Random Variables

Paolo Zacchia

Probability and Statistics

Lecture 1

Sample Spaces

Statistics deals with uncertain "events." It is thus necessary to formalize notions about the **probability** for such events.

However, one needs first to characterize what are the "events." To do this, mathematical probability borrows from set theory.

Definition 1

Sample Space. The set $\mathbb{S}$ collecting all possible outcomes associated with a certain phenomenon is called the **sample space**.

A sample space provides, for all possible occurrences of a given phenomenon, the finest-grained characterization that is possible.

Sample Spaces: examples

Countable sample spaces (can be enumerated)

- Tossing a coin: $\mathbb{S}_{coin} = \{Head, Tail\}$
- Exam grades: $\mathbb{S}_{exam} = \{A, B, C, D, E, F\}$
- E-mails received: $\mathbb{S}_{emails} = \{0, 1, 2, \dots\} = \mathbb{N}_0$

Uncountable sample spaces (cannot be enumerated)

- Individual income: $\mathbb{S}_{income} = \mathbb{R}_+$
- Individual wealth: $\mathbb{S}_{wealth} = \mathbb{R}$
- Household wealth: $\mathbb{S}_{household} = \mathbb{S}_{wealth.1} \times \mathbb{S}_{wealth.2} = \mathbb{R}^2$ (e.g. if a household is composed of two individuals, 1 and 2)

Events

Any combination of a sample space's elements is an **event**.

Definition 2

Events. A subset of a sample space $\mathbb{S}$, including $\mathbb{S}$ itself, is an **event**.

Examples:

- Coin tossing: $\mathbb{A}_{null} = \varnothing$, $\mathbb{A}_{head} = \{Head\}$, $\mathbb{A}_{tail} = \{Tail\}$, $\mathbb{A}_{full} = \mathbb{S}_{coin} = \{Head, Tail\}$;
- Exam pass/fail: $\mathbb{A}_{passing} = \{A, B, C\}$, $\mathbb{A}_{failing} = \{D, E, F\}$;
- Any bracket – a segment of the (positive) real line – of the income or wealth space.

Set properties of events

Standard set operations such as: **union** ($\cup$), **intersection** ($\cap$) and **complementation** ($\mathbb{A}^c$), extend to events.

Theorem 1

Properties of Events. *Let $\mathbb{A}_S$, $\mathbb{B}_S$ and $\mathbb{C}_S$ be any three events associated with the sample space $\mathbb{S}$. The following properties hold.*

a. *Commutativity:*

$$\mathbb{A}_S \cup \mathbb{B}_S = \mathbb{B}_S \cup \mathbb{A}_S$$
$$\mathbb{A}_S \cap \mathbb{B}_S = \mathbb{B}_S \cap \mathbb{A}_S$$

b. *Associativity:*

$$\mathbb{A}_S \cup (\mathbb{B}_S \cup \mathbb{C}_S) = (\mathbb{A}_S \cup \mathbb{B}_S) \cup \mathbb{C}_S$$
$$\mathbb{A}_S \cap (\mathbb{B}_S \cap \mathbb{C}_S) = (\mathbb{A}_S \cap \mathbb{B}_S) \cap \mathbb{C}_S$$

c. *Distributive Laws:*

$$\mathbb{A}_S \cap (\mathbb{B}_S \cup \mathbb{C}_S) = (\mathbb{A}_S \cap \mathbb{B}_S) \cup (\mathbb{A}_S \cap \mathbb{C}_S)$$
$$\mathbb{A}_S \cup (\mathbb{B}_S \cap \mathbb{C}_S) = (\mathbb{A}_S \cup \mathbb{B}_S) \cap (\mathbb{A}_S \cup \mathbb{C}_S)$$

d. *DeMorgan's Laws:*

$$(\mathbb{A}_S \cup \mathbb{B}_S)^c = \mathbb{A}_S^c \cap \mathbb{B}_S^c$$
$$(\mathbb{A}_S \cap \mathbb{B}_S)^c = \mathbb{A}_S^c \cup \mathbb{B}_S^c$$

Partitions

A sample space can be split between events that cannot happen at the same time.

Definition 3

Disjoint Events. Two events $\mathbb{A}_1$ and $\mathbb{A}_2$ are **disjoint** or **mutually exclusive** if $\mathbb{A}_1 \cap \mathbb{A}_2 = \varnothing$. The events in a collection $\mathbb{A}_1, \mathbb{A}_2, \ldots$ are **pairwise disjoint** or **mutually exclusive** if $\mathbb{A}_i \cap \mathbb{A}_j = \varnothing$ for all pairs $i \neq j$.

Definition 4

Partition. The events in a collection $\mathbb{A}_1, \mathbb{A}_2, \ldots$ form a **partition** of the sample space $\mathbb{S}$ if they are pairwise disjoint and $\cup_{i=1}^{Z} \mathbb{A}_i = \mathbb{S}$ if the collection is of finite dimension Z; $\cup_{i=1}^{\infty} \mathbb{A}_i = \mathbb{S}$ if the collection has an infinite number of elements.

Examples: $\mathbb{A}_{passing}$ and $\mathbb{A}_{failing}$, all brackets of a tax system.

Sigma Algebra

The following concept is central in the axiomatic definition of **probability functions**.

Definition 5

Sigma Algebra. Given some set $\mathbb{S}$, a **sigma algebra** (σ**-algebra**) or **Borel field** is a collection of subsets of $\mathbb{S}$, which is denoted as $\mathcal{B}$, that satisfies the following properties:

a. $\varnothing \in \mathcal{B}$;

b. for any subset $\mathbb{A} \in \mathcal{B}$, it is $\mathbb{A}^c \in \mathcal{B}$;

c. for any *countable* sequence of subsets $\mathbb{A}_1, \mathbb{A}_2, \cdots \in \mathcal{B}$, it holds that $\cup_{i=1}^{\infty} \mathbb{A}_i \in \mathcal{B}$.

Note that properties **b.** and **c.** together with DeMorgan's Law also imply $\cap_{i=1}^{\infty} \mathbb{A}_i \in \mathcal{B}$ for any appropriate countable sequence of subsets.

Sigma Algebra: examples

Countable sample spaces (can be enumerated)

- Trivial σ-algebrae: $\mathcal{B} = \{\varnothing, \mathbb{S}\}$
- Any partition of $\mathbb{S}$, and all possible unions of its elements

Uncountable sample spaces (cannot be enumerated)

- If $\mathbb{S} = \mathbb{R}$: $\mathcal{B}$ would contain all segments of the kind

$$[a, b], \quad (a, b], \quad [a, b), \quad (a, b)$$

for any two $a, b \in \mathbb{R}$ with $a \leq b$, plus all their unions and intersections.

- If $\mathbb{S} = \mathbb{R}^K$ with $K > 1$, similarly constructed *connected* sets.

Sigma Algebra: counterexample

- Suppose that $\mathcal{B}'$ contains all the finite disjoint unions of sets of the following form, for any two $a, b \in \mathbb{R}$ with $a \leq b$.

$$(-\infty, a], \quad (a, b], \quad (b, \infty), \quad \varnothing, \quad \mathbb{R}$$

- Note that this is based off simple partitions of $\mathbb{R}$.

- However, $\cup_{i=1}^{\infty} \left(0, \frac{i-1}{i}\right] = (0, 1) \notin \mathcal{B}'$, which contradicts the definition of sigma algebra.

- Problem: one cannot characterize events that are obtained as the union of more primitive events in $\mathcal{B}'$!

Probability function

The following *axiomatic* definition is due to A. N. Kolmogorov.

Definition 6

Probability Function. Given a sample space $\mathbb{S}$ and an associated σ-algebra $\mathcal{B}$, a **probability function** $\mathbb{P}$ is a function with domain $\mathcal{B}$ that satisfies the three **axioms of probability**:

a. $\mathbb{P}(\mathbb{A}) \geq 0 \ \forall \mathbb{A} \in \mathcal{B}$;

b. $\mathbb{P}(\mathbb{S}) = 1$;

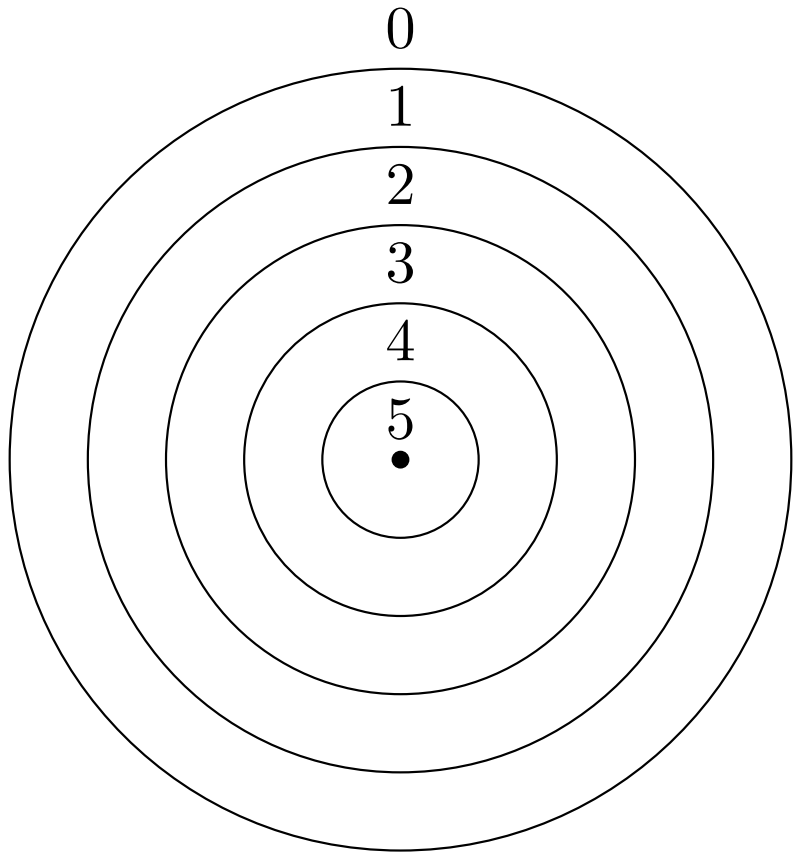
c. given a *countable* sequence of *pairwise disjoint* subsets written as $\mathbb{A}_1, \mathbb{A}_2, \cdots \in \mathcal{B}$, then $\mathbb{P}\left(\cup_{i=1}^{\infty} \mathbb{A}_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A}_i)$.

The construction of probability functions proceeds as follows.

- For countable sample spaces: assign a number $p(s) \geq 0$ to each element $s \in \mathbb{S}$, such that $\sum_{s \in \mathbb{S}} p(s) = 1$.
- For uncountable sample spaces: use proper **measures**.

Example: the naïve dart thrower

Consider dart players who score points depending on how close they get to the center of the dartboard.



Suppose a player hits every point completely at random. What is the probability that he scores a given number of points i?

Probability measures on the dartboard

- A probability function for an uncountable sample space!
- This has to be proportional to the *measure* of the areas of each "ring" in the dartboard – and the outside area too.
- Let the distance of each consecutive "ring" from the center be $(I - i + 1)\, r$, where I is the maximum number of points (say 5) and r the distance between two contiguous rings.
- Suppose that the area outside the dartboard "measures" T. The probability of scoring zero is $\mathbb{P}(0) = T/\left(T + \pi I^2 r^2\right)$.
- The probability of scoring $0 < i \leq I$ points is as follows.

$$\mathbb{P}(i) = \underbrace{\pi r^2 \left[(I + 1 - i)^2 - (I - i)^2\right]}_{= \text{area of the } i\text{-th "ring"}} \times \underbrace{\left(T + \pi I^2 r^2\right)^{-1}}_{= \text{total area}}$$

Properties of probability functions (1/4)

Theorem 2

Properties of Probability Functions (1). *If $\mathbb{P}$ is some probability function and $\mathbb{A}$ is a set in $\mathcal{B}$, the following properties hold:*

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

b. $\mathbb{P}(\mathbb{A}) \leq 1$;

c. $\mathbb{P}(\mathbb{A}^c) = 1 - \mathbb{P}(\mathbb{A})$.

