Chapter 6-7

Joint Distributions and Expectations

Yingbo Li

Clemson University

MATH 4000 / 6000

Recap

Distribution (pdf) of a function of a continuous rv X: if function $g(x)$ is

1. monotonic,
2. differentiable,

on the range of X, then the rv defined by $Y = g(X)$ has pdf

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

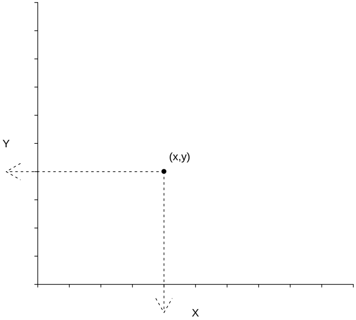
- When conditions are not satisfied,
  (1) Identify the range of Y.
  (2) Find cdf $F_Y(y)$ as a function of $F_X(\cdot)$, for any y in the range of Y.
  (3) Take derivative to get $f_Y(y)$.
  (4) Note that for any y not in the range of Y, $f_Y(y) = 0$.

Joint cdf

Definition

We have a pair of rv's (either discrete or continuous) X and Y. The joint cumulative probability distribution function of X and Y is defined by

$$\begin{aligned} F_{X,Y}(x,y) &= P[X \leq x, Y \leq y] \\ &= P[(X,Y) \text{ lies south-west of the point } (x,y)] \end{aligned}$$



Properties of joint cdf

- For one rv: marginal cdf

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

$$F_Y(y) = F_{X,Y}(\infty, y)$$

- Joint probabilities

$$P(X > x, Y > y) = 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y)$$

Question

Use joint cdf $F(x, y)$ to represent $P(x_1 < X \leq x_2, y_1 < Y \leq y_2)$.

(a) $F(x_2, y_2) + F(x_1, y_1) - F(x_1, y_2) - F(x_2, y_1)$

(b) $F(x_2, y_2) - F(x_1, y_1) - F(x_1, y_2) - F(x_2, y_1)$

(c) $F(x_2, y_2) - F(x_1, y_1)$

(d) none of the above

Example: two discrete random variables

Draw two socks at random, without replacement, from a drawer full of twelve colored socks:

6 black, 4 white, 2 purple

Let B be the number of Black socks, W the number of White socks drawn.

Then the distributions of B and W are given by:

	0	1	2
P(B=k)	$\frac{6}{12}\cdot\frac{5}{11}=\frac{15}{66}$	$2\cdot\frac{6}{12}\cdot\frac{6}{11}=\frac{36}{66}$	$\frac{6}{12}\cdot\frac{5}{11}=\frac{15}{66}$
P(W=k)	$\frac{8}{12}\cdot\frac{7}{11}=\frac{28}{66}$	$2\cdot\frac{4}{12}\cdot\frac{8}{11}=\frac{32}{66}$	$\frac{4}{12}\cdot\frac{3}{11}=\frac{6}{66}$

Note - $P(B=k)=\frac{\binom{6}{k}\binom{6}{2-k}}{\binom{12}{2}}$ and $P(W=k)=\frac{\binom{4}{k}\binom{8}{2-k}}{\binom{12}{2}}$

Draw two socks at random, without replacement, from a drawer full of twelve colored socks: 6 black, 4 white, 2 purple. Let B be the number of Black socks, W the number of White socks drawn.

The *joint distribution* is given by: $p_{B,W}(b,w) = P(B=b, W=w)$

		W = 0	W = 1	W = 2	
B	0	$\frac{1}{66}$	$\frac{8}{66}$	$\frac{6}{66}$	$\frac{15}{66}$
	1	$\frac{12}{66}$	$\frac{24}{66}$	0	$\frac{36}{66}$
	2	$\frac{15}{66}$	0	0	$\frac{15}{66}$
		$\frac{28}{66}$	$\frac{32}{66}$	$\frac{6}{66}$	

$$P(B=b, W=w) = \begin{cases} 1/66 & \text{If b=0,w=0} \\ 8/66 & \text{If b=0,w=1} \\ 6/66 & \text{If b=0,w=2} \\ 12/66 & \text{If b=1,w=0} \\ 24/66 & \text{If b=1,w=1} \\ 15/66 & \text{If b=2,w=0} \end{cases}$$

$$P(B=b, W=w) = \frac{\binom{6}{b}\binom{4}{w}\binom{2}{2-b-w}}{\binom{12}{2}}, \text{ for } 0 \le b, w \le 2 \text{ and } b+w \le 2$$

