Multivariate Probability

CMPUT 267: Basics of Machine Learning

§2.2-2.4

Outline

1. Multiple Random Variables
2. Independence
3. Expectations and Moments

Multiple Variables

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

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

Example: $X =$ number with $\mathcal{X} = \{1,2,3,4,5,6\}$ and $Y =$ position, with $\mathcal{Y} = \{\text{left, right}\}$

May ask questions like $P(X = 1, Y = \text{left})$ or $P(X \geq 4, Y = \text{left})$

Joint Distribution

We typically 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) = 50/100	P(X=0, Y=1) = 1/100
X=1	P(X=1, Y=0) = 10/100	P(X=1, Y=1) = 39/100

Is this joint distribution valid?

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) = 50/100	P(X=0, Y=1) = 1/100
X=1	P(X=1, Y=0) = 10/100	P(X=1, Y=1) = 39/100

- **Exercise**: Check if $\sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} p(x, y) = 1$

Is this joint distribution valid?

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) = 50/100	P(X=0, Y=1) = 1/100
X=1	P(X=1, Y=0) = 10/100	P(X=1, Y=1) = 39/100

- **Exercise**: Check if $\sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} p(x, y) = 1$
- $\sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} p(x, y) = 1/2 + 1/100 + 1/10 + 39/100 = 1$

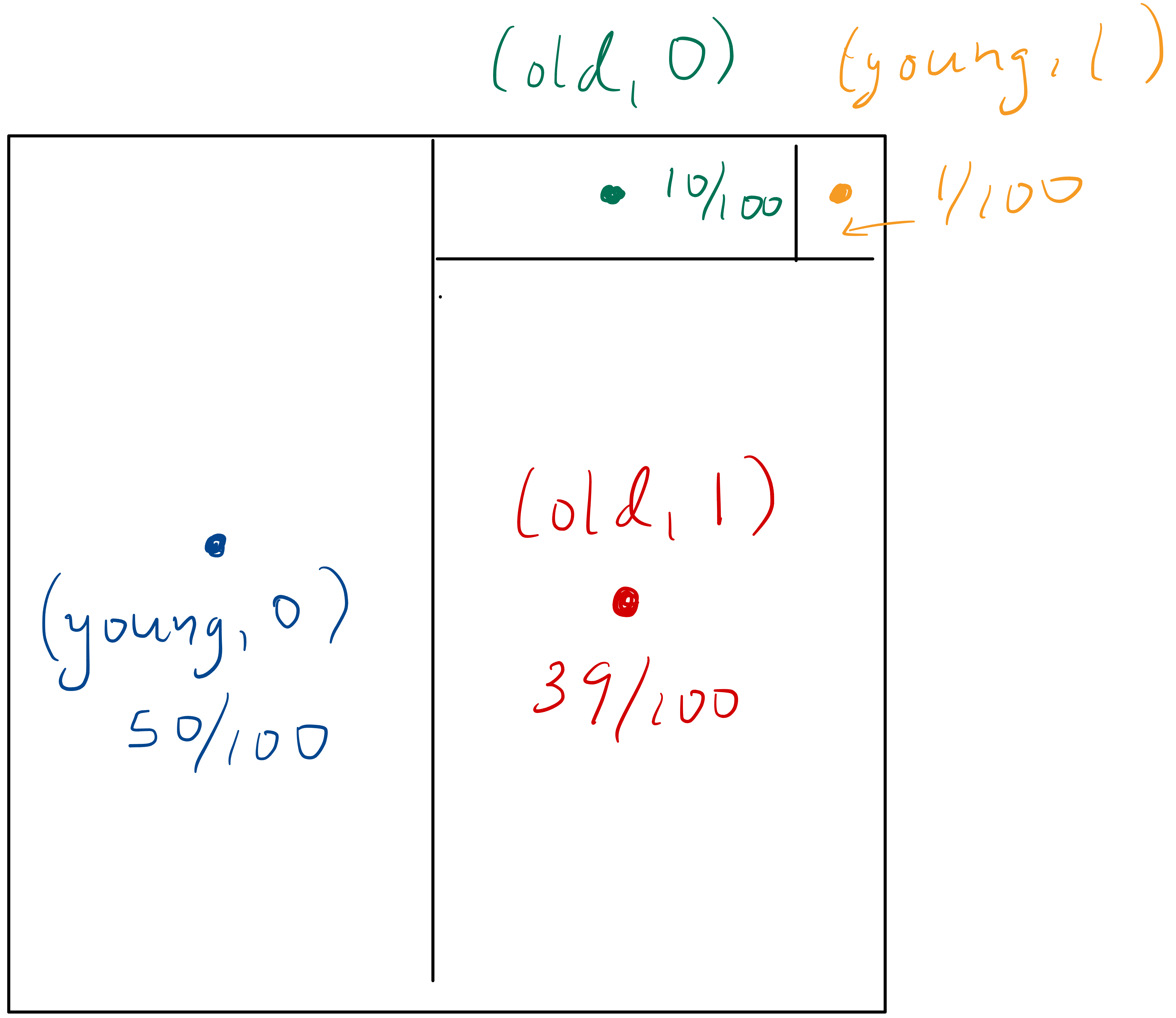
Visualizing the joint table

$$X \in \{\text{young}, \text{old}\}$$

$$Y \in \{0,1\}$$

Shows relative proportion
of each outcome

If I were to throw a dart at this
rectangle and it hit random locations*,
then we would see
(young, 0) half of the time
(young, 1) a 100th of the time
(old, 0) a 10th of the time
(old, 1) almost 4/10 ths of the time



*I'm really bad at darts

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) = 50/100	P(X=0, Y=1) = 1/100
X=1	P(X=1, Y=0) = 10/100	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** $P(Y = 1 \mid X = 0)$? (Prob. that person we know is young has arthritis)

Marginal Distribution for Y

$$p(Y=0) = \sum_{x \in \mathcal{X}} p(x,0) = \sum_{x \in \{\text{young,old}\}} p(x,0)$$

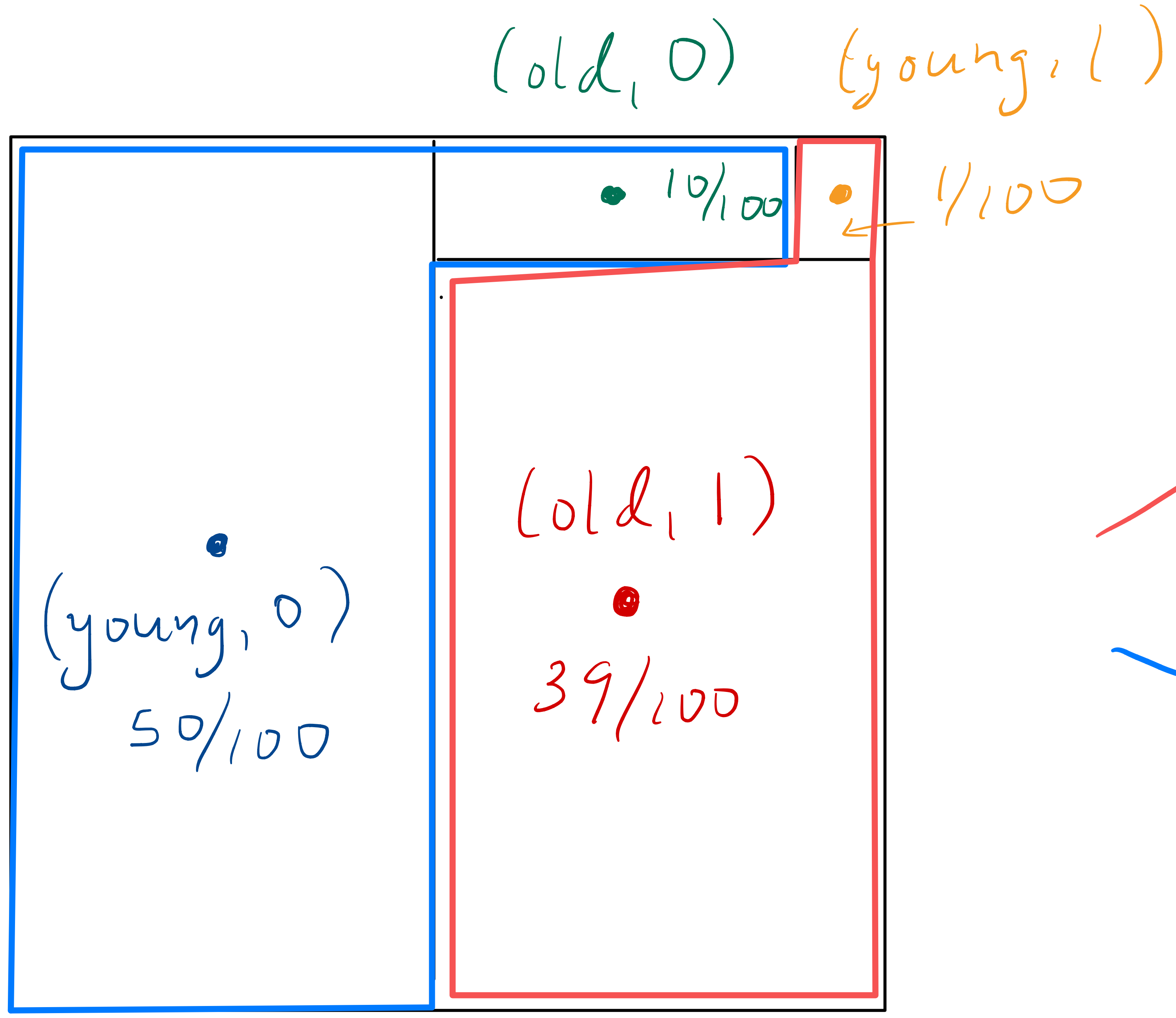
$$p(Y=1) = \sum_{x \in \mathcal{X}} p(x,1) = \sum_{x \in \{\text{young,old}\}} p(x,1)$$

Joint $p(x,y)$

Marginals = Area of subspace in joint event space

More generically

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



$p(Y=1) = \frac{39}{100} + \frac{1}{100} = 0.4$

$P(Y=0) = \frac{50}{100} + \frac{10}{100} = 0.6$

Another Exercise

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) = 50/100	P(X=0, Y=1) = 1/100
X=1	P(X=1, Y=0) = 10/100	P(X=1, Y=1) = 39/100

- **Exercise**: Compute marginal $p(x) = \sum_{y \in \{0,1\}} p(x, y)$

Another Exercise

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) = 50/100	P(X=0, Y=1) = 1/100
X=1	P(X=1, Y=0) = 10/100	P(X=1, Y=1) = 39/100

- **Exercise**: Compute marginal $p(x=1) = \sum_{y \in \{0,1\}} p(x=1, y) = 49/100$,

$p(x=0) = 1 - p(x=1) = 51/100$

Marginal distributions

- For two random variables X, Y,
- If they are discrete we have $p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$
- If they are continuous we have $p(x) = \int_{\mathcal{Y}} p(x, y) dy$
- If X is discrete and Y is continuous then $p(x) = \int_{\mathcal{Y}} p(x, y) dy$
- If X is continuous and Y is discrete then $p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$

Marginals for more than two variables

- The formulas extend naturally for more than two variables (see notes)
- We will almost always marginalize out over one variable

Question: Why do we write p for $p(x)$ and $p(x, y)$?

