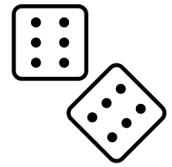
Expectation, Two Random Variables

Contents

- expectation
- mean and variance
- joint, marginal, and conditional pmfs
- joint, marginal and conditional cdfs

Expectation

Introduction to Expectation



- discrete random variable (pmf): $X \sim p(x)$

- expected value:

$$E[X] = \sum_i x_i \; p(x_i) \qquad\qquad E[X] = \sum_{x \in \mathcal{X}} x \; p(x)$$

- example: X is an unfair dice:

$$p(x) = \begin{cases} 0.1 & \text{if } x = 1 \\ 0.1 & \text{if } x = 2 \\ \dots & \\ 0.1 & \text{if } x = 5 \\ 0.5 & \text{if } x = 6 \end{cases}$$

$$E[X] = 1 \times 0.1 + 2 \times 0.1 + \cdots + 5 \times 0.1 + 6 \times 0.5 = 4.5$$

- what's the average of n samples of the random variable?

366666261216653666616256665663535622256265661656664666646164356625632...

$$\frac{1}{100} \sum_{i=1}^{100} X_i \approx 4.62 \qquad\qquad \frac{1}{1000} \sum_{i=1}^{1000} X_i \approx 4.56$$

Expectation of a function of a random variable $Z = g(X)$

For a discrete RV X, we show that

$$\mathbb{E}[g(X)] = \sum_i g(x_i)\mathbb{P}(X = x_i).$$

Expectation of a function of a random variable $Z = g(X)$

First observe that

$$\mathsf{P}(Z = z_k) = \sum_{i:g(x_i)=z_k} \mathsf{P}(X = x_i).$$

Then

$$\begin{aligned}
\mathsf{E}[Z] &= \sum_k z_k \mathsf{P}(Z = z_k) \\
&= \sum_k z_k \left(\sum_{i:g(x_i)=z_k} \mathsf{P}(X = x_i) \right) \\
&= \sum_k \left(\sum_{i:g(x_i)=z_k} z_k \, \mathsf{P}(X = x_i) \right) \\
&= \sum_k \left(\sum_{i:g(x_i)=z_k} g(x_i) \, \mathsf{P}(X = x_i) \right) \\
&= \sum_i g(x_i) \, \mathsf{P}(X = x_i),
\end{aligned}$$

Expectation of a function of a random variable $Y = g(X)$

For a continuous RV X with density $f(x)$, it can be shown that

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx.$$

Expectation of a function of a random variable $Y = g(X)$

Here we derive the formula when $Y = g(X)$ and g takes only finitely many values, say $y_1, y_2, \ldots y_n$.

First write

$$\mathsf{P}(Y = y_j) = \mathsf{P}(g(X) = y_j) = \int_{\{x:g(x)=y_j\}} f(x)\,dx.$$

Then

$$\begin{aligned}
\mathsf{E}[Y] &= \sum_j y_j \mathsf{P}(Y = y_j) \\
&= \sum_j y_j \int_{\{x:g(x)=y_j\}} f(x)\,dx \\
&= \sum_j \int_{\{x:g(x)=y_j\}} y_j f(x)\,dx \\
&= \sum_j \int_{\{x:g(x)=y_j\}} g(x) f(x)\,dx \\
&= \int_{-\infty}^{\infty} g(x) f(x)\,dx,
\end{aligned}$$

Expectation

- discrete random variable (pmf): $X \sim p(x)$

- the *expectation* or *expected value* of $g(X)$ is defined as

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

- continuous random variable (pdf): $X \sim f(x)$

- the *expectation* or *expected value* of $g(X)$ is defined as

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Example: Entropy

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

- you run an experiment to see if sun rises in the east tomorrow

$X = 0$ is the event the sun rises in the east

$X = 1$ something else happens

$$X = \begin{cases} 0 & \text{with probability } 0.9999 \\ 1 & \text{with probability } 0.0001 \end{cases}$$

