Probability & Statistics II

Contents

Week 1

Discrete random variables

A discrete random variable X has the following properties:

- X can assume any value x from a set of discrete values with a certain probability.
- The discrete set of values can be finite or countably infinite.
- The probability that X takes the particular value x is denoted by $P(X = x)$. The collection of the probabilities of all possible values x is called the *probability distribution* of X.
- The function $p_X(x) = P(X = x)$ is called the *probability mass function (pmf)* of X.

Properties of probability mass functions

The pmf p_X has the following important two properties:

(i). If X takes discrete values from the set D, then $\sum_{x\in D} p_X(x) = 1$.

(ii). $p_X(x) \geq 0$ for any $x \in D$.

For any $S \subseteq D$, we also have $P(X \in S) = \sum_{x\in S} p_X(x)$

Expectation & Variance

Definition. The expectation of a discrete random variable X is defined by

$$E[X] := \sum_x x\, P(X = x) \qquad , \text{ if } \sum_x |x|\, P(X = x) < \infty\,,$$

where the sum is taken over all possible values of the random variable X.

Some of the other terms for "expectation of X" used in the mathematical literature (and in these notes) are "mean value of X", just "mean", "first moment of X", or in rough terms "average value of X".

Theorem 1. *Let X be a discrete random variable. Consider a new random variable $Y = g(X)$, where g is a function $g\colon \mathbb{R} \to \mathbb{R}$.*
If $\sum_x |g(x)|P(X = x) < \infty$, then

$$E(Y) = E[g(X)] = \sum_x g(x)P(X = x), \tag{1}$$

where the sum is taken over all possible values of the random variable X.

Remark. By the definition of the expectation, $E(Y) = \sum_y yP(Y = y)$. Formula (1) allows one to compute $E(Y)$ without first computing the probabilities $P(Y = y)$.

Definition. The variance of random variable X is defined by

$$\mathrm{Var}(X) := E[(X - \mu)^2], \quad \text{where } \mu = E[X] \text{ is the expected value of } X.$$

Exercise. *Prove that* $\mathrm{Var(X)} := \mathrm{E(X^2)} - \mu^2$.

Continuous random variables

A continuous random variable X has the following properties:

- X can assume any value x from a continuum of values.
- The set of values is infinite (since X can assume any real number from a subset of $\mathbb{R}$).

Definition. A random variable X is *continuous* if there is a function f_X such that for all intervals $[a, b]$, we have

$$P(X \in [a, b]) = \int_a^b f_X(x)dx.$$

We call the function f_X the *probability density function* of X (or *pdf* of X).

Properties of probability density functions

The probability density function f_X has the following important two properties:

(i). $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,dx = 1.$

(ii). $f_X(x) \geq 0$ for all x.

Cumulative distribution functions

Definition. The cumulative distribution function $F_X(x)$ of a random variable X is defined by

$$F_X(x) := P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt.$$

Recall (from calculus) that this formula can be differentiated at all points for which the probability density function is continuous. Hence for these points

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

Exercise. *Show that $F_X(x)$ is a monotone increasing function of x with $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.*

Expectation

Definition. Let X be a continuous random variable with probability density function $f_X(x)$. If $\int_{-\infty}^{\infty} |x| f_X(x)dx < \infty$, then the expectation of X is defined by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx.$$

Theorem 2. *Let X be a continuous random variable with probability density function $f_X(x)$ and $Y = g(X)$ be a new random variable, where $g : \mathbb{R} \mapsto \mathbb{R}$ is a function. If $\int_{-\infty}^{\infty} |g(x)| f_X(x)dx < \infty$, then*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx. \tag{2}$$

Remark. By the definition of the expectation, $E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy$. Formula (2) allows one to compute $E(Y)$ without first computing the probabilities $f_Y(y)$.

1 Conditional Probability

This section will start as a small revision on events and then will cover new material.

1.1 Conditional Probability of Events (Revision)

Definition. Suppose A and B are events and that $P(B) > 0$. The conditional probability of A given B is defined by

$$P(A \mid B) \stackrel{\text{def}}{=} \frac{P(A \cap B)}{P(B)}.$$

Remark. The condition that $P(B) > 0$ is needed: the definition does not make sense without this. Conditioning on events which have probability zero is a common way of "proving" false results!

There are two key reasons we care about conditional probability:
– we want to know the conditional probability itself,
– we use it as a tool for calculating actual "unconditional" probabilities.

Typically, when doing the former, we either calculate directly the conditional probability from the definition, or we use the Bayes Theorem (see Introduction to Probability). To do the latter, the following theorem (see Introduction to Probability) is very useful. Before stating it we recall the following definition.

Definition. A partition of a space of elementary outcomes Ω is a collection of events $B_1, B_2, B_3, \ldots, B_n$ which are pairwise disjoint and whose union is the whole set Ω.
In other words, $B_1, B_2, B_3, \ldots, B_n$ form a partition if:
(i) $B_i \cap B_j = \emptyset$ for all $i \neq j$,
(ii) $\cup_{j=1}^{n} B_j = \Omega$.

The first condition says that no two (or more) of the events can occur together. The second condition says that at exactly one event one must occur.

Theorem 3 (Theorem of Total Probability)**.** *Suppose that A is an event, that $B_1, B_2, \ldots, B_n$ form a partition and that $P(B_i) > 0$ for all $i = 1, \ldots, n$. Then*

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i).$$

Proof. Can be found in Introduction to Probability. □

Remark. This theorem is useful when the conditional probabilities are easy to calculate.

Example. *Suppose that we roll two fair dice. What is the probability the product is even?*

Answer. Let A be the event that the product is even.
Let B_1 be the event that the first die is odd.
Let $B_2 = B_1^c$ be the event the first die is even.
These events are a partition so we can use the Theorem of Total Probability.
First we calculate $P(B_1)$ and $P(B_2)$. Obviously these are both $1/2$.
Now what is $P(A \mid B_1)$? Since the first die is odd the product is even if and only if the second die is even, which has probability $1/2$. Thus $P(A \mid B_1) = 1/2$.

Next $P(A \mid B_2)$. Since the first die is even the product is always even whatever the second die is. Hence $P(A \mid B_2) = 1$.

Putting these together we see that

$$P(A) = P(A \mid B_1)\, P(B_1) + P(A \mid B_2)\, P(B_2) = \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{3}{4}.$$

□

Example. *Alice and Bob play the following game. They roll a fair dice. If it comes up 1 or 2 Alice wins, if it comes up 6 Bob wins, if it comes up 3,4,5 they play again. What is the probability Alice wins?*

Answer. Let A be the event Alice wins. Let B_1 be the event the first roll is a 1 or 2, B_2 be the event that the first roll is a 6 and B_3 the event that the first roll is a 3,4 or 5. Obviously B_1, B_2, B_3 form a partition, and $P(B_1) = 1/3$, $P(B_2) = 1/6$, and $P(B_3) = 1/2$.

If B_1 occurs, then Alice wins: i.e., $P(A \mid B_1) = 1$.

If B_2 occurs then Bob wins (so Alice does not): i.e., $P(A \mid B_2) = 0$.

If B_3 occurs, then we observe that this is just as if the first roll never happened: i.e., $P(A \mid B_3) = P(A)$.

We apply the Theorem of Total Probability:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + P(A \mid B_3)P(B_3) = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{6} + P(A) \cdot \frac{1}{2}.$$

By rearranging we see that $P(A) = 2/3$. □

1.2 Conditional distribution of Discrete random variables

Suppose that X is a discrete random variable and that B is an event with $P(B) > 0$. Just as with conditional probability we can ask what do we know about X if we are told that B occurs. Indeed, we can define a new random variable $Y = X \mid B$ where

$$P(Y = x) \stackrel{\text{def}}{=} P(X = x \mid B) = \frac{P(\{X = x\} \cap B)}{P(B)}. \tag{3}$$

Example. *We toss a fair coin twice. What is the distribution of the number of heads given that the first toss is a head? What is the distribution of the number of heads given that the first toss is tail?*

Answer. Let X count the number of heads, let B_1 be the event that the first toss is a head and B_2 be the event that the first toss is a tail. Then $P(B_1) = \frac{1}{2}$ and $P(B_2) = \frac{1}{2}$.

The distribution of $X \mid B_1$ is

$$P(X = 0 \mid B_1) = 0 \qquad P(X = 1 \mid B_1) = 1/2 \qquad P(X = 2 \mid B_1) = 1/2.$$

Similarly we compute the distribution of $X \mid B_2$ as

$$P(X = 0 \mid B_2) = 1/2 \qquad P(X = 1 \mid B_2) = 1/2 \qquad P(X = 2 \mid B_2) = 0$$

□

Since $X \mid B$ is a random variable we can talk about the expectation of X given B which we write $E(X \mid B)$: this is defined in the obvious way as

$$E(X \mid B) = \sum_x x\, P(X = x \mid B). \tag{4}$$

Example. *In the previous example: what is the expected number of heads given that the first toss is a head? What if the first toss is a tail?*

Answer. Thus

$$E(X \mid B_1) = 0 \cdot 0 + \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2 = \frac{3}{2}$$

and we see that

$$E(X \mid B_2) = 1/2.$$

□

In the same way, suppose that we have two random variables X and Y and that we are told the value of Y. Then, for any value y of Y for which $P(Y = y) > 0$, we can consider the conditional distribution of $X|\{Y = y\}$ as above in (3) and its conditional expectation as in (4), but this time the event $B = \{Y = y\}$:

$$E\,(X|Y = y) = \sum_x x\, P(X = x \mid Y = y) = \sum_x x\, \frac{P(X = x, Y = y)}{P(Y = y)}. \tag{5}$$

We will now look at a theorem, similar to the theorem of total probability, but now for expectations.