Marginal Distributions

Note that the column and row sums are the distributions of B and W respectively.

$$P(B=b) = P(B=b, W=0) + P(B=b, W=1) + P(B=b, W=2)$$
$$P(W=w) = P(B=0, W=w) + P(B=1, W=w) + P(B=2, W=w)$$

These are the marginal distributions of B and W. In general,

$$P(X=x) = \sum_y P(X=x, Y=y) = \sum_y P(X=x|Y=y)P(Y=y)$$

Joint distribution of two continuous random variables

Definition

Random variables X and Y are jointly continuous if there exists a function $f(x, y)$ such that

1. *Non-negative $f(x, y) \geq 0$, for any $x, y \in \mathbb{R}$, and*
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$

$f_{X,Y}(x, y)$ is called the joint probability density function of X and Y.

- For any set $C \subset \mathbb{R}^2$,

$$P[(X, Y) \in C] = \iint_{(x,y)\in C} f(x, y)\, dx\, dy$$

- Connection between joint pdf and joint cdf

$$F(a, b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y)\, dx\, dy$$

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

Marginal pdfs

Marginal probability density functions are defined in terms of "integrating out" one of the random variables.

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

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

Question

Which of the following can be obtained if the joint pdf $f_{X,Y}(x,y)$ is known?

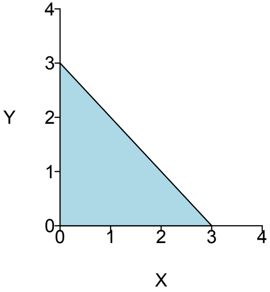
(a) Joint cdf $F_{X,Y}(x,y)$

(b) Marginal cdf's $F_X(x), F_Y(y)$.

(c) Expected values $E[X], E[Y]$.

(d) *all above*

Example 1: let X and Y be drawn uniformly from the triangle below. Find the joint pdf $f_{X,Y}(x,y)$.



Since the joint density is constant, then

$$f(x,y) = \begin{cases} c & \text{for } x \geq 0, y \geq 0 \text{ and } x+y \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Because

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy \\ &= \iint_{x \geq 0, y \geq 0, x+y \leq 3} c\ dx\ dy \\ &= c \times \text{area of the triangle} = c \times \frac{3 \times 3}{2} \end{aligned}$$

Therefore, $c = \frac{2}{9}$.

Recap

Joint cdf of two rv's X and Y:

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

- Probability of (X,Y) in a rectangle

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2)$$

$$= F(x_2,y_2) + F(x_1,y_1) - F(x_1,y_2) - F(x_2,y_1)$$

- Marginal cdf's

$$F_X(x) = F_{X,Y}(x,\infty), \quad F_Y(y) = F_{X,Y}(\infty,y)$$

Joint distribution of two discrete rv's

- Joint pmf

$$p_{X,Y}(x,y) = P(X = x, Y = y)$$

- Marginal pmf's

$$p_X(x) = \sum_{y:p(x,y)>0} p_{X,Y}(x,y), \quad p_Y(y) = \sum_{x:p(x,y)>0} p_{X,Y}(x,y)$$

Joint distribution of two continuous rv's

- Joint pdf
  - Non-negative $f_{X,Y}(x,y) \geq 0$, for any $x, y \in \mathbb{R}$
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy = 1$
  - For any set $C \subset \mathbb{R}^2$,

$$P[(X,Y) \in C] = \iint_{(x,y)\in C} f_{X,Y}(x,y)\ dx\ dy$$

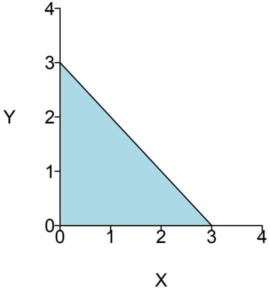
- Marginal pdf's

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx$$

Let X and Y have the following joint pdf

$$f(x,y) = \begin{cases} \frac{2}{9} & \text{for } x \geq 0, y \geq 0 \text{ and } x+y \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the marginal pdf $f_X(x)$.



For $x \in [0,3]$,

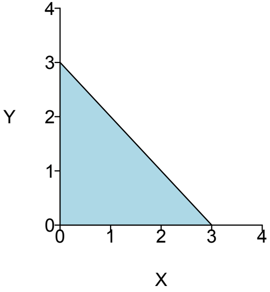
$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_0^{3-x} \frac{2}{9}\, dy = \frac{2}{9}(3-x)$$

Be careful about the range of Y given $X = x$.

$$f_X(x) = \begin{cases} \frac{2}{9}(3-x) & \text{for } x \in [0,3] \\ 0 & \text{otherwise} \end{cases}$$

Question

In the previous example, find the marginal pdf of Y.

