Review for Quiz

Chapter 2 (Probability)
Chapter 3 (Estimation):
Bias, Variance, Concentration Inequalities

CMPUT 296: Basics of Machine Learning

Logistics

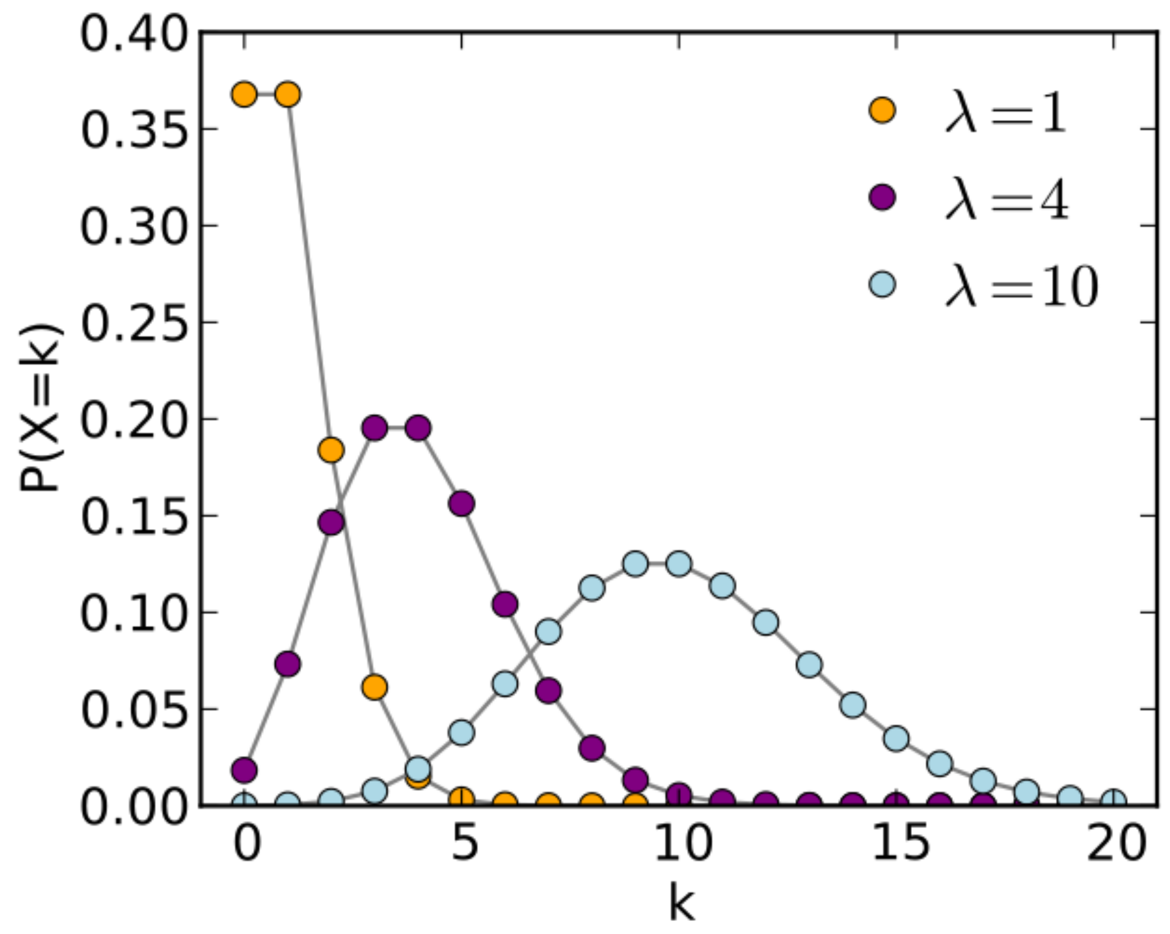
- Quiz during class on Thursday; come to regular zoom lecture earlier
- TAs will go over Assignment 1 in Lab on Wednesday
- Any questions/issues with starting Assignment 2?

