# Probability Theory and Mathematical Statistics

## Part 1: Probability Theory

Geoffrey Decrouez
E-mail: ggdecrouez@hse.ru
Office: room 314

# Describing a random experiment $\mathcal{E}$

A random experiment $\mathcal{E}$ is an experiment in which the outcome or result cannot be predicted with certainty. To assign probabilities to a RE, we need two ingredients: a *sample space* and a $\sigma$-algebra.

**Definition 1.** *A set $\Omega$ of points $\omega$ representing all the possible outcomes of $\mathcal{E}$ is called a sample space.*

**Example 1.** *(i) Toss two coins. We are interested in the observed face of each coin.*

*(ii) Toss two coins. We are interested in the number of heads obtained.*

*(iii) Toss one coin. Number of tosses until we get the first H.*

**Remarks.**

(i) The same experiment can have different sample spaces.

(ii) Different experiments can have equivalent sample spaces (can you think of an example?).

**Definition 2.** *Let $\mathcal{F}$ denote a collection of subsets of $\Omega$. A subset $A \in \mathcal{F}$ is called an event, and so $\mathcal{F}$ is a collection of events. An event $A$ occurs if $\mathcal{E}$ results in an $\omega \in A$.*

Events are important: later we will assign probabilities to all $A \in \mathcal{F}$. But these probabilities cannot be arbitrary, they need to satisfy some properties [$\rightarrow$ axioms of Probability Theory]. Moreover, we need $\mathcal{F}$ to satisfy some conditions [definition 4].

There are two special events: the null event, denoted $\emptyset$, cannot happen, and the certain event $\Omega$, which always happens. We can have $A = \{\omega\}$. [A singleton.]

**Exercise 1.** *Roll a die. Consider events* $A = \{Obtain\ an\ odd\ number\}$ *and* $B = \{Obtain\ a\ number\ larger\ or\ equal\ to\ 5\}$. *Describe* $\Omega$ *for this experiment, and express* $A$ *and* $B$ *as subsets of* $\Omega$.

# Manipulating events

We will need notation from set theory.

- $\#A$ denotes the number of elements in a (finite) set $A$.
- $A$ is a *subset* of $B$ if every element of a set $A$ is also an element of a set $B$. We write $A \subset B$.
- $\omega \in A \Leftrightarrow \{\omega\} \subset A$.
- *Union of events* $A \cup B = \{\omega \in \Omega \mid \omega \in A \text{ or } \omega \in B\}$. $A \cup B$ is the event that $A$ or $B$ or both occur.
- *Intersection of events* $A \cap B = \{\omega \in \Omega \mid \omega \in A \text{ and } \omega \in B\}$. $A \cap B$ is the event that both $A$ and $B$ occur.
- Complementary event $\bar{A} = \{\omega \in \Omega \mid \omega \notin A\}$ occurs if and only if $A$ doesn't.

**Definition 3.** *(i) Two events $A_1$ and $A_2$ are mutually exclusive or disjoint if $A_1 \cap A_2 = \emptyset$ (they cannot occur simultaneously).*

*(ii) Events $A_1, A_2, A_3, \dots$ are mutually exclusive if they are pairwise mutually exclusive:*

$$A_i \cap A_j = \emptyset \textit{ for any } i \neq j \,.$$

*(iii) Events $A_1, A_2, A_3, \dots$ are exhaustive if*

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \ldots \cup A_n = \Omega \,.$$

**Properties of event operations.** Let $A, B$ and $C$ be subsets of $\Omega$. Then we have

- *Commutative laws.*

$$A \cup B = B \cup A \quad \text{and} \quad A \cap B = B \cap A\,.$$

- *Associative laws.*

$$(A \cup B) \cup C = A \cup (B \cup C)\,.$$
$$(A \cap B) \cap C = A \cap (B \cap C)\,.$$

- *Distributive laws.*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)\,.$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)\,.$$

- *De Morgan's laws.*

$$\overline{(A \cup B)} = \bar{A} \cap \bar{B}\,.$$
$$\overline{(A \cap B)} = \bar{A} \cup \bar{B}\,.$$

**Example 2.** *Roll a die.* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *Let* $A = \{5\}$ *and* $B = \{1, 3, 5\}$. *Then*

$$A \subset B \qquad A \cap B = A \qquad A \cup B = B$$

*If in addition we define* $C = \{2, 4, 6\}$, *then* $A \cap C = \emptyset$. *Moreover,*

$$B \cup C = \Omega \quad so \quad C = \bar{B}\,.$$

We are now well equipped to express more complicated event, such as

- None of $A, B, C$ occurred: $\bar{A} \cap \bar{B} \cap \bar{C} = \overline{(A \cup B \cup C)}$ (Use De Morgan's law).
- $A$ occurred and at least one of $B$ and $C$ occurred: $A \cap (B \cup C)$.
- Using De Morgan, $\overline{A \cap (B \cup C)} = \bar{A} \cup \overline{B \cup C} = \bar{A} \cup (\bar{B} \cap \bar{C})$.

**Exercise 2.** *Re-express* $(A \cap B) \cup (B \cap C) \cup (A \cap C)$ *as a union of disjoint events. [Hint: draw a diagram.]*

It becomes clear that if $A$ and $B$ are events, then their union and complement should also be events (in other words, we should be able to assign probabilities to them). The same goes for their intersection, which is automatic from De Morgan's laws. In fact, we need a bit more, we need to be allowed to take *countable* unions of events. When dealing with countably many elements, we use the symbol $\sigma$.

**Definition 4.** *A class $\mathcal{F}$ of subsets of $\Omega$ is called a $\sigma$-algebra if the following holds*

*(i)* $\Omega \in \mathcal{F}$

*(ii)* $A \in \mathcal{F} \Rightarrow \bar{A} \in \mathcal{F}$

*(iii)* $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$*.*

Note that $\emptyset \in \mathcal{F}$ since $\emptyset = \bar{\Omega}$ and using (i) and (ii).

If $A_1$ and $A_2$ are in $\mathcal{F}$, then $A_1 \cup A_2 \in \mathcal{F}$: take $A_3 = A_4 = \ldots = \emptyset$. Also, $A_1 \cap A_2 = \overline{(\bar{A}_1 \cup \bar{A}_2)} \in \mathcal{F}$.

**Definition 5.** *The pair $(\Omega, \mathcal{F})$ is called a measurable space.*

**Examples of $\sigma$-algebra.**

(i) Trivial $\sigma$-algebra $\mathcal{F} = \{\emptyset, \Omega\}$.

(ii) If $\Omega = \{a, b\}$ has $|\Omega| = 2$ elements, then $\mathcal{F} = \{\emptyset, \{a\}, \{b\}, \Omega\}$, the collection of all possible events, is obviously a $\sigma$-algebra. This $\sigma$-algebra has a special name: the *power set*. It is usually denoted $2^{\Omega}$ or $\mathcal{P}(\Omega)$. It contains $|\mathcal{F}| = 2^{|\Omega|} = 4$ distinct events.

(iii) If $\Omega = \{a, b, c, d\}$, then the set of all subsets $\mathcal{P}(\Omega)$ has $2^{|\Omega|} = 16$ elements.

$$\mathcal{P}(\Omega) = \{\emptyset, \{a\}, \dots, \{d\}, \{a, b\}, \dots, \{c, d\}, \{a, b, c\}, \dots, \{b, c, d\}, \Omega\}\,.$$

However, it is not the only $\sigma$-algebra on $\Omega$, $\mathcal{F} = \{\emptyset, \{a, b\}, \{c, d\}, \Omega\}$ also does the job.

(iv) There is an extension when $\Omega$ has countably many elements (e.g. $\Omega = \mathbb{N}$).

(v) What about $\Omega = \mathbb{R}$? The power set of $\mathbb{R}$ is a $\sigma$-algebra. However, it is too big, it contains monsters. [We get back to this point later.]

# Axioms of Probability Theory

Let $(\Omega, \mathcal{F})$ be a measurable space. A *probability* is a real-valued function $\mathbf{P}$ that assigns to each event $A \in \mathcal{F}$ a number $\mathbf{P}(A)$, called probability of the event $A$, such that:

(i) $\mathbf{P}(A) \geqslant 0$

(ii) $\mathbf{P}(\Omega) = 1$

(iii) If a sequence of events $\{A_n\}$ is mutually exclusive, then

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)\,.$$

Claims (i), (ii) and (iii) are called *axioms*. We take them for granted. From these three axioms, we can build a whole theory. They constitute the foundations of modern probability (Introduced by Kolmogorov in 1933). Motivation for using (i), (ii) and (iii) as building blocks of probability theory comes from the frequency interpretation of probability. $(\Omega, \mathcal{F}, \mathbf{P})$ is called a *probability space*.

**Example 3.** *Back to the roll of a die.* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *Then*

$$\mathbf{P}(\{1\}) = \mathbf{P}(\{2\}) = \mathbf{P}(\{3\}) = \mathbf{P}(\{4\}) = \mathbf{P}(\{5\}) = \mathbf{P}(\{6\}) = 1/6\,.$$

At this stage, a few remarks are in order.

(a) On the same measurable space $(\Omega, \mathcal{F})$, we can have different probability distributions. In Exemple 6, what happens if you roll a biased die?

(b) To completely specify $\mathbf{P}$, we do not need to know $\mathbf{P}(A)$ for all events in $\mathcal{F}$. That's good, since the total number of events can be huge. Instead, it is usually sufficient to know $\mathbf{P}(A)$ for a few events only.

(c) What happens to axiom (iii) when the sequence of events $\{A_n\}$ is not mutually exclusive?

(d) What happens if axiom (iii) holds true only for finitely many events $\{A_n\}$?

## Properties of probability.

Let $A, B \in \mathcal{F}$.

(i) $\mathbf{P}(\emptyset) = 0$

(ii) $\mathbf{P}(\bar{A}) = 1 - \mathbf{P}(A)$

(iii) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$

(iv) $\mathbf{P}(A) \leq 1$

(v) $\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap \bar{B})$

(vi) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$

Property (v) is fundamental, and central to many proofs in probability theory. It even has a name: Law of Total Probability (LTP). We will encounter this law in a more general form later on.

# Note on classical probability

Refers to the case when $\Omega$ is a finite set, and that all outcomes are equally likely.

If $n = |\Omega|$ denotes the number of elements of $\Omega$, then $\mathbf{P}(\{\omega\}) = 1/n$.

For any event $A$, we have

$$\mathbf{P}(A) = \sum_{\omega \in A} \frac{1}{n} = \frac{\#A}{n} = \frac{\#A}{\#\Omega} \,.$$

In this case, computing the probability of $A$ a requires the computation of the number of elements in $A$ (using combinatorics).

We will not focus on the classical scheme in this course.

# Conditional probability

**Introductive example.** Suppose there are 20 tulip bulbs very similar in appearance. We are told that

- 8 tulips bloom early (call this event E).
- 12 will bloom late (L)
- 13 are red (R)
- 7 are yellow (Y)

| | E | L | Total |
|---|---|---|---|
| R | 5 | 8 | 13 |
| Y | 3 | 4 | 7 |
| Total | 8 | 12 | 20 |

Select a bulb at random, assuming that each bulb is equally likely. The sample space $\Omega$ consists in the 20 bulbs, and

$$\mathbf{P}(E) = 8/20, \quad \mathbf{P}(R) = 13/20, \quad \mathbf{P}(R \cap E) = 5/20\,.$$

Suppose now we are told the selected bulb will bloom early. What is the probability that the bulb will produce a red tulip?

$$\mathbf{P}(R \mid E) = 5/8 = \frac{\#R \cap E}{\#E} = \frac{(\#R \cap E)/(\#\Omega)}{(\#E)/(\#\Omega)} = \frac{\mathbf{P}(R \cap E)}{\mathbf{P}(E)}\,.$$

This leads to the following definition:

**Definition 6.** *The conditional probability of $A$ given $B$, denoted $\mathbf{P}(A \mid B)$, is*

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

*provided $\mathbf{P}(B) > 0$.*

Note that for a fixed $B$, $\mathbf{P}(\cdot \mid B)$ is also a probability distribution (it satisfies the three axioms of probability). Can you prove this?

**Definition 7.** *Multiplication rule.*

$$\begin{aligned} \mathbf{P}(A \cap B) &= \mathbf{P}(A \,|\, B)\,\mathbf{P}(B) \\ &= \mathbf{P}(B \,|\, A)\,\mathbf{P}(A) \end{aligned}$$

The multiplication rule can be extended to three or more events:

$$\begin{aligned} \mathbf{P}(A \cap B \cap C) &= \mathbf{P}(C \,|\, A \cap B)\,\mathbf{P}(A \cap B) \\ &= \mathbf{P}(C \,|\, A \cap B)\,\mathbf{P}(B \,|\, A)\,\mathbf{P}(A)\,. \end{aligned}$$

**Exercise 3.** *Policy holders in an insurance company are such that 60% are with auto policies, 40% with homeowner policies, and 20% with both. A person is selected at random from the policy holders. Consider the following events*

$$\begin{aligned} A_1 &= \{He/She\ has\ only\ auto\ policy\} \\ A_2 &= \{He/She\ has\ only\ homeowner\ policy\} \\ A_3 &= \{He/She\ has\ both\} \\ A_4 &= \{He/She\ has\ neither\} \\ B &= \{He/She\ is\ to\ renew\ at\ least\ one\ policy\}\,. \end{aligned}$$

*Furthermore, it is known that*

$$\mathbf{P}(B \,|\, A_1) = 0.6, \quad \mathbf{P}(B \,|\, A_2) = 0.7, \quad \mathbf{P}(B \,|\, A_3) = 0.8, \quad \mathbf{P}(B \,|\, A_4) = 0\,.$$

*What is the conditional probability that the person will renew at least one of the auto and homeowner policies given that he/she currently has an auto or homeowner policy?*

# Independent Events

We say that two events $A$ and $B$ are independent if the occurrence of one does not change the likelihood of the other. Namely, $\mathbf{P}(A \,|\, B) = \mathbf{P}(A)$, or $\mathbf{P}(B \,|\, A) = \mathbf{P}(B)$. Since one has

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

we get $\mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B)$. We take this as a definition of independence.

**Definition 8.** *Two events $A$ and $B$ are independent if and only if*

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

**Warning:** Do not confuse *disjoint* and *independent* events.

*Disjoint* events is a set theory concept: the occurrence of one event is incompatible with the occurrence of the other. If you roll a die, then $A = \{\text{number is less than } 2\}$ and $B = \{\text{number is larger than } 5\}$ are disjoint. There is no probability in this definition.

*Independence* is a measure theoretic concept (think of a probability in terms of a measure): learning that one event occurs does not provide information about whether the other event occurred. In fact, two disjoint events cannot be independent, except in the trivial case, since if $A \cap B = \emptyset$, then $\mathbf{P}(A \cap B) = \mathbf{P}(\emptyset) = 0 \neq \mathbf{P}(A)\,\mathbf{P}(B)$.

**Theorem 1.** *If $A$ and $B$ are independent, then*

*(i)* $A$ *and* $\bar{B}$ *are independent*

*(ii)* $\bar{A}$ *and* $B$ *are independent*

*(iii)* $\bar{A}$ *and* $\bar{B}$ *are independent*

**Exercise 4.** *A home audio system is composed of a tuner, a CD player, an amplifier and two speakers. The home audio system is said to be working when either the tuner or CD player, the amplifier, and at least one speaker are working. Assume that these components are working independently of each other. Write down the probability of the event* $A = \{Home\ audio\ system\ is\ working\ \}$ *in terms of the events* $C_i = \{Component\ i\ is\ working\}$*. What about the probability of failure of the audio system?*

**Example 4.** *An urn contains 4 balls numbered 1,2,3 and 4. One ball is drawn at random from the urn. Let* $A = \{1,2\}$*,* $B = \{1,3\}$ *and* $C = \{1,4\}$*.*

*Then obviously* $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = 1/2$*. Moreover,*

$$\begin{aligned}\mathbf{P}(A \cap B) &= 1/4 = \mathbf{P}(A)\,\mathbf{P}(B)\\ \mathbf{P}(A \cap C) &= 1/4 = \mathbf{P}(A)\,\mathbf{P}(C)\\ \mathbf{P}(B \cap C) &= 1/4 = \mathbf{P}(B)\,\mathbf{P}(C)\,.\end{aligned}$$

*In other words,* $A, B$ *and* $C$ *are pairwise independent. However,*

$$\mathbf{P}(A \cap B \cap C) = 1/4 \neq 1/8 = \mathbf{P}(A)\,\mathbf{P}(B)\,\mathbf{P}(C)\,.$$

*Something is missing for the complete independence of* $A, B$ *and* $C$*. This leads us to the following definition:*

**Definition 9.** *Events $A_1, A_2, \ldots, A_n$ are mutually independent if, for any subcollection $\{j_1, \ldots, j_n\} \subset \{1, \ldots, n\}$,*

$$\mathbf{P}(A_{j_1} \cap \ldots \cap A_{j_n}) = \mathbf{P}(A_{j_1}) \times \ldots \times \mathbf{P}(A_{j_n}) \,.$$

Mutual independence implies pairwise independence, but the converse is not true.

Note that for mutually independent events $A_1, \ldots, A_n$, the following collection of events are also mutually independent:

$$\begin{gathered} \bar{A}_1, A_2, A_3, \ldots, A_n \\ \bar{A}_1, \bar{A}_2, A_3, \ldots, A_n \\ A_1 \cap A_2, A_3, \ldots, A_n \\ A_1 \cup A_2, A_3, \ldots, A_n \end{gathered}$$

**Exercise 5.** *Show that* $A \cup B$ *and* $C$ *are independent if* $A, B$ *and* $C$ *are mutually independent.*

*Prove this as well for* $\bar{A}$ *and* $B \cap \bar{C}$*. Is this true for* $\bar{A}, B$ *and* $\bar{C}$*?*

# Law of Total Probability

**Definition 10.** *A partition of $\Omega$ is a collection of disjoint and exhaustive events $A_1, A_2, \ldots, A_n$, that is*

- $A_i \cap A_j = \emptyset$ *for all* $i \neq j$
- $\bigcup_i A_i = \Omega$ .

Note that the simplest partition is of the form $(A, \bar{A})$.

