Probability and Statistics

Omid Safarzadeh

December 20, 2021

Table of contents

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

Joint and Marginal Distribution

- So far, our interest has been on events involving a single random variable only. In other words, we have only considered "univariate models."
- Multivariate models, on the other hand, involve more than one variable.
- Consider an experiment about health characteristics of the population. Would we be interested in one characteristic only, say weight? Not really. There are many important characteristics.

Definition 1.1

An n-dimensional random vectoris a function from a sample space Ω into $\mathbb{R}^n$, n-dimensional Euclidean space.

- Suppose, for example, that with each point in a sample space we associate an ordered pair of numbers, that is, a point $(x, y) \in \mathbb{R}^2$, where $\mathbb{R}^2$ denotes the plane. Then, we have defined a two-dimensional (or bivariate) random vector (X, Y).

Joint and Marginal Distribution

Example 1.1

Consider the experiment of tossing two fair dice. The sample space has 36 equally likely points. For example:

$$\begin{aligned} (3,3) &: \text{ both dices show a 3,} \\ (4,1) &: \text{ first dice shows a 4 and the second die a 1.} \end{aligned}$$

- Now, let

$$X = \text{sum of the two dice} \quad \& \quad Y = |\text{difference of the two dice}|.$$

Then,

$$\begin{aligned} (3,3) &:: X = 6 \quad \text{and} \quad Y = 0, \\ (4,1) &:: X = 5 \quad \text{and} \quad Y = 3, \end{aligned}$$

and so we can define the bivariate random vector (X, Y) thus.

Joint and Marginal Distribution

- What is, $P(X = 5 \text{ and } Y = 3)$? One can verify that the two relevant sample points in Ω are (4,1) and (1,4). Therefore, the event $\{X = 5 \text{ and } Y = 3\}$ will only occur if the event $\{(4,1),(1,4)\}$ occurs. Since each of these sample points in Ω are equally likely,

$$P(\{(4,1),(1,4)\}) = \frac{2}{36} = \frac{1}{18}.$$

  Thus,

$$P(X = 5 \text{ and } Y = 3)\frac{1}{18}.$$

- For example, can you see why

$$P(X = 7, Y \leq 4) = \frac{1}{9}?$$

  This is because the only sample points that yield this event are (4,3), (3,4), (5,2) and (2,5).
- Note that from now on we will use $P(\text{event a, event b})$ rather that $P(\text{event a and event b})$.

Joint and Marginal Distribution

Definition 1.2

Let (X, Y) be a discrete bivariate random vector. Then the function $f(x, y)$ form $\mathbb{R}^2$ into $\mathbb{R}$, defined by $f(x, y) = P(X = x, Y = y)$ is called the joint probability mass function or joint pmf (X,Y). If it is necessary to stress the fact that f is the joint pmf of the vector (X, Y) rather than some vector, the notation $f_{X,Y}(x, y)$ will be used.

Joint and Marginal Distribution

- As before, we can use the joint pmf to calculate the probability of any event defined in terms of (X, Y). For $A \subset \mathbb{R}^2$,

$$P((X,Y) \in A) = \sum_{\{x,y\} \in A} f(x,y).$$

- We could, for example, have $A = \{(x,y) : x = 7 \text{ and } y \leq 4\}$. Then,

$$P((X,Y) \in A) = P(X = 7, Y \leq 4) = f(7,1) + f(7,3) = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

- Expectations are also dealt with in the same way as before. Let $g(x,y)$ be a real-valued function defined for all possible values (x,y) of the discrete random vector (X,Y). Then, $g(X,Y)$ is itself a random variable and its expected value is

$$E[g(X,Y)] = \sum_{(x,y) \in \mathbb{R}^2} g(x,y) f(x,y).$$

Example 1.2

For the (X, Y) whose joint pmf is given in the above Table, what is the expected value of XY? Letting $g(x, y) = xy$, we have

$$E[XY] = 2 * 0 * \frac{1}{36} + 4 * 0 * \frac{1}{36} + ... + 8 * 4 * \frac{1}{36} + 7 * 5 * \frac{1}{18} = 13\frac{1}{18}.$$

- As before,

$$E[ag_1(X, Y) + bg_2(X, Y) + c] = aE[g_1(X, Y)] + E[bg_2(X, Y)] + c.$$

- One very useful result is that any non-negative function from $\mathbb{R}^2$ into $\mathbb{R}$ that is nonzero for at most a countable number of (x, y) pairs sums to 1 is the joint pmf for some bivariate discrete random vector (X, Y).

