Estimation: Sample Averages, Bias, and Concentration Inequalities

CMPUT 267: Basics of Machine Learning

Winter 2024

Jan 18 2024

Logistics

- Assignment 1a has been released
- due **Friday, January 26**

Outline

1. Recap
2. Variance and Correlation
3. Estimators
4. Concentration Inequalities
5. Consistency

Recap

- **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)
    - **Conditioning** on a random variable gives a new distribution over others
- **Bayes' Rule**

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

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

- Suppose you have a biased coin: the probability that it comes up heads is not 0.5. Instead, there is a bias - there is a probability to *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 $P(Z = 0.8) = 0.1$.
- Let X and Y be two consecutive flips of the coin

Questions:

- What other outcome space could we consider?
- What kind of distribution is this?
- What other kinds of distributions could we consider?
- Are X and Y independent?
- Are X and Y conditionally independent given Z?

(Ex 9 in the course text)

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 $P(X = Heads)$?

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

$$= P(X = h \mid Z = 0.3) p(Z = 0.3)$$

$$+ P(X = h \mid Z = 0.5) p(Z = 0.5)$$

$$+ P(X = h \mid Z = 0.8) p(Z = 0.8)$$

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

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 $P(X = Heads, Y = Heads) = P(X = Heads)p(Y = Heads)$?

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

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

Example: Coins (4)

- Let Z be the bias of the coin, with $\mathcal{Z} = \{0.3, 0.5, 0.8\}$ 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 $P(X = x)$
  - 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

Why is independence and conditional independence important?

- 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 $f : \mathcal{X} \rightarrow \mathbb{R}$ 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

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

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

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

$$\mathbb{E}[Y] = \mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x) = f(0)p(0) + f(1)p(1) = .5 \times 3 + .5 \times 10 = 6.5$$

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.

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

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

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

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.

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

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

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

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

$$+ \ (1)\Big(p(X = 4) + p(X = 5) + p(X = 6)\Big) \qquad = -1(0.5) + 1(0.5)$$

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

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

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

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?

Linearity of Expectation

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

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

Linearity of Expectation

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

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

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

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

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

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

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

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

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:
  - $\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \mathbb{E}[Y]$
- **Question:** How would you prove these?

