

# Introduction to Probability

Lecture 1: Conditional probabilities and Bayes' theorem

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology

email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk



UNIVERSITY OF  
CAMBRIDGE

# Outline

---

Logistics, motivation, background

Conditional probability

Bayes' Theorem

Independence



## Lecturers

---



Mateja Jamnik



Thomas Sauerwald



### Rough syllabus:

- Introduction to probability: 1 lecture
- Discrete and continuous random variables: 6 lectures
- Moments and limit theorems: 3 lectures
- Applications/statistics: 2 lectures

### Recommended reading:

- **Ross, S.M. (2014). A First course in probability. Pearson (9th ed.).**
- **Dekking, F.M., et. al. (2005) A modern introduction to probability and statistics. Springer.**
- Bertsekas, D.P. & Tsitsiklis, J.N. (2008). Introduction to probability. Athena Scientific.
- Grimmett, G. & Welsh, D. (2014). Probability: an Introduction. Oxford University Press (2nd ed.).



## Why probability?

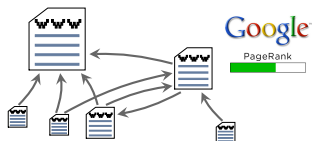
---

- Gives us mathematical tools to deal with uncertain events.
- It is used everywhere, especially in applications of machine learning.
- Machine learning: use **probability** to compute predictions about and from data.
- Probability is not statistics:
  - Both about random processes.
  - Probability: logically self-contained, few rules for computing, one correct answer.
  - Statistics: messier, more art, get experimental data and try to draw probabilistic conclusions, no single correct answer.

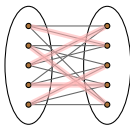


# Applications of probability

## Ranking Websites

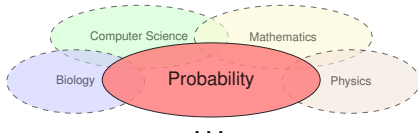


## Matching



$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

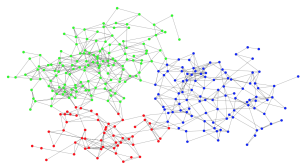
## Finance



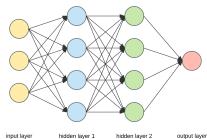
## Medicine



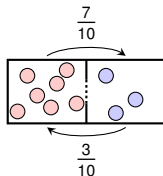
## Data Mining



## Deep Learning



## Particle Processes



- Set theory
  - Counting: product rule, sum rule, inclusion-exclusion
  - Combinatorics: permutations
  - Probability space: sample space, event space
  - Axioms
  - Union bound
- 
- Look for revision material of above on the course website:  
<https://www.cl.cam.ac.uk/teaching/2526/IntroProb/>

# Outline

---

Logistics, motivation, background

Conditional probability

Bayes' Theorem

Independence



## Definition

### Conditional probability

Consider an experiment with sample space  $S$ , and two events  $E$  and  $F$ . Then, the (conditional) probability of event  $E$  given  $F$  has occurred (denoted  $\mathbf{P}[E|F]$ ) with  $\mathbf{P}[F] > 0$  is defined by

$$\mathbf{P}[E|F] = \frac{\mathbf{P}[E \cap F]}{\mathbf{P}[F]} = \frac{\mathbf{P}[EF]}{\mathbf{P}[F]}$$

Sample space: all possible outcomes consistent with  $F$  (i.e.,  $S \cap F = F$ )

Event space: all outcomes in  $E$  consistent with  $F$  (i.e.,  $E \cap F$ )

Note: we assume that all outcomes are equally likely

$$\mathbf{P}[E|F] = \frac{\# \text{ outcomes in } E \cap F}{\# \text{ outcomes in } F} = \frac{\frac{\# \text{ outcomes in } E \cap F}{\# \text{ outcomes in } S}}{\frac{\# \text{ outcomes in } F}{\# \text{ outcomes in } S}} = \frac{\mathbf{P}[E \cap F]}{\mathbf{P}[F]}$$



## Example

### Example

Two dice are rolled yielding value  $D_1$  and  $D_2$ . Let  $E$  be event that  $D_1 + D_2 = 4$ .

1. What is  $\mathbf{P}[E]$ ?
2. Let event  $F$  be  $D_1 = 2$ . What is  $\mathbf{P}[E|F]$ ?

Answer

1.  $|S| = 36$ ,  $E = \{(1, 3), (2, 2), (3, 1)\}$ , thus  $\mathbf{P}[E] = \frac{3}{36} = \frac{1}{12}$ .
2.  $S = \{(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6)\}$ ,  $E = \{(2, 2)\}$ , thus  $\mathbf{P}[E|F] = \frac{1}{6}$



Chain rule

Rearranging the definition of conditional probability gives us:

$$\mathbf{P} [ EF ] = \mathbf{P} [ E|F ] \mathbf{P} [ F ]$$

Generalisation of the Chain rule:

Multiplication rule

$$\mathbf{P} [ E_1 E_2 \cdots E_n ] = \mathbf{P} [ E_1 ] \mathbf{P} [ E_2 | E_1 ] \mathbf{P} [ E_3 | E_2 E_1 ] \cdots \mathbf{P} [ E_n | E_1 \cdots E_{n-1} ]$$

## Example

### Example

An ordinary deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. What is the probability that each pile has exactly 1 ace?

Answer

Define:

$E_1 = \text{ace}\heartsuit$  is in any one pile

$E_2 = \text{ace}\heartsuit$  and  $\text{ace}\spadesuit$  are in different piles

$E_3 = \text{ace}\heartsuit$ ,  $\text{ace}\spadesuit$  and  $\text{ace}\clubsuit$  are in different piles

$E_4 =$  all aces are in different piles

$$\mathbf{P} [ E_1 E_2 E_3 E_4 ] = \mathbf{P} [ E_1 ] \mathbf{P} [ E_2 | E_1 ] \mathbf{P} [ E_3 | E_1 E_2 ] \mathbf{P} [ E_4 | E_1 E_2 E_3 ]$$

We have  $\mathbf{P} [ E_1 ] = 1$ . For rest we consider complement of next ace being in the same pile and thus have:

$$\mathbf{P} [ E_2 | E_1 ] = 1 - \frac{12}{51}$$

$$\mathbf{P} [ E_3 | E_1 E_2 ] = 1 - \frac{24}{50}$$

$$\mathbf{P} [ E_4 | E_1 E_2 E_3 ] = 1 - \frac{36}{49}$$

Thus:

$$\mathbf{P} [ E_1 E_2 E_3 E_4 ] = \frac{39 \cdot 26 \cdot 13}{51 \cdot 50 \cdot 49} \approx 0.105$$



# Outline

---

Logistics, motivation, background

Conditional probability

**Bayes' Theorem**

Independence



## Law of total probability

The law of total probability (a.k.a. Partition theorem)

For events  $E$  and  $F$  where  $\mathbf{P}[F] > 0$ , then for any event  $E$

$$\mathbf{P}[E] = \mathbf{P}[EF] + \mathbf{P}[EF^c] = \mathbf{P}[E|F]\mathbf{P}[F] + \mathbf{P}[E|F^c]\mathbf{P}[F^c]$$

In general, for disjoint events  $F_1, F_2, \dots, F_n$  s.t.  $F_1 \cup F_2 \cup \dots \cup F_n = S$ ,

$$\mathbf{P}[E] = \sum_{i=1}^n \mathbf{P}[E|F_i]\mathbf{P}[F_i]$$

Intuition:

Want to know probability of  $E$ . There are two scenarios,  $F$  and  $F^c$ . If we know these and the probability of  $E$  conditioned on each scenario, we can compute the probability of  $E$ .



## Lightbulb example

### Example

There are 3 boxes each containing a different number of light bulbs. The first box has 10 bulbs of which 4 are dead, the second has 6 bulbs of which 1 is dead, and the third box has 8 bulbs of which 3 are dead. What is the probability of a dead bulb being selected when a bulb is chosen at random from one of the 3 boxes (each box has equal chance of being picked)?

Answer

Let event  $E$  = "dead bulb is picked", and  $F_1$  = "bulb is picked from first box",  $F_2$  = "bulb is picked from second box" and  $F_3$  = "bulb is picked from third box". We know:

$$\mathbf{P}[E|F_1] = \frac{4}{10}, \mathbf{P}[E|F_2] = \frac{1}{6}, \mathbf{P}[E|F_3] = \frac{3}{8}$$

We need to compute  $\mathbf{P}[E]$ , and we know that  $\mathbf{P}[F_i] = \frac{1}{3}$ :

$$\mathbf{P}[E] = \sum_{i=1}^n \mathbf{P}[E|F_i] \mathbf{P}[F_i] = \frac{4}{10} \frac{1}{3} + \frac{1}{6} \frac{1}{3} + \frac{3}{8} \frac{1}{3} = \frac{113}{360} \approx 0.31$$



## Bayes' theorem

How many spam emails contain the word "Dear"?

$$\mathbf{P}[E|F] = \mathbf{P}[\text{"Dear"}|\text{spam}]$$

But how about what is the probability that an email containing "Dear" is spam?

$$\mathbf{P}[F|E] = \mathbf{P}[\text{spam}|\text{"Dear"}]$$

### Bayes' theorem

For any events  $E$  and  $F$  where  $\mathbf{P}[E] > 0$  and  $\mathbf{P}[F] > 0$ ,

$$\mathbf{P}[F|E] = \frac{\mathbf{P}[E|F]\mathbf{P}[F]}{\mathbf{P}[E]}$$

and in expanded form,

$$\mathbf{P}[F|E] = \frac{\mathbf{P}[E|F]\mathbf{P}[F]}{\mathbf{P}[E|F]\mathbf{P}[F] + \mathbf{P}[E|F^c]\mathbf{P}[F^c]} = \frac{\mathbf{P}[E|F]\mathbf{P}[F]}{\sum_{i=1}^n \mathbf{P}[E|F_i]\mathbf{P}[F_i]}$$

using the Law of Total Probability. Note that all events  $F_i$  must be mutually exclusive (non-overlapping) and exhaustive (their union is the complete sample space).



## Example – Do it at home

### Example

60% of all email in 2022 is spam. 20% of spam contains the word "Dear". 1% of non-spam contains the word "Dear". What is the probability that an email is spam given it contains the word "Dear"?

Answer

- Let event  $E$  = "Dear", event  $F$  = spam.
- $\mathbf{P}[F] = 0.6$  thus  $\mathbf{P}[F^c] = 0.4$ .
- $\mathbf{P}[E|F] = 0.2$ .
- $\mathbf{P}[E|F^c] = 0.01$ .
- Compute  $\mathbf{P}[F|E]$ .

$$\begin{aligned}\mathbf{P}[F|E] &= \frac{\mathbf{P}[E|F]\mathbf{P}[F]}{\mathbf{P}[E|F]\mathbf{P}[F] + \mathbf{P}[E|F^c]\mathbf{P}[F^c]} = \\ &= \frac{(0.2)(0.6)}{(0.2)(0.6) + (0.01)(0.4)} \approx 0.968\end{aligned}$$



## Bayes' terminology

The diagram shows the equation for Bayes' theorem with callouts for each term:

$$\mathbf{P}[F|E] = \frac{\mathbf{P}[E|F] \cdot \mathbf{P}[F]}{\mathbf{P}[E]}$$

Callouts:

- posterior:  $\mathbf{P}[F|E]$
- likelihood:  $\mathbf{P}[E|F]$
- prior:  $\mathbf{P}[F]$
- normalisation constant:  $\mathbf{P}[E]$

$F$ : hypothesis,  $E$ : evidence

$\mathbf{P}[F]$ : "prior probability" of hypothesis

$\mathbf{P}[E|F]$ : probability of evidence given hypothesis (likelihood)

$\mathbf{P}[E]$ : calculated by making sure that probabilities of all outcomes sum to 1 (they are "normalised")

## Confusion matrix (error matrix)

Used in classification tasks for predicting output error.

		True condition	
		Condition positive $F$	Condition negative $F^c$
Predicted condition	Predicted condition positive $E$	<b>True positive</b> $\mathbf{P}[E F]$	<b>False positive</b> $\mathbf{P}[E F^c]$
	Predicted condition negative $E^c$	<b>False negative</b> $\mathbf{P}[E^c F]$	<b>True negative</b> $\mathbf{P}[E^c F^c]$



## Medical testing example

### Example

- A test is 98% effective at detecting the disease COVID-19 ("true positive").
- The test has a "false positive" rate of 1%.
- 0.5% of the population has COVID-19.
- What is the likelihood you have COVID-19 if you test positive?

Answer

- Let  $E$ : test positive,  $F$ : actually have COVID-19.
- Need to find  $\mathbf{P}[F|E]$ .
- We know:
  - $\mathbf{P}[E|F] = 0.98$
  - $\mathbf{P}[E|F^c] = 0.01$
  - $\mathbf{P}[F] = 0.005$  thus  $\mathbf{P}[F^c] = 0.995$

- Thus

$$\begin{aligned}\mathbf{P}[F|E] &= \frac{\mathbf{P}[E|F]\mathbf{P}[F]}{\mathbf{P}[E|F]\mathbf{P}[F] + \mathbf{P}[E|F^c]\mathbf{P}[F^c]} = \\ &= \frac{(0.98)(0.005)}{(0.98)(0.005) + (0.01)(0.995)} \approx 0.33\end{aligned}$$



## Bayesian intuition

- 33% chance of having COVID-19 after testing positive may seem surprising.
- But the space of facts is now **conditioned** on a positive test result (people who test positive and have COVID-19 **and** people who test positive and don't have COVID-19).

	$F$ yes disease	$F^c$ no disease
$E$ test+	True positive $\mathbf{P}[E F] = 0.98$	False positive $\mathbf{P}[E F^c] = 0.01$
$E^c$ test-	False negative $\mathbf{P}[E^c F] = 0.02$	True negative $\mathbf{P}[E^c F^c] = 0.99$

- But what is a chance of having COVID-19 if you test and it comes back negative?

$$\mathbf{P}[F|E^c] = \frac{\mathbf{P}[E^c|F]\mathbf{P}[F]}{\mathbf{P}[E^c|F]\mathbf{P}[F] + \mathbf{P}[E^c|F^c]\mathbf{P}[F^c]} \approx 0.0001$$

- We update our beliefs with Bayes' theorem:  
I have 0.5% chance of having COVID-19. I take the test:
  - Test is positive: **I now have 33% chance of having COVID-19.**
  - Test is negative: **I now have 0.01% chance of having COVID-19.**
- So it makes sense to take the test.



# Outline

---

Logistics, motivation, background

Conditional probability

Bayes' Theorem

Independence



### Independence

Two events  $E$  and  $F$  are independent if and only if

$$\mathbf{P}[EF] = \mathbf{P}[E]\mathbf{P}[F]$$

Otherwise, they are called dependent events.

In general,  $n$  events  $E_1, E_2, \dots, E_n$  are mutually independent if for every subset of these events with  $r$  elements (where  $r \leq n$ ) it holds that

$$\mathbf{P}[E_a E_b \cdots E_r] = \mathbf{P}[E_a]\mathbf{P}[E_b] \cdots \mathbf{P}[E_r]$$

Therefore for 3 events  $E, F, G$  to be independent, we must have

$$\mathbf{P}[EFG] = \mathbf{P}[E]\mathbf{P}[F]\mathbf{P}[G]$$

$$\mathbf{P}[EF] = \mathbf{P}[E]\mathbf{P}[F]$$

$$\mathbf{P}[EG] = \mathbf{P}[E]\mathbf{P}[G]$$

$$\mathbf{P}[FG] = \mathbf{P}[F]\mathbf{P}[G]$$



## Independence of complement

Notice an equivalent definition for independent events  $E$  and  $F$  ( $\mathbf{P}[F] > 0$ )

$$\mathbf{P}[E|F] = \mathbf{P}[E]$$

Proof:

$$\mathbf{P}[E|F] = \frac{\mathbf{P}[EF]}{\mathbf{P}[F]} = \frac{\mathbf{P}[E]\mathbf{P}[F]}{\mathbf{P}[F]} = \mathbf{P}[E]$$

Independence of complement

If events  $E$  and  $F$  are independent, then  $E$  and  $F^c$  are independent:

$$\mathbf{P}[EF^c] = \mathbf{P}[E]\mathbf{P}[F^c]$$

Proof:

$$\begin{aligned}\mathbf{P}[EF^c] &= \mathbf{P}[E] - \mathbf{P}[EF] = \mathbf{P}[E] - \mathbf{P}[E]\mathbf{P}[F] = \\ &= \mathbf{P}[E](1 - \mathbf{P}[F]) = \mathbf{P}[E]\mathbf{P}[F^c]\end{aligned}$$



## Example

### Example

Each roll of a die is an independent trial. We have two rolls of  $D_1$  and  $D_2$ . Let event  $E : D_1 = 1$ ,  $F : D_2 = 6$  and event  $G : D_1 + D_2 = 7$  (thus  $G = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ ).

1. Are  $E$  and  $F$  independent?
2. Are  $E$  and  $G$  independent?
3. Are  $E, F, G$  independent?

Answer

1. Yes, since  $\mathbf{P}[E] = \frac{1}{6}$ ,  $\mathbf{P}[F] = \frac{1}{6}$  and  $\mathbf{P}[EF] = \frac{1}{36}$ .
2. Yes, since  $\mathbf{P}[E] = \frac{1}{6}$ ,  $\mathbf{P}[G] = \frac{6}{36} = \frac{1}{6}$  and  $\mathbf{P}[EG] = \frac{1}{36}$ .
3. No, since  $\mathbf{P}[EFG] = \frac{1}{36} \neq \frac{1}{6} \frac{1}{6} \frac{1}{6}$ .



### Conditional independence

Two events  $E$  and  $F$  are called conditionally independent given a third event  $G$  if

$$\mathbf{P}[EF|G] = \mathbf{P}[E|G]\mathbf{P}[F|G]$$

Or equivalently,

$$\mathbf{P}[E|FG] = \mathbf{P}[E|G]$$

Notice that:

- Dependent events can become conditionally independent.
- Independent events can become conditionally dependent.
- Knowing when conditioning breaks or creates independence is a big part of building complex probabilistic models.



### Example

Each roll of a die is an independent trial. We have two rolls of  $D_1$  and  $D_2$ . Let event  $E : D_1 = 1$ ,  $F : D_2 = 6$  and event  $G : D_1 + D_2 = 7$  (thus  $G = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ ).

1. Are  $E$  and  $F$  independent?
2. Are  $E$  and  $F$  independent given  $G$ ?

Answer

1. Yes, since  $\mathbf{P}[E] = \frac{1}{6}$ ,  $\mathbf{P}[F] = \frac{1}{6}$  and  $\mathbf{P}[EF] = \frac{1}{36}$ .
2. No, since  $\mathbf{P}[E|G] = \frac{1}{6}$  and  $\mathbf{P}[F|G] = \frac{1}{6}$ , but  $\mathbf{P}[EF|G] = \frac{1}{6} \neq \mathbf{P}[E|G]\mathbf{P}[F|G]$ .

## Summary of conditional probability

---

Conditioning on event  $G$ :

Name of rule	Original rule	Conditional rule
1st axiom of probability	$0 \leq \mathbf{P}[E] \leq 1$	$0 \leq \mathbf{P}[E G] \leq 1$
Complement	$\mathbf{P}[E] = 1 - \mathbf{P}[E^c]$	$\mathbf{P}[E G] = 1 - \mathbf{P}[E^c G]$
Chain rule	$\mathbf{P}[EF] = \mathbf{P}[E F]\mathbf{P}[F]$	$\mathbf{P}[EF G] = \mathbf{P}[E FG]\mathbf{P}[F G]$
Bayes' theorem	$\mathbf{P}[F E] = \frac{\mathbf{P}[E F]\mathbf{P}[F]}{\mathbf{P}[E]}$	$\mathbf{P}[F EG] = \frac{\mathbf{P}[E FG]\mathbf{P}[F G]}{\mathbf{P}[E G]}$



# Introduction to Probability

Lecture 2: Random variables, probability mass function, expectation

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology  
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk



UNIVERSITY OF  
CAMBRIDGE

# Outline

---

Random variable

Probability mass function

Cumulative distribution function

Expectation



## What is a random variable?

### Random variable

A random variable  $X$  is a function from the sample space to the real numbers.

- We can interpret  $X$  as a quantity whose value depends on the outcome of an experiment (some probabilistic process).
  - Roll two dice,  $X$ : sum of dice
  - Toss 3 coins,  $X$ : number of heads
  - Give a student a test,  $X$ : score
  - Stock market index
- Or can think of  $X$  as a variable in a programming language that takes on values, has a type, and has a domain over which it is applicable.
- Many different types of RV: indicator, binary, choice, Bernoulli, etc.
- Random variable can be **discrete** or continuous:
  - $X$  has finitely many possible values: discrete.
  - $X$  has every integer as a possible value: discrete.
  - $X$  amount of time it takes to finish a race: continuous (possible value:  $\{t : 0 \leq t < \infty\} = [0, \infty)$ ).



### Example

We toss 3 fair coins. Let a **random variable**  $X$  be the total number of heads on the 3 coins. What are the probabilities of  $X$  taking on the following values:  $X = 0$ ,  $X = 1$ ,  $X = 2$ ,  $X = 3$ ,  $X \geq 4$ ?

Answer

1.  $\mathbf{P}[X = 0] = \frac{1}{8}$  where set of outcomes is  $\{(T, T, T)\}$
2.  $\mathbf{P}[X = 1] = \frac{3}{8}$  where set of outcomes is  $\{(H, T, T), (T, H, T), (T, T, H)\}$
3.  $\mathbf{P}[X = 2] = \frac{3}{8}$  where set of outcomes is  $\{(H, H, T), (T, H, H), (H, T, H)\}$
4.  $\mathbf{P}[X = 3] = \frac{1}{8}$  where set of outcomes is  $\{(H, H, H)\}$
5.  $\mathbf{P}[X \geq 4] = 0$  where set of outcomes is  $\{\}$

## Random variables are NOT events

random variables  $\neq$  events

### Tossing 3 fair coins example

$X = x$	$\mathbf{P}[X = x]$	Set of outcomes	Possible event $E$
$X = 0$	$\frac{1}{8}$	$\{(T, T, T)\}$	Toss 0 heads
$X = 1$	$\frac{3}{8}$	$\{(H, T, T), (T, H, T), (T, T, H)\}$	Toss exactly 1 head
$X = 2$	$\frac{3}{8}$	$\{(H, H, T), (T, H, H), (H, T, H)\}$	Event where $X = 2$ Toss exactly 2 heads
$X = 3$	$\frac{1}{8}$	$\{(H, H, H)\}$	Toss 0 tails
$X \geq 4$	0	$\{\}$	Toss 4 or more heads

We can define events by condition of the value of a random variable (RV takes on values that satisfy a numerical test).



