Lecture 12: Independent RVs

Lisa Yan
July 23, 2018

“Woah, we're half way there
Woah, livin' on a prayer”
– Bon Jovi, 1986
Announcements

Midterm tomorrow! 😊

Tuesday, July 24, 7-9pm
Hewlett 201
  ◦ Closed book, closed computer, no calculators
  ◦ Two 8.5” x 11” pages of notes (front and back) allowed
  ◦ Covers up to and including Friday 7/20’s lecture

PS4 due next Monday (7/30)
Previously proved properties of expectation

\[
E \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i]
\]

\[
E[aX + bY + c] = aE[X] + bE[Y] + c
\]

\[
Var(aX + b) = a^2 Var(X)
\]

For non-negative random variables \(X\) and \(Y\):

\[
E[X] = \sum_{k=1}^{\infty} P(X \geq k)
\]

\[
E[Y] = \int_{0}^{\infty} P(Y \geq y) \, dy
\]

(Friday’s lecture was the hardest math we’ll do in this class)
Expectation of Binomial, revisited

Consider independent coin flips of a coin with probability $p$ of resulting in heads.

Let $X =$ number of successes in $n$ coin flips. $X \sim \text{Bin}(n,p)$

Prove that $E[X] = np$.

Proof:

Let $X_i =$ indicator for $i$-th trial being heads:

$$E[X_i] = P(X_i = 1) = p$$

Then $X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$, and

$$E \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np$$
Expectation of negative binomial, revisited

Consider independent coin flips of a coin with probability $p$ of resulting in heads.

Let $X =$ number of trials until $r$ successes. $X \sim \text{NegBin}(r, p)$

Prove that $E[X] = r/p$.

Proof:

Let $X_i =$ # of trials to get a first success after the (i-1)-th success.

\[
X_i \sim \text{Geo}(p) \quad E[X_i] = 1/p
\]

Then $X = X = X_1 + X_2 + \cdots + X_r = \sum_{i=1}^{r} X_i$, and

\[
E \left[ \sum_{i=1}^{r} X_i \right] = \sum_{i=1}^{r} E[X_i] = \sum_{i=1}^{r} \frac{1}{p} = \frac{r}{p}
\]
Joint distributions

Discrete RVs:

Joint PMF:

\[ p_{X,Y}(a, b) = P(X = a, Y = b) \]

Marginal PMFs:

\[ p_X(a) = P(X = a) = \sum_y p_{X,Y}(a, y) \]
\[ p_Y(a) = P(Y = a) = \sum_x p_{X,Y}(x, b) \]

Continuous RVs:

Joint PDF:

\[ P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x, y) \, dy \, dx \]

Marginal PDFs:

\[ f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a, y) \, dy \]
\[ f_Y(b) = \int_{-\infty}^{\infty} f_{X,Y}(x, b) \, dy \]
Goals for today

Independent RVs (aka Joint distributions, pt. 2)
- Discrete case
- Continuous case
- Sum of independent RVs
- Convolution
Independent discrete variables

Two discrete random variables $X$ and $Y$ are defined as independent if:

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

Event interpretation: all events $X = x$ and $Y = y$ need to be independent

If two variables are not independent, they are called dependent

Intuitively: knowing the value of $X$ tells us nothing about the distribution of $Y$ (and vice versa).
Coin flips

Flip a coin with probability $p$ of “heads” $n+m$ times.

- Let $X =$ # of heads in first $n$ flips. \quad $X \sim \text{Bin}(n,p)$
  
- Let $Y =$ # of heads in next $m$ flips. \quad $Y \sim \text{Bin}(m,p)$

Are $X$ and $Y$ independent?

Solution:

\[
P(X = x, Y = y) = \binom{n}{x} p^x (1 - p)^{n-x} \binom{m}{y} p^y (1 - p)^{m-y}
= P(X = x)P(Y = y)
\]

Yes, $X$ and $Y$ are independent.

Let $Z =$ # of total heads in $n + m$ flips. Are $X$ and $Z$ independent?

No, $X$ and $Z$ are dependent. Counterexample: When $Z = 0$, $X = 0$. 
If the joint distribution of $X$ and $Y$ is separable into the product of two functions (for all values of $X$ and $Y$), then $X$ and $Y$ are independent random variables.
Web server requests

Let $N = \#$ of requests you receive per day.
- Suppose $N \sim \text{Poi}(\lambda)$.
- Each request comes from a human (w.p. $p$) or a bot (w.p. $1 - p$), independently
- Define: $X = \#$ requests from humans/day
  $Y = \#$ requests from bots/day

Are $X$ and $Y$ independent?