Joint and Marginal Distribution

Example 1.3

Define $f(x, y)$ by

$$f(0,0) = f(0,1) = 1/6,$$

$$f(1,0) = f(1,1) = 1/3,$$

$$f(x,y) = 0 \qquad \text{for any other } (x,y)$$

Joint and Marginal Distribution

- Suppose we have a multivariate random variable (X, Y) but are concerned with, say, $P(X = 2)$ only.
- We know the joint pmf $f_{X,Y}(x, y)$ but we need $f_X(x)$ in this case.

Theorem 1.1

Let (X, Y) be a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$. Then the marginal pmfs of X and Y, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y)$$

- **Proof**: For any $\bar{x} \in \mathbb{R}$, let $A_{\bar{x}} = \{(\bar{x}, y) : -\infty < y < \infty\}$. That is, $A_{\bar{x}}$ is the line in the plane with first coordinate equal to $\bar{x}$. Then, for any $\bar{x} \in \mathbb{R}$,

$$f_X(\bar{x}) = P(X = \bar{x}, -\infty < Y < \infty)$$

$$= P((X, Y) \in A_{\bar{x}}) = \sum_{(x,y) \in A_{\bar{x}}} f_{X,Y}(x, y)$$

$$= \sum_{y \in \mathbb{R}} f_{X,Y}(\bar{x}, y).$$

Joint and Marginal Distribution

Example 1.4

Now we can compute the marginal distribution for X and Y from the joint distribution given in the above Table. Then

$$f_Y(0) = f_{X,Y}(2,0) + f_{X,Y}(4,0) + f_{X,Y}(6,0)$$

$$+f_{X,Y}(8,0) + f_{X,Y}(10,0) + f_{X,Y}(12,0)$$

$$= 1/6.$$

As an exercise, you can check that,

$$f_Y(1) = 5/18, \quad f_Y(2) = 2/9, \quad f_Y(3) = 1/6, \quad f_Y(4) = 1/9, \quad f_Y(5) = 1/18.$$

Notice that $\sum_{y=0}^{5} f_Y(y) = 1$, as expected, since these are the only six possible values of Y.

Joint and Marginal Distribution

- Now, it is crucial to understand that the marginal distribution of X and Y, described by the marginal pmfs $f_X(x)$ and $f_Y(y)$, do not completely describe the joint distribution of X and Y.
- These are, in fact, many different joint distributions that have the same marginal distributions.
- The knowledge of the marginal distributions only does not allow us to determine the joint distribution (except under certain assumptions).

Joint and Marginal distribution

Example 1.5

Define a joint pmf by

$$f(0,0) = 1/12, \quad f(1,0) = 5/12, \quad f(0,1) = f(1,1) = 3/12,$$

$$f(x,y) = 0 \qquad \text{for all other values.}$$

- Then,

$$f_Y(0) = f(0,0) + f(1,0) = 1/2,$$
$$f_Y(1) = f(0,1) + f(1,1) = 1/2,$$
$$f_X(0) = f(0,0) + f(0,1) = 1/3,$$

and

$$f_X(1) = f(1,0) + f(1,1) = 2/3.$$

Example 1.5 cont.

- Now consider the marginal pmfs for the distribution considered in Example (1.3).

$$f_Y(0) = f(0,0) + f(1,0) = 1/6 + 1/3 = 1/2,$$
$$f_Y(1) = f(0,1) + f(1,1) = 1/6 + 1/3 = 1/2,$$
$$f_X(0) = f(0,0) + f(0,1) = 1/6 + 1/6 = 1/3,$$

and

$$f_X(1) = f(1,0) + f(1,1) = 1/3 + 1/3 = 2/3.$$

- We have the same marginal pmfs but the joint distributions are different!

Joint and Marginal distribution

- Consider now the corresponding definition for continuous random variables.

Definition 1.3

A function $f(x, y)$ from $\mathbb{R}^{\nvDash}$ to $\mathbb{R}$ is called a joint probability density function or joint pdf of the continuous bivariate random vector (X, Y) if, for every $A \subset \mathbb{R}^2$,

$$P((X, Y) \in A) = \int \int_A f(x, y) dxdy.$$

- The notation $\int \int_A$ means that the integral is evaluated over all $(x, y) \in A$.
- Naturally, for real valued functions $g(x, y)$,

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dxdy.$$

- It is important to realise that the joint pdf is defined for all $(x, y) \in \mathbb{R}^2$. The pdf may equal 0 on a large set A if $P((X, Y) \in A) = 0$ but the pdf is still defined for the points in A.

