Probability Theory

SynFerLo

July. 21, 2021

1 Terminology

1.1 Random variable

A Random variable is a function from a set of outcomes to the real line, attaching numbers to outcomes.

1.2 IID

Independence: The RVs $(X_1, X_2, ...Xn)$ is said to be independent if the occurance of any one, X_i, does not influence and is not influenced by the occurence of any other RV in the set, $X_j, j \neq i,$.

Identical Distribution: The density of RVs are identical

$$f(x_1; \theta) = f(x_2; \theta) = \cdots = f(x_n; \theta)$$

1.3 Event Space

Outcomes Set: S, includes all possible distinct outcomes. For example, the outcome set of casting a dice can be written as $S = \{1, 2, 3, 4, 5, 6\}$

Event space: $\mathfrak{F}$, is a set whose elements are the events of interest as well as the related events, those we get by combining the events of interest using set theoretic operations, e.g., $\overline{A}, \overline{B}, A \cap B, A \cup B, (\overline{A}_1 \cap \overline{A}_2)$, etc. $\mathfrak{F}$ is a subset of $\boldsymbol{S}$.

$$\text{for } A \in \mathfrak{F},\ \text{B } \in \mathfrak{F} \text{ and} A \cap B = \emptyset,\ \text{then } \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

For modeling purpose, we need to broaden the event to include not just elementary outcomes but also combinations of them.

Two extreme event spaces:

(1)$\mathfrak{F}_0 = \{\boldsymbol{S}, \emptyset\}$: the **trivial event space.**

(2)$\mathcal{P}(S) = \{A : A \subset \boldsymbol{S}\}$, i.e., the **power set**: the set of all subsets of $\boldsymbol{S}$.

Notice: we cannot always use the power set of $\boldsymbol{S}$ as the appropriate event space, because (1) if $\boldsymbol{S}$ is countable and has N elements, $\mathcal{P}(\boldsymbol{S})$ has 2^N elements, it contain too many elements; (2) when the outcomes set is uncountable, such as

$$\boldsymbol{S} = \{x : 0 \leq x \leq 1, x \in \mathbb{R}\}$$

the power set includes subsets which cannot be considered as events and thus cannot be assigned probabilities.

To circumvent these difficulties, we use a field or σ-field, which ensures if A and B are events then any other events which arise when we combine these with set theoretic operations are also elements of the same event space.

Field: A collection $\mathfrak{F}$ of subsets of $\boldsymbol{S}$, is said to be a field if it satisfies the conditions:

1. $\boldsymbol{S} \in \mathfrak{F}$
2. if $A \in \mathfrak{F}$ then $\overline{A}$ also belong to $\mathfrak{F}$
3. if $A, B \in \mathfrak{F}$, then $(A \cup B) \in \mathfrak{F}$.

This means that $\mathfrak{F}$ is non-empty, closed under complementation, finite unions and finite intersections.

Event: is a subset of the outcomes set $\boldsymbol{S}$, i.e., if $A \subset \boldsymbol{S}, A$ is an event.

Special Events:

1. Sure event: whatever the outcome, $\boldsymbol{S}$ occurs. $\boldsymbol{S}$ is always a subset of itself, i.e., $S \subset S$.

2. Impossible event: $\emptyset$

3. Any two events A and B are said to be **mutually exclusive** if

$$A \cap B = \emptyset$$

4. The events $A_1, A_2, ..., A_m$ is said to constitute a **partition** of $\boldsymbol{S}$ if they are:

(1) mutually exclusive, i.e., $A_i \cap A_j = \emptyset, \forall j \neq j, i, j = 1, 2, ..., m$ and

(2) exhaustive, i.e., $\bigcup_{i=1}^{m} A_i = \boldsymbol{S}$.

1.4 σ-field

A collection $\mathfrak{F}$ of subsets of $\boldsymbol{S}$, is said to be a σ-field if it satisfies the conditions:

1. $\boldsymbol{S} \in \mathfrak{F}$
2. if $A \in \mathfrak{F}$, then $\overline{A} \in \mathfrak{F}$
3. if $A_i \in \mathfrak{F}$ for $i = 1, 2, ...n, ...$ the set $\cup_{i=1}^{\infty} A_i \in \mathfrak{F}$.
4. from 2 and 3, we can deduce that

$$\cap_{i=1}^{\infty} A_i \in \mathfrak{F}, \text{ since } \overline{\cup_{i=1}^{\infty} A_i} = \cap_{i=1}^{\infty} \overline{A}_i$$

Notice: a σ-field is non-empty and closed under countable unions and intersections.

1.5 Borel σ-field

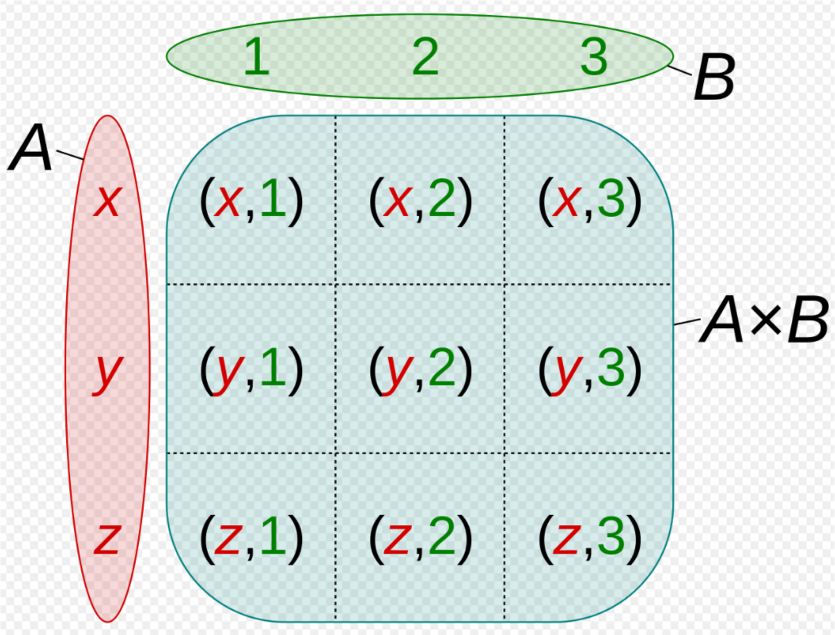
Borel field, or Borel σ-field, is the most important σ -field defined on the real line $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$.