Solution:
1. Can you find distributions for $X$ and $Y$ given a particular value $N = n$?
   - $X \mid N = n \sim \text{Bin}(n,p)$
   - $Y \mid N = n \sim \text{Bin}(n,1-p)$

2. Try finding an expression for $P(X = i, Y = j)$ using the law of total probability on the mutually exclusive events $N = i + j$ or $N \neq i + j$.

$$P(X = i, Y = j) = P(X = i, Y = j \mid N = i + j)P(N = i + j)$$
$$+ P(X = i, Y = j \mid N \neq i + j)P(N \neq i + j)$$

0
Web server requests

Let $N =$ # of requests you receive per day.

- Suppose $N \sim \text{Poi} (\lambda)$.
- Each request comes from a human (w.p. $p$) or a bot (w.p. $1 - p$), independently.
- Define: $X =$ # requests from humans/day
  $Y =$ # requests from bots/day

Are $X$ and $Y$ independent?

Solution:

2. Joint distribution: $P(X = i, Y = j) = P(X = i, Y = j | N = i + j)P(N = i + j)$

3. Plug in known PMFs.

4. Separate $j$ terms and $i$ terms.
Web server requests

Let $N =$ # of requests you receive per day.

- Suppose $N \sim \text{Poi}(\lambda)$.
- Each request comes from a human (w.p. $p$) or a bot (w.p. $1 - p$), independently.
- Define: $X =$ # requests from humans/day $\quad X \mid N = n \sim \text{Bin}(n,p)$
  $Y =$ # requests from bots/day $\quad Y \mid N = n \sim \text{Bin}(n,1-p)$

Are $X$ and $Y$ independent? 

**Yes!!! Furthermore, they are also Poisson!**

**Solution:**

4. Separate $j$ terms and $i$ terms.

$$P(X = i, Y = j) = e^{-\lambda} \frac{\lambda^i p^i (1-p)^j}{i! j!}$$

5. Try to form valid marginal probabilities.

$$e^{-\lambda} \frac{(\lambda p)^i}{i!} \frac{(\lambda (1-p))^j}{j!}$$

$$= \frac{(\lambda p)^i}{i!} e^{-\lambda p} e^{-\lambda(1-p)} \frac{(\lambda (1-p))^j}{j!}$$

$$= P(X = i) P(Y = j)$$

, where $X \sim \text{Poi}(\lambda p)$ and $Y \sim \text{Poi}(\lambda(1-p))$.  

???
Independent continuous variables

Two continuous random variables $X$ and $Y$ are defined as independent if:

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b), \quad -\infty < a, b < \infty$$

Equivalently, for all $a, b$:

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

More generally, independent if joint density factors separately:

$$f_{X,Y}(x, y) = g(x)h(y)$$

Where $g(x)$ and $h(y)$ do not have to be true densities.
Joint density \textbf{fun} time

Are X and Y independent in the following cases?

1. \( f_{X,Y}(x, y) = 6e^{-3x}e^{-2y} \), where \( 0 < x, y < \infty \)

   \textbf{Yes!} \( g(x) = 3e^{-3x} \), \( h(y) = 2e^{-2y} \), and \( X \sim \text{Exp}(3), Y \sim \text{Exp}(2) \) (valid PDFs)

2. \( f_{X,Y}(x, y) = 4xy \), where \( 0 < x, y < 1 \)

   \textbf{Yes!} \( g(x) = 2x \), \( h(y) = 2y \)

3. \( f_{X,Y}(x, y) = 4xy \), where \( 0 < x + y < 1 \)

   \textbf{No!} Cannot capture constraint on \((x + y)\) in factorization
The joy of meetings

Two people set up a meeting time.

- Each arrives independently at a time uniformly distributed between 12pm and 12:30pm.
- Define: $X = \# \text{ minutes past 12pm person 1 arrives}$
  $Y = \# \text{ minutes past 12pm person 2 arrives}$

$P(\text{first to arrive waits > 10 minutes for the other})$?

