Review Session 3 – Multiple Random Variables

References/suggested reading

(i) Casella & Berger's *Statistial Inference*, chapter 4.

1 Joint and Marginal Distributions

Last week, we discussed random variables, their transformations, and associated quantities. Here, we will extend these notions to multiple dimensions. We start with the notion of a random vector, which is the multivariate generalization of a random variable.

Definition 1.1 (random vector)**.** An n-dimensional *random vector* is a function from a sample space $\mathcal{S}$ into $\mathbb{R}^n$.

Example 1.2
Consider the experiment of tossing two fair dice. The sample space for this experiment has 36 equally likely points. With each of these 36 points, we can associate two numbers X and Y, where X is the sum of the two dice and Y is the absolute difference of the two dice. Then, (X, Y) is a bivariate random vector.

In the ensuing discussion, we will define notions such as pmf/pdf, expectations, transformations, etc. mostly for bivariate random vectors (X, Y). The analogous notions for higher dimensions follow straightforwardly. Let's start with the case where X, Y are discrete.

Definition 1.3 (discrete random vector)**.** The random vector (X, Y) is called *discrete* if it only has a countable number of possible values.

We can now extend the notion of a pmf to multiple dimensions.

Definition 1.4 (joint pmf)**.** Associated with a discrete bivariate random vector (X, Y) is the *joint probability mass function (pmf)* which is a function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x, y) = \mathbb{P}(X = x, Y = y).$$

We often denote it as $f_{X,Y}(x, y)$ to emphasize it is the joint pmf of the random vector (X, Y) (as opposed to some other vector).

Just as in the univariate setting, the joint pmf of (X, Y) completely defines the probability distribution of the random vector (X, Y). It can further be used to compute the probability of any event defined in terms of (X, Y). More concretely, for any subset $A \subseteq \mathbb{R}^2$, we have

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

where we interpret this sum as being a countable sum since $f(x, y) > 0$ for only a countable number of values. Following this analogy, the law of the unconscious statistician also has an multivariate analogue:

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

Given a probability model for a random vector (X, Y), we may often only be interested in events involving only one of the random variables in the vector, i.e. we want to compute $\mathbb{P}(X = 2)$. The variable X is itself a random variable (being a function $\mathcal{S} \to \mathbb{R}$), so we can speak about its own pmf $f_X(x) = \mathbb{P}(X = x)$. We will call this the *marginal pmf* of X within the context of the joint distribution of vector (X, Y). The marginal pmf can be obtained from the joint pmf by summing out the other variable(s):

Theorem 1.5

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, $f_X(x) := \mathbb{P}(X = x)$ and $f_Y(y) = \mathbb{P}(Y = y)$ are given by

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

The above sums of course are understood as being over a countable set of y's or x's.

One commonly misunderstood point is that the marginal distributions of X and Y, described by the marginal pmf's $f_X(x)$ and $f_Y(y)$, do not completely describe the joint distribution of X and Y. Let's see an example of this.

Example 1.6 (same marginals, different joint pmf)

Define a joint pmf by

$$f(0,0) = \frac{1}{12}, f(1,0) = \frac{5}{12}, f(0,1) = f(1,1) = \frac{3}{12}, f(x,y) = 0 \text{ for all other values.}$$

We can verify the marginal pmf of Y is $f_Y(0) = f(0,0) + f(1,0) = \frac{1}{2}$ and $f_Y(1) = f(0,1) + f(1,1) = \frac{1}{2}$. Meanwhile, the marginal pmf of X is $f_X(0) = \frac{1}{3}$ and $f_X(1) = \frac{2}{3}$. Now, consider the alternative pmf $g(x, y)$ defined by

$$g(0,0) = g(0,1) = \frac{1}{6}, g(1,0) = g(1,1) = \frac{1}{3}, g(x,y) = 0 \text{ for all other values.}$$

Then, we see that the marginal pmfs $g_X(x), g_Y(y)$ satisfy $g_X(x) = f_X(x)$ and $g_Y(y) = f_Y(y)$ for all x, y. But, clearly $g_{X,Y}$ and $f_{X,Y}$ are distinct joint distributions.

Note that the converse point holds: if two random vectors have the same joint distribution, then their marginals coincide in distribution.

If X and Y are continuous random variables, then the joint pdf, expectation, and marginal pdf can be defined analogously. In this case, we call (X, Y) a *continuous* random vector.

