Probability, continued

CMPUT 296: Basics of Machine Learning

§2.2-2.4

Recap

- Probabilities are a means of **quantifying uncertainty**
- A probability distribution is defined on a measurable space consisting of a **sample space** and an **event space**.
- **Discrete** sample spaces (and random variables) are defined in terms of **probability mass functions** (PMFs)
- **Continuous** sample spaces (and random variables) are defined in terms of **probability density functions** (PDFs)

Logistics

Now available on eClass:

- Videos and slides for last week
- Discussion forum!
- Thought Question 1 (due **Thursday, September 17**)
- Assignment 1 (due **Thursday, September 24**)

TA office hours:

- Ehsan: **Wednesdays 3-4pm**
  - or 3-5pm on "tutorial" weeks
- Liam: **Fridays 11am-12pm**

Outline

1. Recap & Logistics
2. Random Variables
3. Multiple Random Variables
4. Independence
5. Expectations and Moments

Random Variables

Random variables are a way of reasoning about a complicated underlying probability space in a more straightforward way.

Example: Suppose we observe both a die's number, and where it lands.

$$\Omega = \{(left, 1), (right, 1), (left, 2), (right, 2), \ldots, (right, 6)\}$$

We might want to think about the probability that we get a large number, without thinking about where it landed.

We could ask about $P(X \geq 4)$, where X = number that comes up.

Random Variables, Formally

Given a probability space $(\Omega, \mathscr{E}, P)$, a **random variable** is a function $X : \Omega \to \Omega_X$ (where Ω_X is some other outcome space), satisfying

$$\{\omega \in \Omega \mid X(\omega) \in A\} \in \mathscr{E} \quad \forall A \in B(\Omega_X).$$

It follows that $P_X(A) = P(\{\omega \in \Omega \mid X(\omega) \in A\})$.

Example: Let Ω be a population of people, and $X(\omega)$ = height, and $A = [5'1'', 5'2'']$.

$$P(X \in A) = P(5'1'' \leq X \leq 5'2'') = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

Random Variables and Events

- A Boolean expression involving random variables defines an event:

  E.g., $P(X \geq 4) = P(\{\omega \in \Omega \mid X(\omega) \geq 4\})$

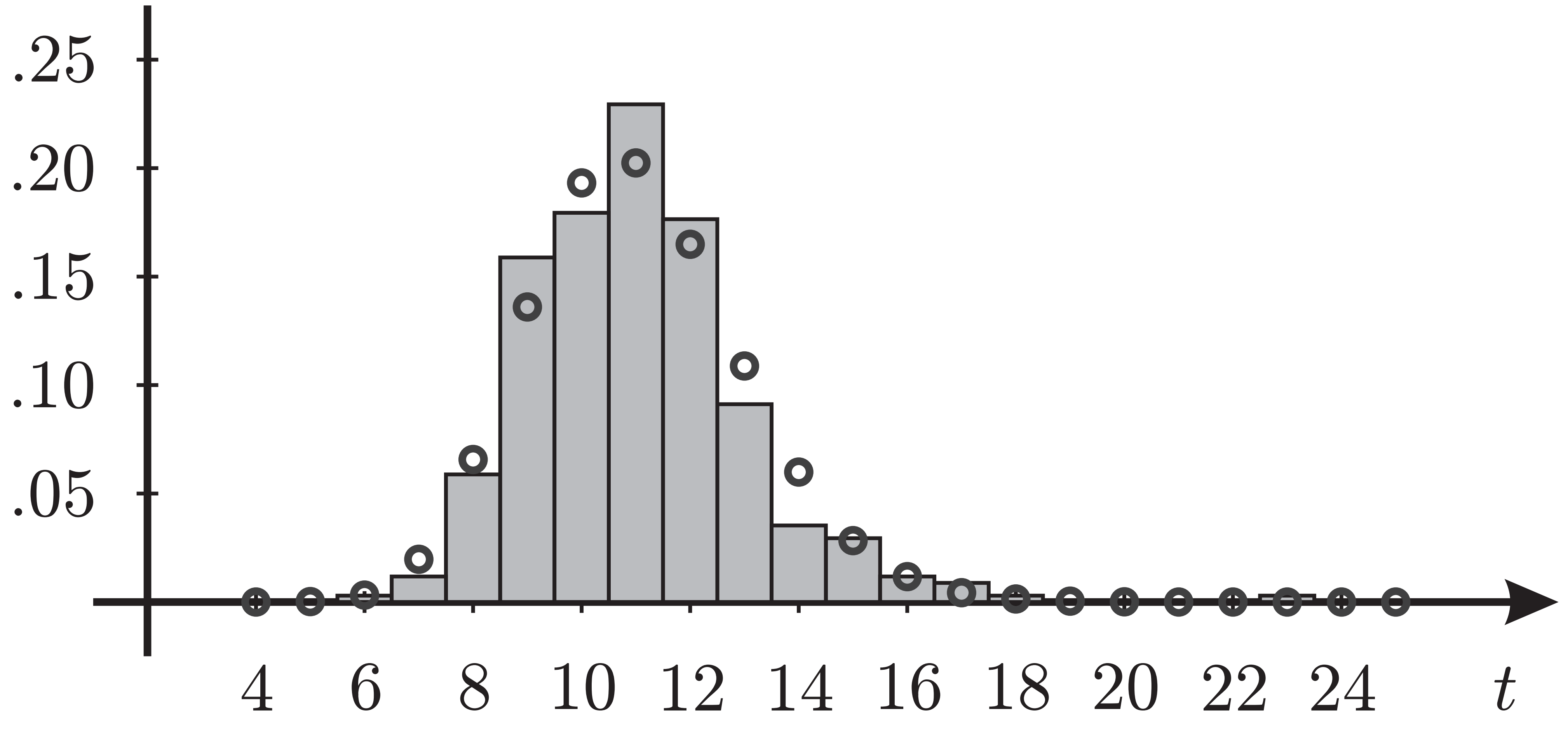
- Similarly, every event can be understood as a Boolean random variable:

$$Y = \begin{cases} 1 & \text{if event } A \text{ occurred} \\ 0 & \text{otherwise.} \end{cases}$$

- From this point onwards, we will exclusively reason in terms of random variables rather than probability spaces.

Example: Histograms

Consider the continuous commuting example again, with observations 12.345 minutes, 11.78213 minutes, etc.



- **Question:** What is the random variable?
- **Question:** How could we turn our observations into a histogram?

What About Multiple Variables?

- So far, we've really been thinking about a single random variable at a time
- Straightforward to define multiple random variables on a single probability space

Example: Suppose we observe both a die's number, and where it lands.

$$\Omega = \{(left,1), (right,1), (left,2), (right,2), \ldots, (right,6)\}$$

$$X(\omega) = \omega_2 = \text{number}$$

$$Y(\omega) = \left\{\begin{cases} 1 & \text{if } \omega_1 = left \\ 0 & \text{otherwise.} \end{cases}\right\} = 1 \text{ if landed on left}$$

$$P(Y = 1) = P(\{\omega \mid Y(\omega) = 1\})$$

$$P(X \geq 4 \wedge Y = 1) = P(\{\omega \mid X(\omega) \geq 4 \wedge Y(\omega) = 1\})$$

Joint Distribution

We typically be model the **interactions** of different random variables.

Joint probability mass function: $p(x, y) = P(X = x, Y = y)$

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

Example: $\mathcal{X} = \{0,1\}$ (young, old) and $\mathcal{Y} = \{0,1\}$ (no arthritis, arthritis)

	Y=0	**Y=1**
X=0	P(X=0,Y=0) = 1/2	P(X=0, Y=1) = 1/100
X=1	P(X=1, Y=0) = 1/10	P(X=1, Y=1) = 39/100

Questions About Multiple Variables

Example: $\mathscr{X} = \{0,1\}$ (young, old) and $\mathscr{Y} = \{0,1\}$ (no arthritis, arthritis)

	Y=0	**Y=1**
X=0	P(X=0,Y=0) = 1/2	P(X=0, Y=1) = 1/100
X=1	P(X=1, Y=0) = 1/10	P(X=1, Y=1) = 39/100

- Are these two variables related at all? Or do they change **independently**?
- Given this distribution, can we determine the distribution over just Y? I.e., what is $P(Y = 1)$? (**marginal distribution**)
- If we knew something about one variable, does that tell us something about the distribution over the other? E.g., if I know $X = 0$ (person is young), does that tell me the **conditional probability** $P(Y = 1 \mid X = 1)$? (Prob. that person we know is young has arthritis)

Conditional Distribution

Definition: Conditional probability distribution

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

This same equation will hold for the corresponding PDF or PMF:

$$p(y \mid x) = \frac{p(x, y)}{p(x)}$$

Question: if $p(x, y)$ is small, does that imply that $p(y \mid x)$ is small?