1.6 Cartesian Product

Define the notion of the Cartesian product of two sets by $A \times B$

$$A \times B = \{(x, y) : x \in A, y \in B\}$$

It is a set of all ordered pairs (x, y) where $x \in A$ and $y \in B$. On example is the Cartesian coordinates of the plane, where X is the set of points on the x-axis, Y is the set of points on the y-axis, and $X \times Y$ is the xy-plane.



1.7 Probability space $(S, \mathfrak{F}, \mathbb{P}(.))$

A **probability space** is a collection of

1. Outcomes set S

2. Event space $\mathfrak{F}$, where $\mathfrak{F}$ is a σ-field of subsets of S.

3. Probability set function $\mathbb{P}(.)$

Notice: the probability function $\mathbb{P}(.)$ statisfies axioms [1]-[3]:

[1]: $\mathbb{P}(S) = 1$, for any outcomes set S

[2]: $\mathbb{P}(A) \geq 0$, for any event $A \in \mathbb{F}$

[3]: Countable additivity. For a countable sequence of mutually exclusive events, i.e., $A_i \in \mathfrak{F}, i = 1, 2, ..., n, ...$ such that $A_i \cap A_j = \emptyset \quad \forall i \neq j, i, j = 1, 2, ..., n, ...$, then

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

1.8 Sampling Space

A sequence of n trials, denoted by $\mathcal{G}_n = \{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n\}$ where $\mathcal{A}_i$, represents the ith trial of the experiment, associated with the product probability space $(\boldsymbol{S}_{(n)}, \mathfrak{F}_{(n)}, \mathbb{P}_{(n)})$, is said to be a sampling space. We use $\mathcal{G}_n$ for sampling space.

1.9 Random Trials

A sequence of trials $\mathcal{G}_n^{IID} := \{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n\}$ which is both **independent** and **identical distributed**, i.e.,

$$\mathbb{P}_{(n)}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap ... \cap \mathcal{A}_k) = \mathbb{P}(\mathcal{A}_1) \cdot \mathbb{P}(\mathcal{A}_2) \cdots \mathbb{P}(\mathcal{A}_k) \text{ for each } k = 2, 3, ..., n$$

is referred to as a sequence of Random trials.

1.10 Statistical Space

It is the combination of <u>a</u> simple product probability space, and a sequence of Random trials,

$$[(\boldsymbol{S}, \mathfrak{F}, \mathbb{P}(.))^n, G_n^{IID}]$$

The more general formulation of a **statistical space:**

$$[(\boldsymbol{S}_{(n)}, \mathfrak{F}_{(n)}, \mathbb{P}_{(n)}), \mathcal{G}_n^{IID}]$$

where each trial $\mathcal{A}_i$ is associated with a different probability space $\{(S_i, \mathfrak{F}_i, \mathbb{P}_i(.))\}$

2 Set operation

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

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

Difference:

$$A - B = A \cap \overline{B} := \{x : x \in A \cap x \notin B\}$$

Symmetric difference:

$$A \Delta B = (A \cap \overline{B}) \cup (\overline{A} \cap B) := \{x : x \in A \cup x \in B \cap x \notin (A \cap B)\}$$

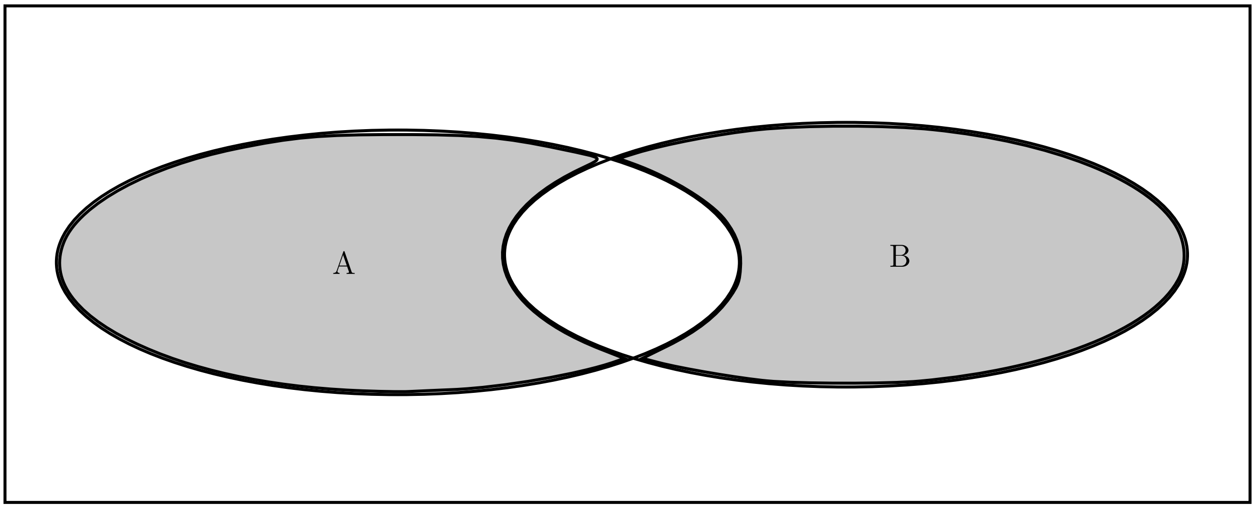


Figure 1: Symmetric difference

3 Useful Probability results

Them 1:

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

Them 2: Continuity property of the probability set function

For $\{A_n\}_{n=1}^{\infty} \in \mathfrak{F}$, if $\lim_{n\to\infty} A_n = A \in \mathfrak{F}$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$

Non-decreasing sequence: A sequence of events $\{A_n\}_{n=1}^{\infty}$ is called non-decreasing if

$$A_1 \subset A_2 \subset \cdots \subset A_n \subset ...$$

It has a property:

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$$

non-increasing sequence:

$$A_1 \supset A_2 \supset \cdots \supset A_n...$$

$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n$$

Thm 3:

$$\mathbb{P}(\bigcap_{k=1}^{n} A_k) \geq 1 - \sum_{k=1}^{n} \mathbb{P}(\overline{A}_k)$$

where $A_k \in \mathfrak{F}, k = 1, 2, ..., n$

Thm: 4:

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

Also, we know

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

Hence,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}, \text{ for } \mathbb{P}(B) > 0$$

This is called Bayes' formula.

