CS145: Probability & Computing

Lecture 11: Marginal and Conditional Probability Densities

Figure credits:
Bertsekas & Tsitsiklis, ***Introduction to Probability****, 2008*
Pitman, ***Probability****, 1999*

CS145: Lecture 11 Outline

- Joint Distribution
- Marginal and Conditional Probability Densities
- Conditional Expectations

Joint Probability Distributions

Definition

The *joint distribution function* of X and Y is

$$F(x, y) = \Pr(X \leq x, Y \leq y).$$

The variables X and Y have joint density function f if for all x, y,

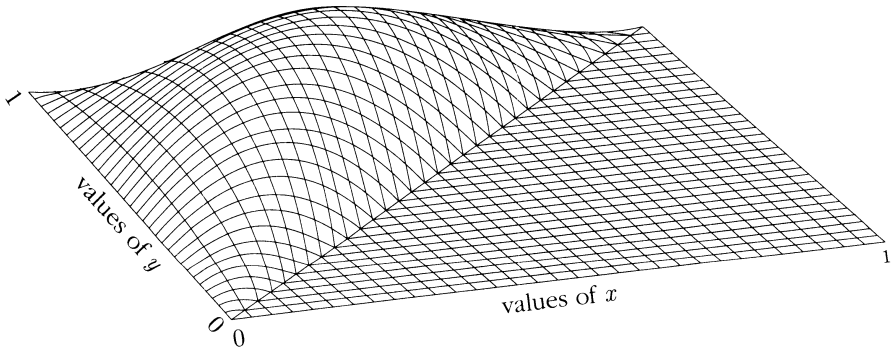
$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)\, du\, dv.$$

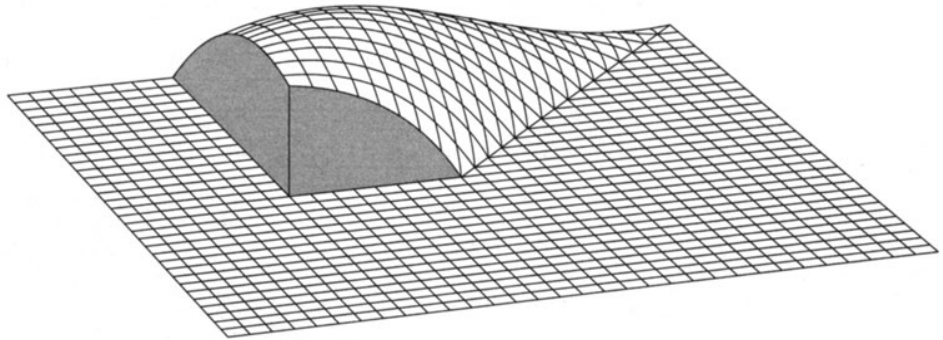
$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

when the derivative exists.

$$\mathbf{P}(x \leq X \leq x+\delta,\ y \leq Y \leq y+\delta) \approx f_{X,Y}(x, y) \cdot \delta^2$$

Example: Joint Distribution





$$f_{XY}(x,y) = 5!x(y-x)(1-y)$$

$$0 < x < y < 1$$

$$\int_0^1 \int_0^y x(y-x)(1-y)\,dxdy = \frac{1}{5!}$$

$$P(X > 0.25, Y > 0.5) =$$

$$\int_{0.5}^1 \int_{0.25}^y f_{XY}(x,y)\,dxdy$$

Pitman's Probability, 1999

2D Uniform Distributions

Density function:

$$f(x,y) = \begin{cases} 1 & x,y \in [0,1]^2 \\ 0 & \text{otherwise.} \end{cases}$$

Probability distribution:

$$F(x,y) = Pr(X \leq x, Y \leq y) = \int_0^{\min[1,x]} \int_0^{\min[1,y]} 1 dx dy$$

$$F(x,y) = \begin{cases} 0 & \text{if } x < 0 \text{ or } y < 0 \\ xy & \text{if } x,y \in [0,1]^2 \\ x & 0 \leq x \leq 1,\ y > 1 \\ y & 0 \leq y \leq 1,\ x > 1 \\ 1 & x > 1,\ y > 1. \end{cases}$$

Independence

Definition

The random variables X and Y are *independent* if for all x, y,

$$\Pr((X \leq x) \cap (Y \leq y)) = \Pr(X \leq x) \Pr(Y \leq y).$$

Two random variables are independent if and only if

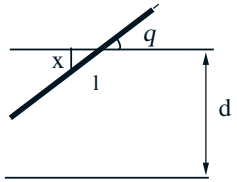
$$F(x, y) = F_X(x) F_Y(y).$$

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

If X and Y are independent then $E[XY] = E[X]E[Y]$.

Buffon's Needle

- Parallel lines at distance d
  Needle of length ℓ (assume $\ell < d$)
- Find $\mathbf{P}$(needle intersects one of the lines)



Θ = smallest angel between the needle and a parallel line, $0 \le \Theta \le \pi/2$

Georges-Louis Leclerc, Comte de Buffon (1707-1787, by François-Hubert Drouais

- $X \in [0, d/2]$: distance of needle midpoint to nearest line
- **Model:** X, Θ uniform, independent $f_{X,\Theta}(x,\theta) = f_X(x) f_\Theta(\theta) = \frac{2}{d}\frac{2}{\pi}$

$$f_{X,\Theta}(x,\theta) = \frac{4}{\pi d} \qquad 0 \le x \le d/2,\ 0 \le \theta \le \pi/2$$

- Intersect if $X \le \frac{\ell}{2}\sin\Theta$

$$\mathbf{P}\left(X \le \frac{\ell}{2}\sin\Theta\right) = \int\int_{x \le \frac{\ell}{2}\sin\theta} f_X(x) f_\Theta(\theta)\, dx\, d\theta$$

