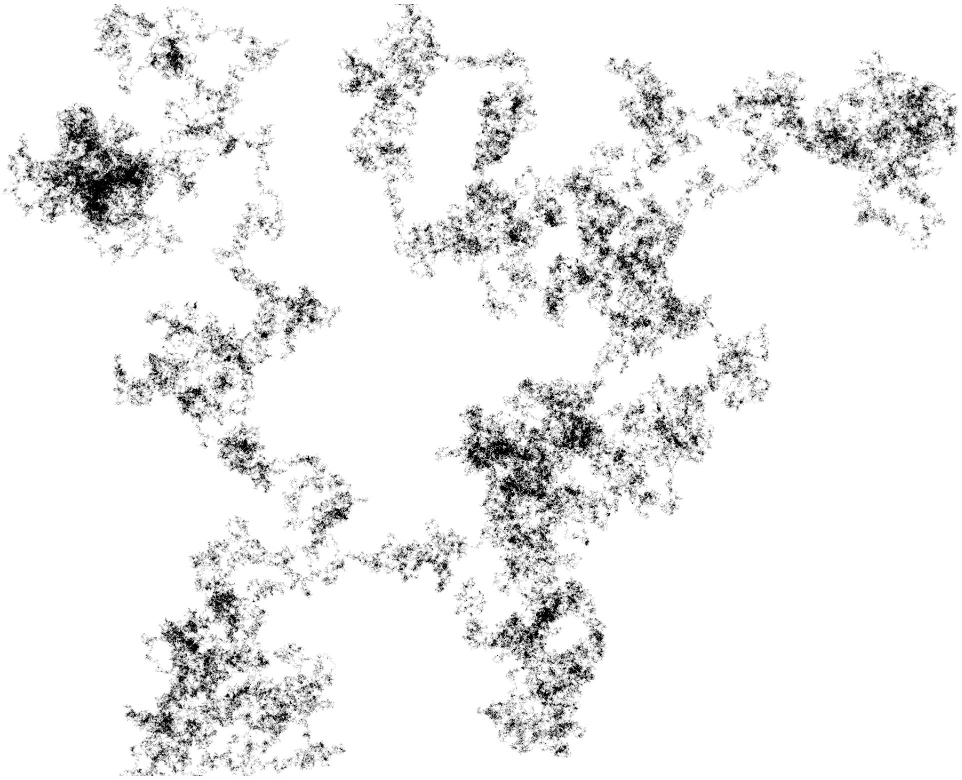
MATH 356 - Honours Probability
Pr. Johanna Nešlehová

Course notes by
Léo Raymond-Belzile
Leo.Raymond-Belzile@mail.mcgill.ca



The current version is that of April 20, 2012

Fall 2011, McGill University

Please signal the author by email if you find any typo.

These notes have not been revised and should be read carefully.

Contents

Chapter 1
Kolmogorov's axioms

We denote by Ω a sample space, that is the set of *all* possible outcomes of a random experiment or a situation whose outcome is unknown, random, etc.

Examples in the discrete (and finite) case include coin toss $\Omega = \{1, 2\}$, the sample space of a coin toss, $\Omega = \{1, 2, 3, 4, 5, 6\}$, a dice toss, $\Omega = \{$ clear, cloudy, rainy $\}$ for weather. We may also look at $\Omega = \{\mathbb{N}\}$, the number of insurance claims, $\Omega = \{[0, \infty)\}$ for water level or $\Omega = \{\mathbb{R}\}$ for temperature. We could of course put upper or lower bounds to these examples in practise, although the point of the exercise is to show that we can have countable or infinite sample space, or some more complicated ones, as $\Omega = \{C([0, t])\}$, the space of continuous functions from 0 to time t.

Probability

Let $A \subseteq \Omega$ be an event. We want to assign probabilities to events.

Terminology

1. $\Omega \subseteq \Omega$: sure event
2. $\emptyset \subseteq \Omega$: impossible event
3. $\omega \subseteq \Omega, \omega \subset \Omega$: singleton, elementary event
4. $A, B \subseteq \Omega$,
   (a) $A \cap B$: A occurred and B occurred as well
   (b) $A \cup B$: A or B occurred
   (c) $A \cap B^{\complement}$: A occurred, but not B
   (d) $A^{\complement}$: complementary event
   (e) $\bigcup_{i=1}^{\infty} A_i$: A_i occurred for at least one $i \in \mathbb{N}$.
   (f) $\bigcap_{i=1}^{\infty} A_i$: A_i occurred for all $i \in \mathbb{N}$.

Example 1.1

Dice toss, $\Omega = \{1, 2, 3, 4, 5, 6\}$, with $A = 1, 3, 5$ and $B = 5, 6$

Definition 1.1 (σ-field)

Let Ω be a sample space. A set of subsets of Ω, $\mathcal{A} \subseteq \mathcal{P}(\Omega)$[1], is called a σ-algebra if

1. $\Omega \in \mathcal{A}$;
2. $A \in \mathcal{A}$, then $A^{\complement} \in \mathcal{A}$;
3. A_1, A_2, A_3 a countable sequence in $\mathcal{A}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$. [2]

Remark

Let $\mathcal{A}$ be a σ-field, then

a) $\emptyset \in \mathcal{A}$ since $\emptyset = \Omega^{\complement}$

b) If $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$, since it is the same as $A \cup B \cup \emptyset \cup \emptyset \dots$ More generally, $A_1, \dots, A_n \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{A}$.

c) If $A, B \in \mathcal{A}$, then by De Morgan's law $A \cap B = \left(A^{\complement} \cup B^{\complement}\right)^{\complement}$

d) If $A_1, A_2, A_3, \dots \in \mathcal{A}$, then $\bigcap_{i=1}^{\infty} A_i = \left(\bigcup_{i=1}^{\infty} A_i^{\complement}\right) \in \mathcal{A}$.

e) $A \setminus B = A \cap B^{\complement} \in \mathcal{A}$.

Example 1.2

1. $\mathcal{P}(\Omega)$ is the largest σ-field;
2. $\{\emptyset, \Omega\}$ is the smallest σ-field
3. Let $A \subseteq \Omega$. The smallest σ-field that contains A is $\{\Omega, \emptyset, A, A^{\complement}\}$

Example 1.3

Let $\Omega = \mathbb{R}$ and let

$$\mathcal{A} = \{A \in \mathbb{R} : A \text{ is countable or } A^{\complement} \text{ is countable}\}$$

$$\mathcal{B} = \{B \in \mathbb{R} : B \text{ is finite or } B^{\complement} \text{ is finite}\}$$

[1] The power set of a set S is defined as the set of all subsets of S, including trivial sets

[2] The σ refers to the fact we take infinite union of sets

We can see that $\mathcal{A}$ is a σ-field, but that $\mathcal{B}$ is not

Proof

1. $\Omega^{\complement} = \emptyset$, so $\Omega \in \mathcal{B}$ and $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^{\complement} \in \mathcal{A}$, and similarly $A \in \mathcal{B} \Rightarrow A^{\complement} \in \mathcal{B}$
3. $A_1, A_2, A_3, \ldots \in \mathcal{A}$. Consider $\bigcup_{i=1}^{\infty} A_i$. Then we have two cases:
   (a) A_i is countable for all $i \in \mathbb{N}$, then the union $\bigcup_{i=1}^{\infty} A_i$ is also countable
   (b) There exists an $i \in \mathbb{N}$, say i_0 such that A_{i_0} is uncountable. Then $(\bigcup_{i=1}^{\infty} A_i)^{\complement} = \bigcap_{i=1}^{\infty} {A_i}^{\complement} \subseteq {A_{i_0}}^{\complement}$, which is countable by definition and indeed in $\mathcal{A}$.
   (c) For $\mathcal{B}$, let's give a counterexample.
   Let $A_1, A_2, \ldots \in \mathcal{B}$ and $A_i = \{i\} \in \mathcal{B}$, but $\bigcup_{i=1}^{\infty} A_i = \mathbb{N} \notin \mathcal{B}$.

□

Note
$\mathcal{A} \neq \mathcal{P}(\mathbb{R})$. Just consider the interval $[0, 1]$.

Theorem 1.2 (σ-field generated by $\mathcal{E}$)
Let $\mathcal{E} \in \mathcal{P}(\mathbb{R})$. Then, there exists a unique smallest σ-field such that

1. $\mathcal{E} \subseteq \sigma(\mathcal{E})$
2. If $\mathcal{A}$ is a σ-field such that $\mathcal{E} \subseteq \mathcal{A}$, then $\sigma(\mathcal{E}) \subseteq \mathcal{A}$.

And therefore $\sigma(\mathcal{E})$ is also called the "σ-field generated by $\mathcal{E}$".

Proof Note that the proof is not constructive.
Let $\sigma(\mathcal{E}) = \bigcap_{i \in I} \mathcal{A}_i$, where $\mathcal{A}_i$ is a σ-field for any i such that $\mathcal{E} \subseteq \mathcal{A}_i$ and I is an arbitrary set of indices.

1. $\mathcal{P}(\Omega) \supseteq \mathcal{E}$ power set is there, so non-empty

2. $\mathcal{E} \subseteq \sigma(\mathcal{E})$ by our requirement, $\mathcal{E} \subseteq \mathcal{A}_i \ \forall\, i \in I$.

3. If $\mathcal{A}$ is a σ-field such that $\mathcal{E} \in \mathcal{A}$, then $\sigma(\mathcal{E}) \subseteq \mathcal{A}$. This proves that it's smallest.

4. $\sigma(\mathcal{E})$ is a σ-field

   (a) $\Omega \in \sigma(\mathcal{E})$ since $\Omega \in \mathcal{A}_i \ \forall\, i \in I \Rightarrow \Omega \in \bigcap_{i \in I} \mathcal{A}_i = \sigma(\mathcal{E})$

   (b) $A \in \sigma(\mathcal{E}) \Rightarrow A \in \mathcal{A}_i \ \forall\, i \in I \Rightarrow A^{\complement} \in \mathcal{A}_i \ \forall\, i \in I \Rightarrow A^{\complement} \in \bigcap_{i \in I} \mathcal{A}_i$

   (c) $A_1, A_2, \ldots \in \sigma(\mathcal{E}) \Rightarrow A_k \in \mathcal{A}_i$ for $k = 1, 2, \ldots \ \forall\, i \in I$. Since they are σ-fields, then $\bigcup_{k=1}^{\infty} A_k \in \mathcal{A}_i \ \forall\, i \in I \Rightarrow \bigcup_{k=1}^{\infty} A_i \in \bigcap_{i \in I} A_i$.

□

Example 1.4

1. Ω countable (or finite)
   $A \subseteq \Omega$, then $A = \bigcup_{\omega : \omega \in A} \{\omega\} \subset \sigma(\mathcal{E})$.
   So say $A = \{2, 4, 6\} \Rightarrow A = \{2\} \cup \{4\} \cup \{6\}$

2. $\Omega = \mathbb{R} \quad (\Omega = \mathbb{R}^n)$
   $\mathcal{E} = \{\{\omega\}, \omega \in \mathbb{R}\}, \sigma(\mathcal{E}) = \{A \subseteq \mathbb{R} : A \text{ or } A^{\complement} \text{ is countable}\}$.

Definition 1.3 (Borel σ-field)

The Borel σ-field on $\mathbb{R}$, denoted by $\mathbb{B}$, is the smallest σ-field that contains

$$\mathcal{E} = \{(a, b], a \leq b \in \mathbb{R}\}$$

Remark

1. If $x \in \mathbb{R}$, then $\{x\} \in \mathbb{B}$;

2. $(a, b) = (a, b] \setminus \{b\}$, therefore $[a, b), [a, b]$, $(-\infty, b)$, $(-\infty, b]$, $[a, \infty)$, (a, ∞) and $(-\infty, \infty)$ are all in $\mathbb{B}$.

3. $O \subseteq \mathbb{R}$ an open set, then $O = \bigcup_{i=1}^{\infty} (a_i, b_i)$ are also in $\mathbb{B}$. Since complement is also by the axioms, then $C \subseteq \mathbb{R}$, the closed set, is also in $\mathbb{B}$.

Notation

Elements of $\mathbb{B}$ are called Borel sets .

Note

Not every subset of $\mathbb{R}$ is a Borel set. Examples include the Vitali sets.

Remark

For more complex sample spaces, (ex. $\Omega = C[0,t]$), then the Borel σ-field is the smallest σ-field generated (contained) by the open subsets of Ω.

Definition 1.4 (Probability)

Let Ω be a sample space and $\mathcal{A}$ a σ-field. The mapping $\mathsf{P} : \mathcal{A} \to \mathbb{R}$ is called a probability (probability measure) if the following three axioms are fulfilled:

1. $\mathsf{P}(A) \geq 0, A \in \mathcal{A}$;
2. $\mathsf{P}(\Omega) = 1$;
3. If $A_1, A_2, A_3, \ldots \in \mathcal{A}$ such that $A_i \cap A_j = \emptyset$ if $i \neq j$, then

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

The triplet $(\Omega, \mathcal{A}, \mathsf{P})$ is called a probability space .

Remark

$\mu : \mathcal{A} \to \overline{\mathbb{R}}$[3], which satisfies $1, 3$ and $\mu(\emptyset) = 0$ is called a measure.

Note

Every probability is a measure. $\mathsf{P}(\emptyset) : \mathsf{P}(\Omega) = 1 = \mathsf{P}(\Omega \cup \emptyset)$, but $\Omega \cap \emptyset = \emptyset$, hence $1 = 1 + \mathsf{P}(\emptyset)$ so this implies $P(\emptyset) = 0$.

Example 1.5

$\Omega = \{\omega_1, \ldots, \omega_n\}, \mathcal{A} = \mathcal{P}(\Omega)$ and $P(\{\omega_i\}) = \frac{1}{n}$ so

$$\mathsf{P}(A) = \mathsf{P}(\bigcup_{\omega \in A} \omega) = \sum_{\omega \in A} \frac{1}{n} = \frac{|A|}{n}$$

for equally likely events.

[3] $\mathbb{R}$ including $\pm\infty$

Example 1.6

$\Omega = [0,1], \mathcal{A} = \{\mathcal{B} \cap [0,1], \mathcal{B} \in \mathbb{B}\}$ and $P(A) = \int_0^1 1_A\{x \in A\}dx$.

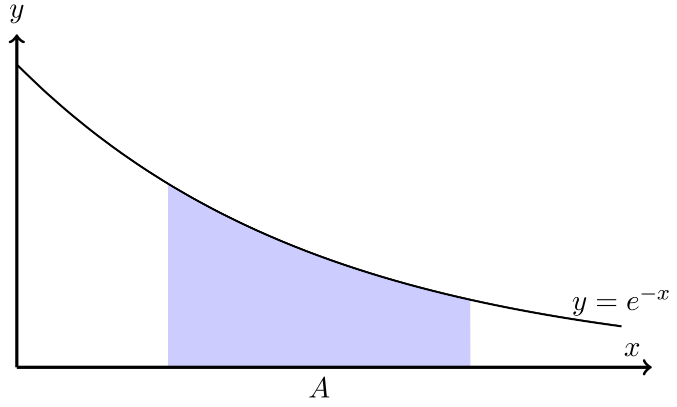


Figure 1: Probability of A for a continuous function; $\mathsf{P}(A) = \int_A e^{-x}dx$

Intuitively, the outcome of the random experiment is $\omega \in \Omega$. Suppose we are able to repeat the experiment independently as often as we like. n repetitions lead to $\omega_1, \dots, \omega_n$ and

$$\mathsf{P}(A) \approx \frac{\#\omega_i \in A}{n} = \frac{\sum_{i=1}^n 1_A(\omega_i)}{n}$$

where $1_A(x)$ denotes the indicator function, namely takes value 1 if $x \in A$ and 0 if $x \notin A$.

Consequences of Kolmogorov's axioms

Lemma 1.5 (Properties of probability spaces)

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space. Then

1. $\mathsf{P}(\emptyset) = 0$;
2. $\mathsf{P}(A^{\complement}) = 1 - \mathsf{P}(A)$ for $A \in \mathcal{A}$;
3. If $A, B \in \mathcal{A}$ are such that $A \subseteq B$, then $\mathsf{P}(B \setminus A) = \mathsf{P}(B) - \mathsf{P}(A)$ and $\mathsf{P}(B) \geq \mathsf{P}(A)$;

4. $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$, for $A, B \in \mathcal{A}$;

5. $\mathsf{P}(A \cup B) \leq \mathsf{P}(A) + \mathsf{P}(B)$ for $A, B \in \mathcal{A}$;

6. Inclusion-exclusion principle or Sieve-formula: if $A_1, \ldots, A_n \in \mathcal{A}$, then

$$\begin{aligned} \mathsf{P}\left(\bigcup_{i=1}^{n} A_n\right) &= \sum_{i=1}^{n} \mathsf{P}(A_i) - \sum_{i<j} \mathsf{P}(A_i \cap A_j) + \sum_{i<j<k} \mathsf{P}(A_i \cap A_j \cap A_k) \\ &\quad - \cdots + (-1)^{n+1} \mathsf{P}(A_1 \cap \ldots \cap A_n) \\ &= \sum_{\emptyset \subseteq I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1}\, \mathsf{P}\left(\bigcap_{i \in I} A_i\right) \end{aligned}$$

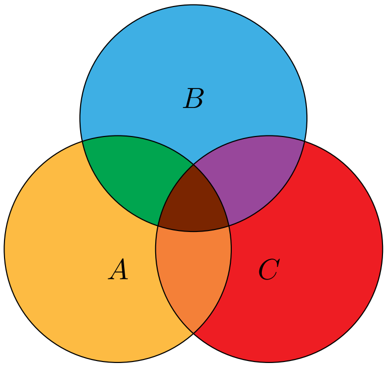


Figure 2: Venn Diagram illustrating the Inclusion-exclusion principle

Proof

2. Given that $A \cup A^{\complement} = \Omega, A \cap A^{\complement} = \emptyset$ and $\mathsf{P}(\Omega) = 1 = P(A \cup A^{\complement})$, then $\overset{(3)}{=} \mathsf{P}(A) + \mathsf{P}(A^{\complement}) \Rightarrow \mathsf{P}(A^{\complement}) = 1 - \mathsf{P}(A)$;

3. Write B as $B = A \cup (B \setminus A), A \cap (B \setminus A) = \emptyset$. Then the probability $\mathsf{P}(B) = \mathsf{P}(A \cup (B \setminus A)) \overset{(3)}{=} \mathsf{P}(A) + \mathsf{P}(B \setminus A) \Rightarrow \mathsf{P}(B \setminus A) = \mathsf{P}(B) - \mathsf{P}(A)$ since $\mathsf{P}(B \setminus A) > 0$ by definition, hence $\mathsf{P}(B) > \mathsf{P}(A)$;

4. In a similar fashion, express $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$. The right hand side terms are pairwise disjoint, so

$$\mathsf{P}(A \cup B) \overset{(3)}{=} \mathsf{P}(A \setminus B) + \mathsf{P}(A \cap B) + \mathsf{P}(B \setminus A).$$

   But $\mathsf{P}(A) = \mathsf{P}((A \setminus B) \cup (A \cap B)) = \mathsf{P}(A \setminus B) + \mathsf{P}(A \cap B)$. We also recall that $\mathsf{P}(A \setminus B) = \mathsf{P}(A) - \mathsf{P}(A \cap B)$ and similarly $\mathsf{P}(B \setminus A) = \mathsf{P}(B) - \mathsf{P}(A \cap B)$, hence $\mathsf{P}(A \cup B) = \mathsf{P}(A) - \mathsf{P}(A \cap B) + \mathsf{P}(A \cap B) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$ which by cancelling is equal to $\mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$.

5. Follows from (4) since $\mathsf{P}(A \cap B) \geq 0$;

6. The proof is by induction on n, starting with $n = 1$ as the trivial case $\mathsf{P}(A) = \mathsf{P}(A_1)$. Suppose the induction hypothesis holds for n, then for $n + 1$, start by splitting the union in two parts:

$$\mathsf{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \mathsf{P}\left(\left\{\bigcup_{i=1}^{n} A_i\right\} \cup A_{n+1}\right)$$

   is by applying (4)

$$= \mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathsf{P}(A_{n+1}) - \mathsf{P}\left(A_{n+1} \cap \left\{\bigcup_{i=1}^{n} A_i\right\}\right)$$

   is the same as the union of intersections of events $A_1, \ldots A_n$ with A_{n+1}

$$= \mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathsf{P}(A_{n+1}) - \mathsf{P}\left(\bigcup_{i=1}^{n} (A_1 \cap A_{n+1})\right)$$

   Then by applying the induction hypothesis to first and last term (regrouping the negative inside the sum), we get

$$\sum_{\emptyset \subseteq I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} \mathsf{P}\left(\bigcap_{i \in I} A_i\right) + \mathsf{P}(A_{n+1})$$
$$+ \sum_{\emptyset \subseteq I \subseteq \{1,\ldots,n\}} (-1)^{|I|+2} \quad \underbrace{\mathsf{P}\left(\bigcap_{i \in I} \{A_i \cap A_{n+1}\}\right)}_{}$$

where the underlined probability is just equal to

$$\mathsf{P}\left(\bigcap_{i\in I\cup\{n+1\}} A_i\right).$$

Then, in this case, either it does contain $n+1$ (second and third terms cancel), else partly, or not at all. In all cases, we get the desired result, that is

$$= \sum_{\emptyset\subseteq I\subseteq\{1,\ldots,n+1\}} (-1)^{|I|+1}\mathsf{P}\left(\bigcap_{i\in I} A_i\right).$$

□

Remark

Note that if $\mathsf{P}(\emptyset) = 0$ and $\mathsf{P}(A) = 0 \not\Rightarrow A = \emptyset$. Just consider $\int_A e^{-x}dx$ where $A = \{x_0\}$, then $\int_{x_0}^{x_0} e^{-x}dx = 0$. Analogously, $\mathsf{P}(\Omega) = 1$, but $\mathsf{P}(A) = 1 \not\Rightarrow A = \Omega$. Consider $[0,\infty)\setminus\{x_0\}$. While Ω is called *sure event*, A is called *almost sure event* if $\mathsf{P}(A) = 1$.

Corollary 1.6 (Bonferroni inequality)

Let $A_1,\ldots,A_n \in \mathcal{A}, n \in \mathbb{N}$. Then

$$\sum_{i=1}^n \mathsf{P}(A_i) - \sum_{i<j}\mathsf{P}(A_i\cap A_j) \leq \mathsf{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathsf{P}(A_i).$$

Proof The proof is similar in style to the proof of Sieve formula, by induction on n. Consider first the right hand side of the inequality. Clearly, $\mathsf{P}(A_i) = \mathsf{P}(A_i)$. Assume as an induction hypothesis that it holds for n. Then for $n+1$,

$$\mathsf{P}\left(\bigcup_{i=1}^{n+1} A_i\right) \leq \mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathsf{P}(A_{n+1}) \leq \sum_{i=1}^{n+1}\mathsf{P}(A_i)$$

For the left hand side of the Bonferroni inequality, $\mathsf{P}(A_i) = \mathsf{P}(A_i)$ and by induction, we get for $n+1$

$$\mathsf{P}\left(\bigcup_{i=1}^{n+1} A_i\right) \geq \sum_{i=1}^{n} \mathsf{P}(A_i) - \sum_{i<j}^{n} \mathsf{P}(A_i \cap A_j) + \mathsf{P}(A_{n+1}) - \sum_{i=1}^{n} (A_i \cap A_{n+1})$$

and remarking the right hand side of the above is equal to

$$\sum_{i=1}^{n+1} \mathsf{P}(A_i) - \sum_{i<j}^{n+1} \mathsf{P}(A_i \cap A_j)$$

we get the desired result. □

Corollary 1.7 (Boole's inequality)

Let $A_1, A_2, \ldots, A_n \in \mathcal{A}$. Then

$$\mathsf{P}(A_1 \cap A_2) \geq 1 - \mathsf{P}({A_1}^{\mathsf{C}}) - \mathsf{P}({A_2}^{\mathsf{C}})$$

and more generally, for $n \in \mathbb{N}$

$$\mathsf{P}\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} \mathsf{P}({A_i}^{\mathsf{C}})$$

Proof

$$\mathsf{P}\left(\left\{\bigcap_{i=1}^{n} A_i\right\}^{\mathsf{C}}\right) = \mathsf{P}\left(\bigcup_{i=1}^{n} {A_i}^{\mathsf{C}}\right) \leq \sum_{i=1}^{n} \mathsf{P}({A_i}^{\mathsf{C}})$$

but also

$$\mathsf{P}\left(\left\{\bigcap_{i=1}^{n} A_i\right\}^{\mathsf{C}}\right) = 1 - \mathsf{P}\left(\bigcap_{i=1}^{n} A_i\right)$$

Rearranging the equalities yield the result. □

Theorem 1.8

Let $\{A_i\}$ be a sequence of events in $\mathcal{A}$. Then

1. If the sequence is increasing, $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots \subseteq A_n \subseteq A_{n+1}$, then
$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \mathsf{P}(A_n);$$
2. If the sequence is decreasing $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n$, then
$$\mathsf{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \mathsf{P}(A_n).$$

Proof (1) Since the sequence is increasing, we can write each set as follows: $A_1 = A_1$, $A_2 = A_1 \cup (A_2 \setminus A_1)$, $A_3 = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2)$, $\ldots$, $A_n = \bigcup_{i=1}^{n}(A_i \setminus A_{i-1})$, and $A_0 = \emptyset$. We have

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} (A_i \setminus A_{i-1})$$

where $(A_i \setminus A_{i-1})$ are pairwise disjoint since $A_{i-1} \subseteq A_i$. Then

$$\begin{aligned}\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i \setminus A_{i-1}) &= \lim_{n\to\infty} \sum_{i=1}^{n} \mathsf{P}(A_i \setminus A_{i-1}) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \underbrace{\{\mathsf{P}(A_i) - \mathsf{P}(A_{i-1})\}}_{\text{telescoping sum}}\end{aligned}$$

which is equal to $\lim_{n\to\infty} \mathsf{P}(A_n)$ and hence $\mathsf{P}(A_1) \le \mathsf{P}(A_2) \le \ldots \le 1$ □

(2) The sequence $\{A_n\}$ is decreasing, then ${A_1}^{\mathsf{C}} \subseteq {A_2}^{\mathsf{C}} \subseteq \ldots$ so by (1)

$$\begin{aligned}\mathsf{P}\left(\bigcup_{i=1}^{\infty} {A_1}^{\mathsf{C}}\right) &= \lim_{n\to\infty} \mathsf{P}({A_n}^{\mathsf{C}}) \\ 1 - \mathsf{P}\left(\bigcap_{i=1}^{\infty} A_i\right) &= 1 - \lim_{n\to\infty} \mathsf{P}(A_n)\end{aligned}$$

Corollary 1.9

Let $\{A_i\}$ be a sequence in $\mathcal{A}$. Then

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

Proof Consider the following sequence: $B_n = \bigcup_{i=1}^{n} A_i$, $n = 1, 2, \ldots$ and $B_1 \subseteq B_2 \subseteq \ldots$. we have that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ so

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \mathsf{P}(B_n) \leq \lim_{n\to\infty} \sum_{i=1}^{n} \mathsf{P}(A_i) = \sum_{i=1}^{\infty} \mathsf{P}(A_i)$$

□

Think about $\mathsf{P}(\bigcap_{i=1}^{\infty} A_i) \geq 1 - \sum_{i=1}^{\infty} \mathsf{P}(A_i)$ and remark that Boole's inequality even hold for ∞.

Definition 1.10 (Limits of sets)

Let $\{A_i\}$ be an arbitrary sequence of events in $\mathcal{A}$.

1. $\liminf\limits_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i = \{\omega : \omega \in A_i \; \forall \; i \text{ but finitely many}\};$
2. $\limsup\limits_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i = \{\omega : \omega \in A_i \text{ for } \infty \text{ many } i\text{'s}\}.$

For sure, we have $\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n$[4]. Furthermore, we get that

$$\left(\liminf_{n\to\infty} A_n\right)^{\complement} = \left(\bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i\right)^{\complement} = \bigcap_{n=1}^{\infty} \left(\bigcap_{i=n}^{\infty} A_i\right)^{\complement} = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i{}^{\complement} = \limsup_{n\to\infty} A_n{}^{\complement}.$$

If $\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n$, then

$$\lim_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n.$$

[4] This idea of lim inf and lim sup refer to the same concept as in analysis, but we are dealing here with sets, not numbers.

Example 1.7

Let $\Omega = \mathbb{R}, \mathcal{A} = \mathbb{B}$ and $A_i = \{(-1)^i\}$.
With this specific example,

$$\liminf_{n\to\infty} A_n = \emptyset \text{ and } \limsup_{n\to\infty} A_n = \{-1, 1\}.$$

We can also look at our previous example of A_n a monotonically increasing sequence, that is $\{A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots\}$. Then

$$\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n.$$

To see this, note that the $\liminf$ is the union of A_n since A_n is always the smallest set in the intersection. For the $\limsup$, we are over and over again intersecting the same set since the union is just the biggest of the elements in the set.

Similarly, we can get for the decreasing sequence of A_n that

$$\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n.$$

It holds that if $\lim_{n\to\infty} A_n$ exists, $\mathsf{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathsf{P}(A_n)$. In general, $\mathsf{P}(\liminf_{n\to\infty} A_n)$ and $\mathsf{P}(\limsup_{n\to\infty} A_n)$ can be computed in some cases. These probabilities are the subject of the so-called 0-1 laws or the lemmas due to Borel and Cantelli.

Chapter 2
Conditional probability

Suppose that we are playing a game (of chance) with the following specifications: 2 coins are flipped independently. If the 2 coins show the same symbol (with probability $\frac{1}{2}$), then player α wins. If the 2 coins show different symbols, then player β wins. This is a fair game. Suppose now that a third player flip the coins and tell us one of the coin showed "tail". Then, it is no longer a fair game. Let us formalize the above.

Let $\Omega = \{(1,1),(0,1),(1,0),(0,0)\}$ and $\mathcal{P}(\Omega)$, $\mathsf{P}((i,j)) = 1/2$ with $i = 0,1; j = 0,1$. With no information, $\mathsf{P}(A) = \mathsf{P}(\{(1,1),(0,0)\}) = 1/2$ and $\mathsf{P}(A)^* = 1/3$. Note that if we are told that the first coin was "tail", then the game is still fair.

Definition 2.1 (Conditional probability)

Let $(\Omega, \mathcal{A}, \mathsf{P})$, B be an event such that $\mathsf{P}(B) > 0$. Then, the conditional probability of (an event) A given B is

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

Example 2.1

$B = \{(1,1),(0,1),(1,0)\}, \mathsf{P}(B) = 3/4$. Then

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{3/4} = \frac{\mathsf{P}(\{1,1\})}{3/4} = \frac{1/4}{3/4} = \frac{1}{3}$$

Example 2.2

Suppose that a teacher carries around an umbrella to go to school. He visits 4 shops before going to his office. The probability that he forgets it in some location A_i is $\mathsf{P}(A_i) = 1/5, i = 1,2,3,4$. B is the event that the umbrella is left in one of the stores. $B = A_1 \cup A_2 \cup A_3 \cup A_4$ and $\mathsf{P}(B) = 4/5$, therefore

$$\mathsf{P}(A_1|B) = \frac{\mathsf{P}(A_1 \cap B)}{4/5} = \frac{1/5}{4/5} = \frac{1}{4}.$$

Additional information changes the sample space and also affect the probability measure. We can make it more precise by observing the following.

Theorem 2.2

Let $(\Omega, \mathcal{A}, \mathsf{P})$ and $B \in \mathcal{A}$, $\mathsf{P}(B) > 0$. Then $\mathsf{P}_B : \mathcal{A} \to \mathbb{R}$ is given by $\mathsf{P}_B(A) = \mathsf{P}(A|B)$. This new mapping is again a probability measure on $(\Omega, \mathcal{A})$.

Proof

1. Consider the probability of Ω:
$$\mathsf{P}_B(\Omega) = \frac{\mathsf{P}(\Omega \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(B)}{\mathsf{P}(B)} = 1;$$
2. $\mathsf{P}_B(A) \geq 0$ since both $\mathsf{P}(A \cap B)$ and $\mathsf{P}(B) \geq 0$;
3. Let $A_1, A_2, A_3, \ldots$ be pairwise disjoint. Then by using the definition of conditional probability and the fact that the A_n are disjoint, hence that the union of $A_i \cap B$ is disjoint:
$$\begin{aligned}\mathsf{P}_B\left(\bigcup_{n=1}^{\infty} A_n\right) &= \frac{\mathsf{P}\left(\{\bigcup_{n=1}^{\infty} A_n\} \cap B\right)}{\mathsf{P}(B)} = \frac{\mathsf{P}\left(\bigcup_{n=1}^{\infty}(A_n \cap B)\right)}{\mathsf{P}(B)} \\ &= \sum_{n=1}^{\infty} \frac{\mathsf{P}\left(A_n \cap B\right)}{\mathsf{P}(B)} = \sum_{n=1}^{\infty} \mathsf{P}(A_n|B) = \sum_{n=1}^{\infty} \mathsf{P}_B(A_n)\end{aligned}$$

□

Remark

Note that if $A \cap B = \emptyset$, then
$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(\emptyset)}{\mathsf{P}(B)} = 0$$

Therefore, the “new” sample space can be B. If $\mathcal{A}_B = \{A \cap B, A \in \mathcal{A}\}$, then $\mathcal{A}_B$ is a σ-field and for $\mathsf{P}_B : \mathcal{A}_B \to \mathbb{R}$, then $(B, \mathcal{A}_B, \mathsf{P}_B)$ is a probability space.

Theorem 2.3

Let $(\Omega, \mathcal{A}, \mathsf{P})$ and $A_1, \ldots, A_n$ events such that $\mathsf{P}(\bigcap_{i=1}^{n} A_i) > 0$. Then

$$\mathsf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathsf{P}(A_1) \cdot \mathsf{P}(A_2|A_1) \cdot \mathsf{P}(A_3|A_1 \cap A_2) \cdots \mathsf{P}(A_n|A_1 \cap \ldots \cap A_{n-1})$$

Proof

$$\mathsf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathsf{P}(A_1) \cdot \frac{\mathsf{P}(A_1 \cap A_2)}{\mathsf{P}(A_1)} \cdot \frac{\mathsf{P}(A_1 \cap A_2 \cap A_3)}{\mathsf{P}(A_1 \cap A_2)} \cdots$$
$$\frac{\mathsf{P}(A_{n-1} \cap A_{n-2} \ldots \cap A_1)}{\mathsf{P}(A_{n-2} \cap A_{n-3} \ldots \cap A_1)} \cdot \frac{\mathsf{P}(A_n \cap A_{n-1} \ldots \cap A_1)}{\mathsf{P}(A_{n-1} \cap A_{n-2} \ldots \cap A_1)}$$

Cancel all but the last numerator to get the result. □

Example 2.3

Consider the following problem involving a random walk problem. Suppose that a coin is flipped and you get -1 if tail, 1, if the result is head. Note that not every sequence is equally likely, that is certain path are impossible. For example, the path (1, -2,1) is not possible.

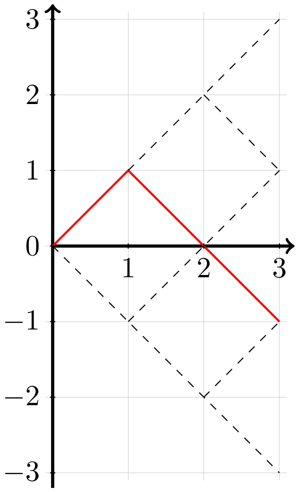


Figure 3: Random walk with a coin toss

We have $\Omega = \{-1, 1\} \times \{-2, 0, 2\} \times \{-3, -1, 1, 3\}$ and suppose as illustrated in red that $(A_1, A_2, A_3) = \{(1, 0, -1)\}$. Then

$$\begin{aligned} A_1 &= \{(t_1, t_2, t_3), t_1 = 1\}, \\ A_2 &= \{(t_1, t_2, t_3), t_2 = 0\}, \\ A_3 &= \{(t_1, t_2, t_3), t_3 = -1\} \end{aligned}$$

and so the probability of this path is

$$\mathsf{P}(A_1 \cap A_2 \cap A_3) = \mathsf{P}(A_1) \cdot \mathsf{P}(A_2 | A_1) \cdot \mathsf{P}(A_3 | A_1 \cap A_2)^{5} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

The previous theorem is of use in stochastic processes and leads to conditional σ-field.

Theorem 2.4 (Law of total probability)

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space, $\{B_i\}$ a collection of pairwise disjoint sets with $\mathsf{P}(B_i) > 0$ such that $\bigcup_{i=1}^{\infty} B_i = \Omega$. In other words, $\{B_i\}$ is a partition of Ω. Then, for any $A \in \mathcal{A}_i$,

$$\mathsf{P}(A) = \sum_{i=1}^{\infty} \mathsf{P}(A|B_i)\mathsf{P}(B_i) = \sum_{i=1}^{\infty} \mathsf{P}(A \cap B_i)$$

Proof Starting from the end terms and going backward, we have

$$\sum_{i=1}^{\infty} \frac{\mathsf{P}(A \cap B_i)}{\mathsf{P}(B_i)} \mathsf{P}(B_i) = \sum_{i=1}^{\infty} \mathsf{P}(A \cap B_i)$$

then by reversing axiom (3) of our definition of probability and after noting that A is in the union for all terms, we get

$$= \mathsf{P}\left(\bigcup_{i=1}^{\infty} (A \cap B_i)\right) = \mathsf{P}\left(A \cap \bigcup_{i=1}^{\infty} B_i\right) = \mathsf{P}(A \cap \Omega) = \mathsf{P}(A)$$

□

[5] Note that $\mathsf{P}(A_3 | A_1 \cap A_2) = \mathsf{P}(A_3 | A_2)$. This is called Markov properties of random chains.

Example 2.4

Consider a random experiment where white and black balls are placed in 3 urns. The following assets for the urn content: $U = (1w, 2b)$, $V = (2w, 1b)$, $W = (3w, 3b)$. Toss a dice: if $\{1, 2, 3\}$, you choose urn $U, \{4\}$ choose V and else if $\{5, 6\}$ choose W. Define A as $A =$"Ball is white" and U, V, W as choosing respectively urns U, V, W. We have the following probabilities:

$\mathsf{P}(U) = 1/2,\ \mathsf{P}(V) = 1/6,\ \mathsf{P}(W) = 1/3$ and $U \cup V \cup W = \Omega,\ U \cap V \cap W = \emptyset$ so we given this partition, we have the following

$$\mathsf{P}(A) = \mathsf{P}(A|U)\mathsf{P}(U) + \mathsf{P}(A|V)\mathsf{P}(V) + \mathsf{P}(A|W)\mathsf{P}(W)$$

$$= \frac{1}{3}\left(\frac{1}{2}\right) + \frac{2}{3}\left(\frac{1}{6}\right) + \frac{1}{2}\left(\frac{1}{3}\right) = \frac{4}{9}$$

Here is an extremely useful corollary of the *Law of total probability*

Theorem 2.5 (Bayes' rule)

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space, $\{B_i\}$ a collection of pairwise disjoint sets with $\mathsf{P}(B_i) > 0$ such that $\Omega = \bigcup_{i=1}^{\infty} B_i$ and $\mathsf{P}(B_i) > 0 \quad \forall \quad i = 1, 2, \ldots$ Then, if $\mathsf{P}(A) > 0$,

$$\mathsf{P}(B_i|A) = \frac{\mathsf{P}(A|B_i)\mathsf{P}(B_i)}{\sum_{j=1}^{\infty} \mathsf{P}(A|B_j)\mathsf{P}(B_j)}.$$

Proof We use the definition of conditional probability twice here and also the *Law of total probability* to get the result

$$\mathsf{P}(B_i|A) = \frac{\mathsf{P}(A \cap B_i)}{\mathsf{P}(A)} = \frac{\mathsf{P}(A|B_i)\mathsf{P}(B_i)}{\sum_{j=1}^{\infty} \mathsf{P}(A|B_j)\mathsf{P}(B_j)}$$

□

Example 2.5 (Paradox of rare disease)

Here are the results of a study performed in Hamburg between 1975 to 1977 about breast cancer in women aged 30 to 40. The study has found that the probability of getting breast cancer is $p = 64/124787 \approx 0,000513$. Let A be the event that a woman aged 30-40 has breast cancer. We denote $\mathsf{P}(A) = p$ and $\mathsf{P}(A^{\complement}) = 1 - p$. The test to test for breast cancer gives event B if the test is positive. The following probabilities have been computed

$$\mathsf{P}(B|A) = 0.99 \quad \mathsf{P}(B^{\complement}|A^{\complement}) = 0.95 \text{ and } \mathsf{P}(B|A^{\complement}) = 1 - \mathsf{P}(B^{\complement}|A^{\complement}) = 0.05$$

Then, by Bayes' rule, the probability of getting positive result and having cancer is

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(B|A)\mathsf{P}(A)}{\mathsf{P}(B|A)\mathsf{P}(A) + \mathsf{P}(B|A^{\complement})\mathsf{P}(A^{\complement})}$$

$$= \frac{0.99p}{0.99p + 0.05(1-p)} \approx 0.01$$

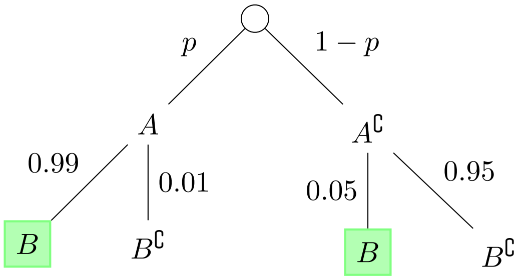


Figure 4: Bayes' rule: considering the situation with a tree

Example 2.6 (Genetic disorder: Haemophilia)

Determine the probability of an unknown woman having the gene for H based on the number of her non-haemophiliac sons (in fact, the gene could be recessive in her sons). Consider first the case of her having one boy.

The event of allele being recessive is denoted by $\bar{H}, \bar{B}$. The two probability trees show the relevant information. Using Bayes' rule,

$$\mathsf{P}(B|\bar{H}) = \frac{(1/2)(1/2)}{(1/2)(1/2) + (1/2)(1)} = 1/3.$$

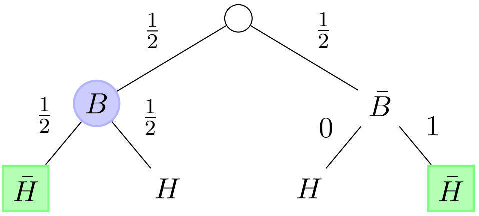


Figure 5: Probability tree for haemophilia with one son

As we will see, having non-haemophiliac sons increase the likeliness of not having the gene. In this case,

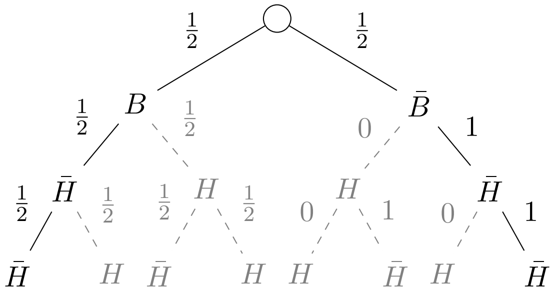


Figure 6: Probability tree for haemophilia with two sons

$$\mathsf{P}(B|\bar{H}^2) = \frac{(1/2)(1/2)(1/2)}{(1/8)+(1/2)} = 1/5.$$

Note that we could have made a smaller tree and spared the big one. If

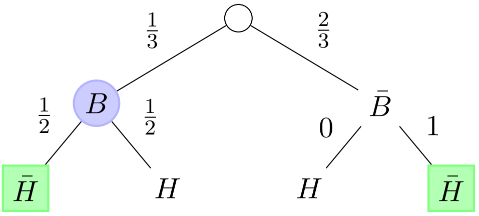
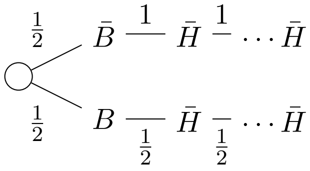


Figure 7: Short probability tree for haemophilia with two sons

we want to find the relation between the number of haemophiliac, we could believe for a minute that it is $\frac{1}{2n+1}$. But considering a larger number of sons (theoretically going to ∞). So the formula is given by



$$\frac{(1/2)(1/2)^n}{(1/2)(1/2)^n + (1/2)} = \frac{1}{1+2^n}$$

All bare down if one son is haemophiliac.

