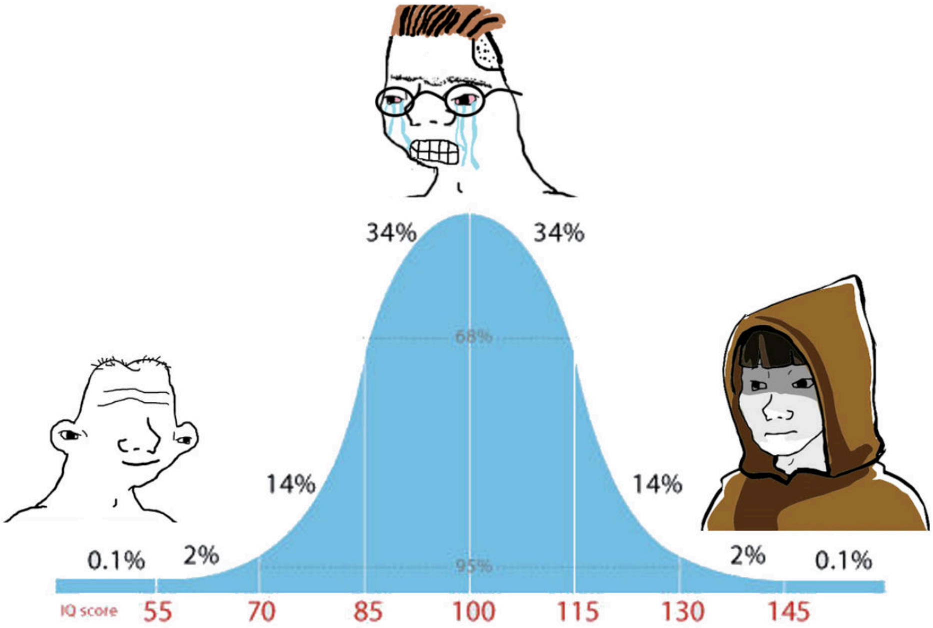
Contents



1. Introduction

1.1. Statistics

The discipline of statistics instructs how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.

Collections of facts are called **data**: statistics provides methods for organizing, summarizing and drawing conclusions based on information contained in the data.

A statistical enquiry will typically focus on a well-defined collection of objects constituting a **population**. When desired information is available for all objects in the population, a **census** is available.

In general, such a situation is hardly possible, either because it would be too expensive or too time consuming to do so or simply because the population has an infinite amount of members. A more reasonable approach is to extract a subset of the population, called **sample** that is both sufficiently small to be able to work with and sufficiently large to capture all the nuances of the population as a whole.

Each object of the population possesses many features, some of which may or may not be of interest. Any feature whose value might change from object to object in the population and that has relevance with respect to a statistical enquiry is called a **variable**.

Variables are generally distinct in **numerical** variables and **categorical** variables. Numerical variables are distinct in **discrete** and **continuous**. Numerical variables are discrete if the set of its possible values is either finite or countably infinite. Numerical variables are continuous if the set of its possible values is uncountably infinite. Categorical variables are distinct in **ordinal** and **nominal**. Categorical variables are ordinal if the set of its possible values obeys an objective hierarchy or ordering of some sort, otherwhise are called **nominal**.

Exercise 1.1.1: Provide an example for each of the four types of variables.

Solution:

- A numerical discrete variable could be the number of items sold in a store, since such number is necessarely an integer (it's not possible to sell, say, half an item, or three quarters of an item). Another example is the number of attempts necessary to win the lottery: it could be infinite, but it's still countable;
- A numerical continuous variable could be the temperature measured in a certain meteorological station, since such value is a real number (it could be approximated to an integer, but it would entail losing much informaton);
- A categorical ordinal variable could be the ranks in an army, such as general, private, captain, etcetera. Such ranks can be arranged in a (very) strict hierarchy, for example corporal is lower than general while corporal is higher than private;
- A categorical nominal variable could be the colors of a dress. It would make little sense to say that, for example, red scores higher than green or that pink scores lower than blue, at least in an objective way.

□

When referring to "statistics", it often entails two distinct concepts. The first one is **descriptive statistics**, that consists in summarizing and describing the data, in general through graphical representations (called **plots**) or through **summary measures**, numbers that on their own represent an aspect of the data as a whole.

The second one is **inferential statistics**, that consists in drawing conclusions about the population as a whole from the sample extracted from such population. In this case, a sampling is a means to an end, not an end in itself.

Given a discrete numerical variable x, let $x_1, x_2, ..., x_n$ be the observations collected from the sample of such variable, with n being the cardinality of the sample. The **sample mean** $\overline{x}$ is a summary measure that describes its average value, and is calculated as:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Given a discrete numerical variable x, let $x_1, x_2, ..., x_n$, be the observations collected from the sample of such variable, arranged from lowest to highest (including duplicates). The **sample median** $\tilde{x}$ is a summary measure that describes the central value, and is calculated as either the middle value of such sequence if n is odd or the average of the two middle values if n is even:

$$\tilde{x} = \begin{cases} \text{The } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ value if } n \text{ is odd} \\ \text{The average of the } \left(\frac{n}{2}\right)^{\text{th}} \text{ and the } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ value if } n \text{ is even} \end{cases}$$

The **sample variance** s^2 is a summary measure that describes how "spread out" are the values of the sample, or equivalently how close its values are to the sample mean, and is defined as:

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n-1} = \frac{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$$

The **sample standard deviation** is defined as the square root of the sample variance:

$$s = \sqrt{s^2}$$

Theorem 1.1.1: Given a discrete numerical variable x, let $x_1, x_2, \dots, x_n$, be the observations collected from the sample of such variable, and let c be a numerical constant. Then:

1. If, for each $1 \leq i \leq n$, the y variable is constructed as $y_i = x_i + c$, it is true that $s_y^2 = s_x^2$;
2. If, for each $1 \leq i \leq n$, the y variable is constructed as $y_i = cx_i$, it is true that $s_y^2 = c^2 s_x^2$;

Where s_x^2 is the sample variance of the "original" variable x and s_y^2 is the sample variance of the "transformed" variable y.

1.2. Probability

Probability provides methods to quantify chance and randomness related to a certain event. Any activity or process having at least one outcome, all being random (not knowable in advance) is called an **experiment**. The set containing all possible outcomes of an experiment, denoted as $\mathcal{S}$, is called **sample space**.

Exercise 1.2.1: Provide some examples of experiments.

Solution:

- The roll of a six-sided dice is an experiment, since the resulting value of the dice is unknown until the dice is rolled. The sample space $\mathcal{S}$ contains 6 elements:

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$

- The drawing of a card from a (standard) deck is an experiment, since the value of the card is unknown until the card is drawn. The sample space $\mathcal{S}$ contains 52 elements:

$$\mathcal{S} = \{A\heartsuit, 2\heartsuit, \dots, Q\heartsuit, K\heartsuit, A\diamondsuit, 2\diamondsuit, \dots, Q\diamondsuit, K\diamondsuit, A\clubsuit, 2\clubsuit, \dots, Q\clubsuit, K\clubsuit, A\spadesuit, 2\spadesuit, \dots, Q\spadesuit, K\spadesuit\}$$

- The gender assigned to the offspring of a couple is an experiment, since their gender is unknown until (roughly) 4 months since conception. The sample space $\mathcal{S}$ contains 8 elements:

$$\mathcal{S} = \{MMM, MMF, MFM, FMM, MMF, FFM, FMF, FFF\}$$

□

Any subset of the sample space is called an **event**. An event can be either **simple** if it's a singleton (it contains a single outcome of the experiment) or **compound** otherwise (it contains multiple outcomes). An event can either occur or not occur, depending on the outcome of the experiment.

Exercise 1.2.2: Provide some examples of events.

Solution:

- Consider the roll of a six-sided dice. The subset $A = \{1, 3, 5\}$ of the sample space $\mathcal{S}$ corresponds to the event “an even number”. It is a compound event;
- Consider the drawing of a card from a deck. The subset $B = \{A\heartsuit, A\diamondsuit, A\clubsuit, A\spadesuit, K\heartsuit, K\diamondsuit, K\clubsuit, K\spadesuit\}$ of the sample space $\mathcal{S}$ corresponds to the event “either an ace or a king of any set”. It is a compound event;
- Consider the gender assigned to the offspring of a couple. The subset $C = \{FFF\}$ of the sample space $\mathcal{S}$ corresponds to the event “exclusively female offspring”. It is a simple event.

□

Being sets, events can be manipulated using set algebra. In particular, given two events A and B:

- The **complement** of A, denoted as A^c, corresponds to the event containing all outcomes not contained in A. That is, A^c occurs if and only if A does not occur. A^c is also called the **complementary event** of A;
- The **intersection** of A and B, denoted as $A \cap B$, corresponds to the event containing all outcomes contained both in A and in B. That is, $A \cap B$ occurs if and only if both A and B occur at the same time;
- The **union** of A and B, denoted as $A \cup B$, corresponds to the event containing all outcomes contained either in A, in B or in both. That is, $A \cup B$ occurs if at most A or B occurs.

Exercise 1.2.3: Provide some examples of complemented, intersected and unified events.

Solution:

- Consider the roll of a six-sided dice. The subset $A = \{1, 2, 3, 4, 5\}$ of the sample space $\mathcal{S}$ corresponds to the event “any number but 6”. It is the complement of the event “exactly six”;
- Consider the drawing of a card from a deck. The subset $B = \{A\heartsuit, A\diamondsuit, A\clubsuit, A\spadesuit, K\heartsuit, K\diamondsuit, K\clubsuit, K\spadesuit\}$ of the sample space $\mathcal{S}$ is actually a union of two smaller events, the first being “an ace of any set” and the second being “a king of any set”;
- Consider the gender assigned to the offspring of a couple. Consider the two events “a male as first born” and “a female as third born”. Their intersection, representing the event “a male as first born and a female as third born” is given by:

$$\{MMM, MMF, MFM, MFF\} \cap \{MMF, MFF, FMF, FFF\} = \{MMF, MFF\}$$

□

The empty set $\emptyset$ denotes the event of having no outcome whatsoever, also called the **null event**. If the intersection of two events is the null event, such events are said to be **mutually esclusive** events, or **disjoint** events. In other words, two events are said to be mutually esclusive if they have no way of happening at the same time.
Modern probability theory, like set theory, is defined axiomatically. Such axioms are also called **Kolmogorov axioms**, and are (supposed to be) the minimum amount of axioms that are needed to construct a theory of probability free of contradictions.
To an event A, it is possible to associate a value called its **probability**, denoted as $P(A)$, that represents a measure of likelihood, certainty or confidence of such event to occur (intuitively, the higher the value of probability, the higher the likelihood). Probabilities obey three axioms, here stated:

1. For any event A, $P(A) \geq 0$. That is, the probability of an event happening is non negative;
2. $P(\mathcal{S}) = 1$. That is, the probability of any even happening at all is fixed as 1;
3. If $A_1, A_2, \dots$ is a collection of countably infinite disjoint events, the following equality holds:

$$P(A_1 \cup A_2 \cup \dots) = \sum_{i=1}^{\infty} P(A_i)$$

That is, given a set of events where no event can occur if at most another one of them occurs, the probability of any such event to occur is the sum of the individual probabilities.

From such axioms, it is possible to derive many useful consequences.

Theorem 1.2.1: $P(\emptyset) = 0$. That is, the null event cannot occur.

Proof: Consider the countably infinite collection of events $\emptyset, \emptyset, \ldots$. By definition, the null event is disjoint with itself, since set algebra gives $\emptyset \cap \emptyset = \emptyset$. The collection $\emptyset, \emptyset, \ldots$ is therefore made up of disjoint events, and by set algebra $\emptyset \cup \emptyset \cup \ldots = \emptyset$, therefore $P(\emptyset \cup \emptyset \cup \ldots) = P(\emptyset)$. Since by axiom 3 $P(\emptyset \cup \emptyset \cup \ldots) = \sum_{i=1}^{\infty} P(\emptyset)$, by transitive property $\sum_{i=1}^{\infty} P(\emptyset) = P(\emptyset)$. Since by axiom 1 the value of $P(\emptyset)$ has to be non negative, such equality can hold exclusively if $P(\emptyset) = 0$. □

Theorem 1.2.2: If $A_1, A_2, \ldots, A_n$ is a collection of finitely many disjoint events, the following equality holds:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

Proof: Consider the countably infinite collection of events $A_1, A_2, \ldots, A_n, A_{n+1} = \emptyset, A_{n+2} = \emptyset, \ldots, \emptyset$, that is, a collection constructed by encoding countably infinitely many null events to the original collection. Applying axiom 3 to such collection gives:

$$P(A_1 \cup A_2 \cup \ldots A_n \cup \emptyset \cup \emptyset \cup \ldots \cup \emptyset) = \sum_{i=1}^{\infty} P(A_i)$$

It is possible to split the summation in two like so:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) + P(\emptyset \cup \emptyset \cup \ldots \cup \emptyset) = \sum_{i=1}^{n} P(A_i) + \sum_{i=n+1}^{\infty} P(\emptyset)$$

But by Theorem 1.2.1, $P(\emptyset) = 0$. Therefore:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) + P(\emptyset \cup \emptyset \cup \ldots \cup \emptyset) = P(A_1 \cup A_2 \cup \ldots \cup A_n) + 0 = \sum_{i=1}^{n} P(A_i)$$

□

Theorem 1.2.3: For any event A, $P(A) + P(A^c) = 1$.

Proof: By definition of complementary event, $A \cup A^c = \mathcal{S}$. They are also disjoint events, since one cannot happen it the other one happened. It is therefore possible to apply Theorem 1.2.2 and state that $\sum_{i=1}^{2} P(A_i) = P(A) + P(A^c) = P(A \cup A^c)$. But, as stated, $A \cup A^c = \mathcal{S}$, and by axiom 2 $P(\mathcal{S}) = 1$. Therefore, by transitive property, $P(A) + P(A^c) = 1$. □

Theorem 1.2.4: For any event A, $0 \leq P(A) \leq 1$.

Proof: By Theorem 1.2.3, $P(A) + P(A^c) = 1$. By axiom 1, both probabilities are greater or equal than 0, therefore, for the equality to hold, both probabilities have to be lower or equal than 1. Combining the two boundaries, $0 \leq P(A) \leq 1$. □

Theorem 1.2.5: For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: By set algebra, the event $A \cup B$ can itself be seen as the union of two disjoint events, A and $A^c \cap B$. It is therefore possible to apply Theorem 1.2.2, resulting in:

$$P(A \cup B) = P(A \cup (A^c \cap B)) = P(A) + P(A^c \cap B)$$

In the same fashion, the event B can be seen as the union of the disjoint events $A \cap B$ and $A^c \cap B$. Applying Theorem 1.2.2 gives:

$$P(B) = P((A \cap B) \cup (A^c \cap B)) = P(A \cap B) + P(A^c \cap B)$$

Moving $P(A \cap B)$ to the left side gives $P(B) - P(A \cap B) = P(A^c \cap B)$. Substituting such expression in the first equation gives $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. □

Theorem 1.2.6: For any three events A, B and C:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Proof: Works similarly as Theorem 1.2.5. □

Theorem 1.2.7 (Boole's inequality): Given any countable set of events $A_1, A_2, ..., A_n$:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$$

It should be stressed that the Kolmogorov axioms simply describe the rules by which probability works, but do not define the probability of any event itself. Infact probabilities can be assigned to any event in any possible way that is constrained by the axioms, but such value can have no bare on reality or on intuition and yet construct a model that is consistent.

Exercise 1.2.4: Provide an example of a probability model that constrasts with reality but obeys Kolmogorov's axioms.