Theorem 4 (Theorem of Total Probability for Expectation)**.** *Suppose that X is a random variable and that $B_1, ..., B_n$ is a partition with $P(B_i) > 0$ for all i. Then*

$$E(X) = \sum_{i=1}^{n} E(X \mid B_i) P(B_i)$$

Proof. By definition

$$E(X) = \sum_x x P(X = x).$$

Since $B_1, ..., B_n$ is a partition (and $P(B_i) > 0$ for all i), the Theorem of Total Probability implies that

$$P(X = x) = \sum_{i=1}^{n} P(X = x \mid B_i) P(B_i).$$

Hence

$$\begin{aligned} E(X) &= \sum_x x P(X = x) \\ &= \sum_x x \sum_{i=1}^{n} P(X = x \mid B_i) P(B_i) \\ &= \sum_{i=1}^{n} \left(\sum_x x P(X = x \mid B_i) \right) P(B_i) \\ &= \sum_{i=1}^{n} E(X \mid B_i) P(B_i). \end{aligned}$$

□

Example. *Alice and Bob play the following game. They roll a fair dice. If it comes up 1 or 2 Alice wins, if it comes up 6 Bob wins, if it comes up 3,4,5 they play again. What is the expected number of dice rolls until the game finishes?*

Answer. Let X be the number of dice rolls until the game finishes, i.e. until either Alice or Bob wins.

Let B_1 be the event the first roll is a $1, 2$ or 6 (Alice or Bob wins)

Let B_2 be the event that the first roll is a $3, 4$ or 5.

Obviously B_1, B_2 form a partition, and $P(B_1) = 1/2$, $P(B_2) = 1/2$.

If B_1 occurs, then the game is finished after one roll, i.e., $E[X \mid B_1] = 1$.

If B_2 occurs then the game does not finish and it is just as if the first roll never happened: i.e., $E[X \mid B_2] = 1 + E[X]$.

We apply the Theorem of Total Probability for Expectation:

$$E(X) = E(X \mid B_1)\, P(B_1) + E(X \mid B_2)\, P(B_2) = 1 \cdot \frac{1}{2} + (1 + E(X))\frac{1}{2}.$$

Solving this equation fo $E(X)$, we see that

$$E(X) = 2.$$

□

2 Jointly distributed Continuous random variables

In many cases we are interested in more than one random variable at the same time. For example, in weather we might be interested in temperature and pressure, or we might be interested in the height and weight of people in a population.

When the two variables are "unrelated" (formally independent) then this is straightforward: we can deal with them individually!

However, in general

random variables are related and their joint behaviour is rather more complicated.

Let us consider the second example. If we take a random person we can view their height and weight as two random variables; let H be the height in metres and W the weight in kilograms. Note that we do **not** expect H and W to be independent: tall people tend to be heavier.

We could ask questions such as

1. $P(1.8 < H < 1.9 \text{ and } 90 < W < 100)$
2. $P(1.5 < H < 1.6 \text{ and } 90 < W < 100)$
3. $P(25H^2 < W)$

The first two of these seem reasonably natural questions the third may look unnatural but it is actually asking what is the probability that a person's BMI is over 25 (one definition of being overweight).

One way of illustrating these probabilities is by plotting each persons height and weight on a graph: the height on the X axis and the weight on the Y axis. Now the first

question is asking for the probability that a person (when plotted) lies in the box with sides given by $H = 1.8$, $H = 1.9$, $W = 90$ and $W = 100$. Similarly the second probability is asking about the box given by $H = 1.4$, $H = 1.5$, $W = 90$ and $W = 100$. Finally the third probability is asking for the probability that a person (when plotted) lies above the curve with equation $W = 25H^2$.

Exercise. *Draw a picture and think about the above.*

2.1 Joint Distributions

Definition. Suppose that X and Y are random variables. We say that X and Y are jointly continuous if there exists a function $f_{X,Y}(x, y)$ such that, for all sets $A \subset \mathbb{R}^2$,

$$P((X, Y) \in A) = \int \int_A f_{X,Y}(s, t)\, dt\, ds.$$

We call $f_{X,Y}$ the *joint probability density function*, or *joint pdf*.

The joint probability density function has properties which are similar to those of the probability density function of one random variable:

(i). $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(s, t)\, dt\, ds = 1.$

Proof. Obvious, since $P(-\infty < X < \infty \text{ and } -\infty < Y < \infty) = 1$. □

(ii). $f_{X,Y}(s, t) \geq 0$ for all s and t.

(Otherwise, we get negative probabilities)

Example. *Suppose that X and Y have joint probability density function given by*

$$f_{X,Y}(x, y) = \begin{cases} c & \text{if } 0 < x < 1,\ 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of c.

Solution. If $0 < x < 1$ then, as we have done before, we need to split the integration into ranges so we can deal with the "case"-definition of the joint probability density function.

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{0} f_{X,Y}(x, y)\, dxdy + ... + \int_{0}^{1}\int_{0}^{1} f_{X,Y}(x, y)\, dxdy + ... + \int_{-\infty}^{\infty}\int_{1}^{\infty} f_{X,Y}(x, y)\, dxdy \\ &= 0 + \int_{0}^{1}\int_{0}^{1} c\, dxdy + 0 \\ &= \int_{0}^{1} \Big[cx\Big]_{x=0}^{1} dy = \int_{0}^{1} c dy = \Big[cy\Big]_{y=0}^{1} = c. \end{aligned}$$

Therefore, we have $c = 1$.

Week 2

Example. *Suppose that X and Y have joint probability density function given by*

$$f_{X,Y}(s,t) = \begin{cases} e^{-s-t}, & \text{if } s,t \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Find:

(i). $P(3 < X < 4 \text{ and } 2 < Y < 5)$

(ii). $P(-1 < X < 3 \text{ and } 2 < Y < 5)$

(iii). $P(Y > X > 0)$.

Answer. (i). By definition

$$P(3 < X < 4 \text{ and } 2 < Y < 5) = \int_3^4 \int_2^5 f_{X,Y}(s,t)\, dt\, ds$$

Now, for all $3 < s < 4$ and $2 < t < 5$, we have $f_{X,Y}(s,t) = e^{-s-t}$ (i.e. we are in the first case of the definition of $f_{X,Y}$). Hence

$$\begin{aligned} P(3 < X < 4 \text{ and } 2 < Y < 5) &= \int_3^4 \int_2^5 f_{X,Y}(s,t)\, dt\, ds \\ &= \int_3^4 \left(\int_2^5 e^{-s-t}\, dt \right) ds \\ &= \int_3^4 \Big[-e^{-s-t} \Big]_{t=2}^5 ds \\ &= \int_3^4 (e^{-s-2} - e^{-s-5})\, ds \\ &= \Big[-e^{-s-2} + e^{-s-5} \Big]_{s=3}^4 \\ &= e^{-5} - e^{-6} - e^{-8} + e^{-9} = 0.004 = 0.4\%. \end{aligned}$$

(ii). For the second probability, again by definition

$$P(-1 < X < 3 \text{ and } 2 < Y < 5) = \int_{-1}^3 \int_2^5 f_{X,Y}(s,t)\, dt\, ds$$

Now, this range covers part of both cases in the definition of $f_{X,Y}$. Thus, we split the range into pieces each of which only contains parts from one of the cases of the definition.

$$\begin{aligned} P(-1 < X < 3 \text{ and } 2 < Y < 5) &= \int_{-1}^3 \int_2^5 f_{X,Y}(s,t)\, dt\, ds \\ &= \int_{-1}^0 \int_2^5 f_{X,Y}(s,t)\, dt\, ds + \int_0^3 \int_2^5 f_{X,Y}(s,t)\, dt\, ds \end{aligned}$$

Throughout the second integral, we have $s > 0$ so $f_{X,Y}(s,t) = e^{-s-t}$ (we are in the first case of the definition of $f_{X,Y}$).

Similarly, throughout the first integral, we have $s < 0$ so $f_{X,Y}(s,t) = 0$ (we are in the second case of the definition of $f_{X,Y}$).

Hence

$$\int_{-1}^{0}\int_{2}^{5} f_{X,Y}(s,t)\,dt\,ds + \int_{0}^{3}\int_{2}^{5} f_{X,Y}(s,t)\,dt\,ds = \int_{-1}^{0}\int_{2}^{5} 0\,dt\,ds + \int_{0}^{3}\int_{2}^{5} e^{-s-t}\,dt\,ds$$

and both of these are integrals we know how to calculate. The difficult step was to split in the appropriate "cases", then the rest of calculations are simple.

$$\begin{aligned}
\int_{0}^{3}\int_{2}^{5} f_{X,Y}(s,t)\,dt\,ds &= \int_{0}^{3}\left(\int_{2}^{5} e^{-s-t}\,dt\right)ds \\
&= \int_{0}^{3}\Big[-e^{-s-t}\Big]_{t=2}^{5} ds \\
&= \int_{0}^{3}(e^{-s-2} - e^{-s-5})\,ds \\
&= \Big[-e^{-s-2} + e^{-s-5}\Big]_{s=0}^{3} \\
&= e^{-2} - e^{-5} - e^{-5} + e^{-8} = e^{-2} - 2e^{-5} + e^{-8} = 0.122 = 12.2\%
\end{aligned}$$

(iii). Finally,

$$P(Y > X > 0) = \int_{0}^{\infty}\int_{0}^{t} f_{X,Y}(s,t)\,ds\,dt = \int_{0}^{\infty}\int_{s}^{\infty} f_{X,Y}(s,t)\,dt\,ds$$

Note we can do the integrals in either order but we do need to make sure we parameterise them correctly (see Calculus II notes).

Note that throughout the range of integration $t > s > 0$ so $f_{X,Y}(s,t) = e^{-s-t}$. Thus

$$\begin{aligned}
P(Y > X > 0) &= \int_{0}^{\infty}\int_{s}^{\infty} f_{X,Y}(s,t)\,dt\,ds \\
&= \int_{0}^{\infty}\int_{s}^{\infty} e^{-s-t}\,dt\,ds \\
&= \int_{0}^{\infty}\Big[-e^{-s-t}\Big]_{t=s}^{\infty} ds \\
&= \int_{0}^{\infty} e^{-2s}\,ds \\
&= \Big[-\frac{1}{2}e^{-2s}\Big]_{s=0}^{\infty} = \frac{1}{2} = 50\%.
\end{aligned}$$

□

Remark. Notice that, if we had done the first expression of the double integral, we'd have

gotten the same answer:

$$\begin{aligned} P(Y > X > 0) &= \int_0^\infty \int_0^t f_{X,Y}(s,t)\, ds\, dt \\ &= \int_0^\infty \int_0^t e^{-s-t}\, ds\, dt \\ &= \int_0^\infty \Big[-e^{-s-t} \Big]_{s=0}^t dt \\ &= \int_0^\infty (-e^{-2t} + e^{-t})\, dt \\ &= \Big[\frac{1}{2} e^{-2t} - e^{-t} \Big]_{t=0}^\infty = \frac{1}{2} = 50\%. \end{aligned}$$

□

2.2 Joint cumulative distribution function

We can also define a *joint cumulative distribution function (or joint cdf)* of two random variable (X, Y) analogously with the cumulative distribution function of one random variable (defined before).

