3. Bivariate Random Variables

Dave Goldsman

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

1/15/20

Lesson 3.1 — Introduction

In this introductory lesson, we'll cover …

- What we mean by bivariate (or joint) random variables.
- The discrete case.
- The continuous case.
- Bivariate cdf's.

In this module, we'll look at what happens when you consider two random variables *simultaneously*.

Example: Choose a person at random. Look at their height and weight (X, Y). Obviously, X and Y will be related somehow.

Discrete Case

Definition: If X and Y are discrete random variables, then (X, Y) is called a **jointly discrete bivariate random variable**.

The **joint (or bivariate) pmf** is

$$f(x, y) \;=\; P(X = x, Y = y), \quad \forall x, y.$$

Properties:

- $0 \leq f(x, y) \leq 1$.
- $\sum_x \sum_y f(x, y) = 1$.
- $A \subseteq \mathbb{R}^2 \Rightarrow P((X, Y) \in A) = \sum\sum_{(x,y)\in A} f(x, y)$.

Example: 3 sox in a box (numbered 1,2,3). Draw 2 sox at random without replacement. $X =$ # of the first sock; $Y =$ # of the second sock. The joint pmf $f(x, y)$ is

$f(x,y)$	$X=1$	$X=2$	$X=3$	$P(Y=y)$
$Y=1$	0	1/6	1/6	1/3
$Y=2$	1/6	0	1/6	1/3
$Y=3$	1/6	1/6	0	1/3
$P(X=x)$	1/3	1/3	1/3	1

$f_X(x) \equiv P(X = x)$ is the **"marginal" pmf of X**.

$f_Y(y) \equiv P(Y = y)$ is the **"marginal" pmf of Y**.

By the Law of Total Probability,

$$P(X=1) \;=\; \sum_{y=1}^{3} P(X=1, Y=y) \;=\; 1/3.$$

In addition,

$$\begin{aligned} &P(X \geq 2, Y \geq 2) \\ &\quad= \sum_{x \geq 2} \sum_{y \geq 2} f(x,y) \\ &\quad= f(2,2) + f(2,3) + f(3,2) + f(3,3) \\ &\quad= 0 + 1/6 + 1/6 + 0 \;=\; 1/3. \quad \square \end{aligned}$$

Continuous Case

Definition: If X and Y are continuous RVs, then (X, Y) is a **jointly continuous bivariate RV** if there exists a magic function $f(x, y)$ such that

- $f(x, y) \geq 0, \forall x, y.$
- $\int\int_{\mathbb{R}^2} f(x, y)\, dx\, dy = 1.$
- $P(A) = P((X, Y) \in A) = \int\int_A f(x, y)\, dx\, dy.$

In this case, $f(x, y)$ is called the **joint pdf**.

If $A \subseteq \mathbb{R}^2$, then $P(A)$ is the volume between $f(x, y)$ and A.

Think of

$$f(x, y)\, dx\, dy \;\approx\; P(x < X < x + dx, y < Y < y + dy).$$

It's easy to see how this generalizes the 1-dimensional pdf, $f(x)$.

Example: Choose a point (X, Y) at random in the interior of the circle inscribed in the unit square, e.g., $C \equiv (x - \frac{1}{2})^2 + (y - \frac{1}{2})^2 \leq \frac{1}{4}$.

Find the pdf of (X, Y).

Since the area of the circle is $\pi/4$,

$$f(x,y) \;=\; \begin{cases} 4/\pi & \text{if } (x,y) \in C \\ 0 & \text{otherwise.} \end{cases} \qquad \Box$$

Application: Toss n darts randomly into the unit square. The probability that any individual dart will land in the circle is $\pi/4$. It stands to reason that the proportion of darts, $\hat{p}_n$, that land in the circle will be approximately $\pi/4$. So you can use $4\hat{p}_n$ to estimate π!

Example: Suppose that

$$f(x,y) = \begin{cases} 4xy & \text{if } 0 \le x \le 1,\, 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the probability (volume) of the region $0 \le y \le 1 - x^2$.

$$\begin{aligned} V &= \int_0^1 \int_0^{1-x^2} 4xy\,dy\,dx \\ &= \int_0^1 \int_0^{\sqrt{1-y}} 4xy\,dx\,dy \\ &= 1/3. \end{aligned}$$

Moral: Be careful with limits! $\square$

Bivariate cdf's

Definition: The **joint (bivariate) cdf** of X and Y is $F(x,y) \equiv P(X \le x, Y \le y)$, for all x, y.

$$F(x,y) = \begin{cases} \sum\sum_{s \le x, t \le y} f(s,t) & \text{discrete} \\ \int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)\,ds\,dt & \text{continuous.} \end{cases}$$

Going from cdf's to pdf's (continuous case):

1-dimension: $f(x) = F'(x) = \frac{d}{dx}\int_{-\infty}^{x} f(t)\,dt$.

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

Properties:

$F(x, y)$ is non-decreasing in both x and y.

$\lim_{x\to-\infty} F(x, y) = \lim_{y\to-\infty} F(x, y) = 0$.

$\lim_{x\to\infty} F(x, y) = F_Y(y) = P(Y \le y)$ **("marginal" cdf of Y).**

$\lim_{y\to\infty} F(x, y) = F_X(x) = P(X \le x)$ **("marginal" cdf of X).**

$\lim_{x\to\infty} \lim_{y\to\infty} F(x, y) = 1$.

$F(x, y)$ is continuous from the right in both x and y.

Example: Suppose

$$F(x,y) \;=\; \begin{cases} 1-e^{-x}-e^{-y}+e^{-(x+y)} & \text{if } x\ge 0,\, y\ge 0 \\ 0 & \text{if } x<0 \text{ or } y<0. \end{cases}$$

The marginal cdf of X is

$$F_X(x) \;=\; \lim_{y\to\infty} F(x,y) \;=\; \begin{cases} 1-e^{-x} & \text{if } x\ge 0 \\ 0 & \text{if } x<0. \end{cases}$$

The joint pdf is

$$\begin{aligned} f(x,y) &= \frac{\partial^2}{\partial x\partial y}F(x,y) \\ &= \frac{\partial}{\partial y}(e^{-x}-e^{-y}e^{-x}) \\ &= e^{-(x+y)} \quad \text{if } x\ge 0,\, y\ge 0. \quad \Box \end{aligned}$$

Lesson 3.2 — Marginal Distributions

We're also interested in the individual (marginal) distributions of X and Y.

Definition: If X and Y are jointly *discrete*, then the **marginal pmf's** of X and Y are, respectively,

$$\boxed{f_X(x) \;=\; P(X=x) \;=\; \sum_y f(x,y)}$$

and

$$\boxed{f_Y(y) \;=\; P(Y=y) \;=\; \sum_x f(x,y).}$$

Example (discrete case): $f(x,y) = P(X = x, Y = y)$.

$f(x,y)$	$X=1$	$X=2$	$X=3$	$P(Y=y)$
$Y=40$	0.01	0.07	0.12	0.2
$Y=60$	0.29	0.03	0.48	0.8
$P(X=x)$	0.3	0.1	0.6	1

By total probability,

$$P(X = 1) = P(X = 1, Y = \text{ any \#}) = 0.3. \quad \Box$$

Example (discrete case): $f(x,y) = P(X = x, Y = y)$.

$f(x,y)$	$X=1$	$X=2$	$X=3$	$P(Y=y)$
$Y=40$	0.06	0.02	0.12	0.2
$Y=60$	0.24	0.08	0.48	0.8
$P(X=x)$	0.3	0.1	0.6	1

Remark: Hmmm.... Compared to the last example, this has the *same marginals* but *different joint* distribution! That's because the joint distribution contains *much more information* than just the marginals.