### Example

Tossing a coin has the probability  $p$  that it comes up heads. Toss a coin 5 times. Let  $X$ : the number of heads in 5 tosses. What is the **range** of  $X$  (i.e., what are the values that  $X$  can take on with non-zero probability)? What is  $\mathbf{P}[X = k]$  where  $k$  is in the range of  $X$ ?

Answer

- Notice that each coin toss is an independent trial.
- Recall  $\mathbf{P}[2 \text{ heads}] = \binom{5}{2}p^2(1-p)^3$ ,  $\mathbf{P}[3 \text{ heads}] = \binom{5}{3}p^3(1-p)^2$ .
- Range of  $X$  :  $\{0, 1, 2, 3, 4, 5\}$
- $\mathbf{P}[X = k] = \binom{5}{k}p^k(1-p)^{5-k}$



# Outline

---

Random variable

Probability mass function

Cumulative distribution function

Expectation



## Probability mass function definition (PMF)

Discrete random variable

A random variable  $X$  is **discrete** if its range has countably many values

$$X = x \text{ where } x \in \{x_1, x_2, x_3, \dots\}$$

Probability mass function

The probability mass function (**PMF**) of a discrete random variable  $X$  is a function  $p(a)$  of  $X$  that maps possible outcomes of a random variable to the corresponding probabilities:

$$p(a) = \mathbf{P}[X = a] = p_X(a)$$

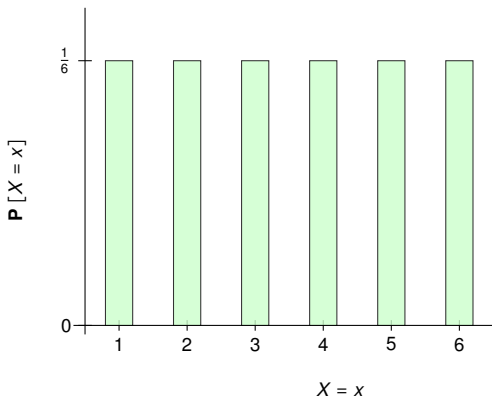
Recall that probabilities must sum to 1:  $\sum_{i=1}^{\infty} p(a_i) = 1$ .



## Example for a single die

- Let  $X$  be a RV representing a single die roll.
- Range of  $X$  :  $\{1, 2, 3, 4, 5, 6\}$ , thus  $X$  is a **discrete** RV.
- PMF of  $X$ :

$$p(x) = \mathbf{P}[X = x] = \begin{cases} \frac{1}{6} & x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

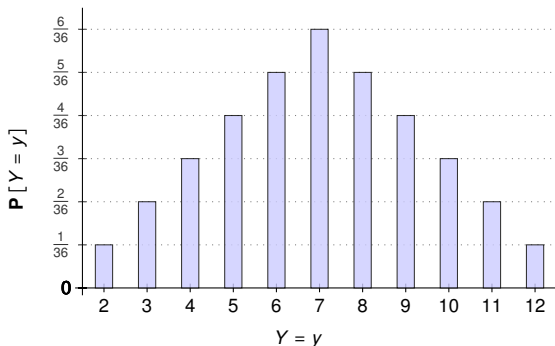


## Example for two dice

- Let  $Y$  be a RV representing the sum of two independent dice rolls.
- Range of  $Y$  :  $\{2, 3, \dots, 11, 12\}$ .
- PMF of  $Y$ :

$$p(y) = \mathbb{P}[Y = y] = \begin{cases} \frac{y-1}{36} & y \in \mathbb{Z}, 2 \leq y \leq 6 \\ \frac{13-y}{36} & y \in \mathbb{Z}, 7 \leq y \leq 12 \\ 0 & \text{otherwise} \end{cases}$$

- Check  $\sum_{y=2}^{12} p(y) = 1$ .



## Properties of PMF

---

Let possible values of  $X = \{a_1, a_2, a_3, \dots\}$ .

1. By Axiom 1:  $0 \leq p(a_i) \leq 1$ .
2.  $p(a) = 0$  if  $a$  is not a possible value.

3. By Axiom 3:  $\sum_{i=1}^{\infty} p(a_i) = 1$ .

$$\sum_{i=1}^{\infty} p(a_i) = \sum_{i=1}^{\infty} \mathbf{P}[X = a_i] = \mathbf{P}\left[\bigcup_{i=1}^{\infty} \{X = a_i\}\right] = \mathbf{P}[S] = 1$$

4. Notice that everything to do with discrete RVs is expressed in terms of (finite or infinite) sum.
5. For continuous RVs, these sums are replaced by integrals.



# Outline

---

Random variable

Probability mass function

Cumulative distribution function

Expectation



## Cumulative distribution function definition (CDF)

Another useful way to analyse probabilities.

### Cumulative distribution function

The cumulative distribution function (CDF) of a random variable  $X$  is defined as

$$F(a) = F_X(a) = \mathbf{P}[X \leq a] \text{ where } -\infty < a < \infty$$

For a **discrete** random variable  $X$ , the CDF is

$$F(a) = \mathbf{P}[X \leq a] = \sum_{\text{all } x \leq a} p(x)$$

Note that for a discrete RV the CDF is a step function, i.e., the value of  $F$  is constant in the intervals  $(x_{i-1}, x_i)$  and then takes a step of size  $p(x_i)$  at  $x_i$ .

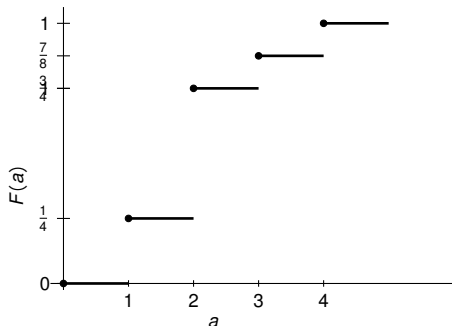


## Example

- Let the PMF for  $X$  be given by  $p(1) = \frac{1}{4}, p(2) = \frac{1}{2}, p(3) = \frac{1}{8}, p(4) = \frac{1}{8}$ .
- Then CDF is:

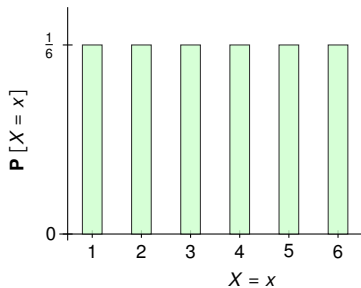
$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{4} & 1 \leq a < 2 \\ \frac{3}{4} & 2 \leq a < 3 \\ \frac{7}{8} & 3 \leq a < 4 \\ 1 & 4 \leq a \end{cases}$$

- Graphical depiction of function:

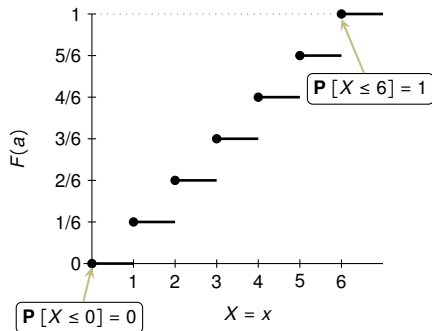


## Example for a single die

PMF of  $X$



CDF of  $X$



## Properties of CDF

---

1.  $0 \leq F(x) \leq 1$  for all  $x$
2.  $\lim_{x \rightarrow -\infty} F(x) = 0$
3.  $\lim_{x \rightarrow \infty} F(x) = 1$
4.  $F(x)$  is a non-decreasing function of  $x$  (if  $x_1 < x_2$  then  $F(x_1) \leq F(x_2)$ )



# Outline

---

Random variable

Probability mass function

Cumulative distribution function

Expectation



### Expectation

The expectation of a discrete random variable  $X$  is defined as

$$\mathbf{E}[X] = \sum_{x:\mathbf{P}[x]>0} x\mathbf{P}[x]$$

- Expectation is the average value of the random variable over many repetitions of the experiment it represents.
- It is the sum over all values of  $X = x$  that have non-zero probability.
- AKA: mean, expected value, weighted average, centre of mass, first moment.



## Example of a die roll

---

What is the expected value of a 6-sided die roll (i.e., what is the average value of a die roll)?

1. Define random variables:

$X = \text{RV for value of roll}$

$$\mathbf{P}[X = x] = \begin{cases} \frac{1}{6} & x \in \{1, \dots, 6\} \\ 0 & \text{otherwise} \end{cases}$$

2. Solve:

$$\mathbf{E}[X] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{7}{2}$$



## Example of school classes

### Example

A school has 3 classes with 5, 10 and 150 students. What is the average class size?

Answer

**Interpretation 1:** Randomly choose a class with equal probability. Thus,  $X$  = size of chosen class

$$\mathbf{E}[X] = 5\left(\frac{1}{3}\right) + 10\left(\frac{1}{3}\right) + 150\left(\frac{1}{3}\right) = \frac{165}{3} = 55$$

**Interpretation 2:** Randomly choose a student with equal probability. Thus,  $Y$  = size of chosen class

$$\mathbf{E}[Y] = 5\left(\frac{5}{165}\right) + 10\left(\frac{10}{165}\right) + 150\left(\frac{150}{165}\right) = \frac{22635}{165} = 137$$

This is a general phenomenon: it occurs because the more students are in a class, the more likely it is that a randomly chosen student would be in that class. As a result, bigger classes are given more weight than smaller classes.



## Example of Roulette Version 1

### Example

A roulette wheel has 36 places numbered from 1 to 36. In addition, 18 of them are coloured red and the other 18 are coloured black. A ball is thrown to take one of 36 places. A gambler can bet:

- on the colour of the place that the ball takes. A correct, either red or black, place wins them a 1 to 1 ratio payout;
- on the number of the place that the ball takes. A correct number wins them a 35 to 1 ratio payout.

What is the expected value if a gambler bets on

1. the colour of the place in the roulette;
2. the number of the place in the roulette that the ball will fall.

Are the two different betting games fair?

\_\_\_\_\_ Answer \_\_\_\_\_



## Example of Roulette Version 1 Cont.

### Example

What is the expected value if a gambler bets on

1. the colour of the place in the roulette;
2. the number of the place in the roulette that the ball will fall.

Are the two different betting games fair?

Answer

1. Let  $E_X$  : bet on colour.

- If win, then  $X = 1$ . Thus  $\mathbf{P}[win_X] = \frac{1}{2}$ .
- If loose, then  $X = -1$ . Thus  $\mathbf{P}[loose_X] = \frac{1}{2}$ .
- Thus,  $\mathbf{E}[X] = (-1)(\frac{1}{2}) + (1)(\frac{1}{2}) = 0$ , This game is "fair".

2. Let  $E_Y$  : bet on number.

- If win, then  $Y = 35$ . Thus  $\mathbf{P}[win_Y] = \frac{1}{36}$ .
- If loose, then  $Y = -1$ . Thus  $\mathbf{P}[loose_Y] = \frac{35}{36}$ .
- Thus,  $\mathbf{E}[Y] = (-1)(\frac{35}{36}) + (35)(\frac{1}{36}) = 0$ , This game is "fair" too.



## Example of Roulette Version 2

### Example

Change the game to add two green places, 0 and 00. Now there are a total of 38 places. Payouts are the same as before. What are the expected values now?

Answer

1. Let  $E_X$  : bet on red colour.

$$\text{Thus, } \mathbf{E}[X] = (-1)\left(\frac{20}{38}\right) + (1)\left(\frac{18}{38}\right) = -\frac{1}{19}.$$

2. Let  $E_Y$  : bet on number 10.

$$\text{Thus, } \mathbf{E}[Y] = (-1)\left(\frac{37}{38}\right) + (35)\left(\frac{1}{38}\right) = -\frac{1}{19}.$$

So, no, these games are not fair, as the gambler would lose  $\pounds \frac{1}{19} = 5.3$  pence per game.



# Introduction to Probability

Lecture 3: Expectation properties, variance, discrete distributions

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology  
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk



UNIVERSITY OF  
CAMBRIDGE

Properties of expectation

Variance

Bernoulli discrete random variable

Binomial discrete random variable



## Properties of expectation: linearity

### Linearity of expectation

Expectations preserve linearity: if  $a$  and  $b$  are constants, then

$$\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$$

**Proof:**

$$\begin{aligned}\mathbf{E}[aX + b] &= \sum_{x:p(x)>0} (ax + b)p(x) \\ &= \left( a \sum_{x:p(x)>0} xp(x) \right) + \left( b \sum_{x:p(x)>0} p(x) \right) \\ &= a\mathbf{E}[X] + b(1) = a\mathbf{E}[X] + b\end{aligned}$$

### Example

Let the event be a roll of a 6-sided die,  $X$  its random variable, and  $Y$  another random variable where  $Y = 3X + 1$ . What are the expected values  $\mathbf{E}[X]$  and  $\mathbf{E}[Y]$ ?

Answer

We know from last time that  $\mathbf{E}[X] = 3.5$ . Thus  $\mathbf{E}[Y] = 3 \cdot 3.5 + 1 = 11.5$ .



## Properties of expectation: additivity

Additivity of expectation

Expectation of a sum is equal to the sum of expectations: if  $X$  and  $Y$  are any random variables on the same sample space then

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

Example

Let the events be rolls of 2 dice, and  $X$  the random variable for the roll of die 1, and  $Y$  for the roll of die 2. What is the expected value of the sum of the rolls of the two dice?

Answer

$$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y] = 3.5 + 3.5 = 7$$



## Properties of expectation: LOTUS

Law of the unconscious statistician (LOTUS)

Let  $X$  be a random variable, and  $Y$  another random variable that is a function of  $X$ , so  $Y = g(X)$ . Let  $p(x)$  be a PMF of  $X$ . Then

$$\mathbf{E}[Y] = \mathbf{E}[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$$

Note how now we no longer need to know PMF of  $Y$ .

- LOTUS is also known as **expected value of a function of a random variable**.
- Note that the properties of expectation let you avoid defining difficult PMFs.
- Let  $X$  be a discrete RV, then:
  - $\mathbf{E}[X^2]$  is know as the **second moment of  $X$** .
  - $\mathbf{E}[X^n]$  is know as the  **$n^{\text{th}}$  moment of  $X$** .



## Second moment example

### Example

Let  $X$  be a discrete random variable that ranges over the values  $\{-1, 0, 1\}$ , and respective probabilities  $\mathbf{P}[X = -1] = 0.2$ ,  $\mathbf{P}[X = 0] = 0.5$  and  $\mathbf{P}[X = 1] = 0.3$ . Let another random variable  $Y = X^2$  (second moment). What is  $\mathbf{E}[Y]$ ?

Answer

Note that  $Y = g(X) = X^2$  and  $\mathbf{E}[Y] = \mathbf{E}[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$ , thus

$$\mathbf{E}[Y] = (-1)^2(0.2) + 0^2(0.5)^2 + (1)^2(0.3) = 0.5$$



# Outline

---

Properties of expectation

Variance

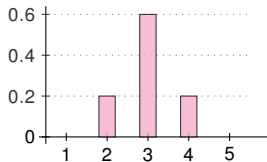
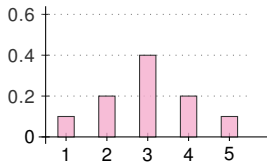
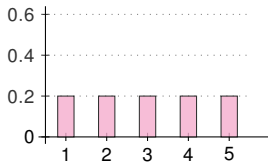
Bernoulli discrete random variable

Binomial discrete random variable



## Spread in the distribution

Expectation is a useful statistic, but it does not give a detailed view of the PMF. Consider these 3 distributions (PMFs).



- Expectation is the same for all distributions:  $\mathbf{E}[X] = 3$ .
- First has the greatest spread, the third has the least spread.
- But the "spread" or "dispersion" of  $X$  in the distribution is very different!
- **Variance**,  $\mathbf{V}[X]$  defines a formal quantification of "spread".
- Several ways to quantify: it uses average square distance from the mean.



## Definition of variance

### Variance

The variance of a discrete random variable  $X$  with expected value (mean)  $\mu$  is:

$$\mathbf{V}[X] = \mathbf{E}[(X - \mu)^2]$$

When computing the variance, we often use a different form of the same equation:

$$\mathbf{V}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

Proof:  $\mathbf{E}[X] = \mu$

$$(X - \mu)^2 = X^2 - 2\mu X + \mu^2$$

$$\begin{aligned}\mathbf{E}[(X - \mu)^2] &= \mathbf{E}[X^2 - 2\mu X + \mu^2] = \mathbf{E}[X^2] - 2\mu\mathbf{E}[X] + \mu^2 \\ &= \mathbf{E}[X^2] - \mu^2 = \mathbf{E}[X^2] - (\mathbf{E}[X])^2\end{aligned}$$

Note:

- $\mathbf{V}[X] \geq 0$
- AKA: Second **central** moment, or square of the standard deviation



## Example with a die roll

### Example

Let  $X$  be the value on one roll of a 6-sided fair die. Recall that  $\mathbf{E}[X] = \frac{7}{2} = 3.5$ . What is  $\mathbf{V}[X]$ ?

Answer

Using  $\mathbf{V}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$ :

$$\mathbf{E}[X^2] = 1^2 \frac{1}{6} + 2^2 \frac{1}{6} + 3^2 \frac{1}{6} + 4^2 \frac{1}{6} + 5^2 \frac{1}{6} + 6^2 \frac{1}{6} = \frac{91}{6}$$

$$\mathbf{V}[X] = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12} = 2.9$$

Using  $\mathbf{V}[X] = \mathbf{E}[(X - \mu)^2] = \mathbf{E}[(X - \mathbf{E}[X])^2]$ :

$$\begin{aligned} \mathbf{V}[X] &= \left(1 - \frac{7}{2}\right)^2 \frac{1}{6} + \left(2 - \frac{7}{2}\right)^2 \frac{1}{6} + \left(3 - \frac{7}{2}\right)^2 \frac{1}{6} + \left(4 - \frac{7}{2}\right)^2 \frac{1}{6} + \\ &+ \left(5 - \frac{7}{2}\right)^2 \frac{1}{6} + \left(6 - \frac{7}{2}\right)^2 \frac{1}{6} = \frac{35}{12} = 2.9 \end{aligned}$$



## Example of spread

### Example

Let  $X$ ,  $Y$  and  $Z$  be discrete random variables with the range  $X : \{10\}$  and probability 1, and  $Y : \{11, 9\}$  and  $Z : \{110, -90\}$  with equal probabilities  $\frac{1}{2}$ . Compute expectation and variance for  $X$ ,  $Y$  and  $Z$ .

$$\text{a) } \mathbf{E}[X] = \sum_x xp(x) = 10 \cdot 1 = \mathbf{10}$$

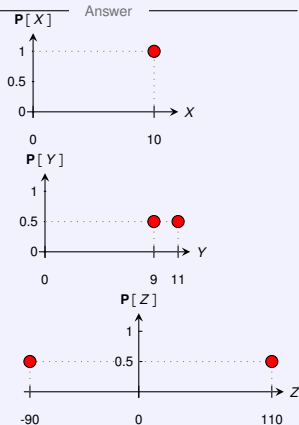
$$\begin{aligned} \mathbf{V}[X] &= \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[(X - 10)^2] \\ &= (X - 10)^2 p(x) = 0^2 \cdot 1 = \mathbf{0} \end{aligned}$$

$$\text{b) } \mathbf{E}[Y] = (11)(0.5) + (9)(0.5) = \mathbf{10}$$

$$\begin{aligned} \mathbf{V}[Y] &= \mathbf{E}[(Y - \mathbf{E}[Y])^2] = \mathbf{E}[(Y - 10)^2] \\ &= (11 - 10)^2(0.5) + (9 - 10)^2(0.5) = \mathbf{1} \end{aligned}$$

$$\text{c) } \mathbf{E}[Z] = (110)(0.5) + (-90)(0.5) = \mathbf{10}$$

$$\begin{aligned} \mathbf{V}[Z] &= \mathbf{E}[(Z - \mathbf{E}[Z])^2] = \mathbf{E}[(Z - 10)^2] = \\ &= (110 - 10)^2(0.5) + (-90 - 10)^2(0.5) \\ &= \mathbf{100^2} = \mathbf{10000} \end{aligned}$$



## Standard deviation

---

- Standard deviation is a kind of average distance of a sample of the mean, i.e., a root mean square (RMS) average.
- Variance is the square of this average distance.

Standard deviation

Standard deviation is defined as a square root of variance:

$$\mathbf{SD} [ X ] = \sqrt{\mathbf{V} [ X ]}$$

Note:

- $\mathbf{E} [ X ]$  and  $\mathbf{V} [ X ]$  are real numbers, not RVs.
- $\mathbf{V} [ X ]$  is expressed in units of the values in the range of  $X^2$ .
- $\mathbf{SD} [ X ]$  is expressed in units of the values in the range of  $X$ .
- For the spread example above:  $\mathbf{SD} [ X ] = 0$ ,  $\mathbf{SD} [ Y ] = 1$ ,  $\mathbf{SD} [ Z ] = 100$ .