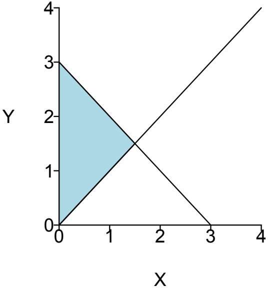


$$f_Y(y) = \begin{cases} \frac{2}{9}(3-y) & \text{for } y \in [0,3] \\ 0 & \text{otherwise} \end{cases}$$

In the previous example, find $P(X < Y)$.

$$f(x,y) = \begin{cases} \frac{2}{9} & \text{for } x \geq 0, y \geq 0 \text{ and } x + y \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Identify the region
$C = \{(x,y) : x \geq 0, y \geq 0 \text{ and } x + y \leq 3, x < y\}$.

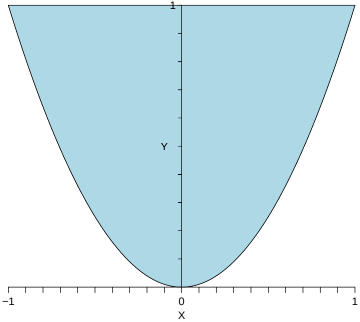


$$\begin{aligned} P(X < Y) &= \iint_{(x,y)\in C} f(x,y)\, dx\, dy \\ &= \int_0^{\frac{3}{2}} \left[\int_x^{3-x} \frac{2}{9}\, dy \right] dx \\ &= \int_0^{\frac{3}{2}} \frac{2}{9}(3-2x)\, dx = \frac{2}{9} \times \left[3x - x^2 \Big|_0^{\frac{3}{2}} \right] \\ &= \frac{2}{9}\left(\frac{9}{2} - \frac{9}{4}\right) = \frac{1}{2} \end{aligned}$$

Example 2

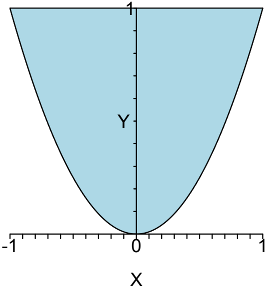
Let $f(x,y) = cx^2y$ for $x^2 \le y \le 1$.
Find:

1) c
2) $P[X \ge Y]$
3) $f_X(x)$ and $f_Y(y)$



Let $f(x,y) = cx^2y$ for $x^2 \le y \le 1$, we can rewrite the bounds as

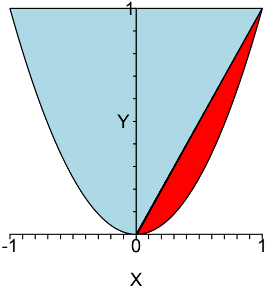
$$0 \le y \le 1, \quad -\sqrt{y} \le x \le \sqrt{y}$$



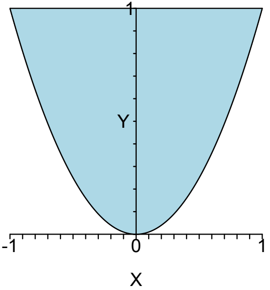
$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy \\
&= \int_0^1 \int_{-\sqrt{y}}^{\sqrt{y}} cx^2y\, dx\, dy \\
&= \int_0^1 \left(cy \frac{x^3}{3} \Bigg|_{x=-\sqrt{y}}^{\sqrt{y}} \right) dy \\
&= \int_0^1 \left(cy^{5/2}/3 + cy^{5/2}/3 \right) dy \\
&= \frac{4cy^{7/2}}{21} \Bigg|_{y=0}^{1} = \frac{4}{21} c \Longrightarrow c = \frac{21}{4}
\end{aligned}$$

