Statistical Methods MATH2715 info

Teaching material is all online!

- On Minerva `http://minerva.leeds.ac.uk`
- On GitHub `https://github.com/luisacutillo78/Statistical-Methods-Lecture-Notes`

Marking Allocation

- You can find your marker information in the red folders which is next to the homework hand-in pigeonholes.

Resources

- Mathematical Statistics and Data Analysis - 3rd ed. (by J. A. Rice);
- Introduction to Statistics - Online Edition -D.M.Lane et al.
- http://www1.maths.leeds.ac.uk/statistics/R/Rintro.pdf;
- https://www.datacamp.com/courses/free-introduction-to-r.

Joint Probability Distribution

- In many situations, there are more than one quantity associated with the experiment
- it is interesting to study the JOINT BEHAVIOUR of these random quantities.

Definition

The joint behaviour of two r.v., X and Y, is determined by the **Cumulative Distribution Function**:

$$F(X, Y) = P(X \leq x, Y \leq y)$$

Regardless of whether X and Y are continuous or discrete.

Example

Titanic

On April 15, 1912 the ocean liner Titanic collided with an iceberg and sank. Out of the 2201 passengers on board, 1495 died. The question as to whether passenger class was related to survival has been discussed extensively.

Passenger Status	Survivors	Fatalities	TOTAL
First Class	203	122	325
Second Class	118	167	285
Third Class	178	528	706
Crew	212	673	885
Total	711	1490	2201

Titanic

question

Choose a Titanic passenger randomly. What's the chance that he/she is a non-crew member who survived ?

There are two random variables in play here,

$$X = \begin{cases} 0 & \text{if passenger survived} \\ 1 & \text{if passenger died} \end{cases}$$

and

$$Y = \begin{cases} 1 & \text{if passenger was in first class} \\ 2 & \text{if passenger was in second class} \\ 3 & \text{if passenger was in third class} \\ 4 & \text{if passenger was a crew member} \end{cases}$$

We need to define the **Joint Probability Mass Function** of (X, Y) !

Joint Probability Distribution for discrete random variables

Given two discrete random variables X and Y, it's important to understand how they behave together.
We define the **joint pmf** of X and Y as,

$$p_{X,Y}(x,y) = P(X = x, Y = y)$$

In the Titanic example we approximate the joint pdf of X and Y as,

Titanic

	X=0	X=1
Y=1	0.09	0.06
Y=2	0.05	0.08
Y=3	0.08	0.24
Y=4	0.10	0.30

Hence, for example, $p_{X,Y}(0,1) = \frac{203}{2201} = 0.09$.

Titanic Example

	X=0	X=1
Y=1	0.09	0.06
Y=2	0.05	0.08
Y=3	0.08	0.24
Y=4	0.10	0.30

Question

Find $P(X + Y \leq 2)$.

$$\begin{aligned} P(X+Y \leq 2) &= \\ P(X=0, Y=1) &+ P(X=0, Y=2) + P(X=1, Y=1) = \\ &= 0.09 + 0.05 + 0.06 = 0.2 \end{aligned}$$

Conditional Distributions

Definition

Let X and Y be two discrete random variables. Then the conditional probability mass function (conditional pmf) that $X = x_i$ given that $Y = y_j$, if $p_Y(y_j) > 0$, is defined as:

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)}$$

Application

Let's see how we can apply this definition in the titanic example,

$$
\begin{aligned}
P(\text{Surv.}|\text{1st Class}) &= p_{X|Y=1}(0) = \frac{p_{X,Y}(0,1)}{p_Y(1)} = \frac{0.09}{0.09+0.06} = 0.6 \\
P(\text{Surv.}|\text{2nd Class}) &= p_{X|Y=2}(0) = \frac{p_{X,Y}(0,2)}{p_Y(2)} = \frac{0.05}{0.05+0.08} = 0.3846 \\
P(\text{Surv.}|\text{3rd Class}) &= p_{X|Y=3}(0) = \frac{p_{X,Y}(0,3)}{p_Y(3)} = \frac{0.08}{0.08+0.24} = 0.25 \\
P(\text{Surv.}|\text{Crew}) &= p_{X|Y=4}(0) = \frac{p_{X,Y}(0,4)}{p_Y(4)} = \frac{0.1}{0.1+0.3} = 0.25
\end{aligned}
$$

So we see that probability of survival across passenger class is decreasing, although we should be careful while making remarks like this and consider other factors present.

Joint Distribution of Continuous Random Variables

Definition

Suppose we have two continuous random variables X and Y. Their **joint density function** is a piecewise continuous function of two variables $f(x, y)$ which has the following properties

- $f_{X,Y}(x, y) \geq 0 \qquad \forall (x, y) \in \mathbb{R}^2$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1.$
- $\forall A \subset \mathbb{R}^2$, $P((X, Y) \in A) = \int \int_A f(x, y)dxdy$

It follows that the "**joint probability distribution function**" $F_{X,Y}$ is defined as,

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)dxdy$$

Example

Industrial production problem

A certain process for producing an industrial chemical yelds a product that contains two main types of impurities. Let X denote the proportion of impurities of Type I and Y denote the proportion of impurities of Type II. Suppose that the joint density of X and Y can be modelled as,

$$f(x,y) = \begin{cases} 2(1-x) & \text{if } 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Compute: $P(0 \le X \le 0.5, 0.4 \le Y \le 0.7)$.

