4. Handling Multiple Random Variables

4.1 Discrete Multivariate Distributions

Whether we are looking at a random sample of data or modeling the relationship between two or more variables, we must have a probability setting that allows us to consider multiple random variables simultaneously.

**** A word of warning: not only will we be discussing a multiple number of random variables but we will also have parameters and "ordinary" function variables. To complicate matters even further, sometimes variables will be "fixed" and act like parameters. This means it is imperative that one keeps track of how one defines the variables and utilizes them appropriately. This cannot be emphasized enough!*

DEFINITION 4.1 Let $\mathcal{S}$ be a sample space. A random vector $(X_1, \ldots, X_k)$ is a k-dimensional function defined on $\mathcal{S}$. In other words, it is a vector of k random variables, all defined (simultaneously) on $\mathcal{S}$.

Example 4.1 Consider a simple random sample, *with replacement*, from a population with size N. Suppose the response could be Yes, No or Unknown, with proportions p_1, p_2, p_3, respectively (and $p_1 + p_2 + p_3 = 1$).

If s is one of the N^n possible samples, we define

$$\begin{aligned} X_1(s) &= \#\text{ of Yes responses in the sample } s, \\ X_2(s) &= \#\text{ of No responses in the sample } s, \\ X_3(s) &= \#\text{ of Unknown responses in the sample } s. \end{aligned}$$

So (X_1, X_2, X_3) is a trivariate random vector with the requirement that $X_1 + X_2 + X_3 = n$.

Assuming $x_1 + x_2 + x_3 = n$, if we count the number of ways to select x_1 Yes's, x_2 No's and x_3 Unknown's, we obtain $(p_1N)^{x_1}(p_2N)^{x_2}(p_3N)^{x_3}$.

There are $\frac{n!}{x_1!x_2!x_3!}$ ways to order these in the sample, and thus

$$\begin{aligned} \mathsf{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) &= \frac{\frac{n!}{x_1!x_2!x_3!}(p_1N)^{x_1}(p_2N)^{x_2}(p_3N)^{x_3}}{N^n} \\ &= \frac{n!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_3^{x_3}, \quad x_i = 0, 1, \ldots, n \text{ and } x_1 + x_2 + x_3 = n. \end{aligned}$$

(Events separated by commas are meant to indicate *intersection*.) This is the trinomial probability function. It does not depend on the population size N.

The example above illustrates that we can, and must be able to, discuss probabilities involving all the variables in the random vector. To do this generally, we start with a *multivariate* version of cdfs.

DEFINITION 4.2 Let $(X_1, \ldots, X_k)$ be a random vector.

i. The (joint) cumulative distribution function for $(X_1, \ldots, X_k)$ is

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \mathsf{P}(X_1 \leq x_1, \ldots, X_k \leq x_k).$$

ii. The (marginal) cumulative distribution function for each X_i is

$$F_{X_i}(x) = \mathsf{P}(X_i \leq x).$$

(Again, events separated by commas are meant to indicate an *intersection*, unless the word "or" is used in which case it would mean a *union*.)

**** The marginal cdf is in fact the same as the cdf defined previously for a single rv. (See Ch. 2.) The term "marginal" merely means it is for part of the jointly distributed random vector and it can be derived from a joint cdf.*

As in the case of a single rv, the multivariate distribution of a discrete random vector is often described in terms of probabilities for distinct values.

DEFINITION 4.3 Let $(X_1, \ldots, X_k)$ be a random vector, where each X_i is *discrete*.

i. The (joint) probability mass function for $(X_1, \ldots, X_k)$ is

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \mathsf{P}(X_1 = x_1, \ldots, X_k = x_k).$$

ii. The (marginal) probability mass function for each X_i is

$$f_{X_i}(x) = \mathsf{P}(X_i = x).$$

Observe that a random vector is discrete (i.e., has countably many values) iff each component is a discrete random variable.

As we will see, the marginal pmfs can be derived from the joint pmf. But the reverse is not generally true.

**** Be sure to distinguish the different pmfs by subscripting them with the random variable names.*

Example 4.1 (cont.) The joint pmf for (X_1, X_2, X_3) is the trinomial probability function given above.

By letting "success" mean "Yes" and "failure" mean "either No or Unknown", we see that X_1 must have binomial(n, p_1) distribution. Thus the marginal pmf for X_1 is

$$f_{X_1}(x) = \binom{n}{x} p_1^x (1-p_1)^{n-x}, x = 0, 1, \ldots, n.$$

Likewise, X_2 has binomial(n, p_2) pmf and X_3 has binomial(n, p_3) pmf.

Note, however, that knowing the marginal distributions *is not enough* to describe the multivariate probabilities.

Example 4.2 (Marked Poisson) A study of falcons involves capturing the birds and checking if they are currently tagged.

Suppose the number of birds captured is $Y \sim \text{Poisson}(\lambda)$, and of these birds X are tagged. If p is the proportion of tagged falcons in the population (assume large) and $Y = y$ is given, then it is reasonable to consider X to represent the number of successes from a random sample of size y. That is,

$$\mathsf{P}(X = x|Y = y) = \binom{y}{x} p^x (1-p)^{y-x}, 0 \leq x \leq y.$$

By the definition of conditional probability, we obtain the joint pmf as

$$\begin{aligned} f_{X,Y}(x,y) &= \mathsf{P}(X = x, Y = y) = \mathsf{P}(X = x|Y = y)\mathsf{P}(Y = y) \\ &= \binom{y}{x} p^x (1-p)^{y-x} 1_{\{0,\ldots,y\}}(x) \frac{\lambda^y}{y!} \mathrm{e}^{-\lambda} 1_{\{0,1,\ldots\}}(y). \end{aligned}$$

Note the indicators that identify the support for (x, y). In particular, $x \leq y$.

A special case to note is if $Y = 0$. Clearly we should have $\mathsf{P}(X = 0|Y = 0) = 1$, and we must interpret the above accordingly. That is, $\binom{0}{0} = \frac{0!}{0!0!} = 1$.

The marginal for Y is of course the Poisson(λ) pmf, but to identify the marginal for X takes some calculation using the partitions rule. But *also note the restriction:* $X \leq Y$.

As we have done before, the "trick" is to reexpress a sum in terms of something we can recognize. Here it will also involve a change of variables. Thus,

$$\begin{aligned}
f_X(x) &= \mathsf{P}(X = x) = \sum_y \mathsf{P}(\{X = x\} \cap \{Y = y\}) \\
&= \sum_{y:y \geq x} \binom{y}{x} p^x (1-p)^{y-x} 1_{\{0,\ldots,y\}}(x) \frac{\lambda^y}{y!} \mathrm{e}^{-\lambda} 1_{\{0,1,\ldots\}}(y) \\
&= \frac{(p\lambda)^x}{x!} \mathrm{e}^{-\lambda} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\
&= \frac{(p\lambda)^x}{x!} \mathrm{e}^{-p\lambda} \sum_{k=0}^{\infty} \frac{((1-p)\lambda)^k \mathrm{e}^{-(1-p)\lambda}}{k!} \\
&= \frac{(p\lambda)^x}{x!} \mathrm{e}^{-p\lambda},
\end{aligned}$$

where we have summed a Poison($(1-p)\lambda$) pmf.

This says that $X \sim$ Poisson$(p\lambda)$.

The last computation demonstrates the relationship between joint pmfs and marginal pmfs. In particular, the marginal pmfs can be derived from the joint pmf.

THEOREM 4.4 Suppose (X, Y) has joint pmf $f_{X,Y}(x, y)$. Then

$$f_X(x) = \sum_y f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_x f_{X,Y}(x, y).$$

More generally, suppose $(X_1, \ldots, X_k)$ has joint pmf $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$. Then

$$f_{X_i}(x_i) = \sum_{x_j, j \neq i} f(x_1, \ldots, x_k).$$

In other words, the marginal pmf is obtained by *summing out* the other variable(s).

PROOF This is the partitions rule, Thm. 1.11, just as in Ex. 4.2. □

Note, however, that one may be able to identify a marginal pmf on the basis of other information without computing as in Thm. 4.4. See, for example, Ex. 4.1.

Ex. 4.2 also involved conditional probabilities. Thinking of them as functions of values of the random variables also leads to the following.

DEFINITION 4.5 Suppose (X, Y) has joint pmf $f_{X,Y}(x, y)$. The conditional probability mass function for X, given $Y = y$, is

$$f_{X|Y}(x|y) = \mathsf{P}(X = x|Y = y) = \begin{cases} \dfrac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0, \\ f_X(x) & \text{if } f_Y(y) = 0. \end{cases}$$

(The second part of this definition, for the case $f_Y(y) = 0$, is needed for completeness but in fact can be any reasonable pmf. Sometimes a "continuous extension" of the first part is sensible.)

Additionally, we have the *multiplication rule*

$$f_{X,Y}(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x) \quad \text{for all } x, y.$$

Observe that $f_Y(y) = 0$ implies $f_{X,Y}(x, y) = 0$.

Example 4.1 (cont.) Consider again the trinomial vector (X_1, X_2, X_3). Since $X_3 = n - X_1 - X_2$, it suffices to replace x_3 with $n - x_1 - x_2$. Then

$$f_{X_1,X_2}(x_1, x_2) = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} p_3^{n-x_1-x_2}, \quad x_1 + x_2 \leq n.$$

As we know $p_3 = 1 - p_1 - p_2$ and

$$f_{X_1}(x_1) = \binom{n}{x_1} p_1^{x_1} (1 - p_1)^{n-x_1}, x_1 = 0, 1, \ldots, n,$$

we obtain

$$\begin{aligned} f_{X_2|X_1}(x_2|x_1) &= \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} = \frac{\frac{n!}{x_1!x_2!(n-x_1-x_2)!} p_1^{x_1} p_2^{x_2} p_3^{n-x_1-x_2}}{\binom{n}{x_1} p_1^{x_1} (1 - p_1)^{n-x_1}} \\ &= \binom{n - x_1}{x_2} \left(\frac{p_2}{1 - p_1}\right)^{x_2} \left(1 - \frac{p_2}{1 - p_1}\right)^{(n-x_1)-x_2}, \end{aligned}$$

for $x_2 = 0, \ldots, n - x_1$. In other words, the binomial$(n - x_1, p_2/(1 - p_1))$ pmf.

Observe that, for each trial, P("No" | "not Yes") $= \frac{p_2}{1-p_1}$. Essentially, if we are given $X_1 = x_1$ then the number of No's in the remainder of the sample is the number of successes where the success rate is the conditional probability of a No, given that the response is not a Yes.

THEOREM 4.6 Suppose (X, Y) have joint discrete distribution. For fixed y, the conditional pmf $f_{X|Y}(x|y)$ is a valid pmf in x.

PROOF Obviously $f_{X|Y}(x|y) \geq 0$. By Def. 4.5 and Thm. 4.4,

$$\sum_x f_{X|Y}(x|y) = \frac{\sum_x f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1.$$

The conclusion follows by Thm. 2.8.i. □

**** This means that whenever you compute or identify a conditional pmf, you should always check that is indeed a pmf (as a function of the variable in front of the vertical bar |).*

The event you condition on (the variable behind |) *acts like a parameter* of the distribution: in general, $f_{X|Y}(x|y_1)$ and $f_{X|Y}(x|y_2)$ give different distributions when $y_1 \neq y_2$.

Thinking conditionally naturally leads to questions of independence.

DEFINITION 4.7 Random variables $X_1, \ldots, X_k$ are independent if

$$\mathsf{P}(X_1 \in A_1, \ldots, X_k \in A_k) = \mathsf{P}(X_1 \in A_1) \cdots \mathsf{P}(X_k \in A_k),$$

for all subsets $A_1, \ldots, A_k$.

Equivalently, $X_1, \ldots, X_k$ are independent if the joint cdf *factors*:

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \prod_{i=1}^{k} F_{X_i}(x_i) \quad \text{for all } x_1, \ldots, x_k.$$

In other words, any k events about the separate variables are independent. Note that this definition is stronger than the one for independence of k specific events, because it is required to hold more generally.

Example 4.3 Roll two fair dice and let X_i be the result showing on dice $\#i$, $i = 1, 2$. We know that $\mathsf{P}(X_1 = j, X_2 = k) = \frac{1}{36}$, $\mathsf{P}(X_1 = j) = \frac{1}{6}$ and $\mathsf{P}(X_2 = k) = \frac{1}{6}$, for all $j, k = 1, \ldots, 6$. By summing,

$$\begin{aligned}
\mathsf{P}(X_1 \in A, X_2 \in B) &= \sum_{j \in A} \sum_{k \in B} \mathsf{P}(X_1 = j, X_2 = k) \\
&= \sum_{j \in A} \sum_{k \in B} \mathsf{P}(X_1 = j)\mathsf{P}(X_2 = k) \\
&= \sum_{j \in A} \mathsf{P}(X_1 = j) \sum_{k \in B} \mathsf{P}(X_2 = k) = \mathsf{P}(X_1 \in A)\mathsf{P}(X_2 \in B).
\end{aligned}$$

Therefore X_1 and X_2 are independent and we say the two dice are independent.

Now let $Y = X_1 + X_2$. It is easy to show that

$$\mathsf{P}(X_1 = j, Y = 7) = \mathsf{P}(X_1 = j)\mathsf{P}(Y = 7)$$

but this does not make X_1 and Y independent. It is quite clear that

$$\mathsf{P}(X_1 = j, Y = k) \neq \mathsf{P}(X_1 = j)\mathsf{P}(Y = k)$$

for any $k \neq 7$ (except those with probability 0). So X_1 and Y are not independent rvs.

THEOREM 4.8 Suppose (X, Y) has joint pmf $f_{X,Y}(x, y)$. Then

i. X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad \text{for all } x, y \text{ (including when equal to 0).}$$

ii. X and Y are independent if and only if

$$f_{X|Y}(x|y) = f_X(x) \quad \text{for all } x \text{ and all } y \text{ with } f_Y(y) > 0.$$

iii. X and Y are independent if and only if $f_{X,Y}(x, y) = g(x)h(y)$ for some nonnegative functions $g(x)$, $h(y)$ (including required indicator functions).

PROOF

i. We have essentially done the "if" part in Ex. 4.3:

$$\mathsf{P}(X \in A, Y \in B) = \sum_{x \in A} \sum_{y \in B} f_{X,Y}(x, y) = \sum_{x \in A} \sum_{y \in B} f_X(x) f_Y(y)$$

$$= \sum_{x \in A} f_X(x) \sum_{y \in B} f_Y(y) = \mathsf{P}(X \in A)\mathsf{P}(Y \in B).$$

The "only if" part is required by the definition of independence.

ii., iii. (exercise.) □

Example 4.1 (cont.) Clearly the three rvs in the trinomial random vector (X_1, X_2, X_3) are not independent since they sum to a fixed constant n: you can compute each from the other two.

Even though the joint pmf $(\frac{n!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_3^{x_3})$ *appears* to factor – why does it not? Hint: there is also an indicator $I_{\{n\}}(x_1 + x_2 + x_3)$. Indeed,

$$\mathsf{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = 0 \quad \text{if } x_1 + x_2 + x_3 \neq n,$$

even if the individual events all have positive probabilities.

What about just X_1 and X_2? Are they independent?

Solution First, we note that $X_1 + X_2 \leq n$, which seems to imply that one rv is going to constrain the other. This observation alone is almost enough to prove dependence. Just note further that each rv has positive chance of equaling n but they cannot both equal n at the same time.

We look more closely, however, at the conditional pmfs. Recall that, given $X_1 = x_1$, the conditional pmf for X_2 is a binomial$(n - x_1, p_2/(1 - p_1))$ pmf. This depends on x_1 so it clearly cannot be the same as the unconditional binomial(n, p_2) pmf that is the marginal pmf for X_2. Hence X_1 and X_2 are not independent.

(We can also observe that the joint pmf does not factor.)

**** Generally, it is not safe to proceed as if the rvs are independent, unless it has been demonstrated that they are or it has explicitly been assumed.*

Whether the rvs are independent or not, we often need to calculate probabilities that do not split the variables into separate events.

Example 4.3 (cont.) Roll two fair dice. We let (X_1, X_2) be the results and Y be the total. To get the pmf for Y, we will use the joint pmf for (X_1, X_2).

$$\begin{aligned} f_Y(y) &= \mathsf{P}(X_1 + X_2 = y) = \sum_{x_1+x_2=y} f_{X_1,X_2}(x_1, x_2) \\ &= \sum_{x_1} \sum_{x_2=y-x_1} f_{X_1,X_2}(x_1, x_2) = \sum_{x_1} f_{X_1,X_2}(x_1, y - x_1) \\ &= \sum_{x_1} \frac{1}{36} 1_{\{1,\ldots,6\}}(x_1) 1_{\{1,\ldots,6\}}(y - x_1) = \frac{6 - |7 - y|}{36} 1_{\{2,\ldots,12\}}(y). \end{aligned}$$

The final equality is obtained by counting the possibilities for which $1 \leq x_1 \leq 6$ and $1 \leq y - x_1 \leq 6$ (check the cases $y < 7$ and $y \geq 7$ separately).

This last example illustrates a very useful relationship for sums of discrete random variables.

THEOREM 4.9 Suppose (X, Y) has joint pmf $f_{X,Y}(x, y)$ and let $T = X + Y$.

i. The pmf for T is given by

$$f_T(t) = \sum_x f_{X,Y}(x, t - x) = \sum_y f_{X,Y}(t - y, y).$$

ii. (Convolution Formula) If X and Y are *independent* then

$$f_T(t) = \sum_x f_X(x) f_Y(t - x) = \sum_y f_X(t - y) f_Y(y).$$

PROOF Again, this was effectively argued in the last example. □

**** The end result must be a function only of the variable on the left (here, t).*

Note: the term "convolution" applies *only when the joint function factors.* In other words it is an operation between two functions (the marginal pmfs) that results in another function.