Language of Probabilities

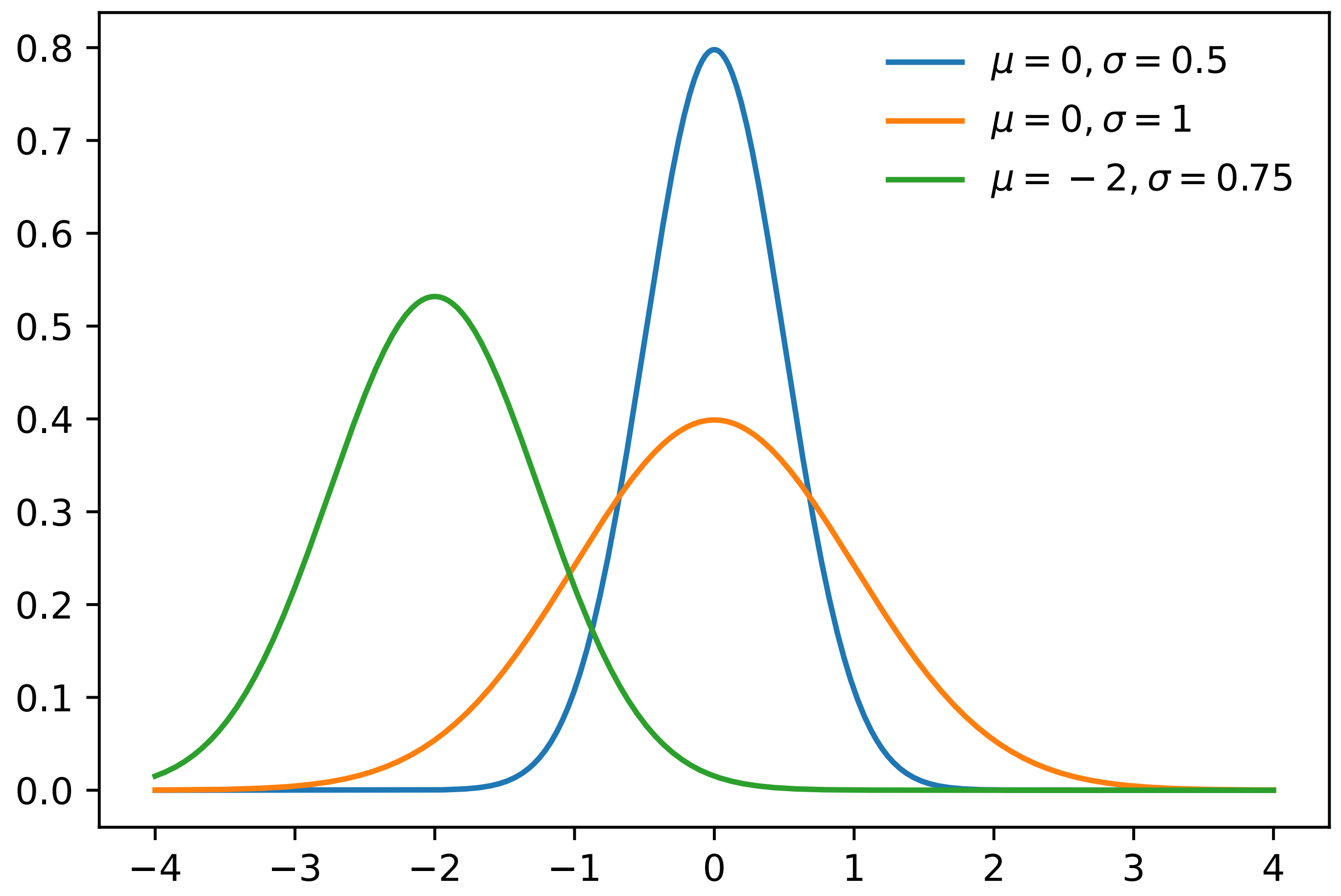
- Define random variables, and their distributions
  - Then can formally reason about them
- Express our beliefs about behaviour of these RVs, and relationships to other RVs
- Examples:
  - $p(x)$ Gaussian means we believe X is Gaussian distributed
  - $p(y \mid X = x)$—or written $p(y \mid x)$— is Gaussian says that conditioned on x, then y is Gaussian; but $p(y)$ might not be Gaussian
  - $p(w)$ and $p(w \mid \text{Data})$

PMFs and PDFs

- Discrete RVs have PMFs
  - outcome space: e.g, $\Omega = \{1,2,3,4,5,6\}$
  - examples pmfs: probability tables, Poisson $p(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$



- Continuous RVs have PDFs
  - outcome space: e.g., $\Omega = [0,1]$
  - example pdf: Gaussian, Gamma



A few questions

- Do PMFs p(x) have to output values between [0,1]?
- Do PDFs p(x) have to output values between [0,1]?
- What other condition(s) are put on a function p to make it a valid pmf or pdf?
- Is the following function a pdf or a pmf?
- $p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$ i.e., $p(x) = \dfrac{1}{b-a}$ for $x \in [a, b]$

How would you define a uniform distribution for a discrete RV

- Imagine $x \in \{1,2,3,4,5\}$
- What is the uniform pmf for this outcome space?
- $$p(x) = \begin{cases} \frac{1}{5} & \text{if } x \in \{1,2,3,4,5\}, \\ 0 & \text{otherwise.} \end{cases}$$

How do you answer the probabilistic question?