PMFs and PDFs of Many Variables

In general, we can consider a d-dimensional random variable $\overrightarrow{X} = (X_1, \ldots, X_d)$ with vector-valued outcomes $\overrightarrow{x} = (x_1, \ldots, x_d)$, with each x_i chosen from some $\mathcal{X}_i$. Then,

Discrete case:

$p : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \to [0,1]$ is a **(joint) probability mass function** if

$$\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \cdots \sum_{x_d \in \mathcal{X}_d} p(x_1, x_2, \ldots, x_d) = 1$$

Continuous case:

$p : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \to [0,\infty)$ is a **(joint) probability density function** if

$$\int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \cdots \int_{\mathcal{X}_d} p(x_1, x_2, \ldots, x_d)\, dx_1 dx_2 \ldots dx_d = 1$$

Marginal Distributions

A **marginal distribution** is defined for a subset of $\vec{X}$ by summing or integrating out the remaining variables. (We will often say that we are "marginalizing over" or "marginalizing out" the remaining variables).

Discrete case: $p(x_i) = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} \sum_{x_{i+1} \in \mathcal{X}_{i+1}} \cdots \sum_{x_d \in \mathcal{X}_d} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$

Continuous: $p(x_i) = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{i-1}} \int_{\mathcal{X}_{i+1}} \cdots \int_{\mathcal{X}_d} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_d$

Question: Can a marginal distribution also be a joint distribution?

Question: Why p for $p(x_i)$ and $p(x_1, \ldots, x_d)$?

- They can't be the same function, they have different domains!

Are these really the same function?

- **No.** They're not the same function.
- But they are **derived** from the **same joint distribution**.
- So for brevity we will write

$$p(y \mid x) = \frac{p(x, y)}{p(x)}$$

- Even though it would be more precise to write something like

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

  - We tell which function we're talking about from context (i.e., arguments)

Chain Rule

From the definition of conditional probability:

$$
\begin{aligned}
p(y \mid x) &= \frac{p(x, y)}{p(x)} \\
\iff p(y \mid x){\color{blue}p(x)} &= \frac{p(x, y)}{p(x)}{\color{blue}p(x)} \\
\iff p(y \mid x)p(x) &= p(x, y)
\end{aligned}
$$

This is called the **Chain Rule**.

Multiple Variable Chain Rule

The chain rule generalizes to multiple variables:

$$p(x, y, z) = p(x, y \mid z)p(z) = p(x \mid y, z)\underbrace{p(y \mid z)p(z)}_{p(y,z)}$$

Definition: **Chain rule**

$$p(x_1, \ldots, x_d) = p(x_d)\prod_{i=1}^{d-1} p(x_i \mid x_{i+1}, \ldots x_d)$$

$$= p(x_1)\prod_{i=2}^{d} p(x_i \mid x_i, \ldots x_{i-1})$$

Bayes' Rule

From the chain rule, we have:

$$p(x, y) = p(y \mid x)p(x)$$
$$= p(x \mid y)p(y)$$

- Often, $p(x \mid y)$ is easier to compute than $p(y \mid x)$
  - e.g., where x is **features** and y is **label**

Definition: Bayes' rule

$$p(y \mid x) = \frac{p(x \mid y)\,p(y)}{p(x)}$$

Posterior: $p(y \mid x)$

Likelihood: $p(x \mid y)$

Prior: $p(y)$

Evidence: $p(x)$

Example: Drug Test

$$p(y \mid x) = \frac{p(x \mid y)\,p(y)}{p(x)}$$

Posterior, Likelihood, Prior, Evidence

Example:

$$p(Test = pos \mid User = T) = 0.99$$
$$p(Test = pos \mid User = F) = 0.01$$
$$p(User = True) = 0.005$$