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

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

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

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

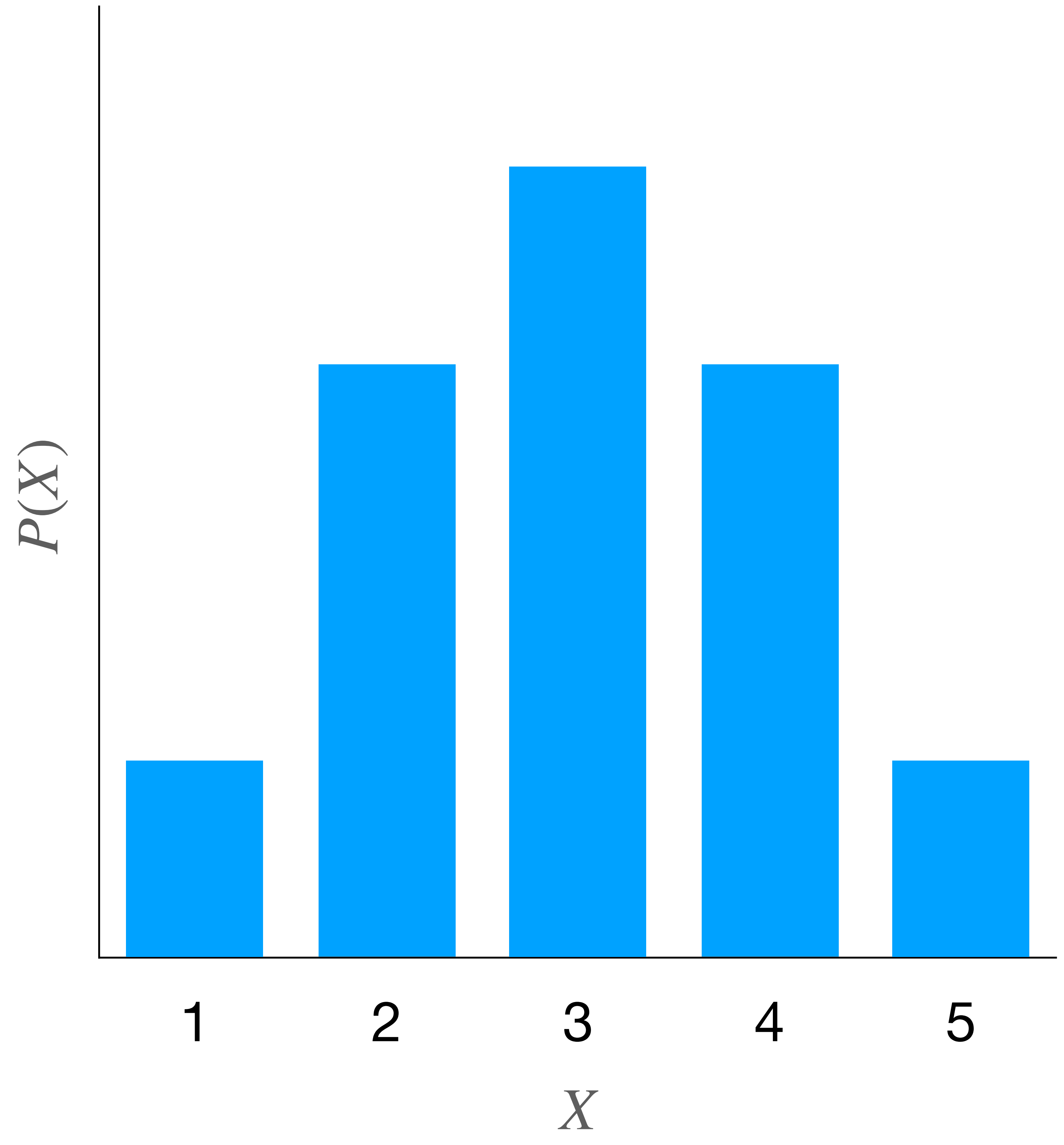
$$= \sum_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}} y p(y \mid x) \right) p(x)$$

