Introduction to statistics and data science

A hands-on approach

Stéphane Puechmorel [†]

October 8, 2025

[†] ENAC Dept. SINA
email: stephane.puechmorel@enac.fr
phone: 0562259503

Outline

Introduction

Introduction to statistics and data science

Why this course ?

- Data are omnipresent in engineering and business intelligence.
- Understanding them can often be challenging.
- Statistics provides the data analyst with a rigorous framework.

Course contents:

- Visual data exploration.
- Concepts from probability theory.
- Inferential statistics.
- Hypothesis testing.

Work organization

- All lectures will include course material and exercises.
- The notions are illustrated using real data and statistical software.
- Should you encounter any difficulties, please do not hesitate to inquire.
- Personal work is essential and must be completed regularly.

Visual exploration of data

An introduction to PowerBI

A report generator

- Power BI is a reporting and data analytics tool. that simplifies data collection, transformation, and visualization tasks.
- It is designed to be extensible and tends has become the de facto standard in the industry.

What PowerBI is not

- While it performs well in terms of data visualization, it is not a full-fledged statistical software.
- The use of a superior data transformation and analysis tool is a highly effective strategy.

i

R

The workhorse of statistics

- A free and open source piece of software.
- R boasts implementation of nearly all developments in statistics.
- The number of free programs written in R is keeping increasing.

A companion to PowerBI

- PowerBI can collect data from a R script.
- PowerBI can transform data using a R script.
- PowerBI can display the output of a R script.

R IN A NUTSHELL

R as a desk calculator

- Open R by double-clicking on the R icon.
- Simply type ”2+3” and observe the result.

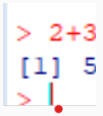

Figure 1: A simple sum

- A value can be stored in a variable with the operator $<-$.

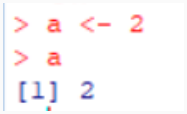

Figure 2: Using a variable.

VECTORS

Vectors

- Vectors are created using the ”c” (concatenate) or range operator ”min:max”.
- Try ”c(0.1, 2, 3)” and ”1:100”.
- If you want a non-unit step, use the ”seq(min, max, step)” command.

```
> seq(0,1,0.1)
 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

Figure 3: User-defined step

- All the elements of a vector must be of the same type: numeric (double, integer, complex), logical, string.

VECTORS

Numeric and boolean types

- By default, a numeric value is encoded as a double precision number on 64 bits.
- A complex number is declared with syntax $x + yi$:

```
> 2+3i
[1] 2+3i
> (2+3i)*(1+2i)
[1] -4+7i
```

Figure 4: Complex numbers

- Booleans are TRUE or FALSE.

```
> c(TRUE, TRUE,FALSE, FALSE)
[1]  TRUE  TRUE FALSE FALSE
```

Figure 5: Booleans

VECTORS

Character strings

- A string is enclosed with double or simple quotes.

```
> c("hello",', world')
[1] "hello"   ", world"
```

Figure 6: Character strings

- Any R object can be converted to an informative string with the function `str()`.

```
> str(a)
 logi [1:4] TRUE TRUE FALSE FALSE
```

Figure 7: Str function.

VECTORS

Operators

- Operators and functions are applied elementwise.

```
> a <- 1:10
> b <- 1:10
> a+b
 [1]  2  4  6  8 10 12 14 16 18 20
> a*b
 [1]   1   4   9  16  25  36  49  64  81 100
> sin(a)
 [1]  0.8414710  0.9092974  0.1411200 -0.7568025 -0.9589243 -0.2794155
 [7]  0.6569866  0.9893582  0.4121185 -0.5440211
```

Figure 8: Operating on vectors.

- This is true also for comparison operators.

```
> c(1,3,5,8) <= c(0,4,5,8)
[1] FALSE  TRUE  TRUE  TRUE
```

Figure 9: Comparison operators.

VECTORS

Accessing elements

- An element in a vector can be referred to by its index.