The conditioning probability can be used to determine whether the occurance of B alters the probability of occurance of A. If not, $\mathbb{P}(A|B) = \mathbb{P}(A)$, we say, they are independent.

Hence we have,

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

Notice: independent and mutually exlusive are not the same. The latter one does not involve probability. If A and B are mutually exlusive,

$$\mathbb{P}(A \cap B) = 0, \text{ since } A \cap B = \emptyset$$

If A and B are independent and $\mathbb{P}(A) > 0 \mathbb{P}(B) > 0$,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) > 0$$

4 cdf and pdf

Density function:

It is used to assign probability to event.

$$f_x(x) := \mathbb{P}(X = x), \quad \forall x \in \mathbb{R}_X$$

For $x \notin \mathbb{R}_X, \quad X^{-1}(x) = \emptyset$ and thus $f_x(x) = 0, \quad \forall x \notin \mathbb{R}_X$.

Note: $(X = x)$ is a shorthand notation for $A_x = \{s : X(s) = x\}$.

4.0.1 For a discrete RV

$$F_X(x_k) = \mathbb{P}(\{s : X(s) \leq x_k\}) = \sum_{i=1}^{k} f_x(x_i), \quad fork = 1, 2, ..., n$$

Properties:

$$f_x(x) \geq 0, \quad \forall x \in \mathbb{R}_X$$

$$\sum_{x_i \in \mathbb{R}_X} f_x(x_i) = 1$$

$$F_X(b) - F_X(a) = \sum_{a < x_i \leq b} f_x(x_i), \quad a < b, \quad a, b \in \mathbb{R}$$

Example:

Bernoulli RV

$$f_x(1) = \theta \text{ and } f_x(0) = (1-\theta)$$

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \theta, & 0 \le x < 1, \\ 1, & 1 \le x \end{cases}$$

4.1 Probability Model

A probability model is a collection of density functions

$$\Phi = \{f_x(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, x \in \mathbb{R}_X\}$$

5 Parameters and Moments

The most efficient way to deal with the unknown parameters θ is to relate them to the moments of the distribution.

5.1 Moments

The **moments** of a distribution are defined in terms of the mathematical expectation of certain functions of the random variable X, generically denoted by $h(X)$

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x; \boldsymbol{\theta}) dx$$

where $E[h(X)]$ is some function of $\boldsymbol{\theta}$

$$E[h(X)] = g(\boldsymbol{\theta})$$

By choosing specific forms of the function $h(X)$, such as:

$$h(X) = X^r, h(X) = |X|^r, r = 1, 2, ..., h(X) = e^{tx}, h(X) = e^{itx},$$

we obtain several functions of the form $g(\boldsymbol{\theta})$ which involve what we call **moments** of $f(x; \boldsymbol{\theta})$.

Note: the best way to handle probability models (postulate a statistical model, estimate $\boldsymbol{\theta}$, test hypotheses about these parameters, $\boldsymbol{\theta}$, etc.) is often via the moments of the postulate probability distribution.

5.1.1 Higher Raw moments

A direct generalization of the **mean** yields the so-called raw moments. For $h(X) := X^r, r = 2, 3, 4...$ the raw moments:

$$\mu'_r(\boldsymbol{\theta}) := E(X^r) = \int_{-\infty}^{\infty} x^r f_x(x; \boldsymbol{\theta}) dx, \quad r = 1, 2, 3, ...$$

Note:

The second raw moment is often useful in deriving the variance. Recall

$$Var(X) = E(X^2) - [E(X)]^2,$$

where $E(X^2)$ is the second raw moment.

Lower moments lemma

If $\mu'_k := E(X^k)$ exists for some positive integer k, then all the raw moments of order less than k also exists, i.e.,

$$E(X^i) < \infty, \quad \forall i = 1, 2, ..., k - 1$$

5.1.2 Moment generating function (MGF)

A particular convenient way to compute the **raw moment** is by way of the MGF using the integral with $h(X) = e^{tx}$

$$m_X(t) := E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx, \text{ for } t \in (-h, h), h > 0$$

Recall, $e^{\theta} = \sum_{i=0}^{\infty} \frac{\theta^i}{i!}$

Moments and MGF:

The rth raw moments is the rth derivative of MGF with respect to t when $t = 0$.

$$\mu_r' = E(X^r) = \left.\frac{d^r m_X(t)}{dt^r}\right|_{t=0} := m_X^{(r)}(0), r = 1, 2, 3, ...$$

Hence, mean is the raw moment, $\mu_1' = \left.\frac{dm_X(t)}{dt}\right|_{t=0} = m_X(0)$

Consider **Poisson** distribution as an example.