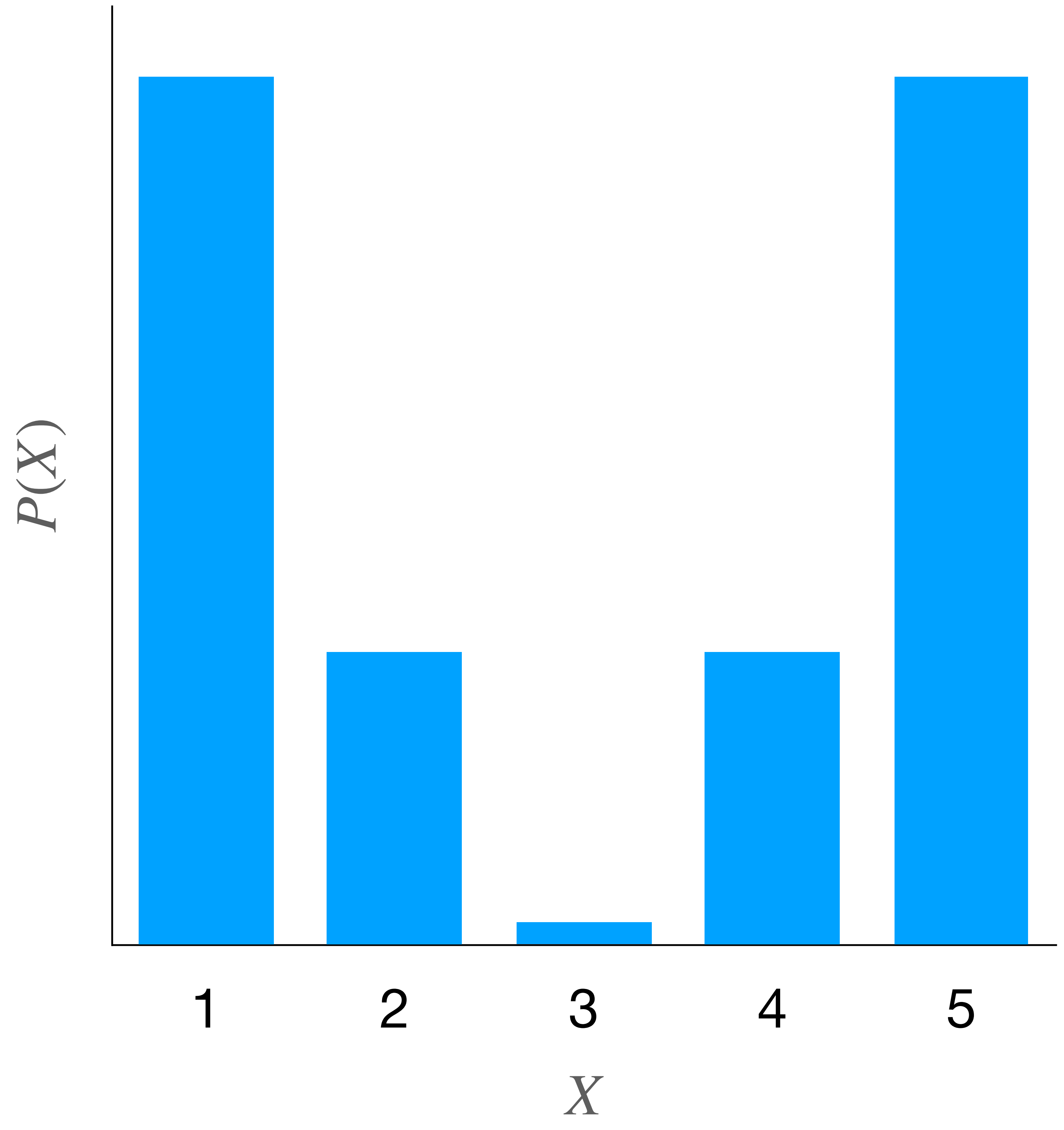
$$= \sum_{x \in \mathcal{X}} \left(\mathbb{E}[Y \mid X = x] \right) p(x) \quad \text{def. E[Y | X = x]}$$

