Econ 240A (1st Half)
Section 3: Fall 2018
Friday, September 14

Fengshi Niu*

Contents

1 Multiple Random Variables **2**

1.1 Definitions 2

1.2 Invertible Transformations 4

2 Normal and related distributions **6**

3 Conditional expectation and conditional distribution **9**

4 Inequalities **15**

*Thanks to Mikkel Sølvsten and Matias D. Cattaneo for letting me use their old notes as an inspiration.

1 Multiple Random Variables

In this Section we extend the discussion of Probability Theory to the context of multiple random variables. Fortunately, most ideas presented until now can be extended relatively easy to this more general setup. For the most part, we will concentrate the discussion to the leading case of two random variables; that is, a bivariate random vector. All ideas presented for this case generalize straightforward to the most general case. Finally, although in most cases the exposition will be restricted to the continuous random variables case, it is important to note that the discrete case is analogous provided that integration signs are replaced by summation signs.

1.1 Definitions

First we revisit and extend accordingly most of the definitions given in the previous Sections.

Definition 1.1. *An K-**dimensional random vector** is a function from a sample space Ω in to $\mathbb{R}^K$.*

From now we specialize our discussion to the case of $K = 2$ random variables or equivalently a random vector $(X, Y)' \in \mathbb{R}^2$. Fortunately, the main ideas can be discussed in the special case of $K = 2$. A more general treatment can be found in any standard textbook, and for the most part it generalizes in the obvious way.

In the next definition we extend the notion of cdf in the bivariate case.

Definition 1.2. *The **joint cumulative distribution function** (joint cdf) of the random vector $(X, Y)' \in \mathbb{R}^2$ is the function $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ given by*

$$F_{X,Y}(x, y) = \mathbb{P}[X \leq x, Y \leq y],$$

for all $(x, y)' \in \mathbb{R}^2$.

It is important to stress that the correspondence theorem generalizes to $\mathbb{R}^K$ and thus it suffices to specify $F_{X,Y}(\cdot, \cdot)$ in order to characterize the underlying joint probability function. Furthermore, there exists necessary and sufficient conditions for a function $F_{X,Y}(\cdot, \cdot)$ to be a joint cdf of some multivariate random variable.

In the next definition we turn to the notion of pmf and pdf for the discrete and continuous case respectively.

Definition 1.3. *Let $(X, Y)'$ be a random vector in $\mathbb{R}^2$ with joint cdf $F_{X,Y}(\cdot, \cdot)$, then*

1. *$(X, Y)' \in \mathbb{R}^2$ is **discrete** if there exists $f_{X,Y} : \mathbb{R}^2 \to [0, 1]$ such that*

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

*for all $(x, y)' \in \mathbb{R}^2$. The function $f_{X,Y}(\cdot, \cdot)$ is the **joint probability mass function** (joint pmf) of $(X, Y)'$.*

2. $(X,Y)' \in \mathbb{R}^2$ *is* ***continuous*** *if there exists* $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ *such that*

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

for all $(x,y)' \in \mathbb{R}^2$*. The function* $f_{X,Y}(\cdot,\cdot)$ *is a* ***joint probability density function*** *(joint pdf) of* $(X,Y)'$*.*

Observe that an immediate consequence of these definitions is that both the pmf and pdf have to sum and integrate to one respectively; that is,

1. If $(X,Y)' \in \mathbb{R}^2$ is discrete, then $f_{X,Y} : \mathbb{R}^2 \to [0,1]$ is a pmf iff

$$\sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} f_{X,Y}(s,t) = 1.$$

2. If $(X,Y)' \in \mathbb{R}^2$ is continuous, then $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ is a pdf iff

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

The definition of expectation is extended in the following way.

Definition 1.4. *If* $(X,Y)' \in \mathbb{R}^2$ *is a discrete (continuous) random vector with joint pmf (pdf)* $f_{X,Y}(\cdot,\cdot)$ *and let* $g : \mathbb{R}^2 \to \mathbb{R}$ *be a (measurable) function, then the* ***expected value*** *of* $g(x,y)$ *is*

$$\mathbb{E}[g(X,Y)] = \begin{cases} \sum_{y=-\infty}^{\infty} \sum_{x=-\infty}^{\infty} g(x,y) \cdot f_{X,Y}(x,y) & \text{if } (X,Y)' \text{ is discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \cdot f_{X,Y}(x,y) \cdot dx \cdot dy & \text{if } (X,Y)' \text{ is continuous} \end{cases}.$$

So far we have discussed the natural extension of the main definitions given previously to the case of bivariate random variables. Now we present a couple of new definitions that are not only very useful, but also follow quiet directly from the discussion given so far.

Definition 1.5. *If* $(X,Y)' \in \mathbb{R}^2$ *is a random vector, then the cdf of* X *is called the* ***marginal cdf*** *of* X*. Similarly, the pmf (pdf) of the discrete (continuous) random variable* X *is called the* ***marginal pmf (marginal pdf)*** *of* X*.*

From these definitions we obtain the following useful facts:

1. If $(X,Y)' \in \mathbb{R}^2$ has joint cdf $F_{X,Y}(\cdot,\cdot)$, then the marginal cdf of X, denoted $F_X(\cdot)$, is given by

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

for $x \in \mathbb{R}$.