A simple graph helps to explain the formula for the pmf of a sum (whether or not it is a convolution).

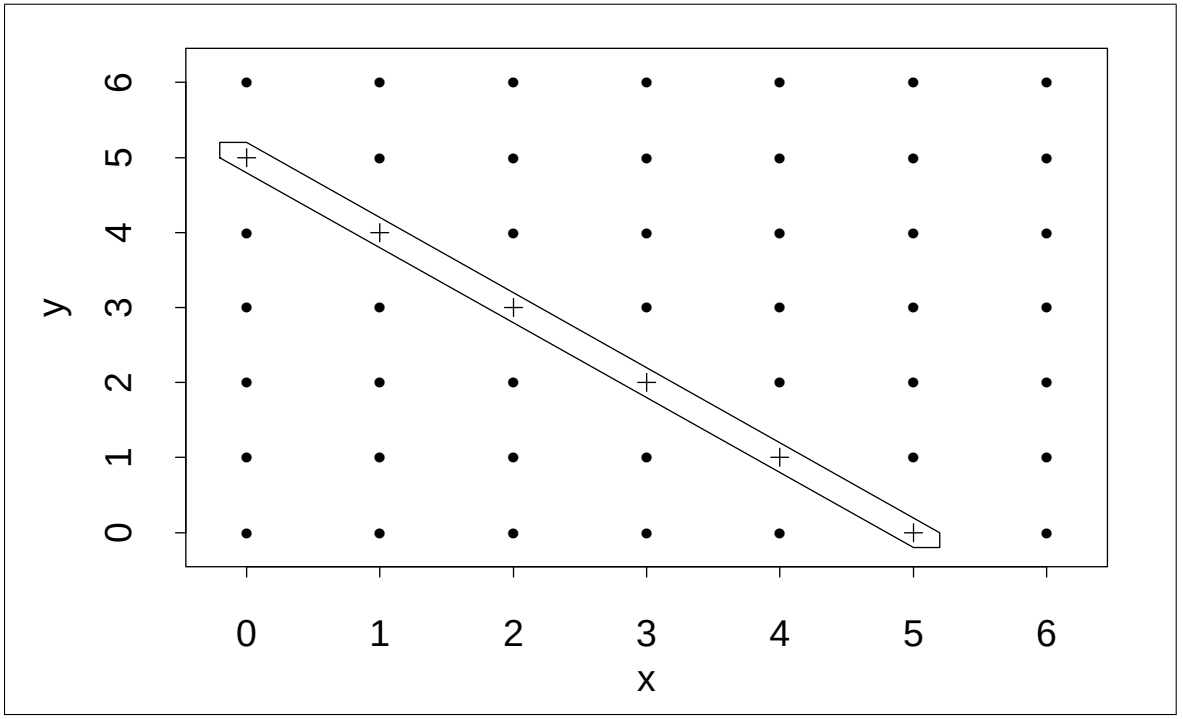

Figure 4.1 Points (x, y) to be summed in the pmf for the sum of two nonnegative random variables in the case for $t = x + y = 5$.

THEOREM 4.10 (Sum of Independent Poisson RVs) Suppose $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are *independent*. Then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

You can remember the parameter (which, for Poisson, is the mean) by $\mathsf{E}(X + Y) = \mathsf{E}(X) + \mathsf{E}(Y) = \lambda + \mu$.

PROOF Let $T = X + Y$. By the convolution formula, for $t = 0, 1, \ldots,$

$$\begin{aligned} f_T(t) &= \sum_x f_X(x) f_Y(t-x) = \sum_{x=0}^{\infty} \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \frac{\mu^{t-x} \mathrm{e}^{-\mu}}{(t-x)!} 1_{\{0,1,\ldots\}}(t-x) \\ &= \frac{(\lambda+\mu)^t \mathrm{e}^{-(\lambda+\mu)}}{t!} \sum_{x=0}^{t} \binom{t}{x} \left(\frac{\lambda}{\lambda+\mu}\right)^x \left(\frac{\mu}{\lambda+\mu}\right)^{t-x} \\ &= \frac{(\lambda+\mu)^t \mathrm{e}^{-(\lambda+\mu)}}{t!}. \end{aligned}$$

(Check that this is a function only of t.) This is the Poisson$(\lambda + \mu)$ pmf. □

Here, the sum of independent rvs has the same "type" of distribution as do the two terms. This is a very nice property, but *not always true*. Other examples with this property include the binomial and negative binomial families if p is fixed (exercise).

The method demonstrated by convolution can be used to find the pmf of any function of discrete rvs X and Y since

$$f_{g(X,Y)}(w) = \mathsf{P}(g(X,Y) = w) = \sum_{g(x,y)=w} f_{X,Y}(x,y).$$

The tricky, but crucial, part is being careful to sum over pairs (x, y) that satisfy the constraint $g(x, y) = w$ (where w is fixed). *The result depends only on w.*

For some problems, however, it may be easier to compute the cdf.

Example 4.4 Let $X \sim$ geometric(p) and $Y \sim$ geometric(q) be independent and define $W = \min(X, Y)$. (The first to have a success "wins".) We recall (Ex. 2.2) that

$$\mathsf{P}(X > x) = (1-p)^x \quad \text{and} \quad \mathsf{P}(Y > y) = (1-q)^y.$$

Since $\min(x, y) > w \iff x > w$ and $y > w$,

$$\mathsf{P}(W > w) = \mathsf{P}(X > w, Y > w) = \mathsf{P}(X > w)\mathsf{P}(Y > w) = ((1-p)(1-q))^w,$$

by independence. (Check that this is function only of w.) This says that W must have geometric$(1-(1-p)(1-q))$ distribution.

4.2 Continuous Multivariate Distributions

We will now reiterate the results of the previous section for continuous rvs. There are, however, several critical differences.

DEFINITION 4.11 The random vector $(X_1, \ldots, X_k)$ has (absolutely) continuous distribution with (joint) pdf $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$ if

$$\mathsf{P}(a_1 \leq X_1 \leq b_1, \ldots, a_k \leq X_k \leq b_k) = \int_{a_k}^{b_k} \cdots \int_{a_1}^{b_1} f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \mathrm{d}x_1 \cdots \mathrm{d}x_k,$$

for all $a_i \leq b_i$, $i = 1, \ldots, k$.

In this case,

$$\mathsf{P}((X_1, \ldots, X_k) \in A) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 1_A(x_1, \ldots, x_k) f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \mathrm{d}x_1 \cdots \mathrm{d}x_k,$$

for any (Borel) subset A of $\mathbb{R}^k$.

The joint cdf is $F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \mathsf{P}(X_1 \leq x_1, \ldots, X_k \leq x_k)$ and

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F_{X_1,\ldots,X_k}(x_1, \ldots, x_k).$$

Once again, we can talk about the "marginal" distributions.

THEOREM 4.12 Suppose (X, Y) has joint pdf $f_{X,Y}(x, y)$. Then X and Y are each (absolutely) continuous with marginal pdfs

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}x.$$

More generally, suppose $(X_1, \ldots, X_k)$ has joint pdf $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$. Then X_i is continuous with pdf

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \cdots \mathrm{d}x_k.$$

That is, the marginal pdf is obtained by *integrating out* the other variable(s).

PROOF Note that

$$\begin{aligned}\int_a^b f_X(x)\mathrm{d}x &= \mathsf{P}(a \le X \le b) = \mathsf{P}(a \le X \le b, -\infty < Y < \infty) \\ &= \int_a^b \Big(\int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y \Big)\mathrm{d}x.\end{aligned}$$

Since this is true for any a and b, it must be that the inner integral on the right is the pdf of X. □

Example 4.5 Suppose (X, Y) has joint pdf

$$f_{X,Y}(x,y) = \tfrac{1}{2}\Big(\lambda^2 \mathrm{e}^{-\lambda(x+y)} + \mu^2 \mathrm{e}^{-\mu(x+y)}\Big) 1_{(0,\infty)}(x) 1_{(0,\infty)}(y).$$

Then the marginal of X (and also of Y) is

$$\begin{aligned} f_X(x) &= \int_0^\infty \tfrac{1}{2}\Big(\lambda^2 \mathrm{e}^{-\lambda(x+y)} + \mu^2 \mathrm{e}^{-\mu(x+y)}\Big) 1_{(0,\infty)}(x) \mathrm{d}y \\ &= \tfrac{1}{2}\Big(\lambda \mathrm{e}^{-\lambda x} + \mu \mathrm{e}^{-\mu x}\Big) 1_{(0,\infty)}(x). \end{aligned}$$

(This integrates to 1, so both it and $f_{X,Y}$ are legitimate pdfs. In fact, it is just the average of two exponential densities.)

Why does Y have the same marginal pdf as X in this case?

Calculating a probability involves a *double* integral. For example,

$$\begin{aligned} \mathsf{P}(X > x, Y > y) &= \int_y^\infty \int_x^\infty f_{X,Y}(s,t) \mathrm{d}s \mathrm{d}t \\ &= \int_y^\infty \int_x^\infty \tfrac{1}{2}\Big(\lambda^2 \mathrm{e}^{-\lambda(s+t)} + \mu^2 \mathrm{e}^{-\mu(s+t)}\Big) \mathrm{d}s \mathrm{d}t \\ &= \tfrac{1}{2}\Big(\mathrm{e}^{-\lambda(x+y)} + \mathrm{e}^{-\mu(x+y)}\Big). \end{aligned}$$

As in the discrete case, the marginal distributions can be derived from the joint distribution, but the joint distribution cannot be derived from the marginals – at least not without additional information such as independence.

One thing that differs from the discrete case is the following. Having marginal pdfs does not imply the joint distribution is absolutely continuous: a joint pdf need not even exist!

Example 4.6 Suppose X has a normal(0,1) distribution and let $Y = X^2$. We know that Y has chi-square(1) distribution which is also continuous.

But there is *no function* $f(x, y)$ that will give the joint cdf $\mathrm{P}(X \leq x, Y \leq y)$ as a double integral. Indeed, it can only be calculated as a single integral using the marginal distribution of X.

There are other, more general examples. Suppose Z is Poisson(1), independent of $X \sim$ normal$(0, 1)$ and define $W = X + Z$. Then (it can be shown) W has a pdf also but (X, W) does not have a joint pdf. (In this case you can back up and rewrite any probability, or expectation, about (X, W) in terms of one about (X, Z). Something like that is often true for the examples one is likely to encounter, but not always.)

Conditional pdfs are defined analogously to the way that conditional pmfs are. Their interpretation, however, is slightly different.

DEFINITION 4.13 Suppose (X, Y) has joint pdf $f_{X,Y}(x, y)$. The conditional probability density function for X, given $Y = y$, is

$$f_{X|Y}(x|y) = \begin{cases} \dfrac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0, \\ f_X(x) & \text{if } f_Y(y) = 0. \end{cases}$$

(The second part of this definition, for the case $f_Y(y) = 0$, is needed for completeness but in fact can be any reasonable pdf.)

Additionally, we have the *multiplication rule*

$$f_{X,Y}(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x) \quad \text{for all } x, y.$$

Furthermore, the conditional distribution of X, given $Y = y$, is determined by conditional probabilities such as

$$\mathsf{P}(a \leq X \leq b|Y = y) = \int_a^b f_{X|Y}(x|y)\mathrm{d}x.$$

Note that $\mathsf{P}(Y = y) = 0$. So, technically, $\mathsf{P}(X \in A|Y = y)$ has not been defined. We can give a heuristic explanation as follows. Let $\delta > 0$ be small and suppose the joint density is continuous. Then

$$\begin{aligned}\mathsf{P}(a \leq X \leq b|y \leq Y \leq y+\delta) &= \frac{\int_y^{y+\delta}\int_a^b f_{X,Y}(x,u)\mathrm{d}x\mathrm{d}u}{\int_y^{y+\delta} f_Y(u)\mathrm{d}u} \\ &\doteq \frac{\delta\int_a^b f_{X,Y}(x,y)\mathrm{d}x}{\delta f_Y(y)} = \int_a^b f_{X|Y}(x|y)\mathrm{d}x.\end{aligned}$$

THEOREM 4.14 Suppose (X, Y) has joint continuous distribution. For *fixed* y, the conditional pdf $f_{X|Y}(x|y)$ is a valid pdf in x.

PROOF This is proven analogously to Thm. 4.6. □

**** You should always check that $f_{X|Y}(x|y)$ is indeed a pdf (as a function of x).*

The variable y in the above *acts like a parameter*: in general, $f_{X|Y}(x|y_1)$ and $f_{X|Y}(x|y_2)$ give different distributions when $y_1 \neq y_2$.

(Conditional distributions can be defined in general for any vector of rvs regardless of type; it is how to compute them that is the crucial problem.)

Example 4.7 Suppose (X, Y) has joint pdf

$$f(x,y) = \frac{8}{\pi}(x^2+y^2), \quad \text{for } x \geq 0,\ y \geq 0,\ x^2+y^2 \leq 1.$$

(Check that this is a legitimate joint pdf – use polar coordinates.)

Find the marginal pdf for X and the conditional pdf for Y, given $X = x$.

Solution For fixed $x \in [0,1]$, we have $0 \leq y \leq \sqrt{1-x^2}$. Therefore,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y = \frac{8}{\pi}\int_0^{\sqrt{1-x^2}} (x^2+y^2)\,\mathrm{d}y \\ &= \frac{8}{3\pi}(1+2x^2)(1-x^2)^{1/2}, \quad 0 \leq x \leq 1. \end{aligned}$$

From this it follows that

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{3(x^2+y^2)}{(1+2x^2)(1-x^2)^{1/2}}, \quad 0 \leq y \leq \sqrt{1-x^2}.$$

Again, we interpret the conditional pdf as a pdf in y, with x fixed like a parameter. In this case *both the form of the distribution and its support depend on* x.

Example 4.8 Recall the Poisson process discussed in Sec. 3.1. Let T_i be the time until the i-th occurrence. What is the joint pdf for (T_1, T_2)?

Solution Given the memoryless nature of the exponential distribution, it is natural to think that if $T_1 = t_1$ then it will be another exponential$(1/\lambda)$ time until the next occurrence. That is, we might anticipate that, given $T_1 = t_1$, T_2 is like $X + t_1$ where $X \sim$ exponential$(1/\lambda)$.

In other words, we anticipate that the conditional pdf is a *shifted* exponential pdf, with location parameter t_1:

$$f_{T_2|T_1}(t_2|t_1) = \lambda \mathrm{e}^{-\lambda(t_2-t_1)} 1_{(t_1,\infty)}(t_2).$$

Since we already know that $T_1 \sim$ exponential$(1/\lambda)$, we can then obtain the joint pdf using the multiplication rule:

$$\begin{aligned} f_{T_1,T_2}(t_1,t_2) = f_{T_1}(t_1) f_{T_2|T_1}(t_2|t_1) &= \lambda \mathrm{e}^{-\lambda t_1} 1_{(0,\infty)}(t_1) \lambda \mathrm{e}^{-\lambda(t_2-t_1)} 1_{(t_1,\infty)}(t_2) \\ &= \lambda^2 \mathrm{e}^{-\lambda t_2} 1_{(0,\infty)}(t_1) 1_{(t_1,\infty)}(t_2). \end{aligned}$$

This argument is legitimate *only* if we know our assumption above about exponential times between occurrences is correct. In fact, the memoryless property is critical and the argument cannot be generalized beyond this example.

We now check whether the conclusion is correct by using the properties of the Poisson process as described before. Recall that intervals of length t have Poisson(λt) number of occurrences and disjoint intervals are independent.

First, $T_1 \leq t_1 < t_2 < T_2$ iff there is exactly one occurrence in the interval $[0, t_1]$ and none in the interval $[t_1, t_2]$. This implies (exercise),

$$\mathsf{P}(T_1 \leq t_1 < t_2 < T_2) = \lambda t_1 \mathrm{e}^{-\lambda t_2} \quad \text{for } 0 < t_1 < t_2,$$

and therefore (exercise),

$$\begin{aligned} F_{T_1,T_2}(t_1, t_2) &= \mathsf{P}(T_1 \leq t_1, T_2 \leq t_2) = \mathsf{P}(T_1 \leq t_1) - \mathsf{P}(T_1 \leq t_1 < t_2 < T_2) \\ &= 1 - \mathrm{e}^{-\lambda t_1} - \lambda t_1 \mathrm{e}^{-\lambda t_2}. \end{aligned}$$

Thus,

$$f_{T_1,T_2}(t_1, t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} F_{T_1,T_2}(t_1, t_2) = \lambda^2 \mathrm{e}^{-\lambda t_2} \quad \text{for } 0 < t_1 < t_2.$$

On the other hand, if $0 < t_2 \leq t_1$ then, using Thm. 3.4,

$$\mathsf{P}(T_1 \leq t_1, T_2 \leq t_2) = \mathsf{P}(T_2 \leq t_2) = 1 - (1 + \lambda t_2)\mathrm{e}^{-\lambda t_2},$$

which confirms that

$$f_{T_1,T_2}(t_1, t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} F_{T_1,T_2}(t_1, t_2) = 0 \quad \text{for } 0 < t_2 \leq t_1.$$

For independence of continuous rvs we have:

THEOREM 4.15 Suppose (X, Y) has joint pdf $f_{X,Y}(x, y)$. Then

i. X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y \text{ (including when 0).}$$

ii. X and Y are independent if and only if

$$f_{X|Y}(x|y) = f_X(x) \quad \text{for all } x \text{ and all } y \text{ with } f_Y(y) > 0.$$

PROOF This is proven analogously to Thm. 4.8. □

**** You actually need only verify that the joint pdf factors somehow (including the indicator functions), say $f_{X,Y}(x, y) = g(x)h(y)$, because if that is the case then you can always rearrange multiplicative constants to ensure that each factor integrates to 1.*

Example 4.9 Suppose $f_{X,Y}(x,y) = ce^{-(x^2+y^2)/2}$ for some $c > 0$. We can write this as

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

which looks like the product of two standard normal densities. Indeed, it must be, because no other function proportional to this would integrate to 1. Therefore, $c = 1/(2\pi)$ and $f_{X,Y}(x,y)$ is the joint pdf of independent normal(0,1) rvs.