Given the density of Poisson distribution:

$$f_x(x) = \frac{e^{-\theta}\theta^x}{x!}$$

The MGF can be written as

$$\begin{aligned} m_X(t) = E(e^{tX}) &= \sum_{x=0}^{\infty} e^{tx} \left(\frac{e^{-\theta}\theta^x}{x!} \right) \\ &= e^{-\theta} \sum_{x=0}^{\infty} \frac{e^{tx}\theta^x}{x!} \\ &= e^{-\theta} \sum_{x=0}^{\infty} \frac{(e^t\theta)^x}{x!} \end{aligned}$$

Recall

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

Hence,

$$m_X(t) = e^{-\theta} e^{e^t\theta} = e^{\theta(e^t-1)}$$

Recall the first moment is the first derivative of the MGF with respect to t. Hence, we can derive the expectation of Poisson distribution

$$\begin{aligned} E(X) &= \left. \frac{dm_X(t)}{dt} \right|_{t=0} \\ &= \left. e^{\theta(e^t-1)}\theta e^t \right|_{t=0} \\ &= \left. \theta e^{\theta(e^t-1)+t} \right|_{t=0} \\ &= \theta \end{aligned}$$

For the variance, $Var(X) = E(X^2) - (E(X))^2$.

$$\begin{aligned}
E(X^2) &= \left.\frac{d^2 m_X(t)}{dt^2}\right|_{t=0} \\
&= \left.\frac{d}{dt}\theta e^{\theta(e^t-1)+t}\right|_{t=0} \\
&= \left.\theta e^{\theta(e^t-1)+t}(\theta e^t + 1)\right|_{t=0} \\
&= \theta(\theta + 1) \\
&= \theta^2 + \theta
\end{aligned}$$

Hence,

$$Var(X) = \theta^2 + \theta - \theta^2 = \theta$$

Uniqueness lemma:

When a MGF exists (it does not always exists), it is unique in the sense that two RVs X and Y that have the same MGF must have the same distribution, and conversely.

Probability integral transformation lemma:

For any continuous RV X, with cdf $F_X(x)$, the RV defined by $Y = F_X(x)$ has a uniform distribution over the range (0,1), i.e.,

$$Y = F_X(x) \sim \boldsymbol{U}(0, 1)$$

Proof:

Derive the MGF for uniform distribution first. Recall the pdf of uniform distribution,

$$f(x) = \frac{1}{b-a}$$

Then, we can write the MGF as the following

$$\begin{aligned}
m_X(t) = E(e^{tx}) &= \int_a^b e^{tx} \frac{1}{b-a} dx \\
&= \frac{1}{b-a} \int_a^b e^{tx} dx \\
&= \frac{1}{b-a} \left(\left. \frac{e^{tx}}{t} \right|_a^b \right) \\
&= \frac{1}{b-a} \left(\frac{e^{bt} - e^{at}}{t} \right) \\
&= \frac{e^t - 1}{t}, \text{ if } (a,b) \text{ is } (0,1)
\end{aligned}$$

Now let's derive the MGF for $Y = F_X(x)$

$$\begin{aligned}
m_Y(t) = E(e^{tY}) &= \int_{-\infty}^{\infty} e^{tF(x)} f(x) dx \\
&= \int_{-\infty}^{\infty} e^{tF(x)} dF \\
&= \left. \frac{e^{tF(x)}}{t} \right|_{-\infty}^{\infty} \\
&= \frac{e^t - 1}{t}
\end{aligned}$$

Note: $F(\infty) = 1$, $F(-\infty) = 0$.

5.1.3 Cumulants

A cumulant generating function is the logarithm of a MGF.

$$\psi_X(t) = \ln(m_X(t)) = \sum_{r=1}^{\infty} \boldsymbol{\kappa}_r \frac{t^r}{r!}, \text{ for } t \in (-h, h), h > 0$$

where $\boldsymbol{\kappa}_r, r = 1, 2, 3, ...$ are referred to as cumulants(or semi-invariants).

$$\boldsymbol{\kappa}_1 = E(X) = \left.\frac{d\psi_X(t)}{dt}\right|_{t=0}$$

$$\boldsymbol{\kappa}_2 = Var(X) = \left.\frac{d^2\psi_X(t)}{dt^2}\right|_{t=0}$$

The relation between the first few cumulants and the raw moments are as follows:

$$\begin{aligned} \boldsymbol{\kappa}_1 &= \mu_1', \\ \boldsymbol{\kappa}_2 &= \mu_2' - (\mu_1')^2, \\ \boldsymbol{\kappa}_3 &= \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3, \end{aligned}$$

Properties: The cumulants are often **preferable** to the moments for several reasons including:

1. In the case of the Normal distribution: $\boldsymbol{\kappa}_r = 0, r = 3, 4, ...$

2. The rth cumulant is rth-order homogeneous: $\boldsymbol{\kappa}_r(\alpha X) = \alpha^r \boldsymbol{\kappa}_r(X), r = 1, 2, ..$

3. The rth cumulant is a function of the moments of order up to r.

4. For **independent** RVs, the cumulant of the sum is the sum of the cumulants:

$$\boldsymbol{\kappa}_r(\sum_{k=1}^{n} X_k) = \sum_{k=1}^{n} \boldsymbol{\kappa}_r(X_k), r = 1, 2, ...$$

Recall the definition of homogeneous function:

$$f(rx, ry) = r^k f(x, y)$$

is called function $f(.)$ is homogeneous of degree k.

5.1.4 Characteristic function(CF)

The existence of the MGF depends on $m_X(t)$ being finite on the interval $(-h, h)$. In such a case, all the moments $E(X^r)$ are **finite.** If $E(X^r)$ is **not finite** for some r, $m_X(t)$ is not finite on the interval. To solve this we define the **characteristic function** (Cramer, 1946):

$$\phi_X(t) := E(e^{itX}) = \int_{-\infty}^{\infty} e^{itX} f(x)dx = m_X(it), \text{ for } i = \sqrt{-1},$$

CF always exist since for all t, $\phi_X(t)$ is bounded:

$$|\phi_X(t)| \leq E(\left|e^{itX}\right|)$$

Hence, in many cases, we can derive the CF using MGF. The CF is related to the moments (when they exist) via

$$\phi_X(t) = \sum_{r=0}^{\infty} \frac{(it)^r}{r!} \mu'_r, \text{ for } t \in (-h, h), h > 0$$

5.2 Problem of moments

The existence and the uniqueness of the moment. In general, the answer is no. However, under certain conditions, the answer is yes.

