Statistics for Data Science

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Logistics

- Instructors: Myrto Limnios and Rajita Chandak
- TAs: Ramzi Dakhmouche, Yann Becker, Regina García Averell.
- Schedule:
  - Lectures: Mondays 10:00–12:00 (CM 1 1) + Tuesdays 08:00–10:00 (CM 1 1)
  - Exercises: Wednesdays 08:00-10:00 (CM 1 5)
- Midterm: Date TBA (Apr 2 or 7), 100 minutes
- Final Exam: TBD, 3 hours
- Course website: `https://go.epfl.ch/MATH-413`
- Ed Forum: See Moodle.
- Lecture notes: =slides (more references on website).

Course Structure

Probabilistic background

1. Probability.
2. Reminder on Basic Probability Distributions.
3. Entropy and Exponential Families.

Sampling Theory

1. Sufficient Statistics.
2. Sampling distributions.
3. Stochastic convergence

Marginal Inference

1. Point Estimation and Likelihood Theory.
2. Hypothesis Testing and Confidence Intervals.
3. Nonparametric marginal inference and smoothing.

Inference with covariates (Regression)

1. Gaussian Linear Models.
2. Generalised Linear Models
3. Nonparametric regression and regularisation.

Disclaimer

- This is intended to be (and will be) a mathematical course.
- There will be some proofs.
  - $\hookrightarrow$ Proofs marked with an $*$ will not be examined, though.
- We will start from first principles and build up our theory.
- Reality is messy, complicated, and does not easily submit to narrative.

What is Statistics?

Learning from Data under Uncertainty

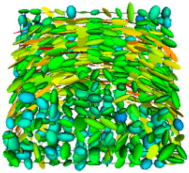
Statistics: a more eloquent attempt

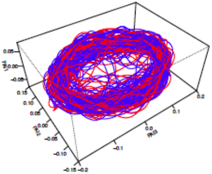
We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

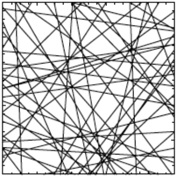
Ronald A. Fisher
(Biologist and Mathematician)

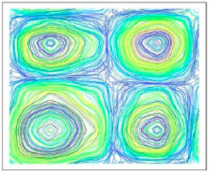
What is Data? Anything and Everything

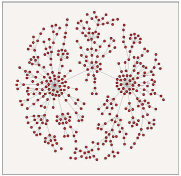
Anything[1]

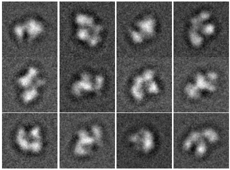


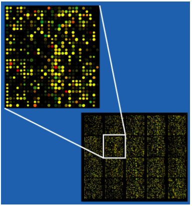


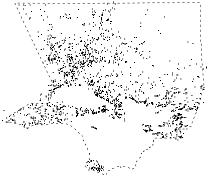












[1] That we can mathematically represent

A subjective view (others might object):

$$\text{Machine Learning} \quad \subset \quad \text{Statistics}$$

(this does *not* mean that all machine learning was invented by statisticians!)

In some ways, machine learning corresponds one of two cultures in statistics[2]:

1. Inference/Modeling/Uncertainty
   - ↪ Focus on interpretability, reduction, and statistical efficiency.
   - ↪ Traditionally linked with science.
2. Prediction/Algorithms/Optimisation
   - ↪ Focus on emulation, automation, and computational efficiency.
   - ↪ Linked more with technology.

The two need not be mutually exclusive –
depends on the problem and intended use!

Cultural differences run deep – philosophy of mind:

what does it mean to know/learn?

[2] Breiman (2001), "Statistical Modeling: The Two Cultures" – make sure to read discussion.

A mathematical model is a simplified representation of reality expressed in mathematical terms that, despite being an approximation, gives a fruitful framework accounting for our empirical observations and generating further conjectures not directly suggested by experience itself

- Building/disputing/calibrating/refining/improving models is at the core of the scientific method
- Inextricably linked with the concept of "theory".
- Constructed combining empirical observations, mathematical considerations, and philosophical principles.
- Can be seen as a vehicle for learning – parsimoniously reduces complexity and diversity of observations and makes accurate predictions.

Uncertainty may stem from many sources:

1. Sampled data (observe the *particular* but not the *general*).
2. Measurement error.
3. Chaos.
4. Intrinsic stochasticity.
5. Fundamental limitations to precision.

⋮

We will use Kolmogorov's axiomatic system of probability theory to mathematically encapsulate uncertainty.

Probability:

1. Process of interest conceptualised as a probability model
2. Use model to learn about probability of potential outcomes.

Statistics:

1. Process of interest conceptualised as a probability model
2. Data viewed as observed outcomes from model
3. Use outcomes to learn about the model.

The Job of the Probabilist

Given a probability model $\mathbb{P}$ on a space Ω find the probability $\mathbb{P}[A]$ that the outcome of the experiment is $A \subset \Omega$.

The Job of the Statistician

Given an outcome of $A \subset \Omega$ (the data) of a probability experiment on Ω, tell me something *interesting** about the (uknown) probability model $\mathbb{P}$ that generated it.

(*something in addition to what was known before observing the outcome A)

Such *interesting* questions can be:

1. Are the data more more consistent with one or another model?
2. Given a family of models, can we determine which model generated the data?
3. What range of models are consistent with a given set of data?
4. How to best answer these questions? (is there even a best way?)