Questions:

1. What is the likelihood?
2. What is the prior?
3. What is $p(User = T \mid Test = pos)$?

Independence of Random Variables

Definition: X and Y are **independent** if:

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

X and Y are **conditionally independent given** Z if:

$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

Example: Coins (Ex.7 in the course text)

- Suppose you have a biased coin: It does not come up heads with probability 0.5. Instead, it is more likely to come up heads.
- Let Z be the bias of the coin, with $\mathscr{Z} = \{0.3, 0.5, 0.8\}$ and probabilities $P(Z = 0.3) = 0.7$, $P(Z = 0.5) = 0.2$ and $P(Z = 0.8) = 0.1$.
  - **Question:** What other outcome space could we consider?
  - **Question:** What kind of distribution is this?
  - **Question:** What other kinds of distribution could we consider?
- Let X and Y be two consecutive flips of the coin
- **Question:** Are X and Y independent?
- **Question:** Are X and Y conditionally independent given Z?

Conditional Independence Is a Property of the Distribution

- Conditional independence is a property of the (joint) distribution
  - It is not somehow objective for all possible distributions

X	Y	Z	p
0	0	0.3	0.245
0	0	0.8	0.02
0	1	0.3	0.105
0	1	0.8	0.08
1	0	0.3	0.105
1	0	0.8	0.08
1	1	0.3	0.045
1	1	0.8	0.32

X	Y	Z	p
0	0	0.3	0.08
0	0	0.8	0.08
0	1	0.3	0.12
0	1	0.8	0.12
1	0	0.3	0.12
1	0	0.8	0.12
1	1	0.3	0.18
1	1	0.8	0.18

Expected Value

The expected value of a random variable is the **weighted average** of that variable over its domain.

Definition: Expected value of a random variable

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} x p(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} x p(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Expected Value with Functions

The expected value of a function $f : \mathscr{X} \to \mathbb{R}$ of a random variable is the **weighted average** of that function's value over the domain of the variable.

Definition: Expected value of a function of a random variable

$$\mathbb{E}[f(X)] = \begin{cases} \sum_{x \in \mathscr{X}} f(x)p(x) & \text{if } X \text{ is discrete} \\ \int_{\mathscr{X}} f(x)p(x)\,dx & \text{if } X \text{ is continuous.} \end{cases}$$

Example:
Suppose you get \$10 if heads is flipped, or lose \$3 if tails is flipped.
What are your winnings **on expectation**?

Conditional Expectations

Definition:

The **expected value of Y conditional on $X = x$** is

$$\mathbb{E}[Y \mid X = x] = \begin{cases} \sum_{y \in \mathcal{Y}} y p(y \mid x) & \text{if } Y \text{ is discrete,} \\ \int_{\mathcal{Y}} y p(y \mid x)\, dy & \text{if } Y \text{ is continuous.} \end{cases}$$

Question: What is $\mathbb{E}[Y \mid X]$?

Properties of Expectations

- Linearity of expectation:
  - $\mathbb{E}[cX] = c\mathbb{E}[X]$ for all constant c
  - $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- Products of expectations of **independent** random variables X, Y:
  - $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Law of Total Expectation:
  - $\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \mathbb{E}[Y]$
- **Question:** How would you prove these?

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} y p(y)$$ def. E[Y]

$$= \sum_{y \in \mathcal{Y}} y \sum_{x \in \mathcal{X}} p(x, y)$$ def. marginal distribution

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y p(x, y)$$ rearrange sums

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y p(y \mid x) p(x)$$ Chain rule