Independence of events

Recall on of the previous examples. $\Omega = \{(1,1),(0,1),(1,0),(0,0)\}$ and $A = \{(1,1),(0,0)\}$, $B = \{(1,1),(0,1),(1,0)\}$ where at least one of the coins show head and $C\{(1,0),(1,1)\}$ if the first coin show heads. Then we have found that $\mathsf{P}(A) = 1/2$ and $\mathsf{P}(A|B) = 1/3$. In contrast, $\mathsf{P}(A|C) = 1/2$. C does not change the probability of A, since it does not provide addition information. This observation leads to

Definition 2.6 (Independent events)

Let $(\Omega, \mathcal{A}, \mathsf{P}), A, B \in \mathcal{A}$. A and B are independent if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$.

Lemma 2.7

If A, B are independent and $\mathsf{P}(B) > 0$, then $\mathsf{P}(A|B) = \mathsf{P}(A)$ (and vice-versa) since $\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}$.

Lemma 2.8

If A and B are independent, then so are A and $B^{\complement}$, $A^{\complement}$ and B, $A^{\complement}$ and $B^{\complement}$.

Proof

$$\mathsf{P}(A^{\complement} \cap B) = \mathsf{P}(B \setminus A) = \mathsf{P}(B) - \mathsf{P}(A \cap B) = \mathsf{P}(B) - \mathsf{P}(A)\mathsf{P}(B)$$

by factoring $\mathsf{P}(B)$, you get $\mathsf{P}(B)\{1 - \mathsf{P}(A)\} = \mathsf{P}(B)\mathsf{P}(A^{\complement})$. The intuition here is that it gives you the same amount of information. □

Note

Independence does not mean empty intersection. They [events] are exclusive so $\mathsf{P}(A \cap B) = 0 \neq \mathsf{P}(A)\mathsf{P}(B)$ unless either $\mathsf{P}(A) = 0$ or $\mathsf{P}(B) = 0$. This will be useful later when we discuss of random vectors.

Definition 2.9

Let $A_i, i \in I$ be a collection of sets in $\mathcal{A}$ (I may be finite, countable or infinitely large). Then

1. $\{A_i, i \in I\}$ are *pairwise independent* if $\forall \; i \neq j; i, j \in I$, it holds that

$$\mathsf{P}(A_i \cap A_j) = \mathsf{P}(A_i)\mathsf{P}(A_j),$$

   that is A_i, A_j are independent.

2. $\{A_i, i \in I\}$ is *independent* if for *any finite* subset $J \subseteq I$, it holds that

$$\mathsf{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathsf{P}(A_i).$$[6]

As for the remark, a similar warning apply to disjoint sets. To see that pairwise disjoint $A_i \cap A_j = \emptyset, i \neq j$ is stronger, note that it implies disjoint, that is $\bigcap_{i \in I} A_i = \emptyset$. For the reverse implication, consider the following counterexample.

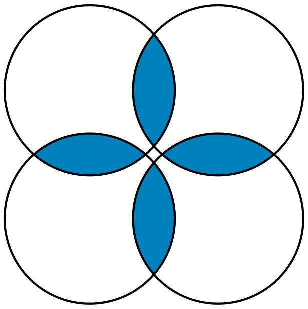


Figure 8: An example of disjoint sets, but not pairwise disjoint.

[6] Here, one should note that (1) does not imply (2). However, independence does imply pairwise independence, as it is a stronger concept

Example 2.7

Consider an urn with 4 marbles, a white, a yellow, a blue and one rainbow. Each can be drawn with probability $p = 1/4$. Denote the events A_ω =“Marble shows colour ω” where $\omega = \{\text{white, yellow, blue}\}$. Then $\mathsf{P}(A_b) = 1/2, \mathsf{P}(A_w) = 1/2, \mathsf{P}(A_y) = 1/2$. If we consider the probability of having both white and yellow on a marble, then $\mathsf{P}(A_w \cap A_y) = 1/4 = \mathsf{P}(A_w)\mathsf{P}(A_y)$ and similarly for all pairs $A_y \cap A_b$ and $A_w \cap A_b$. Therefore, A_w, A_y, A_b are pairwise independent. But if we consider rainbow marble, that is one showing all three colours, we have

$$\mathsf{P}(A_w \cap A_y \cap A_b) = \frac{1}{4} \neq \frac{1}{8} = \mathsf{P}(A_w)\mathsf{P}(A_y)\mathsf{P}(A_b)$$

We have to conclude that pairwise independence does not imply independence. The following other example is illustrative too, although being artificial.

Example 2.8

If I is finite, considering $J = I$ is not enough. We could conclude from this observation that getting the condition for all may well be luck. Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ with $\mathcal{P}(\Omega)$ be our sample space. We have

$$\begin{aligned}
&\mathsf{P}(\{\omega_1\}) = \sqrt{2}/2 - 1/4\\
&\mathsf{P}(\{\omega_2\}) = \mathsf{P}(\{\omega_4\}) = 1/4\\
&\mathsf{P}(\{\omega_3\}) = 3/4 - \sqrt{2}/2.
\end{aligned}$$

Consider now the following events:

$E_1 = \{\omega_1, \omega_3\}, E_2 = \{\omega_2, \omega_3\}, E_3 = \{\omega_3, \omega_4\}$. Now if we look at the probability of all events, we get

$$\begin{aligned}
\mathsf{P}(E_1 \cap E_2 \cap E_3) = \mathsf{P}(\{\omega_3\}) &= \frac{3}{4} - \frac{\sqrt{2}}{2}\\
&= \left(\frac{1}{2}\right)\left(1 - \frac{\sqrt{2}}{2}\right)\left(1 - \frac{\sqrt{2}}{2}\right) = \mathsf{P}(E_1)\mathsf{P}(E_2)\mathsf{P}(E_3)
\end{aligned}$$

But if we look at pairs

$$\mathsf{P}(E_1 \cap E_2) = 3/4 - \sqrt{2}/2 \neq 1/2 - \sqrt{2}/4 = \mathsf{P}(E_1)\mathsf{P}(E_2).$$

It is therefore important to look at all possible combinations before claiming independence.

Chapter 3
Laplace experiments

We recall the equally likely events from the introduction.

Definition 3.1

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space such that

- $\Omega = \{\omega_1, \ldots, \omega_n\}$; Ω is finite
- $\mathcal{P}(\Omega)$
- $\mathsf{P}(\{\omega_i\}) = 1/n, i = 1, \ldots, n$

Then $(\Omega, \mathcal{A}, \mathsf{P})$ is called a Laplace experiment and P is called a *Laplace distribution.*

$$\mathsf{P}(A) = \sum_{i:\omega_i \in A} \frac{1}{n} = \frac{|A|}{n} = \frac{\#\text{ favorable outcomes}}{\#\text{ all outcomes}}.$$

Example 3.1 (Chevalier de Méré, (1607-1685))

Suppose that you are playing with three dices and you want to look at the probabilities that 3 numbers sum to 11 and to 12. de Méré noticed that the probability of getting a sum of 11 seemed to be higher than just playing 12, which contradicted Laplace's definition of equally likely events. In this example, $\Omega = ((\omega_1, \omega_2, \omega_3), \omega_i \in \{1, 2, \ldots, 6\}, i = 1, 2, 3)$ and we have $\mathsf{P}(\{(\omega_1, \omega_2, \omega_3)\}) = 1/6^3$ and two events, namely A={sum is 11} and B={sum is 12}. Let's look at the possibilities, examining also the number of permutations, denoted $\#A_i$ and $\#B_i$. It is now easily seen, as Poincarré pointed out to his friend that not all the combinations have the same number of permutations, and in a sense his observation was right as

$$\mathsf{P}(A) = 27/6^3 \approx 0,125 > 0,116 \approx 25/6^3.$$

Lemma 3.2 (The multiplicative rule)

Let $(\omega_1, \ldots, \omega_k)$ be an *ordered* k-tuple (called k-permutation) and suppose

A_i	$\#A_i$	B_i	$\#B_i$
1+4+6	6	1+5+6	6
1+5+5	3	2+4+6	6
2+3+6	6	2+5+5	3
2+4+5	6	3+3+6	3
3+3+5	3	3+4+5	6
3+4+4	3	4+4+4	1

Figure 9: Summing three dices to get 11 or 12

that $\omega_1 \in A_1$ and $|A_i| = n_1, \omega_2 \in A_2, \ldots, |A_2| = n_2, \ldots, \omega_k \in A_k, \ldots, |A_k| = n_k$. Then, there are exactly $\prod_{i=1}^{k} n_i$ such k-permutations.

Example 3.2

Consider $(\omega_1, \ldots, \omega_k)$ with $\omega_i \in \{1, \ldots, k\}$ and $\omega_i \neq \omega_j$. There are $k!$ such permutations. Here, $\Omega = \{\omega_1, \ldots, \omega_n\}, |\mathcal{P}(\Omega)| = 2^n$. To see this, just consider the binary representation in a table.

Example 3.3

Consider an urn with N numbered (distinguishable) marbles $1, \ldots, N$ and of which k marbles with replacement (drawn, then placed back).

We have $\Omega = \{(\omega_1, \ldots, \omega_k), \omega_i \in \{1, \ldots, N\}\}$. Then $|\Omega| = N^k$

Example 3.4

Urn as above, k marbles are drawn without replacement.
$\Omega = \{(\omega_1, \ldots, \omega_k), \omega_i \in \{1, \ldots, N\}, \omega_i \neq \omega_j$ for $i \neq j\}$. Then

$$|\Omega| = N(N-1)\cdots(N-(k-1)) = \frac{N!}{(N-k)!}$$

Example 3.5 (Birthday match)

Here $\Omega = \{(i_1, \ldots, i_n), i_j \in \{1, \ldots, 365\}\}$ and $|\Omega| = (365)^n$. We are interested in the probability of having at least one overlap for a given group. Let A be “there is no overlap” (that is, we look at the complement of having at least one overlap). Then for $i_j \neq i_k$:

$$\mathsf{P}(A) = \frac{365 \cdot 364 \cdot 363 \cdots (365-n+1)}{365^n} = \frac{\frac{365!}{(365-n)!}}{365^n}$$

Then for $n \approx 20, \mathsf{P}(A) \approx 0.589$ so overlap happens close to 41% ot the time. Similarly, if $n = 30$, we have $\mathsf{P}(A) \approx 0.32$ implying $\mathsf{P}(A^{\complement}) \approx 0.68$. By enlarging the number of people in the sample to only 60 people, we have $\mathsf{P}(A^{\complement}) \approx 0.99412$.

Lemma 3.3 (k-combination)

A k-combination is a sequence of k-numbers in which the ordering is irrelevant. There are exactly $\binom{n}{k}$ k-combinations out of n elements. Here $\binom{n}{k}$ is called the binomial coefficient and is equal to

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Example 3.6 (Crazy cloakroom lady)

Suppose a lady in charge of the cloakroom decides to mix the numbers corresponding to the clothes. Than, for number $1, \ldots, n$, we are interested in the probability of being lucky and getting a tie, that is the event A_i that there is a fixed point of the permutation ($\pi_i = i$).

The probability of getting one fixed point is equal to $\mathsf{P}(\bigcup_{i=1}^{n} A_i)$ and $\Omega =$the set of all permutations $1, \ldots, n$. The probability $\mathsf{P}(A_i) = (n-1)!/n!$ and $\mathsf{P}(A_i \cap A_j) = (n-2)!/n!$ so by the Inclusion-exclusion principle, we have $\mathsf{P}\left(\bigcup_{i=1}^{n} A_n\right)$ is equal to

$$\begin{aligned}
&= \sum_{i=1}^{n} \mathsf{P}(A_i) - \sum_{i<j} \mathsf{P}(A_i \cap A_j) + \sum_{i<j<k} \mathsf{P}(A_i \cap A_j \cap A_k) \\
&\qquad - \cdots + (-1)^{n+1} \mathsf{P}(A_1 \cap \ldots \cap A_n) \\
&= n\frac{(n-1)!}{n!} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} + \cdots + \frac{(-1)^{n+1}}{n!} \\
&= 1 - \frac{1}{2!} + \frac{1}{3!} + \cdots + (-1)^{n+1}\frac{1}{n!} \\
&= \sum_{i=1}^{n} (-1)^{i+1} \frac{1}{i!}
\end{aligned}$$

It may seem paradoxical that as $n \to \infty$, the probability of getting back a coat does not go to zero. Indeed, P(no fixed point) equals

$$1-\mathsf{P}\left(\bigcup_{i=1}^{n} A_n\right) = 1-\sum_{i=1}^{n}(-1)^{i+1}\frac{1}{i!} = \sum_{i=0}^{n}(-1)^{i}\frac{1}{i!}$$

As $n \to \infty$ the probability

$$\mathsf{P} \to \sum_{i=0}^{\infty}\frac{1}{i!} = \frac{1}{e} \approx 0.368$$

Remark (About the binomial coefficient)

We have the following properties for the Pascal triangle

1. $\binom{n}{0} = \binom{n}{n}$;
2. $\binom{n}{1} = n$;
3. $\binom{n}{k} = \binom{n}{n-k}$ (symmetric);
4. $\binom{n-1}{k-1} + \binom{n-1}{k} = \binom{n}{k}$;
5. Binomial theorem :

$$(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}$$

and in particular $\sum_{k=0}^{n}\binom{n}{k} = 2^n$.

Note that $\binom{n}{k}$ is also the number of k-combinations drawn out of n balls *without replacement*. There are $\binom{n+k-1}{k}$ k-combinations *with replacement*. To see this, let us have a look at the table where the different draws are expressed as follow: each draw has zero corresponding to the number of 0 not distinguishable and 1 not distinguishable. We have $n-1$ ones and k zeros. We get the expression

$$\underbrace{000}_{\text{draw 1}}\ 1\ \underbrace{00}_{\text{draw 2}}\ 1\ \underbrace{0}_{\text{draw 3}}\ 1\ \underbrace{0000}_{\text{draw 4}} \rightsquigarrow \frac{(n-1+k)!}{k!(n-1)!}$$

$$\begin{array}{ccccccccc} & & & & \binom{0}{0} & & & & \\ & & & \binom{1}{0} & & \binom{1}{1} & & & \\ & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & \\ & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & \\ \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} \end{array}$$

Figure 10: Illustration of the Pascal triangle

Lemma 3.4

A set of size n can be partitioned into k categories (subsets) of size $n_1, \ldots n_k$ (where $n_1 + \ldots + n_k = n$) in exactly $\frac{n!}{n_1! \ldots n_k!}$ ways, or equally $\binom{n}{n_1 \ldots n_k}$. Here, $\binom{n}{n_1 \ldots n_k}$ is called the multinomial coefficient.

Proof

$$\binom{n}{n_1} \cdot \binom{n-n_1}{n_2} \cdot \binom{n-n_1-n_2}{n_3} \cdots \binom{n-n_1-\cdots-n_{k-1}}{n_k}$$

are the quantities and what can placed into each bin. By cancelling the terms, we get

$$\frac{n!}{n_1!(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \cdots \frac{n_k}{n_k!0!} = \frac{n!}{n! \cdots n_k!}$$

□

Lemma 3.5 (Multinomial theorem)

$$(x_1 + \cdots + x_k)^n = \sum_{\substack{n_1, \ldots, n_k \geq 0 \\ n_1 + \cdots + n_k = n}} \frac{n!}{n_1! \cdots n_k!} x_1^{n_1} \cdots x_k^{n_k}$$

of which the binomial theorem is a special case.

Chapter 4

Discrete probability measures

Definition 4.1 (Discrete probability measure)

Let $S \subseteq \Omega, S$ countable (S is finite or countably infinite) and let $p_s \in [0,\infty), s \in S$ be a collection of numbers such that $\sum_{s\in S} p_s = 1$.

Then $\mathsf{P}\colon \mathcal{P}(\Omega) \to [0,1]$ such that

$$\mathsf{P}(A) = \sum_{s\in S\cap A} p_s = \sum_{s\in S} p_s 1_A(s)$$

is a probability measure on $(\Omega, \mathcal{P}(\Omega))$. It is called a discrete[7] probability measure with *support* S.

The function $F : \Omega \to \mathbb{R}$ given by

$$f(\omega) = \begin{cases} 0 & \omega \notin S \\ p_s & \omega = s \in S \end{cases}$$

is called the *probability mass function* or PMF (density of P with respect to the counting measure).

Remark

- $p_s = \mathsf{P}(\{s\}), s \in S$;
- $\Omega = S$
- The PMF defines P uniquely

If $f : \Omega \to [0,\infty)$ is such that $S = \{\omega \in \Omega : f(\omega) > 0\}$ is countable and $\sum_{\omega\in S} f(\omega) = 1$, then f defines a probability measure on $\Omega, \mathcal{P}(\Omega))$ by $\mathsf{P} : \mathcal{P}(\Omega) \to [0,1]$ and $A = \sum_{\omega\in S} f(\omega) 1_A(\omega)$ (that is, we get an injective function from the PMF to the probability measure).

Notation

A discrete probability measure if often called a *discrete distribution*

[7] Refers to the countability of the support

Example 4.1 (Dirac distribution)

$\Omega, \omega_o \in \Omega$ fixed and $f : \Omega \to [0,1]$ and

$$\omega \to \begin{cases} 0 & \text{if } \omega \neq \omega_o \\ 1 & \text{if } \omega = \omega_o \end{cases}$$

The support of f is $\{\omega_o\}$. The corresponding probability measure is such that

$$\mathsf{P}(A) = \begin{cases} 0 & \text{if } \omega \notin A \\ 1 & \text{if } \omega \in A \end{cases}$$

The Dirac distribution describes a deterministic experiment (sure outcome). It is also denoted by $\mathcal{E}_{\omega_o}$.

Example 4.2 (Discrete uniform distribution)

$\Omega = \{\omega_1, \ldots \omega_n\},\ f(\omega_i) = 1/n$ where $i = 1, \ldots, n$.

Example 4.3 (Bernoulli distribution)

$\Omega = \{0, 1\}$, $f(0) = 1 - p$ and $f(1) = p$ for $p \in [0, 1]$.

The Bernoulli distribution is denoted $\mathcal{B}(p)$ or $\mathcal{B}(1, p)$. It describes the outcome of a "0,1" experiment, where 0="failure" and 1="success". Examples include the coin toss, gender, presence or absence of some characteristic. If we were to be interested in a sequence of Bernoulli trials, then we would look at

Example 4.4 (Binomial distribution)

$\Omega = \{0, \ldots, n\} = S$ and

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 1, \ldots, n$. Indeed, $\sum_{k=0}^{n} f(k) = 1$ (by the binomial theorem). The binomial distribution is denoted $\mathcal{B}(n, p)$. It describes the number of successes in n independent Bernoulli trials.

We have $\Omega^* = \{(\omega_1, \ldots, \omega_n), \omega_i \in \{0, 1\}\}$ where $A_k =$ successes out of $n = \{(\omega_1, \ldots, \omega_n) \in \Omega^*, \sum_{i=1}^{n} \omega_i = k\}$. $\mathsf{P}(\{1, \ldots, 1, 0, \ldots 0)\} = p^k (1-p)^{n-k}$

which is equal to $\left(\frac{1}{2}\right)^n$ if $p = \frac{1}{2}$ and similarly for all $(\omega_1, \ldots, \omega_n)$. with k 1's and $(n-k)$ 0's with $\mathsf{P}(A_k) = \binom{n}{k} p^k (1-p)^{n-k}$. Here is an example. Given that a mother carries the gene for haemophilia, suppose she has n sons. The probability that exactly k of them are haemophiliac.

$$\binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k}$$

Another example arises from quality control. In Hamburg harbour, shipments of coffee arrive and are stored in gigantic storage house. Bags are chosen at random and quality of the shipment is evaluated by digging the bag with a small spade to retrieve 50 beans and considering each of them. This procedure allows to asset the quality of the shipment.

Alternatively, consider quality-control in a manufacture of light bulbs. Each hour, 10 bulbs are selected ar random and tested. If at least one is defective bulb is discovered, the production stops. Suppose the probability of a light bulb being defective is 2% Then having no defective in the observation happens with probability $\binom{10}{0}(0.02)^0(0.98)^{10} \approx 0.82$.

Example 4.5 (Hypergeometric distribution)

Let $N \ldots$ size of the population, $K \ldots$ defectives in the population, $n \ldots$ size of the sub-population (sample) from the entire population and $\Omega = \{0, \ldots n\}$ and

$$f(k) = \begin{cases} 0 & k > \min(n,k), k < \max(0, K+n-N) \\ \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} & \text{otherwise.} \end{cases}$$

If $N \to \infty, K \to \infty$ such that $K/N \to p \in [0,1]$, then if n is fixed

$$f(k) \approx \binom{n}{k} p^k (1-p)^{n-k}.$$

Let us look closer at the above. We have

$$f(k) = \frac{\dfrac{K!}{k!(K-k)!} \quad \dfrac{(N-K)!}{(n-k)!(N-K-n-k)!}}{\dfrac{N!}{n!(N-n)!}}$$

and for the denominator expanding yields

$$\binom{n}{k}\frac{K(K-1)\dots(K-k+1)}{N(N-1)\dots(N-k+1)}$$

and for the denominator

$$\frac{(N-K)(N-K-1)\dots(N-K-n-k+1)}{(N-k)\dots(N-n+1)}$$

which tends to our starting result.

Example 4.6 (Poisson distribution)

The Poisson distribution (denoted $\mathcal{P}(\lambda)$ has a countable support $\Omega = \{0, 1, \dots\}$ with

$$f(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad \lambda > 0$$

which is clearly non-negative so

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda}$$

is a Taylor series that converges to $e^{\lambda}e^{-\lambda}$ so this is a valid probability. We could see the Poisson as the limit of the binomial, just as the binomial can be seen to be a limit of the hypergeometric. Our motivation here is that $n \to \infty, p \to 0$ such that $np \to \lambda > 0$. Then,

$$\binom{n}{k}p^k(1-p)^{n-k} = \frac{n(n-1)\dots(n-k+1)}{k!n^k}(np)^k\left(1-\frac{np}{n}\right)^n\left(1-\frac{np}{n}\right)^{-k}$$

and as n tends to infinity and np to λ, we get $(k!)^{-1}\lambda^k(e^{-\lambda})$.

The Poisson distribution describes a distribution of rare events. For example, the number of gene that mutate due to radiation or the probability for each car of getting an accident today in the city of Montreal. An interesting fact about the historic of the Poisson distribution is that it served originally to measure the death of soldiers due to horse kick in the Prussian army.

Example 4.7 (Geometric distribution)

Has support $\Omega = \{0, 1, \ldots\}$ with

$$f(k) = p^k(1-p)$$

where $p \in (0,1)$. Non-negativity is obvious since the sum gives the geometric series

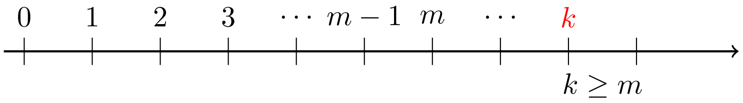
$$\sum_{k=0}^{\infty} f(k) = \left(\sum_{k=0}^{\infty} p^k\right)(1-p) = 1$$

We can think of the geometric distribution as an infinite sequence of independent trials, for example in a coin toss experiment the number of heads before seeing a tail. The Geometric distribution describes the number of failures until the first success.[8]

The Geometric distribution arises in fields like extreme-value theory. One could possibly study the dikes in the Netherlands (Delta Works) and wonder how high the dikes should be. Consider the heights given annually from the data sample and cut the data into yearly partition, looking at the maximum for each year. with $p \ldots$ maximum yearly sea height $\leq H$ and $n - p \ldots$ the max sea height $> H$. We are interested in the number of years until the first spillover. This event has geometric distribution supposing yearly flow. Now, reverse the question: if we want to have spillovers only every say 125,000 years, what height should the dikes be? If we extend this concept to get r spillovers, we then get the negative binomial distribution, but before proceeding we look at a feature of the geometric distribution, the
Lack of memory property.
The info that nothing happened before $m-1$ does not affect the probability (in our example, the maxima would have to be independent). Therefore,

[8] There is discordance in the literature: sometimes, the support is given to be $\{1, \ldots\}$ and the distribution is defined to be the number of trial before the first success

Figure 11: Lack of memory property



the probability $\mathsf{P}(\{k\}|\{m, m-1, \dots\})$

$$= \frac{p^k(1-p)}{\sum_{i=m}^{\infty} p^i(1-p)} = \frac{p^k(1-p)}{p^m \sum_{i=0}^{\infty} p^i(1-p)} = \frac{p^k(1-p)}{p^m} = p^{k-m}(1-p)$$

remains unchanged.

Example 4.8 (Negative binomial distribution)

Consider our Bernoulli trials: 000011001011. We are waiting for the r success with $\Omega = \{0, 1, \dots\}, r \geq 1$. Look at the number k of failures before the r success

$$f(k) = \binom{k+r-1}{k} p^k(1-p)^r$$

where $p \in (0, 1)$. This distribution is linked to the Poisson distribution: to see this, consider the car accident example, where we assumed the probability of failure to be the same. Not every car is as likely to be involved in an accident. If we assign different probabilities, such that we get a distribution, namely the Negative binomial, that is used mostly in insurance.

Chapter 5

Probability measures on $(\mathbb{R}, \mathbb{B})$

Remark

Any discrete probability measure discussed so far (Bernoulli, binomial, hypergeometric, Poisson, geometric, NB) can be viewed as a probability measure on $(\mathbb{R}, \mathbb{B})$. For example, the Poisson distribution on $(\mathbb{R}, \mathbb{B})$ has support $S = \{0, 1, \ldots\}$ and PMF

$$f(k) = \begin{cases} \dfrac{\lambda^k}{k!} e^{-\lambda} & x \in S \\ 0 & x \notin S \end{cases}$$

just by embedding the distribution in $(\mathbb{R}, \mathbb{B})$.

Example 5.1

Consider the interval $[0, 1]$ and suppose we fix $n \in \mathbb{N}$ large and we look at the discrete uniform distribution on $\{\frac{i}{n}, i \in \{1, \ldots, n\}\}$.

$$\mathsf{P}_n\left(\left\{\frac{i}{n}\right\}\right) = \frac{1}{n}$$

so $\mathsf{P}_n((a, b]),\ 0 \leq a < b \leq 1$

$$= \frac{1}{n}\left|i : a < \frac{i}{n} \leq b\right| = \frac{1}{n}\Big[\lfloor bn \rfloor - \lfloor an \rfloor\Big]$$

is by adding and subtracting the same terms

$$= \frac{1}{n}(bn - an - bn + an + \lfloor bn \rfloor - \lfloor an \rfloor) = b - a + \frac{\varepsilon_n}{n}$$

where $\varepsilon_n = (\lfloor bn \rfloor - bn) + (an - \lfloor an \rfloor)$. Observe that $|\varepsilon_n| \leq 2$ (in fact ≤ 1 by the triangle inequality). Hence, $\lim_{n \to \infty} \mathsf{P}_n((a, b]) = b - a$.

Definition 5.1

A probability distribution function F is any function $F : \mathbb{R} \to \mathbb{R}$ such that

1. F is non-decreasing;

2. F is right-continuous;

3. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$[9]

Example (continued)

$F : \mathbb{R} \to \mathbb{R}$ and choose

$$F(x) = \begin{cases} 0 & x < 0 \\ x & x \in [0,1) \\ 1 & x \geq 1 \end{cases}$$

clearly is an example of distribution function (it is even continuous)

$$\mathsf{P}_n((a,b]) \to b - a = F(b) - F(a) \quad \text{for } a, b.$$

Theorem 5.2

For every distribution function F, there exists a unique probability measure on $(\mathbb{R}, \mathbb{B})$ such that $\mathsf{P}([a,b]) = F(b) - F(a)$.

PROOF The proof would take several class and it is therefore suggested you take *Advanced Probability*. Recall that $\mathbb{B} = \sigma((a,b], a \leq b)$.

The generator $\mathcal{E} = \{(a,b] : a \leq b\}$ is $\cap$-stable. That is

$$E_1, E_2 \in \mathcal{E} \Rightarrow E_1 \cap E_2 \in \mathcal{E}.$$

Nevertheless, here are two important results from measure theory.

- If P is specified on $\mathcal{E}$, then P can be extended to a probability measure on $\sigma(\mathcal{E})$[10]

- If $\mathcal{E}$ is $\cap$-stable and $\mathsf{P}_1, \mathsf{P}_2$ are probability measures on $\sigma(\mathcal{E})$ such that $\mathsf{P}_1(E) = \mathsf{P}_2(E)$ for all $E \in \mathcal{E}$, then $\mathsf{P}_1 = \mathsf{P}_2$(or, in other words, $\mathsf{P}_1(A) = \mathsf{P}_2(A)\ \forall A \in \sigma(\mathcal{E})$).[11]

[9] This means $\lim_{h\to 0, h\geq 0} F(x+h) = F(x)$ for all $x \in \mathbb{R}$.

[10] Under suitable assumptions on $\mathcal{E}$(it need to be a ring of sets)

[11] If you are interested to look further, the book *Measure and integration* by Heinz Bauer is recommended

□

Remark

The theorem 5.2 holds also for functions $G : \mathbb{R} \to \mathbb{R}$ such that (1) & (2) holds (from Definition 5.1). Then, there exists precisely one measure μ_G so that $\mu_G((a,b]) = G(b) - G(a)$. In particular, if $G(x) = x, x \in \mathbb{R}$ then $\mu_G((a,b]) = b - a$ (the identity mapping). μ_G is called a Lebesque measure (denoted λ).

Theorem 5.3

If P is a probability measure on $(\mathbb{R}, \mathbb{B})$, then the function F given by $F(x) = \mathsf{P}((-\infty, x]), x \in \mathbb{R}$ is a distribution function.

Proof Consider $F : \mathbb{R} \to \mathbb{R}$.

1. F increasing, for $x \leq y$ then $(-\infty, x] \subseteq (-\infty, y]$ and

$$F(y) = \mathsf{P}((-\infty,]) \leq \mathsf{P}((-\infty, x]) = F(x)$$

2. F is right-continuous: $h_n, h_n \to 0, h_n \geq 0 \;\forall\; n \in \mathbb{N}$ and w.l.o.g. is monotone. Then $(-\infty, x + h_n] \subseteq (-\infty, x + h_{n+1}] \subseteq \ldots$. We have a decreasing sequence of sets so

$$\bigcap_{n=1}^{\infty} (-\infty, x + h_n] = (-\infty, x]$$

and

$$\begin{aligned} F(x) = \mathsf{P}((-\infty, x]) = \mathsf{P}(\bigcap_{n=1}^{\infty} (-\infty, x + h_n]) \\ = \lim_{n \to \infty} \mathsf{P}((-\infty, x + h_n]) = \lim_{n \to \infty} F(x + h_n) \end{aligned}$$

3. Consider the interval $[-\infty, x_n] \supset [-\infty, x_{n+1}] \supset \ldots$ and without loss of generality x_n decreasing; $x_n \to -\infty$. Then $\cap_{n=1}^{\infty} (-\infty, x_n] = \emptyset$. So

$$\mathsf{P}((-\infty, x_n)) = 0 = \lim_{n \to \infty} \mathsf{P}((-\infty, x_n]) = \lim_{n \to \infty} F(x_n)$$

Then, if we have another sequence $y_n \to \infty$ increasing and $(-\infty, y_n] \subseteq (-\infty, y_{n+1}] \subseteq \ldots$ so $\cup_{n=1}^{\infty}(-\infty, y_n] = \mathbb{R}$ so $\mathsf{P}(\cup_{n=1}^{\infty}(-\infty, y_n]) = 1$ and $\mathsf{P} = \lim_{n\to\infty} F(y_n)$.

□

Corollary 5.4

If F is a distribution function, then there exists a *unique* probability measure on $(\mathbb{R}, \mathbb{B})$ such that $\mathsf{P}((a,b]) = F(b) - F(a)$. Conversely, if P is a probability measure on $(\mathbb{R}, \mathbb{B})$, then $F : \mathbb{R} \to \mathbb{R}$ and $F(x) = \mathsf{P}((-\infty, x])$ is a distribution function and such that $\mathsf{P}((a,b]) = F(b) - F(a)$.

Proof There is a one-to-one correspondence between the Probability measure on $(\mathbb{R}, \mathbb{B})$ and the distribution function: everything is encoded in F. Consider $(a,b] = (-\infty, b] \setminus (-\infty, a]$ for $a \leq b$. We have

$$\mathsf{P}((-\infty, b] \setminus (-\infty, a]) = \mathsf{P}((-\infty, b]) - \mathsf{P}((-\infty, a]) = F(b) - F(a)$$

□

Lemma 5.5 (Properties of F)

Let F be a distribution function of a probability measure P on $(\mathbb{R}, \mathbb{B})$. Then

1. $\mathsf{P}((-\infty, x)) = F(x^-)$
2. $\mathsf{P}(\{x\}) = F(x) - F(x^-)$
3. F is continuous $\Leftrightarrow \mathsf{P}(\{x\}) = 0 \;\; \forall\, x \in \mathbb{R}$
4. $\mathsf{P}((a,b]) = 0 \Leftrightarrow F(b) - F(a)$

Proof We prove (1) and (2), the others follow directly.

(1) $(-\infty, x) = \bigcup_{n=1}^{\infty}(-\infty, x - \frac{1}{n}]$, so we have

$$\mathsf{P}((-\infty, x)) = \lim_{n\to\infty} \mathsf{P}\left(\left(-\infty, x - \frac{1}{n}\right)\right) = \lim_{n\to\infty} F\left(x - \frac{1}{n}\right) = F(x^-)$$

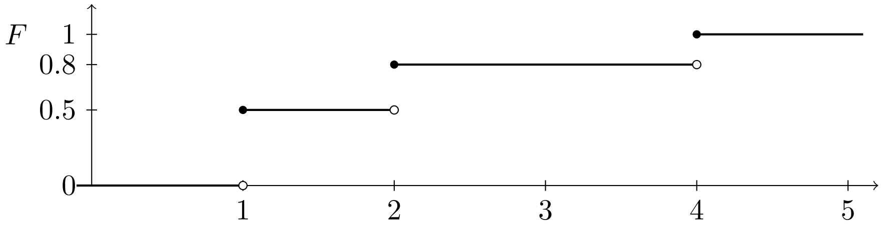


Figure 12: The cadlag function F

(2)

$$\mathsf{P}(\{x\}) = \mathsf{P}((-\infty, x]) - \mathsf{P}((-\infty, x)) = F(x) - F(x^-)$$

□

Remark

F is a cadlag function (i.e. "continue à droite, limite à gauche")

Example 5.2

If P is a discrete probability measure on $(\mathbb{R}, \mathbb{B})$ with support S and PMF f. To observe this, $\mathsf{P}(B) = \sum_{x \in S \cap B} f(x)$.

$$F(x) = \mathsf{P}((-\infty, x]) = \sum_{\substack{s \in S \\ s \leq x}} f(s)$$

where f is the jump function. Here is a concrete example: $S = \{1, 2, 4\}$ with $f(1) = 0.5, f(2) = 0.3$ and $f(4) = 0.2$.

In general

1. F is a jump function;
2. The jumps occur at points in S;
3. The size of a jump at $s \in S$ is $f(s)$.

Remark

A probability measure is called continuous if its distribution function is continuous. In general, F has at most countably many discontinuities (jump). The proof is in the book. [1] p.44

In general,

$$F = p_1 F_1 + p_2 F_2 + p_3 F_3$$

where $p_1, p_2, p_3 \geq 0$ are weight and $p_1 + p_2 + p_3 = 1$ and where F_1 is a DF of a discrete PM, F_2 is continuous, but such that $F_2'(x) = 0$ for almost all $x \in \mathbb{R}$ or equally $\lambda\{x : F_2'(x) = 0\} = 0$. Examples of such functions include the Devil's staircase or the Cantor function. Finally, F_3 is the DF of probability measure that are absolutely continuous. Note that the above is a difficult result (decomposition of measure) from measure theory. We look at the case $p_3 : p_1, p_2 = 0$ in the next section.

Chapter 6

Probability measures on $(\mathbb{R}, \mathbb{B})$ with (Riemann) density

Recall that the function $f : [a, b] \to \mathbb{R}$ is Riemann integrable if and only if

1. f is bounded;
2. f is almost everywhere continuous, i.e.

$$\lambda\{x : f \text{ is discontinuous in } x\} = 0.$$

We make the following observation

Definition 6.1 (($\mathcal{R}$) probability density)

A function $f : \mathbb{R} \to \mathbb{R}$ is called a $\mathcal{R}$ probability density function (($\mathcal{R}$) PDF) if

1. $f(x) \geq 0, x \in \mathbb{R}$;
2. f is $\mathcal{R}$-integrable on $\mathbb{R}$ and $\int_{-\infty}^{\infty} f(t)dt = 1$.

Theorem 6.2 (Continuous distribution function)

Let f be a ($\mathcal{R}$)PDF, then $F : \mathbb{R} \to \mathbb{R}$ given by $F(x) = \int_{-\infty}^{x} f(t)dt$ is a continuous distribution function. Furthermore, $F'(x) = f(x)$ whenever f is continuous at x (the density is the derivative of the distribution function). In addition,

$$\mathsf{P}((a, b]) = F(b) - F(a) = \int_a^b f(t)dt$$

It is easily seen that this is a probability.

Remark

If F has a ($\mathcal{R}$) PDF f, *id est* for all $x \in \mathbb{R}$, $F(x) = \int_{-\infty}^{x} f(t)dt$, then F is continuous, but *not conversely!* However, it can be shown that F is absolutely continuous if it has ($\mathcal{R}$) PDF. If a distribution function is AC, then $\exists f^*$ such that

$$F(x) = \int_{-\infty}^{x} f^* t d\lambda$$

Heuristic interpretation of f

Recall that if P is discrete with PMF f, then $\mathsf{P}(\{x\}) = f(x)$. If P has a ($\mathcal{R}$) density f, then F is continuous and hence $\mathsf{P}(\{x\}) \neq f(x), \mathsf{P}(\{x\}) = 0 \ \forall \ x \in \mathbb{R}$, implying no outcome is favoured. But

$$\mathsf{P}((x-\varepsilon, x+\varepsilon]) = \int_{x-\varepsilon}^{x+\varepsilon} f(t)dt \overset{MVT}{\approx} f(x_\varepsilon) \cdot 2\varepsilon \approx f(x) \cdot 2\varepsilon.$$

Example 6.1 (Uniform distribution)

$\mathcal{U}[a,b]$ on $[a,b] : a \leq b$ on $\mathbb{R}$. The density is given by

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

Indeed, $\int_{-\infty}^{\infty} f(t)dt = \int_a^b \frac{1}{b-a}dt = 1$. The corresponding DF is

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \displaystyle\int_a^x \frac{dt}{b-a} = \frac{x-a}{b-a} & \text{if } x \in [a,b) \\ 1 & \text{if } x \geq b \end{cases}$$

with the special case $a = 0$ and $b = 1$ is called the standard uniform distribution.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \in [0,1) \\ 1 & \text{if } x \geq 1 \end{cases}$$

Example 6.2 (Exponential distribution)

Consider a rate $\lambda > 0$. We say the random variable X has exponential distribution (denoted $X \sim \mathcal{E}$) if the ($\mathcal{R}$) density is

$$f(x) = \lambda e^{-\lambda x} 1_{(x>0)}.$$

given that

$$\int_{-\infty}^{\infty} f(t)dt = \int_0^{\infty} \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_0^{\infty} = 1$$

is clearly non-negative. The corresponding distribution function is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \displaystyle\int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

Suppose that we have the following situation. Waiting for an event of a certain type (insurance claims, earthquake,etc.)

> The probability of exactly one such event will happen in an interval of length h is $\lambda h + o(h)$, where $o(h) = \frac{g(h)}{h} \to 0$ as $h \to 0$.
>
> The probability that there are two or more events occurring in an interval of length h is $o(h)$.
>
> The number of events in disjoint intervals are independent.

We will make it more precise and show it follows a Poisson. Let $N(t) \ldots$ the number of events in $[0, t]$. Chop in n intervals of equal length n and let $n \to \infty$. Then $\mathsf{P}(\{N(t) = k\}) = \mathsf{P}(A) + \mathsf{P}(B)$, where $A \ldots$ there are exactly k intervals $\left[\frac{(i-1)t}{n}, \frac{it}{n}\right]$ with 1 event and $(n-k)$ intervals with no events. $B \ldots (N(t) = k)$ and there exists at least one interval with 2 or more events. Therefore $A \cap B = \emptyset$ and

$$\mathsf{P}(B) \leq \mathsf{P}\left(\bigcup_{i=1}^n B_i\right) \subseteq \sum_{i=1}^n \mathsf{P}(B_i) = \sum_{i=1}^n o\left(\frac{t}{n}\right) = \frac{o\left(\dfrac{t}{n}\right)t}{\left(\dfrac{t}{n}\right)}$$

so $\lim_{n\to\infty} \mathsf{P}(B) = 0$ and B is negligible and we have

$$\mathsf{P}(A) = \binom{n}{k} \left\{\lambda \frac{t}{n} + o\left(\frac{t}{n}\right)\right\}^k \left\{1 - \lambda\left(\frac{t}{n}\right) - o\left(\frac{t}{n}\right)\right\}^{n-k}.$$

$$n\lambda\frac{t}{n} + \frac{o\left(\dfrac{t}{n}\right)t}{\dfrac{t}{n}} \xrightarrow[n\to\infty]{} \lambda t \longrightarrow \lim_{n\to\infty} \mathsf{P}(A) = \frac{1}{k!} e^{-\lambda t} (\lambda t)^k$$

We had derived this goes to Poisson so

$$\mathsf{P}(\{N(t)=k\}) = \frac{1}{k!}e^{-\lambda t}(\lambda t)^k$$

Now, suppose you wait until the first event. If T is the time at which the first event occurred,

$$\mathsf{P}(T \le t) = \mathsf{P}([0,t]) = 1 - \mathsf{P}(T > t) = 1 - (\mathsf{P}(N(t)=0)) = 1 - e^{-\lambda t}$$

for $t > 0$. In other words, the exponential distribution describes the waiting time for the first occurrence in a Poisson process.