Definition. Suppose X and Y are random variables. Then the *joint distribution function* $F_{X,Y}$ is defined by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

If X and Y are jointly continuous with joint probability density function $f_{X,Y}$ then

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(s,t)\, dt\, ds$$

Remark. It is easy to see that

$$F_{X,Y}(\infty, \infty) = P(X < \infty \text{ and } Y < \infty) = 1$$

and

$$F_{X,Y}(-\infty, -\infty) = P(X < -\infty \text{ and } Y < -\infty) = 0$$

Also, the function $F_{X,Y}$ is increasing in the sense that whenever $x_1 \leq x_2$ and $y_1 \leq y_2$ then

$$F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2).$$

Just as in the one random variable pdf/cdf case, we can recover the joint density function from the joint distribution function.

Theorem 5. *Suppose X and Y are jointly continuous random variables with joint cumulative distribution function $F_{X,Y}$. Then the joint probability density function $f_{X,Y}$ is given by*

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

Proof. By definition

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)dtds.$$

Hence

$$\begin{aligned}\frac{\partial^2}{\partial x\partial y}F_{X,Y}(x,y) &= \frac{\partial^2}{\partial x\partial y}\int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)dtds \\ &= \frac{d}{dy}\int_{-\infty}^{y} f_{X,Y}(x,t)dt && \text{(by the fundamental theorem of calculus)} \\ &= f_{X,Y}(x,y) && \text{(by the fundamental theorem of calculus)}\end{aligned}$$

as claimed. □

2.3 Marginal distributions

In this section, we aim to find $f_X(x)$ and $f_Y(y)$ given that we know $f_{X,Y}$. The following property of the joint cdf allows one to find $F_X(x)$ and $F_Y(y)$ if $F_{X,Y}$ is known.

Lemma 6. *Let (X,Y) be two random variables with joint cumulative distribution function $F_{X,Y}$, and F_X, F_Y be the cumulative distribution functions of X and Y respectively. Then*

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

Proof. Note that the events $\{X \le x\} = \{X \le x, Y \le \infty\}$. Indeed, the left hand side of this equality is the event that "X is no larger than x" and the right hand side is the event "X is no larger than x and Y is arbitrary" which is the same thing. Hence

$$F_X(x) = P(X \le x) = P(X \le x, Y < \infty) = F_{X,Y}(x,\infty).$$

The second formula is proved similarly. □

Exercise. *Prove the second formula in the above lemma.*

Theorem 7. *Suppose X and Y have joint probability density function $f_{X,Y}$. Then X and Y are continuous random variables. The density functions f_X of X is given by*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy.$$

and similarly

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

The functions f_X and f_Y are called the **marginal probability density functions** *or* marginal pdfs.

Proof. We present two different, but equivalent, proofs to this theorem. The first one was covered in the lectures, the second one we only present here. Of course, you can use any one you prefer.

Proof (I). By lemma 6

$$F_X(x) = F_{X,Y}(x,\infty)$$

and by the definition of the joint probability density function

$$F_X(x) = F_{X,Y}(x,\infty) = P(X \le x, Y < \infty) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dt\,ds = \int_{-\infty}^{x}\left(\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dt\right)ds$$

Hence

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{d}{dx}\int_{-\infty}^{x}\left(\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dt\right)ds = \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt,$$

where the last equality follows by the fundamental theorem of calculus.

Proof (II). Recall that a function $f(x)$ is a probability density function of a random variable X iff for any (a,b)

$$P(X \in (a,b)) = \int_a^b f(x)dx.$$

Note that the events $\{X \in (a,b)\}$ and $\{X \in (a,b),\ Y \in (-\infty,\infty)\}$ are identical and hence, by the definition of $f_{X,Y}$, we have

$$\begin{aligned}\int_a^b f(x)dx &= P(X \in (a,b)) \\ &= P(X \in (a,b),\ Y \in (-\infty,\infty)) = \int_a^b\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right)dx,\end{aligned}$$

which implies $f(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$. Hence the proof of the theorem. □

Exercise. *Prove the second formula in the above Theorem.*

Example. *Suppose that X and Y have joint probability density function given by*

$$f_{X,Y}(x,y) = \begin{cases} c & \text{if } 0 < x < 1,\ 0 < y < 1, \\ 0 & \text{otherwise.}\end{cases}$$

Find the marginal probability density function f_X of X.

Solution. The marginal density f_X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt.$$

Now if $x \le 0$ or $x \ge 1$ then, regardless of t, $f_{X,Y}(x,t) = 0$ and so, in this case, the integral is zero. I.e., if $x \le 0$ or $x \ge 1$ then $f_X(x) = 0$.

If $0 < x < 1$ then, as we have done before, we need to split the integration into ranges so we can deal with the "case"-definition of the joint probability density function.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt \\ &= \int_{-\infty}^{0} f_{X,Y}(x,t)\,dt + \int_0^1 f_{X,Y}(x,t)\,dt + \int_1^{\infty} f_{X,Y}(x,t)\,dt \\ &= 0 + \int_0^1 c\,dt + 0 = \Big[ct\Big]_{t=0}^{1} = c.\end{aligned}$$

Thus the marginal density f_X is given by

$$f_X(x) = \begin{cases} c & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

To find c, we use $1 = \int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^1 c dx = c$. Therefore,

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(We recognize this as the probability density function of a Uniform distribution but that is not part of the question.)

Example. *Suppose that X and Y have joint probability density function given by*

$$f_{X,Y}(x,y) = \begin{cases} e^{-x-y} & \textit{if } x, y \geq 0 \\ 0 & \textit{otherwise} \end{cases}$$

Find the marginal density f_X.

Solution. The marginal density f_X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt.$$

Now if $x < 0$ then, regardless of t, $f_{X,Y}(x,t) = 0$ and so, in this case, the integral is zero. I.e., if $x < 0$ then $f_X(x) = 0$.

If $x > 0$ then, as we have done before, we need to split the integration into ranges so we can deal with the "case"-definition of the joint probability density function.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt \\ &= \int_{-\infty}^{0} f_{X,Y}(x,t)\,dt + \int_0^{\infty} f_{X,Y}(x,t)\,dt \\ &= 0 + \int_0^{\infty} e^{-x-t}\,dt \\ &= \Big[-e^{-x-t} \Big]_{t=0}^{\infty} = e^{-x}. \end{aligned}$$

Thus the marginal density f_X is given by

$$f_X(x) = \begin{cases} e^{-x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(We recognize this as the probability density function of the Exp(1) random variable but that is not part of the question.)

Example. *Suppose that X and Y have joint probability density function given by*

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x-y} & \text{if } x \geq y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the marginal density f_Y.

Solution. The marginal density f_Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(s,y)\, ds.$$

Now if $y < 0$ then, regardless of s, $f_{X,Y}(s,y) = 0$ and so, in this case, the integral is zero. I.e., if $y < 0$ then $f_Y(y) = 0$.

If $y > 0$ then, as before, we need to split the integration into ranges so we can deal with the "case"-definition of the joint probability density function.

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(s,y)\, ds \\ &= \int_{-\infty}^{y} f_{X,Y}(s,y)\, ds + \int_{y}^{\infty} f_{X,Y}(s,y)\, ds && \text{(true for any function } f_{X,Y}) \\ &= \int_{-\infty}^{y} 0\, ds + \int_{y}^{\infty} 2e^{-s-y}\, ds && \text{(see below for reason)} \\ &= 0 + \Big[-2e^{-s-y} \Big]_{s=y}^{\infty} = 2e^{-2y}. \end{aligned}$$

The third line follows because:

- in the second integral s ranges from y to ∞: thus $s > y > 0$ and so $f_{X,Y}(s,y) = 2e^{-s-y}$.
- in the first integral s ranges from $-\infty$ to y: thus $s < y$ and so $f_{X,Y}(s,y) = 0$ (note it does not matter if $s < 0$ or $s > 0$: whenever $s < y$ we are in the second case of the definition of $f_{X,Y}$).

Thus the marginal density f_X is given by

$$f_Y(y) = \begin{cases} 2e^{-2y} & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(We recognize this as the probability density function of an Exp(2) random variable but that is not part of the question.)

Exercise. *Find the marginal density f_X in this second example.*

Statistical remarks

(1). Inferring the marginal density of one random variable X from a joint probability density function of a couple (X,Y) is a fundamental tool used in Statistics.

(2). In this course, we will see this when testing a hypothesis. It is natural to need to combine *two (or more)* random variables to obtain another one for the purpose of the test. However when transforming the couple (U,V) of random variables, we obtain another couple (X,Y) of random variables, hence to isolate the random variable X (which is our aim), we need to obtain its marginal probability density function.

2.4 Expectation over a joint distribution, Covariance & Correlation

Let (X, Y) be two random variables with joint probability density function $f_{X,Y}$ and let $g(x, y)$ be a function of two variables taking real values. We can then consider a new random variable $g(X, Y)$. We aim to find the expectation $E\left(g(X, Y)\right)$.

In the sequel, the following important formula shall be used (no proof will be given). If a function $g(x, y)$ is such that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|g(x, y)|f_{X,Y}(x, y)dxdy < \infty$ then

$$E\left(g(X, Y)\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy. \tag{6}$$

Remark. This formula is an analogue of the equivalent one-dimensional case.

Remark. You are reminded that if $G(x, y)$ is a "good" function then its double integral can be computed as follows:

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x, y)dxdy &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} G(x, y)dx\right)dy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} G(x, y)dy\right)dx.\end{aligned}$$

Important examples of expectations over joint distributions are the so-called *raw moment*

$$E(X^k Y^m)$$

and the so-called *central moment*

$$E[(X - \mu_1)^k (Y - \mu_2)^m)],$$

where $\mu_1 = E(X)$ and $\mu_2 = E(Y)$.