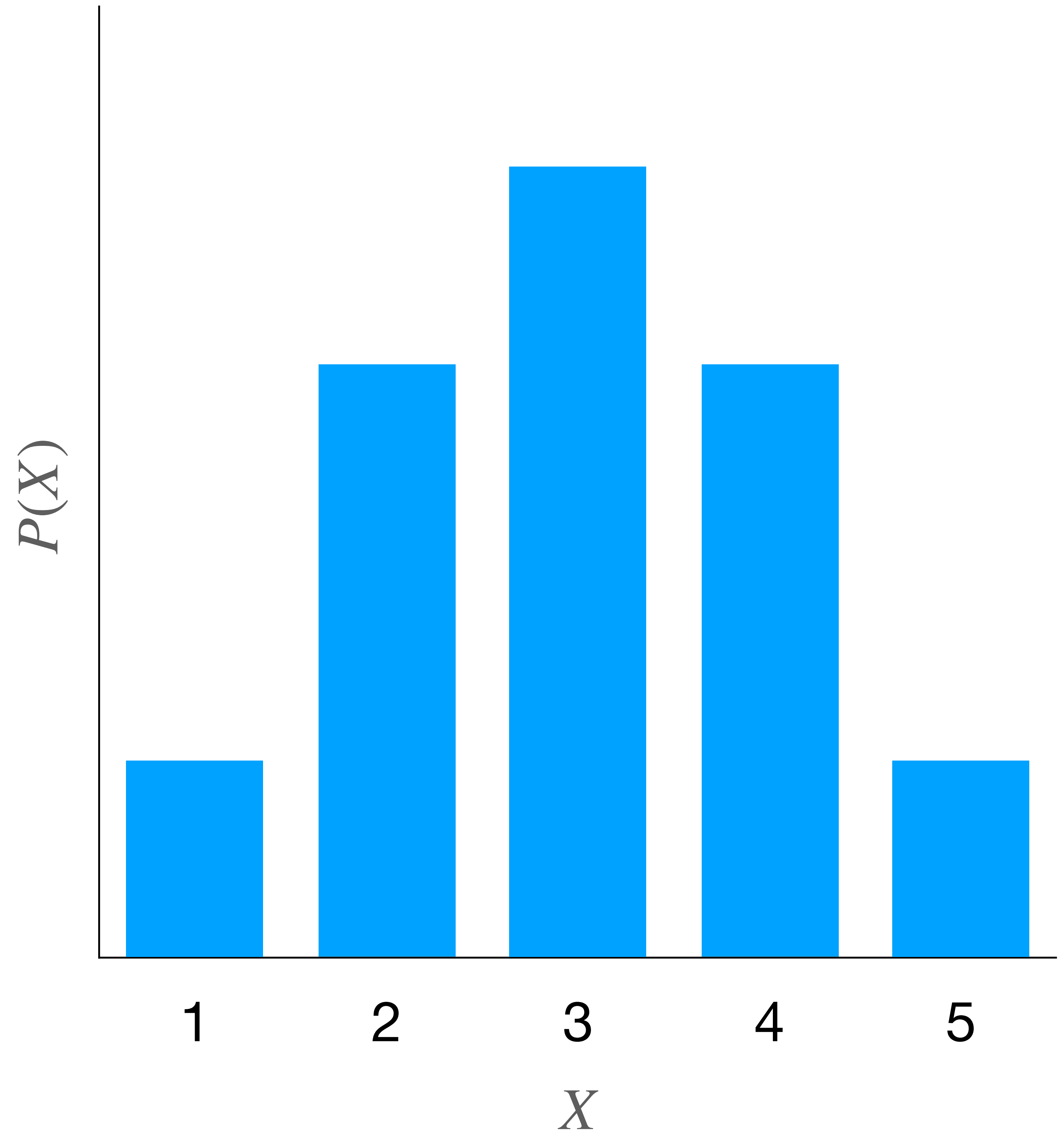
$$= \sum_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}} y p(y \mid x) \right) p(x)$$