- **Property 1:**  $V[X] = E[X^2] - (E[X])^2$
- **Property 2:** variance is **not linear**:  $V[aX + b] = a^2V[X]$

Proof:

$$\begin{aligned}V[aX + b] &= E[(aX + b)^2] - (E[aX + b])^2 \\&= E[a^2X^2 + 2abX + b^2] - (aE[X] + b)^2 \\&= a^2E[X^2] + 2abE[X] + b^2 - (a^2(E[X])^2 + 2abE[X] + b^2) \\&= a^2E[X^2] - (a^2(E[X])^2) = a^2(E[X^2] - (E[X])^2) \\&= a^2V[X]\end{aligned}$$

## Summary of expectation and variance for discrete RV

---

$$\mathbf{E}[X] = \sum_{x:\mathbf{P}[x]>0} x\mathbf{P}[x] = \sum_x xp(x)$$

### Properties of Expectation

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

$$\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$$

$$\mathbf{E}[g(X)] = \sum_x g(x)p_X(x)$$

### Properties of Variance

$$\mathbf{V}[X] = \mathbf{E}[(X - \mu)^2]$$

$$\mathbf{V}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

$$\mathbf{V}[aX + b] = a^2\mathbf{V}[X]$$



- There is deluge of classic RV abstractions that show up in problems.
- They give rise to significant discrete distributions.
- If problem fits, use precalculated (parametric) PMF, expectation, variance and other properties by providing parameters of the problem.
- We will cover the following RVs:
  1. Bernoulli
  2. Binomial
  3. Poisson
  4. Geometric
  5. Negative Binomial
  6. Hypergeometric



# Outline

---

Properties of expectation

Variance

**Bernoulli discrete random variable**

Binomial discrete random variable



### Bernoulli discrete random variable

A Bernoulli RV  $X$  maps "success" of an experiment to 1 and "failure" to 0. It is AKA indicator RV, boolean RV.  $X$  is "Bernoulli RV with parameter  $p$ ", where  $\mathbf{P}[\text{"success"}] = p$  and so PMF  $p(1) = p$ .

$$\mathbf{X} \sim \mathbf{Ber}(p)$$

$$\text{Range: } \{0, 1\}$$

$$\text{PMF: } \mathbf{P}[X = 1] = p(1) = p$$

$$\mathbf{P}[X = 0] = p(0) = 1 - p$$

$$\text{Expectation: } \mathbf{E}[X] = p$$

$$\text{Variance: } \mathbf{V}[X] = p(1 - p)$$

Examples: coin toss, random binary digit, if someone likes a film, the gender of a newborn baby, pass/fail of you taking an exam.



## Bernoulli examples

### Example

You watch a film on Netflix. At the end you click "like" with probability  $p$ . Define a RV representing this event.

Answer

- $X$ : 1 if "like"-d
- $X \sim \text{Ber}(p)$
- $\mathbf{P}[X = 1] = p, \mathbf{P}[X = 0] = 1 - p$

### Example

Two fair 6-sided dice are rolled. Define a random variable  $X$  for a successful roll of two 6's, and failure for anything else.

Answer

- $X$ : 1 if "success" of rolling two 6's
- $X \sim \text{Ber}(p)$
- $\mathbf{P}[X = 1] = \frac{1}{36}$



# Outline

---

Properties of expectation

Variance

Bernoulli discrete random variable

**Binomial discrete random variable**



# Binomial

## Binomial discrete random variable

A Binomial RV  $X$  represents the number of successes in  $n$  successive independent trials of a Bernoulli experiment.  $X \sim \text{Bin}(n, p)$  is a Binomial RV, where  $p$  is the probability of success in a given trial:

$$X \sim \text{Bin}(n, p)$$

is distributed as a

Range:  $\{0, 1, \dots, n\}$

with parameters

Probability that  $X$  takes on the value  $k$

$$\mathbf{P}[X = k] = p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Expectation:  $\mathbf{E}[X] = np$

Variance:  $\mathbf{V}[X] = np(1-p)$

Probability Mass Function for a Binomial

Examples: # heads in  $n$  coin tosses, # of 1's in randomly generated length  $n$  bit string

Note: by Binomial theorem (revision), we can prove  $\sum_{k=0}^n \mathbf{P}[X = k] = 1$ .



## Binomial example

### Example

Let  $X$  be the number of heads after a coin is tossed three times:  
 $X \sim \text{Bin}(3, 0.5)$ . What is the probability of each of the different values of  $X$ ?

Answer

$$\mathbf{P}[X = 0] = p(0) = \binom{3}{0} p^0 (1-p)^3 = \frac{1}{8}$$

$$\mathbf{P}[X = 1] = p(1) = \binom{3}{1} p^1 (1-p)^2 = \frac{3}{8}$$

$$\mathbf{P}[X = 2] = p(2) = \binom{3}{2} p^2 (1-p)^1 = \frac{3}{8}$$

$$\mathbf{P}[X = 3] = p(3) = \binom{3}{3} p^3 (1-p)^0 = \frac{1}{8}$$

$$\mathbf{P}[X = 9] = p(9) = 0$$



## Binomial RV is sum of Bernoulli RVs

---

Let  $X$  be a Bernoulli RV:  $X \sim \text{Ber}(p)$ . Let  $Y$  be a Binomial RV:  $Y \sim \text{Bin}(n, p)$ .  
Binomial RV = sum of  $n$  independent Bernoulli RVs:

$$Y = \sum_{i=1}^n X_i, \quad X_i \sim \text{Ber}(p)$$

$$\mathbf{E}[Y] = \mathbf{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{E}[X_i] = np$$

Note:  $\text{Ber}(p) = \text{Bin}(1, p)$



## Another example

### Example

An off-licence sells cases of wine, each containing 20 bottles. The probability that a bottle is bad is 0.05. The off-licence gives a money-back guarantee that the case will contain no more than one bad bottle. What is the probability that the off-licence will have to give money back?

Answer

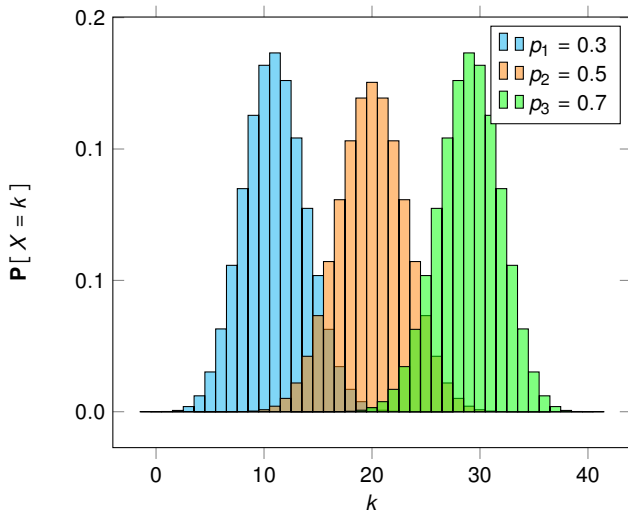
- $X$ : # of bad bottles in a case (20 bottles)
- $\mathbf{P}[\text{have to give money back}] = \mathbf{P}[X \geq 2] = 1 - \mathbf{P}[X = 0] - \mathbf{P}[X = 1]$
- $X$  is a binomial RV with parameters  $X \sim \text{Bin}(n = 20, p = 0.05)$ .
- Bernoulli trial: check if a bottle is bad
- $\mathbf{P}[\text{success}] = \mathbf{P}[\text{bottle is bad}] = 0.05$   
 $\mathbf{P}[\text{failure}] = \mathbf{P}[\text{bottle is good}] = 0.95$
- Recall, when  $X \sim \text{Bin}(n, p)$  then  $\mathbf{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}$  thus

$$\begin{aligned}\mathbf{P}[X \geq 2] &= 1 - \mathbf{P}[X = 0] - \mathbf{P}[X = 1] \\ &= 1 - \binom{20}{0} 0.05^0 0.95^{20} - \binom{20}{1} 0.05^1 0.95^{19} = 0.26\end{aligned}$$



## Visualising Binomial PMFs

$X \sim \text{Bin}(40, 0.3)$ ;  $X \sim \text{Bin}(40, 0.5)$ ;  $X \sim \text{Bin}(40, 0.7)$



# Introduction to Probability

Lecture 4: More discrete distributions – Poisson, Geometric,  
Negative Binomial, Hypergeometric

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology  
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk



UNIVERSITY OF  
CAMBRIDGE

Poisson discrete random variable

Geometric discrete random variable

Negative binomial discrete random variable

Hypergeometric discrete random variable



## Preliminaries:

### The natural exponent $e$

$e$  is a mathematical constant AKA the Euler number.  $e$  is very important for exponential functions. Here are some important identities:

$$e \approx 2.71828$$

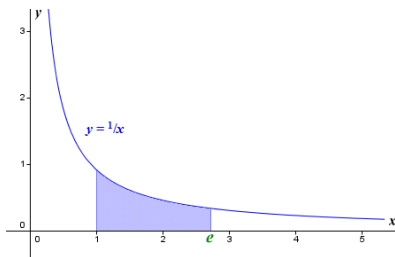
$$e = \sum_{n=0}^{\infty} \frac{1}{n!}$$

$$e = \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{n} \right)^n$$

$$e^{-\lambda} = \lim_{n \rightarrow \infty} \left( 1 - \frac{\lambda}{n} \right)^n$$

$$e^r = \lim_{n \rightarrow \infty} \left( 1 + \frac{r}{n} \right)^n$$

$$e^r = \sum_{n=0}^{\infty} \frac{r^n}{n!}$$



## Binomial RV example: large $n$ , small $p$

We are trying to predict footfall in a store. We know, based on previous data, that on average 8 people enter the store per hour. What is the probability of  $k$  people entering the store in the next 1 hour?

### 1. Break an hour into **minutes**.

What if 2 people enter in the same minute?

- At each **minute**, independent Bernoulli trial with 1 for a person entering the store and 0 for nobody entering the store.
- $X$  is a Binomial RV: # people entering in an hour, so  $\mathbf{E}[X] = np = \lambda = 8$ .
- $X \sim \text{Bin}(n = 60, p = \frac{\lambda}{n})$ , so  $\mathbf{P}[X=k] = \binom{n}{k} p^k (1-p)^{n-k} = \binom{60}{k} \left(\frac{8}{60}\right)^k \left(1 - \frac{8}{60}\right)^{n-k}$

### 2. Break an hour into **milliseconds**.

What if 2 people enter in the same millisecond?

- At each **millisecond**, independent Bernoulli trial: 1 for enter, 0 for not enter.
- $X$  is a Binomial RV: # people entering in an hour, so  $\mathbf{E}[X] = np = \lambda = 8$ .
- $X \sim \text{Bin}(n = 3600000, p = \frac{\lambda}{n})$ , so  $\mathbf{P}[X=k] = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$

### 3. Break an hour into **infinitely small units**.

- At each **unit**, independent Bernoulli trial: 1 for enter, 0 for not enter.
- $X$  is a Binomial RV: # people entering in an hour, so  $\mathbf{E}[X] = np = \lambda = 8$ .
- $X \sim \text{Bin}(n, p = \frac{\lambda}{n})$ , thus  $\mathbf{P}[X=k] = \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$



## Computing Binomial in the limit

$$\begin{aligned} \mathbf{P}[X = k] &= \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \stackrel{\text{expand}}{=} \lim_{n \rightarrow \infty} \frac{n!}{k!(n-k)!} \frac{\lambda^k \left(1 - \frac{\lambda}{n}\right)^n}{n^k} \\ &\stackrel{\text{rearrange}}{=} \lim_{n \rightarrow \infty} \frac{n!}{n^k(n-k)!} \frac{\lambda^k \left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} \stackrel{\text{def of } e}{=} \lim_{n \rightarrow \infty} \frac{n!}{n^k(n-k)!} \frac{\lambda^k e^{-\lambda}}{\left(1 - \frac{\lambda}{n}\right)^k} \\ &\stackrel{\text{expand}}{=} \lim_{n \rightarrow \infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k e^{-\lambda}}{\left(1 - \frac{\lambda}{n}\right)^k} \end{aligned}$$

$$\text{as } n \rightarrow \infty \quad \frac{n(n-1)\cdots(n-k+1)}{n^k} \approx \frac{n^k}{n^k} = 1$$

$$\left(1 - \frac{\lambda}{n}\right)^k \approx 1^k = 1$$

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda} \text{ because } e^{-\lambda} = \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n \text{ thus we have}$$

$$\mathbf{P}[X = k] = (1) \frac{\lambda^k e^{-\lambda}}{k!} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Therefore, in our store footfall example: the probability of  $k$  people entering the store in the next 1 hour is:

$$\mathbf{P}[X=k] = \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$



### Poisson discrete random variable

A Poisson RV  $X$  approximates Binomial where  $n$  is large,  $p$  is small, and  $\lambda = np$  is "moderate". Thus we no longer need to know  $n$  and  $p$ , we only need to provide **rate**  $\lambda$ .  $X$  is the number of successes over the duration of the experiment.

$$X \sim \text{Pois}(\lambda)$$

$$\text{Range: } \{0, 1, 2, \dots\}$$

$$\text{PMF: } \mathbf{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\text{Expectation: } \mathbf{E}[X] = \lambda$$

$$\text{Variance: } \mathbf{V}[X] = \lambda$$

Examples: # earthquakes in a given year, # goals scored during a 90 minute football game, # misprints per page in a book, # emails per day.

**Key idea:** Divide time into a **large number** of small increments. Assume that during each increment, there is some **small probability** of the event happening (independent of other increments).



## Earthquake example

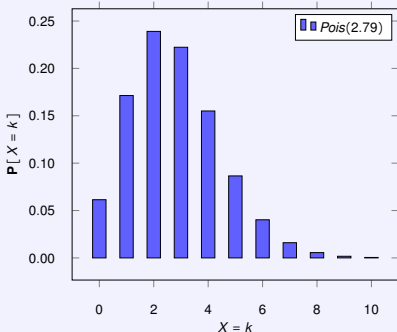
### Example

Suppose there are an average of 2.79 major earthquakes in the world each year. What is the probability of getting 3 major earthquakes next year?

Answer

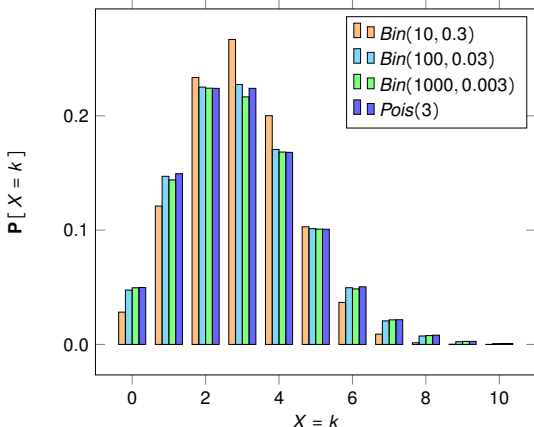
Define RVs:  $\lambda = 2.79, k = 3, X \sim \text{Pois}(2.79)$

$$\text{Solve: } \mathbf{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda} = \mathbf{P}[X = 3] = \frac{2.79^3}{3!} e^{-2.79} \approx 0.23$$



## Poisson paradigm

- Poisson approximates Binomial when  $n$  is large,  $p$  is small, and  $\lambda = np$  is "moderate".
- Different interpretations of "moderate". Commonly accepted ranges are:
  - $n > 20$  and  $p < 0.05$
  - $n > 100$  and  $p < 0.1$
- Poisson is Binomial in the limit:  $\lambda = np$  where  $n \rightarrow \infty, p \rightarrow 0$ .



$$\text{PMF: } = k \in \{0, 1, 2, \dots, \infty\}; \mathbf{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\mathbf{E}[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} =$$

$$= \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \quad (\text{let } i = k - 1)$$

$$= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$



$$\begin{aligned}\mathbf{E}[X^2] &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \quad (\text{let } i = k - 1) \\ &= \lambda \sum_{i=0}^{\infty} (i+1) \frac{\lambda^i}{i!} e^{-\lambda} = \lambda \left( \underbrace{\sum_{i=0}^{\infty} i \frac{\lambda^i}{i!} e^{-\lambda}}_{\text{same as before}} + \underbrace{\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda}}_{\text{sum of PMFs}=1} \right) = \\ &= \lambda(\lambda + 1) \quad \text{thus}\end{aligned}$$

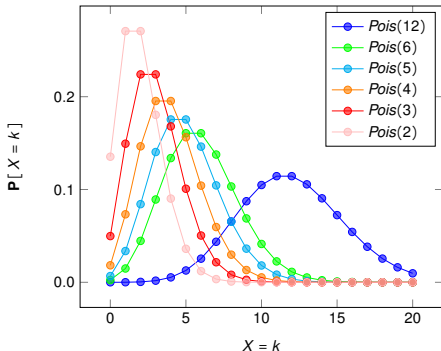
$$\mathbf{V}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda$$

$$\mathbf{E}[X^k] = \lambda \mathbf{E}[(X+1)^{k-1}] \quad \leftarrow \text{useful generalisation}$$



## Bernoulli, Poisson, and random processes

- A Poisson process is a model for a series of discrete events where the **average time** between events is known, but the exact timing of events is random.
  - The arrival of an event is independent of the event before (waiting time between events is memoryless).
  - The average rate (events per time period) is constant.
  - Two events cannot occur at the same time: each sub-interval of a Poisson process is a Bernoulli trial that is either a success or a failure.
- Example: your website goes down on average twice per 60 days; calling a help centre; movements of stock price...



# Outline

---

Poisson discrete random variable

**Geometric discrete random variable**

Negative binomial discrete random variable

Hypergeometric discrete random variable



### Geometric discrete random variable

$X$  is a geometric RV if  $X$  is a number of independent Bernoulli trials until the **first** success, and  $p$  is the probability of success on each Bernoulli trial.

$$X \sim \text{Geo}(p)$$

$$\text{Range: } \{1, 2, \dots\}$$

$$\text{PMF: } \mathbf{P}[X = n] = (1 - p)^{n-1} p$$

$$\text{Expectation: } \mathbf{E}[X] = \frac{1}{p}$$

$$\text{Variance: } \mathbf{V}[X] = \frac{1 - p}{p^2}$$

Examples: tossing a coin ( $\mathbf{P}[\textit{head}] = p$ ) until first heads appears, generating bits with  $\mathbf{P}[\textit{bit} = 1] = p$  until first 1 is generated.



PMF ( $E_i$  is the event that the  $i$ -th trial succeeds):

$$\begin{aligned}\mathbf{P}[X = n] &= \mathbf{P}[E_1^c E_2^c \dots E_{n-1}^c E_n] = \\ &= \mathbf{P}[E_1^c] \mathbf{P}[E_2^c] \dots \mathbf{P}[E_{n-1}^c] \mathbf{P}[E_n] = \\ &= (1 - p)^{n-1} p\end{aligned}$$

CDF ( $\mathbf{P}[X > n]$  is the probability that at least the first  $n$  trials fail):

$$\begin{aligned}\mathbf{P}[X \leq n] &= 1 - \mathbf{P}[X > n] = \\ &= 1 - \mathbf{P}[E_1^c E_2^c \dots E_n^c] = \\ &= 1 - \mathbf{P}[E_1^c] \mathbf{P}[E_2^c] \dots \mathbf{P}[E_n^c] = \\ &= 1 - (1 - p)^n\end{aligned}$$



## Die example

### Example

You roll a fair 6-sided die until it comes up with #6. What is the probability that it will take 3 rolls?

Answer

Let  $X$  be a RV for # of rolls. Probability for any # on die is  $\frac{1}{6}$ .

Define RVs:  $X \sim \text{Geo}(\frac{1}{6})$ , want  $\mathbf{P}[X = 3]$ .

$$\begin{aligned}\text{Solve: } \mathbf{P}[X = 3] &= (1 - p)^{n-1} p \text{ where } n = 3, p = \frac{1}{6} \\ &= \left(1 - \frac{1}{6}\right)^{3-1} \frac{1}{6} = \left(\frac{5}{6}\right)^2 \frac{1}{6} = \frac{25}{216}\end{aligned}$$

$$\mathbf{P}[X = 3] = \mathbf{P}[\text{not } 6, \text{not } 6, 6] = \frac{5}{6} \frac{5}{6} \frac{1}{6} = \frac{25}{216}$$

$$\mathbf{P}[X = n] = \mathbf{P}[\text{not } 6(n-1) \text{ times}, 6] = \left(\frac{5}{6}\right)^{n-1} \left(\frac{1}{6}\right)$$



# Outline

---

Poisson discrete random variable

Geometric discrete random variable

**Negative binomial discrete random variable**

Hypergeometric discrete random variable



## Negative binomial

### Negative binomial discrete random variable

$X$  is a negative binomial RV if  $X$  is the number of independent Bernoulli trials until  $r$  successes and  $p$  is the probability of success on each trial.

$$X \sim \text{NegBin}(r, p)$$

$$\text{Range: } \{r, r + 1, \dots\}$$

$$\text{PMF: } \mathbf{P}[X = n] = \binom{n-1}{r-1} (1-p)^{n-r} p^r$$

$$\text{Expectation: } \mathbf{E}[X] = \frac{r}{p}$$

$$\text{Variance: } \mathbf{V}[X] = \frac{r(1-p)}{p^2}$$

Examples: tossing a coin until  $r$ -th heads appears, generating bits until the first  $r$  1's are generated.

**Note:**  $\text{Geo}(p) = \text{NegBin}(1, p)$ .



Example (not real life!)

A PhD student is expected to publish 2 papers to graduate. A conference accepts each paper randomly and independently with probability  $p = 0.25$ . On average, how many papers will the student need to submit to a conference in order to graduate?

Answer

Let  $X$  be # submissions required to get 2 acceptances. Thus  $X \sim \text{NegBin}(r = 2, p = 0.25)$ . So,

$$\mathbf{E}[X] = \frac{r}{p} = \frac{2}{0.25} = 8$$



## Adding NegBin example

### Example

Let  $X \sim \text{NegBin}(m, p)$  and  $Y \sim \text{NegBin}(n, p)$  be two independent RVs. Define a new RV as  $Z = X + Y$ . Find PMF of  $Z$ .

Answer

- Need to show that  $Z \sim \text{NegBin}(m + n, p)$ .
- Consider the sequence of independent events tossing a coin with  $\mathbf{P}[\text{heads}] = p$ .
- Let  $X$  be a RV for # of coin tosses until  $m$  heads are observed. Thus  $X \sim \text{NegBin}(m, p)$ .
- Now, continue to toss a coin after  $m$  heads are observed, until  $n$  more heads are observed. Thus, for this part of the sequence,  $Y \sim \text{NegBin}(n, p)$ .
- Looking at it from the beginning we tossed independently the coin until we observed  $m + n$  heads, thus  $Z = X + Y$  and thus  $Z \sim \text{NegBin}(m + n, p)$ .
- Note: if  $X_1, X_2, \dots, X_m$  are  $m$  independent  $\text{Geo}(p)$  RVs, then the RV  $X = X_1 + X_2 + \dots + X_m$  has  $\text{NegBin}(m, p)$  distribution.