```
> a<-1:10
> a[1]
[1] 1
> a[2:4]
[1] 2 3 4
> a[c(2,6,8)]
[1] 2 6 8
```

Figure 10: Accessing elements.

- A selection by booleans is also possible.

```
> a[c(TRUE, FALSE, TRUE)]
[1]  1  3  4  6  7  9 10
> a[a >= 5]
[1]  5  6  7  8  9 10
```

Figure 11: Boolean selection.

Data frames

A convenient object.

- Data frames are arrays holding observed values.
- Observed characteristics are in columns, observations in rows.
- The columns are named and can be referred to by their names.

```
> dataset <- data.frame(id=1:100,val=0:0.01:99)
> str(dataset["id"])
'data.frame':   100 obs. of  1 variable:
 $ id: int  1 2 3 4 5 6 7 8 9 10 ...
> str(dataset$val)
 int [1:100] 0 1 2 3 4 5 6 7 8 9 ...
```

Figure 12: A data frame.

- Data frames are used by PowerBI to communicate with R.

PASSENGERS DATASET

Data collection

- The dataset, collected from the site `https://data.europa.eu/data/datasets?locale=en` is stored in the file `avia_paoc_monthly.csv` that can be retrieved on ecampus in the folder `Datasets`.
- Launch PowerBI and select `Get data from other sources`.
- Select `Text/CSV` and open the file.
- A table view of the dataset appears.

PASSENGERS DATASET

Data transformation

- PowerBI offers the opportunity to transform the data.

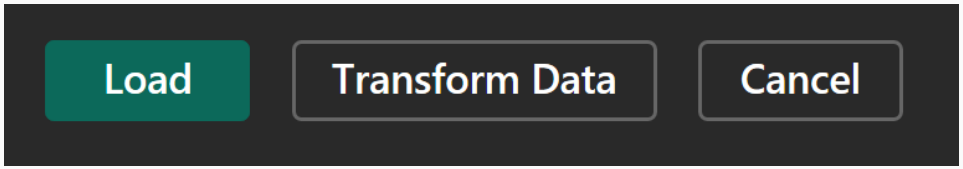

Figure 13: Data transform.

- We are only interested in the total number of passengers carried by country, so all the rows below row 32 must be removed.
- Select `Keep rows`, then `Keep top rows` and finally enter 32.

PASSENGERS DATASET

Final transformations

- In the first column, a description of the data includes at the end a geographic information.
- Select the first column, right click on it, the select `replace values`.
- Copy all the text in the first cell, except the final two-characters country code.
- Keep the replacement text empty and click `replace`.

PASSENGERS DATASET

Transposing datasets

- Datasets must have their values in columns.
- Select the tab `Transform`, then `Transpose`.
- Return to the `Home` tab and select `Use first row as headers`.
- The label of the first column is not very informative. Right click on it and select `Rename`.

PASSENGERS DATASET

Cleaning data

- A dataset may have blank row. Check yours and remove them if needed (`Remove rows` tab).
- Some values may be missing. In this dataset, they are coded as ":".
- The column "MK" has such values. Right click on it, then use `Replace values` to change ":" by "0".
- finally, change the type of the column to `Whole number`.

Applying transforms

- Select the `Close and apply` to apply the transforms.
- It is time to save your work (tab `File`) !

PASSENGERS DATASET

Visual exploration of the dataset

- Return to the `Home` tab, and start populating the report.
- Text boxes or visual can be added.
- Select a text box and enter the text you wish to start your report.
- Select `New visual`. A new box appears (by default a bar chart).
- Select `Date` in the `Data` right pane and drag the `Month` field to the visual. This defines the ”x” axis.
- You can add data to your visual by selecting columns in the `Data` pane or by dragging them.
- Try changing the visual and see what happens.

PASSENGERS DATASET

Data aggregation

- In the ”passengers” dataset, there is a one-to-one mapping between a month and the number of passengers.
- What happens if we decide to plot against the quarter ?
- Uncheck `Month` in `Date` on the right pane and select `Quarter`.
- The visual now displays the total number of passengers for each quarter.
- In the `Y axis` part of the `visualization` pane, try to change `Sum` to `Average`.
- The aggregation of values is an important feature in PowerBI.

PASSENGERS DATASET

Measures

- A measure is a value that can be computed from the dataset.
- It may be viewed as a practical implementation of a random variable, a concept introduced later in the course.
- Since measures are computed on-the-fly, they always reflect the most up-to-Date dataset.
- Using them is a must in PowerBI.
- Create a `New quick measure` on the ”passengers” dataset that sums the amount of carried passengers by quarter for a given country (select the one you prefer).
- Use a visual to plot the result.
- Custom measures can be used, provided you know the ”DAX” language.

PowerBI and R

R as a data source

- Open the `Get data` tab, then select `R`.
- In the script box, type:

```
dataset <- data.frame(x=1:100,y=runif(n=100,min=0,max=10))
```

- Observe the result in PowerBI. What happens if the name `dataset` for the ouput variable is changed ?
- Create two data frames in the R script: what happens in PowerBI?

POWERBI AND R

R as a data processor

- Open the `Get data` tab, then select `Run R script`.
- In the script box, type:

```
1 output <- dataset
2 output$z <- as.numeric(dataset$y)+1
```

- A new column named `z` was added and its values are those of column `y` plus 1.

POWERBI AND R

Visuals

- On the `visualizations` pane, select the `R` icon.
- Drag the `y` field in the visual.
- In the script box, type:

```
1 hist(dataset$y)
```

- An histogram is displayed in the visual.
- This may be useful when a given visual, like the histogram, is not implemented in PowerBI or requires a plugin.

Probability theory

A BIT OF HISTORY

Games of chance

- According to popular belief, knights returning from the Crusades introduced a dice game named ”Hazard.” Although the rules were complicated, the game attracted many players.
- The mathematician Gerolamo Cardano, in the middle of the 16th century, starts to investigate the odds of winning in a game of chance. He gives the first definition of the probability of an event as the ratio of the number of favorable outcomes to the total number of outcomes.
- One century later, Pierre de Fermat and Blaise Pascal lay the first theory of probability when answering a question from Antoine Gombaud, a famous Parisian gambler.

A BIT OF HISTORY

A modern view

- In 1933, the Russian mathematician Andrey Kolmogorov gives the first axiomatic description of probability theory.
- This is the approach we will take in the sequel.

Challenges

- To fully understand Kolmogorov's contribution, we must consider cases in which counting the number of favorable outcomes is insufficient.
- For example, take a situation where the measure of interest is a real number, such as a current or a pressure.
- The value itself is not discrete; only the number of times it falls within a given interval can be observed.
- The Fermat-Pascal approach must be extended to these cases.

From intuition to axioms

Sample space

- Let's consider a very simple experiment: a coin is tossed, and the outcome observed.
- What are the possible results ?
- Of course, "head" and "tails" !
- The **sample space** of an experiment is the set of all possible results.
- In the coin tossing example, it is just the set $\Omega = \{\text{heads}, \text{tails}\}$.
- If a die is rolled, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- In a physical system, Ω may be a subset of the real numbers, or something even more complicated.

From intuition to axioms

Events

- If two coins are tossed, the sample space is $\Omega = \{HH, HT, TH, TT\}$.
- An **event** is a statement about the outcome of the experiment, like ”at least one tails was observed”.
- The previous event is the subset $\{HT, TH, TT\}$ of Ω.
- In probability theory, an **event** is a subset of the **sample space**.
- The empty set $\emptyset$ and the sample space Ω are always considered as events.

From intuition to axioms

A bit of logic

- Assertions about events are also events:
  - If A, B are events, so is $A \cup B$.
  - If A is an event, the complementary set $\bar{A}$ is an event.
- For the two coins tossing experiment, if $A = \{HH\}, B = \{HT\}$, then the event $A \cup B = \{HH, HT\}$ is "the first toss is a head", while the event $\bar{A} = \{HT, TH, TT\}$ is "at least one tails was observed".