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

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

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,

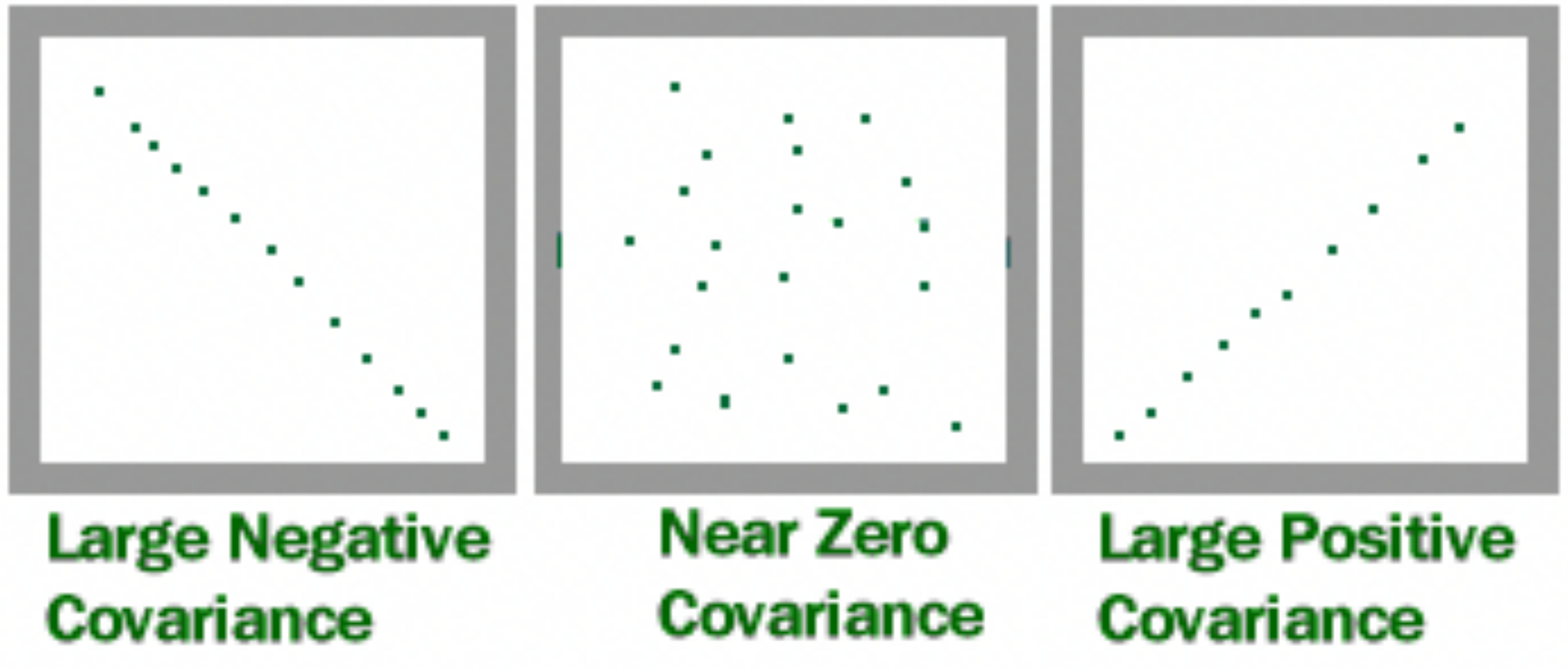
$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

(**why?**)

