

# Chapter 4 and beyond: Learning about distributions from finite data

## Part 1, Mon 11 July

Jess Kunke

MATH/STAT 394: Probability I (Summer 2022 A-term)

# Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

# Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

## Announcements + clarifications

$$\text{pmf } P_X(k) = \begin{cases} \frac{1}{3} & k=0 \\ \frac{1}{3} & k=1 \\ \frac{1}{3} & k=2 \\ 0 & \text{else} \end{cases}$$

- ▶ HW 4
- ▶ Piecewise function notation

	M	Tu	W	Th	Fr
Week 4	Jmsn on lecture		lecture		lecture about final on remote
Week 5	lecture			Final	
	HW 4 due	on remote			395 starts

$$E[X]_{kt}$$

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

2nd centered

# Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

## Expectation of a function of a RV

From Chapter 3:

If we know the distribution of a RV  $X$  and now we are interested in a RV  $Y = g(X)$  for some function  $g$ , we know how to compute  $E[Y]$ :

### Theorem

Let  $X$  be a RV that takes values in  $\mathcal{X}$  and  $g : \mathcal{X} \rightarrow \mathbb{R}$  be some function.

$$E[g(X)] = \sum_{k \in \mathcal{X}} g(k)p(k) \quad \text{if } X \text{ is discrete with pmf } p,$$

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx \quad \text{if } X \text{ is continuous with pdf } f.$$

Now we will cover how to derive the *distribution* of  $Y$  from the dist. of  $X$

## Invertible functions

Main idea: map values of  $Y = g(X)$  back to  $X$

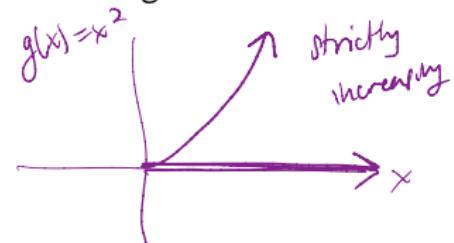
- One concept that might come to mind is the inverse: a map or function  $g : A \rightarrow B$  is **invertible** if for every  $y \in B$  there is a unique  $x \in A$  such that  $y = g(x)$
- Any monotonic (strictly increasing/decreasing) function  $g$  is invertible
- e.g.  $g(x) = x^2$  is invertible on  $[0, \infty)$

Let  $X \in \mathcal{X} = \{1, 2, 3\}$  w pmf  $p_X(a) = \begin{cases} 1/4 & a=1 \\ 1/2 & a=2 \\ 1/4 & a=3 \end{cases}$

Suppose  $Y = X^2 \Rightarrow Y = \{1, 4, 9\}$

$X$	$g$	$Y$	$1-1$ map $g$ is invertible
1	$\longrightarrow$	1	
2	$\longrightarrow$	4	
3	$\longrightarrow$	9	$g^{-1}(y) = \sqrt{y}$

$$\begin{aligned} P_Y(y) &=? \\ P(Y=1) &= P(X=1) = 1/4 \\ P(Y=4) &= P(X=2) = 1/2 \\ P(Y=9) &= P(X=3) = 1/4. \end{aligned}$$

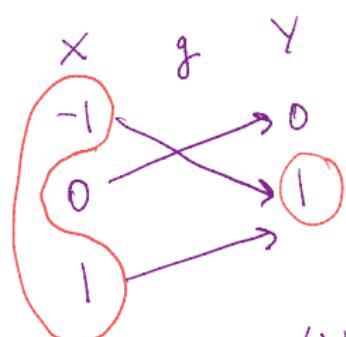


$$P(Y=y) = P(X = g^{-1}(y)).$$

What if  $g$  is not invertible?