Let $B$ be some event. Then

$$B = B \cap \Omega = B \cap (\bigcup_i A_i) = \bigcup_i (B \cap A_i) \, .$$

It follows from the third axiom that

$$\mathbf{P}(B) = \mathbf{P}\left(\bigcup_i (B \cap A_i)\right) = \sum_i \mathbf{P}(B \cap A_i) = \sum_i \mathbf{P}(B \mid A_i)\mathbf{P}(A_i) \, .$$

This brings us to the law of total probability:

**Law of Total Probability (LTP).**

If $A_1, A_2, \ldots, A_n$ form a partition of $\Omega$, then

$$\mathbf{P}(B) = \sum_i \mathbf{P}(B \cap A_i) = \sum_i \mathbf{P}(B \mid A_i)\mathbf{P}(A_i)\,.$$

**Exercise 6.** *Suppose we have a medical test to test whether or not a patient has a disease. Suppose*

- *If the patient has the disease, the result of the test is positive with probability 0.95*
- *If the patient is healthy, the test is positive with probability 0.01*

*Suppose that 5 in 1000 suffer from this disease. What is the probability that a (randomly chosen) patient will test positive?*

# Bayes Theorem

Same set up as for the LTP: suppose we have a partition $A_1, A_2, \ldots, A_n$ of $\Omega$, and some event $B$. Then

$$\begin{aligned}
\mathbf{P}(A_i \mid B) &= \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)} \text{ (definition of conditional probability)} \\
&= \frac{\mathbf{P}(B \mid A_i)\mathbf{P}(A_i)}{\mathbf{P}(B)} \text{ (multiplication rule)} \\
&= \frac{\mathbf{P}(B \mid A_i)\mathbf{P}(A_i)}{\sum_{k=1}^{n} \mathbf{P}(B \mid A_k)\mathbf{P}(A_k)}
\end{aligned}$$

This is Bayes Theorem.

**Exercise 7.** *Consider a multiple choice exam that has $m$ choices of answer each question. Assume that the probability that a student knows the correct answer to a question is $p$. A student that doesn't know the correct answer marks an answer at random. Suppose that the answer marked to a particular question was correct. What is the probability that the student knows the right answer?*

# Random Variables

In a random experiment, we are sometimes (often) interested in some function of the outcome, rather than the actual outcome itself.

A random variable is a quantity $X = X(\omega)$ whose value depend on the outcome of our random experiment.

**Definition 11.** *Consider a random experiment $\mathcal{E}$ with sample space $\Omega$. A (real-valued) function $X$ on $\Omega$ such that for any real $x$ the set*

$$\{X \leq x\} = \{\omega \in \Omega \mid X(\omega) \leq x\}$$

*is an event is called a Random Variable (RV). The state space $S_X$ of $X$ is the set of possible values of $X$,*

$$S_X = \{x \mid X(\omega) = x, \quad \omega \in \Omega\}.$$

**Remarks:**

(a) The term RV is unfortunate since $X$ is neither random nor a variable.

(b) RVs are denoted using capital letters ($X, Y$ and $Z$ most commonly), and the values they take by lower case letters (e.g. $x, y, z$). Thus, $X$ is a function and $x$ a real number.

(c) $X$ is not necessarily a one-to-one function: there may be several outcomes $\omega \in \Omega$ such that $X(\omega) = x$.

**Example 5.** *Consider the toss of three (unbiased) coins. The sample space of this experiment is* $\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$*. Let* $X$ *be the number of heads. The possible values of* $X$ *are 0,1,2 and 3 (state space* $S_X = \{0, 1, 2, 3\}$*), and* $X(\{htt\}) = X(\{tht\}) = X(\{tth\}) = 1$*.*

*Alternatively, since we are only interested in the number of heads, we can take* $\Omega = \{0, 1, 2, 3\} = S_X$*. [*$\rightarrow$ *Not always necessarily a good thing to do – why?]*

The previous example shows that we can view the possible values of $x$ as new omegas.

What about a probability on this new space?

So far we have learnt how to deal with probabilities $\mathbf{P}$ defined on $(\Omega, \mathcal{F})$. Now what?

We need to consider two types of RVs: discrete and continuous. We start with discrete.

**Definition 12.** *A random variable $X$ is said to be discrete if its state space $S_X$ is finite, or countably infinite. [Usually a subset of $\{0, 1, 2, 3, \ldots\}$.]*

To specify a probability on $S_X$, we only need to know the probability of particular outcomes

$$p_X(x) := \mathbf{P}(X = x) := \mathbf{P}(\{\omega \mid X(\omega) = x\}) = \sum_{\omega \mid X(\omega)=x} \mathbf{P}(\{\omega\}) \, .$$

**Definition 13.** *For a discrete random variable $X$ with state space $S_X$, $p_X(x) = \mathbf{P}(X = x)$ for $x \in S_X$ is called the probability mass function or pmf of $X$.*

The pmf of $X$ 'stores' all information you need about $X$. In other words, knowing $p_X$ is enough to know everything you want about $X$.

We shift our interest from the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ towards the newly defined probability $p_X$ of the random variable $X$.

**Exercise 8.** *Roll a four-sided die twice. Let a random variable $X =$ the larger of the two face numbers appeared if they are different and the common value if they are the same. Describe the sample space $\Omega$ of this experiment, specify the probability $\mathbf{P}$ on it, and the probability $p_X$ induced by $X$.*

We have:

$$\begin{aligned}1 = \mathbf{P}(\Omega) &= \mathbf{P}(\{\omega \mid X(\omega) \in S_X\}) \\ &= \mathbf{P}(\{\omega \mid X(\omega) = x_1 \text{ or } x_2 \text{ or } x_3 \text{ or } \dots \}) \\ &= \sum_{x \in S_X} \mathbf{P}(\{\omega \mid X(\omega) = x\}) \text{ (union of disjoint events)} \\ &= \sum_{x \in S_X} \mathbf{P}(X = x) = \sum_{x \in S_X} p_X(x).\end{aligned}$$

Obviously, $p_X(x) \geq 0$, $\forall x \in S_X$ since by definiiton $p_X(x) = \mathbf{P}(\{\omega \mid X(\omega) = x\})$.

**Theorem 2.** *The pmf $p_X(x)$ of a discrete random variable $X$ satisfies the following properties*

*(i)* $p_X(x) \geq 0$, *for all* $x \in S_X$

*(ii)* $\sum_{x \in S_X} p_X(x) = 1$

**Remark:** Two RVs can have the same pmf (we also say the same law) without being equal!

**Example 6.** *Roll two dice, one red and one black. Let*

$$\begin{aligned} X &= \textit{ number obtained on the red die} \\ Y &= \textit{ number obtained on the black die}\,. \end{aligned}$$

*Here*

$$\begin{aligned} \Omega &= \{(1,1),\ldots,(6,6)\} \\ S_X &= \{1,2,3,4,5,6\} \\ S_Y &= \{1,2,3,4,5,6\} = S_X\,. \end{aligned}$$

*Obviously*

$$p_X(k) = p_Y(k) = 1/6, \qquad k = 1,\ldots,6\,.$$

*Thus $X$ and $Y$ have the same law, but we do not have $X = Y$, which would mean $X(\omega) = Y(\omega)$ for all $\omega$. [In words – what does $X(\omega) = Y(\omega)$ mean?]*

Another way to specify the distribution of a RV is via its cumulative distribution function.

**Definition 14.** *The cumulative distribution function or cdf $F_X(x)$ of a discrete random variable $X$ is a function from $\mathbb{R}$ to $[0,\ 1]$ defined by*

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

We notice that for a discrete RV, the cdf $F_X(x) = \sum_{y \leq x} p_X(y)$ is a pure jump function, with $p_X(x)$ being the jump in $F_X$ at $x$.

Other important properties of the cdf can easily extracted:

(i) $0 \leq F_X(x) \leq 1$ since it is a probability ($F_X(x) = \mathbf{P}(X \leq x)$).

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

(iii) For any $x < y$, $\mathbf{P}(x < X \leq y) = F_X(y) - F_X(x)$.

Indeed, $\{X \leq y\}$ can be expressed as a union of two disjoint events $\{X \leq x\} \cup \{x < X \leq y\}$, so that $\mathbf{P}(X \leq x) + \mathbf{P}(x < X \leq y) = \mathbf{P}(X \leq y)$.

(iv) $F_X(x)$ is non-decreasing.

Indeed, for any $x < y$, $F_X(y) - F_X(x) = \mathbf{P}(x < X \leq y) \geq 0$.

(v) $F_X(x)$ is right-continuous.

So, for discrete RVs, we can specify its distribution by means of its pmf or cdf, and knowing one, we can always compute the other. So why bother? Because for continuous RVs, pmf does not work, while cdf works. [Details coming up soon.]

# Bernoulli trial

In the simplest case, a random experiment has only two possible outcomes: a 'success' and a 'failure'. Such a random experiment is called a *Bernoulli trial*.

**Example 7.** *Coin tossing, die rolling (e.g. 'success'=1), any random experiment where 'success' corresponds to the occurrence on a certain event $A$.*

Let $p$ be the probability of a success. A Bernoulli RV corresponds to the number of success in a single Bernoulli trial, so

$$X = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p\,. \end{cases}$$

The pmf of X is thus $p_X(1) = p = 1 - p_X(0)$. We say that $X$ has a Bernoulli distribution and we write

$$X \sim \mathcal{B}(p)\,.$$

# Sequence of Bernoulli trials

Consider now a sequence of $n$ independent Bernoulli trials (the outcome of one trial does not affect the outcome of another trial). We are interested in different aspects of the observed sequence of successes and failures, such as

- The number of successes (Binomial distribution).
- The number of failures before the first success (Geometric distribution).

The sample space $\Omega$ can be taken to be the set of all sequences of the form

$$\omega = S\,F\,S\,S\,F\ \ldots\ F\,F\,S\,,$$

and

$$\mathbf{P}(\{\omega\}) = p^{\#\text{successes}}(1-p)^{\#\text{failures}}\,.$$

Consider first $X =$number of successes. What is the distribution of $X$?

**Example 8.** *Take $n = 3$. The probability for each possible outcome of $X$ is given below*

| $X$ | *outcome* | *probability* |
|---|---|---|
| *3* | $S\,S\,S$ | $p^3$ |
| *2* | $S\,S\,F$ | $p^2(1-p)$ |
| | $S\,F\,S$ | $p^2(1-p)$ |
| | $F\,S\,S$ | $p^2(1-p)$ |
| *1* | $S\,F\,F$ | $p(1-p)^2$ |
| | $F\,S\,F$ | $p(1-p)^2$ |
| | $F\,F\,S$ | $p(1-p)^2$ |
| *0* | $F\,F\,F$ | $(1-p)^3$ |

*We see that*

$$p_X(0) = (1-p)^3\,, \quad p_X(1) = 3p(1-p)^2\,, \quad p_X(2) = 3p^2(1-p)\,, \quad p_X(3) = p^3\,.$$

The distribution of $X$ can be summarized into

$$p_X(x) = N(3,x)\, p^x (1-p)^{3-x}\,, \qquad \text{for } x = 0, 1, 2, 3\,.$$

The coefficient $N(3,x)$ gives the number of ways of selecting $x$ positions for the $x$ successes in the $n = 3$ trials. To generalize the previous distribution for any $n$, we need an expression for $N(n,x)$.

We can verify that

$$N(n,x) = \binom{n}{x} = \frac{n!}{(n-x)!\, x!}\,.$$

[How many ways can you choose $k$ objects out of $n$ distinct objects if order matters? if order does not matter?]

We say that $X$ has Binomial distribution with parameters $n \geq 1$ and $p \in [0,\, 1]$ if its pmf is given by

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}\,, \quad x = 0, 1, \ldots, n\,,$$

and we write

$$X \sim \mathrm{Bi}(n, p)\,.$$

[Check that the pmf sums to one.]

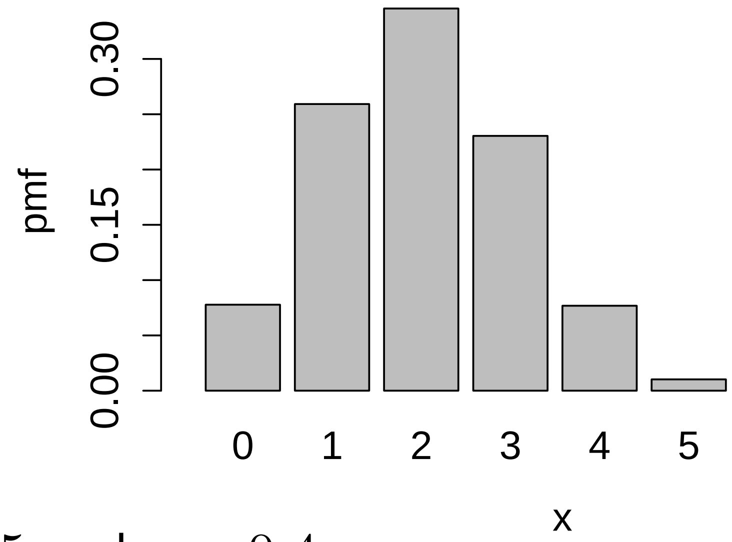
**Exercise 9.** *Tay-Sachs disease is a rare disease that progressively destroys neurons in the brain and spinal cord. This condition is inherited in a recessive pattern, meaning that a child must have a copy of the mutated gene from each parent to have the disease. Parents with only one copy of the mutated gene do not show symptoms of the disease. If such parents have 4 children, what is the probability that at least one of them present signs of the condition?*

## Shape of the Binomial distribution

Look at the ratio of successive binomial probabilities

$$r(x) = \frac{p_X(x)}{p_X(x-1)} = \frac{\binom{n}{x}}{\binom{n}{x-1}} \frac{p^x(1-p)^{n-x}}{p^{x-1}(1-p)^{n-x+1}} = \frac{(n+1)/x - 1}{1/p - 1}\,,$$

and compare this ratio with the value 1. The binomial distribution has a single peak, and typically looks like.



This one is for $n = 5$ and $p = 0.4$.

# Geometric distribution

Back to our random experiment consisting in the sequence of independent Bernoulli trails, with probability of success $p$.

Define $T$ to be the number of failures before the first success. What is the distribution of $T$?

First, note that this number can be arbitrarily large. The set of possible values of $T$ is thus $S_T = \{0, 1, 2, 3, \ldots\}$. Then

$$\begin{aligned} p_T(0) &= \mathbf{P}(T = 0) = \mathbf{P}(S) = p \\ p_T(1) &= \mathbf{P}(T = 1) = \mathbf{P}(F\,S) = (1 - p)\,p \\ p_T(2) &= \mathbf{P}(T = 2) = \mathbf{P}(F\,F\,S) = (1 - p)^2 p\,, \end{aligned}$$

and more generally,

$$p_T(t) = (1 - p)^t\, p\,, \quad t = 0, 1, 2, \ldots$$

We say that $T$ has a Geometric distribution with parameter $p$ if its pmf is given by

$$p_T(t) = (1-p)^t p\,, \quad t = 0, 1, 2, \dots\ ,$$

and we write

$$X \sim \mathrm{G}(p)\,.$$

[Check that the pmf sums to one.]

**Remark:** The Geometric distribution can be defined in a slightly different way, by counting the number of trials until the first success instead of the number of failures before it. Can you derive the pmf in this case?

**Lack of memory property.** Given the first $x$ trials are failures, what is the distribution of the 'residual time' $T - x$ until the occurrence of the first success?

# Negative binomial distribution

What about the distribution of the number of failures before the occurrence of the $r$-th success (call it $X$)?

Suppose we observe $x$ failures, and $r$ successes,

$$\omega = F\,F\,F\,F\,F\,F\,F\,F\,F\,F\,S\,S\,S\,S\,S\,|\,S\,,$$

corresponding to the event $\{X = x\}$. There are $\binom{x+r-1}{r-1}$ of such arrangements, leaving the final $r$-th success. The probability of this sequence is $(1-p)^x p^r$. Thus

$$p_X(x) = \mathbf{P}(X = x) = \binom{x+r-1}{r-1}(1-p)^x p^r\,, \quad x = 0, 1, 2, 3, \ldots\,.$$

We say that $X$ has a *negative binomial* distribution, and we write

$$X \sim \mathrm{NB}(r, p)\,.$$

# Poisson distribution

The Poisson distribution is the ‘continuous analogue’ of Bernoulli trials. A Poisson RV counts the number of successes occurring in continuous time. [$\rightarrow$ But it is still a discrete RV!]

Suppose we have $n$ independent Bernoulli trials. Consider time slots of length $1/n$, so that we are interested in counting the number of successes occurring from 0 to 1.

During each time slot, suppose that the probability of a success is proportional to $1/n$: the longer you wait, the more likely you obtain a success.

Let $X$ be the number of successes occurring from 0 to 1. Then

$$X \sim \mathrm{Bi}(n, \lambda/n)\,,$$

so that

$$\mathbf{P}(X = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^{x} \left(1 - \frac{\lambda}{n}\right)^{n-x}, \quad x = 0, 1, \ldots\,.$$

Now, shrink the length of time for each interval by taking $n$ larger and larger. In the limit,

$$\begin{aligned}\lim_{n\to\infty}\mathbf{P}(X=x) &= \lim_{n\to\infty}\frac{n!}{n^x(n-x)!}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\\ &= \frac{\lambda^x}{x!}e^{-\lambda}\,,\end{aligned}$$

since

$$\frac{n!}{n^x(n-x)!} = \frac{n(n-1)\ldots(n-x+1)}{n^x}\to 1 \text{ as } n\to\infty\,.$$

A discrete random variable X is said to have a Poisson distribution with parameter $\lambda>0$ if

$$p_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}\,, \qquad x=0,1,2,\ldots\,,$$

and we write

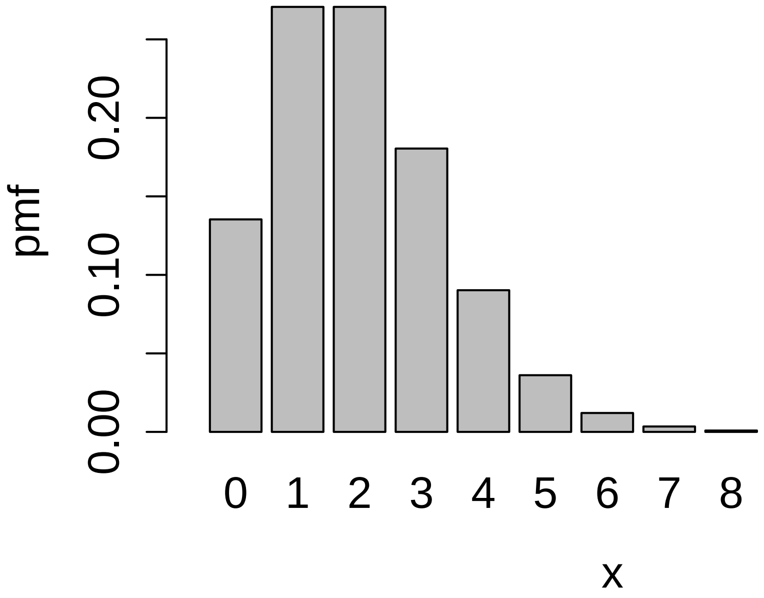
$$X\sim \mathrm{P}(\lambda)\,.$$

[Check that the pmf sums to one.]