2. If $(X,Y)' \in \mathbb{R}^2$ is discrete with joint pmf $f_{X,Y}(\cdot,\cdot)$, then X is discrete with pmf $f_X : \mathbb{R} \to [0,1]$, given by
$$f_X(x) = \sum\nolimits_{t=-\infty}^{\infty} f_{X,Y}(x,t),$$
for $x \in \mathbb{R}$.

3. If $(X,Y)' \in \mathbb{R}^2$ is continuous with joint pdf $f_{X,Y}(\cdot,\cdot)$, then X is continuous with pdf $f_X : \mathbb{R} \to \mathbb{R}$, given by
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,t) \cdot dt,$$
for $x \in \mathbb{R}$.

Using these results, we can see that in fact marginal distributions can be recovered from joint distributions. However, it is important to note that in general the converse is not true; that is, it is not possible to construct the joint cdf of a random vector from the marginal cdf of its components. In the following definition we give conditions under which we can recover joint distributions from marginal distributions.

Definition 1.6. *Let $(X,Y)' \in \mathbb{R}^2$ be a random vector with joint cdf $F_{X,Y}$ and marginal cdfs F_X and F_Y. Then the random variables X and Y are **independent** iff*
$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y).$$

This definition gives two very important results:

1. If $(X,Y)' \in \mathbb{R}^2$ is a discrete (continuous) random vector with pmf (pdf) $f_{X,Y}(\cdot,\cdot)$ and marginal pmfs (pdfs) $f_X(\cdot)$ and $f_Y(\cdot)$, then X and Y are **independent** if
$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y).$$

2. If $(X,Y)' \in \mathbb{R}^2$ is a random vector with cdf $F_{X,Y}$, X and Y are independent, and $V = g(X)$ and $W = h(Y)$ then
$$F_{V,W}(x,y) = F_V(x) \cdot F_W(y).$$
that is, V and W are also independent.

1.2 Invertible Transformations

In this section we briefly discuss an important formula to do change of variables for multiple random variables. As usual, we present the discussion for the particular case of $K = 2$ but this generalize in the obvious way to an arbitrary random vector.

Before presenting the main theorem for this section, we explain the basic setup. Let $\mathbf{X} = [X_1, X_2]'$ be random vector with density function $f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,X_2}(x_1,x_2)$. Consider the random variable $\mathbf{Y} = [Y_1, Y_2]' = \mathbf{g}(X_1, X_2) = [g_1(X_1, X_2), g_2(X_1, X_2)]'$, where $\mathbf{g} : \mathbb{R}^2 \to \mathbb{R}^2$ is assumed to be a continuously differentiable, one-to-one mapping. Recall from calculus that the derivative matrix is given by

$$\mathrm{D}\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_1} \\ \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} \end{bmatrix},$$

and the Jacobian is $|\mathrm{D}\mathbf{g}(\mathbf{x})|$. Since $\mathbf{g}$ is one-to-one, we know that there exists an inverse function given by $\mathbf{g}^{-1} : \mathbb{R}^2 \to \mathbb{R}^2$ and also that $|\mathrm{D}\mathbf{g}^{-1}(\mathbf{y})| = |\mathrm{D}\mathbf{g}(\mathbf{x})|^{-1}$.

Now we are in conditions to present the main theorem.

Theorem 1.1. *(*INVERTIBLE TRANSFORMATIONS*) Let $\mathbf{X} = [X_1, X_2]'$ be a continuous random variable with density function $f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,X_2}(x_1, x_2)$, and let $\mathbf{Y} = [Y_1, Y_2]' = \mathbf{g}(X_1, X_2) = \mathbf{g}(\mathbf{X})$ where $\mathbf{g} : \mathbb{R}^2 \to \mathbb{R}^2$ is assumed to be a continuously differentiable, one-to-one mapping. Then,*

$$f_{\mathbf{Y}}(\mathbf{y}) = \left| D\mathbf{g}^{-1}(\mathbf{y}) \right| \cdot f_{\mathbf{X}}\left(\mathbf{g}^{-1}(\mathbf{y})\right),$$

where $f_{\mathbf{Y}}(\mathbf{y})$ is the density function of the random vector $\mathbf{Y} = \mathbf{g}(\mathbf{X})$.

Next we present a simple example where we apply this formula.