# Outline

---

Poisson discrete random variable

Geometric discrete random variable

Negative binomial discrete random variable

Hypergeometric discrete random variable



## Hypergeometric

### Hypergeometric discrete random variable

$X$  is a hypergeometric RV that samples  $n$  objects, **without replacement**, with  $i$  successes (random draw for which the object drawn has a specified feature), from a finite population of size  $N$  that contains exactly  $m$  objects with that feature.

$$X \sim \text{Hyp}(N, n, m)$$

$$\text{Range: } \{0, 1, \dots, n\}$$

$$\text{PMF: } \mathbf{P}[X = i] = \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}}$$

$$\text{Expectation: } \mathbf{E}[X] = n \frac{m}{N}$$

$$\text{Variance: } \mathbf{V}[X] = n \frac{m}{N} \left(1 - \frac{m}{N}\right) \left(1 - \frac{n-1}{N-1}\right)$$

Example: an urn has  $N$  balls of which  $m$  are white and  $N - m$  are black; we take a random sample **without replacement** of size  $n$  and measure  $X$ : # of white balls in the sample.



### Example

A street has 40 houses of which 5 houses are inhabited by families with an income below the poverty line. In a survey, 7 houses are sampled at random from this street. What is the probability that: (a) none of the 5 families with income below poverty line are sampled? (b) 4 of them are sampled? (c) no more than 2 are sampled? (d) at least 3 are sampled?

Answer

Let  $X$ : # of families sampled which are below the poverty line.

$X \sim \text{Hyp}(N = 40, n = 7, m = 5)$ .

$$(a) \mathbf{P}[X = 0] = \frac{\binom{5}{0}\binom{40-5}{7-0}}{\binom{40}{7}} = \frac{\binom{35}{7}}{\binom{40}{7}} \approx 0.36$$

$$(b) \mathbf{P}[X = 4] = \frac{\binom{5}{4}\binom{40-5}{7-4}}{\binom{40}{7}}$$

$$(c) \mathbf{P}[X \leq 2] = \mathbf{P}[X = 0] + \mathbf{P}[X = 1] + \mathbf{P}[X = 2]$$

$$(d) \mathbf{P}[X \geq 3] = 1 - \mathbf{P}[X \leq 2]$$

## Summary of discrete RV

	$Ber(p)$	$Bin(n, p)$	$Pois(\lambda)$	$Geo(p)$	$NegBin(r, p)$	$Hyp(N, n, m)$
PMF	$\mathbf{P}[X=1]=p$	$\mathbf{P}[X=k]=\binom{n}{k}p^k(1-p)^{n-k}$	$\mathbf{P}[X=k]=\frac{\lambda^k}{k!}e^{-\lambda}$	$\mathbf{P}[X=n]=(1-p)^{n-1}p$	$\mathbf{P}[X=n]=\binom{n-1}{r-1}(1-p)^{n-r}p^r$	$\mathbf{P}[X=i]=\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$
$\mathbf{E}[X]$	$p$	$np$	$\lambda$	$\frac{1}{p}$	$\frac{r}{p}$	$n\frac{m}{N}$
$\mathbf{V}[X]$	$p(1-p)$	$np(1-p)$	$\lambda$	$\frac{1-p}{p^2}$	$\frac{r(1-p)}{p^2}$	$n\frac{m}{N}(1-\frac{m}{N})(1-\frac{n-1}{N-1})$
Descr.	1 experiment with prob $p$ of success	$n$ independent trials with prob $p$ of success	# successes over experiment duration, $\lambda = np$ rate of success	# independent trials until first success	# independent trials until $r$ successes	# successes of drawing item with a feature (without replacement) in a sample of size $n$ from a population of size $N$ with $m$ items with the feature

# Introduction to Probability

Lecture 5+: Continuous random variables

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology

email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk



UNIVERSITY OF  
CAMBRIDGE

## Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

Exponential random variable

Normal (Gaussian) random variable



## From discrete to continuous RV

---

- So far, all RV were discrete: can only take on integer values.
- If RV need to take on values in the real number domain ( $\mathbb{R}$ ), then continuous random variable.
- Examples of continuous RV: Uniform RV, Exponential RV, Normal RV.
- Continuous RV are just like discrete RV, except that every **sum** becomes an **integral**.
- Example of possible values of continuous RV  $X$ :

$$(0, 1) = \{x \in \mathbb{R}; 0 < x < 1\}$$

$$[0, 1] = \{x \in \mathbb{R}; 0 \leq x \leq 1\}$$

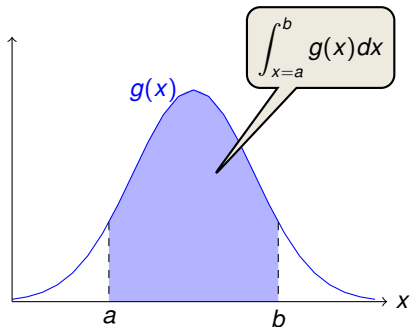
$$[0, 1) = \{x \in \mathbb{R}; 0 \leq x < 1\}$$

$$(-\infty, \infty) = \text{all real numbers}$$

- Examples:
  - $X$ : price of a stock
  - $X$ : time that a machine works before breakdown
  - $X$ : error in an experimental measurement



## Integrals revision



Integral = area under a curve =  $\int_{x=a}^b g(x) dx = G(x) \Big|_a^b = G(b) - G(a)$   
where  $G(x)$  is the antiderivative for  $g(x)$ .

Some examples:

$$\int_a^b x^2 dx = \frac{x^3}{3} \Big|_a^b = \frac{b^3 - a^3}{3}$$

$$\int a dx = ax + C$$

$$\int \frac{1}{x} dx = \ln|x| + C$$

$$\int e^x dx = e^x + C$$

- The most important property of discrete RV was probability mass function (PMF) denoting the probability of the RV taking on a certain value.
- But in the continuous world this is impossible:  
What is the probability that a newborn child weighs **exactly** 3.215438765432532 kg? **NONE**
- Real values are defined with infinite precision, thus the probability that a RV takes on a specific value is not meaningful when the RV is continuous.
- We need a function that says how likely is it that a RV takes on a particular value relative to other values that it could take on: **probability density function**.

## Definition of continuous RV

### Continuous random variable

A random variable  $X$  is continuous if there is a **probability density function (PDF)**,  $f(x) \geq 0$  such that for  $-\infty < x < \infty$ :

$$\mathbf{P}[a \leq X \leq b] = \int_a^b f(x) dx$$

To preserve the axioms that guarantee that  $\mathbf{P}[a \leq X \leq b]$  is a probability, the following properties must hold:

$$0 \leq \mathbf{P}[a \leq X \leq b] \leq 1$$

$$\mathbf{P}[-\infty < X < \infty] = 1 \quad \left( = \int_{-\infty}^{\infty} f(x) dx \right)$$

- Note: we also write  $f(x)$  as  $f_X(x)$ .
- In continuous world, every RV has a PDF: its relative value wrt to other possible values.
- Integrate  $f(x)$  to get probabilities.



## Comparing PMF and PDF

**Discrete** random variable  $X$

Probability mass function (PMF):

$$p(x)$$

Compute probability:

$$\mathbf{P}[X = x] = p(x)$$

$$\mathbf{P}[a \leq X \leq b] = \sum_{x=a}^b p(x)$$

**Continuous** random variable  $X$

Probability density function (PDF):

$$f(x)$$

Compute probability:

$$\mathbf{P}[a \leq X \leq b] = \int_{x=a}^b f(x) dx$$

Both are measures of how **likely** is  $X$  to take on a value.



## Computing probability example

### Example

Let  $X$  be a continuous RV with PDF:

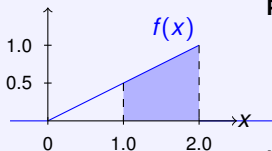
$$f(x) = \begin{cases} \frac{1}{2}x & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

What is  $\mathbf{P}[X \geq 1]$ ?

Answer

Method 1: integrate

$$\begin{aligned} \mathbf{P}[1 \leq X < \infty] &= \int_1^{\infty} f(x) dx = \int_1^2 \frac{1}{2} x dx \\ &= \frac{1}{2} \left( \frac{1}{2} x^2 \right) \Big|_1^2 = \frac{1}{2} \left( 2 - \frac{1}{2} \right) = \frac{3}{4} \end{aligned}$$



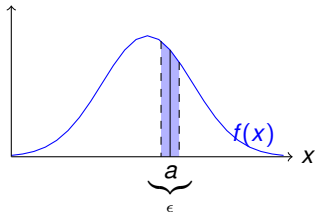
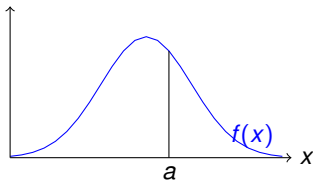
Method 2: areas of triangles

$$\text{area} = \frac{(2)(1)}{2} - \frac{(1)(0.5)}{2} = 1 - \frac{1}{4} = \frac{3}{4}$$



- $f(x)$  is NOT a probability, it is probability density:

$$\mathbf{P}[X = a] = \int_a^a f(x)dx = 0 \neq f(a)$$



$$\mathbf{P}\left[a - \frac{\epsilon}{2} \leq X \leq a + \frac{\epsilon}{2}\right] = \int_{a-\frac{\epsilon}{2}}^{a+\frac{\epsilon}{2}} f(x)dx \approx \text{width} \times \text{height} = \epsilon f(a)$$

$$\text{Thus, } \mathbf{P}[X = a] = \lim_{\epsilon \rightarrow 0} \epsilon f(a) = 0.$$

- $\mathbf{P}[a \leq X \leq b] = \mathbf{P}[a < X \leq b] = \mathbf{P}[a \leq X < b] = \mathbf{P}[a < X < b]$

## PDF and probability example

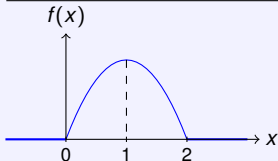
### Example

Let  $X$  be a continuous RV with PDF:

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{when } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the value of the constant  $C$ ? What is  $\mathbf{P}[X > 1]$ ?

Answer



$C$  is a normalisation constant. We know that PDF must sum to 1:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^0 f(x) dx + \int_0^2 f(x) dx + \int_2^{\infty} f(x) dx \\ &= \int_{-\infty}^0 0 dx + \int_0^2 C(4x - 2x^2) dx + \int_2^{\infty} 0 dx \\ &= \int_0^2 C(4x - 2x^2) dx = C \left( 2x^2 - \frac{2x^3}{3} \right) \Big|_0^2 = C \left( 8 - \frac{16}{3} \right) = C \frac{8}{3} \end{aligned}$$

$$\text{Thus } C = \frac{3}{8}$$



## PDF and probability example cont.

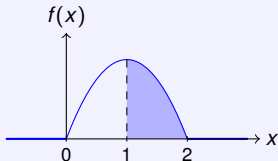
### Example

Let  $X$  be a continuous RV with PDF:

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{when } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the value of the constant  $C$ ? What is  $\mathbf{P}[X > 1]$ ?

Answer



$$\begin{aligned} \mathbf{P}[X > 1] &= \int_1^{\infty} f(x) dx = \int_1^2 f(x) dx + \int_2^{\infty} 0 dx \\ &= \int_1^2 \frac{3}{8}(4x - 2x^2) dx = \frac{3}{8} \left( 2x^2 - \frac{2x^3}{3} \right) \Big|_1^2 = \\ &= \frac{3}{8} \left( \left( 8 - \frac{16}{3} \right) - \left( 2 - \frac{2}{3} \right) \right) = \frac{1}{2} \end{aligned}$$



# Outline

---

Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

Exponential random variable

Normal (Gaussian) random variable



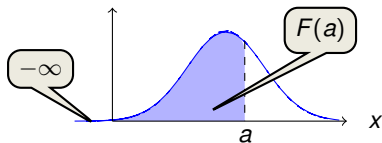
## Cumulative distribution function

- Since PDF is not a probability, we need to solve an integral every single time we want to calculate a probability.
- To save effort, cumulative distribution function (CDF) computes this:  
 $F(a) = F_X(a) = \mathbf{P}[X \leq a]$  where  $-\infty < a < \infty$ .
- Recall: CDF for *discrete* RV is  $F(a) = \sum_{\text{all } x \leq a} p(x)$

### Cumulative distribution function for a continuous RV

For a continuous random variable  $X$  with PDF  $f(x)$ , the cumulative distribution function (CDF) is:

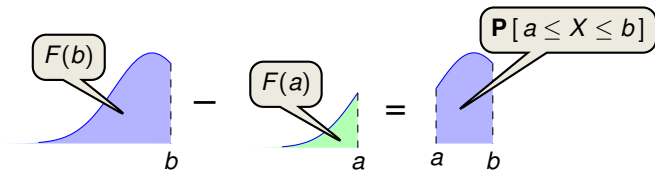
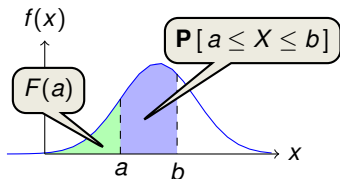
$$F_X(a) = \mathbf{P}[X \leq a] = \int_{-\infty}^a f(x) dx$$



- While PDF is not a probability, CDF is.
- If you learn to use CDFs, you can avoid integrating the PDF.
- It is a matter of convention that CDF is probability that a RV takes on a value **less than** (or equal to) the input value as opposed to greater than.
- Useful examples of using CDF:

<b>Probability question</b>	<b>Solution</b>	<b>Explanation</b>
$\mathbf{P}[X \leq a]$	$F(a)$	Definition of CDF
$\mathbf{P}[X < a]$	$F(a)$	Note that $\mathbf{P}[X = a] = 0$
$\mathbf{P}[X > a]$	$1 - F(a)$	$\mathbf{P}[X \leq a] + \mathbf{P}[X > a] = 1$
$\mathbf{P}[a < X < b]$	$F(b) - F(a)$	$F(a) + \mathbf{P}[a < X < b] = F(b)$

## Computing CDF



$$\begin{aligned} F(b) - F(a) &= \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx \\ &= \left( \int_{-\infty}^a f(x) dx + \int_a^b f(x) dx \right) - \int_{-\infty}^a f(x) dx \\ &= \int_a^b f(x) dx = \mathbf{P}[a < X < b] = \mathbf{P}[a \leq X \leq b] \end{aligned}$$



### Discrete RV $X$

$$\mathbf{E}[X] = \sum_x xp(x)$$

$$\mathbf{E}[g(X)] = \sum_x g(x)p(x)$$

### Continuous RV $X$

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} xf(x)dx$$

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

### Both continuous and discrete RVs

$$\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$$

Linearity of expectation

$$\mathbf{V}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

Properties of

$$\mathbf{V}[aX + b] = a^2\mathbf{V}[X]$$

variance



# Outline

---

Continuous random variables

Cumulative distribution function, expectation, variance

**Uniform random variable**

Exponential random variable

Normal (Gaussian) random variable



## Uniform continuous RV

### Uniform continuous random variable

A uniform continuous random variable  $X$  is defined as follows:

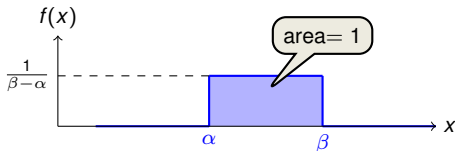
$$X \sim \text{Uni}(\alpha, \beta)$$

Range:  $[\alpha, \beta]$ , sometimes  $(\alpha, \beta)$

$$\text{PDF: } f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{when } \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Expectation: } \mathbf{E}[X] = \frac{\alpha + \beta}{2}$$

$$\text{Variance: } \mathbf{V}[X] = \frac{(\beta - \alpha)^2}{12}$$



- Notice that the density  $\frac{1}{\beta - \alpha}$  is exactly the same regardless of the value of  $x$ . This makes it **uniform**.
- The PDF is  $\frac{1}{\beta - \alpha}$  since it is a constant such that the integral over all possible inputs evaluates to 1.



## Public transport example

### Example

The University bus arrives at the Computer Lab bus stop at 7:00, 7:15 and so on at 15 minute intervals. You arrive at the bus stop a time uniformly distributed in the interval between 1pm and 1:30pm. What is the probability that you wait less than 5 minutes for the bus?

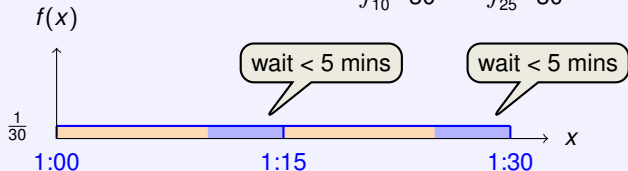
Answer

Let  $X$  be a RV for the time you arrive after 1pm to the bus stop.

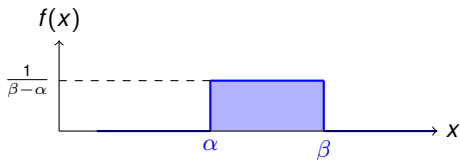
Define RVs:  $X \sim Uni(0, 30)$

Solve:

$$\mathbf{P}[10 < X < 15] + \mathbf{P}[25 < X < 30] = \int_{10}^{15} \frac{1}{30} dx + \int_{25}^{30} \frac{1}{30} dx = \frac{5}{30} + \frac{5}{30} = \frac{1}{3}$$



## Expectation for Uniform RV – convince yourself



$$\begin{aligned} \mathbf{E}[X] &= \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_{\alpha}^{\beta} x \cdot \frac{1}{\beta - \alpha} dx \\ &= \frac{1}{\beta - \alpha} \frac{1}{2} x^2 \Big|_{\alpha}^{\beta} = \frac{1}{\beta - \alpha} \frac{1}{2} (\beta^2 - \alpha^2) \\ &= \frac{1}{2} \frac{(\beta + \alpha)(\beta - \alpha)}{\beta - \alpha} = \frac{\alpha + \beta}{2} \end{aligned}$$



# Outline

---

Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

**Exponential random variable**

Normal (Gaussian) random variable



## Exponential continuous RV

### Exponential continuous random variable

An exponential random variable  $X$  represents the time until an event (first success) occurs. It is parametrised by  $\lambda > 0$ , the constant rate at which the event occurs.

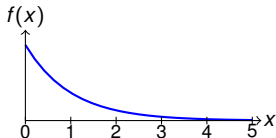
$$X \sim \text{Exp}(\lambda)$$

Range:  $[0, \infty)$

$$\text{PDF: } f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{when } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Expectation: } \mathbf{E}[X] = \frac{1}{\lambda} \quad (\text{time})$$

$$\text{Variance: } \mathbf{V}[X] = \frac{1}{\lambda^2}$$



- Examples: time until next earthquake, time for request to reach web server, time until end of mobile phone contract.
- Note that  $\lambda$  is the same as the one in the Poisson RV.
- Poisson RV counts # of events that occur in a fixed interval, exponential RV measures the amount of time until the next event occurs.



## Pandemic example

### Example

Major pandemics occur once every 100 years. What is the probability of a major pandemic in the next 5 years? What is the standard deviation of years until the next pandemic?

Answer

Let  $X$  be a RV for the time when the next pandemic happens.  
Let a unit of time be 1 year.

**Define RVs:**  $X \sim \text{Exp}(\lambda)$ ,  $\mathbf{E}[X] = \frac{1}{\lambda} = 100$ , thus  $\lambda = \frac{1}{100} = 0.01$   
 $X \sim \text{Exp}(\lambda = 0.01)$ .

**Solve:** Compute  $\mathbf{P}[X < 5]$ ,  $\mathbf{SD}[X]$ .

$$\begin{aligned}\mathbf{P}[X < 5] &= \int_0^5 0.01 e^{-0.01x} dx && \text{(remember that } \int e^{cx} dx = \frac{1}{c} e^{cx}\text{)} \\ &= 0.01 \frac{1}{-0.01} e^{-0.01x} \Big|_0^5 \\ &= -(e^{-0.05} - e^0) \approx 0.049\end{aligned}$$

$$\mathbf{SD}[X] = \sqrt{\mathbf{V}[X]} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda} = 100 \text{ years}$$



## CDF of Exponential RV – convince yourself

### CDF for Exponential RV

If  $X$  is an exponential continuous random variable,  $X \sim \text{Exp}(\lambda)$ , then its cumulative distribution function CDF (where  $x \geq 0$ ) is

$$F(x) = 1 - e^{-\lambda x}$$

Proof:

$$\begin{aligned} F(x) &= \mathbf{P}[X \leq x] = \int_0^x \lambda e^{-\lambda x} dx \\ &= \lambda \frac{1}{-\lambda} e^{-\lambda x} \Big|_0^x \\ &= -1(e^{-\lambda x} - e^{-\lambda 0}) \\ &= 1 - e^{-\lambda x} \end{aligned}$$



# Outline

---

Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

Exponential random variable

**Normal (Gaussian) random variable**



## Normal continuous RV

### Normal continuous random variable

A normal random variable  $X$ , parametrised over mean  $\mu$  and variance  $\sigma^2$  is defined as

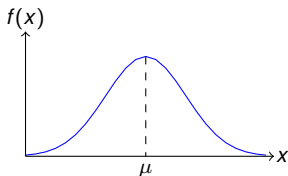
$$\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$$

Range:  $(-\infty, \infty)$

$$\text{PDF: } f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Expectation:  $\mathbf{E}[X] = \mu$

Variance:  $\mathbf{V}[X] = \sigma^2$

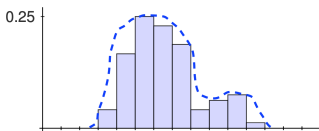


- The most important random variable type, AKA **Gaussian** RV and **Bell curve**.
- Generated from summing independent RV, thus occurs often in nature (cf. Central Limit Theorem in Lecture 8).
- Used to model entropic (conservative) distribution of data with mean and variance.

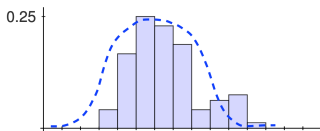


## Normal RV paradigm

Goal: translate problem statement into a RV – **model real life situation** with probability distributions (e.g., height distribution in a class).



Perfect fit!  
But what about another class?  
Overfit?



Same mean and variance!  
Generalises well.

Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ . PDF of  $X$ :

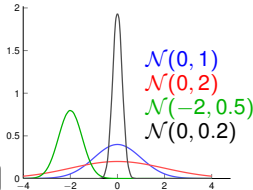
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

exponential tail

symmetric around  $\mu$

normalising constant

variance  $\sigma^2$  manages spread



## Walking example

### Example

You spent  $X$  minutes walking to the department every day. The average time you spend is  $\mu = 10$  minutes. The variance from day to day of the time spent to get to the department is  $\sigma^2 = 2$  minutes<sup>2</sup>. Suppose  $X$  is normally distributed. What is the probability you spend  $\geq 12$  minutes travelling to the department?