Example (A Probabilist and a Statistician Flip a Coin)

Let $Y_1, ..., Y_{10}$ denote the results of flipping a coin ten times, with

$$Y_i = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails} \end{cases}, \quad i = 1, ..., 10.$$

A plausible model is $Y_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$. We record the outcome

$$(0, 0, 0, 1, 0, 1, 1, 1, 1, 1).$$

Probabilist Asks:

- Probability of outcome as function of θ?
- Probability of k-long run?
- If keep tossing, how many k-long runs? How long until k-long run?
- What about the sum of observations? How does it behave? How does it scale?

Example (A Probabilist and a Statistician Flip a Coin (cont'd))

Statistician Asks:

- Is the coin fair?
- What is a good guess of the value of θ on the basis of the observations?
- What range of θ is plausible on the basis of the observations?
- How much error do we make when trying to decide the above from the observations?
- How does our answer change if the observations are perturbed?
- Is there a "best" solution to the above problems?
- How sensitive are our answers to departures from the model?
- How do our "answers" behave as $\#$ tosses $\longrightarrow \infty$?
- How many tosses would we need until we can get "accurate answers"?

Example (A Probabilist and a Statistician invest in the stock market)

Let $Y_1, ..., Y_t$ denote the price of a certain stock, say AAPL, over a period of t days A plausible model for the stock price is $Y_i \sim \text{Black-Scholes}(\mu, \sigma)$. Probabilist

Asks:

- Probability of the price crossing \$300 in the next year?
- What about any known function of the stock prices in a week? How does it behave? How does it scale?

Example (A Probabilist and a Statistician invest in the stock market (cont'd))

Statistician Asks:

- Is the mean price of AAPL equal to 200 in the last 5 years?
- What is a good guess of the value of σ on the basis of the observations?
- What range of μ is plausible on the basis of the observations?
- How much error do we make when trying to estimate the parameters from the observations?
- How does our answer change if we know our observations are noisy?
- Is there a "best" solution to the above problems?
- How sensitive are our answers to departures from the model?
- How do our "answers" behave as $t \longrightarrow \infty$?

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

Seems too simple?

→ Spans essence of most of the ideas used in the most complex of problems!

→ In principle, $\mathcal{Y}^n$ and Θ can be quite complicated, though:

- Almost always $\mathcal{Y}^n \subseteq \mathbb{R}^n$.
- Typically $\Theta \in \mathbb{R}^p$, some fixed p. Sometimes Θ is a function space. These are the parametric vs nonparametric regimes.

Prototypical Statistical Inference Tasks

1. **Estimation**. Given realisation $(Y_1, \dots, Y_n)^\top$ from $F(y_1, ..., y_n; \theta)$, how can we produce an educated guess for the unknown true parameter θ?

2. **Hypothesis Testing**. Given two disjoint regions Θ_0 and Θ_1, which is more plausible to contain the true θ that generated our observation $(Y_1, \dots, Y_n)^\top$?

3. **Confidence Intervals**. Instead of estimating a unique θ that may have generated $(Y_1, \dots, Y_n)^\top$, how can we give a whole range of θ that are plausible on the basis of $(Y_1, \dots, Y_n)^\top$?

Additional tasks that can be formulated as versions/extensions of the above are:

- **Prediction**. Given data $(Y_1, \dots, Y_n)^\top$ from a distribution $F(y_1, ..., y_n; \theta)$ where θ is unknown, predict a future outcome from the same distribution.

- **Classification**. Given observations $(Y_1^{(i)}, \dots, Y_n^{(i)})^\top$ from various distributions $F(y_1, ..., y_n; \theta_i)$ (where $i = 1, ..., k$) depending on unknown parameters, and given a new observation Y, declare which of these distribution generated the observation Y.

Can distinguish between two broad types of inference settings:

1. **Marginal Inference.** Here $(Y_1, ..., Y_n)^\top$ has i.i.d. entries each from the same distribution $F(y; \theta)$ with the same parameter θ.
   - In other words, all observations were obtained under identical experimental conditions, and thus depend in the same way on the same unknown θ.
2. **Regression.** Here $(Y_1, ..., Y_n)^\top$ has independent entries, each with distribution $F(y; \theta_i)$ of the same family but with different parameters.
   - Each observation was generated under slightly different experimental conditions. They depend in a similar way on different θ_i.
   - These θ_i correspond to different experimental conditions, say x_i.
   - Each x_i is called a covariate/feature, and is an input that the experimenter can vary. They are known (non-random). The index i reminds us that it corresponds to the ith observation Y_i.
   - Usually θ_i is postulated to have a special relationship to x_i, for example $\theta_i = \exp\{\alpha + \beta x_i\}$, for (α, β) uknown parameters.
   - The goal, then, is to understand how the distribution of Y depends on covariates/features x.

Example

1. **Marginal Inference.** Here we have independent realisations $(Y_1, ..., Y_n)$ each from the same distribution F_θ.

   - For instance, Y_i represents the outcome of flipping the same coin independently, and we wish to understand the probability of success.
   - Then we can model $Y_1, \dots, Y_n \overset{iid}{\sim}$Bernoulli$(\theta)$.

2. **Regression.** Here we have independent realisations $(Y_1, ..., Y_n)$ each from a distribution F_{θ_i} of the same family but with different parameters.

   - For instance, Y_i represents the voting intention of the ith voter on a referendum. The corresponding feature may be his/her income level x_i.
   - We may model the probability of the ith voter voting 1 as Bernoulli(θ_i) with

     $$\theta_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

     for (α, β) a pair of unknown parameters.
   - Note the parsimony: even though each θ_i is different, there are not n unknown parameters but only 2 unknown parameters α, β.

Introduction to probability

Algebra of Events

Random experiment: process whose outcome is uncertain.
Outcomes and any statement involving them must be expressed via *set theory*.

