Probability and Statistics

Omid Safarzadeh

December 20, 2021

Table of contents

*__Acknowledgement__: This slide is prepared based on Casella and Berger, 2002

Sample space

Sample space: The set of all possible outcomes of a particular experiment is called the *sample space* for the experiment, which generally denoted by Ω.

Example 1.1

In tossing a fair coin, we get either "heads" or "tails" , so the sample space is:

$$\Omega = \{H, T\}.$$

what is the sample space of a fair dice? Whats is the sample space of a credit risk problem ? Whats is the sample space of a Classification problem, dogs vs cats ?

Event

Event: An event is any collection of possible outcomes of an experiment, which is, any subset of Ω (including Ω itself).

Example 1.2

Some possible events in roll of a dice are:

$$E_1 = \{2, 4, 6\} : \text{obtaining an even number },$$

$$E_2 = \{2, 3, 5\} : \text{obtaining a prime number},$$

The sample space in tossing a coin is given by

$$\Omega = \{HH, HT, TH, TT\}.$$

The event that we get at least one tails is, then, given by

$$E = \{HT, TH, TT\}.$$

Set theory Operations

- **Union**: The union of A and B is the set of elements that belong to either A or B or both.
- **Intersection**: The intersection of A and B is the set of elements that belong to both A and B.
- **Complement**: The complement of A is the set of all elements that are not in A.

$$A^c = \{x : x \notin A\}.$$

Axiomatic Foundations

Definition (1.2.1): A collection of subsets Ω is called sigma algebra (or Borel field) , denoted by $\mathbb{B}$, if it satisfies the following three properties.

1. $\varnothing \in \mathbb{B}$ (the empty set is an element of $\mathbb{B}$).
2. If $A \in \mathbb{B}$, then $A^c \in \mathbb{B}$ ($\mathbb{B}$ is closed under complement).
3. If $A_1, A_2, ... \in \mathbb{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathbb{B}$.

Note that $\varnothing$ is in the sigma algebra by (1). We know that $\varnothing^c = \Omega$, so by (2), Ω is in $\mathbb{B}$.

Axiomatic Foundations

if $A_1, A_2, ... \in \mathbb{B}$ then $A_1^c, A_2^c, ... \in \mathbb{B}$, by(2). Now, by (3), $\cup_{i=1}^{\infty} A_i^c \in \mathbb{B}$. However, using De Morgan's Law,

$$\left(\cup_{i=1}^{\infty} A_i^c \right)^c = \cap_{i=1}^{\infty} A_i.$$

Then, by (2), $\cap_{i=1}^{\infty} A_i \in \mathbb{B}$ and $\mathbb{B}$ is closed under countable intersections, as well. Note that sigma algebra is sometimes also denoted as $\sigma - algebra$ and $\mathbb{F}$.

Axiomatic Foundations

Example 2.1

if Ω is finite or countable, then we can define for a given sample space Ω

$$\mathbb{B}=(\text{all subsets of } \Omega \text{ including } \Omega \text{ itself }).$$

- Take, for example, $\Omega=\{1,2,3\}$. Then the sigma algebra defined in the previous bullet point is given by

$$\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\},\varnothing.$$

- if Ω is uncountable, then it becomes difficult to describe $\mathbb{B}$. Intuitively, it is chosen to contain any set of interest.

Axioms of Probability

- **Definition (1.2.2)**: Given a sample space Ω and an associated sigma algebra $\mathbb{B}$, a probability function is a function P with domain $\mathbb{B}$ that satisfies
  1. $P(A) \geq 0$ for all $A \in \mathbb{B}$.
  2. $P(\Omega) = 1$.
  3. If $A_1, A_2, ... \in \mathbb{B}$ are pairwise disjoint, then

$$P\Big(\cup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

- These three points are usually referred to as the Axioms of Probability or the Kolmogorov's Axioms.
- Now, any function $P(.)$ that satisfies the Kolmogorov Axioms is a valid probability function.

Axioms of Probability

Example 2.2

The experiment consists of tossing a fair coin.Therefore, $\Omega = \{H, T\}$. The probability function is

$$P(\{H\}) = P(\{T\}),$$

as the coin is fair.