Covariance

Definition: The covariance of two random variables is

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

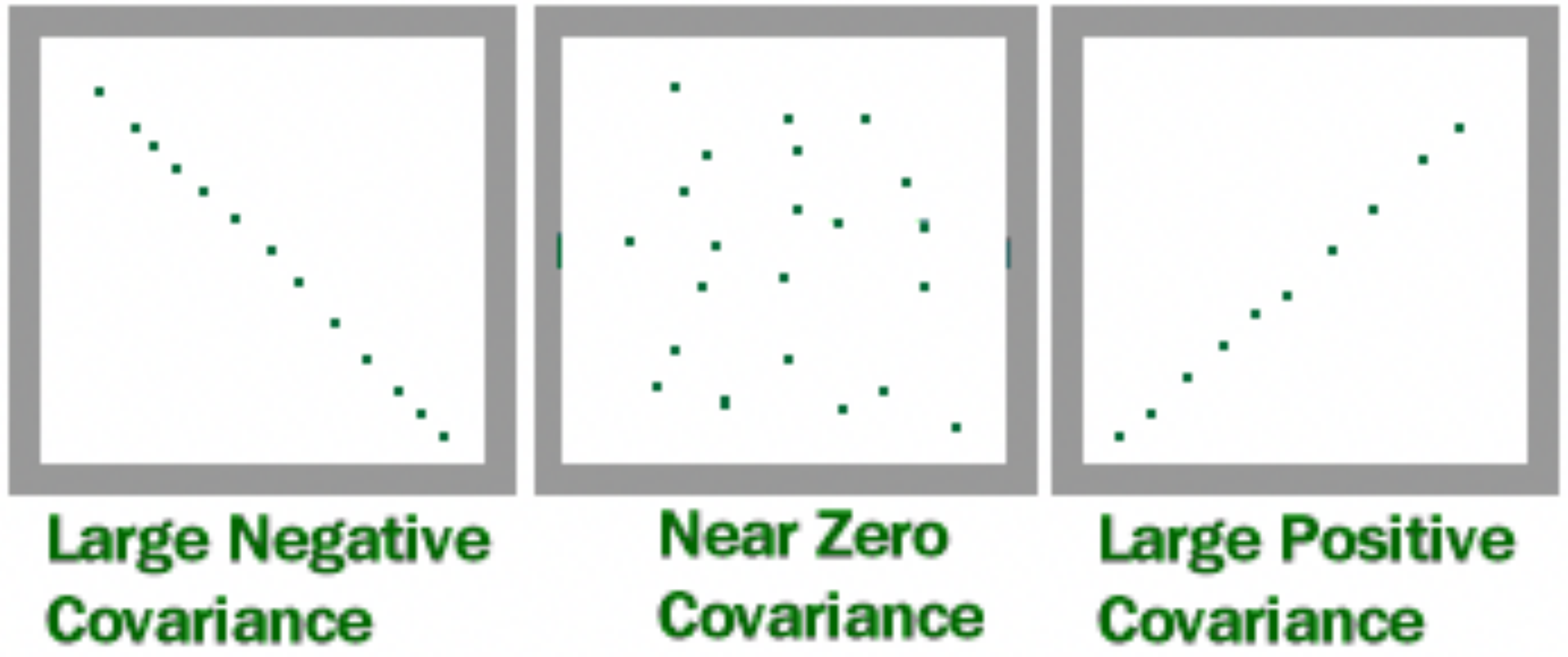


Question: What is the range of $\text{Cov}(X,Y)$?

Correlation

Definition: The correlation of two random variables is

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



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

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

Independence and Decorrelation

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

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

- 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$\{-2, -1, 0, 1, 2\}$, $Y = X^2$
    - $\mathbb{E}[XY] = .2(-2 \times 4) + .2(2 \times 4) + .2(-1 \times 1) + .2(1 \times 1) + .2(0 \times 0)$
    - $\mathbb{E}[X] = 0$
    - So $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$