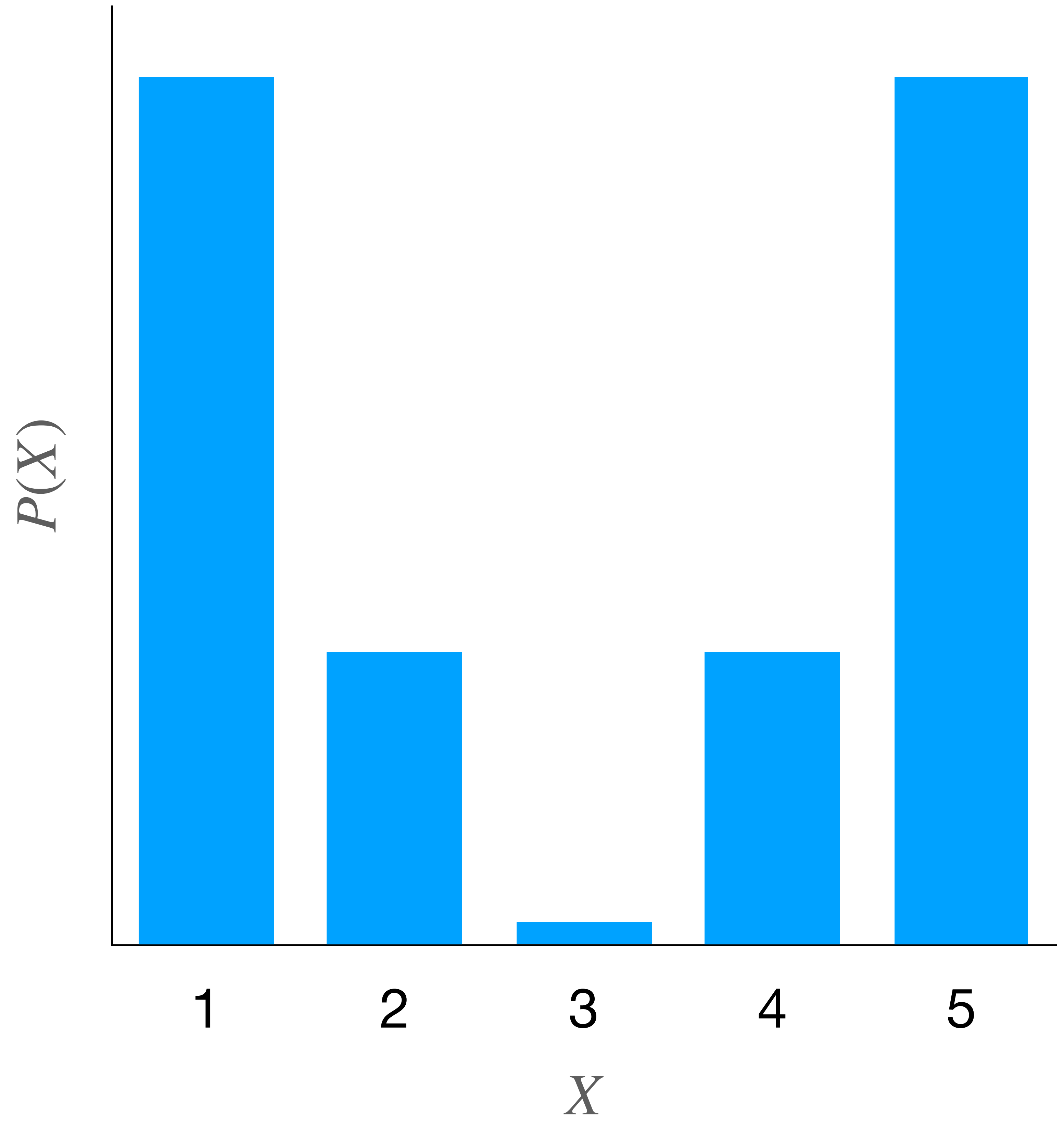
$$= \sum_{x \in \mathcal{X}} \left(\mathbb{E}[Y \mid X = x] \right) p(x)$$ def. E[Y | X = x]

$$= \sum_{x \in \mathcal{X}} \left(\mathbb{E}[Y \mid X = x] \right) p(x)$$

$$= \mathbb{E}\left(\mathbb{E}[Y \mid X] \right) \blacksquare$$ def. expected value of function

Expected Value is a Lossy Summary





$\mathbb{E}[X] = 3$

$\mathbb{E}[X^2] \simeq 10$

$\mathbb{E}[X] = 3$

$\mathbb{E}[X^2] \simeq 12$

Variance

Definition: The **variance** of a random variable is

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

i.e., $\mathbb{E}[f(X)]$ where $f(x) = (x - \mathbb{E}[X])^2$.