Now suppose $f_{X,Y}(x,y) = ce^{-(x^2-xy+2y^2)/2}$. What can we say about the relationship between X and Y?

Solution Completing the square (in x), we rearrange it to be a density in x that possibly also depends on y, multiplied by a density in y that depends only on y:

$$f_{X,Y}(x,y) = ce^{-(x-(y/2))^2/2-7y^2/8}$$

$$= \left(\frac{1}{\sqrt{2\pi}}e^{-(x-(y/2))^2/2}\right)\left(\frac{1}{\sqrt{2\pi}\sqrt{4/7}}e^{-7y^2/8}\right)2c\pi\sqrt{4/7}.$$

The first factor is the normal$(y/2, 1)$ pdf, as a function of x (with mean $y/2$). Integrating x out leaves us with the second factor which is the normal$(0, 4/7)$ pdf, and so we also have $2c\pi\sqrt{4/7} = 1$.

We conclude, therefore, that $Y \sim$ normal$(0, 4/7)$ and $X|\{Y = y\} \sim$ normal$(y/2, 1)$. X and Y are not independent.

Likewise, $X \sim$ normal$(0, 8/7)$ and $Y|\{X = x\} \sim$ normal$(x/4, 1/2)$ (exercise). (By the way, the shorthand "$Y|\{X = x\}$" is *not a random variable*. It merely means we are talking about a conditional distribution.)

**** This shows that sometimes you can reason out the marginal and conditional pdfs without ever doing an integral!*

Can $f(x, y) = ce^{-(x^2-3xy+2y^2)/2}$ be a joint pdf for some c? Why or why not?

While we are on the subject of independence, we mention the following "obvious" point, which is valid regardless of whether the rvs are discrete or continuous or anything else.

THEOREM 4.16 Suppose $X_1, \ldots, X_k$ are independent rvs. Then $g_1(X_1)$, $\ldots, g_k(X_k)$ are independent for any functions g_i, $i = 1, \ldots, k$.

PROOF Given any subsets $B_i \subset \mathbb{R}$, let $A_i = g_i^{-1} B_i$, $i = 1, \ldots, k$. By the definition of g_i^{-1}, we know $X_i \in A_i \iff g_i(X_i) \in B_i$.

By Def. 4.7,

$$\{X_1 \in A_1\} = \{g_1(X_1) \in B_1\}, \ldots, \{X_k \in A_k\} = \{g_k(X_k) \in B_k\}$$

are independent events (for all subsets B_i) and therefore $g_1(X_1), \ldots, g_k(X_k)$ are independent random variables, again by Def. 4.7. □

Once again, sums of random variables are important, and especially sums of independent random variables. So we have a continuous version of the convolution formula.

THEOREM 4.17 Suppose (X, Y) has joint pdf $f_{X,Y}(x, y)$ and let $T = X + Y$.

i. The pdf for T is given by

$$f_T(t) = \int_{-\infty}^{\infty} f_{X,Y}(x, t-x)\mathrm{d}x = \int_{-\infty}^{\infty} f_{X,Y}(t-y, y)\mathrm{d}y.$$

ii. (Convolution Formula) If X and Y are independent then

$$f_T(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t-x)\mathrm{d}x = \int_{-\infty}^{\infty} f_X(t-y) f_Y(y)\mathrm{d}y.$$

As in the discrete case "convolution" is an operation between two functions and only applies in the independence case.

**** Beware: the convolution formula differs, depending on whether the rvs are continuous or discrete. Pay attention to context!*

PROOF We start by obtaining the cdf $\mathsf{P}(T \leq t)$. Computing this probability requires integrating $f_{X,Y}(x,y)$ over the subset $\{(x,y) : x+y \leq t\}$. That is (see Fig. 4.2 on the next slide),

$$\begin{aligned}\mathsf{P}(T \leq t) &= \int_{-\infty}^{\infty}\int_{-\infty}^{t-x} f_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x = \int_{-\infty}^{\infty}\int_{-\infty}^{t} f_{X,Y}(x,u-x)\mathrm{d}u\mathrm{d}x \\ &= \int_{-\infty}^{t}\left(\int_{-\infty}^{\infty} f_{X,Y}(x,u-x)\mathrm{d}x\right)\mathrm{d}u,\end{aligned}$$

where we have made a change of variables $u = y + x$ and then exchanged the order of integration.

This suffices since it indicates the inner integral is the pdf for T.

The second part of the theorem comes from applying the independence of X and Y to the first part. □

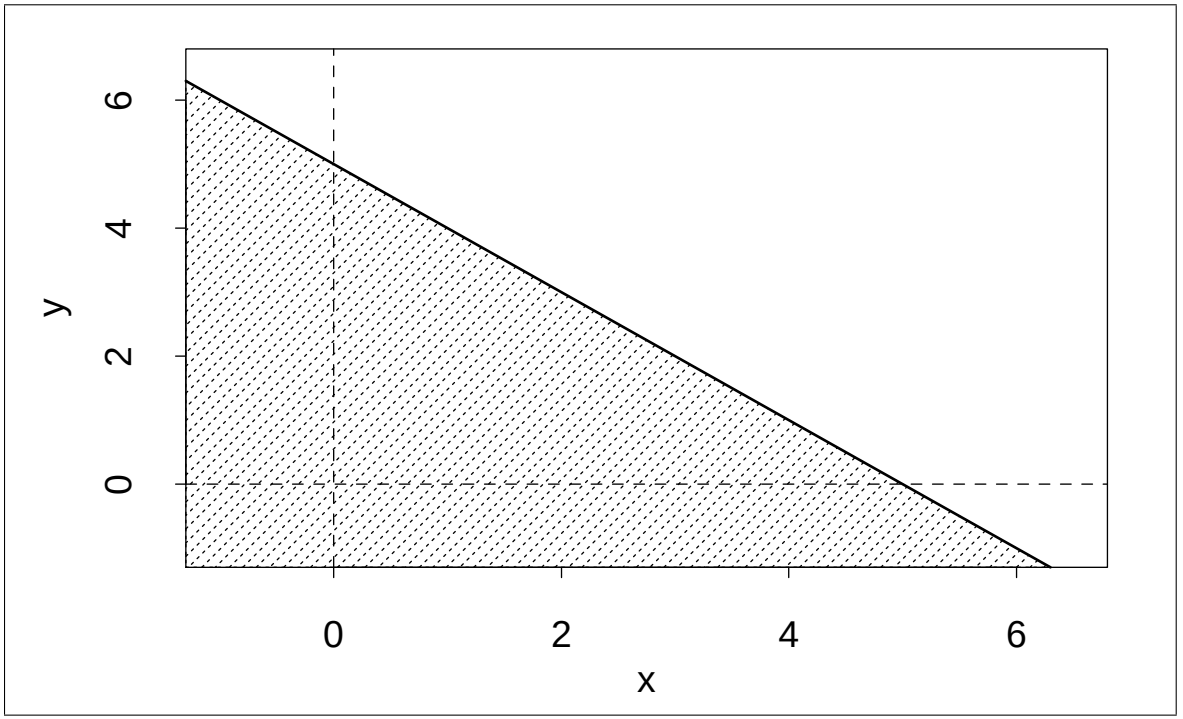

Figure 4.2 Region to integrate over for $\mathrm{P}(X + Y \leq t)$, for $t = 5$.

Example 4.10 Suppose $X \sim$ exponential(β) and $Y \sim$ gamma(α, β), and they are independent. Let $T = X + Y$.

Note that $0 < Y < T$ since both X and Y are positive. Then, for $t > 0$,

$$\begin{aligned} f_T(t) &= \int_{-\infty}^{\infty} \frac{1}{\beta} \mathrm{e}^{-x/\beta} 1_{(0,\infty)}(x) \frac{1}{\Gamma(\alpha)\beta^{\alpha}} (t-x)^{\alpha-1} \mathrm{e}^{-(t-x)/\beta} 1_{(0,\infty)}(t-x) \mathrm{d}x \\ &= \frac{\mathrm{e}^{-t/\beta}}{\Gamma(\alpha)\beta^{\alpha+1}} \int_0^t (t-x)^{\alpha-1} \mathrm{d}x = \frac{\mathrm{e}^{-t/\beta}}{\alpha\Gamma(\alpha)\beta^{\alpha+1}} \int_0^t \alpha x^{\alpha-1} \mathrm{d}x \\ &= \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} t^{\alpha} \mathrm{e}^{-t/\beta}. \end{aligned}$$

Therefore, $T \sim$ gamma$(\alpha + 1, \beta)$.

Now suppose $X_1, X_2, \ldots, X_n$ are independent exponential(β) random variables. Since exponential(β) = gamma$(1, \beta)$, we can iterate the above to conclude $X_1 + X_2 + \cdots + X_n \sim$ gamma(n, β).

Example 4.11 Let $V \sim \text{normal}(\lambda, \sigma^2)$ and $W \sim \text{normal}(\mu, \sigma^2)$, independent, and let $X = V + W$. Find the distribution of X.

Solution By the convolution formula,

$$f_X(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(v-\lambda)^2/(2\sigma^2)} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-v-\mu)^2/(2\sigma^2)} \mathrm{d}v.$$

To complete the square (in v) in the exponent, we note a basic identity:

$$\begin{aligned} (v-a)^2 + (v-b)^2 &= 2v^2 - 2(a+b)v + \tfrac{(a+b)^2}{2} + a^2 + b^2 - \tfrac{(a+b)^2}{2} \\ &= 2\left(v - \tfrac{a+b}{2}\right)^2 + \tfrac{(a-b)^2}{2}. \end{aligned}$$

Therefore,

$$\begin{aligned} f_X(x) &= \frac{1}{2\sqrt{\pi}\sigma} \mathrm{e}^{-(x-\lambda-\mu)^2/(4\sigma^2)} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}\sigma} \mathrm{e}^{-(v-(x+\lambda-\mu)/2)^2/\sigma^2} \mathrm{d}v \\ &= \frac{1}{2\sqrt{\pi}\sigma} \mathrm{e}^{-(x-\lambda-\mu)^2/(4\sigma^2)}. \end{aligned}$$

This says $X \sim \text{normal}(\lambda + \mu, 2\sigma^2)$.

Note that we also could have simply looked at the integrand after completing the square to see that the answer would be a normal density – and then identified the parameters for X immediately.

- When it turns out that the pdf or pmf has have a *known form*, you can often directly deduce what the parameters must be just by looking at the expression. Nevertheless, it is a good idea to keep track of constants as you compute because it will help ensure you get the right parameter values in the end.
- Thm. 4.17 can be extended to *linear combinations* of X and Y simply by modifying the proof slightly. See Cor. 4.19 below.
- Much more general, however, is the problem of how to handle all kinds of transformations of the vector (X, Y). The basic method, as always, is to *identify the cdf* and then derive the pdf or pmf, as appropriate, from that. For example, if $W = g(X, Y)$ then we would find $F_W(w) = \mathsf{P}(g(X, Y) \leq w)$.
- But if (X, Y) have a *joint pdf* and $g(x, y)$ is *sufficiently smooth* then we would expect W to have a continuous distribution with some pdf. And so we would want to find the function f_W satisfying

$$\mathsf{P}(g(X, Y) \leq w) = \int_{-\infty}^{w} f_W(u)\mathrm{d}u.$$

One very useful tool is given by the next result.

THEOREM 4.18 Suppose (X, Y) has pdf $f_{X,Y}$, and let $U = g(X, Y)$ and $V = h(X, Y)$. Assume the transformation $(x, y) \to (g(x, y), h(x, y))$ is *1-1 and differentiable* on a set A such that $\mathrm{P}((X, Y) \in A) = 1$.

Then (U, V) has pdf satisfying

$$f_{U,V}(u, v) = f_{X,Y}(x, y)\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}u\mathrm{d}v} \quad \text{expressed as } \textit{a function of } (u, v),$$

where we interpret the differentials in terms of the Jacobian matrix:

$$\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}u\mathrm{d}v} = \left|\det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}\right|.$$

(Compare to the one-variable case Cor. 2.12 in Chapter 2.)

PROOF This is based on the classic theorem about change of variables in multivariable calculus. The point is that we want these two integrals to match:

$$\int_{-\infty}^{b}\int_{-\infty}^{a} f_{U,V}(u, v)\mathrm{d}u\mathrm{d}v = \iint_{g^{-1}(-\infty,a]\cap h^{-1}(-\infty,b]} f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y.$$

Both sides have the value $\mathrm{P}(U \le a, V \le b)$; they are just computed using different variables. The theorem explains how the integrands are related. □

Some pointers about Thm. 4.18.

- Watch your variables. The end result needs to be in terms of the variables used for the pdf you are finding, and no others.
- If $g(x, y)$ is not 1-1 or not differentiable then return to the basic method of finding the (joint) cdf first.
- Even if you only want the pdf for U, you still have to get a joint pdf and then integrate out the other variable ($V = h(X, Y)$). However, you can choose whatever function h makes the calculation more convenient, including simply $h(x, y) = x$.
- Recall that $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$.
- The theorem can be extended to higher dimensions in the obvious way, although computing the determinant is harder.
- Be aware that the individual differentials ($\mathrm{d}x$, $\mathrm{d}y$, etc.) are *not quantities or functions*. They are simply a device to help remember the theorem.

Example 4.8 (cont.) Let T_i be the time until the i-th occurrence in a Poisson process. We found earlier that

$$f_{T_1,T_2}(t_1,t_2) = \lambda^2 \mathrm{e}^{-\lambda t_2} 1_{(0,\infty)}(t_1) 1_{(t_1,\infty)}(t_2).$$

Let $W_1 = T_1$, $W_2 = T_2 - T_1$. These are the "waiting times". Here, $t_1 = w_1$ and $t_2 = w_1 + w_2$. Since

$$\frac{\mathrm{d}t_1 \mathrm{d}t_2}{\mathrm{d}w_1 \mathrm{d}w_2} = \left| \det \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \right| = 1,$$

we obtain

$$\begin{aligned} f_{W_1,W_2}(w_1,w_2) &= f_{T_1,T_2}(t_1,t_2) = \lambda^2 \mathrm{e}^{-\lambda(w_1+w_2)} 1_{(0,\infty)}(w_1) 1_{(0,\infty)}(w_2) \\ &= \left(\lambda \mathrm{e}^{-\lambda w_1} 1_{(0,\infty)}(w_1)\right)\left(\lambda \mathrm{e}^{-\lambda w_2} 1_{(0,\infty)}(w_2)\right). \end{aligned}$$

Therefore, W_1 and W_2 are *independent* exponential$(1/\lambda)$ rvs.

In fact, this exercise can be extended to the whole process: if $W_1 = T_1$ and $W_i = T_i - T_{i-1}$, $i = 2, 3, \ldots$, then $W_1, W_2, W_3, \ldots$ are all independent exponential$(1/\lambda)$ rvs.

Note that this agrees with Ex. 3.2 and Ex. 4.10 as well: $T_i \sim$ gamma$(i, 1/\lambda)$ and (apparently) T_i and W_{i+1} are independent so $T_{i+1} \sim$ gamma$(i+1, 1/\lambda)$.

Thm. 4.17 and the last example are special cases of linear transformations. Based on Thm. 4.18, we can state the following more general result.

COROLLARY 4.19 Suppose (X, Y) has joint pdf $f_{X,Y}(x, y)$. Let $W = aX + bY$ and $Z = cX + dY$.

i. If $b \neq 0$ then (X, W) has joint pdf $f_{X,W}(x, w) = \frac{1}{|b|} f_{X,Y}(x, (w - ax)/b)$ and

$$f_W(w) = \int_{-\infty}^{\infty} \frac{1}{b} f_{X,Y}(x, (w - ax)/b)\, \mathrm{d}x.$$

ii. If $ad \neq bc$ then (W, Z) has joint pdf

$$f_{W,Z}(w, z) = \frac{1}{|ad - bc|} f_{X,Y}\left(\frac{dw - bz}{ad - bc}, \frac{az - cw}{ad - bc}\right).$$

Why would (W, Z) *not* have a joint pdf in the case $ad = bc$? Hint: what would the 2-dimensional range of (W, Z) look like in this case, and what would happen if you tried doing a double integral over that range?

We now look at a couple examples with nonlinear transformations.

Example 4.5 (cont.) Suppose (X, Y) has joint pdf

$$f_{X,Y}(x,y) = \tfrac{1}{2}\Big(\lambda^2 \mathrm{e}^{-\lambda(x+y)} + \mu^2 \mathrm{e}^{-\mu(x+y)}\Big) 1_{(0,\infty)}(x) 1_{(0,\infty)}(y).$$

Let $T = X + Y$ and $W = X/(X+Y)$. Now we have $x = wt$, $y = (1-w)t$ and

$$\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}t\mathrm{d}w} = \left|\det \begin{pmatrix} w & t \\ 1-w & -t \end{pmatrix}\right| = t.$$

So

$$f_{T,W}(t,w) = t f_{X,Y}(wt,(1-w)t) = \tfrac{t}{2}\Big(\lambda^2 \mathrm{e}^{-\lambda t} + \mu^2 \mathrm{e}^{-\mu t}\Big) 1_{(0,\infty)}(t) 1_{(0,1)}(w).$$

Since this factors, we can also conclude T and W are independent with

$$f_T(t) = \tfrac{1}{2}\Big(\lambda^2 t \mathrm{e}^{-\lambda t} + \mu^2 t \mathrm{e}^{-\mu t}\Big) 1_{(0,\infty)}(t)$$

and $W \sim \text{uniform}(0,1)$.