Properties of Variances

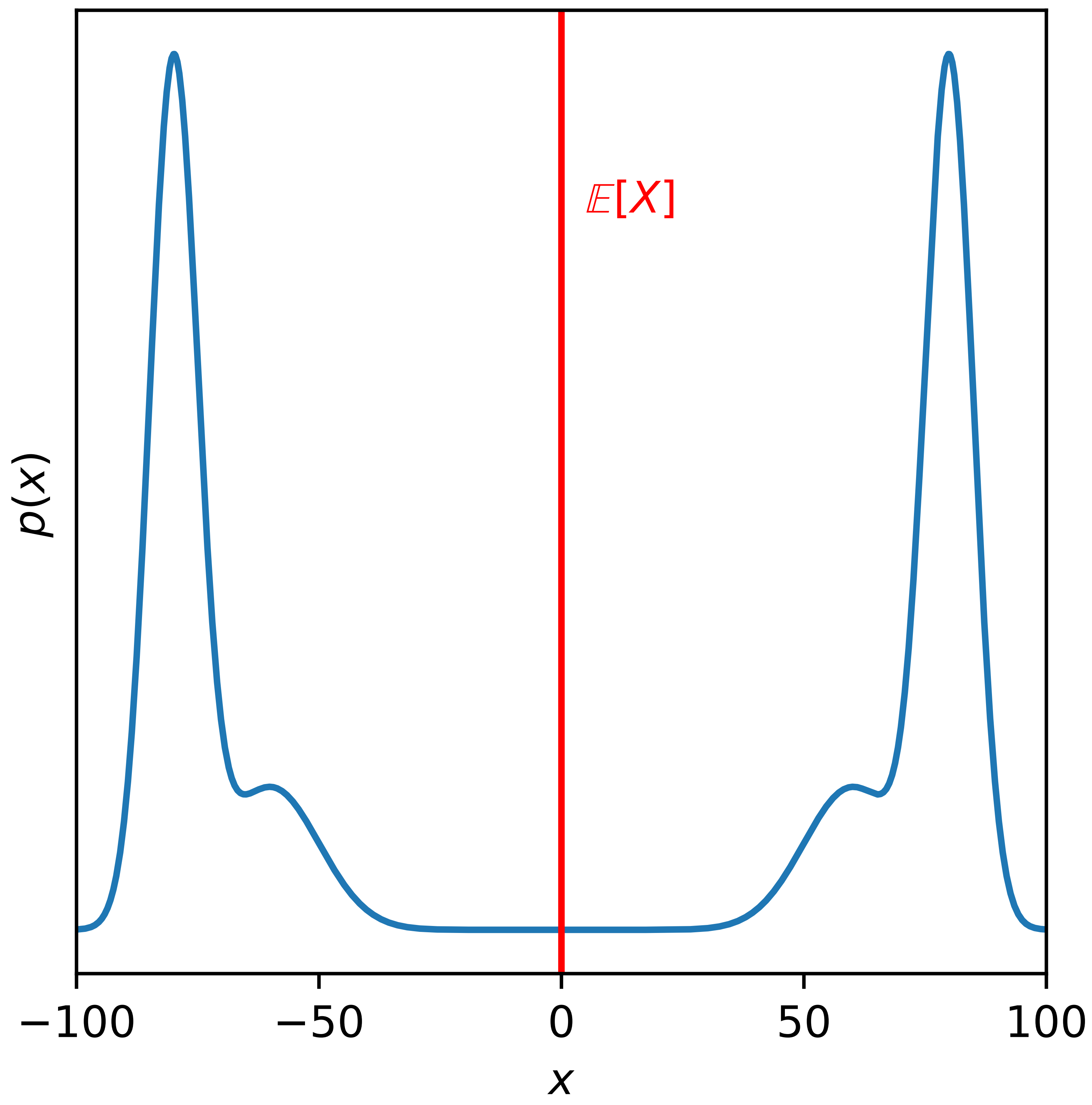
- Var$[c] = 0$ for constant c
- Var$[cX] = c^2$Var$[X]$ for constant c
- Var$[X + Y] =$ Var$[X] +$ Var$[Y] + 2$Cov$[X, Y]$
- For **independent** X, Y,
  Var$[X + Y] =$ Var$[X] +$ Var$[Y]$ (**why?**)

Estimation

Estimators

Definition: An **estimator** is a procedure for estimating an unobserved quantity based on data.

Example: Estimating $\mathbb{E}[X]$ for r.v. $X \in \mathbb{R}$.



Questions:

random variable!

Suppose we can observe a different variable Y. Is Y a good estimator of $\mathbb{E}[X]$ in the following cases? Why or why not?

1. $Y \sim$ Uniform$[0,10]$
2. $Y = \mathbb{E}[X] + Z,$ where $Z \sim$ Uniform$[0,1]$
3. $Y = \mathbb{E}[X] + Z,$ where $Z \sim N(0,100^2)$
4. $Y = X$
5. How would you estimate $\mathbb{E}[X]$?

Bias

Definition: The **bias** of an estimator $\hat{X}$ is its expected difference from the true value of the estimated quantity X:

$$\text{Bias}(\hat{X}) = \mathbb{E}[\hat{X} - X]$$

- Bias can be positive or negative or zero
- When $\text{Bias}(\hat{X}) = 0$, we say that the estimator $\hat{X}$ is **unbiased**

Questions:

What is the **bias** of the following estimators of $\mathbb{E}[X]$?