Solution: Consider the toss of a coin. Such action can be conceived as an experiment, since whose side the coin is gonna land when tossed is unknown until the coin lands. Only two events are possible, heads or tails; since a coin cannot land on both sides at the same time, such events are disjoint.
It is a known fact that the probability of both events is 0.5, and indeed such assignment respects all of three axioms. But by choosing the assignment, say, 0.2 to the landing on heads and 0.8 to the landing on tails, no axiom is violated, even though such an assignment has very little resonance with experience or common sense.
This does not mean that probabilities can be assigned at libitum, since they still ought to comply with the axioms. For example, assigning 0.4 to the probability of the coin to land on heads and 0.3 to the probability of the coin to land on tails won't do, since axiom 2 would be violated. As another example, assigning 1.5 to the probability of the coin to land on heads and -0.5 to the probability of the coin to land on tails would violate axiom 1, and therefore invalid. □

The appropriate or correct assignment depends on how one *interprets* probability, that is to say how one intends the link between the mathematical treatment of probability and the physical world. This quest is just as philosophical as mathematical.

One possible and often invoked interpretation of probability is the **objective** interpretation, also called **frequentist** interpretation. Consider an experiment that can be repeatedly performed in an identical and independent fashion, and let A be an event consisting of a fixed set of outcomes of the experiment. If the experiment is performed n times, the event A will occur $n(A)$ times (with $0 \leq n(A) \leq n$) and will not occur $n - n(A)$ times. The ratio $n(A)/n$ is called the **relative frequency** of occurrence of the event A in the sequence of n attempts.
Empirical data suggests that the relative frequency fluctuates considerably if n is a small number, while tends to stabilize itself as n grows. Ideally, repeating such experiment infinitely many times, it would be possible to obtain a "perfect" frequency, called **limiting relative frequency**. The objective interpretation of probability states that this limiting relative frequency is indeed the probability of A to occurr.
This interpretation of probability is said to be objective in the sense that it rests on a property of the experiment and not on the concerns of the agent performing it (ideally, two agents performing the same experiment the same number of times would obtain the same relative limiting frequency, and therefore the same probability).
This interpretation has limited appliability, since not all events can be performed n number of times to draw similar conclusions. In situations such as these, it makes more sense to interpret probability in a **subjective** way, which can be thought of as the "degree of confidence" with which an agent believes an event to occur.
The simplest situation to model is the one where to each simple event $E_1, E_2, ..., E_N$ is assigned the same value of probability $P(E_i)$:

$$1 = \sum_{i=1}^{N} P(E_i) \Rightarrow P(E_i) = \frac{1}{N}$$

That is, if there are N equally likely outcomes, the probability of one of such outcomes to happen is $1/N$.
More generally, consider an event A containing $N(A)$ number of outcomes. Then the task of computing probabilities reduces itself to **counting**:

$$P(A) = \sum_{E_i \in A} P(E_i) = \sum_{E_i \in A} \frac{1}{N} = \frac{N(A)}{N}$$

Given two events A and B with $P(B) > 0$, the probability of A to occur given that B occurred is called the **conditional probability** of A given B, and is given as:

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

Theorem 1.2.8 (Law of total probability): Let $A_1, A_2, ..., A_n$ be a finite partition of a sample space $\mathcal{S}$ such that no event has assigned zero probability, and let B be any event in $\mathcal{S}$. Then:

$$P(B) = P(B \mid A_1)P(A_1) + ... + P(B \mid A_n)P(A_n) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

Theorem 1.2.9 (Bayes' theorem): Let $A_1, A_2, ..., A_n$ be a finite partition of a sample space $\mathcal{S}$. Each event A_j has a probability $P(A_j)$, also called its **prior probability**, that is non zero. Let B be any event in $\mathcal{S}$ whose probability is non zero. The probability $P(A_j \mid B)$, also called the **posterior probability**, is given as:

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{\infty} P(B \mid A_i)P(A_i)}$$

Exercise 1.2.5: An electronics store sells three different brands of DVD players. Of its DVD player sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3. Each manufacturer offers a 1 -year warranty on parts and labor. It is known that 25% of brand 1's DVD players require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.

1. What is the probability that a randomly selected purchaser has bought a brand 1 DVD player that will need repair while under warranty?
2. What is the probability that a randomly selected purchaser has a DVD player that will need repair while under warranty?
3. If a customer returns to the store with a DVD player that needs warranty repair work, what is the probability that it is a brand 1 DVD player?

Consider two events, A and B, the second happening after the first. The fact that B occurred may or may not influence the probability of A to occur. If the probability of A to happen is the same whether or not B happened, that is to say if $P(A)$ and $P(A \mid B)$ are equal, The event A is said to be **independent** of B. Otherwise, it's said to be **dependent** of B.

Theorem 1.2.10: Event independence is symmetric. In other words, given two events A and B, if A is independent of B, then B is independent of A.

