Multivariate Distributions

PSE 4.1-4.2

So far, we have considered the case with a single (or "univariate" or "scalar") random variable. However, we will often be interested in how multiple random variables are related to each other; for example, a person's education and their wages. In this section, we will extend our previous discussion to multivariate random variables (or "random vectors"). Some of this will be mild extensions to what we have already done, but there are also some new issues that arise in this case. These notes will also mostly focus on the case with "bivariate" (i.e., two) random variables. The extension to three or more random variables is straightforward.

Definition. A (k-dimensional) **random vector** is a function from the sample space to $\mathbb{R}^k$.

Thus, bivariate random variables are a function from the sample space to $\mathbb{R}^2$. We will typically represent these using upper case letters like (X, Y) or (X_1, X_2). As before, we will use lower case letters such as (x, y) or (x_1, x_2) to indicate particular values that they could take.

Example: Recall the sample space of flipping a coin three times is $S = \{HHH, HHT, HTH, ...\}$. We can define the bivariate random variables X to be the number of heads on the first two rolls, and Y to be the number of heads on the last two rolls. These are both random variables; for example $X(\{HHH\}) = 2$ and $Y(\{HHH\}) = 2$.

cdfs, pmfs, and pdfs

PSE 4.3-4.4

We can define cdfs, pmfs, and pdfs for random vectors in essentially analogous ways to the case of a single random variable. All of these fully describe the joint distribution of a random vector.

Definition. The **joint cdf** of two random variables X and Y is the function

$$\mathrm{F}_{XY}(x, y) := \mathrm{P}(X \leq x, Y \leq y)$$

Joint cdfs satisfy the same sorts of properties as in the univariate case: (i) $\mathrm{F}_{XY}(x, y)$ is weakly increasing in both of its arguments, and (ii) $0 \leq \mathrm{F}_{XY}(x, y) \leq 1$ for all possible values of x and y.

Definition. If X and Y are both discrete random variables, then we can define the **joint probability mass function** $\pi(x, y) := \mathrm{P}(X = x, Y = y)$

Joint pmfs satisfy similar properties to the pmf of a scalar random variable. In particular,

- $\pi(x, y) \geq 0$ for all x and y. In words: pmfs are positive

- $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \pi(x, y) = 1$. In words: the sum of the pmf across all possible values of x and y is equal to 1.

Definition. If X and Y are both continuous random variables, then we can define the **joint probability density function** (and given that $\mathrm{F}_{XY}(x, y)$ is differentiable), $f_{XY}(x, y) := \frac{\partial^2}{\partial x \, \partial y} \mathrm{F}_{XY}(x, y)$

Joint pdfs also satisfy similar properties to the pdf of a scalar random variable. In particular,

- $f_{XY}(x, y) \geq 0$ for all possible x and y. In words: pdfs are positive.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \, dx \, dy = 1$. In words: the integral of the pdf across all possible values of X and Y is equal to one.
- $\int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) \, du \, dv = \mathrm{F}_{XY}(x, y)$. In words: Given the pdf, one can recover the cdf by integrating the pdf across all values of X and Y that are less than the particular values x and y.

Marginal distribution

PSE 4.6

One point that is worth clarifying is that if we know the joint distribution of X and Y (e.g., $F_{XY}(x, y)$), this is more information than knowing just the distribution of X and/or Y by themselves (i.e., $F_X(x)$ and/or $F_Y(y)$) — it is common to refer to these latter distributions as **marginal distributions**. In particular, notice that

$$\mathrm{F}_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(X \leq x, Y \leq \infty)$$

The above expression implies that, if we know the joint distribution of X and Y, then we can recover the marginal distribution of X. A symmetric argument would also imply that we could also recover the marginal distribution of Y.

It is also possible to recover marginal pmfs and pdfs from a joint pmf/pdf. In particular,

$$\begin{aligned} \pi_X(x) &= \sum_{y \in \mathcal{Y}} \pi_{XY}(x, y) \quad \text{discrete case} \\ f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y) \, dy \quad \text{continuous case} \end{aligned}$$

These steps are referred to as "integrating out" Y. It is worth briefly explaining this result, where I'll focus on the continuous case. We have that

$$f_X(x) = \frac{d}{dx} \mathrm{F}_X(x) = \frac{d}{dx} \mathrm{F}_{XY}(x, \infty) = \frac{d}{dx} \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_{XY}(u, v) \, du \, dv = \int_{-\infty}^{\infty} f_{XY}(x, v) \, dv$$

Here the last equality passes the derivative through the first integral; the limits of integration of the inner integral depend on x though, but from Leibniz's rule, we have that $\frac{d}{dx}\int_{-\infty}^{x} f_{XY}(u,v)\,du = f_{XY}(x,v)$.

Expectation

PSE 4.7

For some function $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}[g(X,Y)] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g(x,y)\pi(x,y) \quad \text{discrete case} \\
\mathbb{E}[g(X,Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\,dx\,dy \quad \text{continuous case}
\end{aligned}$$

There is a very useful property related to marginal distributions and expectations that we have just been discussing concerning the **expected value of sums of random variables**. In particular,

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

In other words, expectations can pass through sums.