Joint and Marginal distribution

- Again, naturally,

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

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

- As before, a useful result is that any function $f(x,y)$ satisfying $f(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$ and

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy,$$

is the joint pdf of some continuous bivariate random vector (X,Y).

Joint and Marginal distribution

- The joint probability distribution of (X, Y) can be completely described using the joint cdf (cumulative distribution function) rather than with the joint pmf or joint pdf.
- The joint cdf is the function $F(x, y)$ defined by

$$F(x, y) = P(X \leq x, Y \leq y) \quad \text{for all (x,y)} \ (x, y) \in \mathbb{R}^2.$$

- Although for discrete random vectors it might not be convenient to use the joint cdf, for continuous random variables, the following relationship makes the joint cdf very useful:

$$F(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(s, t) ds dt.$$

- From the bivariate Fundamental Theorem of Calculus,

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

at continuously points of $f(x, y)$. This relationship is very important.

Conditional Distributions and Independence

- We have talked a little bit about conditional probabilities before. Now we will consider conditional distributions.
- The idea is the same. If we have some extra information to make better inference.
- Suppose we are sampling from a population where X is the height (in kgs) and Y is the weight (in cms). What is $P(X > 95)$? Would we have a better/more relevant answer if we knew that the person in question has $Y = 202cms$? Usually, $P(X > 95|Y = 202)$ is supposed to be much larger than $P(X > 95|Y = 165)$.
- Once we have the joint distribution for (X, Y), we can calculate the conditional distributions, as well.
- Notice that now we have three distribution concepts: marginal distribution, conditional distribution and joint distribution.

Conditional Distributions and Independence

Definition 2.1

Let (X, Y) be a discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $P(X = x) = f_X(x) > 0$, the conditional pmf of Y given that $X = x$ is the function of y denoted by $f(y|X)$ and defined by

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

$Y = y$ is the function of x denoted by $f(x|y)$ and defined by

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

- Can we verify that, say, $f(y|x)$ is a pmf? First, since $f(x, y) \geq 0$ and $f_X(x) > 0$. $f(y|x) \geq 0$ for every y. Then,

$$\sum_y f(y|x) = \frac{\sum_y f(x, y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$$

Conditional Distributions and Independence

Example 2.1

Define the joint pmf of (X, Y) by

$$f(0,10) = f(0,20) = \frac{2}{18}, \quad f(1,10) = f(1,30) = \frac{3}{18},$$

$$f(1,20) = \frac{4}{18} \quad \text{and} \quad f(2,30) = \frac{4}{18},$$

while $f(x, y) = 0$ for all other combinations of (x, y).

- Then,

$$f_X(0) = f(0,10) + f(0,20) = \frac{4}{18},$$

$$f_X(1) = f(1,10) + f(1,20) + f(1,30) = \frac{10}{18},$$

$$f_X(2) = f(2,30) = \frac{4}{18}.$$

EXample 2.1 cont.

- Moreover,

$$f(10|0) = \frac{f(0,10)}{f_X(0)} = \frac{2/18}{4/18} = \frac{1}{2},$$

$$f(20|0) = \frac{f(0,20)}{f_X(0)} = \frac{2/18}{4/18} = \frac{1}{2},$$

Therefore, given the knowledge that $X = 0$, Y is equal to either 10 or 20, with equal probability.

Conditional Distributions and Independence

- In addition,

$$f(10|1) = f(30|1) = \frac{3/18}{10/18} = \frac{3}{10},$$

$$f(20|1) = \frac{4/18}{10/18} = \frac{4}{10},$$

$$f(30|2) = \frac{4/18}{4/18} = 1.$$

  Interestingly, when $X = 2$, we know for sure that Y will be equal to 30.

- Finally,

$$P(Y > 10|X = 1) = f(20|1) + f(30|1) = \frac{7}{10},$$

$$P(Y > 10|X = 0) = f(20|0) = \frac{1}{2},$$

etc...

- The analogous definition for continuous random variables is given nest.

