leq 10 \end{cases}$$

f) We apply the formula for the mode of continuous rv, obtaining:

$$m_o = 5 + 1\frac{60}{60+40} = 5.60$$

g) The median of a continuous statistical variable is computed as follows:

$$m_e = 5 + \frac{1}{80}(\frac{310}{2} - 110) = 5.6525$$

Thus, $m_e = 5.65$.

Problem 1.3
The observation of a bidimensional random variable (X,Y) led to the results in the following table.

Observation	1	2	3	4	5
x_i	0.1	1.1	2.0	0.5	1.3
y_i	−1.1	2.9	4.8	0.5	3.0

Find:

a) the mean value of X, $\bar{x}$.
b) the variance of X, s^2.
c) correlation coefficient ρ_{XY}.

Resolution

a) $\bar{x} = \frac{1}{5}(0.1+1.1+2.0+0.5+1.3) = 1.0$.

b) $s_x^2 = \frac{1}{5}\sum_{i=1}^{5} f_i x_i^2 - [\bar{x}]^2 = 1.432 - 1^2 = 0.432$. Note that:

$$\frac{1}{5}\sum_{i=1}^{5} f_i x_i^2 = \frac{1}{5}\left(0.1^2 + 1.1^2 + 2.0^2 + 0.5^2 + 1.3^2\right) \simeq 1.432$$

c) $\rho_{XY} = \dfrac{\overline{(XY)} - \bar{x}\bar{y}}{\sqrt{s_x s_y}} \simeq 0.987.$

Note that

$$\overline{(XY)} = \frac{1}{5}((0.1)\cdot(-1.1) + (1.1)\cdot(2.9) + (2)\cdot(4.8) + (0.5)\cdot(0.5) + (1.3)\cdot(3.0)) = 3.366$$

and

$$s_y^2 = \frac{1}{5}\sum_{i=1}^{5} f_i y_i^2 - (\bar{y})^2 = 4.3016$$

where $\bar{y} = \frac{1}{5}(-1.1+2.9+4.8+0.5+3.0) = 2.02$ and $\frac{1}{5}\sum_{i=1}^{5} f_i y_i^2 = 8.382$.

1.3 Proposed Exercises

Exercise 1.1
The observation of a statistical variable X led to the following set of observations

$$7,2,3,1,4,2,6,6,5,3,5,4$$

Sketch the graph of the general mean formula

$$m(q) = \Big[\frac{1}{N}\sum_{i=1}^{N} x_i^q\Big]^{1/q}$$

for the values $q = -\infty, -1, 0, +1, +\infty$.

Exercise 1.2
From a given experience were drawn the following results: 0, 1, a, 4, a, a, 3.

a) Knowing that the mean is $\mu = a$, determine the value of the constant $a \in \mathbb{R}$. Choose the right option.

A) $a = 0$.
B) $a = 1$.
C) $a = 3$.
D) $a = 2$.

b) Compute the standard deviation s.

A) $s = 1.195$.
B) $s = 2.041$.
C) $s = 1.429$.
D) None of the above.

c) Determine the Fisher skewness coefficient, γ_1.

A) $\gamma_1 = 0.5$.
B) $\gamma_1 = 0$.
C) $\gamma_1 = -0.5$.
D) None of the above.

Exercise 1.3
For a given statistical variable, the following results were observed: 11, 13, 17, 10, 12, 17, 18, 11, 13, 16, 15.

a) The arithmetic mean, $\bar{x}$, is:

A) $\bar{x} = 13.392$.
B) $\bar{x} = 13.909$.
C) $\bar{x} = 13.649$.
D) $\bar{x} = 13.000$.

b) The standard deviation s is:

A) $s = 7.174$.
B) $s = 14.165$.
C) $s = 2.678$.
D) $s = 0.000$.

c) The median, m_e, is:

A) $m_e = 13$.
B) $m_e = 13.5$.
C) $m_e = 12$.
D) $m_e = 12.5$.

d) The geometric mean, m_g, is:

A) $m_g = 13.392$.
B) $m_g = 13.909$.
C) $m_g = 13.649$.
D) $m_g = 13.000$.

e) The harmonic mean, m_h, is:

A) $m_h = 13.392$.
B) $m_h = 13.909$.
C) $m_h = 13.649$.
D) $m_h = 13.000$.