Example 1.1. *Let $\mathbf{X} = [X_1, X_2]'$ be a continuous random vector where $X_1 \sim Exponential(1)$ and $X_2 \sim Exponential(1)$, independent. Let $\mathbf{Y} = \mathbf{g}(X_1, X_2) = [X_1 + X_2, X_1 - X_2]'$. Then, observe that*

$$\mathbf{g}^{-1}(y_1, y_2) = \left[\frac{y_1 + y_2}{2}, \frac{y_1 - y_2}{2}\right]',$$

which implies that $\left|D\mathbf{g}^{-1}(\mathbf{y})\right| = -\frac{1}{2}$. Hence, we conclude that the density of the random vector $\mathbf{Y}$ is given by

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \left| D\mathbf{g}^{-1}(\mathbf{y}) \right| \cdot f_{\mathbf{X}}\left(\mathbf{g}^{-1}(\mathbf{y})\right) \\ &= -\frac{1}{2} \cdot f_{X_1}\left(g_1^{-1}(\mathbf{y})\right) \cdot f_{X_2}\left(g_2^{-1}(\mathbf{y})\right) \\ &= -\frac{1}{2} \cdot f_{X_1}\left(\frac{y_1 + y_2}{2}\right) \cdot f_{X_2}\left(\frac{y_1 - y_2}{2}\right) \\ &= -\frac{1}{2} \cdot \exp\left\{-\frac{y_1 + y_2}{2}\right\} \cdot \mathbb{I}\left\{\frac{y_1 + y_2}{2} > 0\right\} \cdot \exp\left\{-\frac{y_1 - y_2}{2}\right\} \cdot \mathbb{I}\left\{\frac{y_1 - y_2}{2} > 0\right\} \\ &= -\frac{1}{2} \cdot \exp\{-y_1\} \cdot \mathbb{I}\{y_1 > |y_2|\}. \end{aligned}$$

Moreover, in this case, we can obtain the marginal distribution of Y_2 given by

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{\mathbf{Y}}(y_1, y_2) \cdot dy_1 = \int_{|y_2|}^{\infty} -\frac{1}{2} \cdot \exp\{-y_1\} \cdot dy_1 = \frac{1}{2} \exp\{-|y_2|\},$$

which implies that $Y_2 \sim DoubleExponential(1, 1)$.

This formula could be come very handy when dealing with otherwise difficult transformations. The following exercise ask you to deduce the convolution formula using this method.

Exercise 1.1. *(*CONVOLUTION FORMULA*) Let $\mathbf{X} = [X_1, X_2]'$ be a continuous random vector with density $f_{\mathbf{X}}(\mathbf{x})$.*

1. *Determine the joint density function of $\mathbf{Y} = [X_1, X_1 + X_2]'$.*
2. *Determine the density function of $W = X_1 + X_2$.*

The convolution formula presented in the previous exercise could have been deduce from first principles. However, this is somehow more tedious and less systematic. In general, the formula of invertible transformations is very useful to deal with all kind of problems and encompasses most of the classical results such as ratio of random variables or other kind of transformations.

2 Normal and related distributions

Here we briefly review the main results concerning some distributions generated by transformation of normal r.v.s. These results will be very helpful when doing hypothesis testing or when constructing confidence intervals for the parameter of interest. In particular, we will review the normal, chi-squared, t, and F distributions. We begin by reviewing some of the basic results of normal distributions.

Theorem 2.1. *(*NORMAL DISTRIBUTION*) Let $\mathbf{X} = (X_1, X_2, ..., X_N)'$ be a random sample from a Normal distribution, that is $X_n \sim iid\ \mathcal{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2)' \in \mathbb{R} \times \mathbb{R}_{++}$, where*

$$f_{X_n}(x_n | \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(x_n - \mu)^2}{\sigma^2} \right\}.$$

Then the following results hold:

1. *Moments:*

$$\begin{aligned} \mathbb{E}_\theta [X_n] &= \mu, \\ \mathbb{V}ar_\theta [X_n] &= \sigma^2. \end{aligned}$$

2. *Linear Combinations ($\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}_{++}$):*

$$\begin{aligned} \alpha + \beta \cdot X_n &\sim \mathcal{N}\left(\beta \cdot \mu + \alpha, \beta^2 \cdot \sigma^2\right) \Longrightarrow \frac{X_n - \mu}{\sigma} = Z_n \sim \mathcal{N}(0, 1), \\ \left(\sum\nolimits_{n=1}^N X_n\right) &\sim \mathcal{N}\left(N \cdot \mu, N \cdot \sigma^2\right). \end{aligned}$$

Now we connect the normal distribution with the chi-squared distribution.

Theorem 2.2. *(*Chi-squared Distribution*) Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Normal distribution, that is* $X_n \sim iid\ \mathcal{N}(\mu, \sigma^2)$*, with* $\theta = (\mu, \sigma^2)' \in \mathbb{R} \times \mathbb{R}_{++}$*. Then the chi-squared distribution (recall that if* $Z \sim \mathcal{N}(0,1)$*, then* $Z^2 \sim \chi_1^2$*) can be obtained as follows:*

$$\sum\nolimits_{n=1}^{N} \left(\frac{X_n - \mu}{\sigma}\right)^2 = \sum\nolimits_{n=1}^{N} Z_n^2 = W \sim \chi_N^2 = Gamma\left(\frac{N}{2}, 2\right),$$

where

$$f_W(w|\,\theta) = \frac{w^{\frac{N}{2}-1} \cdot \exp\left\{-\frac{w}{2}\right\}}{\Gamma\left(\frac{N}{2}\right) \cdot 2^{\frac{N}{2}}} \cdot \mathbb{I}(w > 0).$$

1. *Observe that* N *indexes a family of chi-squared distributions. This parameter is usually called* ***degrees of freedom****.*