Proof.

The observation that $\mathbb{A}$ and $\mathbb{A}^c$ form a partition of $\mathbb{S}$ and therefore it is $\mathbb{P}(\mathbb{A}) + \mathbb{P}(\mathbb{A}^c) = \mathbb{P}(\mathbb{S}) = 1$ proves **c.** – thus **a.** and **b.** follow. □

Properties of probability functions (2/4)

Theorem 3

Properties of Probability Functions (2). *If $\mathbb{P}$ is some probability function and $\mathbb{A}$, $\mathbb{B}$ are sets in $\mathcal{B}$, the following properties hold:*

a. $\mathbb{P}(\mathbb{B} \cap \mathbb{A}^c) = \mathbb{P}(\mathbb{B}) - \mathbb{P}(\mathbb{A} \cap \mathbb{B})$;

b. $\mathbb{P}(\mathbb{A} \cup \mathbb{B}) = \mathbb{P}(\mathbb{A}) + \mathbb{P}(\mathbb{B}) - \mathbb{P}(\mathbb{A} \cap \mathbb{B})$;

c. *if* $\mathbb{A} \subset \mathbb{B}$, *it is* $\mathbb{P}(\mathbb{A}) \leq \mathbb{P}(\mathbb{B})$.

Proof.

To prove **a.** note that $\mathbb{B}$ can be expressed as the union of two disjoint sets $\mathbb{B} = \{\mathbb{B} \cap \mathbb{A}\} \cup \{\mathbb{B} \cap \mathbb{A}^c\}$, thus $\mathbb{P}(\mathbb{B}) = \mathbb{P}(\mathbb{B} \cap \mathbb{A}) + \mathbb{P}(\mathbb{B} \cap \mathbb{A}^c)$. To show **b.** decompose the union of $\mathbb{A}$ and $\mathbb{B}$ as $\mathbb{A} \cup \mathbb{B} = \mathbb{A} \cup \{\mathbb{B} \cap \mathbb{A}^c\}$ – again two disjoint sets; hence by **a.** the following holds.

$$\mathbb{P}(\mathbb{A} \cup \mathbb{B}) = \mathbb{P}(\mathbb{A}) + \mathbb{P}(\mathbb{B} \cap \mathbb{A}^c) = \mathbb{P}(\mathbb{A}) + \mathbb{P}(\mathbb{B}) - \mathbb{P}(\mathbb{A} \cap \mathbb{B})$$

Finally, **c.** follows from **a.** as $\mathbb{A} \subset \mathbb{B}$ implies that $\mathbb{P}(\mathbb{A} \cap \mathbb{B}) = \mathbb{P}(\mathbb{A})$, so $\mathbb{P}(\mathbb{B} \cap \mathbb{A}^c) = \mathbb{P}(\mathbb{B}) - \mathbb{P}(\mathbb{A}) \geq 0$. □

Properties of probability functions (3/4)

Theorem 4

Properties of Probability Functions (3). *If $\mathbb{P}$ is some probability function, the following properties hold:*

a. $\mathbb{P}(\mathbb{A}) = \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A} \cap \mathbb{C}_i)$ *for any* $\mathbb{A} \in \mathcal{B}$ *as well as any partition* $\mathbb{C}_1, \mathbb{C}_2, \ldots$ *of the sample space such that* $\mathbb{C}_i \in \mathcal{B}$ *for all* $i \in \mathbb{N}$*;*

b. $\mathbb{P}(\cup_{i=1}^{\infty} \mathbb{A}_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(\mathbb{A}_i)$ *for any sets* $\mathbb{A}_1, \mathbb{A}_2, \ldots$ *such that* $\mathbb{A}_i \in \mathcal{B}$ *for all* $i \in \mathbb{N}$*.*

Proof.

Regarding **a.** note that, by the Distributive Laws of events, it is

$$\mathbb{A} = \mathbb{A} \cap \mathbb{S} = \mathbb{A} \cap \left(\bigcup_{i=1}^{\infty} \mathbb{C}_i \right) = \bigcup_{i=1}^{\infty} (\mathbb{A} \cap \mathbb{C}_i)$$

where the intersection sets of the form $\mathbb{A} \cap \mathbb{C}_i$ are pairwise disjoint as the $\mathbb{C}_i$ sets are. Hence, **a.** follows from the third axiom of probability functions. **(Continues...)**

Properties of probability functions (4/4)

Theorem 4

Proof.

(Continued.) To establish **b.** construct another collection of *pairwise disjoint* events $\mathbb{A}_1^*, \mathbb{A}_2^*, \ldots$ such that $\cup_{i=1}^{\infty} \mathbb{A}_i = \cup_{i=1}^{\infty} \mathbb{A}_i^*$ and

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} \mathbb{A}_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} \mathbb{A}_i^*\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(\mathbb{A}_i^*\right) \leq \sum_{i=1}^{\infty} \mathbb{P}\left(\mathbb{A}_i\right)$$

where the second equality follows from the pairwise disjoint property. Such additional collection of events can be obtained as:

$$\mathbb{A}_1^* = \mathbb{A}_1, \quad \mathbb{A}_i^* = \mathbb{A}_i \cap \left(\bigcup_{j=1}^{i-1} \mathbb{A}_j\right)^c = \mathbb{A}_i \cap \left(\bigcap_{j=1}^{i-1} \mathbb{A}_j^c\right) \text{ for } i = 2, 3, \ldots$$

that are by construction pairwise disjoint and satisfy $\cup_{i=1}^{\infty} \mathbb{A}_i = \cup_{i=1}^{\infty} \mathbb{A}_i^*$. Furthermore, by construction $\mathbb{A}_i^* \subset \mathbb{A}_i$ and thus $\mathbb{P}\left(\mathbb{A}_i^*\right) \leq \mathbb{P}\left(\mathbb{A}_i\right)$ for every i, leading to the inequality $\sum_{i=1}^{\infty} \mathbb{P}\left(\mathbb{A}_i^*\right) \leq \sum_{i=1}^{\infty} \mathbb{P}\left(\mathbb{A}_i\right)$. □

Conditional probability

Definition 7

Conditional Probability. Consider a sample space $\mathbb{S}$, an associated σ-algebra $\mathcal{B}$, and any two events $\mathbb{A}, \mathbb{B} \in \mathcal{B}$ such that $\mathbb{P}(\mathbb{B}) > 0$. The **conditional probability** of $\mathbb{A}$ **given** $\mathbb{B}$ is written as $\mathbb{P}(\mathbb{A}|\mathbb{B})$ and is defined as follows.

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

- Conditional probability is about defining the probability for an event $\mathbb{A}$ when restricting (re-defining) the sample space to a specific event $\mathbb{B}$.

- To help intuition, we speak about "the probability of $\mathbb{A}$ when $\mathbb{B}$ is *fixed*."

Conditional probability: examples (1/3)

- Let us return to the grades example and suppose that:

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 0.3$$
$$\mathbb{P}(D) = \mathbb{P}(F) = 0.05$$
$$\mathbb{P}(E) = 0$$

- ... hence, it is $\mathbb{P}(passing) = 0.9$ and $\mathbb{P}(failing) = 0.1$.

- Note that $A \subset \mathbb{A}_{passing}$, so $\mathbb{P}(A \cap passing) = \mathbb{P}(A)$.

- Hence:

$$\mathbb{P}(A|\, passing) = \frac{\mathbb{P}(A \cap passing)}{\mathbb{P}(passing)} = \frac{0.3}{0.9} = \frac{1}{3}$$

  and similarly $\mathbb{P}(D|\, failing) = \mathbb{P}(F|\, failing) = 0.5$.

Conditional probability: examples (2/3)

- Let us return to the "naïve dart thrower" example.
- Since it is $\mathbb{P}(i > 0) = \pi I^2 r^2 / \left(T + \pi I^2 r^2\right)$, it is:

$$\begin{aligned}\mathbb{P}(i \,|\, i > 0) &= \frac{\mathbb{P}(i \cap i > 0)}{\mathbb{P}(i > 0)} \\ &= I^{-2}\left[(I + 1 - i)^2 - (I - i)^2\right]\end{aligned}$$

  i.e. the probability of scoring i, *conditional* on hitting the dartboard.

- Suppose that the player learns to consistently score *at least* $3 < I$ points: the new probabilities are conditional.

$$\mathbb{P}(i \,|\, i > 2) = \frac{\mathbb{P}(i \cap i > 2)}{\mathbb{P}(i > 2)} = \frac{\left[(I + 1 - i)^2 - (I - i)^2\right]}{(I - 2)^2}$$

Conditional probability: examples (3/3)

- After the partial takeup of a preemptive medical treatment:

$$\begin{aligned}\mathbb{P}(taker \cap healthy) &= 0.40\\ \mathbb{P}(taker \cap sick) &= 0.20\\ \mathbb{P}(hesitant \cap healthy) &= 0.15\\ \mathbb{P}(hesitant \cap sick) &= 0.25\end{aligned}$$

- . . . and therefore, it is $\mathbb{P}(taker) = 0.60$, $\mathbb{P}(hesitant) = 0.40$, $\mathbb{P}(healthy) = 0.55$ and $\mathbb{P}(sick) = 0.45$.
- It is *not* easier to find a sick person among those who took the treatment than among those who hesitated.

$$\mathbb{P}(sick|\, taker) = \frac{\mathbb{P}(taker \cap sick)}{\mathbb{P}(taker)} = \frac{1}{3}$$

$$\mathbb{P}(sick|\, hesitant) = \frac{\mathbb{P}(hesitant \cap sick)}{\mathbb{P}(hesitant)} = \frac{5}{8}$$

Bayes' Rule

- One can recast the definition of conditional probability by swapping $\mathbb{A}$ and $\mathbb{B}$.

$$\mathbb{P}\left(\mathbb{B}\middle|\,\mathbb{A}\right) = \frac{\mathbb{P}\left(\mathbb{B} \cap \mathbb{A}\right)}{\mathbb{P}\left(\mathbb{A}\right)}$$

- Combining the above with the reverse expression delivers the following result (again $\mathbb{A}$ and $\mathbb{B}$ can be swapped).

$$\mathbb{P}\left(\mathbb{A}\middle|\,\mathbb{B}\right) = \frac{\mathbb{P}\left(\mathbb{B}\middle|\,\mathbb{A}\right)\mathbb{P}\left(\mathbb{A}\right)}{\mathbb{P}\left(\mathbb{B}\right)}$$

- This is known as **Bayes' Rule**: a powerful expression that relates conditional probabilities to one another.

Bayes' Theorem

Bayes' Rule extends to multiple pairwise disjoint events.

Theorem 5

Bayes' Theorem. *Let* $\mathbb{A}_1, \mathbb{A}_2, \ldots$ *be a partition of the sample space* $\mathbb{S}$, *and* $\mathbb{B}$ *some event* $\mathbb{B} \subset \mathbb{S}$. *For* $i = 1, 2, \ldots$ *the following holds.*

$$\mathbb{P}\left(\mathbb{A}_i \middle| \mathbb{B}\right) = \frac{\mathbb{P}\left(\mathbb{B} \middle| \mathbb{A}_i\right) \mathbb{P}\left(\mathbb{A}_i\right)}{\sum_{j=1}^{\infty} \mathbb{P}\left(\mathbb{B} \middle| \mathbb{A}_j\right) \mathbb{P}\left(\mathbb{A}_j\right)}$$

Proof.