Definition 2.2

Let (X, Y) be a continuous bivariate random vector with joint pdf $f(X, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the conditional pdf of Y given that $X = x$ is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

For any y such that $f_Y(y) > 0$, the conditional pdf of X given that $Y = y$ is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

- Note that for discrete random variables, $P(X = x) = f_X(x)$ and $P(X = x, Y = y) = f(x, y)$. Then Definition (2.1) is actually parallel to the definition of ($P(Y = y|X = x)$ in Definition (2.1). The same interpretation is not valid for continuous random variables since ($P(X = x) = 0$ for every x. However, replacing pmfs with pdfs lead to Definition (2.2).
- The conditional expected value of $g(Y)$ given $X = x$ is given by

$$E[g(Y)|x] = \sum_y g(y)f(y|x) \quad \text{and} \quad E[g(Y)|x] = \int_{-\infty}^{\infty} g(y)f(y|x)dx,$$

  in the discrete and continuous cases, respectively.
- The conditional expected value has all of the properties of the usual expected value listed in Theorem (2.1)

Conditional Distributions and Independence

Definition 2.3

Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called independent random variables if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f(x, y) = f_X(x)f_Y(y). \tag{1}$$

Conditional Distributions and Independence

- Now, in the case of independence, clearly,

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y).$$

- We can either start with the joint distribution and check independence for each possible value of x and y, or start with the assumption that X and Y are independent and model the joint distribution accordingly. In this latter direction, our economic intuition might have to play an important role.
- "Would information on the value of X really increase our information about the likely value of Y?"

Example 2.2

Consider the discrete bivariate random vector (X, Y), with joint pmf given by

$$f(10,1) = f(20,1) = f(20,2) = 1/10,$$

$$f(10,2) = f(10,3) = 1/5 \quad \text{and} \quad f(20,3) = 3/10.$$

- The marginal pmfs are then given by

$$f_X(10) = f_X(20) = 0.5 \quad \text{and} \quad f_Y(1) = 0.2, f_Y(2) = 0.3 \quad \text{and} \quad f_Y(3) = 0.5.$$

Example 2.2 cont.

- Now, for example,

$$f(10,3) = \frac{1}{5} \neq \frac{1}{2}\frac{1}{2} = f_X(10)f_Y(3),$$

although

$$f(10) = \frac{1}{10} = \frac{1}{2}\frac{1}{5} = f_X(10)f_Y(1).$$

- Do we always have to check all possible pairs, one by one???

Example 2.3

Let X be the number of living parents of a student randomly selected from an elementary school in Kansas city and Y be the number of living parents of a retiree randomly selected from Sun City. Suppose, furthermore, that we have

$$f_X(0) = 0.01 \quad f_X(1) = 0.09 \quad f_X(2) = 0.9,$$

$$f_Y(0) = 0.7 \quad f_Y(1) = 0.25 \quad f_Y(2) = 0.05.$$

- It seems reasonable that X and Y will be independent: knowledge of the number of parents of the student does not give us any information on the number of parents of the retiree and vice versa. Therefore, we should have

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

Example 2.3 cont.

- Then, for example

$$f_{X,Y}(0,0) = 0.007, \quad f_{X,Y}(0,1) = 0.0025,$$

  etc.

- We can thus calculate quantities such as,

$$P(X = Y) = f(0,0) + f(1,1) + f(2,2)$$

$$= 0.01 * 0.7 + 0.09 * 0.25 + 0.9 * 0.05 = 0.0745.$$

Theorem 2.1

Let X and Y be independent random variables.

1. *For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$; that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.*
2. *Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

- **Proof**: *Exercise!*

Theorem 2.2

Let $X \sim N(\mu_X, \sum_X^2$ and $Y \sim N(\mu_Y, \sum_y^2)$ be independent normal variables. Then the random variable $Z = X + Y$ has a $N(\mu_X + \mu_Y, \sum_X^2 + \sum_Y^2)$ distribution.

- **Proof**: *Exercise!*

Bivariate Transformations

Theorem 3.1

If $X \sim Poisson(\theta)$, $Y \sim Poisson(\lambda)$ and X and Y are independent, then $X + Y \sim Poisson(\theta + \lambda)$

Theorem 3.2

Let $X \sim N(\mu_x, \sum_x^2)$ and $Y \sim N(\mu_Y, \sum_Y^2)$ be independent normal variables. Then the random variable $Z = X + Y$ has a $N(\mu_X + \mu_Y, \sum_X^2 + \sum_Y^2)$ distribution.

- Then,

$$U = X + Y \sim N(0, 2).$$

- What about V? Define $Z = -Y$ and notice that

$$Z = -Y \sim N(0, 1).$$

- Then, by Theorem (3.2)

$$V = X - Y = X + Z \sim N(0, 2),$$

as well.