2. *Moments:*

$$\begin{aligned} \mathbb{E}_\theta[W] &= N, \\ \mathbb{V}ar_\theta[W] &= 2 \cdot N. \end{aligned}$$

3. *Linear Combinations (*$k \in \mathbb{R}_{++}$*). Recall the two main properties of the Gamma distribution: Let* $\{Y_n\}_{n=1}^N \sim iid\ Gamma(\alpha_n, \beta)$*, then*

$$\begin{aligned} \left(\sum\nolimits_{n=1}^{N} Y_n\right) &\sim Gamma\left(\left(\sum\nolimits_{n=1}^{N} \alpha_n\right), \beta\right), \\ k \cdot Y_n &\sim Gamma(\alpha_n, k \cdot \beta), \end{aligned}$$

and the results follows directly for the special case of the chi-squared distribution.

The next two distributions are the really new ones for us. Observe that, for reasons that will become clear later, we only present the results for scalar random variables. In particular, we do not discuss linear combinations of this random variables (e.g., convolution or scalar multiplication).

Theorem 2.3. *(*t distribution*) Let* Z *be a random variable with standard normal distribution, and let* X *be a random variable with chi-squared distribution with* v *degrees of freedom. Furthermore, assume that* Z *and* X *are independent. Then the t-distribution can be obtained as follows:*

$$t = \frac{Z}{\sqrt{\frac{X}{v}}} \sim \mathcal{T}_v$$

that is, has a t-distribution with v *degrees of freedom where (let* $t = W$*)*

$$f_W(w) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi} \cdot \Gamma\left(\frac{v}{2}\right)} \cdot \left(1 + \frac{w^2}{v}\right)^{-\frac{v+1}{2}}$$

1. Moments:

$$\begin{aligned}\mathbb{E}_\theta[W] &= 0, \textit{ for } v > 1, \\ \mathbb{V}ar_\theta[W] &= \frac{v}{v-2}, \textit{ for } v > 2.\end{aligned}$$

The following exercise gives an important result widely used in hypothesis testing and confidence interval estimation.

Exercise 2.1. *Let* $\mathbf{X} = (X_1, X_2, ..., X_N)'$ *be a random sample from a Normal distribution, that is* $X_n \sim iid\ \mathcal{N}(\mu, \sigma^2)$, *with* $\theta = (\mu, \sigma^2)' \in \mathbb{R} \times \mathbb{R}_{++}$. *Show that*

$$\sqrt{N}\frac{\bar{X}-\mu}{S} \sim \mathcal{T}_{N-1}.$$

(Hint: check Theorem 1.2 in Section Notes 4).

Finally, we present the last distribution of interested.

Theorem 2.4. *(*F DISTRIBUTION*) Let* X *be a random variable with chi-squared distribution with* n *degrees of freedom and let* Y *be a random variable with chi-squared distribution with* m *degrees of freedom. Furthermore, assume that* X *and* Y *are independent. Then the F-distribution can be obtained as follows:*

$$F = \frac{\frac{X}{n}}{\frac{Y}{m}} \sim \mathcal{F}_{n,m},$$

that is, has a F-distribution with n *and* m *degrees of freedom where (let* $F = W$*)*

$$f_W(w) = \frac{n \cdot \Gamma\left(\frac{n+m}{2}\right)}{m \cdot \Gamma\left(\frac{n}{2}\right) \cdot \Gamma\left(\frac{m}{2}\right)} \cdot \left(\frac{n}{m}w\right)^{\frac{n}{2}-1} \cdot \left(1 + \frac{n}{m}w\right)^{-\frac{n+m}{2}} \cdot \mathbb{I}(w > 0).$$

1. Moments:

$$\begin{aligned}\mathbb{E}_\theta[W] &= \frac{m}{m-2}, \textit{ for } m > 2, \\ \mathbb{V}ar_\theta[W] &= 2 \cdot \left(\frac{m}{m-2}\right)^2 \cdot \frac{n+m-2}{n(m-4)}, \textit{ for } m > 4.\end{aligned}$$

2. Particular case:

$$(\mathcal{T}_v)^2 \sim \mathcal{F}_{1,m}.$$

Exercise 2.2. *Prove the particular case* $(\mathcal{T}_v)^2 \sim \mathcal{F}_{1,m}$.

3 Conditional expectation and conditional distribution

Finally, we turn to a new concept based on the notion of conditional probability. This is given in the next definition.

Definition 3.1. *If $(X,Y)' \in \mathbb{R}^2$ is a discrete (continuous) random vector with joint pmf (pdf) $f_{X,Y}(\cdot,\cdot)$ and marginal pmfs (pdfs) $f_X(\cdot)$ and $f_Y(\cdot)$. For any $x \in \mathbb{R}$, a* ***conditional pmf (pdf) of*** *Y* ***given*** *$X = x$, is any function $f_{Y|X}(\cdot|\, x) : \mathbb{R} \to \mathbb{R}_+$ such that*

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

for $x \in \mathbb{R}$.

It can be shown that the following remarks hold:

1. If $(X,Y)' \in \mathbb{R}^2$ is discrete and $f_X(x) > 0$, then $f_{Y|X}(\cdot|\, x)$ is unique and

$$\sum_{y=-\infty}^{\infty} f_{Y|X}(y|\, x) = 1,$$

for $x \in \mathbb{R}$.