Note that

$$r(x) = \frac{p_X(x)}{p_X(x-1)} = \frac{\lambda^x}{x!}\, e^{-\lambda}\, \frac{(x-1)!}{\lambda^{x-1}}\, e^{\lambda} = \frac{\lambda}{x}\,, \qquad \text{for } x = 1, 2, \ldots$$

so the pmf of a Poisson RV has a single peak located at $x = \lambda$, increases as a function of $x$ for $X \leq \lambda$ and decreases for $x \geq \lambda$. Typically,



This one is for $\lambda = 2$.

We already know we can approximate binomial probabilities using a Poisson distribution. [$\rightarrow$ This is how we introduced it!] More generally,

**Poisson Theorem.** If $X_n \sim \mathrm{Bi}(n, p_n)$ and $np_n \to \lambda$ as $n \to \infty$, then for all $x = 0, 1, 2, \ldots$

$$p_{X_n}(x) \to p_X(x)\,, \qquad \text{where} \qquad X \sim \mathrm{P}(\lambda)\,.$$

In other words, the distribution of the number of rare events [i.e. for small $p_n$] is approximatively Poisson.

# Uniform distribution

The uniform distribution formalizes the idea of equally likely outcomes.

Suppose there are only finitely many possible outcomes. A *uniform* random variable $X$ taking values in $S_X = \{m, m+1, \ldots, n\}$ has pmf

$$p_X(x) = \mathbf{P}(X = x) = \frac{1}{n - m + 1}\,, \quad x = m, \ldots, n\,.$$

We write

$$X \sim \mathrm{U}(m, n)\,.$$

In particular, if $X \sim \mathrm{U}(1, n)$, then $p_X(x) = 1/n$. For example, if you roll a fair die, the number obtained $\sim \mathrm{U}(1, 6)$.

[Picture of the uniform pmd and cdf.]

# Expectation

Recall the frequency interpretation of probability: when performing the same experiment $n$ times in the same conditions, we observes empirically that for an event $A$

$$\frac{n_A}{n} \approx \text{constant} = \mathbf{P}(A)\,,$$

where $n_A$ denotes the number of times event $A$ occurs.

Similarly, if we observe $n$ realizations $X_j$ of a RV, then it is also empirically observed that the *mean*

$$\frac{\sum_{j=1}^{n} X_j}{n} \approx \text{constant}$$

Suppose $X$ can only take finitely many values $x_1, \ldots, x_n$. Then

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n} X_j &= \frac{1}{n}(x_1 n_{\{X=x_1\}} + \ldots + x_n n_{\{X=x_n\}}) \\ &= \sum_{j=1}^{n} x_j \frac{n_{\{X=x_j\}}}{n} \\ &\approx \sum_{j=1}^{n} x_j p_X(j)\,,\end{aligned}$$

which leads to the following definition

**Definition 15.** *The expected value or mean or first moment of a discrete random variable $X$ with state space $S_X$, denoted $\mathbf{E}\,(X)$ or $\mu_X$ is*

$$\mu_X = \mathbf{E}\,(X) = \sum_{x \in S_X} x p_X(x)\,.$$

This can be an infinite series, and so it should converge..

Note that in the case of a discrete sample space $\Omega$,

$$\mathbf{E}\left(X\right) = \sum_x x p_X(x) = \sum_x x \sum_{\omega | X(\omega)=x} \mathbf{P}(\{\omega\}) = \sum_{\omega\in\Omega} X(\omega)\mathbf{P}(\{\omega\})\,.$$

This expression is useful to derive the expected value of a transform $\psi(X)$ of $X$, for some (real-valued) function $\psi$.

Let $Y = \psi(X)$. Then

$$\mathbf{E}\left(\psi(X)\right) = \mathbf{E}\left(Y\right) = \sum_y y p_Y(y)\,.$$

It seems that we need to compute the pmf of $Y$. Well, using the expression above, we see that

$$\begin{aligned}\mathbf{E}\left(Y\right) &= \sum_{\omega\in\Omega} Y(\omega)\,\mathbf{P}(\{\omega\}) \\ &= \sum_{\omega\in\Omega} \psi(X(\omega))\,\mathbf{P}(\{\omega\})\,.\end{aligned}$$

$$\begin{aligned}\mathbf{E}\,(Y) &= \sum_x \sum_{\omega|\psi(\omega)=x} \psi(X(\omega))\,\mathbf{P}(\{\omega\}) \\ &= \sum_x \psi(x) \sum_{\omega|\psi(\omega)=x} \mathbf{P}(\{\omega\}) \\ &= \sum_x \psi(x)\,p_X(x)\,.\end{aligned}$$

**Theorem 3.** *Let $X$ be a discrete RV with state space $S_X$, pmf $p_X(x)$, and let $\psi$ be some real-valued function. Then,*

$$\mathbf{E}\,(\psi(X)) = \sum_{x\in S_X} \psi(x)\,p_X(x)\,.$$

*(Again, provided the sum converges..)*

**Exercise 10.** *Let $X$ be a discrete RV with pmf*

$$p_X(x) = 1/3\,, \quad x \in S_x = \{-1, 0, 1\}\,,$$

*and let $\psi(x) = x^2$. Find $\mathbf{E}\left(\psi(X)\right)$.*

**Properties of expectation.**

(i) For a constant $c$, $\mathbf{E}\left(c\right) = c$.

Indeed, $\mathbf{E}\left(c\right) = \sum_x x p_X(x) = c \cdot 1 = c$.

(ii) $\mathbf{E}\left(\lambda X\right) = \lambda \mathbf{E}\left(X\right)$ for any $\lambda \in \mathbb{R}$.

In particular, $\mathbf{E}\left(-X\right) = -\mathbf{E}\left(X\right)$.

Consequence: Expectation is linear. Let $f(x) = \lambda x + c$. Then $\mathbf{E}\left(f(X)\right) = \mathbf{E}\left(\lambda X + c\right) = \lambda \mathbf{E}\left(X\right) + c = f(\mathbf{E}\left(X\right))$.

*Warning:* $\mathbf{E}\left(f(X)\right) \neq f(\mathbf{E}\left(X\right))$ in general.

(iii) $|\mathbf{E}\left(X\right)| \leq \mathbf{E}\left|X\right|$ (triangle inequality).

(iv) If $X$ can only take non-negative integer values $0, 1, 2, 3, \ldots$, then

$$\mathbf{E}\left(X\right) = \sum_{n=0}^{\infty} (1 - F_X(n)) \, .$$

# Describing a distribution

Recall that the pmf and cdf completely specify a distribution. What if we are interested in a brief summary of a given distribution?

Using words or just a picture is nice but not (precise) enough. Instead, we need a few numbers which give a reasonable idea of what the distribution looks like.

We have already seen one example: the mean $\mathbf{E}(X)$, which can be interpreted as the centre of mass of our distribution.

Other features to be summarized are for example spread (how likely are we to deviate from the mean?), or symmetry (is the distribution skewed?).

The most widespread measures of spread are the *variance* and *standard deviation*.

**Definition 16.** *The variance of $X$ is defined by*

$$Var(X) = \mathbf{E}(X - \mathbf{E}(X))^2 \,,$$

*and the standard deviation is $\sigma_X = \sqrt{Var(X)}$.*

*Why?* By definition, denoting $\mu_X = \mathbf{E}\,(X)$, and assuming that $X$ takes only finitely many values,

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{x \in S_X} (x - \mu_X)^2 p_X(x) \\ &= (x_1 - \mu_X)^2 p_X(x_1) + \ldots + (x_n - \mu_X)^2 p_X(x_n) \,,\end{aligned}$$

which is a weighted sum of the squares of the differences $(x_1 - \mu_X), \ldots, (x_k - \mu_X)$. Var$(X)$ indeed measures the variability of $X$ about its mean: the smaller Var$(X)$ and the more often $X$ stay near $\mathbf{E}\,(X)$, whereas a large value of Var$(X)$ is an indicator that $X$ varies about $\mathbf{E}\,(X)$ a lot.

So why bother introducing $\sigma_X$? Because the standard deviation and $X$ are expressed in the same unit.

**Example 9.** *Suppose the pmf of $X$ is given by*

$$p_X(-1) = p_X(0) = p_X(1) = 1/3\,.$$

*Then $\mathbf{E}\,(X) = 0$ and Var$(X) = \mathbf{E}\,(X^2) = 2/3$. Suppose now we are given another random variable $Y$ with pmf*

$$p_Y(-2) = p_Y(0) = p_Y(2) = 1/3\,.$$

*Then $\mathbf{E}\,(Y) = 0$ and Var$(Y) = \mathbf{E}\,(Y^2) = 8/3 >$Var$(X)$. Why? [Draw a picture.]*

**Properties of variance.**

(i) Var$(X) \geq 0$.

Indeed, by definition $\mathrm{Var}(X) = \mathbf{E}\,(X - \mu_X)^2 = \sum (x - \mu_X)^2 p_X(x)$ (sum of positive terms).

(ii) Var$(X) = \mathbf{E}\,(X^2) - (\mathbf{E}\,(X))^2$.

By definition,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbf{E}\,(X - \mathbf{E}\,(X))^2 \\
&= \mathbf{E}\,(X^2 - 2X\mathbf{E}\,(X) + (\mathbf{E}\,(X))^2) \\
&= \mathbf{E}\,(X^2) - 2\mathbf{E}\,(X)\mathbf{E}\,(X) + (\mathbf{E}\,(X))^2 \\
&= \mathbf{E}\,(X^2) - (\mathbf{E}\,(X))^2
\end{aligned}$$

Note that since the variance is non-negative, we get $\mathbf{E}\,(X^2) \geq (\mathbf{E}\,(X))^2$.

$\mathbf{E}\,(X^2)$ is called the *second moment* of $X$.

(iii) If $Y = aX + b$, then Var$(Y) = a^2 \mathrm{Var}(X)$, and $\sigma_Y = |a|\sigma_X$.

Indeed, $\mu_Y = a\mu_X + b$. Next

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbf{E}\,(Y - \mu_Y)^2 \\ &= \mathbf{E}\,(aX + b - (a\mu_X + b))^2 \\ &= \mathbf{E}\,(a^2(X - \mu_X)^2) \\ &= a^2\,\mathbf{E}\,(X - \mu_X)^2 \\ &= a^2\,\mathrm{Var}(X)\,.\end{aligned}$$

(iv) If $X$ has mean $\mu$ and standard deviation $\sigma$, then

$$Z = \frac{X - \mu}{\sigma}$$

has mean 0 (centered) and variance 1 (standardised).

**Example 10.** *Mean and variance of the Uniform distribution.*

*Let* $X \sim U(1, m)$. *Then*

$$\begin{aligned}
\mathbf{E}\,(X) &= \sum_{x=1}^{m} x\,\frac{1}{m} = \frac{1}{m}\sum_{x=1}^{m} x = \frac{1}{m}\,\frac{m(m+1)}{2} = \frac{m+1}{2}\,,\\
\mathbf{E}\,(X^2) &= \sum_{x=1}^{m} x^2\,\frac{1}{m} = \frac{1}{m}\,\frac{m(m+1)(2m+1)}{6} = \frac{(m+1)(2m+1)}{6}\,,\\
Var(X) &= \frac{1}{12}(m^2-1)\,.
\end{aligned}$$

# Continuous random variables

**Definition 17.** *Let $X$ be a RV. If $X$ can take any value in an interval, we say that $X$ is a continuous random variable.*

*More formally, the state space $S_X$ of $X$ is uncountable (usually $S_X = \mathbb{R}$).*

**Example 11.** *Temperature, height, weight, amount of rainfalls, time to failure, waiting time, etc.*

It is not possible to assign directly a probability to every possible value of a continuous RV (there are too many of them). In Part 1, we underlined the importance of $\sigma$-algebras, and we already mentioned that the power set of $\mathbb{R}$ contains monsters. For example, working with $\mathcal{P}(\mathbb{R})$ would result in the impossibility to define a uniform distribution on the interval $[0, 1]$, which is a disaster!

Instead, we assign probabilities to intervals, and we need *all* open intervals $(a,\, b)$ to be there. Taking the *smallest* $\sigma$-algebra containing such intervals is meaningful. This is the so-called *Borel $\sigma$-algebra*, denoted $\mathcal{B}(\mathbb{R})$, or simply $\mathcal{B}$.

Note that since all $A_n = (x - n,\, x) \in \mathcal{B}(\mathbb{R})$, for $n = 1, 2, \ldots$ and any given $x$, then

$$\begin{aligned} A &= \bigcup_{n=1}^{\infty} A_n = (-\infty, x) \in \mathcal{B}(\mathbb{R}) \\ \bar{A} &= [x, \infty) \in \mathcal{B}(\mathbb{R}) \,. \end{aligned}$$

Moreover, all $B_n = (y,\, y + n) \in \mathcal{B}(\mathbb{R})$ as well, so that

$$\begin{aligned} B &= \bigcup_{n=1}^{\infty} B_n = (y,\, \infty) \in \mathcal{B}(\mathbb{R}) \\ \bar{B} &= (-\infty, y] \in \mathcal{B}(\mathbb{R}) \\ \bar{A} \cap \bar{B} &= [x,\, y] \in \mathcal{B}(\mathbb{R}) \,, \end{aligned}$$

and in particular, taking $x = y$, we see that $\{x\} \in \mathcal{B}(\mathbb{R})$.

When you see $A \in \mathcal{B}(\mathbb{R})$, think of $A$ as an interval!

To specify the distribution of $X$, we thus want to specify the probability of intervals ⇒ cdf!

**Definition 18.** *Let $X$ be a continuous RV. A function $f_X(x)$ such that*

$$F_X(x) = \mathbf{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)du\,,$$

*is called a probability density function (pdf) of $X$.*

**Properties.**

(i) By definition, at points $x$ where $F_X(x)$ is differentiable,

$$f_X(x) = \frac{dF_X(x)}{dx}\,.$$

(ii) $F_X(x)$ is non-decreasing ⇒ $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.

(iii) $\int_{-\infty}^{\infty} f_X(u)du = \lim_{x\to\infty} F_X(x) = 1$.

(iv) For $x < y$,

$$\begin{aligned}\mathbf{P}(x < X \leq y) &= \mathbf{P}(X \leq y) - \mathbf{P}(X \leq x) \\ &= \int_{-\infty}^{y} f_X(u)du - \int_{-\infty}^{x} f_X(u)du \\ &= \int_{x}^{y} f_X(u)du\,,\end{aligned}$$

which is the area under $f_X(x)$ bounded by $x$ and $y$.

Consequence: $\mathbf{P}(X = x) = \mathbf{P}(x \leq X \leq x) = \int_x^x f_X(u)du = 0$ if $f_X$ has no mass at $x$ (and thus $F_X$ is continuous at $x$). Moreover,

$$\mathbf{P}(x < X \leq y) = \mathbf{P}(x \leq X \leq y) = \mathbf{P}(x \leq X < y) = \mathbf{P}(x < X < y)\,.$$

**Important:** $F_X$ is a probability. $f_X$ is *not* a probability, it does *not* need to be less than 1.

But then... if $f_X(x)$ is *not* the probability of the event $\{X = x\}$, then what is it?

Let's have a look at the probability that $X$ takes values in a small neighborhood around $x$

$$\left[x - \frac{\epsilon}{2}, x + \frac{\epsilon}{2}\right], \qquad \epsilon > 0 \text{ small.}$$

Then

$$\begin{aligned} \mathbf{P}(X \approx x) &= \mathbf{P}\left(X \in \left[x - \frac{\epsilon}{2}, x + \frac{\epsilon}{2}\right]\right) \\ &= \int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} f_X(u)du \\ &\approx f_X(x)\epsilon\,, \end{aligned}$$

where we used the fact that $f_X$ is approximately constant in a small neighbourhood of $x$.

**Infinitesimal notation:** $\mathbf{P}(X \in du) = f_X(u)du = dF_X(u)$, so that

$$\mathbf{P}(x \le X \le y) = \int_x^y \mathbf{P}(X \in du) = \int_x^y dF_X(u) = \int_x^y f_X(u)du\,.$$

**Example 12.** *Let $X$ be a continuous RV with density*

$$f_X(x) = \begin{cases} 0 & \textit{if} \quad x < 0 \\ 4x & \textit{if} \quad 0 \leq x < 1/2 \\ 1/3 & \textit{if} \quad 1/2 \leq x < 2 \\ 0 & \textit{if} \quad x \geq 2 \end{cases}$$

*(Check that $f_X$ can be regarded as a pdf)*

*The cdf of $X$ is $F_X(x) = \mathbf{P}(X \leq x)$, can be found to be*

$$F_X(x) = \begin{cases} \int_{-\infty}^{x} 0du = 0 & \textit{if} \quad x < 0 \\ \int_{-\infty}^{0} 0du + \int_{0}^{x} 4udu = 2u^2 & \textit{if} \quad 0 \leq x < 1/2 \\ \int_{-\infty}^{0} 0du + \int_{0}^{1/2} 4udu + \int_{1/2}^{x} \frac{1}{3}du = \frac{1}{2} + \frac{1}{3}(x - \frac{1}{2}) & \textit{if} \quad 1/2 \leq x < 2 \\ \int_{-\infty}^{0} 0du + \int_{0}^{1/2} 4udu + \int_{1/2}^{2} \frac{1}{3}du + \int_{2}^{x} 0du = 1 & \textit{if} \quad x \geq 2\,. \end{cases}$$

*[Plots required.]*

Note that we can use either $f_X$ or $F_X$ to compute various probabilities of $X$.