- observe that $\Omega = \{H\} \cup \{T\}$. Then, from Axiom 2 we must have

$$P(\{H\} \cup \{T\}) = 1.$$

- Since $\{H\}$ and $\{T\}$ are disjoint,

$$P(\{H\} \cup \{T\}) = P(\{H\}) + P(\{T\}).$$

So,

$$P(\{H\} \cup \{T\}) = P(\{H\}) + P(\{T\}) = 1.$$

Axioms of Probability

- Our intuition and the Kolmogorov Axioms together tell us that $P(\{H\} = P(\{T\}) = 1/2$.
- However, any non-negative probabilities that add up to one would have been valid, say $P(\{H\}) = 1/9$ and $P(\{T\}) = 8/9$. The reason we chose equal probabilities is our knowledge that the coin is fair!

The Foundation of Probabilities Functions

- **Theorem (1.2.1)**: If P is a probability function and A is any set in $\mathbb{B}$, then
  1. $P(\varnothing) = 0$ where $\varnothing$ is the empty set.
  2. $P(A) \leq 1$.
  3. $P(A^c) = 1 - P(A)$.
- **Theorem (1.2.2)**: if P is a probability function and A and are any sets in $\mathbb{B}$, then
  1. $P(B \cap A^c) = P(B) - P(A \cap B)$.
  2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
  3. If $A \subset B$, then $P(A) \leq P(B)$.

Conditional Probability and Independence

- **Definition (1.3.1)**: If A and B are events in Ω, and $P(B) > 0$, then the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1}$$

- In words, given that B has occurred, what is the probability that A will occur?
- By definition,

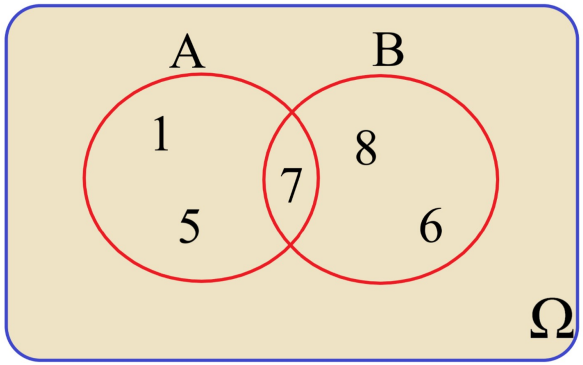
$$P(B|B) = 1,$$

  as B has already occurred.
- If A and B are disjoint sets, then, by (1) $P(A \cap B) = 0$ and

$$P(A|B) = P(B|A) = 0.$$

- In fact, what happens in the conditional probability calculation is that B becomes the sample space.
- It is straightforward to verify that the probability function $P(.|B)$ satisfies Kolmogorov's Axioms, for any B for which $P(B) > 0$.

Conditional Probability



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{7}{7+8+6} = \frac{1}{3}$$

Figure: Conditional Probability.

Bayes Rule

- Observe that since

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

we have

$$P(A|B)P(B) = P(A \cap B) = P(B|A)(PA) \Rightarrow P(A|B) = P(B|A)\frac{P(A)}{P(B)} \quad (2)$$

- Which is known as Bayaes's Rule.

Bayes Theorem

Theorem (1.3.1): Let $A_1, A_2, ...$ be a partition of the sample, and let B be any set. Then, for each $i = 1, 2, ...$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)(PA_j)}.$$

This is actually not much more different that (2) since

$$\sum_{j=1}^{\infty} P(B|A_j)P(A_j) = \sum_{j=1}^{\infty} P(A_j \cap B) = P(B)$$

given that $A_1, A_2, ...$ is a partition of the sample space

Independence

Definition (1.3.2): Two events, A and B, are statistically independent if

$$P(A \cap B) = P(A)P(B).$$

Theorem (1.3.2): If A and B are independent events, then the following pairs are also independent:

1. A and B^c
2. A^c and B
3. A^c and B^c

Random Variables

Definition (1.4.1): A random variable is a function from a sample space Ω into the real numbers.