Bivariate Transformations

Theorem 3.3

Let $X \perp\!\!\!\perp Y$ be two random variables. Define $U = g(X)$ and $V = h(Y)$, where $g(x)$ is a function only of x and $h(y)$ is a function only of y. Then $U \perp\!\!\!\perp V$.

- **Proof**: *Exercise!*

Hierarchical Models and Mixture Distribution

- Now comes a very useful theorem which you will, most likely, use frequently in the future.
- Remember that $E[X[Y]]$ is a function of y and $E[X|Y]$ is a random variable whose value depends on the value of Y.

Theorem 4.1

If X and Y are two random variables, then

$$E_X[X] = E_Y\{E_{X|Y}[X|Y]\},$$

provided that the expectations exist.

- It is important to notice that the two expectations are with respect to two different probability densities, $f_X(.)$ and $f_{X|Y}(.|Y = y)$.
- This result is widely known as the Law of Iterated Expectations.

Hierarchical Models and Mixture Distribution

Definition 4.1

A random variable X is said to have a mixture distributionif the distribution of X depends on a quantity that also has a distribution.

- Therefore, the mixture distribution is a distribution that is generated through a hierarchical mechanism.

Hierarchical Models and Mixture Distribution

Example 4.1

Now, consider the following hierarchical model:

$$\begin{aligned} X|Y &\sim \text{binomial}(Y, p), \\ Y|\Lambda &\sim \text{Poisson}(\Lambda), \\ \Lambda &\sim \text{exponential}(\beta), \end{aligned}$$

- Then,

$$\begin{aligned} E_X[X] = E_Y\{E_{X|Y}[X|Y]\} = E_Y[pY] \\ = E_\Lambda\{E_{Y|\Lambda}[pY|\Lambda]\} = pE_\Lambda\{E_{Y|\Lambda}[Y|\Lambda]\} \\ = pE_\Lambda[\Lambda] = p\beta, \end{aligned}$$

Example 4.1 cont.

Which is obtained by successive application of the Law of Iterated Expectations.

- Note that in this example we considered both discrete and continuous random variables. This is fine.

Theorem 4.2

For any two random variables X and Y,

$$Var_X(X) = E_Y[Var_{X|Y}(X|Y)] + Var_Y\{E_{X|Y}[X|Y]\}$$

- **Proof**: *Exercise!*

Example 4.2

Consider the following generalisation of the binomial distribution, where the probability of success varies according to a distribution.

- Specifically,

$$\begin{aligned} X|P &\sim \text{binomial}(n, P), \\ P &\sim \text{beta}(\alpha, \beta), \end{aligned}$$

- Then

$$E_X[X] = E_P\{E_{X|P}[X|P]\} = E_P[nP] = n\frac{\alpha}{\alpha+\beta},$$

  where the last result follows from the fact that for $P \sim \text{beta}(\alpha, \beta)$, $E[P] = \alpha/(\alpha+\beta)$.

Example 4.3

Now, let's calculate the variance of X. By Theorem (4.2),

$$Var_X(X) = Var_p\{E_{X|P}[X|P]\} + E_P[Var_{X|P}(X|P)].$$

- Now, $E_{X|P}[X|P] = nP$ and since $P \sim$beta$(\alpha + \beta)$,

$$Var_P(E_{X|P}[X|P]) = Var_p(nP) = n^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

- Moreover, $Var_{X|P}(X|P) = nP(1-P)$, due to $X|P$ being a *binomial* random variable.

Covariance and Correlation

- Let X and Y be two random variables. To keep notation concise, we will use the following notation.

$$E[X] = \mu_X, \quad E[Y] = \mu_Y, \quad Var(X) = \sum_X^2 \quad and \quad Var(Y) = \sum_Y^2.$$

Definition 5.1

The covariance of X and Y is the number defined by

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Definition 5.2

The correlation of X and Y is the number defined by

$$\rho_{XY} = \frac{Cov(X, Y)}{\sum_x \sum_y},$$

which is also called the correlation coefficient.

Covariance and Correlation

- If large(small) values of X tend to be observed with large(small) values of Y, then will be positive.
- Why so? Within the above setting, when $X > \mu_X$ then $Y > \mu_Y$ is likely to be true whereas when $X < \mu_X$ then $Y < \mu_Y$ is likely to be true. Hence

$$E[(X - \mu_X)(Y - \mu_Y)] > 0.$$

- Similarly, if large(small) values of X tend to be observed with small(large) values of Y, then $Cov(X, Y)$ will be negative.