What is $\mathbf{P}(1/4 < X \leq 3/2)$?

(i) Using pdf ...

$$\begin{aligned} &= \int_{1/4}^{1/2} f_X(u)du + \int_{1/2}^{3/2} f_X(u)du = \int_{1/4}^{1/2} 4udu + \int_{1/2}^{3/2} \frac{1}{3}du \\ &= 2\left[\frac{1}{4} - \frac{1}{16}\right] + \frac{1}{3}\left[\frac{3}{2} - \frac{1}{2}\right] \\ &= 17/24 \\ &\approx 0.708\,. \end{aligned}$$

(ii) Using cdf..

$$\begin{aligned} \mathbf{P}(1/4 < X \leq 3/2) &= F_X(3/2) - F_X(1/4) \\ &= \frac{1}{2} + \frac{1}{3}(\frac{3}{2} - \frac{1}{2}) - 2\frac{1}{16} \\ &\approx 0.708\,. \end{aligned}$$

## Comparison between discrete and continuous RVs

| Discrete RV | Continuous RV |
|---|---|
| $X$ has at most countably many values | $X$ has uncountably many possible values |
| pmf $\quad p_X(x) = \mathbf{P}(X = x)$ | pdf $\quad f_X(x) \geq 0$ (not a proba) |
| $\sum_{x \in S_X} p_X(x) = 1$ | $\int_{-\infty}^{+\infty} f_X(x) dx = 1$ |
| cdf $\quad F_X(x) = \sum_{y \leq x} p_X(y)$ | cdf $\quad F_X(x) = \int_{-\infty}^{x} f_X(u) du$ |
| $F_X$ is discontinuous, with jumps<br>at possible values of $X$ | $F_X$ is continuous if $f_X$ has no mass |
| $\mathbf{P}(x < X \leq y) = \sum_{x < u \leq y} p_X(u)$<br>$= F_X(y) - F_X(x)$ | $\mathbf{P}(x < X \leq y) = \int_{x}^{y} f_X(u) du$<br>$= F_X(y) - F_X(x)$ |

# Summarising a continuous RV

For continuous RVs, definitions associated with the mathematical expectation are the same as those for discrete RVs, except that integrals are used to replace summations.

**Definition 19.** *(i) The expectation or mean or first moment of $X$ is*

$$\mu_X = \mathbf{E}\,(X) = \int x\, f_X(x)\, dx\,.$$

*(ii) The variance of $X$ is*

$$\sigma^2 = \, Var(X) = \mathbf{E}\,(X - \mu_X)^2 = \int (x - \mu_X)^2 f_X(x)\, dx\,.$$

*(iii) The standard deviation of $X$ is*

$$\sigma_X = \sqrt{Var(X)}\,.$$

(iv) The $(100\,p)$-*th percentile* of $X$ is a number $\pi_p$ such that

$$p = \int_{-\infty}^{\pi_p} f_X(x)dx = F_X(\pi_p)$$

For continuous cdfs, $\pi_p = F_X^{-1}(p)$. [Picture.]

To cover cases where $F_X$ is not continuous, we need the more formal definition $F_X(\pi_p - 0) < p \leq F_X(\pi_p)$. [Picture.]

Examples of percentiles include

(a) $p = 0.5 \Rightarrow \pi_{0.5} =: m$ is the *median* of a RV.

(b) The $\pi_p$ corresponding to $p = 0.25, 0.5, 0.75$ are the *quartiles* of a RV.

Quantiles are useful summary of distributions, and are often used in statistics.

**Remarks:**

(i) Connection with expected value of a discrete RV.

Partition $S_X$ into $n$ intervals of length $\epsilon$. Then

$$\mathbf{E}\,(X) = \int_{S_X} x\, f_X(x)dx \approx \sum_{i=1}^{n} x_i f_X(x_i)\epsilon \approx \sum_{i=1}^{n} x_i\, \mathbf{P}(x_i \le X < x_i + \epsilon)\,.$$

(ii) If $X \ge 0$, an alternative expression for the mean is

$$\mathbf{E}\,(X) = \int_0^\infty (1 - F_X(x))dx\,.$$

Compare this expression with the alternative expression of $\mathbf{E}\,(X)$ obtained in the discrete case.

# Uniform distribution

By analogy to the discrete case, a *continuous uniform* RV has a constant pdf over $S_X$ (obviously, $S_X$ needs to be bounded).

Let $a < b$ be real numbers. A continuous random variable $U$ with pdf

$$f_U(u) = \begin{cases} \dfrac{1}{b-a} & \text{if} \quad u \in (a,\, b) \\ 0 & \text{elsewhere,} \end{cases}$$

is said to have a uniform distribution over the interval $(a,\, b)$, and we write

$$U \sim \mathcal{U}(a,\, b)\,.$$

The cdf of $U$ is then

$$F_U(u) = \mathbf{P}(U \leq u) = \int_a^u \frac{1}{b-a}\, dv = \frac{u-a}{b-a} \;\text{ if }\; a < u < b,$$

and $F_U(u) = 0$ for $u < a$, $F_U(u) = 1$ for $u > b$. [Plots required.]

Moments of $U$

$$\mathbf{E}\,(U) = \int_a^b \frac{u}{b-a}du = \frac{1}{b-a}\left[\frac{u^2}{2}\right]_a^b = \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{a+b}{2}\,.$$

$$\mathbf{E}\,(U^2) = \int_a^b \frac{u^2}{b-a}du = \frac{1}{b-a}\left[\frac{u^3}{3}\right]_a^b = \frac{1}{3}\frac{b^3-a^3}{b-a}\,.$$

$$\mathrm{Var}(U) = \mathbf{E}\,(U^2) - (\mathbf{E}\,(U))^2 = \frac{(b-a)^2}{12} \text{ (after simplifications).}$$

An important special case is the standard uniform distribution $\mathcal{U}(0,1)$. It plays a central role in computer generation of random numbers. Namely, any random number generated by a computer program is generated through generation of a standard uniform number.

**Exercise 11.** *Suppose a value $x$ is chosen at random in the interval $[2,\, 6]$, so that $X \sim \mathcal{U}(2,\, 6)$. This value divides the interval $[2,\, 6]$ into two subintervals of respective lengths $X - 2$ and $6 - X$. Introduce*

$$Y = \max(X - 2,\, 6 - X)\,,$$

*corresponding to the length of the largest interval. Distribution of $Y$?*

# Exponential distribution

Recall that a geometric RV counts the number of failures (on a discrete time grid) before the occurrence of the first success.

The exponential distribution is the continuous version of a geometric RV. It models the waiting time until the occurrence of a given event.

Let $T =$ waiting time to the first event (taking values in $[0, \infty)$).

First consider a fine time grid (step$=1/n$), then proceed to the limit.

The probability of success on a Bernoulli trial is $\lambda/n$. By time $t$, we have completed $nt$ trials.

The probability that there is no success in a time period $[0, t]$ is then

$$\left(1 - \frac{\lambda}{n}\right)^{nt} \to e^{-\lambda t} \text{ as } n \to \infty\,,$$

so that

$$\mathbf{P}(T > t) = e^{-\lambda t}\,,$$

and

$$\mathbf{P}(T \le t) = 1 - e^{-\lambda t}\,, \qquad t \ge 0\,, \quad \lambda > 0\,.$$

A continuous random variable $T$ with such a cdf is called *exponential*, and we write

$$T \sim \mathrm{Exp}(\lambda)\,.$$

The pdf of $T$ can easily be calculated to be

$$f_T(t) = \lambda e^{-\lambda t}\,, \qquad t \ge 0\,, \quad \lambda > 0\,.$$

[Picture.]

Moments of the exponential distribution?

## Lack of memory property of the exponential distribution

Suppose that the life of a certain light bulb has an exponential distribution with a mean life of 500 hours. Let $T$ denote the life of this component. Then

$$\mathbf{P}(T > t) = \int_t^\infty \frac{1}{500} e^{-u/500} du = e^{-t/500} \, .$$

Let $t_0 > 0$. The probability that the light bulb lasts an additional $t_0$ hours, given that it has lasted $t$ hours already...

$$\mathbf{P}(T > t + t_0 \mid T > t) = \frac{\mathbf{P}(T > t + t_0)}{\mathbf{P}(T > t)} = \frac{e^{-(t+t_0)/500}}{e^{-t/500}} = e^{-t_0/500} = \mathbf{P}(T > t_0)$$

... is the same as the probability of lasting $t_0$ hours when new!

This is what we refer to as the *memoryless property of the exponential distribution*.

(Compare with the geometric distribution.)

# Gamma distribution

The Gamma distribution models the waiting time until the arrival of the $r$-th occurrence of an event. As such, it is the continuous version of the negative binomial distribution.

The expression of its density involves the *Gamma function*, defined as

$$\Gamma(t) = \int_0^\infty y^{t-1} e^{-y} dy\,, \text{ for } t > 0\,.$$

A RV X is said to have a *Gamma distribution* with parameters $r > 0$ and $\lambda > 0$ if its pdf is given by

$$f_X(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda\, x}\,, \quad x \geq 0\,,$$

and we write

$$X \sim \gamma(r, \lambda)\,.$$

**Remarks:**

(i) $\Gamma(1) = \int_0^\infty e^{-y} dy = 1$. If $t > 1$,

$$\begin{aligned}\Gamma(t) &= \int_0^\infty y^{t-1} e^{-y} dy \\ &= \left[-y^{t-1} e^{-y}\right]_0^\infty + (t-1) \int_0^\infty y^{t-2} e^{-y} dy \quad \text{(IBP)} \\ &= (t-1) \int_0^\infty y^{t-2} e^{-y} dy \\ &= (t-1)\Gamma(t-1)\,.\end{aligned}$$

(ii) Consequence: if $n$ is a positive integer, then $\Gamma(n) = (n-1)!$. For this reason, the Gamma function is also called the *generalized factorial*.

(iii) $f_X \geq 0$ is indeed a pdf,
$\int_0^\infty f_X(x) dx = \frac{1}{\Gamma(r)} \int_0^\infty \lambda^r x^{r-1} e^{-\lambda x} dx = \frac{1}{\Gamma(r)} \int_0^\infty y^{r-1} e^{-y} dy = 1$.

(iv) Moments of the Gamma distribution: $\mathbf{E}(X) = r/\lambda$, $\mathrm{Var}(X) = r/\lambda^2$.

# Normal distribution

Perhaps the most famous probability distribution. It is also known as the Gaussian distribution.

Like the Poisson distribution, the normal distribution arises as a limit (as a cumulative effect of a large number of small independent random factors).

Also, it possesses a number of unique properties which makes it easy to handle (analytically), and very popular to model real phenomena.

**Definition 20.** *A continuous random variable $X$ with pdf*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty\,,$$

*is said to have a normal distribution with parameters $\mu$ and $\sigma^2 > 0$, and we write*

$$X \sim \mathcal{N}(\mu, \sigma^2)\,.$$

An important special case is $\mathcal{N}(0,1)$, called the *standard normal distribution*. The pdf of $Z \sim \mathcal{N}(0,1)$ is

$$f_Z(x) := \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,, \quad -\infty < x < \infty\,.$$

$\phi(x)$ is called the *standard normal density*.

The cdf of $Z$ is

$$F_Z(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du\,, \quad -\infty < x < \infty\,.$$

$\Phi(x)$ is called the *standard normal cdf*.

Check that $\phi$ is a pdf indeed

Note that $\phi(x)$ is an even (symmetric) function, $\phi(x) = \phi(-x)$.

Thus

$$\Phi(-x) = 1 - \Phi(x)\,.$$

**Important:** If for some constants $\mu \in \mathbb{R}$ and $\sigma > 0$ and RVs $Z \sim \mathcal{N}(0,1)$ and $X = \sigma Z + \mu$, then

$$\begin{aligned}
F_X(x) &= \mathbf{P}(X \leq x) \\
&= \mathbf{P}(\sigma Z + \mu \leq x) \\
&= \mathbf{P}(Z \leq \frac{x-\mu}{\sigma}) \\
&= \int_{-\infty}^{(x-\mu)/\sigma} \phi(u)du\,,
\end{aligned}$$

and the pdf of $X$ is thus

$$f_X(x) = \frac{dF_X(x)}{dx} = \left(\frac{x-\mu}{\sigma}\right)' \phi\left(\frac{x-\mu}{\sigma}\right)$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R},$$

so that $X \sim \mathcal{N}(\mu, \sigma^2)$.

Summarizing, we only need the standard normal pdf/cdf to evaluate any probability for any normal RV:

$$\text{If } X \sim \mathcal{N}(\mu, \sigma^2), \text{ then } Z := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

$\mu$ is the *location* parameter and $\sigma > 0$ the scale parameter [$\rightarrow$ connection with moments?].

**Example 13.** *If $X \sim \mathcal{N}(25, 36)$, find a constant $c$ such that*

$$\mathbf{P}(|X - 25| \leq c) = 0.9544\,.$$

*We want*

$$\mathbf{P}(-c \leq X - 25 \leq c) = \mathbf{P}\left(-\frac{c}{6} \leq \frac{X - 25}{6} \leq \frac{c}{6}\right) = 0.9544\,.$$

*Thus*

$$\begin{aligned} \phi\left(\frac{c}{6}\right) - \phi\left(-\frac{c}{6}\right) &= \phi\left(\frac{c}{6}\right) - \left(1 - \phi\left(\frac{c}{6}\right)\right) \\ &= 2\phi\left(\frac{c}{6}\right) - 1 \\ &= 0.9544\,, \end{aligned}$$

*so that $\phi\left(\frac{c}{6}\right) = 0.9772$, and looking up tables gives $c = 12$.*

## Skewness and tail thickness

To summarize key features of a distribution so far, we have the mean, variance, standard deviation, and the median. Why stop here? Higher order moments provide further insight in the shape of a distribution.

The third central moment

$$\nu_3 = \mathbf{E}\left[(X - \mu_x)^3\right]$$

is an indicator of skewness. Indeed, if the distribution has a long positive tail, then $(X - \mu_X)^3$ has large positive values but does not have large negative values. Thus $\nu_3 > 0$ and the distribution is *positively skewed*. Similarly, if the distribution has a long negative tail, $\nu_3 < 0$ and the distribution is negatively skewed.

**Definition 21.** *The coefficient of skewness is obtained by standardising to remove the scale effect*

$$Skew(X) = \frac{\nu_3}{\sigma_X^3} = \mathbf{E}\left(\frac{X - \mu_X}{\sigma_X}\right)^3 .$$

Of course, if the pdf of $X$ is symmetric, then $\mathrm{Skew}(X) = 0$ since $(x - \mu_X)^3 f_X(x)$ is an odd function around $\mu$. In particular

$$Z \sim \mathcal{N}(0,1)\,, \qquad \mathrm{Skew}(Z) = 0\,.$$

The fourth central moment $\nu_4 = \mathbf{E}\left[(X - \mu_X)^4\right]$ is an indicator of the peakedness and the length of the tails of a distribution.

**Definition 22.** *The coefficient of kurtosis is obtained by standardising to remove the scale effect*

$$Kurt(X) = \frac{\nu_4}{\sigma_X^4} - 3\,.$$

Why removing 3?

Let's have a look at the value of $\nu_4$ for the $\mathcal{N}(0,1)$ distribution. We derived previously

$$\mathbf{E}\,(Z^4) = 3\,\mathbf{E}\,(Z^2) = 3\,.$$

So, by removing 3 in the definition of kurtosis, we make the normal distribution a reference distribution.

$$Z \sim \mathcal{N}(0,1)\,, \qquad \mathrm{Kurt}(Z) = 0\,.$$

- If $\mathrm{Kurt}(X) < 0$, the distribution is likely to be flatter, and have lighter tails than $\mathcal{N}$.
- If $\mathrm{Kurt}(X) > 0$, the distribution is likely to be more peaked, and have heavier tails than $\mathcal{N}$.

# Pareto distribution

The Pareto distribution is a *heavy-tailed* distribution, meaning that its tails are thicker than the normal distribution: large values are more likely to occur than under a normal distribution model. The pdf and cdf of a Pareto distribution are

$$
\begin{aligned}
f_X(x) &= \frac{\alpha k^\alpha}{x^{\alpha+1}} \\
F_X(x) &= 1 - \left(\frac{k}{x}\right)^\alpha ,
\end{aligned}
$$

for $k \le x < \infty$ and $k > 0$, representing the lowest value $X$ can take.

Under which condition is the mean of $X$ finite? What about its variance? Well, provided $\alpha > n$,

$$
\mathbf{E}(X^n) = \int_k^\infty x^n \frac{\alpha k^\alpha}{x^{\alpha+1}} dx = \alpha k^\alpha \int_k^\infty x^{n-\alpha-1} dx = \frac{\alpha k^n}{\alpha - n} ,
$$

from which we get $\mathbf{E}(X) = \alpha k/(\alpha - 1)$ and $\mathrm{Var}(X) = \alpha k^2/[(\alpha-1)^2(\alpha-2)]$.

# Functions of a continuous RV

## One-to-one transforms

Let $X$ be some RV with state space $S_X$. We are interested in the distribution of $Y = g(X)$ (state space $S_Y$), for some strictly increasing or decreasing function $g$. We will learn two techniques for finding the distribution of $Y$, namely

(i) Distribution function technique

(ii) Change of variable technique

Let $y \in S_Y$. Suppose for now that $g$ is strictly increasing. Then

$$\begin{aligned} F_Y(y) = \mathbf{P}(Y \leq y) &= \mathbf{P}(g(X) \leq y) \\ &= \mathbf{P}(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y))\,. \end{aligned}$$

**Example 14.** *A spinner is mounted at the point $(0, 1)$. Let $\theta$ be the smallest angle between the $y$-axis and the spinner. Assume that $\theta$ is the value of a random variable $\Theta \sim \mathcal{U}(-\pi/2, \pi/2)$. Let $(X, 0)$ be the intersection point of the $x$-axis and the spinner. Find the pdf of $X$.*

*First, note that $X = \tan \Theta$, which takes values in $S_X = \mathbb{R}$ (and $\Theta = \arctan(X)$). The cdf of $X$ is*

$$\begin{aligned} F_X(x) = \mathbf{P}(X \le x) &= \mathbf{P}(\Theta \le \arctan(X)) \\ &= F_\Theta(\arctan(x)) \\ &= \frac{1}{\pi}\left(\arctan(x) + \frac{\pi}{2}\right), \quad -\infty < x < \infty \,. \end{aligned}$$

*Differentiating gives*

$$f_X(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty \,.$$

*This distribution is known as the Cauchy distribution.*

We now generalize the procedure of the previous example. So far, we have for $y \in S_y$ and $g$ strictly increasing,

$$F_Y(y) = F_X(g^{-1}(y)) = \int_{x_1}^{g^{-1}(y)} f_X(u) du\,, \text{ where } x_1 < x < x_2.$$

Differentiating this expression yields

$$f_Y(y) = F_y'(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}\,, \quad y \in S_Y\,.$$

Similarly, if $g$ is strictly decreasing,

$$f_Y(y) = F_y'(y) = -f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}\,, \quad y \in S_Y\,.$$

In general, if $g$ is a continuous monotonic function,

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|\,, \quad y \in S_Y\,.$$

This expression is referred to as the *change of variable formula*.