**Solution:**

WTF: $P(X + 10 < Y) + P(Y + 10 < X) = 2P(X + 10 < Y)$ \hspace{1cm} (by symmetry)

$2P(X + 10 < Y) = 2 \int_{x+10<y} f_{x,y}(x,y) dx \, dy$

$= 2 \int_{x+10<y} f_x(x)f_y(y) dx \, dy$ \hspace{1cm} (independence)

$= 2 \int_{x+10<y} \left(\frac{1}{30}\right)^2 dx \, dy$ \hspace{1cm} \(X \sim \text{Unif}(0,30), \ Y \sim \text{Unif}(0,30)\)
The joy of meetings

Two people set up a meeting time.

- Each arrives independently at a time uniformly distributed between 12pm and 12:30pm.
- Define: \( X = \) # minutes past 12pm person 1 arrives
  \( Y = \) # minutes past 12pm person 2 arrives

\[ P(\text{first to arrive waits > 10 minutes for the other})? \]

Solution:

\[
2P(X + 10 < Y) = 2 \int \int_{x+10<y} \left(\frac{1}{30}\right)^2 \, dx \, dy \quad (X \sim \text{Unif}(0,30), \, Y \sim \text{Unif}(0,30))
\]

\[
= 2 \int_{y=10}^{30} \int_{x=0}^{y-10} \left(\frac{1}{30}\right)^2 \, dx \, dy = \frac{2}{30^2} \int_{y=10}^{30} (y - 10) \, dy
\]

\[
= \frac{2}{30^2} \left[ \frac{y^2}{2} - 10y \right]_{y=10}^{30} = \frac{2}{30^2} \left[ \left( \frac{30^2}{2} - 300 \right) - \left( \frac{10^2}{2} - 100 \right) \right] = \frac{4}{9}
\]
Defects on a hard drive

A single point defect is uniformly distributed over a disk of radius R.

Are X and Y independent?

Solution:

\[
f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dy = \frac{1}{\pi R^2} \int_{y: x^2 + y^2 \leq R^2} \, dy
\]

\[
= \frac{1}{\pi R^2} \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \, dy \quad \text{if } 0 \leq x \leq R
\]

\[
f_X(x) = \frac{2 \sqrt{R^2 - x^2}}{\pi R^2} \quad \text{if } 0 \leq x \leq R
\]

\[
f_Y(y) = \frac{2 \sqrt{R^2 - y^2}}{\pi R^2} \quad \text{if } 0 \leq y \leq R, \text{ by symmetry}
\]

\[
f_{X,Y}(x,y) = \begin{cases} 
\frac{1}{\pi R^2} & \text{if } x^2 + y^2 \leq R^2 \\
0 & \text{otherwise}
\end{cases}
\]

\[
f_X(x) f_Y(y) \neq f_{X,Y}(x,y)
\]

No. X and Y are dependent.
Defects on a hard drive

A single point defect is uniformly distributed over a disk of radius R.

Calculate the expected distance of the defect from the center, i.e., \( E[D] \), where \( D = \sqrt{X^2 + Y^2} \)

Solution:

\[
F_D(a) = \frac{a^2}{R^2}
\]

\[
E[D] = \int_0^R P(D > a)da = \int_0^R (1 - F_D(a))da
\]

\[
= \int_0^R \left(1 - \frac{a^2}{R^2}\right)da
\]

\[
= \left[a - \frac{a^3}{3R^2}\right]_0^R = \frac{2}{3}R
\]
Independence of Multiple Variables

$n$ discrete random variables $X_1, X_2, \ldots, X_n$ are independent iff:

$$P(X_1 = x_1, X_2 = x_2 \ldots, X_k = x_k) = \prod_{i=1}^{k} P(X_i = x_i), \text{ for all subsets } x_1, \ldots, x_k$$

$n$ continuous random variables $X_1, X_2, \ldots, X_n$ are independent iff:

$$P(X_1 \leq x_1, X_2 \leq x_2 \ldots, X_k \leq x_k) = \prod_{i=1}^{k} P(X_i \leq x_i), \text{ for all subsets } x_1, \ldots, x_k$$

def i.i.d. RVs – independent and identically distributed random variables

Note the difference:

$$X = Y: \quad P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \lambda e^{-\lambda x} dx \quad (\text{let } X \sim \text{Exp}(\lambda), Y \sim \text{Exp}(\lambda))$$

$$X, Y \text{ i.i.d.: } \quad P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} \lambda e^{-\lambda x} \lambda e^{-\lambda y} dxdy$$
Independence is Symmetric!

Let $X_1, X_2, \ldots,$ be a sequence of independent and identically distributed (i.i.d.) continuous random variables.