- For continuous RV X with a uniform distribution and outcome space [0,10], what is the probability that X is greater than 7?

$$\Pr(X > 7) = \int_7^{10} p(x)dx = \int_7^{10} \frac{1}{10}dx$$

$$= \frac{1}{10}\int_7^{10} dx = \frac{1}{10}x\,|_7^{10}$$

$$= \frac{3}{10}$$

-

Multivariate Setting

- Conditional distribution, $p(y \mid x) = \dfrac{p(x, y)}{p(x)}$, Marginal $p(y) = \sum_{x \in \mathcal{X}} p(x, y)$
- Chain Rule $p(x, y) = p(y \mid x)p(x) = p(x \mid y)p(y)$
- Bayes Rule $p(y \mid x) = \dfrac{p(x \mid y)p(y)}{p(x)}$
- Law of total probability $p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x)$
- **Question**: How do you get the law of total probability from the chain rule?

Conditional Expectations

Definition:

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

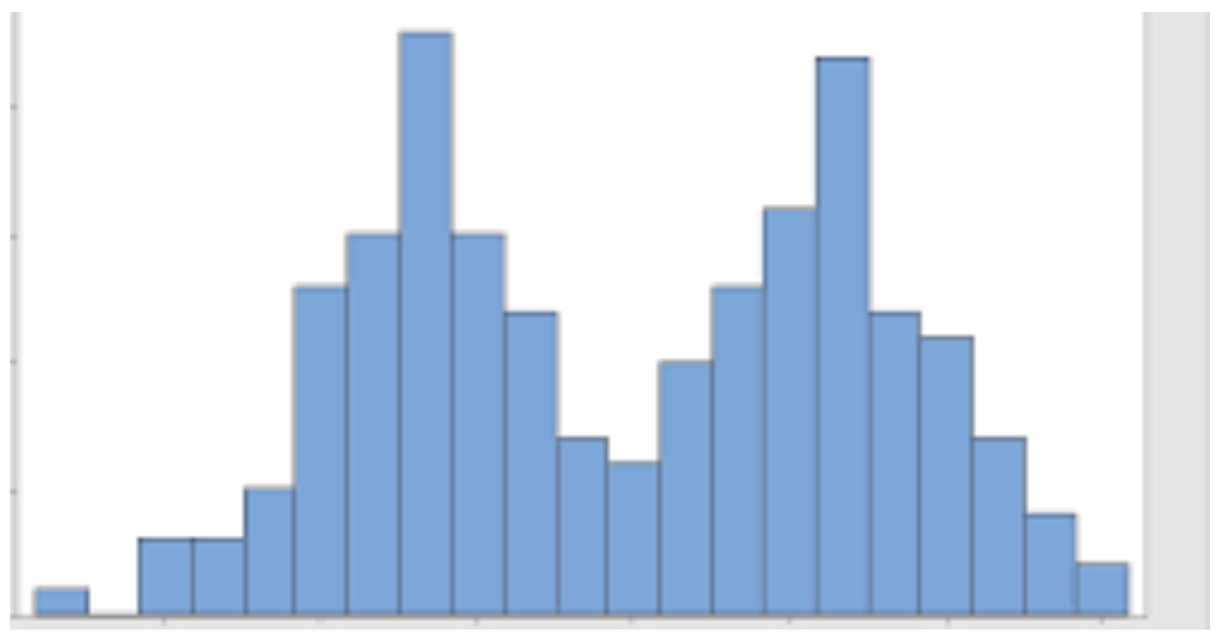
$$\mathbb{E}[Y \mid X = x] = \begin{cases} \sum_{y \in \mathcal{Y}} y p(y \mid x) & \text{if } Y \text{ is discrete,} \\ \int_{\mathcal{Y}} y p(y \mid x) \, dy & \text{if } Y \text{ is continuous.} \end{cases}$$