Answer

$$X \sim \mathcal{N}(\mu = 10, \sigma^2 = 2)$$

$$\mathbf{P}[X \geq 12] = \int_{12}^{\infty} f(x) dx = \int_{12}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

**Cannot be solved analytically!**

That is, no closed form for the integral of the Normal PDF. (But...)



## Properties for Normal RV

Let  $X \sim \mathcal{N}(\mu, \sigma^2)$  with CDF  $\mathbf{P}[X \leq x] = F(x)$ .

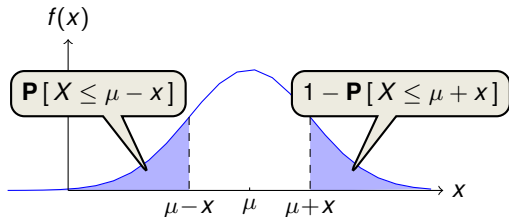
- Linear transformations of Normal RVs are also Normal RVs.

$$\text{If } Y = aX + b, \text{ then } Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

Proof outline:

- $\mathbf{E}[Y] = \mathbf{E}[aX + b] = a\mathbf{E}[X] + b = a\mu + b$  (linearity of expectation)
  - $\mathbf{V}[Y] = \mathbf{V}[aX + b] = a^2\mathbf{V}[X] = a^2\sigma^2$
  - $Y$  is also Normal.
- The PDF of a Normal RV is symmetric about the mean  $\mu$ .

$$F(\mu - x) = 1 - F(\mu + x)$$



## Computing probabilities with Normal RV

---

Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ . How do we compute CDF,  $\mathbf{P}[X \leq x] = F(x)$ ?

- We cannot analytically solve the integral (it has no closed form).
- But we can solve numerically using a function  $\Phi$ , which is a **precomputed** function:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

CDF of the Standard Normal,  $Z$



## Z: Standard Normal RV

### Standard Normal random variable Z

The Standard Normal continuous random variable Z is defined as

$$Z \sim \mathcal{N}(0, 1)$$

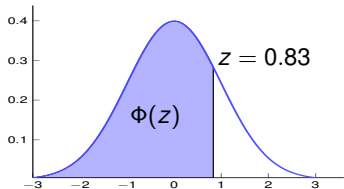
Expectation:  $\mathbf{E}[Z] = \mu = 0$  (zero mean)

Variance:  $\mathbf{V}[Z] = \sigma^2 = 1$  (unit variance)

- Not a new distribution: a special case of the Normal ( $\mathcal{N}(\mu, \sigma^2) = \mu + \sigma\mathcal{N}(0, 1)$ ).
- CDF of Z defined as  $\mathbf{P}[Z \leq z] = \Phi(z)$ .

**Table A.3** Standard Normal Curve Areas (cont.)

$\Phi(z) = P(Z \leq z)$



$$\mathbf{P}[Z \leq 0.83] = \Phi(0.83) = 0.7967$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633



### Example

You spent  $X$  minutes walking to the department every day. The average time you spend is  $\mu = 10$  minutes. The variance from day to day of the time spent to get to the department is  $\sigma^2 = 2$  minutes<sup>2</sup>. Suppose  $X$  is normally distributed. What is the probability you spend  $\geq 12$  minutes travelling to the department?

Answer

$$X \sim \mathcal{N}(\mu = 10, \sigma^2 = 2)$$

(But  $\mathbf{P}[X \geq 12] = \int_{12}^{\infty} f(x)dx$  has no analytic solution.)

1. Compute  $z = \frac{(x-\mu)}{\sigma}$ :

$$\begin{aligned}\mathbf{P}[X \geq 12] &= 1 - F_x(12) \\ &= 1 - \Phi\left(\frac{12 - 10}{\sqrt{2}}\right) \\ &\approx 1 - \Phi(1.41)\end{aligned}$$

2. Look up  $\Phi(z)$  in table:

$$\begin{aligned}1 - \Phi(1.41) &\approx 1 - 0.9207 \\ &= 0.0793\end{aligned}$$



# Introduction to Probability

Lecture 6: Marginals and Joint Distributions

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology

email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk



Experiments often involve **several** random variables, and some of them may **influence** each other.

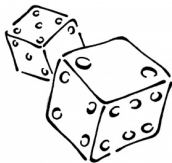
To this end, we will introduce:

- Joint/Marginal distribution of two (or more) variables
- Independence of two (or more) variables
- Covariance of two variables

For simplicity, we will mainly focus on **discrete** random variables.



## Warm-Up Exercise



### Example

Let  $X_1, X_2 \in \{1, 2, \dots, 6\}$  be two independent rolls of an unbiased die. Let  $S := X_1 + X_2$  and  $M := \max\{X_1, X_2\}$ . List the elements of the event  $\{S = 7, M \leq 5\}$  and deduce the probability.

Answer

The elements are  $\{(2, 5), (3, 4), (4, 3), (5, 2)\}$ . Since each of these elements has a probability of  $1/6 \cdot 1/6 = 1/36$ , the sought probability is  $4/36 = 1/9$ .

## Joint Probability

### Joint Probability Mass Function

The **joint probability mass function** of two **discrete** random variables  $X$  and  $Y$  is the function  $p : \mathbb{R}^2 \rightarrow [0, 1]$ , defined by:

$$p_{X,Y}(a, b) = \mathbf{P}[X = a, Y = b] \quad \text{for } -\infty < a, b < \infty.$$

### Joint Distribution Function

The **joint distribution function** of two (**discrete or continuous**) random variables  $X$  and  $Y$  is the function  $F : \mathbb{R}^2 \rightarrow [0, 1]$ , defined by:

$$F_{X,Y}(a, b) = \mathbf{P}[X \leq a, Y \leq b] \quad \text{for } -\infty < a, b < \infty.$$

### Marginal Distribution

Given a joint distribution  $F_{X,Y}$  of two random variables  $X, Y$ , one obtains the **marginal distribution** of  $X$  for any  $a$  as follows:

$$F_X(a) = \mathbf{P}[X \leq a] = \lim_{b \rightarrow \infty} F_{X,Y}(a, b).$$

Joint Distribution contains (much) more information than the two marginals!



## Discrete Example 1

### Example

Let  $X_1, X_2 \in \{1, 2, \dots, 6\}$  be independent rolls of an unbiased die. Let  $S := X_1 + X_2$  and  $M := \max\{X_1, X_2\}$ . Compute the **joint probability mass function**  $p$  of  $S$  and  $M$  and the **marginal distributions** of  $S$  and  $M$ .

Answer

$a$	$b$						$p_S(a)$
	1	2	3	4	5	6	
2	1/36	0	0	0	0	0	1/36
3	0	2/36	0	0	0	0	2/36
4	0	1/36	2/36	0	0	0	3/36
5	0	0	2/36	2/36	0	0	4/36
6	0	0	1/36	2/36	2/36	0	5/36
7	0	0	0	2/36	2/36	2/36	6/36
8	0	0	0	1/36	2/36	2/36	5/36
9	0	0	0	0	2/36	2/36	4/36
10	0	0	0	0	1/36	2/36	3/36
11	0	0	0	0	0	2/36	2/36
12	0	0	0	0	0	1/36	1/36
$p_M(b)$	1/36	3/36	5/36	7/36	9/36	11/36	1

$(X_1, X_2) = (4, 3)$  or  
 $(X_1, X_2) = (3, 4)$

$$p_{S,M}(7, 4) = \frac{1}{36} + \frac{1}{36}$$

$\{S = 8\} =$   
 $\{S = 8, M = 4\}$   
 $\cup \{S = 8, M = 5\}$   
 $\cup \{S = 8, M = 6\}$

$$p_S(8) = \frac{1}{36} + \frac{2}{36} + \frac{2}{36}$$



## Discrete Example 2

### Example

Suppose an urn contains balls numbered  $1, 2, \dots, N$ . We draw  $1 \leq n \leq N$  balls uniformly and **without replacement** from the urn. Let  $X_i \in \{1, 2, \dots, N\}$  be the number of the ball drawn in the  $i$ -th step. What is the marginal distribution of  $X_i$ ?

Answer

We first compute the **joint distribution**. For distinct  $a_1, a_2, \dots, a_n$ ,

$$\begin{aligned} p(a_1, a_2, \dots, a_n) &= \mathbf{P}[X_1 = a_1, X_2 = a_2, \dots, X_n = a_n] \\ &= \frac{1}{N(N-1)\cdots(N-n+1)}. \end{aligned}$$

Fix  $i$  and consider the **marginal distribution** of  $X_i$ :

$$\begin{aligned} p_{X_i}(k) &= \sum_{a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n} p(a_1, \dots, a_{i-1}, k, a_{i+1}, \dots, a_n) \\ &= (N-1)(N-2)\cdots(N-n+1) \cdot \frac{1}{N(N-1)\cdots(N-n+1)} \end{aligned}$$

The  $X_i$ 's are **not** independent, yet their marginals are **identical!**

Same argument applies to the **hypergeometric distribution**, with balls of two different colours.



## Joint Distributions of Continuous Variables

Definition

Random variables  $X$  and  $Y$  have a **joint continuous distribution** if for some function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  and for all numbers  $a_1 \leq b_1$  and  $a_2 \leq b_2$ ,

$$\mathbf{P}[a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y) dx dy.$$

The function  $f$  has to satisfy  $f(x, y) \geq 0$  for all  $x$  and  $y$ , and  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ . We call  $f$  the **joint probability density**.

As in one-dimensional case we switch from  $F$  to  $f$  by **differentiating** (or **integrating**):

$$F(a, b) = \int_{-\infty}^a \int_{-\infty}^b f(x, y) dx dy \quad \text{and} \quad f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$



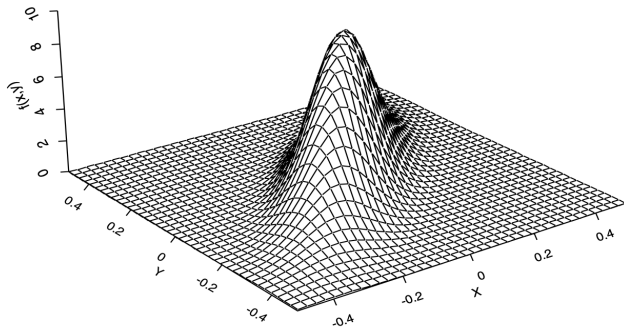
## Example of a Joint Distribution of Continuous Random Variables

- Consider the density:

$$f(x, y) = \frac{30}{\pi} \cdot e^{-50x^2 - 50y^2 + 80xy},$$

where  $-\infty < x, y < \infty$ .

- This is an example of a so-called **bivariate normal probability density function**.



Source: Modern Introduction to Statistics



## Dealing with Continuous Variables

### Example (1/2)

Suppose that the joint probability density of  $X$  and  $Y$  is given by

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y} & \text{for } 0 < x < \infty, 0 < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Compute (i)  $\mathbf{P}[X > 1, Y < 1]$  and (ii)  $\mathbf{P}[X < Y]$ .

Answer \_\_\_\_\_

(i) We first compute:

$$\begin{aligned} \mathbf{P}[X > 1, Y < 1] &= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y} dx dy \\ &= \int_0^1 2e^{-2y}(-e^{-x}) \Big|_1^\infty dy = \int_0^1 2e^{-2y}(0 - (-e^{-1})) dy \\ &= e^{-1} \int_0^1 2e^{-2y} dy = e^{-1} \cdot (1 - e^{-2}). \end{aligned}$$



## Dealing with Continuous Variables (cont.)

### Example (2/2)

Suppose that the joint probability density of  $X$  and  $Y$  is given by

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y} & \text{for } 0 < x < \infty, 0 < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Compute (i)  $\mathbf{P}[X > 1, Y < 1]$  and (ii)  $\mathbf{P}[X < Y]$ .

Answer

(ii) We have:

$$\begin{aligned} \mathbf{P}[X < Y] &= \int_0^{\infty} \int_0^y 2e^{-x}e^{-2y} dx dy \\ &= \int_0^{\infty} 2e^{-2y}(1 - e^{-y}) dy \\ &= \int_0^{\infty} 2e^{-2y} dy - \int_0^{\infty} 2e^{-3y} dy \\ &= 1 - \frac{2}{3} = \frac{1}{3}. \end{aligned}$$



# Introduction to Probability

Lecture 7: Independence, Covariance and Correlation

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology

email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk



## Independence of Random Variables

This definition covers the **discrete** and **continuous** case!

Definition of Independence

Two random variables  $X$  and  $Y$  are **independent** if for all values  $a, b$ :

$$\mathbf{P}[X \leq a, Y \leq b] = \mathbf{P}[X \leq a] \cdot \mathbf{P}[Y \leq b].$$

For two **discrete** random variables, an equivalent definition is:

$$\mathbf{P}[X = a, Y = b] = \mathbf{P}[X = a] \cdot \mathbf{P}[Y = b].$$

This is useless for continuous random variables.

Remark

Using the **joint probability distribution**, the above is equivalent to for all  $a, b$ ,

$$F(a, b) = F_X(a) \cdot F_Y(b).$$

All these definitions extend in the natural way to **more than two** variables!



### Factorisation

The definition of independence of  $X$  and  $Y$  implies the following **factorisation** formula: for any “suitable” sets  $A$  and  $B$ ,

$$\mathbf{P}[X \in A, Y \in B] = \mathbf{P}[X \in A] \cdot \mathbf{P}[Y \in B]$$

For **continuous** distributions one obtains by differentiating both sides in the formula for the joint distribution:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

### Example

Let  $X$  and  $Y$  be two independent variables. Let  $I = (a, b]$  be any interval and define  $U := \mathbf{1}_{X \in I}$  and  $V := \mathbf{1}_{Y \in I}$ . Prove  $U$  and  $V$  are independent.

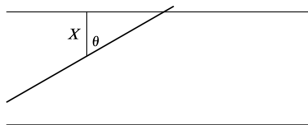
$$\begin{aligned} \mathbf{P}[U = 0, V = 1] &= \mathbf{P}[X \in I^c, Y \in I] \\ &= \mathbf{P}[X \notin I] \mathbf{P}[Y \in I] = \mathbf{P}[U = 0] \mathbf{P}[V = 1]. \end{aligned}$$

Verification for other combination of values is similar ( $U = 1, V = 0$  and  $U = 0, V = 0$ , etc).

## Buffon's Needle Problem (1/2)



Georges-Louis Leclerc de Buffon 1707–1788 (Source Wikipedia)



Source: Ross, Probability 8th ed.

- A table is ruled with equidistant, parallel lines a distance  $D$  apart.
- A needle of length  $L$  is thrown randomly on the table.
- **What is the probability that the needle will intersect one of the two lines?**

Let  $X$  be the distance of the **middle point** of the needle to the closest parallel line. Needle intersects a line if hypotenuse of the triangle is less than  $L/2$ , i.e.,

$$\frac{X}{\cos(\theta)} < \frac{L}{2} \quad \Leftrightarrow \quad X < \frac{L}{2} \cos(\theta).$$

We assume that  $X \in [0, D/2]$  and  $\theta \in [0, \pi/2]$  are **independent** and **uniform**.

Can be thought of as: 1. Sample the middle point of needle, 2. Sample the angle.



## Buffon's Needle Problem (2/2)

Let us compute the probability that the line intersects:

$$\begin{aligned} \mathbf{P} \left[ X < \frac{L}{2} \cdot \cos(\theta) \right] &= \iint_{x < (L/2) \cos y} f_{X,\theta}(x, y) dx dy \\ &= \iint_{x < (L/2) \cos y} f_X(x) f_\theta(y) dx dy \\ &= \frac{4}{\pi D} \int_0^{\pi/2} \int_0^{L/2 \cos(y)} dx dy \\ &= \frac{4}{\pi D} \int_0^{\pi/2} \frac{L}{2} \cos(y) dy \\ &= \frac{2L}{\pi D}. \end{aligned}$$

This gives us a method to estimate  $\pi$ !



## Covariance

### Definition of Covariance

Let  $X$  and  $Y$  be two random variables. The **covariance** is defined as:

$$\mathbf{Cov}[X, Y] = \mathbf{E}[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])].$$

Interpretation:

- If  $\mathbf{Cov}[X, Y] > 0$  and  $X$  has a realisation larger (smaller) than  $\mathbf{E}[X]$ , then  $Y$  will likely have a realisation larger (smaller) than  $\mathbf{E}[Y]$ .
- If  $\mathbf{Cov}[X, Y] < 0$ , then it is the other way around.

### Alternative Formula

Using the linearity of expectation rule, one has the equivalent definition:

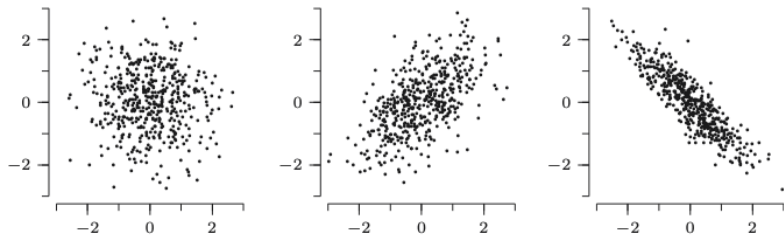
$$\mathbf{Cov}[X, Y] = \mathbf{E}[X \cdot Y] - \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

- Note that  $\mathbf{Cov}[X, X] = \mathbf{V}[X]$ .
- Two variables  $X, Y$  with  $\mathbf{Cov}[X, Y] > 0$  are **positively correlated**.
- Two variables  $X, Y$  with  $\mathbf{Cov}[X, Y] < 0$  are **negatively correlated**.
- Two variables  $X, Y$  with  $\mathbf{Cov}[X, Y] = 0$  are **uncorrelated**.



## Illustration of 3 Cases for Cov $[X, Y]$

500 outcomes of randomly generated pairs of RVs  $(X, Y)$  with different joint distributions



**Fig. 10.1.** Some scatterplots.  
Source: Textbook by Dekking

1. What is the covariance (positive, negative, neutral)?
  - Left: set of sampled points has a circular shape, so uncorrelated.
  - Middle: looks like ellipsoids with  $y = x$  as main axis, so positively correlated.
  - Right: looks like ellipsoids with  $y = -x$  as main axis, so negatively correlated.
2. Where is the covariance the largest (in magnitude)?
  - Right: points more closely concentrated, hence correlation is largest.

## Independence implies Uncorrelated

### Example

Let  $X$  and  $Y$  be two **independent** random variables. Then  $X$  and  $Y$  are **uncorrelated**, i.e.,  $\mathbf{Cov}[X, Y] = 0$ .

Answer

We give a proof for the discrete case:

$$\begin{aligned}\mathbf{E}[X \cdot Y] &= \sum_i \sum_j a_i \cdot b_j \cdot \mathbf{P}[X = a_i, Y = b_j] \\ &= \sum_i \sum_j a_i \cdot b_j \cdot \mathbf{P}[X = a_i] \cdot \mathbf{P}[Y = b_j] \\ &= \left( \sum_i a_i \cdot \mathbf{P}[X = a_i] \right) \cdot \left( \sum_j b_j \cdot \mathbf{P}[Y = b_j] \right) \\ &= \mathbf{E}[X] \cdot \mathbf{E}[Y].\end{aligned}$$



## Uncorrelated may not imply Independence

### Example

Find a (simple) example of two random variables  $X$  and  $Y$  which are **uncorrelated but dependent**.

Answer

- Let  $X$  be uniformly sampled from  $\{-1, 0, +1\}$  and  $Y := \mathbf{1}_{X=0}$ .  
 $\Rightarrow X \cdot Y = 0$  (for all outcomes), and thus

$$\mathbf{E}[X \cdot Y] = 0.$$

- Further,  $\mathbf{E}[X] = 0$  (and  $\mathbf{E}[Y] = 1/3$ ), and hence:

$$\mathbf{Cov}[X, Y] = \mathbf{E}[X \cdot Y] - \mathbf{E}[X] \cdot \mathbf{E}[Y] = 0.$$

- On the other hand,  $\mathbf{P}[X = 0] = 1/3$  and  $\mathbf{P}[Y = 0] = 2/3$ , and thus

$$1 = \mathbf{P}[X \cdot Y = 0] > \mathbf{P}[X = 0] \cdot \mathbf{P}[Y = 0] = 2/9.$$



### Variance of Sum Formula

- For any two random variables  $X, Y$ ,

$$\mathbf{V}[X + Y] = \mathbf{V}[X] + \mathbf{V}[Y] + 2 \cdot \mathbf{Cov}[X, Y].$$

- Hence if  $X$  and  $Y$  are **uncorrelated** variables,

$$\mathbf{V}[X + Y] = \mathbf{V}[X] + \mathbf{V}[Y].$$

Generalisation of the case where  $X$  and  $Y$  are even **independent**!

- For any random variables  $X_1, X_2, \dots, X_n$ :

$$\mathbf{V}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{V}[X_i] + 2 \cdot \sum_{i=1}^n \sum_{j=i+1}^n \mathbf{Cov}[X_i, X_j].$$

## Computing Variances of Sums of Uncorrelated Variables

### Example

Recall the example where  $X \in \{-1, 0, +1\}$  uniformly and  $Y := \mathbf{1}_{X=0}$ . Compute  $\mathbf{V}[X + Y]$ .

Answer

- We first compute  $\mathbf{V}[X]$ :

$$\mathbf{V}[X] = \frac{1}{3} \cdot (-1)^2 + \frac{1}{3} \cdot 0^2 + \frac{1}{3} \cdot 1^2 = \frac{2}{3}.$$

- Now for  $\mathbf{V}[Y]$ :

$$\begin{aligned}\mathbf{V}[Y] &= \frac{1}{3} \cdot \left(1 - \frac{1}{3}\right)^2 + \frac{2}{3} \left(0 - \frac{1}{3}\right)^2 \\ &= \frac{2}{9}.\end{aligned}$$

⇒ Hence:

$$\begin{aligned}\mathbf{V}[X + Y] &= \mathbf{V}[X] + \mathbf{V}[Y] + 2 \cdot \mathbf{Cov}[X, Y] \\ &= \frac{2}{3} + \frac{2}{9} + 0 = \frac{8}{9}.\end{aligned}$$



## Correlation Coefficient: Normalising the Covariance

The definition of covariance is **not scaling invariant**:

- If  $X$  increases by a factor of  $\alpha$ , then  $\mathbf{Cov}[X, Y]$  increases by a factor of  $\alpha$ .
- ⇒ Even if  $X$  and  $Y$  both increase by  $\alpha$ , then  $\mathbf{Cov}[X, Y]$  will change.  
(Exercise: It changes by?)