We need to integrate over the region which is indicated in red below, where

$$x^2 \le y \le 1, \quad x \ge y$$



$$\begin{aligned}
P(X \ge Y) &= \int_0^1 \int_{x^2}^{x} \frac{21}{4} x^2 y \, dy \, dx \\
&= \frac{21}{4} \int_0^1 \left(\frac{x^2 y^2}{2} \Big|_{x^2}^{x} \right) dx \\
&= \frac{21}{4} \int_0^1 \left(\frac{x^4}{2} - \frac{x^6}{2} \right) dx \\
&= \frac{21}{4} \left(\frac{x^5}{10} - \frac{x^7}{14} \right) \Big|_0^1 \\
&= \frac{21}{4} \left(\frac{1}{10} - \frac{1}{14} \right) \\
&= \frac{21}{4} \left(\frac{2}{70} \right) = 0.15
\end{aligned}$$



$$f_X(x) = \int_{x^2}^{1} \frac{21}{4} x^2 y \, dy$$

$$= \frac{21}{4} \left(\left. \frac{x^2 y^2}{2} \right|_{x^2}^{1} \right)$$

$$= \frac{21}{8} \left(x^2 - x^6 \right), \text{ for } x \in (-1, 1)$$

$$f_Y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{21}{4} x^2 y \, dx$$

$$= \frac{21}{4} \left(\left. \frac{x^3 y}{3} \right|_{-\sqrt{y}}^{\sqrt{y}} \right)$$

$$= \frac{21}{4} \left(2 \frac{y^{5/2}}{3} \right)$$

$$= \frac{7}{2} y^{5/2}, \text{ for } y \in (0, 1)$$

Joint distribution of two continuous rv's

- Joint pdf
  - Non-negative $f_{X,Y}(x,y) \geq 0$, for any $x, y \in \mathbb{R}$
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy = 1$
  - For any set $C \subset \mathbb{R}^2$,

$$P[(X,Y) \in C] = \iint_{(x,y)\in C} f_{X,Y}(x,y)\ dx\ dy$$

- Between joint cdf and joint pdf

$$F_{X,Y}(a,b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f_{X,Y}(x,y)\ dx\ dy$$

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

- Marginal pdf's

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx$$

Independent random variables

Definition

Random variables X and Y are independent if any real sets $A, B \subset \mathbb{R}$,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Random variables X and Y are independent **if and only if**

- Cdf: for any $x, y \in \mathbb{R}$

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

- If both are discrete, pmf: for any $x, y \in \mathbb{R}$

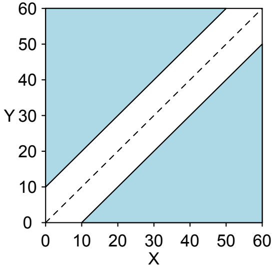
$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

- If both are continuous, pdf: for any $x, y \in \mathbb{R}$

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

Example: a man and a woman decide to meet at a certain location. If each of them independently arrives at a time uniformly distributed between 12 noon and 1 P.M., find the probability that the first to arrive has to wait longer than 10 minutes.

Let rv's X, Y be the time they arrive (uniform between 0 to 60 minutes).



$$
\begin{aligned}
& P(|X-Y|>10) \\
&= P(Y>X+10)+P(Y<X-10) \\
&= 2P(Y>X+10) \\
&= 2\iint_{y>x+10} f_{X,Y}(x,y)\,dx\,dy \\
&= 2\iint_{y>x+10} f_X(x)f_Y(y)\,dx\,dy \\
&= 2\int_{10}^{60}\int_{0}^{y-10}(1/60)^2\,dx\,dy = 25/36
\end{aligned}
$$

Independent random variables

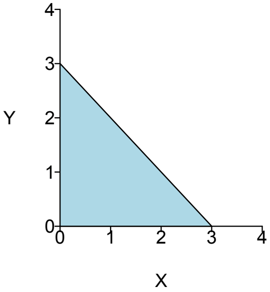
The continuous (discrete) rv's X and Y are independent **if and only if** their joint probability density (mass) function can be expressed as

$$f_{X,Y}(x, y) = g(x)h(y), \quad -\infty < x, y < \infty$$

Let X and Y be drawn uniformly from the triangle below, i.e., their joint pdf is

$$f(x,y) = \begin{cases} \frac{2}{9} & \text{for } x \geq 0, y \geq 0 \text{ and } x+y \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Are they independent?



Denote indicator function,

$$I(x,y) = \begin{cases} 1 & \text{for } x \geq 0, y \geq 0 \text{ and } x+y \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

Then for any $x, y \in \mathbb{R}$, $f(x,y) = \frac{2}{9}\ I(x,y)$.
So NOT independent.
We can use marginal pdf's $f_X(x), f_Y(y)$ to double check.