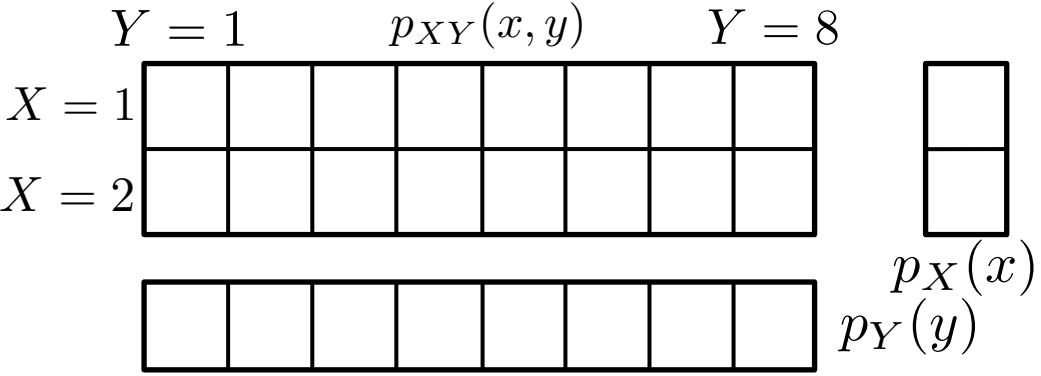
$$= \frac{4}{\pi d}\int_0^{\pi/2}\int_0^{(\ell/2)\sin\theta} dx\, d\theta$$

$$= \frac{4}{\pi d}\int_0^{\pi/2} \frac{\ell}{2}\sin\theta\, d\theta = \frac{2\ell}{\pi d}$$

CS145: Lecture 11 Outline

- Joint Distribution
- Marginal and Conditional Probability Densities
- Conditional Expectations

Reminder: Discrete Marginals



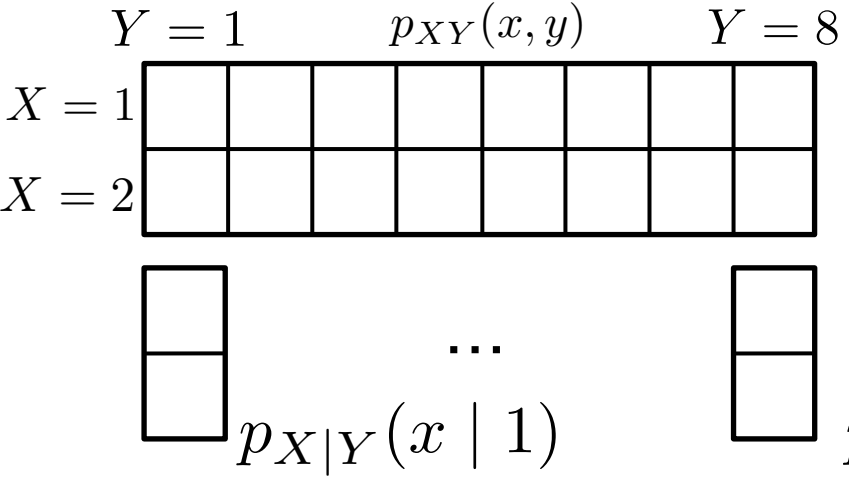
- The *joint probability mass function* of two variables:

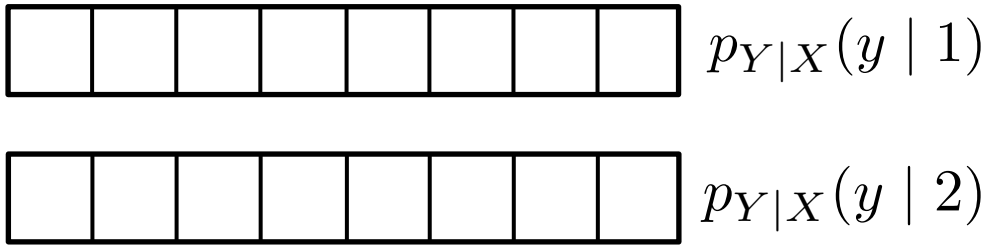
$$p_{XY}(x, y) = P(X = x \text{ and } Y = y)$$

- The range of each variable defines a partition of the sample space, so the *marginal distributions* can be computed from the joint distribution:

$$p_X(x) = P(X = x) = \sum_y p_{XY}(x, y)$$
$$p_Y(y) = P(Y = y) = \sum_x p_{XY}(x, y)$$

Reminder: Discrete Conditionals





- By the definition of conditional probability:

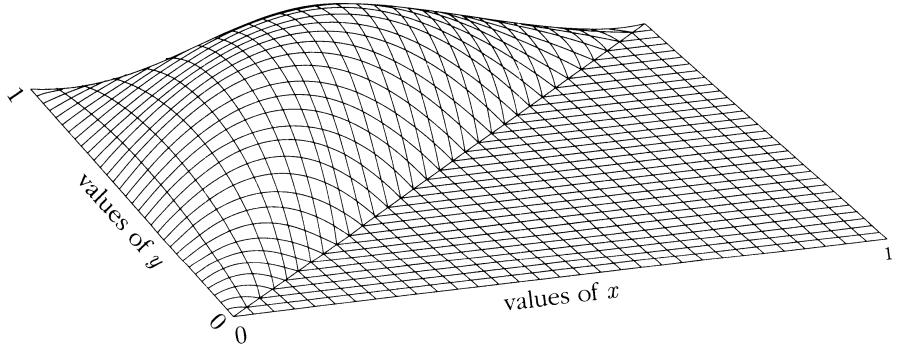
$$P(X = x \mid Y = y) = \frac{P(X = x \text{ and } Y = y)}{P(Y = y)}$$

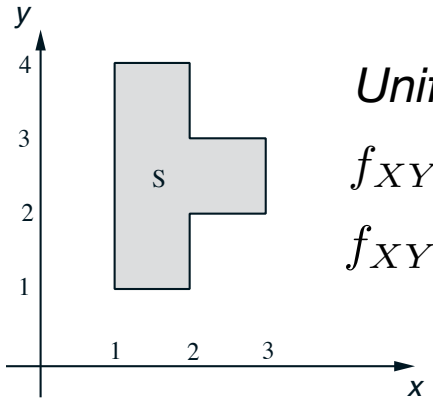
- The *conditional probability mass function* is then:

$$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{XY}(x, y)}{\sum_{x'} p_{XY}(x', y)}$$

Joint Probability Density Functions

$$f_{XY}(x,y) = 5!x(y-x)(1-y)$$





Uniform on set S:

$f_{XY}(x,y) = \frac{1}{4}$ if $(x,y) \in S$.