Covariance and Correlation

- Correlation normalises covariance by the standard deviations and is, therefore, a more informative measure.
- If Cov(X, Y)=50 while Cov(W, Z)=0.9, this does not necessarily mean that there is a much stringer relationship between X and Y. For example, if Car(X)=Var(Y)=100 while Var(W)=Var(Z)=1, then

$$\rho_{XY} = 0.5 \quad \rho_{WZ} = 0.9.$$

Theorem 5.1

For any random variables X and Y,

$$Cov(X, Y) = E[XY] - \mu_X \mu_Y.$$

- **Proof**:

$$\begin{aligned} Cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y. \end{aligned}$$

Theorem 5.2

If $X \perp\!\!\!\perp Y$, then Cov(X, Y)$= \rho_{XY} = 0$.

- **Proof**: Since $X \perp\!\!\!\perp Y$, by Theorem (2.1), Then

$$Cov(X, Y) = E[XY] - \mu_X \mu_Y = \mu_X \mu_Y - \mu_X \mu_Y = 0,$$

and consequently,

$$\rho_{XY} = \frac{Cov(X, Y)}{\sum_X \sum_Y} = 0.$$

- It is crucial to note that although $X \perp\!\!\!\perp Y$ implies that $Cov(X, Y) = \rho_{XY} = 0$, the relationship does not necessarily hold in the reverse direction.

Covariance and Correlation

Theorem 5.3

If X and Y are any two random variables and a and b are any two constants, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

If X and Y are independent random variables, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y).$$

- **Proof**: *Exercise!*

- Note that if two random variables, X and Y, are positively correlated, then

$$Var(X + Y) > Var(X) + Var(Y),$$

  whereas if X and Y are negatively correlated, then

$$Var(X + Y) < Var(X) + Var(Y).$$

- For positively correlated random variables, large values in one tend to be accompanied by large values in the other. Therefore, the total variance is magnified.
- Similarly, for negatively correlated random variables, large values in one tend to be accompanied by small values in the other. Hence, the variance of the sum is dampened.

Bivariate Normal Distribution

- We now introduce the bivariate normal distribution.

Definition 6.1

Let $-\infty < \mu_X < \infty$, $-\infty < \mu_Y < \infty$, $\sum_X > 0$, $\sum_Y > 0$ and $-1 < \rho < 1$. The bivariate normal pdf with means μ_X and μ_y, variances $\sum_X^2$ and $\sum_Y^2$, and correlation ρ is the bivariate pdf given by

$$f_{X,Y}(x, y) = \frac{1}{2\pi \sum_X \sum_Y \sqrt{1-\rho^2}}$$

$$x\ exp\{-\frac{1}{2(1-\rho^2)}[u^2 - 2\rho uv + v^2]\},$$

where $u = (\frac{y-\mu_Y}{\sum_Y})$ and $v = (\frac{x-\mu_X}{\sum_X})$, while $-\infty < x < \infty$ and $-\infty < y < \infty$.

Bivariate Normal Distribution

- More concisely, this would be written as

$$\binom{X}{Y} \sim N\{\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sum_X^2 & \rho \sum_X \sum_Y \\ \rho \sum_X \sum_Y & \sum_Y^2 \end{pmatrix}\}.$$

- In addition, starting from the bivariate distribution, one can show that

$$Y|X = x \sim N\{\mu_Y + \rho \sum_Y (\frac{x - \mu_X}{\sum_X}, \sum_Y^2 (1 - \rho^2)\},$$

and, likewise,

$$X|Y = y \sim N\{\mu_X + \rho \sum_X (\frac{y - \mu_Y}{\sum_Y}, \sum_X^2 (1 - \rho^2)\}.$$

- Finally, again, starting from the bivariate distribution, it can be shown that

$$X \sim N(\mu_X, \sum_X^2) \quad \text{and} \quad Y \sim N(\mu_Y, \sum_Y^2).$$

- Therefore, joint normality implies conditional and marginal normality. However, this does not go in the opposite direction; marginal or conditional normality does not necessarily imply joint normality.

Bivariate Normal Distribution

- The normal distribution has another interesting property.
- Remember that although independence implies zero covariance, the reverse is not necessarily true.
- The normal distribution is an exception to this: if two normally distributed random variables have zero correlation (or, equivalently, zero covariance) then they are independent.
- Why? Remember that independence is a property that governs all moments, not just the second order ones (such as variance or covariance).
- However, as the preceding discussion reveals, the distribution of a bivariate normal random variable is entirely determined by its mean and covariance matrix. In other words, the first and second order moments are sufficient to characterise the distribution.
- Therefore, we do not have to worry about any higher order moments. Hence, zero covariance implies independence in this particular case.