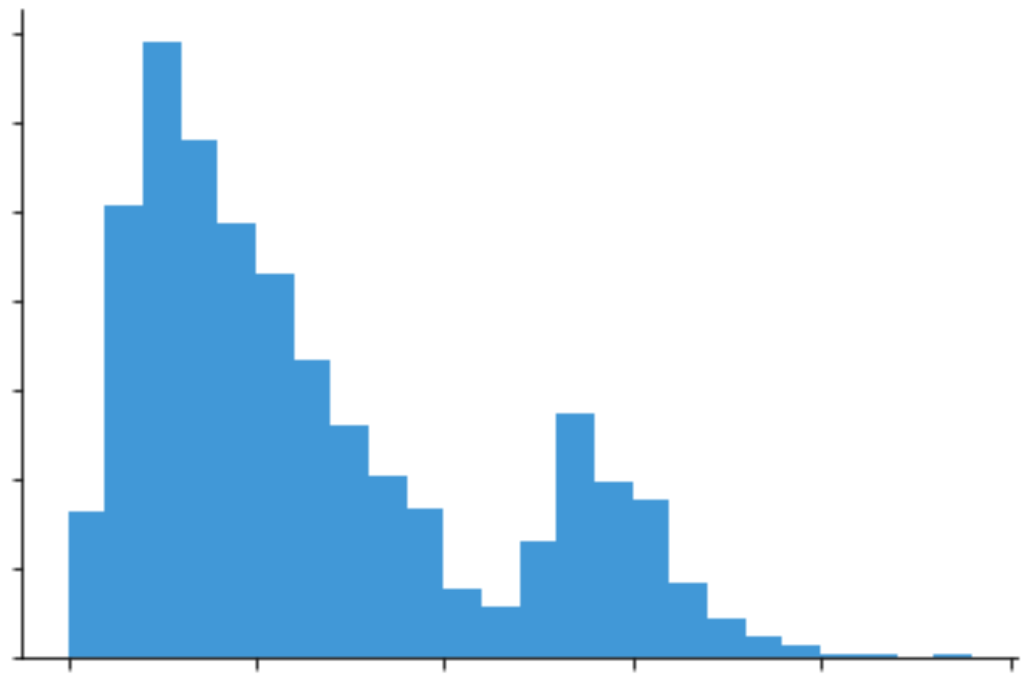
Conditional Expectation Example

- 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
- $p(y \mid X = 0)$ is different from $p(y \mid X = 1)$
- $\mathbb{E}[Y \mid X = 0]$ is different from $\mathbb{E}[Y \mid X = 1]$
  - e.g. $\mathbb{E}[Y \mid X = 0] = 70$ is different from $\mathbb{E}[Y \mid X = 1] = 150$

Conditional Expectation Example (cont)

- $p(y \mid X = 0)$ $\quad\quad$ $p(y \mid X = 1)$





- $\mathbb{E}[Y \mid X = 0]$ 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 $p(y \mid X = 1)$

What if Y is dollars earned?

- Y is now a continuous RV
- What is $p(y \mid x)$?

What if Y is dollars earned?

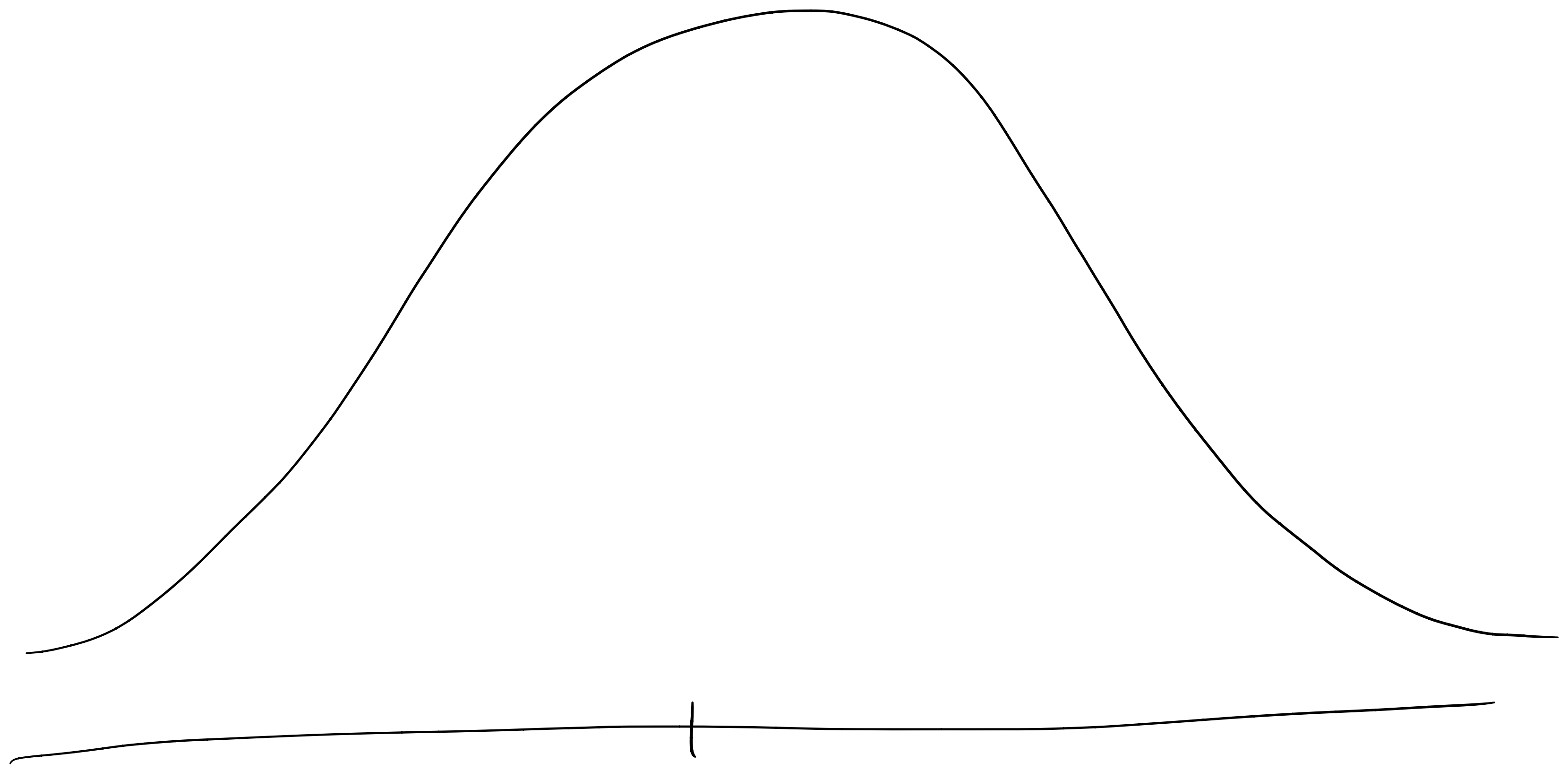
- Y is now a continuous RV
- Notice that $p(y \mid x)$ is defined by $p(y \mid X = 0)$ and $p(y \mid X = 1)$
- What might be a reasonable choice for $p(y \mid X = 0)$ and $p(y \mid X = 1)$?