Example 4.12 Suppose X and Y are independent gamma rvs with the same scale parameter γ and shape parameters α and β, respectively. We want the distribution of X/Y.

Solution Let $R = X/Y$ and $W = Y$. Now, $X = RW$ and

$$\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}r\mathrm{d}w} = \left|\det\begin{pmatrix} w & r \\ 0 & 1 \end{pmatrix}\right| = w.$$

So

$$f_{R,W}(r,w) = w f_{X,Y}(rw,w) = w\frac{(rw)^{\alpha-1}\mathrm{e}^{-rw/\gamma}}{\Gamma(\alpha)\gamma^{\alpha}}\frac{w^{\beta-1}\mathrm{e}^{-w/\gamma}}{\Gamma(\beta)\gamma^{\beta}}.$$

Next, we want to integrate out w. The part depending on w looks like another gamma, but with a scale parameter that depends on $1+r$:

$$\begin{aligned} f_{R,W}(r,w) &= \frac{r^{\alpha-1}}{\Gamma(\alpha)\Gamma(\beta)}\frac{w^{\alpha+\beta-1}\mathrm{e}^{-(1+r)w/\gamma}}{\gamma^{\alpha+\beta}} \\ &= \left(\frac{\Gamma(\alpha+\beta)r^{\alpha-1}}{\Gamma(\alpha)\Gamma(\beta)(1+r)^{\alpha+\beta}}\right)\left(\frac{w^{\alpha+\beta-1}\mathrm{e}^{-w/(\gamma/(1+r))}}{\Gamma(\alpha+\beta)(\gamma/(1+r))^{\alpha+\beta}}\right) = f_R(r)f_{W|R}(w|r). \end{aligned}$$

The first factor is the density of R (see Ex. 3.5) and the second is the conditional density of W, given $R = r$.

An important special case of this example is when $X \sim$ chi-square(m) and $Y \sim$ chi-square(n). Then (recall Def. 3.10) $\alpha = m/2$, $\beta = n/2$ and $\gamma = 2$.

Letting $T = \frac{X/m}{Y/n} = \frac{n}{m}R$ (such as an F-statistic in an analysis of variance), we find

$$f_T(t) = \frac{m}{n} f_R(mt/n) = \frac{m\Gamma((m+n)/2)(mt/n)^{(n/2)-1}}{n\Gamma(m/2)\Gamma(n/2)(1+(mt/n))^{(m+n)/2}}.$$

This is Snedecor's F-distribution.

A related example is the joint distribution of $T = X + Y$ and $U = \frac{X}{X+Y}$, which turn out to be independent (exercise).

4.3 Expectations and Conditional Expectations

Expectations involving multiple random variables are defined in an obvious way for discrete and continuous joint distributions.

DEFINITION 4.20 Suppose (X, Y) has joint pmf or pdf $f_{X,Y}$.

i. For any function $g(X, Y)$ for which the following exists,

$$\mathsf{E}(g(X,Y)) = \begin{cases} \sum_x \sum_y g(x,y) f_{X,Y}(x,y) & \text{if } f(x,y) \text{ is a pmf,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y & \text{if } f(x,y) \text{ is a pdf.} \end{cases}$$

ii. The extension to expectation of $g(X_1, \ldots, X_k)$ for jointly distributed $(X_1, \ldots, X_k)$ is done in like manner using the joint pmf/pdf.

This definition is not so much about *existence* as it is about *how to compute* expectations when there is a joint pmf or a joint pdf. While we do not dwell on it in this course, expectations are of course suitably defined even when there is not a joint pmf or pdf for the random vector.

Note the following, as well.

- Existence can be inferred if either $g(x, y)$ is a nonnegative function (in which case the expectation could be infinite) or $|g(X, Y)|$ has finite expectation (in which case $g(X, Y)$ will also). The point, as always, is to avoid getting $\infty - \infty$.
- Assuming existence, order of summation/integration does not matter, nor does rewriting the expression in some other way in order to simplify computation.
- A function of X, say $h(X)$, is also a function of (X, Y) and so the expectation of $h(X)$ can be computed from the joint distribution for (X, Y). This can actually be simpler than first finding the pmf/pdf for X and using that.
- As before, expectations of indicator random variables are probabilities: $\mathsf{E}(1_{(X,Y)\in B}) = \mathsf{P}((X, Y) \in B)$.
- The general definition is based on discrete approximations to the random variable $g(X, Y)$, and taking limits.

Not surprisingly, this definition enjoys all the properties ascribed to expectations earlier.

THEOREM 4.21 All the properties of expectation given in Thm. 2.17 and Thm. 2.18 hold when X is replaced with (X, Y) (or $(X_1, \ldots, X_k)$).

In particular, the linearity property holds. For example,

$$\mathsf{E}(ag(X) + bh(Y)) = a\mathsf{E}(g(X)) + b\mathsf{E}(h(Y)).$$

Likewise, if $g(x, y) \leq h(x, y)$ for all (x, y) then the monotonicity property

$$\mathsf{E}(g(X, Y)) \leq \mathsf{E}(h(X, Y))$$

holds.

Indeed, one can think of expectation as the unique *operator* (converting functions of $(X_1, \ldots, X_n)$ to real values) that has the properties of *linearity, monotonicity and convergence of limits* (under appropriate constraints for the last), subject to the requirement that $\mathsf{E}(1_{(X,Y)\in B}) = \mathsf{P}((X, Y) \in B)$.

Example 4.5 (cont.) Suppose (X, Y) has joint pdf

$$f_{X,Y}(x,y) = \tfrac{1}{2}\Big(\lambda^2 \mathrm{e}^{-\lambda(x+y)} + \mu^2 \mathrm{e}^{-\mu(x+y)}\Big) 1_{(0,\infty)}(x) 1_{(0,\infty)}(y).$$

Find $\mathsf{E}((X+Y)^2)$ *directly from the joint pdf*, instead of from the pdf of $T = X + Y$.

Solution We can do this in parts.

$$\mathsf{E}(X^2) = \int_0^\infty \int_0^\infty \frac{1}{2}\Big(\lambda^2 x^2 \mathrm{e}^{-\lambda(x+y)} + \mu^2 x^2 \mathrm{e}^{-\mu(x+y)}\Big)\mathrm{d}x\mathrm{d}y = \frac{1}{2}\Big(\frac{2}{\lambda^2} + \frac{2}{\mu^2}\Big).$$

By symmetry, $\mathsf{E}(Y^2) = \mathsf{E}(X^2)$.

$$\mathsf{E}(XY) = \int_0^\infty \int_0^\infty \frac{1}{2}\Big(\lambda^2 xy \mathrm{e}^{-\lambda(x+y)} + \mu^2 xy \mathrm{e}^{-\mu(x+y)}\Big)\mathrm{d}x\mathrm{d}y = \frac{1}{2}\Big(\frac{1}{\lambda^2} + \frac{1}{\mu^2}\Big).$$

Therefore,

$$\mathsf{E}((X+Y)^2) = \mathsf{E}(X^2) + 2\mathsf{E}(XY) + \mathsf{E}(Y^2) = \frac{3}{\lambda^2} + \frac{3}{\mu^2}.$$

(Compare this with $\mathsf{E}(T^2)$ computed from the pdf for T obtained earlier.)

Example 4.13 A lottery cage has N balls, numbered $1, 2, \ldots, N$. Suppose 2 balls are randomly drawn from the cage (in known order), with values X_1 and X_2. Find $\mathsf{E}(X_1 X_2)$.

Solution Since $X_1 \neq X_2$,

$$\begin{aligned}\mathsf{E}(X_1X_2) &= \frac{1}{N(N-1)}\sum_{k=1}^{N}\sum_{m=1}^{N} km1_{m\neq k} = \frac{1}{N(N-1)}\Big(\sum_{k=1}^{N}\sum_{m=1}^{N} km - \sum_{k=1}^{N} k^2\Big) \\ &= \frac{1}{N(N-1)}\Big(\Big(\frac{N(N+1)}{2}\Big)^2 - \frac{N(N+1)(2N+1)}{6}\Big) \\ &= \frac{(N+1)(3N+2)}{12},\end{aligned}$$

where we have used well-known results for summing $1, 2, \ldots, N$ and $1^2, 2^2, \ldots, N^2$.

(Note also the subtraction in the calculation that avoids trying to deal with the restriction $m \neq k$ directly in the double sum.)

Example 4.7 (cont.) Consider the random pair (X, Y) with joint pdf

$$f(x,y) = \frac{8}{\pi}(x^2+y^2), \quad \text{for } x \geq 0,\ y \geq 0,\ x^2+y^2 \leq 1.$$

Recall that $f_X(x) = \frac{8}{3\pi}(1+2x^2)(1-x^2)^{1/2}$, $0 \leq x \leq 1$. To find $\mathsf{E}(X)$ we could integrate $\int_0^1 x f_X(x)\,\mathrm{d}x$.

Or, what might be a little less daunting, we can compute, using polar coordinates,

$$\begin{aligned}\mathsf{E}(X) &= \int_0^1\int_0^1 x\,\frac{8}{\pi}(x^2+y^2)1_{x^2+y^2\leq 1}\,\mathrm{d}x\mathrm{d}y \\ &= \frac{8}{\pi}\int_0^{\pi/2}\int_0^1 r\cos(\theta)\,r^2\,r\mathrm{d}r\mathrm{d}\theta = \frac{8}{5\pi}.\end{aligned}$$

Similarly, using the fact that X and Y have the same distribution (why?),

$$\mathsf{E}(X^2) = \frac{\mathsf{E}(X^2+Y^2)}{2} = \frac{4}{\pi}\int_0^1\int_0^1 (x^2+y^2)^2 1_{x^2+y^2\leq 1}\,\mathrm{d}x\mathrm{d}y = \cdots = \frac{1}{3}.$$

then $\mathsf{var}(X) = \frac{1}{3} - \frac{64}{25\pi^2}$.

As a check, we calculate $\mu_X = 0.50930$ and $\sigma_X = 0.27194$, which are sensible since $0 \leq X \leq 1$.

THEOREM 4.22 Suppose $(X_1, \ldots, X_k)$ has joint distribution. Then $X_1, \ldots, X_k$ are independent if and only if

$$\mathsf{E}(g_1(X_1) \times \cdots \times g_k(X_k)) = \mathsf{E}(g_1(X_1)) \times \cdots \times \mathsf{E}(g_k(X_k)))$$

whenever these expectations exist. ($g_1(X_1), \ldots, g_k(X_k)$ are also independent.)

In particular, if X and Y are independent and have finite means then $\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y)$.

PROOF If X and Y are independent then we have (for example, in the continuous case with $k = 2$)

$$\mathsf{E}(g(X)h(Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y$$

$$= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)\mathrm{d}y = \mathsf{E}(g(X))\mathsf{E}(h(Y)).$$

On the other hand, if this equality holds for all g and h we simply apply it to arbitrary indicator functions:

$$\mathsf{P}(X \in A, Y \in B) = \mathsf{E}(1_A(X)1_B(Y))$$

$$= \mathsf{E}(1_A(X))\mathsf{E}(1_B(Y)) = \mathsf{P}(X \in A)\mathsf{P}(Y \in B).$$

The general result has essentially the same proof. □

**** Note, however, that* $\mathsf{E}(XY) = \mathsf{E}(X)\mathsf{E}(Y)$ *does not imply* X *and* Y *are independent. The theorem says that* $\mathsf{E}(g(X)h(Y))$ *must factor for all sensible functions* $g(x)$ *and* $h(y)$.

Example 4.14 (Annuities) An insurance company promises to pay its customer \$100 a month, plus a fixed cost of living increase, every month until he dies. Suppose his time to death is a negative binomial(k, p) random variable and the cost of living increase is based on a financial index which, at contract time, has a normal(μ, σ^2) distribution. These may be assumed independent.

What is the expected payout?

Solution Let T be the number of months until death and C be the cost of living increase (in %). The total payout is $P = 100(T + CT(T-1)/2)$. Thus, since T and C are independent,

$$\begin{aligned}\mathsf{E}(P) &= 100(\mathsf{E}(T) + \mathsf{E}(C)\mathsf{E}(T(T-1))/2) \\ &= 100\left(\frac{k(1-p)}{p} + \frac{\mu k(k+1)(1-p)^2}{2p^2}\right),\end{aligned}$$

where the first and second moments of T were obtained from the negative binomial distribution.

Example 4.15 Suppose $U \sim \text{gamma}(2, \beta)$ and $V \sim \text{gamma}(3, \beta)$, independent. What are the mean and variance of $R = \frac{U}{V}$?

Solution First express $R = U \times \frac{1}{V}$ so that it is a *product* of separate functions.

From Thm. 3.12, $\text{E}(U) = 2\beta$ and $\text{E}(U^2) = 6\beta^2$.

The same argument in the proof of that theorem can be applied to deduce that $\text{E}(\frac{1}{V}) = \frac{1}{2\beta}$ and $\text{E}(\frac{1}{V^2}) = \frac{1}{2\beta^2}$.

Hence,

$$\text{E}(R) = \text{E}(U)\text{E}\left(\frac{1}{V}\right) = 1 \quad \text{and} \quad \text{E}(R^2) = \text{E}(U^2)\text{E}\left(\frac{1}{V^2}\right) = 3.$$

This gives $\text{var}(R) = 2$.

**** $\text{E}(R)$ is not equal to $\text{E}(U)/\text{E}(V)$.*

Recall that β is the *scale* parameter for the gamma distribution. Consequently, the ratio U/V ought to not depend on β for this example. We see from above that, in fact, its mean and variance do not depend on β.

Now we turn to defining and using conditional expectations. Simply put, conditional expectations are expectations for conditional distributions.

They are, in fact, *more important than ordinary expectations*. They can be used to explain how one random variable is related to others or to predict one from others. They can even be used to compute ordinary expectations.

DEFINITION 4.23 Suppose X, given $Y = y$, has conditional pmf or pdf $f_{X|Y}(x|y)$.

i. For any function $g(X, Y)$ for which the following exists,

$$\mathsf{E}(g(X,Y)|Y=y) = \begin{cases} \sum_x g(x,y) f_{X|Y}(x|y) & \text{if } f_{X|Y} \text{ is a pmf,} \\ \int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y) \mathrm{d}x & \text{if } f_{X|Y} \text{ is a pdf.} \end{cases}$$

Something that is *apparent* from the definition above is that

$$\mathsf{E}(g(X,Y)|Y=y) = \mathsf{E}(g(X,y)|Y=y).$$

In other words, *with $Y = y$ given we may substitute y for Y* in the function $g(X, Y)$, which is the "natural" thing to do.

Example 4.13 (cont.) A lottery cage has N balls, numbered $1, 2, \ldots, N$. Suppose 2 balls are randomly drawn from the cage (in known order), with values X_1 and X_2.

Now find $\mathsf{E}(X_2|X_1 = k)$.

Solution We note first that if we are given $X_1 = k$, then X_2 can be any other value with equal probabilities. Thus, the conditional pmf for X_2, given $X_1 = k$, is simply

$$f_{X_2|X_1}(m|k) = \frac{1}{N-1} 1_{\{1,\ldots,k-1,k+1,\ldots,N\}}(x).$$

Therefore, we can compute

$$\begin{aligned} \mathsf{E}(X_2|X_1 = k) &= \sum_{m \neq k} m f_{X_2|X_1}(m|k) = \frac{1}{N-1}\left(\left(\sum_{m=1}^{N} m\right) - k\right) \\ &= \frac{1}{N-1}\left(\frac{N(N+1)}{2} - k\right). \end{aligned}$$

Example 4.7 (cont.) Once again, consider the random pair (X, Y) with joint pdf

$$f(x,y) = \frac{8}{\pi}(x^2+y^2), \quad \text{for } x \geq 0,\ y \geq 0,\ x^2+y^2 \leq 1$$

and conditional pdf (found earlier)

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{3(x^2+y^2)}{(1+2x^2)(1-x^2)^{1/2}}, \quad 0 \leq y \leq \sqrt{1-x^2}.$$

Recall that we treat this as a *function of* y, with x *fixed*.

The k-th moment for the conditional pdf is therefore

$$\mathsf{E}(Y^k|X=x) = \int_0^{\sqrt{1-x^2}} y^k f_{Y|X}(y|x)\,\mathrm{d}y = \cdots = \frac{3(k+1+2x^2)(1-x^2)^{k/2}}{(k+1)(k+3)(1+2x^2)}.$$

(Check: this equals 1 for $k = 0$.)

In particular, $\mathsf{E}(Y|X=x) = \frac{(1+x^2)(1-x^2)^{1/2}}{4(1+2x^2)}$ and $\mathsf{E}(Y^2|X=x) = \frac{(1-x^4)}{15(1+2x^2)}$.

DEFINITION 4.24 The function $\mu_Y(x) = \mathrm{E}(Y|X = x)$ is called the regression of Y on X, or simply, the regression function.

The regression function can be interpreted as describing (in an average way) how Y is related to X, or it can be interpreted as providing a prediction for Y when the value of X is known. By far, the major objective of statistics is to find a regression function or something analogous.

The regression function can be linear, as it is in Ex. 4.13, or it can be nonlinear as in Ex. 4.7.

Example 4.9 (cont.) Suppose $f_{X,Y}(x,y) = \frac{\sqrt{7}}{4\pi}\mathrm{e}^{-(x^2-xy+2y^2)/2}$. We saw that $X|\{Y = y\} \sim \text{normal}(y/2, 1)$. So the regression of X on Y, given by $\mathrm{E}(X|Y = y) = y/2$, is linear with no intercept.