Formula (6) can be used for finding such moments, e.g.

$$E[(X - \mu_1)^k (Y - \mu_2)^m] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \mu_1)^k (y - \mu_2)^m f_{X,Y}(x, y)dxdy.$$

Definition. Covariance of two random variables X and Y is

$$\mathrm{Cov}(X, Y) = E\left[(X - E(X))(Y - E(Y))\right].$$

Definition. Correlation (or the coefficient of correlation) of two random variables X and Y is

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov(X, Y)}}{\sqrt{\mathrm{Var(X)Var(Y)}}}.$$

A standard notation for correlation is $\rho(X, Y) \equiv \mathrm{Corr}(X, Y)$.

Remark. The covariance and even more so the correlation of two random variables is commonly used as measures of their inter-dependence.

- If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.
- Hence Independence implies that Covariance (and so Correlation) is zero.
- However it is not true that zero Correlation implies Independence (see next section).

Week 3

Lemma 8. *For any two random variables X and Y we have*

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

Proof 1. Denote $E(X) = \mu_1$ and $E(Y) = \mu_2$. Then

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E\left[(X - E(X))(Y - E(Y))\right] \\
&= E[(X-\mu_1)(Y-\mu_2)] \\
&= E[XY - \mu_1 Y - \mu_2 X + \mu_1\mu_2] \\
&= E(XY) - E(\mu_1 Y) - E(\mu_2 X) + \mu_1\mu_2 \\
&= E(XY) - \mu_1 E(Y) - \mu_2 E(X) + \mu_1\mu_2 \\
&= E(XY) - \mu_1\mu_2 - \mu_2\mu_1 + \mu_1\mu_2 = E[XY] - \mu_1\mu_2.
\end{aligned}$$

□

Proof 2. We prove this for any X and Y continuous random variables (*Exercise.* Prove this for discrete random variables). Given that $E(X) = \mu_1$ and $E(Y) = \mu_2$ are constants, we have:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E\left[(X - E(X))(Y - E(Y))\right] \\
&= E[(X-\mu_1)(Y-\mu_2)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_1)(y-\mu_2)f_{X,Y}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_{X,Y}(x,y)dxdy - \mu_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dxdy \\
&\quad - \mu_1\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dxdy + \mu_1\mu_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f_{X,Y}(x,y)dxdy
\end{aligned}$$

Then by switching the order of integration as follows, we get

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E(XY) - \mu_2\int_{-\infty}^{\infty}x\left(\int_{-\infty}^{\infty}f_{X,Y}(x,y)dy\right)dx \\
&\quad - \mu_1\int_{-\infty}^{\infty}y\left(\int_{-\infty}^{\infty}f_{X,Y}(x,y)dx\right)dy + \mu_1\mu_2 \\
&= E(XY) - E(Y)\int_{-\infty}^{\infty}xf_X(x)dx - E(X)\int_{-\infty}^{\infty}yf_Y(y)dy + E(X)E(Y) \\
&= E(XY) - E(X)E(Y)
\end{aligned}$$

□

Exercise. (Check first-year probability & statistics notes) *Suppose that X, Y, Z are random variables. If $a \in \mathbb{R}$, prove that*

(i).

$$\mathrm{Cov(aX, Y)} = \mathrm{Cov(X, aY)} = \mathrm{a\,Cov(X, Y)}$$

(ii).

$$\mathrm{Cov(a + X, Y)} = \mathrm{Cov(X, a + Y)} = \mathrm{Cov(X, Y)}$$

(iii).

$$\mathrm{Cov(X, Y + Z)} = \mathrm{Cov(X, Y)} + \mathrm{Cov(X, Z)}$$

2.5 Conditional distribution of Continuous random variables

We would like to do something like we did in the discrete case and define the conditional distribution. There, we defined the random variable $X|\{Y = y\}$, but this is not possible for continuous random variables, since $P(Y = y) = 0$ for any y. However, we can get round this by conditioning on Y being "very close" to y, instead of Y being equal to y. If we do this sensibly we arrive at the following definition.

Definition. Suppose X and Y are jointly continuous random variables with joint probability density function $f_{X,Y}$. Then the conditional probability density function of X given $Y = y$ is

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

whenever $f_Y(y) \neq 0$. (As usual f_Y is the marginal probability density function of Y.)

Remark. Similarly, $f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ whenever $f_X(x) \neq 0$.

Example. *Suppose that X and Y have joint probability density function given by*

$$f_{X,Y}(x, y) = \begin{cases} e^{-x-y} & \text{if } x, y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the conditional probability density function $f_{X|Y=y}$.

Answer. Firstly, we obtain the marginal density f_Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx.$$

Now if $y < 0$, then $f_{X,Y}(x, y) = 0$ and so, in this case, the integral is zero. So, if $y < 0$ then $f_Y(y) = 0$.

If $y > 0$ then, as we have done before, we need to split the integration into ranges so we can deal with the "case"-definition of the joint probability density function.

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx \\ &= \int_{-\infty}^{0} f_{X,Y}(x, y)\, dx + \int_{0}^{\infty} f_{X,Y}(x, y)\, dx \\ &= 0 + \int_{0}^{\infty} e^{-x-y}\, dx \\ &= \Big[-e^{-x-y} \Big]_{x=0}^{\infty} = e^{-y}. \end{aligned}$$

By definition, if $y > 0$, we have

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

and is not defined for $y \leq 0$ (as $f_Y(y)$ would be zero).

If $x < 0$ then $f_{X,Y}(x, y) = 0$ so $f_{X|Y=y}(x) = 0$.

Finally, if $x > 0$ then

$$f_{X|Y=y}(x) = \frac{e^{-x-y}}{f_Y(y)} = \frac{e^{-x-y}}{e^{-y}} = e^{-x}$$

Hence, for any given $y > 0$, we have

$$f_{X|Y=y}(x) = \begin{cases} e^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Example. *Suppose X and Y have joint probability density function*

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x-y} & \text{if } x > y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the conditional probability density function $f_{X|Y=y}$.

Solution. Let us first compute $f_Y(y)$. As usual,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$

and we see that $f_Y(y) = 0$ if $y \leq 0$. If $y > 0$ then

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_{y}^{\infty} 2e^{-x-y}\,dx = -2e^{-x-y}\Big|_{x=y}^{x=\infty} = 2e^{-2y}.$$

By definition, if $y > 0$, we have

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

and is not defined for $y \leq 0$ (as $f_Y(y)$ would be zero).

If $x < y$ then $f_{X,Y}(x,y) = 0$ so $f_{X|Y=y}(x) = 0$.

Finally, if $x > y > 0$ then

$$f_{X|Y=y}(x) = \frac{2e^{-x-y}}{f_Y(y)} = \frac{2e^{-x-y}}{2e^{-2y}} = e^{-x+y}$$

Hence, for any given $y > 0$, we have

$$f_{X|Y=y}(x) = \begin{cases} e^{-x+y} & \text{if } x > y \\ 0 & \text{otherwise.} \end{cases}$$

□

Recall that if X is a random variable with probability density function f_X then

$$E(X) = \int_{-\infty}^{\infty} s f_X(s)\,ds.$$

More generally, if g is any function then

$$E(g(X)) = \int_{-\infty}^{\infty} g(s) f_X(s)\,ds.$$

Note, in both cases, we can ignore ranges of s where the density is zero. So, for example, if f_X is only non-zero between 0 and 1 then we can replace the range of integration above by the range 0 to 1.

Having just defined the probability density function of $X|\{Y = y\}$, whenever $f_Y(y) \neq 0$, we can use the following formula to find the conditional expectation of X given that $Y = y$, for all y such that $f_Y(y) \neq 0$:

$$E(X|Y = y) = \int_{-\infty}^{\infty} s\, f_{X|Y=y}(s)\, ds = \int_{-\infty}^{\infty} s\, \frac{f_{X,Y}(s, y)}{f_Y(y)}\, ds.$$

Example (previous example continued...)**.** *Suppose X and Y have joint probability density function*

$$f_{X,Y}(x, y) = \begin{cases} 2e^{-x-y} & \text{if } x > y > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of X given that Y takes the value 5?

Answer. We use the conditional probability density function of X given $Y = 5$, namely we substitute $y = 5$. Hence, we have

$$f_{X|Y=5}(x) = \begin{cases} e^{5-x} & \text{if } x > 5 \\ 0 & \text{otherwise.} \end{cases}$$

This implies that

$$E(X|Y = 5) = \int_{-\infty}^{\infty} x\, f_{X|Y=5}(x)\, dx = \int_{5}^{\infty} x\, e^{5-x}\, dx$$

Using integration by parts we get

$$E(X|Y = 5) = \left[-x\, e^{5-x}\right]_{x=5}^{\infty} + \int_{5}^{\infty} e^{5-x}\, dx = 0 + 5 - \left[e^{5-x}\right]_{x=5}^{\infty} = 5 - 0 + 1 = 6.$$

□

3 Independence

3.1 Independence of events (Revision)

Definition. Two events A, $B \in \mathcal{F}$ are *independent* if $P(A \cap B) = P(A)P(B)$. More generally, events $A_1, A_2, ..., A_n$ are *mutually independent* if for any $1 \leq i_1 < i_2 < ...i_m \leq n$

$$P(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_m}).$$

Exercise.
(a). *Prove that* $A \cap B^c = A \setminus (A \cap B)$.
(b). *Prove that if* A *and* B *are independent events then* A *and* B^c *are independent.*

3.2 Independence of random variables

We start by defining what we mean by the independence of two random variables.

Definition. Two random variables X and Y are independent if, for any intervals (a, b) and (c, d), the events

$$\{X \in (a, b)\} \qquad \text{and} \qquad \{Y \in (c, d)\}$$

are independent: i.e.,

$$P(X \in (a, b) \text{ and } Y \in (c, d)) = P(X \in (a, b))P(Y \in (c, d)) \tag{7}$$

Remark. Another way to define independence is: X and Y are independent if, for any measurable sets A and B, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent. This may look like a stronger requirement but in fact these definitions are equivalent. We shall not prove this fact.

Theorem 9. *Suppose that* X *and* Y *have joint density function* $f_{X,Y}$ *and marginal density functions* f_X *and* f_Y. *Then* X *and* Y *are independent if and only if*

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

for all x *and* y.