- A possible outcome ω of a random experiment is called an elementary event.
- The set of all possible outcomes, Ω, is assumed non-empty ($\Omega \neq \emptyset$).
- An event is a subset $A \subset \Omega$ of Ω. An event A "is realised" (or "occurs") whenever the outcome of the experiment is an element of A.
- The union of two events F_1 and F_2, written $F_1 \cup F_2$ occurs if and only if either of F_1 or F_2 occurs. Equivalently, $\omega \in F_1 \cup F_2$ if and only if $\omega \in F_1$ or $\omega \in F_2$,

$$F_1 \cup F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ or } \omega \in F_2\}$$

- The intersection of two events F_1 and F_2, written $F_1 \cap F_2$ occurs if and only both F_1 and F_2 occur. Equivalently, $\omega \in F_1 \cap F_2$ if and only if $\omega \in F_1$ and $\omega \in F_2$,

$$F_1 \cap F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ and } \omega \in F_2\}$$

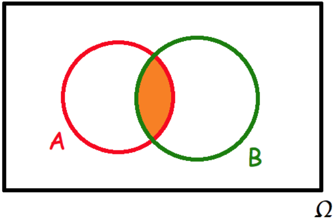
- Unions and intersections of several events, $F_1 \cup \ldots \cup F_n$ and $F_1 \cap \ldots \cap F_n$ are defined iteratively from the definition for unions and intersections of pairs.

- The complement of an event F, denoted F^c, contains all the elements of Ω that are not contained in F,

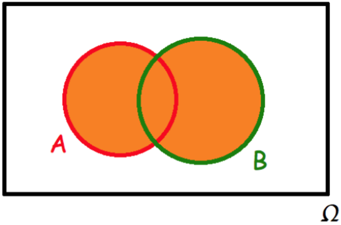
$$F^c = \{\omega \in \Omega : \omega \notin F\}.$$

- Two events F_1 and F_2 are called disjoint if the contain no common elements, that is $F_1 \cap F_2 = \emptyset$.
- A partition $\{F_n\}_{n\geq 1}$ of Ω is a collection of events such that $F_i \cap F_j = \emptyset$ for all $i \neq j$, and $\cup_{n\geq 1} F_n = \Omega$.
- The difference of two events F_1 and F_2 is defined as $F_1 \setminus F_2 = F_1 \cap F_2^c$. It contains all the elements of F_1 that are not contained in F_2. Notice that the difference is not symmetric: $F_1 \setminus F_2 \neq F_2 \setminus F_1$.
- It can be checked that the following properties hold true
  - (i) $(F_1 \cup F_2) \cup F_3 = F_1 \cup (F_2 \cup F_3) = F_1 \cup F_2 \cup F_3$
  - (ii) $(F_1 \cap F_2) \cap F_3 = F_1 \cap (F_2 \cap F_3) = F_1 \cap F_2 \cap F_3$
  - (iii) $F_1 \cap (F_2 \cup F_3) = (F_1 \cap F_2) \cup (F_1 \cap F_3)$
  - (iv) $F_1 \cup (F_2 \cap F_3) = (F_1 \cup F_2) \cap (F_1 \cup F_3)$
  - (v) $(F_1 \cup F_2)^c = F_1^c \cap F_2^c$ and $(F_1 \cap F_2)^c = F_1^c \cup F_2^c$

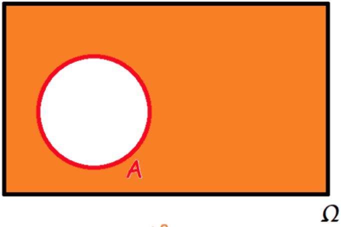
Venn Diagrams



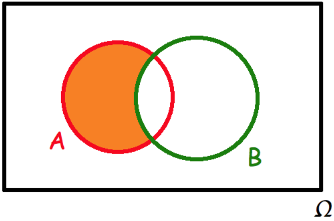
$A \cap B$



$A \cup B$



A^c



$A \setminus B$

Probability measure $\mathbb{P}$: real function defined over the events of Ω, assigning a probability to any event.

Interpreted as a measure of how certain we are that the event will occur.

Postulated to satisfy the following properties (known as axioms of probability):

1. $\mathbb{P}(F) \geq 0$, for all events F.
2. $\mathbb{P}(\Omega) = 1$.
3. If an event F is a countable union $F = \cup_{n \geq 1} F_n$ of disjoint events $\{F_n\}_{n \geq 1}$,

$$\mathbb{P}\left(F\right) = \sum_{n \geq 1} \mathbb{P}(F_n).$$

The following properties are immediate consequences of the probability axioms:

- $\mathbb{P}(F^c) = 1 - \mathbb{P}(F)$.
- $\mathbb{P}(F_1 \cap F_2) \leq \min\{\mathbb{P}(F_1), \mathbb{P}(F_2)\}$.
- $\mathbb{P}(F_1 \cup F_2) = \mathbb{P}(F_1) + \mathbb{P}(F_2) - \mathbb{P}(F_1 \cap F_2)$.
- Continuity from below: let $\{F_n\}_{n\geq 1}$ be nested events, such that $F_j \subseteq F_{j+1}$ for all j, and let $F = \cup_{n\geq 1} F_n$. Then $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- Continuity from above: let $\{F_n\}_{n\geq 1}$ be nested events, such that $F_j \supseteq F_{j+1}$ for all j, and let $F = \cap_{n\geq 1} F_n$. Then $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- If $\Omega = \{\omega_1, ..., \omega_K\}$, $K < \infty$, is a finite set, then for any event $F \subseteq \Omega$, we have $\mathbb{P}(F) = \sum_{j:\omega_j \in F} \mathbb{P}(\omega_j)$.