Exercise

- Find the sample space Ω associated with the roll of a die.
- What are the respective subsets A, B of Ω corresponding to the observations "the result is greater than or equal to 3","the result is an even number"?
- Describe the events $A \cup B$, $\bar{A}$.

PROBABILITIES

σ-algebras

- Given a sample space Ω, a σ-algebra of events is a set $\mathcal{T}$ of subsets from Ω such that:
  - $\Omega \in \mathcal{T}$.
  - If $A \in \mathcal{T}$, then $\bar{A} \in \mathcal{T}$.
  - For any **countable** sequence A_n in $\mathcal{T}$, the union $\bigcup_n A_n$ is in $\mathcal{T}$.

Probability measure

- Given a σ-algebra $\mathcal{T}$ on Ω, a probability P is a mapping that assigns to $A \in \mathcal{T}$ a real number $P(A) \in [0, 1]$ such that:
  - $P(\Omega) = 1$.
  - For any countable sequence A_n of **pairwise disjoint** events in $\mathcal{T}$,

$$P\left(\bigcup_n A_n\right) = \sum_n P(A_n). \tag{1}$$

SOME PROPERTIES

Probability space
A sample space Ω together with a σ-algebra of events $\mathcal{T}$ and a probability P is called a **probability space**, denoted by $(\Omega, \mathcal{T}, P)$.

Set operations
Let $(\Omega, \mathcal{T}, P)$ be a probability space. Let A, B be events. Then:

- $P\left(\bar{A}\right) = 1 - P(A)$.
- $P\left(A \cup B\right) = P(A) + P(B) - P\left(A \cap B\right)$.
- If $B \subset A$, then $P\left(A - B\right) = P(A) - P(B)$, where $A - B = A \cap \bar{B}$.

THE DIE ROLL

Defining a Probability

- $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{T} = \mathcal{P}(\Omega)$, the σ-algebra of all subsets.
- How to find a probability on $\mathcal{T}$?
- It is required that $P(\Omega) = 1$.
- If the die is fair, then the probability to observe any number from Ω must be the same.
- Since:

$$1 = P(\Omega) = \sum_{i=1}^{6} P(\{i\}), \tag{2}$$

  one deduces $P(\{i\}) = 1/6, i = 1 \ldots 6$.
- The probability of any event can then be obtained by summing the probabilities of its elements.

EXERCISE

A card game

- A gambler randomly draws a hand of 5 cards from a deck of 52. Can you find a sample space describing this experiment ?
- If the probabilities of all hands are equal, what is the probability of having a four of a kind ?
- Hint: the number of ways to select k elements in a set of n is :

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}. \tag{3}$$

Conditional probability

Motivation

- In many cases, there is some information about the outcome of an experiment.
- As an example, for a die roll, it may be ”the result is odd”.
- The sample space is thus reduced, and the probabilities must be rescaled accordingly.
- For the die example, it is $\{1, 3, 5\}$, and the probability to draw a 3 is $1/3$.

Bayes' formula

The conditional probability of A knowing B, denoted $P(A|B)$, is given by the Bayes' formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{4}$$

Around Bayes' formula

Useful formulas

- Let (B_n) be a countable sequence of pairwise disjoint events such that $\bigcup_n B_n = \Omega$. For any event A, one has the formula of total probabilities:

$$P(A) = \sum_n P(A|B_n)\, P(B_n)\,. \tag{5}$$

- Given events $B_1 \ldots B_n, A$, one has:

$$P(A) = P(A|B_1, \ldots, B_n)\, P(B_n|B_{n-1}, \ldots B_1) \ldots P(B_1)\,. \tag{6}$$

- Conditioning can be reversed using the next formula, provided $P(B) \neq 0$:

$$P(B|A) = \frac{P(A|B)\, P(B)}{P(A)} \tag{7}$$

Exercise