$f_{XY}(x,y) = 0$ otherwise.

- The *joint probability density function (PDF)* of two variables is defined so:

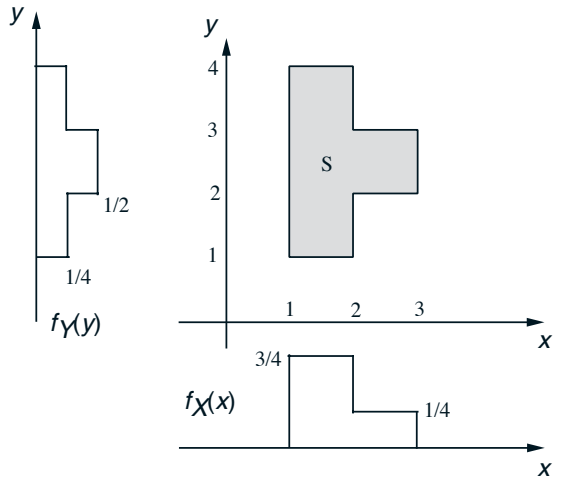
$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x,y)\, dxdy$$

$$P(x \leq X \leq x+\delta, y \leq Y \leq y+\delta) \approx f_{XY}(x,y)\delta^2$$

- To define a proper distribution, the PDF must be *normalized*:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x,y)\, dxdy = 1$$

Marginal Probability Density Functions



Uniform on set S:

$f_{XY}(x,y) = \frac{1}{4}$ if $(x,y) \in S$.

$f_{XY}(x,y) = 0$ otherwise.

- The *marginal probability density functions (PDF)* of *X* and *Y* equal:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\, dy \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\, dx$$

- The marginals are defined to compute consistent probabilities:

$$P(y_1 \le Y \le y_2) = \int_{y_1}^{y_2} f_Y(y)\, dy = \int_{y_1}^{y_2} \int_{-\infty}^{+\infty} f_{XY}(x,y)\, dx dy$$

Marginal Distributions

Definition

Given a joint distribution function $F_{X,Y}(x,y)$ the *marginal distribution function* of X is

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx$$

and the corresponding marginal density functions is

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

Example: Uniform Distributions

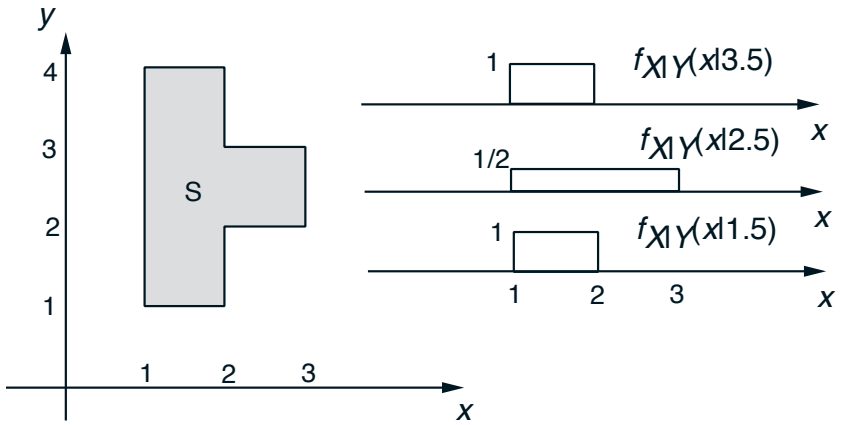
$$f_{X,Y}(x,y) = \begin{cases} 1 & x,y \in [0,1]^2 \\ 0 & \text{otherwise.} \end{cases} \qquad f_X(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Conditional Probability Density Functions

Uniform on set S:

$f_{XY}(x,y) = \frac{1}{4}$ if $(x,y) \in S$.

$f_{XY}(x,y) = 0$ otherwise.



- The marginal PDF of Y: $f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\, dx$
- The *conditional probability density function (PDF)* of *X* given *Y*:

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

$$P(x \le X \le x+\delta \mid y \le Y \le y+\delta) \approx f_{X|Y}(x \mid y)\delta$$

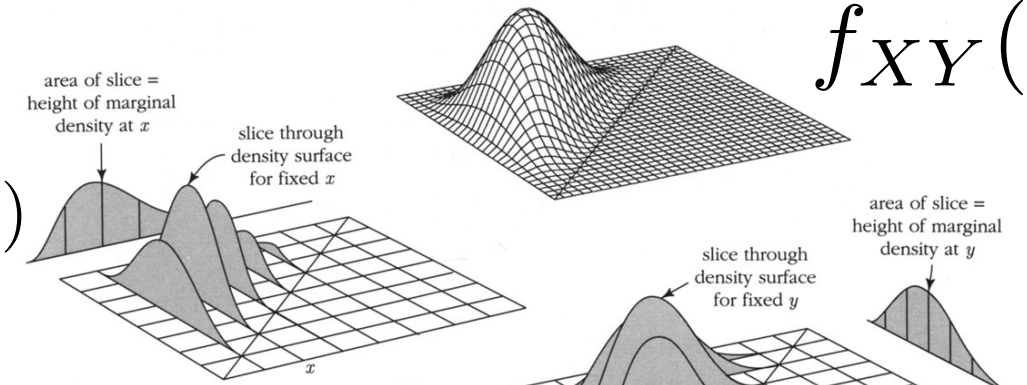
Conditioning on Continuous Observations

$$
\begin{aligned}
F(X \mid Y = y) &= \int_{-\infty}^{X} f_{X|Y=y}(X) = Pr(X \leq x \mid Y = y) \\
&= \lim_{dy \to 0} Pr(X \leq x \mid y \leq Y \leq y + dy) \\
&= \lim_{dy \to 0} \frac{Pr(X \leq x \text{ AND } y \leq Y \leq y + dy)}{Pr(y \leq Y \leq y + dy)} \\
&= \lim_{dy \to 0} \frac{\frac{1}{dy} \int_{y}^{y+dy} \int_{-\infty}^{x} f_{X,Y}(x, y) dx dy}{\frac{1}{dy} \int_{y}^{y+dy} f_Y(y) dy} \\
&= \frac{\int_{-\infty}^{x} f_{X,Y}(x, y) dx}{f_Y(y)} = \int_{-\infty}^{x} \frac{f_{X,Y}(x, y)}{f_Y(y)} dx
\end{aligned}
$$