Definition 1.7 (joint pdf, expectation, marginal pdf)**.** A function $f : \mathbb{R}^2 \to \mathbb{R}$ is called a *joint probability density function* or *joint pdf* of the continuous bivariate random vector (X, Y) if, for every $A \subset \mathbb{R}^2$:

$$\mathbb{P}((X, Y) \in A) = \int \int_A f(x, y)\, dx\, dy.$$

Here, the notation "$\int \int_A$" simply means that the limits of integration are set so that the function is integrated over all $(x, y) \in A$. You may recall from your multivariate calculus course that it is not always obvious what these limits should be for complicated non-rectangular sets A. We'll later see how to transform these multiple integrals into a more convenient form in these situations.

The *expectation of a function* of a continuous random vector $g(X, Y)$ is defined as

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

The *marginal pdf's* of X and Y are obtained by integrating out the other variable:

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

Similar to the characterization of a univariate pdf, we have that any function $f(x,y)$ satisfying $f(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$ and

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

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

Recall that the cumulative distribution function (cdf) was another way of characterizing the distribution of a random variable. This can also be generalized to a random vector.

Definition 1.8 (joint cdf)**.** The joint cdf of a random vector (X,Y) (which may be either continuous or discrete) is the function

$$F(x,y) := \mathbb{P}(X \leq x, Y \leq y).$$

For a continuous bivariate random vector, we have the relationship:

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t)\, dt\, ds \implies \frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y).$$

Remark 1.9. You might find it peculiar that we defined F by integrating with respect to Y first, and then X. What if we had integrated in the other order:

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

It turns out this is the same multivariate integral (and thus equal to $F(x,y)$) due to a result known as Tonelli's theorem, which allows us to switch the order of integration. You will learn about Tonelli's theorem in your probability theory course. This also allows us to conclude that the pdf can be obtained from the joint cdf by differentiating in a different order:

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

You might recall this latter fact as Clairaut's theorem from multivariate calculus.

2 Conditional Distributions and Independence

Often times when two random variables, (X,Y), are observed, the values of the two variables are related. For example, suppose that, in sampling from a human population, X is a person's height and Y is the same person's weight. Then, knowledge about X, e.g. $X = 73$ inches, would confer us probably knowledge about Y, i.e. $Y > 150$ pounds. On the other hand, sometimes knowledge about X may give us no information about Y. Much like the conditional probabilities discussed in the previous review section, we can capture this through *conditional distributions*.

Definition 2.1 (conditional pmf)**.** Let (X,Y) be a discrete bivariate random vector with joint pmf $f(x,y)$ and maginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $\mathbb{P}(X = x) = f_X(x) > 0$, the *conditional pmf of* Y *given* $X = x$ is the function

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

Similarly, the conditional pmf of X given $Y = y$ is given by $f(x|y) = \mathbb{P}(X = x | Y = y) = \frac{f(x,y)}{f_Y(y)}$.

We can verify that $f(y|x)$ is a valid univariate pmf, i.e. $f(y|x) \geq 0$ for every y and $\sum_y f(y|x) = 1$.

To obtain the analogue for continuous random variables X and Y, we need to be more careful. $\mathbb{P}(X = x) = 0$ for every x since X is continuous. Yet in actuality a value of X is observed. However, we can still define a conditional pdf by taking the ratio of the joint pdf and the marginal pdf, which both capture relative likelihoods near their respective points.

Definition 2.2 (conditional pdf)**.** 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* $X = x$ is the function

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

Similarly, the conditional pdf of X given $Y = y$ is the function $f(x|y) := \frac{f(x,y)}{f_Y(y)}$ for y such that $f_Y(y) > 0$.

Like before, we can verify $f(y|x)$ and $f(x|y)$ are indeed pdf's. It then follows that we have the law of the unconscious statistician for conditional distributions:

$$\mathbb{E}[g(Y)|X=x] = \begin{cases} \sum_y g(y)f(y|x) & X,Y \text{ are discrete} \\ \int_{-\infty}^{\infty} g(y)f(y|x)\,dy & X,Y \text{ are continuous.} \end{cases}$$

This will serve useful as the *conditional expectation*, $\mathbb{E}[Y|X]$, is a quantity which appears in many applications (e.g., linear regression, Bayes estimation).

We can also define the *conditional variance* of Y given $X=x$, which is the variance of the conditional distribution:

$$\mathrm{Var}(Y|X=x) = \mathbb{E}[Y^2|X=x] - (\mathbb{E}[Y|X=x])^2.$$

Note that the conditional distribution of Y given $X=x$ is possibly a different probability distribution for each value of x. Thus, the conditional "$Y|X$" really refers to a family of probability distributions, one for each $x \in \mathcal{X}$.

Along this same line of thought, the conditional expectation $\mathbb{E}[g(Y)|X=x]$ is now a function of x.

