FOUNDATION YEAR 2025-2026

COURSE - DATA ANALYSIS

Lesson 4

Niccolò Maffezzoli, Ca' Foscari University of Venice

January 22, 2026

Practical info

Mid-Term Exam: 09/02/26

Exercises on paper. Bring (scientific) calculator. No other electronic equipment allowed. No phone.

Final Exam: 02/03/26

Exercises on paper. Bring (scientific) calculator. No other electronic equipment allowed. No phone.

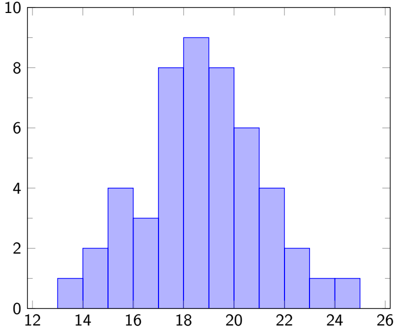
Coding project DUE 22/02/26 at 23:59

- Download dataset of your choice.
- Max 2 people per team. Send me team names by email BEFORE 02/02/26.
- Make an analysis in python. chatGPT is not allowed. I will find out :)
- Hand in all codes (in .zip folder) and a .PDF report on your analyses and results. Files: surname1_surname2.zip and surname1_surname2.pdf

How can we show our data in a compact way ? Histograms

In a movie theater the age of 49 people watching at a movie are the following:

Age (x_i)	frequency (f_{x_i})
13	1
14	2
15	4
16	3
17	8
18	9
19	8
20	6
21	4
22	2
23	1
24	1



Question

Why are the data distributed in this way ? Why aren't there as many 13-yr old as 20-yr old ?

Statistical distributions

- A statistical distribution describes the behavior of a random variable.
- The variable can be discrete or continuous.
- Commonly used distributions include:
  - Uniform Distribution
  - Binomial Distribution
  - Poisson Distribution
  - Gaussian (or Normal) Distribution
- The choice of distribution depends on the nature of the data and the underlying laws governing the phenomenon we are describing.
- The distribution can be characterized by its mean μ, variance σ^2, standard deviation σ, and shape parameters.

Uniform Distribution

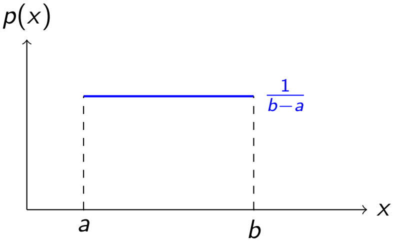
Definition

The uniform distribution is a probability distribution that describes a situation where **any value in a given range is equally likely to be observed.** The distribution is often denoted by the symbol $U(a, b)$, where a and b are the minimum and maximum values of the range.

Example

We roll a 6-sided dice 60 times. What are the possible outcomes ? What is the distribution of the results ?

Uniform Distribution



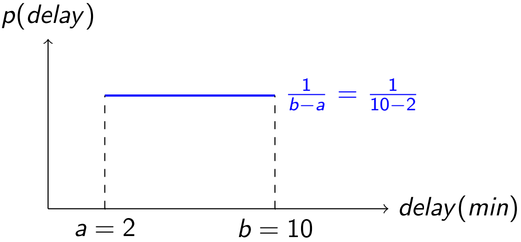
The probability density function (PDF) of the uniform distribution is given by:

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{elsewhere} \end{cases} \tag{1}$$

- The **mean** of the Uniform distribution is $\mu = (a + b)/2$.
- The **variance** of the Uniform distribution is $\sigma^2 = (b - a)^2/12$.
- The **standard deviation** of the Uniform distribution is $\sigma = (b - a)/\sqrt{12}$.

Uniform Distribution - Example

You take a bus to work and every day the bus is uniformly late between 2 and 10 minutes. On any given day, how long can you expect to wait, and with what standard deviation ?

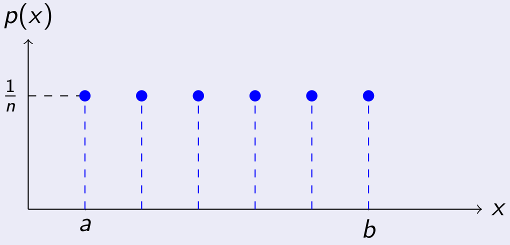


- $\mu = (a+b)/2 = (2+10)/2 = 6min$
- $\sigma = (b-a)/\sqrt{12} = (10-2)/\sqrt{12} = 2.3min$

Discrete uniform distribution

There is a finite number of values which are equally likely to be observed, i.e. every one of n values has equal probability 1/n.
It is the "dice" example. There are n=6 possible outcomes (1, 2, 3, 4, 5, 6) and every one of them has equal probability p=1/n=1/6.