5.2.1 Lemma 1 (existence)

A sufficient (not certainly necessary) condition for the existence of moments is that the **support** of the RV X is a **bounded interval**, i.e., $\mathbb{R}_X := [a, b]$, where $-\infty < a < b < \infty$. In this case, **all moments exist:**

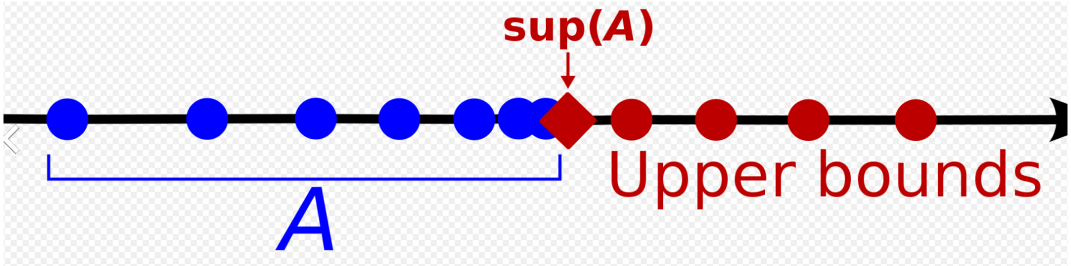
$$\mu'_k = \int_a^b x^r f(x)dx < \infty, \forall k = 1, 2, ...$$

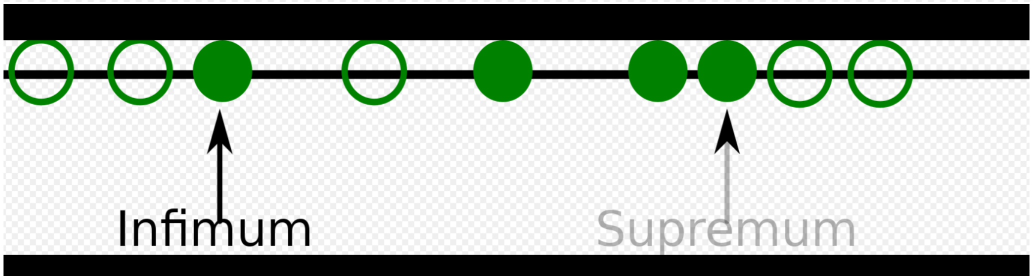
5.2.2 Lemma 2 (uniqueness)

The moments $\{\mu'_k, k = 1, 2, ...\}$(assuming they exist) determine the distribution function **uniquely** if:

$$\lim_{n\to\infty}\left(\sup\left[(2n)^{-1}(\mu'_{2n})^{\frac{1}{2n}}\right]\right) < \infty$$

Note: sup for supremum (the least upper bound), and inf for infimum (the largest lower bound).





5.3 Higher Central Moments

The notion of the **variance** can be extended to define the **central moments** using the sequence of functions $h(X) := (X - E(X))^r, r = 3, 4, ...$

$$\mu_r(\boldsymbol{\theta}) := \int_{-\infty}^{\infty}(x-\mu)^r f(x;\boldsymbol{\theta})dx, r = 2, 3, ...$$

We normally derive the central moments by using the relationship with the raw moments and the cumulants.

$$\mu_2 = \mu_2' - (\mu_1')^2 \qquad \kappa_2 = \mu_2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 \qquad \kappa_3 = \mu_3$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \qquad \kappa_4 = \mu_4 - 3\mu_2^2$$

5.3.1 Symmetry

A RV X with density $f(x)$ is said to be symmetric about a point a if

$$f(a-x) = f(a+x), \forall x \in \mathbb{R}_X$$

$$F_X(a-x) + F_X(a+x) = 1, \forall x \in \mathbb{R}_X$$

5.3.2 Skewness

It gives us some ideas about the shape/possible asymmetry of a density function around the mean is the **skewness** coefficient .

$$\textbf{Skewness:}\alpha_3(X) = \frac{\mu_3}{(\sqrt{\mu_2})^3}$$

Note, $\sqrt{\mu_2} = (Var(X))^{\frac{1}{2}} = SD$.

If the distribution is symmetric around the mean, $\alpha_3 = 3$, **the converse does not hold!**

5.3.3 Kurtosis

Kurtosis measures the peakedness of a density in relation to the shape of the tail.

$$\textbf{Kurtosis:}\alpha_4(X) = \frac{\mu_4}{(\mu_2)^2}$$

In the case of Normal distribution $\alpha_4 = 3$, and it is referred to as a mesokurtic distribution.

If a distribution is flatter than this, we call it platykurtic. And if it is steeper than this, we call it leptokurtic

5.3.4 Quantile function

$F(x_p) = p$ The value p is know as the pth **percentile**, and the value x_p the corresponding quantile.

$$x_{\frac{1}{4}} = F^{-}(0.25), \quad x_{\frac{3}{4}} = F^{-}(0.75)$$

where $F^{-}(.)$ is known as the **quantile function**.

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

For example, In the case of standard Normal distribution,

$$x_{\frac{1}{4}} = -0.6745, \quad x_{\frac{3}{4}} = 0.6745$$

For an arbitrary Normal distribution,

$$x_{\frac{1}{4}} = \mu - 0.6745\sigma, \quad x_{\frac{3}{4}} = \mu + 0.6745\sigma$$

5.3.5 Interquartile range

$IQR(X) := (x_{\frac{3}{4}} - x_{\frac{1}{4}})$

5.4 Mean

For $h(X) := X$, where X takes values in $\mathbb{R}_X$, we can write the **mean** of the distribution

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_x(x; \boldsymbol{\theta}) dx \quad \text{for continuous variables,}$$

$$E(X) = \sum_{x_i \in \mathbb{R}_X} x_i \cdot f_x(x_i; \boldsymbol{\theta}) \quad \text{for discrete random variables.}$$

5.4.1 Properties

1. $E(c) = c$
2. $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$

5.5 Variance

For $h(X) := [X - E(X)]^2$ the integral yields the variance:

$$Var(X) := E[(X - E(X))^2] = \int_{-\infty}^{\infty} [x - \mu]^2 f_x(x; \boldsymbol{\theta}) dx$$

In the case of **discrete** RVs the integral is replaced by the **summation**.