Conditional Probability and Independence

Suppose we don't know the precise outcome $\omega \in \Omega$ that has occurred, but we are told that $\omega \in F_2$ for some event F_2, and are asked to now calculate the probability that $\omega \in F_1$ also, for some other event F_1.

- For any pair of events F_1, F_2 such that $\mathbb{P}(F_2) > 0$, we define the conditional probability of F_1 given F_2 to be

$$\mathbb{P}(F_1|F_2) = \frac{\mathbb{P}(F_1 \cap F_2)}{\mathbb{P}(F_2)}.$$

- Let G be an event and $\{F_n\}_{n\geq 1}$ be a partition of Ω such that $\mathbb{P}(F_n) > 0$ for all n. We then have:
  - Law of total probability: $\mathbb{P}(G) = \sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)$
  - Bayes' theorem: $\mathbb{P}(F_j|G) = \frac{\mathbb{P}(F_j \cap G)}{\mathbb{P}(G)} = \frac{\mathbb{P}(G|F_j)\mathbb{P}(F_j)}{\sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)}$
- The events $\{G_n\}_{n\geq 1}$ are called independent if and only if for any finite sub-collection $\{G_{i_1}, \ldots, G_{i_K}\}$, $K < \infty$, we have:

$$\mathbb{P}(G_{i_1} \cap \cdots \cap G_{i_K}) = \mathbb{P}(G_{i_1}) \times \mathbb{P}(G_{i_2}) \times \ldots \times \mathbb{P}(G_{i_K})$$

Random variables: numerical summaries of the outcome of a random experiment.

They allow us to not worry too much about precise structure of outcome $\omega \in \Omega$

We can concentrate on range of a random variable, rather than consider Ω.

- A random variable is a real function $X : \Omega \to \mathbb{R}$.
- We write $\{a \leq X \leq b\}$ to denote the event

  $$\{\omega \in \Omega : a \leq X(\omega) \leq b\}.$$

  More generally, if $A \subset \mathbb{R}$ is a generic subset, we write $\{X \in A\}$ to denote the event

  $$\{\omega \in \Omega : X(\omega) \in A\}.$$

- If we have a probability measure defined on the events of Ω, then X induces a new probability measure on subsets of the real line. This is described by the distribution function (or cumulative distribution function) $F_X : \mathbb{R} \to [0, 1]$ of a random variable X (or the law of X),

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

- By its definition, a distribution function satisfies the following properties:

  (i) $x \leq y \;\Rightarrow\; F_X(x) \leq F_X(y)$

  (ii) $\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$

  (iii) $\lim_{y\downarrow x} F_X(y) = F_X(x)$, that is, F_X is right-continuous.

  (iv) $\lim_{y\uparrow x} F_X(y)$ exists, that is, F_X is left-limited.

  (v) $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.

  (vi) $\mathbb{P}(X > a) = 1 - F(a)$.

  (vii) Let $D_X := \{x \in \mathbb{R} : F_X(x) - \lim_{y\uparrow x} F_X(y) > 0\}$ be the set of points where F_X is not continuous.

    - D_X is a countable set.
    - If $\mathbb{P}(\{X \in D_F\}) = 1$ then X is called a *discrete* random variable (equivalently, X has a finite or countable range, with probability 1).
    - If $D_X = \emptyset$ then X is called a *continuous* random variable (the distribution function F_X is continuous).
    - It may very well happen that a random variable may be neither discrete nor continuous.

Given a probability $\alpha \in (0, 1)$, which is the (smallest) real number x such that $\mathbb{P}[X \leq x] = \alpha$?

Let X be a random variable and F_X be its distribution function. We define the quantile function of X (or equivalently of F_X) to be the function

$$F_X^- : (0, 1) \to \mathbb{R}$$

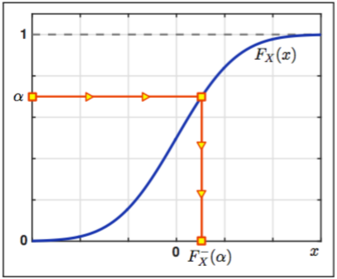
$$F_X^-(\alpha) = \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}.$$

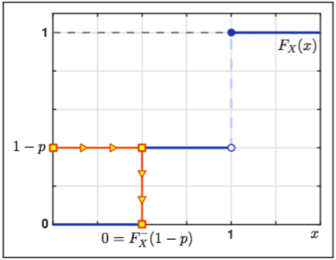
- If F_X is strictly increasing and continuous, then $F_X^- = F_X^{-1}$

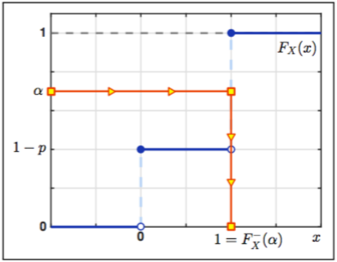
Given an $\alpha \in (0, 1)$, the α-quantile of X (or equivalently of F_X) is the real number

$$q_\alpha = F_X^-(\alpha).$$

The pictures that say it all:







Lemma

Let $Y \sim Unif(0,1)$ and let F be a distribution function. Then, the distribution function of the random variable $X = F^{-}(Y)$ is given precisely by F.

- Can be used to generate realisations from any distribution
  - *Provided we can generate realsations from uniform on* $[0,1]$.
  - *Can do this with binary expansions and Bernoulli draws.*
  - *Reduces problem to infinite coin flipping.*

Partial Converse

Let X be a random variable with strictly increasing and continuous distribution function F_X. Then, $F_X(X) \sim Unif(0,1)$

The probability mas function (or frequency function) $f_X : \mathbb{R} \to [0,1]$ of a discrete random variable X is defined as

$$f_X(x) = \mathbb{P}(X = x)$$

and the set $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$ is the support of X.
By its definition, a probability mass function satisfies

(i) $\mathbb{P}(X \in A) = \sum_{t \in A \cap \mathcal{X}} f_X(t)$, for $A \subseteq \mathbb{R}$.