However, sometimes, the conditional distribution of Y given $X=x$ will be the same for different values of x. In fact, if this is the case, this univariate distribution will coincide with the marginal distribution of Y (i.e., have pdf/pmf $f_Y(y)$). This is akin to the way we defined independence in the previous review session.

Definition 2.3 (independence)**.** Let (X,Y) be a bivariate random vector with joint pdf or pmf $f(x,y)$ and marginal pdf/pmf $f_X(x)$ and $f_Y(y)$. Then, X and Y are called *independent random variables* if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$:

$$f(x,y) = f_X(x)f_Y(y).$$

Now, what does this say about the conditional pdf/pmf $f(y|x)$? As expected, we have

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

Similarly, $f(x|y) = f_X(x)$.

This notion of independence can be difficult to check if the marginals $f_X(x)$ and $f_Y(y)$ are tedious to compute. Luckily, there is a much simpler criterion which is much faster to verify.

Theorem 2.4

Let (X,Y) be a bivariate random vector with joint pdf or pmf $f(x,y)$. Then, X and Y are independent random variables iff there exist functions $g(x)$ and $h(y)$ such that, for every $x,y \in \mathbb{R}$,

$$f(x,y) = g(x)h(y)$$

Proof. We will show this for continuous X,Y. The proof for discrete X,Y is identical with integrals replaced by sums. The "only if" part is clear since we can let $g(x) = f_X(x)$ and $h(y) = f_Y(y)$. Conversely, suppose $f(x,y) = g(x)h(y)$ for all $x,y \in \mathbb{R}$. Then, we can verify that the marginals pdf satisfty

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{-\infty}^{\infty} g(x)h(y)\,dy = g(x)\cdot\int_{-\infty}^{\infty} h(y)\,dy$$
$$f_Y(y) = h(y)\cdot\int_{-\infty}^{\infty} g(x)\,dx.$$

Then,

$$\begin{aligned} f(x,y) &= g(x)h(y) \cdot 1 \\ &= g(x)h(y) \int_{\mathbb{R}^2} f(x,y)\, d(x,y) \\ &= g(x)h(y) \int_{\mathbb{R}^2} g(x)h(y)\, dx\, dy \\ &= \left(g(x) \cdot \int_{\mathbb{R}} h(y)\, dy \right) \left(h(y) \cdot \int_{\mathbb{R}} g(x)\, dx \right) \\ &= f_X(x) f_Y(y). \end{aligned}$$

Thus, X, Y are independent. ■

Here is another way to check for independence just by looking at the support of the joint pdf.

Theorem 2.5

If $f(x,y)$ is a joint pdf or pmf and the set where $f(x,y) > 0$ cannot be given as a cross-product (i.e., $A \times B$ for some $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}$), then it turns out that X and Y with joint pdf or pmf $f(x,y)$ are *not* independent.

Proof. If X, Y were independent random variables, then it would be clear that $f(x,y) > 0$ on the set $\{(x,y) : x \in A, y \in B\}$ where $A = \{x : f_X(x) > 0\}$ and $B = \{y : f_Y(y) > 0\}$. In fact, this must be the only set of x's where $f(x,y) > 0$ since $f(x,y) > 0 \implies f_X(x), f_Y(y) > 0$. Thus, $f(x,y) > 0$ on $A \times B$, a contradiction. ■

Two random variables X, Y are independent if, intuitively, information about one of them does not affect the realization of the other. This intuition carries through for many of the familiar properties of a random variable:

Theorem 2.6

Let X and Y be independent random variables.

1. For any $A \subset \mathbb{R}, B \subset \mathbb{R}$,
$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$
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,
$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$
The converse is true if the above holds for all bounded almost surely continuous g, h.

Proof. We can show the second part by splitting up the double integral into a product of two univariate integrals:

$$\mathbb{E}[g(X)h(Y)] = \int_{\mathbb{R}^2} g(x)h(y)f(x,y)\, dx\, dy = \left(\int_{\mathbb{R}} h(y) f_Y(y)\, dy \right) \cdot \left(\int_{\mathbb{R}} g(x) f_X(x)\, dx \right) = \mathbb{E}[h(Y)]\mathbb{E}[g(X)].$$

The first part can then be shown from the second part by letting $g(x) = \mathbf{1}\{x \in A\}$, $h(y) = \mathbf{1}\{y \in B\}$ (i.e., the indicator functions over the sets A and B) so that

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{E}[\mathbf{1}\{x \in A\} \cdot \mathbf{1}\{y \in B\}] = \mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)] = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

■

One immediate application of the previous theorem is that the mgf of the sum of independent random variables $X + Y$ is the product of the mgf's:

Theorem 2.7 (mgf of independent sum factors)