Proof:

$$\begin{aligned}
\mathbb{E}[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f_{XY}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{XY}(x,y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{XY}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} x \underbrace{\left(\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy\right)}_{=f_X(x)} dx + \int_{-\infty}^{\infty} y \underbrace{\left(\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\right)}_{=f_Y(y)} dy \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

Conditional distributions

PSE 4.8-4.9

Often, we will be interested in distributions of some random variable conditional on another random variable taking some particular value. Typically, we refer to the first variable as the "outcome" and denote it by Y; the second variable is given different names depending on the context but typically is denoted by X. For example, we might be interested in the wage distribution separately for men and women.

To start with, consider the case where X is discrete. In this case, the **conditional cdf** of Y given $X = x$ is given by

$$\mathrm{F}_{Y|X}(y|x) := \mathrm{P}(Y \leq y|X = x)$$

It is helpful to view this as a function of y and x; that is, if you change either of these, it changes the value of the cdf. For example, this allows for the fraction of people with wages less than \$15 per hour to be different for men and women. In the case where X is continuous, we must be a bit more careful conditioning on $X = x$ (because $\mathrm{P}(X = x) = 0$). In this case, we define the conditional distribution function as

$$\mathrm{F}_{Y|X}(y|x) := \lim_{\epsilon \downarrow 0} \mathrm{P}(Y \leq y|x - \epsilon \leq X \leq x + \epsilon)$$

which is the probability that Y is less than y conditional on X being in an arbitrarily small neighborhood of x.

We can also define conditional pmf/pdfs:

Case 1: X is discrete

- $\pi_{Y|X}(y|x) := \mathrm{P}(Y = y|X = x)$ when Y is discrete
- $f_{Y|X}(y|x) := \frac{\partial}{\partial y}\mathrm{F}_{Y|X}(y|x)$ when Y is continuous

  These are both exactly what you would expect. In particular, in the continuous case, just like for the unconditional pdf we studied earlier, to obtain the pdf we take the derivative of the cdf.

Case 2: X is continuous

- $\pi_{Y|X}(y|x) := \lim_{\epsilon \downarrow 0} \mathrm{P}(Y = y|x - \epsilon \leq X \leq x + \epsilon)$ when Y is discrete
- $f_{Y|X}(y|x) := \frac{\partial}{\partial y}\mathrm{F}_{Y|X}(y|x) = \frac{f_{YX}(y,x)}{f_X(x)}$ when Y is continuous

  As you would expect, the conditional pmf amounts to calculating the probability that Y takes the particular value y given that X falls in a small neighborhood of x. In the continuous case, as before we get the conditional pdf by taking the derivative of the conditional cdf. The second equality is not completely obvious and it is worth spending a moment practicing working with conditional cdfs/pmfs. We'll show the second equality over the next two results.

Proposition: If $F_{XY}(x,y)$ is differentiable with respect to x and $f_X(x) > 0$, then $F_{Y|X}(y|x) = \frac{\frac{\partial}{\partial x}F_{XY}(x,y)}{f_X(x)}$.

Proof:

$$\begin{aligned}
\mathrm{F}_{Y|X}(y|x) &= \lim_{\epsilon \downarrow 0} \mathrm{P}(Y \le y | x - \epsilon \le X \le x + \epsilon) \\
&= \lim_{\epsilon \downarrow 0} \frac{\mathrm{P}(Y \le y, x - \epsilon \le X \le x + \epsilon)}{\mathrm{P}(x - \epsilon \le X \le x + \epsilon)} \\
&= \left(\lim_{\epsilon \downarrow 0} \frac{F_{XY}(x + \epsilon, y) - F_{XY}(x - \epsilon, y)}{\epsilon} \right) \left(\lim_{\epsilon \downarrow 0} \frac{\epsilon}{F_X(x + \epsilon) - F_X(x - \epsilon)} \right) \\
&= \left(\frac{\partial}{\partial x} F_{XY}(x, y) \right) \left(\frac{1}{f_X(x)} \right) = \frac{\frac{\partial}{\partial x} F_{XY}(x, y)}{f_X(x)}
\end{aligned}$$

where the first equality holds by the definition of conditional cdf, the second equality holds by the definition of conditional probability, the third equality holds by (i) re-writing the numerator and denominator in terms of the joint cdf, (ii) by multiplying and dividing by ϵ, and (iii) because the limit of a product is equal to the product of the limits, and the last line holds by re-writing the limit terms as derivatives.

The previous proposition showed that the conditional cdf can be written in terms of the joint cdf and marginal pdf. Next, we provide a similar result for the conditional pdf.

Proposition: For continuous X and Y, the conditional pdf of Y given $X = x$ is $f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$ for any x such that $f_X(x) > 0$.