Two factories, denoted F_0, F_1 are producing components. The end customer receives a component that comes from F_0 with probability 0.3. Furthermore, the probability of the component being defective is 0.1 when coming from factory F_0 and 0.06 when coming from factory F_1. If the recived component is defective, what is the probability that it has been produce by F_1 ?

Random variables

Definition
Given two sample spaces E, F equipped with respective σ-algebras of events $\mathcal{T}, \mathcal{F}$, a mapping $X\colon E \to F$ is said to be a random variable if, for any event $A \in \mathcal{F}$, $X^{-1}(A)$ is an event of $\mathcal{T}$.

Example
Consider a die roll with $E = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{T} = \{E, \emptyset, \{2, 4, 6\}, \{1, 3, 5\}\}$. Let $F = \{0, 1\}, \mathcal{F} = \mathcal{P}(F)$. The mapping X that associates to an even number the value 0 and 1 to an odd number is a random variable. However, Y that maps any value less than 3 to 0 and the others to 1 is not.

THE LAW OF A RANDOM VARIABLE

Définition

- In the previous example, the probability of the event $X = 0$ is $1/2$, the probability of $X^{-1}(0) = \{2, 4, 6\}$.
- If X is a random variable on a probability space $(E, \mathcal{T}, P)$ with values in F, the law of X is the probability on $\mathcal{F}$ defined by:

$$P_X(A) = P\left(X^{-1}(A)\right). \tag{8}$$

- In many cases, the outcome of an experiment is impossible to observe, but some agregated value is, e.g. the pressure or the temperature of a gas.
- Formally, this is a random variable.
- Only its law has some interest.

Discrete random variables

The law of a discrete random variable

- A random variable is said to be discrete if it takes its values in a finite or infinite countable set.
- The law of such a random variable X with values in a set E is entirely determined by its **distribution** $p(x) = P_X(\{x\}), x \in E$.

Expected value

- Let X be a discrete random variable with values in $E \subset \mathbb{R}$. Its expected value is defined as:

$$E[X] = \sum_{x \in E} xp(x), \tag{9}$$

provided:

$$\sum_{x \in E} |x| p(x) < +\infty. \tag{10}$$

Exercise

A game of chance

- Two dice are rolled and their values are added.
- Show that the sum is a random variable X with values in $\{2, \ldots, 12\}$.
- Find the distribution of X.
- If you bet at the start, you'll win five times your original bet if $X \geq 10$. Is this rule fair?

Exercise

Bernoulli random variable

- A bernoulli random variable X can take only two values: 0 and 1.
- If $P(X = 1) = p$, can you compute $P(X = 0)$?
- Compute $E[X]$.

Binomial random variable

- A binomial random variable X with parameters $n \in \mathbb{N}, p \in [0, 1]$ has distribution:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0 \dots n. \tag{11}$$

- Find $E[X]$.

Expected value

Properties

- If X, Y are discrete random variables with values in $E \subset \mathbb{R}$ and λ is a real number, then:

$$E[\lambda X + Y] = \lambda E[X] + E[Y]. \tag{12}$$

- The expected value is monotone. If $X \leq Y$, then:

$$E[X] \leq E[Y] \tag{13}$$

Moments

The expected value in a general sense

- If $g\colon \mathbb{R} \to \mathbb{R}$ is piecewise continuous and

$$\sum_{x \in E} |g(x)| p(x) < +\infty, \tag{14}$$

then one defines:

$$E\left[g(X)\right] = \sum_{x \in E} g(x) p(x) \tag{15}$$

- The variance of the discrete random variable X is:

$$V(X) = E\left[\left(X - E\left[X\right]\right)^2\right] \tag{16}$$

- It measures the dispersion of X around its expected value.

Moments

Higher order moments

- Let n be an integer. The n-th moment of X is, if it exists:

$$E\left[X^n\right]. \tag{17}$$

- The moment generating function is the function:

$$t \mapsto m_X(t) = E\left[e^{tX}\right] \tag{18}$$

- If its domain contains 0, then:

$$E\left[X^n\right] = m^{(n)}(0), \tag{19}$$

where $m^{(n)}$ denotes the n-th derivative of m_X.

CONDITIONING

Independence

- By the Bayes' formula:

$$P\left(X = x | Y = y\right) = \frac{P\left(X = x, Y = y\right)}{P(Y = y)}. \tag{20}$$

- If $P\left(X = x, Y = y\right) = P\left(X = x\right) P\left(Y = y\right)$, then $P\left(X = x | Y = y\right) = P\left(X = x\right)$.
- In such a case, X, Y are said to be **independent**.
- If X, Y are independent, then $E\left[XY\right] = E[X]E[Y]$.
- If $X_1, \ldots, X_n$ are pairwise independent, then:

$$V\left(X_1 + \cdots + X_n\right) = V\left(X_1\right) + \cdots + V\left(X_n\right). \tag{21}$$

EXERCISE

A pair of dice... again !

- Let X, Y be two independent random variables corresponding to the respective values of two die throws. Compute :

$$E[X], E[Y], E[X+Y].$$

- Same question with the variances.

A taste of estimation theory

- Let $X_1, \ldots X_n$ be a sequence of independent random variables and let:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Compute $E\left[\bar{X}\right], V\left(\bar{X}\right)$. What happens if $n \to +\infty$?

Poisson distribution

A model for random events

- Given a time interval of length $[t_0, t_1]$, one counts the number of occurrences X of an event.
- The random variable X is said to have a Poisson distribution with rate λ if:

$$P(X = k) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}, \tag{22}$$

  where $T = t_1 - t_0$.
- $E[X] = V(X) = \lambda T$.
- The Poisson distribution models a situation where the occurrences of an event are independent and occur at constant rate.

Exercise

Distribution of a sum

- The moment generating function (MGF) of a random variable is the function $m_X(t) = E\left[e^{tX}\right]$. It is characteristic of a distribution. Prove that if X, Y are independent, $m_{X+Y}(t) = m_X(t)m_Y(t)$.
- Compute the MGF of X with Poisson distribution of rate λ.
- Let X, Y be independent random variables with Poisson distributions of respective rates λ, μ. Show that $X + Y$ has Poisson distribution of rate $\lambda + \mu$.

Selecting events

Let X be a random variable with Poisson distribution of rate λ. The underlying events are selected at random with probability p and counted, yielding a random variable Y. Prove that Y has Poisson distribution with rate $p\lambda$.

Real random variables

Definition

- Let $(\Omega, \mathcal{T}, P)$ be a measure space. A mapping $X\colon \Omega \to \mathbb{R}$ is said to be a real random variable if, for any $t \in \mathbb{R}$, $\{\omega | X(\omega) \leq t\}$ is an event in $\mathcal{T}$.
- If X is a real random variable, its cumulative distribution function (CDF) is the mapping:

$$F_x\colon t \in R \mapsto P\left(X \leq t\right). \tag{23}$$

- $\lim_{t\to-\infty} F_x(t) = 0$, $\lim_{t\to+\infty} F_x(t) = 1$.
- F_x is right continuous and admits a limit to the left.
- If F_x is differentiable, its derivative is called the density of X.

Real random variables

Expected value

- Let X be a real random variable with density p_X.
- If $g\colon \mathbb{R} \to \mathbb{R}$ is piecewise continuous, then one defines:

$$E\left[g(X)\right] = \int_{\mathbb{R}} g(t)p_X(t)dt \tag{24}$$

provided:

$$\int_{\mathbb{R}} |g(t)|p_X(t)dt < +\infty. \tag{25}$$

- As special cases:

$$\begin{aligned} P\left(X \in [a,b]\right) &= E\left[1_{[a,b]}(X)\right] = \int_a^b p_X(t)dt \\ F_X(t) &= \int_{-\infty}^t p_X(t)dt. \end{aligned} \tag{26}$$

Real random variables

Joint CDF and density

- If X, Y are real random variables, the CDF of the couple (X, Y) is the mapping:

$$F_{X,Y} \colon (s,t) \in \mathbb{R}^2 \mapsto P\left(X \leq s, Y \leq t\right). \tag{27}$$

- Taking the derivative with respect to s, t allows defining the joint density:

$$p_{X,Y}(x,y) = \frac{\partial^2}{\partial_s \partial_t} F_{X,Y}(t,s)|_{x,y}. \tag{28}$$

- X, Y are said to be independent if:

$$F_{X,Y}(s,t) = F_X(s) F_Y(t). \tag{29}$$

Marginal densities

- Let X, Y be a couple of real random variables with joint density $p_{X,Y}$.
- The density of X (resp. Y) can be obtained as:

$$p_X(x) = \int_{\mathbb{R}} p_{X,Y}(x,t)dt \text{ (resp.) } p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(s,y)ds \tag{30}$$

- The conditional density of X knowing Y is defined as:

$$p\left(X|Y=y\right)(s) = \frac{p_{X,Y}(s,y)}{p_Y(y)}. \tag{31}$$

- If X, Y are independent, $p(X|Y=y)(s) = p_X(s)$ or, equivalently, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$.

CHARACTERISTIC FUNCTION

Definition

- Let X be a real random variable. Its characteristic function is:

$$\phi_X : t \in \mathbb{R} \mapsto E\left[e^{itX}\right]. \tag{32}$$

- Using the density p_X, it can be computed as:

$$\phi_X(t) \int_{\mathbb{R}} e^{itx} p_X(x) dx. \tag{33}$$

Properties

- If two random variables have the same characteristic function, they have the same distribution.
- If X, Y are independent, $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.
- $E\left[X^k\right] = i^{-k}\phi_X^{(k)}(0), k \in \mathbb{N}$.

UNIFORM DISTRIBUTION

- The uniform distribution on an interval $[a, b]$ has density:

$$p(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b], \\ 0 & \text{otherwise.} \end{cases} \tag{34}$$

- If X has uniform distribution on $[a, b]$, then:

$$E[X] = \frac{a+b}{2}, \; V(X) = \frac{(b-a)^2}{12}, \\ \phi_X(t) = \frac{e^{itb} - e^{ita}}{t(b-a)}. \tag{35}$$

Normal distribution

Density

- Let $\mu \in \mathbb{R}, \sigma > 0$. A real random variable X is said to have a normal distribution $\mathcal{N}(\mu, \sigma^2)$ if its density is:

$$p_X(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}. \tag{36}$$

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then:

$$Y = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1). \tag{37}$$

Normal distribution

Moments

If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$:

$$E\left[X\right] = \mu,\ V\left(X\right) = \sigma^2,$$
$$\phi_X(t) = \exp\left(i\mu t - \sigma^2 t^2/2\right). \tag{38}$$

Cumulative distribution function

- There is no closed-form expression for the CDF of a $\mathcal{N}(\mu, \sigma^2)$ distribution, but it can be evaluated numerically.
- If Φ is the CDF of a $\mathcal{N}(0, 1)$ distribution, then the CDF of a $\mathcal{N}(\mu, \sigma^2)$ distribution is:

$$\Phi\left(\frac{x-\mu}{\sigma}\right). \tag{39}$$

CENTRAL LIMIT THEOREM

Convergence in distribution

A sequence $(X_n)_{n\mathbb{N}}$ of real random variables is said to converge in distribution to a random variable X if:

$$\forall t \in \mathbb{R}, \lim_{n \to +\infty} P\left(X_n \leq t\right) = P\left(X \leq t\right). \tag{40}$$

Theorem (Central limit)

If $(X_n)_{n\in\mathbb{N}}$ is a sequence of independent, identically distributed real random variables, then the sequence of random variables $\sqrt{n}\left(\bar{X}_n - \mu\right)$ converges in distribution to $X \sim \mathcal{N}(0, \sigma^2)$, with:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \mu = E\left[X_i\right], \sigma^2 = V\left(X_i\right). \tag{41}$$