Equivalently,

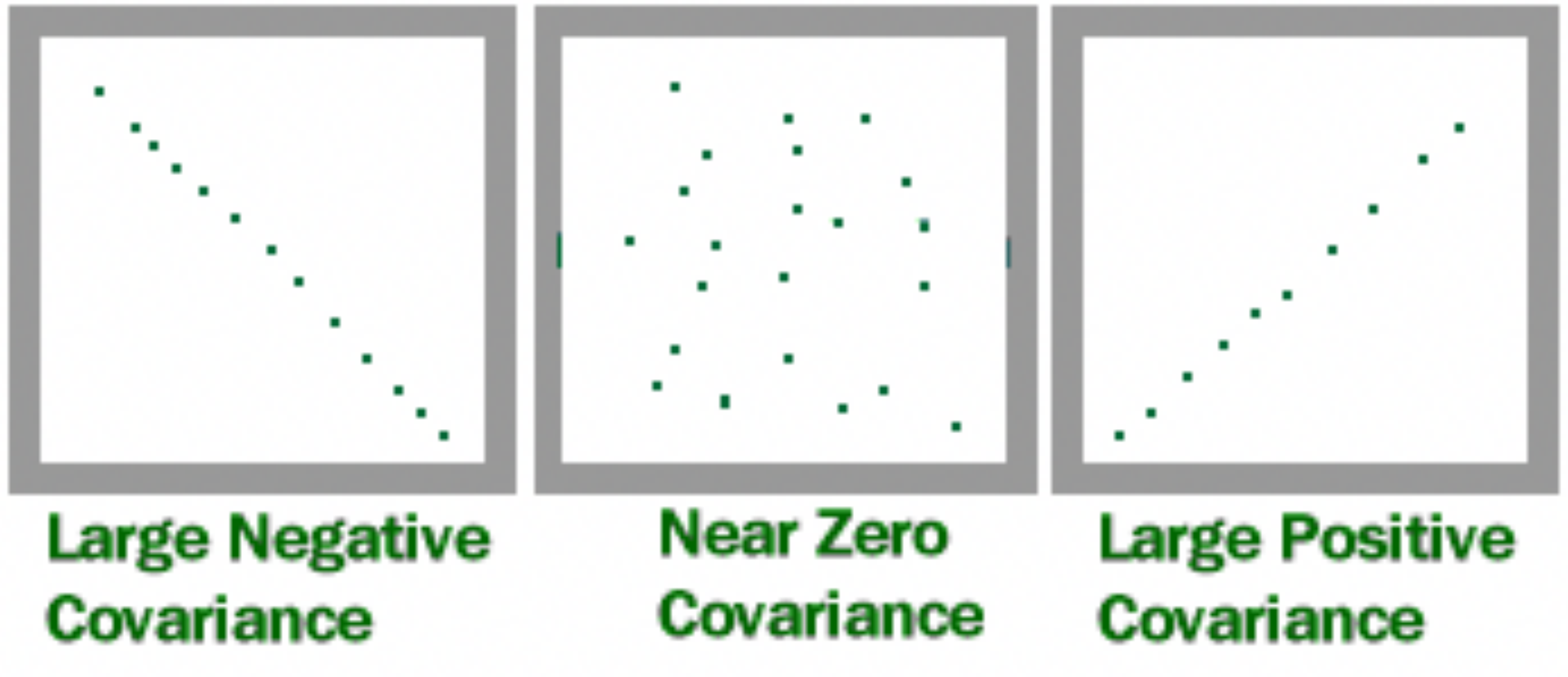
$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

(**why?**)