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

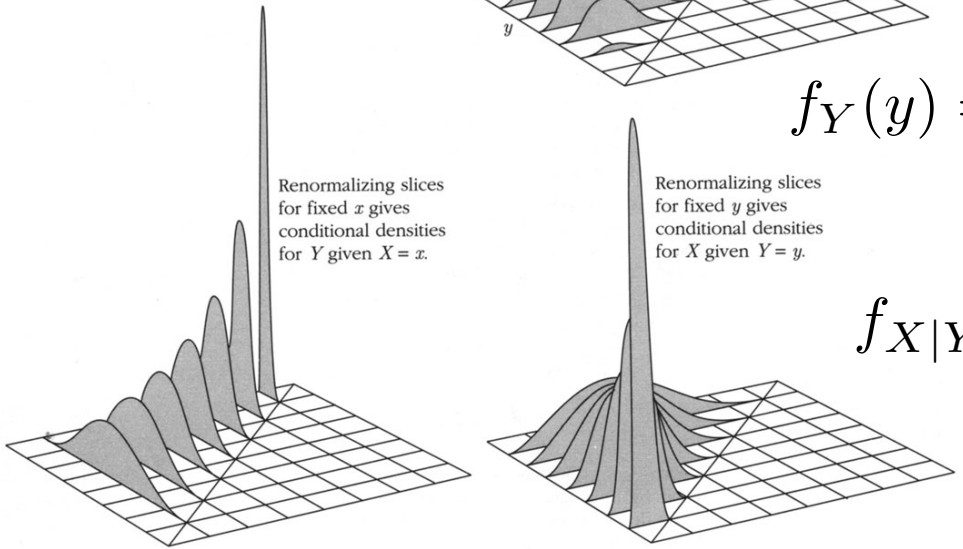
Joint, Marginal, & Conditional Distributions

$f_{XY}(x, y)$



$f_X(x)$

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



$f_{Y|X}(y \mid x)$

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

Pitman's Probability, 1999

Continuous Inference from Discrete Data

y= concentration of virus

$X = \{0, 1\}$ - tested negative/positive

Test's specifications give $P(X \mid Y)$.

$$P(Y \leq y \mid X = x) = \frac{\int_{Y \leq y} f(y) P(X{=}x \mid Y{=}y) dy}{P(X{=}x)}$$

$$f(y \mid X = x) = \frac{f(y) P(X{=}x \mid Y{=}y)}{P(X{=}x)}$$

Variants of Bayes Rule

Infer discrete X from discrete Y:

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_{Y|X}(y \mid x)}{p_Y(y)}$$

$$p_Y(y) = \sum_x p_X(x)p_{Y|X}(y \mid x)$$

Example:

- $X = 1, 0$: airplane present/not present
- $Y = 1, 0$: something did/did not register on radar

Infer continuous X from continuous Y:

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

$$f_Y(y) = \int_x f_X(x)f_{Y|X}(y \mid x)\,dx$$

Example: X: some signal; "prior" $f_X(x)$
Y: noisy version of X
$f_{Y|X}(y \mid x)$: model of the noise

Infer discrete X from continuous Y:

$$p_{X|Y}(x \mid y) = \frac{p_X(x)f_{Y|X}(y \mid x)}{f_Y(y)}$$

$$f_Y(y) = \sum_x p_X(x)f_{Y|X}(y \mid x)$$

Example:

- X: a discrete signal; "prior" $p_X(x)$
- Y: noisy version of X
- $f_{Y|X}(y \mid x)$: continuous noise model

Infer continuous X from discrete Y:

$$f_{X|Y}(x \mid y) = \frac{f_X(x)p_{Y|X}(y \mid x)}{p_Y(y)}$$

$$p_Y(y) = \int_x f_X(x)p_{Y|X}(y \mid x)\,dx$$

Example:

- X: a continuous signal; "prior" $f_X(x)$ (e.g., intensity of light beam);
- Y: discrete r.v. affected by X (e.g., photon count)
- $p_{Y|X}(y \mid x)$: model of the discrete r.v.

Example: Hard Drive Lifetimes

- Suppose 90% of hard drives in some laptop computer model have exponentially distributed lifetime param θ_0

$$f_{Y|X}(y \mid 0) = \theta_0 e^{-\theta_0 y} \qquad p_X(0) = 0.9$$

- However, 10% of hard drives have a manufacturing defect that gives them a shorter lifetime $\theta_1 > \theta_0$

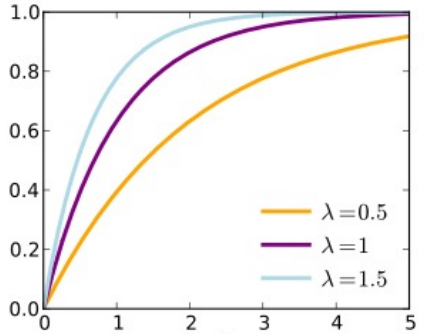
$$f_{Y|X}(y \mid 1) = \theta_1 e^{-\theta_1 y} \qquad p_X(1) = 0.1$$

- Recall mean of exponential distribution:

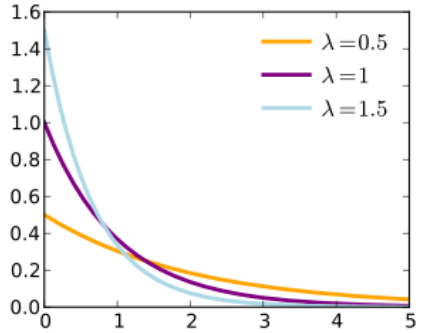
$$E[Y \mid X = 0] = \frac{1}{\theta_0} > \frac{1}{\theta_1} = E[Y \mid X = 1]$$

$$E[X] = \int_0^\infty x\theta e^{-\theta x} dx = [-xe^{-\theta x} - \tfrac{1}{\theta} e^{-\theta x}]_{x=0}^{\infty} = \tfrac{1}{\theta}$$