Problem Solution

Problem Solution

Solution

$$
\begin{aligned}
P(0 \leq X \leq 0.5, 0.4 \leq Y \leq 0.7) &= \int_0^{0.5} \int_{0.4}^{0.7} 2(1-x)dydx \\
&= \int_0^{0.5} (0.7-0.4) \times 2(1-x)dx \\
&= 0.3 \times \int_0^{0.5} 2(1-x)dx \\
&= 0.3 \times \left[-(1-x)^2\right]_0^{0.5} \\
&= 0.3 \times \left[-(1-0.5)^2 + (1-0)^2\right] \\
&= 0.3 \times [1-0.25] \\
&= 0.3 \times 0.75 \\
&= 0.225
\end{aligned}
$$

Marginal pdf

Definition

If X and Y are continuous random variables, with joint pdf $f_{X,Y}$ then the individual or marginal pdf's f_X and f_Y are given by,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

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

Industrial production problem

For the industrial production example considered previously, compute f_X:

Industrial production problem

For the industrial production example considered previously, compute f_X:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\ &= \int_{-\infty}^{\infty} 2(1-x)1_{\{0\leq x\leq 1, 0\leq y\leq 1\}}dy \\ &= \int_{0}^{1} 2(1-x)1_{\{0\leq x\leq 1\}}dy \\ &= 2(1-x)1_{\{0\leq x\leq 1\}} \end{aligned}$$

hence:

$$f_X(x) = \begin{cases} 2(1-x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Industrial production problem

For the industrial production example considered previously, compute f_Y:

Industrial production problem

For the industrial production example considered previously, compute f_Y:

$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx \\
&= \int_{-\infty}^{\infty} 2(1-x)1_{\{0\le x\le 1, 0\le y\le 1\}}dx \\
&= \int_{0}^{1} 2(1-x)1_{\{0\le y\le 1\}}dx \\
&= 1_{\{0\le y\le 1\}}\left[-(1-x)^2\right]_0^1 \\
&= 1_{\{0\le y\le 1\}} \times 1
\end{aligned}
$$

hence:

$$
f_Y(y) = \begin{cases} 1 & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}
$$

Conditional pdf

Suppose we are interested in the behaviour of the random variable Y given $X = x$.

Conditional pdf

Let X and Y be continuous random variables with joint pdf $f_{X,Y}$ and marginal pdf's f_X and f_Y. Then for any x such that $f_X(x) > 0$, we define the **conditional pdf** of Y given $X = x$ as,

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

for all $y \in \mathbb{R}$.

Now we can calculate, for example:

$$P(Y \leq 0.5|X = 0.3) = \int_{-\infty}^{0.5} f_{Y|X=0.3}(y)dy$$

Independence of random variables

We recall that, two events A and B are said to be independent if,

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

We also recall that two discrete random variables X and Y are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Independence for continuous rv

Let X and Y be two continuous random variables. We say X and Y are independent if

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

$\forall x, y \in \mathbb{R}$.

We can verify that the above definition is same as saying

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

$\forall x$ such that $f_X(x) > 0$ and $\forall y \in \mathbb{R}$.

convolution: discrete case

Assume X and Y are discrete rv taking values on integers with joint pmf $p(x,y)$. Let $Z = X + Y$. Note that $Z = z$ whenever $X = x$ and $Y = z - x$. It follows that:

$$P(Z = z) = p_Z(z) = \sum_{-\infty}^{\infty} p(x, z - x)$$

if X and Y are independent then:

$$P(Z = z) = p_Z(z) = \sum_{-\infty}^{\infty} p_X(x) p_Y(z - x)$$

This sum is called the **convolution** of the sequences p_X and p_Y.

Functions of Jointly distributed rv

convolution: continuous case

Assume X and Y are continuous rv. Let $Z = X + Y$. Let's start computing the cdf $F(Z)$:

$$F_Z(z) = \int\int_{\{(x,y):x+y\leq z\}} f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f(x,y)dxdy$$

Make a change of variable from y to $v = x + y$ and then reverse the order of integration: if X and Y are independent then:

$$F_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f(x,v-x)dvdy = \int_{-\infty}^{z-x}\int_{-\infty}^{\infty} f(x,v-x)dvdy$$

Finally, differentiate to find the density: $f_Z(z) = \int_{-\infty}^{\infty} f(x,z-x)dx$

Again, if X and Y are independent, the result is a **convolution**: $f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$

Homeworks for next week workshop

Exercise 1

Let X and Y be independent exponential random variables with the same parameter λ. Find the distribution of their sum: $Z = X + Y$.

Exercise 2

Let X and Y be independent standard normal random variables. That is, N(0, 1). Find the distribution of the ratio: $Z = Y/X$.

Exercise 3

The joint distribution of X and Y is given in the following table.

	X=0	X=1	X=3
Y=-1	0.11	0.03	0.00
Y=2.5	0.03	0.09	0.16
Y=3	0.15	0.15	0.06
Y=4.7	0.04	0.16	0.02

Find $P(Y - X \leq 2)$, $P(2 \leq Y \leq 4|X = 1)$.

Homeworks for next week workshop

Exercise 4

The joint distribution of X and Y is given in the following table:

	X=-1	X=-2	X=2	X=3
Y=-3	0.14	0.14	0.01	0.05
Y=-1	0.15	0.06	0.06	0.04
Y=1	0.03	0.10	0.11	0.11

Find $P(Y + X \leq 0)$, $P(-2 \leq X \leq 2 | Y = -1)$.