What if Y is dollars earned?

- Notice that $p(y \mid x)$ is defined by $p(y \mid X=0)$ and $p(y \mid X=1)$

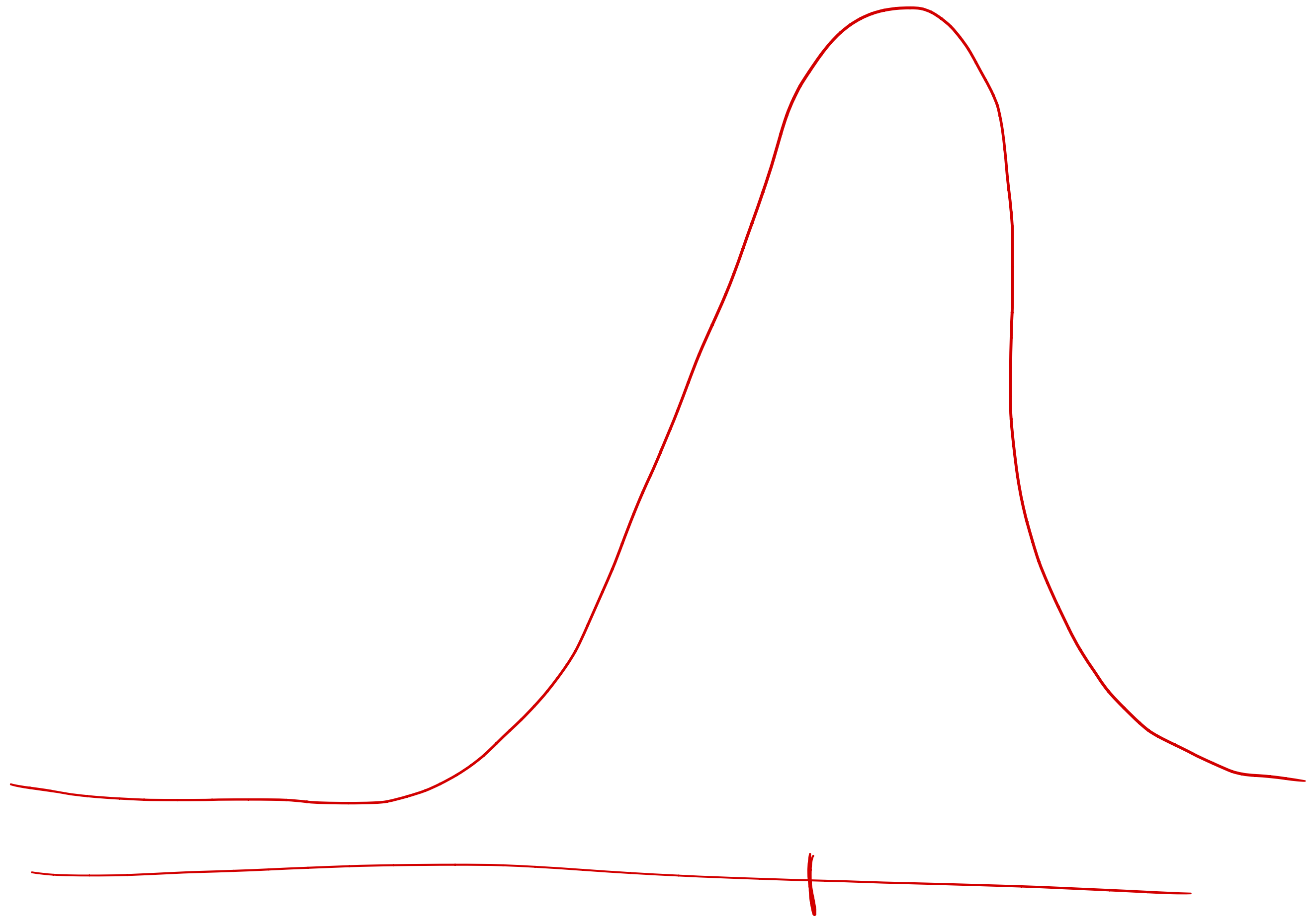
$p(y \mid X=0) = \mathcal{N}(\mu_0, \sigma_0^2)$