Exponential Distributions:



$$F_Y(y) = 1 - e^{-\lambda y}$$



$$f_Y(y) = \lambda e^{-\lambda y}$$

Example: Hard Drive Lifetimes

- Suppose 90% of hard drives in some laptop computer model have exponentially distributed lifetime param θ_0

$$f_{Y|X}(y \mid 0) = \theta_0 e^{-\theta_0 y} \qquad p_X(0) = 0.9$$

- However, 10% of hard drives have a manufacturing defect that gives them a shorter lifetime $\theta_1 > \theta_0$

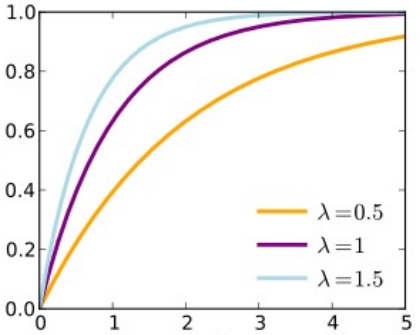
$$f_{Y|X}(y \mid 1) = \theta_1 e^{-\theta_1 y} \qquad p_X(1) = 0.1$$

- If your hard drive has *operated for t seconds and has not yet failed*, what is the probability it is defective?

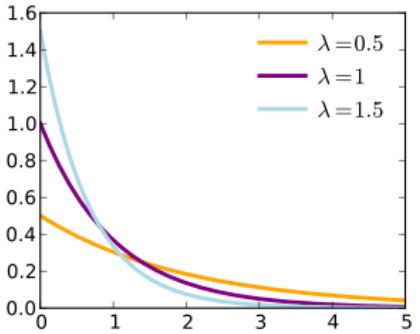
$$P(X = 1 \mid Y > t) = \frac{P(Y > t \mid X = 1)P(X = 1)}{P(Y > t)}$$

$$= \frac{0.1e^{-\theta_1 t}}{0.1e^{-\theta_1 t} + 0.9e^{-\theta_0 t}}$$

Exponential Distributions:



$$F_Y(y) = 1 - e^{-\lambda y}$$



$$f_Y(y) = \lambda e^{-\lambda y}$$

Example: Hard Drive Lifetimes

- Suppose 90% of hard drives in some laptop computer model have exponentially distributed lifetime param θ_0

$$f_{Y|X}(y \mid 0) = \theta_0 e^{-\theta_0 y} \qquad p_X(0) = 0.9$$

- However, 10% of hard drives have a manufacturing defect that gives them a shorter lifetime $\theta_1 > \theta_0$

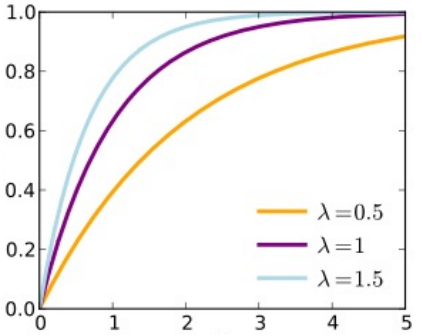
$$f_{Y|X}(y \mid 1) = \theta_1 e^{-\theta_1 y} \qquad p_X(1) = 0.1$$

- If your hard drive *fails after exactly t seconds of operation*, what is the probability it is defective?

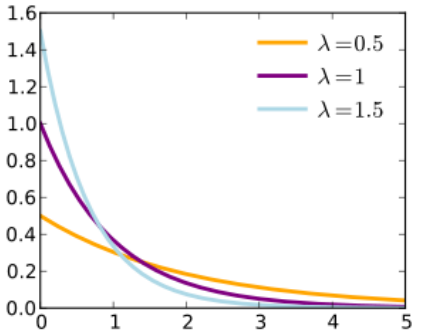
$$P(X = 1 \mid Y = t) = \frac{f_{Y|X}(y \mid 1) p_X(1)}{f_Y(y)}$$

$$= \frac{0.1\theta_1 e^{-\theta_1 t}}{0.1\theta_1 e^{-\theta_1 t} + 0.9\theta_0 e^{-\theta_0 t}}$$

Exponential Distributions:



$$F_Y(y) = 1 - e^{-\theta y}$$

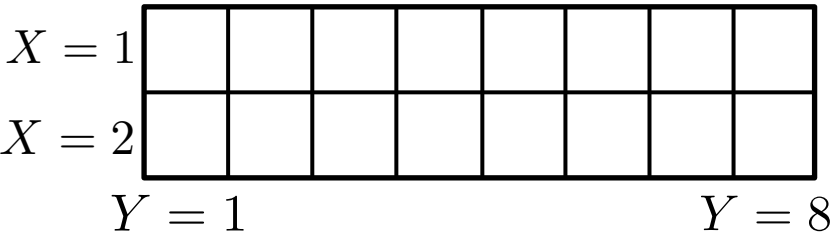


$$f_Y(y) = \theta e^{-\theta y}$$

CS145: Lecture 11 Outline

- Joint Distribution
- Marginal and Conditional Probability Densities
- Conditional Expectations

Joint Probability Distributions



In this example, N=2 and M=8, and the joint PMF is a 2x8 matrix.