Experiment	**Random Variable**
Toss two dice	X = sum of the numbers
Toss a coin 10 times	X = number of heads in 10 tosses

Random Variables

- Suppose the sample space is $\Omega = \{\omega_1, ..., \omega_n\}$ and the original probability function is P. Define the new random variable

$$X : \Omega \to \chi, \qquad \chi = \{x_1, ..., x_m\}$$

- Define the new probability function for X ad P_X where

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\}).$$

- P_X is an induced probability function, as it is defined in terms of the original probability function, P.
- If χ is uncountable, the induced probability function is defined in a slightly different way. Namely, for any set $A \subset X$,

$$P_X(X \in A) = P(\{\omega \in \Omega : X(_\omega) \in A\}).$$

Random Variables

- In both cases, it is possible to show that the induced probability function satisfies the Kolmogorov Axioms.
- Note that the general convention in the literature is to assign capital letters to random variables and lower case letters to the particular value they take. Hence, for example, the number of people who answer "yes" in the survey could be X while a particular value, say 12, would be x.

Probability Function

- Suppose the sample space is $\Omega = \{\omega_1, ..., \omega_n\}$and the original probability function is P. Define the new random variable

$$X : \Omega \to X, \qquad X = \{x_1, ..., x_m\}$$

- Define the new probability function for X ad P_X where

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\}).$$

- P_X is an induced probability function, as it is defined in terms of the original probability function, P.
- If X is uncountable, the induced probability function is defined in a slightly different way. Namely, for any set $A \subset X$,

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

Probability Function

Example 2.3

Consider again the experiment of tossing a fair coin three times. Define the random variable X to be the number of heads obtained in the three tosses. A complete enumeration of the value of X for each point in the sample space is:

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

Probability Function

The range for the random variable X is $X = \{0, 1, 2, 3\}$. Assuming that all eight points ω have probability $\frac{1}{8}$, by simply counting in the above display we see that the induced probability function on X is given by

x	0	1	2	3
P_X(X=x)	1/8	3/8	3/8	1/8

For example,

$$P_X(X = 1) = P(HTT, THT, TTH) = \frac{3}{8}$$

Distribution Functions

- All random variables are associated with a distribution function. This distribution function includes all information about the randomness of the variable.
- **Definition (1.5.1)**: The cumulative distribution function or cdf of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \leq x), \qquad \text{for all } x.$$

- When we write $P_X(X \leq x)$, we mean the probability that the random variable X takes a value equal to or smaller than x. The subscript X in $P_X(.)$ denotes that this probability is obtained with respect to the probability distribution of X.

Distribution Functions

- In this particular case, this is too clear and so a bit redundant. However, if we consider $Y = f(X)$, then the notation provides clarification because we will have to deal with $P_X(Y \leq y)$.
- Note that the cdf is also generally denoted as simply the 'distribution."

Distribution Functions

Example 2.4

Consider the experiment of tossing three fair coins, and let X =number of heads observed. The cdf of X is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0, \\ 1/8 & \text{if } 0 \leq x < 1, \\ 1/2 & \text{if } 1 \leq x < 2, \\ 7/8 & \text{if } 2 \leq x < 3, \\ 1 & \text{if } 3 \leq x < \infty. \end{cases}$$

- Note that, $F_X(x)$ is defined for all possible values of $x \in \chi$. Hence,

$$P_X(x \leq 2.5) = P(X = 0, 1 or 2) = 7/8.$$

Distribution Functions

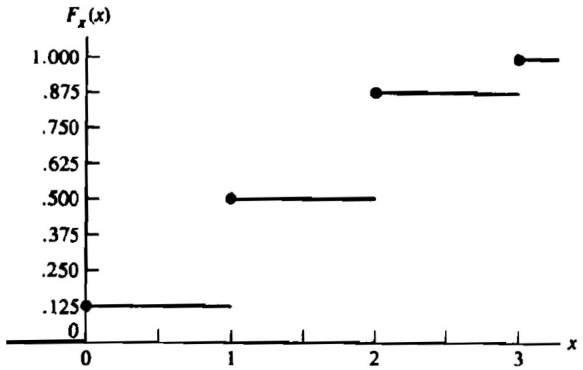


Figure: 1.5.1. from Casella and Berger (2002, p.30). Cdf of example 1.5.1

Distribution Functions

Theorem (1.5.1): The function $F_X(x)$ is a cdf id and only if the following three conditions hold:

1. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to+\infty} F_X(x) = 1$.
2. $F_X(x)$ is a non-decreasing function of x.
3. $F_X(x)$ is right-continuous; that is, for every number x_0, $\lim_{x\downarrow x_0} F_x(x) = F(x_0)$.