Definition: If X and Y are jointly *continuous*, then the **marginal pdf's** of X and Y are, respectively,

$$\boxed{f_X(x) \;=\; \int_{\mathbb{R}} f(x,y)\,dy \quad \text{and} \quad f_Y(y) \;=\; \int_{\mathbb{R}} f(x,y)\,dx.}$$

Example:

$$f(x,y) \;=\; \begin{cases} e^{-(x+y)} & \text{if } x \geq 0,\, y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then the marginal pdf of X is

$$f_X(x) \;=\; \int_{\mathbb{R}} f(x,y)\,dy \;=\; \int_0^\infty e^{-(x+y)}\,dy \;=\; e^{-x},\ \text{if } x \geq 0. \qquad \square$$

Example:

$$f(x,y) = \begin{cases} \frac{21}{4}x^2y & \text{if } x^2 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Note ***funny limits*** where the pdf is positive, i.e., $x^2 \le y \le 1.$

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy = \int_{x^2}^{1} \frac{21}{4}x^2y\,dy = \frac{21}{8}x^2(1-x^4), \quad -1 \le x \le 1.$$

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{21}{4}x^2y\,dx = \frac{7}{2}y^{5/2}, \quad 0 \le y \le 1. \quad \Box$$

Lesson 3.3 — Conditional Distributions

Recall conditional probability: $P(A|B) = P(A \cap B)/P(B)$ if $P(B) > 0$.

Suppose that X and Y are jointly discrete RVs. Then if $P(X = x) > 0$,

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

$P(Y = y|X = 2)$ defines the probabilities on Y given that $X = 2$.

Definition: If $f_X(x) > 0$, then the **conditional pmf/pdf of Y given $X = x$** is

$$\boxed{f_{Y|X}(y|x) \;\equiv\; \frac{f(x,y)}{f_X(x)}.}$$

Remark: We usually just write $f(y|x)$ instead of $f_{Y|X}(y|x)$.

Remark: Of course, $f_{X|Y}(x|y) = f(x|y) = \frac{f(x,y)}{f_Y(y)}$.

Georgia Tech

Discrete Example: $f(x,y) = P(X = x, Y = y)$.

$f(x,y)$	$X = 1$	$X = 2$	$X = 3$	$f_Y(y)$
$Y = 40$	0.01	0.07	0.12	0.2
$Y = 60$	**0.29**	**0.03**	**0.48**	**0.8**
$f_X(x)$	0.3	0.1	0.6	1

Then, for example,

$$f(x|y = 60) = \frac{f(x,60)}{f_Y(60)} = \frac{f(x,60)}{0.8} = \begin{cases} \frac{29}{80} & \text{if } x = 1 \\ \frac{3}{80} & \text{if } x = 2 \\ \frac{48}{80} & \text{if } x = 3. \end{cases} \quad \square$$

Old Continuous Example:

$$
\begin{aligned}
f(x,y) &= \frac{21}{4}x^2 y, \quad \text{if } x^2 \le y \le 1. \\
f_X(x) &= \frac{21}{8}x^2(1-x^4), \quad \text{if } -1 \le x \le 1. \\
f_Y(y) &= \frac{7}{2}y^{5/2}, \quad \text{if } 0 \le y \le 1.
\end{aligned}
$$

Then the conditional pdf of Y given $X = x$ is

$$
f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{21}{4}x^2 y}{\frac{21}{8}x^2(1-x^4)} = \frac{2y}{1-x^4}, \quad \text{if } x^2 \le y \le 1.
$$

So, for example,

$$f\big(y|1/2\big) \;=\; \frac{f(\frac{1}{2},y)}{f_X(\frac{1}{2})} \;=\; \frac{\frac{21}{4}\cdot\frac{1}{4}y}{\frac{21}{8}\cdot\frac{1}{4}\cdot(1-\frac{1}{16})} \;=\; \frac{32}{15}y, \quad \text{if } \tfrac{1}{4}\le y\le 1. \quad \Box$$

Note that $2/(1-x^4)$ is a *constant* with respect to y, and we can check to see that $f(y|x)$ is a legit conditional pdf:

$$\int_{\mathbb{R}} f(y|x)\,dy \;=\; \int_{x^2}^{1}\frac{2y}{1-x^4}\,dy \;=\; 1. \quad \Box$$

Georgia Tech

Typical Problem: Given $f_X(x)$ and $f(y|x)$, find $f_Y(y)$.

Game Plan: Find $f(x,y) = f_X(x)f(y|x)$ and then $f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx$.

Example: Suppose $f_X(x) = 2x$, for $0 < x < 1$. Given $X = x$, suppose that $Y|x \sim \text{Unif}(0,x)$. Now find $f_Y(y)$.

Solution: $Y|x \sim \text{Unif}(0,x)$ implies that $f(y|x) = 1/x$, for $0 < y < x$. So,

$$\begin{aligned} f(x,y) &= f_X(x)f(y|x) \\ &= 2x \cdot \frac{1}{x} \quad \text{for } 0 < x < 1 \text{ and } 0 < y < x \\ &= 2 \quad 0 < y < x < 1 \text{ (still have funny limits).} \end{aligned}$$

Thus,

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx = \int_y^1 2\,dx = 2(1-y),\ 0 < y < 1. \quad \square$$

Lesson 3.4 — Independent Random Variables

Recall that two events are independent if $P(A \cap B) = P(A)P(B)$.

Then

$$P(A|B) \;=\; \frac{P(A \cap B)}{P(B)} \;=\; \frac{P(A)P(B)}{P(B)} \;=\; P(A).$$

And similarly, $P(B|A) = P(B)$.

Now we want to define independence for random variables, i.e., the outcome of X doesn't influence the outcome of Y (and vice versa).

Definition: X and Y are **independent** RVs if, for all x and y,

$$\boxed{f(x,y) \;=\; f_X(x)f_Y(y).}$$

Equivalent definitions:

$$F(x,y) \;=\; F_X(x)F_Y(y), \;\; \forall x,y$$

or

$$P(X \le x, Y \le y) \;=\; P(X \le x)P(Y \le y), \;\; \forall x,y.$$

If X and Y aren't independent, then they're **dependent**.

Nice, Intuitive Theorem: X and Y are independent if and only if $f(y|x) = f_Y(y)\ \forall x,y$.

Proof:

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

Similarly, X and Y independent implies $f(x|y) = f_X(x)$.

Example (discrete): $f(x,y) = P(X = x, Y = y)$.

$f(x,y)$	$X = 1$	$X = 2$	$f_Y(y)$
$Y = 2$	0.12	0.28	0.4
$Y = 3$	0.18	0.42	0.6
$f_X(x)$	0.3	0.7	1

X and Y are independent since $f(x,y) = f_X(x)f_Y(y), \forall x, y$. $\Box$

Example (continuous): Suppose $f(x, y) = 6xy^2$, $0 \le x \le 1$, $0 \le y \le 1$. After some work (which can be avoided by the next theorem), we can derive

$$\begin{aligned} f_X(x) &= 2x, \;\; \text{if } 0 \le x \le 1 \quad \text{and} \\ f_Y(y) &= 3y^2, \;\; \text{if } 0 \le y \le 1. \end{aligned}$$

X and Y are independent since $f(x, y) = f_X(x) f_Y(y)$, $\forall x, y$. $\square$

Easy way to tell if X and Y are independent....

Theorem: X and Y are independent iff $f(x, y) = a(x)b(y)$, $\forall x, y$, for some functions $a(x)$ and $b(y)$ (not necessarily pdf's).

So if $f(x, y)$ factors into separate functions of x and y, then X and Y are independent.

But if there are *funny limits*, this messes up the factorization, so in that case, X and Y will be dependent — watch out!

Example: $f(x, y) = 6xy^2$, $0 \le x \le 1$, $0 \le y \le 1$. Take

$$a(x) \;=\; 6x, \;\; 0 \le x \le 1, \text{ and } b(y) \;=\; y^2, \;\; 0 \le y \le 1.$$

Thus, X and Y are independent (as above). $\square$

Example: $f(x,y) = \frac{21}{4}x^2y, \;\; x^2 \le y \le 1.$

Funny (non-rectangular) limits make factoring into marginals impossible. Thus, X and Y are *not* independent. $\Box$

Example: $f(x,y) = \frac{c}{x+y}, 1 \le x \le 2, 1 \le y \le 3.$

Can't factor $f(x,y)$ into functions of x and y separately. Thus, X and Y are *not* independent. $\Box$

Now that we can figure out if X and Y are independent, what can we do with that knowledge?