Proof:

$$\begin{aligned}
f_{Y|X}(y|x) &= \frac{\partial}{\partial y} F_{Y|X}(y|x) \\
&= \frac{\partial}{\partial y} \frac{\frac{\partial}{\partial x} F_{XY}(x, y)}{f_X(x)} \\
&= \frac{\frac{\partial^2}{\partial x \partial y} F_{XY}(x, y)}{f_X(x)} \\
&= \frac{f_{X,Y}(x, y)}{f_X(x)}
\end{aligned}$$

where the first equality uses the definition of conditional pdf, the second equality uses the result from the previous proposition, the third equality holds immediately, and the last equality holds by the definition of the joint pdf.

Independence

PSE 4.11

Recall that two events A and B are said to be independent if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. We can extend this definition to random variables. In particular, two random variables X and Y are said to be **statistically independent** if, for all x and y, $F_{XY}(x,y) = F_X(x)F_Y(y)$. We will often use the notation $X \perp\!\!\!\perp Y$ to indicate two statistically independent random variables.

This definition immediately implies that, if X and Y are independent, then $f_{XY}(x,y) = f_X(x)f_Y(y)$ (just take derivatives of the cdfs in the definition). In addition, it also (essentially) immediately implies that $f_{Y|X}(y|x) = f_Y(y)$. Often, I find this last expression as the most natural way to think about independence — it says that, if you know the value that the random variable X takes, it does not provide any information about the distribution of Y (or the value that Y is more or less likely to take).

There are a few useful properties of independent random variables that are worth mentioning. First, if X and Y are independent, then for any functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ (and assuming the moments exist), $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$.

Proof:

$$\begin{aligned}
\mathbb{E}[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} g(x)f_X(x)\,dx \int_{-\infty}^{\infty} h(y)f_Y(y)\,dy \\
&= \mathbb{E}[g(X)]\mathbb{E}[h(Y)]
\end{aligned}$$

Interestingly, the same argument works in reverse. In particular, if for all functions g and h, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)h(Y)]$, it implies that $F_{XY}(x,y) = F_X(x)F_Y(y)$ for all possible values of x and y. To see this, take $g(x) = \mathbb{1}\{x \leq a\}$ and $h(y) = \mathbb{1}\{y \leq b\}$ for constants a and b and recall that $\mathrm{P}(X \leq a, Y \leq b) = \mathbb{E}[\mathbb{1}\{X \leq a\}\mathbb{1}\{Y \leq b\}]$ and $\mathrm{P}(X \leq a) = \mathbb{E}[\mathbb{1}\{X \leq a\}]$. This can be a useful way to show that two random variables are independent of each other.

Side-Comment: The previous argument used that $\mathrm{P}(X \leq a) = \mathbb{E}[\mathbb{1}\{X \leq a\}]$. To see this, notice that

$$\begin{aligned}
\mathbb{E}[\mathbb{1}\{X \leq a\}] &= \int_{-\infty}^{\infty} \mathbb{1}\{x \leq a\}f_X(x)\,dx \\
&= \int_{-\infty}^{a} \underbrace{\mathbb{1}\{x \leq a\}}_{=1 \text{ in this region}} f_X(x)\,dx + \int_{a}^{\infty} \underbrace{\mathbb{1}\{x \leq a\}}_{=0 \text{ in this region}} f_X(x)\,dx \\
&= \int_{-\infty}^{a} f_X(x)\,dx \\
&= \mathrm{P}(X \leq a)
\end{aligned}$$

where the third equality holds because $\mathbb{1}\{x \leq a\} = 1$ for values of x below a, and it equals 0 for values of x above a.

Another useful property concerns moment generating functions. If X and Y are independent with mgfs $M_X(t)$ and $M_Y(t)$, then the mgf of $Z = X + Y$ is $M_Z(t) = M_X(t)M_Y(t)$. The proof of this result is not complicated, but I'll leave it as an exercise for you. This suggests that moment generating functions can be very useful for working with sums of independent random variables; we'll use it to prove the central limit theorem soon. This is also a useful result in econometrics research on measurement error and panel data which can sometimes be translated into problems involving sums of independent random variables.

A distinct concept from independence but one that is worth mentioning here is the following. Two random variables X and Y are said to be **identically distributed** if $F_X(x) = F_Y(x)$ for all x.

Covariance and correlation

PSE 4.12

The **covariance** between two random variables X and Y is a summary measure of how they "co-move". It is defined as

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

This is a natural definition of covariance. It is also helpful to note that it can be rewritten as $\mathrm{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ which is often more useful in calculations. Covariance shows up naturally in many expressions in statistics, but like variance, it can be hard to interpret a particular value of the covariance. The issue, again similar to variance, is that the units are unusual — they are in the units of X times the units of Y.

The **correlation** between X and Y is defined as

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{s.d.}(X)\mathrm{s.d.}(Y)}$$

Correlations are bounded between -1 and 1 and do not have units. If $\mathrm{cov}(X, Y) = 0$ then $\mathrm{corr}(X, Y) = 0$, and X and Y are said to be **uncorrelated**. Two independent random variables will also be uncorrelated (you can show this using the definition of covariance above), but two uncorrelated random variables will not necessarily be independent; the textbook provides an example of this is PSE 4.12.