Maybe multiple values of  $X$  map to the same value ( $y$ ) of  $Y$ , so that  $g^{-1}(y)$  is a set, not a single number

$$x \in X = \{-1, 0, 1\} \quad \text{pmf } P_X(a) = \begin{cases} 1/4 & a = -1 \\ 1/2 & a = 0 \\ 1/4 & a = 1 \end{cases}$$
$$Y = X^2 \quad Y = \{0, 1\}$$



$$P(Y=0) = P(X=0) = 1/2$$

$$\begin{aligned} P(Y=1) &= P(X=\pm 1) = P(X=1) + P(X=-1) \\ &= 1/4 + 1/4 = 1/2. \end{aligned}$$

$$P(Y=y) = P\left(X \underbrace{\in \{k \in X : g(k)=y\}}_{\text{now a set, not necessarily a single pre-image}}\right)$$

now a set, not necessarily a single pre-image

## Images and pre-images of sets

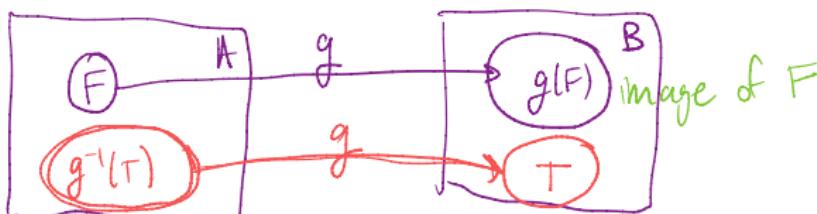
### Definition

Let  $A, B$  be two sets and  $g : A \rightarrow B$ . The **image** of  $F \subseteq A$  under  $g$  is defined as

$$g(F) = \{g(x) : x \in F\} \subseteq B.$$

The **pre-image** of  $T \subseteq B$  under  $g$  is

$$g^{-1}(T) = \{x \in A : g(x) \in T\} \subseteq A.$$



- ▶ The notation  $g^{-1}$  is the same we use for the inverse of  $g$  when it is defined, but here we are not assuming  $g$  is invertible
- ▶ The pre-image of a set always exists even if the inverse does not exist
- ▶ If  $g$  is invertible, then  $g^{-1}(T)$  is the image of  $T$  under the inverse map  $g^{-1}$
- ▶ These definitions apply on sets not on variables
- ▶ If there is no element that maps onto  $T$ , then  $g^{-1}(T) = \emptyset$

## Summary: transformations of a discrete RV

### Lemma

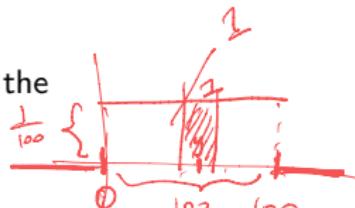
Let  $X$  be a discrete RV, let  $g : \mathbb{R} \rightarrow \mathbb{R}$ , and let  $Y = g(X)$ . The pmf of  $Y$  is

$$p_Y(y) = P(g(X) = y) = P(X \in g^{-1}(\{y\})) = \sum_{\substack{k: g(k)=y \\ k \in \mathcal{X}}} p_X(k).$$

- ▶ Why is this result specifically for discrete RVs?

## Discrete transformation of a continuous RV

For continuous RVs  $X$  let's start by considering the case that the transformation  $g(X)$  is discrete.



### Example

Suppose that a student's score  $X$  is continuous and uniformly distributed on  $[0, 100]$ :  $X \sim \text{Unif}[0, 100]$ . A teacher rounds the students' scores to the nearest integer, e.g. if  $4.5 \leq X < 5.5$ , then the rounded score  $Y$  equals 5.

What is the pmf of the rounded scores  $Y$ ?

$g(x) = \text{nearest integer to } x$

for  $y \in \{1, \dots, 99\}$ ,  $g: [y - 1/2, y + 1/2) \mapsto y$

$P(Y=y) = P(X \in [y - 1/2, y + 1/2)) = \int_{y-1/2}^{y+1/2} f_X(x) dx$

$= \int_{y-1/2}^{y+1/2} \frac{1}{100} dx = \frac{y+1/2}{100} - \frac{y-1/2}{100} = \frac{1}{100}$

$P(Y=0) = P(X \in [0, 0.5)) = \int_0^{0.5} \frac{1}{100} dx = \frac{1}{200}$

$P(Y=100) = P(X \in [99.5, 100]) = \int_{99.5}^{100} \frac{1}{100} dx = \frac{1}{200}$

pdf of  $\text{Unif}[0, 100]$

cont.	discrete
$X$	$Y$
0	0
0.5	1
1.5	2
...	...
98.5	99
99.5	100

## Discrete transformation of a continuous RV

More generally we have the following result, which is the same as before:

### Lemma

Let  $X$  be a continuous RV and  $g : \mathbb{R} \rightarrow \mathcal{Y}$  be a function that maps  $\mathbb{R}$  onto a discrete set  $\mathcal{Y}$ .