More than two random variables

☞ Random variables $X_1, X_2, \ldots, X_n$ are *independent* if any real sets $A_1, A_2, \ldots, A_n \subset \mathbb{R}$,

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$$

Random variables $X_1, X_2, \ldots, X_n$ are independent **if and only if**

- Cdf: for any $x_1, x_2, \ldots, x_n \in \mathbb{R}$

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

- If both are discrete, pmf: for any $x_1, x_2, \ldots, x_n \in \mathbb{R}$

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$$

- If both are continuous, pdf: for any $x_1, x_2, \ldots, x_n \in \mathbb{R}$

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

Sums of continuous random variables

If X, Y have a joint density $f(x,y)$, then $X+Y$ has the following density

$$\begin{aligned} f_{X+Y}(z) &= \int_{-\infty}^{\infty} f(x, z-x)\ dx \\ &= \int_{-\infty}^{\infty} f(z-y, y)\ dy \end{aligned}$$

If X and Y are independent, we can use the convolution formula

$$\begin{aligned} f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\ dx \\ &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\ dy \end{aligned}$$

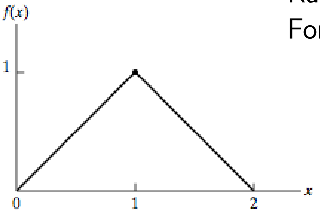
Cdf F_{X+Y} is called the *convolution* of the distributions F_X and F_Y.

Why?

$$
\begin{aligned}
F_{X+Y}(z) &= \iint_{x+y\le z} f_{X,Y}(x,y)\ dy\ dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_{X,Y}(x,y)\ dy\ dx \\
f_{X+Y}(z) &= \frac{d}{dz}F_{X+Y}(z) \\
&= \int_{-\infty}^{\infty}\left\{\frac{d}{dz}\int_{-\infty}^{z-x} f_{X,Y}(x,y)\ dy\right\}\ dx \\
&= \int_{-\infty}^{\infty} f_{X,Y}(x,z-x)\ dx
\end{aligned}
$$

Sum of two independent Uniforms: Triangular distribution

Let X and Y have independent Unif$(0,1)$ distribution. Find pdf of $X+Y$.



Range of $X+Y$ is $(0,2)$.

For any $0<z\le 1$,

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\ dx$$

$$= \int_{0}^{z} 1\ dx = z$$

For any $1<z<2$,

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\ dx$$

$$= \int_{z-1}^{1} 1\ dx = 2-z$$

$$f_{X+Y}(z) = \begin{cases} z & \text{if } 0<z\le 1 \\ 2-z & \text{if } 1<z<2 \\ 0 & \text{otherwise} \end{cases}$$

Sum of independent Exponentials (same λ)

Example: suppose $X, Y \overset{\text{ind}}{\sim} \text{Exp}(\lambda)$. What distribution does $X+Y$ have?

Range of $X+Y$ is $(0,\infty)$.
For any $z>0$,

$$
\begin{aligned}
f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx \\
&= \int_{\color{red}0}^{\color{red}z} \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)}\, dx \\
&= \int_0^z \lambda^2 e^{-\lambda z}\, dx \\
&= \lambda^2 x e^{-\lambda z}\Big|_0^z = \lambda^2 z e^{-\lambda z}
\end{aligned}
$$

$$
f_{X+Y}(z) = \begin{cases} \lambda^2 z e^{-\lambda z} & \text{if } z>0 \\ 0 & \text{otherwise} \end{cases}
$$

Recall: Gamma distribution pdf is:

$$
f_Z(z|\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\lambda z}
$$

Therefore, $X+Y \sim \text{Gamma}(2,\lambda)$.

Sum of independent random variables

Random variables X and Y are independent, then

X	Y	$X+Y$
$\mathsf{N}(\mu_1, \sigma_1^2)$	$\mathsf{N}(\mu_2, \sigma_2^2)$	$\mathsf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
$\mathsf{G}(\alpha_1, \lambda)$	$\mathsf{G}(\alpha_2, \lambda)$	$\mathsf{G}(\alpha_1 + \alpha_2, \lambda)$
$\mathsf{Poi}(\lambda_1)$	$\mathsf{Poi}(\lambda_2)$	$\mathsf{Poi}(\lambda_1 + \lambda_2)$
$\mathsf{Bin}(n_1, p)$	$\mathsf{Bin}(n_2, p)$	$\mathsf{Bin}(n_1 + n_2, p)$

Sum of independent Chi-squared distributions

Question

Suppose $X \sim \chi^2(n_1)$, and $Y \sim \chi^2(n_2)$ independently. What distribution does $X + Y$ have?