In fact, we can easily generalize this. Suppose $f_{X,Y}(x,y) = K\mathrm{e}^{-(ax^2+bxy+cy^2)/2}$ for some constant K, and a, b, c satisfying $a > 0$, $c > 0$ and $b^2 < 4ac$. By the same reasoning as earlier,

$$f_{Y|X}(y|x) = K_1\mathrm{e}^{-c(y+bx/(2c))^2},$$

for some K_1, which is a normal pdf and tells us $\mathrm{E}(Y|X = x) = -bx/(2c)$.

Example 4.16 In the random walk model of finance, if a stock has price X today then its value in t days is XY, where $Y \sim$ lognormal$(0, \sigma^2 t)$ and Y is independent of X. (That is, $\log Y \sim$ normal$(0, \sigma^2 t)$.) Let t be fixed.

Suppose, however, there an option on the stock whose value cannot exceed a "strike price" $X + K$ (fixed parameter K). This means the option will have value $V = \min(XY, X + K)$ in t days.

What is a sensible way to price the option, given the current stock price? That is, what is the regression of the value V on the initial price X? *Beware:* this is a little complicated.

Solution To separate the randomness from the parameters, let $Z = \frac{\log Y}{\sqrt{\sigma^2 t}}$. Then Z is standard normal and independent of X, and we can express the stock value as

$$V = \min(X \mathrm{e}^{\sigma\sqrt{t}Z}, X + K).$$

Hence,

$$f_{Z|X}(z|x) = f_Z(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}.$$

Let $\Phi(z)$ denote the standard normal cdf. Note that V will take the strike price only if $Z \geq h(X) \stackrel{\text{def}}{=} \log(1 + K/X)/\sigma\sqrt{t}$.

Now we compute

$$\begin{aligned}
\mathsf{E}(V|X=x) &= \int_{-\infty}^{\infty} \min(xe^{\sigma\sqrt{t}z}, x+K)\frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}}\mathrm{d}z \\
&= \int_{-\infty}^{h(x)} x\mathrm{e}^{\sigma\sqrt{t}z}\frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}}\mathrm{d}z + \int_{h(x)}^{\infty}(x+K)\frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}}\mathrm{d}z \\
&= xe^{-\sigma^2 t/2}\Phi(h(x)-\sigma\sqrt{t}) + (x+K)(1-\Phi(h(x))).
\end{aligned}$$

This is obviously nonlinear in x.

(Our assumptions actually are too simple in reality.)

Another "obvious" result is the following, which explains what happens when conditioning on an independent random variable.

THEOREM 4.25 Suppose X and Y are independent. Then, subject to existence,

$$\mathsf{E}(g(X)|Y=y) = \mathsf{E}(g(X)) \quad \text{and} \quad \mathsf{E}(h(Y)|X=x) = \mathsf{E}(h(Y)).$$

In particular, the regression of Y on X has constant value $\mathsf{E}(Y)$ and the regression of X on Y has constant value $\mathsf{E}(X)$.

PROOF (exercise) □

At this point we want to broaden our understanding of conditional expectation.

- Suppose $h(y) = \mathsf{E}(g(X)|Y = y)$ is our "prediction" of $g(X)$ when we know that $Y = y$. In practice, however, Y may not yet be known or observed, or we are simply thinking in advance. In such a case, we consider the *random variable* $h(Y)$ as what the prediction would be once we have a value for Y.
- For the sake of notation, we use $\mathsf{E}(g(X)|Y)$ to denote the random variable $h(Y)$. (But $\mathsf{E}(g(X)|Y = Y)$ is not meaningful. Why not?)
- Indeed, $\mathsf{E}(g(X)|Y)$ has many uses, not the least of which is that we can use it, along with the distribution of Y, to find $\mathsf{E}(g(X))$ without needing the marginal distribution of X.
- Or, more generally, if we know $\mathsf{E}(g(X,Y)|Y)$ and the distribution of Y then we can find $\mathsf{E}(g(X,Y))$.

To see the last two points, first note that $\mathsf{E}(g(X,Y)|Y)$ is a function *just of* Y. (Since that is all that is given.)

THEOREM 4.26 (Iterated Expectation) Suppose Y has pmf or pdf f_Y. For any function $g(x,y)$ for which $\mathsf{E}(g(X,Y))$ exists,

$$\mathsf{E}(g(X,Y)) = \begin{cases} \sum_y \mathsf{E}(g(X,Y)|Y=y) f_Y(y) & \text{if } f_{X|Y} \text{ is a pmf,} \\ \int_{-\infty}^{\infty} \mathsf{E}(g(X,Y)|Y=y) f_Y(y) \mathrm{d}y & \text{if } f_{X|Y} \text{ is a pdf.} \end{cases}$$

To express this more succinctly, we write

$$\mathsf{E}(g(X,Y)) = \mathsf{E}(\mathsf{E}(g(X,Y)|Y)),$$

which is true regardless of the nature of the joint distribution for (X,Y).

This is also called "conditioning on Y".

PROOF We observe that $f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y)$ for all x, y (even if $f_Y(y) = 0$). Thus in the continuous case, for example,

$$\begin{aligned} \mathsf{E}(g(X,Y)) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y) \mathrm{d}x \right) f_Y(y) \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathsf{E}(g(X,Y)|Y=y) f_Y(y) \mathrm{d}y. \end{aligned}$$

(Effectively, we are integrating over x first instead of over y first.) □

**** Thm. 4.26 is a powerful result.* In spite of its apparent complexity, using this can greatly simplify computation of expectations by removing the need for explicit integration/summation in the context of multiple rvs, and can even make them trivial.

Example 4.2 (cont.) In the marked Poisson example, we had $Y \sim \text{Poisson}(\lambda)$ and $X|\{Y = y\} \sim \text{binomial}(y, p)$.

So $\mathsf{E}(X|Y = y) = yp$ and

$$\mathsf{E}(X) = \sum_y \mathsf{E}(X|Y = y) f_Y(y) = \sum_y yp f_Y(y) = p\mathsf{E}(Y) = p\lambda.$$

Or, much more directly, we write $\mathsf{E}(X|Y) = Yp$ and

$$\mathsf{E}(X) = \mathsf{E}(\mathsf{E}(X|Y)) = \mathsf{E}(Yp) = p\mathsf{E}(Y) = p\lambda.$$

Likewise, using the second moments of the binomial and Poisson distribution (recall $\mu_2' = \sigma^2 + \mu^2$),

$$\begin{aligned}\mathsf{E}(X^2) &= \mathsf{E}(\mathsf{E}(X^2|Y)) = \mathsf{E}(Yp(1-p) + (Yp)^2) \\ &= p(1-p)\lambda + p^2(\lambda + \lambda^2) = p\lambda + p^2\lambda^2.\end{aligned}$$

Thus, $\text{var}(X) = p\lambda$.

(Check this with the Poisson$(p\lambda)$ pmf for X obtained earlier in Sec. 4.1.)

THEOREM 4.27

i. Suppose $\mathsf{E}(g(X,Y))$ exists and let $q(y) = \mathsf{E}(g(X,Y)|Y=y)$. Then $q(y) = \mathsf{E}(g(X,y)|Y=y)$ and $\mathsf{E}(g(X,Y)) = \mathsf{E}(q(Y))$.

ii. (Pull-Out) $\mathsf{E}(h(Y)g(X)|Y=y) = h(y)\mathsf{E}(g(X)|Y=y)$ and $\mathsf{E}(h(Y)g(X,Y)|Y=y) = h(y)\mathsf{E}(g(X,y)|Y=y)$.

iii. $\mathsf{E}(g(X)h(Y)) = \mathsf{E}(h(Y)\mathsf{E}(g(X)|Y))$ for any functions g and h such that the expectations exist.

In other words, conditioning on Y is like holding Y constant, even if only temporarily. (Property iii. above is actually how conditional expectation is defined in general: $\mathsf{E}(g(X)|Y)$ is the function of Y such that the equality in iii. holds for all suitable functions $h(y)$.)

PROOF We indicate the proof for the continuous case. The discrete case is similar.

i. By definition, $q(y) = \int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y)\mathrm{d}x = \mathsf{E}(g(X,y)|Y=y)$. The second equality is just a restatement of iterated expectation.

ii., iii. These follow from Thm. 4.26 and i. (exercise) □

Example 4.1 (cont.) Suppose (X_1, X_2, X_3) has trinomial(n, p_1, p_2, p_3) distribution. We saw that $X_2|\{X_1 = x_1\} \sim$ binomial$(n - x_1, p_2/(1 - p_1))$.

Now find $\mathsf{E}(X_1X_2)$.

Solution First observe that $\mathsf{E}(X_2|X_1) = (n - X_1)p_2/(1 - p_1)$, according to the conditional distribution of X_2, given X_1.

Next, we calculate

$$\begin{aligned} \mathsf{E}(X_1(n - X_1)) &= n\mathsf{E}(X_1) - (\mathsf{var}(X_1) + (\mathsf{E}(X_1)^2) \\ &= n^2p_1 - np_1(1 - p_1) - n^2p_1^2 \\ &= n(n - 1)p_1(1 - p_1). \end{aligned}$$

Hence,

$$\mathsf{E}(X_1X_2) = \mathsf{E}(X_1\mathsf{E}(X_2|X_1)) = \mathsf{E}(X_1(n - X_1)p_2/(1 - p_1)) = n(n - 1)p_1p_2.$$

Compare this approach to the restricted double sum (with $x_1 + x_2 \leq n$) required otherwise.

Example 4.17 (Linear Regression) Suppose $Y|\{X = x\} \sim \text{normal}(\beta_0+\beta_1 x, \sigma^2)$. If also $X \sim \text{normal}(\mu_X, \sigma_X^2)$ then

$$\mu_Y = \mathsf{E}(\mathsf{E}(Y|X)) = \mathsf{E}(\beta_0 + \beta_1 X) = \beta_0 + \beta_1 \mu_X$$

and

$$\begin{aligned} \mathsf{E}(Y^2) &= \mathsf{E}(\mathsf{E}(Y^2|X)) = \mathsf{E}(\sigma^2 + (\beta_0 + \beta_1 X)^2) \\ &= \sigma^2 + \beta_1^2\sigma_X^2 + (\beta_0 + \beta_1\mu_X)^2. \end{aligned}$$

Hence $\text{var}(Y) = \sigma^2 + \beta_1^2\sigma_X^2$.

Note that even though the variance of the conditional distribution does not depend on x, it is not $\text{var}(Y)$.

Actually, *the normality assumption was not required* for either expectation above. We just used the mean and variance of the conditional distribution for Y, given X, and the first two moments of X.

Example 4.18 (Insurance Claims) Suppose the number of claims in a particular month is $N \sim \text{Poisson}(\lambda)$ and the claim values are independent exponential(β) rvs and also independent of N.

If $N = n$ then by Ex. 4.10 the sum Y of the n claims is a gamma(n, β) rv. That is, $Y|\{N = n\} \sim \text{gamma}(n, \beta)$ with mean $\mathsf{E}(Y|N = n) = n\beta$. (Define $Y = 0$ if $n = 0$ and note that $\mathsf{E}(Y|N = n) = n\beta$ still holds in that case.)

Thus,

$$\mathsf{E}(Y) = \mathsf{E}(\mathsf{E}(Y|N)) = \mathsf{E}(N\beta) = \lambda\beta$$

and

$$\mathsf{E}(Y^2) = \mathsf{E}(\mathsf{E}(Y^2|N)) = \mathsf{E}(N(N+1)\beta^2) = (\lambda^2 + 2\lambda)\beta^2.$$

And consequently, $\text{var}(Y) = 2\lambda\beta^2$.

(Note how we avoided being explicit about the *joint* distribution of (N, Y), which in this case is neither discrete nor continuous.)

Since variances are so important in statistics, we look at how they are defined in the context of conditioning.

DEFINITION 4.28 The conditional variance of Y, given $X = x$, is the *variance of the conditional distribution* of Y, given $X = x$, if it exists. It is denoted $\text{var}(Y|X = x)$ or $\sigma^2_Y(x)$ and is a function of x.

The *random variable* $\text{var}(Y|X) = \sigma^2_Y(X)$ is the value of the function $\sigma^2_Y(x) = \text{var}(Y|X = x)$ with X in place of x. (Again, the notation $\text{var}(Y|X = X)$ is not sensible.)

The shortcut formula works, of course:

$$\text{var}(Y|X = x) = \text{E}(Y^2|X = x) - (\text{E}(Y|X = x))^2.$$

Two special cases are given in the next result.

THEOREM 4.29 $\text{var}(h(X)|X = x) = 0$, whereas if X and Y are independent then $\text{var}(h(Y)|X = x) = \text{var}(h(Y))$.

PROOF (Exercise) □

Example 4.1 (cont.) Suppose (X_1, X_2, X_3) has trinomial(n, p_1, p_2, p_3) distribution $(p_1 + p_2 + p_3 = 1)$.

We saw that $X_2|\{X_1 = x_1\} \sim$ binomial$(n - x_1, p_2/(1 - p_1))$. Therefore, using what we know about binomial distributions,

$$\text{var}(X_2|X_1 = x_1) = (n - x_1)\frac{p_2}{1 - p_1}\left(1 - \frac{p_2}{1 - p_1}\right) = \frac{(n - x_1)p_2p_3}{(1 - p_1)^2}.$$

Example 4.9 (cont.) Suppose $f_{X,Y}(x, y) = K\mathrm{e}^{-(ax^2+bxy+cy^2)/2}$ for some constant K, and a, b, c satisfying $a > 0$, $c > 0$ and $b^2 < 4ac$.

We pointed out earlier that

$$f_{Y|X}(y|x) = K_1\mathrm{e}^{-c(y+bx/(2c))^2} = K_1\mathrm{e}^{-(y-(-bx/2c))^2/(2/2c)},$$

for some K_1, which is a normal$(\frac{-bx}{2c}, \frac{1}{2c})$ pdf. Thus, var$(Y|X = x) = \frac{1}{2c}$ for all x.

Analogous to iterated expectation we have the following very useful formula for variance.

THEOREM 4.30 (Variance Partition) $\text{var}(Y) = \text{var}(\text{E}(Y|X)) + \text{E}(\text{var}(Y|X))$.

PROOF By Def. 4.28, $\text{var}(Y|X = x)$ is the second central moment of the conditional distribution of Y, given $X = x$. That is,

$$\text{var}(Y|X = x) = \text{E}(Y^2|X = x) - (\text{E}(Y|X = x))^2.$$

Taking the expectation of the random version of this we get

$$\text{E}(\text{var}(Y|X)) = \text{E}(\text{E}(Y^2|X)) - \text{E}((\text{E}(Y|X))^2) = \text{E}(Y^2) - \text{E}((\text{E}(Y|X))^2).$$

In addition, $\text{E}(Y|X)$ is a rv with mean $\text{E}(\text{E}(Y|X)) = \text{E}(Y)$ and variance

$$\text{var}(\text{E}(Y|X)) = \text{E}((\text{E}(Y|X))^2) - (\text{E}(Y))^2.$$

Adding these together gives the result. □

Note the *balance* of the partition expression. But you do have to be careful to interpret it properly.

Example 4.18 (cont.) We have that $N \sim \text{Poisson}(\lambda)$ and (in random terms) $Y|N \sim \text{gamma}(N, \beta)$ (where $N = 0 \implies Y = 0$). So $\text{E}(Y|N) = N\beta$ and

$$\text{var}(\text{E}(Y|N)) = \text{var}(N\beta) = \lambda\beta^2.$$

Meanwhile, $\text{var}(Y|N) = N\beta^2$ so

$$\text{E}(\text{var}(Y|N)) = \text{E}(N\beta^2) = \lambda\beta^2.$$

Hence $\text{var}(Y) = \text{var}(\text{E}(Y|X)) + \text{E}(\text{var}(Y|X)) = 2\lambda\beta^2$.

Compare with the earlier computation.

Example 4.17 (cont.) (Linear Regression) We assume $\text{E}(Y|X) = \beta_0 + \beta_1 X$ and $\text{var}(Y|X) = \sigma^2$. Hence

$$\begin{aligned} \text{var}(Y) &= \text{var}(\text{E}(Y|X)) + \text{E}(\text{var}(Y|X)) = \text{var}(\beta_0 + \beta_1 X) + \text{E}(\sigma^2) \\ &= \beta_1^2\sigma_X^2 + \sigma^2. \end{aligned}$$

Compare with the earlier computation.

- Like iterated expectation, the variance partition formula involves "conditioning" on one variable. It is most helpful when the conditional means and variances are known, as well as the marginal distribution of the variable you are conditioning on.
- The partitioning of the sums of squares in a statistical Analysis of Variance (ANOVA) is an analog of Thm. 4.30: total variation = variation of the regression/means model + expected variation of the data given the model.
- Indeed, Thm. 4.30 is a special case of a far more general theorem about *orthogonal projections*, of which the Pythagoras theorem is the simplest example, and which is an important aspect of *Hilbert Space theory*.

We have a final result that gives another reason to be looking at conditional expectations.

THEOREM 4.31 (Best Predictor) Assume X and Y are jointly distributed and Y has a finite second moment.

The function of X that best predicts Y, in the sense of minimizing $\mathsf{E}((Y - g(X))^2)$ (mean squared prediction error), is $g(X) = \mathsf{E}(Y|X)$.

