Estimation: Sample Averages, Bias, and Concentration Inequalities

CMPUT 296: Basics of Machine Learning

Textbook §3.1-3.3

Logistics

Reminders:

- Thought Question 1 (due **Thursday, September 17**)
- Assignment 1 (due **Thursday, September 24**)

New:

- Group Slack channel: **#cmput296-fall20** (on Amii workspace)

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
- X is **independent** of Y: conditioning on X does **not** give a new distribution over Y
  - X is **conditionally independent** of Y given Z: $P(Y \mid X, Z) = P(Y \mid Z)$
- The **expected value** of a random variable is an **average** over its values, **weighted** by the probability of each value

Outline

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

Variance

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

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

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

Equivalently,

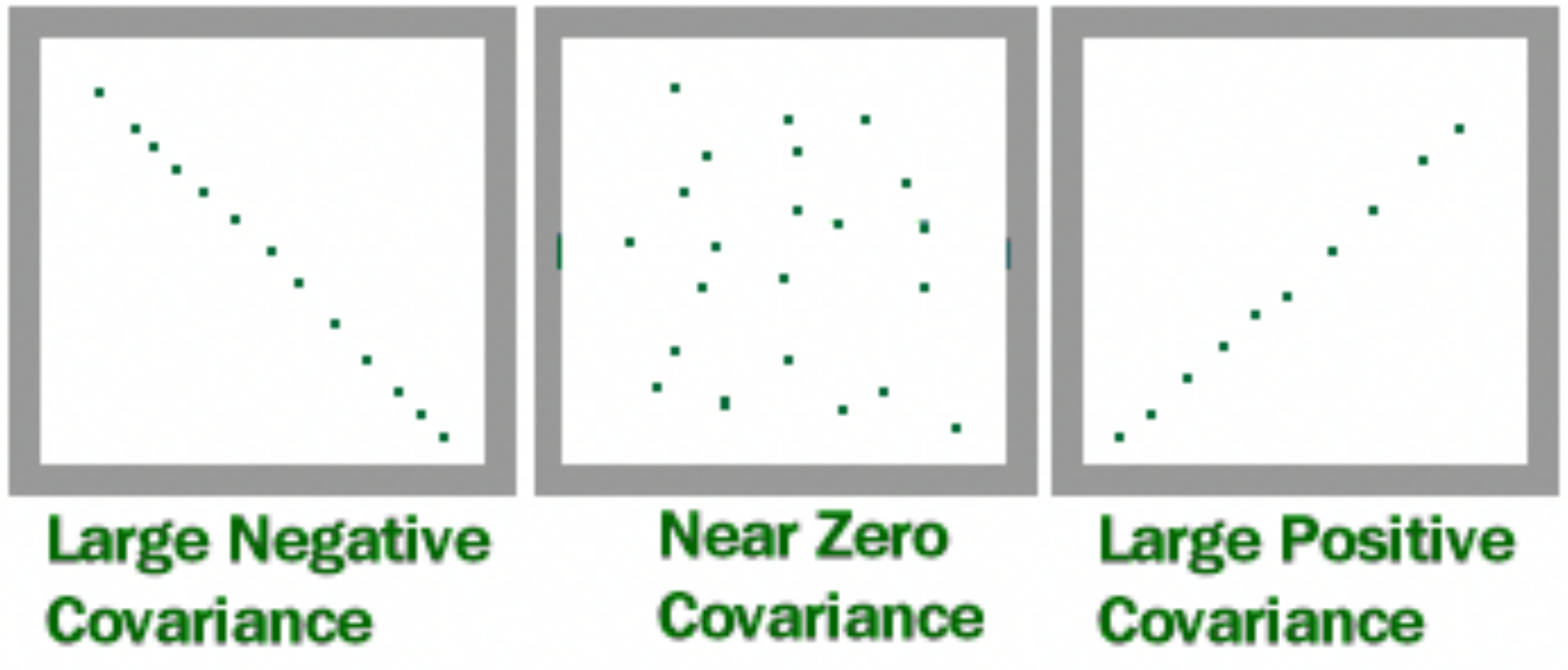
$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

(**why?**)