Multivariate Distribution

- Let $\mathbf{X} = (X_1, ..., X_n)$. Then the sample space for $\mathbf{X}$ is a subset of $\mathbb{R}^n$, the n-dimensional Euclidian space.
- If this sample space is countable, then $\mathbf{X}$ is a discrete random vector and its joint pmf is given by

$$f(\mathbf{x}) = f(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n) \text{ for each } (x_1, ..., x_n) \in \mathbb{R}^n.$$

- For any $A \subset \mathbb{R}^n$,

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}).$$

- Similarly, for the continuous random vector, we have the joint pdf given by $f(\mathbf{x}) = f(x_1, ..., x_n)$ which satisfies

$$P(\mathbf{X} \in A) = \int ... \int_A f(\mathbf{x}) d\mathbf{x} = \int ... \int_A f(x_1, ..., x_n) dx_1 ... dx_n.$$

- Note that $\int ... \int_A$ is an n-fold integration, where the limits of integration are such that the integral is calculated over all points $\mathbf{x} \in A$.

Multivariate Distribution

- Let $g(\mathbf{x}) = g(x_1, ..., x_n)$. be a real-valued function defined on the sample space of $\mathbf{X}$. Then, for the random variable $g(\mathbf{X})$,

$$\text{(discrete)} \quad : \quad E[g(\mathbf{X}] = \sum_{x \in \mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x}),$$

$$\text{(continuous)} \quad : \quad E[g(\mathbf{X}] = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x}) dx.$$

- The marginal pdf or pmf of $(X_1, ..., X_k)$, the first k coordinates of $(X_1, ..., X_n)$, is given by

$$\text{(discrete)} \quad : \quad f(x_1, ..., x_k) = \sum_{(x_{k+1}, \dots, x_n) \in \mathbb{R}^{n-k}} f(x_1, ..., x_n),$$

$$\text{(discrete)} \quad : \quad f(x_1, ..., x_k) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x_1, ..., x_n), dx_{k+1} ... dx_n,$$

for every $(x_1, ..., x_k) \in \mathbb{R}^k$.

Multivariate Distribution

Definition 7.1

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be random vectors with joint pdf or pmf $f(\mathbf{x}_1, ..., \mathbf{x}_n)$. Let $f_{\mathbf{X}_i}(\mathbf{x}_i)$ denote the marginal pdf of pmf of $\mathbf{X}_i$. Then, $\mathbf{X}_1, ...\mathbf{X}_n$ are called mutually independent random vectors if, for every $(\mathbf{x}_1, ...\mathbf{x}_n)$,

$$f(\mathbf{x}_1, ..., \mathbf{x}_n) = f_{\mathbf{X}_1}(\mathbf{x}_1)......f_{\mathbf{X}_n}(\mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i)$$

- If the $\mathbf{X}_i$s are all one-dimensional, then $X_1, ..., X_n$ are called mutually independent random variables.

Theorem 7.1

Generalisation of Theorem 3.3: *Let* $\mathbf{X}_1, ..., \mathbf{X}_n$ *be independent random vectors. Let* $g_i(\mathbf{x}_i)$ *be a function of* $\mathbf{x}_i$, $i = 1, ..., n$. *Then, the random variables* $U_i = g_i(\mathbf{X}_i)$, $i = 1, ..., n$, *are mutually independent.*

Inequalities

- We will now cover some basic inequalities used in statistics and econometrics.
- Most of the time, more complicated expressions can be written in terms of simpler expressions. Inequalities on these simpler expressions can then be used to obtain an inequality, or often a bound, on the original complicated term.
- This part is based on Sections 3.6 and 4.7 in Casella & Berger.

- We start with one of the most famous probability inequalities.

Theorem 8.1

Chebychev's Inequality: *Let X be a random variable and let $g(x)$ be a non-negative function. Then, for any $r > 0$,*

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r}.$$

Inequalities

- **Proof**: Using the definition of the expected value,

$$
\begin{aligned}
E[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\
&\geq \int_{\{x: g(x) \geq r\}} g(x) f_X(x) dx \\
&\geq r \int_{\{x: g(x) \geq r\}} f_X(x) dx \\
&= rP(g(X) \geq r).
\end{aligned}
$$

- Hence,

$$E[g(X)] \geq rP(g(X) \geq r),$$

 implying

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r}.$$

Inequalities

- This result comes in very handy when we want to turn a probability statement into an expectation statement. For example, this would be useful if we already have some moment existence assumptions and we want to prove a result involving a probability statement.