Proof: If A is independent of B, then $P(A \mid B) = P(A)$. Applying Theorem 1.2.9 gives:

$$P(A \mid B) = \cancel{P(A)} = \frac{P(B \mid A)\cancel{P(A)}}{P(B)} \Rightarrow P(B \mid A) = P(B)$$

Which, by definition, means that B is independent of A as well. □

An equivalent definition of independent events is as follows. Given two independent events A and B, by the previous definition $P(A) = P(A \mid B)$, so:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A)P(B)$$

Event independence can be extended to a situation with more than two events. Given a collection of n events $A_1, A_2, ..., A_n$, such events are said to be **mutually independent** if for every $k = 2, 3, ..., n$ and for every subset of indices $i_1, i_2, ..., i_k$:

$$P(A_{i,1} \cap A_{i,2} \cap ... \cap A_{i,k}) = P(A_{i,1}) \cdot P(A_{i,2}) \cdot ... \cdot P(A_{i,k})$$

1.3. Discrete random variables

As already stated, Kolmogorov axioms define the properties of probability but do not offer a method for assigning them to events. The simplest approaches, such as assigning the same probability to each event, are far to weak to model reality. A more powerful concept to be introduced which can help model probability is the **random variable**.

A random variable can be conceived as a mapping from the sample space to the real line. In other words, a random variable is a function that assigns a probability to any possible event of the sample space. Given a sample space $\mathcal{S}$, a random variable X for such sample space is defined as $X : \mathcal{S} \mapsto \mathbb{R}$, and the probability of such variable to assume a certain value x of the sample space is denoted as $P(X = x)$.

Random variables fall in two broader categories: **discrete** and **continuous**. A random variable is said to be discrete if the set of values it can assume is either finite or countably infinite. A random variable is said to be continuous if the two following properties apply:

1. Its set of possible values consists either of all numbers in a single (possibly infinite) interval on the real line or all numbers in a disjoint union of such intervals;

2. The probability of the random variable to assume a specific value is always zero.

The set of values that a random variable can assume is called its **support**.
The **probability mass function** (abbreviated as pmf) of a discrete random variable X, denoted as $p(X)$, is a function that assigns a probability to each possible value that such random variable can assume. More formally, given a random variable X, for each value x of its sample space the pmf of X is defined as:

$$p(x) = P(X = x) = P(\omega : \omega \in \mathcal{S}, X(\omega) = x)$$

The **cumulative distribution function** (abbreviated as cdf) of a discrete random variable X, denoted as $F(X)$, is defined as the probability of such random variable to assume a value less than or equal to a threshold. More formally, given a random variable X, for each value x of its sample space the cdf of X is defined as:

$$F(x) = P(X \leq x) = \sum_{y : y \leq x} p(y)$$

Let X be a discrete random variable with support D and probability mass function $p(X)$. The **expected value** (or **mean value**) of X, denoted as $E(X)$ or μ_X is given by:

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

When the variable X is known, the pedix X in μ_X is omitted.
The expected value of a random variable is the equivalent of the mean with respect to populations.

Exercise 1.3.1: Let X be the Apgar score of a randomly selected child born at a certain hospital during the next year, with pmf as follows:

X	0	1	2	3	4	5	6	7	8	9	10
$p(X)$	0.002	0.001	0.002	0.005	0.02	0.04	0.18	0.37	0.25	0.12	0.01

What is the expected value of X?

Solution:

$$E(X) = \sum_{x \in D} x \cdot p(x) = 0 \cdot 0.002 + 1 \cdot 0.001 + 2 \cdot 0.002 + 3 \cdot 0.005 + 4 \cdot 0.02 + 5 \cdot 0.04 + 6 \cdot 0.18 + \\ 7 \cdot 0.37 + 8 \cdot 0.25 + 9 \cdot 0.12 + 10 \cdot 0.01 = 7.15$$

□

The expected value is oblivious to whether its argument is a random variable or a function whose input is a random variable. In other words, let X be a discrete random variable with support D and probability mass function $p(X)$, and let $h(X)$ be a function whose argument is (the realization of) the random variable X. The expected value of $h(X)$ is still defined as:

$$E(h(X)) = \mu_{h(X)} = \sum_{x \in D} h(x) \cdot p(x)$$

Theorem 1.3.1: Let X be a discrete random variable with support D and probability mass function $p(X)$. Given two coefficients a and b, the following equality holds:

$$E(aX + b) = aE(X) + b$$

Proof: Let D be the support of X and $p(X)$ its probability mass function. Let Y be the random variable $Y = aX + b$. The probability of Y to assume a particular value $ax + b$ does not depend on a and b, but only on x. Therefore, $P(Y = ax + b) = P(X = x)$. Computing the expected value for Y gives:

$$E(Y) = \sum_{x \in D} (ax + b) \cdot P(Y = ax + b) = \sum_{x \in D} (ax + b) \cdot P(X = x) = \sum_{x \in D} ax \cdot P(X = x) + b \cdot P(X = x) =$$
$$a \sum_{x \in D} x \cdot P(X = x) + b \sum_{x \in D} P(X = x) = aE(X) + b \cdot 1 = aE(X) + b$$

Where $\sum_{x \in D} P(X = x) = 1$ stems from the fact that said summation is the sum of the probabilities of the entire sample space, which is 1 by definition. □

Theorem 1.3.2: Let X and Y be two random variables. $E[X + Y] = E[X] + E[Y]$.