This follows from Bayes' Rule for $\mathbb{A} = \mathbb{A}_i$ and by observing that:

$$\mathbb{P}\left(\mathbb{B}\right) = \sum_{j=1}^{\infty} \mathbb{P}\left(\mathbb{B} \cap \mathbb{A}_j\right) = \sum_{j=1}^{\infty} \mathbb{P}\left(\mathbb{B} \middle| \mathbb{A}_j\right) \mathbb{P}\left(\mathbb{A}_j\right)$$

from Theorem 4 and the definition of conditional probability. □

Bayes' Rule: example

- Let us return to the previous example about the imperfect takeup of an imperfect preemptive medical treatment.

- A simple application of Bayes' rule is the following.

$$\mathbb{P}\left(taker|\, sick\right) = \mathbb{P}\left(sick|\, taker\right) \frac{\mathbb{P}\left(taker\right)}{\mathbb{P}\left(sick\right)}$$

$$= \frac{4}{9}$$

- Thus, one can calculate the reverse conditional probabilities without knowing the probabilities of intersected events!

Independent events

Definition 8

Statistical independence (*two events*). Two events $\mathbb{A}$ and $\mathbb{B}$ are **statistically independent** if the following holds.

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

- Note: this implies $\mathbb{P}(\mathbb{A}|\,\mathbb{B}) = \mathbb{P}(\mathbb{A})$! Fixing $\mathbb{B}$ does not affect the probability of $\mathbb{A}$.

Definition 9

Mutual statistical independence (*multiple events*). The events of any collection $\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_N$ are **mutually independent** if, for any subcollection $\mathbb{A}_{i_1}, \mathbb{A}_{i_2}, \ldots, \mathbb{A}_{i_{N'}}$ with $N' \leq N$, the following holds.

$$\mathbb{P}\left(\bigcap_{j=1}^{N'} \mathbb{A}_{i_j}\right) = \prod_{j=1}^{N'} \mathbb{P}\left(\mathbb{A}_{i_j}\right)$$

Independence and complementary events

Theorem 6

Independence and Complementary Events. *Consider any two independent events* $\mathbb{A}$ *and* $\mathbb{B}$*. It can be concluded that the following pairs of events are independent too:*

a. $\mathbb{A}$ *and* $\mathbb{B}^c$*;*

b. $\mathbb{A}^c$ *and* $\mathbb{B}$*;*

c. $\mathbb{A}^c$ *and* $\mathbb{B}^c$*.*

Proof.

Case **a.** follows from the definition of independence (second equality):

$$\begin{aligned}\mathbb{P}\left(\mathbb{A} \cap \mathbb{B}^c\right) &= \mathbb{P}\left(\mathbb{A}\right) - \mathbb{P}\left(\mathbb{A} \cap \mathbb{B}\right) \\ &= \mathbb{P}\left(\mathbb{A}\right) - \mathbb{P}\left(\mathbb{A}\right)\mathbb{P}\left(\mathbb{B}\right) \\ &= \mathbb{P}\left(\mathbb{A}\right)\left[1 - \mathbb{P}\left(\mathbb{B}\right)\right] \\ &= \mathbb{P}\left(\mathbb{A}\right)\mathbb{P}\left(\mathbb{B}^c\right)\end{aligned}$$

Cases **b.** and **c.** are analogous. □

Random variables

- Probability functions are defined for generic σ-algebrae but sometimes these can be complicated to work with.

- It is often convenient to re-formulate probability functions so that the function's domain is a subset of $\mathbb{R}$.

Definition 10

Random Variables. A **random variable** X is a function from the the sample space $\mathbb{S}$ onto the set of real numbers $X : \mathbb{S} \to \mathbb{R}$.

- They are denoted with upper case italic letters like X.

- A specific **realization** of a random variable (e.g. the value of X that occurs, or is presumed to occur in the real world) is denoted with a lower case italic letter like x.

Random variables: examples

- X_{coin} is a random variable mapping $\{Head, Tail\} \to \{0, 1\}$ or vice versa, or any other mapping one may find suitable.

- X_{coin} has two realizations.

$$x_{coin} = \begin{cases} 1 & \text{if } Tail \\ 0 & \text{if } Head \end{cases}$$

- X_{grade} is a random variable mapping $\{A, B, C, D, E, F\}$ to, say, numeric scores for GPA calculation.

- X_{emails} maps the sample space $\mathbb{N}_0$ onto itself.

- X_{income} also maps the sample space $\mathbb{R}_+$ onto itself.

- X_{wealth} likewise maps the sample space $\mathbb{R}$ onto itself.

Probability distributions

- Random variables re-formulate the sample space, but what about the probability functions?

Definition 11
Cumulative Probability Distribution. Given a random variable X, a **cumulative (probability) distribution function** (typically abbreviated as **c.d.f.**) is a function $F_X(x)$ which is defined as follows.

$$F_X(x) = \mathbb{P}(X \leq x) \quad \text{for all } x \in \mathbb{R}$$

- Conventionally, the subscript X in F_X associates the c.d.f. to the random variable X.

- The domain of a c.d.f. is typically the map of the union of multiple events in the original sample space.

Example: two coins (1/3)

- Suppose we are tossing not one, but two coins. Thus:

$$\mathbb{S}_{2.coins} = \{Head\,\&\,Head,\ Head\,\&\,Tail,\ Tail\,\&\,Head,\ Tail\,\&\,Tail\}$$

  where & separates the outcomes of first vs. second coin.

- The random variable $X_{2.coins} \in \{0, 1, 2\} \in \mathbb{R}$ here counts the number of "tails."

$$\begin{aligned} X\,(Head\,\&\,Head) &= 0 \\ X\,(Head\,\&\,Tail) &= 1 \\ X\,(Tail\,\&\,Head) &= 1 \\ X\,(Tail\,\&\,Tail) &= 2 \end{aligned}$$

Example: two coins (2/3)

- Suppose that the coins are balanced: at every attempt, the chances to get tail or head are equal.

- Thus all elements in $\mathbb{S}_{2.coins}$ have equal probability!

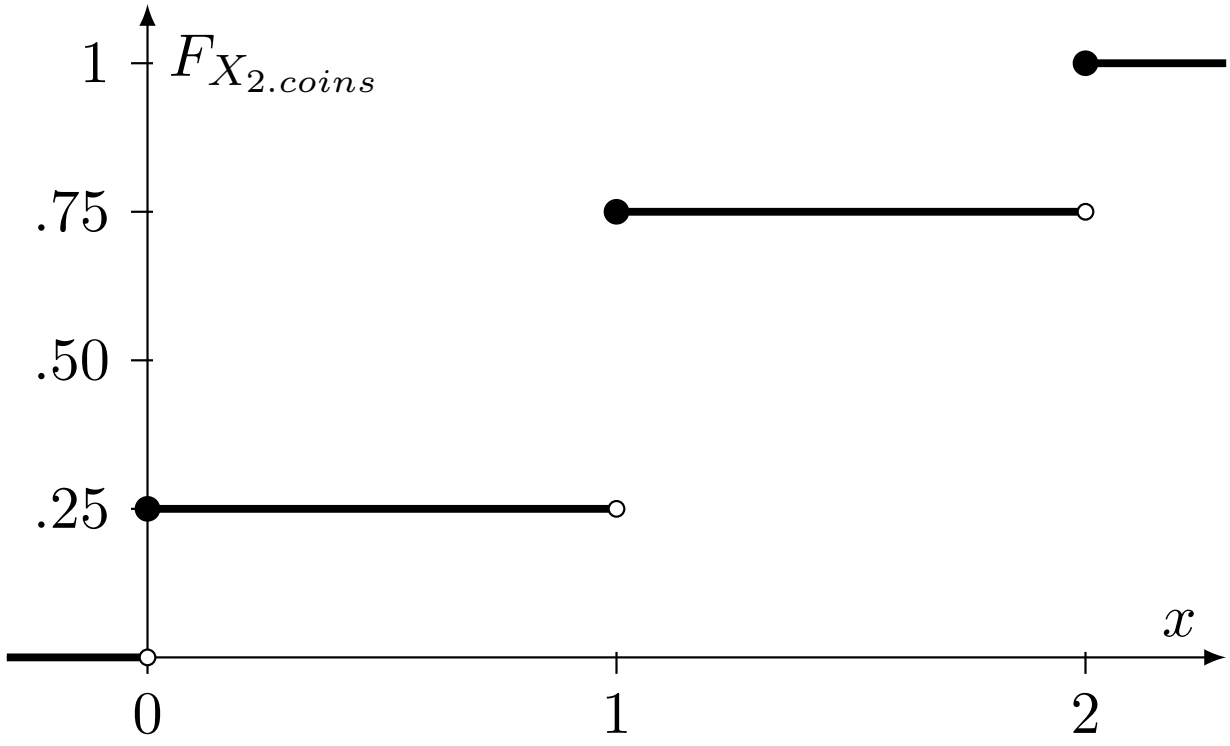
- It follows that $\mathbb{P}(X_{2.coins} = 0) = \mathbb{P}(X_{2.coins} = 2) = 0.25$ and $\mathbb{P}(X_{2.coins} = 1) = 0.50$.

- The associated c.d.f. is the following.

$$F_{X_{2.coins}}(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0) \\ .25 & \text{if } x \in [0, 1) \\ .75 & \text{if } x \in [1, 2) \\ 1 & \text{if } x \in [2, \infty) \end{cases}$$

Example: two coins (3/3)

- This c.d.f. is represented graphically below.
- Observe the "discrete steps" when a new tail count is hit.



Properties of cumulative distributions

Theorem 7

Properties of Probability Distribution Functions. *A function $F(x)$ can be a (cumulative) probability distribution function if and only if the following three conditions hold:*

a. $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$*;*

b. $F(x)$ *is a nondecreasing function of* x*;*

c. $F(x)$ *is right-continuous, that is* $\lim_{x \downarrow x_0} F(x) = F(x_0) \ \forall x_0 \in \mathbb{R}$.

Proof.

(Outline.) Necessity follows directly from the definition of Probability Functions. Sufficiency requires some reverse engineering, showing how for each Probability Distribution Function with the above properties, one can find an appropriate sample space $\mathbb{S}$, an associated probability function $\mathbb{P}$ and a relative random variable X. $\square$

Discrete and continuous random variables

Definition 12

Types of Random Variables. A random variable X is **continuous** if $F_X(x)$ is a continuous function of x, while it is **discrete** if $F_X(x)$ is a step function of x.