Another useful result is that, for two random variables X and Y,

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X, Y)$$

Thus, the variance of the sum of random variables is related to the sum of their variances, but it also depends on their covariance. For an example, suppose that X and Y are outcomes of two different dice rolls. If they are independent, then the variance of the sum of the dice rolls is just equal to the sum of the variances. If they are somehow dependent (for simplicity: consider an extreme case where the roll of one die fully determines the outcome of the roll of the other die; that is, there is a mechanism so that $X = Y$). In this case, the variance of the some of the dice rolls will include an additional (in our example) positive covariance term, making the variance of the sum higher. A rough intuition is that, if the dice rolls are correlated, the chances of “extreme” sums (like 12) is more likely than in the case with independent rolls, which increases the variance.

Cauchy-Schwarz

PSE 4.13

One useful inequality (that I’ll just state without proof) with mulitple random variables is the **Cauchy-Schwarz** inequality: For any random variables X and Y, $\mathbb{E}|XY| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$.

A useful implication of the Cauchy-Schwarz inequality is that, if $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$ (i.e, the second moments of X and Y exist), then $\text{cov}(X, Y)$ also exists. To see this, recall that $\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. The Cauchy-Schwarz inequality implies that $\mathbb{E}[XY]$ exists, and we showed previously (as an implication of Jensen’s inequality) that higher order moments existing implies that lower order moments (like $\mathbb{E}[X]$ and $\mathbb{E}[Y]$) also exist.

Conditional expectation

Next, we’ll introduce what is perhaps the central object of interest in first-year econometrics, the conditional expectation. The **conditional expectation** of Y given $X = x$ is the expected value of the conditional distribution $F_{Y|X}(y|x)$. It is given by

$$\begin{aligned}
\mathbb{E}[Y|X = x] &= \sum_{y \in \mathcal{Y}} y f_{Y|X}(y|x) \quad \text{when } Y \text{ is discrete} \\
\mathbb{E}[Y|X = x] &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x) \quad \text{when } Y \text{ is continuous}
\end{aligned}$$

Observe that the conditional expectation $\mathbb{E}[Y|X = x]$ depends on x only, as we’ve averaged over various values of Y. Accordingly, we can define a function that evaluates $\mathbb{E}[Y|X = x]$ over different values of x:

Definition. The conditional expectation function (CEF) of Y given X is $m(x) := \mathbb{E}[Y|X = x]$.

This is rightly viewed as a function of x. For example, the average wages of men and women could be different.

Alternatively, we can also use the CEF to define a new random variable, denoted $\mathbb{E}[Y|X]$.

Definition. $\mathbb{E}[Y|X] = m(X)$, where $m(x) := \mathbb{E}[Y|X = x]$.

Written this way, $\mathbb{E}[Y|X]$ should itself be viewed as a random variable; that is, as a transformation of the random variable X. For example, if X is discrete, then $\mathbb{E}[Y|X]$ takes value $m(x_j) = \mathbb{E}[Y|X = x_j]$ with probability π_j.

Relative to, say, covariance, one of the most useful features of conditional expectations is that they easily generalize to the case where X is a vector. Just to be explicit about this, I'll momentarily write X_1, X_2, X_3, etc. Then, $m(x_1, x_2, x_3) := \mathbb{E}[Y|X_1 = x_1, X_2 = x_2, X_3 = x_3]$ is the expectation of Y conditional on X_1, X_2, and X_3 taking on the particular values x_1, x_2, and x_3. In economics, we are often interested in what are called **partial effects**. The partial effect of X_1 on Y is defined as:

$$PE_1(x_1, x_2, x_3) := \frac{\partial \mathbb{E}[Y|X_1 = x_1, X_2 = x_2, X_3 = x_3]}{\partial x_1} \quad \text{if } X_1 \text{ continuous}$$

$$PE_1(x_1, x_2, x_3) := \mathbb{E}[Y|X_1 = x_1, X_2 = x_2, X_3 = x_3] - \mathbb{E}[Y|X_1 = (x_1 - 1), X_2 = x_2, X_3 = x_3] \quad \text{if } X_1 \text{ discrete}$$

The partial effect of X_1 on Y should be interpreted as how much the outcome (Y) changes, on average, when X_1 increases by 1 unit, holding X_2 and X_3 fixed. Notice that it is a function of X_1, X_2, and X_3 (i.e., the partial effect depends on the values of the other X's as well as the value of X_1). An example of a partial effect would be how much wages increase on average when age increases by year holding a person's occupation constant; notice that the average effect of being a year older can depend on the particular age (it might be different going from 21 to 22 than from 61 to 62) as well as the particular occupation (e.g., the partial effect of age could be different for a professional baseball player relative to a teacher).

Law of iterated expectations

PSE 4.15

When we treat $\mathbb{E}[Y|X]$ as a random variable, it allows us to entertain some interesting ideas. For one, we can think of the expectation of $\mathbb{E}[Y|X]$. In fact, one of the most famous and useful results for working with conditional expectations, comes from this idea.