- 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 $p(x, y)$, $p(x)$ and $p(y)$
- Even though it would be more precise to write something like $p(x, y)$, $p_x(x)$ and $p_y(y)$
- We can tell which function we're talking about from context (i.e., arguments)

Now let’s consider PMFs and PDFs for more than two variables

PMFs and PDFs of Many Variables

In general, we can consider a d-dimensional random variable $X = (X_1, \ldots, X_d)$ with vector-valued outcomes $\mathbf{x} = (x_1, \ldots, x_d)$, 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

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

PMFs and PDFs of Many Variables

In general, we can consider a d-dimensional random variable $X = (X_1, \ldots, X_d)$ with vector-valued outcomes $\mathbf{x} = (x_1, \ldots, x_d)$, 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

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

Rules of Probability Already Covered the Multidimensional Case

Outcome space is $\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2 \times \ldots \times \mathscr{X}_d$

Outcomes are multidimensional variables $\mathbf{x} = [x_1, x_2, \ldots, x_d]$

Discrete case:

$p : \mathscr{X} \to [0,1]$ is a **(joint) probability mass function** if $\sum_{\mathbf{x} \in \mathscr{X}} p(\mathbf{x}) = 1$

Continuous case:

$p : \mathscr{X} \to [0,\infty)$ is a **(joint) probability density function** if $\int_{\mathscr{X}} p(\mathbf{x})\, d\mathbf{x} = 1$

But useful to recognize that we have multiple variables

Conditional Distribution

Definition: Conditional probability distribution

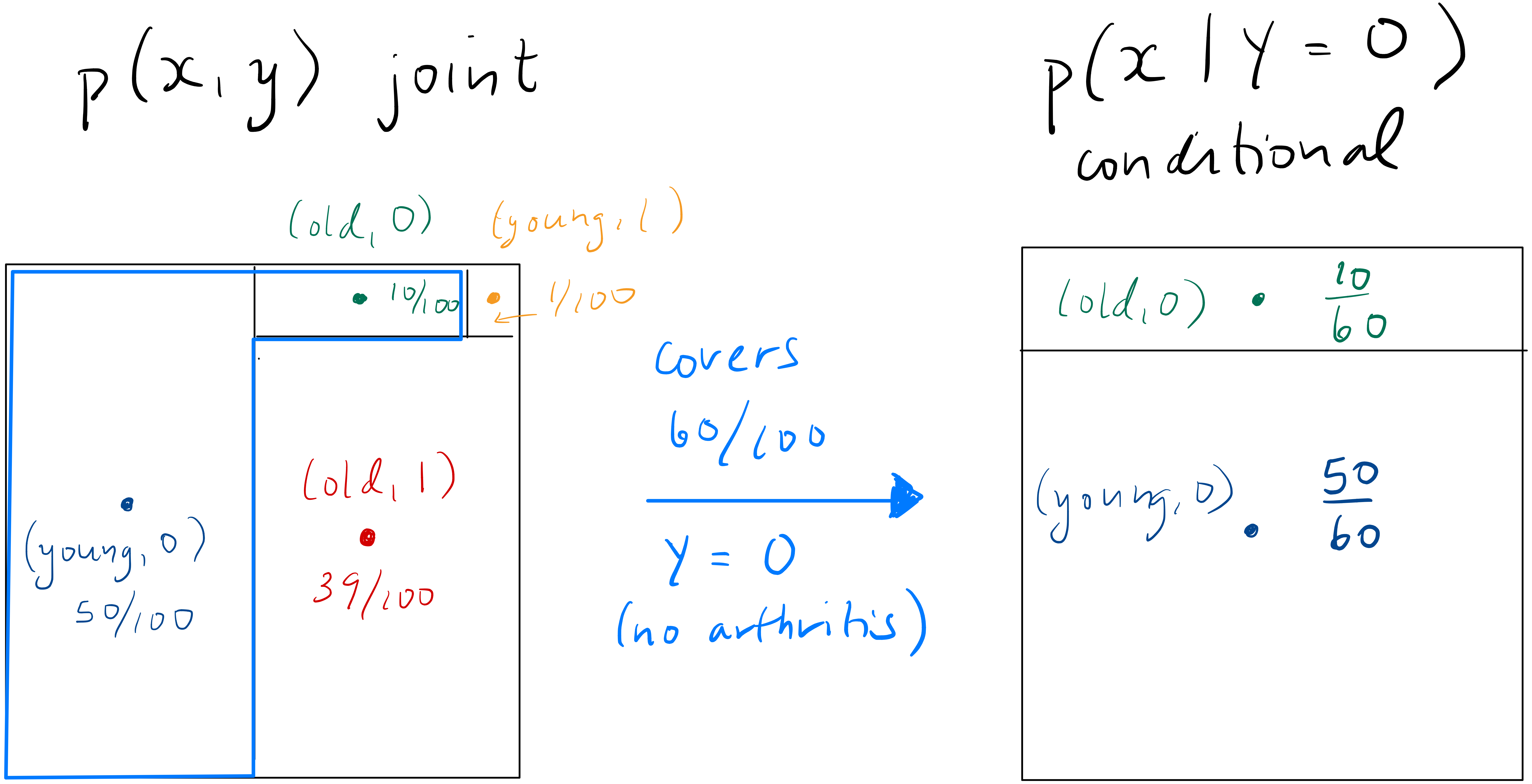
$$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

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?

Visualizing the conditional distribution



$$P(X = \text{young} \mid Y = 0) = P(X = \text{young}, Y = 0)/P(Y = 0) = (50/100)/(60/100) = 50/60$$

Announcements

- The first Reading Exercises is due next Thursday, at 11:59 pm
- You get two attempts, and we use the attempt with the highest mark
- eClass has some math and probability exercises, with solutions
- This course will remain heavy on math, because ML is math-heavy
  - One of the goals of this course is to get you more comfortable with math
  - It is a language, and like learning any language, it hurts the brain but gets better with practice! You can and will learn it

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**

$$p(x_1, \ldots, x_d) = p(x_1 \mid x_2, \ldots, x_d)p(x_2 \mid x_3, \ldots, x_d)\ldots p(x_{d-1} \mid x_d)p(x_d)$$

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

The Order Does Not Matter

The RVs are not ordered, so we can write

$$
\begin{aligned}
p(x, y, z) &= p(x \mid y, z)p(y|z)p(z) \\
&= p(x \mid y, z)p(z|y)p(y) \\
&= p(y \mid x, z)p(x|z)p(z) \\
&= p(y \mid x, z)p(z|x)p(x) \\
&= p(z \mid x, y)p(y|x)p(x) \\
&= p(z \mid x, y)p(x|y)p(y)
\end{aligned}
$$

All of these probabilities are equal

Bayes' Rule

From the chain rule, we have:

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

- 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

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

Bayes' Rule

- Bayes’ rule is typically used to reason about our beliefs, given new information
- Example: a scientist might have a belief about the prevalence of cancer in smokers (Y), and update with new evidence (X)
- In ML: we have a belief over our estimator (Y), and we update with new data that is like new evidence (X)

Definition: Bayes' rule

$$\underbrace{\boxed{p(y \mid x)}}_{\text{Posterior}} = \frac{\underbrace{\boxed{p(x \mid y)}}_{\text{Likelihood}}\,\underbrace{\boxed{p(y)}}_{\text{Prior}}}{\underbrace{\boxed{p(x)}}_{\text{Evidence}}}$$

Example: Disease Test

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

Example:

$$p(Test = pos \mid Dis = T) = 0.99$$
$$p(Test = pos \mid Dis = F) = 0.03$$
$$p(Dis = T) = 0.005$$

Questions:

1. What is $p(Dis = F)$?
2. What is $p(Dis = T \mid Test = pos)$?

Mapping to the formula, let
X be Test
Y be presence of the Disease

Example: Disease Test

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

Example:

$$p(Test = pos \mid Dis = T) = 0.99$$
$$p(Test = pos \mid Dis = F) = 0.03$$
$$p(Dis = T) = 0.005$$

Questions:

1. What is $p(Dis = F)$?
2. What is $p(Dis = T \mid Test = pos)$?

$$p(Dis = F) = 1 - p(Dis = T) = 1 - 0.005 = 0.995$$

Example:
Disease Test

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

Example:

$$p(Test = pos \mid Dis = T) = 0.99$$
$$p(Test = pos \mid Dis = F) = 0.03$$
$$p(Dis = T) = 0.005$$

Questions:

1. What is $p(Dis = F)$?
2. What is $p(Dis = T \mid Test = pos)$?

$$p(Dis = T \mid Test = pos) = \frac{p(Test = pos \mid Dis = T)p(Dis = T)}{p(Test = pos)}$$

Need to compute this part

Example: Disease Test

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

Example:

$$p(Test = pos \mid Dis = T) = 0.99$$
$$p(Test = pos \mid Dis = F) = 0.03$$
$$p(Dis = T) = 0.005$$

Questions:

1. What is $p(Dis = F)$?
2. What is $p(Dis = T \mid Test = pos)$?

$$\begin{aligned} p(Test = pos) &= \sum_{d \in \{T,F\}} p(Test = pos, d) \\ &= p(Test = pos, D = F) + p(Test = pos, D = T) \\ &= p(Test = pos \mid D = F)p(D = F) + p(Test = pos \mid D = T)p(D = T) \\ &= 0.03 \times 0.995 + 0.99 \times 0.005 = 0.0348 \end{aligned}$$

Example: Disease Test

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

Example:

$$p(Test = pos \mid Dis = T) = 0.99$$
$$p(Test = pos \mid Dis = F) = 0.03$$
$$p(Dis = T) = 0.005$$

Questions:

1. What is $p(Dis = F)$?
2. What is $p(Dis = T \mid Test = pos)$?

$$p(Test = pos) = 0.0348$$

$$p(Dis = T \mid Test = pos) = \frac{p(Test = pos \mid Dis = T)p(Dis = T)}{p(Test = pos)} = \frac{0.99 \times 0.005}{0.0348} \approx 0.142$$

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?

Example: Coins (2)

- Now imagine I told you $Z = 0.3$ (i.e., probability of heads is 0.3)
- Let X and Y be two consecutive flips of the coin
- What is $P(X = Heads \mid Z = 0.3)$? What about $P(X = Tails \mid Z = 0.3)$?
- What is $P(Y = Heads \mid Z = 0.3)$? What about $P(Y = Tails \mid Z = 0.3)$?
- Is $P(X = x, Y = y \mid Z = 0.3) = P(X = x \mid Z = 0.3)P(Y = y \mid Z = 0.3)$?

Example: Coins (3)

- Now imagine we do not know Z
  - e.g., you randomly grabbed it from a bin of coins with probabilities $P(Z = 0.3) = 0.7$, $P(Z = 0.5) = 0.2$ and $P(Z = 0.8) = 0.1$
- What is $P(X = Heads)$?

$$P(X = Heads) = \sum_{z \in \{0.3, 0.5, 0.8\}} P(X = Heads \,|\, Z = z) p(Z = z)$$

$$= P(X = Heads \,|\, Z = 0.3) p(Z = 0.3)$$

$$+ P(X = Heads \,|\, Z = 0.5) p(Z = 0.5)$$

$$+ P(X = Heads \,|\, Z = 0.8) p(Z = 0.8)$$

$$= 0.3 \times 0.7 + 0.5 \times 0.2 + 0.8 \times 0.1 = 0.39$$

Example: Coins (4)

- Now imagine we do not know Z
  - e.g., you randomly grabbed it from a bin of coins with probabilities $P(Z = 0.3) = 0.7$, $P(Z = 0.5) = 0.2$ and $P(Z = 0.8) = 0.1$
- Is $P(X = Heads, Y = Heads) = P(X = Heads)p(Y = Heads)$?
  - For brevity, lets use h for Heads

$$P(X = h, Y = h) = \sum_{z \in \{0.3,0.5,0.8\}} P(X = h, Y = h \,|\, Z = z)p(Z = z)$$

$$= \sum_{z \in \{0.3,0.5,0.8\}} P(X = h \,|\, Z = z)P(Y = h \,|\, Z = z)p(Z = z)$$

Example: Coins (4)

- $P(Z = 0.3) = 0.7$, $P(Z = 0.5) = 0.2$ and $P(Z = 0.8) = 0.1$
- Is $P(X = Heads, Y = Heads) = P(X = Heads)p(Y = Heads)$?

$$
\begin{aligned}
P(X = h, Y = h) &= \sum_{z \in \{0.3,0.5,0.8\}} P(X = h, Y = h \,|\, Z = z)p(Z = z) \\
&= \sum_{z \in \{0.3,0.5,0.8\}} P(X = h \,|\, Z = z)P(Y = h \,|\, Z = z)p(Z = z) \\
&= P(X = h \,|\, Z = 0.3)P(Y = h \,|\, Z = 0.3)p(Z = 0.3) \\
&+P(X = h \,|\, Z = 0.5)P(Y = h \,|\, Z = 0.5)p(Z = 0.5) \\
&+P(X = h \,|\, Z = 0.8)p(Y = h \,|\, Z = 0.8)p(Z = 0.8) \\
&= 0.3 \times 0.3 \times 0.7 + 0.5 \times \times 0.5 \times 0.2 + 0.8 \times 0.8 \times 0.1 \\
&= 0.177 \neq 0.39 * 0.39 = 0.1521
\end{aligned}
$$

Example: Coins (4)

- 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 $P(Z = 0.8) = 0.1$.
- Let X and Y be two consecutive flips of the coin
- **Question:** Are X and Y conditionally independent given Z?
  - i.e., $P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$
- **Question:** Are X and Y independent?
  - i.e. $P(X = x, Y = y) = P(X = x)P(Y = y)$