$p(y \mid X=1) = \mathcal{N}(\mu_1, \sigma_1^2)$



$\mu_0 = 300$

Non-fiction



$\mu_1 = 100$

Fiction

Exercise

- Come up with an example of X and Y, and give possible choice for $p(y \mid x)$
- Do you need to know $p(x)$ to specify $p(y \mid x)$?

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:
  - $\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \mathbb{E}[Y]$

Linearity of Expectation for any X and Y

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

Properties of Expectations for X and Y independent

$$
\begin{aligned}
\mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) xy \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(y \mid x) p(x) xy \\
&= \sum_{x \in \mathcal{X}} xp(x) \sum_{y \in \mathcal{Y}} p(y \mid x) y \\
&= \sum_{x \in \mathcal{X}} xp(x) \mathbb{E}[Y \mid x] \\
&= \sum_{x \in \mathcal{X}} xp(x) \mathbb{E}[Y] \qquad \text{since X and Y independent} \\
&= \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}
$$

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

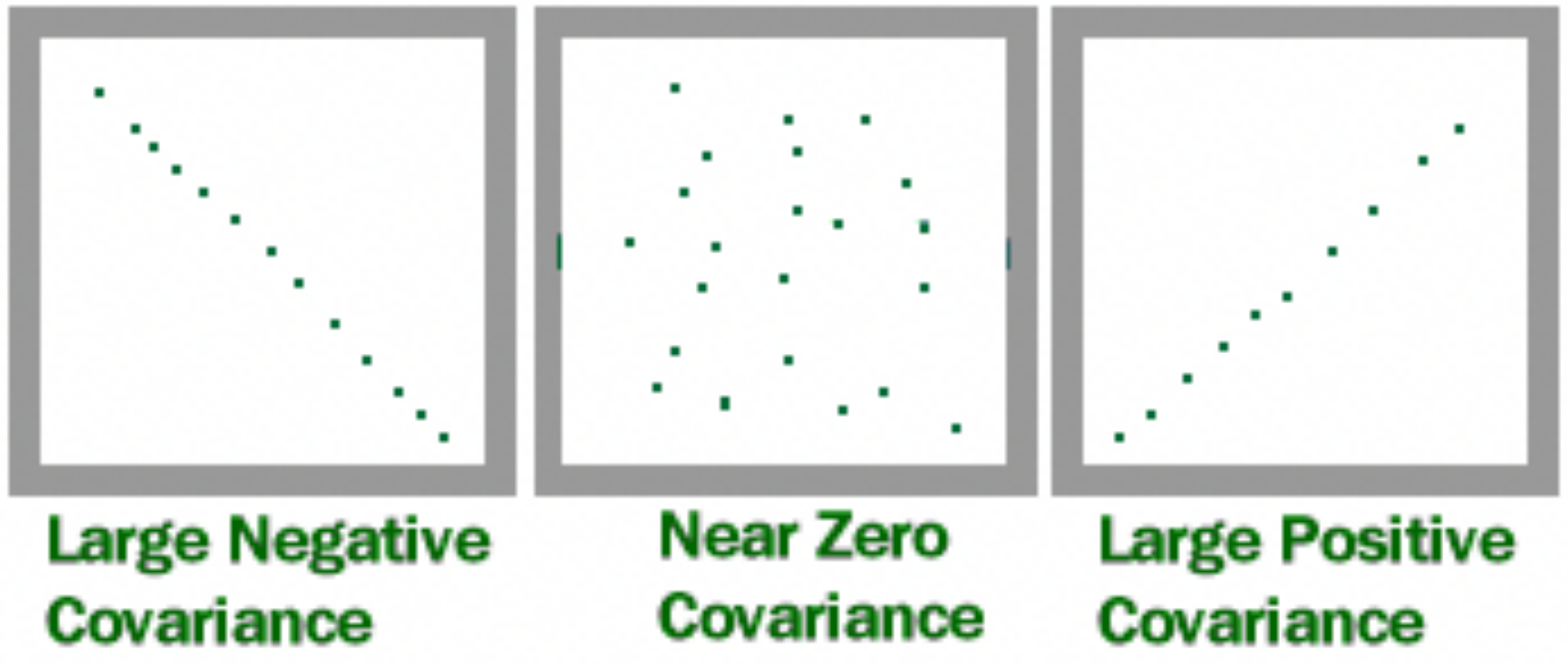
Equivalently,

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

Covariance

Definition: The covariance of two random variables is

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



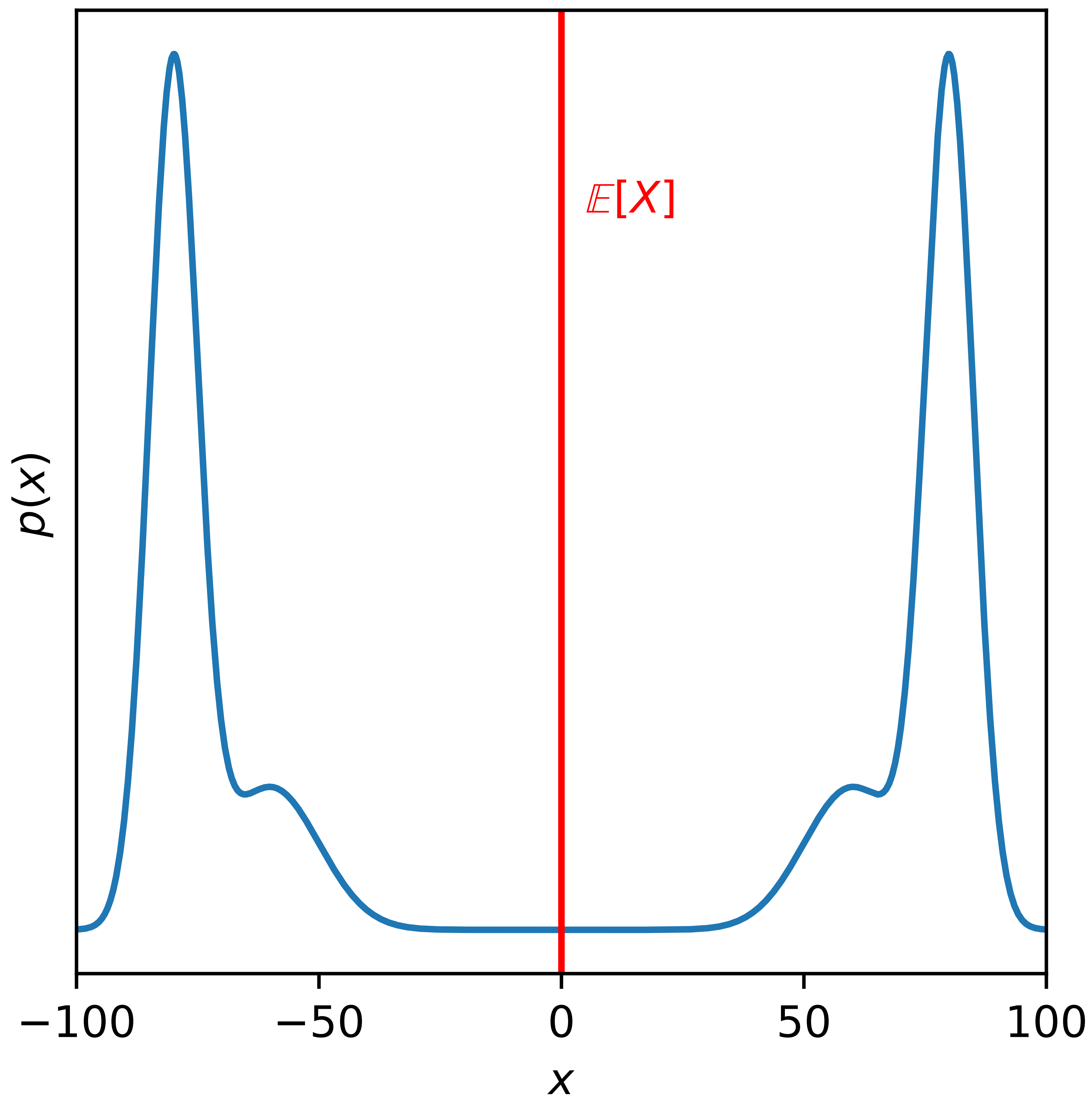
Properties of Variances

- 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, because Cov$[X, Y] = 0$
  Var$[X + Y] =$ Var$[X] +$ Var$[Y]$

Estimators

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

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



random variable!

Questions:

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 $Z \sim N(0,100^2)$
3. $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$, for $X_i \sim p$

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**

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.

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

Bias

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

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

- Bias can be positive or negative or zero
- When $\mathrm{Bias}(\hat{X}) = 0$, 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. $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$

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 $\delta, \epsilon > 0$
- $\mathrm{Var}[\bar{X}] = \frac{1}{n}\sigma^2$ means that with "enough" data, $\Pr\left(\left|\bar{X} - \mu\right| < \epsilon\right) > 1 - \delta$ for *any* $\delta, \epsilon > 0$ that we pick
- 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 $\text{Var}[X_i] \leq \frac{1}{4}(b-a)^2$
- 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

Sample Complexity

Definition:
The **sample complexity** of an estimator is the number of samples required to guarantee an expected error of at most ϵ with probability $1 - \delta$, for given δ and ϵ.