(ii) $F_X(x) = \sum_{t \in (-\infty, x] \cap \mathcal{X}} f_X(t)$, for all $x \in \mathbb{R}$.

(iii) An immediate corollary is that $F_X(x)$ is piecewise constant with jumps at the points in $\mathcal{X}$.

A continuous random variable X has probability density function $f_X : \mathbb{R} \to [0, +\infty)$ if

$$F_X(b) - F_X(a) = \int_a^b f_X(t)dt.$$

for all real numbers $a < b$. By its definition, a probability density satisfies

(i) $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$

(ii) $f_X(x) = F_X'(x)$, whenever f_X is continuous at x.

(iii) Note that $f_X(x) \neq \mathbb{P}(X = x) = 0$. In fact, it can be $f(x) > 1$ for some x. It can even happen that f is unbounded.

Transformed Mass Functions

Let X be discrete, taking values in $\mathcal{X}$, and define $Y = g(X)$. Then, Y takes values in $\mathcal{Y} = g(\mathcal{X})$ and

$$\begin{aligned} F_Y(y) = \mathbb{P}[g(X) \leq y] &= \sum_{x \in \mathcal{X}} f_X(x) \mathbf{1}\{g(x) \leq y\}, \qquad \forall y \in \mathcal{Y} \\ f_Y(y) = \mathbb{P}[g(X) = y] &= \sum_{x \in \mathcal{X}} f_X(x) \mathbf{1}\{g(x) = y\}, \qquad \forall y \in \mathcal{Y}. \end{aligned}$$

Let X be continuous, taking values in $\mathcal{X} \subseteq \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ a transformation that is

1. monotone,
2. continuously differentiable,
3. with non-vanishing derivative.

If $Y = g(X)$, then Y takes values in $\mathcal{Y} = g(\mathcal{X})$ and

$$f_Y(y) = \left| \frac{\partial}{\partial y} g^{-1}(y) \right| f_X(g^{-1}(y)), \qquad y \in \mathcal{Y}.$$

Random Vectors and Joint Distributions

A random vector $\mathbf{X} = (X_1, \dots, X_d)^\top$ is a finite collection of random variables (arranged as the coordinates of a vector)

We may want to make probabilistic statements on the joint behaviour of all these random variables.

- The joint distribution function of a random vector $\boldsymbol{X} = (X_1, \dots, X_d)^\top$ is defined as:

$$F_{\boldsymbol{X}}(x_1, \dots, x_d) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d).$$

- Correspondingly, one defines the
  - joint frequency function, if the $\{X_i\}_{i=1}^d$ are all discrete,

$$f_{\boldsymbol{X}}(x_1, \dots, x_d) = \mathbb{P}(X_1 = x_1, \dots, X_d = x_d).$$

  - the joint density function, if there exists $f_{\boldsymbol{X}} : \mathbb{R}^d \to [0, +\infty)$ such that:

$$F_{\boldsymbol{X}}(x_1, \dots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_{\boldsymbol{X}}(u_1, \dots, u_d) du_1 \dots du_d$$

In this case, when $f_{\boldsymbol{X}}$ is continuous at the point $\boldsymbol{x}$,

$$f_{\boldsymbol{X}}(x_1, \dots, x_d) = \frac{\partial^d}{\partial x_1 \dots \partial x_d} F_{\boldsymbol{X}}(x_1, \dots, x_d)$$

Marginal Distributions

Given the joint distribution of the random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$, we can isolate the distribution of a single coordinate, say X_i.

- discrete case, the marginal frequency function of X_i is given by

$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_d} f_{\boldsymbol{X}}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)$$

- In the continuous case, the marginal density function of X_i is given by

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_d) dy_1 \ldots dy_{i-1} dy_{i+1} dy_d .$$

- More generally, we can define the joint frequency/density of a random vector formed by a subset of the coordinates of $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$, say the first k
  - Discrete case: $f_{X_1,\ldots,X_k}(x_1, ..., x_k) = \sum_{x_{k+1}} \cdots \sum_{x_d} f_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)$.
  - Continuous case
    $f_{X_1,\ldots,X_k}(x_1, ..., x_k) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d) dx_{k+1} \ldots dx_d$.
- i.e. to marginalise we integrate/sum out the remaining random variables from the overall joint density/frequency.
- Marginals do not uniquely determine the joint distribution.

Conditional Distributions

We may wish to make probabilistic statements about the potential outcomes of one random variable, if we already know the outcome of another.

For this we need the notion of a conditional density/frequency function.

If $(X_1, ..., X_d)$ is a continuous/discrete random vector, we define the conditional probability density/frequency function of $(X_1, ..., X_k)$ given $\{X_{k+1} = x_{k+1}, ..., X_d = x_d\}$ as

$$f_{X_1,\ldots,X_k|X_{k+1},\ldots,X_d}(x_1, ..., x_k|x_{k+1}, ..., x_d) = \frac{f_{X_1,\ldots,X_d}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)}{f_{X_{k+1},\ldots,X_d}(x_{k+1}, ..., x_d)}$$

provided that $f_{X_{k+1},\ldots,X_d}(x_{k+1}, ..., x_d) > 0$.