The Distribution Changes Based on What We Know

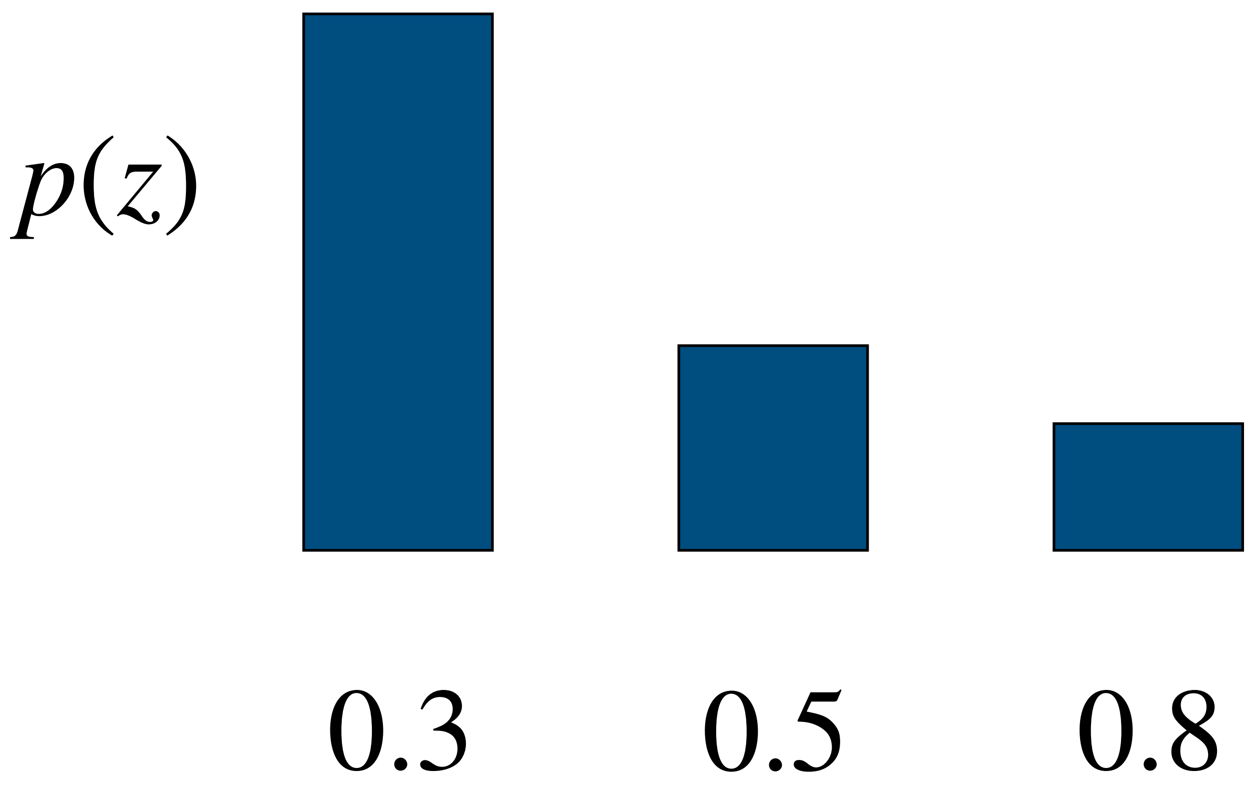
- The coin has some true bias z
- If we **know** that bias, we reason about $P(X = x \mid Z = z)$
  - Namely, the probability of x **given** we know the bias is z
- If we **do not know** that bias, then **from our perspective** the coin outcomes follows probabilities $P(X = x)$, which is a weighted average over three different worlds (in each world the coin bias is different)
  - The world still flips the coin with bias z
- Conditional independence is a property of the distribution we are reasoning about, not an objective truth about outcomes

A bit more intuition

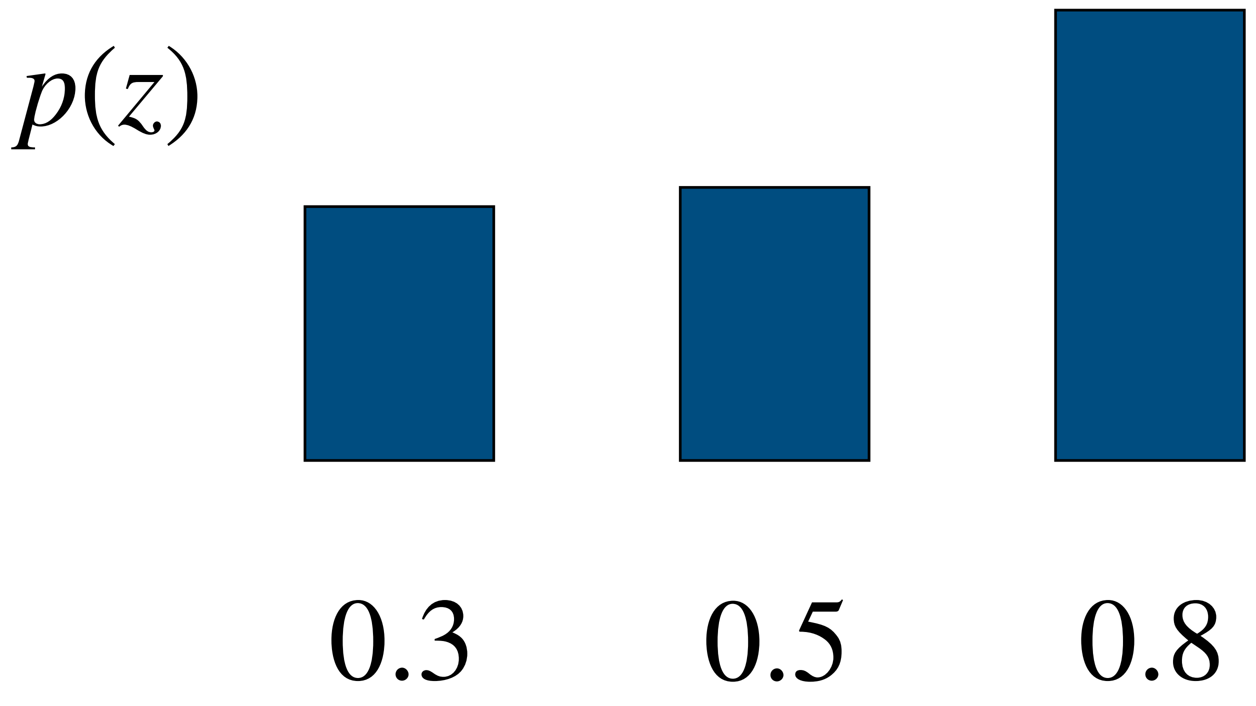
- If we **do not know** that bias, then **from our perspective** the coin outcomes follows probabilities $P(X = x, Y = y)$
  - and X and Y are correlated
- If we know $X = h$, do we think it's more likely $Y = h$? i.e., is $P(X = h, Y = h) > P(X = h, Y = t)$?

How is this relevant to us?

- Let's imagine you want to infer (or learn) the bias of the coin, from data
  - data in this case corresponds to a sequence of flips $X_1, X_2, \ldots, X_n$
- You can ask: $P(Z = z \mid X_1 = H, X_2 = H, X_3 = T, \ldots, X_n = H)$



See 10 Heads
and 2 Tails



More uses for independence and conditional independence

- If I told you X = roof type was **independent** of Y = house price, would you use X as a feature to predict Y?
- Imagine you want to predict Y = Has Lung Cancer and you have an indirect correlation with X = Location since in Location 1 more people smoke on average. If you could measure Z = Smokes, then X and Y would be **conditionally independent** given Z.
  - Suggests you could look for such causal variables, that explain these correlations
- We will see the utility of conditional independence for learning models

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

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

Relationship to Population Average and Sample Average

- Or Population Mean and Sample Mean
- Population Mean = Expected Value, Sample Mean estimates this number
  - e.g., Population Mean = average height of the entire population
- For RV X = height, p(x) gives the probability that a randomly selected person has height x
- Sample average: you randomly sample n heights from the population
  - implicitly you are sampling heights proportionally to p
- As n gets bigger, the sample average approaches the true expected value

Connection to Sample Average

- Imagine we have a biased coin, p(x = 1) = 0.75, p(x = 0) = 0.25
- Imagine we flip this coin 1000 times, and see (x = 1) 700 times
- The sample average is

$$\frac{1}{1000}\sum_{i=1}^{1000} x_i = \frac{1}{1000}\left[\sum_{i:x_i=0} x_i + \sum_{i:x_i=1} x_i\right] = 0 \times \frac{300}{1000} + 1 \times \frac{700}{1000} = = 0 \times 0.3 + 1 \times 0.7 = 0.7$$

- The true expected value is

$$\sum_{x \in \{0,1\}} p(x)x = 0 \times p(x = 0) + 1p(x = 1) = 0 \times 0.25 + 1 \times 0.75 = 0.75$$

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 **in expectation**?

Expected Value Example

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

X is the outcome of the coin flip, 1 for heads and 0 for tails

$$f(x) = \begin{cases} -3 & \text{if } x = 0 \\ 10 & \text{if } x = 1 \end{cases}$$

$Y = f(X)$ is a new random variable

$$\mathbb{E}[Y] = \mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x) = f(0)p(0) + f(1)p(1) = .5 \times -3 + .5 \times 10 = 3.5$$

One More Example

Suppose X is the outcome of a dice role

$$f(x) = \begin{cases} -1 & \text{if } x \leq 3 \\ 1 & \text{if } x \geq 4 \end{cases}$$

$Y = f(X)$ is a new random variable. We see $Y = -1$ each time we observe 1, 2 or 3.

We see $Y = 1$ each time we observe 4, 5, or 6.

$$\begin{aligned}\mathbb{E}[Y] = \mathbb{E}[f(X)] &= \sum_{x \in \mathcal{X}} f(x)p(x) \\ &= (-1)\Big(p(X=1) + p(X=2) + p(X=3)\Big) \\ &\quad + (1)\Big(p(X=4) + p(X=5) + p(X=6)\Big)\end{aligned}$$

One More Example

Suppose X is the outcome of a dice role

$$f(x) = \begin{cases} -1 & \text{if } x \leq 3 \\ 1 & \text{if } x \geq 4 \end{cases}$$

$Y = f(X)$ is a new random variable. We see $Y = -1$ each time we observe 1, 2 or 3.

We see $Y = 1$ each time we observe 4, 5, or 6.

$$\mathbb{E}[Y] = \mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x) = \sum_{y \in \{-1,1\}} yp(y)$$