**Example 15.** *Suppose $X$ has pdf $f_X(x) = 3(1-x)^2$ for $0 < x < 1$ and 0 elsewhere. Find the pdf of $Y = (1-X)^3$.*

*(i)* $g(x) = (1-x)^3$ *which is strictly decreasing on* $(0,\,1)$.

*(ii) The domain* $0 < x < 1$ *is mapped onto* $0 < y < 1$.

*(iii) Inverse transform is* $X = 1 - Y^{1/3} = g^{-1}(Y)$, *so that* $g^{-1}(y) = 1 - y^{1/3}$.

*(iv) The change of variable technique gives*

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| \\ &= 3\left(1 - (1 - y^{1/3})\right)^2 \cdot \left| -\frac{1}{3} y^{-2/3} \right| \\ &= \left(1 - 2(1 - y^{1/3}) + (1 - y^{1/3})^2\right) y^{-2/3} \\ &= 1 \;\; (after\ simplifications). \end{aligned}$$

*So that* $Y \sim \mathcal{U}(0,1)$.

## Non monotonic transforms

The distribution function technique is highly recommended for this situation. Do not use the change of variable formula!

The idea is to partition $S_Y$ into intervals over which $g$ is monotonic.

**Example 16.** *Let $X$ have the pdf $f_X(x) = x^3/3$ for $-1 < x < 2$. Find the pdf of $Y = X^2$.*

*(i) The support of $Y$ is $0 \leq y < 4$.*

*(ii) For $0 \leq y < 1$, the cdf of $Y$ is*

$$\begin{aligned} F_Y(y) &= \mathbf{P}(X^2 \leq y) \\ &= \mathbf{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{x^2}{3}\, dx \\ &= \frac{2}{9}\, y^{3/2}\,. \end{aligned}$$

(iii) For $1 \le y < 4$, the cdf of $Y$ is

$$\begin{aligned} F_Y(y) &= \mathbf{P}(X^2 \le y) \\ &= \mathbf{P}(-1 \le X \le \sqrt{y}) \\ &= \int_{-1}^{\sqrt{y}} \frac{x^2}{3}\, dx \\ &= \frac{1}{9}\, y^{3/2} + \frac{1}{9}\,. \end{aligned}$$

(iv) Therefore, the pdf of $Y$ is

$$f_Y(y) = F_Y'(y) = \begin{cases} \sqrt{y}/3 & \text{if } 0 \le y < 1 \\ \sqrt{y}/6 & \text{if } 1 \le y < 4\,. \end{cases}$$

[Picture.]

# Bivariate distributions

A random variable $X = X(\omega)$ is a *measurement* of the outcome $\omega$. In many situations, we are interested in more than one measurement. For example,

*Patient*: age, sex, weight,...

*Wind*: speed, direction,...

In the simplest case, these two measurements are two numerical characteristics of an outcome pair

$$(X, Y) = (X(\omega), Y(\omega)) \in \mathbb{R}^2 \, .$$

$(X, Y)$ is called a *bivariate random variable*.

**Example 17.** *Toss a coin twice and let*

$$\begin{aligned} X &= \textit{ number of heads in the first toss } (0 \textit{ or } 1) \\ Y &= \textit{ total number of heads } (0, 1 \textit{ or } 2) \end{aligned}$$

How to describe the (joint) distribution of $(X, Y)$?

# Discrete bivariate distributions

$(X, Y)$ can only take values $(x_1, y_1), (x_2, y_2), \ldots$. Denote by $S_{X,Y}$ the set of all possible values that the pair $(X, Y)$ can take [the state space]. In Example 18, $S_{X,Y} = \{(0,0), (1,1), (0,1), (1,2)\}$.

**Definition 23.** *Let $(X, Y)$ be a discrete bivariate RV with state space $S_{X,Y}$. The joint pmf of $(X, Y)$ is defined as*

$$p_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y)\,,$$

*and satisfies the same properties as the pmf of a univariate RV:*

*(i)* $0 \le p_{X,Y}(x, y) \le 1$.

*(ii)* $\sum\sum_{(x,y)\in S_{X,Y}} p_{X,Y}(x, y) = 1$.

*(iii)* $\mathbf{P}((X, Y) \in B) = \sum\sum_{(x,y)\in B} p_{X,Y}(x, y)$, *for* $B \subset S_{X,Y}$.

**Definition 24.** *The joint cdf of $X$ and $Y$ is defined as*

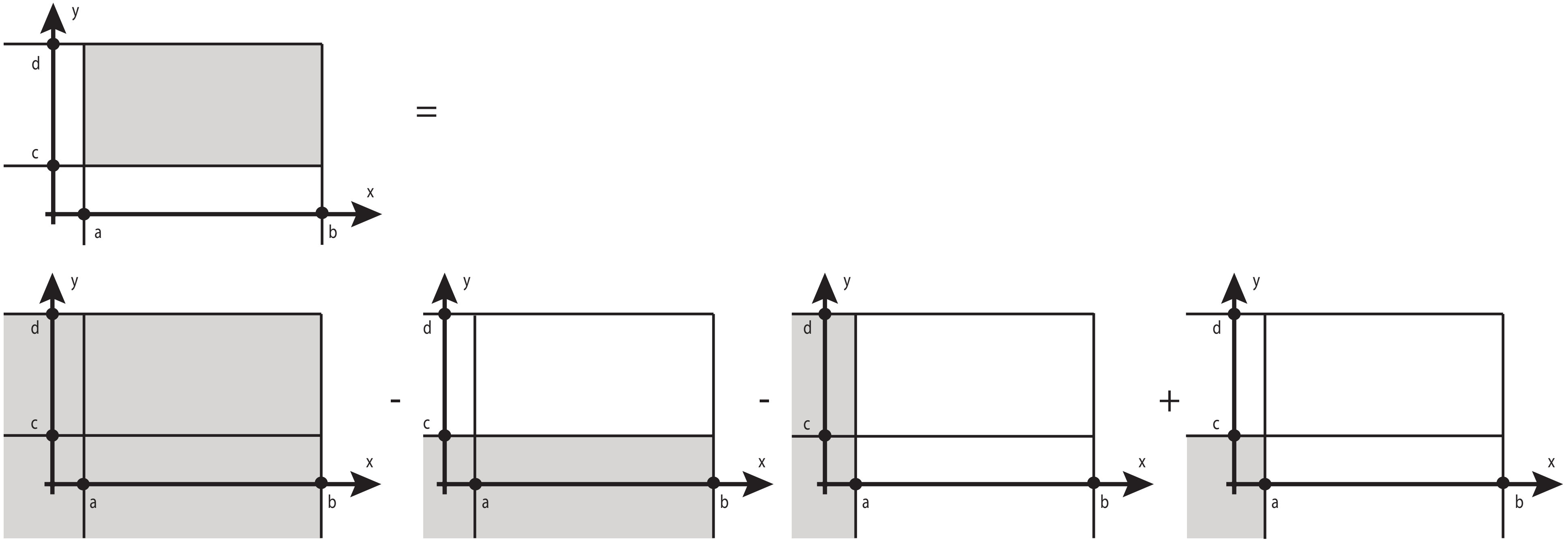
$$\begin{aligned} F_{X,Y}(x,y) &= \mathbf{P}(X \leq x, Y \leq y) \\ &= \sum_{j \,|\, x_j \leq x} \sum_{j \,|\, y_j \leq y} \mathbf{P}(X = x_j, Y = y_j) \\ &= \sum_{j \,|\, x_j \leq x} \sum_{j \,|\, y_j \leq y} p_{X,Y}(x_j, y_j)\,. \end{aligned}$$

Recall, for a univariate random variable $X$ with cdf $F_X$ we have

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

What about bivariate distribution?

Instead of intervals, we have rectangles...



$$\mathbf{P}(a < X \leq b, c < Y \leq d) = F_{X,Y}(b,d) - F_{X,Y}(b,c) - F_{X,Y}(a,d) + F_{X,Y}(a,c) \geq 0\,.$$

From the bivariate cdf $F_{X,Y}$, we can get the *marginal* univariate cdfs $F_X$ and $F_Y$. The joint behaviour of $X$ and $Y$ contains information about individual behaviours...

$$\begin{aligned} F_X(x) &= \mathbf{P}(X \leq x) \\ &= \mathbf{P}(X \leq x, Y < \infty) \\ &= \lim_{y \to \infty} \mathbf{P}(X \leq x, Y < y) \\ &= \lim_{y \to \infty} F_{X,Y}(x, y) \\ &= F_{X,Y}(x, \infty)\,. \end{aligned}$$

Similarly

$$F_Y(y) = F_{X,Y}(\infty, y)\,.$$

But not the other way around! Knowing individual behaviours provides absolutely no information about how $X$ and $Y$ are related!

*Note:* $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$.

**Definition 25.** *Let $(X, Y)$ be a discrete bivariate RV with space $S_{X,Y}$. The marginal pmf of $X$ is defined by*

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ &= \sum_y \mathbf{P}(X = x, Y = y) \\ &= \sum_y p_{X,Y}(x, y)\,, \quad s \in S_X\,. \end{aligned}$$

*The sums run over values of $y$ such that $(x, y) \in S_{X,Y}$.*

*Similarly for $Y$.*

## Independence

Recall, for events $A$ and $B$, independence is defined as $\mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B)$. Similarly, $X$ and $Y$ are said independent if

$$F_{X,Y}(x,y) = F_X(x)\,F_Y(y)\,.$$

In other words, events $\{X \le x\}$ and $\{Y \le y\}$ are independent for all $x, y$.

If $F_{X,Y}(x,y) = F_X(x)\,F_Y(y)$ holds, then for any sets $M, N \subset \mathbb{R}$,

$$\mathbf{P}(X \in M, Y \in N) = \mathbf{P}(X \in M)\,\mathbf{P}(Y \in N)\,,$$

that is, any events $A$ and $B$ such that $A$ is defined in terms of $X$ ($A = \{X \in M\}$) and $B$ in terms of $Y$ ($B = \{Y \in N\}$) are independent.

For discrete RVs

$$X \text{ and } Y \text{ are independent } \Leftrightarrow\ p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$$

*Proof?*

**Example 18.** *Roll a pair of unbiased dice. Let $X$ be the smaller and $Y$ the larger outcome on the dice.*

*The set of possible values is*

$$\begin{aligned} S_{X,Y} &= \{(1,1),\dots,(1,6),(2,2),\dots,(2,6),\dots(6,6)\} \\ &= \{(x,y) \mid x \le y,\ x = 1,\dots,6 \textit{ and } y = 1,\dots,6\}\,. \end{aligned}$$

*The joint distribution of $X$ and $Y$ is*

$$p_{X,Y}(x,y) = \begin{cases} 1/36 & \textit{if } 1 \le x = y \le 6 \\ 2/36 & \textit{if } 1 \le x < y \le 6\,, \end{cases}$$

*where $x$ and $y$ are integer.*

*What about marginal distributions? First,*

$$S_X = S_Y = \{1,2,3,4,5,6\}\,.$$

*We have, for example,*

$$\begin{aligned} p_X(4) &= \mathbf{P}(X=4) \\ &= p_{X,Y}(4,4) + p_{X,Y}(4,5) + p_{X,Y}(4,6) \\ &= 5/36\,. \end{aligned}$$

*and*

$$\begin{aligned} p_Y(4) &= \mathbf{P}(Y=4) \\ &= p_{X,Y}(1,4) + p_{X,Y}(2,4) + p_{X,Y}(3,4) + p_{X,Y}(4,4) \\ &= 7/36\,. \end{aligned}$$

*For $X$ and $Y$ to be independent, we need $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$ for all $x, y$. Clearly $X$ and $Y$ are not independent since*

$$p_{X,Y}(4,4) = 1/36 \neq p_X(4)\,p_Y(4)\,.$$

**Example 19.** *Let the joint pmf of $X$ and $Y$ be*

$$p_{X,Y}(x,y) = \frac{x\,y^2}{30}\,, \quad x = 1,2,3 \textit{ and } y = 1,2\,.$$

*The marginal pmfs are*

$$p_X(x) = \sum_{y=1}^{2} \frac{x\,y^2}{30} = \frac{x}{6}\,, \quad x = 1,2,3\,.$$

$$p_Y(y) = \sum_{x=1}^{3} \frac{x\,y^2}{30} = \frac{y^2}{5}\,, \quad y = 1,2\,.$$

*And for all $(x,y) \in S_{X,Y}$,*

$$p_{X,Y}(x,y) = \frac{x}{6}\,\frac{y^2}{5} = p_X(x)\,p_Y(y)\,.$$

$\Rightarrow$ *$X$ and $Y$ are independent.*

# Continuous bivariate distributions

Definitions are the same as in the discrete case, except that $\int$ replace $\sum$.

In particular, the *joint probability density function* (joint pdf) of two continuous random variables $X$ and $Y$ is an integrable function $f_{X,Y}(x,y)$ satisfying

(i) $f_{X,Y}(x,y) \geq 0$.

(ii) $\int \int f_{X,Y}(x,y) dx dy = 1$.

(iii) $\mathbf{P}((X,Y) \in B) = \int \int_B f_{X,Y}(x,y) dx dy$ for $B \subset \mathbb{R}^2$.

The joint cdf of $X$ and $Y$ is

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbf{P}(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v)\, du\, dv \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\, du\, dv\,. \end{aligned}$$

The joint pdf $f_{X,Y}(x,y)$ satisfies

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} = \frac{\partial}{\partial y}\left(\frac{\partial F_{X,Y}(x,y)}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial F_{X,Y}(x,y)}{\partial y}\right),$$

if the partial derivatives exist.

Recall, if $X$ is a univariate RV, then

$$f_X(x) = \frac{dF_X(x)}{dx} \approx \frac{1}{h}\mathbf{P}(x - h/2 \leq X \leq x + h/2),$$

so that

$$\mathbf{P}(X \in \text{ small interval around } x) \approx f_X(x) \times (\text{length of the interval}).$$

What about in the bivariate case? Interpretation of $f_{X,Y}(x,y)$?

Similarly,

$$f_{X,Y}(x,y) = \frac{\partial}{\partial x}\left(\frac{\partial F_{X,Y}(x,y)}{\partial y}\right),$$

where

$$\frac{\partial F_{X,Y}(x,y)}{\partial y} \approx \frac{1}{h}\left[F(x,y+h/2) - F(x,y-h/2)\right].$$

Thus $f_{X,Y}(x,y)$ is roughly equal to

$$\frac{1}{h}\left[\frac{F(x+\frac{h}{2},y+\frac{h}{2}) - F(x-\frac{h}{2},y+\frac{h}{2})}{h} - \frac{F(x+\frac{h}{2},y-\frac{h}{2}) - F(x-\frac{h}{2},y-\frac{h}{2})}{h}\right].$$

Rearranging terms gives (cf slide 6)

$$f_{X,Y}(x,y) \approx \mathbf{P}\left(x-\frac{h}{2} < X \le x+\frac{h}{2},\, y-\frac{h}{2} < Y \le y+\frac{h}{2}\right),$$

and we see that

$$\mathbf{P}((X,Y) \in \text{small set around } (x,y)) \approx f_{X,Y}(x,y) \times \text{area of this small set}\,.$$

For a large set, we just integrate the pdf over this set. If $a < b$ and $c < d$,

$$\mathbf{P}(a < X \leq b,\, c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y) dx dy\,.$$

The *marginal* pdfs of $X$ and $Y$ are respectively

$$f_X(x) = \int f_{X,Y}(x,y) dy$$
$$f_Y(y) = \int f_{X,Y}(x,y) dx\,.$$

Of course, we can talk about independent RVs in the continuous case as well, and we have the following result

$$X \text{ and } Y \text{ are independent } \Leftrightarrow f_{X,Y}(x,y) = f_X(x)\, f_Y(y) \quad \text{for all } (x,y) \in S_{X,Y}\,.$$

# Mathematical expectations of $\psi(X, Y)$

Let $(X, Y)$ be a (discrete or continuous) bivariate RV, and $\psi(x, y)$ a function in $\mathbb{R}$.

Extending the definition of expectation involving a single RV, we get

$$\mathbf{E}\,\psi(X, Y) = \sum \sum \psi(x, y)\, p_{X,Y}(x, y)\,,$$

if $(X, Y)$ is discrete, and

$$\mathbf{E}\,\psi(X, Y) = \int \int \psi(x, y)\, f_{X,Y}(x, y)\, dx\, dy\,,$$

if $(X, Y)$ is continuous.

(provided the sums or integrals converge..)

**Example 20.** *Let*

$$(X,Y)=\begin{cases}(-1,\,0) & \textit{with probability} 1/3\\ (1,\,0) & \textit{with probability} 1/3\\ (0,\,1) & \textit{with probability} 1/3\,.\end{cases}$$

*Then* $XY=0$ *always so* $\mathbf{E}\left(XY\right)=0$*. Moreover,* $\mathbf{E}\left(X\right)=-1\cdot(1/3)+0\cdot(1/3)+1\cdot(1/3)=0$*, so that* $\mathbf{E}\left(XY\right)=\mathbf{E}\left(X\right)\mathbf{E}\left(Y\right)$*. But if* $X=0$*, then* $Y=1$ *always, so that* $X$ *and* $Y$ *are not independent.*

**Exercise 12.** *Let $X$ and $Y$ have the joint pdf*

$$f_{X,Y}(x,y) = 2\,, \qquad 0 \le x \le y \le 1\,.$$

*The support of $X$ and $Y$ is not rectangular, so that $X$ and $Y$ are not independent. Find $\mathbf{P}\left(X \le \frac{1}{2}\right)$, $\mathbf{E}\left(X\right)$ and $\mathbf{E}\left(Y\right)$.*

# Covariance and correlation coefficient

Here we discuss tools for describing the relationship between two random variables $X$ and $Y$.