PROOF First, let $g(x)$ be any function and let $h(x) = g(x) - \mathsf{E}(Y|X = x)$. We note that

$$(Y - g(X))^2 - (Y - \mathsf{E}(Y|X))^2 = h^2(X) - 2h(X)(Y - \mathsf{E}(Y|X)).$$

Since

$$\mathsf{E}(h(X)(Y - \mathsf{E}(Y|X))|X) = h(X)(\mathsf{E}(Y|X) - \mathsf{E}(Y|X)) = 0,$$

we have $\mathsf{E}(h(X)(Y - \mathsf{E}(Y|X))) = 0$ and

$$\mathsf{E}((Y - g(X))^2) - \mathsf{E}((Y - \mathsf{E}(Y|X))^2) = \mathsf{E}((g(x) - \mathsf{E}(Y|X))^2) \geq 0.$$

Thus $\mathsf{E}(Y|X)$ is as good a predictor as any. □

- A mathematical interpretation of this result is that rvs with finite second moment form a *Hilbert space*, and the conditional expectation of Y, given X, is the projection of Y onto the subspace of functions of X with finite second moment. (This result actually provides another way to define conditional expectation generally, except it has to be extended for the case that the second moment of Y is not finite.)
- An engineering interpretation is that $\mathsf{E}(\cdot|X)$ is the best *filter* that depends only on X.
- We should also point out that "best" here means minimum mean *squared* prediction error, which is not the only way to optimize. For example, one could instead minimize the mean absolute prediction error, $\mathsf{E}(|Y - g(X)|)$. It turns out that the *conditional median* optimizes this criterion.
- We examined a special case of Thm. 4.31 in Ex. 1.12 to illustrate how well conditional probabilities can predict. (X and Y were indicator rvs.)

Conditioning on more than one variable is the natural extension of the definitions above. We briefly describe it here.

For example, the conditional pdf/pmf for $(X_{k+1}, \ldots, X_m)$, given $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$, is the ratio of the joint pdf/pmf for $(X_1, \ldots, X_m)$ to that for $(X_1, \ldots, X_k)$.

$$f_{X_{k+1},\ldots,X_m|X_1,\ldots,X_k}(x_{k+1}, \ldots, x_m|x_1, \ldots, x_k) = \frac{f_{X_1,\ldots,X_m}(x_1, \ldots, x_m)}{f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)}.$$

Then, conditional probabilities and expectations for $(X_{k+1}, \ldots, X_m)$, given $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$, are computed from that conditional distribution.

4.4 Covariance and Correlation

Conditional distributions are one (relatively complete) way to describe the relationship between random variables. Conditional expectations (and variances) reduce that information to something more practical and easier to visualize. However, conditioning can be complex.

What can also be useful is a moment-like description of relationships. This is discussed in the present section. While simpler, it is limited to describing the extent to which a relationship is *linear*, as we will see.

DEFINITION 4.32 Assume X and Y are jointly distributed and their variances exist. Let μ_X, μ_Y and σ_X^2, σ_Y^2 be the means and variances.

i. The covariance of X and Y is $\text{cov}(X,Y) = \text{E}((X-\mu_X)(Y-\mu_Y))$.

ii. The correlation of X and Y is $\text{corr}(X,Y) = \text{cov}(X,Y)/(\sigma_X\sigma_Y)$.

If $\text{corr}(X,Y) = 0$ we say X and Y are uncorrelated. Otherwise, they are correlated.

**** Note the comma in the notation; this is about a pair of rvs.*

- Note that correlation is *unitless* but covariance is not.
- Intuitively, the covariance will be positive if X and Y tend to be above their respective means together and below their respective means together. If they tend to be opposite then the covariance will be negative.
- In addition, the covariance is a measure of how much above or below their means they are.
- The correlation is re-scaled according to how much the two variables individually differ from their means and so it basically indicates the degree to which the variables act (linearly) together.
- There is no definition of correlation for more than two random variables. However, you can get the correlation between each pair, often expressed in terms of a matrix. There are also notions of *partial correlation* and other ways to quantify relationships among more than two variables (not covered in this course).

THEOREM 4.33 (Properties of Covariance and Correlation) Assume X and Y are jointly distributed and their variances exist.

i. $\text{cov}(X, Y) = \text{cov}(Y, X)$ and $\text{corr}(X, Y) = \text{corr}(Y, X)$.

ii. $\text{cov}(X, Y) = \text{E}(XY) - \mu_X\mu_Y = \text{E}((X - a)(Y - \mu_Y))$ for any real a.

iii. $\text{cov}(X, X) = \text{var}(X)$.

iv. $\text{cov}(aX + b, Y) = a\,\text{cov}(X, Y)$ and $\text{corr}(aX + b, Y) = \text{sign}(a)\,\text{corr}(X, Y)$.

PROOF These are all evident from Def. 4.32 and the linearity of expectation. For example,

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

□

Part ii. of Thm. 4.33 gives several choices for computing the covariance. Which choice to use would depend on which variable(s) are easy to center or shift before taking an expectation.

Example 4.1 (cont.) Suppose $(X_1, X_2, X_3) \sim \text{trinomial}(n, p_1, p_2, p_3)$.

We have seen that $\mathsf{E}(X_i) = np_i$ and $\mathsf{E}(X_1X_2) = n(n-1)p_1p_2$.

Thus, we obtain

$$\mathsf{cov}(X_1, X_2) = \mathsf{E}(X_1X_2) - \mathsf{E}(X_1)\mathsf{E}(X_2) = -np_1p_2.$$

The covariance is negative because, if X_1 is large then X_2 is necessarily small, and vice versa. (Recall that $X_1 + X_2 \leq n$.)

Since also $\mathsf{var}(X_i) = np_i(1-p_i)$, we can compute

$$\mathsf{corr}(X_1, X_2) = \frac{\mathsf{cov}(X_1, X_2)}{\sqrt{\mathsf{var}(X_1)}\sqrt{\mathsf{var}(X_2)}} = -\sqrt{\frac{p_1p_2}{(1-p_1)(1-p_2)}}.$$

This value can range from -1 to 0. (It cannot be less than -1 because $p_1+p_2 < 1$.) It is interesting to note that this correlation does not depend on n.

Example 4.5 (cont.) Suppose (X, Y) has joint pdf

$$f_{X,Y}(x,y) = \tfrac{1}{2}\Big(\lambda^2 e^{-\lambda(x+y)} + \mu^2 e^{-\mu(x+y)}\Big) 1_{(0,\infty)}(x) 1_{(0,\infty)}(y).$$

We found $\mathsf{E}(X) = \mathsf{E}(Y) = \frac{1}{2\lambda} + \frac{1}{2\mu}$ and $\mathsf{E}(XY) = \frac{1}{2\lambda^2} + \frac{1}{2\mu^2}$.

Thus,

$$\mathsf{cov}(X,Y) = \frac{1}{2\lambda^2} + \frac{1}{2\mu^2} - \Big(\frac{1}{2\lambda} + \frac{1}{2\mu}\Big)^2 = \frac{1}{4}\Big(\frac{1}{\lambda} - \frac{1}{\mu}\Big)^2,$$

which is positive as long as $\lambda \neq \mu$.

Here is a heuristic explanation as to why the covariance depends on the (absolute) difference between $1/\lambda$ and $1/\mu$.

The structure of the pdf suggests that, with probability $1/2$, X and Y behave like independent exponential$(1/\lambda)$ rvs and with probability $1/2$, X and Y behave like independent exponential$(1/\mu)$ rvs.

Supposing that $1/\lambda$ is small and $1/\mu$ is large, then X and Y are (usually) small together half the time and large together half the time. In other words, they are positively correlated.

Example 4.17 (cont.) (Linear Regression) We assume

$$\mathsf{E}(Y|X) = \beta_0 + \beta_1 X \quad \text{and} \quad \mathsf{var}(Y|X) = \sigma^2.$$

We know from above that $\mu_Y = \beta_0 + \beta_1 \mu_X$ and $\sigma_Y^2 = \sigma^2 + \beta_1^2 \sigma_X^2$.

Again conditioning on X, we calculate

$$\begin{aligned}\mathsf{cov}(X,Y) &= \mathsf{E}(\mathsf{E}(XY|X)) - \mu_X \mu_Y \\ &= \mathsf{E}(X(\beta_0 + \beta_1 X)) - \mu_X(\beta_0 + \beta_1 \mu_X) = \beta_1 \sigma_X^2.\end{aligned}$$

Also,

$$\mathsf{corr}(X,Y) = \frac{\mathsf{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\beta_1 \sigma_X}{\sqrt{\sigma^2 + \beta_1^2 \sigma_X^2}} = \beta_1 \frac{\sigma_X}{\sigma_Y}.$$

The correlation will be close to 1 or -1 if the conditional variance σ^2 is small, relative to $\beta_1^2 \sigma_X^2$.

We can interpret σ^2 as measuring the scatter (noise) of Y away from the regression relationship (signal). If the noise is much greater than the relationship, however, the correlation will be small.

Again note that, *under our assumption* on the conditional expectation and variance, only the first two moments of the distributions are relevant, not the type of distribution.

THEOREM 4.34 (More Properties of Covariance and Correlation) Assume X, Y and W are jointly distributed and their variances exist.

i. $\text{cov}(X+W,Y) = \text{cov}(X,Y) + \text{cov}(W,Y)$.

ii. $\text{var}(aX+bY) = a^2\,\text{var}(X) + b^2\,\text{var}(Y) + 2ab\,\text{cov}(X,Y)$.

iii. If X and Y are independent then $\text{cov}(X,Y) = 0$ and $\text{corr}(X,Y) = 0$.

iv. $|\,\text{cov}(X,Y)| \leq \sigma_X\sigma_Y$ and $|\,\text{corr}(X,Y)| \leq 1$.

v. $|\,\text{corr}(X,Y)| = 1$ if and only if $\text{P}(Y = aX+b) = 1$ for some a and b, where a has the same sign as does $\text{corr}(X,Y)$.

PROOF i., ii., iii. (exercise).

iv. Choose $a = \sigma_Y$ and $b = -\sigma_X$ and apply ii. to get

$$0 \leq \text{var}(\sigma_Y X - \sigma_X Y) = 2\sigma_X^2\sigma_Y^2 - 2\sigma_X\sigma_Y\,\text{cov}(X,Y).$$

This says $\text{cov}(X,Y) \leq \sigma_X\sigma_Y$.

By choosing $b = \sigma_X$ instead, we get $\text{cov}(X,Y) \geq -\sigma_X\sigma_Y$.

These two inequalities are equivalent to $|\,\text{corr}(X,Y)| \leq 1$.

v. Furthermore, considering the case of equality, we have $\text{var}(\sigma_Y X - \sigma_X Y) = 0$ if and only if $\text{cov}(X, Y) = \sigma_X \sigma_Y$, which in turn is equivalent to $\text{corr}(X, Y) = 1$.

By Thm. 2.18.v., $\text{E}((W - \mu_W)^2) = 0$ if and only if $\text{P}(W = \mu_W) = 1$. So, in this case, $W = \sigma_Y X - \sigma_X Y$ is degenerate (i.e., a constant) and Y is a linear function of X with positive slope.

Likewise, if $\text{corr}(X, Y) = -1$ then Y is a linear function of X with negative slope. □

**** Perfect correlation means the relationship between the variables is exactly linear.*

So in general the correlation is just measuring *the degree* to which the dependence between the two random variables *is linear*.

This has the advantage that it focuses on the most useful and recognizable form of relationship, but it has the disadvantage that it *could be missing other types of dependence* and could therefore be misleading.

The correlation is 0 iff $\text{E}(XY) = \mu_X \mu_Y$.

**** However, zero correlation does not necessarily imply independence.*

Example 4.19 Suppose

$$f_{X,Y}(x,y) = \frac{y^2}{4\beta^4} e^{-x/\beta} 1_{(-x,x)}(y) 1_{(0,\infty)}(x).$$

That is, $X \sim \text{gamma}(4,\beta)$ and $f_{Y|X}(y|x) = \frac{3y^2}{2x^3} 1_{(-x,x)}(y)$.

Since $\text{E}(Y) = \text{E}(XY) = 0$, we must have $\text{cov}(X,Y) = 0$. Thus X and Y are uncorrelated, even though the pdf clearly does not factor and they are not independent.

Indeed, $\text{E}(X) = 4\beta$, $\text{E}(|Y|) = 3\beta$ and $\text{E}(X|Y|) = 15\beta^2$, so

$$\text{cov}(X,|Y|) = \text{E}(X|Y|) - \text{E}(X)\text{E}(|Y|) = 3\beta^2 > 0.$$

In addition, $\text{var}(X) = 4\beta^2$ and $\text{E}(|Y|^2) = 12\beta^2$ so that

$$\text{var}(|Y|) = 3\beta^2 \quad \text{and} \quad \text{corr}(X,|Y|) = \frac{\sqrt{3}}{2} \doteq .86603.$$

Thus, while there is no linear component to the relationship between X and Y, there certainly is a linear relationship between X and $|Y|$. In other words, a linear prediction of Y from X would be useless, but we *can* predict $|Y|$ linearly.

(Note: for this example, $\text{E}(Y|X) = 0$ also.)

We have indicated that the correlation measures the degree to which a relationship is linear, but how would one actually predict linearly? Linear predictors are always convenient, after all.

THEOREM 4.35 (Best Linear Predictor) Assume X and Y are jointly distributed and their variances exist.

The function $aX+b$ that minimizes $\mathsf{E}((Y-(aX+b))^2)$ is $\mu_Y+\frac{\mathsf{cov}(X,Y)}{\sigma_X}\left(\frac{X-\mu_X}{\sigma_X}\right)$.

PROOF It suffices to show that the best linear predictor of $Y-\mu_Y$ is $\frac{\mathsf{cov}(X,Y)}{\sigma_X^2}(X-\mu_X)$. By Thm. 4.33.iii. and Thm. 4.34.ii.,

$$\mathsf{E}(((Y-\mu_Y)-(a(X-\mu_X)+b_*))^2)=\sigma_Y^2-2a\,\mathsf{cov}(X,Y)+a^2\sigma_X^2+b_*^2.$$

Obviously, $b_*=0$ is required to minimize this.

Then the expression is a quadratic in a that can easily be minimized to get $a=\frac{\mathsf{cov}(X,Y)}{\sigma_X^2}$. We then get $b=\mu_Y-a\mu_X$ for the best linear predictor of Y. □

If the regression function $\mathsf{E}(Y|X)$ is *linear* in X then the best linear predictor is the best overall predictor, namely, $\mathsf{E}(Y|X)$ (see Thm. 4.31). But otherwise the two predictors are not the same.

Our final result for this section states another of those concepts that seems so reasonable, we are tempted to take it for granted.

THEOREM 4.36 Let X be any rv and suppose $g(x)$ and $h(x)$ are *nondecreasing and nonconstant* functions on an interval containing the support of X.

Then, assuming existence of the appropriate expectations, $U = g(X)$ and $V = h(X)$ are *positively correlated*.

PROOF We need only to show $\text{cov}(g(X), h(X)) > 0$. Let $\mu = \text{E}(g(X))$ and $\lambda = \text{E}(h(X))$. Since h is nondecreasing there exists a such that $h(x) \leq \lambda$ for $x < a$ and $h(x) \geq \lambda$ for $x \geq a$. Since g is also nondecreasing, it follows that

$$(g(x) - g(a))(h(x) - \lambda) \geq 0, \quad \text{for all } x.$$

Hence,

$$\begin{aligned} \text{cov}(g(X), h(X)) &= \text{E}((g(X) - g(a) + g(a) - \mu)(h(X) - \lambda)) \\ &= \text{E}((g(X) - g(a))(h(X) - \lambda)) + (g(a) - \mu)\text{E}((h(X) - \lambda)) \\ &> 0. \end{aligned}$$

The last inequality is strict because $g(X)$ and $h(X)$ are not constants, so $(g(X) - g(a))(h(X) - \lambda) > 0$ with positive probability. □

Thus, for example, if X is a nonnegative random variable with at least four finite moments then X and X^2 are positively correlated, while X and e^{-X} are negatively correlated (because X and $-\mathrm{e}^{-X}$ are positively correlated by Thm. 4.36).

On the other hand, even though nonnegative X and X^2 have a *perfect* (and 1-1) relationship, $\text{corr}(X, X^2) < 1$ because the relationship is not strictly linear.

4.5 Bivariate Normal Distribution

Although we have seen special cases, we now discuss this very important bivariate distribution in some generality. We start by returning to the linear regression example.

Example 4.17 (cont.) Suppose we again make the normality assumption: $Y|\{X = x\} \sim \text{normal}(\beta_0 + \beta_1 x, \sigma^2)$ and $X \sim \text{normal}(\mu_X, \sigma_X^2)$. Thus the joint pdf is

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y|x) = \frac{1}{2\pi\sigma\sigma_X} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{1}{2}\left(\frac{y-\beta_0+\beta_1 x}{\sigma}\right)^2},$$

with the *characteristic quadratic function* in the exponent (and no other dependence on the variables).

By our "completing the squares" argument (see before), any joint pdf $f(x,y)$ for which $\log f(x,y)$ is a quadratic function in x and y must look like this example. (Otherwise there would be extra terms that prevent the density from integrating to a finite value.)

Recall that $\mu_Y = \beta_0 + \beta_1\mu_X$, $\sigma_Y^2 = \sigma^2 + \beta_1^2\sigma_X^2$ and $\text{corr}(X,Y) = \frac{\beta_1\sigma_X}{\sigma_Y}$.

Define $\rho = \text{corr}(X,Y)$ and assume $|\rho| \neq 1$. We may expand the exponent in the pdf as

$$\begin{aligned}
&-\frac{1}{2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\left(\frac{y-\mu_Y}{\sigma}\right) - \beta_1\left(\frac{x-\mu_X}{\sigma}\right)\right)^2\right) \\
&= -\frac{\sigma_Y^2}{2\sigma^2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\beta_1\sigma_X}{\sigma_Y}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right) \\
&= -\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right).
\end{aligned}$$

This leads to the following definition.