Let X and Y be independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then the moment generating function of the random variable $Z = X + Y$ is given by

$$M_Z(t) = M_X(t)M_Y(t)$$

3 Bivariate Transformations

Next, we will focus on bivariate transformations $(X, Y) \mapsto (U, V)$ where $U = g_1(X, Y)$ and $V = g_2(X, Y)$ are univariate transformations of the random vector (X, Y). Let's suppose that the transformation $(x, y) \mapsto (g_1(x, y), g_2(x, y))$ is bijective and differentiable. We have that if $B \subseteq \mathbb{R}^2$, then $(U, V) \in B$ iff $(X, Y) \in A$ where $A := \{(x, y) : (g_1(x, y), g_2(x, y)) \in B\}$. Thus,

$$\mathbb{P}((U, V) \in B) = \mathbb{P}((X, Y) \in A) \implies \int_B f_{U,V}(u, v)\, d(u, v) = \int_A f_{X,Y}(x, y)\, d(x, y).$$

Thus, the joint pdfs $f_{U,V}(u, v)$ and $f_{X,Y}(x, y)$ must be related through the change of variables procedure for $(x, y) \mapsto (u, v)$. In one dimension, recall from the previous review session that this involved the derivative of the transformation g. However, now there are two transformations g_1 and g_2.

In the "Linear Algebra" review session, we introduced the determinant of a matrix A as the volume of a parallelogram spanned by the rows of A. When A is the *inverse Jacobian matrix* the determinant is

$$\det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \frac{\partial x}{\partial u} \cdot \frac{\partial y}{\partial v} - \frac{\partial y}{\partial u} \cdot \frac{\partial x}{\partial v} =: J.$$

Then, the rectangle of differentials in X, Y space with area $dx \times dy$ goes to a parallelogram in U, V space with area roughly the determinant of the Jacobian J. Thus, we should have $d(x, y) = |J| \cdot d(u, v)$ and

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v))|J|,$$

where $(u, v) \mapsto (h_1(u, v), h_2(u, v)$ is the inverse transformation of $(x, y) \mapsto (g_1(x, y), g_2(x, y))$.

Remark 3.1. Sometimes it may occur that there is only one function $U = g_1(X, Y)$ of interest. If another convenient function $V = g_2(X, Y)$ can be chosen so that the resulting transformation from (X, Y) to (U, V) is one-to-one on the support of $f_{X,Y}(x, y)$, then the joint pdf of (U, V) can be derived using the result from before and the marginal pdf of U can be obtained by integrating the joint pdf. For example, if $U = XY$, we can choose $V = X$ or $V = Y$ so that the resulting transformation $(X, Y) \mapsto (U, V)$ is one-to-one.

Remark 3.2. We clearly generalized the pdf transformation law from the previous review session here. Yet you may notice that we did not make any explicit requirement that g_1 or g_2 be monotone, unlike in the univariate setting. Does this mean the univariate pdf transformation law holds for non-monotonic functions? It turns out that every continuous bijective function is in fact monotonic. Try to prove this or draw a picture to convince yourself. Thus, since we assume the transformation is differentiable and bijective, we are in fact working with monotonic transformations when using this result in one dimension.

What about the transformation of independent random variables? If X, Y are independent, then any modification of them $g(X), h(Y)$ should also be independent.

Theorem 3.3

Let X and Y be independent random variables. Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then the random variables $U = g(X)$ and $V = h(Y)$ are independent.

Proof. We have the joint cdf of (U, V) is

$$F_{U,V}(u, v) = \mathbb{P}(g(X) \le u, h(Y) \le v) = \mathbb{P}(g(X) \le u)\mathbb{P}(h(Y) \le v) \implies f_{U,V}(u, v) = \frac{\partial^2}{\partial u \partial v} F_{U,V}(u, v) = \tilde{g}(u) \cdot \tilde{h}(v).$$

Thus, since the joint pdf $f_{U,V}(u, v)$ factors, U, V are independent. ■

Example 3.4 (ratio of Gaussians)