Proposition: The law of iterated expectations $\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right]$. In other words, the expectation value of $\mathbb{E}[Y|X]$ (where the expectation is over the distribution of X) recovers the (unconditional) expectation of Y.

Proof: We prove it for the case in which both X and Y are continuous random variables. The other cases are analogous. Let $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$ (just to make sure conditional expectations are well-defined); then

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[Y|X]\right] &= \int_{\mathcal{X}} f_X(x)\,\mathbb{E}[Y|X=x]\,dx \\
&= \int_{\mathcal{X}} f_X(x)\left\{\int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\,dy\right\} dx \\
&= \int_{\mathcal{X}} \cancel{f_X(x)}\left\{\int_{-\infty}^{\infty} y\,\frac{f_{XY}(x,y)}{\cancel{f_X(x)}}\,dy\right\} dx \\
&= \int_{-\infty}^{\infty} y\,\underbrace{\left\{\int_{\mathcal{X}} f_{XY}(x,y)\,dx\right\}}_{=f_Y(y)} dy = \int_{-\infty}^{\infty} y\, f_Y(y)\,dy = \mathbb{E}[Y]
\end{aligned}$$

The law of iterated expectations will be quite useful for us later this semester and next semester.

One helpful intuition for the law of iterated expectations is that, although $\mathbb{E}[Y|X=x]$ could vary perhaps arbitrarily across different values of X, knowing conditional expectations for all values of X does pin down the overall expectation.

Example: Suppose that Y is individual i's height and X is an indicator for whether they are a child or an adult. Then the law of iterated expectations tells us that the average height in the population can be obtained by averaging together the mean height among children with the mean height among adults. Suppose that 75% of the population are adults. Then the law of iterated expectations reads as:

$$\mathbb{E}[height] = .75 \cdot \mathbb{E}[height|adult] + .25 \cdot \mathbb{E}[height|child]$$

The above is probably the most useful version of the law of iterated expectations, but there are slightly more general versions that can sometimes be useful. One that we use below is that

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

Conditional variance

PSE 4.16

We can analogously define a **conditional variance** function $Var(Y|X=x) = E[(Y-E[Y|X=x])^2|X=x]$ from the conditional distribution $F_{Y|X=x}$. An analog to the law of iterated expectations exists for the conditional variance, which is sometimes called the **law of total variance**

Proposition: (law of total variance) $Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$.

Proof: To start with, write $Y - \mathbb{E}[Y] = (Y - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X] - \mathbb{E}[Y])$. Thus,

$$\begin{aligned}\text{var}(Y) &= \mathbb{E}[(Y - \mathbb{E}[Y])^2] \\ &= \mathbb{E}\left[((Y - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X] - \mathbb{E}[Y]))^2\right] \\ &= \underbrace{\mathbb{E}[(Y - \mathbb{E}[Y|X])^2]}_{A} + \underbrace{\mathbb{E}[(\mathbb{E}[Y|X] - \mathbb{E}[Y])^2]}_{B} + 2\underbrace{\mathbb{E}\left[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - \mathbb{E}[Y])\right]}_{C}\end{aligned}$$

Consider each term above in turn:

$$\begin{aligned}A &= \mathbb{E}\left[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]\right] \\ &= \mathbb{E}[\text{var}(Y|X)]\end{aligned}$$

where the first equality uses the law of iterated expectations. For the second, term recall that, by the law of iterated expectations, the mean of $\mathbb{E}[Y|X]$ is $\mathbb{E}[Y]$. Thus

$$B = \text{var}(\mathbb{E}[Y|X])$$

Finally, for the last term, the law of iterated expectations implies that

$$\begin{aligned}C &= \mathbb{E}\Big[\mathbb{E}\left[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - \mathbb{E}[Y])|X\right]\Big] \\ &= \mathbb{E}\Big[(\mathbb{E}[Y|X] - \mathbb{E}[Y])\underbrace{\mathbb{E}\left[(Y - \mathbb{E}[Y|X])|X\right]}_{=(\mathbb{E}[Y|X]-\mathbb{E}[Y|X])=0}\Big]\end{aligned}$$

where the second equality holds because $\mathbb{E}[Y]$ is non-random, and $\mathbb{E}[Y|X]$ is non-random conditional on X, thus they can both be moved outside of the inside expectation.

Example: Recall the height example from the law of iterated expectations. The law of total variance reveals that the variance of heights in the population overall is *greater* than what we would get by just averaging the variances of each subgroup. That is:

$$\text{var}(height) > .75 \cdot \text{var}(height|adult) + .25 \cdot \text{var}(height|child)$$

The reason is that var($height$) involves making comparisons directly between the heights of children and adults, which are not captured in $\text{var}(Y|X = x)$ for either value of x. The law of total variance tells us exactly what correction we would need to make, which is to add the second term $\text{var}(\mathbb{E}[Y|X])$. Remarkably, the correction required just depends on the *average* height within each group $\mathbb{E}[Y|X = x]$, as well as the proportion of adults vs. children: $\text{P}(X = x)$.