Lesson 3.5 — Consequences of Independence

Definition/Theorem (**two-dimensional Unconscious Statistician**): Let $h(X,Y)$ be a function of the RVs X and Y. Then

$$\boxed{\mathrm{E}[h(X,Y)] \;=\; \begin{cases} \sum_x \sum_y h(x,y)f(x,y) & \text{discrete} \\ \int_{\mathbb{R}} \int_{\mathbb{R}} h(x,y)f(x,y)\,dx\,dy & \text{continuous.} \end{cases}}$$

Theorem: *Whether or not* X and Y are independent,

$$\boxed{\mathrm{E}[X+Y] \;=\; \mathrm{E}[X]+\mathrm{E}[Y].}$$

Proof (continuous case):

$$
\begin{aligned}
\mathrm{E}[X+Y] &= \int_{\mathbb{R}}\int_{\mathbb{R}} (x+y) f(x,y)\,dx\,dy \quad \text{(2-D LOTUS)} \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} x f(x,y)\,dx\,dy + \int_{\mathbb{R}}\int_{\mathbb{R}} y f(x,y)\,dx\,dy \\
&= \int_{\mathbb{R}} x \int_{\mathbb{R}} f(x,y)\,dy\,dx + \int_{\mathbb{R}} y \int_{\mathbb{R}} f(x,y)\,dx\,dy \\
&= \int_{\mathbb{R}} x f_X(x)\,dx + \int_{\mathbb{R}} y f_Y(y)\,dy \\
&= \mathrm{E}[X] + \mathrm{E}[Y]. \quad \square
\end{aligned}
$$

One can generalize this result to more than two random variables.

Corollary: If $X_1, X_2, \ldots, X_n$ are RVs, then

$$\mathrm{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{E}[X_i].$$

Proof: Induction. $\square$

Theorem: If X and Y are *independent*, then $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Proof (continuous case):

$$\begin{aligned}
\mathrm{E}[XY] &= \int_{\mathbb{R}}\int_{\mathbb{R}} xy f(x,y)\,dx\,dy \quad \text{(2-D LOTUS)} \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} xy f_X(x) f_Y(y)\,dx\,dy \;\; (X \text{ and } Y \text{ are indep}) \\
&= \left(\int_{\mathbb{R}} x f_X(x)\,dx\right)\left(\int_{\mathbb{R}} y f_Y(y)\,dy\right) \\
&= \mathrm{E}[X]\mathrm{E}[Y]. \quad \Box
\end{aligned}$$

Remark: The above theorem is *not* necessarily true if X and Y are *dependent*. See the upcoming discussion on covariance.

Georgia Tech

Theorem: If X and Y are *independent*, then

$$\boxed{\mathrm{Var}(X+Y) \;=\; \mathrm{Var}(X)+\mathrm{Var}(Y).}$$

Proof:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}[(X+Y)^2]-(\mathrm{E}[X+Y])^2 \\
&= \mathrm{E}[X^2+2XY+Y^2]-\big(\mathrm{E}[X]+\mathrm{E}[Y]\big)^2 \\
&= \mathrm{E}[X^2]+2\mathrm{E}[XY]+\mathrm{E}[Y^2]-\Big\{\big(\mathrm{E}[X]\big)^2+2\mathrm{E}[X]\mathrm{E}[Y]+\big(\mathrm{E}[Y]\big)^2\Big\} \\
&= \mathrm{E}[X^2]+2\mathrm{E}[X]\mathrm{E}[Y]+\mathrm{E}[Y^2]-\big(\mathrm{E}[X]\big)^2-2\mathrm{E}[X]\mathrm{E}[Y]-\big(\mathrm{E}[Y]\big)^2 \\
&\quad \text{(since } X \text{ and } Y \text{ are independent)} \\
&= \mathrm{E}[X^2]-\big(\mathrm{E}[X]\big)^2+\mathrm{E}[Y^2]-\big(\mathrm{E}[Y]\big)^2. \quad \Box
\end{aligned}$$

Remark: The assumption of independence really is important here. If X and Y aren't independent, then the result might not hold!

Can generalize…

Corollary: If $X_1, X_2, \ldots, X_n$ are *independent* RVs, then

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

Proof: Induction. $\square$

Corollary: If $X_1, X_2, \ldots, X_n$ are *independent* RVs, then

$$\mathrm{Var}\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X_i).$$

Lesson 3.6 — Random Samples

Definition: $X_1, X_2, \ldots, X_n$ form a **random sample** if

- X_i's are all *independent*.
- Each X_i has the same pmf/pdf $f(x)$.

Notation: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ (**"independent and identically distributed"**).

Example/Theorem: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$, with $\mathrm{E}[X_i] = \mu$, and $\mathrm{Var}(X_i) = \sigma^2$. Define the **sample mean** as $\bar{X} \equiv \sum_{i=1}^{n} X_i/n$. Then

$$\mathrm{E}[\bar{X}] \;=\; \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] \;=\; \frac{1}{n}\sum_{i=1}^{n}\mu \;=\; \mu.$$

So the mean of $\bar{X}$ is the same as the mean of X_i. $\square$

Meanwhile, how about the *variance* of the sample mean?

$$
\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \quad (X_i\text{'s indep}) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \;=\; \sigma^2/n.
\end{aligned}
$$

So the mean of $\bar{X}$ is the same as the mean of X_i, but the *variance decreases*! This makes $\bar{X}$ a great *estimator* for μ (which is usually unknown in practice); the result is referred to as the **Law of Large Numbers**. Stay tuned.

Lesson 3.7 — Conditional Expectation

The Next Few Lessons:

- Conditional expectation — definition and examples.
- “Double” expectation — a very cool theorem.
- Honors Class: First-step analysis.
- Honors Class: Random sums of random variables.
- Honors Class: The standard conditioning argument and its applications.

Consider the usual definition of expectation. (E.g., what's the average weight of a male?)

$$\mathrm{E}[Y] = \begin{cases} \sum_y y f(y) & \text{discrete} \\ \int_{\mathbb{R}} y f(y)\, dy & \text{continuous.} \end{cases}$$

Now suppose we're interested in the average weight of a 6' tall male.

$f(y|x)$ is the conditional pmf/pdf of Y given $X = x$.

Definition: The **conditional expectation** of Y given $X = x$ is

$$\boxed{\mathrm{E}[Y|X = x] \equiv \begin{cases} \sum_y y f(y|x) & \text{discrete} \\ \int_{\mathbb{R}} y f(y|x)\, dy & \text{continuous.} \end{cases}}$$

Note that $\mathrm{E}[Y|X = x]$ is a function of x.

Discrete Example:

$f(x,y)$	$X=0$	$X=3$	$f_Y(y)$
$Y=2$	0.11	**0.34**	0.45
$Y=5$	0.00	**0.05**	0.05
$Y=10$	0.29	**0.21**	0.50
$f_X(x)$	0.40	**0.60**	1

The *unconditional* expectation is

$$\mathrm{E}[Y] \;=\; \sum_y y f_Y(y) \;=\; 2(0.45)+5(0.05)+10(0.50) \;=\; 6.15.$$

But conditional on $X = 3$, we have

$$f(y|x=3) \;=\; \frac{f(3,y)}{f_X(3)} \;=\; \frac{f(3,y)}{0.60} \;=\; \begin{cases} \frac{34}{60} & \text{if } y = 2 \\ \frac{5}{60} & \text{if } y = 5 \\ \frac{21}{60} & \text{if } y = 10. \end{cases}$$

So the expectation conditional on $X = 3$ is

$$\begin{aligned} \mathrm{E}[Y|X=3] &= \sum_y y f(y|3) \\ &= 2(34/60) + 5(5/60) + 10(21/60) \\ &= 5.05. \end{aligned}$$

This compares to the unconditional expectation $\mathrm{E}[Y] = 6.15$. So information that $X = 3$ pushes the conditional expected value of Y down to 5.05. $\square$

Old Continuous Example:

$$f(x,y) = \frac{21}{4}x^2y, \quad \text{if } x^2 \le y \le 1.$$

Recall that

$$f(y|x) = \frac{2y}{1-x^4} \quad \text{if } x^2 \le y \le 1.$$

Thus,

$$\mathrm{E}[Y|x] = \int_{\mathbb{R}} y f(y|x)\,dy = \frac{2}{1-x^4}\int_{x^2}^{1} y^2\,dy = \frac{2}{3}\cdot\frac{1-x^6}{1-x^4}.$$

So, e.g., $\mathrm{E}[Y|X=0.5] = \frac{2}{3}\cdot\frac{63}{64}/\frac{15}{16} = 0.70.$ $\Box$

Lesson 3.8 — Double Expectation

Theorem (double expectation):

$$\boxed{\mathrm{E}[\mathrm{E}(Y|X)] \; = \; \mathrm{E}[Y].}$$

Remarks: Yikes, what the heck is this!?