Covariance

Definition: The covariance of two random variables is

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

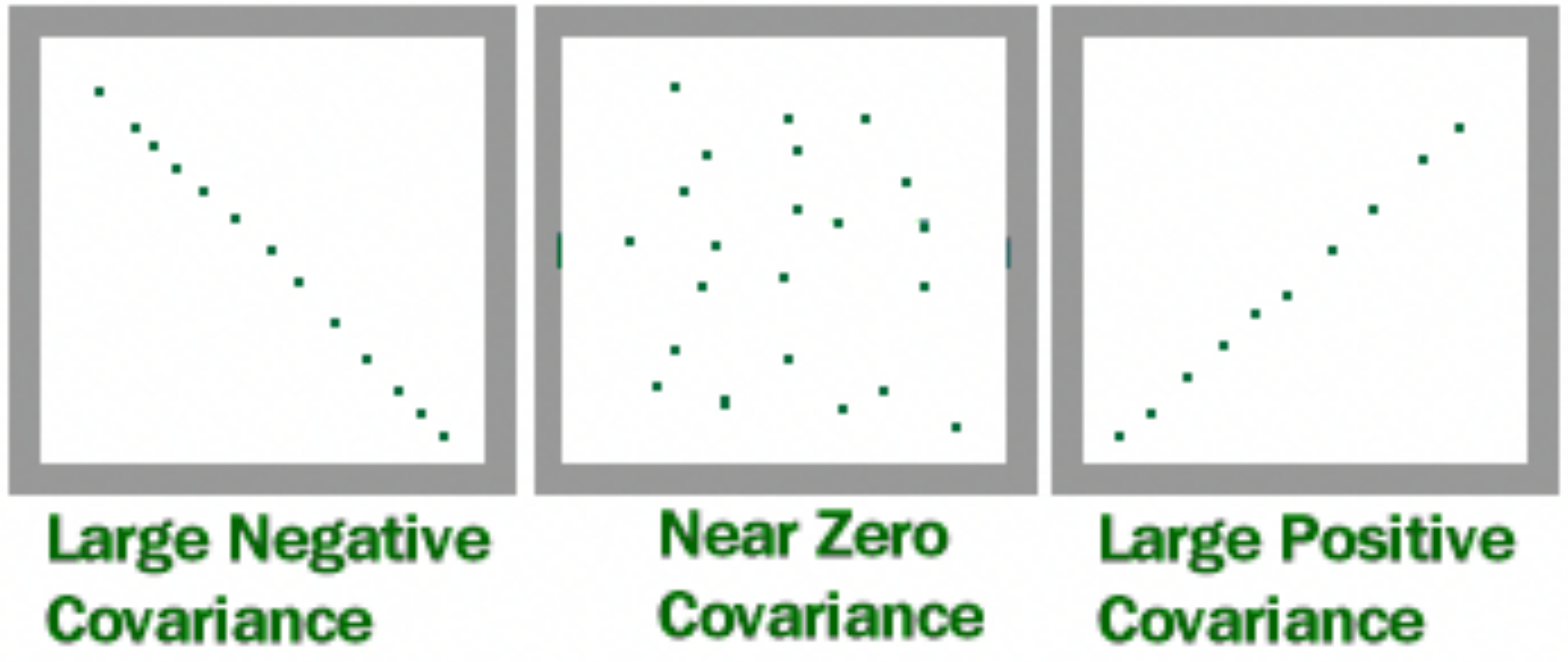


Question: What is the range of $\mathrm{Cov}(X, Y)$?

Correlation

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

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



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

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

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 $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$