Proof. Part I. Suppose $f_{X,Y}(s, t) = f_X(s)f_Y(t)$ for all s, t. Then for any intervals (a, b), $(c, d) \subset \mathbb{R}$ we have

$$\begin{aligned} P(X \in (a, b) \text{ and } Y \in (c, d)) &= \int_a^b \int_c^d f_{X,Y}(s, t)\, ds\, dt \\ &= \int_a^b \int_c^d f_X(s) f_Y(t)\, ds\, dt \\ &= \int_a^b f_X(s)\, ds \times \int_c^d f_Y(t)\, dt \\ &= P(X \in (a, b))P(Y \in (c, d)) \end{aligned}$$

Since this is true for all intervals (a, b) and (c, d) we see that X and Y are independent.

Part II. We now have to show that (7) implies $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. So suppose that for any (a,b) and (c,d) equation (7) holds. Take $(a,b) = (-\infty, x)$ and $(c,d) = (-\infty, b)$. Then

$$P(X \in (-\infty, x) \text{ and } Y \in (-\infty, y)) = P(X \in (-\infty, x)) \times P(Y \in (-\infty, y))$$

which is the same as

$$P(X < x,\ Y < y) = P(X < x)P(Y < y). \tag{I}$$

The probabilities in the last formula are distribution functions and so we can re-write the last line as

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

But then

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} \left(F_X(x)F_Y(y)\right) \\ &= \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}\left(F_X(x)F_Y(y)\right)\right) = \frac{\partial}{\partial x}\left(F_X(x)\frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}\right) \\ &= \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}\frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = f_X(x)f_Y(y). \end{aligned}$$

□

Remark. Strictly speaking, (I) and (II) above are not the same thing. The reason for that is that in general $F_X(x) = P(X \le x) \neq P(X < x)$ (and similarly for $F_{X,Y}(x,y)$, $F_Y(y)$). However, in this theorem we deal with continuous random variables and therefore $F_X(x) = P(X < x)$, etc.

Using Theorem 9, we also have the following result.

Corollary 10. *If X and Y are independent then their joint probability density function $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ so*

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

In other words the distribution of X does not depend on the value of Y.

Example. *Suppose that X and Y have joint density function $f_{X,Y}$ given by*

$$f_{X,Y}(x,y) = \begin{cases} 6e^{-2x-3y} & \text{if } x > 0 \text{ and } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Are the random variables X and Y independent?

Answer. Find the marginal densities f_X and f_Y.

We know that the marginal density f_X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt$$

If $x \leq 0$ then, $f_{X,Y}(x,y) = 0$ (regardless of y; we are in the second case of the function definition) so the integral, and thus the marginal density $f_X(x) = 0$.

Now suppose $x > 0$. As usual we split this into pieces

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\infty}^{0} f_{X,Y}(x,y)\,dy + \int_{0}^{\infty} f_{X,Y}(x,y)\,dy$$

In the first integral y ranges from $-\infty$ to 0 and so $f_{X,Y}(x,y) = 0$ (we are in the second case of the function definition).

In the second integral y ranges from 0 to ∞. Thus $f_{X,Y}(x,y) = 6e^{-2x-3y}$ (we are in the first case of the function definition).

Thus,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{0} f_{X,Y}(x,y)\,dy + \int_{0}^{\infty} f_{X,Y}(x,y)\,dy \\ &= \int_{-\infty}^{0} 0\,dy + \int_{0}^{\infty} 6e^{-2x-3y}\,dy \\ &= 0 - \left[2e^{-2x-3y}\right]_{y=0}^{\infty} \\ &= 2e^{-2x} \end{aligned}$$

Thus

$$f_X(x) = \begin{cases} 2e^{-2x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

We also know that the marginal density f_Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(s,y)\,ds$$

If $y < 0$ then, regardless of s $f_{X,Y}(x,y) = 0$ (we are in the second case of the function definition) so the integral, and thus the marginal density $f_Y(y) = 0$.

Now suppose that $y > 0$. As usual we split this into pieces:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_{-\infty}^{0} f_{X,Y}(x,y)\,dx + \int_{0}^{\infty} f_{X,Y}(x,y)\,dx$$

In the first integral x ranges from $-\infty$ to 0 and so $f_{X,Y}(x,y) = 0$ (we are in the second case of the function definition).

In the second integral x ranges from 0 to ∞ and therefore $f_{X,Y}(x,y) = 6e^{-2x-3y}$ (we are in the first case of the function definition).

Putting this all together we see that

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_{-\infty}^{0} f_{X,Y}(x,y)\,dx + \int_{0}^{\infty} f_{X,Y}(x,y)\,dx \\ &= 0 + \int_{0}^{\infty} 6e^{-2x-3y}\,dx = \left[-\frac{6}{2}e^{-2x-3y} \right]_{x=0}^{\infty} = 3e^{-3y} \end{aligned}$$

Thus

$$f_Y(y) = \begin{cases} 3e^{-3y} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Therefore, we see that $f_X(x) \cdot f_Y(y) = 0 = f_{X,Y}(x,y)$ if either $x \leq 0$ or $y \leq 0$.

The only combination of values that leads to a non-zero value is when both $x > 0$ and $y > 0$, for which we get $f_X(x) \cdot f_Y(y) = 6e^{-2x-3y} = f_{X,Y}(x,y)$.

Hence, overall, we conclude that X and Y are independent.

Week 4

Sometimes it may be difficult to use the above theorem because we need to find the marginal densities and we may not be able to do the necessary integration. The following theorem lets us prove independence without calculating the marginal densities.

Theorem 11. *Suppose that X and Y have joint density function $f_{X,Y}$. Then X and Y are independent if and only if there are functions g and h such that*

$$f_{X,Y}(x,y) = g(x)h(y) \text{ for all } x \text{ and } y.$$

Proof. Part I. Suppose that $f_{X,Y}(x,y) = g(x)h(y)$. Set $H = \int_{-\infty}^{\infty} h(t)\,dt$ and $G = \int_{-\infty}^{\infty} g(s)\,ds$. We shall first prove that $GH = 1$. This will be used below. We have

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(s,t)\,dt\,ds = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(s)h(t)\,dt\,ds \\ &= \int_{-\infty}^{\infty} g(s)\,ds \int_{-\infty}^{\infty} h(t)\,dt = GH \end{aligned}$$

Next, let us compute $f_X(x)$ and $f_Y(y)$.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt = \int_{-\infty}^{\infty} g(x)h(t)\,dt = g(x)\int_{-\infty}^{\infty} h(t)\,dt = Hg(x).$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(s,y)\,ds = \int_{-\infty}^{\infty} g(s)h(y)\,ds = h(y)\int_{-\infty}^{\infty} g(s)\,ds = Gh(y).$$

It remains to note that

$$f_X(x) \times f_Y(y) = Hg(x) \times Gh(y) = GHg(x)h(y) = g(x)h(y) = f_{X,Y}(x,y).$$

By Theorem of independence using probability density functions, we can now state that X and Y are independent.

Part II. The other direction is trivial. If X and Y are independent then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ and we can put $g(x) = f_X(x)$ and $h(y) = f_Y(y)$. □

Exercise. *Try using Theorem 11 instead to answer the last example.*

Example. *Suppose X and Y have joint density function*

$$f_{X,Y}(x,y) = \begin{cases} 2x & \text{if } 0 < x < 1 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

Solution. Let

$$g(x) = \begin{cases} 2x & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$h(y) = \begin{cases} 1 & \text{if } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}.$$

If $x \notin (0,1)$ or $y \notin (0,1)$ then $g(x)h(y) = 0$ which equals $f_{X,Y}(x,y)$ in this case.
If $x \in (0,1)$ and $y \in (0,1)$ then

$$g(x)h(y) = 2x = f_{X,Y}(x,y).$$

In all cases, $g(x)h(y) = f_{X,Y}(x,y)$, so X and Y are independent.

Remark. One nice thing is that we don't need to justify how we get g and h: all we need to do is find them.

Remark. Also note that it was important that $g(x)h(y) = f_{X,Y}(x,y)$ **for all** x **and** y.

Example. *Suppose X and Y have joint density function*

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x-y} & \textit{if } y > x > 0 \\ 0 & \textit{otherwise.} \end{cases}$$

Are X and Y independent?

Remark. It may look like the above factorizes: it does not because of the mixed "case"–conditions *(if $y > x > 0$)* in the definition.

There are various ways to solve this problem.

Solution (Solution 1)**.** Make use of the Theorem of independence of random variables using probability density functions, and look at the marginal densities.

Remark. This solution is far from being the simplest one or the shortest one. However, computing the marginal densities is in itself a useful exercise and this is one of the reasons we discuss it hear.

See below two more solutions to this example and read carefully the remark to the next example.

Note that when calculating the marginal densities, because of the multiple cases in the definition of the density $f_{X,Y}$, we have to split the integration into several parts.

For $x > 0$, we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt \\ &= \int_{-\infty}^{x} f_{X,Y}(x,t)\,dt + \int_{x}^{\infty} f_{X,Y}(x,t)\,dt \\ &= \int_{-\infty}^{x} 0\,dt + \int_{x}^{\infty} 2e^{-x-t}\,dt \\ &= 0 + \left[-2e^{-x-t}\right]_{t=x}^{\infty} = 2e^{-2x}. \end{aligned}$$

Similarly, for $y > 0$, we have

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(s,y)\,ds \\ &= \int_{-\infty}^{0} f_{X,Y}(s,y)\,ds + \int_{0}^{y} f_{X,Y}(s,y)\,ds + \int_{y}^{\infty} f_{X,Y}(s,y)\,ds \\ &= \int_{-\infty}^{0} 0\,ds + \int_{0}^{y} 2e^{-s-y}\,ds + \int_{y}^{\infty} 0\,ds \\ &= 0 + \left[-2e^{-s-y}\right]_{s=0}^{y} + 0 = 2e^{-y} - 2e^{-2y} = 2e^{-y}(1-e^{-y}). \end{aligned}$$

Thus for $y > x > 0$ we see that $f_X(x)f_Y(y) = 2e^{-2x} \times 2e^{-y}(1-e^{-y}) \neq f_{X,Y}(x,y)$ so X and Y are not independent. □

Solution (Solution 2)**.** Consider any point (x,y), with $x > 0$ and $y > 0$. Note that then $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt > 0$ since $f_{X,Y}(x,t) > 0$ when $t > x$. Similarly, $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(s,y)\,ds > 0$ since $f_{X,Y}(s,y) > 0$ when $s \in (0,y)$. So, the product $f_X(x)f_Y(y) > 0$ for any such (x,y), and thus $0 = f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$ if $0 < y < x$. Hence X and Y are not independent.