1. $Y \sim \text{Uniform}[0,10]$
2. $Y = \mathbb{E}[X] + Z$, where $Z \sim \text{Uniform}[0,1]$
3. $Y = \mathbb{E}[X] + Z$, where $Z \sim N(0,100^2)$
4. $Y = X$

Independent and Identically Distributed (i.i.d.) Samples

- We usually won't try to estimate anything about a distribution based on only a single sample
- Usually, we use **multiple samples** from the **same distribution**
  - *Multiple samples:* This gives us more information
  - *Same distribution:* We want to learn about a single population
- One additional condition: the samples must be **independent** (**why?**)

Definition: When a set of random variables $X_1, X_2, \ldots$ are all independent, and each has the same distribution $X \sim F$, we say they are **i.i.d.** (independent and identically distributed), written

$$X_1, X_2, \ldots \overset{i.i.d.}{\sim} F.$$

Estimating Expected Value via the Sample Mean

Example: We have n i.i.d. samples from the same distribution F,

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} F,$$

with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ for each X_i.

We want to estimate μ.

Let's use the **sample mean** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate μ.

Question: Is this estimator **unbiased**?

Question: Are **more samples** better? Why?

Estimating Expected Value via the Sample Mean

Example: We have n i.i.d. samples from the same distribution F,

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} F,$$

with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ for each X_i.

We want to estimate μ.

Let's use the **sample mean** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate μ.

Question: Is this estimator **unbiased**?

Question: Are **more samples** better? Why?

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu\,. \quad \blacksquare$$

Variance of the Estimator

- Intuitively, more samples should make the estimator "closer" to the estimated quantity
- We can formalize this intuition partly by characterizing the **variance $\mathrm{Var}[\hat{X}]$ of the estimator itself**.
  - The variance of the estimator should decrease as the number of samples increases
- **Example:** $\bar{X}$ for estimating μ:
  - The variance of the estimator shrinks linearly as the number of samples grows.

Variance of the Estimator

- Intuitively, more samples should make the estimator "closer" to the estimated quantity
- We can formalize this intuition partly by characterizing the **variance $\text{Var}[\hat{X}]$ of the estimator itself**.
  - The variance of the estimator should decrease as the number of samples increases
- **Example:** $\bar{X}$ for estimating μ:
  - The variance of the estimator shrinks linearly as the number of samples grows.

$$\begin{aligned}
\text{Var}[\bar{X}] &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Xi\right] \\
&= \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{1}{n^2}n\sigma^2 \quad = \frac{1}{n}\sigma^2.
\end{aligned}$$

Concentration Inequalities

- We want to obtain a confidence interval around our estimate - we want the difference from the expected value to be small, and be consistently small.
- We would like to be able to claim $\Pr\left(\left|\bar{X}-\mu\right| < \epsilon\right) > 1-\delta$ for some $\delta, \epsilon > 0$
- This tells us that $\mathbb{E}[\bar{X}] \in \{\bar{X}-\epsilon, \bar{X}+\epsilon\}$ with a large probability, $1-\delta$
- Confidence level: δ, width of interval: ϵ
- $\Pr\left(\left|\bar{X}-\mu\right| < \epsilon\right) > 1-\delta$ for *any* $\delta, \epsilon > 0$ that we pick (**why?**)
- $\mathrm{Var}[\bar{X}] = \frac{1}{n}\sigma^2$ means that with "enough" data we can get close to the expected value.
- Suppose we have $n = 10$ samples, and we know $\sigma^2 = 81$; so $\mathrm{Var}[\bar{X}] = 8.1$.
- **Question:** What is $\Pr\left(\left|\bar{X}-\mu\right| < 2\right)$?

Variance Is Not Enough

Knowing $\mathrm{Var}[\bar{X}] = 8.1$ is **not enough** to compute $\Pr(|\bar{X} - \mu| < 2)$!