$$Var(X) := \sum_{i=1}^{n} (X_i - \mu)^2 f_x(x; \boldsymbol{\theta})$$

5.5.1 Properties

1. $Var(c) = 0$
2. $Var(aX_1 + bX_2) = a^2 Var(X_1) + b^2 Var(X_2) + 2abCov(X_1, X_2)$

Bienayme's lemma If $X_1, X_2, ..., X_n$ are Independent Distributed RVs:

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i)$$

5.6 Standard Deviation

Std. Deviation is the square root of the variance.

$$SD(X) = [Var(X)]^{\frac{1}{2}} = \sigma$$

Note: When we need to render a RV free of its units of measurement, we divide it by its SD, i.e., we define the standardized variable:

$$X^* := \frac{X}{[Var(X)]^{\frac{1}{2}}}, \text{ where } Var(X^*) = 1$$

5.7 Chebyshev's inequality

Let X be a RV with bounded variance:

$$\mathbb{P}(|X - E(X)| > \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}, \text{ for any } \varepsilon > 0$$

5.8 General Chebyshev's Inequality

Let $X(.) : S \to \mathbb{R}_X := (0, \infty)$ be a positive RV, and let $g(.) : (0, \infty) \to (0, \infty)$ be a positive and increasing function. Then for each $\varepsilon > 0$:

$$\mathbb{P}(g(X) \geq \varepsilon) \leq \frac{E(g(X))}{g(\varepsilon)}$$

5.9 Markov's inequality

Let X be a RV such that $E(|X|^p) < \infty$ for $p > 0$:

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{E(|X|^p)}{\varepsilon^p}$$

5.10 Jensen's inequality

Let $\varphi(.) : \mathbb{R} \to \mathbb{R}$ be a convex function, i.e.,:

$$\lambda\varphi(x) + (1-\lambda)\varphi(y) \geq \varphi(\lambda x + (1-\lambda)y), \quad \lambda \in (0,1), \quad , x, y \in \mathbb{R}.$$

Assuming that $E(|X|) < \infty$, then

$$\varphi(E(X)) \leq E(\varphi(X))$$

5.11 Minkowski's inequality

Let X and Y be RVs such that $E(|X|^p) < \infty$, and $E(|Y|^p) < \infty$, where $1 \leq p < \infty$ then:

$$E(|X+Y|^p)^{\frac{1}{p}} \leq E(|X|^p)^{\frac{1}{p}} + E(|Y|^p)^{\frac{1}{p}}$$

6 Sampling Model

6.1 Joint Distribution

For Discrete RVs:

The **joint density** function is defined by:

$$f(.,.) : \mathbb{R}_X \times \mathbb{R}_Y \to [0,1]$$
$$f(x,y) = \mathbb{P}\left\{s : X(s) = x, Y(s) = y\right\}, (x,y) \in \mathbb{R}_X \times \mathbb{R}_Y$$

For Continuous RVs:

The joint distribution function:

$$F_{XY}(.,.) : \mathbb{R}^2 \to [0,1],$$

$$F_{XY}(x,y) = \mathbb{P}\{s : X(s) \leq x, Y(s) \leq y\} = P_{XY}((-\infty, x) \times (-\infty, y)), \quad (x,y) \in \mathbb{R}^2$$

The joint cdf can also be defined on intervals of the form $(a,b]$

$$\mathbb{P}\{s : x_1 < X(s) \leq x_2, y_1 < Y(s) \leq y_2\} = F(x_2,y_2) - F(x_1,y_2) - F(x_2,y_1) + F(x_1,y_1)$$

6.1.1 joint density

Bivariate

Assuming that $f(x,y) > 0$ exists, the joint density function is defined via:

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v) du dv$$

In the case where $F(x,y)$ is differentiable at (x,y) we can derive the **joint density** by partial differentiation:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y},$$

at all continuity points of $f(x,y)$.

Properties

1. $f(x,y) \geq 0 \forall (x,y) \in \mathbb{R}_X \times \mathbb{R}_Y$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx \cdot dy = 1$
3. $F_{XY}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x,y) dx dy$
4. $f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$
5. $\sum\limits_{i=1}^{\infty} \sum\limits_{j=1}^{\infty} f(x_i, y_j) = 1, F(x_k, y_m) = \sum\limits_{i=1}^{k} \sum\limits_{j=1}^{m} f(x_i, y_j)$

n-random variables:

1. $f(x_1, x_2, ..., x_n) \geq 0, \forall (x_1, x_2, ..., x_n) \in \mathbb{R}_X^n$
2. $\int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(x_1, x_2, ..., x_n) dx_1 dx_2 \cdots dx_n = 1$
3. $F(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(u_1, u_2, ..., u_n) du_1 du_2 ... du_n$

6.2 Joint moments

We define the **joint product moments** of order (k, m) by:

$$\mu'_{km} = E\left\{X^k Y^m\right\}, k, m = 0, 1, 2, ...,$$

and the **joint central moments** of order (k, m) are defined by:

$$\mu_{km} = E\left\{(X - E(X))^k (Y - E(Y))^m\right\}, k, m = 0, 1, 2, ...$$

$$\begin{aligned}
\mu'_{10} &= E(X), & \mu_{10} &= 0 \\
\mu'_{01} &= E(Y), & \mu_{01} &= 0 \\
\mu'_{20} &= E(X)^2 + Var(X), & \mu_{20} &= Var(X) \\
\mu'_{02} &= E(Y)^2 + Var(Y), & \mu_{02} &= Var(Y) \\
\mu'_{11} &= E(XY), & \mu_{11} &= E[(X - E(X))(Y - E(Y))]
\end{aligned}$$

The most widely used joint moment is the **covariance**, defined by

$$\mu_{11} := Cov(X, Y) = E\left\{[X - E(X)][Y - E(X)]\right\}$$

Properties:

1. $Cov(X, Y) = E(XY) - E(X) \cdot E(Y)$

2. $Cov(X, Y) = Cov(Y, X)$

3. $Cov(aX + bY, Z) = aCov(X, Y) + bCov(Y, Z)$, for $(a, b) \in \mathbb{R}^2$

6.3 Marginal Distribution

The **marginal distribution** is derived via a limiting process of the join cdf:

$$F_X(x) = \lim_{y\to\infty} F(x,y) \text{ and } F_Y(y) = \lim_{x\to\infty} F(x,y)$$