- sure enough, the sun rises in the east tomorrow

$$\log_2 \left(\frac{1}{0.9999} \right) \approx 0.0001 \text{ bits}$$

- imagine it doesn't rise in the east

$$\log_2 \left(\frac{1}{0.0001} \right) \approx 13.2 \text{ bits}$$

- how much information did we learn about the world *on average*?

$$H(X) = E\left[\log_2 \left(\frac{1}{p(X)} \right) \right]$$

Mean, moments and variance

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

- the *mean* (or first moment) of a random variable X is defined as

$$E[X] \qquad (\text{i.e, } g(X) = X)$$

- discrete random variable (pmf):

$$E[X] = \sum_{x \in \mathcal{X}} xp(x)$$

- continuous random variable (pdf):

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

- the *second moment* of a random variable X is defined as

$$E[X^2] \qquad (\text{i.e, } g(X) = X^2)$$

- the *variance* of a random variable X is defined as

$$\text{var}(X) = E[(X - E[X])^2]$$

- the *standard deviation* of a random variable X is the square root of the variance

Properties of Expectation

- $E[c] = c$ for any constant c
- linearity: $E[ag_1(X) + bg_2(X)] = E[g_1(X)] + E[g_2(X)]$ for any constants a, b
- transforms: if $Y = g(X)$, then $E[Y] = E[g(X)]$
  - note: usually easier to compute $E[g(X)]$ vs. computing $p(y)$ and then $E[Y]$
- expectation can be infinite. example $f(x) = \frac{1}{x^2}$ for $x \geq 1$

Bounding Probabilities Using Expectation

- Markov:
- if X is a non-negative RV, i.e, $X \geq 0$, then for any $\alpha \geq 0$

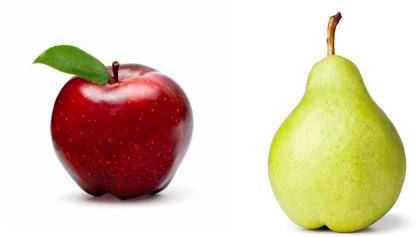
$$\mathbb{P}(X \geq \alpha E[X]) \leq \frac{1}{\alpha}$$

- Chebyshev Inequality:
- Let X be a RV with $E[X]$ and $\text{Var}(X) = \sigma^2$, then for any $\alpha > 1$

$$\mathbb{P}(|X - E[X]| \geq \alpha\sigma) \leq \frac{1}{\alpha^2}$$

Two Random Variables

Apple/Pear classifier



$y = -1$ if pear, $y = 1$ if apple, 0 other fruit

$x \in \{0.5, 1, 2\}$ is top to bottom ratio

$\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{m}$

- model features x and label y as joint random variables X and Y

- X and Y are completely described by their *joint* pmf

$$p(x, y) = \mathbb{P}(X = x, Y = y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

$p(x,y)$		x		
		0.5	1	2
y	-1	$\frac{1}{8}$	$\frac{1}{8}$	0
	0	0	$\frac{1}{2}$	0
	1	0	$\frac{1}{8}$	$\frac{1}{8}$

- usual ML assumption: training/test samples are *i.i.d.* from some joint distribution.

$$(X, Y) \overset{i.i.d.}{\sim} p(x, y)$$

Joint pmfs

- discrete X and Y are completely described by their *joint* pmf

$$p(x,y) = \mathbb{P}(X = x, Y = y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

$$\sum_{y \in Y} \sum_{x \in X} p(x,y) = 1$$

$p(x,y)$	x = 0.5	1	2
y = −1	$\frac{1}{8}$	$\frac{1}{8}$	0
0	0	$\frac{1}{2}$	0
1	0	$\frac{1}{8}$	$\frac{1}{8}$