2. If $(X,Y)' \in \mathbb{R}^2$ is continuous, then $f_{Y|X}(\cdot|\, x)$ and $f_X(\cdot)$ can be chosen such that

$$\int_{y=-\infty}^{\infty} f_{Y|X}(y|\, x) \cdot dy = 1,$$

for $x \in \mathbb{R}$.

Observe that conditional densities have a natural interpretation in the case of discrete random variables. In particular, assuming that $(X,Y)' \in \mathbb{R}^2$ is discrete and $f_X(x) > 0$, we have

$$f_{Y|X}(y|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[X = x]} = \mathbb{P}[Y = y|\, X = x].$$

However, in the case of continuous random variables, this interpretation breaks down, since

$$f_{Y|X}(y|\, x) = \mathbb{P}[Y = y|\, X = x]$$

is not well-defined (recall that if X is continuous, then $\mathbb{P}[X = x] = 0$ for all $x \in \mathbb{R}$). Consequently, defining conditional densities requires additional work based on a measure theoretical approach, which of course exceeds the scope of this class. Nevertheless, we will abstract from these technicalities and we will continue to work with conditional densities even in the case of continuous random variables. In general, we will have

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

provided that $f_X(x) > 0$.

Observe that as a direct consequence of the definition of independence of random variables, we have that if X and Y are independent then

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

that is, the conditional and marginal pmfs (pdfs) are equal. Notice that, in particular, this gives us a way of checking if two random variables are statistically independent.

The following exercise uses the concept of independence and also requires some nontrivial work with algebra.

Exercise 3.1. *Let $X \sim Poisson(\lambda)$, and $U_n \sim Uniform\,[0,1]$, $n = 1, 2, ..., N$, independent random variables. Let $M_N = \min_{1 \leq n \leq N} \{U_n\}$ denote the minimum of the N independent uniform $[0,1]$ random variables. Show that*

$$\mathbb{E}[\exp\{\lambda \cdot M_N\}] = \frac{\mathbb{P}[X \geq N]}{\mathbb{P}[X = N]}.$$

Conditional Expectation

Now we turn to work with conditional random variables. Observe that $Y|\,X = x$ is also a random variable and thus we can, in principle, compute moments. So we have analogous concepts as those presented before in the two definitions.

Definition 3.2. *Let $(X,Y)' \in \mathbb{R}^2$ be a discrete (continuous) random vector and let $g : \mathbb{R} \to \mathbb{R}$ be a (measurable) function. For any $x \in \mathbb{R}$, a **conditional expected value of** $g(y)$ **given** $X = x$ is*

$$\mathbb{E}[g(Y)|\,X = x] = \begin{cases} \sum_{y=-\infty}^{\infty} g(y) \cdot f_{Y|X}(y|\,x) & \text{if } (X,Y)' \text{ is discrete} \\ \int_{-\infty}^{\infty} g(y) \cdot f_{Y|X}(y|\,x) \cdot dy & \text{if } (X,Y)' \text{ is continuous} \end{cases}.$$

Definition 3.3. *Let $(X,Y)' \in \mathbb{R}^2$ be a random vector and let $g : \mathbb{R} \to \mathbb{R}$ be a (measurable) function. The **conditional variance of** Y **given that** $X = x$ is*

$$\mathbb{V}ar[g(Y)|\,X = x] = \begin{cases} \sum_{y=-\infty}^{\infty} (g(y) - \mathbb{E}[g(Y)|\,x])^2 \cdot f_{Y|X}(y|\,x) & \text{if } (X,Y)' \text{ is discrete} \\ \int_{-\infty}^{\infty} (g(y) - \mathbb{E}[g(Y)|\,x])^2 \cdot f_{Y|X}(y|\,x) \cdot dy & \text{if } (X,Y)' \text{ is continuous} \end{cases}.$$

The previous definitions do not use the formal (measure-theoretic) definition of conditional expectation. Although this formal definition is beyond the scope of this class, it is important to highlight that the object $\mathbb{E}[Y|\,X]$ exists under some conditions in general and it is in fact a random variable. It is important to notice that these expressions could be misleading because of the notation. After we compute these moments, what we obtain are functions that depend on *only one* random variable: X. Note that these functions do not depend on Y. Moreover, two important and very useful implications of these remarks are presented in the following exercise:

Exercise 3.2. *(*PROPERTIES OF CONDITIONAL EXPECTATION*)Let $(X,Y)' \in \mathbb{R}^2$ be a random vector and let $h, g : \mathbb{R} \to \mathbb{R}$. Show that:*

1. $\mathbb{E}[h(X) \cdot g(Y)|\, X] = h(X) \cdot \mathbb{E}[g(Y)|\, X]$.
2. $\mathbb{V}ar[h(X) \cdot g(Y)|\, X] = (h(X))^2 \cdot \mathbb{V}ar[g(Y)|\, X]$.

The next theorem is one of the most important and widely used theorem in both probability and statistics.