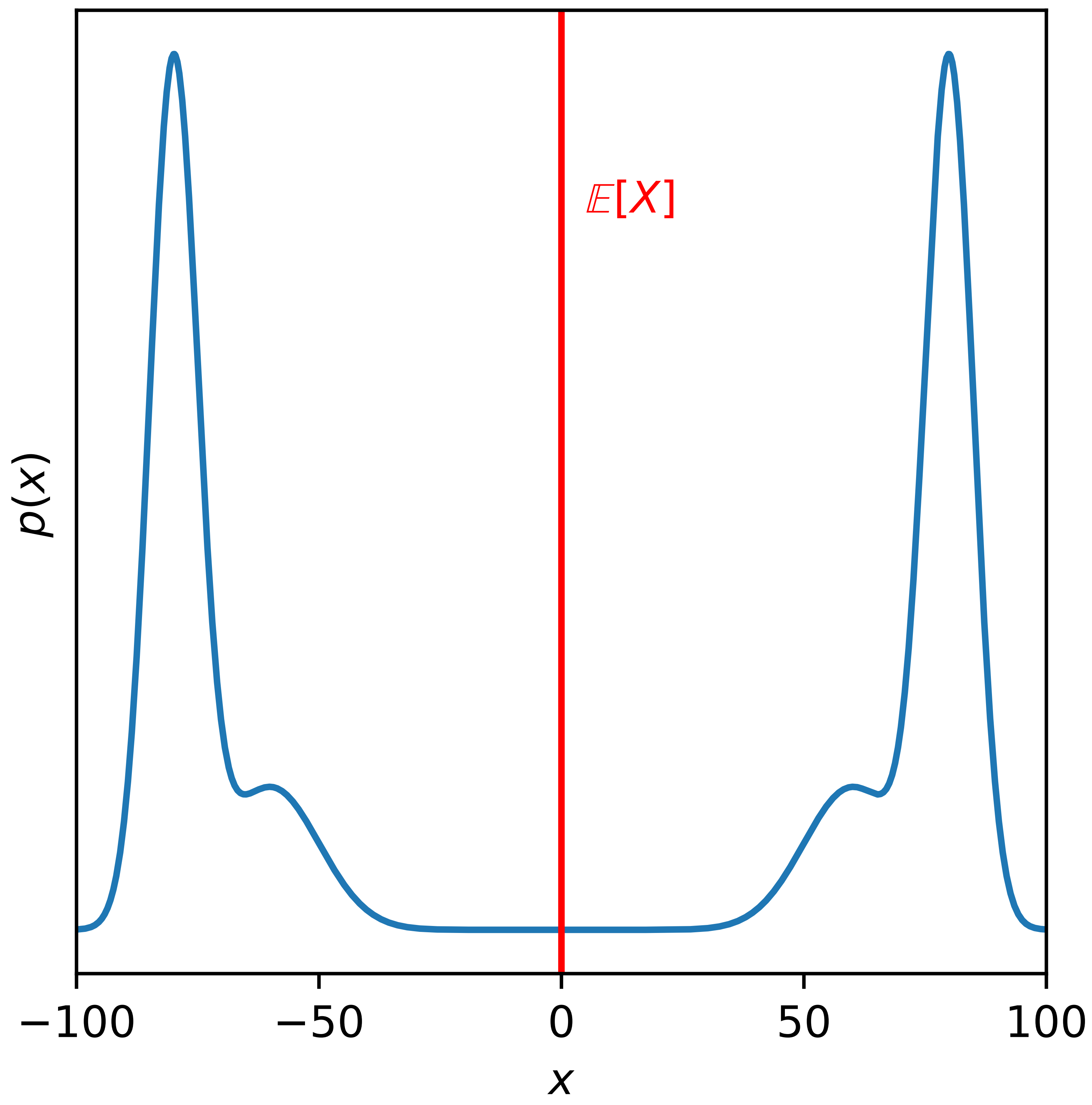
Properties of Variances

- $\text{Var}[c] = 0$ 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?**)

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. $Y \sim$ 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. $Y = X$

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 are $X_1, X_2, \ldots$ 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?

$$\begin{aligned}
\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu \\
&= \mu . \qquad \blacksquare
\end{aligned}$$

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.

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

Concentration Inequalities

- We would like to be able to claim $\Pr\left(\left|\bar{X} - \mu\right| < \epsilon\right) > 1 - \delta$ for some $\delta, \epsilon > 0$
- $\text{Var}[\bar{X}] = \frac{1}{n}\sigma^2$ means that with "enough" data, $\Pr\left(\left|\bar{X} - \mu\right| < \epsilon\right) > 1 - \delta$ for *any* $\delta, \epsilon > 0$ that we pick (**why?**)
- Suppose we have $n = 10$ samples, and we know $\sigma^2 = 81$; so $\text{Var}[\bar{X}] = 8.1$.
- **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