Let X, Y be independent $\mathcal{N}(0,1)$ random variables. Consider the transformation $U = X/Y$ and $V = |Y|$ (for $Y = 0$, we can let $(U, V) = (1, 1)$ or any value since $\mathbb{P}(Y = 0) = 0$ and so this case is negligible). This transformation is not one-to-one, but is one-to-one when restricted to either positive or negative values of y. Let

$$A_1 = \{(x, y) : y > 0\}, A_2 = \{(x, y) : y < 0\}, A_0 = \{(x, y) : y = 0\}$$

which partition $\mathcal{A} = \mathbb{R}^2$. For either A_1 or A_2, if $(x, y) \in A_i$, $v = |y| > 0$ and for a fixed $v = |y|$, $u = x/y$ can be any real number. Thus,

$$\mathcal{B} = \{(u, v) : v > 0\}$$

The inverse transformations from $\mathcal{B}$ to A_1 and $\mathcal{B}$ to A_2 are given by $(u, v) \mapsto (uv, v)$ and $(u, v) \mapsto (-uv, -v)$. The determinants of both Jacobians are v. Then, using the fact that

$$f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-x^2/2} e^{-y^2/2}$$

we have using our transformation law:

$$f_{U,V}(u, v) = \frac{v}{\pi} e^{-(u^2+1)v^2/2}, u \in \mathbb{R}, v \in (0, \infty)$$

From this the marginal pdf of U can be computed to be (via a change-of-variables and recognizing the integrand is proportional to an exponential pdf):

$$f_U(u) = \int_0^\infty \frac{v}{\pi} e^{-(u^2+1)v^2/2}\, dv = \frac{1}{\pi(u^2 + 1)}, u \in \mathbb{R}$$

You may recognize this last formula as the pdf of a *Cauchy distribution*. Thus, the ratio of two independent standard normal random variables is a Cauchy random variable.

4 Hierarchical Models and Mixture Distributions

In the cases we have seen thus far, a random variable has a single distribution, possibly depending on parameters. More generally, we can think of these parameters as also following a distribution, giving a hierarchical model. This is a setting where it is particularly important to make correct use of conditional and joint distributions. Here are two useful identities that often come in play:

Theorem 4.1 (law of iterated expectation, or tower property)

If X, Y are any two random variables, then

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]],$$

provided that the expectations exist.

Proof. Letting $f(x, Y)$ denote the joint pdf of X, Y, we have

$$\mathbb{E}[X] = \int \int x f(x, y)\, dx\, dy = \int \left[\int x f(x|y)\, dx \right] f_Y(y)\, dy = \mathbb{E}[\mathbb{E}[X|Y]].$$

■

Theorem 4.2 (law of total variance)

For any two random variables X, Y

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y])$$

provided that the expectations exist.

Proof. This is another one of those proofs where we decompose a square error as a sum of two squares:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(X - \mathbb{E}[X|Y] + \mathbb{E}[X|Y] - \mathbb{E}[X])^2] = \mathbb{E}[(X - \mathbb{E}(X|Y))^2] + \mathbb{E}[(\mathbb{E}(X|Y) - \mathbb{E}[X])^2].$$

We obtained the last term by expanding the square and applying the linearity of expectation. Why did the cross term disappear? We can finish by verifying

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}(X|Y))^2] &= \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y]] = \mathbb{E}[Var(X|Y)] \\ \mathbb{E}[(\mathbb{E}(X|Y) - \mathbb{E}[X])^2] &= \mathrm{Var}(\mathbb{E}[X|Y]).\end{aligned}$$

■

Example 4.3 (binomial-poisson hierarchy)

An insect lays a large number of eggs, each surviving with probability p. On the average, how many eggs will survive? Let the "large number" of eggs laid be a random variable $Y \sim$ Poisson(λ) and let the number of survivors be a random variable X. Assuming each egg's survival is independent, we have $X|Y \sim$ binomial(Y, p). This gives a *hierarchical model*. We have

$$\begin{aligned}\mathbb{P}(X = x) &= \sum_{y=0}^{\infty} \mathbb{P}(X = x, Y = y) \\ &= \sum_{y=0}^{\infty} \mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y) \\ &= \sum_{y=x}^{\infty} \left(\binom{y}{x} p^x (1-p)^{y-x} \right) \cdot \left(\frac{e^{-\lambda}\lambda^y}{y!} \right) \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{t!} \qquad (t = y - x) \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p}\end{aligned}$$

so that $X \sim$ Poisson(λp). Thus, $\mathbb{E}[X] = \lambda p$ to answer the original question.

We can also determine this much more quickly from the law of iterated expectation

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[pY] = p\lambda$$

Example 4.4

Consider a generalization of Example 4.3, where instead of one mother insect there are a large number of mothers and one mother is chosen at random. Let X be the number of survivors in a litter; then

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

where the last stage of the hierarchy accounts for the variability across different mothers. The mean of X is then

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[pY] = \mathbb{E}[\mathbb{E}[pY|\Lambda]] = \mathbb{E}[p\Lambda] = p\beta$$

5 Covariance and Correlation

We've discussed the absence or presence of a relationship between two random variables by speaking of independence vs. non-independence. However, we'd like a more refined quantity that describes the strength of the relationship. This leads us to covariance and correlation. Intuitively, it describes whether greater or lesser values of one variable correspond to greater or lesser values of another variable. For example, if X is the weight of a human and Y is the height of that same human, then we might expect X and Y to generally increase and decrease together across a sample of people, but not all the time.