Independent Random Variables

The random variables $X_1, \ldots, X_d$ are called independent if and only if for all $x_1, \ldots, x_d \in \mathbb{R}$

$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = F_{X_1}(x_1) \times \ldots \times F_{X_d}(x_d).$$

Equivalently, $X_1, \ldots, X_d$ are independent if and only if, for all $x_1, \ldots, x_d \in \mathbb{R}$

$$f_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = f_{X_1}(x_1) \times \ldots \times f_{X_d}(x_d).$$

For two random variables X and Y, we denote their independence as $X \perp\!\!\!\perp Y$.

Note that when random variables are independent, conditional distributions reduce to the corresponding marginal distributions.

Knowing the value of one of the random variables gives us no information about the distribution of the rest.

Conditionally Independent Random Variables

The random vector X in $\mathbb{R}^d$ is called conditionally independent of the random vector Y given the random vector Z, written

$$X \perp\!\!\!\perp_Z Y \quad \text{or} \quad X \perp\!\!\!\perp Y \,|\, Z,$$

if and only if, for all $x_1, \dots, x_d \in \mathbb{R}$

$$F_{X_1,\dots,X_d|Y,Z}(x_1, \dots, x_d) = F_{X_1,\dots,X_d|Z}(x_1, \dots, x_d).$$

Equivalently, if and only if, for all $x_1, \dots, x_d \in \mathbb{R}$

$$f_{X_1,\dots,X_d|Y,Z}(x_1, \dots, x_d) = f_{X_1,\dots,X_d|Z}(x_1, \dots, x_d).$$

Knowing Y in addition to knowing Z gives us no more information about X.

Consequence: if X is conditionally independent of Y given Z, then

$$F_{X,Y|Z} = F_{X|Y,Z}F_{Y|Z} = F_{X|Z}F_{Y|Z}$$

Consequence: $X \perp\!\!\!\perp_Z Y \iff Y \perp\!\!\!\perp_Z X$

Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable bijection,

$$g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \dots, g_n(\boldsymbol{x})), \qquad \boldsymbol{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n.$$

Let $X = (X_1, \dots, X_n)^\top$ have joint density $f_{\boldsymbol{X}}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$, and define $\boldsymbol{Y} = (Y_1, \dots, Y_n)^\top = g(\boldsymbol{X})$. Then, $\boldsymbol{Y}$ takes values in $\mathcal{Y}^n = g(\mathcal{X}^n)$, and

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y}))\left| \det \left[J_{g^{-1}}(\boldsymbol{y}) \right] \right|, \qquad \text{for } \boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathcal{Y}^n,$$

and zero otherwise, whenever $J_{g^{-1}}(\boldsymbol{y})$ is well-defined. Here, $J_{g^{-1}}(\boldsymbol{y})$ is the Jacobian of g^{-1}, i.e. the $n \times n$ matrix-valued function,

$$J_{g^{-1}}(\boldsymbol{y}) = \begin{bmatrix} \frac{\partial}{\partial y_1} g_1^{-1}(\boldsymbol{y}) & \dots & \frac{\partial}{\partial y_n} g_1^{-1}(\boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial y_1} g_n^{-1}(\boldsymbol{y}) & \dots & \frac{\partial}{\partial y_n} g_n^{-1}(\boldsymbol{y}) \end{bmatrix}.$$

Example (Convolution of densities)

Let X and Y be independent, continuous random variables with densities f_X and f_Y. The density of $X + Y$ is the *convolution* of f_X with f_Y:

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u-v) f_Y(v) dv.$$

Define $g : \mathbb{R}^2 \to \mathbb{R}^2$, $\quad (x, y) \overset{g}{\mapsto} (x + y, y) \quad (u, v) \overset{g^{-1}}{\mapsto} (u - v, v)$.
The Jacobian of the inverse is

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

and its determinant is 1. It follows that

$$f_{X+Y,Y}(u, v) = f_{X,Y}(u - v, v) = f_X(u - v) f_Y(v),$$

and we integrate out v to find the marginal f_{X+Y}:

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u-v) f_Y(v) dv.$$

Moments

Expectation

The expectation (or expected value) of a random variable X formalises the notion of the "average" value taken by that random variable.

- For continuous variables:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x) dx.$$

- For discrete variables:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, f_X(x), \qquad \mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

The expectation satisfies the following properties:

- Linearity: $\mathbb{E}[X_1 + \alpha X_2] = \mathbb{E}[X_1] + \alpha\, \mathbb{E}[X_2]$.
- $\mathbb{E}[h(X)] = \sum_{x \in \mathcal{X}} h(x) f_X(x)$ (discrete case)
  or
  $\mathbb{E}[h(x)] = \int_{-\infty}^{+\infty} h(x) f(x) dx$ (continuous case).

Let $\boldsymbol{X} = (X_1, \dots, X_d)^\top$ be a random vector in $\mathbb{R}^d$ with joint density function $f_{\boldsymbol{X}}(x_1, \dots, x_d)$. For any $g : \mathbb{R}^d \to \mathbb{R}$, we define

$$\mathbb{E}\left\{g(X_1, \dots, X_d)\right\} = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g(x_1, \dots, x_d) f_{\boldsymbol{X}}(x_1, \dots, x_d) dx_1 \dots dx_d.$$

Similarly, in the discrete case,

$$\mathbb{E}\left\{g(X_1, \dots, X_d)\right\} = \sum_{x_1 \in \mathcal{X}_1} \dots \sum_{x_d \in \mathcal{X}_d} g(x_1, \dots, x_d) f_{\boldsymbol{X}}(x_1, \dots, x_d).$$

The mean vector or a random vector $\boldsymbol{X} = (X_1, \dots, X_d)$ is defined as

$$\mathbb{E}[\boldsymbol{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

i.e. it is the vector of means.