Theorem 3.1. *(*LAW OF ITERATED EXPECTATIONS*) Let $(X,Y)' \in \mathbb{R}^2$ be a random vector and suppose that the corresponding moments exist. Then $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|\, X]]$.*

Proof. We proof this result for the continuous case. The discrete case is analogous. Observe that

$$\begin{aligned}
\mathbb{E}[Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \cdot f_{X,Y}(x,y) \cdot dx \cdot dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \cdot f(y|\, x) f_X(x) \cdot dx \cdot dy \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} y \cdot f(y|\, x)\, dy\right] \cdot f_X(x) \cdot dx \\
&= \int_{-\infty}^{\infty} [\mathbb{E}[Y|\, x]] \cdot f_X(x) \cdot dx \\
&= \mathbb{E}[\mathbb{E}[Y|\, X]],
\end{aligned}$$

which concludes the proof. □

This result is very useful when correctly applied. In general, the previous theorem works only when both sides agree. A typical error in applications of this result is to try compute a marginal expectation that does not exists by using this result. The following exercise ask you to provide one such example.

Exercise 3.3. *(*FAILURE OF LIE*) Give an example where $\mathbb{E}[Y]$ does not exists but $\mathbb{E}[\mathbb{E}[Y|\, X]]$ is well-defined.*

The next theorem provides a very interesting result, which heavily relies on the Law of Iterated Expectations. This result is used in estimation theory and will appear again in this course.

Theorem 3.2. *(*MEAN SQUARED ERROR PREDICTION*) Let $(X,Y)' \in \mathbb{R}^2$ be a random vector and let $\mathcal{G}$ be the space of all (measurable squared integrable) functions of X. Consider the problem of predicting Y with a function $g(\cdot) \in \mathcal{G}$. Assuming a quadratic loss function (denoted by Mean Squared Error, MSE), the best predictor is given by:*

$$\mathbb{E}[Y|\, X] = \arg\min_{g(\cdot)\in\mathcal{G}} \mathbb{MSE}[g(X)] \equiv \arg\min_{g(\cdot)\in\mathcal{G}} \mathbb{E}\left[(Y - g(X))^2\right],$$

with an associated loss

$$\mathbb{MSE}\left[g^*\left(X\right)\right]=\mathbb{E}\left[\left(Y-g^*\left(X\right)\right)^2\right]=\mathbb{E}\left[\mathbb{V}ar\left[Y|\,X\right]\right].$$

Proof. This theorem follows by using the law of iterated expectations. Observe that the problem is

$$\min_{g(\cdot)\in\mathcal{G}}\mathbb{MSE}\left[g\left(X\right)\right]\equiv\min_{g(\cdot)\in\mathcal{G}}\mathbb{E}\left[\left(Y-g\left(X\right)\right)^2\right]$$

Notice that in this case we cannot use calculus (why?). However, we can solve this highly dimensional problem by just using the "plugging in the solution" technique. We have

$$\begin{aligned}\mathbb{MSE}\left[g\left(X\right)\right] &\equiv \mathbb{E}\left[\left(Y-g\left(X\right)\right)^2\right]\\ &= \mathbb{E}\left[\left(\left[Y-\mathbb{E}\left[Y|\,X\right]\right]+\left[\mathbb{E}\left[Y|\,X\right]-g\left(X\right)\right]\right)^2\right]\\ &= \mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|\,X\right]\right)^2\right]+\mathbb{E}\left[\left(\mathbb{E}\left[Y|\,X\right]-g\left(X\right)\right)^2\right]+2\cdot\mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|\,X\right]\right)\cdot\left(\mathbb{E}\left[Y|\,X\right]-g\left(X\right)\right)\right],\end{aligned}$$

but for the third term we have (using the law of iterated expectations)

$$\mathbb{E}\left[\mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|\,X\right]\right)\cdot\left(\mathbb{E}\left[Y|\,X\right]-g\left(X\right)\right)\Big|\,X\right]\right]=\mathbb{E}\left[\left(\mathbb{E}\left[Y|\,X\right]-g\left(X\right)\right)\cdot\underbrace{\mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|\,X\right]\right)|\,X\right]}_{=0}\right],$$

and therefore the third term drops out. Consequently, we have

$$\begin{aligned}\min_{g(\cdot)\in\mathcal{G}}\mathbb{MSE}\left[g\left(X\right)\right] &\equiv \min_{g(\cdot)\in\mathcal{G}}\mathbb{E}\left[\left(Y-g\left(X\right)\right)^2\right]\\ &= \min_{g(\cdot)\in\mathcal{G}}\left\{\mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|\,X\right]\right)^2\right]+\mathbb{E}\left[\left(\mathbb{E}\left[Y|\,X\right]-g\left(X\right)\right)^2\right]\right\}\\ &= \mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|\,X\right]\right)^2\right]+\min_{g(\cdot)\in\mathcal{G}}\mathbb{E}\left[\left(\mathbb{E}\left[Y|\,X\right]-g\left(X\right)\right)^2\right],\end{aligned}$$

and it follows that the solution is (why?)

$$g^*\left(X\right)=\mathbb{E}\left[Y|\,X\right].$$

Moreover, we see that

$$\begin{aligned}\mathbb{MSE}\left[g^*\left(X\right)\right] &= \mathbb{E}\left[\left(Y-g^*\left(X\right)\right)^2\right]\\ &= \mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|\,X\right]\right)^2\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|\,X\right]\right)^2\Big|\,X\right]\right]\\ &= \mathbb{E}\left[\mathbb{V}ar\left[Y|\,X\right]\right],\end{aligned}$$

which concludes the proof. □

In the next theorem we present three results relating conditional and unconditional variances. The following theorem provides two of the most important results for conditional expectations and conditional variance.