Remark. In fact, Solution 2 proves the following general statement: if the domain where $f_{X,Y}(x,y) > 0$ is not a direct product of two subset of $\mathbb{R}$, then X and Y are not independent.Can you see that?

Solution (Solution 3)**.** Since both X and Y can take any positive values but $Y > X$, they cannot be independent.

More precisely, $P(X \in [a,b]) > 0$ for any $b > a > 0$ and $P(Y \in [c,d]) > 0$ for any $d > c > 0$. However, $P(X \in [a,b] \text{ and } Y \in [c,d]) = 0$ if $d < a$. Hence, Y and X are not independent.

Example. *Suppose X and Y have joint density function*

$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{if } 0 < y < 1,\ 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

Remark. In this example, the approach suggested in Solution 1 leads to the correct answer. Can you see that the approach suggested in solutions 2 and 3 doesn't work without additional computations?

Exercise. *For the last example, invent a solution which doesn't require the knowledge of f_X and f_Y.*

Hint. Note that if X and Y are independent random variables, then the ratio $\varphi(y) = f_{X,Y}(x_1,y)/f_{X,Y}(x_2,y)$ doesn't depend on y. Check whether that is the case in this example.

3.3 Expectation of the product of independent random variables

Exercise. *Suppose that X and Y are two random variables.*

(i). *Prove that if X and Y are independent then*

$$E(g(X)h(Y)) = E(g(X))\,E(h(y)),$$

where g, h are any functions.

(ii). *Deduce that if X and Y are independent then*

$$E(X^k Y^m) = E(X^k)\,E(Y^m),$$

(iii). *Deduce that if X and Y are independent random variables then $Corr(X, Y) = 0$.*

3.4 Distribution of a sum of independent random variables

If two random variables are independent, we can obtain the distribution of their sum as follows.

Definition. Suppose $f(x)$ and $g(y)$ are two continuous functions. Then, the convolution $f * g$ is the function given by

$$(f * g)(z) = \int_{-\infty}^{+\infty} f(z-y)g(y)dy = \int_{-\infty}^{+\infty} f(x)g(z-x)dx$$

The following theorem states that if X and Y are independent, then the density of their sum is the convolution of their densities.

Theorem 12. *Suppose X and Y are two independent random variables with probability density functions $f_X(x)$ and $f_Y(y)$, respectively. Then, their sum $Z = X+Y$ is a random variable with probability density function given by*

$$f_Z(z) = (f_X * f_Y)(z)$$

Proof. Left as an exercise. Not examinable. □

In order to appreciate the above result, we present the following example.

Example. *Suppose we choose independently X and Y to be two Exponential(λ) random variables. Use their convolution to find the probability density function of their sum $Z = X + Y$.*

Answer. We have

$$f_X(x) = f_Y(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{otherwise .} \end{cases}$$

The probability density function of their sum is given by

$$\begin{aligned} f_Z(z) = (f_X * f_Y)(z) &= \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)dy \\ &= \int_0^{\infty} \lambda e^{-\lambda y} f_X(z-y)dy \end{aligned}$$

Note that $0 < z - y$ if and only if $y < z$, hence:

For $z < 0$, we have $f_Z(z) = 0$.

For $z > 0$, we have

$$f_Z(z) = \int_0^z \lambda^2 e^{-\lambda y} e^{-\lambda(z-y)} dy = \int_0^z \lambda^2 e^{-\lambda z} dy = \lambda^2 z e^{-\lambda z}$$

Overall, we have

$$f_Z(z) = \begin{cases} \lambda^2 z e^{-\lambda z}, & \text{if } z > 0, \\ 0 & \text{otherwise .} \end{cases}$$

4 Distributions of functions of random variables

There are three main methods to find the distribution of a function of one or more random variables. These are to use the CDF, to transform the probability density function directly or to use moment generating functions. We shall study these in turn and along the way find some results which are used in statistics.

4.1 Method of cumulative distribution functions

We first give an example before discussing the general method.

Example. *Suppose the random variable Y has a probability density function*

$$f_Y(y) = \begin{cases} 3y^2, & \textit{if } 0 < y < 1 \\ 0, & \textit{otherwise.} \end{cases}$$

What is the probability density function of $U = 2Y + 3$?

Answer. The range of U is $3 < U < 5$ and

$$F_U(u) = P(U \le u) = P(2Y + 3 \le u) = P\left(Y \le \frac{u-3}{2}\right)$$
$$= \int_0^{\frac{u-3}{2}} f_Y(y)\, dy = \int_0^{\frac{u-3}{2}} 3y^2\, dy = \left(\frac{u-3}{2}\right)^3$$

so

$$F_U(u) = \begin{cases} 0 & u < 3 \\ \left(\frac{u-3}{2}\right)^3 & 3 \le u \le 5 \\ 1 & u > 5 \end{cases}$$

and

$$f_U(u) = \frac{dF(u)}{du} = \begin{cases} \frac{3}{8}(u-3)^2 & 3 \le u \le 5 \\ 0 & \text{otherwise} \end{cases}$$

□

The general method

Suppose that $X_1, X_2, \ldots, X_n$ are random variables with pdf $f_{X_1,\ldots,X_n}(x_1, ..., x_n)$ and $Y = g(X_1, ..., X_n)$, where $g : \mathbb{R}^n \mapsto \mathbb{R}$ is a function of n random variables. How can we compute the pdf $f_Y(y)$?

Answer

1. Find the region in $\mathbb{R}^n$ where $g(x_1, ..., x_n) \leq y$. Let it be $\mathcal{D}$.
2. Compute $F_Y(y) = P(Y \leq y) = \int ... \int_{\mathcal{D}} f_{X_1,\ldots,X_n}(x_1, ..., x_n)dx_1...dx_n$. (We integrate the probability density function over the region $\mathcal{D}$.)
3. Find the density function $f_Y(y)$ by differentiating $F_Y(y)$: $f_Y(y) = \frac{dF_Y(y)}{dy}$.

Exercise (This is in fact a theorem – however, it can be viewed as an application of the method of CDFs)**.** *If the random variable $Z \sim N(0, 1)$ then $Z^2 \sim \chi^2(1)$.*

Answer. One way to prove this is by using the method of cumulative distribution functions. For $Z \sim N(0, 1)$, we have

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad -\infty < z < \infty.$$

Then if $U = Z^2$, we have

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(Z^2 \leq u) \\ &= P(-\sqrt{u} \leq Z \leq \sqrt{u}) \\ &= \int_{-\sqrt{u}}^{\sqrt{u}} f_Z(z)dz \\ &= \Phi(\sqrt{u}) - \Phi(-\sqrt{u}). \end{aligned}$$

So if we differentiate both sides with respect to u we find

$$\begin{aligned} f_U(u) &= f_Z(\sqrt{u})\left(\frac{1}{2\sqrt{u}}\right) + f_Z(-\sqrt{u})\left(\frac{1}{2\sqrt{u}}\right) \\ &= \frac{1}{2\sqrt{u}}\left[f_Z(\sqrt{u}) + f_Z(-\sqrt{u})\right] \\ &= \frac{1}{2\sqrt{u}}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u}{2}\right) + \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u}{2}\right)\right] \\ &= \frac{1}{\sqrt{2\pi u}}\exp\left(-\frac{u}{2}\right) \end{aligned}$$

Since this is not yet in the desirable form to conclude that U is a $\chi^2(1)$ random variable, we rearrange the above expression as follows. Given that $\Gamma(1/2) = \sqrt{\pi}$, we can rewrite this as

$$f_U(u) = \frac{(1/2)^{1/2}\, u^{(1/2)-1}}{\Gamma(1/2)} e^{-u/2}$$

and so U has a $Ga(1/2, 1/2)$ or $\chi^2(1)$ distribution.

Remark. If you are unsure about the last step of the proof above, recall from your list of continuous distributions the probability density function of a $Gamma(\alpha, \beta)$ random variable.

4.2 Method of direct transformation of one random variable

We first see the theorem of transformation of one random variable by a monotone function. We do not prove this theorem in this course; the proof can be found in any standard probability textbook.

Theorem 13. *Let X be a continuous random variable with probability density function f_X and support $I = [a, b]$. Let $g : I \to \mathbb{R}$ be a continuously differentiable monotonic function with inverse function $g^{-1} : J \to I$, where $J = g(I)$. Then, the probability density function f_Y of $Y = g(X)$ is given by*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| & \text{if } y \in J \\ 0 & \text{otherwise.} \end{cases}$$

Example. *Suppose X has the density*

$$f_X(x) = \begin{cases} \frac{\theta}{x^{\theta+1}}, & x > 1 \\ 0, & \text{otherwise.} \end{cases}$$

where θ is a positive parameter. This is an example of a Pareto distribution. What is the probability density function of $Y = \ln X$?

Answer. In this case, we have $y = g(x) = \ln x$. As the support of X, i.e. the range on which the density is non-zero, is $x > 1$ the support of Y is $y > 0$. The inverse transformation is

$$x = g^{-1}(y) = e^y$$

and

$$\frac{d}{dy}g^{-1}(y) = e^y.$$

Therefore

$$f_Y(y) = \begin{cases} \frac{\theta}{(e^y)^{\theta+1}}\, e^y = \theta e^{-y\theta}, & y > 0 \\ 0, & \text{otherwise} \end{cases}$$

and so Y has the exponential distribution, $Y \sim \text{Exp}(\theta)$.

Week 5

4.3 Method of direct transformation of two random variables

A theorem analogous to Theorem 13 allows one to compute the joint pdf of two (or indeed n) transformed random variables. We don't prove this theorem in this course. However, we describe the method for two random variables. You are supposed to know this method and to be able to use it.We remark that a similar procedure is used for $N > 2$ random variables.

Let X_1, X_2 be two random variables with joint probability density function $f_{X_1,X_2}(x_1, x_2)$ with support $A = \{(x_1, x_2) : f(x_1, x_2) > 0\}$. We are interested in the random variables $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$, where the transformation $y_1 = g_1(x_1, x_2)$ and

$y_2 = g_2(x_1, x_2)$ is a 1-1 transformation of A onto B (in other words, B is the image of A). So there is an inverse transformation $x_1 = f_1(y_1, y_2)$ and $x_2 = f_2(y_1, y_2)$.