Covariance

Definition: The covariance of two random variables is

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\,.$$

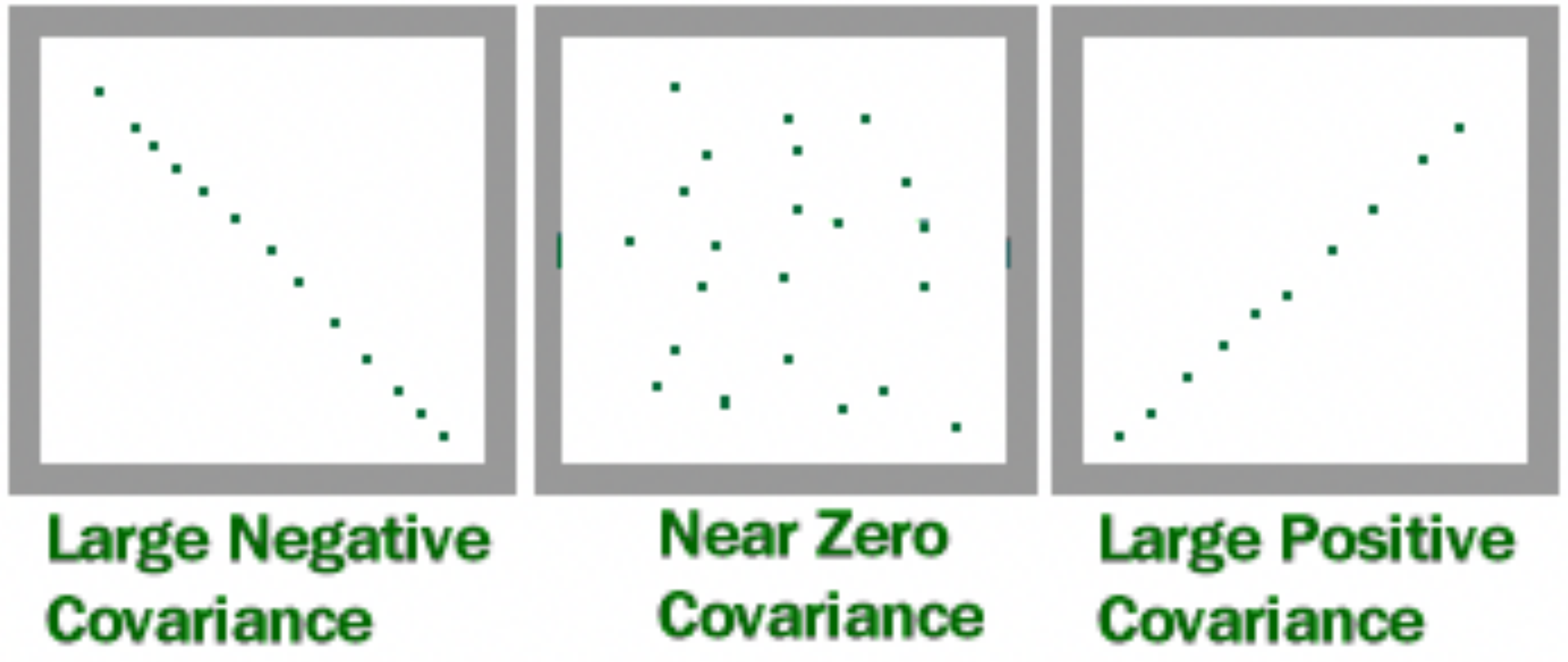


Question: What is the range of $\text{Cov}(X, Y)$?

Correlation

Definition: The **correlation** of two random variables is

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



Question: What is the range of $\text{Corr}(X, Y)$?

hint: $\text{Var}(X) = \text{Cov}(X, X)$

Properties of Variances

- $\text{Var}[c] = 0$ for constant c
- $\text{Var}[cX] = c^2\text{Var}[X]$ for constant c
- $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$
- For **independent** X, Y,
  $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$ (**why?**)

Independence and Decorrelation

- Independent RVs have zero correlation (**why?**)

  hint: $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

- Uncorrelated RVs (i.e., $\text{Cov}(X, Y) = 0$) **might be dependent** (i.e., $p(x, y) \neq p(x)p(y)$).
  - Correlation (**Pearson's correlation coefficient**) shows linear relationships; but can miss nonlinear relationships
  - **Example:** $X \sim \text{Uniform}\{-2, -1, 0, 1, 2\}$, $Y = X^2$
    - $\mathbb{E}[XY] = .2(-2 \times 4) + .2(2 \times 4) + .2(-1 \times 1) + .2(1 \times 1) + .2(0 \times 0)$
    - $\mathbb{E}[X] = 0$
    - So $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$

Summary

- **Random variables** are functions from sample to some value
  - Upshot: A random variable takes different values with some probability
- The value of one variable can be informative about the value of another (because they are both functions of the same sample)
  - Distributions of multiple random variables are described by the **joint** probability distribution (joint PMF or joint PDF)
  - You can have a new distribution over one variable when you **condition** on the other
- The **expected value** of a random variable is an **average** over its values, **weighted** by the probability of each value
- The **variance** of a random variable is the expected squared distance from the mean
- The **covariance** and **correlation** of two random variables can summarize how changes in one are informative about changes in the other.