Theorem 3.3. *Let $(X, Y)' \in \mathbb{R}^2$ be a random vector and suppose that the corresponding moments exists. Then:*

1. $\mathbb{V}ar\left[Y|\, X\right] = \mathbb{E}\left[Y^2|\, X\right] - \left(\mathbb{E}\left[Y|\, X\right]\right)^2$.

2. $\mathbb{V}ar\left[Y\right] = \mathbb{E}\left[\mathbb{V}ar\left[Y|\, X\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[Y|\, X\right]\right]$.

3. $\mathbb{V}ar\left[\mathbb{E}\left[Y|\, X\right]\right] \leq \mathbb{V}ar\left[Y\right]$.

Proof. To see (1), note that

$$\begin{aligned}
\mathbb{V}ar\left[Y|\, X\right] &= \int_{-\infty}^{\infty} \left(y - \mathbb{E}\left[Y|\, X\right]\right)^2 \cdot f_{Y|X}\left(y|\, X\right) \cdot dy \\
&= \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y|\, X\right]\right)^2 \middle|\, X\right] \\
&= \mathbb{E}\left[Y^2 \middle|\, X\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y|\, X\right]\right)^2 \middle|\, X\right] - 2 \cdot \mathbb{E}\left[Y \cdot \mathbb{E}\left[Y|\, X\right]|\, X\right] \\
&= \mathbb{E}\left[Y^2 \middle|\, X\right] + \left(\mathbb{E}\left[Y|\, X\right]\right)^2 - 2 \cdot \mathbb{E}\left[Y|\, X\right] \cdot \mathbb{E}\left[Y|\, X\right] \\
&= \mathbb{E}\left[Y^2 \middle|\, X\right] - \left(\mathbb{E}\left[Y|\, X\right]\right)^2.
\end{aligned}$$

To see (2), note that (using the results in MSE prediction Theorem)

$$\begin{aligned}
\mathbb{E}\left[\left(Y - g\left(X\right)\right)^2\right] &= \mathbb{E}\left[\left(\left[Y - \mathbb{E}\left[Y|\, X\right]\right] + \left[\mathbb{E}\left[Y|\, X\right] - g\left(X\right)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y|\, X\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y|\, X\right] - g\left(X\right)\right)^2\right],
\end{aligned}$$

and letting $g\left(X\right) = \mathbb{E}\left[Y\right]$ we obtain (using LIE)

$$\begin{aligned}
\mathbb{V}ar\left[Y\right] &\equiv \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y|\, X\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y|\, X\right] - \mathbb{E}\left[Y\right]\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(Y - \mathbb{E}\left[Y|\, X\right]\right)^2 \middle|\, X\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[Y|\, X\right]\right] \\
&= \mathbb{E}\left[\mathbb{V}ar\left[Y|\, X\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[Y|\, X\right]\right].
\end{aligned}$$

To see (3), note that $\mathbb{V}ar\left[Y|\, X\right] \geq 0$ (a.s.) and so is $\mathbb{E}\left[\mathbb{V}ar\left[Y|\, X\right]\right]$. Thus,

$$\mathbb{V}ar\left[Y\right] \geq \mathbb{V}ar\left[\mathbb{E}\left[Y|\, X\right]\right].$$

which concludes the proof. □

In the next exercise you can try these and some other related definitions and concepts.

Exercise 3.4. *(*PREDICTION*) Let (Y, X) be a bivariate random vector with distributions as follows:*

$$X \sim Gamma\,(\alpha, \beta) \qquad and \qquad Y|\,X = x \sim Poisson\,(x)$$

1. *Find the best predictor of Y based on X. Is it equal to the best linear predictor?*

2. *Find the mean squared error of the best predictor of Y based on X.*

3. *Determine the density of Y.*

4. *Determine the conditional density of X given $Y = y$.*

5. *Find the best predictor of X based on Y. Is it equal to the best linear predictor?*

A specially useful and well-known measure of linear association between two variables is called covariance of X and Y. Also, sometimes an unit free notion is constructed using the covariance of two variables. This alternative measure is called correlation of X and Y. These two concepts are defined as follows.

Definition 3.4. *Let $(X, Y)' \in \mathbb{R}^2$ be a random vector. Then the **covariance of** X **and** Y is defined as*

$$\mathbb{C}ov\,[X, Y] = \mathbb{E}\,[(X - \mathbb{E}\,[X]) \cdot (Y - \mathbb{E}\,[Y])]\,.$$

*Also, the **correlation of** X **and** Y is defined as*

$$corr\,[X, Y] \equiv \rho_{XY} = \frac{\mathbb{C}ov\,[X, Y]}{\sqrt{\mathbb{V}ar\,[X]} \cdot \sqrt{\mathbb{V}ar\,[Y]}}.$$

In the following Theorem we list a number of useful properties of this measures of linear association.