Discrete uniform distribution



- $\mu = (a + b)/2$
- $\sigma^2 = (n^2 - 1)/12, with\ n = b - a + 1$

Exercise

Calculate the mean, variance and standard deviation of the distribution of outputs from a dice experiment.

- $\mu =?$
- $\sigma^2 =?$
- $\sigma =?$

Exercise

Calculate the mean, variance and standard deviation of the distribution of outputs from a dice experiment.

- $\mu =?$
- $\sigma^2 =?$
- $\sigma =?$

Solution

- $\mu = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} \simeq 3.5$
- $\sigma^2 = \frac{n^2-1}{12} = \frac{36-1}{12} = \frac{35}{12} \simeq 2.9$
- $\sigma = \sqrt{\frac{35}{12}} \simeq 1.7$

Binomial Distribution

Definition

- The binomial distribution described processes with a given number of identical trials, with two possible outcomes. The 'success' and 'failure' outcomes have probability p and *(1-p)*.
- The basic process is repeated n time - n is called number of trials.
- The distribution gives the probability of k successes out of these n trials.

Binomial PDF

$$P(k, p, n) = p^k (1-p)^{n-k} \frac{n!}{k!(n-k)!} \tag{2}$$

- k is the number of successes.
- n is the number of trials.
- p is the probability of success.
- $x!$ is called factorial. $x! = x \cdot (x-1) \cdot (x-2) \cdot (x-3) \cdot ... \cdot 3 \cdot 2 \cdot 1$.
  For example, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$
- Properties of factorial: $x! = x \cdot (x-1)!$; $0! = 1$

Binomial Distribution - examples

Example 1

Suppose, according to the latest police reports, 80% of all petty crimes are unresolved, and in your town, three petty crimes are committed. What is the probability that one of the three crimes will be resolved?

- The probability of success $p = 0.2$ (20% of cases are solved)
- Number of fixed trials $n = 3$ (number of petty crimes)
- Number of successes $k = 1$ (number of petty crimes solved)

$$P(k, p, n) = p^k (1-p)^{n-k} \frac{n!}{k!(n-k)!}$$

$$P(k = 1, p = 0.2, n = 3) = (0.2)^1 (1-0.2)^{3-1} \frac{3!}{1!(3-1)!} = 0.2 \cdot 0.8^2 \cdot 3 = 0.384$$

Binomial Distribution - examples

Example 2

In a glass bottle factory, the production line has an efficiency of 99.5% in producing a glass without defects. During a shift, 100 bottles are produced. What is the probability that all of them are produced correctly ?

- The probability of success $p = 0.995$ (99.5% efficiency)
- Number of fixed trials $n = 100$ (number of bottles produced)
- Number of successes $k = 100$ (number of correctly produced bottles)

$$P(k, p, n) = p^k (1-p)^{n-k} \frac{n!}{k!(n-k)!}$$

$$P(k = 100, p = 0.995, n = 100) = (0.995)^{100} (1 - 0.995)^{100-100} \frac{100!}{100!(100-100)!}$$

$$= 0.995^{100} = 0.61 = 61\%$$

Binomial Distribution - examples

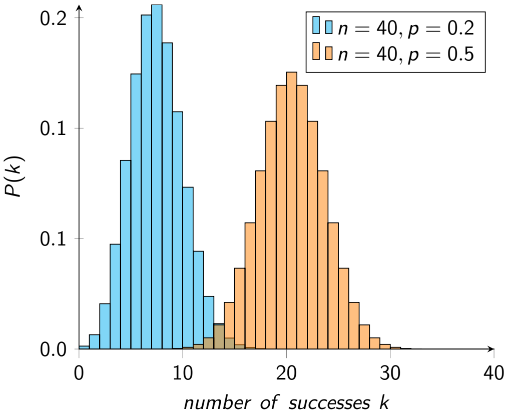
Example 3 (page 27 Barlow)

In an experiment into extrasensory perception, a person guesses the symbol on a card. There are five different symbols so she has a 20% chance of guessing right by chance. She guesses six cards - what is the probability of getting more than half correct by chance ?

Example 4 (page 27 Barlow)

You are trying to measure the tracks of cosmic ray particles using spark chambers, which are 95% efficient. You make the sensible decision that at least 3 points are needed to define a track. How efficient at detecting tracks would a stack of three chambers be ? Would using four of five chambers give a significant improvement ?

Binomial Distribution



- The **mean** of the Binomial distribution is $\mu = np$.
- The **variance** of the Binomial distribution is $\sigma^2 = np(1-p)$.
- The **standard deviation** of the Binomial distribution is $\sigma = \sqrt{np(1-p)}$.