### Correlation Coefficient

Let  $X$  and  $Y$  be two random variables. The **correlation coefficient**  $\rho(X, Y)$  is defined as:

$$\rho(X, Y) = \frac{\mathbf{Cov}[X, Y]}{\sqrt{\mathbf{V}[X] \cdot \mathbf{V}[Y]}}$$

If  $\mathbf{V}[X] = 0$  or  $\mathbf{V}[Y] = 0$ , then it is defined as 0.

### Properties:

1. The correlation coefficient is **scaling-invariant**, i.e.,  
 $\rho(X, Y) = \rho(\alpha \cdot X, \beta \cdot Y)$  for any  $\alpha, \beta > 0$ .
2. For any two random variables  $X, Y$ ,  $\rho(X, Y) \in [-1, 1]$ .



## Range of the Correlation Coefficient

### Example

Verify that the correlation coefficients' range satisfies  $\rho(X, Y) \in [-1, 1]$ .

Answer

- We will only prove  $\rho(X, Y) \geq -1$  (the other direction follows in analogous way).
- Let  $\sigma_x^2$  and  $\sigma_y^2$  denote the variances of  $X$  and  $Y$ , and  $\sigma_x$  and  $\sigma_y$  their standard deviations.
- Then:

$$\begin{aligned} 0 &\leq \mathbf{V} \left[ \frac{X}{\sigma_x} + \frac{Y}{\sigma_y} \right] \\ &= \mathbf{V} \left[ \frac{X}{\sigma_x} \right] + \mathbf{V} \left[ \frac{Y}{\sigma_y} \right] + 2 \mathbf{Cov} \left[ \frac{X}{\sigma_x}, \frac{Y}{\sigma_y} \right] \\ &= \frac{\mathbf{V}[X]}{\mathbf{V}[X]} + \frac{\mathbf{V}[Y]}{\mathbf{V}[Y]} + 2 \cdot \frac{\mathbf{Cov}[X, Y]}{\sigma_x \cdot \sigma_y} \\ &= 2 \cdot (1 + \rho(X, Y)). \end{aligned}$$



# Introduction to Probability

Lecture 8: Basic Inequalities and Law of Large Numbers

Mateja Jamnik, [Thomas Sauerwald](#)

University of Cambridge, Department of Computer Science and Technology  
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

Easter 2026



# Outline

---

Introduction

Markov's Inequality and Chebyshev's Inequality

Weak Law of Large Numbers

# Board Games Involving Dice

- Games with One Die: 



- Games with Two Dice:  

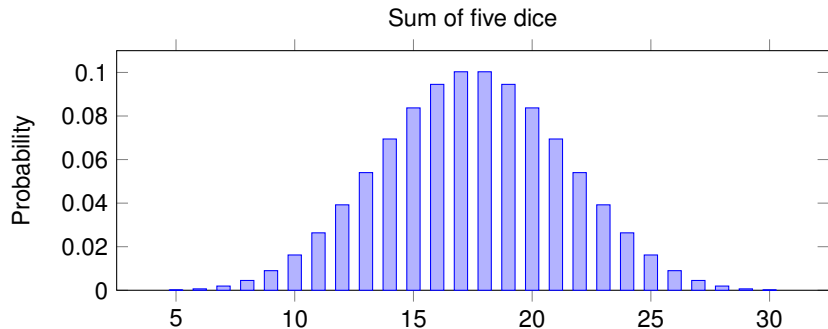
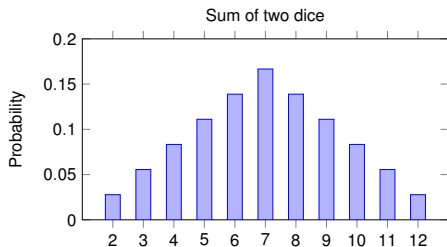
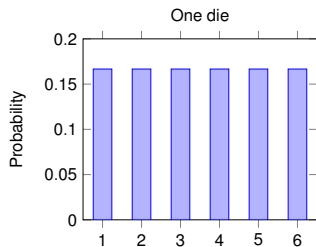


- Games with Five Dice: 



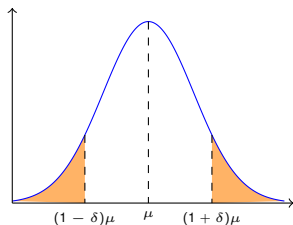
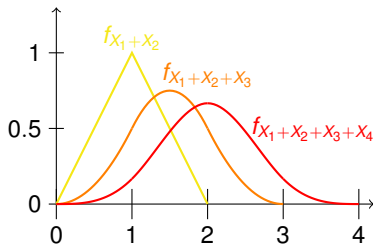
Source: All images from Wikipedia.

# Joint Distributions of Sums



## Motivation

We will study sums of independent and identically distributed variables. How does their distribution look like, and how well do they concentrate around the expectation?



1. Markov's inequality
2. Chebyshev's inequality
3. Law of Large Numbers
4. **Central Limit Theorem**

Re-use concepts from previous lectures:

1. Independence (Random Var.) (Lec. 1, 7)
2. Expectation and Variance (Lec. 2, 3)
3. Normal Distribution (Lec. 5)
4. Sums of Random Variables (Lec. 6)

# Outline

---

Introduction

Markov's Inequality and Chebyshev's Inequality

Weak Law of Large Numbers

## Markov's Inequality

### Markov's Inequality

For any **non-negative** random variable  $X$  with finite  $\mathbf{E}[X]$ , it holds for any  $a > 0$ ,

$$\mathbf{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

Markov's inequality is a so-called **tail-bound**: it upper bounds the probability that the random variable **exceeds** its mean



A. Markov (1856-1922)

### Comments:

- Markov's inequality can be rewritten as: for any  $\delta > 0$ ,

$$\mathbf{P}[X \geq \delta \cdot \mathbf{E}[X]] \leq 1/\delta.$$

- **Advantage**: Very basic inequality, we only need to know  $\mathbf{E}[X]$
- **Downside**: For many distributions, the tail bound might be quite loose
- Proof is similar to the proof of Chebyshev's inequality (Exercise!)

## Applying Markov's Inequality

### Example 2

Consider throwing an unbiased, six-sided dice 120 times and let  $X$  denote the number of times we obtain a six.

1. Derive an upper bound on  $\mathbf{P}[X \geq 30]$ .
2. Can you also derive an upper bound on  $\mathbf{P}[X \leq 10]$ ?

Answer

1. First compute  $\mathbf{E}[X] = 1/6 \cdot 120 = 20$ . Then by Markov:

$$\mathbf{P}[X \geq 30] \leq \frac{20}{30} = \frac{2}{3}.$$

2. Consider now the second bound.

- Define a new random variable  $Y := 120 - X$ .
- ⇒ This random variable is also non-negative (as  $X \leq 120$ ).
- Applying Markov's inequality (equivalent version) to  $Y$  yields:

$$\begin{aligned}\mathbf{P}[X \leq 10] &= \mathbf{P}[Y \geq 110] = \mathbf{P}\left[Y \geq \frac{110}{100} \cdot \mathbf{E}[Y]\right] \\ &\leq \frac{100}{110} = \frac{10}{11}.\end{aligned}$$

## Chebyshev's Inequality

### Chebyshev's Inequality

For **any** random variable  $X$  with finite  $\mathbf{E}[X]$  and  $\mathbf{V}[X]$ , for any  $a > 0$ ,

$$\mathbf{P}[|X - \mathbf{E}[X]| \geq a] \leq \mathbf{V}[X]/a^2.$$



P. Chebyshev (1821-1894)

### Comments:

- can be rewritten as:

The " $\mu \pm$  a few  $\sigma$ " rule. Most of the probability mass is within a few standard deviations from  $\mu$ .

$$\mathbf{P}[|X - \mathbf{E}[X]| \geq \sqrt{\delta \cdot \mathbf{V}[X]}] \leq 1/\delta.$$

- Unlike Markov, Chebyshev's inequality is two-sided and also holds for random variables with **negative** values
- In most cases, Chebyshev's inequality yields much **stronger bounds** than Markov (however, it requires knowledge not only of  $\mathbf{E}[X]$  but also  $\mathbf{V}[X]$ !)
- Chebyshev's inequality is also known as **Second Moment Method**

## Derivation of Chebychev's inequality

### Proof

- We will give a **self-contained** proof for a continuous random variable  $X$  (the case for discrete  $X$  is analogous).
- Write down the definition of  $\mathbf{V}[X]$  and then lower bound:

$$\begin{aligned}\mathbf{V}[X] &= \mathbf{E}[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f_X(x) dx \\ &\geq \int_{|x - \mu| \geq a} (x - \mu)^2 \cdot f_X(x) dx \\ &\geq \int_{|x - \mu| \geq a} a^2 \cdot f_X(x) dx \\ &= a^2 \cdot \int_{|x - \mu| \geq a} f_X(x) dx \\ &= a^2 \cdot \mathbf{P}[|X - \mu| \geq a].\end{aligned}$$

- Dividing both sides by  $a^2$  yields the result.

**Exercise:** Can you find a proof that uses Markov's inequality?

## Example: Chebychev is (usually) much stronger than Markov

### Example 3

Throw an unbiased coin  $n$  times and let  $X$  be the total number of heads. In an experiment, with  $n$  large, we would usually expect a number of heads that is close to the expectation. Can we justify that?

Answer

$X \sim \text{Bin}(n, 1/2)$  so  $\mathbf{E}[X] = n \cdot \frac{1}{2}$ .

- **Markov's inequality:** For any  $\delta > 0$ ,

$$\mathbf{P}[X \geq (1 + \delta) \cdot \mathbf{E}[X]] \leq \frac{1}{1 + \delta}$$

- **Chebychev's inequality:**

$\Rightarrow$  We have  $\mathbf{V}[X] = np(1 - p) = n \cdot 1/2 \cdot 1/2$ . For any  $\delta > 0$ ,

$$\begin{aligned} \mathbf{P}[X \geq (1 + \delta) \cdot \mathbf{E}[X]] &= \mathbf{P}[X - \mathbf{E}[X] \geq \delta \cdot \mathbf{E}[X]] \\ &\leq \mathbf{P}[|X - n/2| \geq \delta \cdot (n/2)] \\ &\leq \frac{n \cdot 1/4}{\delta^2 (n/2)^2} = \frac{1}{\delta^2 n} \end{aligned}$$

# Outline

---

Introduction

Markov's Inequality and Chebyshev's Inequality

**Weak Law of Large Numbers**

## Law of Large Numbers

= independent and identically distributed

### The Weak Law of Large Numbers

Let  $\bar{X}_n := 1/n \cdot \sum_{i=1}^n X_i$ , where the  $X_i$ 's are **i.i.d.** with finite expectation  $\mu$  and finite variance  $\sigma^2$ . Then, for any  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left[ |\bar{X}_n - \mu| > \epsilon \right] = 0$$

$$\forall \epsilon > 0: \forall \delta > 0: \exists N > 0: \forall n \geq N: \mathbf{P} \left[ |\bar{X}_n - \mu| > \epsilon \right] \leq \delta$$

- “Power of Averaging”: repeated samples allow us to estimate  $\mu$
- A similar statement holds even if the  $X_i$ 's are not identically distributed
- There is also a **strong law of large numbers**:

$$\mathbf{P} \left[ \lim_{n \rightarrow \infty} \bar{X}_n = \mu \right] = 1.$$

*“For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one’s goal.”*



J. Bernoulli (1655-1705)

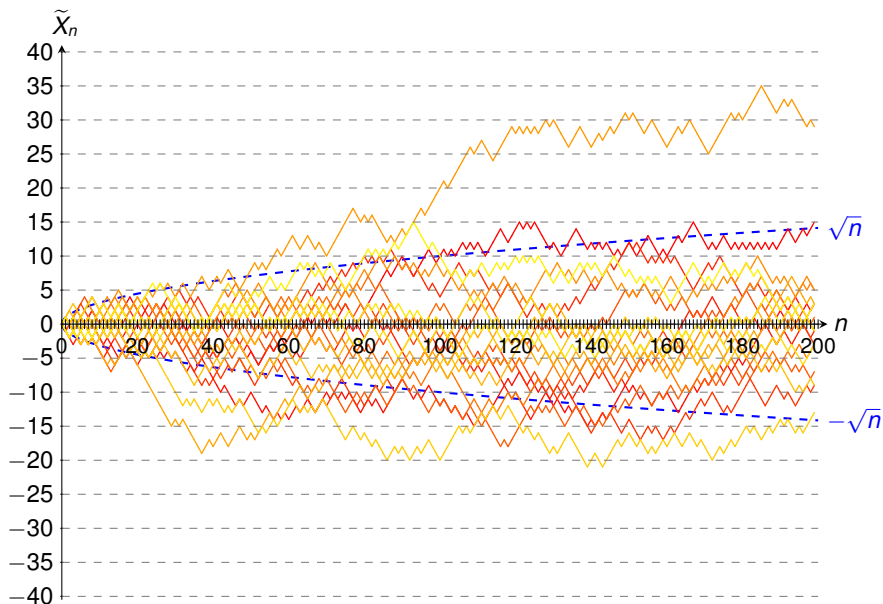
## Illustration of Weak Law of Large Numbers (1/4)

---

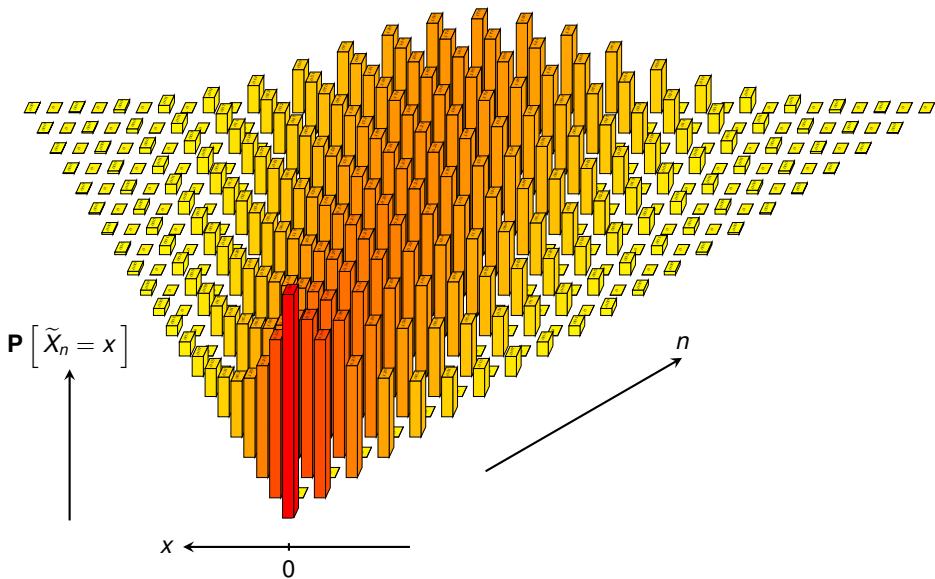
- Let  $X_i$  be independent random variables taking values  $\in \{-1, +1\}$  with probability  $1/2$  each
- Consider  $\tilde{X}_n := \sum_{i=1}^n X_i$  for any  $n = 0, 1, \dots, 200$

How does a “typical” realisation look like?

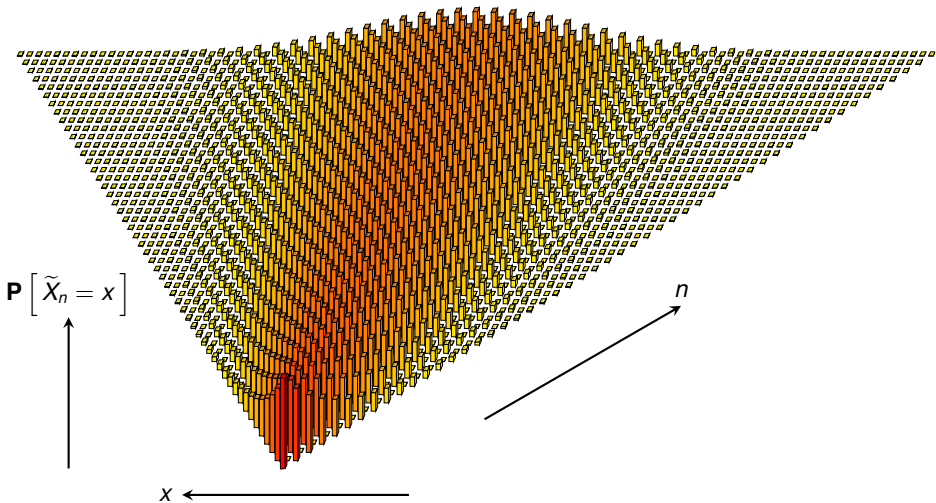
## Illustration of Weak Law of Large Numbers (2/4)



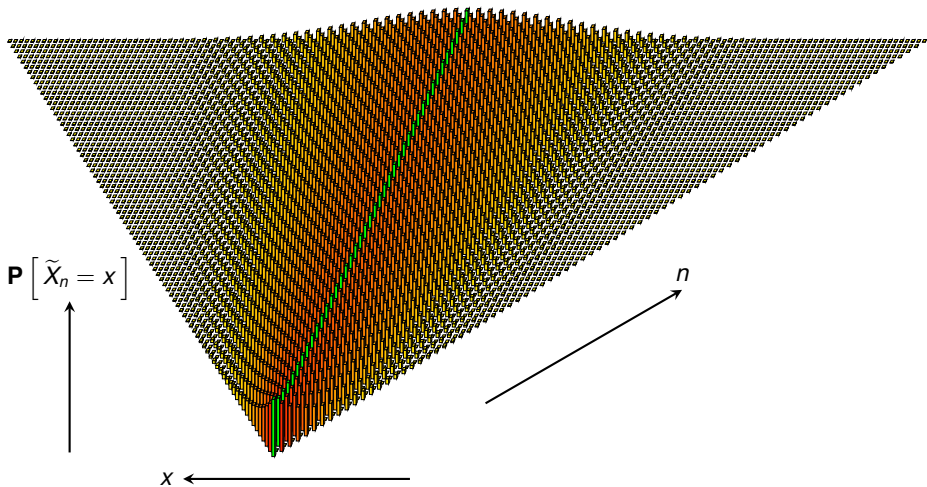
## Plot of the Distributions for $n = 0, 1, \dots, 20$



# Plot of the Distributions for $n = 0, 1, \dots, 50$



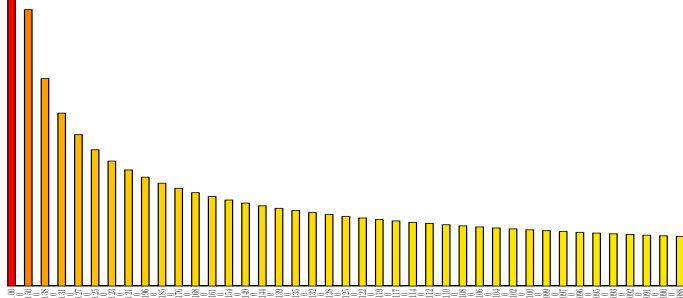
# Plot of the Distributions for $n = 0, 1, \dots, 80$



## Interlude: Approximation of $\mathbf{P}[\tilde{X}_n = 0]$

### Exercise

Try to find an expression for  $\mathbf{P}[\tilde{X}_n = 0]$ . Using Stirling's approximation for  $n!$ , conclude that  $\mathbf{P}[\tilde{X}_n = 0] = \Theta(1/\sqrt{n})$  for even integers  $n$ .



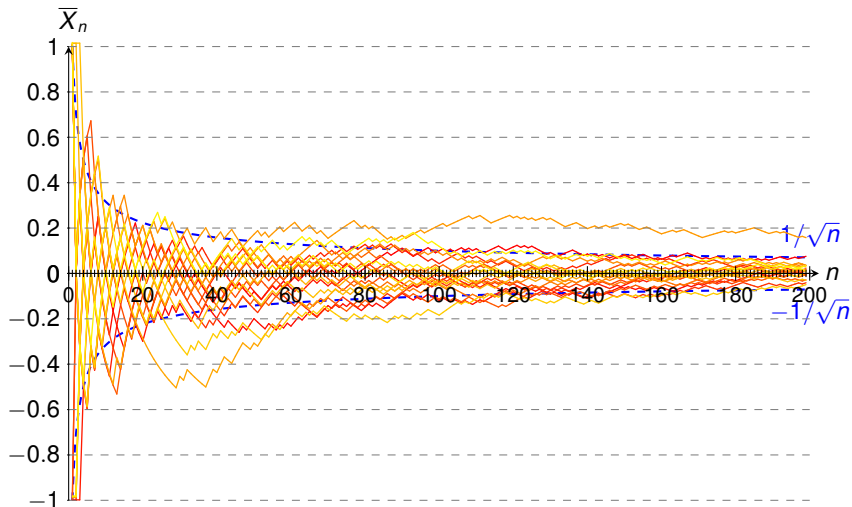
## Illustration of Weak Law of Large Numbers (3/4)

- Let  $X_i$  be independent random variables taking values  $\in \{-1, +1\}$  with probability  $1/2$  each
- Consider  $\tilde{X}_n := \sum_{i=1}^n X_i$  for any for any  $n = 0, 1, \dots, 200$

This does **not** converge!

Consider now the **average (sample mean)**:  $\bar{X}_n := 1/n \cdot \sum_{i=1}^n X_i$ .

## Illustration of Weak Law of Large Numbers (4/4)



## Proof of the Weak Law of Large Numbers

### The Weak Law of Large Numbers

Let  $\bar{X}_n := 1/n \cdot \sum_{i=1}^n X_i$ , where the  $X_i$ 's are **i.i.d.** with finite expectation  $\mu$  and finite variance  $\sigma^2$ . Then, for any  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left[ |\bar{X}_n - \mu| > \epsilon \right] = 0$$

#### Proof

- Let  $\bar{X}_n := 1/n \cdot \sum_{i=1}^n X_i$
- Then  $\mathbf{E} [\bar{X}_n] = \mu$  and  
 $\mathbf{V} [\bar{X}_n] = 1/n^2 \cdot \mathbf{V} [\sum_{i=1}^n X_i] = 1/n^2 \cdot \sum_{i=1}^n \mathbf{V} [X_i] = 1/n \cdot \sigma^2$ .
- Applying **Chebyshev's inequality** yields:

$$\mathbf{P} \left[ \left| \bar{X}_n - \mathbf{E} [\bar{X}_n] \right| > \epsilon \right] \leq \frac{1}{\epsilon^2} \cdot \mathbf{V} [\bar{X}_n] = \frac{\sigma^2}{n\epsilon^2}.$$

- For any (fixed)  $\epsilon > 0$ , the right hand side vanishes as  $n \rightarrow \infty$ .  
(Let  $\epsilon > 0, \delta > 0$ . Pick  $N = \frac{\sigma^2}{\epsilon^2 \cdot \delta}$ . Then for any  $n \geq N$ , the probability above is smaller than  $\delta$ .)