Examples:

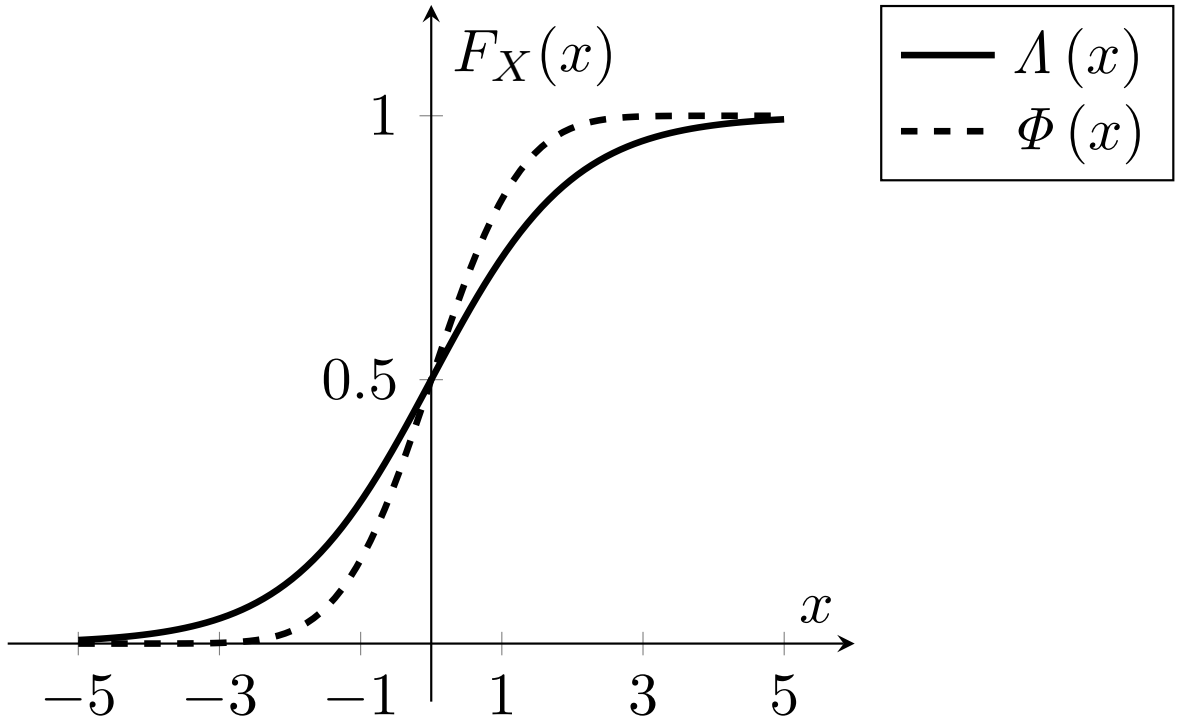
- Discrete: $X_{2.coins}$, X_{grades} and similar ones;
- Continuous: the **standard logistic distribution**;

$$F_X(x) = \Lambda(x) = \frac{1}{1 + \exp(-x)}$$

- Continuous: the **standard normal distribution**.

$$F_X(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

Standard logistic and normal distributions



The standard logistic assigns a higher probability, relative to the standard normal, to realizations that are more distant from zero.

Mass and density functions

Definition 13

Probability Mass Function. Given a *discrete* random variable X, its probability **mass** function $f_X(x)$ (which is often abbreviated as **p.m.f.**) is defined as follows.

$$f_X(x) = \mathbb{P}(X = x) \quad \text{for all } x \in \mathbb{R}$$

Definition 14

Probability Density Function. Given a *continuous* random variable X, its probability **density** function $f_X(x)$ (which is often abbreviated as **p.d.f.**) is defined as the function that satisfies the following relationship.

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \quad \text{for all } x \in \mathbb{R}$$

Note: if a the c.d.f. of a continuous random variable is differentiable everywhere in $\mathbb{R}$, the associated density function is $f_X(t) = \frac{\partial F_X(t)}{\partial x}$.

Support of random variables

Definition 15

Support of a random variable. Given a random variable X which is either discrete or continuous, its **support** $\mathbb{X}$ is defined as the set

$$\mathbb{X} \equiv \{x : x \in \mathbb{R}, f_X(x) > 0\}$$

where $f_X(x)$ is the probability mass *or* density function associated with X, as appropriate.

In general:

- the support of **discrete** random variables is a **countable** set (corresponding with a countable sample space);
- whereas the support of **continuous** random variables is an **uncountable** set (thus corresponding with an uncountable sample space).

Mass function and support

- As the support of a discrete random variable is countable, a p.m.f. has an easy interpretation as a transposition of the underlying probability function.

- Thus:

$$\mathbb{P}\left(a \leq X \leq b\right) = F_X\left(b\right) - F_X\left(a\right) = \sum_{t=a}^{b} f_X\left(t\right)$$

hence:

$$\mathbb{P}\left(X \leq b\right) = F_X\left(b\right) = \sum_{t=\inf \mathbb{X}}^{b} f_X\left(t\right)$$

and:

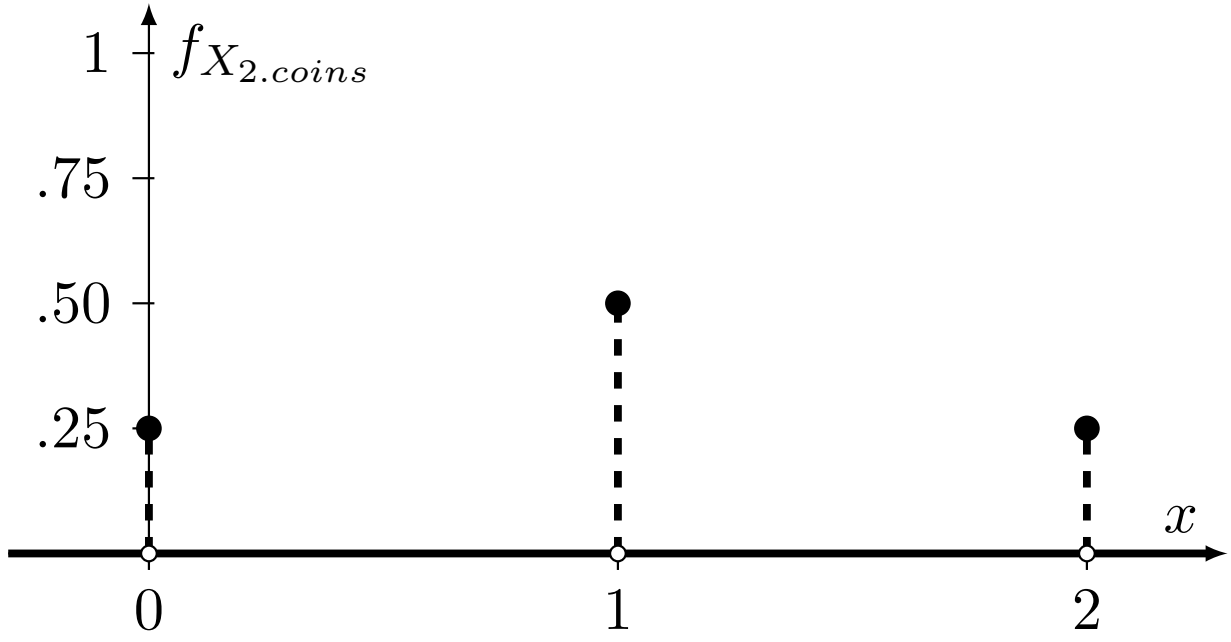
$$\mathbb{P}\left(X \in \mathbb{X}\right) = \sum_{t \in \mathbb{X}} f_X\left(t\right) = 1$$

which connects a p.m.f. with the c.d.f. $F_X\left(x\right)$.

Example: two coins, revisited

The p.m.f. associated with $X_{2.coins}$ is as follows (with figure).

$$f_{X_{2.coins}}(x) = \begin{cases} .25 & \text{if } x = 0 \\ .50 & \text{if } x = 1 \\ .25 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$



Density function and support

- For density functions instead the support is an uncountable set, and the interpretation of $f_X(x) \geq 0$ is subtler.

- This is no probability: x has "measure zero" in the support.

- In this case, the definition of c.d.f. implies:

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(t)\, dt$$

hence:

$$\mathbb{P}(X \in \mathbb{X}) = \int_{\mathbb{X}} f_X(t)\, dt = 1$$

a probabilistic interpretation for *segments* of $\mathbb{R}$.

- Unlike mass functions, density functions can generally take values larger than one: their probabilistic interpretation is based on the above integral formulations.

Standard logistic and normal densities (1/2)

- The **density** function associated with the standard logistic distribution $\Lambda(x)$ is:

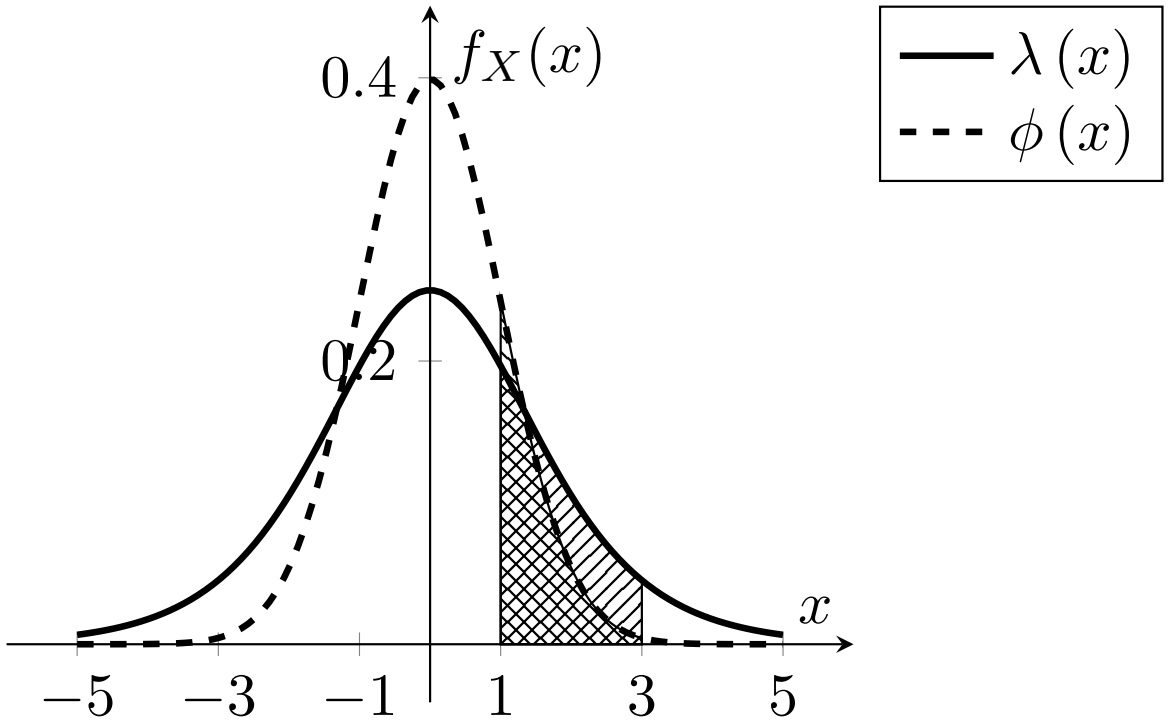
$$\lambda(x) = \frac{d\Lambda(x)}{dx} = \frac{\exp(-x)}{[1+\exp(-x)]^2}$$

- and the **density** function corresponding with the standard normal distribution is as follows.

$$\phi(x) = \frac{d\Phi(x)}{dx} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

- More general (i.e. non-standard) **parameterized** versions of these distributions exist.

Standard logistic and normal densities (2/2)



- Note: the shaded areas represent the probability that x falls in the $[1, 3]$ interval for either distribution.

Properties of mass and density functions

Theorem 8

Properties of mass and density functions. *A function $f_X(X)$ is an appropriate probability mass or density function of a given random variable X if and only if:*

a. $f_X(X) \geq 0$ *for all* $x \in \mathbb{R}$;

b. $\sum_{x \in \mathbb{X}} f_X(x) = 1$ *or* $\int_{\mathbb{X}} f_X(x)\,dx = 1$ *respectively for mass and density functions.*

Proof.

(Outline.) Necessity follows directly by the definitions of c.d.f., p.m.f. and p.d.f.; sufficiency follows by Theorem 7 after having constructed the associated cumulative distribution $F_X(X)$. □

Mixing discrete and continuous

- Certain distributions are continuous in some parts of their support, and discrete in other parts.