$$p(Y = -1) = p(X = 1) + p(X = 2) + p(X = 3) = 0.5$$

$$p(Y = 1) = p(X = 4) + p(X = 5) + p(X = 6) = 0.5$$

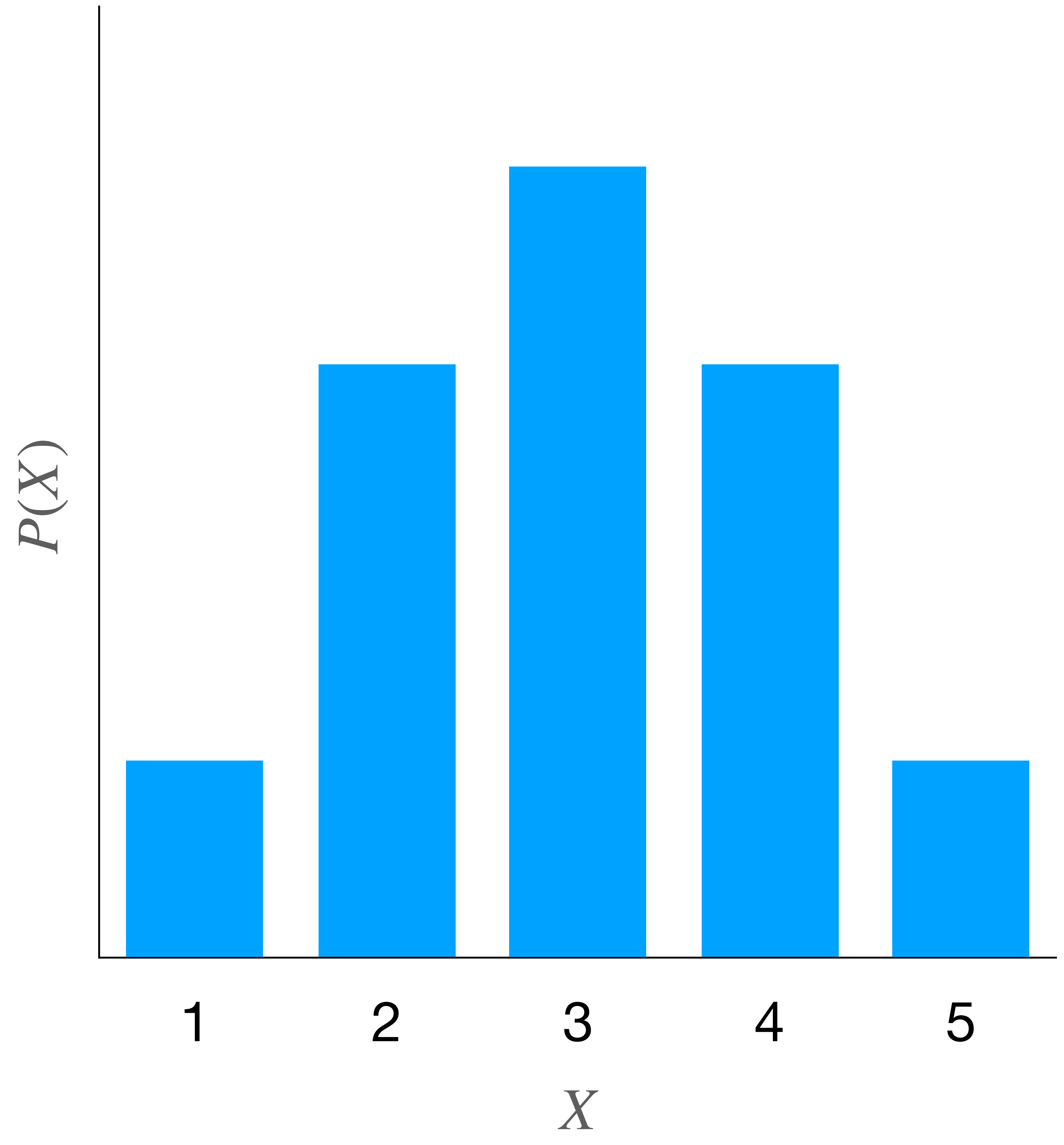
$$= (-1)\Big(p(X = 1) + p(X = 2) + p(X = 3)\Big)$$

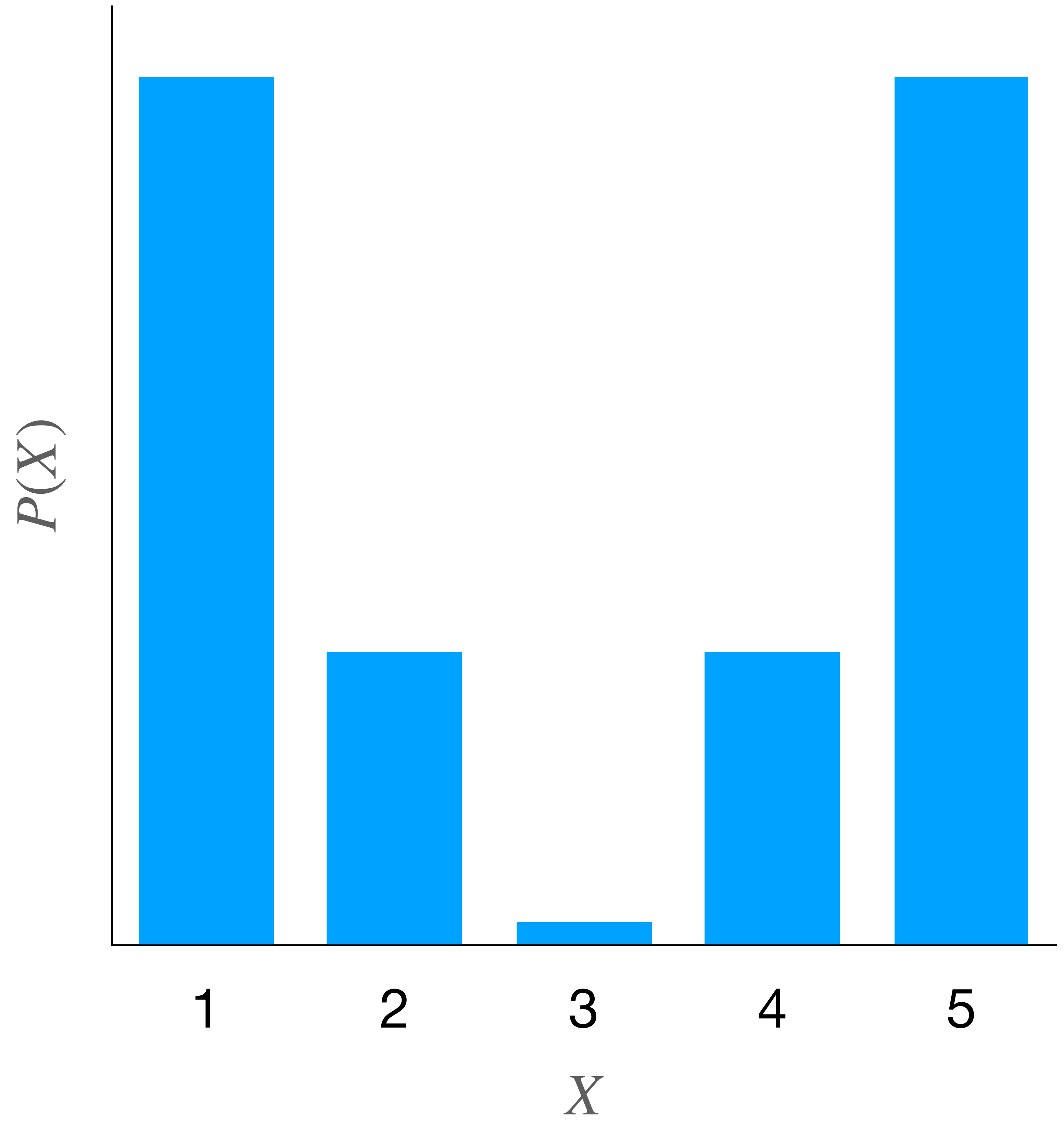
$$+ \ (1)\Big(p(X = 4) + p(X = 5) + p(X = 6)\Big)$$

$$= -1(0.5) + 1(0.5)$$

Summing over x with p(x) is equivalent, and can be simpler (no need to infer p(y))

Expected Value is a Lossy Summary





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

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

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

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

Conditional Expectations

Definition:

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

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

Another way to Write Conditional Expectations

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

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

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

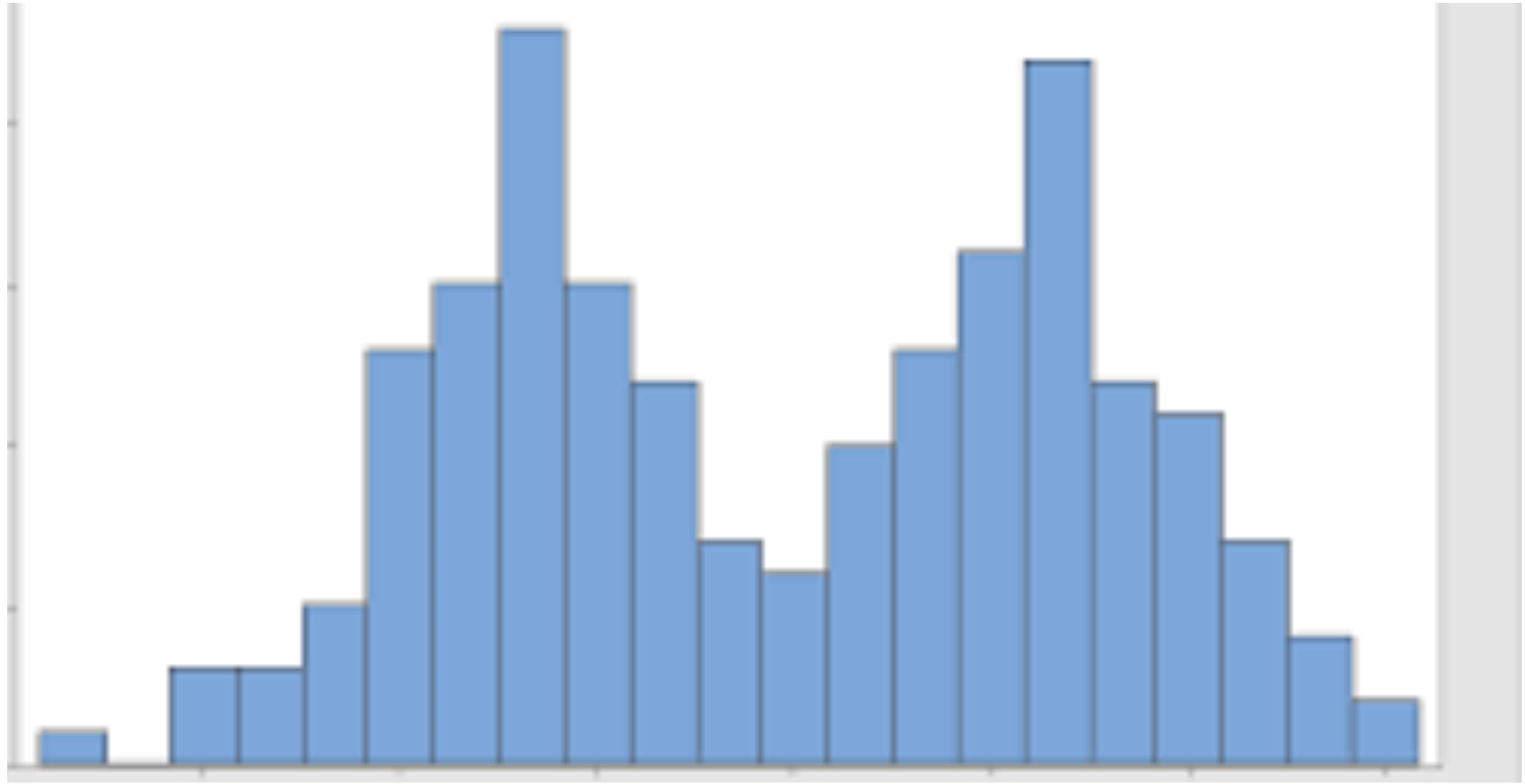
Conditional Expectation Example

- X is the type of a book, 0 for fiction and 1 for non-fiction
  - $p(X = 1)$ is the proportion of all books that are non-fiction
- Y is the number of pages
  - $p(Y = 100)$ is the proportion of all books with 100 pages
- $\mathbb{E}[Y|X = 0]$ is different from $\mathbb{E}[Y|X = 1]$
  - e.g. $\mathbb{E}[Y|X = 0] = 70$ is different from $\mathbb{E}[Y|X = 1] = 150$

Conditional Expectation Example (cont)