### Example 4

Suppose that, instead of the expectation  $\mu$ , we want to estimate the probability of an **event**, e.g.,

$$p := \mathbf{P}[X \in (a, b)], \text{ where } a < b.$$

How can we use the **Law of Large Numbers**?

Answer

- Let  $X_1, X_2, \dots, X_n \sim X$ . For each  $1 \leq i \leq n$ , define:

$$Y_i = \begin{cases} 1 & \text{if } X_i \in (a, b), \\ 0 & \text{otherwise.} \end{cases}$$

- We have:

$$\mathbf{E}[Y_i] = \mathbf{P}[X_i \in (a, b)] \cdot 1 + \mathbf{P}[X_i \notin (a, b)] \cdot 0 = p.$$

- Similarly,  $\mathbf{V}[Y_i] = p(1 - p)$
- The random variables  $Y_1, Y_2, \dots, Y_n$  are i.i.d., so we can apply the Law of Large Numbers to  $\bar{Y}_n$ .

# Introduction to Probability

Lectures 9: Central Limit Theorem

Mateja Jamnik, [Thomas Sauerwald](#)

University of Cambridge, Department of Computer Science and Technology  
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

Easter 2026



UNIVERSITY OF  
CAMBRIDGE

# Outline

---

Recap: Weak Law of Large Numbers

Central Limit Theorem

Illustrations

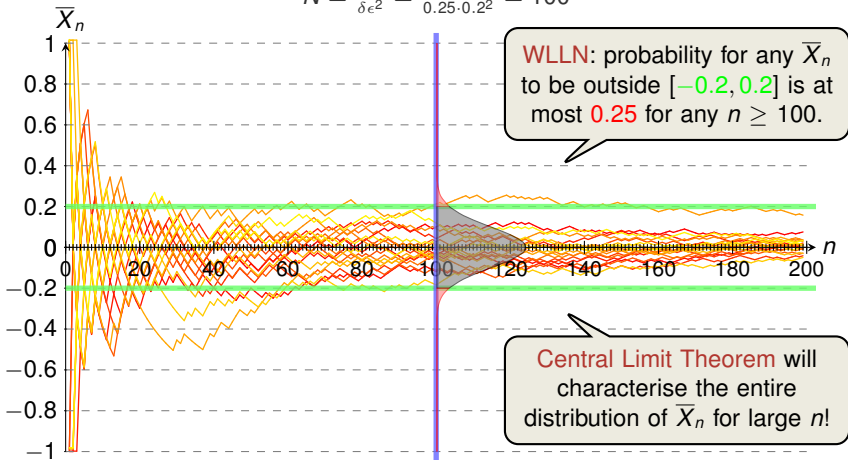
Examples

## Weak Law of Large Numbers (4/4)

Weak Law of Large Numbers: For any  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} \mathbf{P} \left[ |\bar{X}_n - \mu| > \epsilon \right] = 0$

$$\Rightarrow \epsilon = 0.2, \delta = 0.25, \exists N: \forall n \geq N: \mathbf{P} \left[ |\bar{X}_n - \mu| > 0.2 \right] \leq 0.25$$

$$N = \frac{\sigma^2}{\delta \epsilon^2} = \frac{1}{0.25 \cdot 0.2^2} = 100$$



# Outline

---

Recap: Weak Law of Large Numbers

Central Limit Theorem

Illustrations

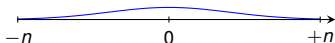
Examples

## Towards the CLT: Finding the Right Scaling

- Let  $X_1, X_2, \dots$  i.i.d. with  $\mu = 0$  and finite  $\sigma^2$

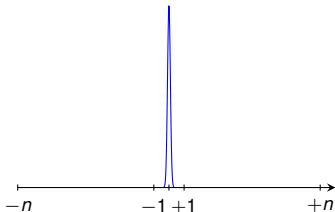
The Sum

- Let  $\tilde{X}_n := \sum_{i=1}^n X_i$  (often denoted by  $S_n$ )
- The variance is  $\mathbf{V}[\tilde{X}_n] = n\sigma^2 \rightarrow \infty$



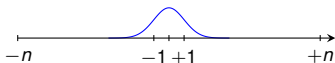
The Sample Average (Sample Mean)

- Let  $\bar{X}_n := \frac{1}{n} \cdot \sum_{i=1}^n X_i$
- The variance is  $\mathbf{V}[\bar{X}_n] = \sigma^2/n \rightarrow 0$



The Proper Scaling (Standardising, see Lec. 5)

- Let  $Z_n := \frac{1}{\sqrt{n} \cdot \sigma} \cdot \sum_{i=1}^n X_i$
- The variance is  $\mathbf{V}[Z_n] = 1$



# Central Limit Theorem



A. de Moivre (1667-1754) P.-S. de Laplace (1749-1827) C. Gauss (1777-1855) A. Lyapunov (1857-1918) C. Lindeberg (1876-1932)

## Central Limit Theorem

Let  $X_1, X_2, \dots$  be any sequence of independent identically distributed random variables with finite expectation  $\mu$  and finite variance  $\sigma^2$ . Let

$$Z_n := \sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n} \cdot \sigma} \cdot \left( \sum_{i=1}^n X_i - n \cdot \mu \right)$$

Then for any number  $a \in \mathbb{R}$ , it holds that

$$\lim_{n \rightarrow \infty} F_{Z_n}(a) = \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a e^{-x^2/2} dx,$$

where  $\Phi$  is the distribution function of the  $\mathcal{N}(0, 1)$  distribution.

In words: the distribution of  $Z_n$  **always** converges to the distribution function  $\Phi$  of the standard normal distribution.

## Comments on the CLT

---

- one of the most remarkable results in probability/statistics
- extremely useful tool in data analysis or physical measurements
  - we may not know the actual distribution in real-world, and CLT says we don't have to(!)
  - adding up independent noises in measurements leads to an error following the Normal distribution
  - applies also to sums of random variables which may be unbounded
- catch: the CLT only holds **approximately**, i.e., for large  $n$

When is the approximation good?

- usually  $n \geq 10$  or  $n \geq 15$  is sufficient in practice
- approximation tends to be worse when threshold  $a$  is far from 0, distribution of  $X_i$ 's asymmetric, bimodal or discrete
- (for a result quantifying the approximation error: Berry-Esseen-Theorem)

# Outline

---

Recap: Weak Law of Large Numbers

Central Limit Theorem

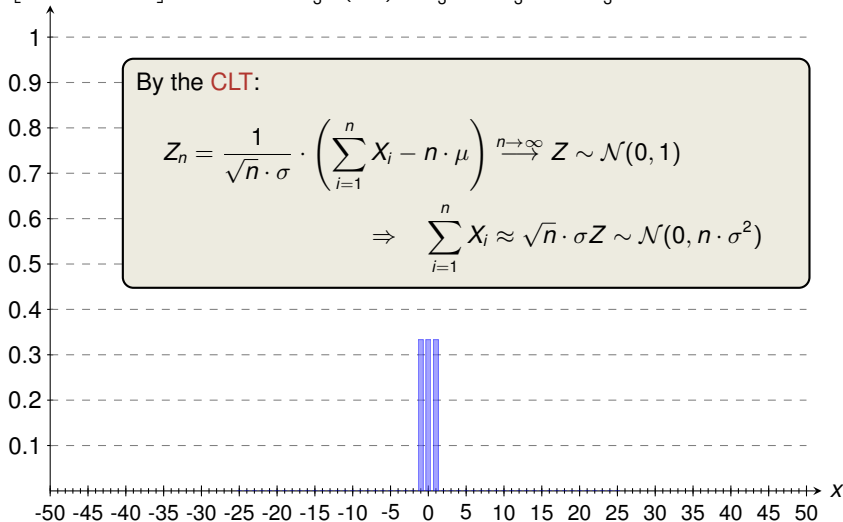
Illustrations

Examples

## Illustration of CLT (1/4)

$$P \left[ \sum_{j=1}^1 X_j = x \right]$$

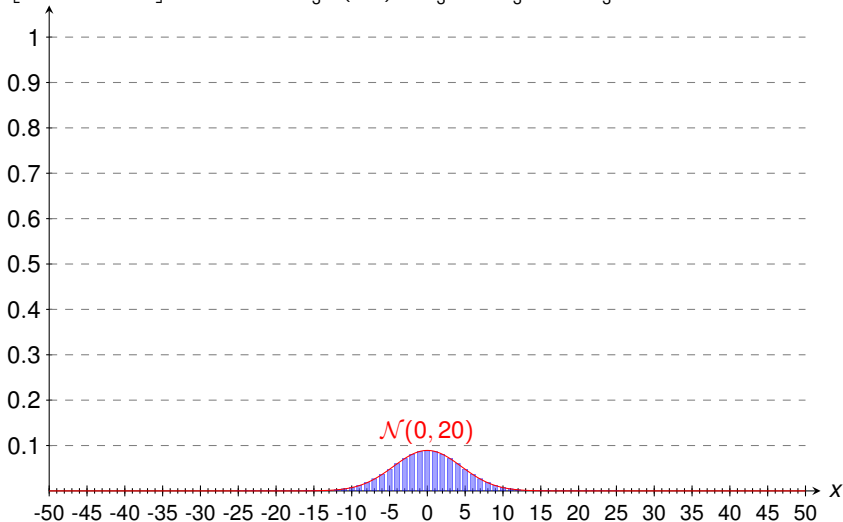
- $\mu = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{3} \cdot (-1)^2 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1^2 = \frac{2}{3}$



## Illustration of CLT (1/4)

$$\mathbf{P} \left[ \sum_{j=1}^{30} X_j = x \right]$$

- $\mu = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{3} \cdot (-1)^2 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1^2 = \frac{2}{3}$

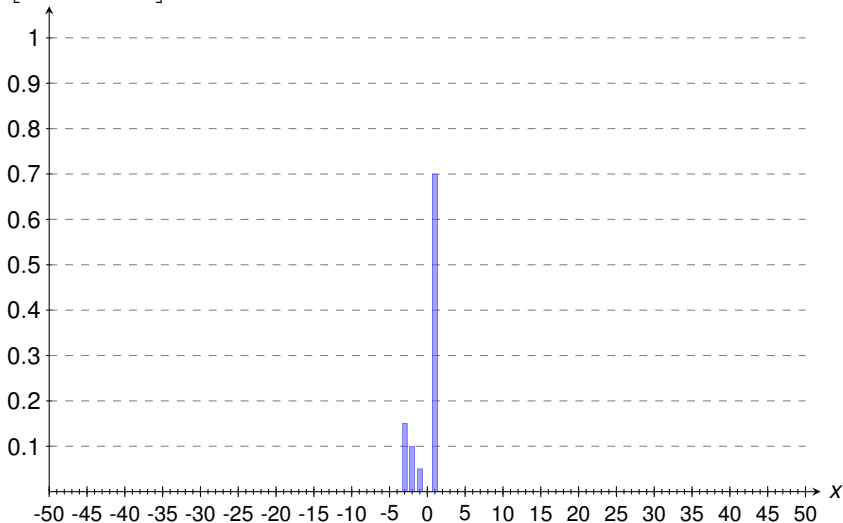


## Illustration of CLT (2/4)

$$\mathbf{P} \left[ \sum_{j=1}^1 X_j = x \right]$$

$$\blacksquare \mu = 0.15 \cdot (-3) + 0.1 \cdot (-2) + 0.05 \cdot (-1) + 0.7 \cdot 1 = 0$$

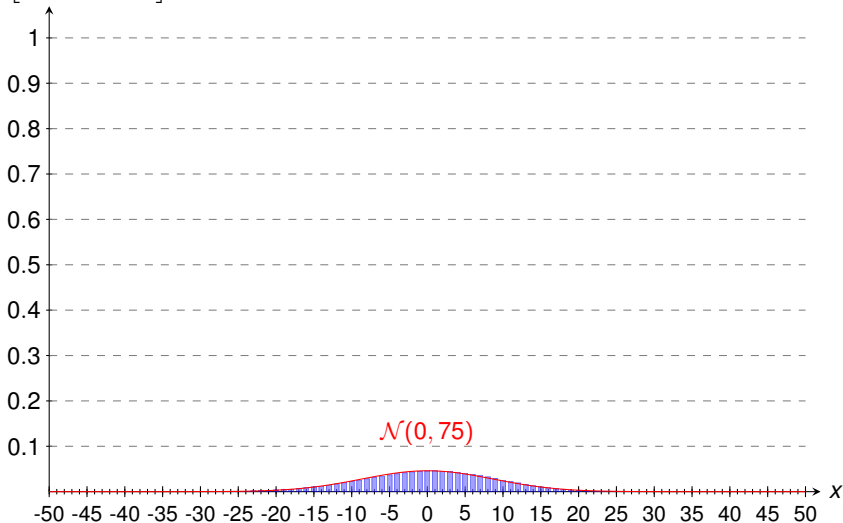
$$\blacksquare \sigma^2 = 0.15 \cdot 9 + 0.1 \cdot 4 + 0.05 \cdot 1 + 0.7 \cdot 1 = 2.5$$



## Illustration of CLT (2/4)

$$\mathbf{P} \left[ \sum_{j=1}^{30} X_j = x \right]$$

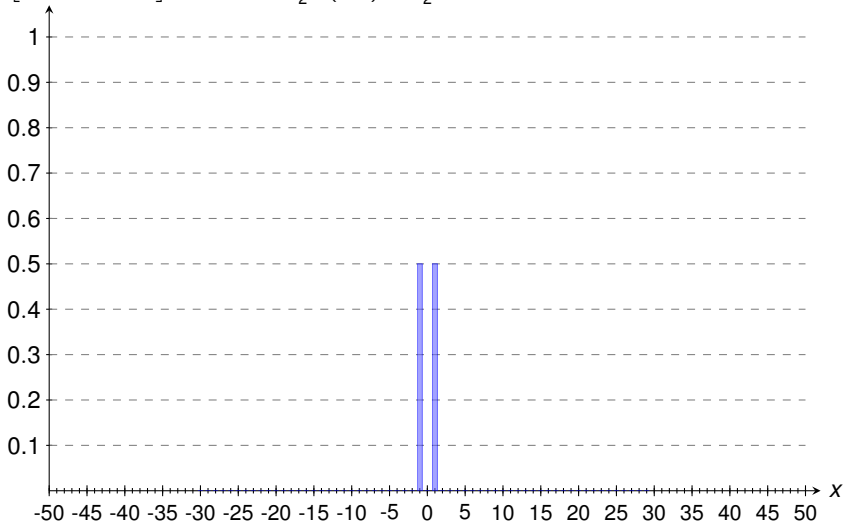
- $\mu = 0.15 \cdot (-3) + 0.1 \cdot (-2) + 0.05 \cdot (-1) + 0.7 \cdot 1 = 0$
- $\sigma^2 = 0.15 \cdot 9 + 0.1 \cdot 4 + 0.05 \cdot 1 + 0.7 \cdot 1 = 2.5$



## Illustration of CLT (3, Part I) (Distribution from Lecture 8)

$$\mathbf{P} \left[ \sum_{j=1}^1 X_j = x \right]$$

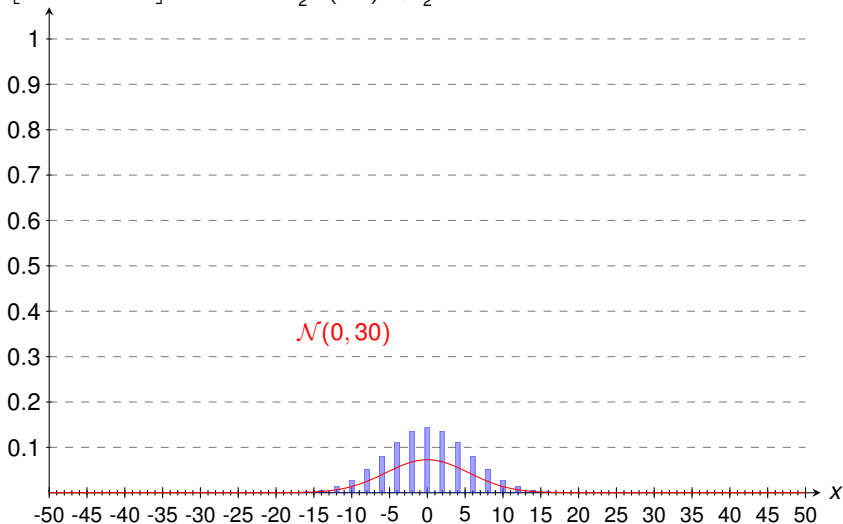
- $\mu = \frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{2} \cdot (-1)^2 + \frac{1}{2} \cdot 1^2 = 1$



## Illustration of CLT (3, Part I) (Distribution from Lecture 8)

$$P \left[ \sum_{j=1}^{30} X_j = x \right]$$

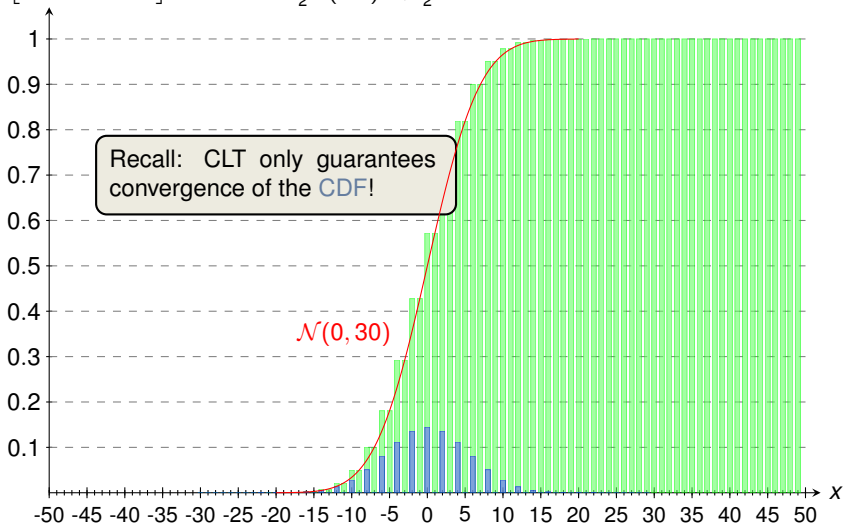
- $\mu = \frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{2} \cdot (-1)^2 + \frac{1}{2} \cdot 1^2 = 1$



## Illustration of CLT (3, Part I) (Distribution from Lecture 8)

$$\mathbf{P} \left[ \sum_{j=1}^{30} X_j = x \right]$$

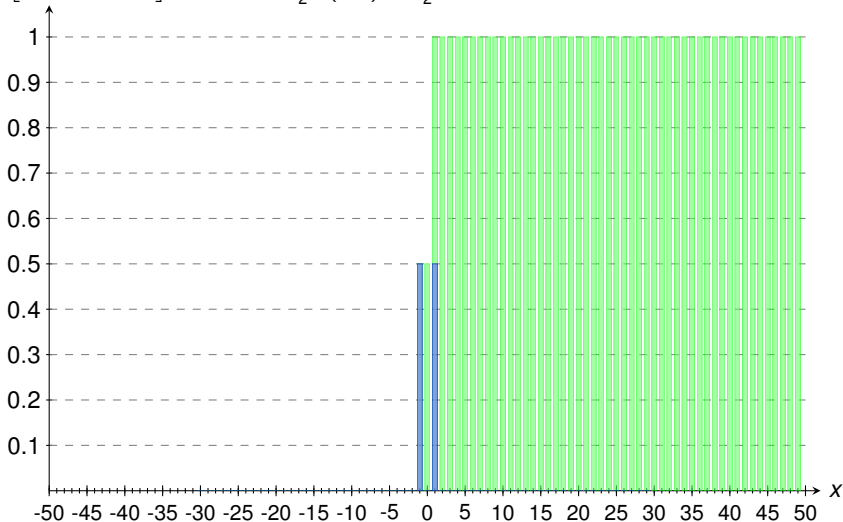
- $\mu = \frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{2} \cdot (-1)^2 + \frac{1}{2} \cdot 1^2 = 1$



## Illustration of CLT (3, Part II) (Distribution from Lecture 8)

$$\mathbf{P} \left[ \sum_{j=1}^1 X_j \leq x \right]$$

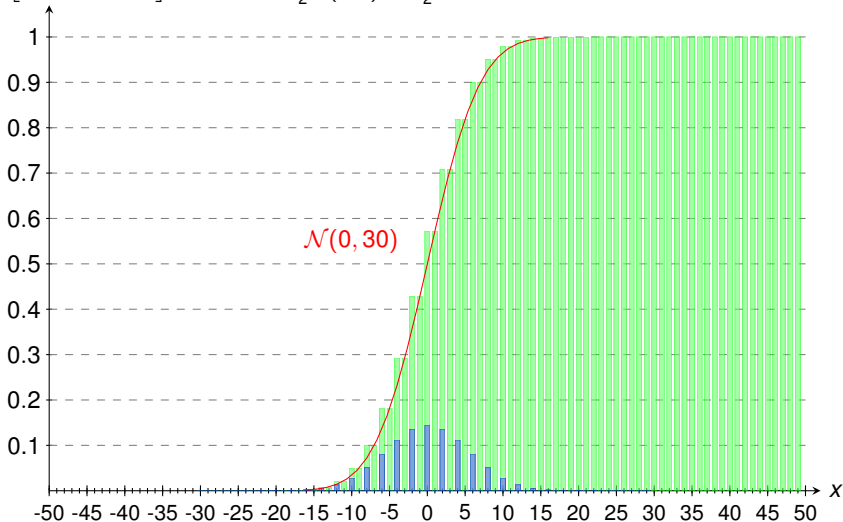
- $\mu = \frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{2} \cdot (-1)^2 + \frac{1}{2} \cdot 1^2 = 1$



## Illustration of CLT (3, Part II) (Distribution from Lecture 8)

$$\mathbf{P} \left[ \sum_{j=1}^{30} X_j \leq x \right]$$

- $\mu = \frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot 1 = 0$
- $\sigma^2 = \frac{1}{2} \cdot (-1)^2 + \frac{1}{2} \cdot 1^2 = 1$



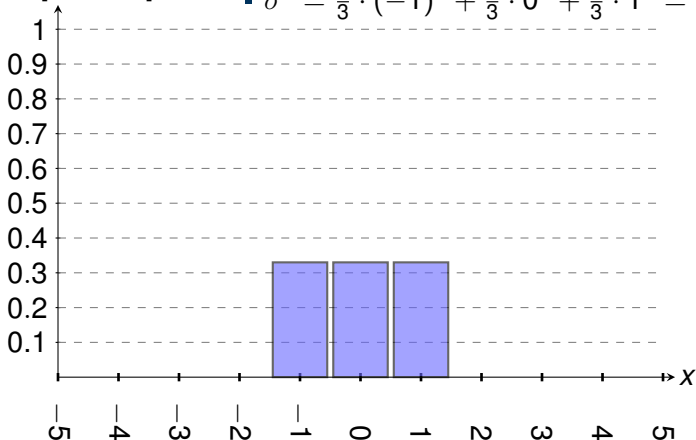
## Illustration of CLT (4, Part I) with Standardising

Please see the full slides for the illustration!