- A unified treatment of these distributions is possible using measure theory.

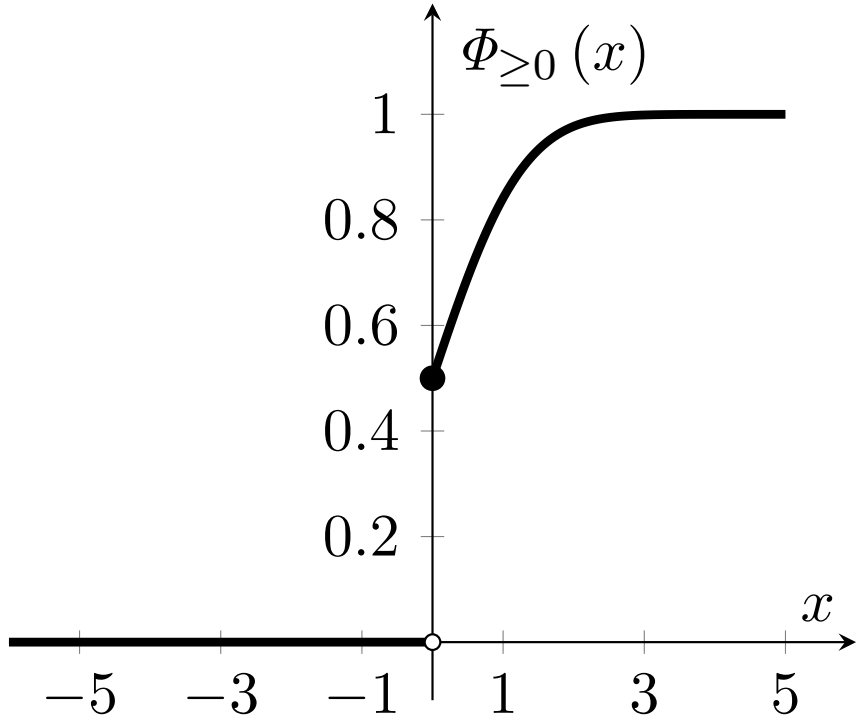
- An example is a **truncated** standard normal distribution.

$$\Phi_{\geq 0}(x) = \begin{cases} 0 & \text{if } x < 0 \\ \Phi(x) & \text{if } x \geq 0 \end{cases}$$

  Interpretation: negative realizations of x are not measured.

- For such distributions, the definitions of mass and density functions are relevant only in subsets of the support.

Truncated standard normal distribution



- Note: it is $\mathbb{P}(X = 0) = 0.5$ and $\mathbb{P}(X < 0) = 0$.

Identical distributions

Definition 16

Identically Distributed Random Variables. Any pair of random variables X and Y sharing a sample space $\mathbb{S}$ and an associated sigma algebra $\mathcal{B}$ are said to be **identically distributed** if, for every event $\mathbb{A} \in \mathcal{B}$, it is $\mathbb{P}\left(X \in X\left(\mathbb{A}\right)\right) = \mathbb{P}\left(Y \in Y\left(\mathbb{A}\right)\right)$.

Theorem 9

Identical Distribution. *Given two random variables X and Y whose primitive sample space is a subset of the real numbers $\mathbb{S} \subseteq \mathbb{R}$, the following two statements are identical:*

a. *X and Y are identically distributed;*

b. *$F_X\left(x\right) = F_Y\left(x\right)$ for every x in the relevant support.*

Proof.

(Outline.) Clearly here **a.** implies **b.** by construction. The reverse is proved by showing that if the two distributions are identical, they also share a probability function defined for some sigma algebra $\mathcal{B}$ of $\mathbb{S}$. $\square$

Transforming random variables

- Sometimes one wants to apply a **transformation** $g(\cdot)$ to a random variable X.

$$Y = g(X)$$

- This is interpreted as follows.

$$\mathbb{P}(Y \in [a,b]) = \mathbb{P}(g(X) \in [a,b])$$

- Interpretation is clearer if $g(\cdot)$ is invertible in the interval of interest – say $[a,b]$, $(a,b]$, $[a,b)$ or (a,b).

$$\mathbb{P}(Y \in [a,b]) = \mathbb{P}\left(X \in g^{-1}([a,b])\right)$$

- Transformations may imply a change in support; e.g. if X has support $\mathbb{X} = \mathbb{R}$, $Y = \exp(X)$ has support $\mathbb{Y} = \mathbb{R}_{++}$.

Transformations: discrete vs. continuous

- When X is discrete, it is easy to calculate the distribution of its transformations.

- For every point $y \in \mathbb{Y}$ (where $\mathbb{Y}$ is in the support of Y):

$$f_Y(y) = f_X\left(g^{-1}(y)\right)$$

- . . . and thus the cumulative distribution for Y would follow.

$$F_Y(y) = \sum_{\inf \mathbb{Y}}^{y} f_Y(y)$$

- Continuous distributions entail more complications.

Cumulative transformed distributions (1/2)

Theorem 10
Cumulative Distribution of Transformed Random Variables. *Let X and $Y = g(X)$ be two random variables that are related by a transformation $g(\cdot)$, $\mathbb{X}$ and $\mathbb{Y}$ their respective supports, and $F_X(x)$ the cumulative distribution of X.*

a. *If $g(\cdot)$ is increasing in $\mathbb{X}$, it is $F_Y(y) = F_X\left(g^{-1}(y)\right)$ for all $y \in \mathbb{Y}$.*

b. *If $g(\cdot)$ is decreasing in $\mathbb{X}$ and X is a continuous random variable, it is $F_Y(y) = 1 - F_X\left(g^{-1}(y)\right)$ for all $y \in \mathbb{Y}$.*

Proof.
(Continues...)

Cumulative transformed distributions (2/2)

Proof.
(Continued.) This is almost tautological: **a.** is shown as:

$$F_Y(y) = \int_{-\infty}^{g^{-1}(y)} f_X(x)\,dx = F_X\left(g^{-1}(y)\right)$$

because an increasing function applied upon some interval preserves its order. The demonstration of **b.** is symmetric:

$$F_Y(y) = \int_{g^{-1}(y)}^{\infty} f_X(x)\,dx = 1 - F_X\left(g^{-1}(y)\right)$$

because a decreasing function applied upon an interval would invert its order and because $\int_{-\infty}^{a} f_X(x)\,dx + \int_{a}^{\infty} f_X(x)\,dx = 1$ if $f_X(x)$ is a density function. □

Transformed density functions: simple (1/2)

Theorem 11

Density of Transformed Random Variables (simple). *Let X and $Y = g(X)$ be two random variables related by a transformation $g(\cdot)$, $\mathbb{X}$ and $\mathbb{Y}$ their respective supports, and $f_X(x)$ the probability density function of X, which is continuous on $\mathbb{X}$. If the inverse of the transformation function, $g^{-1}(\cdot)$, is continuously differentiable on $\mathbb{Y}$, the probability density function of Y can be calculated as follows.*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| & \text{if } y \in \mathbb{Y} \\ 0 & \text{if } y \notin \mathbb{Y} \end{cases}$$

Proof.

(Continues...)

Transformed density functions: simple (2/2)

Theorem 11

Proof.

(Continued.) Increasing and decreasing functions are monotone; hence, since $g^{-1}(\cdot)$ is continuously differentiable on $\mathbb{Y}$, for all $y \in \mathbb{Y}$:

$$f_Y(y) = \frac{d}{dy} F_Y(y)$$

$$= \begin{cases} f_X\left(g^{-1}(y)\right) \frac{d}{dy} g^{-1}(y) & \text{if } g(\cdot) \text{ is increasing} \\ -f_X\left(g^{-1}(y)\right) \frac{d}{dy} g^{-1}(y) & \text{if } g(\cdot) \text{ is decreasing} \end{cases}$$

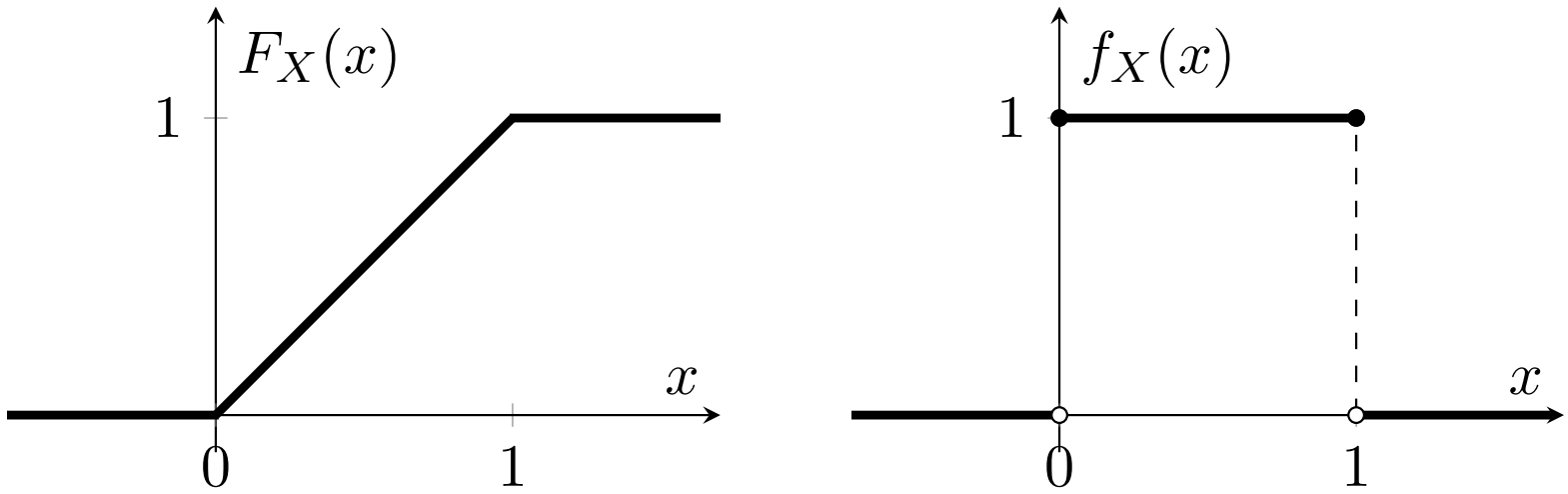
by Theorem 10 and the chain rule. □

Example: uniform-to-exponential (1/2)

Consider the **uniform distribution on the unit interval**: a random variable X with support $\mathbb{X} = [0, 1]$, c.d.f.

$$F_X(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0] \\ x & \text{if } x \in (0, 1) \\ 1 & \text{if } x \in [1, \infty) \end{cases}$$

and p.d.f. $f_X(x) = \mathbb{1}[x \in [0, 1]]$.