DEFINITION 4.37 (X,Y) has bivariate normal$(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ distribution if its joint pdf is

$$f_{X,Y}(x,y) = \frac{\mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

for all (x,y) in the real plane.

See Fig. 4.3 for an example with $\rho = .7$.

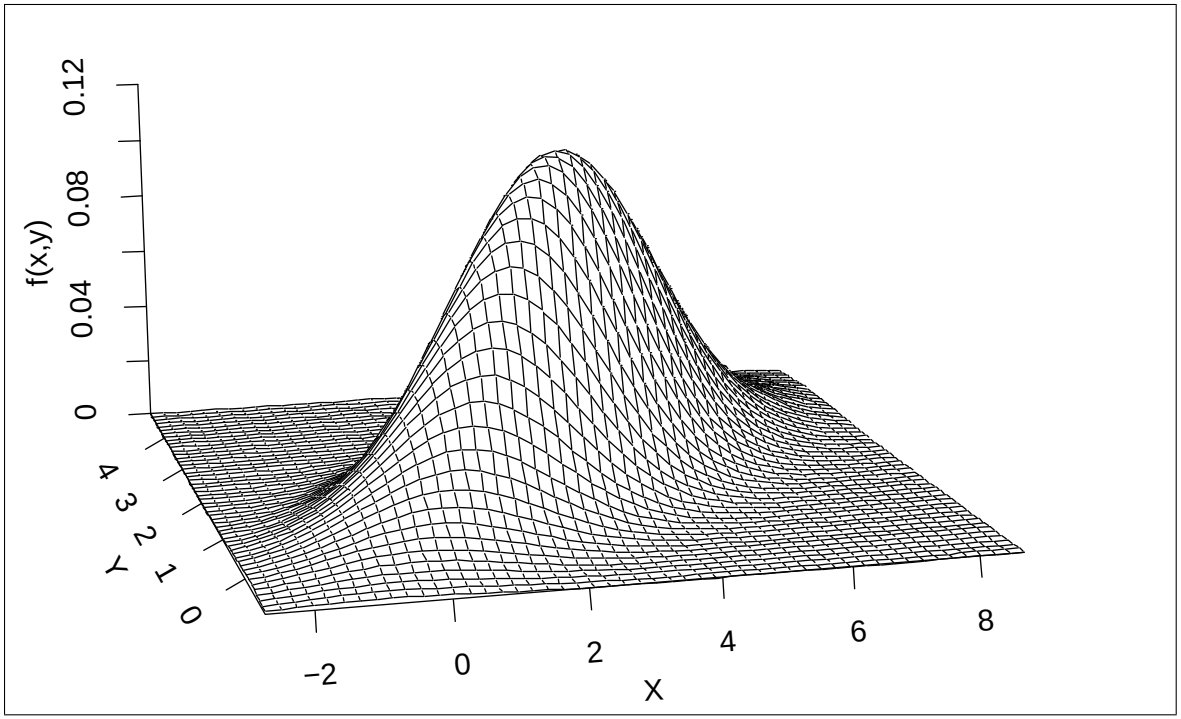

Figure 4.3 Bivariate normal density function, with $\mu_X = 3$, $\mu_Y = 2$, $\sigma_X^2 = 4$, $\sigma_Y^2 = 1$ and $\rho = .7$.

Note that if X and Y are *independent normal rvs*, then their joint density has the form in Def. 4.37 with $\rho = 0$ and so (X, Y) is bivariate normal.

On the other hand, if not independent, X and Y could have *normal marginal* distributions but *not be bivariate normal*, which is very specialized.

More important than the formula are the properties of this distribution.

THEOREM 4.38 (Properties of the Bivariate Normal Distribution)

Assume (X, Y) has bivariate normal$(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ distribution.

i. $\rho = \text{corr}(X, Y)$.

ii. X has normal(μ_X, σ_X^2) distribution and Y has normal(μ_Y, σ_Y^2) distribution.

iii. $Y|\{X = x\}$ is normal$(\beta_0 + \beta_1 x, \sigma^2)$ for every x, where $\beta_0 = \mu_Y - \beta_1 \mu_X$, $\beta_1 = \rho\sigma_Y/\sigma_X$ and $\sigma^2 = (1 - \rho^2)\sigma_Y^2$.
$(X|\{Y = y\}$ is likewise conditionally normal – see below.)

iv. X and Y are independent if and only if $\rho = 0$.

v. If $a_1 b_2 - a_2 b_1 \neq 0$ then $(U, V) = (a_1 X + a_2 Y + a_3, b_1 X + b_2 Y + b_3)$ has bivariate normal distribution.

vi. $Z_1 = \frac{1}{\sqrt{2(1+\rho)}}\left(\left(\frac{X-\mu_X}{\sigma_X}\right) + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)$ and $Z_2 = \frac{1}{\sqrt{2(1-\rho)}}\left(\left(\frac{X-\mu_X}{\sigma_X}\right) - \left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)$ are *independent* standard normal rvs.

Remember that $Y|\{X = x\}$ is not a random variable but merely refers to the conditional distribution of Y given $X = x$.

PROOF i., ii. and iii. These follow from the previous example and the fact (mentioned in the example) that no other quadratic expression in the exponent would result in a proper joint density.

iv. The joint pdf factors iff $\rho = 0$.

v. Using the method of bivariate transformations given in Thm. 4.18, it is easy to see that the joint pdf for (U, V) again has the form $ce^{-g(u,v)}$ where $g(u, v)$ is a quadratic function of u and v.

The condition $a_1b_2 - a_2b_1 \neq 0$ is required to ensure a nonzero determinant of the Jacobian for the transformation.

vi. By iv., Z_1 and Z_2 are jointly bivariate normal. They are thus individually normal by ii. It is easy to check that they each have mean $= 0$ and variance $= 1$.

Applying Thm. 4.34.i. (twice), we find that $\text{cov}(Z_1, Z_2) = 0$ and hence they are independent by part iii. □

What happens in the case $\rho = \text{corr}(X, Y)$ is equal to 1 or -1?

The family of bivariate normal distributions is an affine family – it is closed under all linear transformations and the contours of the density are elliptical. This is a generalization of the location-scale family for a single normal rv. Few useful multivariate distributions have this property, and it turns out to have important benefits for statistical analysis.

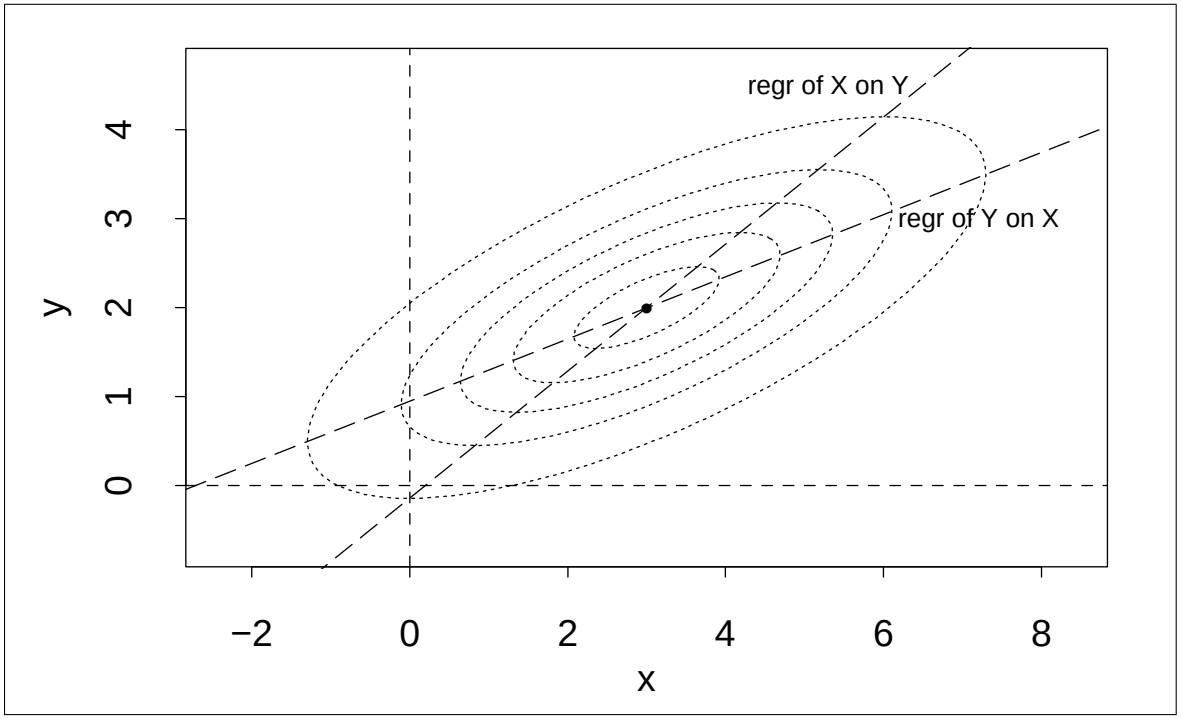

Figure 4.4 Contour plot of bivariate normal(3,2,4,1,.7) density function, with the graphs of the regression of Y on X and of X on Y. The contour lines contain, starting with the innermost, 10%, 30%, 50%, 70% and 90% of the probability.

From Thm. 4.38.ii. we see that the regression of Y on X is linear. Likewise, the regression of X on Y is linear.

But these regressions *do not have the same graph* in a plot of Y vs. X.

- The regression of X on Y is $x = \mu_X + \rho\sigma_X(y - \mu_Y)/\sigma_Y$ which has a graph given by $y = \mu_Y + \sigma_Y(x - \mu_X)/(\rho\sigma_X)$.
- The slope $(\frac{\sigma_Y}{\rho\sigma_X})$ of this is greater than the slope $(\frac{\rho\sigma_Y}{\sigma_X})$ of the regression of Y on X. See Fig. 4.4.
- This is because the regression of Y on X is concerned with the best prediction of Y (vertical fit), while the regression of X on Y is concerned with the best prediction of X (horizontal fit).
- Neither regression is the major axis of the density's elliptical contours. In fact the term "regression" refers to how the regression of Y on X is a *movement back* toward the constant graph $y = \mu_Y$.

As a consequence, one must be careful to interpret the relationship between X and Y appropriately.

Example 4.20 Log_{10} of Nile River inflows for January and February, 1874–1988.

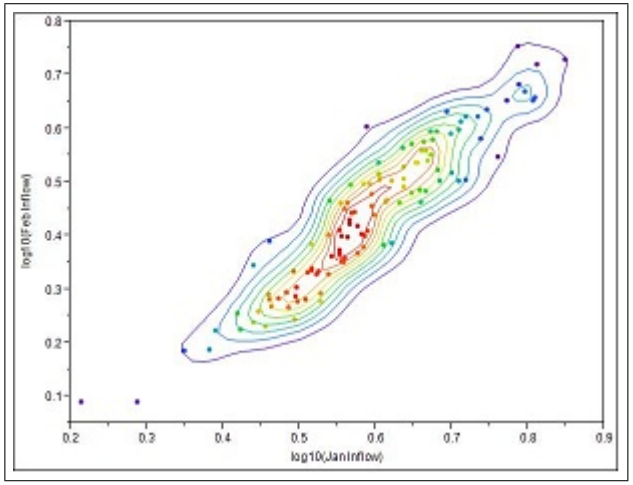

Figure 4.5 Estimated bivariate contours for Nile data.

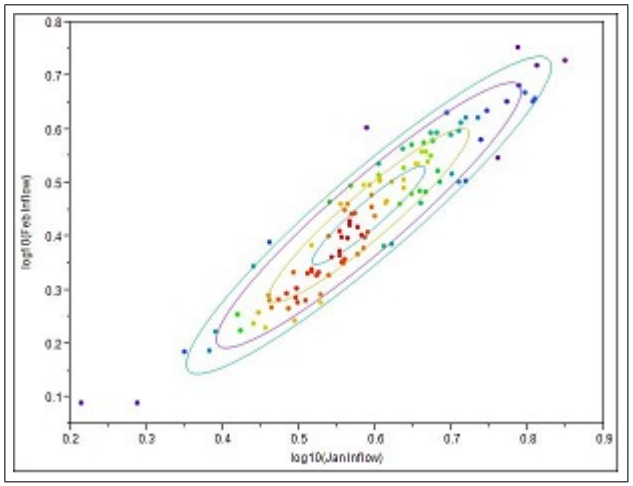

Figure 4.6 Estimated bivariate normal contours for Nile data.

4.6 Mixtures and Hierarchical Models

One can extend various models by *mixing* them in useful ways. Essentially, this amounts to interpreting the parameter(s) as random also.

CONCEPT 4.39 A mixture distribution is one that can be thought of as the distribution that results when the parameter is chosen at random according to some other distribution. That is, it is the marginal distribution of the data when the data and parameter are jointly distributed.

A hierarchical model is a mixture model defined in a sequential fashion either for ease of defining the model or for statistical purposes or both.

These notions are obviously somewhat vague – they really are no more than descriptive of how we envision and utilize certain kinds of models.

The main point is that the concept sets up the possibility for taking advantage of conditioning.

Example 4.5 (cont.) Suppose (X, Y) has joint pdf

$$f_{X,Y}(x,y) = \tfrac{1}{2}\left(\lambda^2 \mathrm{e}^{-\lambda(x+y)} + \mu^2 \mathrm{e}^{-\mu(x+y)}\right) 1_{(0,\infty)}(x) 1_{(0,\infty)}(y).$$

This is a simple mixture of two bivariate density functions,

$$f_1(x,y) = \lambda^2 \mathrm{e}^{-\lambda(x+y)} \quad \text{and} \quad f_2(x,y) = \mu^2 \mathrm{e}^{-\mu(x+y)},$$

each chosen with probability $1/2$.

Both bivariate distributions are for independent exponential rvs; they just have different means.

Note, however, that if you only observe the pair (X, Y) then you will not know (for certain) *which of the two* bivariate distributions it was sampled from.

Example 4.21 (Contaminated Normal) Many kinds of data are commonly assumed to be normal in distribution. But it is also quite usual for this assumption to be unreasonable.

Sometimes one may think of the data as having random "outliers" or values that, for whatever reason, are unusual. If the "usual" data have one normal distribution then the outliers might have another.

The actual data observed are thus a mixture of these two types.

Specifically, let $\phi(x)$ be the standard normal pdf. Assume that most of the data, say proportion $1-\delta$, come from a normal(μ, σ^2) distribution, which has pdf $\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$.

However a small proportion δ of the data are outliers that come from normal distribution with very different mean and/or standard deviation, say λ and τ, respectively.

Since you do not know which is the case for an individual datum then the observed data have density

$$f(x) = (1-\delta)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) + \delta\frac{1}{\tau}\phi\left(\frac{x-\lambda}{\tau}\right).$$

See Fig. 4.7 and Fig. 4.8.

We can interpret this as saying the experiment first selects the parameter at random, (μ, σ^2) with probability $1 - \delta$ and (λ, τ) with probability δ, and then selects the rv from the corresponding choice of distribution.

But, again, the observer does not know which distribution was chosen.

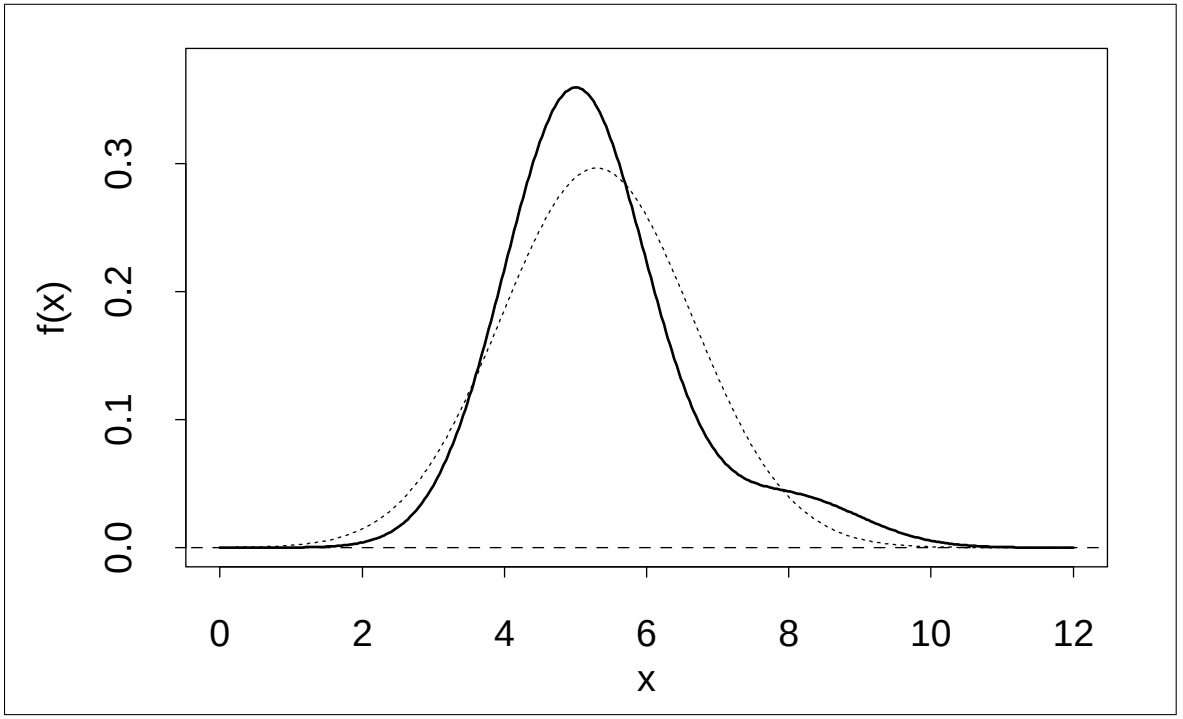

Figure 4.7 Contaminated normal density with outliers having mean three standard deviations away. The dotted overlay is the normal density with equivalent mean and variance.