Let X be a discrete random variable with support D and probability mass function $p(X)$. The **variance** of X, denoted as $V(X)$ or σ_X^2 is given by:

$$V(X) = \sigma_X^2 = \sum_{x \in D} (x - E(X))^2 \cdot p(x) = E\left((X - E(X))^2\right)$$

When the variable X is known, the pedix X in σ_X^2 is omitted.
The square root of the variance is called the **standard deviation**:

$$SD(X) = \sigma_X = \sqrt{V(X)}$$

The variance and the standard deviation measure how a random variable is "spread out", in the sense of how much the values of the support of said variable are detached from its expected value.

Lemma 1.3.1: Let X be a discrete random variable with support D and probability mass function $p(X)$. The following equality holds:

$$V(X) = \left(\sum_{x \in D} x^2 \cdot p(x)\right) - (E(X))^2 = E(X^2) - (E(X))^2$$

Theorem 1.3.3: Let X be a discrete random variable with support D and probability mass function $p(X)$. Given two coefficients a and b, the following equality holds:

$$V(aX + b) = a^2 V(X)$$

Proof: Let Y be the random variable $Y = aX + b$. From Theorem 1.3.1, $E(Y) = E(aX + b) = aE(X) + b$. Substituting this expression in the variance one gives:

$$V(Y) = E\left((Y - E(Y))^2\right) = E\left((ax + \cancel{b} - aE(X) - \cancel{b})^2\right) = \sum_{x \in D} (ax - aE(X))^2 P(Y = ax + b) =$$
$$\sum_{x \in D} a^2 (x - E(X))^2 P(X = x) = a^2 \sum_{x \in D} (x - E(X))^2 P(X = x) = a^2 V(X)$$

□

Theorem 1.3.4: Let X and Y be two random variables. $V[X + Y] = V[X] + V[Y]$ if X and Y are independent.

1.4. Known discrete random variables

Some specific discrete random variables have been studied extensively, mostly because they model very well many phenomena in the real world. For this reason, such random variables have proper names. To denote that a random variable X has the same distribution as a known random variable F, the notation $X \sim F$ is used.

1.4.1. Bernoulli random variable

A discrete random variable X is distributed as a **Bernoulli random variable** of parameter $p \in [0, 1]$ (denoted as $X \sim B(p)$) if it can assume exclusively the values 1 and 0, with probabilities p and $1 - p$ respectively. The pdf and cdf of a Bernoulli random variable X of parameter p are therefore as follows:

$$p(x) = P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \qquad F(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Bernoulli random variables model experiments that have two mutually exclusive results: success ($X = 1$) or failure ($X = 0$), with nothing in between.

Theorem 1.4.1.1: The expected value and variance of a random variable $X \sim B(p)$ are as follows:

$$E(X) = p \qquad V(X) = p(1 - p)$$

Proof:

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = 0 + p = p$$

$$V(X) = (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p = \\ p^2(1 - p) + p(1 - p)^2 = (p^2 + p(1 - p))(1 - p) = \\ (\cancel{p^2} + p - \cancel{p^2})(1 - p) = p(1 - p)$$

□

1.4.2. Binomial random variable

Let $Y_1, Y_2, ..., Y_n$ be n independent and identically distributed Bernoulli random variables (all having the same parameter p). Let X be the random variable defined as the sum of all said variables:

$$X = \sum_{i=1}^{n} Y_i = Y_1 + Y_2 + ... + Y_n$$

The random variable X defined as such is distributed as a **binomial random variable** of parameters p and n (denoted as $X \sim Bi(n, p)$).

Since a specific realization of X is a sum of 0s and 1s, a realization k is simply the number of Bernoulli variables that define X that had assumed value 1. The pdf and cdf of a binomial random variable X of parameters n and p are therefore as follows:

$$p(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1 - p)^{n - x} & \text{if } \begin{cases} x \in \mathbb{N} \\ x \leq n \end{cases} \\ 0 & \text{otherwise} \end{cases} \qquad F(x) = P(X \leq x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1 - p)^{n - k}$$

Theorem 1.4.2.1: The expected value and variance of a random variable $X \sim Bi(p, n)$ are as follows:

$$E(X) = np \qquad V(X) = np(1 - p)$$

Proof: This result can be proved by applying Theorem 1.3.2 and Theorem 1.3.4 (the latter can be applied since the Bernoulli random variables that constitue X are independent).

$$E(X) = E(Y_1 + Y_2 + ... + Y_n) = E(Y_1) + E(Y_2) + ... + E(Y_n) = nE(Y_i) = np$$

$$V(X) = V(Y_1 + Y_2 + ... + Y_n) = V(Y_1) + V(Y_2) + ... + V(Y_n) = nV(Y_i) = np(1 - p)$$

Where $E(Y_i)$ and $V(Y_i)$ are retrieved from Theorem 1.4.1.1. □

Binomial random variables model experiments composed by many mutually exclusive results.