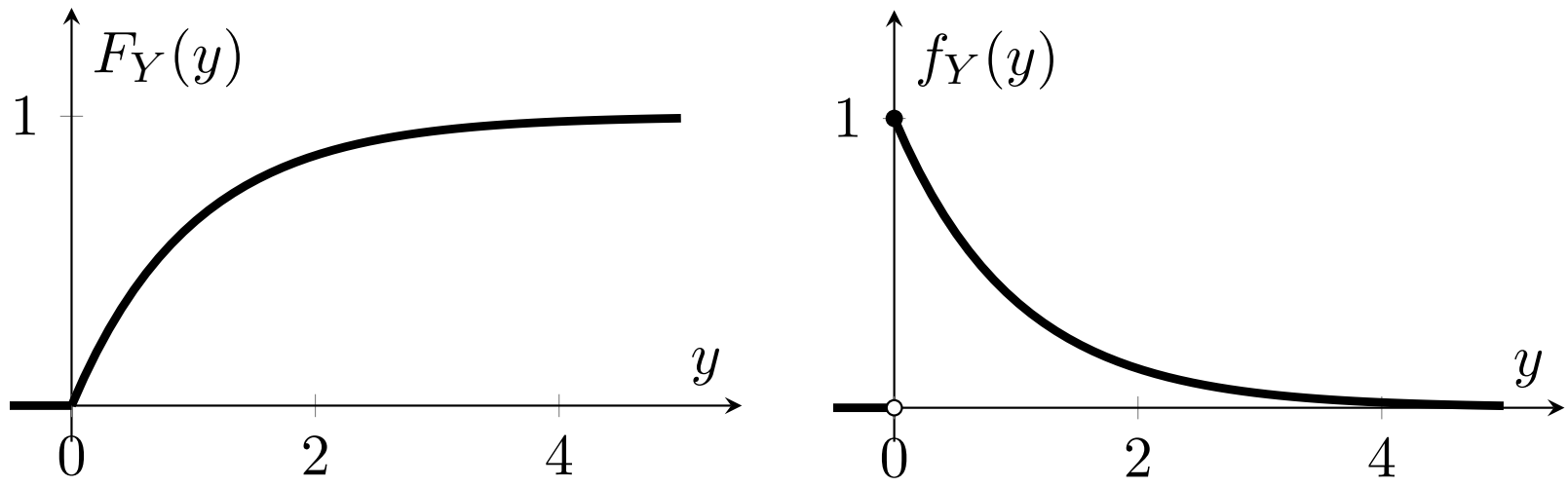
Note: cumulative distribution function $F_X(x)$ on the left, density function $f_X(x)$ on the right

Example: uniform-to-exponential (2/2)

Next, apply to X the transformation $Y = -\log X$: this returns the **exponential distribution with unit parameter**. Notice that the inverse transformation is $X = \exp(-Y)$, the support of Y is $\mathbb{Y} = \mathbb{R}_+$; Y has c.d.f.

$$F_Y(y) = 1 - \exp(-y)$$

while its p.d.f. is $f_Y(y) = \exp(-y)$ both defined for $y > 0$.



Note: cumulative distribution function $F_Y(y)$ on the left, density function $f_Y(y)$ on the right

Transformed density functions: composite

Theorem 12

Density of Transformed Random Variables (composite). *Let X and $Y = g(X)$ be two random variables related by some transformation $g(\cdot)$, $\mathbb{X}$ and $\mathbb{Y}$ their respective supports, and $f_X(x)$ the probability density function of X. Suppose further that there exists a partition of X's support, $\mathbb{X}_0, \mathbb{X}_1, \ldots, \mathbb{X}_K$ such that $\cup_{i=0}^{k} \mathbb{X}_i = \mathbb{X}$, $\mathbb{P}(x \in \mathbb{X}_0) = 0$, and $f_X(x)$ is continuous on each $\mathbb{X}_i$. Finally, suppose that there is a sequence of functions $g_1(x), \ldots, g_k(x)$, each associated with one set in $\mathbb{X}_1, \ldots, \mathbb{X}_K$, satisfying the following conditions for $i = 1, \ldots, K$:*

i. $g(x) = g_i(x)$ *for every* $x \in \mathbb{X}_i$;

ii. $g_i(x)$ *is monotone in* $\mathbb{X}_i$;

iii. $\mathbb{Y} = \{y : y = g_i(x) \text{ for some } x \in \mathbb{X}_i\}$, *that is the image of* $g_i(x)$ *is always equal to the support of* Y;

iv. $g_i^{-1}(y)$ *exists and is continuously differentiable in* $\mathbb{Y}$.

Then the density of Y can be calculated as follows.

$$f_Y(y) = \begin{cases} \sum_{i=1}^{K} f_X\left(g_i^{-1}(y)\right) \left| \frac{d}{dy} g_i^{-1}(y) \right| & \text{if } y \in \mathbb{Y} \\ 0 & \text{if } y \notin \mathbb{Y} \end{cases}$$

Example: squaring the standard normal

Let X follow the *standard normal distribution* $\Phi(x)$, and allow for the transformation $Y = X^2$: this is **not** monotone in $\mathbb{X} = \mathbb{R}$, but it is decreasing in $\mathbb{X}_1 = \mathbb{R}_{--}$, increasing in $\mathbb{X}_2 = \mathbb{R}_{++}$, while in both sets it maps onto $\mathbb{Y} = \mathbb{R}_{++}$. Also, $\mathbb{P}(X = 0) = 0$. Thus:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(-\sqrt{y}\right)^2}{2}\right) \left|-\frac{1}{2\sqrt{y}}\right| \\ &\quad + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\sqrt{y}\right)^2}{2}\right) \left|\frac{1}{2\sqrt{y}}\right| \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} \exp\left(-\frac{y}{2}\right) \end{aligned}$$

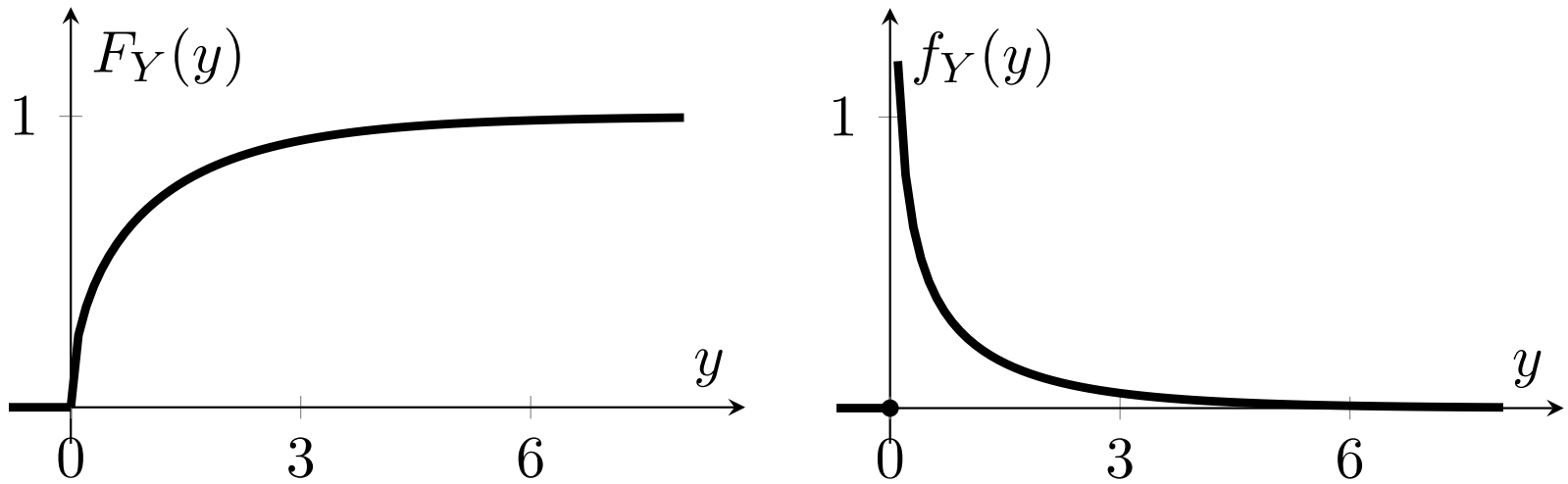
where in the first line:

- the first term accounts for $\mathbb{X}_1 = \mathbb{R}_{--}$ with $g_1^{-1}(y) = -\sqrt{y}$;
- the second term accounts for $\mathbb{X}_2 = \mathbb{R}_{++}$ with $g_2^{-1}(y) = \sqrt{y}$.

The chi-squared distribution with one d.f.

The distribution obtained by squaring the standard normal is a specific kind of a **chi-squared** (χ^2) distribution, that with **one degree of freedom**. Its c.d.f. obtains by integrating the p.d.f.:

$$F_Y(y) = \int_0^y \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{t}} \exp\left(-\frac{t}{2}\right) dt.$$



Note: cumulative distribution function $F_Y(y)$ on the left, density function $f_Y(y)$ on the right

- Chi-squared distributions are central in statistical inference.

Quantile functions

Definition 17

Quantile Function. The *quantile* function associated with a random variable X is the following function with argument $p \in (0,1)$.

$$Q_X(p) = \inf\{x \in \mathbb{X} : p \leq F_X(x)\}$$

- $Q_X(p)$ corresponds with the inverse of $F_X(x)$ if the latter is strictly increasing.

- otherwise (if $F_X(x)$ is flat on segments of the support of X) the quantile function returns a "pseudo-inverse" that has the property

$$\mathbb{P}(Q_X[F_X(X)] \leq Q_X(p)) = \mathbb{P}(X \leq Q_X(p))$$

for all $p \in (0,1)$, by construction.

Example: non-strictly-monotonic c.d.f. (1/2)

- Consider a random variable with the following distribution.

$$F_X(x) = \begin{cases} [1 + \exp(-x)]^{-1} & \text{if } x \in (-\infty, 0] \\ 0.5 & \text{if } x \in (0, 3) \\ [1 + \exp(-x + 3)]^{-1} & \text{if } x \in [3, \infty) \end{cases}$$

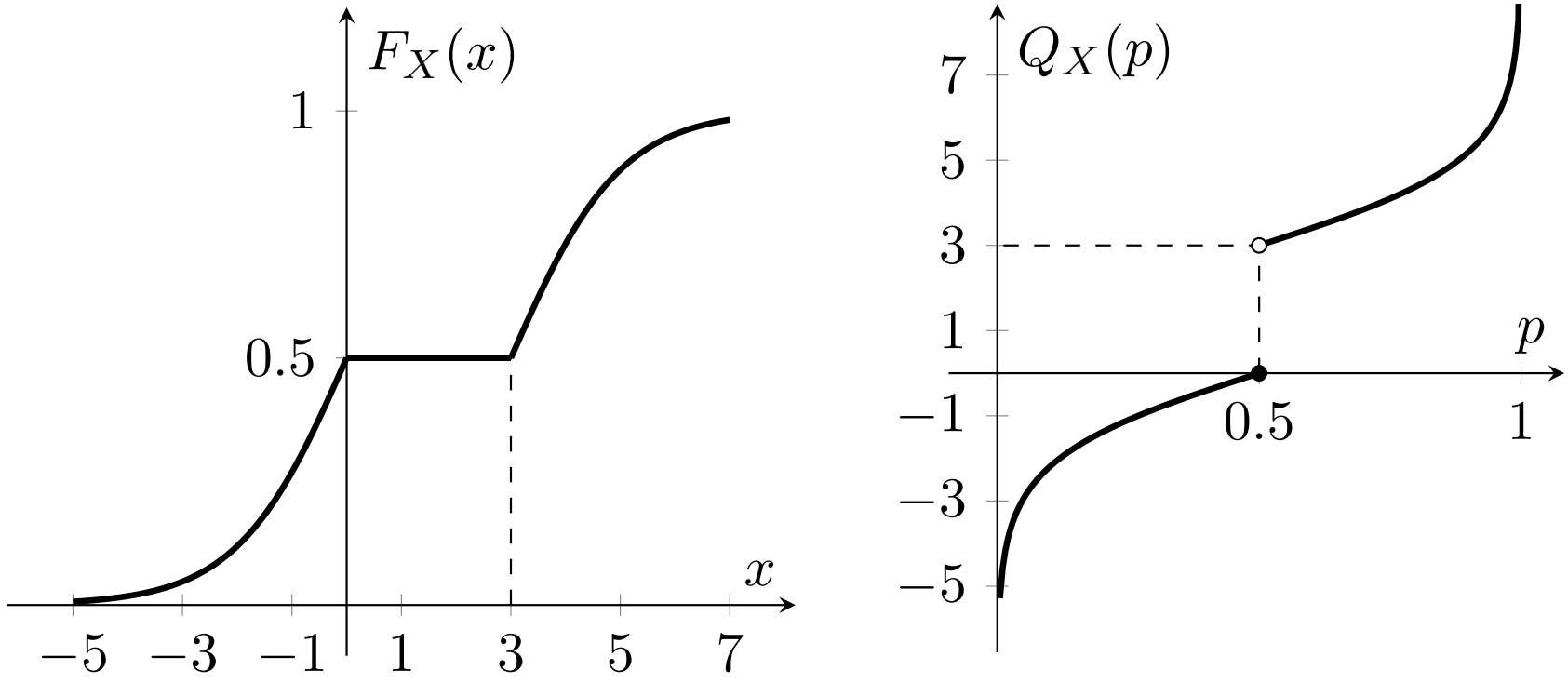
- This is similar to the standard logistic, but it is "stretched" over the $(0, 3)$ interval of the support, where it is flat.