EXPONENTIAL DISTRIBUTION

- The exponential distribution with rate $\lambda > 0$ has density:

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0, \\ 0, x < 0. \end{cases} \tag{42}$$

- If X is exponentially distributed with rate λ :

$$E[X] = \frac{1}{\lambda}, V(X) = \frac{1}{\lambda^2}, \phi_X(t) = \frac{\lambda}{\lambda - it}. \tag{43}$$

- The exponential distribution is memoryless:

$$P(X > t + s | X > s) = P(X > t), s, t \geq 0. \tag{44}$$

VECTOR RANDOM VARIABLES

CDF and density

- Let $(\Omega, \mathcal{T}, P)$ be a measure space. Let $X\colon \Omega \to \mathbb{R}^n$ be a mapping. X is said to be a vector (or multivariate) random variable iff:

$$\{\omega | X_1(\omega) \leq t_1, \ldots, X_n(\omega) \leq t_n\} \in \mathcal{T},\ (t_1, \ldots t_n) \in \mathbb{R}^n. \tag{45}$$

- The CDF of X is:

$$F_x(t_1, \ldots, t_n) = P\left(X_1 \leq t_1, \ldots, X_n \leq t_n\right). \tag{46}$$

- Taking the partial derivatives yields the density:

$$p_X(t_1, \ldots, t_n) = \frac{\partial^n F_X(t_1, \ldots, t_n)}{\partial t_1 \ldots \partial t_n}. \tag{47}$$

VECTOR RANDOM VARIABLES

Moments

- The expected value of a vector random variable X is just the vector of the expected values of the coordinates:

$$E[X] = (E[X_1], \ldots, E[X_n]). \tag{48}$$

- The density of $X_i, i = 1 \ldots n$ can be computed by integration:

$$p_{X_i}(x) = \int_{\mathbb{R}^{n-1}} p_X(t_1, \ldots, x, \ldots t_n)\, dt_1, \ldots dt_n, \tag{49}$$

where x occurs in the i-th position.

VECTOR RANDOM VARIABLES

Moments

- Higher moments are tensors, generally difficult to compute.
- The covariance matrix of two random variables is:

$$Cov(X, Y) = E\left[XY^t\right] - E\left[X\right] E\left[Y\right]^t. \tag{50}$$

- The density of a vector normal distribution $\mathcal{N}(\mu, \Sigma)$ is:

$$P_X(x) = \frac{1}{(2\pi)^{n-2} \det(\Sigma)^{-1/2}} \exp\left(\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right), \tag{51}$$

with $\mu = E\left[X\right]$, $\Sigma = Cov(X, X)$.

VECTOR RANDOM VARIABLES

Characteristic function

- The characteristic function of a vector random variable X is defined as:

$$\phi_X(t_1, \ldots, t_n) = E\left[e^{i(t_1 X_1 + \cdots + t_n X_n)}\right]. \tag{52}$$

- For a $\mathcal{N}(\mu, \Sigma)$ distribution, it is:

$$\phi_X(t) = \exp\left(i\mu^t t - \frac{1}{2} t^t \Sigma t\right). \tag{53}$$

VECTOR RANDOM VARIABLES

Central limit theorem

- If $X_1, \ldots X_n$ are independent, identically distributed vector random variables, then $\sqrt{n}\left(\bar{X}_n - \mu\right)$ converges in distribution to $\mathcal{N}\left(0, \Sigma\right)$ with:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \Sigma = Cov(X_i, X_i). \tag{54}$$

- In practice, almost all problems in data analysis can be solved using the vector CLT.

Estimation

INTRODUCTORY EXAMPLE

Delays dataset

- Download the file ”delays.csv” and open it in PowerBI.
- Use the R visual tab to create an histogram:

```
1 hist(dataset$Column3, breaks=100, freq=FALSE)
```

- Negative delays are not of interest for us. Use the ”Filters” pane to keep only postive ones.
- Create a new measure and type:

```
avg_delays = AVERAGEX(FILTER(delays,delays[Column3]>=0),
                                    delays[Column3])
```

- Finally, visualize the newly created measure by selecting the ’Card’ visual.

INTRODUCTORY EXAMPLE

Modeling the delays

-