- What do we mean by $p(y\,|\,X=0)$? How might it differ from $p(y\,|\,X=1)$

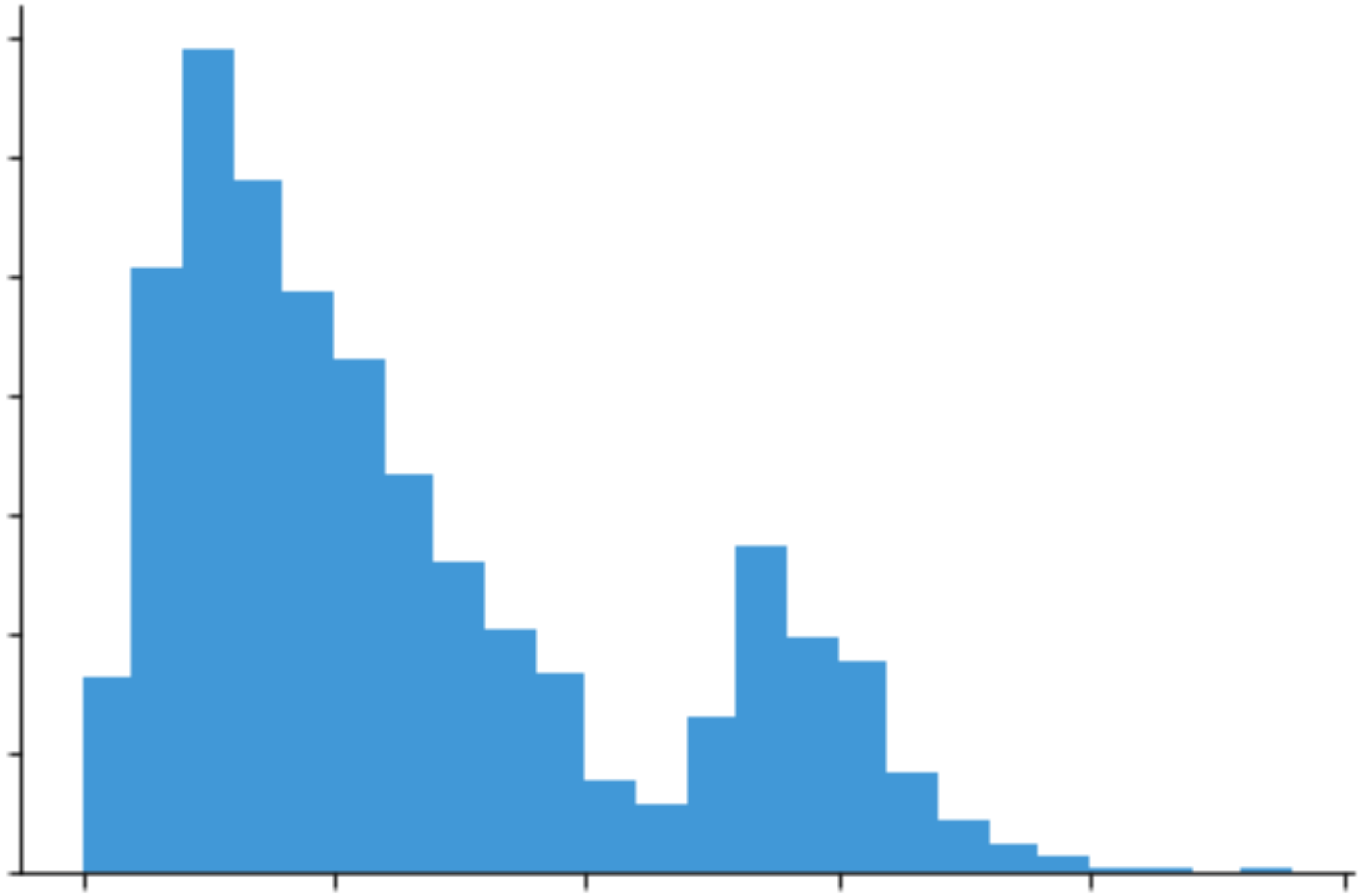
p(y) for X = 0, fiction books



Lots of shorter books

Lots of medium length books

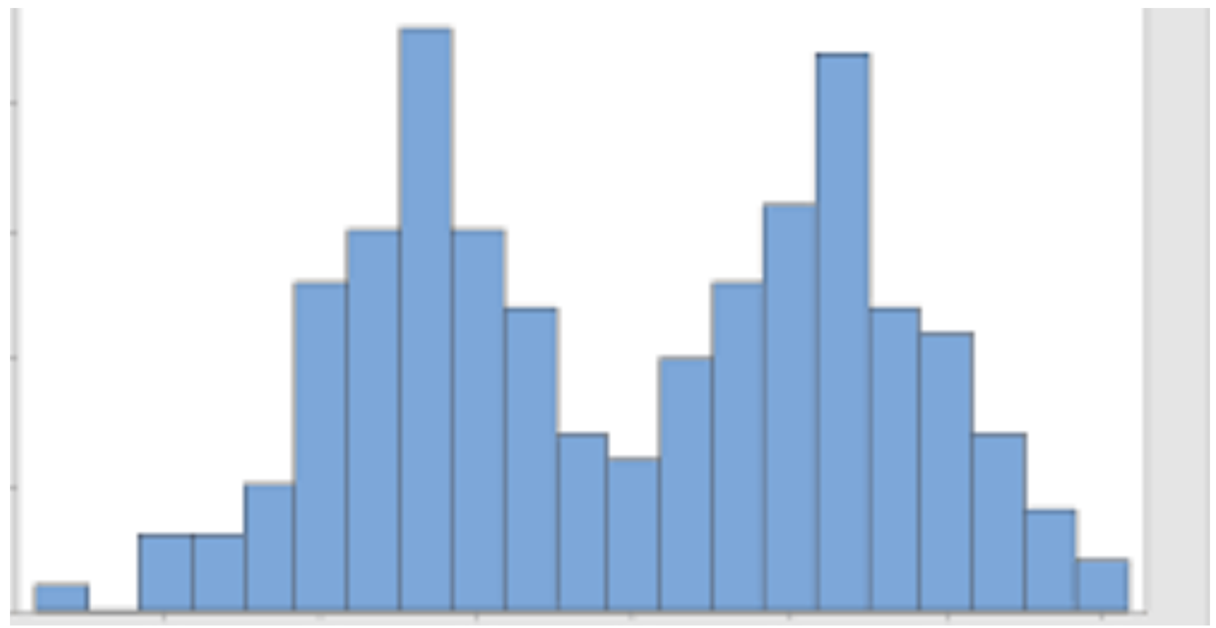
p(y) for X = 1, nonfiction books

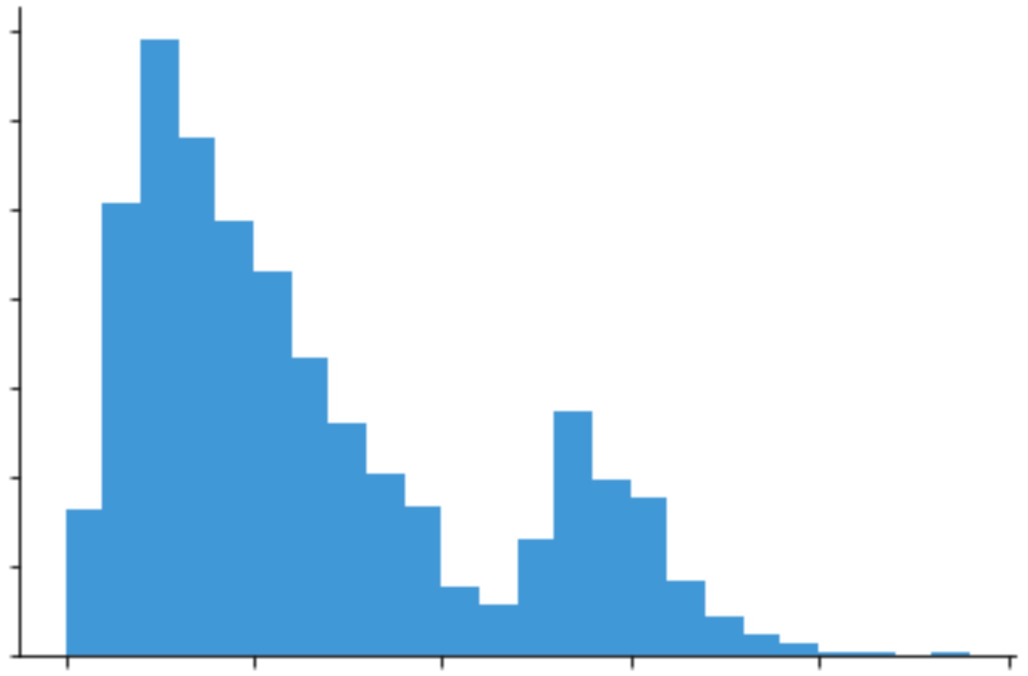


A long tail, a few very long books

Conditional Expectation Example (cont)

- What do we mean by $p(y \mid X = 0)$? How might it differ from $p(y \mid X = 1)$





- $\mathbb{E}[Y \mid X = 0]$ is the expectation over Y under distribution $p(y \mid X = 0)$
- $\mathbb{E}[Y \mid X = 1]$ is the expectation over Y under distribution $p(y \mid X = 1)$

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 $\mathbb{E}[Y \mid X]$?

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 $\mathbb{E}[Y \mid X]$?

Answer: $Z = \mathbb{E}[Y \mid X]$ is a random variable, $z = \mathbb{E}[Y \mid X = x]$ is an outcome

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?

Linearity of Expectation

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)(x+y) \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y)(x+y) \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y)x + \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y)y
\end{aligned}$$

$$\begin{aligned}
\sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y)x &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)x \\
&= \sum_{x\in\mathcal{X}} x \sum_{y\in\mathcal{Y}} p(x,y) \qquad \triangleright\ p(x) = \sum_{y\in\mathcal{Y}} p(x,y) \\
&= \sum_{x\in\mathcal{X}} xp(x) \\
&= \mathbb{E}[X]
\end{aligned}$$

Linearity of Expectation

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)(x+y) \\
&= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)(x+y) \\
&= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)x + \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)y \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

$$\begin{aligned}
\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)x &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)x \\
&= \sum_{x\in\mathcal{X}} x \sum_{y\in\mathcal{Y}} p(x,y) \qquad \triangleright\ p(x) = \sum_{y\in\mathcal{Y}} p(x,y) \\
&= \sum_{x\in\mathcal{X}} xp(x) \\
&= \mathbb{E}[X]
\end{aligned}$$

What if the RVs are continuous?

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)(x+y) \\
&= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)(x+y) \\
&= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)x + \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)y \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \int_{\mathcal{X}\times\mathcal{Y}} p(x,y)(x+y)d(x,y) \\
&= \int_{\mathcal{Y}}\int_{\mathcal{X}} p(x,y)(x+y)dxdy \\
&= \int_{\mathcal{Y}}\int_{\mathcal{X}} p(x,y)xdxdy + \int_{\mathcal{Y}}\int_{\mathcal{X}} p(x,y)ydxdy \\
&= \int_{\mathcal{X}} x\int_{\mathcal{Y}} p(x,y)dydx + \int_{\mathcal{Y}} y\int_{\mathcal{X}} p(x,y)dxdy \\
&= \int_{\mathcal{X}} xp(x)dx + \int_{\mathcal{Y}} yp(y)dy \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}
$$

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:
  - $\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \mathbb{E}[Y]$
- Notice: $f(x) = E[Y|X = x]$
  $\mathbb{E}[f(X)] = \mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \mathbb{E}[Y]$

$$
\begin{aligned}
\mathbb{E}[Y] &= \sum_{y\in\mathcal{Y}} yp(y) && \text{def. E[Y]}\\
&= \sum_{y\in\mathcal{Y}} y \sum_{x\in\mathcal{X}} p(x,y) && \text{def. marginal distribution}\\
&= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} yp(x,y) && \text{rearrange sums}\\
&= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} yp(y \mid x)p(x) && \text{Chain rule}\\
&= \sum_{x\in\mathcal{X}} \left(\sum_{y\in\mathcal{Y}} yp(y \mid x) \right) p(x) && \\
&= \sum_{x\in\mathcal{X}} \left(\mathbb{E}[Y \mid X = x]\right) p(x) && \text{def. E[Y | X = x]}\\
&= \sum_{x\in\mathcal{X}} \left(\mathbb{E}[Y \mid X = x]\right) p(x) && \\
&= \mathbb{E}\left(\mathbb{E}[Y \mid X]\right) \blacksquare && \text{def. expected value of function}
\end{aligned}
$$