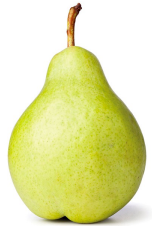
- given a joint pmf, $p(x)$ and $p(y)$ are called *marginals*

$$p(x) = \sum_{y \in \mathcal{Y}} p(x,y) \text{ for all } x \in \mathcal{X}$$

- X, Y are independent if and only if

$$p(x,y) = p(x)p(y) \text{ for all } x \in \mathcal{X},\ y \in \mathcal{Y}$$

Conditional pmfs



x is top to bottom ratio

$y = -1$ if pear, $y = 1$ if apple, 0 other fruit

- the *conditional pmf* of X given Y is defined as

$$p(x|y) = \frac{p(x,y)}{p(y)} \text{ for } p(y) \neq 0 \text{ and } x \in X$$

- example:

y \ x	0.5	1	2
-1	$\frac{1}{8}$	$\frac{1}{8}$	0
0	0	$\frac{1}{2}$	0
1	0	$\frac{1}{8}$	$\frac{1}{8}$

- Bayes rule:

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y' \in \mathcal{Y}} p(x,y')} = \frac{p(x|y)p(y)}{\sum_{y' \in \mathcal{Y}} p(x|y')p(y')}$$

- independence implies:

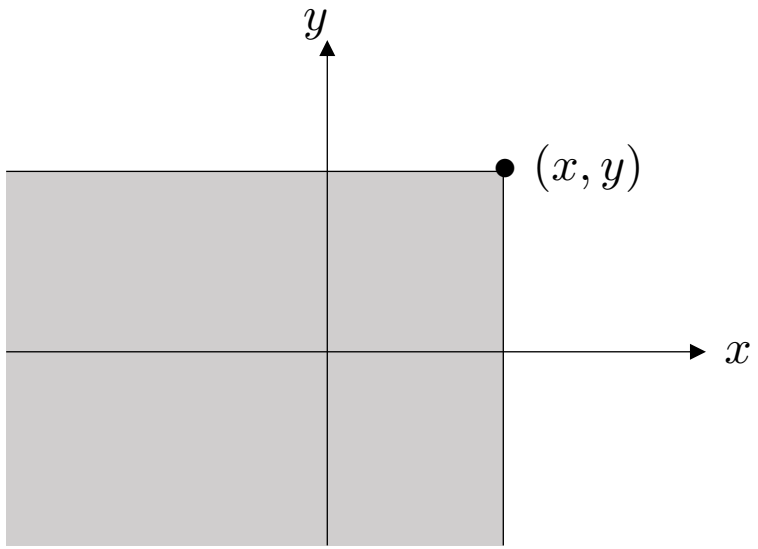
$$p(y|x) = p(y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

Joint cdfs

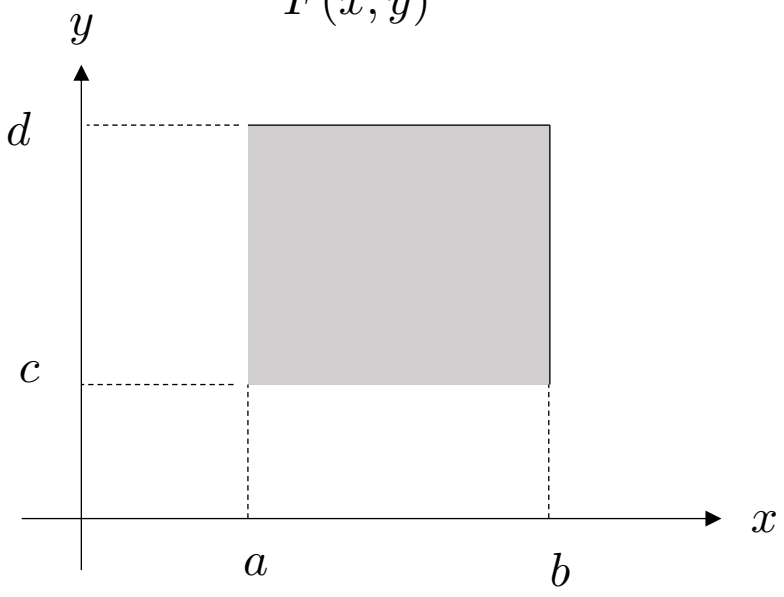
- let X and Y be two random variables
- X and Y are completely described by their *joint* cdf

$$F(x,y) = \mathbb{P}(X \leq x, Y \leq y) \text{ for all } x \in \mathbb{R}, y \in \mathbb{R}$$