1.4.3. Poisson random variable

Let Y a binomial random variable, and let $\lambda \in \mathbb{R}^+$ be the product of its parameters n and p. By applying the double limit $n \to \infty, p \to 0$ while keeping their product constant a new random variable X is constructed, called a **Poisson random variable** (denoted as $X \sim \text{Pois}(\lambda)$). The pdf and cdf of a Poisson random variable X of parameter λ are therefore as follows:

$$p(x) = P(X = x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases} \qquad F(x) = P(X \leq x) = \sum_{k \in \mathbb{N}, k \leq x} \frac{\lambda^k}{k!} e^{-\lambda}$$

Theorem 1.4.3.1: The expected value and variance of a random variable $X \sim \text{Pois}(\lambda)$ are as follows:

$$E(X) = \lambda \qquad V(X) = \lambda$$

Proof: Let $Y \sim Bi(n, p)$ be a random variable to which the double limit $n \to \infty, p \to 0$ is applied, and let $\lambda = np$. This results in:

$$E(X) = \lim_{\substack{n \to \infty \\ p \to 0}} E(Y) = \lim_{\substack{n \to \infty \\ p \to 0}} np = \lambda \qquad V(X) = \lim_{\substack{n \to \infty \\ p \to 0}} V(Y) = \lim_{\substack{n \to \infty \\ p \to 0}} np(1-p) = \lambda(1 - 0) = \lambda$$

□

Exercise 1.4.3.1: Let X be the number of traps occurring in a particular type of transistor. Suppose $X \sim \text{Pois}(2)$; what is the probability of retrieving 3 traps? What is the probability of retrieving 3 or less traps?

Solution:

$$p(3) = P(X = 3) = \frac{2^3}{3!} e^{-2} = \frac{8}{6} e^{-2} \approx 0.18$$

$$F(3) = P(X \leq 3) = \sum_{k \in \mathbb{N}, k \leq 3} \frac{2^k}{k!} e^{-2} =$$

$$\frac{2^0}{0!} e^{-2} + \frac{2^1}{1!} e^{-2} + \frac{2^2}{2!} e^{-2} + \frac{2^3}{3!} e^{-2} =$$

$$e^{-2} \left(\frac{1}{1} + \frac{2}{1} + \frac{4}{2} + \frac{8}{6} \right) = e^{-2} \frac{19}{3} \approx 0.86$$

□

The Poisson distribution model events where the size of the population is very large and the probability of the event to occur is very small. This is why the Poisson distribution is used to model *rare events*, events that have a very slim, but still relevant, probability to occur in a certain span of time. More formally, a rare event can be modeled as such if the following properties hold:

1. There exist a parameter $\alpha > 0$ such that for any short time interval of length Δt, the probability that exactly one event occurs is $\alpha \Delta t \cdot o(\Delta t)$, where $o(\Delta t)$ is a little-o of Δt;
2. The probability of more than one event occurring during Δt is $o(\Delta t)$. In other words, it is much more likely that a single event happens during Δt than multiple events occur;

3. The number of events occurring during the time interval Δt is independent of the number that occur prior to this time interval.

The probability mass function of a Poisson distribution can be adapted in this sense if, instead of the expected value λ, one is given α, the expected number of events occurring in a unitary time interval, and a time interval Δt. The probability of k events to occur in a time slice Δt is then as follows:

$$p_k(\Delta t) = \frac{(\alpha \Delta t)^k}{k!} e^{-\alpha \Delta t}$$

The occurrence of events over time as described is called a **Poisson process** and the parameter α specifies the *rate* of said process.

1.4.4. Hypergeometric random variable

Let N be the size of a population of individuals, each of them having associated either a value of 1 (success) or 0 (failure). Let M be the number of individuals whose value is 1, and therefore $N - M$ is the number of individuals whose number is 0. Let $n \leq N$ be the size of a sample extracted from the population. The random variable X whose values are the number of successes (of 1s) found in a sample of size n is said to distributed as an **hypergeometric random variable** (denoted $X \sim H(n, N, M)$). The pdf of an hypergeometric random variable X of parameters M, N and n is therefore as follows:

$$p(x) = P(X = x) = \begin{cases} \dfrac{\dbinom{M}{x}\dbinom{N-M}{n-x}}{\dbinom{N}{n}} & \text{if} \max(0, n - N + M) \leq x \leq \min(n, M) \\ 0 & \text{otherwise} \end{cases}$$

The binomial $\binom{M}{x}$ is the number of ways it is possible to extract a sample where there are x individuals whose value is 1, while the binomial $\binom{N-M}{n-x}$ is the number of ways it is possible to extract a sample where there are $n - x$ individuals whose value is 0. The binomial $\binom{N}{n}$ is the number of combinations of elements of N of size n (without any requirement on the number of individuals having a particular value).
The constraint $x \leq \min(n, M)$ denotes that the number of observed successes cannot be greater than the size of the entire sample (and, of course, cannot be greater than the size of the entire population).

Exercise 1.4.4.1: A university IT office received 20 service orders for issues with printers, out of which 8 were laser printers and 12 were inkjet printers. A sample of 5 of these service orders were selected to perform a customer satisfaction survey. What is the probability that, out of those 5, 2 were inkjet printers?