Variance

Definition: The variance of a random variable is

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

Equivalently,

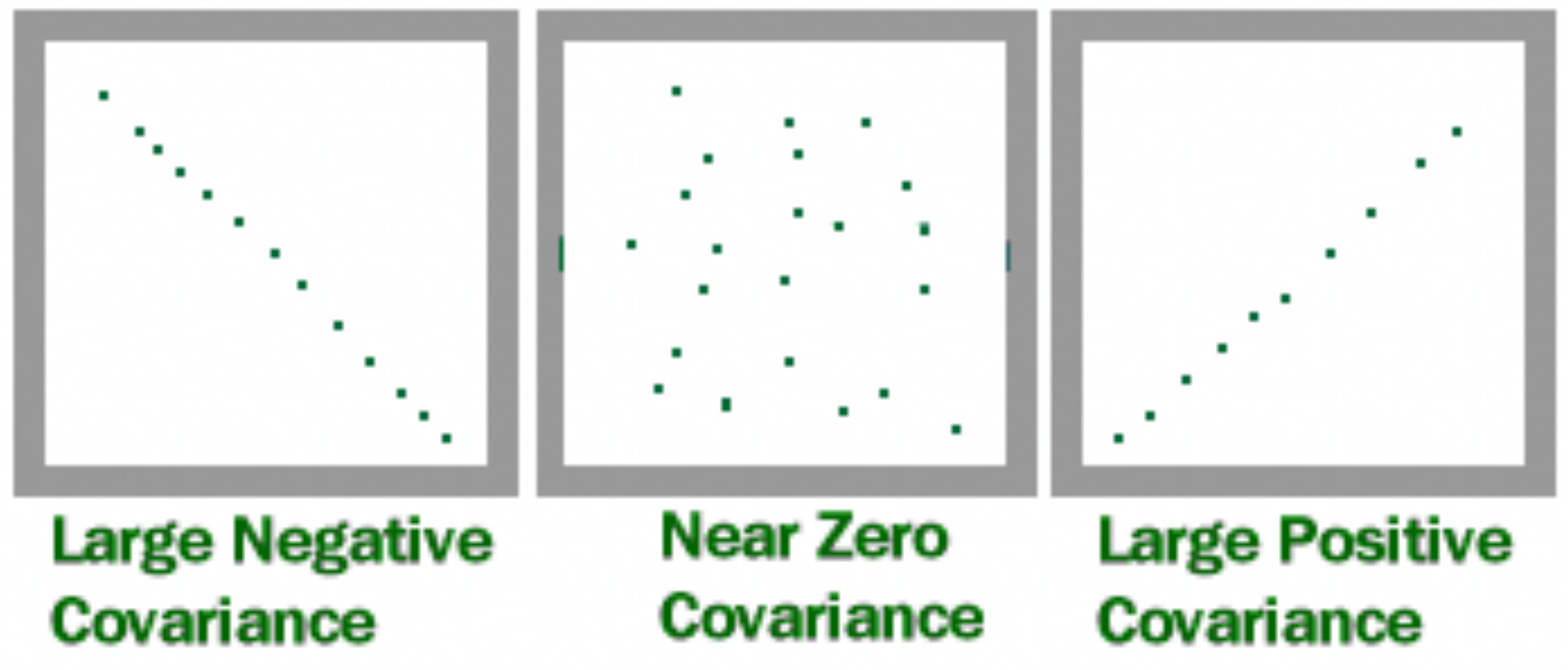
$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

(**Exercise:** Show that this is true)

Covariance

Definition: The covariance of two random variables is

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

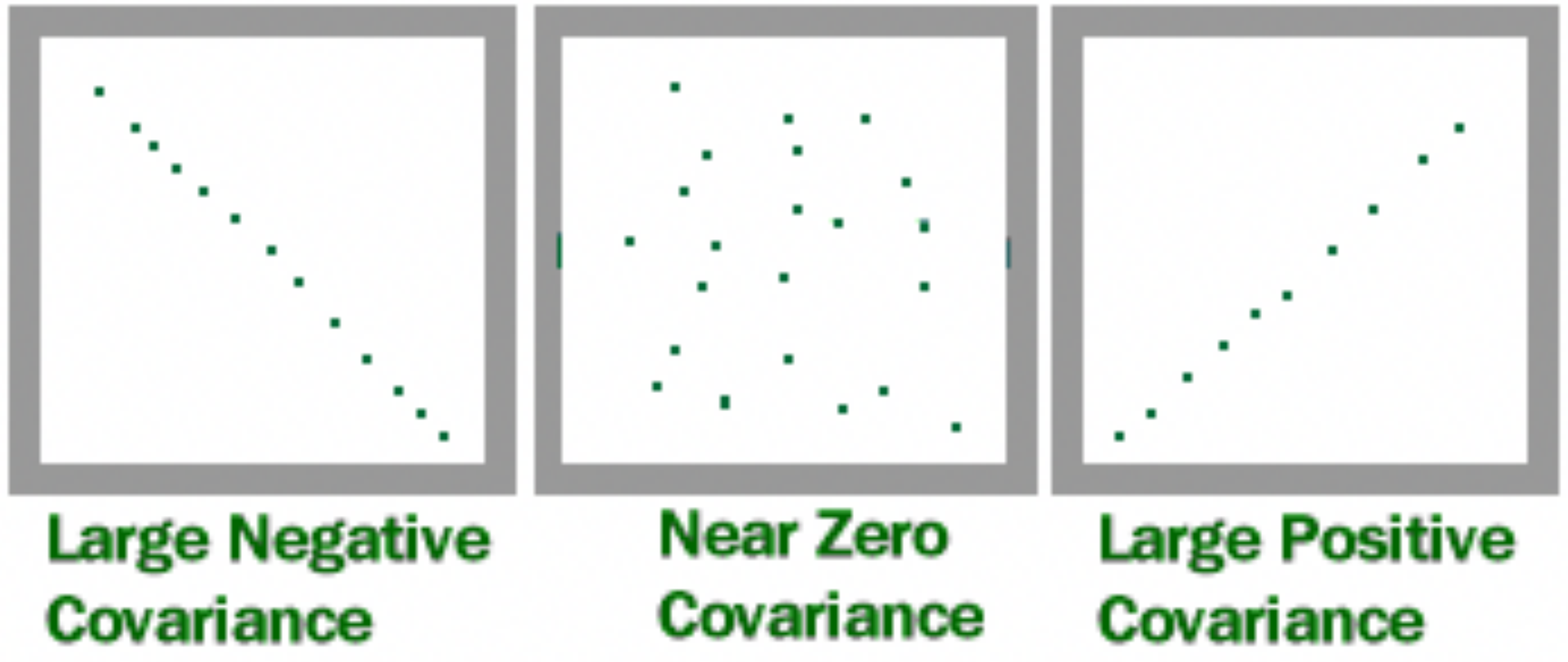


Question: What is the range of $\mathrm{Cov}(X, Y)$?

Correlation

Definition: The correlation of two random variables is

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



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

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

Properties of Variances

- Var$[c] = 0$ for constant c
- Var$[cX] = c^2$Var$[X]$ for constant c
- Var$[X + Y] =$ Var$[X] +$ Var$[Y] + 2$Cov$[X, Y]$
- For **independent** X, Y,
  Var$[X + Y] =$ Var$[X] +$ Var$[Y]$ (**why?**)
    - Recall if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

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,
  $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$
  - Recall if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
  - $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Independence and Decorrelation

- Independent RVs have zero correlation
- Uncorrelated RVs (i.e., 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:** $X \sim$ Uniform$\{-2, -1, 0, 1, 2\}$, $Y = X^2$
    - $\mathbb{E}[XY] = .2(-2 \times 4) + .2(2 \times 4) + .2(-1 \times 1) + .2(1 \times 1) + .2(0 \times 0) = 0$
    - $\mathbb{E}[X] = 0$
    - So Cov$(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$

Summary

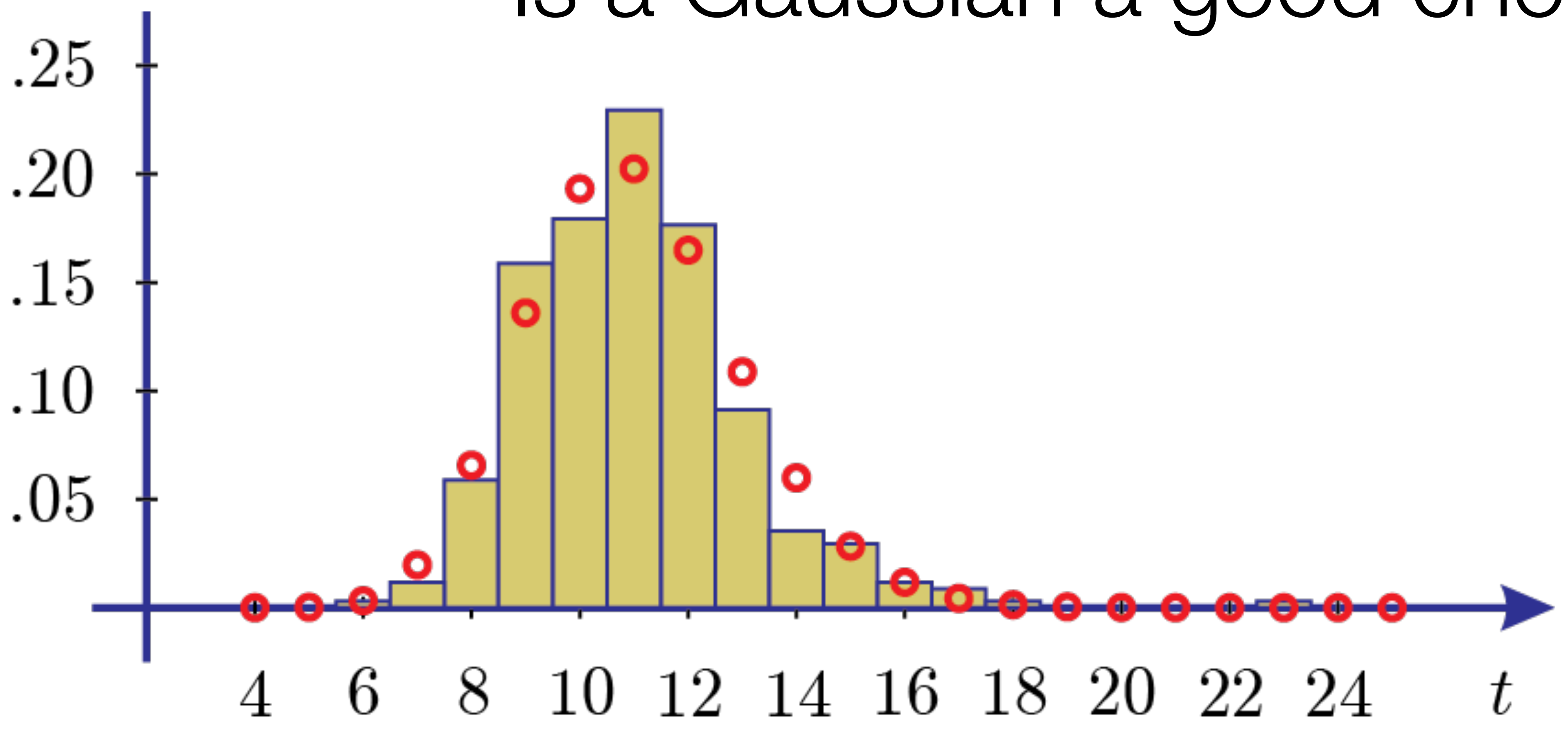
- **Random variables** takes different values with some probability
- The value of one variable can be informative about the value of another
  - Distributions of multiple random variables are described by the **joint** probability distribution (joint PMF or joint PDF)
  - You can have a new distribution 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.

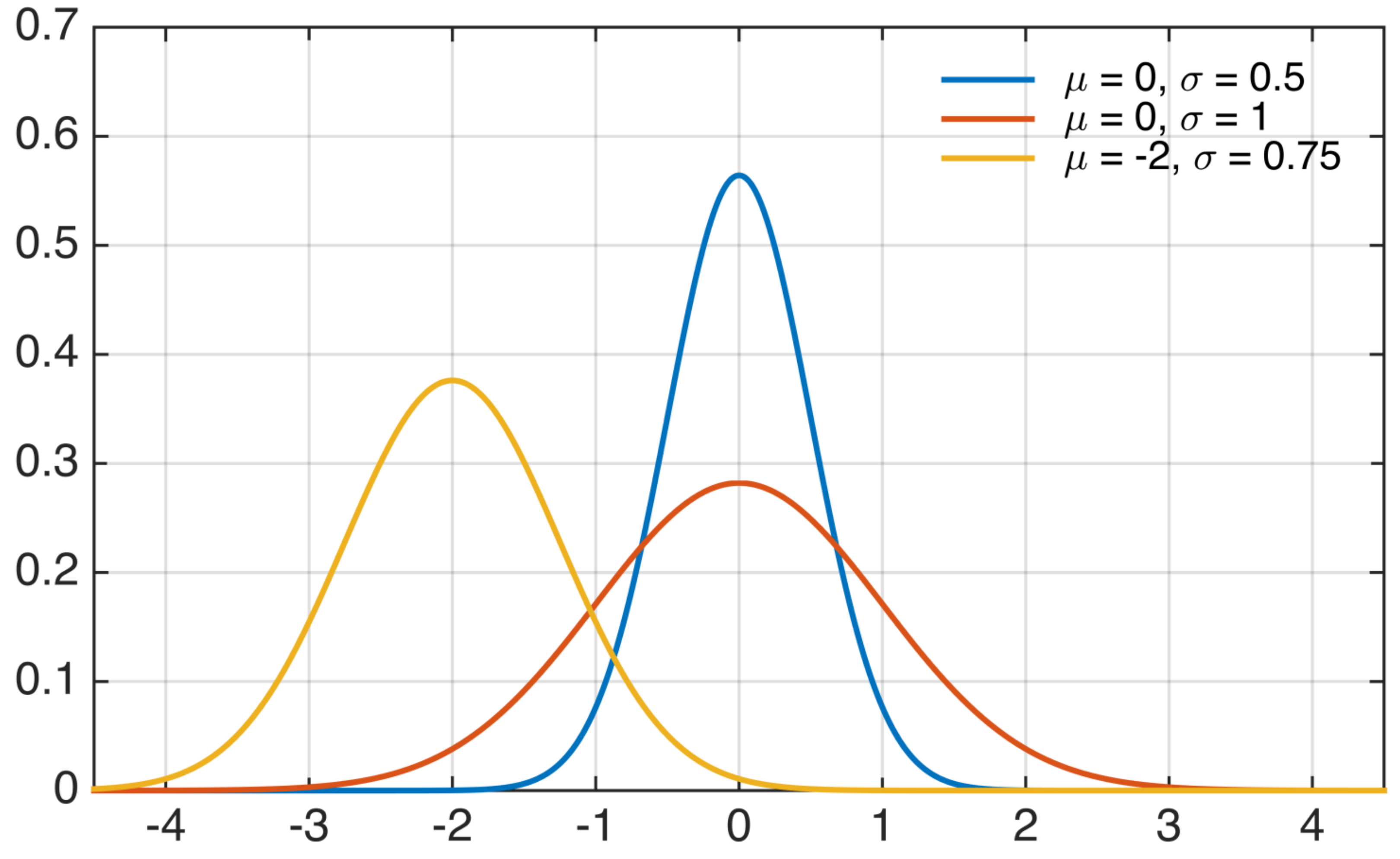
Exercise applying your knowledge

- Let’s revisit the commuting example, and assume we collect continuous commute times
- We want to model commute time as a Gaussian