- The corresponding quantile function is as follows.

$$Q_X(p) = \begin{cases} \log(x) - \log(1 - x) & \text{if } x \in (0, 0.5] \\ \log(x) - \log(1 - x) + 3 & \text{if } x \in (0.5, 1) \end{cases}$$

- At $p = 0.5$ the quantile function is discontinuous and equals zero: the infimum of the interval where the c.d.f. is flat.

Example: non-strictly-monotonic c.d.f. (2/2)



This figure represents the c.d.f. (left) and the quantile function (right) of the random variable described in this example.

Cumulative Transformation

Theorem 13

Cumulative Transformation. *For any continuous random variable X with cumulative distribution denoted as $F_X(x)$, the transformation $P = F_X(X)$ follows a uniform distribution on the unit interval.*

Proof.

By the properties of quantile functions (they are monotone increasing by definition, etc.), for all $p \in (0, 1)$ it holds that:

$$\begin{aligned}
\mathbb{P}(P \leq p) &= \mathbb{P}(F_X(X) \leq p) \\
&= \mathbb{P}(Q_X[F_X(X)] \leq Q_X(p)) \\
&= \mathbb{P}(X \leq Q_X(p)) \\
&= F_X(Q_X(p)) \\
&= p
\end{aligned}$$

The fourth and fifth lines follow from the definition and continuity of $F_X(x)$. Since by construction $F_P(p) = 0$ for $p \leq 0$ and $F_P(p) = 1$ for $p \geq 0$, P follows a uniform distribution on the interval $(0, 1)$. □

Uncentered moments

Definition 18

Uncentered Moments. The r-th uncentered moment of a random variable X with support $\mathbb{X}$, denoted as $\mathbb{E}\left[X^r\right]$, is defined as follows for some positive integer r and for discrete random variables:

$$\mathbb{E}\left[X^r\right] = \sum_{x \in \mathbb{X}} x^r f_X\left(x\right)$$

and as follows in the case of continuous random variables.

$$\mathbb{E}\left[X^r\right] = \int_{\mathbb{X}} x^r f_X\left(x\right) dx$$

Note: by definition, it is always $\mathbb{E}\left[X^0\right] = 1$.

- The most important uncentered moment is that for $r = 1$: $\mathbb{E}\left[X\right]$. It is called the **mean**.

- It is the "central" value of X in a probabilistic sense.

Centered moments

Definition 19

Centered Moments. The r-th centered moment of a random variable X with support $\mathbb{X}$, denoted as $\mathbb{E}\left[(X - \mathbb{E}\left[X\right])^r\right]$, is defined as follows for some positive integer r and for discrete random variables:

$$\mathbb{E}\left[(X - \mathbb{E}\left[X\right])^r\right] = \sum_{x \in \mathbb{X}} (x - \mathbb{E}\left[X\right])^r f_X(x)$$

and as follows in the case of continuous random variables.

$$\mathbb{E}\left[(X - \mathbb{E}\left[X\right])^r\right] = \int_{\mathbb{X}} (x - \mathbb{E}\left[X\right])^r f_X(x) \, \mathrm{d}x$$

- The most important centered moment is the one for $r = 2$: $\mathbb{V}\mathrm{ar}\left[X\right] \equiv \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right]$. It is called the **variance**.

- This is a measure of "dispersion" of the distribution of X around the mean. It is never negative.

Skewness

- The *standardized* centered moment of order $r = 3$ is called **skewness**.

$$\mathbb{S}\text{kew}\left[X\right] \equiv \frac{\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^3\right]}{\left(\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]\right)^{\frac{3}{2}}} \gtreqless 0$$

- It measures the degree of *asymmetry* of a distribution. It is a positive number for asymmetric distributions "skewed" to the *right* of the mean.

- At the same time, it is a negative number for asymmetric distributions "skewed" to the *left* of the mean.

- It equals exactly zero for distributions that are symmetric around the mean.

Kurtosis

- The *standardized* centered moment of order $r = 4$ is called **kurtosis**.

$$\mathbb{K}\text{urt}\left[X\right] \equiv \frac{\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^4\right]}{\left(\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]\right)^2} \geq 0$$

- Like the variance, it is always nonnegative.

- It measures the overall "thickness" of the distribution – the relative frequency of realizations of X that are distant from the mean.

- The kurtosis gives more weight to *extreme* deviations from the mean than the variance does.

From centered to uncentered moments

- Observe that the variance can be conveniently expressed in terms of uncentered moments.

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[X\right] &= \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}\left[X\right]\mathbb{E}\left[X\right] + \mathbb{E}\left[X\right]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2\end{aligned}$$

- This is true more generally for all centered moments.

$$\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^r\right] = \sum_{i=0}^{r} \binom{r}{i} (-1)^{r-i}\,\mathbb{E}\left[X^i\right]\mathbb{E}\left[X\right]^{r-i}$$

- This is convenient: uncentered moments can be **calculated** via *moment generating* or *characteristic* functions (more on this soon).

Example: tossing an unbalanced coin

- Consider an **unbalanced** coin: let $x_{coin} = 1$ for *Head* while $x_{coin} = 0$ for *Tail*, with $f_{X_{coin}}(1) = 0.6$ and $f_{X_{coin}}(0) = 0.4$.

- The mean is calculated as:

$$\begin{aligned}\mathbb{E}[X_{coin}] &= 1 \cdot f_{X_{coin}}(1) + 0 \cdot f_{X_{coin}}(0) \\ &= 1 \cdot 0.6 + 0 \cdot 0.4 \\ &= 0.6\end{aligned}$$

- . . . whereas the variance is calculated as follows.

$$\begin{aligned}\mathbb{V}\mathrm{ar}[X_{coin}] &= (1 - \mathbb{E}[X_{coin}])^2 \cdot f_{X_{coin}}(1) + \\ &\quad + (0 - \mathbb{E}[X_{coin}])^2 \cdot f_{X_{coin}}(0) \\ &= (0.4)^2 \cdot 0.6 + (-0.6)^2 \cdot 0.4 \\ &= 0.24\end{aligned}$$

Example: tossing many unbalanced coins (1/3)

- Now let $X_{n.coins}$ be the random variable which counts the outcome of n such experiments with an unbalanced coin.
- In total, there are 2^n possible sequences of outcomes, the interest though falls on the total number of *heads*.
- The outcomes that deliver exactly x heads are counted via the binomial coefficient $\binom{n}{x} = n!/x!\,(n-x)!$.
- For each such outcome, a head still occurs with probability 0.6 and a tail with probability 0.4.
- The p.m.f. for $X_{n.coins}$ is thus as follows.

$$f_{X_{n.coins}}(x) = \binom{n}{x} \cdot 0.6^x \cdot 0.4^{n-x} = \frac{n!}{x!\,(n-x)!} \cdot 0.6^x \cdot 0.4^{n-x}$$

Example: tossing many unbalanced coins (2/3)

The mean of $X_{n.coins}$ is calculated as follows.

$$
\begin{aligned}
\mathbb{E}\left[X_{n.coins}\right] &= \sum_{x=0}^{n} x \binom{n}{x} \cdot 0.6^x \cdot 0.4^{n-x} \\
&= \sum_{x=1}^{n} n \binom{n-1}{x-1} \cdot 0.6^x \cdot 0.4^{n-x} \\
&= 0.6 \cdot n \sum_{x=1}^{n} \binom{n-1}{x-1} \cdot 0.6^{x-1} \cdot 0.4^{n-1-x+1} \\
&= 0.6 \cdot n \underbrace{\sum_{y=0}^{n} \binom{n-1}{y} \cdot 0.6^y \cdot 0.4^{n-1-y}}_{=1} \\
&= 0.6 \cdot n
\end{aligned}
$$

Note: in the fourth line, it is $y = x - 1$.

Example: tossing many unbalanced coins (3/3)

The second uncentered moment of $X_{n.coins}$ thus obtains as:

$$\begin{aligned}
\mathbb{E}\left[X_{n.coins}^2\right] &= \sum_{x=0}^{n} x^2 \binom{n}{x} \cdot 0.6^x \cdot 0.4^{n-x} \\
&= \sum_{x=1}^{n} xn \binom{n-1}{x-1} \cdot 0.6^x \cdot 0.4^{n-x} \\
&= n \sum_{y=0}^{n} (y+1) \binom{n-1}{y} \cdot 0.6^{y+1} \cdot 0.4^{n-y-1} \\
&= 0.6 \cdot n \sum_{y=0}^{n} y \binom{n-1}{y} \cdot 0.6^y \cdot 0.4^{n-y} + \\
&\quad + 0.6 \cdot n \sum_{y=0}^{n} \binom{n-1}{y} \cdot 0.6^y \cdot 0.4^{n-y} \\
&= 0.6 \cdot n \cdot [0.6 \cdot (n-1) + 1]
\end{aligned}$$

hence, $\mathbb{V}\mathrm{ar}\left[X_{n.coins}\right] = \mathbb{E}\left[X_{n.coins}^2\right] - \mathbb{E}\left[X_{n.coins}\right]^2 = 0.24 \cdot n.$

Example: moments of the uniform distribution

- Let X follow the uniform distribution on the $[0, 1]$ interval.
- The mean is obtained as:

$$\mathbb{E}[X] = \int_0^1 x dx = \left. \frac{x^2}{2} \right|_0^1 = \frac{1}{2}$$

- while the second uncentered moment is:

$$\mathbb{E}\left[X^2\right] = \int_0^1 x^2 dx = \left. \frac{x^3}{3} \right|_0^1 = \frac{1}{3}$$

- thus the variance is as follows.

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Example: moments in the exponential case (1/2)

- Now let $Y = -\log(X)$: therefore, Y follows the exponential distribution with unit parameter as examined earlier.

- The mean is obtained as:

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^\infty y \exp(-y)\, dy \\ &= -y \exp(-y)|_0^\infty + \int_0^\infty \exp(-y)\, dy \\ &= 1\end{aligned}$$

  where the second line applies integration by parts.

- Note: $\lim_{M\to\infty} -y \exp(-y)|_0^M = 0$ and $\int_0^\infty \exp(-y)\, dy = 1$.

Example: moments in the exponential case (2/2)

- The second uncentered moment is:

$$\begin{aligned}\mathbb{E}\left[Y^2\right] &= \int_0^\infty y^2 \exp\left(-y\right) dy \\ &= -y^2 \exp\left(-y\right)\Big|_0^\infty + 2\int_0^\infty y \exp\left(-y\right) dy \\ &= 2\end{aligned}$$

  Once again, this applies integration by parts and standard calculations.

- We can thus derive the variance.