**Definition 26.** *The covariance of $X$ and $Y$ is*

$$\sigma_{X,Y} = Cov(X,Y) = \mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)\right] ,$$

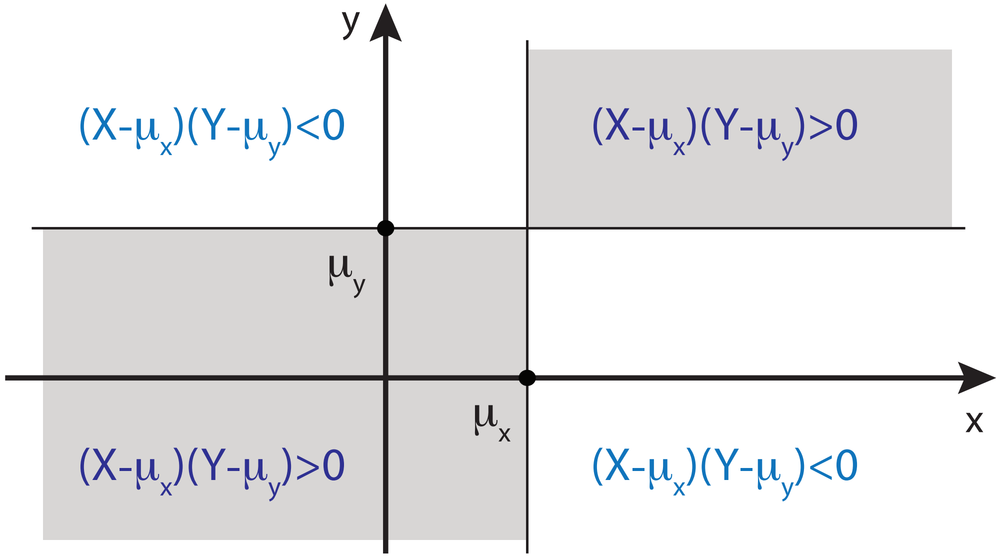
*where $\mu_X = \mathbf{E}\left(X\right)$, and $\mu_Y = \mathbf{E}\left(Y\right)$.*

It is sometimes more convenient to work with

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathbf{E}\left(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y\right), \\ &= \mathbf{E}\left(XY\right) - \mathbf{E}\left(\mu_X Y\right) - \mathbf{E}\left(\mu_Y X\right) + \mu_X\mu_Y , \\ &= \mathbf{E}\left(XY\right) - \mathbf{E}\left(X\right)\mathbf{E}\left(Y\right). \end{aligned}$$

Note that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$ and $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

## What does the value of $\mathrm{Cov}(X, Y)$ indicate?



- If a large **positive**/negative value of $X$ makes a large **positive**/negative value of $Y$ more likely, there is a **positive relationship** between $X$ and $Y$ and $\mathrm{Cov}(X, Y) > 0$.

- If a large **positive**/negative value of $X$ makes a large **negative**/positive value of $Y$ more likely, there is a **negative relationship** between $X$ and $Y$ and $\mathrm{Cov}(X, Y) < 0$.

So $\mathrm{Cov}(X, Y)$ is an 'indicator' of relationship between $X$ and $Y$, and also scale. Its size depends on the scale (variance) of the two variables, and thus has no upper or lower boundary.

Covariance provides the *direction* of the relationship.

To remove the scale effect, we need to *standardize* the covariance,

**Definition 27.** *The correlation coefficient of $X$ and $Y$ is*

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X\,\sigma_Y}\,.$$

We will show later that $-1 \leq \rho \leq 1$. The correlation coefficient provides information about the direction *and* strength of the relationship between the variables.

**Remark:** The variance of $X + Y$ in the general case is

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)\,.$$

Indeed,

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbf{E}\left[(X+Y-(\mu_X+\mu_Y))^2\right] \\
&= \mathbf{E}\,(X-\mu_X)^2 + 2\mathbf{E}\,(X-\mu_X)(Y-\mu_Y) + \mathbf{E}\,(Y-\mu_Y)^2 \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)\,.
\end{aligned}$$

**Definition 28.** *If $\rho = 0$ (equiv. $Cov(X,Y) = 0$), the variables $X$ and $Y$ are said to be uncorrelated.*

With this definition, we have

If $X$ and $Y$ are uncorrelated, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

## Properties of covariance and correlation

(i) Let $a \in \mathbb{R}$, then $\mathrm{Cov}(X, a) = \mathbf{E}\,(X - \mu_X)(a - a) = 0$.

(ii) Let $a, b \in \mathbb{R}$

$$\begin{aligned}\mathrm{Cov}(aX,\, bY) &= \mathbf{E}\,[(aX - a\mu_x)(bY - b\mu_Y)]\\ &= a\,b\,\mathbf{E}\,(X - \mu_X)(Y - \mu_Y)\\ &= a\,b\,\mathrm{Cov}(X, Y)\,.\end{aligned}$$

(iii)

$$\begin{aligned}\mathrm{Cov}(X + a,\, Y + b) &= \mathbf{E}\,[(X + a - (\mu_X + a))(Y + b - (\mu_Y + b))]\\ &= \mathrm{Cov}(X, Y)\,.\end{aligned}$$

(iv)

$$\begin{aligned}\mathrm{Cov}(X,\, aX + b) &= \mathbf{E}\,[(X - \mu_X)(aX + b - (a\mu_X + b))]\\ &= \mathbf{E}\,[(X - mu_X)\,a\,(X - \mu_X)] = a\,\mathrm{Var}(X)\,.\end{aligned}$$

## Correlation and independence

Suppose $X$ and $Y$ are two independent continuous RVs, anf $g$ and $h$ two functions. Then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ and

$$\begin{aligned}\mathbf{E}\left[g(X)h(Y)\right] &= \int\int g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \left(\int g(x)f_X(x)dx\right)\cdot\left(\int h(y)f_Y(y)dy\right) \\ &= \mathbf{E}\left(g(X)\right)\mathbf{E}\left(h(Y)\right).\end{aligned}$$

Thus, if $X$ and $Y$ are independent, then $\mathbf{E}\left[g(X)h(Y)\right] = \mathbf{E}\left(g(X)\right)\mathbf{E}\left(h(Y)\right)$. In particular, $\mathbf{E}\left(XY\right) = \mathbf{E}\left(X\right)\mathbf{E}\left(Y\right)$.

Moreover, take $g(x) = (x - \mu_X)$ and $h(y) = (y - \mu_Y)$, we get

$$\mathbf{E}\left(X - \mu_X\right)(Y - \mu_Y) = 0 = \mathrm{Cov}(X, Y).$$

We just showed that

$$X \text{ and } Y \text{ are independent } \Rightarrow X \text{ and } Y \text{ are uncorrelated}$$

However, in general the converse is not true..

**Example 21.** *Let $X$ and $Y$ have the joint pmf*

$$p_{X,Y}(x,y) = 1/3 \textit{ for } (x,y) = (0,1),(1,0),(2,1)\,.$$

*Since the support is not 'rectangular', $X$ and $Y$ must be dependent. However, $\mu_X = 1$, $\mu_Y = 2/3$, and $Cov(X,Y) = \mathbf{E}\,(XY) - 2/3 = 0$. So $X$ and $Y$ are uncorrelated even though they are not independent.*

We have the following important result

**Theorem 4.** *The correlation coefficient $\rho$ is always between -1 and 1*

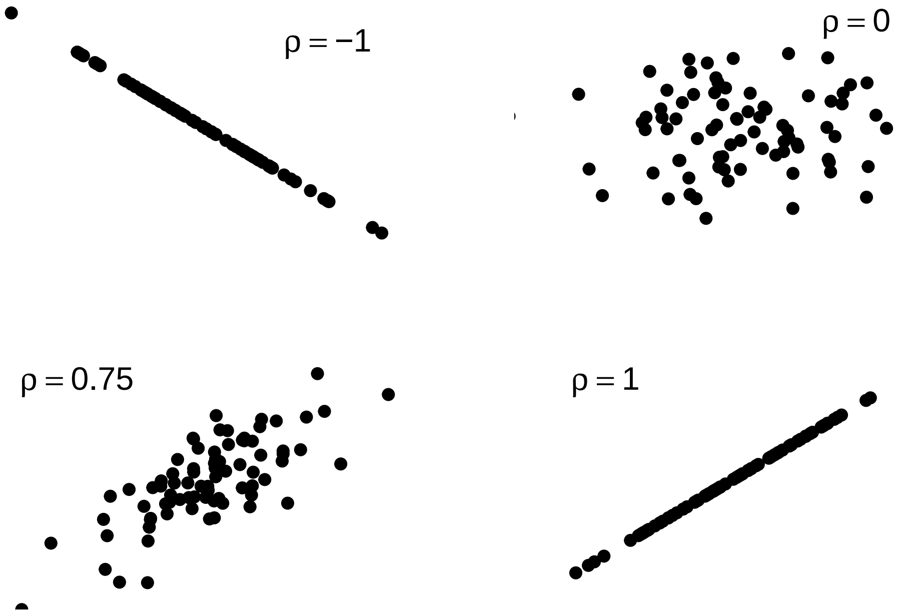
$$|\rho| \leq 1\,.$$

*Moreover, $\rho$ measures the amount of linear relationship between two RVs.*

What happens when $\rho = \pm 1$?

In that case, $g(a^*, b^*) = 0 = \mathbf{E}[(Y - a^* X - b)^2]$. Since $(Y - a^* X - b)^2 \geq 0$, then necessarily $Y = a^* X + b^*$ (linear relationship).

$$Y = \rho \frac{\sigma_Y}{\sigma_X}(X - \mu_X) + \mu_Y \text{ if } \rho = \pm 1\,.$$



# Conditional distributions

For discrete random variables $X$ and $Y$ with joint pmf $p_{X,Y}(x,y)$, and events $A = \{X = x\}$ and $B = \{Y = y\}$, what is the probability of occurrence of $A$ knowing that $B$ occurs? Sounds familiar..

$$\mathbf{P}(A \mid B) = \mathbf{P}(X = x \mid Y = y) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(X = x, \, Y = y)}{\mathbf{P}(Y = y)} \, .$$

Hence the definition

**Definition 29.** *The conditional pmf of $X$ given $Y = y$ is*

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \, .$$

*Similarly, we have*

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \, .$$

Note that for a fixed $y$, $p_{X|Y}(x \mid y)$ regarded as a function of $x$, satisfies all the properties of a pmf,

(i) $p_{X|Y}(x \mid y) \geq 0$.

(ii) $\sum_{x \in S_X} p_{X|Y}(x \mid y) = 1$.

(iii) $\mathbf{P}(a < X < b \mid Y = y) = \sum_{x|\, a<x<b} p_{X|Y}(x \mid y)$,

and similarly for $p_{Y|X}(y \mid x)$, regarded as a function of $y$.

**Definition 30.** *The conditional mean of $X$ given $Y = y$ is*

$$\mathbf{E}(X \mid Y = y) = \sum_{x \in S_X} x \, p_{X|Y}(x \mid y) = \textit{a function of } y =: \eta(y)\,.$$

*The conditional variance of $X$ given $Y = y$ is*

$$\mathit{Var}(X \mid Y = y) = \sum_{x \in S_X} x^2 \, p_{X|Y}(x \mid y) - (\mathbf{E}(X \mid Y = y))^2 = \textit{a function of } y =: \zeta(y)\,.$$

**Remarks:**

(i) If $\psi$ and $\phi$ are two functions in $\mathbb{R}$, then

$$\mathbf{E}(\psi(X) \mid Y = y) = \sum_{x \in S_X} \psi(x)\, p_{X|Y}(x \mid y)\,.$$

$$\mathbf{E}(\phi(Y) \mid X = x) = \sum_{y \in S_Y} \phi(y)\, p_{Y|X}(y \mid x)\,.$$

(ii) Back to independence.

Recall, events $A$ and $B$ are independent iff $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ or, using conditional probability, $\mathbf{P}(A \mid B) = \mathbf{P}(A)$.

Similarly, if $X$ and $Y$ are independent, then $p_{X|Y}(x \mid y) = p_X(x)$ and $p_{Y|X}(y \mid x) = p_Y(y)$.

Thus $\mathbf{E}(\psi(X) \mid Y = y) = \mathbf{E}(\psi(X))$.

If $X$ and $Y$ are continuous RVs, the situation is more tricky. What is the conditional distribution of $X$ given $Y = y$? The event $\{Y = y\}$ has zero probability..

$$\mathbf{P}(X \leq x, | \, Y = y) = \frac{\mathbf{P}(X \leq x, \, Y = y)}{\mathbf{P}(Y = y)} = \frac{0}{0} = \dots .$$

In fact, we can make the above meaningful by taking small intervals

$$\begin{aligned}
\text{Conditional Proba} &= \frac{\mathbf{P}\left(x - \frac{h}{2} < X \leq x + \frac{h}{2} \, , y - \frac{h}{2} < Y \leq y + \frac{h}{2}\right)}{\mathbf{P}\left(y - \frac{h}{2} < Y \leq y + \frac{h}{2}\right)} \\
&\approx \frac{f_{X,Y}(x,y) \, h^2}{f_Y(y) \, h} \\
&= \frac{f_{X,Y}(x,y)}{f_Y(y)} \, h \, .
\end{aligned}$$

Which leads to the following definition.

**Definition 31.** *The conditional pdf of $X$ given $Y = y$ is defined by*

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \,.$$

**Remarks:**

(i) The conditional cdf is $F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(u \mid y) du$.

(ii) The conditional mean is $\mathbf{E}\,(X \,|\, Y = y) = \int x\, f_{X|Y}(x \mid y) dx$.

(iii) The conditional variance Var$(X|Y = y) = \mathbf{E}(X^2 \,|\, Y = y) - (\mathbf{E}(X \,|\, Y = y))^2$.

## Two important properties

(i) Tower property

$$\mathbf{E}\left(X\right)=\mathbf{E}\left(\mathbf{E}\left(X\,|\,Y\right)\right)$$
$$\mathbf{E}\left(Y\right)=\mathbf{E}\left(\mathbf{E}\left(Y\,|\,X\right)\right).$$

(ii) Conditional variance

$$\mathrm{Var}(X)=\mathbf{E}\left(\mathrm{Var}(X\,|\,Y)\right)+\mathrm{Var}(\mathbf{E}\left(X\,|\,Y\right)).$$
$$\mathrm{Var}(Y)=\mathbf{E}\left(\mathrm{Var}(Y\,|\,X)\right)+\mathrm{Var}(\mathbf{E}\left(Y\,|\,X\right)).$$

**Exercise 13.** *The number of emissions from a radioactive source that occur in the time period $[0, t]$ follows a Poisson distribution with parameter $\lambda$. Each of the emissions is detected by a Geiser counter with probability $p$, or missed with probability $(1 - p)$. What is the expected number of emissions detected in $[0, t]$?*

# Bivariate normal distribution

**Definition 32.** *If the pdf of $(X, Y)$ is given by*

$$\phi_\rho(x, y) := f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{x^2 - 2\rho xy + y^2}{1-\rho^2}\right\},$$

*where $|\rho| < 1$, then we say that $(X, Y)$ has the standard bivariate normal distribution with parameter $\rho$, and we write*

$$(X, Y) \sim \mathcal{N}_2(\rho).$$

*Check:* marginal distribution of $Y$?

$$\begin{aligned} f_Y(y) &= \int f_{X,Y}(x, y) dx \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int \exp\left\{-\frac{1}{2}\frac{x^2 - 2\rho xy + y^2}{1-\rho^2}\right\} dx \end{aligned}$$

$$f_Y(y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-y^2/2} \int \exp\left\{-\frac{1}{2}\frac{x^2 - 2\rho xy + \rho^2 y^2}{1-\rho^2}\right\} dx\,.$$

Let $u = (x - \rho y)/\sqrt{1-\rho^2}$. Then

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} \frac{1}{2\pi} e^{-y^2/2} \int e^{-u^2/2} \sqrt{1-\rho^2} du \\ &= \frac{1}{2\pi} e^{-y^2/2} \int e^{-u^2/2} du \\ &= \frac{1}{2\pi} e^{-y^2/2}\,. \end{aligned}$$

Thus

$$X \sim \mathcal{N}(0,\, 1)\,.$$

$$Y \sim \mathcal{N}(0,\, 1)\,.$$

Good. But then, is the converse true? No!