Multivariate normal distribution

PSE 5.1-5.2, 5.9

The multi-variate normal distribution generalizes the normal distribution to a random vector $X = (X_1, X_2, \dots X_k)$. In this case, we would write $X \sim N(\mu, \mathbf{\Sigma})$ where μ is a $k \times 1$ vector and $\mathbf{\Sigma}$ is a $k \times k$ variance matrix. In the case where $k = 2$, the variance matrix is given by

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho \cdot \sigma_1 \cdot \sigma_2 \\ \rho \cdot \sigma_1 \cdot \sigma_2 & \sigma_2^2 \end{pmatrix} = (\sigma_1, \sigma_2) \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} (\sigma_1, \sigma_2)'$$

where $\sigma_j^2 = \text{var}(X_j)$ and $\rho = \text{corr}(X_1, X_2)$. In other words, the distribution of (X_1, X_2) is fully determined by (μ_1, μ_2), (σ_1^2, σ_2^2), and ρ.

There are a number of useful properties of multi-variate normal distributions that are provided in chapter 5 of the textbook. I am going to briefly mention several that I know we will use at some point, but we make come back and pick up other properties as needed.

Property. If X and Y are jointly normal with $\begin{pmatrix} X \\ Y \end{pmatrix} = N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right)$

$$\text{Then:} \quad X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY})$$

Property. If $Z \sim N(0, \mathbf{I}_r)$ (i.e., multivariate standard normal), then $Z'Z \sim \chi_r^2$ (that is, $Z'Z$ follows a chi-squared distribution with r degrees of freedom).

Property. If $X \sim N(0, \mathbf{A})$ where $\mathbf{A}$ is an $r \times r$ positive definite matrix, then $X'A^{-1}X \sim \chi_r^2$

The intuition for this property is that, if $\mathbf{A}$ is positive definite, then it has positive definite inverse $\mathbf{A}^{-1}$ which itself has a corresponding square root matrix, which we can denote $\mathbf{A}^{-1/2}$ (which is positive definite, hence symmetric). Thus, $X'\mathbf{A}^{-1}X = \underbrace{(\mathbf{A}^{-1/2}X)'}_{\sim N(0,\mathbf{I}_r)}\underbrace{\mathbf{A}^{-1/2}X}_{\sim N(0,\mathbf{I}_r)}$.

Example: Income, education, and age in the U.S.

To conclude this section, let us return to the example that we used previously on income and education (and we'll also add age) in the U.S.

```r
# load useful packages
library(haven)
library(dplyr)
library(ggplot2)
```

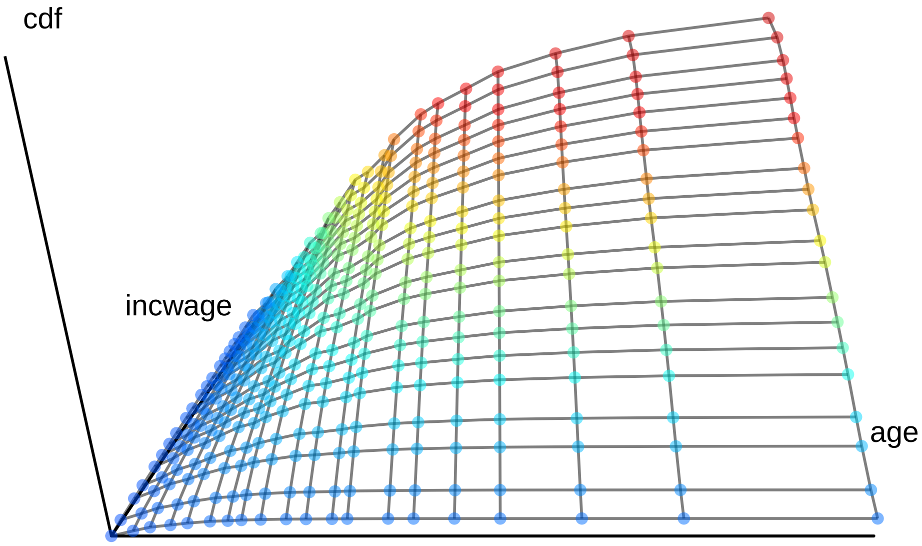
```
# load data
load("us_data.RData")
```

Conditional distributions

Joint distributions/pmfs/pdfs are typically hard objects to display. We could potentially make a three-dimensional plot of, say, the joint cdf of income and age. This is not something that I do often, but I found an R package that can plot bivariate cdfs, and I'll briefly show you how to do this.