Hint: recall that $\chi^2(n) = G(n/2, 1/2)$.

$$X \sim \mathsf{G}\left(\frac{n_1}{2}, \frac{1}{2}\right), Y \sim \mathsf{G}\left(\frac{n_2}{2}, \frac{1}{2}\right)$$

$$X + Y \sim \mathsf{G}\left(\frac{n_1 + n_2}{2}, \frac{1}{2}\right) \sim \chi^2\left(n_1 + n_2\right)$$

Difference of independent Normal random variables

Question

Suppose $X \sim \mathsf{N}(\mu_1, \sigma_1^2)$, and $Y \sim \mathsf{N}(\mu_2, \sigma_2^2)$ independently. What distribution does $X - Y$ have?

Hint: find the distribution of $W = -Y$ first.
Since $g(y) = -y$ is monotonic and differentiable on $\mathbb{R}$,

$$f_W(w) = f_Y(y) \left| \frac{dy}{dw} \right| = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(-w-\mu_2)^2}{2\sigma_2^2}}$$

$$W \sim \mathsf{N}(-\mu_2, \sigma_2^2)$$

Since X and W are also independent,

$$X - Y = X + W \sim \mathsf{N}\left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2\right)$$

Recap

Random variables X and Y are independent if any real sets $A, B \subset \mathbb{R}$,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Random variables X and Y are independent **if and only if**

- Cdf: for any $x, y \in \mathbb{R}$

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

- For any $x, y \in \mathbb{R}$, the pmf / pdf

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

If X and Y are independent continuous rv's, then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\ dx$$

Expected value of $g(X, Y)$

Recap: expectation of rv $g(X)$

- Discrete case $E[g(X)] = \sum_{\text{all } x} g(x)f(x)$
- Continuous case $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$

Suppose $g(X, Y)$ is a real-valued function of rv's X and Y, then

- Discrete case

$$E[g(X, Y)] = \sum_{\text{all } x} \sum_{\text{all } y} g(x, y)f(x, y)$$

- Continuous case

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

Example: let X and Y be rv's with joint pdf $f(x,y)$. Find $E(X+Y)$

$$\begin{aligned}
E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)\ dx\ dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\ dy\ dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\ dx\ dy \\
&= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x,y)dy\right]dx + \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x,y)dx\right]dy \\
&= \int_{-\infty}^{\infty} xf_X(x)dx + \int_{-\infty}^{\infty} yf_Y(y)dy \\
&= E(X)+E(Y)
\end{aligned}$$

Expectation of sums of two rv's

$$E(X+Y) = E(X) + E(Y)$$

- It's not difficult to show that if either (or both) of the X, Y is discrete, this formula still holds.
- This results does not require X and Y to be independent.
- This can be generalized to n rv's

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

Question

What's the expected value of $X - Y$?

$$E(X-Y) = E[X + (-Y)] = E(X) + E(-Y) = E(X) - E(Y)$$

Example: suppose that n people throw their hats into the center of a room. The hats are mixed up, and each person randomly selects one. Find the expected number of people that select their own hat.

Let X denote the total number of matches.

$$X_i = \begin{cases} 1 & \text{if the } i \text{ th person selects his own hat} \\ 0 & \text{otherwise} \end{cases}$$

$$X = X_1 + X_2 + \cdots + X_n$$

For any $i = 1, 2, \ldots, n$,

$$E(X_i) = 1/n \Longrightarrow$$

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = n/n = 1$$

Let X and Y be INDEPENDENT rv's with joint pdf $f(x,y)$. Find $E(XY)$.

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\ dx\ dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\ dx\ dy \\ &= \int_{-\infty}^{\infty} yf_Y(y)\left[\int_{-\infty}^{\infty} xf_X(x)dx\right] dy \\ &= E(X)\int_{-\infty}^{\infty} yf_Y(y)dy \\ &= E(X)E(Y) \end{aligned}$$

Note: (1) this formula only holds when X and Y are independent.
(2) This is not a sufficient condition for independence.