Exercise 1.4
Consider that a given statistical variable led to the following results:
1, 2, 1, 3, 2, 1, 0, 2, 0, 3, 4, 2, 1, 2, 1, 3, 4, 4, 1, 3.

a) The arithmetic mean, $\bar{x}$, is:

A) $\bar{x} = 1$.

B) $\bar{x} = 2$.
C) $\bar{x} = 3$.
D) None of the above.

b) The median m_e is:

A) $m_e = 1$.
B) $m_e = 2$.
C) $m_e = 3$.
D) None of the above.

c) The mode, m_o, is:

A) $m_o = 1$.
B) $m_o = 2$.
C) $m_o = 3$.
D) None of the above.

Exercise 1.5
Consider the following absolute frequency distribution, f_i, for salaries of workers of a company:

Salary	$[50,100[$	$[100,150[$	$[150,200[$	$[200,250[$	$[250,300[$
f_i	2	12	35	26	5

a) the arithmetic mean, $\bar{x}$, of the salaries.
b) the median, m_e, of the salaries.
c) the standard deviation, s, of the salaries.
d) the mode, m_o, of the salaries.

Exercise 1.6
In a test done to a given medicine the following results were obtained:

Treatment period x_i	$0-2$	$2-4$	$4-6$	$6-8$	$8-10$
Number of cures f_i	1	2	8	5	4

a) Compute the meadian, m_e, of the treatment period.

A) $m_e = 5.9$.
B) $m_e = 5.5$.
C) $m_e = 5.75$.
D) None of the above.

b) Determine the arithmetic mean $\bar{x}$ of the treatment period.

A) $\bar{x} = 5.5$.
B) $\bar{x} = 5.75$.
C) $\bar{x} = 5.9$.
D) None of the above.

Exercise 1.7
In a class of first year students' of the degree course in Mathematics, at the University Universitas, the following distribution of ages is observed:

Age (in years)	17	18	19	20	21	22
Number of students	1	10	7	5	5	2

a) The mode, m_o, of the ages is:

A) $m_o = 18$ years.
B) $m_o = 10$ years.
C) $m_o = 22$ years.
D) $m_o = 2$ years.

b) The arithmetic mean, $\bar{x}$, of the ages is:

A) $\bar{x} = 20.0$ years.
B) $\bar{x} = 19.3$ years.
C) $\bar{x} = 19.5$ years.
D) $\bar{x} = 18.0$ years.

c) The variance, s^2, of the ages is:

A) $s^2 = 1.81$ years^2.
B) $s^2 = 1.35$ years^2.
C) $s^2 = 0.59$ years^2.
D) $s^2 = 3.96$ years^2.

Exercise 1.8
The results of an exam of Statistics are represented in the following table where f_i and I_i are, respectively, the absolute frequency and the interval of class $i = 1, \ldots, 5$.

Classification I_i	$[0,4[$	$[4,8[$	$[8,12[$	$[12,16[$	$[16,20[$
Absolute frequency f_i	5	12	22	7	4

a) The arithmetic mean, $\bar{x}$, of the classifications is:

A) $\bar{x} = 4.158$.
B) $\bar{x} = 10.00$.
C) $\bar{x} = 9.440$.
D) None of the above.

b) The standard deviation s of the classifications is given by:

A) $s = 10.0$.
B) $s = 4.158$.
C) $s = 9.440$
D) None of the above.

Exercise 1.9
The analysis of prices of a meal in the restaurants of a particular city led to the following table:

Price	Absolute frequency
500 – 1000	6
1000 – 2000	9
2000 – 2500	17
2500 – 6000	4

Determine:

a) the arithmetic mean of the prices, $\bar{x}$.
b) the median of the prices, m_e.
c) the standard-deviation of the prices, s.
d) the mode of the prices, m_o.