Then the RV  $Y$  is discrete and for any  $k \in \mathcal{Y}$ ,

$$P(Y = k) = P(X \in g^{-1}(\{k\})).$$



$$= \int$$

## Summary so far

- If  $X$  is discrete and  $g$  is any function, then the pmf of  $Y = g(X)$  is

$$P(Y = y) = P(X \in g^{-1}(\{y\})) \quad \text{for } y \in g(\mathcal{X}),$$

where  $g^{-1}(\{y\}) = \{k \in \mathcal{X} : g(k) = y\}$  is the pre-image of  $y$  under  $g$ .

- If  $X$  is continuous but  $g$  is a discrete function ( $g : \mathbb{R} \rightarrow \mathcal{Y}$  with  $\mathcal{Y}$  discrete), then we have the same result except that we integrate over subsets of  $\mathbb{R}$  instead of summing over subsets of a discrete space  $\mathcal{X}$ :

$$P(Y = y) = P(X \in g^{-1}(\{y\})),$$

where  $g^{-1}(\{y\}) = \{x \in \mathbb{R} : g(x) = y\}$  is the pre-image of  $y$  under  $g$ .

## Continuous transformation of a continuous RV

The cdf method

- ▶ For a continuous RV, the pdf has no interpretation as a probability distribution
- ▶ It is easier to compute the cdf of  $Y = g(X)$ , then differentiate to get the pdf
- ▶ We will first illustrate the idea, then detail the method if
  1.  $g$  is invertible
  2.  $g$  is not invertible on  $\mathbb{R}$  but invertible on some intervals of  $\mathbb{R}$  that form a partition of  $\mathbb{R}$

## The cdf method: Example

### Example

Let  $X$  be a continuous RV with pdf  $f_X$ . What is the pdf of  $Y = aX + b$  for some constants  $a, b$  with  $a > 0$ ?

$$\begin{aligned} \text{cdf } F_Y(y) &= P(Y \leq y) = P(ax + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

We've written the cdf of  $Y$  in terms of the cdf of  $X$ .

$$\text{pdf } f_Y(y) = F'_Y(y) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right).$$

What if  $a < 0$ ?

$$\begin{aligned} \text{cdf } F_Y(y) &= P(Y \leq y) = P(ax + b \leq y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - P\left(X \leq \frac{y-b}{a}\right) \\ &= 1 - F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

We've written the cdf of  $Y$  in terms of the cdf of  $X$ .

$$\text{pdf } f_Y(y) = F'_Y(y) = \frac{-1}{a} \cdot f_X\left(\frac{y-b}{a}\right). \quad \text{For } a \neq 0, \quad f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

bc  $X$  continuous.

## The cdf method: Derivation

*increasing*  
g invertible, strictly ~~decreasing~~

$$\text{cdf } F_Y(t) = P(Y \leq t) = P(X \leq g^{-1}(t)) = F_X(g^{-1}(t))$$

$$\text{pdf } f_Y(t) = \frac{d}{dt} F_Y(g^{-1}(t)) = (g^{-1}(t))' F_X'(g^{-1}(t))$$

finding derivative of  $g^{-1}(t)$ :

$$g(g^{-1}(t)) = t \quad \downarrow \frac{d}{dt}$$

$$(g^{-1}(t))' \cdot g'(g^{-1}(t)) = 1$$

$$(g^{-1}(t))' = \frac{1}{g'(g^{-1}(t))}.$$

$$\downarrow$$

$$= \frac{1}{g'(g^{-1}(t))} f_X(g^{-1}(t)).$$

g invertible, strictly decreasing

$$F_Y(t) = P(Y \leq t) = P(g(X) \leq t) = P(X \geq g^{-1}(t)) = 1 - F_X(g^{-1}(t))$$

$$f_Y(t) = -\frac{d}{dt} F_Y(g^{-1}(t)) = -\frac{1}{g'(g^{-1}(t))} f_X(g^{-1}(t)).$$

## The cdf method: summary

Previous considerations can be summarized by the following lemma:

### Lemma

Let  $X$  be a continuous RV with pdf  $f_X$ .

Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be differentiable and strictly monotonic.