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

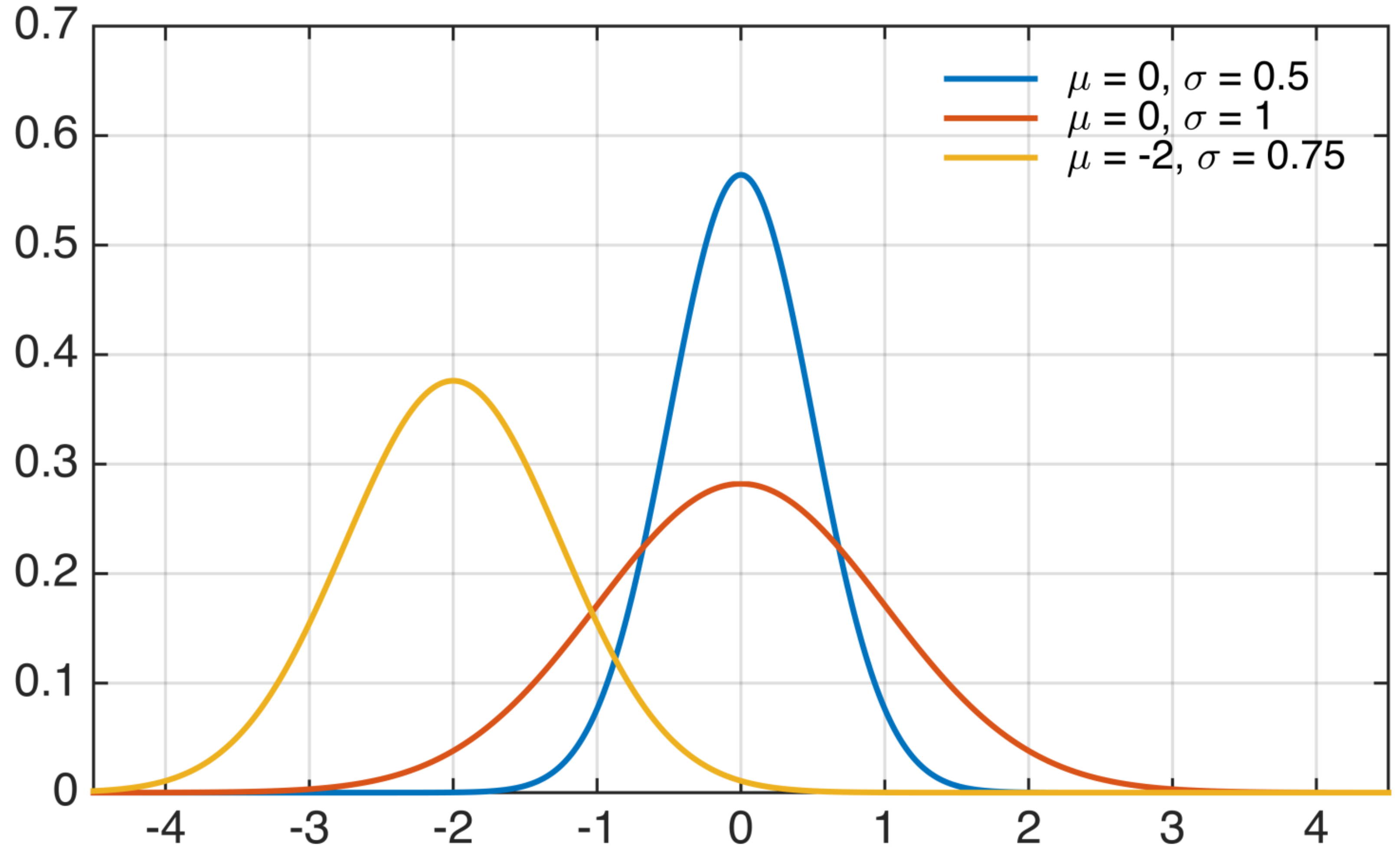
- What parameters do I have to specify (or learn) to model commute times with a Gaussian?
- Is a Gaussian a good choice?