The expected value (averaged over all X's) of the conditional expected value (of $Y|X$) is the plain old expected value (of Y).

Think of the outside expected value as the expected value of $h(X) = \mathrm{E}(Y|X)$. Then LOTUS miraculously gives us $\mathrm{E}[Y]$.

Believe it or not, sometimes it's easier to calculate $\mathrm{E}[Y]$ indirectly by using our double expectation trick.

Proof (continuous case): By the Unconscious Statistician,

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}(Y|X)] &= \int_{\mathbb{R}} \mathrm{E}(Y|x) f_X(x)\, dx \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y f(y|x)\, dy \right) f_X(x)\, dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} y f(y|x) f_X(x)\, dx\, dy \\
&= \int_{\mathbb{R}} y \int_{\mathbb{R}} f(x,y)\, dx\, dy \\
&= \int_{\mathbb{R}} y f_Y(y)\, dy \\
&= \mathrm{E}[Y]. \quad \Box
\end{aligned}
$$

Old Example: Suppose $f(x, y) = \frac{21}{4}x^2y$, if $x^2 \le y \le 1$.

Find $\mathrm{E}[Y]$ *two ways.*

By previous examples, we know that

$$
\begin{aligned}
f_X(x) &= \frac{21}{8}x^2(1 - x^4), \quad \text{if } -1 \le x \le 1 \\
f_Y(y) &= \frac{7}{2}y^{5/2}, \quad \text{if } 0 \le y \le 1 \\
\mathrm{E}[Y|x] &= \frac{2}{3} \cdot \frac{1 - x^6}{1 - x^4}.
\end{aligned}
$$

Solution #1 (old, boring way):

$$\mathrm{E}[Y] \;=\; \int_{\mathbb{R}} y f_Y(y)\,dy \;=\; \int_0^1 \frac{7}{2} y^{7/2}\,dy \;=\; \frac{7}{9}.$$

Solution #2 (new, exciting way):

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}[\mathrm{E}(Y|X)] \\
&= \int_{\mathbb{R}} \mathrm{E}(Y|x) f_X(x)\,dx \\
&= \int_{-1}^1 \left(\frac{2}{3}\cdot\frac{1-x^6}{1-x^4}\right)\left(\frac{21}{8}x^2(1-x^4)\right)dx \\
&= 7/9.
\end{aligned}$$

Notice that both answers are the same (good)! $\square$

Lesson 3.9 — Honors Class: First-Step Analysis

Example: **"First-step" method** to find the mean of $Y \sim \text{Geom}(p)$. Think of Y as the number of coin flips before H appears, where $P(\text{H}) = p$.

Furthermore, consider the first step of the coin flip process, and let $X = \text{H}$ or T denote the outcome of the first toss. Based on the result X of this first step, we have

$$\begin{aligned}
\text{E}[Y] &= \text{E}[\text{E}(Y|X)] \\
&= \sum_x \text{E}[Y|x] f_X(x) \\
&= \text{E}[Y|X = \text{T}]P(X = \text{T}) + \text{E}[Y|X = \text{H}]P(X = \text{H}) \\
&= (1 + \text{E}[Y])(1 - p) + (1)(p) \quad \text{(start from scratch if } X = \text{T).}
\end{aligned}$$

Solving, we get $\text{E}[Y] = 1/p$ (which is the correct answer)! $\square$

Georgia Tech

Example: Consider a sequence of coin flips. What is the expected number of flips Y until "HT" appears for the first time?

Clearly, $Y = A + B$, where A is the number of flips until the first "H" appears, and B is the number of subsequent flips until "T" appears for the first time after the sequence of H's begins.

For instance, the sequence TTTHHT corresponds to $Y = A + B = 4 + 2 = 6$.

In any case, it's obvious that A and B are iid Geom($p = 1/2$), so by the previous example, $\mathrm{E}[Y] = \mathrm{E}[A] + \mathrm{E}[B] = (1/p) + (1/p) = 4$. $\square$

This example didn't involve first-step analysis (besides using the expected value of a geometric RV). But the next related example will....

Example: Again consider a sequence of coin flips. What is the expected number of flips Y until "HH" appears for the first time?

For instance, the sequence TTHTTHH corresponds to $Y = 7$ tries.

Using an enhanced first-step analysis, we see that

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}[Y|\mathrm{T}]P(\mathrm{T}) + \mathrm{E}[Y|\mathrm{H}]P(\mathrm{H}) \\
&= \mathrm{E}[Y|\mathrm{T}]P(\mathrm{T}) \\
&\qquad + \big\{\mathrm{E}[Y|\mathrm{HH}]P(\mathrm{HH}|\mathrm{H}) + \mathrm{E}[Y|\mathrm{HT}]P(\mathrm{HT}|\mathrm{H})\big\}P(\mathrm{H}) \\
&= \big(1 + \mathrm{E}[Y]\big)(0.5) + \big\{(2)(0.5) + \big(2 + \mathrm{E}[Y]\big)(0.5)\big\}(0.5) \\
&\qquad \text{(since we have to start over once we see a T)} \\
&= 1.5 + 0.75\,\mathrm{E}[Y].
\end{aligned}$$

Solving, we obtain $\mathrm{E}[Y] = 6$, which is perhaps surprising given the result from the previous example. $\square$

Georgia Tech

Lesson 3.10 — Honors Class: Random Sums of Random Variables

Bonus Theorem (expectation of sum of a random number of RVs):

Suppose that $X_1, X_2, \ldots$ are independent RVs, all with the same mean.

Also suppose that N is a nonnegative, integer-valued RV that's independent of the X_i's. Then

$$\mathrm{E}\left[\sum_{i=1}^{N} X_i\right] = \mathrm{E}[N]\mathrm{E}[X_1].$$

Remark: You have to be very careful here. In particular, note that $\mathrm{E}\big[\sum_{i=1}^{N} X_i\big] \neq N\mathrm{E}[X_1]$, since the LHS is a number and the RHS is random.