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[Y\right] &= \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[Y\right]^2 \\ &= 2 - 1 \\ &= 1\end{aligned}$$

Moments of linear transformations

Let $Y = a + bX$ a **linear** transformation of a random variable X. How are the moments of X and Y related to one another?

- As for the **mean**:

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[a + bX] \\ &= a + b\,\mathbb{E}[X]\end{aligned}$$

- whereas with regard to the **variance** it is:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[Y] &= \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] \\ &= \mathbb{E}\left[b^2 (X - \mathbb{E}[X])^2\right] \\ &= b^2\,\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= b^2\,\mathbb{V}\mathrm{ar}[X]\end{aligned}$$

- These follow from the linear properties of sums/integrals.

Moments of non-linear transformations

And if Y & X are related by a **non-linear** function $Y = g(X)$?

- No exact result, but by **Jensen's Inequality** one has:

$$\begin{aligned} \mathbb{E}[g(X)] &\leq g(\mathbb{E}[X]) \quad \text{if } g(\cdot) \text{ is a } \textbf{concave} \text{ function;} \\ \mathbb{E}[g(X)] &\geq g(\mathbb{E}[X]) \quad \text{if } g(\cdot) \text{ is a } \textbf{convex} \text{ function.} \end{aligned}$$

- Furthermore, a **Taylor approximation** around $\mathbb{E}[X]$:

$$\begin{aligned} g(X) &\approx g(\mathbb{E}[X]) + g'(\mathbb{E}[X])[X - \mathbb{E}[X]] \\ &= [g(\mathbb{E}[X]) - g'(\mathbb{E}[X])\mathbb{E}[X]] + g'(\mathbb{E}[X])X \end{aligned}$$

hints that $\mathbb{E}[g(X)] \approx g(\mathbb{E}[X])$ is a **bad** approximation; the rearrangement of terms in the second line however suggests that the following is a **good** approximation.

$$\mathbb{V}\text{ar}[g(X)] \approx [g'(\mathbb{E}[X])]^2 \mathbb{V}\text{ar}[X]$$

Markov's Inequality

Theorem 14

Markov's Inequality. *Given a nonnegative random variable* $X \in \mathbb{R}_+$ *and a constant* $k > 0$*, it must be* $\mathbb{P}[X \geq k] \leq \mathbb{E}[X]/k$.

Proof.

Apply the decomposition

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{+\infty} x f(x)\, dx \\
&\geq \int_k^{+\infty} x f(x)\, dx \\
&\geq k \int_k^{+\infty} f(x)\, dx \\
&= k\, \mathbb{P}[X \geq k]
\end{aligned}$$

with the first equality requiring X to be nonnegative. □

Čebyšëv's Inequality

Theorem 15

Čebyšëv's Inequality. *Given a random variable* $Y \in \mathbb{R}$ *and a number* $\delta > 0$*, it must be* $\mathbb{P}\left[|Y - \mathbb{E}\left[Y\right]| \geq \delta\right] \leq \mathbb{V}\mathrm{ar}\left[Y\right]/\delta^2$*.*

Proof.

Rephrase Markov's inequality setting $X = \left(Y - \mathbb{E}\left[Y\right]\right)^2$ and $k = \delta^2$, and notice that:

$$\begin{aligned}
\mathbb{P}\left[|Y - \mathbb{E}\left[Y\right]| \geq \delta\right] &\leq \mathbb{P}\left[\left(Y - \mathbb{E}\left[Y\right]\right)^2 \geq \delta^2\right] \\
&\leq \frac{\mathbb{E}\left[\left(Y - \mathbb{E}\left[Y\right]\right)^2\right]}{\delta^2} \\
&= \frac{\mathbb{V}\mathrm{ar}\left[Y\right]}{\delta^2}
\end{aligned}$$

as postulated. □

Minimum squared error of prediction (1/3)

- To build intuition about the relationship between the mean and the variance, consider the following problem.

$$\min_{\widehat{X}} \mathbb{E}\left[\left(X - \widehat{X}\right)^2\right]$$

- This is about *guessing* ("predicting") the realization of X using a single value $\widehat{X}$.

- The minimand is the **mean squared error** (MSE). This punishes big mistakes more than small mistakes.

- (This is an ubiquitous exercise in statistics! Understanding this simple example helps.)

Minimum squared error of prediction (2/3)

Note that:

$$\begin{aligned}
\mathbb{E}\left[\left(X-\widehat{X}\right)^2\right] &= \mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]+\mathbb{E}\left[X\right]-\widehat{X}\right)^2\right] \\
&= \mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)^2\right]+\mathbb{E}\left[\left(\mathbb{E}\left[X\right]-\widehat{X}\right)^2\right] \\
&\quad +2\underbrace{\mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)\left(\mathbb{E}\left[X\right]-\widehat{X}\right)\right]}_{=0} \\
&= \mathbb{V}\mathrm{ar}\left[X\right]+\mathbb{E}\left[\left(\mathbb{E}\left[X\right]-\widehat{X}\right)^2\right]
\end{aligned}$$

because the third term in the second line must be zero:

$$\mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)\left(\mathbb{E}\left[X\right]-\widehat{X}\right)\right]=\left(\mathbb{E}\left[X\right]-\widehat{X}\right)\mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)\right]=0$$

both $\mathbb{E}\left[X\right]$ and $\widehat{X}$ are "constants" and $\mathbb{E}\left[\left(X-\mathbb{E}\left[X\right]\right)\right]=0$.

Minimum squared error of prediction (3/3)

- Hence:

$$\min_{\widehat{X}} \mathbb{E}\left[\left(X-\widehat{X}\right)^2\right] = \min_{\widehat{X}} \mathbb{V}\mathrm{ar}\left[X\right] + \mathbb{E}\left[\left(\mathbb{E}\left[X\right]-\widehat{X}\right)^2\right]$$

- . . . and since $\mathbb{V}\mathrm{ar}\left[X\right]$ is a constant, the **optimal predictor** is the mean!

$$\widehat{X}^{optimal} = \mathbb{E}\left[X\right]$$

- A prediction based on the mean returns a residual MSE that is equal to the variance $\mathbb{V}\mathrm{ar}\left[X\right]$.

- Therefore, the variance is the prediction error "that cannot be removed."

Moment generating function

Definition 20

Moment generating function. Given some random variable X with support $\mathbb{X}$, the **moment-generating** function $M_X(t)$ is defined, for $t \in \mathbb{R}$, as the expectation of the transformation $g(X) = \exp(tX)$, so long as it exists. For discrete random variables this is:

$$M_X(t) = \mathbb{E}[\exp(tX)] = \sum_{x \in \mathbb{X}} \exp(tx) f_X(x)$$

while for continuous random variables, it is as follows.

$$M_X(t) = \mathbb{E}[\exp(tX)] = \int_{\mathbb{X}} \exp(tx) f_X(x) \, dx$$

- A moment generating function is often abbreviated as m.g.f.

Moment generation

Theorem 16

Moment generation. *If a random variable X has an associated moment generating function $M_X(t)$, its r-th uncentered moment can be calculated as the r-th derivative of the moment generating function evaluated at $t = 0$.*

$$\mathbb{E}[X^r] = \left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0}$$

Proof.

Note that, for all $r = 1, 2, \ldots$:

$$\frac{d^r M_X(t)}{dt^r} = \frac{d^r}{dt^r}\,\mathbb{E}[\exp(tX)] = \mathbb{E}\left[\frac{d^r}{dt^r}\exp(tX)\right] = \mathbb{E}[X^r \exp(tX)]$$

so long as the r-th derivative with respect to t can pass through the expectation operator. If so, $\mathbb{E}[X^r \exp(tX)] = \mathbb{E}[X^r]$ for $t = 0$. □

Example: m.g.f. for coin experiments

- Recall X_{coin} for an *unbalanced* coin. Its m.g.f. is:

$$\begin{aligned} M_{X_{coin}}(t) &= \exp(t \cdot 1) \cdot f_{X_{coin}}(1) + \exp(t \cdot 0) \cdot f_{X_{coin}}(0) \\ &= 0.6 \cdot \exp(t) + 0.4 \end{aligned}$$

- while for $X_{n.coins}$, it is as follows:

$$\begin{aligned} M_{X_{n.coins}}(t) &= \sum_{x=0}^{n} \binom{n}{x} \exp(tx) \cdot 0.6^x \cdot 0.4^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} [0.6 \cdot \exp(t)]^x \cdot 0.4^{n-x} \\ &= [0.6 \cdot \exp(t) + 0.4]^n \end{aligned}$$

this obtains by another application of the binomial formula.

Example: m.g.f. for the uniform distribution

- Let X follow the uniform distribution on the $[0, 1]$ interval. The m.g.f. is calculated as follows.

$$\begin{aligned} M_X(t) &= \int_0^1 \exp(tx)\, dx \\ &= \frac{1}{t} \exp(tx) \Big|_0^1 \\ &= \frac{1}{t} [\exp(t) - 1] \end{aligned}$$

- Note: this might seem ill-defined at $t = 0$, but applying the following **Taylor expansion** for $t = 0$:

$$\exp(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!}$$

into the m.g.f. formula, reveals that actually $M_X(0) = 1$.

Example: m.g.f. for the exponential case

- Let $Y = -\log(X)$: as previously, Y follows the exponential distribution with unit parameter.

- The m.g.f. is calculated as follows.

$$\begin{aligned} M_Y(t) &= \int_0^\infty \exp\left((t-1)\,y\right) dy \\ &= \lim_{M\to\infty} \; -\frac{1}{1-t} \exp\left(-(1-t)\,y\right)\Big|_0^M \\ &= \frac{1}{1-t} \end{aligned}$$

- This m.g.f. **only exists** for $t < 1$! In fact, the integral that defines it diverges if $t \geq 1$.

M.g.f. of transformations

- Let $Y = a + bX$ again a **linear** transformation of a random variable X.

- If they exist, the two m.g.f.s of X and Y are related to one another as follows.

$$\begin{aligned} M_Y(t) &= \mathbb{E}\left[\exp(tY)\right] \\ &= \mathbb{E}\left[\exp(ta + tbX)\right] \\ &= \exp(at)\,\mathbb{E}\left[\exp(btX)\right] \\ &= \exp(at)\,M_X(bt) \end{aligned}$$

- Getting exact results for non-linear transformations is not as straightforward.

The role of m.g.f.s

- Why are m.g.f.s important? As already seen, they allow to calculate moments, often more easily.

- There is another reason: a m.g.f. is unique to a distribution – i.e. each distribution $F_X(x)$ has a **distinct** m.g.f. $M_X(t)$.

- The proof of this result is outside this lecture's scope.

- However, a *sequence of moments* does *not* uniquely identify a distribution, i.e. different distributions can have identical moments. Exception: distributions with *bounded support*.

- Problem: m.g.f.s *may not exist* (for some value of t even).

- However, the closely related **characteristic functions** are also unique to distributions, and they **always** exist.

Characteristic functions

Definition 21

Characteristic function. Given some random variable X with support $\mathbb{X}$, the characteristic function $\varphi_X(t)$ is defined, for $t \in \mathbb{R}$ and for discrete random variables:

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = \sum_{x \in \mathbb{X}} \exp(itx) f_X(x)$$

while for continuous random variables it is:

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = \int_{\mathbb{X}} \exp(itx) f_X(x) \, dx$$

where i is the imaginary unit.

- Similarly to the case of m.g.f.s, one can calculate moments via characteristic functions.

$$\mathbb{E}[X^r] = \frac{1}{i^r} \cdot \left. \frac{d^r \varphi_X(t)}{dt^r} \right|_{t=0} \quad \text{for } r \in \mathbb{N}$$