For example, consider the bivariate exponential distribution:

$$F(x,y) = (1-e^{-\alpha x})(1-e^{-\beta y}), \alpha, \beta, x, y > 0$$

The marginal distribution:

$$\begin{aligned} F_X(x) = \lim_{y\to\infty} F(x,y) &= \lim_{y\to\infty}(1-e^{-\alpha x})(1-e^{-\beta y}) \\ &= 1-e^{-\alpha x} \\ F_Y(y) = \lim_{x\to\infty} F(x,y) &= \lim_{x\to\infty}(1-e^{-\alpha x})(1-e^{-\beta y}) \\ &= 1-e^{-\beta y} \end{aligned}$$

6.4 Defined with density funciton

6.4.1 Continuous RV

$$F_X(x) = \lim_{y\to\infty} F(x,y) = \lim_{y\to\infty} \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dydx = \int_{-\infty}^{x}\left[\int_{-\infty}^{\infty} f(x,y)dy\right]dx$$

And,

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy, x \in \mathbb{R}_X$$

$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)dx, y \in \mathbb{R}_Y$$

6.4.2 Discrete RV

$$f_x(x) = \sum_{i=1}^{\infty} f(x, y_i), X \in \mathbb{R}_X$$

$$f_y(y) = \sum_{i=1}^{\infty} f(x_i, y), y \in \mathbb{R}_y$$

7 Conditional Distribution

7.1 Discrete RVs

Two RVs X and Y

$$A = \{Y = y\}, B = \{X = x\}$$

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

$$\mathbb{P}(Y = y, X = x) = f(x, y)$$

$$\mathbb{P}(Y = y|X = x) = f(y|x)$$

7.2 Continuous RVs

Two RVs X and Y

$$A = \{X \leq x\}, B = \{Y \leq y\}$$

The conditional cdf:

$$F_{Y|X}(y|X = x) = \lim_{h \to 0^+} \frac{\mathbb{P}(Y \leq y, x \leq X \leq x + h)}{\mathbb{P}(x \leq X \leq x + h)} = \int_{-\infty}^{y} \frac{f(x, u)}{f_x(x)} du$$

Note:

This suggest that we could indeed define the conditional density function but we **should**

not interpret it as **assigning** probabilities because:

$$f(.|x) : \mathbb{R}_Y \to [0, \infty)$$

7.2.1 Properties:

1. $f(y|x) \geq 0, \quad \forall y \in \mathbb{R}_Y$
2. $\int_{-\infty}^{\infty} f(y|x)dy = 1$
3. $F(y|x) = \int_{-\infty}^{y} f(u|x)du$

For example:

Consider the case where the joint density function:

$$f(x, y) = 8xy, \quad 0 < x < y, 0 < y < 1$$

The marginal densities of x and y can be derived from the joint density:

$$\begin{aligned} f_x(x) &= \int_x^1 (8xy)dy = 4xy^2|_{y=x}^{y=1} = 4x(1 - x^2), \quad 0 < x < 1 \\ f_y(y) &= \int_0^y (8xy)dx = 4x^2y|_{x=0}^{x=y} = 4y^3, \quad 0 < y < 1 \end{aligned}$$

Then, we can derive the conditional densities:

$$\begin{aligned} f(y|x) &= \frac{8xy}{4x(1 - x^2)} = \frac{2y}{(1 - x^2)}, \quad x < y < 1, 0 < x < 1 \\ f(x|y) &= \frac{8xy}{4y^3} = \frac{2x}{y^2}, \quad 0 < x < y, 0 < y < 1 \end{aligned}$$

Note:

The range of X and Y is constrained by each other.

7.3 Continuous and Discrete RVs

It turns out that a most convenient way to specify such a joint distribution is via the conditional density

Consider the case where $F(x, y)$ is the joint cdf of the RV (X, Y) where X is **discrete** and Y is **continuous**. Let $\mathbb{R}_X = \{x_1, x_2, ...\}$ be the range of values of the RV X. The joint cdf is completely determined by the sequence of pairs of a marginal probability and the associated conditional density:

$$(f_x(x_k), f(y|x_k), \forall y_k \in \mathbb{R}_X)$$

The only difficulty in this result is how to specify the conditional density. It is defined by:

$$f(y|x_k) = \frac{1}{f_x(x_k)} \frac{d[F(x_k, y) - F(x_k - 0, y)]}{dy},$$

where the notation $(x_k - 0)$ indicates taking the derivative from the left, such that:

$$F(x, y) = \sum_{x_k \leq x} f_x(x_k) \int_{-\infty}^{y} f(u|x_k) du$$

Similarly, the marginal distribution of the random variable Y is defined by:

$$F_Y(y) = \sum_{x_k \in \mathbb{R}_X} f_x(x_k) \int_{-\infty}^{y} f(u|x_k) du$$

7.4 Conditional moments

8 Useful Densities

Note: Different density functions can have very similar shape with specific parameter. The best way to distinguish between them is via **index measures based on moments** which are invariant to scale and location parameter changes (skewness, kurtosis).

For example, in the case of modeling data referring to exam scores it is often more realistic to use the **Beta** and **not** the **Normal** distribution because all scores can be easily expressed in the interval [0,1]; the Normal distribution has support $(-\infty, \infty)$.