The other interpretation of the exponential distribution is the lifetime of a component. Note that

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \frac{\mathsf{P}((x, x+\varepsilon])}{\mathsf{P}((x,\infty))} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \frac{1 - e^{-\lambda(x+\varepsilon)} - 1 + e^{-\lambda x}}{e^{-\lambda x}} = \lim_{\varepsilon \to 0} \frac{1 - e^{-\lambda \varepsilon}}{\varepsilon} = \lambda$$

or in a more transparent way

$$\frac{F(x+\varepsilon) - F(x)}{\varepsilon} \cdot \frac{1}{1 - F(x)} = \frac{f(x)}{1 - F(x)} = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}} = \lambda$$

Here, λ is the hazard rate which is the probability of an event just after x given that nothing happened in $[0,x]$. The exponential distribution describes a lifetime of an object that does not age.

Definition 6.3 (Absolutely continuous)

F is absolutely continuous on $[a,b]$ if for $k = 1, \ldots, n$

$$\Big(\forall\, \varepsilon > 0\Big)\ \Big(\exists \delta > 0\Big)\ \Big(\forall\, x_k \le y_k, [x_k, y_k] \subseteq [a,b]\Big)$$
$$\left(\sum_{k=1}^{n} |y_k - x_k| < \delta \Rightarrow \sum_{k=1}^{n} |F(x_k) - F(y_k)| < \varepsilon\right)$$

In other words, F is AC if it is AC on any $[a,b] \Leftrightarrow F(x) = \int_{-\infty}^{x} f d\lambda$

Example 6.3 (Weibull distribution)

Let $\alpha, \beta > 0$. Then

$$f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}} 1_{(x>0)}$$

which integrates to 1 by a change of variable. The hazard rate is given by

$$\frac{f(x)}{1-F(x)} = \alpha\beta x^{\beta-1}.$$

This distribution can describe components that age or are in the so called "burn-in" phenomenon. For different values of β, we have

$\beta > 1 \ldots$ hazard rate is increasing in x

$\beta = 1 \ldots$ exponential distribution, hazard rate is constant

$\beta < 1 \ldots$ hazard rate is decreasing in x.

Example 6.4 (Normal (Gaussian) distribution)

Perhaps one of the best known, the Normal distribution has

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$$

for $x \in \mathbb{R}$. The fact that the function integrate to one is a consequence of the gamma function. Indeed,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} dx = 2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} dx$$

using the symmetry of the bell curve. By a change of variable $t = x^2/2 \Rightarrow \sqrt{2t} = x$ and $dt = xdx$, we get

$$= 2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t} \frac{1}{\sqrt{2t}} dt = \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} t^{\frac{1}{2}-1} e^{-t} dt = \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1$$

Remember that the gamma function is defined to be

$$\Gamma(a) = \int_{0}^{\infty} t^{a-1} e^{-t} dt.$$

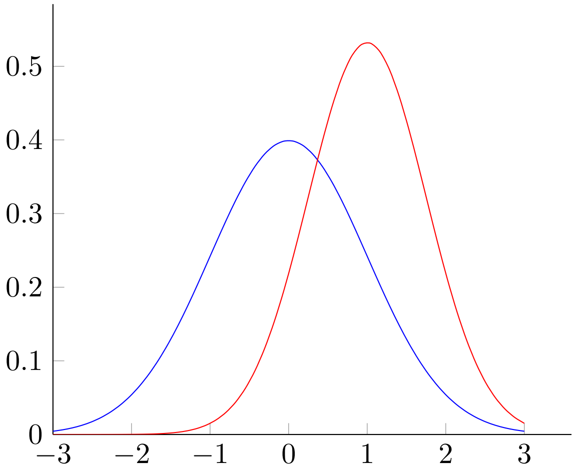


Figure 13: Gaussian bell curve with standard parameters and $\mu = 1, \sigma = 0.75$

The Gaussian distribution may be viewed as the limit of a (rescaled) Binomial distribution (from the Central Limit theorem). It is worth noting that

$$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} dx$$

is not in closed form.

The general normal distribution $\mathcal{N}(\mu, \sigma^2)$ has parameters $\mu \in \mathbb{R}, \sigma > 0$ and a probability distribution function

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Chapter 7

Random variables and their distributions

Recall that $(\Omega, \mathcal{A})$ is called a measurable space. Suppose another measurable set $(\mathcal{X}, \mathcal{B})$. We first start with a little motivation. Consider a toss with 2 distinguishable dice. The sample space is $\Omega = (i, j); i, j \in \{1, \ldots, 6\}$ and $\mathcal{P}(\Omega)$, but this time we are interested in the sum of the face values.

With our knowledge of the dice toss, $\Omega^* = \{2, \ldots, 12\}$ and $\mathcal{A}^* = \mathcal{P}(\Omega^*)$ is a new probability that we recognize to be not anymore a Laplace experiment. In fact, $\mathsf{P}^*(\{2\}) = \mathsf{P}(\{(1,1)\}) = 1/36$ and $\mathsf{P}^*(\{3\}) = \mathsf{P}(\{(1,2),(2,1)\}) = 1/18$. We will introduce a mapping $X : \Omega \to \Omega^*$ or $(i, j) \to i + j$ and

$$\mathsf{P}^*(\{k\}) = \mathsf{P}(\{i, j) : i + j = k\}) = \mathsf{P}(\{i, j) : X(i, j) = k\})$$

In general, $X : \Omega \to \mathcal{X}$. We could consider for example a conversion from canadian dollars to euros. We have for any such mapping

$$\mathsf{P}^*(B) = \mathsf{P}(\omega \in \Omega : X(\omega) \in B)$$

notice that we have by definition of probability that $\mathsf{P} : \mathcal{A} \to \mathbb{R}$, which leads to the following problem.

$$\{\omega \in \Omega : X(\omega) \in B\} = X^{-1}(B)$$

may not be in $\mathcal{A}$, *i.e.* we may not be able to compute $\mathsf{P}(X^{-1}(B))$.

Definition 7.1 ($\mathcal{A} - \mathcal{B}$ measurable mapping)

A mapping $X : \Omega \to \mathcal{X}$ where $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$ are measurable spaces is called $\mathcal{A} - \mathcal{B}$ measurable if

$$\forall \;\; B \in \mathcal{B} : X^{-1}(B) \in \mathcal{A}.$$

The common notation is $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ although it is confusing. Whether B is measurable depends on $\mathcal{A}$ and $\mathcal{B}$.[12]

[12] This concept is linked to continuity on topological spaces.

Example 7.1 (Coin toss)

$\Omega = \{(0,0),(1,0),(0,1),(1,1)\}$ and $X : \Omega \to \mathbb{R}, (i,j) \to i+j$. We are interested in the number of tails, so 0 =“heads” and 1 =“tails”. Now some remarks before we proceed. The measurability depends on the σ-field. Hence, X is $\mathcal{P}(\Omega) - \mathbb{B}$ measurable. $\mathcal{A} = \{\emptyset, \Omega, \{(1,0),(0,1)\}\}$, is a σ-field on Ω, but X is NOT $\mathcal{A} - \mathcal{B}$ measurable, that is $X^{-1}(\{2\}) = \{(1,1)\} \notin \mathcal{A}$. However, $\mathcal{B} = \{\emptyset, \mathbb{R}, \{1\}, \mathbb{R} \setminus \{1\}\}$ is a σ-field on $\mathbb{R}$, but sufficiently core so that X is $\mathcal{A} - \mathcal{B}$ measurable

$$\begin{aligned}
X^{-1}(\emptyset) &= \emptyset \in \mathcal{A} \\
X^{-1}(\mathbb{R}) &= \Omega \in \mathcal{A} \\
X^{-1}(\{1\}) &= \{(1,0),(0,1)\} \in \mathcal{A} \\
X^{-1}(\mathbb{R} \setminus \{1\}) &= \{(0,0),(1,1)\} \in \mathcal{A}
\end{aligned}$$

So measurability depends on the mapping and the σ-field.

Terminology

If $X : \Omega \to \mathbb{R}$ is $\mathcal{A} - \mathbb{B}$ measurable, then X is called a random variable and for $X : \Omega \to \mathbb{R}^d$ and $\mathcal{A} - \mathbb{B}^d$ measurable, then X is called a random vector.

$$(X_1, \ldots, X_d) : \Omega \to \mathbb{R}^d \quad \omega \to (X_1(\omega), \ldots, X_d(\omega))$$

Theorem 7.2 (Conditions for measurability)

Let $\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$ be measurable spaces and $\mathcal{B} = \sigma(\mathcal{E})$ for $\mathcal{E} \subseteq \mathcal{P}(\mathcal{X})$. Then

$$X \text{ is } \mathcal{A} - \mathcal{B} \text{ measurable } \Leftrightarrow \forall\, E \in \mathcal{E} : X^{-1}(E) \in \mathcal{A}$$

Proof

$\Rightarrow$ We have $\forall\ E \in \mathcal{E} : X^{-1}(E) \in \mathcal{A}$ hold if X is $\mathcal{A} - \mathcal{B}$ measurable because $\mathcal{E} \subseteq \sigma(\mathcal{E}) = \mathcal{B}$.

$\Leftarrow$ We will use the *good sets principle.*

A good set is

$$\mathcal{B}^* = \{B \subseteq \mathcal{X} : X^{-1}(B) \in \mathcal{A}\}$$

By assumption $\mathcal{E} \subseteq \mathcal{B}^*$ and need to show that $B \subseteq \mathcal{B}^*$. To do this, we will show that $\mathcal{B}^*$ is a σ-field.

1. $\mathcal{X} \in \mathcal{B}^*$, $X^{-1}(\mathcal{X}) = \Omega \in \mathcal{A}$
2. $B \in \mathcal{B}^* \Rightarrow B^{\complement} \in \mathcal{B}^*$.

$$\begin{aligned} X^{-1}(B^{\complement}) &= \{\omega \in \Omega : X(\omega) \notin B\} \\ &= \{\omega \in \Omega : X(\omega) \in B\}^{\complement} \\ &= \{X^{-1}(B)\}^{\complement} \end{aligned}$$

   is in $\mathcal{A}$ because $B \in \mathcal{B}^*$ and $\mathcal{A}$ is a σ-field.

3. $B_1, B_2 \ldots \in \mathcal{B}^* \Rightarrow \bigcup_{i=1}^{\infty} X^{-1}(B_i) \in \mathcal{B}^*$.

We claim that
$$X^{-1}\left(\bigcup_{i=1}^{\infty}(B_i)\right) = \bigcup_{i=1}^{\infty} X^{-1}(B_i).$$

First,

$$\begin{aligned} \omega \in X^{-1}\left(\bigcup_{i=1}^{\infty}(B_i)\right) &\Leftrightarrow X(\omega) \in \bigcup_{i=1}^{\infty}(B_i) \\ &\Leftrightarrow X(\omega) \in B_i \text{ for at least one } i \\ &\Leftrightarrow \omega \in X^{-1}(B_i) \text{ for at least one } i \end{aligned}$$

Hence we have
$$X^{-1}\left(\bigcup_{i=1}^{\infty}(B_i)\right) = \bigcup_{i=1}^{\infty} \underbrace{X^{-1}(B_i)}_{\in \mathcal{A}}$$
given that B_i is a good set and hence the union is in $\mathcal{A}$ because $\mathcal{A}$ is a σ-field.

Hence $\mathcal{E} \subseteq \mathcal{B}^*$, $\mathcal{B}^*$ is a σ-field and therefore $\sigma(\mathcal{E}) \subseteq \mathcal{B}^*$. □

Corollary 7.3

$X : \Omega \to \mathbb{R}$ is $\mathcal{A} - \mathbb{B}$ measurable if and only if

$$X^{-1}((a, b]) = \{\omega \in \Omega : X(\omega) \in (a, b]\} \in \mathcal{A}$$

for all $a < b \in \mathbb{R}$

Note that the Borel σ-field generated by open-closed sets can arise from any of the following:

$$\begin{aligned}\mathbb{B} : \sigma((a,b), a < b \in \mathbb{R}) &= \sigma((-\infty, a), a \in \mathbb{R} \\ &= \sigma((-\infty, a], a \in \mathbb{R}) \\ &= \sigma((a, \infty), a \in \mathbb{R}) \\ &= \sigma([a, \infty), a \in \mathbb{R}).\end{aligned}$$

Example 7.2

Suppose that $X : \Omega \to \mathbb{R}$ is $\mathcal{A} - \mathcal{B}$ measurable and such that

$$X^{-1}(\{0\}) = \{\omega : X(\omega) = 0\} = \emptyset.$$

Then

$$Y : \Omega \to \mathbb{R}; \omega \to \frac{1}{X(\omega)}$$

is also a random variable (i.e. $\mathcal{A} - \mathcal{B}$ measurable). This is because if $a \in \mathbb{R}$ is arbitrary, $Y^{-1}((-\infty, a]) \in \mathcal{A}$. To handle this expression, we deal with it in subcases. So $Y^{-1}((-\infty, a])$ is equal to

Indeed,

$$\begin{aligned}&= \{\omega : \frac{1}{X(\omega)} \leq a\} \\ &= \left\{\omega : \frac{1}{X(\omega)} \leq a \;\&\; X(\omega) > 0\right\} \cup \left\{\omega : \frac{1}{X(\omega)} \leq a \;\&\; X(\omega) < 0\right\} \\ &= \{\omega : 1 \leq a \;\&\; X(\omega) > 0\} \cup \{\omega : 1 \geq aX(\omega) \;\&\; X(\omega) < 0\}\end{aligned}$$

which is equal to either of these if respectively $a = 0, a > 0$ and $a < 0$

$$\begin{cases}\{\omega : X(\omega) < 0\} \\ \left(\omega : \frac{1}{a} \leq X(\omega)\} \cap \{\omega : X(\omega) > 0\}\right) \cup \left(\omega : \frac{1}{a} \geq X(\omega)\} \cap \{\omega : X(\omega) < 0\}\right) \\ \left(\omega : \frac{1}{a} \geq X(\omega)\} \cap \{\omega : X(\omega) > 0\}\right) \cup \left(\omega : \frac{1}{a} \leq X(\omega)\} \cap \{\omega : X(\omega) < 0\}\right)\end{cases}$$

are all in $\mathcal{A}$ and $\mathcal{A}$ being a σ-field, it is closed under unions and intersection, so we are good.

Lemma 7.4

Suppose $X : \Omega \to \mathcal{X}$ and $Y : \mathcal{X} \to \mathcal{Y}$ is $\mathcal{B} - \mathcal{C}$ measurable. Then, the convolution $Y \circ X : \Omega \to \mathcal{Y}, \omega \mapsto Y(X(\omega))$ is $\mathcal{A} - \mathcal{C}$ measurable.

Proof $\forall\, C \in \mathcal{C}$, $(Y \circ X)^{-1}(C) \in \mathcal{A}$, but

$$\begin{aligned}(Y \circ X)^{-1} &= \{\omega : Y(X(\omega)) \in C\} \\ &= \{\omega : X(\omega) \in Y^{-1}(C)\}\end{aligned}$$

Given that $Y^{-1}(C) \in \mathcal{B}$ because Y is $\mathcal{B} - \mathcal{C}$ measurable, we get that $\{\omega : X(\omega) \in Y^{-1}(C)\} \in \mathcal{A}$ because X is $\mathcal{A} - \mathcal{B}$ measurable. □

Lemma 7.5

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous. Then f is $\mathbb{B}^d - \mathbb{B}$ measurable.

Lemma 7.6

Let $\mathbf{X} = (X_1, \ldots, X_d) : \Omega \to \mathbb{R}^d$ and $\omega \to (X_1(\omega), \ldots, X_d(\omega))$, then $\mathbf{X}$ is $\mathcal{A} - \mathbb{B}^d$ measurable if and only if X_i is $\mathcal{A} - \mathbb{B}$ measurable for $i = 1, \ldots, d$
The proof is left as an exercise.
Hint: use the fact that X_1 is a projection and that projection are continuous.

Definition 7.7

Let $(\Omega, \mathcal{A}, \mathsf{P})$ and $(\mathcal{X}, \mathcal{B})$ be respectively a probability space and a measurable space. Let also $X : \Omega \to \mathcal{X}$ be $\mathcal{A} - \mathcal{B}$ measurable. Then

$$\mathsf{P}^X : \mathcal{B} \to \mathbb{R},\ B \to \mathsf{P}^X(B) = \mathsf{P}(X^{-1}(B)) = \mathsf{P}(\{\omega : X(\omega) \in B\})$$

is a probability measure on $(\mathcal{X}, \mathcal{B})$. It is called the image measure of P or *distribution* of X.

Let us stop a moment to do a reality check:

- P^X is well-defined because X is $\mathcal{A} - \mathcal{B}$ measurable;

- $\mathsf{P}^X(\mathcal{X}) = \mathsf{P}(\{\omega : X(\omega) \in \mathcal{X}\}) = \mathsf{P}(\Omega) = 1$;
- for $B_1, B_2, \ldots$ disjoint,

  $$\begin{aligned} \mathsf{P}^X\left(\bigcup_{i=1}^{\infty} B_i\right) &= \mathsf{P}(X^{-1}(\bigcup_{i=1}^{\infty} B_i)) \\ &= \mathsf{P}(\bigcup_{i=1}^{\infty} X^{-1}(B_i)) = \sum_{i=1}^{\infty} \mathsf{P}(X^{-1}(B_i)) = \mathsf{P}^X(B_i) \end{aligned}$$

  given that $X^{-1}(B_i)$ are pairwise-disjoint.

Notation

- P^X is also denoted by $\mathcal{L}(X)$ ("law" of X);
- If Q is a probability measure on $(\mathcal{X}, \mathcal{B})$ such that $\mathsf{P}^X = Q$, then $X \sim Q$ (X has distribution Q). For example, $X : \Omega \to \mathbb{R},\ X \sim \mathcal{N}(0,1)$ means that P^X is the standard normal distribution.

Example 7.3

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability measure and $A \in \mathcal{A}$ an event with

$$X : \Omega \to \mathbb{R},\ \omega \mapsto \begin{cases} 0 & \text{if } \omega \notin A \\ 1 & \text{if } \omega \in A \end{cases}$$

In other words, $X = 1_A$. Clearly, X is a random variable since

$$\begin{cases} \emptyset & \text{if } B \cap \{0,1\} = \emptyset \\ A & \text{if } B \cap \{0,1\} = \{1\} \\ A^{\complement} & \text{if } B \cap \{0,1\} = \{0\} \\ \Omega & \text{if } B \cap \{0,1\} = \{0,1\} \end{cases}$$

is in $\mathcal{A}$ and

$$\begin{aligned} \mathsf{P}^X(\{0\}) &= \mathsf{P}(\omega : X(\omega) = 0) = \mathsf{P}(A^{\complement}) = 1 - \mathsf{P}(A) \\ \mathsf{P}^X(\{1\}) &= \mathsf{P}(\omega : X(\omega) = 1) = \mathsf{P}(A) \end{aligned}$$

and $X \sim \mathcal{B}(\mathsf{P}(A))$ (Bernoulli).

Important observations

1. $X : \Omega \to \mathcal{X}$ measurable. Then X describes a new random experiment $(\mathcal{X}, \mathcal{B}, \mathsf{P}^X)$. For example,

$$\left.\begin{aligned} \Omega = \Big\{(i,j), i,j \in \{1,\ldots,6\}\Big\} \\ \mathcal{A} = \mathcal{P}(\Omega) \\ \mathsf{P}(\{i,j\}) = 1/36 \end{aligned}\right\} \quad X : \Omega \to \{2,\ldots,12\};\ (i,j) \to i+j$$

   The new random experiment describes the sum of face values $\mathcal{X}$: $\{2,\ldots,12\}, \mathcal{B} = \mathcal{P}(\mathcal{X})$,

$$\mathsf{P}^X(\{k\}) = \mathsf{P}((i,j) : i+j=k) = \frac{|(i,j) : i+j=k|}{36}.$$

2. Every random experiment $(\Omega, \mathcal{A}, \mathsf{P})$ can be described by some X: $\mathcal{X} = \Omega, \mathcal{B} = \mathcal{A}, X(\omega) = \omega$ (X is the identity map) and

$$\mathsf{P}^X(A) = \mathsf{P}(\omega : X(\omega) \in \mathcal{A}) = \mathsf{P}(\omega : \omega \in \mathcal{A}) = \mathsf{P}(A)$$

3. Consequently, we will describe random experiments by mappings. With the old way, we would specify for example $\Omega = [0,\infty), \mathbb{B}$ and $\mathsf{P} \sim \mathcal{N}(150,4)$. With the new way, the above would just be X a random variable (r.v.), $X : \Omega \to \mathbb{R}$ and $X \sim \mathcal{N}(150,4)$.[13]

We can deal with many r.v. at the time.

Note

A measurable mapping $X : \Omega \to \mathcal{X}$ is also called a random variable in the context where we are interested in P^X. A real-valued random variable is a measurable mapping $\Omega \to \mathbb{R}$.

Notation

A random variable is called *discrete* if X can take only countably many values i.e. $\text{ran}(X) = \{X(\omega), \omega \in \Omega\}$ is countable. Then P^X is discrete with support range(X) or ran(X).

[13] Capital letters are reserved for random variables exclusively.

Notation

A real-valued random variable is *continuous* if its distribution P^X has a density[14]. The distribution P^X of a real-valued random variable X is uniquely described by its distribution function F_X[15]:

$$F_X(x) = \mathsf{P}^X((-\infty, x]) = \mathsf{P}(\{\omega : X(\omega) \leq x\}) = \mathsf{P}(X \leq x).$$

By continuity, also note about F_X

1. If X is a real-valued r.v. , then the following probabilities

   $$\mathsf{P}(X < x),\ \mathsf{P}(X > x),\ \mathsf{P}(X \in [a, b]),\ \mathsf{P}(X \in B),\ \ldots,\ \mathsf{P}(X \geq x)$$

   are to be understood in an analogous fashion:

   $$\mathsf{P}(X \in (a, b]) = \mathsf{P}^X((a, b]) = \mathsf{P}(\omega : a < X(\omega) \leq b).$$

2. In statistics and probability, it is not X that is of interest, but rather P^X (namely, its distribution) Therefore

   - $X = Y \Leftrightarrow X(\omega) = Y(\omega)\ \forall\ \omega \in \Omega$;
   - $X = Y$ almost surely $\Leftrightarrow \mathsf{P}(\omega : X(\omega) = Y(\omega)) = 1$. This is a fairly strong assumption, but is often useful.
   - $X \overset{d}{=} Y$ or $X \overset{\mathcal{L}}{=} Y \Leftrightarrow \mathsf{P}^X(B) = \mathsf{P}^Y(B)\ \forall\ B \in \mathcal{B}$. In other words, X is equal to Y in distribution.

[14] It is important to note that it has nothing to do with continuity of X as a mapping

[15] The last expression is an abbreviation of $\mathsf{P}(\{\omega : X(\omega) \leq x\})$

Chapter 8

Functions of random variables

Let X be a r.v. , $X : \Omega \to \mathbb{R}$. if $g : \mathbb{R} \to \mathbb{R}$ is measurable ($\mathbb{B}-\mathbb{B}$ measurable), then $Y = g(X)$ is also a real valued r.v. . The most important examples of measurable function g's are the continuous functions (there are others, such as $g(x) = \text{sign}(x)$.) The distribution of Y, P^Y, is given by its distribution function (DF). Hence

$$\begin{aligned} F_Y(y) &= \mathsf{P}^Y((-\infty, x]) = \mathsf{P}(\omega : Y(\omega) \le y) \\ &= \mathsf{P}(\omega : g(X(\omega)) \le y) = \mathsf{P}\Big(\omega : X(\omega) \in g^{-1}((-\infty, y])\Big) \\ &= \mathsf{P}^X(x : g(x) \le y) \end{aligned}$$

Notation

For a real-valued r.v. X: if X is discrete, then P^X is uniquely determined by its PMF f_X. If X is continuous, P^X has density f_X.

Consider $X : \Omega \to R, g : \mathbb{R} \to \mathbb{R}$. What is then the distribution of $g(x)$?

$$\mathsf{P}(g(x) \le y) = \mathsf{P}^X(x : g(x) \le y) = \mathsf{P}(\omega : g(X(\omega)) \le y)$$

where P^X is on $\mathbb{R}$ and P on Ω.

Example 8.1

X is discrete and takes values $\{-2, -1, 0, 1, 2\}$. For each value, we have $f_X(i) = \mathsf{P}(X = i) = \mathsf{P}(\omega : X(\omega) = i)$.

We have the following: $f_X(-2) = 1/5, f_X(-1) = 1/6, f_X(0) = 1/5, f_X(1) = 1/15, f_X(2) = 11/30$. Now let $Y = X^2 = g(X)$ and $g(X) : \mathbb{R} \to \mathbb{R}, x \mapsto x^2$. We have $\{Y(\omega) : \omega \in \Omega\} = \{0, 1, 4\}$

$$f_Y(0) = \mathsf{P}(Y = 0) = \mathsf{P}(X^2 = 0) = \mathsf{P}(X = 0) = 1/5$$

$$f_Y(1) = \mathsf{P}(Y = 1) = \mathsf{P}(X \in \{-1, 1\}) = 1/6 + 1/15 = 7/30$$

$$f_Y(4) = \mathsf{P}(Y=4) = \mathsf{P}^X(\{-2,2\} = 1/5 + 11/30 = 17/30.$$

Example 8.2

$X \sim \mathcal{P}(\lambda)$ and $Y = e^x = g(X), g : \mathbb{R} \to \mathbb{R}, x \to e^x$ and

$$\{Y(\omega) : \omega \in \Omega\} = \{e^{X(\omega)}, \omega \in \Omega\} = \{e^0, e^1, e^2, \dots\}$$

is the support. Now $f_Y(e^k) = \mathsf{P}(Y = e^k) = \mathsf{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ since the transformation this time is injective.

Observation

If X is a discrete real-valued random variable and $g : \mathbb{R} \to \mathbb{R}$ measurable, then $Y = g(x)$ is also discrete and its PMF can be computed as follows :

$$f_Y(y) = \mathsf{P}(Y = y) = \mathsf{P}(g(X) = y) = \sum_{x:g(x)=y} f_X(x) = \mathsf{P}(X = g^{-1}(y))$$

if g is one-to-one. Note that the sum involves countably many summands.

Example 8.3

Let X a real-valued r.v. with DF F_X. We are interested in a new r.v. $Y = aX + b, b \in \mathbb{R}, a > 0$. This is referred to as a change in location and scale.

$$\begin{aligned} F_Y(y) &= \mathsf{P}(Y \le y) = \mathsf{P}(aX + b \le y) \\ &= \mathsf{P}(\omega \in \Omega : aX(\omega) + b \le y) = \mathsf{P}\left(X \le \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

When X has a density f_X, then Y also has a density $f_Y(y) = f_X\left(\frac{y-b}{a}\right)\frac{1}{a}$. In the special case when $X \sim \mathcal{N}(0,1)$, then $Y \sim \mathcal{N}(b, a^2)$ since

$$f_Y(y) = \frac{1}{\sqrt{2\pi}a}e^{-\frac{(y-b)^2}{2a^2}}$$

Example 8.4 (Chi-square distribution)

$X \sim \mathcal{N}(0,1), Y = X^2, y \ge 0$. $\mathsf{P}(Y \le y) = \mathsf{P}(X^2 \le y) = \mathsf{P}(-\sqrt{y} \le x \le \sqrt{y})$

so X has a density

$$\int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 2\int_{0}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

with a change of variable, $x^2 = t, 2xdx = dt$ so

$$= 2\int_0^y \frac{1}{\sqrt{2\pi}} e^{-\frac{t}{2}} \frac{1}{2\sqrt{t}} dt = \int_0^y \frac{1}{\sqrt{2\pi t}} e^{-\frac{t}{2}} dt$$

So $\mathsf{P}(Y \leq y) = \int_{-\infty}^{y} f_Y(t)dt$. If $y < 0$, then

$$\mathsf{P}(Y \leq y) = (X^2 \leq y) = 0 = \int_{-\infty}^{y} f_Y(t)dt$$

Put together, Y has density f_Y. This distribution of Y is called the chi-square distribution with one degree of freedom (denoted $Y \sim \chi_1^2$).

Example 8.5

Consider X with density f_X which has the property that

$$f_X(x) = f_X(-x) \;\forall\, x \in \mathbb{R}$$

(i.e. f_X is symmetric around 0). $Y = |x|$ so $\mathsf{P}(Y \leq y) = \mathsf{P}(|X| < y) = 0$ if $y < 0$. Now, if $y \geq 0$,

$$\begin{aligned} \mathsf{P}(Y \leq y) = \mathsf{P}(|X| \leq y) &= \mathsf{P}(-y \leq x \leq y) \\ = \int_{-y}^{y} f_X(t)dt &= \int_{-y}^{0} f_X(t)dt + \int_0^y f_X(t)dt \end{aligned}$$

which by symmetry is just

$$2\int_0^y f_X(t)dt$$

so $f_Y(y) = 2f_X(x)1_{(x>0)}$ so $\forall\, y$,

$$\mathsf{P}(Y \leq y) = 2\int_{-\infty}^{y} f_Y(t)dt, y \in \mathbb{R}$$

hence Y has density f_Y.

Observation

If X has density, say f_X, then

$$\mathsf{P}(g(x) \leq y) = \int_{\{x:g(x)\leq y\}} f_X(t)dt.$$

Example 8.6

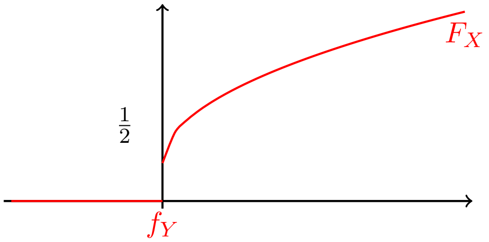
Let $X \sim \mathcal{N}(0,1), Y = \max(x,0) = x^+$ cannot be discrete and we will show that it cannot have density. The probability $\mathsf{P}(Y \leq y) = \mathsf{P}\max(X,0) \leq y)$. Let us exhaust the possible cases:

$y < 0$: $\mathsf{P}(x^+ \leq y) = 0$

$y = 0$: $\mathsf{P}(x^+ \leq 0) = \mathsf{P}(x \leq 0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{2} = \mathsf{P}(x^+ = 0)$

$y > 0$: $\mathsf{P}(x^+ \leq y) = \mathsf{P}(x \leq y) = F_X(y)$

The morale is that you should think first before looking at such problems. Y is neither discrete nor continuous since by the left F_Y is not continuous.



Theorem 8.1 (Transformation theorem for a random variable)

Let X be a real-valued r.v. with density f_X and such that there exists $(a,b), a, b \in \mathbb{R}$ so that

1. $\mathsf{P}(X \in (a,b)) = 1$;
2. f_X is continuous on (a,b).

Then if $g : (a,b) \to \mathbb{R}$ is strictly monotone and its inverse $g^{-1} : \{g(x) : x \in (a,b)\} \to (a,b)$ is *continuously differentiable*, the random variable $Y = g(X)$

has density

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| 1_{y \in \text{ range } g(x)}$$

Proof Assume that g is increasing (the other side is left as an exercise). We have $g : (a,b) \to (\alpha,\beta)$ the range of g which by assumption is continuous and strictly increasing. We want to show that $\mathsf{P}(Y \le y) = \int_{-\infty}^{y} f_Y(t)dt$ for $y \in \mathbb{R}$.

1. First case: $y \in (\alpha,\beta)$ so

$$\begin{aligned}
\mathsf{P}(Y \le y) &= \mathsf{P}(g(x) \le y) = \mathsf{P}(g(x) \le y, x \in (a,b)) \\
&= \mathsf{P}(x \le g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} f_X(x)dx \\
&= \int_{a}^{g^{-1}(y)} f_X(x)dx \xrightarrow{x=g^{-1}(t)} \\
&= \int_{g(a)=\alpha}^{y} f_X\left(g^{-1}(t)\right)\left|x\left(g^{-1}(t)\right)'\right| dt \\
&= \int_{-\infty}^{y} f_Y(t)dt
\end{aligned}$$

2. Second case $y \le \alpha$ so

$$\begin{aligned}
\mathsf{P}(Y \le y) &= \mathsf{P}(g(x) \le \alpha) = \mathsf{P}(g(x) \le \alpha, x \in (a,b)) = 0 \\
&= \int_{-\infty} y f_Y(t)dt
\end{aligned}$$

3. Third case $y \ge \beta$

$$\begin{aligned}
\mathsf{P}(Y \le y) &= \mathsf{P}(g(x) \le y, x \in (a,b)) = \mathsf{P}(X \in (a,b)) = 1 \\
&= \int_{a}^{b} f_X(x)dx = \int_{\alpha}^{\beta} f_Y(t)dt \\
&= \int_{-\infty}^{y} f_Y(t)dt
\end{aligned}$$

if g is decreasing, the proof is analogous, we are only doing changes of variables. □

Example 8.7

$X \sim \mathcal{E}(\alpha)$ so $f_X(x) = \alpha e^{-\alpha x} 1_{(x>0)}$. Fix $\beta > 0$, $X^{\frac{1}{\beta}} = Y$ where $(a,b) = (0,\infty)$. The mapping $g(0,\infty) \to (0,\infty)$, is defined $x \mapsto x^{\frac{1}{\beta}}$ and the inverse map $g^{-1}(0,\infty) \to (0,\infty), y \mapsto y^\beta$ so $(g^{-1})'(y) = \beta y^{\beta-1} > 0$. and

$$f_Y(y) = \alpha e^{-\alpha y^\beta} \beta y^{\beta-1} 1_{(y>0)}$$

Notice that $Y \sim$ Weibull. Also note that we must watch out for the absolute value $|(g^{-1})'(y)|$.

Example 8.8 (Log-normal distribution)

$X \sim \mathcal{N}(\mu,\sigma^2), e^x$. The interval of $(a,b) = (-\infty,\infty)$ and $g : (-\infty,\infty) \to (0,\infty), e \mapsto e^x$. Now, $g^{-1}(0,\infty) \to (-\infty,\infty), y \mapsto \log y, g^{-1}(y) = \frac{1}{y}$. The distribution function of Y is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{(\log y-\mu)^2}{2\sigma^2}} \frac{1}{y} 1_{(y>0)}$$

and $Y \sim$ log-normal distribution, a distribution used in actuarial science. It has certain interesting settings, notably in the case one wants to study the number of claim that we want non-negative, but meanwhile close to normal.

Example 8.9

$X \sim$ Uniform(0,1),$Y = -\log(X)$.

For this example, $f_X(x) = 1_{(x\in[0,1])}$, the interval $(a,b) = (0,1)$. The mappings are $g : (0,1) \to (0,\infty), x \mapsto -\log(x)$ and the inverse map $g^{-1}(0,\infty) \to (0,1), y \mapsto e^{-y}$ and the derivative $(g^{-1})' = -e^{-y}$

$$f_Y(y) = 1|-e^{-y}|1_{(y>0)} = e^{-y}1_{(y>0)}$$

so $Y \sim \mathcal{E}(1)$.

Chapter 9

Moments of a distribution

Suppose a dice is toss in a game. If you toss k, then your gains are $10 \times k\$$. Then the expected win is

$$35 = \sum_{k=1}^{6} k \cdot 10 \cdot \frac{1}{6} = \frac{10}{6} \sum_{k=1}^{6} k,$$

estimated by the arithmetic mean of the dice toss.

Definition 9.1 (Expected value)

Let X be a discrete real-valued r.v. with PMF f_X. The expected value of X is given by

$$\mathsf{E}X = \sum_{x \in \{X(\omega):\omega \in \Omega\}} x f_X(x)$$

provided

$$\sum_{x \in \{X(\omega):\omega \in \Omega\}} |x| f_X(x) < \infty.$$

If the above is not finite, then $\mathsf{E}X$ does not exist.

Let us revisit our example: $X \in \{10, 20, \ldots, 60\}$ and

$$\mathsf{P}(X) = \begin{cases} 1/6, & k \in \{10, 20, \ldots, 60\} \\ 0 & \text{otherwise} \end{cases}.$$

Example 9.1

If $X \sim \mathcal{B}(n,p)$, then

$$\begin{aligned}
\mathsf{E}X &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} k \frac{n!}{(n-k)!(k-1)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=0}^{n} k \frac{(n-1)!}{(n-1-(k-1))!(k-1)!} p^k (1-p)^{n-1-(k-1)}
\end{aligned}$$

Let now $k^* = k-1$; we recover the binomial so this is just equal to np.
If $X \sim \mathcal{P}(\lambda)$

$$f_X(k) = \begin{cases} \dfrac{\lambda^k}{k!} e^{-\lambda} & k \in \{0,1,\ldots\} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathsf{E}X = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{k-1!} = \lambda e^{-\lambda} \sum_{k^*=0}^{\infty} \frac{\lambda^{k^*}}{k^*!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

If $X \sim$ Geo (p), $f_X(k) = p^k(1-p)$ for $k \in \{0,1,\ldots\}$ and 0 otherwise. Then $\mathsf{E}X = \sum_{k=0}^{\infty} kp^k(1-p)$ by taking the geometric distribution and differentiating term by term, hence $\mathsf{E}X = \frac{p}{1-p}$. We will see later that in such a case, the arithmetic mean converge to the expected value by the Law of large numbers.

Example 9.2

Suppose X is a random variable such that

$$ff_X = \begin{cases} (-1)^{k+1} \dfrac{3^k}{k} = \dfrac{2}{3^k} & k = 1,2,\ldots \\ 0 & \text{otherwise} \end{cases}$$

and look at

$$\sum_{k=1}^{\infty} \frac{2}{3^k} = \frac{2}{3} \sum_{k=0}^{\infty} \frac{1}{3^k} = \frac{2}{3} \frac{1}{1-\frac{1}{3}} = 1$$

We could think a moment that $\mathsf{E}X$ is equal to

$$\sum_{k=1}^{\infty}(-1)^{k+1}\frac{2}{3^k}\frac{3^k}{k} = 2\sum_{k=1}^{\infty}\frac{-1^{k+1}}{k} = 2\log 2$$

but $\sum_{k=1}^{\infty}\frac{3^k}{k}\frac{2}{3^k}$ diverges. So $\mathsf{E}X$ is undefined. Note that since it is not absolutely convergent, we can always rearrange as we wish the summation .

Note

$\mathsf{E}X$ depends only on the distribution of X.

Motivation $\sum x\mathsf{P}(X = x)$ makes no sense if X is not discrete. Fix $\varepsilon > 0$ and let the sum be

$$\approx \sum_{i=-\infty}^{\infty} i\varepsilon\mathsf{P}(i\varepsilon < X \leq (i+1)\varepsilon).$$

If X has a $(\mathcal{R})$ density,

$$f_X = \sum_{i=-\infty}^{\infty} i\varepsilon\int_{i\varepsilon}^{(i+1)\varepsilon} f_X(x)dx \approx = \int_{-\infty}^{\infty} xf_X(x)dx.$$

Definition 9.2 (Expected value of random variables with $\mathcal{R}$ density)

Let X be a r.v. with $\mathcal{R}$ density f_X. Then the expected value of X is given as $\mathsf{E}X = \int_{-\infty}^{\infty} xf_X(x)dx$ provided $\int_{-\infty}^{\infty}|x|f_X(x)dx < \infty$, otherwise $\mathsf{E}X$ is not defined.

Example 9.3

Let $X \sim \mathcal{E}(\lambda)$, $f_X(x) = \lambda e^{-\lambda x}1_{(x:>0)}$. In this case, integrating by parts

$$\mathsf{E}X = \int_0^{\infty}\underbrace{x}_{u,u'=1}\underbrace{e^{-\lambda x}\lambda dx}_{v',v=e^{-\lambda x}} = -xe^{\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}dx = \frac{1}{\lambda}\int_0^{\infty}\lambda e^{-\lambda x}dx = \frac{1}{\lambda}$$

since this is a PDF $f_X(x)$. It is worth noting this trick as it will come often in practise.

Let now $X \sim \mathcal{N}(\mu, \sigma^2)$ so $f_X = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. First, we verify that expectation is defined, that is

$$\int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx < \infty.$$

By a change of variable. $t = \frac{x-\mu}{\sigma}, dt = \frac{dx}{\sigma}$ and from the triangle inequality $|\sigma t + \mu \leq \sigma|t| + |\mu|$ so

$$\int_{-\infty}^{\infty} |\sigma t + \mu| \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2}} dt \leq \sigma \int_{-\infty}^{\infty} \frac{|t|}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2}} dt + |\mu| \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2}} dt}_{=1}.$$

Using symmetry, this is just

$$= \frac{2\sigma}{\sqrt{2\pi}} \int_{0}^{\infty} t e^{-\frac{t^2}{2}} dt = \frac{2\sigma}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-y} dy + |\mu|$$

by setting $y = \frac{t^2}{2}$. Since this integral is the one of the density of the exponential distribution with $\lambda = 1$, we obtain a finite result. Indeed,

$$\mathsf{E}X = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} (\sigma t + \mu) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2}} dt$$

which by linearity is just

$$\sigma \underbrace{\int_{-\infty}^{\infty} \frac{t}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2}} dt}_{=0} + \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2}} dt}_{=1} = \mu$$

since density is symmetric around μ.

Lemma 9.3 (Expectation of symmetric distributions)

Suppose that X has $\mathcal{R}$ density f_X such that

1. $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$;
2. $\exists\, a \in \mathbb{R}$ such that $\forall\, x \in \mathbb{R}, f_X(x+a) = f_X(a-x)$

Then $\mathsf{E}X = a$

Proof

$$\begin{aligned}\int_{-\infty}^{\infty} x f_X(x)dx &= \underbrace{\int_{-\infty}^{a} x f_X(x)dx}_{t=a-x} + \underbrace{\int_{a}^{\infty} x f_X(x)dx}_{t=x-a} \\ &= \int_{-\infty}^{0} (a-t) f_X(a-t)dt + \int_{0}^{\infty} (a+t) f_X(a+t)dt \\ &= a\left\{\int_{-\infty}^{0} f_X(a-t)dt + \int_{0}^{\infty} f_X(a+t)dt\right\} \\ &\qquad - \left\{\int_{-\infty}^{0} t f_X(a-t)dt + \int_{0}^{\infty} t f_X(a+t)dt\right\}\end{aligned}$$

Now, we can use the assumption to cancel the second part, revert the change of variable to get the probability density and

$$a\int_{-\infty}^{\infty} f_X(x)dx = a$$

□

Example 9.4

Let $X \sim$ Cauchy and $f_X = 1/\pi(1+x^2), x \in \mathbb{R}$. It is well-known indeed that the Cauchy distribution is a heavy-tail distribution and indeed

$$\int_{-\infty}^{\infty} |x| \frac{1}{\pi} \frac{1}{1+x^2} dx = \frac{2}{\pi} \int_{0}^{\infty} \frac{x}{1+x^2} dx$$

is divergent. One can easily see this using R to generate random values; we get outliers that are way off zero. As a matter of fact, the Cauchy distribution is simply the Student's t distribution with 1 degree of freedom. The Student's t resemble very much the Normal distribution, except that it is more heavy-tailed and that the number of its existing moments is determined by the number of degrees of freedom: only the $m-1$ first moments exist. Hence, for $m = 2$, only the expectation exists and in the case of the Cauchy, the distribution has no moments. As $m \to \infty$, the Student's distribution is seen to converge to the Normal.[16]

[16] The Student's t is in fact the ratio between two independent r.v., one that is normal

Remark

Later, when we see the *Law of large numbers*, we will see that $X_1, X_2, X_3, \ldots$ a sequence of independent random variable, then the arithmetic mean

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p} \mathsf{E}X.$$

Theorem 9.4 (Expectation of functions of random variables)

Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$ Borel-measurable. Then if $Y = g(X)$

(i) If X is discrete and $\mathsf{E}Y$ exists, $\mathsf{E}Y = \mathsf{E}g(x) = \sum g(x) f_X(x)$.