Then the pdf of  $Y = g(X)$  exists and is given by

$$f_Y(y) = \begin{cases} \left| \frac{1}{g'(g^{-1}(y))} \right| f_X(g^{-1}(y)) & \text{if } y \in g(\mathbb{R}), \\ 0 & \text{otherwise.} \end{cases}$$

Note:

- ▶ It is preferable to remember the method rather than the lemma because the method is more flexible (see next slides)

## The cdf method: further considerations

1. What if  $g$  is not defined on all of  $\mathbb{R}$ ?

- $g$  only needs to be defined on a subset  $B \subseteq \mathbb{R}$  s.t.  $P(X \in B) = 1$
- For  $y \notin g(B)$ , define  $f_Y(y) = 0$



$$F_X(x) = x$$

Example

Let  $X \sim \text{Unif}([0, 1])$  and  $g : x \mapsto -\frac{1}{\lambda} \log(1-x)$ , where  $\lambda > 0$ . What is the distribution of  $Y = g(X)$ ?

$$P(X \in [0, 1]) = 1.$$

on  $[0, 1]$ ,

$$P(Y \leq t) = P\left(-\frac{1}{\lambda} \log(1-x) \leq t\right) = P(\log(1-x) \geq -\lambda t)$$

$$= P(1-x \geq e^{-\lambda t}) = P(x \leq 1-e^{-\lambda t})$$

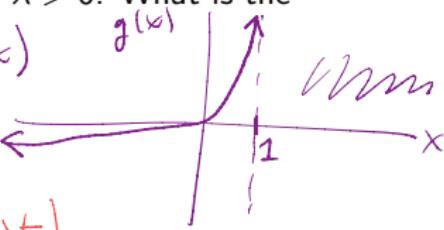
$$= F_X(1-e^{-\lambda t})$$

cdf of  $X$  is  $F_X(x) = x$ .

$$= 1 - e^{-\lambda t}.$$

$$f_Y(y) = \lambda e^{-\lambda t} \quad y > 0$$

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda t} & y > 0 \\ 0 & \text{otherwise} \end{cases}$$



$g$  is monotonically  
strictly nonnegative

$\Rightarrow$  pdf  $f_Y(y) = 0 \quad \forall y < 0$   
so we need to find

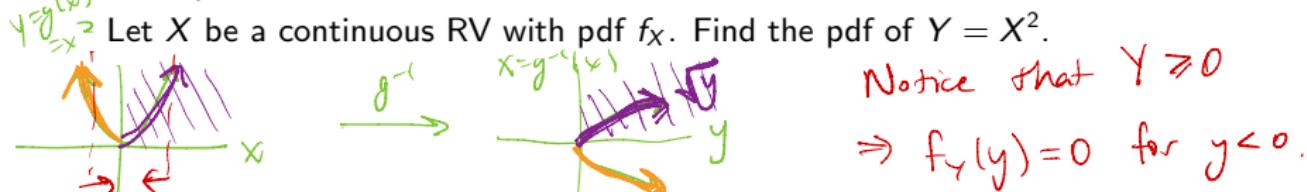
$f_Y(y)$  for  $y \geq 0$ .

The cdf method: further considerations :  $f_y(y) = \begin{cases} \frac{1}{2\sqrt{y}} f_x(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_x(-\sqrt{y}) & y \geq 0 \\ 0 & y < 0 \end{cases}$

## 2. What if $g$ is not invertible?

- ▶ Partition  $\mathbb{R}$  into intervals  $[a_i, a_{i+1}]$  such that  $g$  is invertible on each interval  $[a_i, a_{i+1}]$
- ▶ Apply previous reasoning on these intervals
- ▶ Combine the results to get the cdf, then differentiate to get the pdf

### Example



$$\begin{aligned}
 F_Y(y) &= P(Y \leq y) = P(X^2 \leq y \text{ } \& \text{ } X < 0) + P(X^2 \leq y \text{ } \& \text{ } X \geq 0) \\
 &= P(-\sqrt{y} \leq X < 0) + P(0 \leq X \leq \sqrt{y}) \\
 &= F_X(0) - F_X(-\sqrt{y}) + F_X(\sqrt{y}) - F_X(0) \\
 &= F_X(\sqrt{y}) - F_X(-\sqrt{y})
 \end{aligned}$$