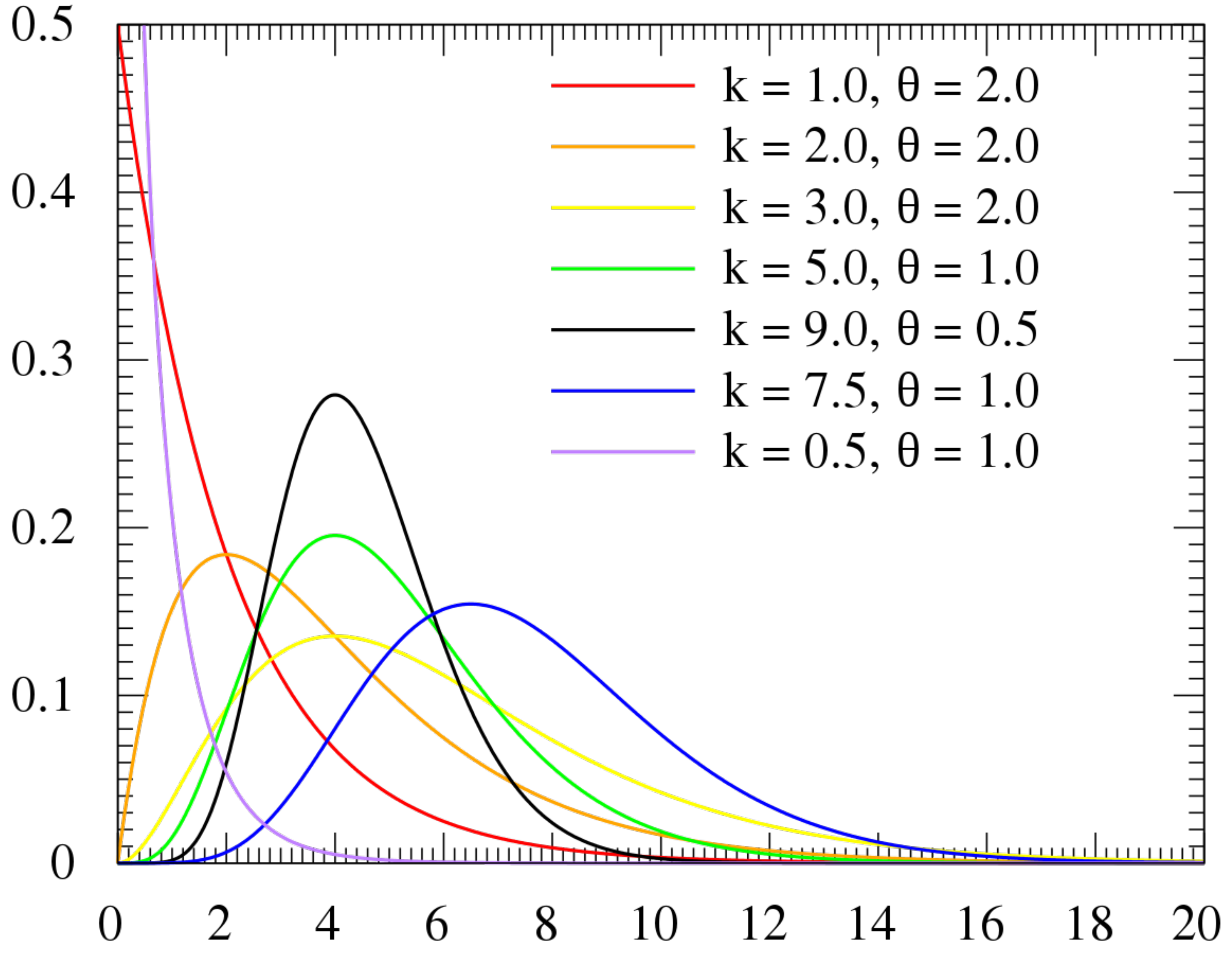


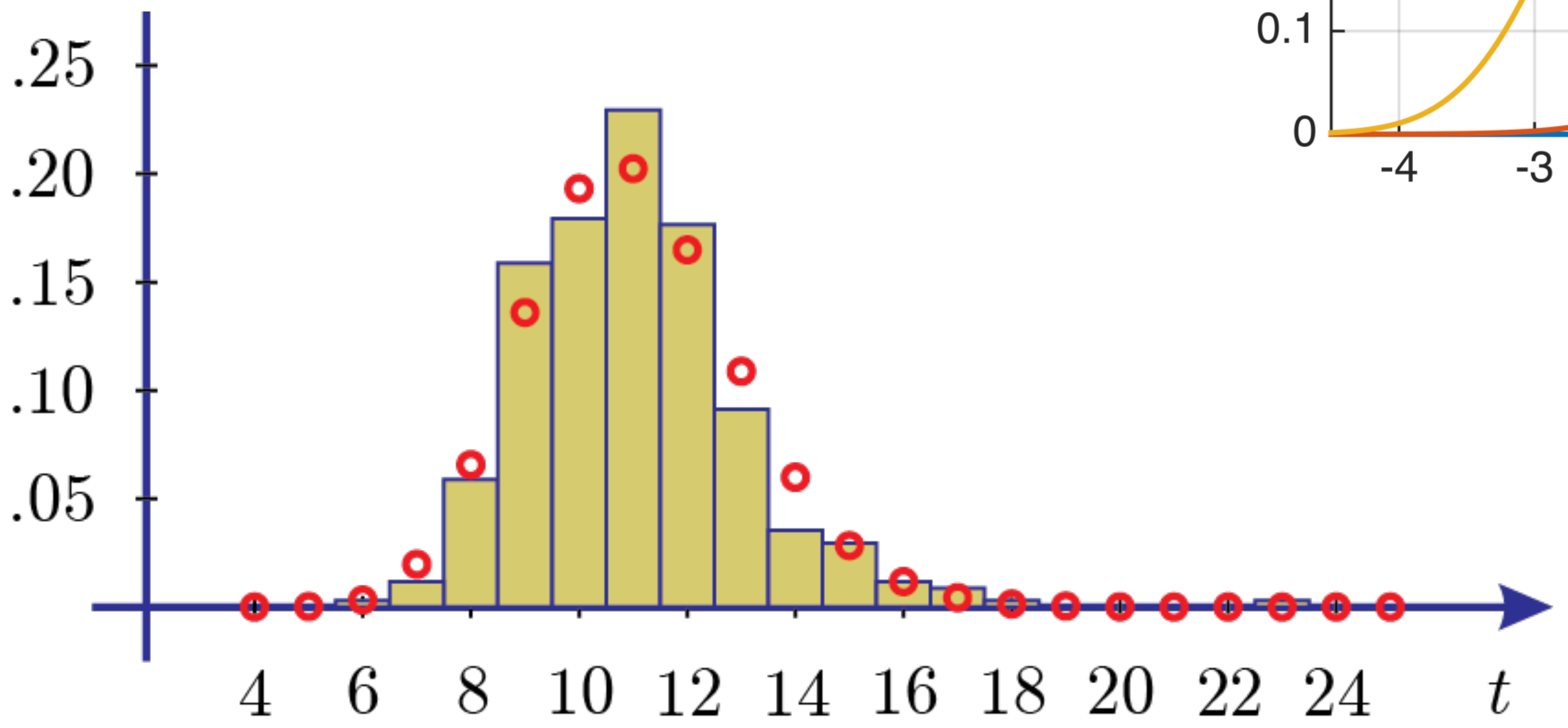


Exercise applying your knowledge

- A better choice is actually what is called a Gamma distribution

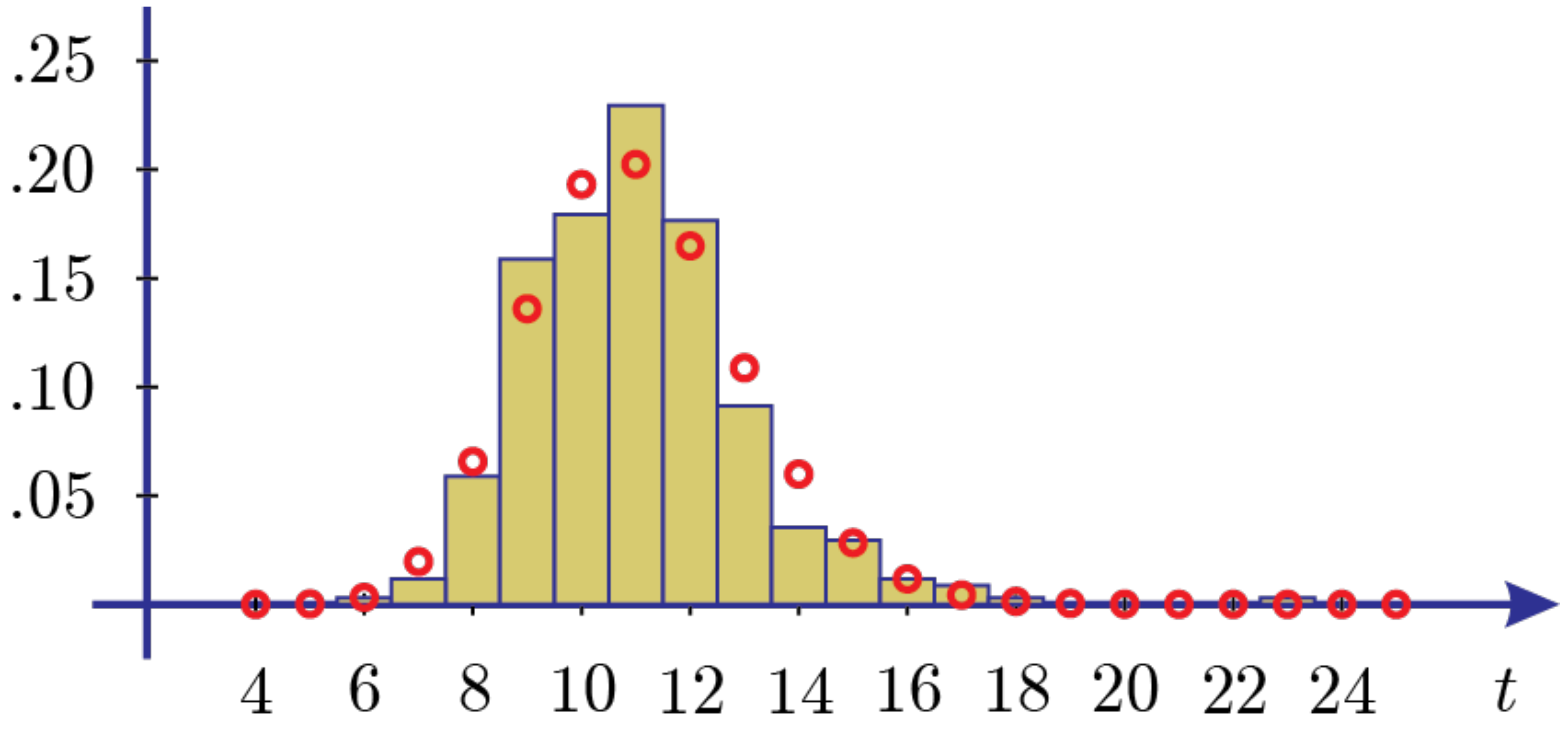






Exercise applying your knowledge

- We can also consider conditional distributions $p(y \mid x)$
- Y is the commute time, let X be the month
- Why is it useful to know $p(y \mid X = \text{Feb})$ and $p(y \mid X = \text{Sept})$?
- What else could we use for X and why pick it?



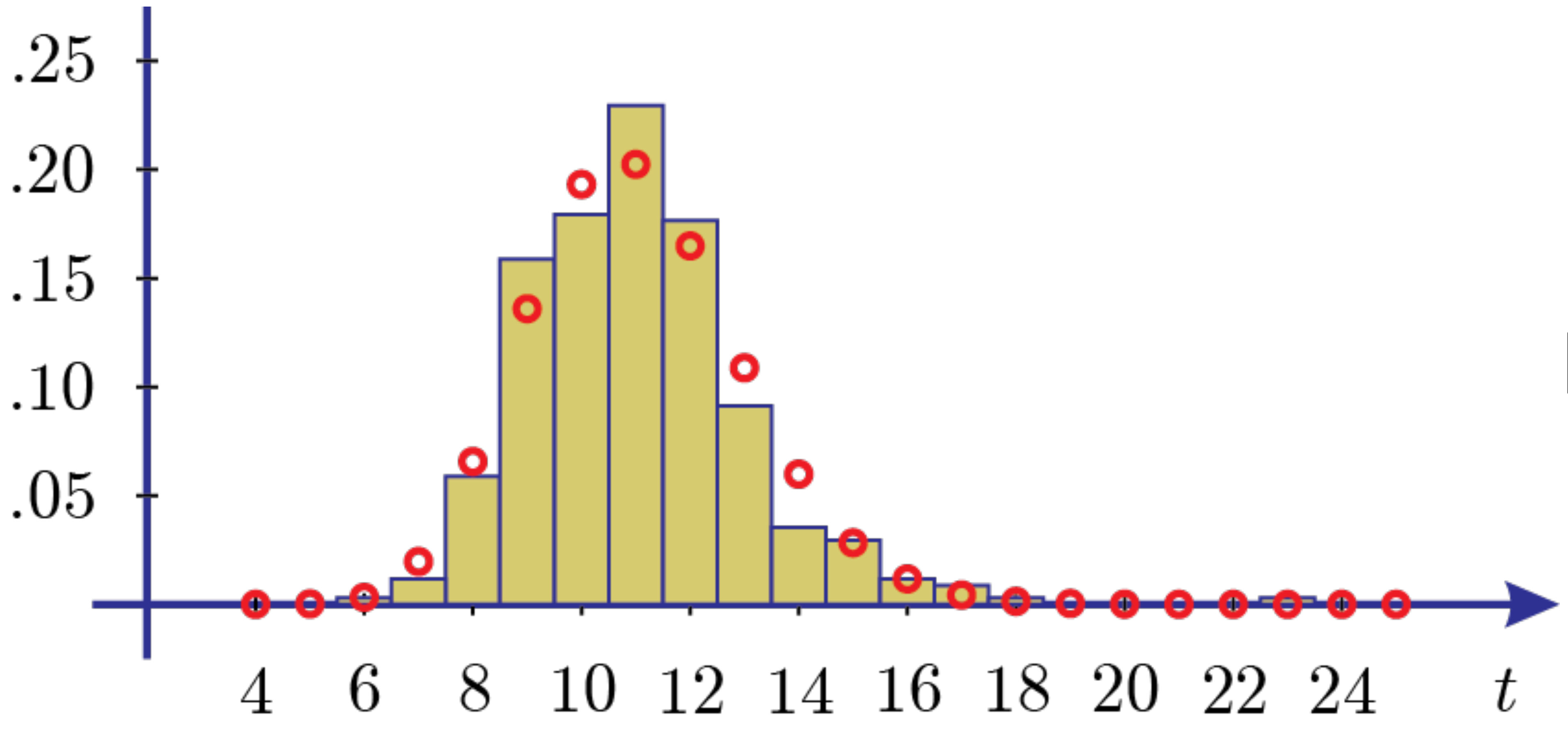
Exercise applying your knowledge

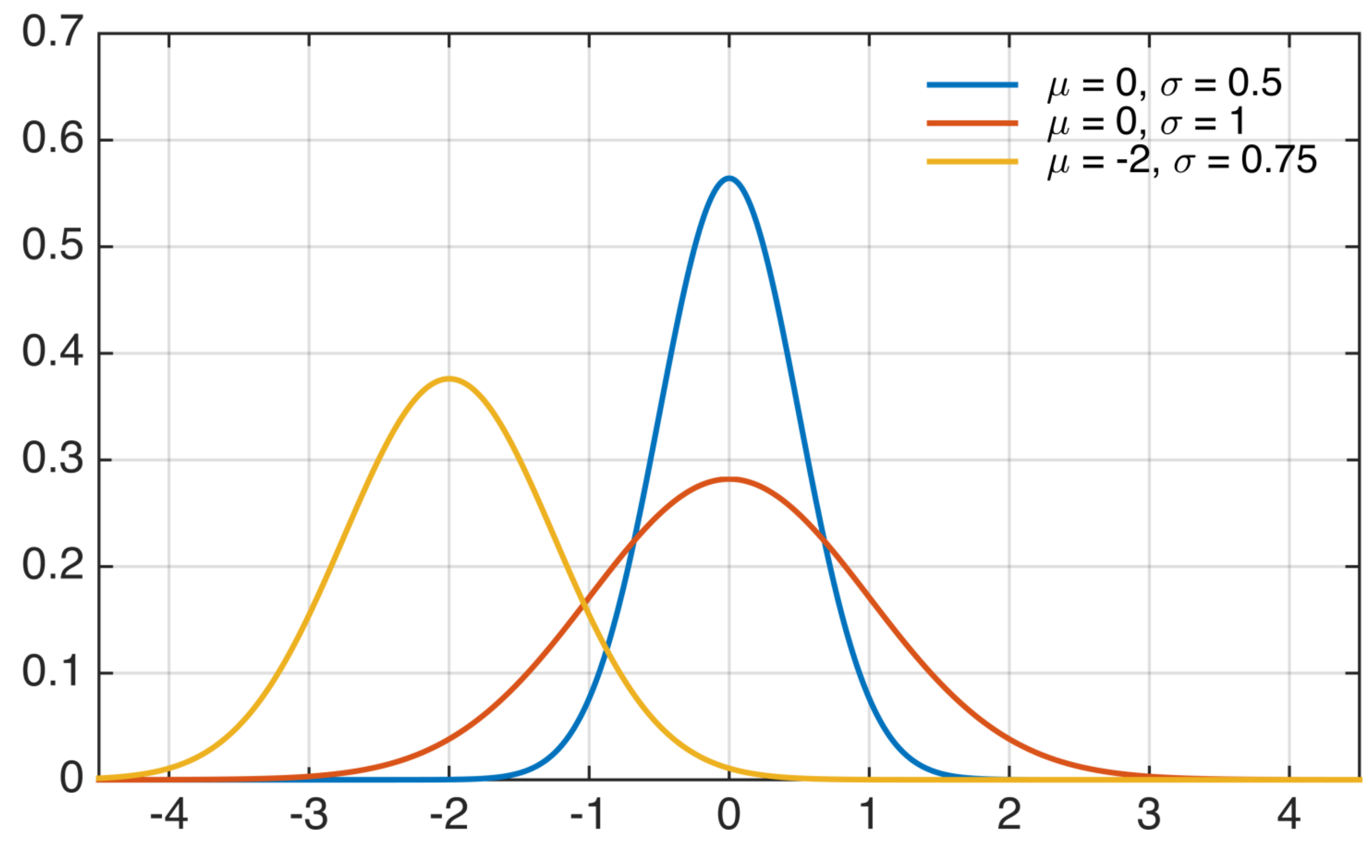
- Let's use a simple X, where it is 1 if it is slippery out and 0 otherwise
- Then we could model two Gaussians, one for the two types of conditions

$$p(y|X=0) = \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$
$$p(y|X=1) = \mathcal{N}\left(\mu_1, \sigma_1^2\right)$$

Gaussian denoted by N

How might μ and σ be different for these two?





Exercise applying your knowledge

- Eventually we will see how to model the distribution over Y using functions of other variables (features) X, e.g,

$$p(y|\mathbf{x}) = \mathcal{N}\left(\mu = \sum_{j=1}^{d} w_i x_i, \sigma^2\right)$$