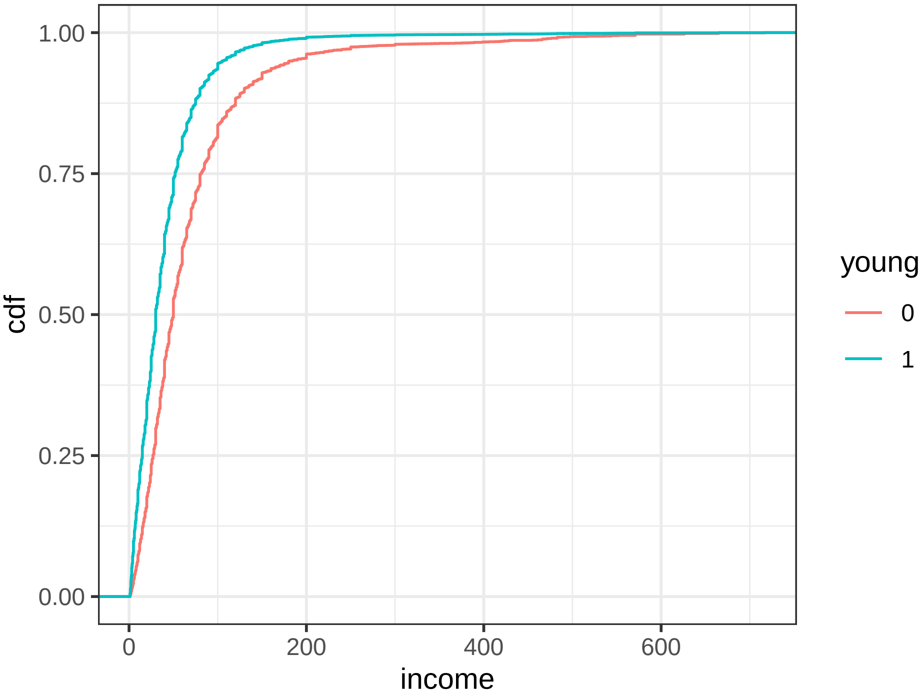
```
# download the gg3d package using the following command
# devtools::install_github("AckerDWM/gg3D")
library(gg3D)

# create grid of values for which to compute joint cdf
# (in principle, you could use all values that show
# up in the data, but that turns out to be fairly computational)
u <- seq(.05,.95,length.out=20)
xy_vals <- expand.grid(quantile(us_data$incwage,u),
                       quantile(us_data$age,u))
colnames(xy_vals) <- c("incwage", "age")
# compute value of cdf at all points in xy_vals
z <- sapply(1:nrow(xy_vals), function(i) {
  mean( 1*(us_data$incwage <= xy_vals[i,]$incwage &
           us_data$age <= xy_vals[i,]$age) )
})
# put into data frame for plotting
plot_df <- cbind.data.frame(xy_vals, z)
# make the plot
theta <- 0 # theta control rotation
phi <- 20   # phi controls tilt
ggplot(plot_df, aes(x=incwage, y=age, z=z)) +
  axes_3D(theta=theta, phi=phi) +
  stat_wireframe(theta=theta, phi=phi, alpha=.5) +
  stat_3D(aes(color=z), theta=theta, phi=phi, alpha=.5) +
  theme_void() +
  theme(legend.position = "none") +
  scale_color_gradientn(colors=plot3D::jet2.col()) +
  labs_3D(labs=c("incwage", "age", "cdf"),
          hjust=c(0,-.2,12), vjust=c(0, 6, 7), angle=c(0, 0, 0))
```



As an alternative to the previous plot, next we'll create a new variable indicating whether a person is young or old and then plot the cdf of income across groups.

```
us_data$young <- as.factor(1*(us_data$age <= 36)) # factor => categorical
ggplot(us_data, aes(x=incwage/1000, color=young)) +
  stat_ecdf() +
  xlab("income") +
  ylab("cdf") +
  theme_bw()
```



which seems to indicate that the income distribution is better for the old than for the young. Next, let's calculate the covariance and correlation between income and age/education.

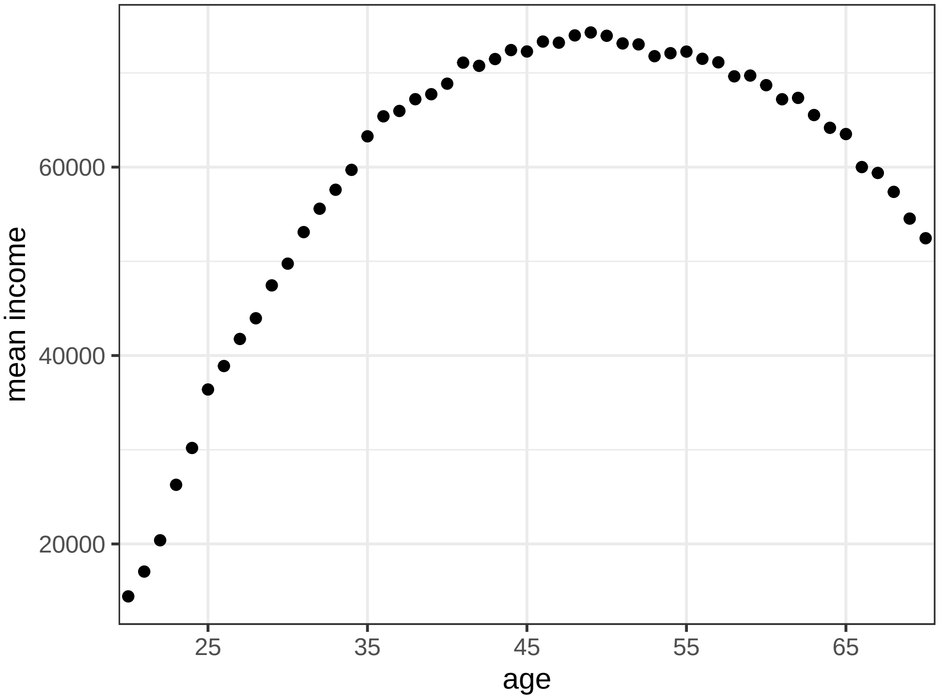
```
cov_inc_age <- cov(us_data$incwage, us_data$age)
cov_inc_edu <- cov(us_data$incwage, us_data$educ)
corr_inc_age <- cor(us_data$incwage, us_data$age)
corr_inc_edu <- cor(us_data$incwage, us_data$edu)
m <- matrix(data=c(cov_inc_age, cov_inc_edu, corr_inc_age, corr_inc_edu),
       nrow=2)
colnames(m) <- c("cov", "corr")
rownames(m) <- c("inc/age", "inc/edu")
round(m,2)
```

```
##               cov corr
## inc/age 173783.05 0.17
## inc/edu  63766.72 0.32
```