Notation 5.1. Henceforth, let $\mathbb{E}[X] =: \mu_X, \mathbb{E}[Y] =: \mu_Y, \operatorname{Var} X =: \sigma_X^2, \operatorname{Var} Y =: \sigma_Y^2$. Also, assume that $\sigma_X^2, \sigma_Y^2 \in (0, \infty)$.

Definition 5.2 (covariance)**.** The *covariance* of X and Y is the number defined by

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \cdot \mu_Y.$$

Definition 5.3 (correlation)**.** The *correlation* of X and Y is the number defined by

$$\rho_{XY} := \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

The value ρ_{XY} is also called the *correlation coefficient*.

Theorem 5.4

If X, Y are independent random variables, then $\operatorname{Cov}(X, Y) = 0$ and $\rho_{XY} = 0$.

Proof. If X, Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \implies \operatorname{Cov}(X, Y) = 0$. ■

A natural question that arises is whether uncorrelated random variables, i.e. $\operatorname{Cov}(X, Y) = 0$, are independent. It turns out that this not the case in general.

Example 5.5 (uncorrelated but dependent random variables)

Let $X \sim \text{Unif}(-1, 1)$ and let $Y = X^2$. Then, X, Y are clearly dependent, but

$$\operatorname{Cov}(X, Y) = \mathbb{E}[X^3] - \mathbb{E}[X^2]\mathbb{E}[X] = 0 - 1/3 \cdot 0 = 0.$$

How is the covariance related to the similarly named variance? Letting $X = Y$, we can see from the formula that $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$. We also have that the term $\operatorname{Cov}(X, Y)$ appears when computing the variance of a linear combination $\operatorname{Var}(aX + bY)$.

Theorem 5.6

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

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var} X + b^2 \operatorname{Var} Y + 2ab \operatorname{Cov}(X, Y)$$

Proof.

$$\mathrm{Var}(aX+bY) = \mathbb{E}[(aX+bY)-(a\mu_X+b\mu_Y))^2] = a^2\mathbb{E}[(X-\mu_X)^2] + b^2\mathbb{E}[(Y-\mu_Y)^2] + 2ab\mathbb{E}[(X-\mu_X)(Y-\mu_Y)].$$

■

Theorem 5.7

For any random variables X, Y,

1. $-1 \leq \rho_{XY} \leq 1$
2. $|\rho_{XY}| = 1$ iff there exist $a \neq 0$ and b such that $\mathbb{P}(Y = aX+b) = 1$. If $\rho_{XY} = 1$, then $a > 0$ and if $\rho_{XY} = -1$, then $a < 0$.

6 Mutual Independence

Most of the results we've derived so far extend naturally to higher dimensions beyond 2. For independence, we have to be more careful. Recall from the previous review session that the mutual independence of multiple events was defined in a rather strong way.

Notation 6.1. We will use boldface letters to denote multiple variates. Thus, we write $\mathbf{X}$ to denote the random variables $X_1, \ldots, X_n$ and $\mathbf{x}$ to denote observations $x_1, \ldots, x_n$.

Definition 6.2 (mutual independence)**.** Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random vectors with joint pdf or pmf $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Let $f_{\mathbf{X}_i}(\mathbf{x}_i)$ denote the marginal pdf or pmf of $\mathbf{X}_i$. Then $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are called *mutually independent random vectors* if, for every $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$,

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

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

Remark 6.3. Note that we did not need to mandate that $f(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}) = f_{\mathbf{X}_{i_1}}(\mathbf{x}_{i_1}) \cdots f_{\mathbf{X}_{i_k}}(\mathbf{x}_{i_k})$ for every subset $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, like we did for defining mutual independence of events. This in fact automatically follows from the above definition. Do you see why? Hint: think about how the marginal pdf $f_{\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_k}}(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k})$ is obtained.

Example 6.4

Pairwise independence of random variables does not imply mutual independence. Consider the probability space $\Omega = [4]$ with probability measure placing mass $1/4$ at each of the four points. The indicator over the events $\{1,2\}, \{2,3\}, \{1,3\}$ are pairwise independent but not mutually independent.