The variance of a random variable X expresses how the realisations of X are spread around its expectation.

$$\mathsf{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \qquad (\text{if } \mathbb{E}[X^2] < \infty).$$

Furthermore, the covariance of a random variable X_1 with another random variable X_2 expresses the degree of linear dependency between the two.

$$\mathrm{cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] \qquad (\text{if } \mathbb{E}[X_i^2] < \infty).$$

The correlation between X_1 and X_2 is defined as

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathsf{Var}(X_1)\,\mathsf{Var}(X_2)}}.$$

Conveys equivalent dependence information to covariance. Advantages: (1) it is invariant to changes of scale, (2) can be be understood in absolute terms (ranges in $[-1,1]$), as a result of the correlation inequality (itself a consequence of the Cauchy-Schwarz inequality):

$$|\mathrm{Corr}(X_1, X_2)| \leq \sqrt{\mathsf{Var}(X_1)\,\mathsf{Var}(X_2)}.$$

Some useful formulae relating expectations, variance, and covariances are:

- $\mathsf{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathrm{cov}(X, X)$
- $\mathsf{Var}(aX + b) = a^2 \, \mathsf{Var}(X)$
- $\mathsf{Var}(\sum_i X_i) = \sum_i \mathsf{Var}(X_i) + \sum_{i \neq j} \mathrm{cov}(X_i, X_j)$
- $\mathrm{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \, \mathbb{E}[X_2]$
- $\mathrm{cov}(aX_1 + bX_2, Y) = a\mathrm{cov}(X_1, Y) + b\mathrm{cov}(X_2, Y)$
- if $\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] < \infty$, then the following are equivalent:
  (i) $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \, \mathbb{E}[X_2]$
  (ii) $\mathrm{cov}(X_1, X_2) = 0$
  (iii) $\mathsf{Var}(X_1 \pm X_2) = \mathsf{Var}(X_1) + \mathsf{Var}(X_2)$

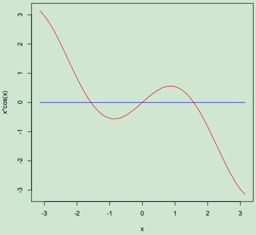
  Independence will imply these three last properties, but none of these properties imply independence.

Example ($\mathrm{Corr}(X, Y) = 0 \not\Rightarrow$ Independence)

Let $X \sim \mathrm{Unif}[-\pi, \pi]$ and define

$$Y = \cos(X).$$

- Clearly X and Y are not independent.
- To the contrary, they are perfectly dependent.
- Their covariance is, nevertheless, zero!



The function $x\ cos(x)$

Concretely, we calculate

$$\mathbb{P}[Y > 0] = 1/2 \qquad \text{and} \qquad \mathbb{P}[Y > 0 | X \in (-\pi, -2)] = 1.$$

Despite this, we have

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \int_{-\pi}^{+\pi} x \cos(x) \frac{1}{2\pi} dx - 0 = 0.$$

Why: Because some non-linear dependencies cannot be detected by covariance...

Example ($\mathrm{Corr}(X,Y)=0 \not\Rightarrow$ Independence)

Let X and Y have joint density

$$f_{XY}(x,y)=\begin{cases}1/\pi & \text{if } x^2+y^2\le 1,\\ 0 & \text{otherwise.}\end{cases}$$

Note that $\mathbb{E}[X]=\mathbb{E}[Y]=0$ by symmetry. Hence, $\mathrm{Cov}(X,Y)=\mathbb{E}[XY]$. But

$$\mathbb{E}[XY]=\iint\limits_{x^2+y^2=1} xy\frac{1}{\pi}dxdy=\iint\limits_{x^2+y^2=1,y\ge 0} xy\frac{1}{\pi}dxdy+\iint\limits_{x^2+y^2=1,y<0} xy\frac{1}{\pi}dxdy$$

The two terms are equal, by symmetry. Moreover,

$$\iint\limits_{x^2+y^2=1,y\ge 0} xy\frac{1}{\pi}dxdy=\frac{1}{\pi}\int_{-1}^{1}x\int_{0}^{1-x^2} ydydx=\frac{1}{\pi}\int_{-1}^{1}x\frac{(1-x^2)^2}{2}dx=0$$

and so the correlation is zero. But X and Y are clearly dependent, since knowing X restricts the possible values of Y.

We can calculate the conditional expectation of a random variable X given that another random variable Y took the value y as

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_{x\in\mathcal{X}} x\,\mathbb{P}[X=x|Y=y], & \text{if } X, Y \text{ are discrete,} \\ \\ \int_{-\infty}^{+\infty} x\, f_{X|Y}(x|y)dx, & \text{if } X, Y \text{ are continuous.} \end{cases}$$

- Precisely the expectation of the conditional distribution.
- Note that $\mathbb{E}[X|Y=y] = q(y)$ results in a function of only y.
- One can plug Y into $q(\cdot)$ and consider $Z = q(Y)$ as a random variable itself.
- Important property/interpretation:

$$\mathbb{E}[X|Y] = \arg\min_{g} \mathbb{E}\left\|X - g(Y)\right\|^2$$

Among all functions[3] of Y, $\mathbb{E}[X|Y]$ best approximates X in mean square.