As we discussed above, the particular value of a covariance is often hard to interpret (because the units are unfamiliar), but both age and education (as we would probably expect) are positively correlated with income.

Next, let's plot the conditional expectation of income as a function of age.

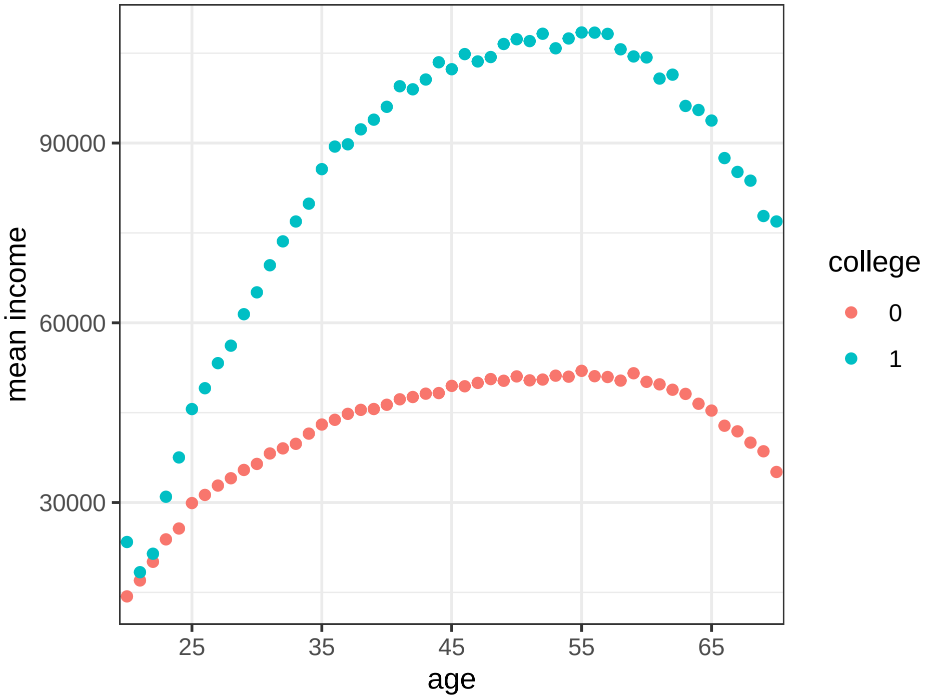
```
plot_df <- dplyr::group_by(us_data, age) %>% summarize(mean_income=mean(incwage))
plot_df <- subset(plot_df, age>=20 & age<=70)
plot_df$age <- as.factor(plot_df$age)
ggplot(plot_df, aes(x=age, y=mean_income)) +
  geom_point() +
  theme_bw() +
  scale_x_discrete(breaks=c("25", "35", "45", "55", "65")) +
  ylab("mean income")
```



This indicates that the partial effect of age on income seems to be much larger among those in their 20's and early 30's relative to those later in their careers.

And, to conclude, let's do the same thing but separately for college graduates and non-college graduates.

```
us_data$col <- 1*(us_data$educ >= 16)
plot_df <- dplyr::group_by(us_data, age, col) %>% summarize(mean_income=mean(incwage))
plot_df <- subset(plot_df, age>=20 & age<=70)
plot_df$age <- as.factor(plot_df$age)
plot_df$college <- as.factor(plot_df$col)
ggplot(plot_df, aes(x=age, y=mean_income, color=college)) +
  geom_point() +
  theme_bw() +
  scale_x_discrete(breaks=c("25", "35", "45", "55", "65")) +
  ylab("mean income")
```



These appear to be notably different income profiles (at least on average) among college graduates and non-college graduates.