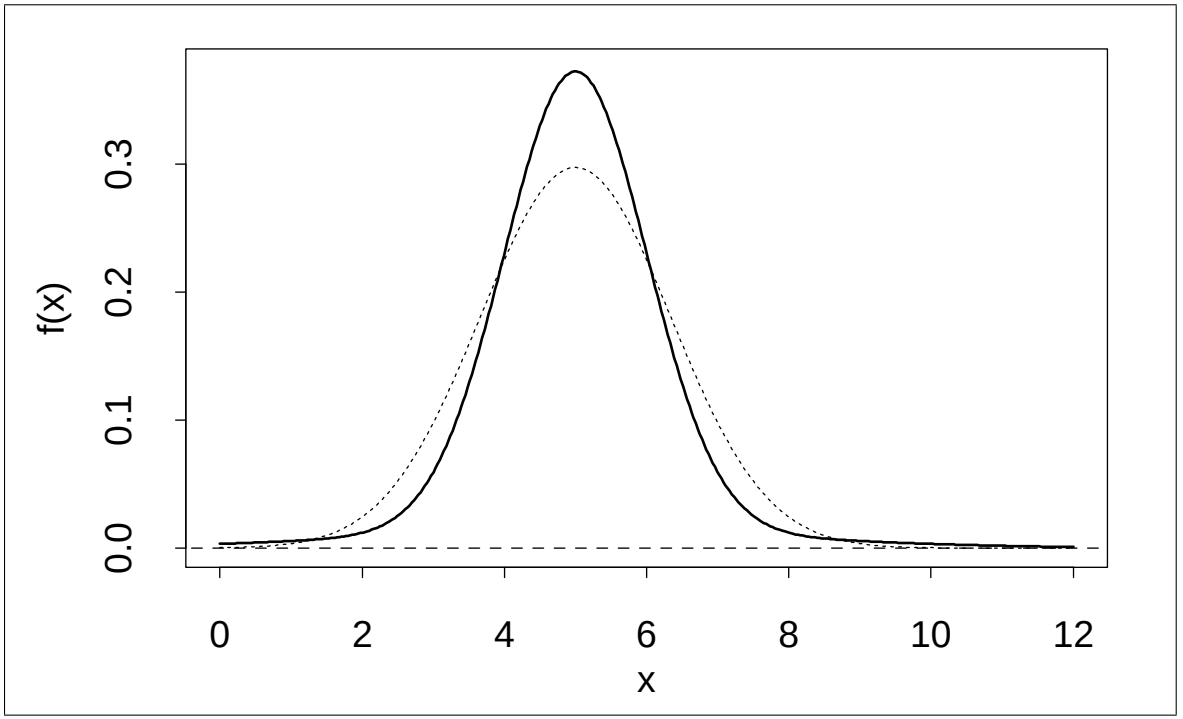

Figure 4.8 Contaminated normal density with outliers having the same mean but three times the standard deviation. The dotted overlay is the normal density with equivalent mean and variance.

Another version of this is when we want to model a bimodal distribution (one whose density has two peaks). Here, the mixing parameter δ is not necessarily small.

Indeed, one may in fact interpret the data as coming from two sub-populations, each of which has its own normal distribution. This occurs frequently with biological data. See Fig. 4.9.

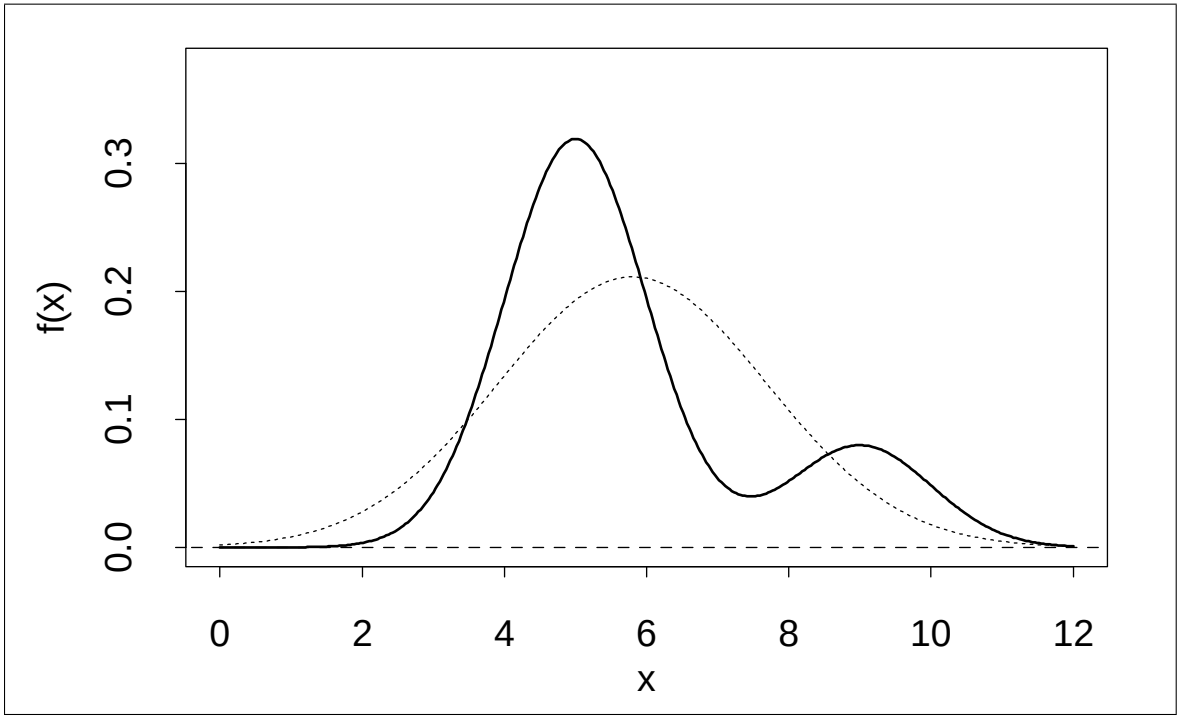

Figure 4.9 Bimodal density resulting from a mixture of two normal densities. The dotted overlay is the normal density with equivalent mean and variance.

Example 4.22 Suppose Y has pdf $f_Y(y) = \frac{\lambda}{5}(1+\lambda y)^2 \mathrm{e}^{-\lambda y}$, for $y > 0$.

This can be represented as a mixture of three gamma densities:

$$f_Y(y) = \left(\frac{1}{5}\right)\lambda \mathrm{e}^{-\lambda y} + \left(\frac{2}{5}\right)\lambda^2 y \mathrm{e}^{-\lambda y} + \left(\frac{2}{5}\right)\tfrac{1}{2}\lambda^3 y^2 \mathrm{e}^{-\lambda y}.$$

In fact, suppose α is a random variable with $\mathsf{P}(\alpha = 1) = 1/5$ and $\mathsf{P}(\alpha = 2) = \mathsf{P}(\alpha = 3) = 2/5$, and suppose further that $\widetilde{Y}|\alpha \sim \text{gamma}(\alpha, 1/\lambda)$. Then the pdf for Y is the same as that for $\widetilde{Y}$.

We can therefore calculate

$$\mathsf{E}(Y) = \mathsf{E}(\mathsf{E}(\widetilde{Y}|\alpha)) = \mathsf{E}(\alpha/\lambda) = \frac{11}{5\lambda}$$

and

$$\mathsf{E}(Y^2) = \mathsf{E}(\mathsf{E}(\widetilde{Y}^2|\alpha)) = \mathsf{E}(\alpha(\alpha+1)/\lambda^2) = \frac{38}{5\lambda^2}.$$

Or even

$$\begin{aligned}\mathsf{var}(Y) &= \mathsf{E}(\mathsf{var}(\widetilde{Y}|\alpha)) + \mathsf{var}(\mathsf{E}(\widetilde{Y}|\alpha)) \\ &= \mathsf{E}(\alpha/\lambda^2) + \mathsf{var}(\alpha/\lambda) = \frac{11}{5\lambda^2} + \frac{14}{25\lambda^2}.\end{aligned}$$

This shows how recognizing that a distribution can be expressed as a mixture could allow for alternative ways to compute probabilities and expectations.

Example 4.2 (cont.) (Marked Poisson) This can be thought of as a hierarchical model where the first stage of the experiment is to select the number of trials $Y \sim \text{Poisson}(\lambda)$ and the second is to select the number of successes $X|Y \sim \text{binomial}(Y, p)$.

We could even make it a three-stage model by letting the first stage be a random selection of the parameters λ and p. This would be appropriate if, for example, the parameters are known to change in some random fashion.

It also is useful for Bayesian statistical analysis when you want to characterize λ and p as random (with a prior distribution). For such analysis, the objective is to find the conditional posterior distribution of the parameters (λ, p), given the data (X, Y).

Example 4.18 (cont.) N is a Poisson(λ) number of insurance claims, each of which has exponential(β) distribution and are independent. We saw that the total claims Y is such that $Y|N \sim \text{gamma}(N, \beta)$. This is again a two stage hierarchical model.

Both of the last two examples are special cases of a compound Poisson rv, i.e., the sum of a Poisson number of independent and identically distributed rvs. This kind of rv appears in various stochastic processes as well.

Example 4.23 Suppose X_1 and X_2 are independent chi-square(1) rvs and $\delta = \pm 1$ w.p. $1/2$ each, independent of X_1 and X_2.

(δ is known as a Rademacher rv and in fact $(\delta + 1)/2 \sim$ Bernoulli$(1/2)$.)

Define $(Y_1, Y_2) = (\delta\sqrt{X_1}, \delta\sqrt{X_2})$.

Recall (Thm. 3.11) that the *square of a standard normal* has chi-square(1) dist.

Thus Y_1, which is equally likely the negative or positive square root of a chi-square(1) rv, must have standard normal distribution. Likewise for Y_2.

But (Y_1, Y_2) is *not bivariate normal* because Y_1 and Y_2 always have the same sign. In fact, $\mathrm{E}(Y_2|Y_1) = \sqrt{\frac{2}{\pi}}\,\mathrm{sign}(Y_1)$, which is *not linear* in Y_1.

In the examples above, we used a discrete rv to mix a continuous one. We could have just as well used a continuous rv to mix a discrete one. In either case, we usually want to describe their joint distribution.

Suppose, for example, that X and Y are jointly distributed but X is discrete and Y is continuous. Sometimes, a convenient description is with the joint "density" function:

$$f_{X,Y}(x,y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathsf{P}(X = x, Y \leq y).$$

We can still factor this into marginal and conditional functions. That is,

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

We can also compute expectations (and probabilities) by

$$\mathsf{E}(g(X,Y)) = \sum_x \int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)\mathrm{d}y.$$

**** We just have to be mindful of the different roles played by the variables.*

Example 4.24 Suppose Y has Poisson distribution with parameter Λ that itself is a random variable with exponential(β) distribution.

Find the (unconditional) pmf for Y.

Solution First,

$$f_\Lambda(\lambda) = \frac{1}{\beta}\mathrm{e}^{-\lambda/\beta} \text{ (pdf)},$$

and

$$f_{Y|\Lambda}(y|\lambda) = \frac{\lambda^y \mathrm{e}^{-\lambda}}{y!} \text{ (conditional pmf)}.$$

Therefore, Y has pmf

$$\begin{aligned} f_Y(y) &= \int_0^\infty \frac{\lambda^y \mathrm{e}^{-\lambda}}{y!}\frac{1}{\beta}\mathrm{e}^{-\lambda/\beta}\,\mathrm{d}\lambda = \frac{1}{\beta(1+1/\beta)^{y+1}}\int_0^\infty \frac{(1+1/\beta)^{y+1}\lambda^y \mathrm{e}^{-(1+1/\beta)\lambda}}{y!}\,\mathrm{d}\lambda \\ &= \frac{1}{\beta+1}\left(\frac{\beta}{\beta+1}\right)^y, \quad \text{for } y = 0, 1, 2\ldots, \end{aligned}$$

where we have used a gamma($y+1, (1+1/\beta)^{-1}$) pdf in the integral.

Hence, $Y \sim$ negative binomial$(1, \frac{1}{\beta+1})$ and $T = Y + 1 \sim$ geometric$(\frac{1}{\beta+1})$.

What is the conditional distribution for Λ, given Y? (exercise)

4.7 Copulas

How might one develop a bivariate distribution that has *given marginal distributions*?

If we have cdfs F_1 and F_2 then $F(x, y) = F_1(x)F_2(y)$ is the joint distribution for some *independent* random variables $X \sim F_1$ and $Y \sim F_2$.

But if we want *dependent* random variables $X \sim F_1$ and $Y \sim F_2$ then we have to determine the nature of the dependence in addition to the marginal distributions.

There are many ways to do this, and what works best will depend on how the model is used. Here we briefly discuss one option that has become popular. It starts with a joint distribution for *dependent uniform* random variables.

DEFINITION 4.40 Suppose $G(u, v)$ is the joint cdf for (U, V) and each of U and V are uniform(0,1) random variables. Then $G(u, v)$ is called a copula.

Observe that the marginal cdfs are uniform(0,1) for a cdf $G(u, v)$ if and only if $G(u, 1) = u$ and $G(1, v) = v$, for $0 \leq u \leq 1$ and $0 \leq v \leq 1$ (and $G(1, 1) = 1$).

Now we extend to a joint distribution with arbitrary *specified* marginal distributions.

THEOREM 4.41 Suppose (U, V) have joint cdf $G(u, v)$ that is a copula. Let Q_1 and Q_2 be the quantile functions for cdfs F_1 and F_2, respectively, and let $X = Q_1(U)$ and $Y = Q_2(V)$.

Then $X \sim F_1$, $Y \sim F_2$ and $(X, Y) \sim F(x, y) = G(F_1(x), F_2(y))$.

PROOF Since U and V have uniform(0,1) distribution, we know from Thm. 2.37 that $X = Q_1(U) \sim F_1$ and $Y = Q_2(V) \sim F_2$. This was based on the fact

$$Q_1(p) \le x \iff p \le F_1(x) \quad \text{and} \quad Q_2(p) \le x \iff p \le F_2(x).$$

By the same fact, the joint cdf for (X, Y) is

$$\begin{aligned} F(x, y) &= \mathsf{P}(Q_1(U) \le x, Q_2(V) \le y) \\ &= \mathsf{P}(U \le F_1(x), V \le F_2(y)) = G(F_1(x), F_2(y)). \end{aligned}$$

□

The modeling problem then becomes one of picking the right kind of copula.

Example 4.25 (Farlie-Morgenstern Copula) Let $\alpha \in [-1,1]$ and (U,V) have joint cdf

$$G(u,v) = uv(1+\alpha(1-u)(1-v)) \quad \text{for } 0<u<1,\ 0<v<1.$$

First, let us verify that this is indeed a copula. Taking derivatives,

$$\begin{aligned} g(u,v) &= \frac{\partial^2}{\partial u\,\partial v} G(u,v) = \left(\frac{\partial^2}{\partial u\,\partial v}\, uv + \alpha \frac{\partial^2}{\partial u\,\partial v}\, u(1-u)v(1-v)\right) \\ &= 1+\alpha(1-2u)(1-2v). \end{aligned}$$

Since $|\alpha(1-2u)(1-2v)| \le 1$ for $0<u<1$ and $0<v<1$, we see that $g(u,v)$ is nonnegative.

Moreover, it is easy to check that $g(u,v)$ integrates to 1 and has marginal uniform(0,1) pdfs. ($G(u,1)=u$ and $G(1,v)=v$.) Hence, $G(u,v)$ is a copula.

By Thm. 4.41, if $F_1(x)$ and $F_2(y)$ are any cdfs on the real line then

$$F(x,y) = F_1(x)F_2(y)(1+\alpha(1-F_1(x))(1-F_2(y)))$$

is a bivariate cdf for some random pair (X,Y).

Observe that $\alpha = 0$ implies X and Y are independent.

Specifically, suppose we want $X \sim$ exponential(1) and $Y \sim$ exponential$(1/2)$, with α as a parameter for dependence. Then

$$F(x,y) = (1-\mathrm{e}^{-x})(1-\mathrm{e}^{-2y})(1+\alpha\mathrm{e}^{-x-2y}),$$

with joint density

$$f(x,y) = \frac{\partial^2}{\partial x\,\partial y}F(x,y) = 2\mathrm{e}^{-x-2y}(1+\alpha(2\mathrm{e}^{-x}-1)(2\mathrm{e}^{-2y}-1)).$$

Copulas are mostly used when you want continuous marginal distributions. In that case, we have the following.

THEOREM 4.42 Suppose $(X,Y) \sim G(F_1(x), F_2(x))$ where $G(u,v)$ is a copula with pdf $g(u,v)$ and F_i has pdf f_i, $i = 1, 2$.

Then (X,Y) has joint pdf

$$f(x,y) = f_1(x)f_2(y)g(F_1(x), F_2(y)).$$

PROOF (exercise.) □

What is the copula cdf $G(u,v)$ and pdf $g(u,v)$ for Ex. 4.23? (exercise.)

Example 4.26 Let $\alpha \in [-1, 1]$ and

$$m = \max_{u \in [0,1]} (u(1-u)|1-2u|) \approx 0.096225.$$

Now consider the joint pdf

$$g(u, v) = 1 + \frac{\alpha}{m^2} u(1-u)(1-2u)v(1-v)(1-2v), \quad 0 \leq u \leq 1, 0 \leq v \leq 1.$$

It is not hard to show that this is the pdf for a copula.

Thus,

$$f(x, y) = \phi(x)\phi(y)g(\Phi(x), \Phi(y))$$

is a joint pdf with standard normal marginal pdfs.