[3] measurable

Important properties of $\mathbb{E}[X|Y]$:

1. Unbiasedness: $\mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[X]$
2. If X independent of Y, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
3. $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$ (taking out known factors)
4. "Tower property": $\mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[X]$
5. Linearity: $\mathbb{E}[aX_1 + X_2|Y] = a\mathbb{E}[X_1|Y] + \mathbb{E}[X_2|Y]$.
6. Monotonicity: $X_1 \leq X_2 \implies \mathbb{E}[X_1|Y] \leq \mathbb{E}[X_2|Y]$

The conditional variance of X given Y is defined as

$$\mathrm{var}[X|Y] = \mathbb{E}\Big[\,(X - \mathbb{E}[X|Y])^2\,\Big|\,Y\Big] = \mathbb{E}[X^2|Y] - \big(\mathbb{E}[X|Y]\big)^2$$

The law of total variance states that

$$\mathrm{var}(X) = \mathbb{E}\left[\mathrm{var}[X|Y]\right] + \mathrm{var}\left(\mathbb{E}[X|Y]\right)$$

Proof:

$$\begin{aligned}
\mathrm{var}(X) &= \mathbb{E}[X^2] - \mathbb{E}^2[X] \\
&= \mathbb{E}\big[\mathbb{E}[X^2|Y]\big] - \mathbb{E}^2\big[\mathbb{E}[X|Y]\big] \\
&= \mathbb{E}\Big[\mathrm{var}[X|Y] + \mathbb{E}^2[X|Y]\Big] - \mathbb{E}^2\big[\mathbb{E}[X|Y]\big] \\
&= \mathbb{E}\big[\mathrm{var}[X|Y]\big] + \mathbb{E}\big[\mathbb{E}^2[X|Y]\big] - \mathbb{E}^2\big[\mathbb{E}[X|Y]\big] \\
&= \mathbb{E}\left[\mathrm{var}[X|Y]\right] + \mathrm{var}\left(\mathbb{E}[X|Y]\right).
\end{aligned}$$

The covariance matrix or a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\top$, say $\boldsymbol{\Omega} = \{\Omega_{ij}\}$, is a $d \times d$ symmetric matrix with entries

$$\Omega_{ij} = \mathrm{cov}(Y_i, Y_j) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])], \quad 1 \leq i \leq j \leq d.$$

That is, the covariance matrix encodes the variances of the coordinates of Y (on the diagonal) and the pairwise covariances between any two coordinates of $\boldsymbol{Y}$ (off the diagonal).

If we write

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y}] = (\mathbb{E}[Y_1], \ldots, \mathbb{E}[Y_d])^\top$$

for the mean vector of $\boldsymbol{Y}$, then

$$\mathbb{E}[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^\top] = \mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

Similarly to the vector case, the expectation of a matrix with random entries is the matrix of expectations of the random entries.

Let $\boldsymbol{Y}$ be a random $d \times 1$ with mean vector $\boldsymbol{\mu}$ be the mean vector and covariance matrix $\boldsymbol{\Omega}$.

- PSD: for any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} \geq 0$.
- If $\boldsymbol{A}$ is a $p \times d$ deterministic matrix, the mean vector and covariance matrix of $\boldsymbol{AY}$ are $\boldsymbol{A\mu}$ and $\boldsymbol{A\Omega A}^\top$, respectively.
- If $\boldsymbol{\beta} \in \mathbb{R}^d$ is a deterministic vector, the variance of $\boldsymbol{\beta}^\top \boldsymbol{Y}$ is $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$.
- If $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^d$ are deterministic vectors, the covariance of $\boldsymbol{\beta}^\top \boldsymbol{Y}$ with $\boldsymbol{\gamma}^\top \boldsymbol{Y}$ is $\boldsymbol{\gamma}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$.

Given X be a non-negative random variable. Then, given any $\epsilon > 0$,

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \text{[Markov]}$$

Let X be a random variable with finite mean $\mathbb{E}[X] < \infty$. Then, given any $\epsilon > 0$,

$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\mathrm{var}[X]}{\epsilon^2} \qquad \text{[Chebyschev]}$$

For any convex[4] function $\varphi : \mathbb{R} \to \mathbb{R}$, if $\mathbb{E}|\varphi(X)| + \mathbb{E}|X| < \infty$, then one has

$$\varphi\Big(\mathbb{E}[X]\Big) \leq \mathbb{E}[\varphi(X)] \qquad \text{[Jensen]}$$

Let X be a real random variable with $\mathbb{E}[X^2] < \infty$. Let $g : \mathbb{R} \to \mathbb{R}$ be a non decreasing function such that $\mathbb{E}[g^2(X)] < \infty$. Then,

$$\mathrm{cov}[X, g(X)] \geq 0 \qquad \text{[Monotonicity and Covariance]}$$

[4] Recall that a function φ is convex if $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for all x, y, and $\lambda \in [0, 1]$.

Let X be a random variable taking values in $\mathbb{R}$. The moment generating function (MGF) of X is defined as

$$M_X(t) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$$

$$M_X(t) = \mathbb{E}\left[e^{tX}\right], \qquad t \in \mathbb{R}.$$

When $M_X(t), M_Y(t)$ exist (are finite) for $t \in I \ni 0$, then:

- $\mathbb{E}[|X|^k] < \infty$ and $\mathbb{E}[X^k] = \frac{d^k M_X}{dt^k}(0)$, for all $k \in \mathbb{N}$.
- $M_X = M_Y$ on I if and only if $F_X = F_Y$
- $M_{X+Y} = M_X M_Y$ when X and Y are independent

Similarly, for a random vector $\boldsymbol{X}$ in $\mathbb{R}^d$, the MGF is

$$M_{\boldsymbol{X}}(\boldsymbol{u}) : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$$

$$M_{\boldsymbol{X}}(\boldsymbol{u}) = \mathbb{E}\left[e^{\boldsymbol{u}^\top \boldsymbol{X}}\right], \qquad \boldsymbol{u} \in \mathbb{R}^d.$$

and has analogous properties.