Theorem 3.4. *Let $(X, Y)' \in \mathbb{R}^2$ be a random vector and assume that the necessary moments exist. Then for any constants $a, c \in \mathbb{R}$ and $b, d \in \mathbb{R}_{++}$ we have:*

1. $\mathbb{C}ov\,[X, Y] = \mathbb{E}\,[X \cdot Y] - \mathbb{E}\,[X] \cdot \mathbb{E}\,[Y]$.

2. *If $X = Y$ (a.s), then $\mathbb{C}ov\,[X, Y] = \mathbb{V}ar\,[X]$.*

3. *If X and Y are independent random variables, then $\mathbb{C}ov\,[X, Y] = 0$.*

4. $\mathbb{C}ov\,[a + b \cdot X, c + d \cdot Y] = b \cdot d \cdot \mathbb{C}ov\,[X, Y]$.

5. $\mathbb{V}ar\,[a \cdot X + b \cdot Y] = a^2 \cdot \mathbb{V}ar\,[X] + b^2 \cdot \mathbb{V}ar\,[Y] + 2 \cdot a \cdot b \cdot \mathbb{C}ov\,[X, Y]$.

6. $|\rho_{XY}| \leq 1$.

It is extremely important to note that two random variables could have a strong relationship (even a perfect or determinist relationship) and still we could have that $\mathbb{C}ov[X, Y] = 0$. This is true because the covariance is only a measure of *linear association*, and of course the relationship between two random variables could be nonlinear.

Exercise 3.5. *Give an example where two random variables, X and Y, are dependent but with $\mathbb{C}ov[X, Y] = 0$.*

4 Inequalities

In this section we discuss two inequalities that will be very useful later in this class. They are also widely use in theoretical developments. First, the following exercise gives a useful ingredient for the main proof.

Exercise 4.1. *(YOUNG'S INEQUALITY) Let $a, b, p, q \in \mathbb{R}_{++}$ and assume that $\frac{1}{p} + \frac{1}{q} = 1$. Then,*

$$a \cdot b \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Theorem 4.1. *(HÖLDER'S INEQUALITY) Let $(X, Y)' \in \mathbb{R}^2$ be a random vector, assume that the necessary moments exist, and let $p, q \in [1, \infty]$ such that*

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

Then

$$\|X \cdot Y\|_1 = \mathbb{E}[|X \cdot Y|] \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} \cdot (\mathbb{E}[|Y|^q])^{\frac{1}{q}} = \|X\|_p \cdot \|Y\|_q.$$

Proof. First, observe that if $X \cdot Y = 0$ (a.s.), then the result is immediate. Also, if. $p = 1$ and $q = \infty$, then it can be argued that

$$\|X \cdot Y\|_1 \leq \|X\|_1 \cdot \|Y\|_\infty,$$

using some results and definitions from functional analysis. Finally, if $p, q \in (1, \infty)$ and $X, Y > 0$ (as), then we have using Young's inequality that

$$\frac{|X \cdot Y|}{\|X\|_p \cdot \|Y\|_q} = \frac{|X|}{\|X\|_p} \cdot \frac{|Y|}{\|Y\|_q} \leq \frac{1}{p} \cdot \left(\frac{|X|}{\|X\|_p}\right)^p + \frac{1}{q} \cdot \left(\frac{|Y|}{\|Y\|_q}\right)^q,$$

which implies that using the properties of the expectation that

$$\frac{\|X \cdot Y\|_1}{\|X\|_p \cdot \|Y\|_q} = \frac{\mathbb{E}[|X \cdot Y|]}{\|X\|_p \cdot \|Y\|_q} \leq \frac{1}{p} \cdot \frac{\mathbb{E}[|X|^p]}{\left(\|X\|_p\right)^p} + \frac{1}{q} \cdot \frac{\mathbb{E}[|Y|^q]}{\left(\|Y\|_q\right)^q} = \frac{1}{p} + \frac{1}{q} = 1,$$

giving the desired result. ∎

Now we use this inequality to derive two of the most useful inequality in statistics.

Exercise 4.2. *(*Cauchy-Schwarz Inequality*) Let* $(X, Y)' \in \mathbb{R}^2$ *be a random vector and assume that the necessary moments exist. Show that*

$$|\mathbb{E}[X \cdot Y]| \leq \mathbb{E}[|X \cdot Y|] \leq \left(\mathbb{E}\left[|X|^2\right]\right)^{\frac{1}{2}} \cdot \left(\mathbb{E}\left[|Y|^2\right]\right)^{\frac{1}{2}}.$$

Exercise 4.3. *(*Covariance Inequality*) Let* $(X, Y)' \in \mathbb{R}^2$ *be a random vector and assume that the necessary moments exist. Show that*

$$(\mathbb{C}ov[X, Y])^2 \leq \mathbb{V}ar[X] \cdot \mathbb{V}ar[Y].$$