(ii) If X has $(\mathcal{R})$ density f_X and (using the monotone density transformation result) g is strictly monotone and continuously differentiable on $(a, b) = \{x : f_X(x) > 0\}$, then if $\mathsf{E}g(X)$ exists,

$$\mathsf{E}Y = \mathsf{E}g(X) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Proof

(i) Denote by $S = \{X(\omega), \omega \in \Omega\} = \{s_1, s_2 \ldots\}$ (countable) and by $T = \{g(X(\omega)), \omega \in \Omega\} = \{t_1, t_2, \ldots\}$. Then, by definition,

$$\begin{aligned}\mathsf{E}Y &= \sum_{k=1}^{\infty} t_k f_Y(t_k) = \sum_{k=1}^{\infty} t_k \underbrace{\mathsf{P}(g(X) = t_k)}_{\mathsf{P}(X = g^{-1}(t_k))} \\ &= \sum_{k=1}^{\infty} t_k \sum_{i : g(s)i) = t_k} s_i \\ &= \sum_{k=1}^{\infty} \sum_{i : g(s)i) = t_k} g(s_i) f_X(s_i)\end{aligned}$$

which is nothing but the sum of all $s_i = \sum_{i=1}^{\infty} g(s_i) f_X(s_i)$.

and one that follows a $\chi^2(m)$ distribution.

(ii)

$$\mathsf{E}Y = \int_{\alpha}^{\beta} y f_X\left(g^{-1}(y)\right)\left|\left(g^{-1}(y)\right)'\right| dy$$

where $(\alpha, \beta) = \{g(X), x \in (a, b)\}$. We note that the derivative is monotone (up or down) and that (α, β) is continuous and strictly monotone. Replacing $g^{-1}(y) = x$, we obtain

$$= \int_a^b g(x) f_X(x) dx$$

□

Example 9.5

Let $X \sim \text{Weibull}(\alpha, \beta)$ and $X \overset{d}{=} Y^{\frac{1}{\beta}}$, $Y \sim \mathcal{E}(\alpha)$. Then

$$\mathsf{E}X = \mathsf{E}Y^{\frac{1}{\beta}} = \int_o^\infty y^{\frac{1}{\beta}} \alpha e^{-\alpha y} dy = \frac{1}{\alpha^{\frac{1}{\beta}}} \int_0^\infty t^{\frac{1}{\beta}+1-1} e^{-t} dt = \alpha^{-\frac{1}{\beta}} \Gamma(\frac{1}{\beta} + 1)$$

Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then, given that $\mathsf{E}X = 0$, $Y \overset{d}{=} \mu + \sigma X \ldots \mathsf{E}Y = \mu$ which entails that $\mathsf{E}Y = \mathsf{E}(\mu + \sigma X) = \mu + \sigma \mathsf{E}X = \mu$ using linearity.

Lemma 9.5 (Properties of the expectation)

In the context of 9.4 (X r.v. $, g : \mathbb{R} \to \mathbb{R}, \mathsf{E}g(X)$ exists), then

i) $\mathsf{E}(g_1(X) + \ldots + g_d(X)) = \mathsf{E}g_1(X) + \ldots + \mathsf{E}g_d(X)$ provided each $\mathsf{E}g_i(X)$ exists. This is just the linearity property of the expectation.

ii) $\mathsf{E}aX = a\mathsf{E}X$

iii) $\mathsf{E}(a) = a$, for $a \in \mathbb{R}$

Definition 9.6 (Moments of a distribution)

Let X be a r.v. $, n \in \mathbb{N}, \alpha \in \mathbb{R}$. If they exist, the following expectations have special names.

1. $\mathsf{E}X^n$ is the n^{th} moment of X

2. $\mathsf{E}|X|^\alpha$ is the α^{th} absolute moment of X.

Example 9.6 (Pareto distribution)

Consider a r.v. with density

$$f_X(x) = \begin{cases} 0 & x \le 1 \\ \dfrac{\beta}{X^{\beta+1}} & \text{otherwise, } \beta > 0 \end{cases}$$

which is the Pareto distribution . Then $\mathsf{E}|X|^\alpha = \beta \int_1^\infty x^{\alpha-\beta-1} dx < \infty$ only if $\beta > \alpha$, i.e. the first moment exists if $\beta > 1$, second moment if $\beta > 2$, etc.

The question that arises is: under what conditions do moments exist?

Example 9.7 (Standard Gaussian distribution moments)

Consider the random variable $X \sim \mathcal{N}(0,1)$. Then $\mathsf{E}(X^k) = 0$ if k is odd and $\mathsf{E}(X^k) = 1 \cdot 3 \cdot 5 \cdots (k-1)$ if k is even. We know that for $\sigma > 0$,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{X^2}{2}\frac{1}{\sigma^2}} dx = 1$$

since this is the PDF of $\mathcal{N}(0,\sigma^2)$. Suppose $a > 0$, then

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{1}{a}}} e^{-\frac{X^2}{2}a} dx \Rightarrow \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}a} dx = \frac{1}{\sqrt{a}} = a^{-\frac{1}{2}}$$

Differentiate both sides w.r.t a. This is legit as both terms inside have majorant so we can differentiate inside

$$\Rightarrow \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{X^2}{2} e^{-\frac{X^2}{2}a} dx = -\frac{1}{2} a^{-\frac{3}{2}}$$

This holds for any a, so set $a = 1$, then

$$\Rightarrow \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} X^2 e^{-\frac{X^2}{2}} dx = 1$$

so $\mathsf{E}X^2 = 1$. Differentiate again

$$\Rightarrow \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \left(\frac{X^2}{2}\right)\left(\frac{X^2}{2}\right) e^{-\frac{X^2}{2}a} dx = \left(\frac{-1}{2}\right)\left(\frac{-3}{2}\right) a^{-\frac{3}{2}}$$

so $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} X^4 e^{-\frac{X^2}{2}} dx = 3$ so $\mathsf{E}X^4 = 3$.

Let's make a small observation: let X r.v. with $\mathsf{P}(|X| < M) = 1$ for $M > 0$. In this case, $\mathsf{E}|X|^t < \infty, 0 < t < \infty$ since $|X|^t < M^t < \infty$.

Theorem 9.7

Suppose X is a random variable such that $\mathsf{E}|X|^t < \infty$. Then $|X|^s < \infty \; \forall \; 0 < s < t$.

Proof For the continuous case, the discrete case is analogous

$$\begin{aligned}
\mathsf{E}|X|^s &= \int_{-\infty}^{\infty} |x|^s f(x)dx = \int_{x:|x|<1} |x|^s f(x)dx + \int_{x:|x|\geq 1} |x|^s f(x)dx \\
&\leq \int_{x:|x|<1} 1 f(x)dx + \int_{x:|x|\geq 1} |x|^t f(x)dx \\
&\leq \int_{-\infty}^{\infty} f(x)dx + \int_{-\infty}^{\infty} |x|^t f(x)dx \\
&= 1 + \mathsf{E}|X|^t < \infty
\end{aligned}$$

□

Observation

Suppose $\mathsf{E}|X|^t < \infty : \int_{-\infty}^{\infty} |x|^t f(x)dx < \infty$. We can rewrite this as

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} |x|^t f(x)dx < \infty.$$

This implies $\int_{x:|X|>n} |x|^t f(x)dx \to 0$ as $n \to \infty$. Furthermore, we can say

$$\int_{x:|X|>n} |x|^t f(x)dx \geq n^t \int_{x:|X|>n} f(x)dx = n^t \mathsf{P}(|X| > n) \to 0 \text{ as } n \to \infty.$$

Theorem 9.8

Let X be a r.v. such that $\mathsf{E}|X|^t < \infty$ for some $0 < t < \infty$. Then $n^t \mathsf{P}(|X| > n) \to 0$ as $n \to \infty$. The converse does not hold in general. Consider X such that $\mathsf{P}(X = n) = \frac{c}{n^2 \log n}$, for $n = 2, 3 \ldots$. and $\sum_{k=2}^{\infty} \frac{c}{n^2 \log n} = 1$ (c is a normalizing constant). Take

$$\mathsf{P}(X > n) \approx \int_n^{\infty} \frac{c}{n^2 \log n} dx \approx \frac{c}{n \log n} \Rightarrow n\mathsf{P}(X > n) \to 0$$

as $n \to \infty$ Here, by $\approx$, we mean that the ratio of both sides go to 1 as $n \to \infty$. However, $\mathsf{E}X = \sum_{k=2}^{\infty} \frac{c}{n \log n} = \infty$.

Theorem 9.9

Let X be a r.v. with CDF F. Then $\mathsf{E}|X| < \infty$ if and only if $\int_{-\infty}^{\mathsf{C}} F(x)dx$ and $\int_0^{\infty}(1 - F(x))dx$ both converge. In that case, $\mathsf{E}(X) = \int_0^{\infty}(1 - F(x)dx - \int_{-\infty}^{\mathsf{C}} F(x)dx$.

Proof We consider again only the continuous case, proceeding by integration by parts.

($\Rightarrow$) If $\mathsf{E}|X| < \infty$, then by a change of variables $u = x, du = dx, dv = f(x)dx, v = F(x)$, we obtain

$$\int_0^n xf(x)dx = nF(n) - \int_0^n F(x)dx = nF(n) - n + \int_0^n (1 - F(x))dx$$

by adding and subtracting n to obtain $-n(1 - F(n)) + \int_0^n (1 - F(x))dx$. Now consider

$$n(1 - F(n)) = n \int_n^{\infty} f(x)dx \le \int_n^{\infty} xf(x)dx \to 0$$

as $n \to \infty$ and $F(n) = \mathsf{P}(X \le n)$ and $1 - F(n) = \mathsf{P}(X > n)$. Take the limits: $\int_0^{\infty} xf(x)dx = \int_0^{\infty}(1 - F(x))dx$. For the other part, we use the same trick again, namely a change of variables.

$$\int_{-\infty}^0 xf(x)dx = \lim_{n\to\infty} \int_{-n}^0 xf(x)dx = \lim_{n\to\infty} \left[nF(-n) - \int_{-n}^0 F(x)dx \right]$$

Consider the part $nF(-n)$

$$= n\int_{-\infty}^{-n} f(x)dx \leq \int_{-\infty}^{-n} |x|f(x)dx < \infty$$

So $nF(-n) \to 0$ as $n \to \infty$ hence by combining both parts, we get

$$\int_{-\infty}^{0} xf(x)dx + \int_{0}^{\infty} xf(x)dx = \mathsf{E}X = \int_{0}^{\infty}(1 - F(x))dx - \int_{-\infty}^{0} F(x)dx$$

($\Leftarrow$) Observe that, from ($\Rightarrow$)

$$\int_{0}^{n} xf(x)dx \leq \int_{0}^{n}(1 - F(x))dx \leq \int_{0}^{\infty}(1 - F(X))dx - \int_{-\infty}^{0} F(x)dx$$

Similarly,

$$\int_{-n}^{0} |x|f(x)dx \leq \int_{n}^{0} F(x)dx \leq \int_{-\infty}^{0} F(x)dx < \infty$$

and taken together, $\mathsf{E}|X| < \infty$. □

Corollary 9.10

Let X be a non-negative r.v. . Then, $\mathsf{E}X$ exists if and only if $\int_0^\infty(1 - F(x))dx < \infty$. In that case, $\mathsf{E}X = \int_0^\infty(1 - F(x))dx$.

Example 9.8

$X \sim \mathcal{E}(\alpha), f(x) = \alpha e^{-\alpha x}, x > 0, \alpha > 0$. So

$$\begin{aligned} 1 - F(X) &= \int_{x}^{\infty} \alpha e^{-\alpha y}dy = e^{-\alpha x} \\ &\Rightarrow \mathsf{E}X = \int_{0}^{\infty} e^{-\alpha x}dx = \frac{1}{\alpha}\int_{0}^{\infty} \alpha e^{-\alpha x}dx = \frac{1}{\alpha}. \end{aligned}$$

Theorem 9.11

Let X be a random variable such that $x^t\mathsf{P}\left(|X| > x\right) \to 0$ as $x \to \infty$. Then $\mathsf{E}|X|^s < \infty$ for $0 < s < t$.

Proof Fix $\varepsilon > 0$. Then, $\exists\, x_0$ such that $\forall\, x \geq x_0$ then

$$x^t \mathsf{P}\left(|X| > x\right) < \varepsilon$$

if and only if $\mathsf{P}\left(|X| > x\right) < \frac{\varepsilon}{x^2}$. We have

$$\mathsf{E}|X|^s = \int_0^\infty \mathsf{P}\left(|X|^s > x\right) dx = \int_0^\infty \mathsf{P}\left(X > x^{\frac{x}{s}}\right) dx$$

Now, using a change of variable $y = x^{\frac{1}{s}} \Leftrightarrow y^s = x, sy^{s-1}dy = dx$. Recall that we had for the expectation of $Y \geq 0$, $\mathsf{E}Y = \int_0^\infty \left(1 - F_Y(y)\right) dy$ where $(1 - F_Y(y)) = \mathsf{P}(Y \geq y)$.

$$\begin{aligned}
&= \int_0^\infty sy^{s-1}\mathsf{P}\left(|X| > y\right) dy \\
&= \int_0^{x_0} sy^{s-1}\mathsf{P}\left(|X| > y\right) dy + \int_{x_0}^\infty sy^{s-1}\mathsf{P}\left(|X| > y\right) dy \\
&\leq \int_0^{x_0} sy^{s-1} dy + \int_{x_0}^\infty sy^{s-1}\frac{\varepsilon}{y^t} dy \\
&= x_0^s + \varepsilon s \int_{x_0}^\infty y^{s-t-1}\frac{\varepsilon}{y^t} dy < \infty
\end{aligned}$$

by assumption, since the integral is finite. □

Lemma 9.12

For a r.v. X, with $\mathsf{E}|X|^t < \infty \Leftrightarrow \sum_{n=1}^\infty \mathsf{P}\left(|X| > n^{\frac{1}{t}}\right) < \infty$. The proof is in the book.

The fact we had this only defined for discrete of Riemann density and since the definition was *ad hoc*, we introduce the Lebesgue integral . Here is an optional section.

Remark

If $g \geq 0$ and measurable, and $\mathsf{E}(g(x)) = 0$, or $\sum_x g(x) f_X(x) = 0$, then $\mathsf{P}(g(x) = 0) = 1$.

Definition 9.13 (Central moments)

Let X be a random variable such that $\mathsf{E}X$ exists. then, $\mathsf{E}(X - \mathsf{E}X)^k$ is called the k^{th} central moment and $\mathsf{E}|X - \mathsf{E}X|^k$ is the k^{th} absolute central moment. In particular, for $k = 2$, $\mathsf{E}(X - \mathsf{E}X)^2$ is called the variance of X and $\sqrt{\mathsf{E}(X - \mathsf{E}X)^2}$ is the standard deviation.

Example 9.9

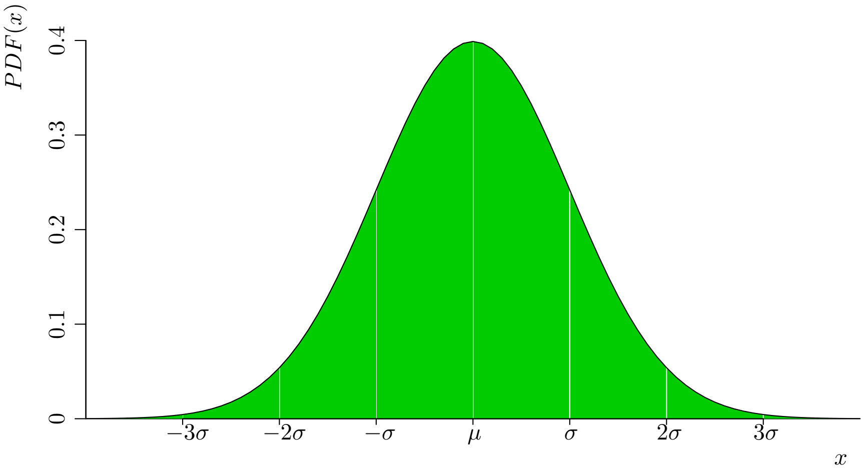
Let $X \sim \mathcal{N}(\mu, \sigma^2)$. For this example, we can compute the variance as follow:

$$\mathsf{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

let now $y = \frac{-\mu}{\sigma}$. Then

$$= \sigma^2 \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \sigma^2$$

Figure 14: Normalcurvisaurus sp.



Example 9.10

Let $X \sim \mathcal{P}(\lambda)$, $\mathsf{E}X = \lambda$. If we look at the variance, we see that

$$\sum_{k=0}^{\infty} (k - \lambda)^2 \frac{\lambda^k}{k!} e^{-\lambda}.$$

If we look at

$$k(k-1)\sum_{k=2}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\lambda^2\frac{\lambda^{k-2}}{(k-2)!}$$

which by shifting is e^{λ} and λ^2.

$$\mathsf{E}(X^2 - X) = \mathsf{E}X^2 - \mathsf{E}X = \mathsf{E}X^2 - \lambda$$

hence $\mathsf{E}X^2 = \lambda^2 + \lambda$. Now, using the next theorem (2), we have $\mathsf{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Theorem 9.14 (Properties of the variance)

Let X be a r.v. with $\mathsf{E}|X| < \infty$. Then

1. $\mathsf{Var}X$ exists if and only if $\mathsf{E}X^2 < \infty$;
2. $\mathsf{Var}X = \mathsf{E}X^2 - (\mathsf{E}X)^2$ (if $\mathsf{E}X^2 < \infty$);
3. $\mathsf{Var}X = 0 \Leftrightarrow \mathsf{P}(X = c) = 1$ for some $c \in \mathbb{R}$.
4. $\mathsf{Var}(aX + b) = a^2\mathsf{Var}X$
5. $\min_{c\in\mathbb{R}} \mathsf{E}(X - c)^2 = \mathsf{Var}X = \mathsf{E}\left(X - \mathsf{E}X\right)^2$

Proof

1. $X^2 = (X - \mathsf{E}X + \mathsf{E}X)^2 \leq 2(X - \mathsf{E}X)^2 + 2(\mathsf{E}X)^2$ since $(a-b)^2 \leq 2a^2 + 2b^2$ and so $\mathsf{E}(X - \mathsf{E}X)^2 \leq 2X^2 + 2(\mathsf{E}X)^2$.
2. The following is an equivalent way to obtain the variance.

$$\begin{aligned}\mathsf{E}\left(X - \mathsf{E}X\right)^2 &= \mathsf{E}\left(X^2 - 2X\mathsf{E}X + (\mathsf{E}X)^2\right)\\ &= \mathsf{E}X^2 + \mathsf{E}\left\{-2\mathsf{E}XX\right\} + \mathsf{E}\left((\mathsf{E}X)^2\right)\\ &= \mathsf{E}X^2 - 2\mathsf{E}X\mathsf{E}X + (\mathsf{E}X)^2\\ &= \mathsf{E}X^2 - (\mathsf{E}X)^2\end{aligned}$$

3. If $f_x(c) = 1$, then $\mathsf{E}X = c$ and $\mathsf{E}X^2 = c^2$ so by (2) $\mathsf{Var}X = c^2 - c^2 = 0$. Sufficiency of the condition. $\mathsf{E}X^2 - (\mathsf{E}X)^2 = \mathsf{E}(g(X)), g(X) \geq 0 \Rightarrow \mathsf{P}\left((X - \mathsf{E}X)^2 = 0\right) = 1 \Leftrightarrow \mathsf{P}(X = \mathsf{E}X = c) = 1$. It is readily seen that c is the mean of X

4. $\mathsf{E}(aX + b - \mathsf{E}(aX + b))^2 = \mathsf{E}(aX + b + a\mathsf{E}X + b)^2 = a^2\mathsf{E}(a^2(X + \mathsf{E}X))^2$

5.

$$\begin{aligned}
\mathsf{E}(X - c)^2 &= \mathsf{E}(X - c + \mathsf{E}X - \mathsf{E}X)^2 \\
&= \mathsf{E}\left((X - \mathsf{E}X)^2 - 2(X - \mathsf{E}X)(\mathsf{E}X - c) + (\mathsf{E}X - c)^2\right) \\
&= \mathsf{E}(X - \mathsf{E}X)^2 - 2\mathsf{E}(X - \mathsf{E}X)(\mathsf{E}X - c) + (\mathsf{E}X - c)^2 \\
&= \mathsf{Var}X + (\mathsf{E}X - c)^2 \geq \mathsf{Var}X.
\end{aligned}$$

□

Definition 9.15 (Quantile function and median)

Let X be a random variable with distribution function F. Then, the generalized inverse of F (or the quantile function) is given buy, for all $u \in [0, 1]$

$$F^{-1}(u) = \inf(x : F(x) \geq u)$$

note that at 0 inf $= -\infty$ and at 1, inf $= \emptyset$ if 1 is not reached. The median of X is given by $m_X = F^{-1}(1/2)$ and $F(m_X) \geq 1/2$, $F(x) < 1/2$ if $x < m_X$.

Another definition is as follow: A median is any $m \in \mathbb{R}$ s.t. $F(m) \geq 1/2$ and $\mathsf{P}(X \geq m) = 1 - F(m-) \geq 1/2$. If you take this last one, the median may not be uniquely defined.

We also have $\mathsf{E}|X - c|$ is minimized if $c = m_X$. The proof of this fact is left as an exercise. If it is finite, $\mathsf{E}|X - m_X|$ is more robust if the first moments are finite. It is not so easy to compute, but the advantage is that the median always exists.

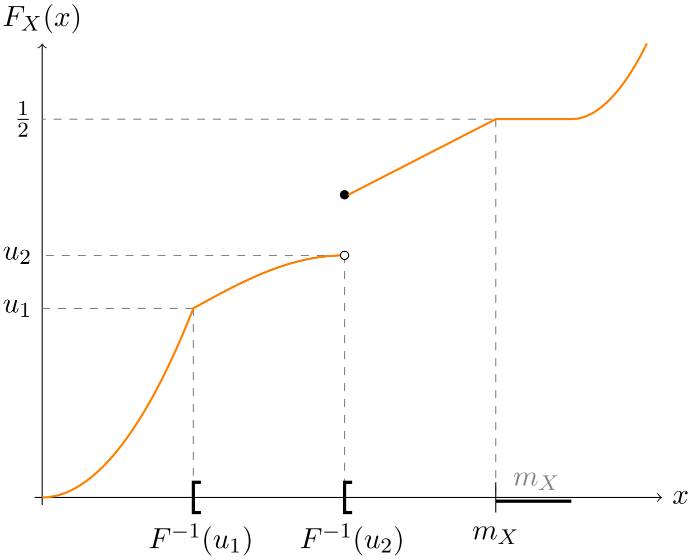


Figure 15: Example of generalized inverse for an arbitrary function.

One can remark the two definitions of the median are used in the figure, illustrating the non-uniqueness of the second definition.

Example 9.11

Let $X \sim$ Cauchy, $f_X(x) = 1/\pi(1+x^2)$ for $x \in \mathbb{R}$ and

$$F_X(x) = \int_{-\infty}^{x} \frac{dy}{\pi(1+y^2)} = \frac{1}{\pi}\left(\arctan x + \frac{\pi}{2}\right)$$

and clearly by inspection $F_X(0) = 1/2$.

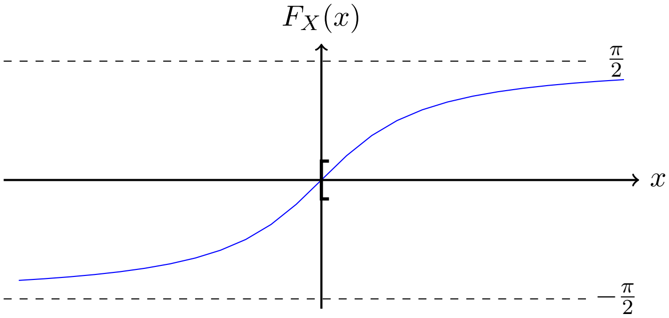


Figure 16: Distribution function of the Cauchy distribution

Example 9.12

X takes the values $1, 2, 3, 10^8$ and $\mathsf{P}(X=1) = \mathsf{P}(X=2) = \mathsf{P}(X=3) = 0.3333$. We have that the expectation is 10001.998, but the median m_X is 2.

Lebesque integration

We aim at defining $\mathsf{E}X$ for any random variable, with the idea that $\mathsf{E}X = \int X d\mathsf{P}$. Recall that for $(\Omega, \mathcal{A})$, a mapping $\mu : \mathcal{A} \to \mathbb{R}$ is called a measure if

1. $\mu(A) \geq 0 \; \forall \; A \in \mathcal{A}$;
2. $\mu(\emptyset) = 0$;
3. For $A_i, A_j, i \neq j, i, j \in [1, \infty)$, then $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$.

The idea is if $X = 1_A$ for any $A \in \mathcal{A}$, then what is $\int 1_A d\mu = \int_A d\mu = \mu(A)$. We go with a

Definition 9.16

X is called simple if the range of X is finite, *i.e.* $\{X(\omega), \omega \in \Omega\}$ is finite. This condition is actually more restrictive than discrete random variables.

Remark

If X is simple, then $X = \sum_{i=1}^{k} a_i \cdot 1_A$ where $a_1, \ldots, a_k \in \bar{\mathbb{R}}$ and A_i are pairwise disjoint and such that $\bigcup_{i=1}^{\infty} A_i = \Omega$. To see this, set $\{X(\omega), \omega \in \Omega\} = \{a_1, \ldots, a_k\}$; then $X(\omega) = \sum_{i=1}^{k} a_i 1_{(X(\omega = a_i)} = 1_{\omega \in X^{-1}(\{a_i\})} = 1_{A_i(\omega)}$ where $A_i = \{\omega : X(\omega) = a_i\} = X^{-1}(\{a_i\})$ with $a_i > 0 \; \forall \; i$.

Definition 9.17

Let X be a non-negative simple function , *i.e.* $X = \sum_{i=1} 6k a_i 1_{A_i}$. Then $\int X d\mu = \sum_{i=1}^{k} a_i \mu(A_i)$, with the convention that $\infty \cdot 0 = 0$. Let us go through the different steps

1. $X = 1_A$ and $\int X d\mu = \mu(A)$;
2. $X = \sum_{i=1}^{k} a_i 1_{A_i}$ and $\int X d\mu = \sum a_i \mu(A_i)$;
3. Chop the range with

$$A_n = X^{-1}\left(\left[\frac{i-1}{2^n}, \frac{i}{2^n}\right)\right), \quad i = 1, \ldots, n \cdot 2^n$$

and $B_n = X^{-1}([n,\infty))$ Now set $X - n = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} 1_{A_{in}} + n1_{B_n}$. If $X(\omega) < \infty$, then for n sufficiently large, there exists i_0 with $X \in \left[\frac{i_0-1}{2^n}, \frac{i_0}{2^n}\right)$ and $X_n(\omega) - X(\omega)| \leq \frac{1}{2^n} \xrightarrow{n\to\infty} 0$ and $X(\omega) = \infty, X_n(\omega) = n \to \infty$ as $n \to \infty$.

Note

If X is a discrete random variable with finite support

$$\mathsf{E}X = \sum_{X \in \{X(\omega):\omega\in\Omega\}} x\mathsf{P}(X = x) = \sum_{x} x\mathsf{P}\left(X^{-1}\{x\}\right) = \int X d\mathsf{P}$$

Here is an amazing

Lemma 9.18

Let X be a non-negative random variable$^{(*)}$ [17]. Then, there exist simple functions $0 \leq X_1 \leq X_2 \leq \ldots$ such that $\forall\, \omega, X(\omega) = \lim_{n\to\infty} X_n(\omega)$

Definition 9.19

If $X \geq 0$ is an arbitrary rv$^{(*)}$, then for $X^+ = \max(X,0)$ and $X^- = \min(X,0)$ we have $X = X^+ - X^-$ and $\int X d\mu = \int X^+ d\mu - \int X^- d\mu$ provided that $\int X^+ d\mu < \infty$ or $\int X^- d\mu < \infty$. If $\int X^+ d\mu < \infty$ and $\int X^- d\mu < \infty$, then the integral $\int X d\mu$ is finite and X is called $\mu-$integrable.

Note

X is $\mu-$integrable, $\int |X| d\mu = \int X^+ + X^- d\mu \overset{*}{=} \int X^+ d\mu + \int X^- d\mu < \infty$, where the last step is the hardest to prove.

Definition 9.20

$\mathsf{E}X = \int X d\mathsf{P}$ provided that X is $\mathsf{P}-$integrable.

Here are some interesting properties of the Lebesque integral.

Theorem 9.21 (Monotone convergence theorem)

If $0 \leq X_1 \leq X_2 \leq \ldots$ be a sequence of non-negative rv*.

Then, $\lim_{n\to\infty} Xn = \sup\{X_n, n \geq 1\} = X$ is a random variable* and $\int X d\mu = \lim_{n\to\infty} \int X_n d\mu$

[17] Generalized random variable with $\pm\infty$)

Theorem 9.22

Let X and Y be μ-integrable. Then

1. $\int_A Xd\mu = \int X1_A d\mu$ is well-defined for any $A \in \mathcal{A}$.
2. $\int Xd\mu \leq \int Yd\mu$ if $X \leq Y[\mathsf{E}X \leq \mathsf{E}Y]$
3. $|\int Xd\mu| \leq \int |X|d\mu[|\mathsf{E}X| \leq \mathsf{E}|X|]$
4. $\int aXd\mu = a\int Xd\mu$ $[\mathsf{E}aX = a\mathsf{E}X]$
5. $\int(X+Y)d\mu = \int Xd\mu + \int Yd\mu$ $[\mathsf{E}(X+Y) = \mathsf{E}X + \mathsf{E}Y]$
6. If X_1, X_2 are non-negative,
$$\int \sum_{i=1}^{\infty} X_i d\mu = \sum_{i=1}^{\infty} \int X_i d\mu$$
by the Monotone Convergence theorem

Theorem 9.23

let $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \rfloor)$ be sample spaces and $X : \Omega \to \mathcal{X}$ measurable, $g : \mathcal{X} \to \mathbb{R}$ measurable. Then

$$\mathsf{E}g(X) = \int g(X)d\mathsf{P} = \int gd\mathsf{P}^X$$

provided that $g(X)$ is integrable.

Observation (Lebesque vs Riemann integral)

f is non-negative and measurable, $f : \Omega \to \mathbb{R}$. Define $\nu : A \to \int_A fd\mu$, where f is called a $\mu-$density. Then ν is a (new) measure on $(\Omega, \mathcal{A})$ and such that for any measurable g

$$\int gd\nu = \int g \cdot fd\mu$$

by the chain-rule.

If X is discrete with PMF f_X and support $S = \{s_1, s_2 \ldots\}$ and we let μ be given by $\mu(B) = |B \cap S|$ for $B \in \mathbb{B}$, then μ is indeed a measure. It is called the counting measure and

$$\mathsf{P}(X \in A) = \mathsf{P}^X(A) = \sum_{s \in S \cap A} f_X(s) = \int_A f_X d\mu$$

$$\mathsf{E}X = \int x d\mathsf{P}^X = \int x f_X d\mu = \sum_{s \in S} s f_X(s)$$

If $f \geq 0$ is measurable and $A \mapsto \int_A f d\mu$ is a new measure. If F is absolutely continuous, then there exists $f \geq 0$ measurable such that $F(B) - F(A) = \int_a^b f d\lambda$ or in other words $F(X) = \int_{-\infty}^x f d\lambda$ and $\mathsf{E}X = \int x d\mathsf{P}^X = \int x f d\lambda$.

$\diamond \diamond \diamond$

Chapter 10
Generating functions

Moments of higher orders are often complicated to compute, furthermore some distributions are not available in closed form, even though they are often encountered (Gaussian distribution). We tackle the problem with moment generating functions.[18]

Definition 10.1

Let X be a random variable. The moment generating function (MGF) of X is given as

$$M_X(s) = \mathsf{E}e^{sx} = \begin{cases} \displaystyle\int_{-\infty}^{\infty} e^{sx} f_X(x)dx & \text{if } X \text{ has a density} \\ \displaystyle\sum_x e^{sx} f_X(x) & \text{if } X \text{ is discrete} \end{cases}$$

provided that $M_X(s)$ exists in a neighbourhood $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$ such that it is finite and defined.

Example 10.1

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, the moment generating function is given by

$$\begin{aligned} M_X(s) = \mathsf{E}e^{sx} &= \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^{\infty} e^{s(\sigma y + \mu)} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}} dy \\ &= e^{s\mu} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2 - s\sigma y}{2}} dy \end{aligned}$$

where the second equality is obtained by making the change of variable $y = \frac{x-\mu}{\sigma}, dy = \frac{dx}{\sigma}$. By completing the square, $y^2 - 2\sigma s y = (y - \sigma s)^2 - \sigma^2 s^2$

[18] There also exist probability generating function, but they are not that useful and only work in the discrete case. The characteristic functions involve complex arguments and will not be dealt with in the course.

hence

$$\begin{aligned} &= e^{s\mu + \frac{\sigma^2 s^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-s\sigma)^2}{2}} dy \\ &= e^{s\mu + \frac{\sigma^2 s^2}{2}} \end{aligned}$$

for $s \in \mathbb{R}$, since we have recovered the distribution function of a normal variable.

Example 10.2

Let $X \sim \mathcal{P}(\lambda)$

$$\mathsf{E}e^{sx} = \sum_{k=0}^{\infty} e^{sk} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^s)^k}{k!} = e^{\lambda e^s - 1}$$

for $s \in \mathbb{R}$ since the sum is $e^{\lambda e^s}$.

Example 10.3

Let $X \sim$ Geo, $\mathsf{P}(X = k) = p^k(1-p), k = 0, 1 \ldots$.

$$M_X(s) = \sum_{k=0}^{\infty} e^{sk} p^k (1-p) = (1-p) \sum_{k=0}^{\infty} (e^s p)^k$$

converge if $e^s p < 1 \Leftrightarrow e^s < 1/p \Leftrightarrow s < \log(1/p)$. so $p \in (0, 1)$. $M_X(s)$ exists for all

$$s < \log(1/p) = (1-p) \frac{1}{1 - e^s p}$$

Other examples will arise in the assignment, for the negative binomial, the double exponential, etc.

Example 10.4

Let $X \sim$ Cauchy, $f_X(x) = 1/\pi(1 + x^2)$ for $x \in \mathbb{R}$. The MGF

$$\int_{-\infty}^{\infty} \frac{e^{sx}}{\pi(1+x^2)} dx$$

is not finite. Even though the MGF is easy to compute, it is not well-defined for all distribution functions.

Heuristic calculation

$$\mathsf{E}e^{sx} = \mathsf{E}\left(\sum_{k=0}^{\infty} \frac{s^k x^k}{k!}\right) \overset{(*)}{=} \sum_{k=0}^{\infty} \frac{s^k \mathsf{E}x^k}{k!}$$

can be differentiated if in the radius of convergence.

$$(\mathsf{E}e^{sx})^{(l)} \overset{(*)}{=} \sum_{k=0}^{\infty} \frac{\mathsf{E}x^k}{k!} \frac{s^{k-l}k!}{(k-l)!}\Big|_0 = \mathsf{E}x^l$$

Suppose X is a r.v. such that $M_X(s)$ exists in $(-\varepsilon, \varepsilon)$ for $\varepsilon > 0$ Then, for any $l \in \mathbb{N}$,

$$(M_X(s))^{(l)}\Big|_0 \mathsf{E}x^l$$

can be shown to converge absolutely. More can be found in Widder book *The Laplace transform* (1946).

Theorem 10.2

If X, Y are r.v. whose MGF exists in $(-\varepsilon, \varepsilon)$ for $\varepsilon > 0$ (the smallest interval of X and Y around 0) and $M_X(s) = M_Y(s)\ \forall\ s \in (-\varepsilon, \varepsilon)$, then $X \overset{d}{=} Y$, *id est* there is a one-to-one correspondence between the moment generating function and the distribution function (if it exists).

Example 10.5

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $M_X(s) = e^{\mu s + \sigma^2 s^2/2}, s \in \mathbb{R}$,

$$M_X'(s) = \left(e^{\mu s + \frac{\sigma^2 s^2}{2}}\right)(\mu s + \sigma^2 s)$$

then $M_X'(0) = \mu = \mathsf{E}X$. Similarly, by the product rule

$$M_X''(x) = \left(e^{\mu s + \frac{\sigma^2 s^2}{2}}\right)(\mu s + \sigma^2 s)^2 + \left(e^{\mu s + \frac{\sigma^2 s^2}{2}}\right)\sigma^2$$

imply $M_X''(0) = \mu^2 + \sigma^2 = \mathsf{E}X^2 \to \mathsf{Var}X = \sigma^2$. If the MGF exists, all moments exist.

Observation

If $M_X(s)$ exists on $(-\varepsilon, \varepsilon)$, then $\mathsf{E}X^k < \infty\ \forall\ k \in \mathbb{N}$. examples include the Student's t, the Cauchy and the Pareto distributions.

Question: If all moments are finite, i.e. i $\mathsf{E}X^k < \infty \ \forall \ k \in \mathbb{N}$, does the MGF $M_X(s) = \sum_{k=0}^{\infty} s^k \mathsf{E}x^k/k!$ exist on $(-\varepsilon, \varepsilon)$? The answer is no.

Example 10.6

Let $X \sim f_x(x) = ce^{-|x|^\alpha}$ for $0 < \alpha < 1, x \in \mathbb{R}$.

$$\mathsf{E}|X|^k = \int_{-\infty}^{\infty} |x|^k ce^{-|x|^\alpha} dx = 2\int_0^\infty x^k ce^{-x^\alpha} dx$$

which by the change of variables $y = x^\alpha, dy = \alpha x^{\alpha-1}dx$

$$= \frac{2c}{\alpha}\int_{-\infty}^{\infty} y^{\frac{k+1}{\alpha}-1} e^{-y} dy = \frac{2c}{\alpha}\Gamma\left(\frac{k+1}{\alpha}\right) < \infty$$

Thus $M_X(s) = \int_{-\infty}^{\infty} e^{sx} ce^{-|x|^\alpha} dx \geq c\int_0^\infty e^{x\left(s-1/x^{t-\alpha}\right)} dx = \infty$.

Example 10.7

Consider X with density

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \dfrac{1}{x\sqrt{2\pi}} e^{-\frac{(\log x)^2}{2}} & \text{otherwise} \end{cases}$$

hence $X \sim$ lognormal (0,1). This distribution arise as follows: for

$$Z \sim \mathcal{N}(\mu, \sigma^2) \ \therefore \ e^Z \sim \mathcal{LN}(\mu, \sigma^2).$$

Let Y have density $g(y) = f(y)(1 = \theta)\sin(2\pi\log y)), |\theta| \leq 1$. You can show g is indeed a density. If we look at the moments of our two distributions:

$$\begin{aligned}\mathsf{E}Y^k &= \int_0^\infty y^k f(y) + y^k f(y)\theta\sin(2\pi\log y)dy \\ &= \mathsf{E}X^t + \int_0^\infty +y^k f(y)\theta\sin(2\pi\log y)dy\end{aligned}$$

and the later is 0 because the sine function is an odd function. Indeed,

$\log y = t$ and

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} e^{kt}\frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}}\theta\sin(2\pi t)dt \\
&= \frac{\theta}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{(t-k)^2}2e^{k^2}2\sin(2\pi+t)dt \\
&= \frac{\theta}{\sqrt{2\pi}}e^{-k^2 2}\int_{-\infty}^{\infty} e^{-z^2}2\sin(2\pi z+k)dz = 0
\end{aligned}
$$

The second equality comes from completing the square and the last from the fact we are integrating $\sin(2\pi z)$. All moments are finite and exists, but the distribution is not the same and furthermore by Exercise 10.6, we know $\exp -(\log x)^2/2$ is even slower then the previous example.

Theorem 10.3

Let X be a random variable such that $\mathsf{E}X^k < \infty \; \forall \; k \in \mathbb{N}$. If $\sum_{k=0}^{\infty} s^k \mathsf{E}x^k/k!$ is absolutely convergent for some $s > 0$ (interchanging the limit and the integral is possible and our heuristic calculation works.). then X has a MGF and the $\{\mathsf{E}X^k, k \in \mathbb{N}\}$ determines the distribution of X uniquely.

Observation

If X has a bounded support, then

$$\mathsf{E}|X|^k = \int_{-\infty}^{\infty} |x|^k f_x(x)dx \leq M^k \int_{-\infty}^{\infty} f_x(x)dx = M^k < \infty$$

Here is a generalization of the moment generating function.

Definition 10.4 (Characteristic function)

Let X be a r.v. The characteristic function of X is given by $\phi_X(t) = \mathsf{E}e^{itx}$ for $t \in \mathbb{R}$. The advantage of this is that is it always defined.

Remark

1. The characteristic function always exists:

   $$\mathsf{E}e^{itx} = \mathsf{E}\cos tx + i\sin tx = \mathsf{E}\cos xt + \mathsf{E}i\sin tx$$

   and both are bounded. Now, $|\mathsf{E}\cos tx| \leq \mathsf{E}|\cos tx| \leq \mathsf{E}1 = 1$ and similarly, $|\mathsf{E}\sin tx| \leq 1$;

2. $\phi_X(t)$ is continuous in t;

3. $\phi_X(t)$ determines the distribution of X uniquely. The convergence of the random variable can be characterized by the convergence of their characteristic function. Some random variables are even defined by their characteristic function; examples such as the elliptical distributions which have application in finance and are nothing but generalization of a more heavy-tailed version of the multinormal distribution.

Chapter 11
Moments inequalities

Theorem 11.1

Let X be a r.v. and $h : \mathbb{R} \to \mathbb{R}$, h measurable and *non-negative.* Furthermore, assume that $\mathsf{E}h(X) < \infty$. Then, for all $M > 0$,

$$\mathsf{P}(h(X) \geq M) \leq \frac{\mathsf{E}h(X)}{M}$$

Contrary to analysis, we are interested by large values of M rather than small ones.