Georgia Tech

Proof (cf. Ross): By double expectation,

$$
\begin{aligned}
\mathrm{E}\Big(\sum_{i=1}^{N} X_i\Big) &= \mathrm{E}\left[\mathrm{E}\Big(\sum_{i=1}^{N} X_i \,\Big|\, N\Big)\right] \\
&= \sum_{n=1}^{\infty} \mathrm{E}\Big(\sum_{i=1}^{N} X_i \,\Big|\, N=n\Big) P(N=n) \\
&= \sum_{n=1}^{\infty} \mathrm{E}\Big(\sum_{i=1}^{n} X_i \,\Big|\, N=n\Big) P(N=n) \\
&= \sum_{n=1}^{\infty} \mathrm{E}\Big(\sum_{i=1}^{n} X_i\Big) P(N=n) \quad (N \text{ and } X_i\text{'s indep}) \\
&= \sum_{n=1}^{\infty} n\mathrm{E}[X_1] P(N=n) \\
&= \mathrm{E}[X_1] \sum_{n=1}^{\infty} nP(N=n). \quad \Box
\end{aligned}
$$

Example: Suppose the number of times we roll a die is $N \sim \text{Pois}(10)$. If X_i denotes the value of the ith toss, then the expected total of all of the rolls is

$$\mathrm{E}\left(\sum_{i=1}^{N} X_i\right) = \mathrm{E}[N]\mathrm{E}[X_1] = 10(3.5) = 35. \quad \Box$$

Theorem: Under the same conditions as before,

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \mathrm{E}[N]\mathrm{Var}(X_1) + (\mathrm{E}[X_1])^2\mathrm{Var}(N).$$

Proof: See, for instance, Ross. $\Box$

Lesson 3.11 — Honors Class: Standard Conditioning Argument

Bonus Theorem/Proof (computing probabilities by conditioning):

Let A be some event, and define the RV Y as the following indicator function:

$$Y \;=\; 1_A \;\equiv\; \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\mathrm{E}[Y] \;=\; \sum_y y f_Y(y) \;=\; P(Y=1) \;=\; P(A).$$

Similarly, for any RV X, we have

$$\mathrm{E}[Y|X=x] \;=\; \sum_y y f(y|x) \;=\; P(Y=1|X=x) \;=\; P(A|X=x).$$

These results suggest an alternative way of calculating $P(A)$....

Theorem: If X is a continuous RV (similar result if X is discrete), then

$$\boxed{P(A) \;=\; \int_{\mathbb{R}} P(A|X=x) f_X(x)\,dx.}$$

Proof:

$$\begin{aligned} P(A) &= \mathrm{E}[Y] \quad \text{(where we take } Y = 1_A\text{)} \\ &= \mathrm{E}[\mathrm{E}(Y|X)] \quad \text{(double expectation)} \\ &= \int_{\mathbb{R}} \mathrm{E}[Y|x] f_X(x)\,dx \quad \text{(LOTUS)} \\ &= \int_{\mathbb{R}} P(A|X=x) f_X(x)\,dx \quad \text{(since } Y = 1_A\text{)}. \quad \Box \end{aligned}$$

Remark: We call this the **"standard conditioning argument."** Yes, it looks complicated. But sometimes you need to take a step backward to go two steps forward!

Example/Theorem: If X and Y are independent continuous RVs, with pdf $f_X(\cdot)$ and cdf $F_Y(\cdot)$, respectively. Then

$$\boxed{P(Y \leq X) \ = \ \int_{\mathbb{R}} F_Y(x) f_X(x)\, dx.}$$

Proof: (Actually, there are many proofs.) Let the event $A = \{Y \leq X\}$. Then

$$\begin{aligned} P(Y \leq X) \ &= \ \int_{\mathbb{R}} P(Y \leq X | X = x) f_X(x)\, dx \\ &= \ \int_{\mathbb{R}} P(Y \leq x | X = x) f_X(x)\, dx \\ &= \ \int_{\mathbb{R}} P(Y \leq x) f_X(x)\, dx \quad (X, Y \text{ are independent}). \qquad \square \end{aligned}$$

Example: If $X \sim \text{Exp}(\alpha)$ and $Y \sim \text{Exp}(\beta)$ are independent RVs, then

$$\begin{aligned} P(Y \le X) &= \int_{\mathbb{R}} F_Y(x) f_X(x)\, dx \\ &= \int_0^\infty (1 - e^{-\beta x}) \alpha e^{-\alpha x}\, dx \\ &= \frac{\beta}{\alpha + \beta}. \quad \square \end{aligned}$$

Remark: Think of X as the time until the next male driver shows up at a parking lot (at rate α / hour) and Y as the time for the next female driver (at rate β / hour). Then $P(Y \le X) = \beta/(\alpha + \beta)$ is the intuitively reasonable probability that the next driver to arrive will be female. $\square$

Example/Theorem: Suppose X and Y are independent continuous RVs, with pdf $f_X(\cdot)$ and cdf $F_Y(\cdot)$, respectively. Define the sum $Z = X + Y$. Then

$$\boxed{P(Z \le z) \;=\; \int_{\mathbb{R}} F_Y(z-x) f_X(x)\,dx.}$$

As expression such as the above for $P(Z \le z)$ is often called a *convolution*.

Proof:

$$\begin{aligned} P(Z \le z) &= \int_{\mathbb{R}} P(X+Y \le z | X = x) f_X(x)\,dx \\ &= \int_{\mathbb{R}} P(Y \le z - x | X = x) f_X(x)\,dx \\ &= \int_{\mathbb{R}} P(Y \le z - x) f_X(x)\,dx \quad (X, Y \text{ are indep}). \qquad \Box \end{aligned}$$

Example: Suppose $X, Y \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$, and let $Z = X + Y$. Then

$$\begin{aligned}
P(Z \le z) &= \int_{\mathbb{R}} F_Y(z-x) f_X(x)\, dx \\
&= \int_0^{\boldsymbol{z}} (1 - e^{-\lambda(z-x)})\lambda e^{-\lambda x}\, dx \\
&\qquad \text{(must have } x \ge 0 \textit{ and } z - x \ge 0) \\
&= 1 - e^{-\lambda z} - \lambda z e^{-\lambda z}, \quad \text{if } z \ge 0.
\end{aligned}$$

Thus, the pdf of Z is

$$\frac{d}{dz} P(Z \le z) = \lambda^2 z e^{-\lambda z}, \quad z \ge 0.$$

This turns out to mean that $Z \sim$ **Gamma**$(2, \lambda)$, aka **Erlang**$_2(\lambda)$. $\square$

You can do the similar kinds of convolutions with discrete RVs. We state the following result without proof (which is straightforward).

Example/Theorem: Suppose X and Y are two independent integer-valued RVs with pmf's $f_X(x)$ and $f_Y(y)$. Then the pmf of $Z = X + Y$ is

$$\boxed{f_Z(z) \;=\; P(Z = z) \;=\; \sum_{x=-\infty}^{\infty} f_X(x) f_Y(z - x).}$$

Example Suppose X and Y are iid Bern(p). Then the pmf of $Z = X + Y$ is

$$\begin{aligned}
f_Z(z) &= \sum_{x=-\infty}^{\infty} f_X(x) f_Y(z-x) \\
&= f_X(0)f_Y(z) + f_X(1)f_Y(z-1) \quad (X \text{ can only} = 0 \text{ or } 1) \\
&= f_X(0)f_Y(z)1_{\{0,1\}}(z) + f_X(1)f_Y(z-1)1_{\{1,2\}}(z) \\
&\qquad (1_{\{\cdot\}}(z) \text{ functions indicate nonzero } f_Y(\cdot)\text{'s}) \\
&= p^0 q^{1-0} p^z q^{1-z} 1_{\{0,1\}}(z) + p^1 q^{1-1} p^{z-1} q^{2-z} 1_{\{1,2\}}(z) \\
&= p^z q^{2-z} \left[1_{\{0,1\}}(z) + 1_{\{1,2\}}(z)\right] \\
&= \binom{2}{z} p^z q^{2-z}, \quad z = 0, 1, 2.
\end{aligned}$$

Thus, $Z \sim \text{Bin}(2, p)$, a fond blast from the past! $\square$

Georgia Tech

Lesson 3.12 — Covariance and Correlation

In the next few lessons we'll cover:

- Basic Concepts of Covariance and Correlation
- Causation
- A Couple of Worked Examples
- Some Useful Theorems

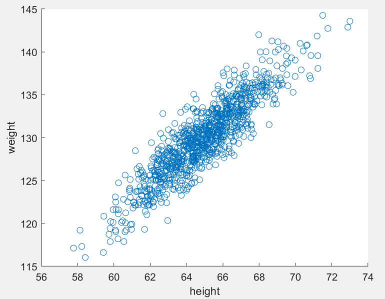
Covariance and correlation are measures used to define the degree of *association* between X and Y if they don't happen to be independent.