We assume that the partial derivatives of the transformation are continuous and the determinant of the Jacobian of the transformation

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

satisfies $J \neq 0$ for $(y_1, y_2) \in B$. Then the joint probability density function of $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ is

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} f_{X_1,X_2}(f_1(y_1, y_2), f_2(y_1, y_2))\,|J|, & \text{for } (y_1, y_2) \in B, \\ 0, & \text{otherwise.} \end{cases}$$

Example. *X_1 and X_2 have joint probability density function*

$$f(x_1, x_2) = \begin{cases} \exp(-(x_1 + x_2)), & \textit{for } x_1 \geq 0,\ x_2 \geq 0 \\ 0, & \textit{otherwise.} \end{cases}$$

Consider the transformation $Y_1 = X_1$ and $Y_2 = X_1 + X_2$ and find the joint probability density function of Y_1 and Y_2.

Answer. The transformation has inverse $x_1 = y_1$, $x_2 = y_2 - y_1$ and by using $A = \{(x_1, x_2) : x_1 \geq 0,\ x_2 \geq 0\}$, we conclude that the set B is given by $B = \{(y_1, y_2) : 0 \leq y_1 \leq y_2 \leq \infty\}$. The Jacobian is

$$J = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

So the joint probability density function of Y_1 and Y_2 is given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \exp(-y_2), & 0 \leq y_1 \leq y_2 \leq \infty \\ 0, & \text{otherwise.} \end{cases}$$

□

Note that, in this example, we started with two random variables that we transformed into two other random variables. If we are only interested in one of the transformed random variables, we can integrate out the other (Find the marginal probability density function of the one we want).

4.4 Method of moment generating functions

4.4.1 Moment generating functions

The moment generating function of a random variable X, written as $M_X(t)$ is defined by

$$M_X(t) := E[e^{tX}]$$

and is defined for t in a region about 0, say for $-h < t < h$ for some h.

- If X is a discrete random variable, then the expectation $E[e^{tX}]$ is given by the sum

$$M_X(t) = \sum_x e^{tx} P(X = x).$$

- If X is a continuous random variable, then the expectation $E[e^{tX}]$ is given by an integral

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx.$$

The moment generating function is useful for the following reasons:

- $M'_X(0) = E[X]$
- $M''_X(0) = E[X^2]$
- $\mathrm{Var(X)} = \mathrm{M''_X}(0) - (\mathrm{M'_X}(0))^2$

We prove these relations for continuous random variables below (similar proof for discrete random variables can be obtained if we use $\sum$ instead of $\int$)

Proof. First note that $M_X(0) = E(e^0) = E(1) = 1$.

Differentiating $M_X(t)$ with respect to t assuming X is continuous we have

$$\begin{aligned} M'_X(t) &= \frac{d}{dt} \int e^{tx} f(x)\, dx \\ &= \int x e^{tx} f(x)\, dx \\ M'_X(0) &= \int x f(x)\, dx \\ &= E[X] \end{aligned}$$

where we assume that we can take $\frac{d}{dt}$ inside the integral. Similarly,

$$\begin{aligned} M''_X(t) &= \frac{d^2}{dt^2} \int e^{tx} f(x)\, dx \\ &= \int x^2 e^{tx} f(x)\, dx \\ M''_X(0) &= \int x^2 f(x)\, dx \\ &= E[X^2]. \end{aligned}$$

Hence $\mathrm{Var(X)} = \mathrm{M''_X}(0) - (\mathrm{M'_X}(0))^2$. □

An alternative related notion is the one of cumulant generating functions.

Definition. The cumulant generating function of a random variable X is given by

$$K_X(t) := \ln\left(M_X(t)\right), \quad \text{where } M_X(t) \text{ is the moment generating function of } X.$$

The above formulae tell us that, if we can calculate the value of the integral (or of the corresponding sum for a discrete random variable) in terms of t, then we can find the moments of X by differentiating $M_X(t)$.

Example. *Consider an exponential random variable X with probability density function $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and zero otherwise. What is the moment generating function, the mean and the variance of X?*

Answer. The moment generating function is

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx \\ &= \int_0^\infty \lambda e^{-(\lambda - t)x}\, dx \\ &= \lambda \left[-\frac{e^{-(\lambda-t)x}}{\lambda - t} \right]_0^\infty \\ &= \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda. \end{aligned}$$

The above holds only for $t < \lambda$ since $\lambda - t$ must be positive for the integral to converge.

Therefore, we have

$$M_X'(t) = \lambda(\lambda - t)^{-2},$$

which gives

$$E[X] = M_X'(0) = \lambda^{-1}.$$

Also,

$$M_X''(t) = 2\lambda(\lambda - t)^{-3},$$

which gives $E(X^2) = M_X''(0) = 2\lambda^{-2}$. Hence,

$$\text{Var(X)} = 2\lambda^{-2} - (\lambda^{-1})^2 = \lambda^{-2}.$$

□

Example. *Suppose X has a Gamma distribution, $Ga(\alpha, \beta)$. What is the moment generating function of X?*

Answer.

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \frac{\beta^\alpha\, x^{\alpha-1}\, \exp(-\beta x)}{\Gamma(\alpha)}\, dx \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x(\beta - t))\, dx \\ &= \frac{\beta^\alpha}{(\beta - t)^\alpha} \int_0^\infty \frac{(\beta - t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x(\beta - t))\, dx \end{aligned}$$

Now the integral is of the probability density function of a $Ga(\alpha, \beta - t)$ random variable and so is equal to one. Thus the moment generating function for a $Ga(\alpha, \beta)$ random variable X is

$$M_X(t) = \left(\frac{\beta}{\beta - t} \right)^\alpha \quad \text{for all } t < \beta.$$

Note we need to have $t < \beta$ to make this work, which is fine, since we are interested in letting $t \to 0$ for the calculation of moments. □

4.4.2 Method of moment generating functions

We are now ready to present the method of transformation of random variables using moment generating functions.

The following theorem (which we won't prove in this course) tells us why we can use the moment generating function to find the distributions of transformed random variables.

Theorem 14. *If X_1 and X_2 are random variables and $M_{X_1}(t) = M_{X_2}(t)$ then X_1 and X_2 have the same distribution.*

Below we see an example where the above theorem can be used to obtain the chi-square distribution as the square of the standard Normal distribution.

Example. *Suppose $Z \sim N(0,1)$ and $Y = Z^2$.*

(a) Find the distribution of Y using the moment generating function technique.

(b) Find the mean and variance of Y using its moment generating function.

Answer. (a). We have that

$$\begin{aligned}
M_Y(t) &= E[e^{tY}] \\
&= E[e^{tZ^2}] \\
&= \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}\, dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2(1-2t)}{2}\right\}\, dz \\
&= (1-2t)^{-1/2} \int_{-\infty}^{\infty} \frac{(1-2t)^{1/2}}{\sqrt{2\pi}} \exp\left\{-\frac{(1-2t)}{2} z^2\right\}\, dz \\
&= (1-2t)^{-1/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-2t)^{-1}}} \exp\left\{-\frac{z^2}{2(1-2t)^{-1}}\right\}\, dz
\end{aligned}$$

Now the function inside the integral is the probability density function of a $N(0,(1-2t)^{-1})$ random variable and so the integral equals to one. Therefore,

$$\begin{aligned}
M_Y(t) &= \left(\frac{1}{1-2t}\right)^{1/2} \\
&= \left(\frac{1/2}{1/2-t}\right)^{1/2}
\end{aligned}$$

which is the moment generating function of a $Ga(1/2, 1/2)$ random variable or equivalently of a $\chi^2(1)$ random variable. Thus the distribution of Y is $\chi^2(1)$.

(b). The mean of Y is given by

$$E(Y) = M_Y'(t)|_{t=0} = \left[\frac{1}{2}\left(\frac{1}{1-2t}\right)^{-1/2} \cdot \frac{2}{(1-2t)^2}\right]_{t=0} = 1.$$

We also have

$$E(Y^2) = M_Y''(t)|_{t=0} = \left[\frac{\partial}{\partial t}\left(\frac{1}{1-2t}\right)^{3/2}\right]_{t=0} = \left[\frac{3}{2}\left(\frac{1}{1-2t}\right)^{1/2} \cdot \frac{2}{(1-2t)^2}\right]_{t=0} = 3,$$

hence, the variance $\mathrm{Var(Y)} = 3 - 1^2 = 2$. □

4.4.3 Sums of independent random variables

The moment generating function is also useful for proving important results for sums of random variables.

Theorem 15. *Suppose $X_1, X_2, \ldots, X_n$ are* independent *random variables with moment generating function $M_{X_i}(t)$, $i = 1, \ldots, n$. Let $Y = \sum_{i=1}^n X_i$ then*

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t).$$

Proof. This is easily proved.

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{t\sum_{i=1}^n X_i}] \\ &= E[e^{tX_1}e^{tX_2}\cdots e^{tX_n}] \\ &= E[e^{tX_1}]E[e^{tX_2}]\cdots E[e^{tX_n}] \qquad \text{(by independence)} \\ &= M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t) \end{aligned}$$

□

Example. *Suppose that $X_1, X_2, \ldots, X_n$ are independent exponential random variables with mean λ^{-1} ($X_i \sim Exp(\lambda)$ for all $i = 1, \ldots, n$). Show that $Y = \sum_{i=1}^n X_i$ has a $Ga(n, \lambda)$ distribution.*

Answer. We showed in a previous example that an exponential distribution has moment generating function

$$M_{X_i}(t) = \frac{\lambda}{\lambda - t} \quad \text{for } \lambda > t.$$

Thus, since the X_i's are independent, we can calculate

$$M_Y(t) = \prod_{i=1}^n \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^n \quad \text{for } \lambda > t,$$

which is the moment generating function of a $Ga(n, \lambda)$ random variable. □

Example. *Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and normally distributed with mean $E[Y_i] = \mu_i$ and variance $\mathrm{Var[Y_i]} = \sigma_i^2$. Define the new random variable U as*

$$U = Y_1 + Y_2 + \cdots + Y_n.$$

Show that U is normally distributed with mean $E[U] = \sum_{i=1}^n \mu_i$ and variance $Var[U] = \sum_{i=1}^n \sigma_i^2$.