Proof We first prove for X discrete and show from right to left.

$$\begin{aligned}\mathsf{E}h(X) = \sum h(x) f_X(x) &= \sum_{x:h(x)\geq M} h(x) f_X(x) + \sum_{x:h(x)<M} h(x) f_X(x) \\ &\geq M \sum_{x:h(x)\geq M} f_X(x) = M \cdot \mathsf{P}(h(X) \geq M)\end{aligned}$$

We first split the sum in two in the first equality, then drop the smallest part of the summation and use M as a lower bound for the second. In full generality now:

$$\begin{aligned}\mathsf{E}h(X) &= \int h(x) d\mathsf{P}^X(x) \\ &= \int 1_{x:h(x)\geq M} h(x) d\mathsf{P}^X(x) + \int 1_{x:h(x)<M} h(x) d\mathsf{P}^X(x) \\ &\geq M \int 1_{x:h(x)\geq M} d\mathsf{P}^X(x) = M \cdot \mathsf{P}(h(X) \geq M)\end{aligned}$$

remarking that $x : h(x) \geq M$ is measurable, so a Borel set. □

Corollary 11.2 (Markov inequality)

Let X be a random variable such that $\mathsf{E}|X|^s < \infty$ for some $s > 0$. Then

$$\mathsf{P}(|X| \geq M) \leq \frac{\mathsf{E}|X|^s}{M^s}$$

$\forall\, M > 0$. This gives a bound on the probability of the tail of the distribution.

Proof For $\mathsf{P}(|X| \geq M) = \mathsf{P}(|X|^s \geq M^s)$. Apply Theorem 11.1 on $h(X) = |X|^s$ to get the result. □

Corollary 11.3 (Chebichev-Bienaymé inequality)

If X is a random variable with $\mathsf{E}X^2 < \infty$, then

$$\mathsf{P}\left(|X - \mathsf{E}X| \geq M\sqrt{\mathsf{Var}X}\right) \leq \frac{1}{M^2}$$

We will use the preceding in the proof of the Law of large numbers.

Proof

$$\mathsf{P}\left(|X - \mathsf{E}X| \geq M\sqrt{\mathsf{Var}X}\right) = \mathsf{P}\left((X - \mathsf{E}X)^2 \geq M^2\mathsf{Var}X\right)$$

Applying Theorem 11.1 on $h(X) = (X - \mathsf{E}X)^2$ yields the ratio $\frac{\mathsf{Var}X}{M^2\mathsf{Var}X} = \frac{1}{M^2}$.
□

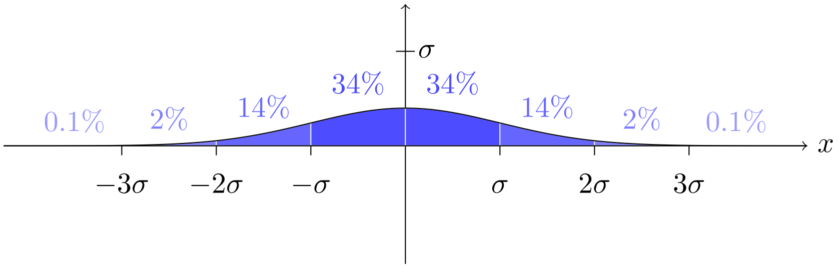


Figure 17: Normal distribution

Example 11.1

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

$$\begin{aligned}\mathsf{P}(|X - \mu| \leq M\sigma) &= \mathsf{P}(-M\sigma \leq X - \mu \leq M\sigma)\\ &= \mathsf{P}\left(-M \leq \frac{X-\mu}{\sigma} \leq M\right) = \Phi(M) - \Phi(-M)\end{aligned}$$

where Φ is the DF of the standard normal. We can implement $\Phi(M)$ in R using `pnorm(M)`. The Gaussian curve has 68.27 % of the density in the interval $[\mu-\sigma, \mu+\sigma]$, 95.45% between $[\mu-2\sigma, \mu+2\sigma]$ and 99.73% within three standard deviation. Now, if we do this calculation using the Chebychev [19] inequality, we get.

$$\mathsf{P}(|X-\mu| \leq M\sigma) = 1 - \mathsf{P}(|X-\mu| \geq M\sigma) \geq 1 - \frac{1}{M^2}$$

For $M = 1 \ldots \mathsf{P} \geq 0$, $M = 2 \ldots \mathsf{P} \geq 3/4$ and for $M = 3 \ldots \mathsf{P} \geq 8/9$, which is not impressive. We can look yet at another example that is artificial, but gives a very sharp bound.

Example 11.2

Fix $M > 0$ and let $\mathsf{P}(X = 0) = 1 - 1/M^2$, $\mathsf{P}(X = \pm 1) = 1/2M^2$. Then, for this r.v.

$$\mathsf{E}X = (-1)\frac{1}{M^2} + (0)\left(1 - \frac{1}{M^2}\right) + (1)\frac{1}{M^2} = 0$$

entails $\mathsf{Var}X = \mathsf{E}X^2$. The probability $\mathsf{P}\left(|X| \geq \frac{M}{M}\right) = \mathsf{P}(|X| \geq 1) = \frac{1}{M^2}$ which is precisely the Chebychev inequality bound.

Theorem 11.4

Let X be a r.v. with $\mathsf{E}X = 0$ and $\mathsf{Var}X = \sigma^2$. Then $\mathsf{P}(X \geq M) \leq \frac{\sigma^2}{\sigma^2+M^2}$ if $M > 0$ and $\mathsf{P}(X \leq -M) \leq \frac{\sigma^2}{\sigma^2+M^2}$ which entails since these are disjoint sets

$$\mathsf{P}(|X| \geq M) = \mathsf{P}(X \geq M \text{ or } X \leq -M) \leq \frac{2\sigma^2}{\sigma^2+M^2}$$

Note that the assumption that $\mathsf{E}X = 0$ is actually not very restrictive since $|X - \mathsf{E}X| = |Y|, Y = X - \mathsf{E}X \Rightarrow \mathsf{E}Y = 0$. It will make calculations much easier in a sense.

[19] It is also found under the name Tchebychev in the literature.

Proof

$$\begin{aligned}
\mathsf{P}(X \geq M) &= \mathsf{P}(X + c \geq M + c) \quad \forall\, c > 0 \\
&\leq \mathsf{P}(|X + c| \geq M + c) = \mathsf{P}\left((X+c)^2 \geq (M+c)^2\right) \\
&\leq \mathsf{E}\frac{(X+c)^2}{(M+c)^2} = \frac{\mathsf{E}X^2 + 2c\mathsf{E}X + c^2}{(M+c)^2} \\
&= \frac{\sigma^2 + c^2}{(M+c)^2} = f(c)
\end{aligned}$$

If we minimize this with respect to c, ie $\min_c f(c)$, we get that the minimum is attained if $c = \sigma^2/M$. In particular,

$$\mathsf{P}(X \geq M) \leq \frac{\sigma^2 + \dfrac{\sigma^4}{M^2}}{\left(\dfrac{M^2+\sigma^2}{M}\right)^2} = \frac{M^2\sigma^2 + \sigma^4}{(M^2+\sigma^2)^2} = \frac{\sigma^2}{M^2+\sigma^2}$$

For the other inequality, you can play with it in a similar fashion

$$\mathsf{P}(X \leq M) = \mathsf{P}(X - c \leq -M - c) \ldots \leq \mathsf{P}((X-c)^2 \leq (M+c)^2)$$

and do the same trick as before. If M is large, we get a more accurate result.
□

Theorem 11.5

Let $\mathsf{E}|X|^4 < \infty$, $\mathsf{E}X = 0$, and $\mathsf{Var}X = \sigma^2$. Then for any $M > 1{,}3$

$$\mathsf{P}(|X| > M\sigma) \leq \frac{\mathsf{E}X^4 - \sigma^4}{\mathsf{E}X^4 + \sigma^4 M^4 - 2M^2\sigma^4}$$

Proof Let $Y = \dfrac{x^2 - \sigma^2}{M^2\sigma^2 - \sigma^2} \Rightarrow \mathsf{E}Y = 0, \mathsf{Var}Y$ also exists. Then

$$\mathsf{P}(Y \geq 1) \leq \frac{\mathsf{Var}X}{1 + \mathsf{Var}X} = \frac{\dfrac{1}{(M^2\sigma^2 - \sigma^2)^2 \, \mathsf{Var}X^2}}{1 + \dfrac{1}{(M^2\sigma^2 - \sigma^2)^2}\mathsf{Var}X^2}$$

$$= \frac{\mathsf{E}X^4 - \sigma^4}{M^4\sigma^4 + 2\sigma^4M^2 + \sigma^4 - \sigma^4 - \mathsf{E}X^4}$$

□

Theorem 11.6 (Lyapunov inequality)

Let X be a r.v. with $\mathsf{E}|X|^n < \infty$ for some $n \in \mathbb{N}$. Then, for any $2 \leq k \leq n$, we have

$$\left(\mathsf{E}|X|^{k-1}\right)^{\frac{1}{k-1}} \leq \left(\mathsf{E}|X|^k\right)^{\frac{1}{k}}$$

Proof Consider the quadratic form

$$Q(u,v) = \int_{-\infty}^{\infty} \left(u|x|^{\frac{k-1}{2}} + v|x|^{\frac{k+1}{2}}\right)^2 d\mathsf{P}^X(x) = \mathsf{E}\left(u|X|^{\frac{k-1}{2}} + v|x|^{\frac{k+1}{2}}\right)^2 \geq 0$$

This means that for all $u, v \in \mathbb{R}$, it is non-negative. We now require $k < n$ for the quadratic form; if we expand the square

$$u^2\mathsf{E}|X|^{k-1} + 2uv\mathsf{E}|X|^k + v^2\mathsf{E}|X|^{k+1} \geq 0$$

$$\Leftrightarrow \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \mathsf{E}|X|^{k-1} & \mathsf{E}|X|^k \\ \mathsf{E}|X|^k & \mathsf{E}|X|^{k+1} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \geq 0$$

is a positive semi-definite matrix. This implies that the determinant

$$\begin{vmatrix} \mathsf{E}|X|^{k-1} & \mathsf{E}|X|^k \\ \mathsf{E}|X|^k & \mathsf{E}|X|^{k+1} \end{vmatrix} = \mathsf{E}|X|^{k+1}\mathsf{E}|X|^{k-1} - \left(\mathsf{E}|X|^k\right)^2 \geq 0$$

and equal by taking the k^{th} power

$$\left(\mathsf{E}|X|^{k+}\mathsf{E}|X|^{k-1}\right)^k \geq \left(\mathsf{E}|X|^k\right)^{2k}$$

Now by induction on k. If $k = 2$,

$$\mathsf{E}X \leq \left(\mathsf{E}|X|^2\right)^{\frac{1}{2}} \Leftrightarrow (\mathsf{E}|X|)^2 \leq \mathsf{E}|X|^2 \Leftrightarrow \mathsf{Var}X \geq 0$$

which is true. Suppose that for $k < n$,

$$\left(\mathsf{E}|X|^{k-1}\right)^k \leq \left(\mathsf{E}|X|^k\right)^{k-1}.$$

At the same time

$$\left(\mathsf{E}|X|^k\right)^{2k} \leq \left(\mathsf{E}|X|^{k+1}\mathsf{E}|X|^{k-1}\right)^k$$

If we multiply the two inequalities, then

$$\left(\mathsf{E}|X|^k\right)^{2k} \cancel{\left(\mathsf{E}|X|^{k-1}\right)^k} \leq \left(\mathsf{E}|X|^k\right)^{k-1} \left(\mathsf{E}|X|^{k+1}\right)^k \cancel{\left(\mathsf{E}|X|^{k-1}\right)^k}.$$

If $\mathsf{E}|X|^n = 0$, it will be degenerate (almost everywhere zero). Assuming it is not the case, we get

$$\left(\mathsf{E}|X|^k\right)^{k+1} \leq \left(\mathsf{E}|X|^{k+1}\right)^k$$

□

Chapter 12
Multivariate random vectors

Definition 12.1 (Random vector)

A random vector $\mathbf{X} = (X_1, \dots, X_n)$ is a measurable function from $(\Omega, \mathcal{A}, \mathsf{P})$ to $\mathbb{R}^d$.

Definition 12.2

A (multivariate) distribution function f is defined by

$$F(X_1, \dots, X_n) = \mathsf{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_d \leq x_d)$$

$\forall\, x_1, \dots, x_n \in \mathbb{R}$.[20] F is a function $F : \mathbb{R}^d \to [0, 1]$.

Properties of F

F characterizes the random behaviour of $\boldsymbol{x}$. Any distribution function F has the following properties:

(i)

$$\lim_{x_1 \to \infty} \dots \lim_{x_d \to \infty} F(x_1, \dots, x_n) = 1;$$

in other words, all components go to 1. We have increasing sequences of events on which we put less and less constraints on Ω.

(ii)

$$\lim_{x_i \to -\infty} F(x_1, \dots, x_n) = 0 \;\forall\; x_i$$

since the intersection of $\emptyset$ with any set is again the emptyset, which has probability zero.

(iii) F is non-decreasing and right continuous in each of its d components. For example, fix $x_2, \dots, x_d \in \mathbb{R}$, then $t \mapsto F(t, x_2, \dots, x_d)$ is right-continuous and non-decreasing. The proof is essentially the same as

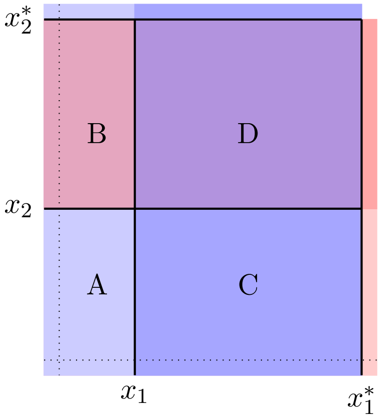
[20] The commas denote intersection of sets.

in one dimension. Note that the three are sufficient in 1-d, but the essential property to construct a distribution function is

(iv) F is $d-$monotone. When $d=2$, this means for $x_1 < x_1^*$ and $x_2 < x_2^*$

$$F(x_1^*, x_2^*) - F(x_1, x_2^*) - F(x_1^*, x_2) + F(x_1, x_2) \geq 0$$

Figure 18: Viewing $2-$monotonicity with sets



The above express $\mathsf{P}(A \cup B \cup C \cup D) - \mathsf{P}(A \cup C) - \mathsf{P}(A \cup B) + \mathsf{P}(A)$ since they are disjoint sets. In terms of areas, this is $(A+B+C+D)-(A+C)-(A+B)+A = D$ hence $\mathsf{P}(x_1 < X < x_1^*, x_2 < X_2 < x_2^*)$. Without (iv), we could build F and get a defective d.

The case of $d=3$ yields

$$\begin{aligned} F(x_1^*, x_2^*, x_3^*) - F(x_1, x_2^*, x_3^*) &+ F(x_1, x_2, x_3^*) - F(x_1, x_2, x_3) \geq 0 \\ - F(x_1^*, x_2, x_3^*) &+ F(x_1, x_2^*, x_3) \\ - F(x_1^*, x_2^*, x_3) &+ F(x_1^*, x_2, x_3) \end{aligned}$$

and in higher dimensions

$$\sum_{\substack{c_i \in \{x_i, x_i^*\} \\ i \in \{1, \ldots, d\}}} (-1)^{v(c)} \times F(c_1, \ldots, c_d)$$

where $v(c)$ is the number of c_i's that have no stars. Conversely, any function $F : \mathbb{R}^d \to [0,1]$ that satisfies (i) → (iv) is a valid distribution function *i.e.* there exists a probability measure that yields F. The proof are exactly as in the univariate case; the key element is that $(x_1, x_1^*] \times (x_2, x_2^*] \times \cdots \times (x_d, x_d^*]$ are precisely the sets of form that generate the Borel class in $\mathbb{R}^d$. Multivariate distributions are the bread and butter of statistician, but let us focus for now however on the bivariate case $(X = X_1, Y = X_2)$.

Definition 12.3

A random vector (X, Y) has a *discrete* distribution if there exists a countable set S such that $\mathsf{P}\{(X,Y) \in S\} = 1$, that is X and Y takes countably many values with probability 1. S is often expressed as a cartesian product $S = S_X \times S_Y$.

The *joint probability mass function* is defined by $f_{(X,Y)} = \mathsf{P}(X = x, Y = y)$ for $x, y \in \mathbb{R}$. Since $S_X = \{x_1, x_2, \ldots\}$ and $S_Y = \{y_1, y_2, \ldots\}$, we have

$$f_{(X,Y)}(x,y) = \begin{cases} 0 & \text{if } x \notin S_X \\ 0 & \text{if } y \notin S_Y \\ p_{ij} & \text{when } x = x_i, y = y_j \end{cases}$$

where $p_{ij} \in [0,1]$ and $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} p_{ij} = 1$.

Example 12.1 (Bivariate Poisson)

Let $X \sim \mathcal{P}(\lambda)$ if and only if $\mathsf{P}(X = x) = e^{-\lambda}\lambda^x/x!$ for $x \in \mathbb{N}, \lambda > 0$. This variable could be the intensity of phone call at 911 in Montreal. We may wish to compare with the number of phone calls on the south shore, which follow another Poisson: we have $Y \sim \mathcal{P}(\mu)$ iff $\mathsf{P}(Y = y) = e^{-\mu}\mu^y/y!$ for $y \in \mathbb{N}, \mu > 0$. Then the joint probability is

$$\mathsf{P}(X = x, Y = y) = \frac{e^{-(\lambda+\mu)}\lambda^x \mu^y}{x!y!} \quad \forall\ (x,y) \in \mathbb{N} \times \mathbb{N}$$

or we could also write it as

$$\mathsf{P}(X = i, Y = j) = \frac{e^{-(\lambda+\mu)}\lambda^i \mu^j}{i!j!} \quad \forall\ (i,j) \in S$$

Example 12.2 (Multinomial distribution)

Suppose a random event can result in $d+1$ different outcomes. This could be classification of the quality of an object, or more simply the distribution of grades. So X_1 is the number of items in bin 1, X_d the number of items in bin D, etc. Assume there are n items in all. $(X_1, \ldots, X_n)$ is said to have a multinomial distribution with parameters $n \in \mathbb{N}$ and $p_1, \ldots, p_d \in [0,1]$ if and only if

$$\begin{aligned}\mathsf{P}(X_1 = x_1, \ldots, X_d = x_d) &= \\ &\binom{n}{X_1 \cdot X_2 \cdots X_d} p_1^{X_1} p_2^{X_2} \cdots p_d^{X_d} \left\{1 - \sum_{i=1}^{d} p_i\right\}^{X_{d+1}}\end{aligned}$$

where $\left(n - \sum_{i=1}^{d} p_i\right) = X_{d+1}$ and we have that $0 \leq p_1 + \cdots p_d \leq 1$ so that $1 - \sum_{i=1}^{d} p_i \in [0,1]$, $X_1, X_2, \ldots, X_{d+1} \in \mathbb{N}$, $\sum_{i=1}^{d+1} X_i = n$ and $p_i =$ probability of falling in bin i.

Remark

In a discrete (bivariate) model, the choice of p_{ij}'s determines the joint distribution.

$$\mathsf{P}\{(X,Y) \in A\} = \mathsf{P}\left\{(X,Y) \in \bigcup\{i,j\}, (i,j) \in A\right\} = \sum_{(i,j) \in A} p_{ij}.$$

Definition 12.4

A random vector $(X_1, \ldots, X_n)$ is said to have a continuous distribution if and only if there exists a function $f : \mathbb{R}^d \to [0, \infty)$ such that

$$\begin{aligned}F(X_1, \ldots, X_n) &= \mathsf{P}(X_1 \leq x_1, \ldots, X_d \leq x_d) \\ &= \int_{-\infty}^{x_d} \cdots \int_{-\infty}^{x_1} f(t_1, \ldots, td) dt_1 \ldots dt_d\end{aligned}$$

If F is absolutely continuous, then

$$f(x_1, \ldots, x_n) = \frac{\partial}{\partial x_1 \ldots \partial x_d} F(X_1, \ldots, X_n).$$

The function f is called the multivariate density function of X. The following properties defines a multivariate probability density function:

(i) $f(X_1,\ldots,X_n)\geq 0$;

(ii) $\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\ldots,x_n)dx_1\ldots dx_d = 1$.

Example 12.3 (A bivariate exponential distribution)

Consider the time between two Poisson, says the waiting time between two calls at 911.

$$X\sim\mathcal{E}(\lambda)\rightarrow f_X(x)=\lambda e^{-\lambda x}, x\in[0,\infty), \lambda>0$$
$$Y\sim\mathcal{E}(\mu)\rightarrow f_Y(y)=\mu e^{-\mu y}, y\in[0,\infty), \mu>0$$

Here is a possible $f_(X,Y)(x,y)=\lambda\mu e^{-\lambda x-\mu y}\ \forall\ (x,y)\in[0,\infty)^2$. "This is called stochastic independence, but as French Canadian, I am not allowed to talk about independence", *dixit* Pr. Christian Genest.

Example 12.4

Define

$$F(x,y)=\begin{cases}0 & \text{if } x<0 \text{ or } y<0\\ \min(x,y) & \text{if } x\in[0,1) \ \&\ y\in[0,1)\\ x & \text{if } x\in[0,1) \ \&\ y\geq 1\\ y & \text{if } x\geq 1 \ \&\ y\in[0,1)\\ 1 & \text{if } x\geq 1 \ \&\ y\geq 1\end{cases}$$

For this example, we see that both random variables are uniform on (0,1), that is $X=\mathcal{U}$ and $Y=\mathcal{U}$ and $\mathsf{P}(X\leq x, Y\leq y)=\mathsf{P}(\mathcal{U}\leq x,\mathcal{U}\leq y)$. If we look at the partial derivatives however, we note $\partial F/\partial x\partial y=0$ almost everywhere and (X,Y) is neither discrete nor does it have a density. Note that F is continuous! Recall the exotic cases we had in 1-dimension; they are much more frequent for higher dimensions and we really must verify it integrates to 1.

Morale Check that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\partial F(x,y)}{\partial x\partial y}dxdy=1$$

It could happen that points cluster on the line and the volume is zero. Consider now a random vector $(X_1, \dots, X_n)$ is provided in joint distribution function $F(X_1, \dots, X_n) = \mathsf{P}(X_1 \leq x_1, \dots, X_d \leq x_d)$. if we wanted to reconstruct the distribution $F_{X_1}(x_1) = \mathsf{P}(X_1 \leq x_1)$, we let all $x_i, i \neq 1$ tend to infinity and this entails

Definition 12.5 (Marginal distribution functions of F)

Let $(X_1, \dots, X_n)$ be a random vector with distribution function F. Then, the marginal distribution functions (margins) of F are F_{X_i} for $i = 1, \dots, d$ where for each $i \in \{1, \dots, d\}$

$$F_{X_i}(x_i) = \lim_{\substack{x_j \to \infty \\ j \neq i}} F(X_1, \dots, X_n), x_i \in \mathbb{R}$$

Let us continue our example.

$$F_X(x) = \lim_{y \to \infty} F(x, y) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \in [0, 1) \\ 1 & \text{if } x \geq 1 \end{cases}$$

and we recover $X \sim \mathcal{U}(0, 1)$ and similarly $Y \sim \mathcal{U}(0, 1)$. Remember that even though X, Y have density, the bivariate distribution does not. We'll stress this a little latter.

Example 12.5

$F(X, Y) = (1 - e^{-x})(1 - e^{-y})$ for $x > 0, y > 0$. Then $\lim_{y \to \infty} F(x, y)$ is equal to 0 if $x < 0$ and $1 - e^{-x}$ if $x \geq 0$. In this case, $X \sim \mathcal{E}(1)$ and $Y \sim \mathcal{E}(1)$s.

Example 12.6

Suppose a vector $(X, Y) \sim \mathcal{M}(p_1, p_2, n)$. The support

$$S = \{(k, l) : k \in \{0, \dots, n\}, l \in \{0, \dots, n\}, k + l \leq n\}$$

The support of X is the projection on the X axis $X \in \{0 \dots 4\}$ and $\mathsf{P}(X = 0) = \mathsf{P}((X, Y) \in \{(0, x), x \in \mathbb{R}\}$.

Lemma 12.6

Let $(X_1, \dots, X_n)$ be a discrete random vector with PMF f and support S.

Then for all $i \in \{1, \dots, d\}$, X_i is discrete with support

$$S_{X_i} = \pi + i(S) = \{x \in \mathbb{R} : \exists s \in S \text{ such that } s_i = x\}$$

Furthermore, for all $x \in S_{X_i}, \mathsf{P}(X_i = x) = \sum_{s \in S, s_i = x} \mathsf{P}(X_i = s_i, \dots, X_d = s_d)$. We could also make it more general.

Remark

$$f_{X_i}(x) = \sum_{\substack{(X_1, \dots, X_n) \in \mathbb{R}^d \\ x_i = x}} f(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_d) \qquad \text{for } i \in \{1, \dots, d\}$$

Let us continue our example. If we look at $\mathsf{P}(X = k), k = \{0, \dots, n\}$, this is $f_x(k)$ which equals

$$\mathsf{P}(X = k) = \sum_{l=0}^{n-k} \binom{n}{k, l, (n-k-l)} p_1^k p_2^l (1 - p_1 - p_2)^{n-k-l}$$

Recall that the multinomial coefficient $\binom{n}{k,l,(n-k-l)} = \frac{n!}{k!l!(n-k-l)!}$ which entails

$$\begin{aligned}
&= \frac{n!}{k!(n-k)!} p_1^k \sum_{l=0}^{n-k} \frac{(n-k)!}{l!(n-k-l)!} p_2^l (1 - p_1 - p_2)^{n-k-l} \\
&= \binom{n}{k} p_1^k (p_2 + 1 - p_1 - p_2)^{n-k} \\
&= \binom{n}{k} p_1^k (1 - p_1)^{n-k}
\end{aligned}$$

Hence $X \sim \mathcal{B}(n, p_1)$ and $Y \sim \mathcal{B}(n, p_2)$.

Lemma 12.7

Let $(X_1, \dots, X_n)$ be a random vector with density f. Then, for all $i \in \{1, \dots, d\}$, X_i has a density given by

$$f_{X_i}(x) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{d-1} f(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_d) \underbrace{dx_1, \dots, dx_d}_{d-1}$$

Proof For the case $d = 2$. Consider wlog X_1. We have

$$\begin{aligned} F_{X_1}(x_1) &= \lim_{x_2\to\infty} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(s,t)dtds \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f(s,t)dtds = \int_{-\infty}^{x_1} f(x_1)(s)ds \end{aligned}$$

If we wanted X_2, swap the integral (since the function integrate to 1 and is positive). □

Example 12.7

Let $f(x,y) = e^{-y-x}, x > 0$ & $y > 0$. If $y \leq 0, f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = 0$. Else,

$$\int_0^{\infty} e^{-x-y}dx = e^{-y}\int_0^{\infty} e^{-x}dx = e^{-y}$$

if y is positive.

Example 12.8 (Margins of Multinormal distribution)

Let $(X_1, \ldots, X_n)$ and define $\mathbf{X} = (X_1 \ldots X_d)^\top$, the expectation vector $\boldsymbol{\mu} = (\mu_1 \ldots \mu_d)^\top$, and the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ positive definite and symmetric. The density is given by

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{|\Sigma|} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

Suppose we consider a simple case with $\boldsymbol{\mu} = \left(\begin{smallmatrix} \mu_1 \\ \mu_2 \end{smallmatrix}\right)$ and $\Sigma = \left(\begin{smallmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{smallmatrix}\right)$, where $\rho \in (-1,1)$ and $\sigma_1, \sigma_2 > 0$. The joint probability density is $f(x,y)$ is

$$= \frac{1}{(2\pi)} \frac{1}{\sigma_1\sigma\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{{\sigma_1}^2} + \frac{(y-\mu_2)^2}{{\sigma_2}^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right)}$$

If we begin to play with the integral, we can obtain $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$

which is

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(- \frac{1}{2(1-\rho^2)} \frac{(x-\mu_1)^2}{{\sigma_1}^2} \right) \\
&\times \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_2 \sqrt{1-\rho^2}} \exp\left(- \frac{1}{2(1-\rho^2){\sigma_2}^2} \left(y - \mu_2 - \frac{\rho\sigma_2}{\sigma_1}(x-\mu_1) \right) \right).
\end{aligned}
$$

Thus, by completing the square on

$$\left(y - \mu_2 - \frac{\rho\sigma_2}{\sigma_1}(x-\mu_1) \right),$$

we get that this part is equal to

$$\left(y - \underbrace{\mu_2 - \frac{\rho\sigma_2}{\sigma_1}(x-\mu_1)}_{\text{constant}} \right)^2 - \frac{\rho^2{\sigma_2}^2}{{\sigma_1}^2}(x-\mu_1)^2 = y - \mu^*$$

The final result will be

$$
\begin{aligned}
f_X(x) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_1} \exp\left(- \frac{1}{2(1-\rho^2)}(x-\mu_1)^2 + \frac{\rho^2{\sigma_2}^2}{2(1-\rho^2){\sigma_2}^2}(x-\mu_1)^2 \right) \\
&\times \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma^*} \exp\left(- \frac{1}{2{\sigma^*}^2}(y-\mu^*) \right) dy \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_1} \exp\left(- \frac{1}{2} \frac{{(x-\mu_1)^2}^2}{\sigma_1} \right)
\end{aligned}
$$

where $\sigma^* = \sigma^2\sqrt{1-\rho^2}$. Note that the second term in the product integrated to 1 as we recovered a density function (namely that of the normal distribution).

Example 12.9

$$f(x,y) = \begin{cases} 1 & \text{if } x \in (0,1) \text{ and } y \in (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

Then the margin

$$f_{X_1}(x) = \begin{cases} 0 & \text{if } x \notin (0,1) \\ \displaystyle\int_0^1 dy & \text{if } x \in (0,1) \end{cases}$$

Hence $X \sim \mathcal{U}(0,1)$ and similarly $Y \sim \mathcal{U}(0,1)$.

Remark

Careful! The joint distribution determines the marginal distribution, but *not conversely*, *id est* the knowledge of the marginal distribution is not enough to reconstruct the joint distribution (an illustration of this fact is to compare Examples 12.4 and 12.9).

Chapter 13

Independence of random variables

Recall that in the case of events, we had independence if $\mathsf{P}(A \cap B) = \mathsf{P}A\mathsf{P}B$. For two random variables X, Y, we need something of the form

$$\mathsf{P}(X \in B_1, Y \in B_2) = \mathsf{P}(X \in B_1)\mathsf{P}(Y \in B_2)$$

for any $B_1, B_2 \in \mathbb{B}$.

Definition 13.1 (Independence)

The random variables X_1 and X_2 are called *independent* if their joint distribution function $F(X_1, X_2)$ is of the form

$$F_{(X_1,X_2)}(x_{,1}, x_2) = F_{X_1}(x_1)F_{X_2}(x_2) \quad \forall\, x_1, x_2 \in \mathbb{R}$$

Example 13.1

Recall the examples of bivariate exponential distribution we had earlier. It is clear that

$$F(x, y) = (1 - e^{-x})(1 - e^{-y})1_{(x \geq 0, y \geq 0)} = F_X(x)F_Y(y)$$

Observation If we have $B_1 = (a, b], B_2 = (c, d]$, then

$$\begin{aligned} \mathsf{P}(X \in (a, b], y \in (c, d]) &= F(b, d) - F(b, c) + F(a, d) + F(a, c) \\ &= (F_X(b) - F_X(a))\,(F_Y(d) - F_Y(c)) \\ &= \mathsf{P}(X \in (a, b]) \times \mathsf{P}(Y \in (c, d]) \end{aligned}$$

if X and Y are independent. That is, if X and Y are independent, then $\mathsf{P}(X \in B_1, Y \in B_2) = \mathsf{P}(X \in B_1)\mathsf{P}(Y \in B_2)$. If $B_1 = (-\infty, x)$ and $B_2 = (-\infty, y]$, then we get the converse.

Definition 13.2

The variables $X_i, i \in I$ where I is an arbitrary index set, are called *indepen-*

dent if for any $k \in \mathbb{N}$ and any collection $i_1, \dots, i_k \in I$, then

$$F_{(X_{i_1},\dots,X_{i_k})}(x_{i_1}, \dots, x_{i_k}) = \prod_{j=1}^{k} F_{X_{ij}}(x_{ij}), \quad (x_{i_1}, \dots, x_{i_k}) \in \mathbb{R}^k \tag{13.1}$$

PROOF Independence clearly implies 13.1 (it is only a special case). We need to show the equation is equivalent to independence.
Let $k \in \{2, \dots, d\}, i_1, \dots, i_k \in \{1, \dots, d\}$. Then $\mathsf{P}(X_{i_1} \leq x_{i_1}, \dots, X_{i_k} \leq x_{i_k})$

$$\begin{aligned}
&= \lim_{\substack{x_j \to \infty \\ j \notin \{i_1,\dots,i_k\}}} \mathsf{P}(X_1 \leq x_1, \dots, X_d \leq x_d) \\
&= \lim_{\substack{x_j \to \infty \\ j \notin \{i_1,\dots,i_k\}}} \prod_{j=1}^{d} F_{X_i}(x_i) \\
&= \lim_{\substack{x_j \to \infty \\ j \notin \{i_1,\dots,i_k\}}} \prod_{j \notin \{i_1,\dots,i_k\}} F_{X_j}(x_j) \left(\prod_{j \in \{i_1,\dots,i_k\}} F_{X_j}(x_j) \right)
\end{aligned}$$

□

Example 13.2 (Farlie-Gumbel-Morgenstein distributions)

Let F_1, F_2, F_3 be some fixed univariate distribution functions.
For $\alpha \in (-1, 1)$, we have $F(x_1, x_2, x_3)$

$$= F_1(x_1)F_2(x_2)F_3(x_3) + \alpha F_1(x_1)\left(1 - F_1(x_1)\right)\left(1 - F_2(x_2)\right)\left(1 - F_3(x_3)\right)$$

which is a valid trivariate distribution function. The margins of F are F_1, F_2, F_3, and they are not independent unless $\alpha = 0$. However, if we look at the distribution of (X_1, X_2), it has distribution function $F_1(x_1)F_2(x_2)$, that is X_1 and X_2 are independent. This holds for all pairs X_i and X_j, for $i, j\ in\{1, 2, 3\}, i \neq j$.

Theorem 13.3 (Conditions for independence)

Let $(X_1, \dots, X_n)$ be a random vector. Then

1. $X_1, \dots, X_n$ are independent if and only if

$$f_{(X_1,\dots,X_n)}(x_1, \dots, x_n) = f_{X_1}(x_1) \times \cdots \times f_{X_d}(x_d) \tag{13.2}$$

   for $x_1, \dots, x_n \in \mathbb{R}$ provided $(X_1, \dots, X_n)$ has a probability mass function $f_{X_1,\dots,X_n}$.

2. $X_1, \dots, X_n$ are independent if and only if 13.2 holds almost everywhere provided $(X_1, \dots, X_n)$ has density $f(X_1, \dots, X_n)$.

Proof Necessity. We have $f_{(X_1,\dots,X_n)}(x_1, \dots, x_n)$

$$\begin{aligned}&\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_{(X_1,\dots,X_n)}(t_1, \dots, t_d) dt_d \dots dt_1 \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_{X_i}(t_i) dt_i \\ &= \prod_{i=1}^{d} \int_{-\infty}^{x_i} f_{X_i}(t_i) dt_i \\ &= \prod_{i=1}^{d} F_{X_i}(x_i)\end{aligned}$$

Sufficiency.

$$\frac{F_{(X_1,\dots,X_n)}(x_1, \dots, x_n)}{\partial x_1 \dots \partial x_d} = \frac{F_{X_1}(x_1)}{\partial x_1} \cdots \frac{F_{X_d}(x_d)}{\partial x_d}$$

$$f_{(X_1,\dots,X_n)}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_d}(x_d)$$

and

$$f_{(X_1,\dots,X_n)}(x_1, \dots, x_n) = \lim_{\varepsilon \to 0} \mathsf{P}\Big(X_1 \in [x_1, X_1 + \varepsilon), \dots, X_d \in [x_d, X_d + \varepsilon)\Big)$$

□

Example 13.3

1. (X, Y) with density

$$e^{-x-y}1_{(x>0)}1_{(y>0)} = e^{-x}1_{(x>0)} \cdot e^{-y}1_{(y>0)}$$

Clearly, X and Y are independent exponential.

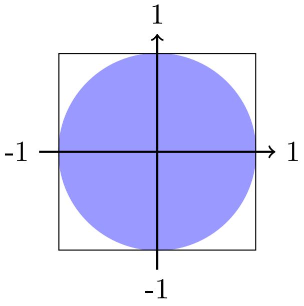
2. $(X, Y) \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{pmatrix}\right)$. If $\rho = 0$, then X and Y are indeed independent.

3.

$$f_{(X,Y)}(x, y) = \binom{n}{x, y, (n-x-y)} p_1^x p_2^y (1-p_1-p_2)^{n-x-y}$$

and 0 otherwise if $x, y \in \{0, \ldots, n\}$ and $x+y \leq n$. Then $X \sim \mathcal{B}(n, p_1)$ and $Y \sim \mathcal{B}(n, p_2)$ are not independent $x, y \in \{0, \ldots, n\}$ and $x, y \leq n$. X and Y are not independent because the support of (X, Y) is not $\{0, \ldots, n\} \times \{0, \ldots, n\}$.

4. Let (X, Y) as in picture have joint distribution in the circle. They cannot be possibly independent.



Theorem 13.4

$X_1, \ldots, X_n$ are independent if and only if for any Borel sets B_1, B_d, we have that

$$\mathsf{P}(X_1 \in B_1, \ldots, X_d \in B_d) = \prod_{i=1}^{d} \mathsf{P}(X_i \in B_i)$$

Corollary 13.5

Suppose $X_1, \ldots, X_n$ are random variables and $g_1, \ldots, g_d$ are (Borel) measurable functions, $g_i : \mathbb{R} \to \mathbb{R}$. Then, if $X_1, \ldots, X_n$ are independent, $g_1(X_1), \ldots, g_d(X_d)$ are also independent. To see this, look at

$$\begin{aligned}\mathsf{P}\left(g_1(X_1) \in B_1, \ldots, g_d(X_d) \in B_d\right) \\ = \mathsf{P}\left(X_1 \in {g_1}^{-1}(B_1), \ldots, X_d \in {g_d}^{-1}(B_d)\right)\end{aligned}$$

remarking that each $g^{-1}(B_i)$ is again a Borel set. For this intersection of events, we apply Theorem 13.4 to get that the above is equal to

$$= \prod_{i=1}^{d} \mathsf{P}\left(X_i \in {g_i}^{-1}(B_i)\right) = \prod_{i=1}^{d} \mathsf{P}\left(g(X_i) \in (B_i)\right)$$

Example 13.4

Consider (X, Y) with density $f(x,y) = \frac{1+xy}{4} 1_{(|x|<1)} 1_{(|y|<1)}$. The marginal

$$f_X(x) = \int_{-1}^{1} \frac{1+xy}{4} dy = \frac{1}{4} \left\{ 2 + x \underbrace{\int_{-1}^{1} y dy}_{\text{odd function}} \right\} = \frac{1}{2} 1_{(x \in (-1,1))}$$

so $X \sim \mathcal{U}(-1,1)$ and by symmetry $Y \sim \mathcal{U}(-1,1)$ and the two are not independent. If we look however at the transformation X^2 and Y^2, we will see that they are independent.

$$\mathsf{P}\left(X^2 \leq u\right) = \mathsf{P}(-\sqrt{u} \leq x \leq \sqrt{u}) = \begin{cases} \int_{-\sqrt{u}}^{\sqrt{u}} \frac{1}{2} dt = \sqrt{u} & \text{if } u \in [0,1] \\ 1 & \text{if } u \geq 0 \\ 0 & \text{if } u < 0 \end{cases}$$

and

$$\mathsf{P}\left(X^2 \leq v\right) = \begin{cases} \sqrt{v} & \text{if } v \in [0,1] \\ 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0 \end{cases} .$$

Thus,

$$\mathsf{P}\left(X^2 \le u, Y^2 \le v\right) = \begin{cases} 0 \text{ if } u < 0, v < 0 \\ \mathsf{P}(-\sqrt{u} \le x \le \sqrt{u}, -\sqrt{v} \le y \le \sqrt{v}) \quad \Diamond \end{cases}$$

where $\Diamond$ is the following:

$$\Diamond = \begin{cases} 1 \text{ if } u \ge 1 \text{ and } v \ge 1 \\ \displaystyle\int_{-\sqrt{u}}^{\sqrt{u}} \int_{-\sqrt{v}}^{\sqrt{v}} \frac{1+xy}{4} dydx = \frac{1}{4}(2\sqrt{u}2\sqrt{v} + \cancel{\int_{-\sqrt{u}}^{\sqrt{u}} \int_{-\sqrt{v}}^{\sqrt{v}} xydydx} = \sqrt{u}\sqrt{v}. \end{cases}$$

If $u \in [0, 1)$ and $v \ge 1$, then $\mathsf{P}(X^2 \le u) = \sqrt{u}$. If $v \in [0, 1)$ and $u \ge 1$, then $\mathsf{P}(Y^2 \le v) = \sqrt{v}$. Hence,

$$\mathsf{P}(\Big(X^2 \le u, Y^2 \le v\Big) = \begin{cases} 0 & \text{if } u < 0 \text{ or } v < 0 \\ \sqrt{u}\sqrt{v} & \text{if } u \in [0,1) \text{ and } v \in [0,1) \\ \sqrt{u} & \text{if } u \in [0,1) \text{ and } v \ge 1 \\ \sqrt{v} & \text{if } u \ge 1 \text{ and } v \in [0,1) \\ 1 & \text{if } u \ge 1, v \ge 1 \end{cases}$$

which indeed equals $\mathsf{P}(X^2 \le u)\mathsf{P}(Y^2 \le v)$ for all $u, v \in \mathbb{R}$. The morale of the story: $X_1, \ldots, X_n$ independent imply $g_1(x_1), \ldots, g_d(x_d)$ independent, but if $X_1, \ldots, X_n$ are dependent, then $g_1(x_1), \ldots, g_d(x_d)$ may or not be independent.

Definition 13.6 (Independently and identically distributed r.v.)

A sequence of random variables $X_1, X_2, X_3 \ldots$ is called independent and identically distributed (i.i.d.) sequence if $\{X_i, i \in \mathbb{N}\}$ are independent and X_i has the same distribution for every i.

Chapter 14

Transformation of random vectors

Let $(X_1, \dots, X_n)$ be a random vector. If we consider a measurable mapping $g : \mathbb{R}^d \to \mathbb{R}^m$ measurable, then $g(X_1, \dots, X_n) = (Y_1, \dots, Y_m)$ is again a random vector.

Example 14.1

Suppose that $X_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_d$, where $\perp\!\!\!\perp$ is the notation for mutually independent. We could be interested in the convolution $X_1 + X_2 + \cdots + X_d$; this could arise for example in the insurance (the sum of all loses due to claims.) If $(X_1, \dots, X_n)$ is discrete, then in this case $(Y_1, \dots, Y_m)$ is also discrete (since $g(X)$ takes at most countably many values). Hence, the distribution of $(Y_1, \dots, Y_m)$ is given by its probability mass function.