**Example 22.** *Let $X \sim \mathcal{N}(0,1)$ and*

$$Y = \begin{cases} X & \textit{with probability } 1/2 \\ -X & \textit{with probability } 1/2\,, \end{cases}$$

*then $Y \sim \mathcal{N}(0,1)$ but $(X, Y)$ is not bivariate normal since the bivariate pdf is non-zero only on the lines $y = \pm x$.*

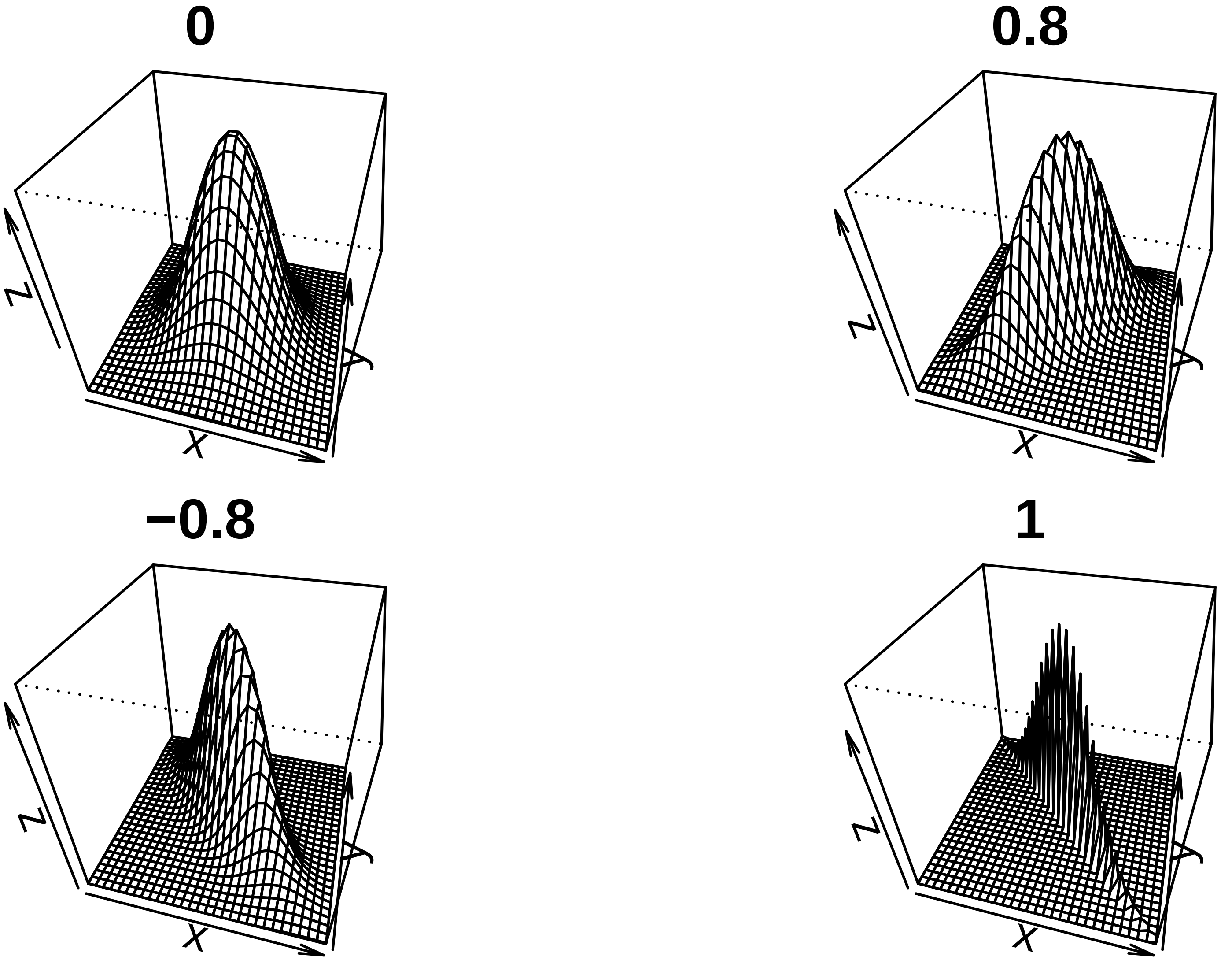
So what? The joint structure is not fixed by the normal marginal distributions (as expected). But then, what does the $\mathcal{N}_2(\rho)$ look like? Why this choice?

The contours of $\phi_\rho(x, y)$ are ellipses with axes inclined at an angle $\pi/4$ to the $x$ and $y$ axes.

If $\rho > 0$, then the major axis lies along $y = x$ and the minor axis along $y = -x$. This means that $X$ and $Y$ tend to be large together, and small together: they are *positively related.*

If $\rho < 0$, then it is the other way around. $X$ and $Y$ are negatively related. [Sounds familiar?]

Plots of the density $\mathcal{N}_2(\rho)$ for different values of $\rho$.



0

0.8

−0.8

1

What about the conditional pdf of $X$ given $Y = y$?

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{x^2 - 2\rho xy + y^2}{1-\rho^2}\right\} \sqrt{2\pi} e^{y^2/2} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{x^2 - 2\rho xy + \rho^2 y^2}{1-\rho^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{(x-\rho y)^2}{1-\rho^2}\right\}. \end{aligned}$$

Still normal! Good.

$$\begin{aligned} &X \,|\, Y = y \sim \mathcal{N}(\rho y, 1-\rho^2)\,. \\ &\mathbf{E}\,(X \,|\, Y = y) = \rho y\,. \end{aligned}$$

Note that if $\rho = 0$, then

$$\phi_0(x,y) = \frac{1}{2\pi}\exp\left\{-\frac{1}{2}(x^2+y^2)\right\} = f_X(x)\, f_Y(y)\,.$$

$X$ and $Y$ are independent!

What is the meaning of $\rho$?

Since $\mathbf{E}\left(X \,|\, Y=y\right)=\rho y$, we have $\mathbf{E}\left(X \,|\, Y\right)=\rho Y$ and

$$\mathbf{E}\left(XY\right)=\mathbf{E}\left(\mathbf{E}\left(X \,|\, Y\right)\right)=\mathbf{E}\left(\rho\, Y\right)=\rho\,\mathbf{E}\left(Y\right)=\rho\,,$$

so

$$\text{correlation coefficient}=\frac{\mathbf{E}\left(XY\right)}{\sigma_X\,\sigma_Y}=\rho\,.$$

$\rho$ is the correlation coefficient between $X$ and $Y$ (not surprising, right?). We have proved the following result

**Theorem 5.** *If $X$ and $Y$ have a bivariate normal distribution with correlation coefficient $\rho$, then*

$$X \textit{ and } Y \textit{ are independent } \Leftrightarrow \rho=0\,.$$

**Important.** This is not true in general, but only for the bivariate normal case: in general, zero correlation does *not* imply independence.

The $\mathcal{N}_2(\rho)$ distribution can be generalised to allow more flexibility. A nice way to do so is to shift the center of the bivariate distribution, and to rotate and stretch/squeeze the main axes of the pdf. If $(X, Y)$ have pdf

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2}{1-\rho^2}\right\},$$

with $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X, \sigma_Y > 0$, $|\rho| < 1$.

In this case we write

$$(X, Y) \sim \mathcal{N}_2(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)\,.$$

What is the meaning of the parameters?

We can show that

$$X \sim \mathcal{N}(\mu_X, \sigma_X) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y)\,,$$

and $\rho$ is the correlation coefficient, so that $\mathrm{Cov}(X, Y) = \rho\,\sigma_X \sigma_Y\,.$

**Remark:** Standardization of the general bivariate RV.

If $(X, Y) \sim \mathcal{N}_2(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, then

$$\left( \frac{X - \mu_X}{\sigma_X}, \, \frac{Y - \mu_Y}{\sigma_Y} \right) \sim \mathcal{N}_2(\rho)\,.$$

The expression of the bivariate normal distribution in its general form given previously is messy. It is a good idea at this point to introduce matrix notation. It will simplify our lives, and it will be helpful to generalise the bivariate case to the multivariate case..

We introduce the covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix} .$$

It plays the role of $\sigma^2$ for univariate distributions. Then

$$\Sigma^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2(1-\rho^2)} \begin{pmatrix} \sigma_Y^2 & -\rho\,\sigma_X\sigma_Y \\ -\rho\,\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} ,$$

and

$$\det(\Sigma) = \sigma_X^2\sigma_Y^2(1-\rho^2)\,.$$

Next, we introduce the mean vector

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

and

$$\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix} .$$

Then the bivariate normal distribution has the form

$$f_{\mathbf{X}}(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left\{ -\frac{1}{2}(\mathbf{X}-\mu)^t \Sigma^{-1} (\mathbf{X}-\mu) \right\} .$$

How would you generalize this expression if $\mathbf{X}$ had $n$ components?

# Random sample

We started this course with a definition of a Random Experiment. One of the key features of a RE is that it can be repeated many times under the same conditions. Each time you run an experiment, you obtain an (independent) observation of a random variable $X$. After $n$ experiments, we thus obtain measures of several RVs

$$X_1,\, X_2,\, \ldots,\, X_n\,.$$

We call $X_1, \ldots, X_n$ a random sample (RS) of size $n$. Each $X_i$ is a copy of a generic $X$, i.e. they have the same distribution.

Product spaces come handy when dealing with repeated experiments. For two sets $A$ and $B$, their product is defined as $A \times B := \{(a, b) \mid a \in A,\, b \in B\}$, and similarly, $A^n := A \times \ldots \times A = \{(a_1, \ldots, a_n) \mid a_j \in A,\, j = 1, 2, \ldots, n\}$. Thus, the sample space of a RE is $\Omega^n$, if $\Omega$ denotes the sample space of one experiment.

Note that Probability Theory has ways to deal with an infinite number of experiments too. It is necessary when dealing with convergence results [later].

*Assumption:* $X_1, \ldots, X_n$ are independent of each other and identically distributed (and we write i.i.d.).

We usually say that $X_1, \ldots, X_n$ is a RS or i.i.d. to refer to the same assumption of independence and identical distribution.

Let the pdf/pmf of $X_i$ be $f_{X_i}(x_i)$, taking values in $S$. Then the joint pdf/pmf of the RS under the independence assumption is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots, f_{X_n}(x_n)\,, \quad x_i \in S \text{ for } i = 1, \ldots, n\,.$$

Let $X_1, \ldots, X_n$ be a RS. We introduce the *sample mean*

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

and the *sample variance*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2\,.$$

First of all, note that both $\bar{X}$ and $S^2$ are RVs! [Are they independent of each other?] What is the mean of $\bar{X}$?

Well, note that if $X_1, \ldots, X_n$ are $n$ independent RVs with respective means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, then the mean and variance of $Y = \sum_{i=1}^n a_i\, X_i$, with $a_i \in \mathbb{R}$ are $\mu_Y = \sum_{i=1}^n a_i\, \mu_i$ and $\sigma_Y^2 = \sum_{i=1}^n a_i^2\, \sigma_i^2$.

This result is obvious for the mean. For the variance, note that independence implies uncorrelated, and thus all the covariance terms vanish,

$$\begin{aligned}\mathrm{Var}(Y) = \mathbf{E}\,(Y - \mu_Y)^2 &= \mathbf{E}\,\left(\sum_i a_i\,(X_i - \mu_i)\right)^2 \\ &= \mathbf{E}\,\sum_i \sum_j a_i\, a_j\,(X_i - \mu_i)(X_j - \mu_j) \\ &= \sum_i \sum_j a_i\, a_j\, \mathbf{E}\,(X_i - \mu_i)(X_j - \mu_j) \\ &= \sum_i a_i^2 \mathbf{E}\,(X_i - \mu_i)^2 = \sum_i a_i^2\, \sigma_i^2\,.\end{aligned}$$

Back to the sample mean,

$$\mathbf{E}\left(\bar{X}\right) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(X_i\right) = \mu_X \, .$$

We say that $\bar{X}$ is an *unbiased* estimator of the mean of $X$. Likewise, it is possible to show (and not very hard) that $S^2$ is an unbiased estimator of the variance of $X$. What about the variance of $\bar{X}$?

Good stuff. However, we want to derive further properties of these estimators (what is their [exact and approximate] distribution? How 'good' are these two estimators? Can we do better in some sense?).

Note that they both involves sums of random variables. We are thus naturally lead to the study of the distribution properties of sums of i.i.d. RVs. Before we can proceed further, we need to introduce a powerful tool to deal with sums of RVs: the *moment generating function*.

# Moment generating function

**Definition 33.** *Let $X$ be a discrete RV (resp. continuous RV) with pmf $p_X(x)$ (resp. with pdf $f_X(x)$) and state space $S_X$. The moment generating function or mgf of $X$, if it exists, is*

$$M_X(t) = \mathbf{E}\left(e^{tX}\right) = \begin{cases} \sum_{x \in S_X} e^{tx}\, p_X(x) & \textit{if } X \textit{ is discrete} \\ \int_{x \in S_X} e^{tx}\, f_X(x)\, dx & \textit{if } X \textit{ is continuous}\,, \end{cases}$$

*for $-h < t < h$, for some positive number $h$.*

An important result is that the pmf *uniquely* determines the mgf, and that the mgf *uniquely* determines the pmf/pdf. In other words,

$$p_X(x) = p_Y(x) \quad \Leftrightarrow \quad M_X(t) = M_Y(t)\,,$$

and similarly in the continuous case. Mgfs thus provide another tool for describing the distribution of a RV. However, the mgf may not exist for some RVs, while the pmf/pdf always exists [well, this is not completely true, but that will do].

**Example 23.** *Let $X$ be a RV with mgf*

$$M_X(t) = \frac{e^t/2}{1 - e^t/2}\,, \qquad t < \ln(2)\,.$$

*What is the pmf of $X$? Recall that in a neighborhood of 0, the Maclaurin's series expansion of $(1-z)^{-1}$ is*

$$\frac{1}{1-z} = 1 + z + z^2 + z^3 + \dots \qquad -1 < z < 1\,.$$

*Therefore, for $t < \ln(2)$,*

$$\begin{aligned} M_X(t) = \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1} &= \frac{e^t}{2}\left(1 + \frac{e^t}{2} + \frac{e^{2t}}{2^2} + \frac{e^{3t}}{2^3} + \dots\right) \\ &= \frac{1}{2}e^t + \frac{1}{2^2}e^{2t} + \frac{1}{2^3}e^{3t} + \dots \end{aligned}$$

*We see that $S_X = \{1, 2, 3, \dots\}$, and that*

$$\mathbf{P}(X = x) = \frac{1}{2^x}\,, \quad x = 1, 2, 3, \dots$$

What about its name? Moment generating function..

Recall the expansion

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}\,, \qquad \forall x \in \mathbb{R}\,.$$

OK. So let's expand the exponential inside the expectation:

$$M_X(t) = \mathbf{E}\left( \sum_{k=0}^{\infty} \frac{(tX)^k}{k!} \right) = \sum_{k=0}^{\infty} \mathbf{E}\left(X^k\right) \frac{t^k}{k!}\,.$$

Moments appear in its expansion! Hence its name.

Next, consider derivatives (with respect to $t$, of course) of the mgf

$$M_X'(t) = \left( \sum_{x \in S_X} e^{tx}\, p_X(x) \right)' = \sum_{x \in S_X} x\, e^{tx}\, p_X(x)\,,$$

so that

$$M_X'(0) = \sum_{x \in S_X} x\, e^{0 \cdot x}\, p_X(x) = \sum_{x \in S_X} x\, p_X(x) = \mathbf{E}\left(X\right).$$

Similarly,

$$M_X''(0) = \sum_{x \in S_X} x^2 \, p_X(x) = \mathbf{E}\left(X^2\right).$$

More generally, the $r$-th derivative of the mgf evaluated at 0 gives the $r$-th moment of $X$, $M_X^{(r)}(0) = \mathbf{E}\left(X^r\right)$.

*The same results hold in the continuous case as well.*

*Another useful property:* if $X_1, \ldots, X_n$ are independent RVs, then the mgf of the sum $Y = X_1 + \ldots + X_n$ is

$$\begin{aligned} M_Y(t) = \mathbf{E}\left(e^{tY}\right) &= \mathbf{E}\left(e^{t(X_1+\ldots+X_n)}\right) \\ &= \mathbf{E}\left(e^{tX_1} \ldots e^{tX_n}\right) \\ &= \mathbf{E}\left(e^{tX_1}\right) \times \ldots \times \mathbf{E}\left(e^{tX_n}\right) \\ &= M_{X_1}(t) \times \ldots \times M_{X_n}(t) \, . \end{aligned}$$

A corollary of the previous result is that if $X_1, \ldots, X_n$ is a RS from a distribution with mgf $M(t)$, then

(i) mgf of $Y = \sum_{i=1}^{n} X_i$ is $M_Y(t) = (M(t))^n$.

(ii) mgf of $\bar{X}$ is $M_{\bar{X}}(t) = (M(t/n))^n$.

**Example 24.** *Let $X_1, \ldots, X_n$ be a RS from the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Then the distribution of the sample mean $\bar{X}$ is $\mathcal{N}(\mu, \sigma^2/n)$.*

*Indeed,*

$$M_{X_i}(t) = \exp\left(\mu\, t + \frac{1}{2}\sigma^2\, t^2\right)\,, \qquad i = 1, \ldots, n\,.$$

*Thus*

$$\begin{aligned} M_{\bar{X}}(t) = (M_X(t/n))^n &= \left[\exp\left(\mu\, t/n + \frac{1}{2}\sigma^2\, \frac{t^2}{n^2}\right)\right]^n \\ &= \exp\left(\mu\, t + \frac{1}{2}\frac{\sigma^2}{n}t^2\right)\,, \end{aligned}$$

*and we recognize the mgf of a $\mathcal{N}(\mu, \sigma^2/n)$ random variable.*

**Example 25.** *Suppose $X_i \sim B(p)$, and that the $X_i$'s are independent of each other. Note that*

$$M_{X_i}(t) = (1-p) + p\,e^t\,.$$

*The mgf of $Y = X_1 + \ldots + X_n$ is thus*

$$M_Y(t) = ((1-p) + p\,e^t)^n\,,$$

*the mgf of a $Bi(n,p)$. Indeed, if $Z \sim Bi(n,p)$,*

$$M_Z(t) = \sum_{z=0}^{n} \binom{n}{z} e^{tz}\,p^z\,(1{-}p)^{n-z} = \sum_{z=0}^{n} \binom{n}{z} (pe^t)^x\,(1{-}p)^{n-z} = ((1{-}p){+}p\,e^t)^n\,.$$

# Inequalities

Computing the mean and variance of sums of independent RVs is relatively straightforward. In fact, it is much simpler than computing the actual distribution of the sum. However, as we shall see, means and variances can give us some quantitative information about the probabilities themselves.

**Markov Inequality.** Let $X$ be a non-negative RV. Then, for all $x > 0$,

$$\mathbf{P}(X \geq x) \leq \frac{\mathbf{E}\,(X)}{x}\,.$$

**Chebyshev Inequality** If $\mu = \mathbf{E}\,(X)$ and $\sigma^2 = \mathrm{Var}(X)$, then

$$\mathbf{P}(|X - \mu| \geq x) \leq \frac{\sigma^2}{x^2}\,, \quad x > 0\,.$$

Of course, you would expect these inequalities to be very crude, as they only make use of the first two moments. And this is indeed the case, we can get much better bounds, for example using higher order moments, or $\mathbf{E}\left(e^{\lambda X}\right)$. For example, if $X \sim \mathcal{N}(\mu, \sigma)$,

| | | |
|---|---|---|
| $k = 1$ | $\mathbf{P}(\lvert X - \mu\rvert \geq \sigma) \leq 1$ | 0.3174 |
| $k = 2$ | $\mathbf{P}(\lvert X - \mu\rvert \geq 2\,\sigma) \leq \frac{1}{4}$ | 0.0456 |
| $k = 2$ | $\mathbf{P}(\lvert X - \mu\rvert \geq 3\,\sigma) \leq \frac{1}{9}$ | 0.0027 |

But still, it can be useful..