Mutually independent random variables behave much like independent random variables. The properties we derived earlier easily generalize:

Theorem 6.5

Let $X_1, \ldots, X_n$ be mutually independent random variables. Let $g_1, \ldots, g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of x_i, for $i = 1, \ldots, n$. Then

$$\mathbb{E}[g_1(X_1) \cdots g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdots \mathbb{E}[g_n(X_n)].$$

Additionally, if $M_{X_1}(t), \ldots, M_{X_n}(t)$ are the mgf's of $X_1, \ldots, X_n$, respectively, and $Z := X_1 + \cdots + X_n$, then the mgf of Z is

$$M_Z(t) = M_{X_1}(t) \cdots M_{X_n}(t).$$

Theorem 6.6

$\mathbf{X}_1, \ldots, \mathbf{X}_n$ are mutually independent random vectors iff there exist functions $g_i(\mathbf{x}_i)$ for $i = 1, \ldots, n$ such that the joint pdf/pmf of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ can be written as

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = g_1(\mathbf{x}_1) \cdots g_n(\mathbf{x}_n).$$

Additionally, if $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are mutually independent random vectors, and if $h_i(\mathbf{x}_i)$ is a function only of $\mathbf{x}_i$ for $i = 1, \ldots, n$, then the random variables $U_i := g_i(\mathbf{X}_i)$ are mutually independent.

7 Inequalities involving multiple random variables

Like in the previous review session, we will finish by discussing some common inequalities, now involving multiple random variables.

Theorem 7.1 (Hölder's Inequality)

Let X, Y be any two random variables, and let p, q satisfy $p^{-1} + q^{-1} = 1$. Then,

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

Theorem 7.2 (Cauchy-Schwarz Inequality)

For any two random variables X, Y

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

Example 7.3 (covariance inequality)

Let X, Y have means μ_X, μ_Y and variances σ_X^2, σ_Y^2, respectively. We can apply Cauchy-Schwarz to get

$$\mathbb{E}[|(X - \mu_X)(Y - \mu_Y)|] \leq (\mathbb{E}[(X - \mu_X)^2])^{1/2} (\mathbb{E}[(Y - \mu_Y)^2])^{1/2} \implies \mathrm{Cov}(X, Y)^2 \leq \sigma_X^2 \sigma_Y^2$$

This also shows the correlation coefficient ρ indeed satisfies $\rho^2 \in [0, 1]$.

Example 7.4 (Liapounov's Inequality)

Setting $Y \equiv 1$ in Cauchy-Schwarz gives $\mathbb{E}[|X|] \leq (\mathbb{E}[|X|^p])^{1/p}$ for $p \in (1, \infty)$. Replacing $|X|$ by $|X|^r$ and letting $s = pr$ (so that $s > r$) we get

$$\mathbb{E}[|X|^r]^{1/r} \leq \mathbb{E}[|X|^s]^{1/s}, 1 < r < s < \infty$$

Theorem 7.5 (Minkowski's Inequality)

Let X, Y be any two random variables. Then for $1 \leq p < \infty$.

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

Theorem 7.6 (covariance and monotone functions inequality)

Let X be any random variable and $g(x), h(x)$ any functions such that $\mathbb{E}[g(X)]$, $\mathbb{E}[h(X)]$, $\mathbb{E}[g(X)h(X)]$ exist.

1. If $g(x)$ is nondecreasing and $h(x)$ is nonincreasing, then

$$\mathbb{E}[g(X)h(X)] \leq \mathbb{E}[g(X)]\mathbb{E}[h(X)]$$

2. If $g(x), h(x)$ are either both nondecreasing or both nonincreasing, then

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

The intuition behind these inequalities is that in the first case, there is a negative correlation between g and h while in the second case, there is a positive correlation.

8 Problems

8.1 Previous Core Competency Problems

Problem 1 (2018 Summer Practice, # 12). Suppose that $U_1, U_2 \overset{i.i.d.}{\sim} U(0,1)$. Let $V_1 := \max(U_1, U_2)$, $V_2 := \min(U_1, U_2)$.

(a) Find $\mathbb{P}(V_1 \geq x, V_2 \leq y)$, where $x, y \in [0,1]$.

(b) Hence or otherwise find the joint density for (V_1, V_2).

(c) Hence or otherwise compute $\mathbb{E}(V_1^2 + V_2^2)$.

Problem 2 (2018 Summer Practice, # 13). Suppose $X_1, X_2, X_3 \overset{i.i.d.}{\sim} N(0,1)$. Let (Y_1, Y_2, Y_3) be defined as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

(a) Find the joint distribution of (Y_1, Y_2, Y_3).

(b) Show that $Y_1^2 + Y_2^2 + Y_3^2 = X_1^2 + X_2^2 + X_3^2$.

(c) Hence or otherwise derive the distribution of $(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2$, where $\overline{X} = \frac{X_1 + X_2 + X_3}{3}$.