Example 14.2

Consider the example from earlier of a toss of two dice. We have (X_1, X_2) and $Y = X_1 + X_2$. Thus, the probability of obtaining a sum of k is given by $P(Y = k) = \sum_{(i,j):i+j=k} \mathsf{P}(X_1 = i, X_2 = j)$. The probability mass function of $(Y_1, \dots, Y_m)$ is given by

$$f_{\mathbf{Y}}(y_1, \dots, y_m) = \sum_{\substack{(X_1,\dots,X_d) \in \mathbb{R}^d: \\ g(X_1,\dots,X_d) = (y_1,\dots,y_m)}} f_{\mathbf{X}}(x_1, \dots, x_d)$$

where $f_{\mathbf{X}}$ is the PMF of $\mathbf{X} = (X_1, \dots, X_d)$ and

$$f_{\mathbf{Y}} = \mathsf{P}(Y_1 = y_1, \dots, Y_m = y_m)$$

but we are really running through the support of $\mathbf{X}$ of

$$f_{\mathbf{Y}}(y_1, \dots, y_d) = \sum_{\substack{(X_1,\dots,X_d) \in \{(x_1,\dots,x_d) > 0\} \\ \text{s.t. } g(x_1,\dots,x_d) = (y_1,\dots,y_m)}} f_{\mathbf{X}}(x_1, \dots, x_d)$$

Example 14.3

Convolution of two binomial: $X_1 \sim \mathcal{B}(n, p)$, $X_2 \sim \mathcal{B}(n, p)$ and $X_1 \perp\!\!\!\perp X_2$, $Y = X_1 + X_2$. We would like to see $Y \sim \mathcal{B}(2n, p)$. To prove this, consider the support of Y, which is $\{0,\dots,2n\}$. Hence $f_Y(x)=0$ if $x\notin\{0,\dots,2n\}$. If however $x\in\{0,\dots,2n\}$,

$$\begin{aligned}
f_Y(x) &= \mathsf{P}(Y=x)=\mathsf{P}(X_1+X_2=x)\\
&=\sum_{\substack{(i,j):i,j\in\{0,\dots,n\}\\ i+j=x}} f_{X_1}(i)f_{X_2}(j)\\
&=\sum_{i=0}^{x} f_{X_1}(i)f_{X_2}(x-i)\\
&=\sum_{i=0}^{x}\binom{n}{i}p^i(1-p)^{n-i}\cdot\binom{n}{x-i}p^{x-i}(1-p)^{n-x+i}\\
&=p^x(1-p)^{2n-x}\sum_{i=0}^{x}\binom{n}{i}\binom{n}{x-i}\\
&=p^x(1-p)^{2n-x}\binom{2n}{x}\sum_{i=0}^{x}\frac{n!n!x!(2n-x)!}{i!(n-i)!(x-i)!(n-x+i)!2n!}\\
&\Rightarrow p^x(1-p)^{2n-x}\binom{2n}{x}\sum_{i=0}^{x}\frac{\binom{x}{i}\binom{2n-x}{n-i}}{\binom{2n}{n}}
\end{aligned}$$

where the sum is the PMF of the hypergeometric distribution(number of defectives in n sample of size $2n$ hence

$$f_Y(x)=p^x(1-p)^{2n-x}\binom{2n}{x}$$

Example 14.4

Say $X_1\sim\mathcal{B}(n,p)$, $X_2\sim\mathcal{B}(n,p)$ and $(X_1,X_2)\sim\mathcal{M}(n,p_1,p_2)$. Then $\mathsf{P}(X_1+X_2=k)=0$ if $k\notin\{0,\dots,n\}$. Note that the support of Y is not anymore the same as before; we look at balls in category 1 and 2). If $k\in\{0,\dots,n\}$

For, $\mathsf{P}(X_1 + X_2 = k)$

$$\begin{aligned}
&= \sum_{i=0}^{k} f_{X_1,X_2}(i, k-i) = \sum_{i=0}^{k} \frac{n!}{i!(k-1)!(n-k)!} p^i p^{k-i} (1-2p)^{n-k} \\
&= p^k (1-2p)^{n-k} \binom{n}{k} \sum_{i=0}^{k} \binom{k}{i} \Rightarrow (1+1)^k \\
&= 2^k p^k (1-2p)^{n-k} \binom{n}{k} \\
&= (2p)^k (1-2p)^{n-k} \binom{n}{k}
\end{aligned}$$

and $X_1 + X_2 \sim \mathcal{B}(n, 2p)$. The second case arise if $(X_1, \ldots, X_n)$ has density $f_{\boldsymbol{X}}$. We must be careful as $(Y_1, \ldots, Y_m) = g(X_1, \ldots, X_n)$ may be arbitrary (it may have a density, be discrete or neither). Computing $\mathsf{P}(Y_1 \le y_1, \ldots, Y_m \le y_m)$ could be done in the following fashion:

$$\underset{\substack{(x_1,\ldots,x_n)\in\mathbb{R}^d: \\ g(X_1,\ldots,X_n)\in\Pi_{i=1}^d[-\infty,y_i)}}{\int \cdots \int} f_{bX}(x_1, \ldots, x_n) dx_1 \ldots dx_d$$

Example 14.5

(X_1, X_2) has $f(X_1, X_2) = e^{-X_1} 1_{(0 < X_2 \le X_1 < \infty)}$ and let $Y = X_1 - X_2$. We look at the distribution function of Y

$$\begin{aligned}
\mathsf{P}(Y \le y) &= \iint_{(X_1,X_2):X_1 - X_2 \le y} e^{-x_1} 1_{(0 < x_2 \le x_1 \le \infty)} dx_1 dx_2 \\
&= \int_0^\infty \int_{x_1 - y}^{x_1} e^{-x_1} dx_2 dx_1
\end{aligned}$$

will be zero if $x_1 \le x_2 - y \Leftrightarrow y \ge 0$. If $y > 0$, two possibilities: either $x_1 < y$

or $x_1 \geq y$

$$
\begin{aligned}
&\int_0^y \int_0^{x_1} e^{-x_1} dx_2 dx_1 + \int_y^\infty \int_{x_1-y}^{x_1} e^{-x_1} dx_2 dx_1 \\
=&\int_0^y e^{x_1} x_1 dx_1 + \int_y^\infty e^{x_1}(x_1 - x_1 + y) dx_1 \\
=&\int_0^y e^{x_1} x_1 dx_1 + y \int_y^\infty e^{x_1} dx_1 \\
=&- x_1 e^{-x_1}\Big|_0^y + \int_0^y e^{-x_1} + ye^{-y} \\
=&- ye^{-y} + ye^{-y} + 1 - e^{-y}
\end{aligned}
$$

and $Y \sim \mathcal{E}(1)$.

Theorem 14.1 (Transformation theorem)

Let $(X_1, \ldots, X_n)$ and $g : A \in \mathbb{R}^d \to b \in \mathbb{R}^d$ (dimensionality of spaces must match) measurable. If

1. $(X_1, \ldots, X_n)$ has density $f_{\boldsymbol{X}}$;
2. $A \in \mathbb{R}^d$ open and such that $\underset{A}{\int \cdots \int} f_{\boldsymbol{X}} dx_1 \ldots dx_d = \mathsf{P}((x_1, \ldots, x_d) \in A) = 1$;
3. The function

$$g(x_1, \ldots, x_n) \mapsto (g_1(x_1, \ldots, x_n), \ldots, g_d(x_1, \ldots, x_n))$$

   is one-to-one.

4. $h : B \to A$ and $(y_1, \ldots y_d) \mapsto h_1(y_1, \ldots y_d), \ldots, h_d(y_1, \ldots y_d)$, namely the inverse mapping of g is such that $\forall\ i \in \{1, \ldots, d\}$, the partial derivative $\frac{\partial h_i}{\partial x_j}$ exist and are continuous $\forall\ j \in \{1, \ldots, d\}$; h is continuously partially differentiable.

5. The Jacobian is not identically zero

$$\begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_d}{\partial y_1} & \cdots & \frac{\partial h_d}{\partial y_d} \end{vmatrix} \neq 0$$

Then $\boldsymbol{Y} = g(X_1, \ldots, X_n)$ has density $f_{bY}(y_1, \ldots y_d)$

$$= \begin{cases} 0 \\ f_{\boldsymbol{X}}\Big(h_1(y_1, \ldots y_d), \ldots, h_d(y_1, \ldots y_d)\Big)|J| \end{cases}$$

if respectively if $(y_1, \ldots y_d) \notin B$ and $(y_1, \ldots y_d) \in B$ where $|J|$ is the absolute value of the Jacobian.

Proof $\mathsf{P}(Y_1 \leq y_1, \ldots, Y_d \leq y_d)$

$$\begin{aligned}
&= \mathsf{P}((X_1, \ldots, X_n) \in g(-\infty, y_1] \times \ldots \times (-\infty, y_d]) \\
&= \int f_{\boldsymbol{X}}(t_1, \ldots, t_d) dt_1 \ldots dt_d \\
&= \int_{-\infty}^{y_1} \ldots \int_{-\infty}^{y_d} f_{\boldsymbol{Y}}(s_1, \ldots, s_d) ds_1 \ldots ds_d \\
&= \int_{-\infty}^{y_1} \ldots \int_{-\infty}^{y_d} f_{\boldsymbol{X}}(h_1(s_1, \ldots, s_d), \ldots, h_d(s_1, \ldots, s_d))|J| ds_1 \ldots ds_d
\end{aligned}$$

by using in the third step the transformation $t_i = h_1(s_1, \ldots, s_d)$ □

Example 14.6

X_1, X_2, X_3 independent and $X_i \sim \mathcal{E}(1)$.

$$f_{\boldsymbol{X}}(x_1, x_2, x_3) = e^{-x_1 - x_2 - x_3} 1_{(x_1>0, x_2>0, x_3>0)}$$

and

$$\boldsymbol{Y} = (Y_1, Y_2, Y_3) = \left(X_1 + X_2 + X_3, \frac{X_1 + X_2}{X_1 + X_2 + X_3}, \frac{X_1}{X_1 + X_2}\right)$$

and $g : (0,\infty)^3 \to \mathbb{R}^3$ and

$$(x_1, x_2, x_3) \mapsto \left(x_1 + x_2 + x_3, \frac{x_1 + x_2}{x_1 + x_2 + x_3}, \frac{x_1}{x_1 + x_2} \right)$$

The inverse function

$$h : (y_1, y_2, y_3) \mapsto (y_1 y_2 y_3, y_1 y_2 - y_1 y_2 y_3, y_1 - y_1 y_2)$$

and the Jacobian is given by

$$J = \det \begin{pmatrix} y_2 y_3 & y_1 y_3 & y_1 y_2 \\ y_2 - y_2 y_3 & y_1 - y_1 y_3 & -y_1 y_2 \\ 1 - y_2 & -y_1 & 0 \end{pmatrix}, \quad |J| = -y_1^2 y_2$$

and using the Transformation theorem $f_{\boldsymbol{Y}}(y_1, y_2, y_3)$

$$\begin{aligned} &= e^{-y_1 y_2 y_3 - y_1 y_2 + y_1 y_2 y_3 - y_1 + y_1 y_2} | - y_1^2 y_2 | 1_{(y_1 y_2 y_3 > 0, y_1 y_2 y_3 < y_1 y_2, y_1 y_2 < y_1)} \\ &= e^{-y_1} y_1^2 y_2 1_{(y_1 > 0)} 1_{(y_2 \in (0,1))} 1_{(y_3 \in (0,1))} \end{aligned}$$

can be factored in three parts

$$\frac{1}{2} e^{-y_1} y_1^2 1_{(y_1 > 0)} \quad 2 y_2 1_{(y_2 \in (0,1))} \quad 1_{(y_3 \in (0,1))}$$

so $Y_1 \perp\!\!\!\perp Y_2 \perp\!\!\!\perp Y_3$ and $Y_1 \sim \Gamma(3), Y_2 \sim$ Beta(2,1) and $Y_3 \sim$ Uniform(0,1).

Example 14.7

Let $X_1 \sim \mathcal{U}(0,1), X_2 \sim \mathcal{U}(0,1), X_1 \perp\!\!\!\perp X_2$ and we are interested in the convolution $Y = X_1 + X_2$ is not from $\mathbb{R}^2 \to \mathbb{R}^2$, but we could invent a mapping $g : (0,1)^2 \to (0,1) \times (0,2)$ and g satisfies the conditions of the theorem and could look at the margin $\boldsymbol{Y} = (Y_1, Y_2) = (X_1, X_1 + X_2)$ and the inverse mapping $h : (y_1, y_2) \mapsto (y_1, y_2 - y_1)$. The Jacobian is given by $\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$ and $|J| = 1$ hence $f_{\boldsymbol{Y}}(y_1, y_2) = 1_{(y_1 \in (0,1))} 1_{y_2 - y_1 \in (0,1)}$. We have $f_{Y_2} = f_{X_1 + X_2}(y)$

$$= \int_0^1 1_{y - y_1 \in (0,1)} dy_1 = \begin{cases} y \in (0,1) & f_{X_1+X_2}(y) = y \\ y \in [1,2) & f_{X_1+X_2}(y) = 2 - y \end{cases}$$

More examples can be found in the book. Let us examine a bit convolutions. Let X, Y be random variables with $X \perp\!\!\!\perp Y$ and $Z = X + Y$. This is encountered in insurance; consider the sum of claims. If (X, Y) is a discrete random vector, then $\mathsf{P}(Z = z) \sum_x \mathsf{P}(X = x)\mathsf{P}(Y = Z - X) = f_X(c) f_Y(z - x)$ and there are only countably many such. If (X, Y) has density $f_X f_Y$, we could invent a mapping $g : \mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \mapsto (x, x + y)$ and $g^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$ has $(u, v) \mapsto (u, v - u)$ and the Jacobian is $\left(\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right)$ and $|J| = 1$. The density of $(X, X + Y)$ is

$$f_{(X,X+Y)}(u, v) = f_X(u) f_Y(v - u) \cdot 1$$

However, our goal is the distribution of $X + Y$ is

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(u) f_Y(z - u) du.$$

We could have looked at the mapping $Y, X + Y = \int_{-\infty}^{\infty} f_Y(u) f_X(z - u) du$. Is this the end of the story? Not really, since we could get a ugly integral. Here is a great trick. If we look at the moment generating function of the characteristic function

$$M_{X+Y}(t) = \mathsf{E}e^{t(X+Y)} = \mathsf{E}e^{tX} \cdot e^{tY} = \mathsf{E}e^{tX}\mathsf{E}e^{tY} = M_X(t) M_y(t)$$

which is

$$\begin{cases} \displaystyle\sum_x \sum_y e^{tx} e^{ty} f_X(x) f_Y(y) \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{tx} e^{ty} f_X(x) f_Y(y) dx dy \end{cases}$$

Example 14.8

The moment generating function of $\mathcal{B}(n, p)$ is $(1 - p + pe^t)^n$. If we are looking at the sum of two independent variables X and Y with $X + Y, X \perp\!\!\!\perp Y, X \sim \mathcal{B}(n, p)$ and similarly $Y \sim \mathcal{B}(n, p)$, then the MGF of $X + Y$ is $(1 - p + pe^t)^{2n}$, which is the MGF of $\mathcal{B}(2n, p)$ and hence $X + Y \sim \mathcal{B}(2n, p)$ (note that we need the p to be the same to get this result).

Example 14.9

$X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\mu)$ and $X \perp\!\!\!\perp Y$. The MGF of X is $M_X(t) = e^{\lambda(e^t-1)}$ for $t \in \mathbb{R}$ and that of Y is $M_Y(t) = e^{\mu(e^t-1)}$. Thus

$$M_{X+Y}(t) = e^{(\mu+\lambda)(e^t-1)}$$

and the convolution $X = Y \sim \mathcal{P}(\mu + \lambda)$.

Example 14.10

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\nu, \tau^2)$, with $X \perp\!\!\!\perp Y$. Then, for $t \in \mathbb{R}$

$$\begin{aligned} M_X(t) &= e^{\mu t + \frac{\sigma^2 t^2}{2}} \\ M_Y(t) &= e^{\nu t + \frac{\tau^2 t^2}{2}} \\ M_{X+Y}(t) &= e^{(\mu+\nu)t + \frac{t^2}{2}(\sigma^2+\tau^2)} \end{aligned}$$

and $X + Y \sim \mathcal{N}(\mu + \nu, \tau^2 + \sigma^2)$.

Remark

If $X_1, \ldots, X_n$ are independent and $M_{X_i}(t)$ exists for $|t| < \varepsilon \; \forall \; i = \{1, \ldots, d\}$ or the minimum satisfies tis condition, then

$$M_{\boldsymbol{X}}(t) = \prod_{i=1}^{d} M_{X_i}(t)$$

This hold for general random variables, however we need to work with the Lebesque integral. Furthermore, if the moment generating function does not exist, we can do a similar trick with the characteristic function

$$C_{X_1,\ldots,X_n}(t) = \mathsf{E} e^{it(X_1,\ldots,X_n)} = \prod_{i=1}^{d} C_{X_i}(t)$$

Here are some more examples

Example 14.11

Suppose that $X \sim \mathcal{P}(\lambda)$; we have seen before that $M_X(t) = e^{\lambda(e^t-1)}$ for

$t \in \mathbb{R}$. But this is also

$$M_X(t) = e^{\lambda(e^t-1)} = e^{n\frac{\lambda}{n}(e^t-1)} = \prod_{i=1}^{d} e^{\frac{\lambda}{n}(e^t-1)}$$

which is the MGF of the $\mathcal{P}\left(\frac{\lambda}{n}\right)$ so $X \xrightarrow{\mathcal{D}} X_1 + \dots + X_n$ where $X_1, \dots, X_n$ are i.i.d. $\mathcal{P}\left(\frac{\lambda}{n}\right)$. This sum is precisely what the Central Limit theorem is approximating. The same trick works for the Negative binomial. This is called *infinite divisibility*.

Example 14.12

Let $X \sim \mathcal{E}(1)$, $Y \sim \mathcal{E}(1)$ and $X \perp\!\!\!\perp Y$. We want to look at $Z = \max(X, Y)$. In this case, we cannot use the transformation theorem since the mapping is not differentiable and we cannot find a continuously differentiable inverse. Also bad is $Z = XY1_{(xy>1)}$ which does not have a density. So we need to look at the distribution

$$\begin{aligned}\mathsf{P}(Z \leq z) &= \mathsf{P}(\max(X, Y) \leq z) = \mathsf{P}(X \leq z, Y \leq Z)\\ &= \mathsf{P}(X \leq z)\mathsf{P}(Y \leq z) = (1 - e^{-z})^2\end{aligned}$$

for $z > 0$ and 0 otherwise. There is no universal recipe.

Chapter 15

Correlation and moments

Consider $(X_1, \ldots, X_d) = \mathbf{X}$, $g : \mathbb{R}^d \to \mathbb{R}$ a measurable function (in the Borel sense). $g(\mathbf{X})$ is a random variable. We first go with a

Definition 15.1

Let $\mathbf{X}$ and g be as above. Then the expectation $\mathsf{E}(g(X_1, \ldots, X_d)$

$$= \begin{cases} \displaystyle\sum_{X_1,\ldots,X_d} g(x_1,\ldots,x_d) f_{\boldsymbol{X}}(x_1,\ldots,x_d) & \text{if } \boldsymbol{X} \text{ is discrete} \\ \displaystyle\int \cdots \int_{-\infty}^{\infty} g(x_1,\ldots,x_d) f_{\boldsymbol{X}}(x_1,\ldots,x_d) d_{X_1} \ldots d_{X_d} & \text{if } \boldsymbol{X} \text{ has density .} \end{cases}$$

provided that $\mathsf{E}|(g(X_1, \ldots, X_d)| < \infty$. To be specific

Definition 15.2 (Moments and central moments)

Suppose $d = 2$ and let (X, Y) be a random vector. Then

1. For $j, k \in \mathbb{N}, \mathsf{E}X^jY^k$ is called the moment of (X, Y) of order $(j + k)$, provided $\mathsf{E}|X|^j|Y|^k < \infty$. Note this is not unique for a moment of one and the same order.

2. For $j, k \in \mathbb{N}, \mathsf{E}(X - \mathsf{E}X)^j(Y - \mathsf{E}Y)^k$ is called the central moment of (X, Y) of order $(j + k)$, provided $\mathsf{E}|X - \mathsf{E}X|^j|Y - \mathsf{E}Y|^k < \infty$. Expectations must be finite; we will see this a little later.

Example 15.1

Suppose the random vector $(X, Y) \sim \mathcal{N}\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right), \rho \in (-1, 1)$ and recall

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

and therefore the second moment

$$\begin{aligned}
\mathsf{E}XY &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{x^2-2\rho xy+y^2}{2(1-\rho)^2}\right)dxdy \\
&= \int_{-\infty}^{\infty} y\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\underbrace{\int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}e^{-\frac{(x-\rho y)^2}{2(1-\rho^2)}}dxdy}_{\text{density of }\mathcal{N}(\rho y,1-\rho^2)} \\
&= \int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\rho ydy \\
&= \rho\underbrace{\int_{-\infty}^{\infty} y^2\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}}_{\mathsf{Var}\text{ of }\mathcal{N}(0,1)} = \rho
\end{aligned}$$

and we get the second line by completing the square.

Example 15.2

Say $(X,Y)\sim\mathcal{M}(n,p_1,p_2)$, that is they follow a Multinomial distribution.

$$\begin{aligned}
\mathsf{E}XY &= \sum\sum_{\substack{k,l\in\{0,\dots,n\}\\ k+l\leq n}} kl\binom{n}{k,l,n-k-l}p_1^k p_2^l(1-p_1-p_2)^{n-k-l} \\
&= \ldots = n(n-1)p_1p_2
\end{aligned}$$

The computation will be done in the assignment.

Theorem 15.3 (Hölder inequality)

Let $p,q>1$ and such that $\frac{1}{p}+\frac{1}{q}=1$. Then, if $\mathsf{E}|X|^p<\infty$ and $\mathsf{E}|Y|^q<\infty$ then

$$\mathsf{E}|XY|\leq\left(\mathsf{E}|X|^p\right)^{\frac{1}{p}}\left(\mathsf{E}|Y|^q\right)^{\frac{1}{q}}$$

and in particular for the case $p=q=2$ the so-called Cauchy-Schwartz inequality if $\mathsf{E}|X|^2<\infty$ and $\mathsf{E}|Y|^2<\infty$

$$\mathsf{E}|XY|\leq\sqrt{\mathsf{E}X^2}\sqrt{\mathsf{E}Y^2}$$

Proof If $\mathsf{E}|X|^p=0\Rightarrow x=0$ almost surely ($\mathsf{P}(X=0)=1$). This entail $\mathsf{E}|XY|=0$ and similarly if $\mathsf{E}|Y|^q=0$. If $\mathsf{E}|X|^p>0$ and $\mathsf{E}|Y|^q>0$, then recall the Young inequality

$$|xy| \leq \frac{|X|^p}{p} + \frac{|y|^q}{q} \qquad \forall\, x, y \in \mathbb{R}$$

We used something similar to show that variance existed if the second moments was finite. Let $\omega \in \Omega$

$$\begin{aligned} \left| \frac{X(\omega)Y(\omega)}{(\mathsf{E}|X|^p)^{\frac{1}{p}} (\mathsf{E}|Y|^q)^{\frac{1}{q}}} \right| &\leq \frac{|X(\omega)|^p}{p\mathsf{E}|X|^p} + \frac{|Y(\omega)|^q}{q\mathsf{E}|Y|^q} \\ \Leftrightarrow \mathsf{E} \left| \frac{X(\omega)Y(\omega)}{(\mathsf{E}|X|^p)^{\frac{1}{p}} (\mathsf{E}|Y|^q)^{\frac{1}{q}}} \right| &\leq \mathsf{E} \left(\frac{|X(\omega)|^p}{p\mathsf{E}|X|^p} + \frac{|Y(\omega)|^q}{q\mathsf{E}|Y|^q} \right) \end{aligned}$$

by first taking expectation of both sides. Now,

$$\begin{aligned} &\mathsf{E}\frac{|X(\omega)|^p}{p\mathsf{E}|X|^p} + \mathsf{E}\frac{|Y(\omega)|^q}{q\mathsf{E}|Y|^q} \\ &= \frac{\mathsf{E}|X|^p}{p\mathsf{E}|X|^p} + \frac{\mathsf{E}|Y|^q}{q\mathsf{E}|Y|^q} \\ &= \frac{1}{p} + \frac{1}{q} = 1 \end{aligned}$$

and noting that the denominator of the left hand side of the equation is fixed because Lebesque integral are monotone and the term is constant; hence we can by linearity take it out of the expectation to get our result since

$$\frac{\mathsf{E}|XY|}{(\mathsf{E}|X|^p)^{\frac{1}{p}} (\mathsf{E}|Y|^q)^{\frac{1}{q}}} \leq 1$$

□

Definition 15.4 (Covariance)

Let (X, Y) be a random vector. Then $\mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y)$ is called the covariance of (X, Y), denoted $\mathsf{Cov}(X, Y)$ provided it exists.

Lemma 15.5

1. $\mathsf{Cov}(X, Y)$ exists if $\mathsf{E}X^2 < \infty$ and $\mathsf{E}Y^2 < \infty$;

2. $\mathsf{Cov}(X,Y) = \mathsf{E}XY - \mathsf{E}X\mathsf{E}Y$.

Proof

1. Using the Cauchy-Schwartz inequality,

$$\begin{aligned}\mathsf{E}|X-\mathsf{E}X||Y-\mathsf{E}Y| &\le \sqrt{\mathsf{E}(X-\mathsf{E}X)^2}\sqrt{\mathsf{E}(Y-\mathsf{E}Y)^2}\\ &= \mathsf{Var}X\mathsf{Var}Y\end{aligned}$$

   so $\mathsf{Var}X < \infty \Leftrightarrow \mathsf{E}X^2 < \infty$.

2.

$$\begin{aligned}\mathsf{Cov}(X,Y) &= \mathsf{E}\Big(X-\mathsf{E}X)(Y-\mathsf{E}Y)\Big)\\ &= \mathsf{E}(XY - Y\mathsf{E}X - X\mathsf{E}Y + \mathsf{E}X\mathsf{E}Y)\\ &= \iint xy - Y\mathsf{E}X - X\mathsf{E}Y + \mathsf{E}X\mathsf{E}Y f_{X,Y}(x,y)dxdy\\ &= \iint xyf_{x,y}(x,y)dxdy - \mathsf{E}X\iint yf_{x,y}(x,y)dxdy\\ &\quad - \mathsf{E}Y\iint xf_{x,y}(x,y)dxdy + \mathsf{E}X\mathsf{E}Y\iint f_{x,y}(x,y)dxdy\\ &= \mathsf{E}(XY) - \mathsf{E}X\mathsf{E}Y - \mathsf{E}X\mathsf{E}Y + \mathsf{E}X\mathsf{E}Y\end{aligned}$$

□

Example 15.3

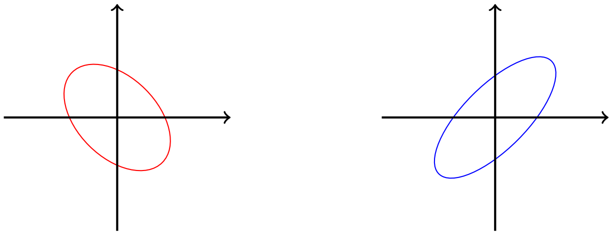
$(X,Y) \sim \mathcal{N}\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1&\rho\\ \rho&1\end{pmatrix}, \rho \in (-1,1)$, then $\mathsf{Cov}(X,Y) = \mathsf{E}XY - \mathsf{E}X\mathsf{E}Y = \rho$. On a computer illustration, we can see that if $\rho > 0$, X and Y are positively correlated and are said to be concordant. For $\rho < 0$, we say that X and Y are discordant.

Definition 15.6 (Pearson linear correlation coefficient)

Let (X,Y) be a random vector such that $\mathsf{E}X^2 < \infty$ and $\mathsf{E}Y^2 < \infty$. Then,

$$\frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}X\mathsf{Var}Y}} = \rho(X,Y)$$

Figure 19: Illustration of discordant and concordant random variables



is called the Pearson linear correlation coefficient. Our motivation is that ρ is a one number summary of the dependence between X and Y.

Lemma 15.7 (Properties of correlation)
Let (X, Y) be such that $\mathsf{E}X^2 < \infty$ and $\mathsf{E}Y^2 < \infty$. Then

1. $|\rho(X, Y)| \leq 1$;
2. If X and Y are independent, $\rho(X, Y) = 0$;
3. $\rho(-X, Y) = -\rho(X, Y)$;
4. $\rho(X, Y) = \rho(Y, X)$; symmetry of the covariance.
5. $\rho = \pm 1 \Leftrightarrow X = aY + b$ almost surely where $b \in \mathbb{R}, a \neq 0 \in \mathbb{R}$ and if $a > 0 \Rightarrow \rho = 1$, else $a < 0 \Rightarrow \rho = -1$.

The proofs are left as exercises.

Downsides of ρ

1. ρ is not always well-defined.
2. $\rho(X, Y) = 0 \nRightarrow X, Y$ are independent.

   Example 15.4
   $X \sim \mathcal{N}(0, 1), Y = X^2$. Then

   $$\mathsf{Cov}(X, Y) = \mathsf{E}XY - \mathsf{E}X\mathsf{E}Y = \mathsf{E}X^3 - \mathsf{E}X\mathsf{E}X^2 = 0 \Rightarrow \rho(X, Y) = 0.$$

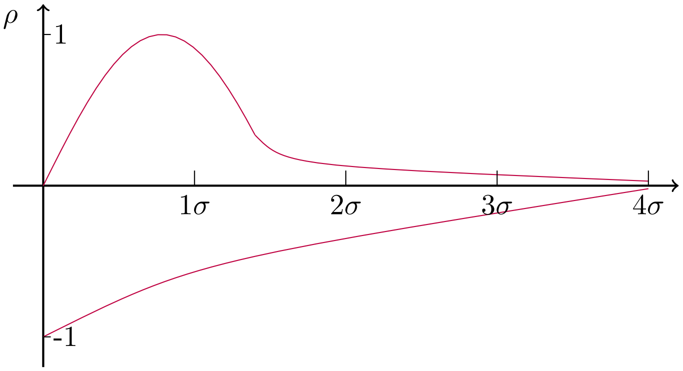
3. If ρ is close to 0, you cannot argue that X and Y are close to being independent.

4. If the marginal distributions of X and Y are given $\rho(X, Y)$ may not attain all values in $[-1, 1]$.

   ### Example 15.5

   This is drawn from the instructor work as a postdoc at ETH Zurich. A client, Swiss RE, which does refinancing for insurance company, asked for some information about possible values of the coefficient of correlation. Here is a possible reason for what is likely to be seen. Suppose as it is often the case in insurance that $X, Y \sim$ Log-normal, $X \sim \text{LN}(0, 1), X = e^Z, Z \sim \mathcal{N}(0, 1)$ and $Y \sim e^W, W \sim \mathcal{N}(0, \sigma^2)$. Then, their correlation coefficient. Then, the result is something of the form even though they are "strongly linked".

Figure 20: Approximate form of the dependence in insurance claim



Remark

If we are interested in $\mathsf{Cov}(X + Y, Z) = \mathsf{Cov}(X, Z) + \mathsf{Cov}(Y, Z)$. Also, $\mathsf{Cov}(a, X) = \mathsf{E}aX - a\mathsf{E}X = 0 \; \forall \; a \in \mathbb{R}$

Example 15.6

Let $X, Y \sim \mathcal{E}(1)$ and X and Y are independent. We are looking at $\rho\left(\frac{X}{X+Y}, \frac{Y}{X+Y}\right)$.

They don't have a density since $\frac{X}{X+Y} = 1 - \frac{Y}{X+Y}$ and would have covariance

$$\frac{\mathsf{Cov}\left(\frac{X}{X+Y}, 1 - \frac{X}{X+Y}\right)}{\sqrt{\mathsf{Var}\left(\frac{X}{X+Y}\right)}\sqrt{\mathsf{Var}\left(1 - \frac{X}{X+Y}\right)}} = -\frac{\mathsf{Cov}\left(\frac{X}{X+Y}, \frac{X}{X+Y}\right)}{\mathsf{Var}\left(\frac{X}{X+Y}\right)} = -1$$

Definition 15.8

Let $(X_1, \dots, X_n)$ be a random vector such that $\mathsf{E}X_i < \infty \ \forall \ i = 1, \dots, d$. Then

$$\mathsf{E}(X_1, \dots, X_n) = \mathsf{E}(\boldsymbol{X}) = \begin{pmatrix} \mathsf{E}X_1 \\ \vdots \\ \mathsf{E}X^d \end{pmatrix}$$

Here $\boldsymbol{X} = (X_1 \dots X_d)^\top$.

Theorem 15.9

Let $(X_1, \dots, X_n)$ such that $\mathsf{E}X_i < \infty \ \forall \ i = 1, \dots, d$. Then. for any $\mathbf{a} = (a_1 \dots a_d)^\top \in \mathbb{R}^{d \times 1}$

$$\mathsf{E}(\mathbf{a}^\top \boldsymbol{X}) = \mathsf{E}(a_1 X_1 + \dots + a_d X_d) = \mathbf{a}^\top \mathsf{E}\boldsymbol{X} = a_1 \mathsf{E}X_1 + \dots + a_d \mathsf{E}X_d$$

Proof Suppose $(X_1, \dots, X_n)$ has a density. Starting with the LHS,

$$\begin{aligned}
\mathsf{E}(\mathbf{a}^\top \boldsymbol{X}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (a_1 x_1 + \dots a_d x_d) f(x_1, \dots, x_n) dx_1 \dots dx_d \\
&= \sum_{i=1}^{d} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} a_i x_i f(x_1, \dots, x_n) dx_1 \dots dx_d \\
&= \sum_{i=1}^{d} \int_{-\infty}^{\infty} a_i x_i \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} a_i x_i f(x_1, \dots, x_n) \underbrace{dx_1 \dots dx_d}_{\text{not } i} dx_i}_{f_{X_i}(x_i)} \\
&= \sum_{i=1}^{d} a_i \mathsf{E}X_i
\end{aligned}$$

□

Theorem 15.10

If $(X_1, \ldots, X_n)$ are such that $\mathsf{E}|X_i| < \infty, i = 1, \ldots, d$ **and** $X_1, \ldots, X_n$ are **independent**, then

$$\mathsf{E}\left(\prod_{i=1}^{d} X_i\right) = \prod_{i=1}^{d} \mathsf{E}X_i$$

Proof Starting with the LHS

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1, \ldots, x_n f_{x_1,\ldots,x_n} dx_1 \ldots dx_d = \prod_{i=1}^{d} \int_{-\infty}^{\infty} x_i f_{X_i}(x_i) dx_i = \prod_{i=1}^{d} \mathsf{E}X_i$$

□

Remark

If $X \perp\!\!\!\perp Y$ and $\mathsf{E}X^2 < \infty$ and $\mathsf{E}Y^2 < \infty$, then

$$\mathsf{Cov}(X, Y) = \mathsf{E}XY - \mathsf{E}X\mathsf{E}Y = \mathsf{E}X\mathsf{E}Y - \mathsf{E}X\mathsf{E}Y = 0$$

Suppose that (X, Y) is a random vector such that $\mathsf{Cov}\Big(g(X), f(Y)\Big) = 0$ for any f and g measurable such that $\mathsf{E}|g(X)| < \infty$ and $\mathsf{E}|f(Y)| < \infty$, then $X \perp\!\!\!\perp Y$. Simply take $g = 1_{B_1}$, $f = 1_{B_2}$ for $B_1, B_2 \in \mathbb{B}$. Then

$$\mathsf{Cov}\Big(g(X), f(Y)\Big) = \mathsf{P}(X \in B_1, Y \in B_2) - \mathsf{P}(X \in B_1)\mathsf{P}(Y \in B_2) = 0.$$

Remark

As a rule of thumb

- $\mathsf{E}(X_1, \ldots, X_n) = \sum_{i=1}^{d} \mathsf{E}X_i$ always hold if well-defined
- $\mathsf{E}(X_1, \ldots, X_n) = \prod_{i=1}^{d} \mathsf{E}X_i$ **only** if $X_1, \ldots, X_n$ independent is sufficient, but not necessary.

Definition 15.11 (Variance-covariance matrix)

Let $(X_1, \ldots, X_n)$ such that $\mathsf{E}X_i^2 < \infty$. Then, the variance-covariance matrix

Σ is a $d \times d$ matrix given by $\Sigma_{i,j} = \mathsf{Cov}(X_i, X_j)$ and

$$\Sigma = \begin{pmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_i, X_j) \\ \mathsf{Cov}(X_j, X_i) & \mathsf{Var}(X_j) \end{pmatrix}$$

Since $\mathsf{Cov}(X_i, X_j) = \mathsf{Cov}(X_j, X_i)$, the matrix is symmetric. An interesting feature of this is that the matrix Σ is positive semi-definite.

Theorem 15.12

Let $(X_1, \dots, X_n)$ be a random vector such that $\mathsf{E}X_i < \infty, i = 1, \dots, d$. Then, for any $\mathbf{a} = (a_1 \dots a_d)^{\perp}\perp \in \mathbb{R}^{d\times 1}$

$$\mathsf{Var}(a_1 X_1 + \cdots + a_d X_d) = \mathsf{Var}(\mathbf{a}^\top \boldsymbol{X}) = \mathbf{a}^\top \boldsymbol{X} \mathbf{a}$$

Remark

Σ is symmetric and positive semi-definite because

$$\mathbf{a}^\top \boldsymbol{X} \mathbf{a} = \mathsf{Var}(\mathbf{a}^\top \boldsymbol{X}) \geq 0.$$

Proof

$$\begin{aligned}
\mathsf{Var}(a_1 X_1 + \cdots + a_d X_d) &= \mathsf{E}((a_1 X_1 + \cdots + a_d X_d - a_1 \mathsf{E}X_1 - \cdots - a_d \mathsf{E}X_d)^2 \\
&= \mathsf{E}\left(\sum_{i=1}^{d} a_i \left(X_i - \mathsf{E}X_i\right)\right)^2 \\
&= \mathsf{E}\left(\sum_{j=1}^{d}\sum_{i=1}^{d} a_i a_j (X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j)\right) \\
&= \sum_{j=1}^{d}\sum_{i=1}^{d} a_i a_j \mathsf{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{d} a_i^2 \mathsf{Var} X_i + \sum_{i \neq j} a_i a_j \mathsf{Cov}(X_i, X_j)
\end{aligned}$$

In particular, if $X_i \perp\!\!\!\perp X_j$, the second sum vanish. □

Corollary 15.13 (Bienaymé identity)

Let $(X_1, \dots, X_n)$ be such that $\mathsf{E}X_i^2 < \infty, i = 1, \dots, d$ and $X_1, \dots, X_n$ are independent.

$$\mathsf{Var}(X_1, \dots, X_n) = \sum_{i=1}^{d} vaX_i = \mathrm{tr}\Sigma$$

Example 15.7

This is useful in random sample. Suppose we have a collection of $X_1, \dots, X_n$ random variables having Bernoulli distribution where $X_i \sim \mathcal{B}(p)$ How do we guess the p?

$$\hat{p_n} = \frac{\#\text{ heads}}{n} = \frac{1}{x}(X_1 + \cdots + X_n)$$

and for the expectation

$$\mathsf{E}\left(\hat{p_n}\right) = \mathsf{E}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}X_i = \frac{1}{n}\sum_{i=1}^{n} p = p$$

Then,

$$\begin{aligned}\mathsf{Var}(\hat{p_n}) &= \mathsf{Var}\left(\frac{1}{n}\left(X_1 + \cdots + X_n\right)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{Var}X_i \\ &= \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n} \xrightarrow{n\to\infty} 0\end{aligned}$$

Chapter 16
Conditional distribution

Correlation $\rho(X, Y)$ may be a good idea to capture the dependence between two random variables, but is is a bit simplistic. We could look at something else, called conditional distributions. Recall the conditional probability, $\mathsf{P}(A \cup B) = \mathsf{P}(A)\mathsf{P}(B)$ if $A \perp\!\!\!\perp B$ and if $\mathsf{P}(B) > 0$, then $\mathsf{P}(A|B) = \mathsf{P}A$. If we convey this idea to random variables X, Y with as our goal to define a conditional distribution of X given $Y = y$.
If (X, Y) is discrete, with probability mass function $f_{X,Y}$,

$$\mathsf{P}(X = x|Y = y) = \frac{\mathsf{P}(X = x, Y = y)}{\mathsf{P}(Y = y)}$$

In terms of PMF, it is just

$$\frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Observe that the above ratio is never zero $\forall\, x \in \mathbb{R}$ and we have

$$\sum_{x\in\mathbb{R}} \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{1}{f_Y(y)} \sum_x f_{X,Y}(x, y) = \frac{f_Y(y)}{f_Y(y)} = 1$$

is itself a probability mass function if viewed as a function of X. This deserves a proper definition.

Definition 16.1 (Conditional distribution)

Let (X, Y) be discrete random variables with PMF $f_{X,Y}$. Then, for any $y \in \mathbb{R}$ such that $\mathsf{P}(Y = y) = F_Y(y) > 0$, the conditional distribution of X given Y is discrete again with PMF given by

$$f_{X|Y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad x \in \mathbb{R}.$$

Similarly, one can define the conditional distribution of Y given $X = x$.

Remark

If X and Y are independent, then $\forall\, y \in \mathbb{R}$ such that $f_Y(y) > 0$

$$f_{X|Y}(x) = \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x), \quad x\ in \mathbb{R}$$

and also conversely.

Example 16.1

Suppose $(X, Y) \sim \mathcal{M}(n, p_1, p_2)$. We have seen that $Y \sim \mathcal{B}(n, p_2)$ and

$$F_{X|Y=y}(x) = \frac{\frac{n!}{x!(n-x-y)!y!}}{\frac{n!}{y!(n-y)!}} \frac{p_1^x p_2^y (1-p_1-p_2)^{n-x-y}}{p_2^y (1-p_2)^{n-y+x-x}}$$

for $y \in \{0, \dots, n\}$ and $x \in \{0, \dots, n-y\}$. If we cancel things, we get

$$\begin{aligned} &\frac{(n-y)!}{x!(n-y-x)!} \left(\frac{p_1}{1-p_2}\right)^x \left(\frac{1-p_1-p_2}{1-p_2}\right)^{n-y-x} \\ &= \frac{(n-y)!}{x!(n-y-x)!} \left(\frac{p_1}{1-p_2}\right)^x \left(1 - \frac{p_1}{1-p_2}\right)^{n-y-x} \end{aligned}$$

and we conclude that given $Y = y, X \sim \mathcal{B}\left(n-y, \frac{p_1}{1-p_2}\right)$. Here is a "dirty calculation". If we are interested in the following $\mathsf{P}(X = x | Y = y)$ supposing (X, Y) has a density $f_{X,Y}$. We can look at

$$\begin{aligned} \mathsf{P}(X = x | Y \in (y-\varepsilon, y+\varepsilon) \xrightarrow{\varepsilon \to 0} &= \frac{\mathsf{P}(X < x, Y \in (y-\varepsilon, y+\varepsilon])}{\mathsf{P}(Y \in y-\varepsilon, y+\varepsilon])} \\ &= \frac{\int_{-\infty}^{x} \int_{y-\varepsilon}^{y+\varepsilon} f_{X,Y}(s,t) dt\, ds / 2\varepsilon}{\int_{y-\varepsilon}^{y+\varepsilon} f_Y(t) dt / 2\varepsilon} \\ &\approx \frac{\int_{-\infty}^{x} f_{X,Y}(s,y) ds}{f_Y(y)} \end{aligned}$$

if all necessary conditions are fulfilled, then by the Mean Value Theorem, the denominator is "precisely" 1. Let now $\varepsilon \to 0$. Then, $\mathsf{P}(X \leq x | Y = y)$ kind of has like a density

$$\frac{\int_{-\infty}^{x} f_{X,Y}(s,y) ds}{f_Y(y)}$$

Definition 16.2 (Conditional distribution)

If (X, Y) has density $f_{X,Y}$, then for any $y \in \mathbb{R}$ such that $f_Y(y) > 0$, the conditional distribution of X given $Y = y$ is defined by its density

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad x \in \mathbb{R}$$

Example 16.2

Building on our previous example, suppose that we have that the random vector $(X, Y) \sim \mathcal{M}(n, p_1, p_2)$, then

- $X|Y = y \sim \mathcal{B}\left(n - y, \frac{p_1}{1-p_2}\right), y \in \{0, \dots, n\}$
- $Y|X = x \sim \mathcal{B}\left(n - x, \frac{p_2}{1-p_1}\right), x \in \{0, \dots, n\}$
- $\mathsf{E}(X|Y = y) = (n - y)\frac{p_1}{1-p_2}$ is a function of y
- $\mathsf{E}(Y|X = x) = (n - x)\frac{p_2}{1-p_1}$ is a function of x

Example 16.3

We have $(X, Y) \sim \mathcal{N}\left(\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\right)$ and

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

and $X|Y = y \sim \mathcal{N}(\rho y, 1 - \rho^2), \mathsf{E}(X|Y = y) = \rho y$ (function of y) and $\mathsf{E}(Y|X = x) = \rho x$ (function of x).