Answer. The moment generating function of Y_i is

$$M_{Y_i}(t) = \exp(\mu_i t + \frac{1}{2}\sigma_i^2 t^2)$$

The Y_i's are independent. Hence

$$\begin{aligned} M_U(t) &= \prod_{i=1}^{n} M_{Y_i}(t) \\ &= \prod_{i=1}^{n} \exp\left(\mu_i t + \frac{1}{2}\sigma_i^2 t^2\right) \\ &= \exp\left(\left(\sum_{i=1}^{n} \mu_i\right)t + \frac{1}{2}\left(\sum_{i=1}^{n} \sigma_i^2\right)t^2\right) \end{aligned}$$

Comparing this with the moment generating function of a normal we see that U is a Normal random variable with mean $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2$. □

Statistical remarks

(1). Obtaining the distribution of transformations of random variables is a fundamental tool used in Statistics. When testing a hypothesis it is natural to need to combine *two (or more)* random variables to obtain another one for the purpose of the test.

(2). In this course, you will see this being applied when deriving the *Student's t* distribution, from a particular combination of a standard Normal random variable and a chi-square random variable.

(3). The above t distribution is applied e.g. when performing a test of hypothesis for the mean of a normal distribution when its variance is unknown.

(4). Also the *chi-square* distribution (obtained in previous section) is applied e.g. when performing a test of hypothesis for the variance of a normal distribution when its mean is unknown (and many other tests that you will encoutner later in the course).

5 Inequalities & Sequences of random variables

In this section we start a very important topic: In many cases we cannot do the computations necessary to compute the probabilities, expectations, etc, that we are interested in. In these cases, it can be very useful to have a rough idea of the size of these things, even if we cannot get a precise answer.

5.1 Inequalities

We start with a simple example: Suppose that we know the expectation of some random variable is small. We cannot in general say that it is unlikely to be large. E.g., it could be a billion with probability 1/2 and minus a billion with probability 1/2. Then its expectation is zero, but half of the times, it is extremely large. However, we can say something if the random variable is non-negative.

Theorem 16 (Markov's Inequality)**.** *Suppose that X is a non-negative random variable. Then for any number $\delta > 0$,*

$$P(X \geq \delta) \leq \frac{E(X)}{\delta}.$$

Proof. Define a random variable Z by setting

$$Z = \begin{cases} 1 & \text{if } X \geq \delta \\ 0 & \text{if } X < \delta. \end{cases}$$

Note that $X \geq \delta Z$. To see this, consider two cases:
(i) If $X \geq \delta$, the inequality holds because $X \geq \delta \equiv \delta Z$;
(ii) If $X < \delta$, the inequality holds because $X \geq 0 \equiv \delta Z$.
But then also $E(X) \geq E(\delta Z) = \delta E(Z)$. Note that

$$E(Z) = P(Z = 1) = P(X \geq \delta)$$

and hence

$$E(X) \geq \delta P(X \geq \delta)$$

which is equivalent to the statement of the Theorem. □

Example. *When a person wants to pass a driving test, they need on average 2.5 attempts. Prove that the chance they need 10 or more attempts is at most* 1/4*.*

Answer. Let X be the number of attempts it takes to pass. We are told that $E(X) = 2.5$. Hence by Markov's Inequality

$$P(X \geq 10) \leq \frac{E(X)}{10} = \frac{2.5}{10} = \frac{1}{4}$$

□

Example. *In the same situation as above what does Markov's Inequality tell you about the probability it takes 2 or more attempts?*

Answer. This time we have

$$P(X \geq 2) \leq \frac{E(X)}{2} = \frac{2.5}{2} = 1.25.$$

This should not be confusing. The probability must be always less than 1, so in this case, Markov's Inequality is true, but not helpful. □

Note the Markov's inequality is **only** a bound - it says that the probability is not *more* than something. It might actually be much smaller than this bound.

Theorem 17 (Chebyshev's Inequality)**.** *Suppose that X is a random variable with mean μ and variance σ^2. Then for any number $\varepsilon > 0$ we have*

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Let $Y = (X - \mu)^2$. Then $E(Y) = E((X - \mu)^2) = \text{Var}(X)$ (this is either the definition of variance or an easy consequence).

We have $\{|X - \mu| \geq \varepsilon\} = \{Y \geq \varepsilon^2\}$ so $P(|X - \mu| \geq \varepsilon) = P(Y \geq \varepsilon^2)$. Now, since Y is a non-negative random variable we can apply Markov's inequality to get

$$P(|X - \mu| \geq \varepsilon) = P(Y \geq \varepsilon^2) \leq \frac{E(Y)}{\varepsilon^2} = \frac{E((X - \mu)^2)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$

□

Example. *Suppose that when you sit an exam, your expected mark is 50 and standard deviation is 10. Show that the probability that you get a first class mark (70 or more) is at most* $1/4$.

Answer. Let X be your mark. We know that X has mean 50 and variance 100. Hence by Chebychev's Inequality

$$P(X \geq 70) \leq P(|X - 50| \geq 20) = P(|X - E(X)| \geq 20) \leq \frac{100}{20^2} = \frac{1}{4}.$$

□

Week 6

5.2 The Law of Large Numbers (LLN)

The theme in this section is the following: "if we add lots of random variables then the "errors" average out."

Before we state and prove the LLN, let us recall the following property of the variance which plays a very important role in the proof.

Lemma 18. *If $X_1, X_2, ..., X_n$ is a sequence of independent random variables with $E(X_i) = \mu_j$, $\mathrm{Var}(X_j) = \sigma_j^2$ then*

$$\mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \sigma_j^2.$$

Proof. See Coursework for the proof. □

Theorem 19 (Law of Large Numbers)**.** *Suppose that $X_1, X_2, \ldots$ is a sequence of independent random variables with mean μ and variance σ^2. Let*

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then for any number $\varepsilon > 0$

$$P(|Y_n - \mu| \leq \varepsilon) \to 1 \qquad \text{as } n \to \infty.$$

Proof. We have

$$E(Y_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu,$$

and

$$\mathrm{Var}(Y_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

where we use two properties of variance: $\mathrm{Var}(cZ) = c^2\mathrm{Var}(Z)$ and Lemma 18.

Hence by Chebyshev's inequality we have

$$P(|Y_n - \mu| > \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Since $\frac{\sigma^2}{n\varepsilon^2}$ tends to zero as $n \to \infty$, so does $P(|Y_n - \mu| > \varepsilon)$. Hence

$$P(|Y_n - \mu| \leq \varepsilon) = 1 - P(|Y_n - \mu| > \varepsilon) \to 1.$$

□

Remark. This is also called the *weak LLN*. It basically says that for some specified "large" n, the average Y_n of the $(X_1, \ldots, X_n)$ is likely to be close to the mean μ. In fact, we can repeat the arguments of the above proof, to also proved a useful estimate:

$$P(|Y_n - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

5.3 Central Limit Theorem

We have seen that the average value $Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ converges to the mean (when the X_k are independent identically distributed random variables with finite variance). The Central Limit Theorem gives a much more precise description of the behaviour of the Y_n. We basically define a "scaled version" of Y_n which has zero mean and variance 1.

Theorem 20 (Central Limit Theorem)**.** *Suppose that $X_1, X_2, X_3, \ldots$ are independent identically distributed random variables with mean μ and variance σ^2. Let*

$$Z_n = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sigma\sqrt{n}}$$

Then Z_n converges, as $n \to \infty$, to a normal random variable with parameters $(0, 1)$ in the sense that, for any s, t, such that $s < t$, we have

$$P(s \leq Z_n \leq t) \to \int_s^t \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \Phi(t) - \Phi(s),$$

where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ is the cumulative distribution function of a standard Normal random variable.

Proof. Can be proved using moment generating functions and such a proof can be found in standard textbooks of probability (this is beyond the scope of this course).

□

Statistical remarks

(1) The Central Limit Theorem (CLT) only tells you about what happens as $n \to \infty$.

(2) However, in Statistics, this is commonly (and very conveniently) used for finite but large values of n.

Suppose that $X_1, X_2, \ldots$ are independent identically distributed random variables with mean μ and variance σ^2. For "large" n, we define their sum by

$$S_n = \sum_{k=1}^{n} X_k.$$

According to the CLT,

the distribution of the random variable $Z_n = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$ *is approximately standard normal.*

Using this we can also characterise the distribution of the average Y_n as n increases, which is a very useful result in statistical applications. Namely, we get that

the distribution of the average $Y_n := \frac{S_n}{n}$ *is approximately* $N(\mu, \frac{\sigma^2}{n})$.

This result justifies also the extensive use of Normal distributions in real-life applications to model data resulting from many different independent factors (roughly independent) or when the distribution of data is unknown.

Finally, we can see from the above statements that

the distribution of the random variable S_n *is approximately* $N(n\mu, n\sigma^2)$.

Example. *Suppose that W_i is the amount (in pounds) that gambler i wins at a visit in a casino, for $i = 1, 2, \ldots$. These amounts are considered to be independent random variables with mean β pounds and variance $4\beta^2$.*

(i). *What is the approximate distribution of the total profit of 100 gamblers?*

(ii). *What is the approximate probability that the total profit of 100 gamblers is negative (i.e. casino wins money), if their individual average profit is $-£5$ for each player (negative profit translates to a loss)?*

Answer. (i). We know that the total profit of 100 customers is given by $T_{100} = W_1 + W_2 + \cdots + W_{100}$ where $W_1, W_2, \ldots$ are their individual profits and we know that they are independent random variables with mean β and variance $4\beta^2$. Hence $E(W_i) = \beta$ and $\text{Var}(W_i) = 4\beta^2$. Therefore $E(T_{100}) = 100\beta$ and $\text{Var}(T_{100}) = 400\beta^2$.

Hence by the approximate Central Limit Theorem

$$T_{100} \approx N\left(100\beta, 400\beta^2\right).$$

(ii). In this case, we have $\beta = -5$. The approximate distribution (from part (i)) is therefore

$$T_{100} \approx N\left(-500, 10000\right).$$

Supposing that $Z \sim N(0, 1)$, this implies that

$$P(T_{100} < 0) \approx P\Big(Z < \frac{500}{100}\Big) = P(Z < 5) = 0.9987$$

Therefore, we do not expect that the casino will lose any money with a 99.87% chance. □