Definition: The **covariance** between X and Y is

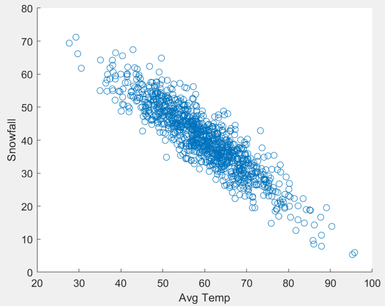
$$\mathrm{Cov}(X,Y) \equiv \sigma_{XY} \equiv \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])].$$

Remark: $\mathrm{Cov}(X,X) = \mathrm{E}[(X-\mathrm{E}[X])^2] = \mathrm{Var}(X)$.

Remark: If X and Y have positive covariance, then X and Y move "in the same direction." Think height and weight.

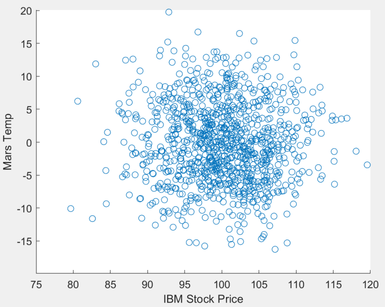


If X and Y have negative covariance, then X and Y move "in opposite directions." Think snowfall and temperature.



If X and Y are *independent*, then of course they have no association with each other. In fact, we'll prove below that independence implies that the covariance is 0 (but not the other way around).

Example: IBM stock price vs. temperature on Mars are independent — at least that's what they want you to believe!



Theorem (easier way to calculate covariance):

$$\text{Cov}(X,Y) \;=\; \text{E}[XY]-\text{E}[X]\text{E}[Y].$$

Proof:

$$\begin{aligned}\text{Cov}(X,Y) &= \text{E}[(X-\text{E}[X])(Y-\text{E}[Y])] \\ &= \text{E}\Big[XY - X\text{E}[Y] - Y\text{E}[X] + \text{E}[X]\text{E}[Y]\Big] \\ &= \text{E}[XY]-\text{E}[X]\text{E}[Y]-\text{E}[Y]\text{E}[X]+\text{E}[X]\text{E}[Y]. \quad \Box\end{aligned}$$

Theorem: X and Y independent implies $\text{Cov}(X,Y)=0$.

Proof: By a previous theorem, X and Y independent implies $\text{E}[XY]=\text{E}[X]\text{E}[Y]$. Then

$$\text{Cov}(X,Y) \;=\; \text{E}[XY]-\text{E}[X]\text{E}[Y] \;=\; \text{E}[X]\text{E}[Y]-\text{E}[X]\text{E}[Y]. \quad \Box$$

Danger Will Robinson! $\text{Cov}(X, Y) = 0$ **does not imply** that X and Y are independent!!

Example: Suppose $X \sim \text{Unif}(-1, 1)$ and $Y = X^2$ (so X and Y are clearly *dependent*).

But

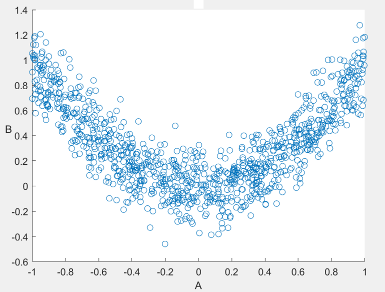
$$\text{E}[X] \;=\; \int_{-1}^{1} x \cdot \frac{1}{2}\, dx \;=\; 0 \;\text{ and}$$

$$\text{E}[XY] \;=\; \text{E}[X^3] \;=\; \int_{-1}^{1} x^3 \cdot \frac{1}{2}\, dx \;=\; 0,$$

so

$$\text{Cov}(X, Y) \;=\; \text{E}[XY] - \text{E}[X]\text{E}[Y] \;=\; 0. \quad ※$$

In fact, here's a graphical illustration of this zero-correlation dependence phenomenon, where we've actually added some normal noise to Y to make it look prettier.



Definition: The **correlation** between X and Y is

$$\rho \;=\; \mathrm{Corr}(X,Y) \;\equiv\; \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \;=\; \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Remark: Covariance has "square" units; correlation is unitless.

Corollary: X, Y independent implies $\rho = 0$.

Theorem: It can be shown that $-1 \le \rho \le 1$.

$\rho \approx 1$ is "high" correlation.
$\rho \approx 0$ is "low" correlation.
$\rho \approx -1$ is "high" negative correlation.

Example: Height is *highly* correlated with weight.
Temperature on Mars has *low* correlation with IBM stock price.

Lesson 3.13 — Correlation and Causation

NOTE! Correlation does not necessarily imply causality! This is a very common pitfall in many areas of data analysis and public discourse.

Example in which correlation does imply causality: Height and weight are positively correlated, and larger height does indeed tend to cause greater weight. □

Example in which correlation does not imply causality: Temperature and lemonade sales have positive corr, and temp has causal influence on lemonade sales. Similarly, temp and overheating cars are positively correlated with a causal relationship. It's also likely that lemonade sales and overheating cars are positively correlated, but there's no causal relationship there. □

Example of a zero correlation relationship with causality! We've seen that it's possible for two dependent RVs to be uncorrelated. □

To prove that X causes Y, one must establish that:

- X occurred before Y;
- The relationship between X and Y is not completely due to random chance; and
- Nothing else accounts for the relationship (which is violated in the lemonade sales / overheating cars example above).

These items can be often be established via mathematical analysis, statistical analysis of appropriate data, or consultation with appropriate experts.

The three examples above seem to give conflicting guidance with respect to the relationship between correlation and causality. How can we interpret these findings in a meaningful way? Here are the takeaways:

- If the correlation between X and Y is (significantly) nonzero, there is some type of relationship between the two items, which may or may not be causal; but this should raise our curiosity.
- If the correlation between X and Y is 0, we are not quite out of the woods with respect to dependence and causality. In order to definitively rule out a relationship between X and Y, it is always highly recommended protocol to, at the very least,
    - Plot data from X and Y against each other to see if there is a nonlinear relationship, as in the uncorrelated-yet-dependent example.
    - Consult with appropriate experts.

Lesson 3.14 — A Couple of Worked Correlation Examples

Discrete Example: Suppose X is the GPA of a UGA student, and Y is their IQ. Here's the joint pmf.

$f(x,y)$	$X=2$	$X=3$	$X=4$	$f_Y(y)$
$Y=40$	0.0	0.2	0.2	0.4
$Y=50$	0.1	0.1	0.0	0.2
$Y=60$	0.4	0.0	0.0	0.4
$f_X(x)$	0.5	0.3	0.2	1

We'll spare the details, but here are the relevant calculations…

$$\begin{aligned}
\mathrm{E}[X] &= \sum_x x f_X(x) \;=\; 2.7, \\
\mathrm{E}[X^2] &= \sum_x x^2 f_X(x) \;=\; 7.9, \quad \text{and} \\
\mathrm{Var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \;=\; 0.61.
\end{aligned}$$

Similarly, $\mathrm{E}[Y] = 50$, $\mathrm{E}[Y^2] = 2580$, and $\mathrm{Var}(Y) = 80$. Finally,

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_x \sum_y xy f(x,y) \\
&= 2(40)(0.0) + 3(40)(0.2) + \cdots + 4(60)(0.0) \;=\; 129, \\
\mathrm{Cov}(X,Y) &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \;=\; -6.0, \quad \text{and} \\
\rho &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \;=\; -0.859. \quad \Box
\end{aligned}$$

Continuous Example: Suppose $f(x,y) = 10x^2y,\ 0 \le y \le x \le 1$.

$$\begin{aligned}
f_X(x) &= \int_0^x 10x^2y\,dy \;=\; 5x^4,\;\; 0 \le x \le 1, \\
\mathrm{E}[X] &= \int_0^1 5x^5\,dx \;=\; 5/6, \\
\mathrm{E}[X^2] &= \int_0^1 5x^6\,dx \;=\; 5/7, \\
\mathrm{Var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \;=\; 0.01984.
\end{aligned}$$

Similarly,

$$
\begin{aligned}
f_Y(y) &= \int_y^1 10x^2y\,dx \;=\; \frac{10}{3}y(1-y^3), \;\; 0 \le y \le 1, \\
\mathrm{E}[Y] &= 5/9, \quad \mathrm{Var}(Y) \;=\; 0.04850, \\
\mathrm{E}[XY] &= \int_0^1 \int_0^x 10x^3y^2\,dy\,dx \;=\; 10/21, \\
\mathrm{Cov}(X,Y) &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \;=\; 0.01323, \\
\rho &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \;=\; 0.4265. \quad \Box
\end{aligned}
$$

Lesson 3.15 — Some Useful Covariance / Correlation Theorems

Theorem: $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$, *whether or not* X and Y are independent.