Define:

- “Record value” = $X_n > X_k$, for all $k = 1, \ldots, n-1$ (i.e., $X_n = \max(X_1, \ldots, X_n)$ )
- $A_k$ = indicator variable that $X_k$ is a “record value”

1. Is $A_{n+1}$ independent of $A_n$?
2. Is $A_n$ independent of $A_{n+1}$?

Solution:

(by symmetry, $P(A_n) = \frac{1}{n}$ and $P(A_{n+1}) = \frac{1}{n+1}$).

$P(A_n A_{n+1}) = \left(\frac{1}{n}\right) \left(\frac{1}{n+1}\right) = P(A_{n+1}) P(A_n)$
Break

Attendance: tinyurl.com/cs109summer2018
Summations of random variables

Suppose $X$, $Y$ are random variables.
Define the random variable $Z = X + Y$.
Expectation easy to calculate:

$$E[Z] = E[X] + E[Y] = \begin{cases} 
\sum_x x p_X(x) + \sum_y y p_Y(y) & X, Y \text{ discrete} \\
\int_{-\infty}^{\infty} x f_X(x)dx + \int_{-\infty}^{\infty} y f_Y(y)dy & X, Y \text{ continuous} 
\end{cases}$$

What about the probability distribution of $Z$?
For discrete $X$ and $Y$:

$$p_Z(z) = P(Z = z) = P(X + Y = z) = \sum_x p_{X,Y}(x, z - x)$$

But what does this mean?
Dance, Dance, Convolution

For any discrete random variables X, Y, and Z, where Z = X + Y,

\[ p_Z(z) = P(X + Y = z) = \sum_x p_{X,Y}(x, z - x) \]

In particular, for independent discrete random variables X and Y,

\[ p_Z(z) = \sum_x p_X(x)p_Y(z - x) = \sum_y p_X(z - y)p_Y(y) \]

p_Z is defined as the convolution of p_X and p_Y.
Dance, Dance, Convolution Extreme

For independent random variables X and Y,

\[ p_Z(z) = \sum_x p_X(x)p_Y(z - x) \quad \text{X, Y discrete} \]

\[ f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx \quad \text{X, Y continuous} \]

Who wants the proof for continuous?
For continuous independent RVs $X$ and $Y$,

1. Prove CDF $F_{X+Y}$ is the convolution of $F_X$ and $F_Y$.

2. Take the derivative of the CDF to prove $f_{X+Y}$ is the convolution of $f_X$ and $f_Y$.

$$F_{X+Y}(a) = P(X + Y \leq a) = \iint_{x+y\leq a} f_X(x)f_Y(y) \, dx \, dy$$

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

$$= \int_{y=-\infty}^{\infty} F_X(a-y)f_Y(y) \, dy \quad \text{(Definition of CDF convolution)}$$

$$f_{X+Y}(a) = \frac{dF_{X+Y}(a)}{da} = \frac{d}{da} \int_{y=-\infty}^{\infty} F_X(a-y)f_Y(y) \, dy$$

$$= \int_{y=-\infty}^{\infty} \frac{\partial}{\partial a} F_X(a-y)f_Y(y) \, dy$$

$$= \int_{y=-\infty}^{\infty} f_X(a-y)f_Y(y) \, dy \quad \text{(Definition of PDF convolution)}$$
Common sums of independent RVs

Convolutions are tedious, so we will now show:
- Sum of independent binomials \(\sim\) binomial
- Sum of independent poissons \(\sim\) poisson
- Sum of independent normals \(\sim\) normal
Sum of independent binomials

Let $X$ and $Y$ be independent RVs, where $X \sim \text{Bin}(n_1, p)$ and $Y \sim \text{Bin}(n_2, p)$.

What is the distribution of $Z = X + Y$?

Solution:

Hint: all $n_1 + n_2$ trials are independent and have same probability of success

Define: $Z = \#$ of successes in $n_1 + n_2$ trials, each trial independent with $P(\text{success}) = p$

$Z \sim \text{Bin}(n_1 + n_2, p)$
Sum of independent poissons

Let $X$ and $Y$ be independent RVs, where

$X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$.

What is the distribution of $Z = X + Y$?