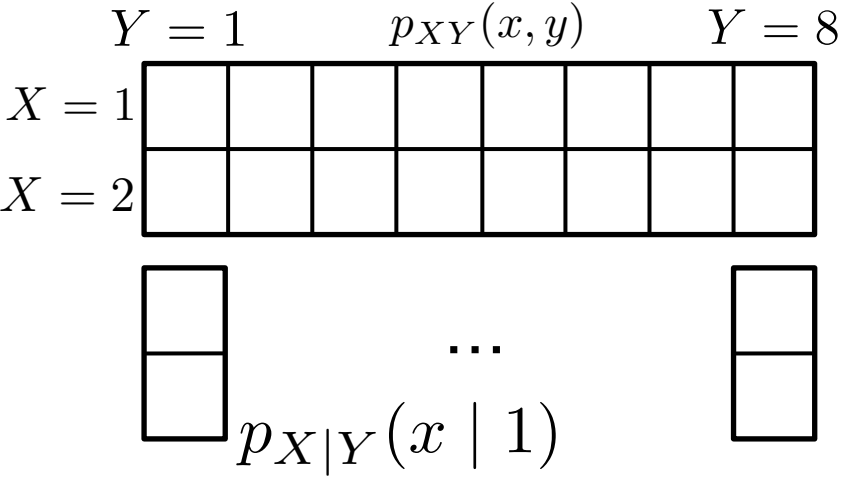
- Consider two random variables *X, Y.*
  Suppose range of X is size N, range of Y is size M.
- The *joint probability mass function* or *joint distribution* of two variables:

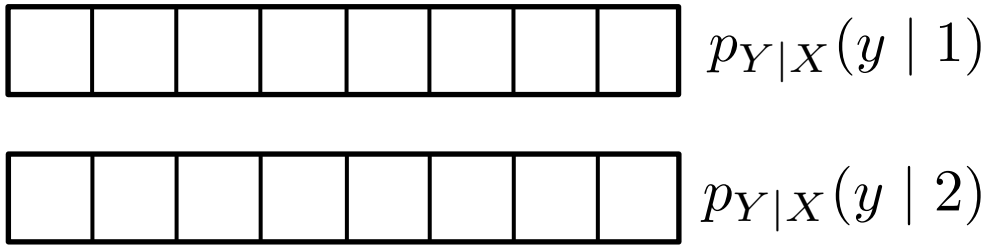
$$p_{XY}(x, y) = P(X = x \text{ and } Y = y)$$

$$p_{XY}(x, y) \geq 0, \qquad \sum_x \sum_y p_{XY}(x, y) = 1.$$

- The joint distribution is uniquely specified by *NM-1* numbers

Conditional Probability Distributions





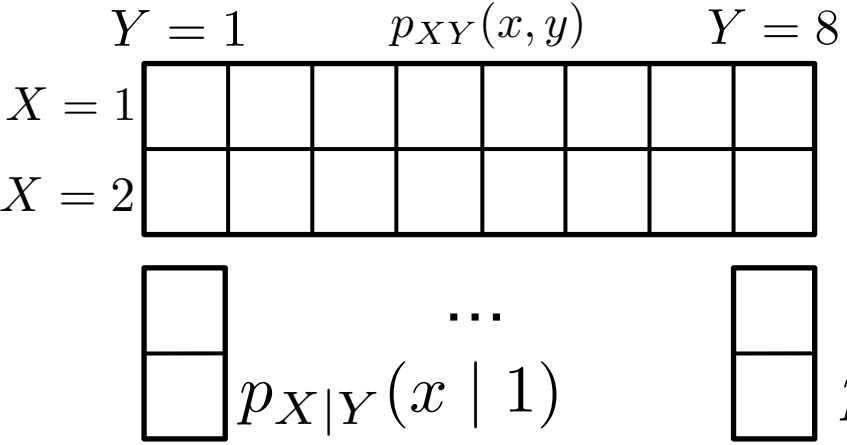
- By the definition of conditional probability:

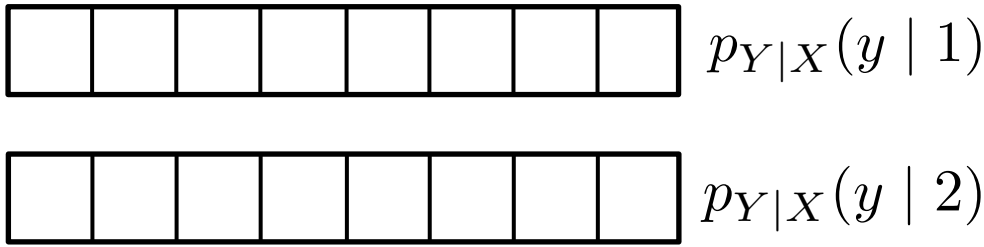
$$P(X = x \mid Y = y) = \frac{P(X = x \text{ and } Y = y)}{P(Y = y)}$$

- The *conditional probability mass function* is then:

$$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{XY}(x, y)}{\sum_{x'} p_{XY}(x', y)}$$

Conditional Expectation





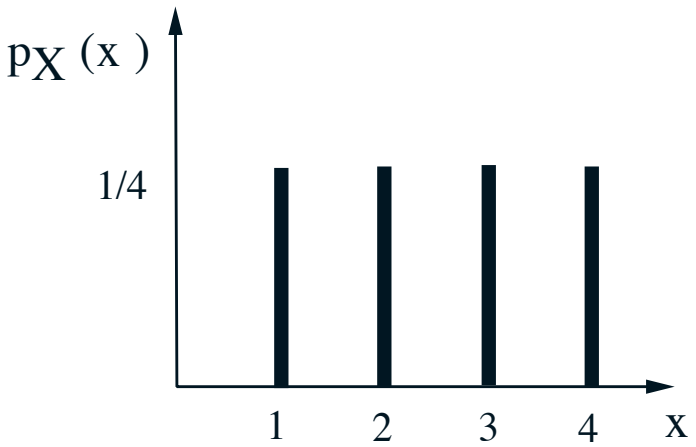
$$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{XY}(x, y)}{\sum_{x'} p_{XY}(x', y)}$$