# Convergence in probability and the LLN

Let $X_1, \ldots, X_n$ be a RS from a distribution with mean $\mu$ and variance $\sigma^2$. We know that

$$\mathbf{E}\,(\bar{X}) = \mu \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\,.$$

Applying Chebyshev inequality to the sample mean $\bar{X}$, we have for all $\epsilon > 0$,

$$\mathbf{P}(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\,\epsilon^2}\,.$$

Therefore,

$$0 \leq \lim_{n\to\infty} \mathbf{P}(|\bar{X} - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\,\epsilon^2} = 0\,,$$

which implies that

$$\lim_{n\to\infty} \mathbf{P}(|\bar{X} - \mu| \geq \epsilon) = 0\,.$$

Equivalently,

$$\forall\epsilon > 0 \qquad \lim_{n\to\infty} \mathbf{P}(|\bar{X} - \mu| \le \epsilon) = 1\,.$$

One says: $\bar{X}$ converges to $\mu$ *in probability* as $n \to \infty$, and we write

$$\bar{X} \overset{P}{\to} \mu\,.$$

This fact is known as the *Law of Large Numbers* (LLN).

## Back to frequency interpretation of probability

Let $A$ be some event, and $A_j =$"event $A$ occurs in the $j$-th replication of simple experiment". Consider

$$X_j = \mathbf{1}(A_j) = \begin{cases} 1 & \text{if } A_j \text{ occurs} \\ 0 & \text{otherwise}\,. \end{cases}$$

Then

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) = \frac{1}{n}(\mathbf{1}(A_1) + \ldots + \mathbf{1}(A_n)) = \frac{n_A}{n} \, .$$

Moreover,

$$\mathbf{E}\,(X_j) = \mathbf{E}\,(\mathbf{1}(A_j)) = \mathbf{P}(A_j) =: p$$

(event $A_j$ occurs with probability $p$). Then LLN says that

$$\forall \epsilon > 0 \qquad \mathbf{P}\left(\left|\frac{n_A}{n} - p\right| > \epsilon\right) \to 0 \qquad \text{as } n \to \infty.$$

Wow! Our probability theory reproduced the main empirical fact about random experiments: the convergence of relative frequencies. We began this course by saying we would like to model just that..

# Convergence in distribution and the CLT

The main use of moment generating functions is not for computing moments of RVs, but in deriving approximations to the distribution of various complex objects, e.g. sums $S_n = X_1 + \ldots + X_n$.

We have already mentioned that mgfs uniquely determine the distribution/law of a RV. In fact, it's even better than that! When mgf $M_X(t)$ is close to mgf $M_Y(t)$ as a function, one can show that the distribution of $X$ will be close to that of $Y$. If

$$M_{X_n} \to M_X(t) \quad \text{as} \quad n \to \infty \,,$$

then

$$F_{X_n}(x) \to F_X(x) \quad \text{for all } x \text{ such that } F_X(x) = F_X(x-0) \,,$$

and one says

$$\text{“}X_n \text{ converges to } X \text{ in distribution”},$$

and we write

$$X_n \overset{d}{\to} X \,.$$

This is another kind of convergence of RVs. In fact, one can show that convergence in probability implies convergence in distribution, but the converse is not true in general.

**LLN revisited.** Let $X_1, \ldots, X_n$ be i.i.d. RVs and $S_n = X_1 + \ldots + X_n$. Denoting $\mathbf{E}(X) = \mu$,

$$\frac{S_n}{n} \xrightarrow{d} \mu \quad \text{as } n \to \infty\,.$$

Now, we look at the LLN through a magnifying glass.

**Central Limit Theorem (CLT).** Let $X_1, \ldots, X_n$ be i.i.d. RVs and $S_n = X_1 + \ldots + X_n$. If $\mathbf{E}(X) = \mu$ and finite $\mathrm{Var}(X) = \sigma^2 > 0$, then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \qquad \text{as} \quad n \to \infty\,.$$

In words, for large $n$, $S_n \overset{d}{\approx} \mathcal{N}(n\mu, n\sigma^2)$, or $\bar{X} \overset{d}{\approx} \mathcal{N}(\mu, \sigma^2/n)$.

**Remarks.**

(i) We have learnt from example 25 that if $X_1, \ldots, X_n$ is a RS from the $\mathcal{N}(\mu, \sigma^2/n)$ distribution, then

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n) \qquad \text{and} \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)\,.$$

Of course, these two results do not hold for general underlying distributions, but the CLT tells us that they still approximately hold if the sample size is large enough. Hence the special status of the normal distribution in the pool of distribution functions! The quality of the approximation will depend on the underlying probability distribution. Usually, take $n > 25$ or 30, but $n$ can be much smaller if the distribution is symmetric and smooth. In any case,

$$\bar{X} \overset{d}{\approx} \mu + \frac{\sigma}{\sqrt{n}}\, Z \qquad Z \sim \mathcal{N}(0, 1)\,.$$

The second term $\frac{\sigma}{\sqrt{n}}\, Z$ is a random error, and typical values are of order $\approx n^{-1/2}$.

(ii) The condition of independence between RVs can be relaxed to some extent without affecting the validity of the CLT.

(iii) The LLN only assumes finite mean. The CLT assumes finite variance as well.

(iv) In more advanced probability courses, we can also investigate the rate of convergence involved in the CLT, provided the third order moment is finite. [$\rightarrow$ Berry-Esseen Theorem.]

And more, if higher order moments are finite! [$\rightarrow$ Edgeworth expansions.]

**Example 26.** *Let $\bar{X}$ be the mean of a RS of size $n = 25$ from a distribution whose pdf if*

$$f_X(x) = \frac{x^3}{4}\,, \qquad 0 < x < 2\,.$$

*Then*

$$\mu_X = \int_0^2 x\,\frac{x^3}{4}\,dx = \left[\frac{x^5}{20}\right]_0^2 = 1.6\,, \quad \sigma_X^2 = \int_0^2 (x - 8/5)^2\,\frac{x^3}{4}\,dx = \ldots = 8/75\,.$$

*Using CLT,*

$$\begin{aligned} \mathbf{P}(1.5 \le \bar{X} \le 1.65) &= \mathbf{P}\left(\frac{1.5 - 1.6}{\sqrt{8/75}/5} \le \frac{\bar{X} - 1.6}{\sqrt{8/75}/5} \le \frac{1.65 - 1.6}{\sqrt{8/75}/5}\right) \\ &= \mathbf{P}(-1.531 \le Z \le 0.765) \\ &\approx \phi(0.765) - \phi(-1.531) \\ &= 0.7150\,, \end{aligned}$$

*where $Z \sim \mathcal{N}(0,1)$.*

# Approximations for discrete distributions

The binomial and Poisson RVs can be expressed as a sum of i.i.d. RVs. The CLT can thus be used to approximate the probabilities associated with these discrete RVs.

**Normal approximation for the binomial distribution.**

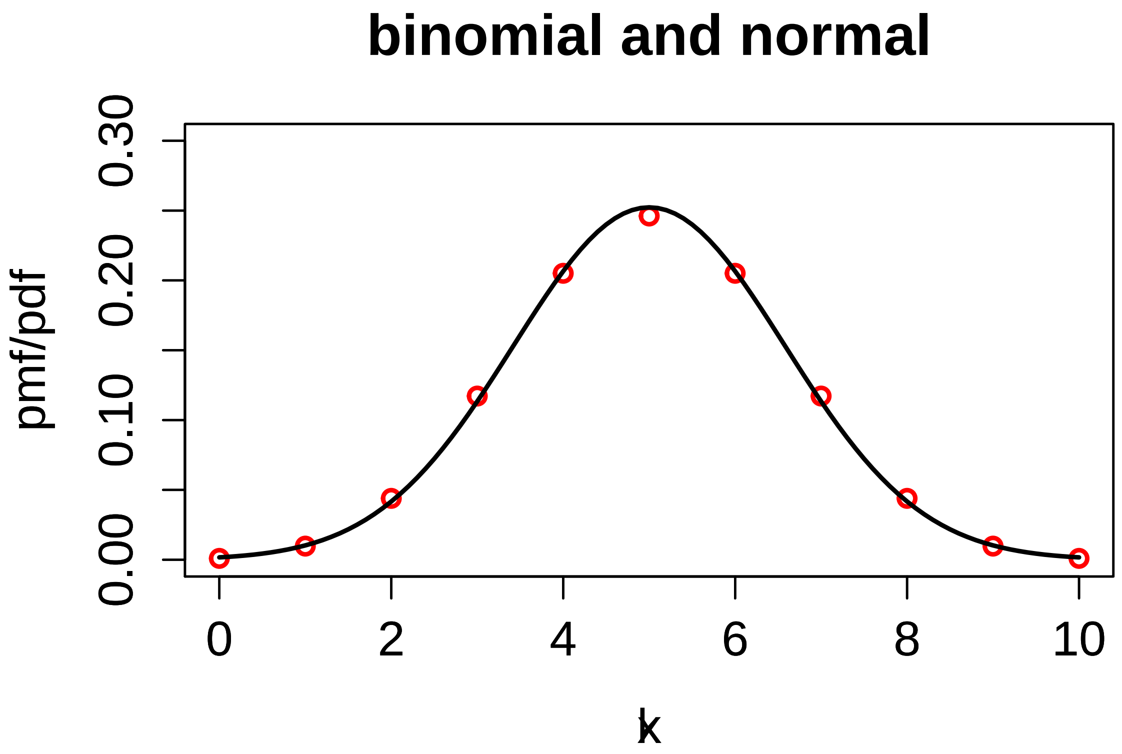
Let $X \sim \mathrm{Bi}(n,p)$. There exist $n$ independent Bernoulli RVs $X_1, \ldots, X_n$, each having common mean $p$ and common variance $p\,(1-p)$ such that $X = X_1 + \ldots + X_n$. We showed this earlier using mgfs. By CLT,

$$\frac{X/n - p}{\sqrt{p(1-p)/n}} = \frac{X - np}{\sqrt{np(1-p)}} \approx \mathcal{N}(0,1)\,.$$

Equivalently,

$$X \overset{d}{\approx} \mathcal{N}(n\,p,\, n\,p\,(1-p))\,.$$

Plot of the $\mathrm{Bi}(n, p)$ and $\mathcal{N}(n\,p,\; n\,p\,(1-p))$ distributions, $n = 10$ and $p = 1/2$.



$\mathbf{P}(X = 4) = \binom{10}{4}\, 0.5^4\, 0.5^6 = 0.2051.$ We approximate this probability with the area under the normal curve which sits on the base $[3.5,\; 4.5]$.

$$
\begin{aligned}
\mathbf{P}(X=4)=\mathbf{P}(3.5\leq X\leq 4.5) &= \mathbf{P}\left(\frac{3.5-5}{\sqrt{2.5}}\leq\frac{X-5}{\sqrt{2.5}}\leq\frac{4.5-5}{\sqrt{2.5}}\right)\\
&\approx \mathbf{P}(-0.9486\leq Z\leq -0.3162)\\
&= \phi(-0.3162)-\phi(-0.9486)\\
&= 0.2025\,.
\end{aligned}
$$

The first step is referred to as *continuity correction*.

Why do we use continuity correction? It usually improves the accuracy of the approximation (remember, you are approximating a *discrete* distribution with a *continuous* one). Let $X\sim \mathrm{Bi}(n,p)$ and $Z\sim\mathcal{N}(np,np(1-p))$,

$$
F_X(x)=\sum_{i=0}^{x}\mathbf{P}(X=i)\approx\sum_{i=0}^{x}\left(F_Z(i+1/2)-F_Z(i-1/2)\right)=F_Z(x{+}1/2){-}F_Z(-0.5)\,,
$$

where $F_Z(-0.5)\approx 0$.

Summarising, in general, for the binomial distribution,

$$\begin{aligned}\mathbf{P}(X=k) &= \mathbf{P}\left(k-\frac{1}{2}\leq X\leq k+\frac{1}{2}\right)\\ &= \mathbf{P}\left(\frac{k-\frac{1}{2}-np}{\sqrt{n\,p\,(1-p)}}\leq\frac{X-\frac{1}{2}-np}{\sqrt{n\,p\,(1-p)}}\leq\frac{k+\frac{1}{2}-np}{\sqrt{n\,p\,(1-p)}}\right).\\ \mathbf{P}(x<X\leq y) &= \mathbf{P}(x+\frac{1}{2}\leq X\leq y+\frac{1}{2})\approx\ldots.\\ \mathbf{P}(x\leq X<y) &= \mathbf{P}(x-\frac{1}{2}\leq X\leq y-\frac{1}{2})\approx\ldots.\end{aligned}$$

*Rule of thumb:* $n$ sufficiently large usually means $np\geq 5$ and $n(1-p)\geq 5$.

## Normal approximation for the Poisson distribution

Let $X \sim \mathrm{P}(\lambda)$. Its mgf is

$$M_X(t) = \exp\left(\lambda\left(e^t - 1\right)\right) = \left[\exp\left(\frac{\lambda}{\lfloor\lambda\rfloor}\left(e^t - 1\right)\right)\right]^{\lfloor\lambda\rfloor},$$

for $\lfloor\lambda\rfloor =$ integer part of $\lambda$. Thus $M_X(t) = M_{X_1}(t) \times \ldots \times M_{X_{\lfloor\lambda\rfloor}}(t)$, and there exists $\lfloor\lambda\rfloor$ independent RVs $X_1, \ldots, X_{\lfloor\lambda\rfloor}$ each $\sim \mathrm{P}(\lambda/\lfloor\lambda\rfloor)$ such that

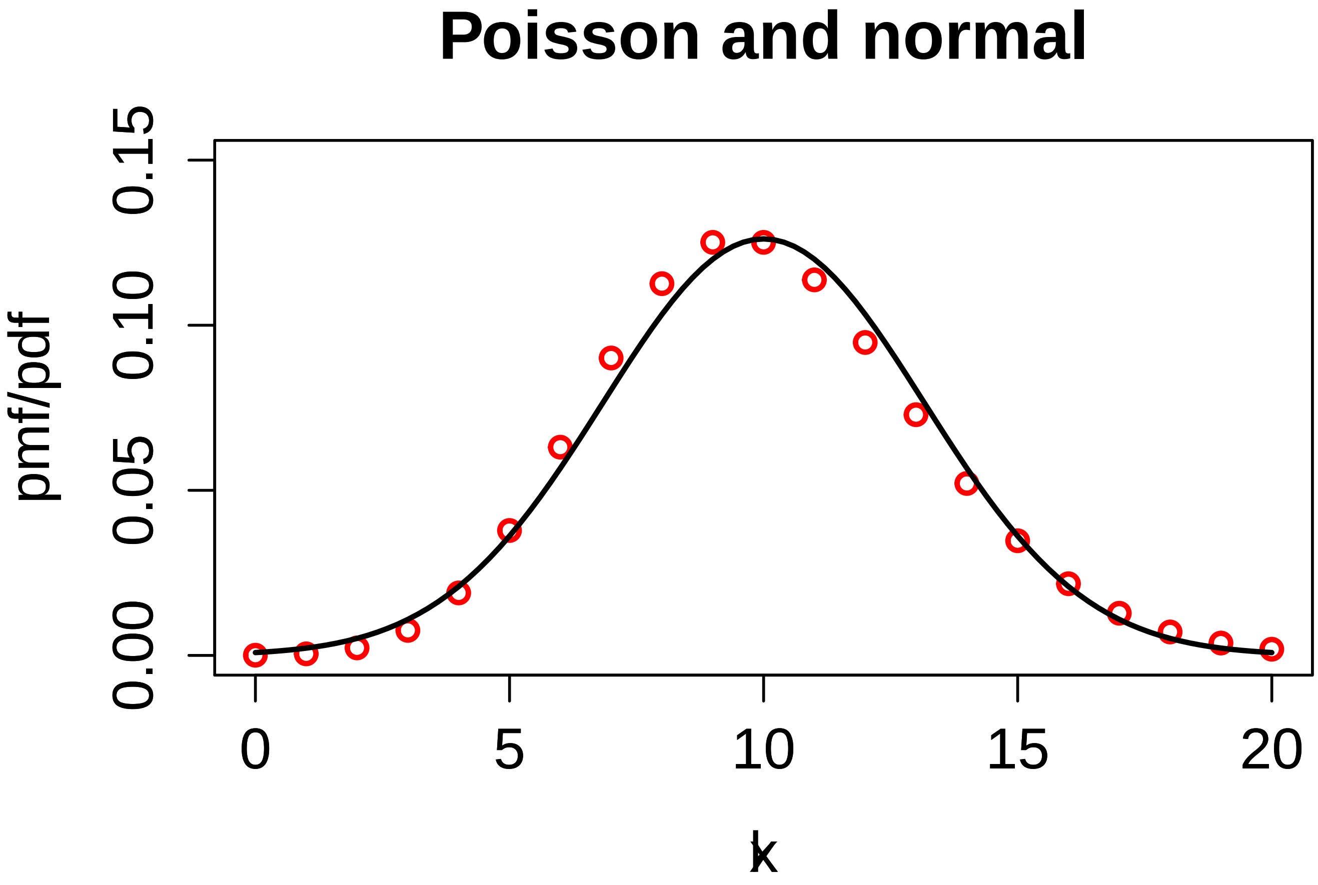
$$X \overset{d}{=} X_1 + \ldots + X_{\lfloor\lambda\rfloor}\,.$$

Hence, by CLT,

$$\frac{X - \lfloor\lambda\rfloor\frac{\lambda}{\lfloor\lambda\rfloor}}{\sqrt{\lfloor\lambda\rfloor\frac{\lambda}{\lfloor\lambda\rfloor}}} = \frac{X-\lambda}{\sqrt{\lambda}} \approx \mathcal{N}(0,1)\,,$$

when $\lfloor\lambda\rfloor$ (and thus $\lambda$) is sufficiently large. Equivalently,

$$X \overset{d}{\approx} \mathcal{N}(\lambda,\lambda)\,, \qquad \text{for } \lambda \text{ large enough.}$$



Continuity correction applies here as well!