$Z \sim \text{Poi}(\lambda_1 + \lambda_2)$

**Proof (for reference):**

$$P(X + Y = n) = \sum_{k=0}^{n} p_X(k)p_Y(n-k)$$

$$(X \text{ and } Y \text{ are independent, convolution})$$

$$= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} = e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!}$$

$$(X \text{ and } Y \text{ are independent, convolution})$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} = \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda_1^k \lambda_2^{n-k}$$

$$(\text{binomial theorem})$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n$$
Sum of independent normals

Let $X$ and $Y$ be independent RVs, where

$X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$.

What is the distribution of $Z = X + Y$?

Solution:

$$Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

(Proof is left to Wikipedia)
Virus Infections

There are 150 computers in a dorm.

- 50 Macs (each independently infected w.p. 0.1)
- 100 PCs (each independently infected w.p. 0.4)

What is \( P(\text{machines infected} \geq 40) \)?

Solution:

Define: \( A = \# \text{ infected Macs} \) \( A \sim \text{Bin}(50, 0.1) \approx X \sim \text{N}(5, 4.5) \)

\( B = \# \text{ infected PCs} \) \( B \sim \text{Bin}(100, 0.4) \approx Y \sim \text{N}(40, 24) \)

WTF: \( P(A + B \geq 40) \approx P(X + Y \geq 39.5) \) (continuity correction)

Define: \( W = X + Y \) \( W \sim \text{N}(5 + 40 = 45, 4.5+24 = 28.5) \)

\[
P(W \geq 39.5) = P \left( \frac{W - 45}{\sqrt{28.5}} \geq \frac{39.5 - 45}{\sqrt{28.5}} \right) \approx 1 - \Phi(-1.03) \approx 0.8485
\]
General sums of independent RVs

Sum of independent binomials $\sim$ binomial

$$X_i \sim \text{Bin}(n_i, p) \Rightarrow \sum_{i=1}^{N} X_i \sim \text{Bin}\left(\sum_{i=1}^{N} n_i, p\right)$$

Sum of independent poissons $\sim$ poisson

$$X_i \sim \text{Poi}(\lambda_i) \Rightarrow \sum_{i=1}^{N} X_i \sim \text{Poi}\left(\sum_{i=1}^{N} \lambda_i\right)$$

Sum of independent normal $\sim$ normal

$$X_i \sim N(\mu_i, \sigma_i^2) \Rightarrow \sum_{i=1}^{N} X_i \sim N\left(\sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2\right)$$
Sum of independent uniform RVs

Let $X$ and $Y$ be independent RVs, where $X \sim \text{Unif}(0,1)$ and $Y \sim \text{Unif}(0,1)$.

What is the distribution of $Z = X + Y$?

**Hint:** no choice but to use definition of convolution

**Solution:**

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy = \int_0^1 f_X(a-y) f_Y(y) dy = \int_0^1 f_X(a-y) dy$$

Note: for different $a$, need bounds on $y$ and $a$ such that $f_X(a-y) = 1$

$$= \begin{cases} 
\int_{y=0}^{a} 1 \, dy = a & 0 \leq a \leq 1 \\
\int_{y=a-1}^{1} 1 \, dy = 2 - a & 1 \leq a \leq 2 \\
0 & \text{otherwise}
\end{cases}$$
Summary of this time

Independent RVs:

\[ p_{X,Y}(x, y) = p_X(x)p_Y(y) \]  \hspace{1cm} \text{(discrete } X \text{ and } Y) \]

\[ f_{X,Y}(x, y) = f_X(x)f_Y(y) \]  \hspace{1cm} \text{(continuous } X \text{ and } Y) \]

\[ F_{X,Y}(x, y) = F_X(x)F_Y(y) \]  \hspace{1cm} \text{(any } X \text{ and } Y) \]

Generally, to prove independence, you just have to prove \textit{separability}.
Summary of this time

Z = X + Y, where X and Y are independent RVs:

\[
p_Z(z) = \sum_{x} p_X(x)p_Y(z - x) \quad \text{ (discrete X and Y)}
\]

\[
f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx \quad \text{ (continuous X and Y)}
\]

Convenient sums of independent RVs:

- Binomial: if same \( p \) of trial success, sum up trial counts \( n_i \)
- Poisson: sum up rates \( \lambda_i \)
  - we proved that if \( X \sim \text{Poi}(\lambda p) \) and \( Y \sim \text{Poi}(\lambda(1-p)) \), \( X + Y \sim \text{Poi}(p) \). (Web Servers)
- Normal: sum up means, sum up variances
  - On homework, you will find distribution of \( aX + bY \) (linear comb. of normal)