Definition 16.3 (Conditional expectation)

Let X, Y be some random variables (not necessarily independent). The conditional expectation $\mathsf{E}(X|Y)$ is defined as $\Psi(Y), \Psi$ measurable and

$$\Psi(y) = \mathsf{E}(X|Y = y) = \begin{cases} \displaystyle\sum_x x\mathsf{P}(X = x|Y = y) & \text{in the discrete case} \\ \displaystyle\int x f_{X|Y=y}(x)dx & \text{in the continuous case} \end{cases}$$

We can think of this as a projection on the space $\mathcal{S}_Y$. In Advanced probability, you will see that one can make Y into a $\sigma-$field. Examples of applications and of use include time series; if we want to predict tomorrow stock market, they are likely to depend on today. For $\mathsf{E}(g(X)|Y)) = \Psi(Y)$ where

$$\Psi(Y) = \mathsf{E}(g(X)|Y=y) = \begin{cases} \displaystyle\sum_x g(x)\mathsf{P}(X=x|Y=y) \\ \displaystyle\int g(x) f_{X|Y=y}(x)dx \end{cases}$$

Example 16.4 (Hierarchical model)

Recall that if $Y \sim \mathcal{P}(\lambda)$, then $\mathsf{E}Y = \mathsf{Var}\,Y = \lambda$. Consider an example involving car accidents. Not all have the same probability: some are bad drivers, some have summer tires (on ice), etc. Consider $\Lambda \sim \Gamma\left(r, \frac{1-p}{p}\right)$, $r \in \mathbb{N}, p \in (0,1)$ where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. We are given

$$f_\Lambda(\lambda) = \frac{1}{\Gamma(r)}\lambda^{r-i}\exp\left(-\frac{\lambda}{\frac{1-p}{p}}\left(\frac{1-p}{p}\right)^r\right)$$

and $X|\Lambda = \lambda \sim \mathcal{P}(\lambda)$. X will be discrete again $\mathsf{P}(X=x)$, for $x \in \{0,1,\ldots\}$ and $\mathsf{P}(X=x|\Lambda=\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$. Before computing this, we need a little more about conditional expectation.

Lemma 16.4 (Properties of conditional expectations)

The following properties hold (provided the expectations exist).

1. $\mathsf{E}(c|Y) = c$;
2. $\mathsf{E}(X|Y) = \mathsf{E}X$ when $X \perp\!\!\!\perp Y$;
3. $\mathsf{E}(a_1 \cdot g_1(X) + a_2 \cdot g_2(X)|Y) = a_1\mathsf{E}(g_1(X)|Y) + a_2\mathsf{E}(g_2(X)|Y)$;
4. If $X \geq 0$ almost surely, then $\mathsf{E}(X|Y) \geq 0$ almost surely;
5. $\mathsf{E}(g(Y)\Psi(X,Y)|Y) = g(Y)\mathsf{E}(\Psi(X,Y)|Y)$;
6. $\mathsf{E}(\Psi(X,Y)|Y=y) = \mathsf{E}(\Psi(X,y)|Y=y)$;
7. $\mathsf{E}(X) = \mathsf{E}(\mathsf{E}(X|Y))$; the iterative expectation formula.

Proof We prove the Law of Iterated expectation, starting from the right hand side.

$$\begin{aligned}
\Psi(y) = \mathsf{E}(X|Y=y) = \int x f_{X|Y=y}(x) \\
&= \int x \frac{f_{X,Y}(x,y)}{f_Y(y)} dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \frac{f_{X,Y}(x,y)dx}{f_Y(y)} f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dx dy \\
&= \int_{-\infty}^{\infty} x \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx}_{f_X(x)} = \int_{-\infty}^{\infty} x f_X(x) dx = \mathsf{E}X
\end{aligned}$$

and in particular, $\mathsf{E}g(X) = \mathsf{E}(\mathsf{E}(g(X)|Y))$ □

Example 16.5 (Hierarchical model (part 2))

Back to our example.

$$\mathsf{P}(X=x) = \mathsf{E}(1_{(X=x)}) = \mathsf{E}(\mathsf{E}(1_{(X=x)}|\Lambda))$$

if we look at the inside, this is

$$\mathsf{E}(1_{(X=x)}|\Lambda=\lambda) = \mathsf{P}(X=x|\Lambda=\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

and it is even continuous; we can replace this inside with Λ random. Therefore

$$\begin{aligned}
\mathsf{E}\left(\frac{e^{-\Lambda}\Lambda^x}{x!}\right) &= \int_0^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} \frac{1}{\Gamma(r)} \lambda^{r-1} e^{-\frac{\lambda}{\frac{1-p}{p}}} \frac{1}{\left(\frac{1-p}{p}\right)^r} d\lambda \\
&= \left(\frac{p}{1-p}\right)^{-r} \frac{1}{x!} \frac{1}{\Gamma(r)} \int_0^{\infty} \lambda^{x+r-1} e^{-\lambda\left(1+\frac{p}{1-p}\right)} d\lambda \\
&= \left(\frac{p}{1-p}\right)^{-r} \frac{1}{x!} \frac{1}{\Gamma(r)} \int_0^{\infty} t^{x+r-1} e^{-t} (1-p)^{x+r} dt
\end{aligned}$$

where the last step follow from the transformation $\frac{\lambda}{1-p} = t$ and therefore, we obtain from the hierarchical model

$$= p^r(1-p)^x \frac{\Gamma(x+r)}{x!\Gamma(r)} = \frac{(x+r-1)!}{x!(r-1)!} p^r(1-p)^x, \quad x \ \in \{0, 1, \dots\}$$

is a Negative Binomial and therefore $\mathsf{E}X = \mathsf{E}(\mathsf{E}(X|\Gamma)) = \mathsf{E}\lambda = r\frac{(1-p)}{p}$.

Example 16.6

For $X \perp\!\!\!\perp Y$, $\mathsf{P}(X \leq Y) = \mathsf{P}(X - Y \leq 0)$. If we adapt (7), $\mathsf{E}(g(X,Y)) = \mathsf{E}(\mathsf{E}(g(X,Y)|Y))$, we can condition and get

$$\begin{aligned} \mathsf{E}(1(X \leq Y)) &= \mathsf{E}(\mathsf{E}(1(X \leq Y)|Y)) \\ \mathsf{E}(\mathsf{E}(1(X \leq Y)|Y = y)) &= \mathsf{E}(1(X \leq y)|Y = y) \\ &= \mathsf{P}(X \leq y|Y = y) = P(X \leq y) = F_X(y) \end{aligned}$$

so $\mathsf{P}(X \leq y) = \mathsf{E}(F_x(Y))$. In particular, if X, Y have densities, we have $\int_{-\infty}^{\infty} F_X(y) f_Y(y) dy$.

Example 16.7

Again, suppose $X \perp\!\!\!\perp Y$. We are interested in
$\mathsf{P}(X + Y \leq z) = \mathsf{E}(\mathsf{E}(1(X + Y \leq z|Y))$ where

$$\begin{aligned} \mathsf{E}(1(X + Y \leq z|Y = y) &= \mathsf{P}(X + Y \leq z|Y = y) \\ &= \mathsf{P}(Z \leq z - y|Y = y) \\ &= \mathsf{P}(X \leq z - y) = F_X(z - y) \end{aligned}$$

by independence so $\mathsf{P}(X + Y \leq Z) = \int_{-\infty}^{\infty} F_X(z - y) f_Y(y) dy$ if Y has a density. If $X \perp\!\!\!\perp Y$, but this time (X, Y) is discrete, if we look at

$$\mathsf{P}(X + Y \leq z) = \mathsf{E}(\underbrace{\mathsf{E}(1(X + Y = z|Y = y)}_{\mathsf{P}(X=z-y)}) = \sum_y \mathsf{P}(X = z - y)\mathsf{P}(Y = y)$$

by making Y a random variable and conditioning. The result is the same we had derived with the Law of total probability.

Theorem 16.5

Let X, Y be random variables such that $\mathsf{E}X^2 < \infty$. Then

$$\mathsf{Var}(X) = \mathsf{E}(\mathsf{Var}(X|Y)) + \mathsf{Var}(\mathsf{E}(X|Y)) \tag{16.3}$$

Example 16.8 (Hierarchical model (part 3))

We return to our example with the Negative binomial. $X|\Lambda = \lambda \sim \mathcal{P}(\lambda)$, $\Lambda \sim \Gamma\left(r\frac{1-p}{p}\right)$. and

$$\begin{aligned}\mathsf{Var}X &= \mathsf{E}(\mathsf{Var}(X|\Lambda)) + \mathsf{Var}(\mathsf{E}(X|\Lambda))\\ &= \mathsf{E}(\lambda + \mathsf{Var}(\lambda)\\ &= r\left(\frac{1-p}{p}\right) + r\left(\frac{(1-p)^2}{p^2}\right)\\ &= r\frac{1-p}{p^2}\end{aligned}$$

Proof We now go to the proof of 16.5. Starting with the right hand side:

$$\begin{aligned}\mathsf{E}(\mathsf{Var}(X|Y)) &= \mathsf{E}\left(\mathsf{E}(X^2|Y) - \{\mathsf{E}(X|Y)\}^2\right)\\ &= \int x^2 f_{X|Y=y}(x)dx - \left(\int x f_{X|Y=y} x dx\right)^2\end{aligned}$$

and from this, we decompose the terms from 16.5

$$\mathsf{Var}\left(\mathsf{E}(X|Y)\right) = \mathsf{E}\left(\{\mathsf{E}(X|Y)\}^2 - \{\mathsf{E}(\mathsf{E}(X|Y))\}^2\right)$$

and

$$\begin{aligned}\mathsf{E}(\mathsf{Var}(X|Y)) + \mathsf{Var}(\mathsf{E}(X|Y)) &= \mathsf{E}(\mathsf{E}(X^2|Y)) - (\mathsf{E}X)^2\\ &= \mathsf{E}\left(\mathsf{E}(X^2|Y)\right) - \mathsf{E}X^2 = \mathsf{E}X^2 - (\mathsf{E}X)^2\end{aligned}$$

□

Suppose $r = \frac{1-p}{p} = \lambda^*$ such that $\mathsf{E}X = \lambda^*$, $Y \sim \mathcal{P}(\lambda^*)$. This is easily seen by plotting the distribution on the computer (the negative binomial is more spread). By conditioning, we increase the variability. These kind of examples arise in Bayesian statistic, a field that takes parameters as random variables.

Example 16.9 (Compound sum)

This is used in actuarial science. Suppose that we have claims for insurance. We have some degree of knowledge of the magnitude of the claims, yet we don't know the number of claims per year. We invent a sequence of i.i.d. random variables $X_1, X_2, X_3, \dots$ (claim size), where $X_i \sim \Gamma$ or Lognormal or Pareto (if claims are high). We are interested in the total amount required to pay at the end of the year, namely $S_N = \sum_{i=1}^{N} X_i$ and N is an integer random variables independent of $X_i\ \forall\ i$. This is a reasonable assumption if we have no catastrophe. Thus, if we want to learn about $\mathsf{P}(S_N \leq z)$, we can look at

$$\mathsf{E}(S_N) = \mathsf{E}(\mathsf{E}(S_N|N))$$

and the inside term is

$$\begin{aligned}\Rightarrow \mathsf{E}(S_N|N=n) &= \mathsf{E}\left(\sum_{i=1}^{n} X_i|N=n\right)\\ &= \mathsf{E}\left(\sum_{i=1}^{n} X_i\right) = n\mathsf{E}X_1\end{aligned}$$

so

$$\mathsf{E}(S_N) = \mathsf{E}(N\mathsf{E}X_1) = \mathsf{E}N\mathsf{E}X_1$$

We can also look at the variance using 16.5

$$\mathsf{Var}(S_N) = \mathsf{E}(\mathsf{Var}(S_N|N)) + \mathsf{Var}(\mathsf{E}(S_N|N))$$

and looking at each term individually, we get

$$\mathsf{Var}(S_N|N=n) = \mathsf{Var}\left(\sum_{i=1}^{n} X_i|N=n\right) = n\mathsf{Var}(X_1)$$

by plugging back,

$$\mathsf{Var}(S_N) = \ (N\mathsf{Var}X_1) + \mathsf{Var}(N\mathsf{E}X_1) = \mathsf{Var}X\mathsf{E}N + (\mathsf{E}X_1)^2\,\mathsf{Var}N.$$

In particular, if $N \sim \mathcal{P}(\lambda)$, we obtain

$$\lambda\left(\mathsf{E}X_i^2 - (\mathsf{E}X_i)^2 + (EX_i)^2\right) = \lambda\mathsf{E}X_i^2$$

The moment generating function is an expectation

$$\mathsf{E}e^{tS_N} = \mathsf{E}\left(\mathsf{E}\left(e^{tS_N}|N\right)\right) = \prod_{i=1}^{n}\mathsf{E}\left(e^{tX_i}\right) = \mathsf{E}\left(\{M_X(t)\}^N\right)$$

Chapter 17

Stochastic convergence and Laws of large numbers

Let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables distributed $\mathcal{N}(\mu, \sigma^2)$. We want to find a mean. One possibility is to take the arithmetic mean

$$\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n)$$

in order to get

$$\mathsf{E}\left(\overline{X_n}\right) = \mu \text{ and } \mathsf{Var}\left(\overline{X_n}\right) = \frac{1}{n^2} \cdot n\mathsf{Var}(X_1) = \frac{\sigma^2}{n}$$

such that $\overline{X_n} \xrightarrow{n\to\infty} \mu$. Recall that X_n is a random variable, so the above is a mapping. Say $Y_1, Y_2, Y_3, \dots$ is another sequence of random variables and we have another rv Y such that $Y_n \to Y$, then $Y_n : \Omega \to \mathbb{R} \; \forall \, n$ and $Y : \Omega \to \mathbb{R}$. Recall that we have *pointwise convergence* if $\forall \, \omega \in \Omega, Y_n(\omega) \to Y(\omega)$. Since we know little about ω, this type of convergence is not really realistic. But we know the distribution and instead, we consider these different type of convergence as $n \to \infty$, we have

1. $Y_n \xrightarrow{\text{a.s.}} Y$: convergence almost surely;
2. $Y_n \xrightarrow{\mathsf{P}} Y$: convergence in probability;
3. $Y_n \rightsquigarrow Y$ or $Y_n \xrightarrow{\mathcal{D}} Y$; convergence in distribution

Definition 17.1 (Convergence in probability and convergence almost surely)

Let $Y_1, Y_2, \dots$ be an arbitrary sequence of random variables. Let Y be an arbitrary random variable. Then

1. $Y_n \to Y$ almost surely $\Leftrightarrow \mathsf{P}\left(\omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\right) = 1$.
2. $Y_n \xrightarrow{\mathsf{P}} Y \Leftrightarrow \; \forall \, \varepsilon > 0, \lim_{n\to\infty} \mathsf{P}(|Y_n - Y| < \varepsilon) = 1$.

Example 17.1 (Marching mountains)

Let $\Omega = [0,1], \mathbb{B} \cup [0,1], \mathsf{P}$ the uniform distribution on $[0,1]$.
Define $X(\omega) = \omega \; \forall \, \omega \in [0,1]$ and $X_n(\omega) = \omega + \omega^n, \omega \in [0,1], n \geq 1$. We

have

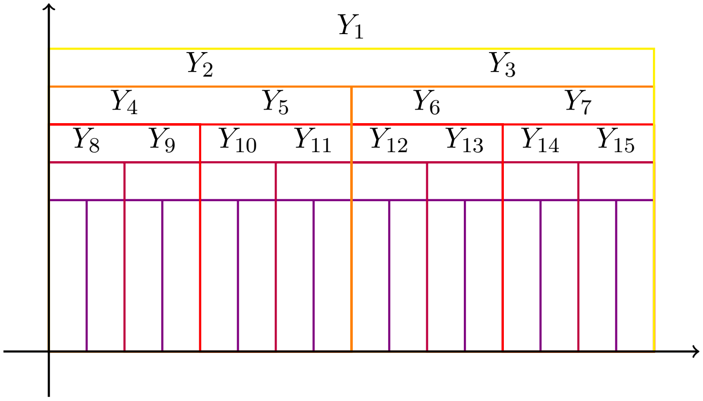
$$\lim_{n\to\infty} X^n(\omega) = \begin{cases} \omega & \text{if } \omega \in [0,1) \\ 2 & \text{if } \omega = 1 \end{cases}$$

and notice $2 \neq \omega$ for $\omega = 1$. We have that $X_n \to X$ almost surely because $\mathsf{P}(\omega : X_n(\omega) \to X(\omega)) = \mathsf{P}([0,1)) = 1$. If we look at the sequence, we have

$$\begin{aligned} Y_1(\omega) &= \omega + 1_{(\omega\in[0,1))} \\ Y_2(\omega) &= \omega + 1_{\left(\omega\in\left[0,\frac{1}{2}\right)\right)} \\ Y_3(\omega) &= \omega + 1_{\left(\omega\in\left[\frac{1}{2},1\right)\right)} \\ Y_4(\omega) &= \omega + 1_{\left(\omega\in\left[0,\frac{1}{4}\right)\right)} \\ &\dots \\ Y_{2^n+k}(\omega) &= \omega + 1_{\left(\omega\in\left[\frac{k}{2^n},\frac{k+1}{2^n}\right)\right)} \end{aligned}$$

If you want to show it converge almost surely, you will find (almost surely)

Figure 21: Marching mountains



that it doesn't hold since

$$Y_n(\omega) = \begin{cases} \omega & \infty \text{ often} \\ \omega + 1 & \infty \text{ often} \end{cases}$$

for $\omega \in [0,1)$ so it doesn't converge. $Y_n \nrightarrow X$ almost surely, but for $\varepsilon < 1$,

$$\lim_{n\to\infty} \mathsf{P}(|Y_{2^n+k} - X| \geq \varepsilon) = \lim_{n\to\infty} \mathsf{P}\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right) = \lim_{n\to\infty} \frac{1}{2^n} = 0$$

and $Y_n \xrightarrow{\mathsf{p}} X$.

Observation

Convergence in probability does not imply convergence almost surely? What about the converse? It indeed is true, but to show it we must investigate further. Consider $X_1, X_2, \ldots$ and X random variables. For $\varepsilon > 0$, define $A_n(\varepsilon) = \{|X_n(\omega) - X(\omega)| > \varepsilon\}$. To have $X_n \xrightarrow{\mathsf{p}} X$ means that

$$\mathsf{P}(A_n(\varepsilon)) \xrightarrow{n\to\infty} 0 \; \forall \; \varepsilon > 0.$$

Convergence almost surely means

$$\mathsf{P}\Big(\limsup_{n\to\infty} A_n(\varepsilon)\Big) = 0 \; \forall \; \varepsilon > 0 \tag{17.4}$$

Set $A = \{\omega : X_n(\omega) \nrightarrow X(\omega)\}$ (for $X_n \xrightarrow{\text{a.s.}} X$ means $\mathsf{P}(A) = 0$. Hence, $X_n(\omega) \nrightarrow X(\omega) \Rightarrow \exists \varepsilon > 0$ such that $|X_n(\omega) - X(\omega)| > \varepsilon$ infinitely often, i.e.

$$\begin{aligned} A &= \bigcup_{\varepsilon>0} \limsup_{n\to\infty} A_n(\varepsilon) = \bigcup_{\varepsilon>0} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n(\varepsilon) \\ &= \{\omega : \omega \in A_n(\varepsilon) \text{ for infinitely many } n\text{'s} \\ &\Rightarrow \bigcup_{k=1}^{\infty} \limsup_{n\to\infty} A_n\left(\frac{1}{k}\right) \end{aligned}$$

and

$$\mathsf{P}(A) = \mathsf{P}\left(\bigcup_{k=1}^{\infty} \limsup_{n\to\infty} A_n\left(\frac{1}{k}\right)\right)$$

First, $\mathsf{P}(\limsup_{n\to\infty} A_n(\varepsilon)) \leq P(A) = 0$. Conversely, if (17.4) holds, then

$$P(A) = \mathsf{P}\left(\bigcup_{k=1}^{\infty} \limsup_{n\to\infty} A_n\left(\frac{1}{k}\right)\right) \leq \sum_{k=1}^{\infty} \mathsf{P}\left(\limsup_{n\to\infty} A_n\left(\frac{1}{k}\right)\right) = 0$$

Observation

By definition, $\limsup_{n\to\infty} A_n(\varepsilon) = \bigcap n = 1^\infty \bigcup_{m=n}^\infty A_m(\varepsilon) \equiv \bigcap_{n=1}^\infty B_n(\varepsilon)$. If I have 0, these $B_n(\varepsilon)$ build a decreasing sequence since we are building a union of less and less sets and

$$\mathsf{P}\left(\limsup_{n\to\infty} A_n(\varepsilon)\right) = \lim_{n\to\infty} \mathsf{P}(B_n(\varepsilon)) \geq \lim_{n\to\infty} P(A_n(\varepsilon))$$

Therefore, if $X_n \to X$ almost surely, $\forall\, \varepsilon > 0$ $\mathsf{P}(\limsup_{n\to\infty} A_n(\varepsilon)) = 0 \Rightarrow \lim_{n\to\infty} \mathsf{P}(A_n(\varepsilon)) = 0 \Rightarrow X_n \xrightarrow{\mathsf{P}} X$.

Theorem 17.2

Let $X_1, X_2, \ldots$ be a sequence of random variables and X another random variable. Then, if $X_n \to X$ almost surely, then $X_n \xrightarrow{\mathsf{P}} X$ as $n \to \infty$, but **not conversely**. Note it is often simpler and nicer to show convergence in probability.

Theorem 17.3

Let $X_1, X_2, \ldots$ be a sequence of random variables and X another random variable. Then,

1. $X_n \to X$ a.s. and $X_n \to Y$ a.s, then $\mathsf{P}(X = Y) = 1$.
2. $X_n \xrightarrow{\mathsf{P}} X$ and $X_n \xrightarrow{\mathsf{P}} Y$, then $\mathsf{P}(X = Y) = 1$.

The limit is almost surely unique.

Proof We prove the second statement. We will show

$$\begin{aligned}\mathsf{P}(X \neq Y) = 0 = \mathsf{P}&\left(\left\{\omega : \exists k > 0, |X(\omega) - Y(\omega)| > \frac{1}{k}\right\}\right)\\ &= \mathsf{P}\left(\bigcup_{k=1}^{\infty}\left\{\omega : |X(\omega) - Y(\omega)| > \frac{1}{k}\right\}\right)\\ &= \lim_{n\to\infty} \mathsf{P}\left(|X - Y| > \frac{1}{k}\right)\end{aligned}$$

since these sets here build a monotone sequence and therefore, we want to show that all probabilities are equal to zero. Fix some $k = k_0$. By the triangle inequality,

$$\begin{aligned}\mathsf{P}\left(|X-Y|>\frac{1}{k_0}\right)&=\mathsf{P}\left(|X-X_n+X_n-Y|>\frac{1}{k_0}\right)\\&=\mathsf{P}\left(|X_n-X|+|X_n-Y|>\frac{1}{k_0}\right)\\&\le\mathsf{P}\left(\left\{|X_n-X|>\frac{1}{2k_0}\right\}\cup\left\{|X_n-Y|+>\frac{1}{2k_0}\right\}\right)\\&\le\mathsf{P}\left(|X_n-X|>\frac{1}{2k_0}\right)+\mathsf{P}\left(|X_n-Y|+>\frac{1}{2k_0}\right)\end{aligned}$$

for every n. Let $n\to\infty$ and hence $\mathsf{P}(|X-Y|>1/k_0)\le 0$. Consequently, $\mathsf{P}(X\neq Y)=\lim_{n\to\infty}0=0$ hence $\lim_{n\to\infty}\mathsf{P}(|X-Y|>1/k)=0$. □

Theorem 17.4 (Weak Law of Large Numbers)

Let $X_1,X_2,\dots$ be a sequence of i.i.d. random variables such that $\mathsf{E}X_1=\mu<\infty$ and $\mathsf{Var}(X_1=\sigma^2<\infty$. Then

$$\overline{X_n}=\frac{1}{n}(X_1+\cdots+X_n)\xrightarrow[n\to\infty]{p}\mu$$

Proof By Chebychev inequality; fix $\varepsilon>0$,

$$\mathsf{P}\left(|\overline{X_n}-\mu|>\varepsilon\right)\le\frac{\mathsf{Var}\left(\overline{X_n}\right)}{\varepsilon^2}=\frac{\sigma^2}{n\cdot\varepsilon^2}\xrightarrow{n\to\infty}0$$

□

Example 17.2

Suppose that we have $X_1,\dots,X_n\sim\mathcal{N}(\mu,\sigma^2)$, then $\overline{X_n}\xrightarrow{\mathsf{p}}\mu$. However, this trick does not always hold; here is a counterexample. If $X_1,\dots,X_n\sim$ Cauchy(0,1), then $\overline{X_n}\sim$ Cauchy(0,1).

Remark

The Weak Law of Large Numbers (WLLN) can be weakened

1. WLLN holds if $\mathsf{E}(X_1)<\infty$. It is however hard to prove (follows from Markov inequality).

2. IF $X_1, X_2, \ldots$ is a sequence of integrable random variables, then $\overline{X_n} \to \mathsf{E}X_1$ almost surely. This is call the Strong Law of Large Numbers and can be proved with a bit of struggle.

3. WLLN holds if $\mathsf{Cov}(X_i, X_j = 0 \ \forall \ i \neq j$ (and $\mathsf{E}X_1 < \infty, \mathsf{Var}X_1 < \infty, X_i \overset{d}{=} X_j$).

Example 17.3 (Monte-Carlo integration)

Here is a crucial example used in numerical analysis. Suppose $g : [0,1] \to \mathbb{R}$, g is continuous and $u \sim \mathcal{U}(0,1)$. We are interested by the integral $\int_0^1 g(x)dx$. We introduce a new variable $X = g(U)$.

$$\mathsf{E}g(Y) = \int_0^1 g(u)du < \infty \leq \|g\|$$

since it is continuous on a closed and bounded interval, but also

$$\mathsf{Var}(g(u)) = \int_0^1 g^2(u)du - \left(\int_0^1 g(u)du\right)^2 \leq 2\|g\|^2$$

Then, we can simply have a sequence to approximate our integral. We have for n large $U_1, \ldots, U_n$ a sequence of independent Uniform(0,1) and compute the approximation $\int_0^1 g(u)du$

$$\frac{1}{n}\sum_{i=1}^{n} g(U_i) \xrightarrow{\mathsf{p}} \int_0^1 g(u)du$$

by Chebychev, we could get an idea of distance and the level of precision.

Theorem 17.5 (Continuous mapping theorem)

Let g be a continuous function and $\{X_n\}, X$ be random variables. Then

1. If $X_n \to X$ almost surely, then $g(X_n \to g(X)$ almost surely;

2. If $X_n \xrightarrow{\mathsf{p}} X$. then $g(X_n) \xrightarrow{\mathsf{p}} g(X)$.

Proof

1. Given that $\mathsf{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$, once you fix ω, you get a sequence of numbers and continuity of the mapping $\mathsf{P}(\{\omega : g(X_n(\omega)) \to g(X(\omega))\}) = 1$ if continuous on $\mathbb{R}$.

2. $\forall\, k > 0, g$ is uniformly continuous on $[-k, k]$ by definition if

   $$(\forall\, \varepsilon > 0)(\exists\, \delta_{(\varepsilon,k)} > 0)(\forall\, x \in [-k, k]) \\ (|x - y| < \delta_{(\varepsilon,k)} \Rightarrow |g(x) - g(y)| < \varepsilon)$$

   equivalent to saying that $|g(x) - g(y)| \geq \varepsilon \Rightarrow |x - y| \geq \delta_{(\varepsilon,k)}$. For all ϵ positive, we have by the Law of total probability

   $$\begin{aligned}&\mathsf{P}(|g(X_n) - g(X)| \geq \varepsilon) \\ &= \mathsf{P}(|g(X_n) - g(X)| \geq \varepsilon, |X| \leq k) + \mathsf{P}(|g(X_n) - g(X)| \geq \varepsilon, |X| > k) \\ &\leq \mathsf{P}(|X_n - X| \geq \delta_{(\varepsilon,k)}, |X| \leq k) + \mathsf{P}(|X| > k) \\ &\leq \mathsf{P}(|X_n - X| \geq \delta_{(\varepsilon,k)}) + \mathsf{P}(|X| > k)\end{aligned}$$

   by first enlarging the probability and secondly using the fact that $\mathsf{P}(A \cap B) \leq \mathsf{P}(B)$. Since both probabilities are non-negative, we obtain

   $$0 \leq \lim_{n\to\infty} \mathsf{P}(|g(X_n) - g(X)| \geq \varepsilon \leq 0 + P(|X| > k) \xrightarrow{k\to\infty} 0$$

   as $n \to \infty$ so the limit must indeed be zero.

□

A simple application of the continuous mapping theorem is the following

Lemma 17.6

If $\{X_n\}, \{Y_n\}, X, Y$ are random variables. Then

1. $X_n \xrightarrow{\mathsf{p}} X$ and $Y_n \xrightarrow{\mathsf{p}} Y$, then $X_n + Y_n \xrightarrow{\mathsf{p}} X + Y$;

2. $X_n \xrightarrow{\mathsf{p}} X$ and $Y_n \xrightarrow{\mathsf{p}} Y$, then $X_n \cdot Y_n \xrightarrow{\mathsf{p}} X \cdot Y$.

Proof

1. $\forall\, \varepsilon > 0$, $\mathsf{P}\left(|X_n + Y_n - X - Y| > \varepsilon\right) \to 0$ as $n \to \infty$. By the triangle inequality, this is less than $\mathsf{P}\left(|X_n - X| + |Y_n - Y| \geq \varepsilon\right)$. If we look at the event

$$\begin{aligned}&\mathsf{P}\left(|X_n - X| \geq \frac{\varepsilon}{2} \text{ or } |Y_n - Y| \geq \frac{\varepsilon}{2}\right)\\ &\quad\leq \mathsf{P}\left(|X_n - X| \geq \frac{\varepsilon}{2}\right) + \mathsf{P}\left(|Y_n - Y| \geq \frac{\varepsilon}{2}\right) \to 0\end{aligned}$$

since $\mathsf{P}(A \cup B) \leq \mathsf{P}A + \mathsf{P}B$.

2. Left as an exercise. Hint: use an $\varepsilon/3$ argument using that

$$|X_nY_n - XY| \leq |X_n - X||Y_n - Y| + |X||Y_n - Y| + |Y||X_n - X|$$

and the case distinction with $|X| \leq k$ and $|X| > k$.

□

Example 17.4

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Recall from your basic intro to statistics that just as we found μ using the arithmetic mean, the sample variance ${S_n}^2$ was found as

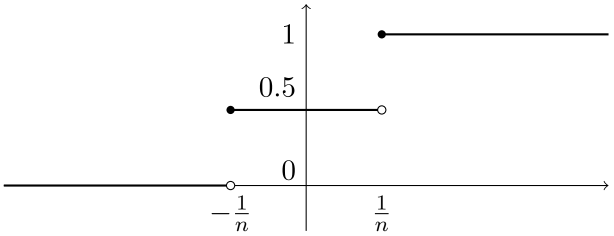
$$ {S_n}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X_n}\right)^2 \tag{17.5}$$

and $\mathsf{E}{S_n}^2 = \sigma^2$ which entails that we can rewrite the above (17.5)

$$\frac{n}{n-1} \frac{1}{n} \sum_{i=1}^{n} {X_i}^2 - \frac{n}{n-1} \left(\overline{X_n}\right)^2$$

and $\frac{n}{n-1}$ converge in probability to 1, hence

$$\frac{1}{n} \sum_{i=1}^{n} {X_i}^2 \xrightarrow[\text{WLLN}]{\mathsf{p}} \mathsf{E}{X_i}^2 \qquad \overline{X_n} \xrightarrow[\text{WLLN}]{\mathsf{p}} \mathsf{E}X_i$$



and by the continuous mapping theorem,

$$\left(\overline{X_n}\right)^2 \xrightarrow{\mathsf{p}} (\mathsf{E}X_i)^2$$

hence ${S_n}^2 \xrightarrow{\mathsf{p}} \mathsf{E}X_i^2 - (\mathsf{E}X_i)^2 = \mathsf{Var}X_i = \sigma^2$ and $\sqrt{{S_n}^2} \xrightarrow{\mathsf{p}} \sigma$ (even if this is a little messier). We say the estimator is weakly consistent.

Definition 17.7 (Convergence in distribution)

Let $\{X_n\}$ be a collection of random variables and X be a r.v. Then $X_n \rightsquigarrow X (X_n$ converge in law or in distribution to X as $n \to \infty)$ if for any *continuity point* x of F_x,

$$F_{X_n}(x) \xrightarrow{n\to\infty} F_X(x).$$

In other words, we have pointwise convergence of the distribution function.

Example 17.5

Suppose that we have $\mathsf{P}(X_n = \frac{1}{n}) = \mathsf{P}(X_n = -\frac{1}{n}) = \frac{1}{2}$. If we draw the distribution function of X_n, recalling that $F_{X_n}(x) = \mathsf{P}(X_n \leq x)$. If we draw the distribution function of X_n, we have

$$F_{X_n}(x) = \begin{cases} 0 & x < 0 \\ \dfrac{1}{2} & x = 0 \\ 1 & x > 0 \end{cases}$$

and this cannot be a distribution function. But $\mathsf{P}(X = 0) = 1$, then $F_{X_n}(x) \to F_X(x)\ \forall\ x \in \mathbb{R}$ if $x \neq 0$ hence by definition $X_n \rightsquigarrow X$. This type of distribution is actually the weakest of all three.

Theorem 17.8

Let $\{X_n\}$ be a collection of random variables and X be a r.v. Then $X_n \xrightarrow{\mathsf{p}} X$ imply that $X_n \rightsquigarrow X$, but not conversely.

Proof ($\Rightarrow$) Let x be a continuity point of F_X. If $\varepsilon > 0$ is sufficiently small. Then

$$\begin{aligned}\mathsf{P}(X_n \leq x) &= \mathsf{P}(X_n \leq x, X \leq x+\varepsilon) + \mathsf{P}(X_n \leq x, X > x+\varepsilon)\\ &\leq \mathsf{P}(X \leq x+\varepsilon) + \mathsf{P}(|X_n - x| > \varepsilon)\end{aligned}$$

Let now $n \to \infty$ to get $\lim_{n\to\infty} F_{X_n}(x) \leq F_X(x+\varepsilon) + 0$.
For the lower bound,

$$\begin{aligned}F_X(x-\varepsilon) &= \mathsf{P}(X \leq x+\varepsilon)\\ &= \mathsf{P}(X \leq x-\varepsilon, X_n > x) + \mathsf{P}(X \leq x-\varepsilon, X_n \leq x)\\ &\leq \mathsf{P}(|X_n - X| > \varepsilon) + F_{X_n}(x)\\ &\leq \lim_{n\to\infty} F_{X_n}(x) \leq \mathsf{P}(X \leq x+\varepsilon) = F_X(x+\varepsilon)\end{aligned}$$

Let $\varepsilon \to 0$ and by continuity

$$F_X(x) \leq \lim_{n\to\infty} F_{X_n}(x) \leq F_X(x).$$

($\Leftarrow$) We go with a counterexample. Suppose we have a random variable $X \sim \mathcal{B}(1, 1/2)$ and X_n is defined as

$$X_n = \begin{cases} X & \text{if } n \text{ is odd;}\\ 1 - X & \text{if } n \text{ is even.}\end{cases}$$

We have for all $n \geq 1$ and $X_n \sim \mathrm{B}\left(\frac{1}{2}\right)$. Clearly, $F_{X_n}(x) \to F_X$ since $F_{X_n}(x) = F_X \;\forall\, x \;\forall\, n$ and $X_n \rightsquigarrow X$. But

$$|X_n - X| = \begin{cases} 0 & \text{for } n \text{ odd}\\ |1 - 2X| & \text{if } n \text{ is even}\end{cases}$$

is equal to 1 almost surely if n is even and $\mathsf{P}(|X_n - X| \geq \varepsilon) \nrightarrow 0$ as $n \to \infty$ and hence X_n does not converge in probability to X. □

Convergence in distribution is a mild concept, but because we are often given the distribution function, this is a fundamental concept.

Remark

$X_n \rightsquigarrow X$ and $X_n \rightsquigarrow Y$, then $X \overset{\mathcal{L}}{=} Y$. This is an appealing result if we have weak convergence, we can construct a probability space and random variable such that it converge almost surely. This is not practical, yet useful for proofs.

Theorem 17.9

Let $\{X_n\}$ be random variables and a constant $c \in \mathbb{R}$. Then, if $X_n \rightsquigarrow c$, then $X_n \xrightarrow{\mathsf{P}} c$.

Proof Let $\varepsilon > 0$ be given. We have

$$\begin{aligned}\mathsf{P}(|X_n - c| > \varepsilon) &= \mathsf{P}(X_n > c + \varepsilon) + \mathsf{P}(X_n < c - \varepsilon) \\ &\leq 1 - \underbrace{\mathsf{P}(X_n \leq \varepsilon + c)}_{\to 1} + \underbrace{\mathsf{P}(X_n \leq c - \varepsilon)}_{\to 0} \to 0\end{aligned}$$

□

Example 17.6 (Order statistics: maximum)

Let $U_1, \ldots, U_n \sim \mathcal{U}(0,1)$ and identically and independently distributed. Let us look at a new series of random variable $M_n = \max(U_1, \ldots, U_n)$ to get an upper bound to know what are bounds of the uniforms to get an approximation of the highest values that can be encountered.

$$\begin{aligned}\mathsf{P}(M_n \leq x) &= \mathsf{P}(U_1 \leq x, \ldots, U_n \leq x) \\ &= \prod_{i=1}^{n} \mathsf{P}(U_i \leq x) \\ &= \begin{cases} 0 & \text{if } x < 0 \\ x^n & \text{if } x \in [0,1] \\ 1 & \text{if } x > 1. \end{cases}\end{aligned}$$

Hence the probability that $M_n \leq x$ is given by

$$\mathsf{P}(M_n \leq x) \to \begin{cases} 0 & \text{if } x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

and $M_n \rightsquigarrow 1 \to M_n \xrightarrow{\mathsf{p}} 1$. But what is the error you make by choosing the maximum instead? $n(1 - M_n) \rightsquigarrow \mathcal{E}(1)$ since

$$\mathsf{P}(n(1-M_n) \leq x) = \mathsf{P}\left(1 - \frac{x}{n} \leq M_n\right) = 1 - \mathsf{P}\left(M_n \leq 1 - \frac{x}{n}\right)$$

$$= \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - \left(1 - \frac{x}{n}\right)^n & \text{if } x \in (0, n) \\ n & \text{if } x \geq n \end{cases} \xrightarrow{n \to \infty} \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-x} & \text{if } x > 0 \end{cases}$$

Theorem 17.10 (Central limit theorem)

We saw that $\overline{X_n} \xrightarrow{\mathsf{p}} \mu$. Consider the proper rescaling $\sqrt{n}(\overline{X_n} - \mu)$, where $\{X_n\}$ is a sequence of i.i.d random variables such that $\mathsf{E}X_1 = \mu < \infty$, and that $\mathsf{Var}(X_1) = \sigma^2 < \infty$.[21] By the Weak law of large numbers, we have $\overline{X_n} \xrightarrow{\mathsf{p}} \mu$. Then

$$\sqrt{n}\left(\frac{\overline{X_n} - \mu}{\sigma}\right) \rightsquigarrow \mathcal{N}(0,1)$$

for n large.

Proof We need to resort to one thing. If we look at our object of interest, we can rewrite in a little different way. Consider

$$\sqrt{n}\left(\frac{\overline{X_n} - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)\right)$$

[21] Without loss of generality, we consider X_1 since all variables have the same distribution.

is such that $\left(\frac{X_i-\mu}{\sigma}\right)$ are independently and identically distributed and

$$\mathsf{E}\left(\frac{X_i-\mu}{\sigma}\right) = \mathsf{E}Y_n = 0 \qquad \mathsf{Var}\left(\frac{X_i-\mu}{\sigma}\right) = \mathsf{Var}Y_n = 1.$$

for Y_1, Y_2 i.i.d.. Since we have no info as in the example, we need to call on the following result which go beyond the scope of what we can prove here. If we have a sequence $\{Z_n\}, Z$ be such that $M_{Z_n}(t) \to M_Z(t)$ (convergence of the moment generating function) for all $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$, then $Z_n \rightsquigarrow Z$.

Assume that Y_i have a MGF [22]. Then

$$M_{\sqrt{n}(\overline{Y_n})}(t) = \mathsf{E}\left(e^{\frac{1}{\sqrt{n}}(Y_1+\ldots+Y_n)t}\right) = \prod_{i=1}^{n} \mathsf{E}\left(e^{\frac{t}{\sqrt{n}}Y_i}\right) = \left(M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

If we look at the moment generating function of Y_1 as a power series, we have

$$\begin{aligned} M_{Y_i}(s) &= M_{Y_i}(0) + M'_{Y_i}(0)\cdot s + M''_{Y_i}(0)\cdot\frac{s^2}{2!} + M'''_{Y_i}(0)\cdot\frac{s^3}{3!}\cdots \\ &= 1 + \mathsf{E}Y_1\cdot s + \mathsf{E}Y_1^2\cdot\frac{s^2}{2!} + \mathsf{E}Y_1^3\cdot\frac{s^3}{3!}\cdots \\ &= 1 + \frac{s^2}{2!} + \mathsf{E}Y_1^3\cdot\frac{s^3}{3!} + \ldots + \mathsf{E}Y_1^k\cdot\frac{s^k}{k!} \end{aligned}$$

by assumption, Now suppose that we have in an interval of convergence

$$M_{Y_1}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + \mathcal{O}\left(n^{-\frac{3}{2}}\right) < 1 + \frac{t^2}{2n} + \mathcal{O}\left(\frac{1}{n}\right)$$

since it goes faster to zero than $\frac{1}{n}$. Hence,

$$M_{\sqrt{n}(\overline{Y_n})}(t) = \left(1 + \frac{t^2}{2n} + \mathcal{O}\left(\frac{1}{n}\right)\right)^n \xrightarrow{n\to\infty} e^{\frac{t^2}{2}}$$

which is precisely the moment generating function of the $\mathcal{N}(0,1)$. □

[22] Otherwise, the same result can be established using the characteristic function; the proof uses the same idea.