- We can also have a continuous cdf.
- **Definition (1.5.2)**: A random variable X is continuous(discrete) if $F_X(x)$ is a continuous(step) function of x.

Distribution Functions

Example 2.5

An example of a continuous cdf is the function

$$F_X(x) = \frac{1}{1+e^{-x}}$$

observe that

$$\lim_{x\to-\infty} F_X(x) = 0 \quad \text{since} \quad \lim_{x\to-\infty} e^{-x} = \infty,$$

$$\lim_{x\to\infty} F_X(x) = 1 \quad \text{since} \quad \lim_{x\to\infty} e^{-x} = 0,$$

$$\frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2} > 0,$$

where the final line proves that $F_X(x)$ is non-decreasing in x. Finally, by definition, $F_X(x)$ is right-continuous as it is continuous in the first place.

- This cdf is a special case of the logistic distribution.

Distribution Functions

- To be a cdf, a function has to possess some key properties.
- **Theorem (1.5.1)**: The function $F_X(x)$ is a cdf id and only if the following three conditions hold:
  1. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to+\infty} F_X(x) = 1$.
  2. $F_X(x)$ is a non-decreasing function of x.
  3. $F_X(x)$ is right-continuous; that is, for every number x_0, $\lim_{x\downarrow x_0} F_x(x) = F(x_0)$.

Distribution Functions

- **Definition (1.5.3)**: The random variables X and Y are identically disturbed if, for every set $A \in \mathbb{B}^1$, $P(X \in A) = P(Y \in A)$.
- **Theorem (1.5.2)**: The following two statements are equivalent:
  1. The random variables X and Y are identically distributed.
  2. $F_X(x) = F_Y(x)$ for every x.
- This is the most important result we have covered so far.

Density and Mass

- **Definition (1.6.1)**:The probability mass functionof a discrete random variable is given by

$$f - X(x) = P(X = x) \quad \text{for all x.}$$

- **Notational convention**: for a given cdf F_X, the corresponding pdf is usually denoted by f_X, the corresponding lower-case letter.
- Similar to cdf being simply called the "distribution", the pdf is sometimes simply called the density.

Density and Mass

Definition (1.6.2): The probability density function or pdf, $f_X(x)$, of a continuous random variable X is the function that satisfies

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

Using the Fundamental Theorem of Calculus, if f_X is continuous then

$$\frac{d}{dz}F_X(z)|_{z=x} = f_X(x).$$

Let's consider some examples for both types of variables.

Example 2.6

For the logistic distribution considered before, we have

$$F_X(x) = \frac{1}{1+e^{-x}}$$

and, hence,

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$

Then, for continuous random variables in general,

$$P(a < X < b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(x)dx - \int_{-\infty}^{a} f_X(x)dx$$

$$= \int_{a}^{b} f_X(x)dx.$$

Density and Mass

- **Theorem (1.6.1)**: A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only of
  1. $f_X(x) \geq 0$ for all x.
  2. $\sum_x f_X(x) = 1$ (discrete) or $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (continuous).
- Some technical details:
  1. From a purely mathematical point of view, any non-negative function with a finite positive integral can be turned into a pdf or pmf. Take, for example, if

$$h(x) = \begin{cases} \geq 0 & \text{for} \quad x \in A \\ 0 & \text{elsewhere} \end{cases}$$

and

$$\int_{x \in A} h(x)dx = \mathrm{K} < \infty, \quad \textit{where } \mathrm{K} > 0,$$

then $f_X(x) = h(x)/K$ is a pdf of a random variable X taking values in A.
  2. In some cases, although $F_X(x)$ exists, $f_X(x)$ may not exist because $F_X(x)$ can be continuous but not differentiable. Therefore, sometimes statistical analysis would be based on $F_X(x)$ and not $f_X(x)$. We will not consider such cases here.

References

Casella, G., & Berger, R. (2002). *Statistical inference*. Cengage Learning. https://books.google.fr/books?id=FAUVEAAAQBAJ