$\frac{d}{dy} \sqrt{y} = \frac{d}{dy} y^{1/2} = \frac{1}{2} y^{-1/2} = \frac{1}{2\sqrt{y}}$

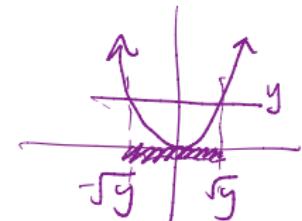
## The cdf method: further considerations

2. What if  $g$  is not invertible?

- ▶ Alternative (ultimately equivalent) approach below

### Example

Let  $X$  be a continuous RV with pdf  $f_X$ . Find the pdf of  $Y = X^2$ .



$$F_Y(y) = P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

From here, we take the derivative just as before & arrive at the same result (the same pdf  $f_Y(y)$ ).

## Summary: transformations of RVs

- If  $X$  is discrete and  $g$  is any function, then the pmf of  $Y = g(X)$  is

$$P(Y = y) = P(X \in g^{-1}(\{y\})) \quad \text{for } y \in g(\mathcal{X}),$$

where  $g^{-1}(\{y\}) = \{k \in \mathcal{X} : g(k) = y\}$  is the pre-image of  $y$  under  $g$ .

- If  $X$  is continuous but  $Y = g(X)$  is discrete ( $g : \mathbb{R} \rightarrow \mathcal{Y}$  with  $\mathcal{Y}$  discrete), then we have the same result except that we integrate over subsets of  $\mathbb{R}$  instead of summing over subsets of a discrete space  $\mathcal{X}$ :

$$P(Y = y) = P(X \in g^{-1}(\{y\})),$$

where  $g^{-1}(\{y\}) = \{x \in \mathbb{R} : g(x) = y\}$  is the pre-image of  $y$  under  $g$ .

- If  $X$  and  $Y = g(X)$  are continuous, then use the cdf method in some form
  - Identify the possible values of  $X$ : only need to account for values of  $g(x)$  on this set
  - Identify the possible values of  $Y$ :  $f_Y(y) = 0$  for all other values in  $\mathbb{R}$
  - Compute the cdf of  $Y$
  - If you need the pdf of  $Y$ , differentiate

# Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

## Motivation

What we've studied so far

- ▶ How can we frame a problem in terms of a probability model?
- ▶ Given a probability model or probability distribution, how can we compute some useful summaries like the expectation and variance?

## Motivation

What we've studied so far

- ▶ How can we frame a problem in terms of a probability model?
- ▶ Given a probability model or probability distribution, how can we compute some useful summaries like the expectation and variance?

But in practice we often don't fully know the distribution

- ▶ When you flip a coin, what if you don't want to assume it's fair but instead estimate  $p$ ?
- ▶ What if you want to estimate average income from survey data?
- ▶ We can often partially write the probability model but
  - ▶ there may be parameters we don't know (e.g.  $X \sim \text{Bern}(p)$  but we don't know  $p$ )
  - ▶ maybe we know the mean and variance but not higher moments

## Example: Bernoulli trials

### Example

Suppose you have a coin and want to estimate its bias. What would you do?

e.g. do 100 flips, count how many heads (or tails)

1) suppose I get 30 H.  $\rightarrow$  may be fair

2) Suppose I get 10 H  $\rightarrow$  still might be  
fair, but less likely...

in case 1, we might estimate  $\hat{P} = \frac{30}{100} = 0.3$

case 2 :

$$\hat{P} = \frac{10}{100} = 0.1$$

## General setting

- ▶ In general, suppose you can frame the problem of interest as a series of independent identically distributed (iid) trials
- ▶ A common estimator of the true mean is the **empirical** or **sample mean**

### Definition

Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} X$  (i.e.  $n$  independent RVs, identically distributed as a RV  $X$ ). The **empirical/sample mean** is defined by

For the coin:  $X_i = \begin{cases} 1 & \text{H} \\ 0 & \text{T} \end{cases}$

$$\bar{X}_n := \frac{X_1 + \dots + X_n}{n}. \quad \bar{X}_n = \frac{\sum_i X_i}{n} = \frac{30}{100} \text{ or } \frac{10}{100} \text{ in}$$

our example.