- Given that I observe *Y=y*, the *conditional expectation* of *X* equals

$$E[X \mid Y = y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x \mid y)$$

- If *X* and *Y* are not independent, observing *Y=y* may change the mean of *X*

Conditional Expectation: Example



$$E[X] = 2.5$$

Given $Y = \{X \geq 2\}$ is observed,

$$p_{X|Y}(x \mid y) =$$

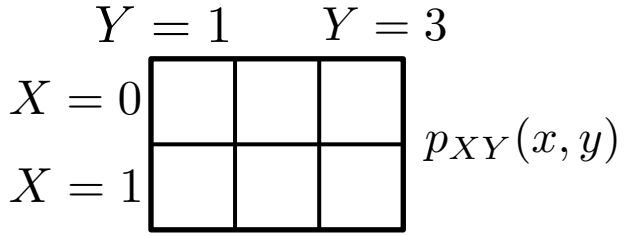
$$E[X \mid Y] = 3$$

- Given that I observe *Y=y*, the *conditional expectation* of *X* equals

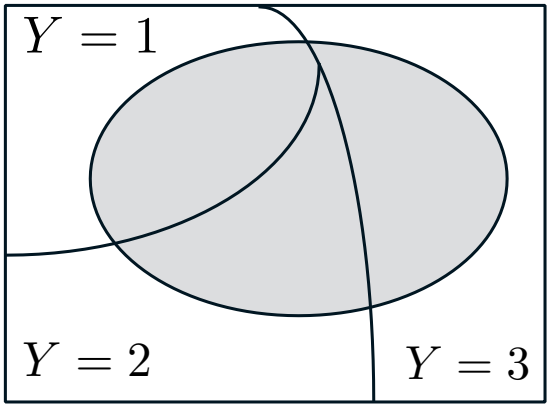
$$E[X \mid Y = y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x \mid y)$$

- If *X* and *Y* are not independent, observing *Y=y* may change the mean of *X*

Total Expectation Theorem



Shaded where X=1,
Unshaded where X=0.



$$p_{X|Y}(x \mid y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{XY}(x, y)}{\sum_{x'} p_{XY}(x', y)}$$

- Applying the definitions of joint, marginal, and conditional distributions:

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{XY}(x, y) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x \mid y) p_Y(y)$$

$$E[X] = \sum_{y \in \mathcal{Y}} p_Y(y) E[X \mid Y = y]$$

Mean is a weighted average of (possibly simpler) conditional means.

Conditional Means are Random Variables

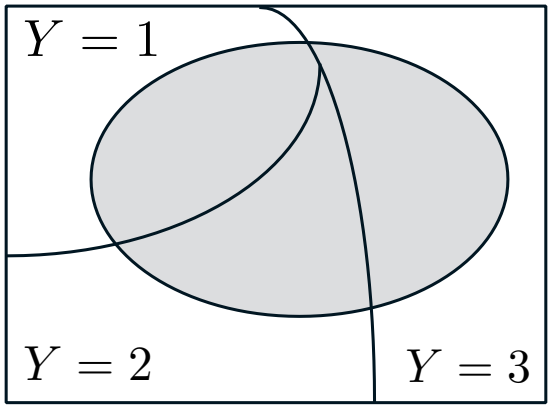
- The quantity *E[X | Y]* is a *random variable* *g(Y)* that takes on the value *g(y)=E[X | Y=y]* when *Y=y* is observed
- This random variable *E[X | Y]* has an expected value, which equals

$$E[E[X|Y]] = \sum_y p_Y(y) E[X \mid Y = y] = \sum_y \sum_x x p_Y(y) p_{X|Y}(x \mid y) = E[X]$$

- This is called the *Law of Iterated Expectations*

$$E[X] = \sum_{y \in \mathcal{Y}} p_Y(y) E[X \mid Y = y]$$

$$E[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y)$$



Mean is a weighted average of (possibly simpler) conditional means.

Example: Class Scores Across Sections

X = average homework score of students divided into 2 sections:

$y = 1$ (10 students); $y = 2$ (20 students)

$$y = 1 : \frac{1}{10} \sum_{i=1}^{10} x_i = 90 \qquad y = 2 : \frac{1}{20} \sum_{i=11}^{30} x_i = 60$$

$$\mathbf{E}[X] = \frac{1}{30} \sum_{i=1}^{30} x_i = \frac{90 \cdot 10 + 60 \cdot 20}{30} = 70$$

$$\mathbf{E}[X \mid Y = 1] = 90, \quad \mathbf{E}[X \mid Y = 2] = 60$$

$$\mathbf{E}[X \mid Y] = \begin{cases} 90, & \text{w.p. } 1/3 \\ 60, & \text{w.p. } 2/3 \end{cases}$$

$$\mathbf{E}[\mathbf{E}[X \mid Y]] = \tfrac{1}{3} \cdot 90 + \tfrac{2}{3} \cdot 60 = 70 = \mathbf{E}[X]$$

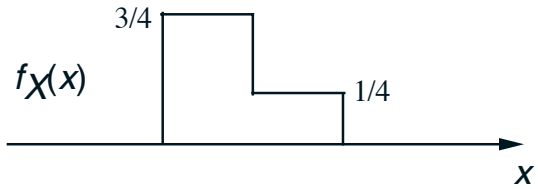
Continuous Iterated Expectations

- The *Law of Iterated Expectations* also applies to continuous variables:

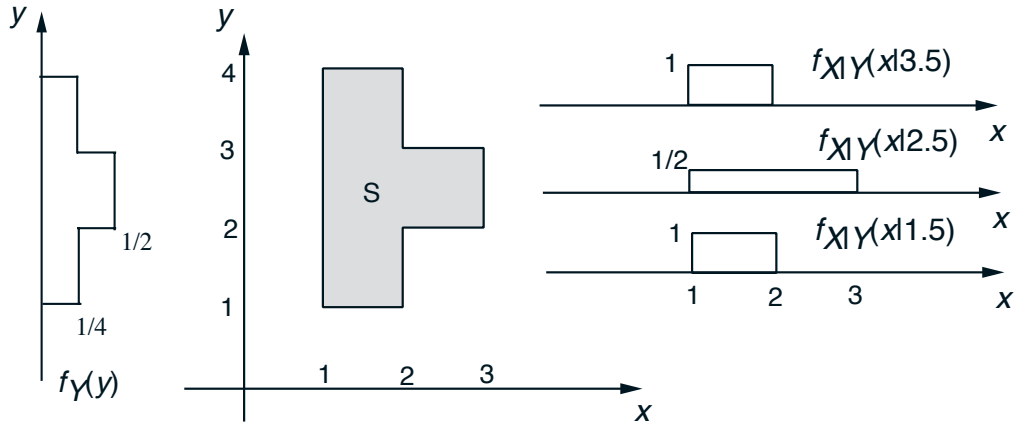
$$E[X \mid Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x \mid y)\, dx$$