  - properties:
    - $F(x,y) \geq 0$
    - nondecreasing in both x, y
    - $\lim_{x\to\infty} F(x,y) = F(y)$
    - independent $\Leftrightarrow F(x,y) = F(x)F(y)$



$F(x,y)$



- example $\mathbb{P}(a < X \leq b, c < Y \leq d)$

$$= F(b,d) - F(a,d) - F(b,c) + F(a,c)$$

Joint pdfs

- X and Y are completely described by their *joint* cdf

$$F(x,y) = \mathbb{P}(X \leq x, Y \leq y) \text{ for all } x \in \mathbb{R}, y \in \mathbb{R}$$

- we can define the joint pdf (if it exists) as a function $f(x,y)$ that satisfies:

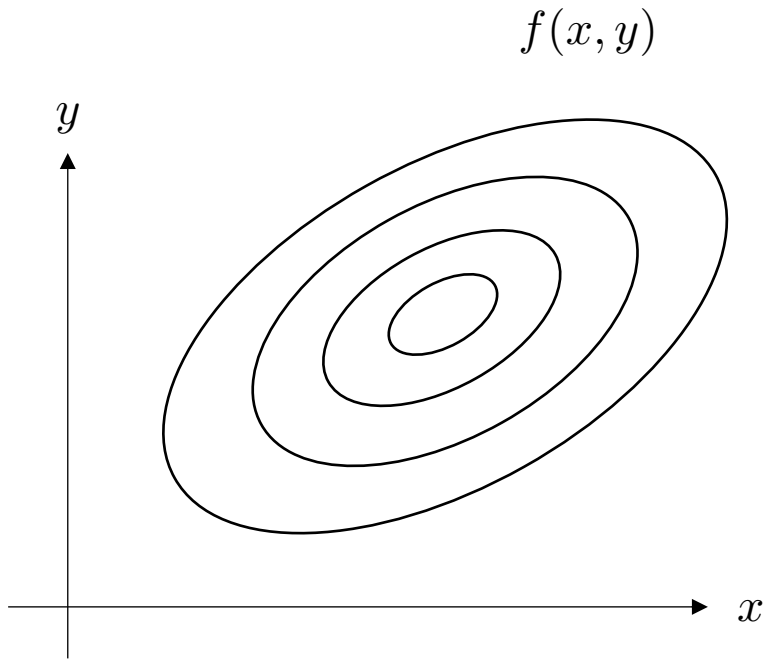
$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v) du dv \text{ for all } x, y \in \mathbb{R}$$

- if X and Y are ***jointly* continuous**, then they are completely described by their *joint* pdf

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} \qquad f(x,y) = \lim_{\Delta x \Delta y \to 0} \frac{\mathbb{P}(x < X \leq x + \Delta x, y < Y \leq y + \Delta y)}{\Delta x \Delta y}$$

Properties of joint pdfs

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



- properties:
  - $f(x,y) \geq 0$
  - $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$
  - $\mathbb{P}((X,Y) \in A) = \int_{(x,y)\in A} f(x,y)dxdy$
- the *marginal pdf* of X can be computed as:

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

- X, Y are independent if and only if

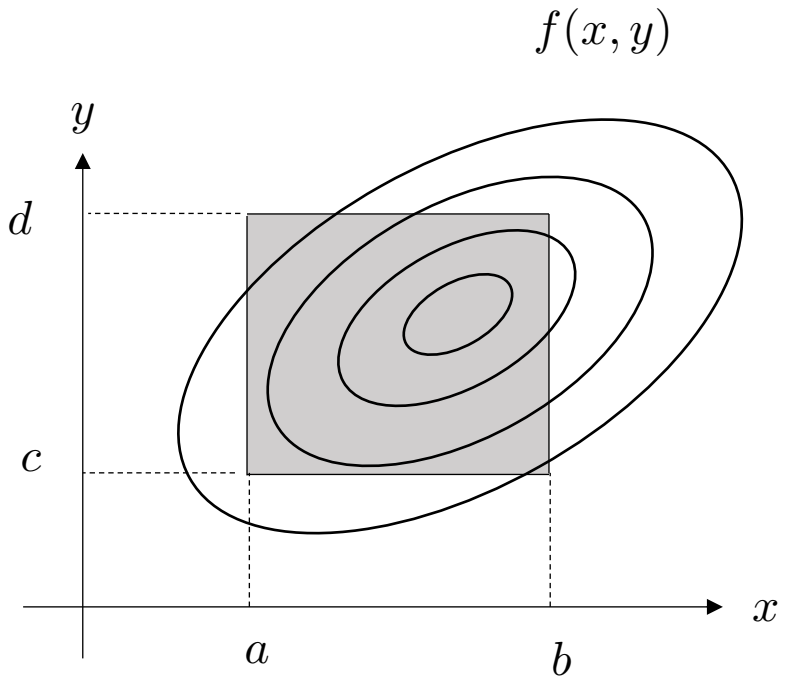
$$f(x,y) = f(x)f(y) \text{ for all } x, y$$

Example

$$\mathbb{P}((X,Y) \in A) = \int_{(x,y)\in A} f(x,y)dxdy$$

$$A = \{(x,y) : a \leq x \leq b, c \leq y \leq d\}$$

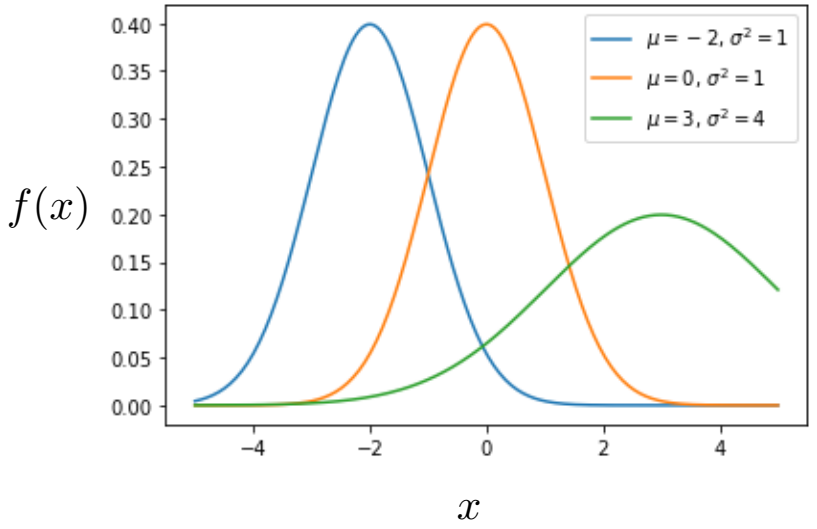
$$\mathbb{P}((X,Y) \in A) = \int_a^b \int_c^d f(x,y)dxdy$$



Example – bivariate Normal

- Gaussian: $X \sim \mathcal{N}(\mu, \sigma^2)$

$$f(x) = \frac{\exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{\sqrt{2\pi}\,\sigma}$$



- Bivariate (jointly) Gaussian:

$$f(x,y) = \frac{\exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}\right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

$$\mu_x := E[X] \qquad \sigma_x^2 := \mathrm{var}(X)$$

$$\mu_y := E[Y] \qquad \sigma_y^2 := \mathrm{var}(Y)$$

$$\rho := \text{correlation}$$