- ▶ In the case that some event  $A$  can happen or not in each trial and you are interested in the average number of times  $A$  happens in  $n$  trials, you have  $X \sim \text{Binom}(n, p)$  and want to estimate  $np$ . What if you know  $n$  but not  $p$ ?
- ▶ We will study the empirical mean as a random variable to understand its properties

## Properties of the empirical mean of iid trials

- Regardless of whether the  $X_i$  are independent, as long as they are identically distributed then we have from linearity of expectation that the mean of the sample mean is the true mean:

$$E[\bar{X}_n] = \frac{1}{n} \sum_i E[X_i] = E[X].$$

$\frac{1}{n} \cdot n E[X]$

We say that the sample mean is **unbiased**.

## Properties of the empirical mean of iid trials

- ▶ Regardless of whether the  $X_i$  are independent, as long as they are identically distributed then we have from linearity of expectation that the mean of the sample mean is the true mean:

$$E[\bar{X}_n] = \frac{1}{n} \sum_i E[X_i] = E[X].$$

We say that the sample mean is **unbiased**.

- ▶ If the  $X_i$  are iid (why do we need iid here?), then

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \frac{1}{n^2} \sum_i \text{Var}[X_i] = \frac{1}{n} \text{Var}[X]. \\ &= \text{Var}\left(\frac{1}{n} \sum_i X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_i X_i\right) \xrightarrow{\text{if indp.}} \underbrace{\frac{1}{n^2} \sum_i \text{Var}(X_i)}_{n \text{ Var}(X)} \end{aligned}$$

As the number of trials increases, what happens to the variance of your estimate? Does that make sense?

## Properties of the empirical mean of iid trials

- ▶ Regardless of whether the  $X_i$  are independent, as long as they are identically distributed then we have from linearity of expectation that the mean of the sample mean is the true mean:

$$E[\bar{X}_n] = \frac{1}{n} \sum_i E[X_i] = E[X].$$

We say that the sample mean is **unbiased**.

- ▶ If the  $X_i$  are iid (why do we need iid here?), then

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_i \text{Var}[X_i] = \frac{1}{n} \text{Var}[X].$$

As the number of trials increases, what happens to the variance of your estimate? Does that make sense?

- ▶ Therefore, as  $n \rightarrow \infty$  it seems that  $\bar{X}_n \rightarrow E[X]$ . We will develop tools to formalize this (both in this course and in MATH/STAT 395).

# Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

## Estimating tail probabilities: Motivation

- ▶ Convergence of the empirical mean will necessarily be stated in terms of probability
- ▶ Namely, we would like to show that as  $n \rightarrow +\infty$ , the probability that  $\bar{X}_n$  differs from  $E[X]$  tends to 0
- ▶ For that we'll need some tools to bound probabilities when we only know e.g. the mean/the variance of a RV
- ▶ That's what concentration inequalities are about

## Concentration inequalities

First, we'll need the following result:

### Theorem (Monotonicity of Expectation)

*If two RVs  $X, Y$  defined on the same probability space  $(\Omega, \mathcal{F}, P)$  have finite means and satisfy that  $P(X \leq Y) = 1$ , then  $E[X] \leq E[Y]$ .*

## Markov's Inequality

What can we say about the probability of  $X$  if we know  $E[X]$ ?

Theorem (Markov's inequality)

If  $X$  is a non-negative RV with finite mean, then for any  $c > 0$ ,

$$P(X \geq c) \leq \frac{E[X]}{c}.$$

Proof: Consider the RV  $\mathbb{1}_{\{X \geq c\}} = \begin{cases} 1 & X \geq c \\ 0 & X < c \end{cases}$

Then  $X \geq c \mathbb{1}_{\{X \geq c\}}$ .

– if  $X \geq c$ , then RHS =  $c \cdot 1 = c$ , so we have the statement  $X \geq c$ .

– if  $X < c$ , then RHS =  $c \cdot 0 = 0$ , so we have  $X \geq 0$ , which is true if  $X$  is non-negative.

Then by monotonicity of expectation (previous slide),  $E[X] \geq E[c \mathbb{1}_{\{X \geq c\}}] = c P(X \geq c)$

Rearrange,  $P(X \geq c) \leq \frac{E[X]}{c}$ .