$$E[X] = E[E[X \mid Y]] = \int_{-\infty}^{+\infty} E[X \mid Y = y] f_Y(y)\, dy$$

- Proof is as before, but replacing PMFs with PDFs, and sums with integrals

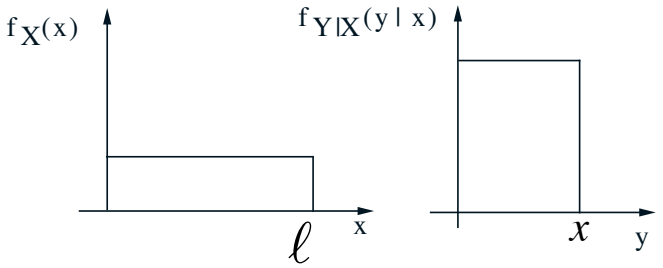


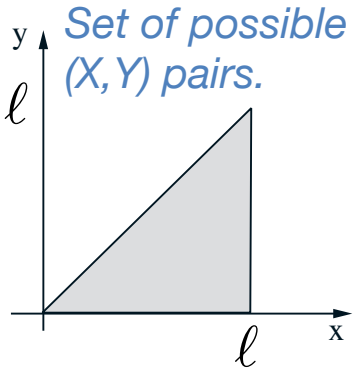
$$E[X] = \frac{3}{4} \cdot 1.5 + \frac{1}{4} \cdot 2.5 = 1.75$$



Example: Stick-Breaking

- Break a stick of length ℓ twice: break at X, uniform in $[0, \ell]$; break again at Y, uniform in $[0, X]$.
- $f_X(x) = 1/\ell$
- $f_{Y|X}(y \mid x) = 1/x$ for $0 \le y \le x$



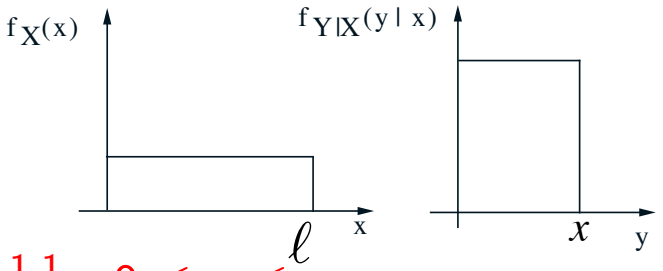


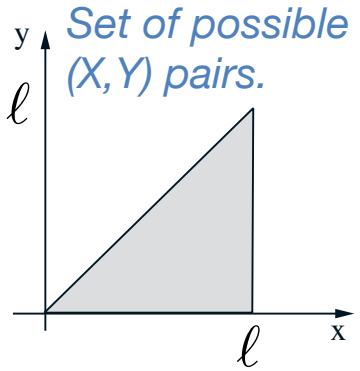
$$E[Y \mid X = x] = \int_0^x y f_{Y|X}(y \mid X = x)dy = \int_0^x \frac{y}{x}dy = \frac{x}{2}$$

$$E[Y] = E[E[Y \mid X]] = \int_0^\ell \frac{x}{2}\frac{1}{\ell}dx = \frac{\ell}{4}$$

Example: Stick-Breaking

- Break a stick of length ℓ twice: break at X, uniform in $[0, \ell]$; break again at Y, uniform in $[0, X]$.
- $f_X(x) = 1/\ell$
- $f_{Y|X}(y \mid x) = 1/x$ for $0 \le y \le x$

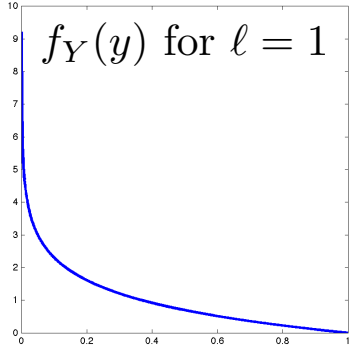




$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x) f_X(x) = \begin{cases} \frac{1}{x}\frac{1}{\ell} & 0 \le y \le x \\ 0 & \text{otherwise.} \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_y^{\ell} \frac{1}{x}\frac{1}{\ell}\,dx = \frac{1}{\ell} \ln \frac{\ell}{y}$$

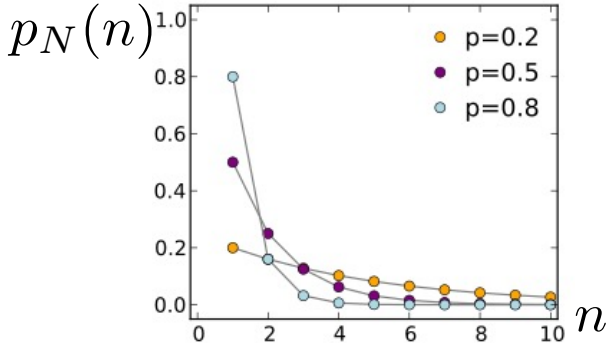
$$E[Y] = \int_0^{\ell} \frac{u}{\ell} \ln \frac{\ell}{u}\,du = \left(\frac{u^2}{2\ell} \ln \frac{\ell}{u} + \frac{u^2}{4\ell}\right)\Big|_0^{\ell} = \frac{\ell}{4}$$



Sums of Random Numbers of Variables

How much money Y do we spend shopping at a random number of stores N?

- N: number of stores visited
  (N is a nonnegative integer r.v.)
- X_i: money spent in store i
  - X_i assumed i.i.d.
  - independent of N
- Let $Y = X_1 + \cdots + X_N$



$$\begin{aligned}\mathbf{E}[Y \mid N = n] &= \mathbf{E}[X_1 + X_2 + \cdots + X_n \mid N = n] \\ &= \mathbf{E}[X_1 + X_2 + \cdots + X_n] \\ &= \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] \\ &= n\,\mathbf{E}[X]\end{aligned}$$

$$\begin{aligned}\mathbf{E}[Y] &= \mathbf{E}[\mathbf{E}[Y \mid N]] \\ &= \mathbf{E}[N\,\mathbf{E}[X]] \\ &= \mathbf{E}[N]\,\mathbf{E}[X]\end{aligned}$$

- $\mathbf{E}[Y \mid N] = N\,\mathbf{E}[X]$