Remark: If X, Y are independent, the covariance term goes away.

Proof: By the work we did on a previous proof,

$$\begin{aligned} \text{Var}(X+Y) &= \text{E}[X^2] - (\text{E}[X])^2 + \text{E}[Y^2] - (\text{E}[Y])^2 \\ &\quad +2(\text{E}[XY] - \text{E}[X]\text{E}[Y]) \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y). \quad \square \end{aligned}$$

Theorem:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \;=\; \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum\sum\nolimits_{i<j} \mathrm{Cov}(X_i, X_j).$$

Proof: Induction.

Corollary: If all X_i's are *independent*, then

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \;=\; \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Theorem: $\text{Cov}(aX, bY + c) = ab\,\text{Cov}(X, Y)$.

Proof:

$$\begin{aligned}
\text{Cov}(aX, bY + c) &= \text{E}[aX \cdot (bY + c)] - \text{E}[aX]\text{E}[bY + c] \\
&= \text{E}[abXY] + \text{E}[acX] - \text{E}[aX]\text{E}[bY] - \text{E}[aX]\text{E}[c] \\
&= ab\,\text{E}[XY] - ab\,\text{E}[X]\text{E}[Y] + ac\text{E}[X] - ac\text{E}[X] \\
&= ab\,\text{Cov}(X, Y). \quad \square
\end{aligned}$$

Theorem:

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i + c\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum \sum_{i<j} a_i a_j \text{Cov}(X_i, X_j).$$

Proof: Put the above two results together. $\square$

Example: $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$.

Example: Suppose $\text{Var}(X) = \text{Var}(Y) = \text{Var}(Z) = 10$, $\text{Cov}(X, Y) = 3$, $\text{Cov}(X, Z) = -2$, and $\text{Cov}(Y, Z) = 0$. Then

$$\begin{aligned} &\text{Var}(X - 2Y + 3Z) \\ &\quad= \text{Var}(X) + 4\text{Var}(Y) + 9\text{Var}(Z) \\ &\qquad\quad -4\text{Cov}(X, Y) + 6\text{Cov}(X, Z) - 12\text{Cov}(Y, Z) \\ &\quad= 14(10) - 4(3) + 6(-2) - 12(0) \;=\; 116. \quad \square \end{aligned}$$

Georgia Tech

Lesson 3.16 — Moment Generating Functions, Revisited

Old Definition: $M_X(t) \equiv \mathrm{E}[e^{tX}]$ is the **moment generating function** (mgf) of the RV X.

Old Example: If $X \sim \mathrm{Bern}(p)$, then

$$M_X(t) \;=\; \mathrm{E}[e^{tX}] \;=\; \sum_x e^{tx} f(x) \;=\; e^{t\cdot 1}p + e^{t\cdot 0}q \;=\; pe^t + q. \quad \Box$$

Old Example: If $X \sim \mathrm{Exp}(\lambda)$, then

$$M_X(t) \;=\; \mathrm{E}[e^{tX}] \;=\; \int_{\mathbb{R}} e^{tx} f(x)\,dx \;=\; \frac{\lambda}{\lambda - t} \quad \text{if } \lambda > t. \quad \Box$$

Old Theorem (why it's called the mgf): Under certain technical conditions,

$$\mathrm{E}[X^k] \;=\; \frac{d^k}{dt^k} M_X(t)\Big|_{t=0}, \quad k = 1, 2, \ldots .$$

New Theorem (mgf of the sum of independent RVs): Suppose $X_1, \ldots, X_n$ are *independent*. Let $Y = \sum_{i=1}^n X_i$. Then

$$M_Y(t) \;=\; \prod_{i=1}^n M_{X_i}(t).$$

Proof:

$$\begin{aligned}
M_Y(t) &= \mathrm{E}[e^{tY}] \\
&= \mathrm{E}[e^{t\sum X_i}] \\
&= \mathrm{E}\Big[\prod_{i=1}^n e^{tX_i}\Big] \\
&= \prod_{i=1}^n \mathrm{E}[e^{tX_i}] \quad (X_i\text{'s independent}) \\
&= \prod_{i=1}^n M_{X_i}(t). \quad \Box
\end{aligned}$$

Corollary: If $X_1, \ldots, X_n$ are iid and $Y = \sum_{i=1}^n X_i$, then

$$M_Y(t) \;=\; [M_{X_1}(t)]^n.$$

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$. Then by a previous example,

$$M_Y(t) \;=\; [M_{X_1}(t)]^n \;=\; (pe^t + q)^n.$$

So what use is a result like this? We can use results such as this with our old friend. . . .

Old Theorem (identifying distributions): *In this class,* each distribution has a unique mgf.

Example/Theorem: The sum Y of n iid Bern(p) RVs is the same as a Bin(n, p) RV.

By the previous example and uniqueness, all we need to show is that the mgf of $Z \sim \mathrm{Bin}(n, p)$ matches $M_Y(t) = (pe^t + q)^n$. To this end, we have

$$\begin{aligned}
M_Z(t) &= \mathrm{E}[e^{tZ}] \\
&= \sum_z e^{tz} P(Z = z) \\
&= \sum_{z=0}^{n} e^{tz} \binom{n}{z} p^z q^{n-z} \\
&= \sum_{z=0}^{n} \binom{n}{z} (pe^t)^z q^{n-z} \\
&= (pe^t + q)^n \quad \text{(by the Binomial Theorem).} \quad \Box
\end{aligned}$$

Example: You can identify a distribution by its mgf.

$$M_X(t) \;=\; \left(\frac{3}{4}e^t + \frac{1}{4}\right)^{15}$$

implies that $X \sim \text{Bin}(15, 0.75)$. $\square$

Old Theorem (mgf of a linear function of X): Suppose X has mgf $M_X(t)$ and let $Y = aX + b$. Then $M_Y(t) = e^{tb} M_X(at)$.

Example:

$$M_Y(t) \;=\; e^{-2t}\left(\frac{3}{4}e^{3t} + \frac{1}{4}\right)^{15} \;=\; e^{bt}(pe^{at} + q)^n \;=\; e^{bt} M_X(at),$$

which implies that Y has the same distribution as $3X - 2$, where $X \sim \text{Bin}(15, 0.75)$. $\square$

Theorem (Additive property of Binomials): If $X_1, \ldots, X_k$ are independent, with $X_i \sim \text{Bin}(n_i, p)$ (where p is the same for all X_i's), then

$$Y \equiv \sum_{i=1}^{k} X_i \sim \text{Bin}\Big(\sum_{i=1}^{k} n_i, p\Big).$$

Proof:

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{k} M_{X_i}(t) \quad \text{(mgf of independent sum)} \\ &= \prod_{i=1}^{k} (pe^t + q)^{n_i} \quad (\text{Binomial}(n_i, p) \text{ mgf}) \\ &= (pe^t + q)^{\sum_{i=1}^{k} n_i}. \end{aligned}$$

This is the mgf of the $\text{Bin}(\sum_{i=1}^{k} n_i, p)$, so we're done. $\square$

Lesson 3.17 — Honors Bivariate Functions of Random Variables

In earlier work, we looked at...