Solution: It is possible to model this situation with an hypergeometric random variable. Since the outcome of interest is the one related to inkjet printers, the parameters of said variable X will be 5 for the sample size, 20 for the population size and 12 for the favorable population size. Therefore, $X \sim (5, 20, 12)$. Evaluating the pdf for $X = 2$ gives:

$$p(2) = P(X = 2) = \frac{\dbinom{12}{2}\dbinom{20-12}{5-2}}{\dbinom{20}{5}} = \frac{\dfrac{12!}{2!(12-2)!}\dfrac{8!}{3!(8-3)!}}{\dfrac{20!}{5!(20-5)!}} = \frac{\dfrac{12 \cdot 11 \cdot \cancel{10!}}{2 \cdot \cancel{10!}}\dfrac{8 \cdot 7 \cdot 6 \cdot \cancel{5!}}{6 \cdot \cancel{5!}}}{\dfrac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 \cdot \cancel{15!}}{120 \cdot \cancel{15!}}} =$$

$$\frac{\cancel{12} \cdot 11 \cdot 8 \cdot 7 \cdot 6}{\cancel{12}} \cdot \frac{\cancel{20} \cdot 6}{\cancel{20} \cdot 19 \cdot 18 \cdot 17 \cdot 16} = \frac{22176}{93024} \approx 0.238$$

□

Theorem 1.4.4.1: The expected value and variance of a random variable $X \sim H(n, N, M)$ are as follows:

$$E(X) = n \cdot \frac{M}{N} \qquad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

The hypergeometric distribution is distinguished from the binomial distribution because the trials are not independent, since each time an individual is "inspected" it is removed from the sample, and therefore the subsequent probabilities are influenced by the outcome (since the number of individuals is decreased). By contrast, in the binomial distribution each trial is independent from the others.

Another similarity between the two comes from observing the equations is Theorem 1.4.4.1. The ratio M/N is the proportion of successes in the population, meaning that it's the probability of picking an element of the entire population that has a value 1. This ratio has the same role that the parameter p has in the binomial distribution. Indeed, substituting M/N with p in said equations gives:

$$E(X) = np \qquad V(X) = \left(\frac{N-n}{N-1}\right) \cdot np(1-p)$$

Where the expected value is identical to the one of a binomial distributed random variable (Theorem 1.4.2.1), while the variance differs for a factor $(N-n)/(N-1)$. Since this factor, called **finite population correction factor**, is always less than 1, the variance of an hypergeometric random variable will always be smaller than a binomial random variable where $p = M/N$.

1.4.5. Geometric random variable

Let X be a random variable that represents the number of (failed) attempts necessary to have a Bernoulli random variable with parameter p to assume value 1. The random variable X is said to distributed as a **geometric random variable** (denoted $X \sim G(p)$). The pdf and cdf of a geometric random variable X of parameter p are therefore as follows:

$$p(x) = P(X = x) = \begin{cases} p(1-p)^x & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases} \qquad F(x) = P(X \leq x) = \sum_{k \in \mathbb{N}, k \leq x} p(1-p)^k$$

The factor $(1-p)^x$ represents the probability of obtaining a failure for exactly x times. This factor is then multiplied by p, which is the probability of obtaining a single success.

A geometric distribution $X \sim G(p)$ has a property called **memorylessness**, expressed mathematically as $P(X > x + y \mid X > y) = P(X > x)$ with x and y positive integers. In other words, the number of attempts necessary for an experiment to have a specific result does not depend on the previous ones.

Theorem 1.4.5.1: The expected value and variance of a random variable $X \sim G(p)$ are as follows:

$$E(X) = \frac{1-p}{p} \qquad V(X) = \frac{1-p}{p^2}$$

Proof: This result can be proven by applying known theorems concerning geometric functions:

$$E(X) = p(1-p)^0 \cdot 0 + p(1-p)^1 \cdot 1 + p(1-p)^2 \cdot 2 + \ldots = p(1-p) + 2p(1-p)^2 + 3p(1-p)^3 + \ldots =$$

$$\sum_{i=0}^{\infty} ip(1-p)^i = p \sum_{i=0}^{\infty} i(1-p)^i = p(1-p) \sum_{i=0}^{\infty} i(1-p)^{i-1} = p(1-p) \left[\frac{d}{dp}\left(-\sum_{i=0}^{\infty} (1-p)^i\right)\right] =$$

$$p(1-p) \frac{d}{dp}\left(-\frac{1}{p}\right) = \cancel{p}(1-p)\left(\frac{1}{p^{\cancel{2}}}\right) = \frac{1-p}{p}$$

Then, applying Lemma 1.3.1:

$$V(X) = E(X^2) - (E(X))^2 = \frac{(2-p)(1-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{2-2p-p+p^2}{p^2} - \frac{1+p^2-2p}{p^2} = \frac{2-3p+\cancel{p^2}-1-\cancel{p^2}+2p}{p^2} = \frac{1-p}{p^2}$$

□

1.4.6. Negative binomial distribution

Let X be a random variable that represents the number of (failed) attempts necessary to have a Bernoulli random variable with parameter p to assume value 1 for r times. The random variable X is said to distributed as a **negative binomial random variable** (denoted $X \sim NB(r,p)$). The pdf of a negative binomial random variable X of parameters r and p is therefore as follows:

$$p(x) = P(X = x) = \begin{cases} \binom{x+r-1}{r-1} p^r (1-p)^x & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

The factor $(1-p)^x$ represents the probability of obtaining a failure for exactly x times. The factor p^r represents the probability of obtaining a success r times. The factor $\binom{x+r-1}{r-1}$ represents the number of ways that $r-1$ successes out of $x+r-1$ attempts can be arranged.

Of course, if r is set to 1 said random variable reduces itself to a geometric random variable:

$$\binom{x+1-1}{1-1} p^1 (1-p)^x = \binom{x}{0} p(1-p)^x = \left(\frac{x!}{0!(x-0)!}\right) p(1-p)^x = \left(\frac{\cancel{x!}}{\cancel{x!}}\right) p(1-p)^x = p(1-p)^x$$