8.1 Bernoulli

Table 1

y	0	1
$f(y;\theta)$	$1-\theta$	θ

$\mathbb{P}(Y = 1) = \theta$, where $0 \leq \theta \leq 1$. Bernoulli density:

$$f(y;\theta) = \theta^y (1-\theta)^{1-y},$$

where $\theta \in [0,1], y = 0,1$

Mean:

$$\mu(\theta) := E(X) = 0 \cdot (1-\theta) + 1 \cdot \theta = \theta$$

Variance:

$$Var(X) = E(X - E[X])^2 = (0-\theta)^2(1-\theta) + (1-\theta)^2\theta = \theta(1-\theta)$$

8.2 Binomial distribution by Bernoulli

18th century

$$f(x,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

where $\theta \in [0,1], x = 0,1, n = 1,2,3,...$ $\binom{n}{x} = \frac{n!}{(n-x)!x!}$

8.3 Normal Distribution

By de Moivre and Laplace in the early 19th century.

$$f(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

where $\theta := (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+, x \in \mathbb{R}$

$$F_X(x;\theta) = \int_{-\infty}^{x} f_x(u)du$$

Mean:

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \\
&= \int_{-\infty}^{\infty} \left(\frac{\sigma z+\mu}{\sigma\sqrt{2\pi}}\right) \text{ exp } \left[-\frac{z^2}{2}\right] (\sigma)dz \\
&= \left(\frac{\sigma}{\sqrt{2\pi}}\right) \int_{-\infty}^{\infty} z \text{ exp } \left[-\frac{z^2}{2}\right] dz + \mu \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right) \text{ exp } \left[-\frac{z^2}{2}\right] dz \\
&= 0 + \mu \cdot 1 \\
&= \mu
\end{aligned}$$

where $z = \left(\frac{x-\mu}{\sigma}\right)$ or $x = \sigma z + \mu$ with $\frac{dx}{dz} = \sigma$.

Variance: $Var(X) = E(X^2) - (E(X))^2 = \sigma^2$, where $E(X^2) = \sigma^2 + \mu^2$

8.4 Uniform Distribution

8.4.1 Continuous form:

$$f_x(x;\theta) = \frac{1}{b-a},$$

where $\boldsymbol{\theta} := (a, b) \in \mathbb{R}^2, a \leq x \leq b$

$$F_X(x;\theta) = \frac{x-a}{b-a},$$

where $\boldsymbol{\theta} := (a,b) \in \mathbb{R}^2, a \leq x \leq b$

Mean:

$$\mu(\theta) = E[X] = \int_{\theta_1}^{\theta_2} \frac{x}{(\theta_2 - \theta_1)} dx = \frac{1}{2} \frac{1}{(\theta_2 - \theta_1)} x^2 \Big|_{\theta_2}^{\theta_1} = \frac{\theta_1 + \theta_2}{2}$$

8.4.2 Discrete form:

$$f_x(x;\theta) = \frac{1}{\theta + 1},$$

where θ is an integer, $x = 0, 1, 2, ..., \theta$. θ is the maximum value of x. For example, if $\theta = 9$, then x can be 0,1,2,3,4,5,6,7,8,9. The density would be

$$f_x(x;9) = \frac{1}{9+1} = \frac{1}{10} = 0.1$$

$$F_X(x;\theta) = \frac{x+1}{\theta + 1},$$

where θ is an integer, $x = 0, 1, 2, ..., \theta$

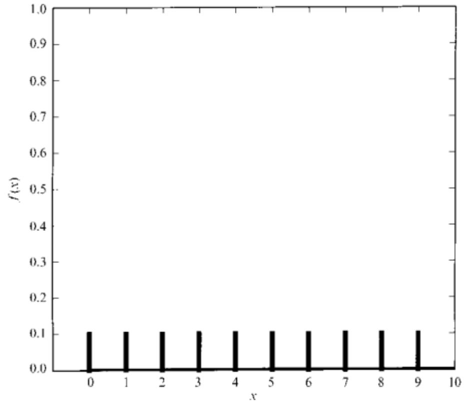


Figure 3.14 Uniform (discrete) density

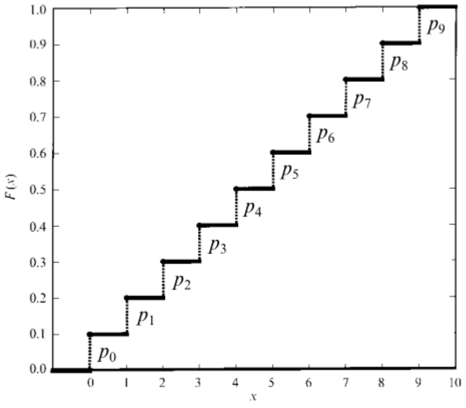


Figure 3.15 Uniform (discrete) cdf

8.5 Poisson (discrete)

8.5.1 Continuous:

$$f_x(x;\theta) = \frac{e^{-\theta}\theta^x}{x!},$$

where $\theta > 0, x = 0, 1, 2, 3, ...$

Mean:

$$\mu(\theta) := E(X) = \sum_{k=0}^{\infty} k \left(\frac{e^{-\theta}\theta^k}{k!} \right) = \theta e^{-\theta} \sum_{k=0}^{\infty} \frac{\theta^{k-1}}{(k-1)!} = \theta, \text{ since } \sum_{k=0}^{\infty} \frac{\theta^{k-1}}{(k-1)!} = e^{\theta}$$

8.5.2 Discrete

$$F_X(x;\theta) = \sum_{k=0}^{x} \frac{e^{-\theta}\theta^x}{k!},$$

where $\theta > 0, x = 0, 1, 2, 3, ...$

8.6 Exponential Distribution

DF:

$$F_X(x;\theta) = 1 - e^{-\theta x},$$

where $\theta > 0, x \in \mathbb{R}_+ := [0, \infty]$

PDF:

$$f_x(x;\theta) = \theta e^{-\theta x},$$

where $\theta > 0, x \in \mathbb{R}_+$

8.7 Cauchy distribution

Cauchy distribution $C(\alpha, \beta)$ has no moments.

$$F(x;\alpha,\beta) = \frac{1}{2} + \left(\frac{1}{\pi} tan^{-1}\left(\frac{x-\alpha}{\beta}\right)\right)$$

$$f(x;\alpha,\beta) = \frac{1}{\beta\left(1+\left(\frac{x-\alpha}{\beta}\right)^2\right)}$$

where $\alpha \in \mathbb{R}, \beta \in \mathbb{R}_+, x \in \mathbb{R}$.