## Markov's inequality

-can we apply markov's inequality here?  
-if yes, then apply it.

### Example

A donut vendor sells on average 1000 donuts per day. Could he sell more than 1400 donuts tomorrow with probability greater than 0.8?

Let  $X = \# \text{ donuts sold per day}$ .  $X \geq 0$ , finite mean

$$P(X \geq 1400) \leq \frac{E[X]}{1400} = \frac{1000}{1400} \approx 0.71$$

no.

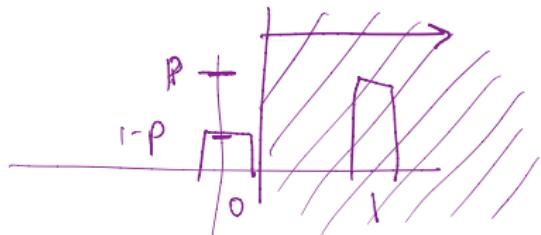
## Markov's inequality

### Example

Let  $X \sim \text{Ber}(p)$  for some  $p \in (0, 1)$ .

- 1. What is  $P(X \geq 0.01)$ ?
- 2. What does Markov's inequality give us?

$$1. P(X \geq 0.01) = P(X=1) = p.$$



$$P(X=0) = 1-p$$

$$P(X=1) = p$$

$$2. P(X \geq 0.01) \leq \frac{E[X]}{0.01} = \frac{p}{0.01} = 100p. \quad \text{could even be } > 1!$$

## Chebyshev's inequality

What can we say about the probability of  $X$  if we know both  $E[X]$  and  $\text{Var}(X)$ ?

### Theorem (Chebyshev's Inequality)

If  $X$  is a RV with finite mean  $\mu$  and finite variance  $\sigma^2$ , then for any  $c > 0$ ,

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

Proof: Define  $Z = (X - \mu)^2$        $Z \geq 0$  & has finite mean  
because  $X$  has finite variance  
&  $\text{Var}(x) = E[(x - \mu)^2]$ .  
 $\therefore$  we can apply Markov's  
inequality to  $Z$ :

$$P(|X - \mu| \geq c) = P(Z \geq c^2) \leq \frac{E[Z]}{c^2} = \frac{\sigma^2}{c^2}. \quad //$$

## Chebyshev's inequality

What can we say about the probability of  $X$  if we know both  $E[X]$  and  $\text{Var}(X)$ ?

### Theorem (Chebyshev's Inequality)

If  $X$  is a RV with finite mean  $\mu$  and finite variance  $\sigma^2$ , then for any  $c > 0$ ,

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

Proof:



### Note:

The event  $\{|X - \mu| \geq c\}$  contains the events  $\{X \geq \mu + c\}$  and  $\{X \leq \mu - c\}$ .  
So we naturally have a bound on  $P(X \geq \mu + c)$ ,  $P(X \leq \mu - c)$

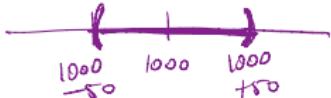
$$P(|X - \mu| \geq c) = P(X \in A \cup B) = P(X \in A) + P(X \in B)$$
$$= P(X \leq \mu - c) + P(X \geq \mu + c)$$

e.g.

$$P(X \geq \mu + c) \leq P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

## Example

We'll define  $X$  as before.



## Example

A donut vendor sells on average 1000 donuts per day with a standard deviation of  $\sqrt{200}$ . Given just this information, provide a bound on

1. the probability that he will sell between 950 and 1050 donuts tomorrow
2. the probability that he will sell at least 1400 donuts tomorrow

$$1. P(950 \leq X \leq 1050) = P(|X - 1000| \leq 50)$$

$$= 1 - P(|X - 1000| \geq 50)$$

complement ( $\& > \geq$   
for cont. RVR)

By Chebyshev's inequality,  $P(|X - 1000| \geq 50) \leq \frac{\text{Var}(X)}{50^2} = \frac{200}{2500} = 0.08$ .

$$\geq 1 - 0.08 = 0.92$$

$$2. P(X \geq 1400) = P\left(X - \frac{1000}{E(X)} \geq 400\right) \leq P(|X - E[X]| \geq 400)$$

$$\leq \frac{\text{Var}(X)}{400^2} = \frac{200}{400^2} = 0.00125.$$