Example 17.7

Suppose that $X \sim \mathcal{B}(n,p) \overset{\mathcal{L}}{=} \sum_{i=1}^n Y_i$, where Y_i are i.i.d. $\mathcal{B}(1,p)$ and we look at $\mathsf{P}(X \leq x)$. We can use our brand new tool to approximate the distribution of Y.

$$\begin{aligned}\mathsf{P}(X \leq x) &= \mathsf{P}\left(\sum_{i=1}^n Y_i \leq x\right) = \mathsf{P}\left(\frac{1}{n}\sum_{i=1}^n Y_i \leq \frac{x}{n}\right)\\ &= \mathsf{P}\left(\sqrt{n}\frac{\overline{Y_n} - p}{\sqrt{p(1-p)}} \leq \sqrt{n}\frac{\frac{x}{n} - p}{\sqrt{p(1-p)}}\right) \approx \Phi\left(\sqrt{n}\frac{\frac{x}{n} - p}{\sqrt{p(1-p)}}\right)\end{aligned}$$

for n large enough.

Example 17.8

Suppose that we have $Y \sim \mathcal{P}(n)$, for $n \in \mathbb{N}$ and $Y \overset{\mathcal{L}}{=} \sum_{i=1}^n X_i$, and X_i are i.i.d. $\mathcal{P}(1)$. In this case, by the Central Limit theorem,

$$\begin{aligned}\mathsf{P}(Y \leq y) &= \mathsf{P}\left(\sum_{i=1}^n X_i \leq y\right) = \mathsf{P}\left(\sqrt{n}\,\frac{\overline{X_n} - 1}{1} \leq \frac{y}{\sqrt{n}} - \sqrt{n}\right)\\ &\approx \Phi\left(\frac{y}{\sqrt{n}} - \sqrt{n}\right)\end{aligned}$$

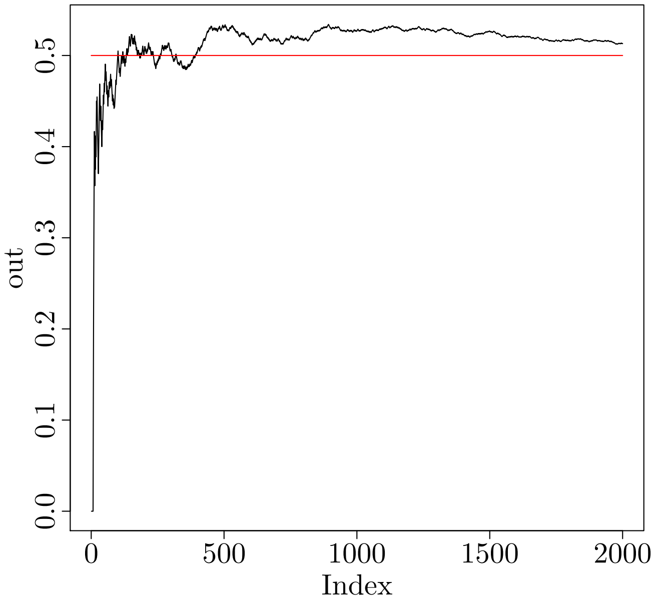
◇◇◇

Using R

R is a free statistics software alimented by a large community of users. It has numerous packages for download, which allow you to perform specialized operations ranging from econometrics to survival analysis. For example, the `stats` package (which is automatically loaded) allow to do hypothesis testing, generate data on random variables or draw distribution functions. Particular instances of those packages can be loaded when needed; you only need to type the command `library()` with the input you require. For more information, we refer the reader to the online documentation. Nevertheless, it you need to look for functionalities and inputs, the following commands are useful:
`> library(help = "stats")` to get all possibilities of a package;
`> help("dnorm")` in case you know the syntax of your command;
`> help.search("normal distribution")`if you are looking for an unknown function, but do not know the exact syntax or its properties and inputs. One can also use the `?"dnorm"` to replace `help()` and `??"normal distribution"`.

R language is quite intuitive and people who are familiar with programming languages should find it is a valuable (yet free) tool, highly customizable. Here are some examples from class and the resulting graphs:

Figure 22: Coin toss experiment repeated 1000 times.



```
coin <- function(N=1000,p=0.5){
X <- rbinom(n=N,size=1,prob=p)
A <- numeric(N)
for(i in 1:N){
A[i] <- mean(X[1:i])
}
A
}

out <- coin(2000)

plot(out,type="l")
lines(c(0,2000),c(0.5,0.5),col=2)
```

```
# Binomial distribution
BiPlot <- function(n=10,p=0.5,col="black")
{
x <- 0:n
y <- dbinom(x,size=n,prob=p)
plot(x,y,col=col)
for(i in 1:(n+1)){
lines(c(x[i],x[i]),c(0,y[i]),col=col)
}
}
par(mfrow=c(1,3))
BiPlot(5,0.5)
BiPlot(10,0.5)
BiPlot(30,0.5)
```

In this case, we see that the command `BiPlot()` takes as input the number of trial and the probability.

```
#Poisson distribution

PoisPlot <- function(lambda=2,N=10,col="black")
{
x <- 0:N
y <- dpois(x,lambda=lambda)
plot(x,y,col=col)
for(i in 1:(N+1)){
lines(c(x[i],x[i]),c(0,y[i]),col=col)
}
}
PoisPlot(lambda=1)
PoisPlot(lambda=2)
PoisPlot(lambda=5)
PoisPlot(lambda=20,N=50)
```

Figure 23: The Binomial distribution

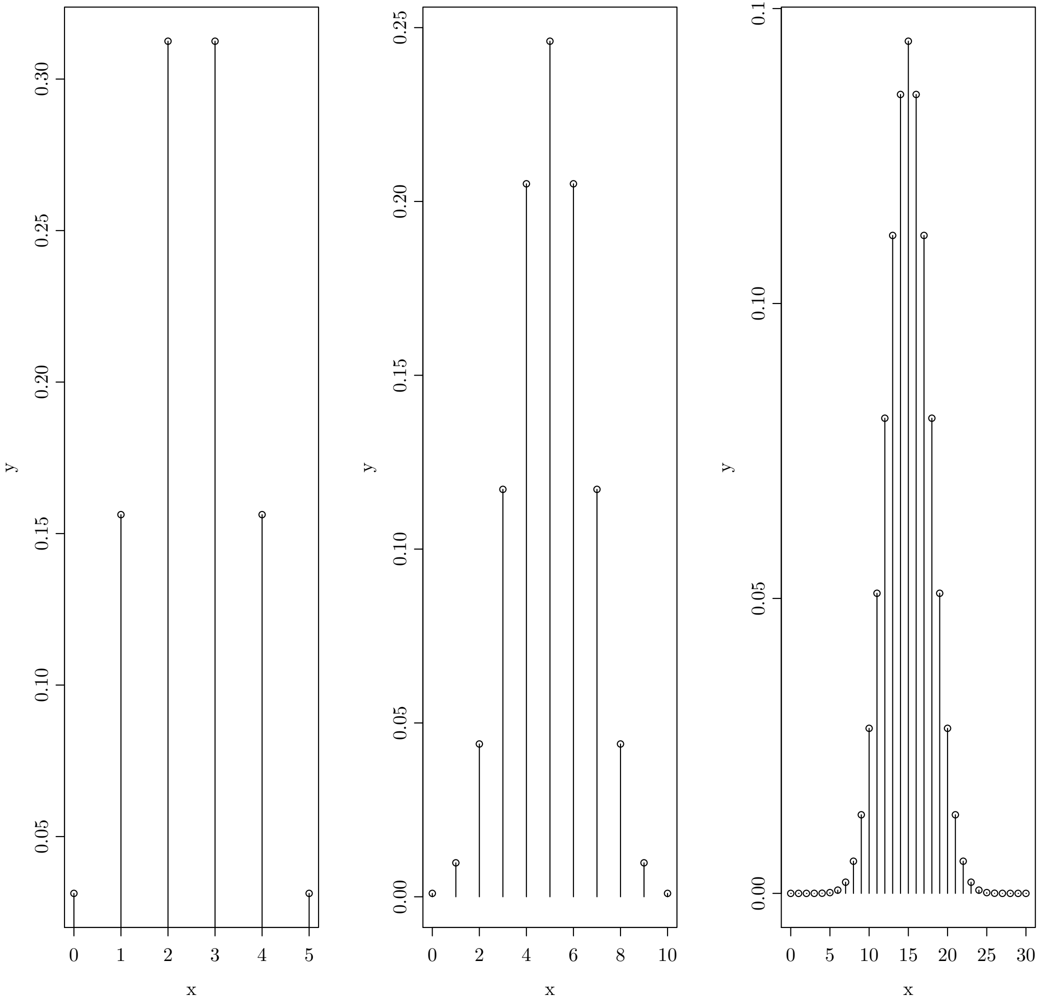


Figure 24: The Poisson distribution with $\lambda = 1$ and $\lambda = 2$

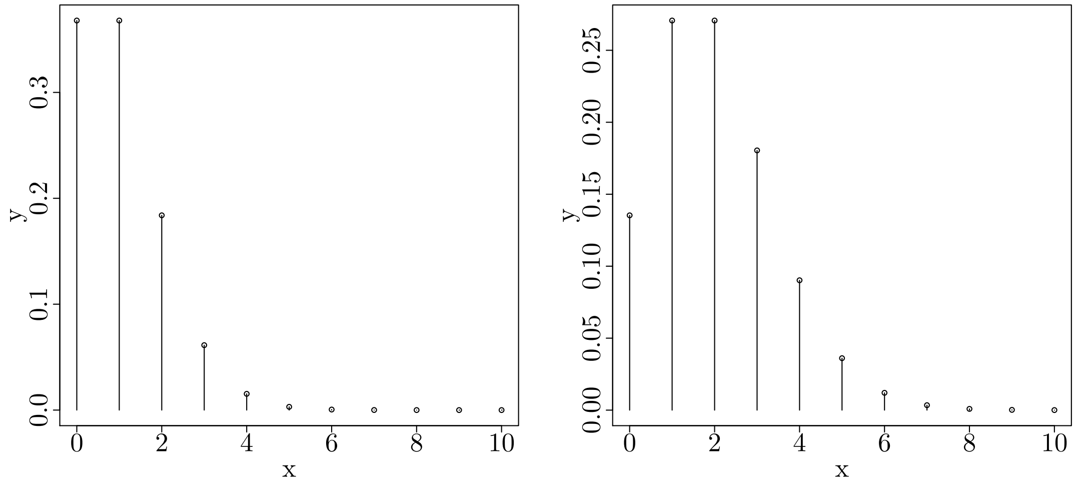
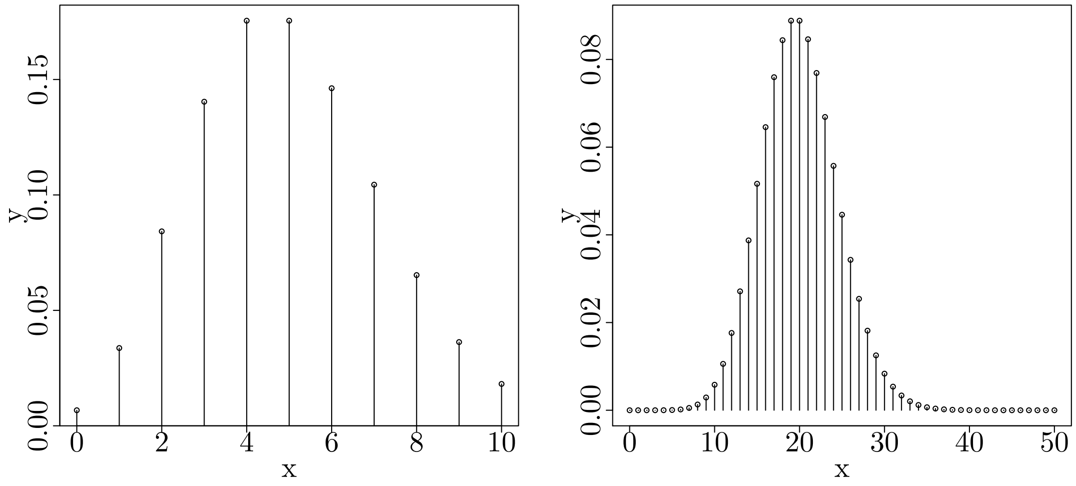


Figure 25: The Poisson distribution with $\lambda = 5, N = 10$ and $\lambda = 20, N = 50$



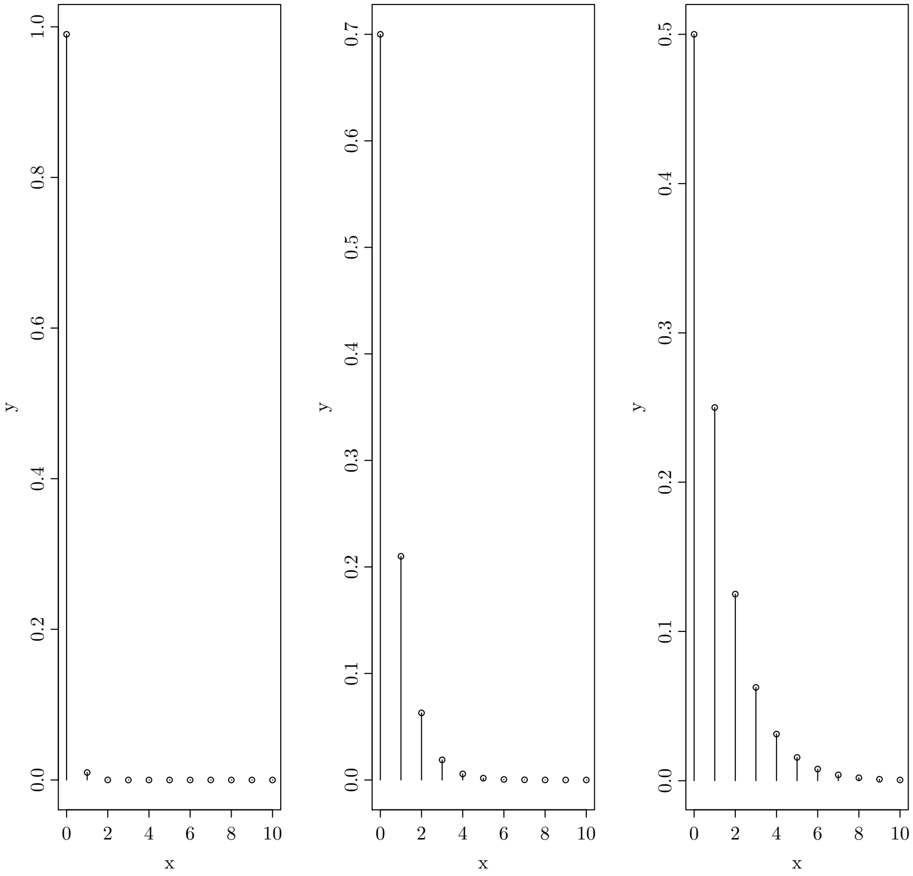


Figure 26: Geometric distribution with probabilities 0.99, 0.7 and 0.5

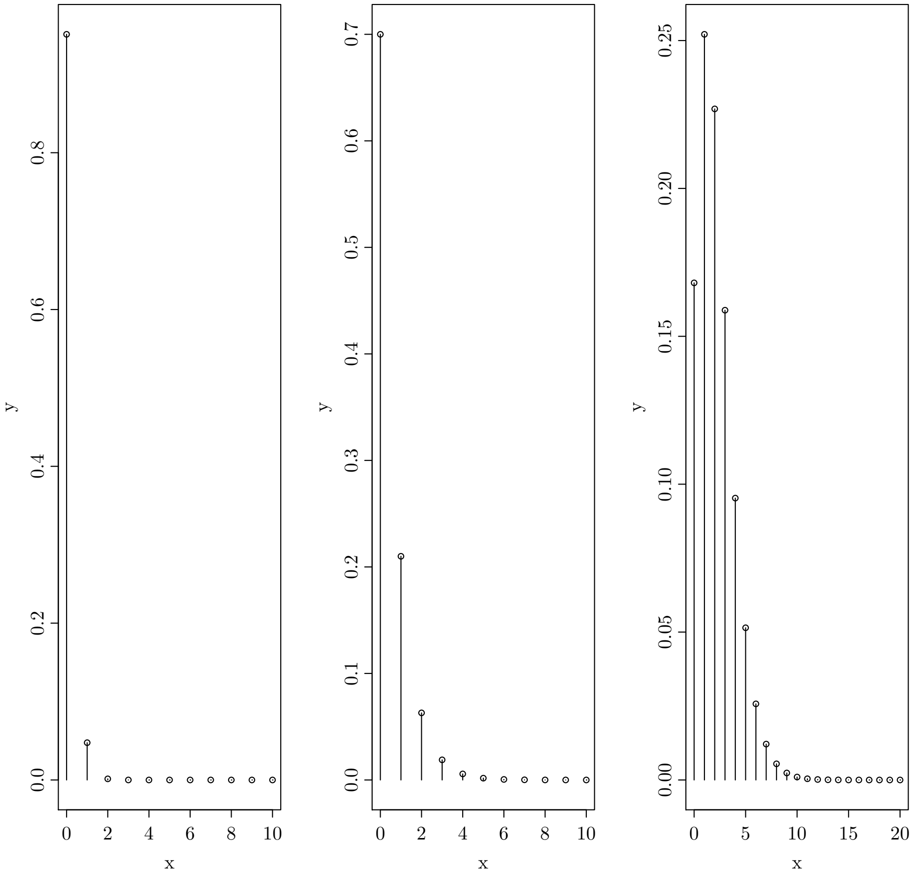


Figure 27: Negative binomial distribution

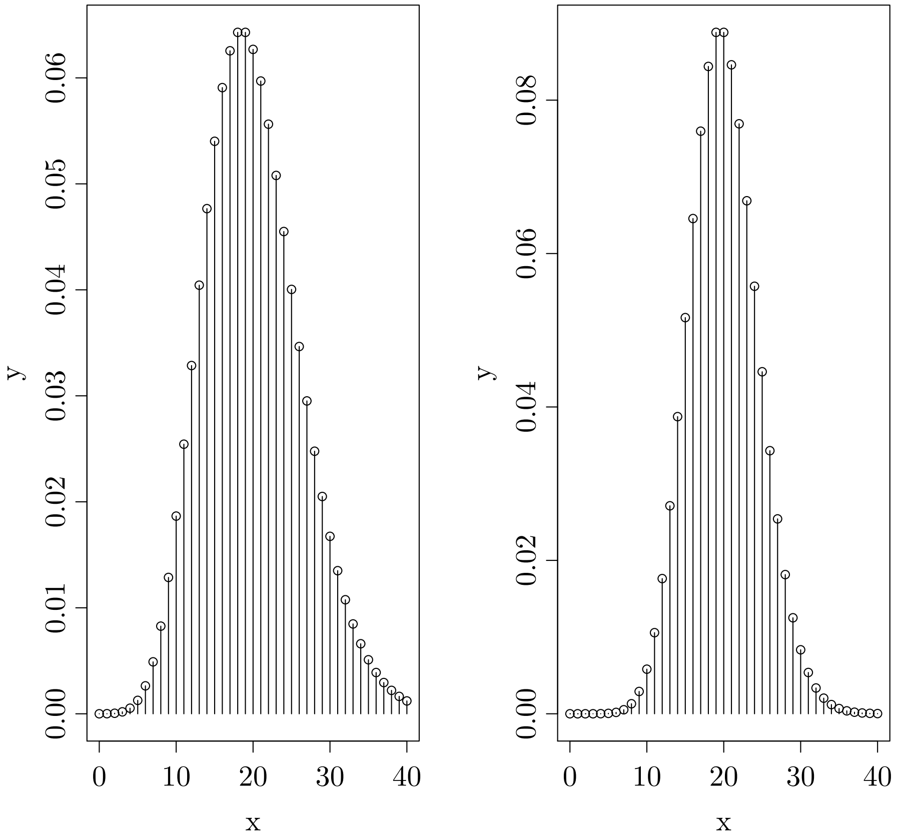


Figure 28: Comparison between the Negative binomial and the Poisson distributions

```
# Geometric distribution

GeoPlot <- function(prob=0.99,N=10,col="black")
{
x <- 0:N
y <- dgeom(x,prob=prob)
plot(x,y,col=col)
for(i in 1:(N+1)){
```

```
lines(c(x[i],x[i]),c(0,y[i]),col=col)
}
}

GeoPlot(prob=0.99)
GeoPlot(prob=0.7)
GeoPlot(prob=0.5)

# Negative Binomial distribution

NBPlot <- function(size=5,prob=0.5,N=10,col="black")
{
x <- 0:N
y <- dnbinom(x,size=size,prob=prob)
plot(x,y,col=col)
for(i in 1:(N+1)){
lines(c(x[i],x[i]),c(0,y[i]),col=col)
}
}
NBPlot(prob=0.99)
NBPlot(size=1,prob=0.7)
NBPlot(size=5,prob=0.7,N=20)
par(mfrow=c(1,2))
NBPlot(size=20,prob=0.5,N=40)
PoisPlot(lambda=20,N=40)
```

Examples of expectations for a dice, and a Normal and Cauchy random variable. The code will display the arithmetic mean for 5000 tries.

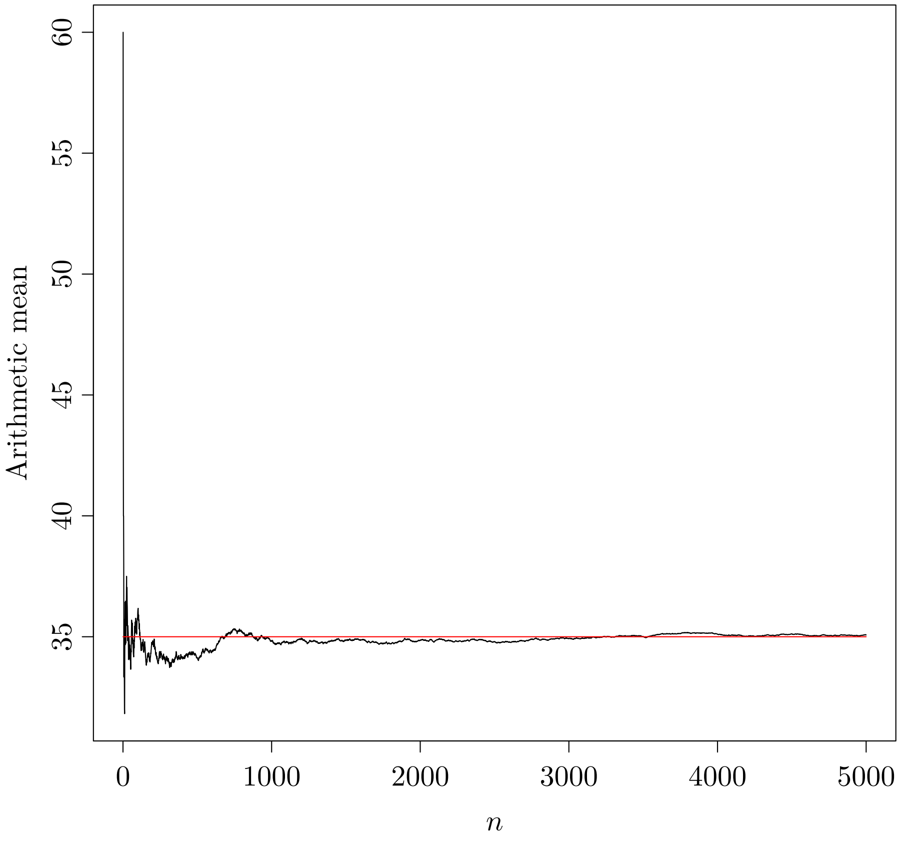


Figure 29: Mean of a dice

```
rdice <- function(n=100){
U <- runif(n=n)
floor(6*U)+1
}

mean.dice <- function(n=5000){
```

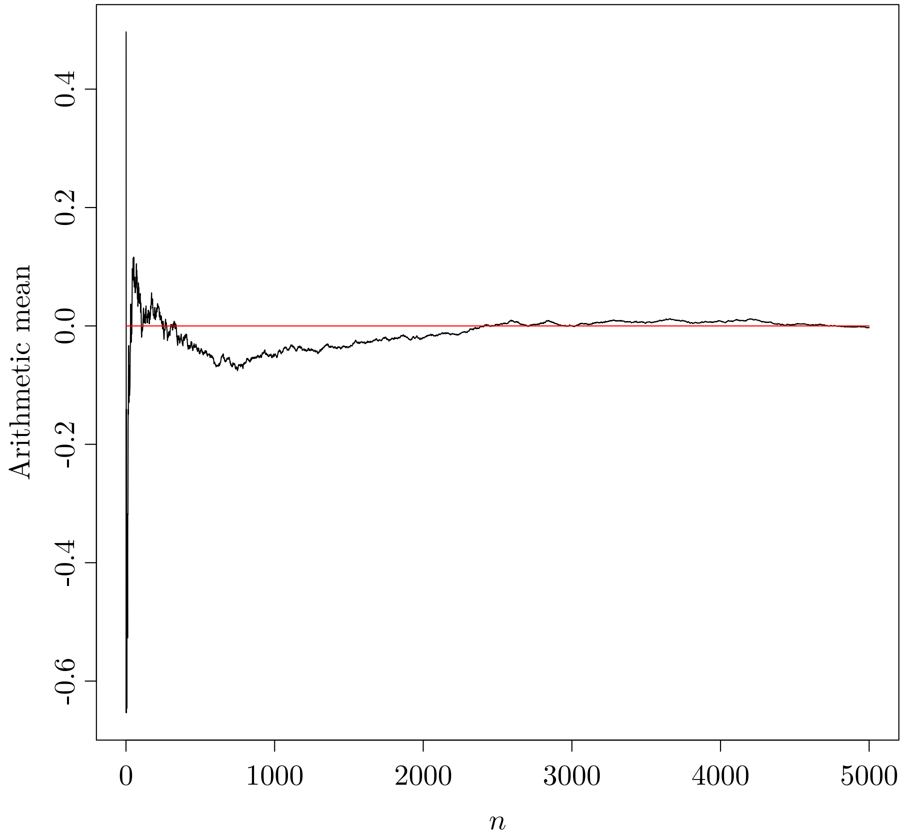


Figure 30: Mean of a Gaussian random variable

```
X <- rdice(n)
m <- cumsum(X*10)
N <- 1:n
plot(m/N,type="l")
lines(c(0,n),c(35,35),col=2)
}
mean.dice()
```

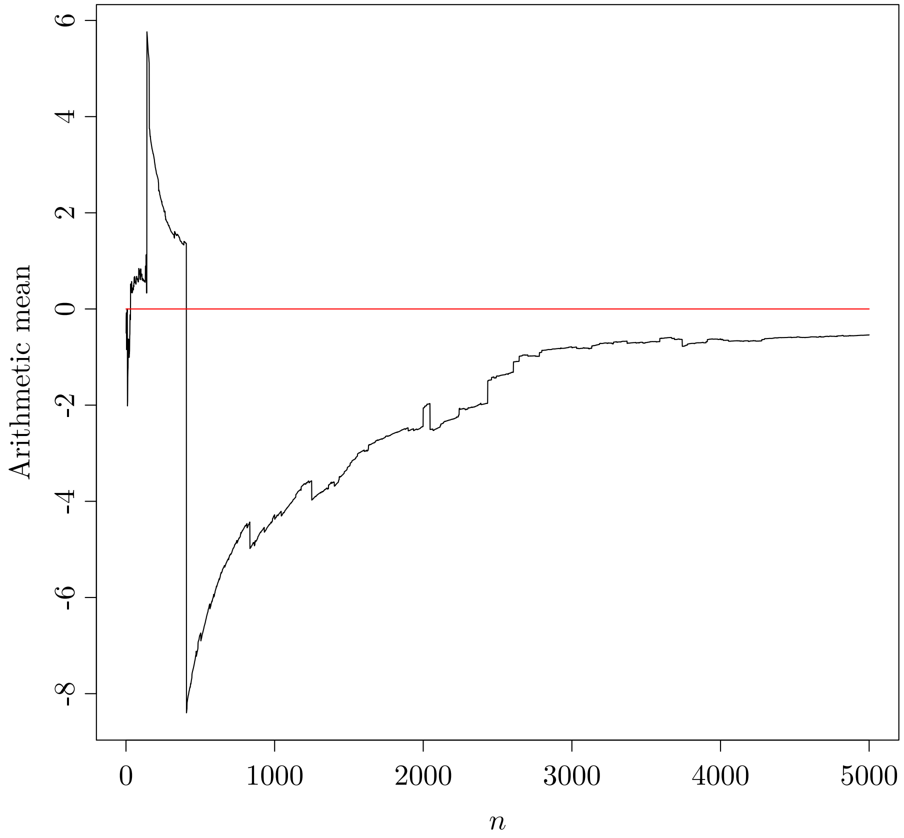


Figure 31: Mean of a Cauchy random variable

```
mean.gauss <- function(n=5000,mu=0,sigma=1){
X <- rnorm(n=n,mean=mu,sd=sigma)
m <- cumsum(X)
N <- 1:n
plot(m/N,type="l")
lines(c(0,n),c(mu,mu),col=2)
}
mean.gauss()
mean.cauchy <- function(n=5000){
X <- rt(n=n,df=1)
m <- cumsum(X)
N <- 1:n
plot(m/N,type="l")
lines(c(0,n),c(0,0),col=2)
}
mean.cauchy()
```

The first two distributions illustrate what happens with an accumulation of points along the line; it has no density since the volume of the line is zero (see Example 12.4). The second distribution stresses out the fact that the support is countable (see Example 12.6). Note that the Quantitative Risk Management library is considered obsolete and is not supported in newer version of R.

```
###### Other distributions #########

U <- runif(1000)
plot(U,U)

library(QRMlib)
require(grDevices)
out <-rcopula.clayton(1000,theta=2,d=2)
X <- qpois(out[,1],lambda=2)
Y <- qbinom(out[,2],prob=0.5,size=10)
plot(X,Y)
hist(X)
hist(Y)
```

Three-dimensional plots of the Bivariate Normal distribution, with contour plot or perspective (3-D)

```
##### Bivariate Normal distribution #######



BiDensPlot(func=dmnorm,mu=c(0,0),Sigma=equicorr(2,0),c(-3, 3),
 ypts = c(-3, 3),type="contour");

BiDensPlot(func=dmnorm,mu=c(0,0),Sigma=equicorr(2,0),c(-3, 3),
 ypts = c(-3, 3),type="persp");
```

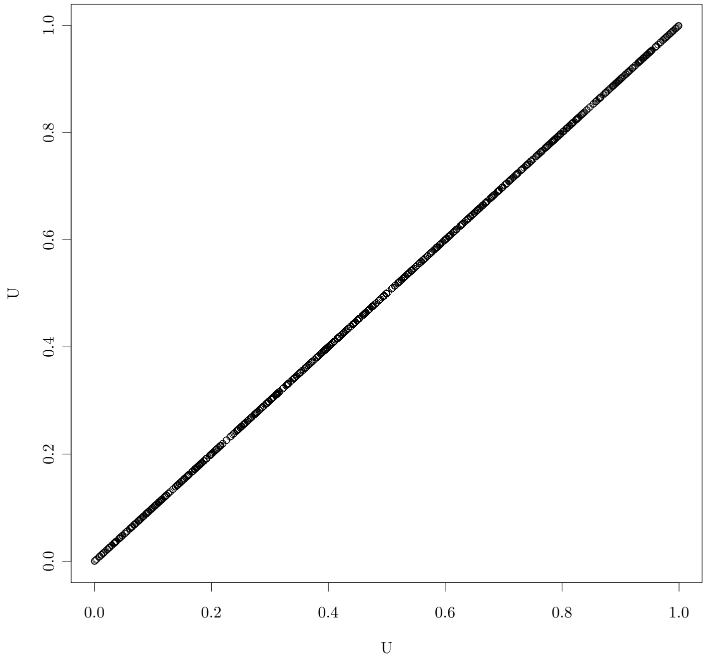


Figure 32: Plot of a Bivariate Uniform random variable

```
par(mfrow=c(1,3))
BiDensPlot(func=dmnorm,mu=c(0,0),Sigma=equicorr(2,0),c(-3, 3),
 ypts = c(-3, 3),type="contour");
BiDensPlot(func=dmnorm,mu=c(0,0),Sigma=equicorr(2,0.4),c(-3, 3),
 ypts = c(-3, 3),type="contour");
BiDensPlot(func=dmnorm,mu=c(0,0),Sigma=equicorr(2,-0.7),c(-3, 3),
 ypts = c(-3, 3),type="contour");

X1 <- rmnorm(1000,rho=0)
X2 <- rmnorm(1000,rho=0.4)
X3 <- rmnorm(1000,rho=-0.7)

plot(X1)
```

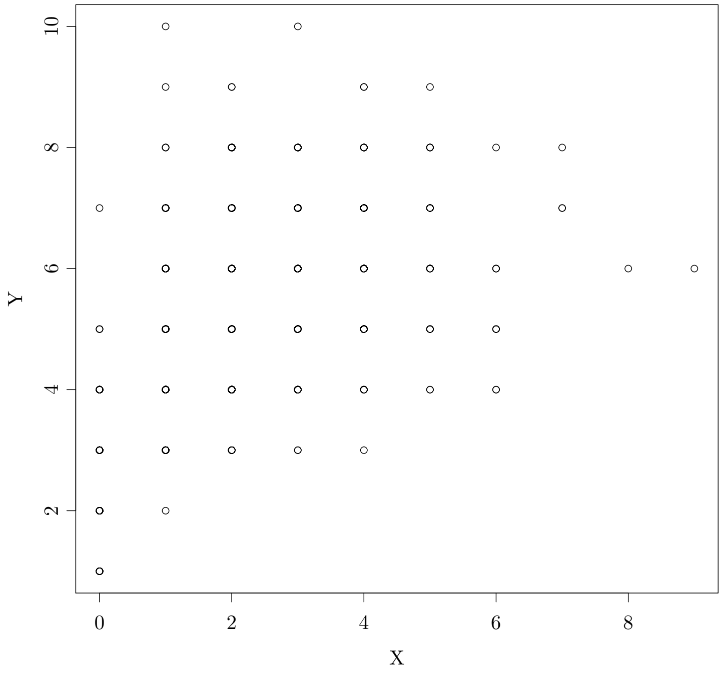
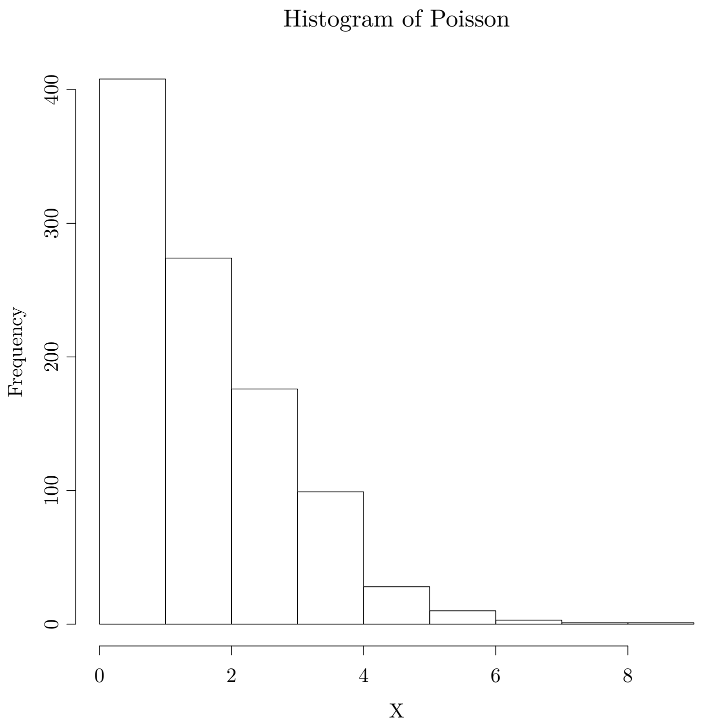


Figure 33: Bivariate Poisson-Binomial random variable

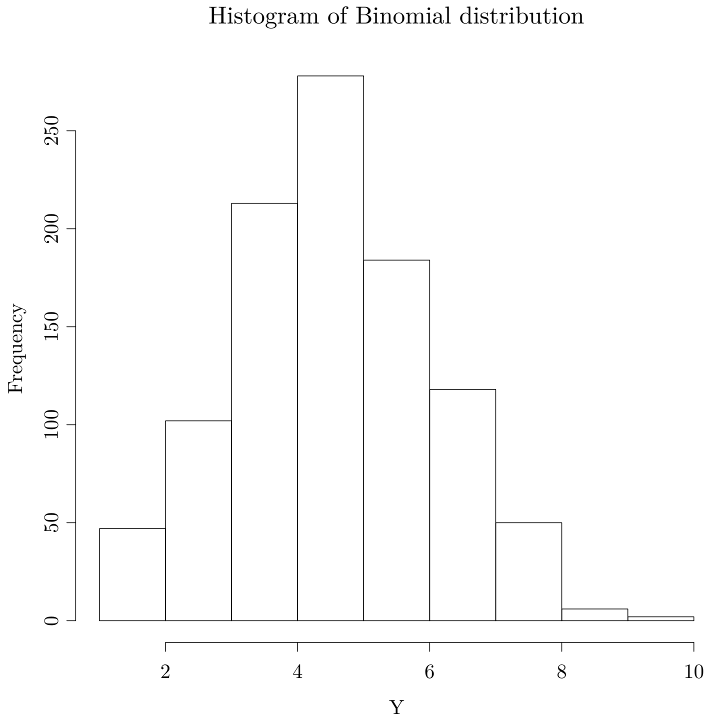
```
plot(X2)
plot(X3)

BiDensPlot(func=dmnorm,mu=c(0,0),Sigma=equicorr(2,0),c(-3, 3),
 ypts = c(-3, 3),type="contour");
BiDensPlot(func=dmnorm,mu=c(0,0),Sigma=equicorr(2,0.7),c(-3, 3),
 ypts = c(-3, 3),type="contour");
BiDensPlot(func=dmnorm,mu=c(1,1),Sigma=equicorr(2,0.7),c(-3, 3),
 ypts = c(-3, 3),type="contour");

X1 <- rmnorm(1000,rho=0)
X2 <- rmnorm(1000,rho=0.7)
X3 <- rmnorm(1000,rho=0.7,mu=c(1,1))
```



```
plot(X1,xlim=c(-5,5),ylim=c(-5,5))
plot(X2,xlim=c(-5,5),ylim=c(-5,5))
plot(X3,xlim=c(-5,5),ylim=c(-5,5))
```


Histogram of Binomial distribution
Frequency
0
50
100
150
200
250
2
4
6
8
10
Y

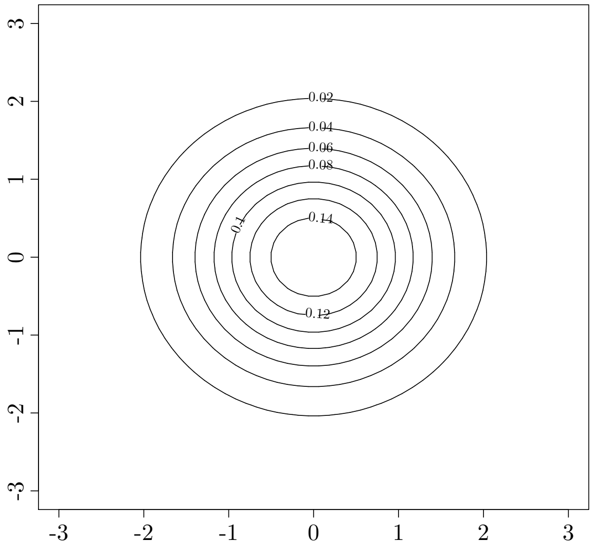


Figure 34: Bivariate Normal distribution-Contour

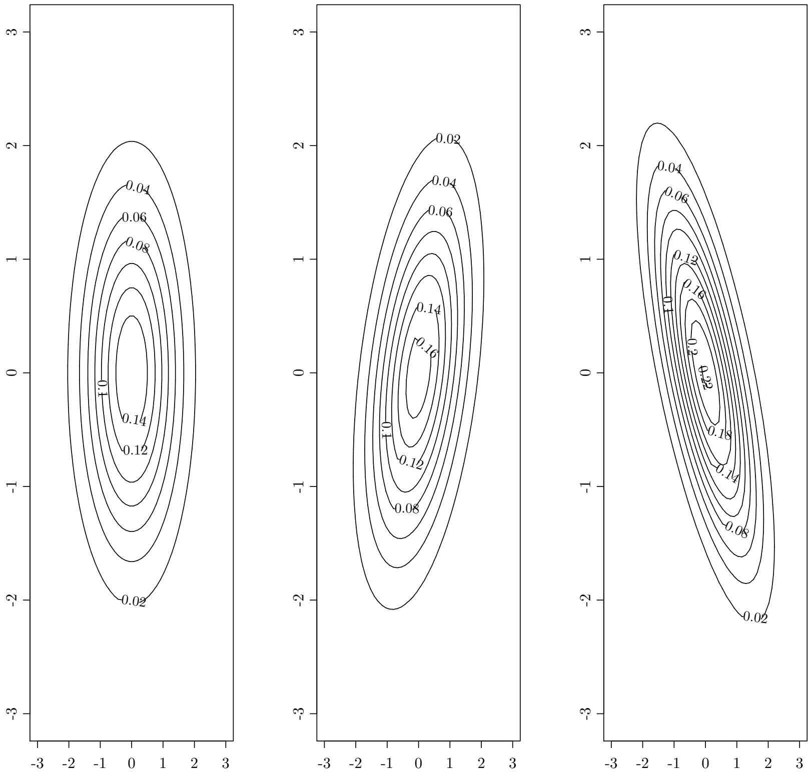


Figure 35: Comparisons between bivariate Normal distributions.

Index

References

[1] Rohatgi, V.K., A.K.M.E. Saleh, *An introduction to Probability and Statistics*, 2nd edition, Wiley, 2001, 716p.

[2] Johanna Nešlehová, *MATH 356: Honours Probability*, Notes taken from September to December 2011 at McGill University.

I am grateful to Pr. Nešlehová for the content of these notes and the great lectures she gave. I also need to credit the following for pictures and images used through this set of notes:

1. Cover image: Daniel L. Lu for the Brownian motion image;
2. Figure 2: Till Tantau for the Venn Diagram;
3. Figure 14: Keven Goulding for the R code of the Normal distribution;
4. Figure 17: Mike Toews for the Gaussian distribution;

All R code in the Using R chapter is the work of Pr. Nešlehová and was compiled using TikZdevice and R 2.11. All other illustrations were made with TikZ, GnuPlot or R.

License