Examples:

$$p(\bar{x}) = \begin{cases} 0.9 & \text{if } \bar{x} = \mu \\ 0.05 & \text{if } \bar{x} = \mu \pm 9 \end{cases} \implies \mathrm{Var}[\bar{X}] = 8.1 \text{ and } \Pr(|\bar{X} - \mu| < 2) = 0.9$$

$$p(\bar{x}) = \begin{cases} 0.999 & \text{if } \bar{x} = \mu \\ 0.0005 & \text{if } \bar{x} = \mu \pm 90 \end{cases} \implies \mathrm{Var}[\bar{X}] = 8.1 \text{ and } \Pr(|\bar{X} - \mu| < 2) = 0.999$$

$$p(\bar{x}) = \begin{cases} 0.1 & \text{if } \bar{x} = \mu \\ 0.45 & \text{if } \bar{x} = \mu \pm 3 \end{cases} \implies \mathrm{Var}[\bar{X}] = 8.1 \text{ and } \Pr(|\bar{X} - \mu| < 2) = 0.1$$

Hoeffding's Inequality

Theorem: Hoeffding's Inequality

Suppose that $X_1, \ldots, X_n$ are distributed i.i.d, with $a \leq X_i \leq b$.
Then for any $\epsilon > 0$,

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Equivalently, $\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}\right) \geq 1 - \delta.$

Chebyshev's Inequality

Theorem: **Chebyshev's Inequality**

Suppose that $X_1, \ldots, X_n$ are distributed i.i.d. with variance σ^2.
Then for any $\epsilon > 0$,

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Equivalently, $\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq \sqrt{\frac{\sigma^2}{\delta n}}\right) \geq 1 - \delta.$

When to Use Chebyshev, When to Use Hoeffding?

- If $a \leq X_i \leq b$, then $\mathrm{Var}[X_i] \leq \frac{1}{4}(b-a)^2$
- Hoeffding's inequality gives $\epsilon = (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}} = \sqrt{\frac{\ln(2/\delta)}{2}}(b-a)\sqrt{\frac{1}{n}}$;

  Chebyshev's inequality gives $\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} \leq \sqrt{\frac{(b-a)^2}{4\delta n}} = \frac{1}{2\sqrt{\delta}}(b-a)\sqrt{\frac{1}{n}}$
- **Hoeffding's inequality** gives a **tighter bound***, but it can only be used on **bounded** random variables
  - $*$ whenever $\sqrt{\frac{\ln(2/\delta)}{2}} < \frac{1}{2\sqrt{\delta}} \iff \delta < \sim 0.232$
- **Chebyshev's inequality** can be applied even for **unbounded** variables

Consistency

Definition: A sequence of random variables X_n **converges in probability** to a random variable X (written $X_n \xrightarrow{p} X$) if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0.$$

Definition: An estimator $\hat{X}$ for a quantity X is **consistent** if $\hat{X} \xrightarrow{p} X$.

Weak Law of Large Numbers

Theorem: Weak Law of Large Numbers

Let $X_1, \ldots, X_n$ be distributed i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$.

Then the **sample mean**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a **consistent estimator** for μ.

Proof:

1. We have already shown that $\mathbb{E}[\bar{X}] = \mu$
2. By Chebyshev,
   $$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$
   for arbitrary $\epsilon > 0$
3. Hence $\lim_{n\to\infty} \Pr\left(\left|\bar{X} - \mu\right| \geq \epsilon\right) = 0$
   for any $\epsilon > 0$
4. Hence $\bar{X} \xrightarrow{p} \mu$. ∎

Summary

- The **variance** $\mathrm{Var}[X]$ of a random variable X is its expected squared distance from the mean
- An **estimator** is a random variable representing a procedure for estimating the value of an unobserved quantity based on observed data
- **Concentration inequalities** let us bound the probability of a given estimator being at least ϵ from the estimated quantity
- An estimator is **consistent** if it **converges in probability** to the estimated quantity