- Functions of a single variable, e.g., what is the expected value of $h(X)$? (LOTUS, from Module 2)
- What is the distribution of $h(X)$? (functions of RVs, from Module 2)
- And sometimes even functions of two (or more) variables. For example, if the X_i's are independent, what's $\text{Var}(\sum_{i=1}^{n} X_i)$? (earlier in Module 3)
- Use a standard conditioning argument to get the distribution of $X + Y$. (earlier in Module 3)

Goal: Now let's give a *general result* on the distribution of functions of *two* random variables, the proof of which is beyond the scope of our class.

Honors Theorem: Suppose X and Y are continuous RVs with joint pdf $f(x,y)$, and $V = h_1(X,Y)$ and $W = h_2(X,Y)$ are functions of X and Y, and

$$X \;=\; k_1(V,W) \quad \text{and} \quad Y \;=\; k_2(V,W),$$

for suitably chosen inverse functions k_1 and k_2.

Then the joint pdf of V and W is

$$g(v,w) \;=\; f\big(k_1(v,w),\, k_2(v,w)\big)\,|J(v,w)|,$$

where $|J|$ is the absolute value of the *Jacobian* (determinant) of the transformation, i.e.,

$$J(v,w) \;=\; \begin{vmatrix} \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \end{vmatrix} \;=\; \frac{\partial x}{\partial v}\,\frac{\partial y}{\partial w} - \frac{\partial y}{\partial v}\,\frac{\partial x}{\partial w}.$$

Corollary: If X and Y are *independent*, then the joint pdf of V and W is

$$g(v,w) \;=\; f_X\big(k_1(v,w)\big)\, f_Y\big(k_2(v,w)\big)\, |J(v,w)|.$$

Remark: These results generalize the 1-D method from Module 2.

You can use this method to find all sorts of cool stuff, e.g., the distribution of $X+Y$, X/Y, etc., as well as the joint pdf of any functions of X and Y.

Remark: Although the notation is nasty, the application isn't really so bad.

Example: Suppose X and Y are iid Exp(λ). Find the pdf of $X+Y$.

We'll set $V = X + Y$ along with the dummy RV $W = X$.

This yields

$$X \;=\; W \;=\; k_1(V,W) \quad \text{and} \quad Y \;=\; V - W \;=\; k_2(V,W).$$

To get the Jacobian term, we calculate

$$\frac{\partial x}{\partial v} = 0, \quad \frac{\partial x}{\partial w} = 1, \quad \frac{\partial y}{\partial v} = 1, \quad \text{and} \quad \frac{\partial y}{\partial w} = -1,$$

so that

$$|J| \;=\; \left| \frac{\partial x}{\partial v}\frac{\partial y}{\partial w} - \frac{\partial y}{\partial v}\frac{\partial x}{\partial w} \right| \;=\; |0(-1) - 1(1)| \;=\; 1.$$

This implies that the joint pdf of V and W is

$$\begin{aligned} g(v,w) &= f\big(k_1(v,w),\, k_2(v,w)\big)\, |J(v,w)| \\ &= f(w, v-w)\cdot 1 \\ &= f_X(w) f_Y(v-w) \quad (X \text{ and } Y \text{ independent}) \\ &= \lambda e^{-\lambda w}\cdot \lambda e^{-\lambda(v-w)}, \quad \text{for } w>0 \text{ and } v-w>0 \\ &= \lambda^2 e^{-\lambda v}, \quad \text{for } 0<w<v. \end{aligned}$$

And, finally, we obtain the desired pdf of the sum V (after carefully noting the region of integration),

$$g_V(v) \;=\; \int_{\mathbb{R}} g(v,w)\,dw \;=\; \int_0^v \lambda^2 e^{-\lambda v}\,dw \;=\; \lambda^2 v e^{-\lambda v}, \quad \text{for } v>0.$$

This is the Gamma$(2,\lambda)$ pdf, which matches our answer from earlier in the current module. $\square$

Honors Example: Suppose X and Y are iid Unif(0,1). Find the joint pdf of $V = X + Y$ and $W = X/Y$.

After some algebra, we obtain

$$X \;=\; \frac{VW}{W+1} \;=\; k_1(V,W) \quad \text{and} \quad Y \;=\; \frac{V}{W+1} \;=\; k_2(V,W).$$

After more algebra, we calculate

$$\frac{\partial x}{\partial v} = \frac{w}{w+1}, \quad \frac{\partial x}{\partial w} = \frac{v}{(w+1)^2}, \quad \frac{\partial y}{\partial v} = \frac{1}{w+1}, \quad \frac{\partial y}{\partial w} = \frac{-v}{(w+1)^2},$$

so that after still more algebra,

$$|J| \;=\; \left| \frac{\partial x}{\partial v}\frac{\partial y}{\partial w} - \frac{\partial y}{\partial v}\frac{\partial x}{\partial w} \right| \;=\; \frac{v}{(w+1)^2}.$$

This implies that the joint pdf of V and W is

$$\begin{aligned}
g(v,w) &= f\big(k_1(v,w),\, k_2(v,w)\big)\,|J(v,w)| \\
&= f\left(\frac{vw}{w+1}, \frac{v}{w+1}\right) \cdot \frac{v}{(w+1)^2} \\
&= f_X\left(\frac{vw}{w+1}\right) f_Y\left(\frac{v}{w+1}\right) \frac{v}{(w+1)^2} \quad (X \text{ and } Y \text{ indep}) \\
&= 1 \cdot 1 \cdot \frac{v}{(w+1)^2}, \text{ for } 0 < x, y < 1 \text{ (since } X, Y \sim \text{Unif(0,1))} \\
&= \frac{v}{(w+1)^2}, \text{ for } 0 < x = \tfrac{vw}{w+1} < 1 \text{ and } 0 < y = \tfrac{v}{w+1} < 1. \\
&= \frac{v}{(w+1)^2}, \; 0 < v < 1 + \min\{\tfrac{1}{w}, w\} \text{ and } w > 0 \;\text{ (after algebra).}
\end{aligned}$$

Note that you have to be careful about the limits of v and w, but this thing really does double integrate to 1! $\square$

We can also get the marginal pdf's. First of all, for the ratio of the uniforms, we get

$$\begin{aligned} g_W(w) &= \int_{\mathbb{R}} g(v,w)\,dv \\ &= \int_0^{1+\min\{1/w,w\}} \frac{v}{(w+1)^2}\,dv \\ &= \frac{\big(1+\min\{1/w,w\}\big)^2}{2(w+1)^2} \\ &= \begin{cases} \frac{1}{2}, & \text{if } w \le 1 \\ \frac{1}{2w^2}, & \text{if } w > 1, \end{cases} \end{aligned}$$

which is a little weird-looking and unexpected to me (it's flat for $w \le 1$, and then decreases to 0 pretty quickly for $w > 1$). $\square$

For the pdf of the sum of the uniforms, we have to calculate $g_V(v) = \int_{\mathbb{R}} g(v,w)\,dw$. But first we need to deal with some inequality constraints so that we can integrate over the proper region, namely,

$$0 \le v \le 1 + \min\{1/w, w\}, \quad 0 \le v \le 2, \quad \text{and} \quad w \ge 0.$$

With a little thought, we see that if $0 \le v \le 1$, then there is no constraint on w except for it being positive. On the other hand, if $1 < v \le 2$, then you can show (it takes a little work) that $v - 1 \le w \le \frac{1}{v-1}$. Thus, we have

$$\begin{aligned} g_V(v) &= \begin{cases} \int_0^\infty g(v,w)\,dw, & \text{if } 0 \le v \le 1 \\ \int_{v-1}^{1/(v-1)} g(v,w)\,dw, & \text{if } 1 < v \le 2 \end{cases} \\ &= \begin{cases} v, & \text{if } 0 \le v \le 1 \\ 2-v, & \text{if } 1 < v \le 2 \end{cases} \quad \text{(after algebra).} \end{aligned}$$

This is a **Triangle(0,1,2)** pdf. Can you see why? Is there an intuitive explanation for this pdf? $\square$

And Now a Word From Our Sponsor...

We are finally done with the most-difficult module of the course. Congratulations and Felicitations!!!

Things will get easier from now on! Happy days are here again! ☺☺☺☺