- We want sample complexity to be small
- Sample complexity is determined by:
  1. The **estimator** itself
     - Smarter estimators can sometimes improve sample complexity
  2. Properties of the **data generating process**
     - If the data are high-variance, we need more samples for an accurate estimate
     - But we can reduce the sample complexity if we can **bias** our estimate **toward the correct value**

Sample Complexity

Definition:
The **sample complexity** of an estimator is the number of samples required to guarantee an expected error of at most ϵ with probability $1 - \delta$, for given δ and ϵ.

For $\delta = 0.05$, **Chebyshev** gives

$$\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} = \frac{1}{\sqrt{0.05}} \frac{\sigma}{\sqrt{n}}$$

$$\iff \quad \epsilon = 4.47 \frac{\sigma}{\sqrt{n}}$$

$$\iff \quad \sqrt{n} = 4.47 \frac{\sigma}{\epsilon}$$

$$\iff \quad n = 19.98 \frac{\sigma^2}{\epsilon^2}$$

With **Gaussian assumption** and $\delta = 0.05$,

$$\epsilon = 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\iff \quad \sqrt{n} = 1.96 \frac{\sigma}{\epsilon}$$

$$\iff \quad n = 3.84 \frac{\sigma^2}{\epsilon^2}$$

Mean-Squared Error

- **Bias:** whether an estimator is correct **in expectation**
- **Consistency:** whether an estimator is correct **in the limit of infinite data**
- **Convergence rate:** how fast the estimator **approaches its own mean**
  - For an **unbiased** estimator, this is also how fast its **error bounds** shrink
- We don't necessarily care about an estimator's being unbiased.
  - Often, what we care about is our estimator's **accuracy in expectation**

Definition: **Mean squared error** of an estimator $\hat{X}$ of a quantity X:

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

different!

Bias-Variance Tradeoff

$$\mathrm{MSE}(\hat{X}) = \mathrm{Var}[\hat{X}] + \mathrm{Bias}(\hat{X})^2$$

- If we can decrease bias without increasing variance, error goes down
- If we can decrease variance without increasing bias, error goes down
- **Question:** Would we ever want to **increase bias**?
- *YES.* If we can increase (squared) bias in a way that **decreases variance more**, then error goes down!
    - **Interpretation:** Biasing the estimator toward values that are **more likely to be true** (based on **prior information)**

Downward-biased Mean Estimation

Example: Let's estimate μ given i.i.d $X_1, \ldots, X_n$ with $\mathbb{E}[X_i] = \mu$ using: $Y = \frac{1}{n+100} \sum_{i=1}^{n} X_i$

This estimator is **biased**:

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{1}{n+100}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n+100}\sum_{i=1}^{n} \mathbb{E}[X_i]$$

$$= \frac{n}{n+100}\mu$$

$$\mathrm{Bias}(Y) = \frac{n}{n+100}\mu - \mu = \frac{-100}{n+100}\mu$$

This estimator has **low variance**:

$$\mathrm{Var}(Y) = \mathrm{Var}\left[\frac{1}{n+100}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{(n+100)^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{(n+100)^2}\sum_{i=1}^{n}\mathrm{Var}[X_i]$$

$$= \frac{n}{(n+100)^2}\sigma^2$$

Estimating μ Near 0

Example: Suppose that $\sigma = 1$, $n = 10$, and $\mu = 0.1$

$\mathrm{Bias}(\bar{X}) = 0$

$$\begin{aligned}\mathrm{MSE}(\bar{X}) &= \mathrm{Var}(\bar{X}) + \mathrm{Bias}(\bar{X})^2 \\ &= \mathrm{Var}(\bar{X}) \\ &= \frac{1}{10}\end{aligned}$$

$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$

$$\begin{aligned}\mathrm{MSE}(Y) &= \mathrm{Var}(Y) + \mathrm{Bias}(Y)^2 \\ &= \frac{n}{(n+100)^2}\sigma^2 + \left(\frac{100}{n+100}\mu\right)^2 \\ &= \frac{10}{110^2} + \left(\frac{100}{110}0.1\right)^2 \\ &\approx 9 \times 10^{-4}\end{aligned}$$

Summary

- **Concentration inequalities** let us bound the probability of a given estimator being at least ϵ from the estimated quantity
- **Sample complexity** is the **number of samples** needed to attain a desired error bound ϵ at a desired probability $1 - \delta$
- The **mean squared error** of an estimator **decomposes** into **bias** (squared) and **variance**
- Using a **biased** estimator can have **lower error** than an unbiased estimator
    - Bias the estimator based some **prior information**
    - *But this only helps if the prior information is **correct,** cannot reduce error by adding in arbitrary bias*