Problem 3 (2019 September, # 4). Let $X_1, \ldots, X_n$ be a random sample (i.i.d.) from a density function f. The corresponding CDF is denoted by F. Denote by $X_{(1)} < \cdots < X_{(n)}$ the order statistics, i.e. a rearrangement of $X_1, \ldots, X_n$ according to their values.

(i) For $n = 2$, derive the density function of $X_{(1)}$ in terms of f and F. [**Hint:** You may want to find the distribution function of $X_{(1)}$ first.]

(ii) For $n = 3$, derive the density function of $X_{(2)}$ in terms of f and F.

(iii) For any n and k, derive the density function of $X_{(k)}$ in terms of f and F.

Problem 4 (2020 May, # 1). Suppose we have a random variable $\xi \sim$ Uniform$(0,1)$. Suppose that conditioning on ξ, we have i.i.d. Bernoulli(ξ) random variables $X_1, X_2, \ldots, X_n, X_{n+1}$, i.e. $P(X_i = 1|\xi) = 1 - P(X = 0|\xi) = \xi$. Calculate

$$P(X_{n+1} = 1 | X_1, \ldots, X_n).$$

Problem 5 (2020 May, # 4). Let Z_1, Z_2, Z_3 be i.i.d. $N(0,1)$ random variables. Let $R = \sqrt{Z_1^2 + Z_2^2 + Z_3^2}$.

(i) Find the distribution of R and write down its density function.

(ii) Suppose that we have two independent random variables $X \sim$ Gamma(α, λ) and $Y \sim$ Gamma(β, λ), where $\alpha, \beta, \lambda > 0$. Let

$$U = X + Y \qquad \text{and} \qquad V = \frac{X}{X+Y}.$$

Find the joint density (p.d.f.) of (U, V) and identify the joint distribution (c.d.f.).

Hint: density function of Gamma$(\alpha, \lambda) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x} / \Gamma(\lambda)$.

Problem 6 (2020 September, # 4). Suppose we generate $U \sim$ Unif$[0,3]$. Let V denote the value of the integer nearest to U (so V takes values in $\{0,1,2,3\}$). Let X denote the rounding error i.e. the absolute distance between U and V.

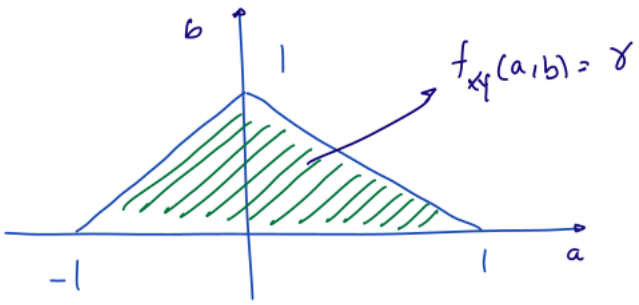
(a) What is the distribution of V?

(b) What is the distribution of the X.

(c) Are X and V independent?

(d) Are X and U independent?

Problem 7 (2021 May, # 4)**.** The joint pdf $f_{X,Y}(a,b)$ of a random variable X and Y is zero outside the triangular region shown below and is equal to a fixed number γ on the triangular region.



Answer the following questions:

(i) Calculate the value of γ.

(ii) Are X and Y independent? Prove your answer and then give an intuitive explanation.

(iii) Calculate $\mathrm{Cov}(X,Y)$. Does the result you obtain make sense?

(iv) Calculate the joint CDF of two random variables $Z = X + Y$ and $W = X - Y$.

8.2 Additional Practice

Problem 8 (Casella & Berger, Exercise 4.17)**.** Let X be an $\exp(1)$ random variable, and define Y to be the integer part of $X + 1$, i.e. $Y = \lfloor X + 1 \rfloor$.

(a) Find the distribution of Y.

(a) Find the conditional distribution of $X - 4$ given $Y \geq 5$.

Problem 9 (Casella & Berger, Exercise 4.54)**.** Find the pdf of $\prod_{i=1}^{n} X_i$ where the X_i's are independent $\mathrm{Unif}([0,1])$ random variables. Hint: try to calculate the cdf, and remember the relationship between uniforms and exponentials.

Problem 10 (Casella & Berger, Exercise 4.59)**.** For any three random variables X, Y, Z with finite variances, prove that

$$\mathrm{Cov}(X,Y) = \mathbb{E}[\mathrm{Cov}(X,Y|Z)] + \mathrm{Cov}(\mathbb{E}[X|Z], \mathbb{E}[Y|Z]),$$

where $\mathrm{Cov}(X,Y|Z)$ is the covariance of X and Y conditional on Z.