Example 8.1

Let $g(x) = (x-\mu)^2/\sigma^2$, where $\mu = E[X]$ and $\sigma^2 = Var(X)$. Let, for convenience, $r = t^2$. Then,

$$P[\frac{(X-\mu)^2}{\sigma^2} \geq t^2] \leq \frac{1}{t^2} E[\frac{(X-\mu)^2}{\sigma^2}] = \frac{1}{t^2}.$$

- This implies that

$$P[|X-\mu| \geq t\sigma] \leq \frac{1}{t^2},$$

and, consequently,

$$P[|X-\mu| < t\sigma] \geq 1 - \frac{1}{t^2}.$$

Example 8.1 cont.

- Therefore, for instance for $t = 2$, we get

$$P[|X - \mu| \geq 2\sigma] \leq 0.25 \quad \text{or} \quad P[|X - \mu| < 2\sigma] \geq 0.75.$$

This says that, there is at least a 75% chance that a random variable will be within 2σ of its mean (independent of the distribution of X).

Inequalities

- This information is useful. However, many times, it might be possible to obtain an even tighter bound in the following sense.

Example 8.2

Let $Z \sim N(0,1)$. Then,

$$P(Z \geq t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx$$

$$\leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-x^2/2} dx$$

$$= \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t}.$$

The second inequality follows from the fact that for $x > t$, $x/t > 1$. Now, since Z has a symmetric distribution, $P(|Z| \geq t) = 2P(Z \geq t)$. Hence,

$$P(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}.$$

Example 8.2 cont.

- Set $t = 2$ and observe that

$$P(|Z| \geq 2) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-2}}{2} = 0.054.$$

and, consequently,

$$P[|X - \mu| < t\sigma] \geq 1 - \frac{1}{t^2}.$$

This is a much tighter bound compared to the one given by Chebychev's Inequality.

- Generally Chebychev's Inequality provides a more conservative bound.

Inequalities

- A related inequality is Markov's Inequality.

Lemma 8.1

If $P(Y \geq 0) = 1$ and $P(Y = 0) < 1$, then, for any $r > 0$,

$$P(Y \geq r) \leq \frac{E[Y]}{r},$$

and the relationship holds with equality if and only if
$P(Y = r) = p = 1 - P(Y = 0)$, where $0 < p \leq 1$.

- The more general form of Chebychev's Inequality, provided in White (2001), is as follows.

White Proposition

Proposition (White 2001): Let X be a random variable such that $E[|X|^r] < \infty$, $r > 0$. Then, for every $t > 0$,

$$P(|X| \geq t) \leq \frac{E[|X|^r]}{t^r}.$$

Setting $r = 2$, and some re-arranging, gives the usual Chebychev's Inequality. If we let $r = 1$, then we obtain Markov's Inequality. See White (2001, pp.29-30).

Theorem 8.2

(Hölder's Inequality): *Let X and Y be any two random variables, and let p and q be such that*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then,

$$E[|XY|] \leq \{E[|X|^p]\}^{1/p}\{E[|Y|^q]\}^{1/q}.$$

- **Proof**: *Exercise!*

Inequalities

Theorem 8.3

(Cauchy-Schwarz Inequality): *For any two random variables X and Y,*

$$E[|XY|] \leq \{E[|X|^2]\}^{1/2}\{E[|Y|^2]\}^{1/2}.$$

- **Proof**: *Set* $p = q = 2$.

Inequalities

Theorem 8.4

(Minkowski's Inequality): *Let X and Y be any two random variables. Then, for* $1 \leq p < \infty$,

$$\{E[|X+Y|^p]\}^{1/p} \leq \{E[|X|^p]\}^{1/p} + \{E[|Y|^p]\}^{1/p}.$$

- **Proof**: *Exercise!*

Definition 8.1

A function $g(x)$ is convexif

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y), \quad \text{for all } x \text{ and } y, \text{ and } 0 < \lambda < 1.$$

The function $g(x)$ is concaveif $-g(x)$ is convex.

- Now we can introduce Jensen's Inequality.

Theorem 8.5

For any random variable X, if $g(x)$ is a convex function, then

$$E[g(X)] \geq g\{E[X]\}.$$

References

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