Recap

Expectation of sum

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

$X_1, X_2, \ldots, X_n$ are independent $\Longrightarrow$ $\not\Longleftarrow$

$$E(X_1 X_2 \cdots X_n) = E(X_1)E(X_2)\cdots E(X_n)$$

Covariance

Definition

Covariance of two rv's X and Y is defined as

$$Cov(X,Y) = E[(X - EX)(Y - EY)]$$

- Simplification

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[XY + \mu_X\mu_Y - X\mu_Y - Y\mu_X]) \\ &= E(XY) - \mu_X\mu_Y \end{aligned}$$

- Recall

$$\begin{aligned} E(XY) &= \int\int xy\ f(x,y)\ dx\ dy \quad \text{if continuous} \\ &= \sum_x \sum_y xy\ f(x,y) \quad \text{if discrete} \end{aligned}$$

Properties of $Cov(X, Y) = E(XY) - E(X)E(Y)$

- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, c) = 0$
- $Cov(X, X) = Var(X)$
- $Cov(aX, bY) = ab\ Cov(X, Y)$

- $Cov(X + a, Y + b) = Cov(X, Y)$

Covariance of sums of rv's

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov\left(X_i, Y_j\right)$$

A special case

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1\leq i<j\leq n} Cov(X_i, X_j)$$

Some more special cases

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$
$$Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$$

Question

Suppose Z_1 and Z_2 are two standard normal rv's. Let

$$X = Z_1 + Z_2,\ Y = Z_1 - Z_2$$

Find $Cov(X, Y)$.

Method 1.

$$\begin{aligned} Cov(X,Y) &= Cov(Z_1 + Z_2, Z_1 - Z_2) \\ &= Cov(Z_1, Z_1) + Cov(Z_2, Z_1) + Cov(Z_1, -Z_2) + Cov(Z_2, -Z_2) \\ &= Cov(Z_1, Z_1) + Cov(Z_2, Z_1) - Cov(Z_1, Z_2) - Cov(Z_2, Z_2) \\ &= Var(Z_1) - Var(Z_2) = 0 \end{aligned}$$

Method 2.

$$\begin{aligned} E(XY) &= E(Z_1^2 - Z_2^2) = E(Z_1^2) - E(Z_2^2) = 0 \\ E(X) &= E(Z_1) + E(Z_2) = 0,\ E(Y) = E(Z_1) - E(Z_2) = 0 \end{aligned}$$

Zero covariance and independence

- X and Y are independent $\Longrightarrow Cov(X, Y) = 0$

  $Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$

- $X_1, X_2, \ldots, X_n$ are independent $\Longrightarrow$

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

- $Cov(X, Y) = 0 \not\Longrightarrow X$ and Y are independent
  Counter example?

Counter example

Let $X \sim$ Unif$(-0.5, 0.5)$ and $Y = X^2$. Find $Cov(X, Y)$, and decide if X and Y and independent.

Covariance

$$E(X) = 0,\ E(XY) = E(X^3) = \int_{-0.5}^{0.5} x^3 dx = \left.\frac{x^4}{4}\right|_{-0.5}^{0.5} = 0$$

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2) = 0$$

Independence: since Y depends on X, so not independent.
(To be more rigorous, we need to show

$$f(x, y) \neq f_X(x) f_Y(y)$$

for some $x, y \in \mathbb{R}$.)

Question

Let $X_1, \ldots, X_n$ be independent rv's having the same variance σ^2, and

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Find $Var(\bar{X})$.

$$\begin{aligned}
Var(\bar{X}) &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(X_i)\right] \\
&= \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}
\end{aligned}$$

Correlation

Since $Cov(X,Y)$ depends on the magnitude of X and Y we would prefer to have a measure of association that is not effected by arbitrary changes in the scales of the random variables.

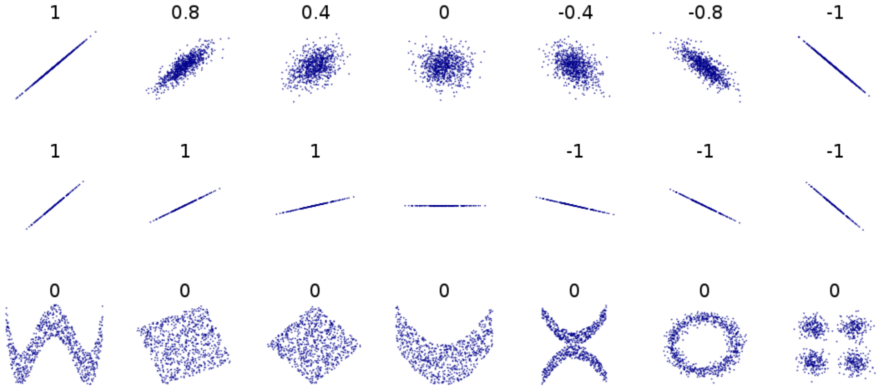
Definition

The most common measure of linear association is correlation which is defined as

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\,\sqrt{Var(Y)}}$$

- range: $-1 < \rho(X,Y) < 1$
- the magnitude (i.e. absolute value) of the $\rho(X,Y)$ measures the strength of the linear association
- the sign determines if it is a positive or negative relationship.

Correlation



Conditional distributions

A, B are two events in a sample space, then conditional probability

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Let X and Y be random variables then
Conditional probability (discrete):

$$\begin{aligned} p_{X|Y}(x|y) &= P(X = x|Y = y) \\ &= \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x, y)}{p_Y(y)} \end{aligned}$$

Conditional density (continuous):

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

Product rule:

$$\begin{aligned} p(x, y) &= p_{X|Y}(x|y)p_Y(y) \\ f(x, y) &= f_{X|Y}(x|y)f_Y(y) \end{aligned}$$

Question

1. The joint probability mass function of X and Y, is given by

$$p(0,0) = 0.5,\ p(0,1) = 0.1,\ p(1,0) = 0.2,\ p(1,1) = 0.2$$

Find the pmt of X given that $Y = 0$.

$$p_Y(0) = p(0,0) + p(1,0) = 0.7$$

$$p_{X|Y}(0|0) = \frac{p(0,0)}{p_Y(0)} = \frac{0.5}{0.7} = \frac{5}{7}$$
$$p_{X|Y}(1|0) = \frac{p(1,0)}{p_Y(0)} = \frac{0.2}{0.7} = \frac{2}{7}$$

Conditional distributions are well-defined distributions

Discrete case

$$
\begin{aligned}
\sum_{\text{all } x} p_{X|Y}(x|y) &= \sum_{\text{all } x} \frac{p(x,y)}{p_Y(y)} \\
&= \frac{\sum_{\text{all } x} p(x,y)}{p_Y(y)} \\
&= \frac{p_Y(y)}{p_Y(y)} = 1
\end{aligned}
$$

Continuous case

$$
\begin{aligned}
\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx &= \int_{-\infty}^{\infty} \frac{f(x,y)}{f_Y(y)}dx \\
&= \frac{\int_{-\infty}^{\infty} f(x,y)dx}{f_Y(y)} \\
&= \frac{f_Y(y)}{f_Y(y)} = 1
\end{aligned}
$$

Conditional, joint and marginal distributions

Product rule:

$$p(x,y) = p(x|y)p_Y(y)$$

$$f(x,y) = f(x|y)f_Y(y)$$

Bayes Theorem:

$$p(x|y) = \frac{p(x,y)}{p_Y(y)} = \frac{p(y|x)p_X(x)}{p_Y(y)}$$

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

Law of Total Probability:

$$p_X(x) = \sum_{\text{all } y} f(x,y) = \sum_{\text{all } y} f(x|y)f_Y(y)$$

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

Conditional distribution and independence

If rv's X and Y are independent, then

$$\text{conditional distribution} = \text{marginal distribution}$$

Discrete case: write pmf as $f(\cdot)$

Continuous case

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

Constrained Sum of Poissons

Let X and Y be independent Poisson random variables with rates λ_1 and λ_2, what is the conditional distribution of X given $X + Y = n$?

$$
\begin{aligned}
P(X = k | X + Y = n) &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\
&= \frac{P(X = k, Y = n - k)}{P(X + Y = n)} \\
&= \frac{P(X = k) P(Y = n - k)}{P(X + Y = n)} \\
&= \frac{e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!}}{e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^n}{n!}} = \frac{\lambda_1^k}{k!} \frac{\lambda_2^{n-k}}{(n-k)!} \frac{n!}{(\lambda_1+\lambda_2)^n} \\
&= \frac{n!}{k!(n-k)!} \frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} = \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}
\end{aligned}
$$

$$
X | X + Y = n \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)
$$

A manufacturing process consists of two stages. The first stage takes Y minutes, and the whole process takes X minutes (which includes the first Y minutes). Suppose that X and Y have a joint pdf

$$f(x,y) = e^{-x}, \quad \text{for } 0 < y < x < \infty$$

If we observe that Y takes 4 minutes, what is the probability that X takes longer than 9 minutes?

$$f_Y(y) = \int_y^\infty e^{-x}\, dx = e^{-y}$$

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{e^{-x}}{e^{-y}} = e^{y-x}$$

$$\begin{aligned} P(X \geq 9|Y=4) &= \int_9^\infty f(x|y=4)\, dx = \int_9^\infty e^{4-x}\, dx \\ &= -e^{4-x}\Big|_9^\infty = e^{-5} = 0.00673 \end{aligned}$$