- $\mu = \frac{1}{3} \cdot (-1) + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 = 0$

- $\sigma^2 = \frac{1}{3} \cdot (-1)^2 + \frac{1}{3} \cdot 0^2 + \frac{1}{3} \cdot 1^2 = \frac{2}{3}$

$P[X_1 = x]$



## Illustration of CLT (4, Part II) with Standardising

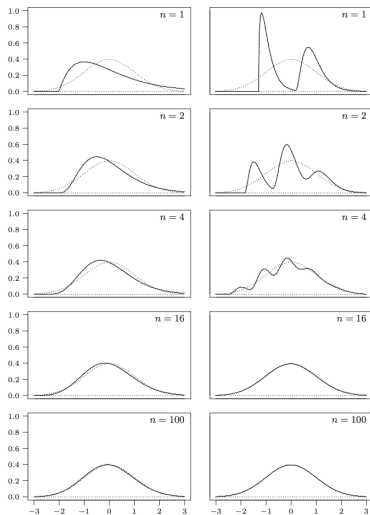


Fig. 14.2. Densities of standardized averages  $Z_n$ . Left column: from a gamma density; right column: from a bimodal density. Dotted line:  $N(0, 1)$  probability density.

Source: Dekking et al., Modern Introduction to Statistics

# Outline

---

Recap: Weak Law of Large Numbers

Central Limit Theorem

Illustrations

Examples

# Recall: Standard Normal Table

Section 5.4 Normal Random Variables 201

TABLE 5.1: AREA  $\Phi(x)$  UNDER THE STANDARD NORMAL CURVE TO THE LEFT OF  $X$

$X$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9988	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9991	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Source: Ross, Probability 8th ed.

**Question:** What if we need  $\Phi(x)$  for negative  $x$ ?

$$Z \sim \mathcal{N}(0, 1) \quad \mathbf{P}[Z \leq x] = \Phi(x)$$

Due to symmetry of density we have  $\Phi(x) = 1 - \Phi(-x)$ .

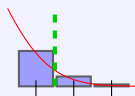
## Normal Approximation of the Binomial Distribution

### Example 1

Suppose you are attending a multiple-choice exam of 10 questions and you are completely unprepared. Each question has 4 choices, and you are going to pass the exam if you **guess** at least 6 correct answers. Use the normal approximation to estimate the probability of passing.

Answer

- Let  $X \sim \text{Bin}(10, 1/4)$ . We are interested in  $\mathbf{P}[X \geq 6]$ .
- Note  $X := \sum_{i=1}^n X_i$ , where each  $X_i \sim \text{Ber}(p)$  and  $n = 10, p = 1/4$ .  
 $\Rightarrow \mu = 1/4$  and  $\sigma^2 = p(1-p) = 3/16$ .
- Applying the **CLT** yields:


$$\begin{aligned} \mathbf{P}[X \geq 6] &= \mathbf{P}\left[\sum_{i=1}^n X_i \geq 6\right] \\ &= \mathbf{P}\left[\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma}} \geq \frac{6 - n\mu}{\sqrt{n\sigma}}\right] \\ &= \mathbf{P}\left[Z_{10} \geq \frac{6 - 2.5}{\sqrt{10} \cdot \sqrt{3/16}}\right] \approx 1 - \Phi(2.56) \approx 0.0052. \end{aligned}$$

**continuity correction:** a better approximation is obtained by  $\mathbf{P}\left[\sum_{i=1}^n X_i \geq 5.5\right] \rightsquigarrow \approx 0.0143$

True value is 0.0197. Error lies in the discretisation!

## A “Reverse” Application of the CLT

### Example 2

Suppose we are sequentially loading one container with packets, whose weights are i.i.d. exponential variables with parameter  $\lambda = 1/2$ . The container has a capacity of 100 weight units. How many packets can we load so that we meet the capacity threshold with at least .95 probability?

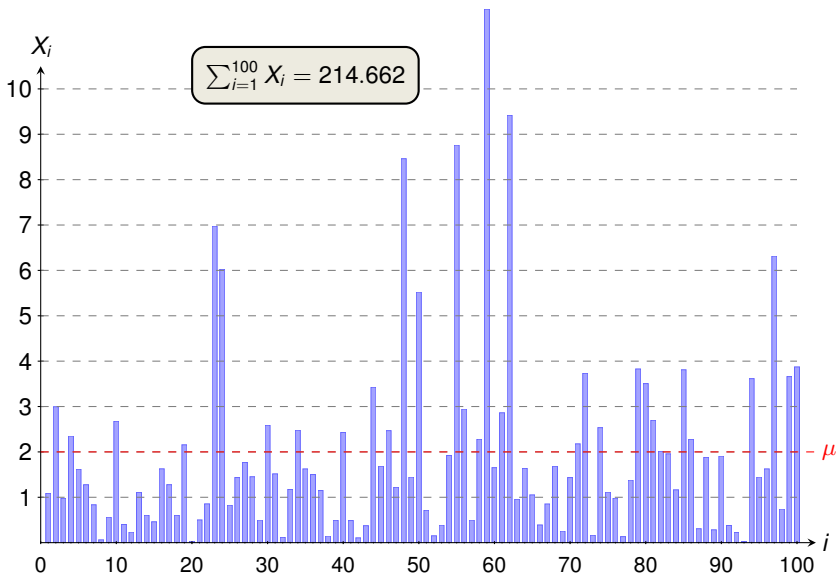
Answer

- We have  $X_1, X_2, \dots, X_n \sim \text{Exp}(1/2)$ , where  $n$  is unknown.
- Recall that  $\mu = \sigma = 2$ .
- By the CLT,

$$\begin{aligned} \mathbf{P} \left[ \sum_{i=1}^n X_i \leq 100 \right] &= \mathbf{P} \left[ \frac{\sum_{i=1}^n X_i - 2n}{2\sqrt{n}} \leq \frac{100 - 2n}{2\sqrt{n}} \right] \\ &\approx \Phi \left( \frac{100 - 2n}{2\sqrt{n}} \right) \stackrel{!}{=} 0.95. \end{aligned}$$

- Using a normal table (looking for value 0.95) yields:  $\frac{100-2n}{2\sqrt{n}} = 1.645$ .
- ⇒ Solving the quadratic gives  $n \leq 39.6$  (so  $n \leq 39$ )

## A Sample of 100 Exponential Random Variables $Exp(1/2)$



## Comparison between Markov, Chebyshev and CLT

### Example 3

Consider  $n = 100$  independent coin flips. Estimate the probability that the number of heads is greater or equal than 75.

Answer

- **Markov:**  $X = \sum_{i=1}^{100} X_i$ ,  $X_i \in \{0, 1\}$  and  $\mathbf{E}[X] = 100 \cdot \frac{1}{2} = 50$ .

$$\mathbf{P}[X \geq 3/2 \cdot \mathbf{E}[X]] \leq 2/3 = 0.666.$$

- **Chebyshev:**  $\mathbf{V}[X] = \sum_{i=1}^{100} \mathbf{V}[X_i] = 100 \cdot (1/2)^2 = 25$ .

$$\mathbf{P}[|X - \mu| \geq 25] \leq \frac{\mathbf{V}[X]}{25^2} = \frac{1}{25} = 0.04.$$

As  $X$  is symmetric, we could deduce probability is at most 0.02.

- **Central Limit Theorem:** First standardise:  $Z_n = \frac{X - n \cdot 1/2}{\sqrt{n \cdot 1/2}}$

$$\mathbf{P}[X \geq 74.5] = \mathbf{P}\left[Z_n \geq \frac{74.5 - n \cdot 1/2}{\sqrt{n \cdot 1/2}}\right] \approx 1 - \Phi(4.9) = 4.79 \cdot 10^{-7}$$

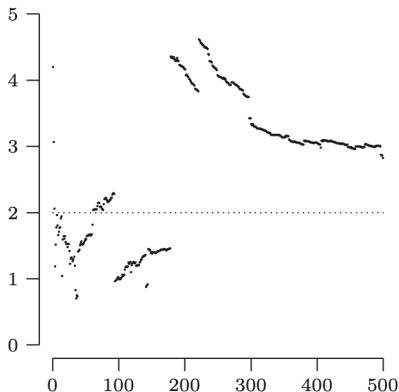
- exact probability is  $2.82 \cdot 10^{-7}$

CLT gives a much better result (but requires i.i.d.)

- **Side Note:** without continuity correction, we have 75 instead 74.5:

- This leads to  $1 - \Phi(5) = 2.86 \cdot 10^{-7}$
- Issue: threshold too large ( $\mathbf{P}[X \geq a] \approx \mathbf{P}[X = a]$ )  $\Rightarrow$  CLT less precise
- In this region, 75 gives a better approximation than 74.5, but for smaller values (e.g.,  $\leq 63$ ) the continuity corrections gives significantly better results.

## A Distribution whose Average does not converge



$\text{Cau}(2, 1)$  distribution, Source: Dekking et al., Modern Introduction to Statistics

The **Cauchy distribution** has “too heavy” tails (no expectation), in particular the average does not converge.

# Introduction to Probability

Lecture 10: Estimators (Part I)

Mateja Jamnik, [Thomas Sauerwald](#)

University of Cambridge, Department of Computer Science and Technology  
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

Easter 2026



UNIVERSITY OF  
CAMBRIDGE

# Outline

---

Introduction

Defining and Analysing Estimators

More Examples

## Introduction

Setting: We can take **random samples** in the form of **i.i.d. random variables**  $X_1, X_2, \dots, X_n$  from an **unknown distribution**.

- Taking enough samples allows us to estimate the **mean** (WLLN, CLT)
- Using indicator variables, we can estimate  $\mathbf{P}[X \leq a]$  for any  $a \in \mathbb{R}$   
↪ in principle we can reconstruct the entire **distribution**
- How can we directly estimate the **variance** or other parameters?  
↪ **estimator**
- How can we **measure** the accuracy of an estimator?  
↪ **bias** (this lecture) and **mean squared error** (next lecture)



### Physical Experiments:

Measurement = Quantity of Interest + Measurement Error

## Empirical Distribution Functions

Definition of Empirical Distribution Function (Empirical CDF)

Let  $X_1, X_2, \dots, X_n$  be i.i.d. samples, and  $F$  be the corresponding distribution function. For any  $a \in \mathbb{R}$ , define

$$F_n(a) := \frac{\text{number of } X_i \in (-\infty, a]}{n}.$$

Remark

The **Weak Law of Large Numbers** implies that for any  $\epsilon > 0$  and  $a \in \mathbb{R}$ ,

$$\lim_{n \rightarrow \infty} \mathbf{P}[|F_n(a) - F(a)| > \epsilon] = 0.$$

Thus by taking enough samples, we can estimate the entire distribution (including its expectation and variance).

## Empirical Distribution Functions (Example 1/2)

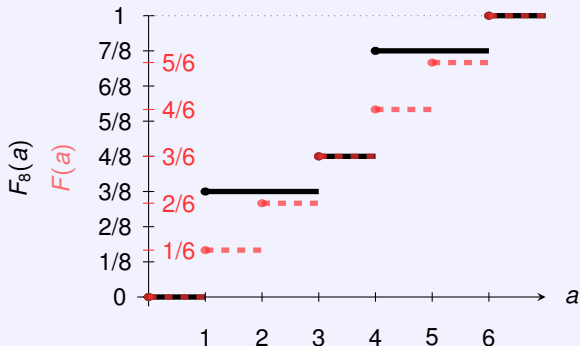
### Example 1

Consider throwing an unbiased dice 8 times, and let the **realisation** be:

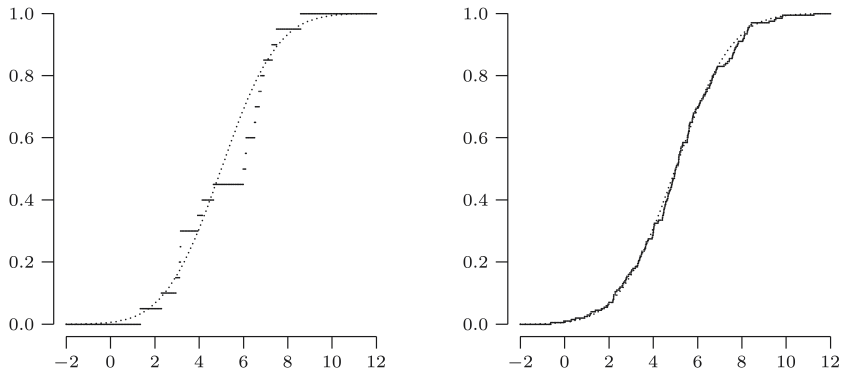
$$(x_1, x_2, \dots, x_8) = (4, 1, 4, 3, 1, 6, 4, 1).$$

What is the Empirical Distribution Function  $F_8(a)$ ?

Answer



## Empirical Distribution Functions (Example 2/2)



Source: Modern Introduction to Statistics

**Figure:** Empirical Distribution Functions of samples from a Normal Distribution  $\mathcal{N}(5, 4)$  ( $n = 20$  left,  $n = 200$  right)

## An Example of an Estimation Problem

Scenario

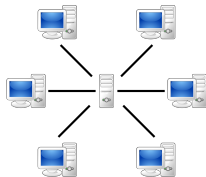
Consider the **packages arriving at a network server**.

- We might be interested in:
  1. number of packets that arrive within a “typical” minute
  2. percentage of minutes during which no packets arrive
- If arrivals occur at random time  $\rightsquigarrow$  number of arrivals during one minute follows a **Poisson distribution** with **unknown** parameter  $\lambda$

Estimator for  $\lambda$

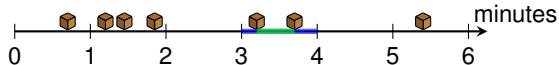
Estimator for  $e^{-\lambda}$

Waiting Time (Lecture 5, Slide 22)



Source: Wikipedia

$$\Delta \sim \text{Exp}(\lambda)$$



$$X \sim \text{Poi}(\lambda)$$

$$\mathbf{P}[X = k] = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

### Definition of Estimator

An **estimate** is a value  $t$  that only depends on the dataset  $x_1, x_2, \dots, x_n$ , i.e.,

$$t = h(x_1, x_2, \dots, x_n).$$

Then  $t$  is a realisation of the random variable

$$T = h(X_1, X_2, \dots, X_n),$$

which is called **estimator**.

### Questions:

- What makes an **estimator** suitable? **unbiased** (later: **mean squared error**)
- Does an **unbiased estimator** always exist? How to compute it?
- If there are several **unbiased** estimators, which one to choose?

# Outline

---

Introduction

Defining and Analysing Estimators

More Examples

## Example: Arrival of Packets (1/3)

- **Samples:** Given  $X_1, X_2, \dots, X_n$  i.i.d.,  $X_i \sim \text{Pois}(\lambda)$
- **Meaning:**  $X_i$  is the number of packets arriving in minute  $i$



### Example 2

Estimate  $\lambda$  by using the sample mean  $\bar{X}_n$ .

Answer

We have

$$\bar{X}_n := \frac{X_1 + X_2 + \dots + X_n}{n},$$

and  $\mathbf{E}[\bar{X}_n] = \mathbf{E}[X_1] = \lambda$ . This suggests the estimator:

$$h(X_1, X_2, \dots, X_n) := \bar{X}_n.$$

Applying the **Weak Law of Large Numbers**:

$$\lim_{n \rightarrow \infty} \mathbf{P} \left[ \left| \bar{X}_n - \lambda \right| > \epsilon \right] = 0 \quad \text{for any } \epsilon > 0.$$

## Example: Arrival of Packets (2/3)

### Example 3a

Define an estimator  $h_1$  for the probability of zero arrivals,  $e^{-\lambda}$ .

Answer

Let  $X_1, X_2, \dots, X_n$  be the  $n$  samples. Let

$$Y_i := \mathbf{1}_{X_i=0}.$$

Then

$$\mathbf{E}[Y_i] = \mathbf{P}[X_i = 0] = e^{-\lambda},$$

and thus we can define an estimator by

$$h_1(X_1, X_2, \dots, X_n) := \frac{Y_1 + Y_2 + \dots + Y_n}{n} = \bar{Y}_n.$$

## Example: Arrival of Packets (3/3)

- Suppose we get the samples  $(x_1, x_2, x_3) = (50, 100, 0)$
- Then  $(y_1, y_2, y_3) = (0, 0, 1)$ , and  $h_1(x_1, x_2, x_3) = \frac{1}{3}$
- This seems **too large!** Also note that for the samples  $(x_1, x_2, x_3) = (1, 1, 0)$ , our estimator would give the same estimate

### Example 3b

Define an estimator  $h_2$  for  $e^{-\lambda}$  based on  $\bar{X}_n$ .

Answer

We saw that  $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$  satisfies  $\mathbf{E}[\bar{X}_n] = \mathbf{E}[X_1] = \lambda$ .

Recall by the **Weak Law of Large Numbers**:

$$\lim_{n \rightarrow \infty} \mathbf{P} \left[ \left| \bar{X}_n - \lambda \right| > \epsilon \right] = 0 \quad \text{for any } \epsilon > 0.$$

This suggests to estimate  $e^{-\lambda}$  by  $e^{-\bar{X}_n}$ . Hence our estimator is

$$h_2(X_1, X_2, \dots, X_n) := e^{-\bar{X}_n}.$$

## Behaviour of the Estimators

---

- Suppose we have  $n = 30$  and we want to estimate  $e^{-\lambda}$
- Consider the **two estimators**  $h_1(X_1, \dots, X_n)$  and  $h_2(X_1, \dots, X_n)$ .

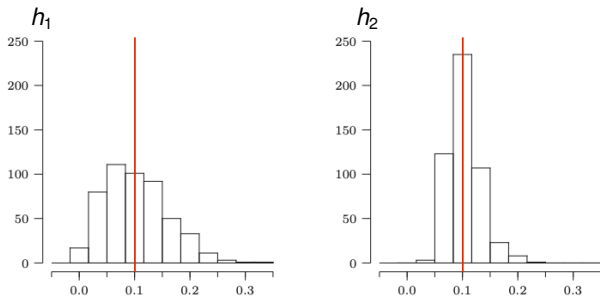
How **good** are these two estimators?

- ⇒ The first estimator can only attain values  $0, \frac{1}{30}, \frac{2}{30}, \dots, 1$
- ⇒ The second estimator can only attain values  $1, e^{-1/30}, e^{-2/30}, \dots$

For most values of  $\lambda$ , both estimators will never return the **exact** value of  $e^{-\lambda}$  on the basis of 30 observations.

## Simulation of the two Estimators

- The **unknown parameter** is  $p = e^{-\lambda} = 0.1$  (i.e.,  $\lambda = \ln 10 \approx 2.30 \dots$ )
- We consider  $n = 30$  minutes and compute  $h_1$  and  $h_2$
- We repeat this 500 times and draw a **frequency histogram** ( $h_1 = \bar{Y}_n$  left,  $h_2 = e^{-\bar{X}_n}$  right)



Source: Modern Introduction to Statistics

Both estimators concentrate around the true value 0.1, but the second estimator appears to be more concentrated.

## Unbiased Estimators and Bias

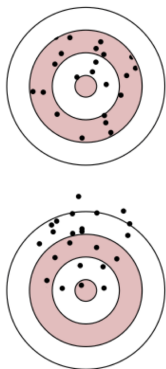
Definition

An **estimator**  $T$  is called an **unbiased estimator** for the parameter  $\theta$  if

$$\mathbf{E}[T] = \theta,$$

irrespective of the value  $\theta$ . The **bias** is defined as

$$\mathbf{E}[T] - \theta = \mathbf{E}[T - \theta].$$



Source: Edwin Leuven (Point Estimation)

Which of the two estimators  $h_1, h_2$  are unbiased?



### Example 4a

Is  $h_1(X_1, X_2, \dots, X_n) = \frac{Y_1 + Y_2 + \dots + Y_n}{n}$  an unbiased estimator for  $e^{-\lambda}$ ?

Answer

Recall we defined  $Y_i := \mathbf{1}_{X_i=0}$ . **Yes**, because:

$$\begin{aligned}\mathbf{E}[h_1(X_1, X_2, \dots, X_n)] &= \frac{n \cdot \mathbf{E}[Y_1]}{n} \\ &= \mathbf{P}[X_1 = 0] \\ &= e^{-\lambda}.\end{aligned}$$

## Bias of the Second Estimator (and Jensen's Inequality)

### Example 4b

Is  $h_2(X_1, X_2, \dots, X_n) = e^{-\bar{X}_n}$  an **unbiased estimator** for  $e^{-\lambda}$ ?

Answer

**No!** (recall:  $\mathbf{E}[X^2] \geq \mathbf{E}[X]^2$ )

- We have

$$\mathbf{E}[e^{-\bar{X}_n}] > e^{-\mathbf{E}[\bar{X}_n]} = e^{-\lambda}$$

- This follows by **Jensen's inequality**, and the inequality is **strict** since  $g : z \mapsto e^{-z}$  is **strictly convex** and  $\bar{X}_n$  is not constant.
- Thus  $h_2(X_1, X_2, \dots, X_n)$  is not unbiased – it has **positive bias**.

$$\lambda g(a) + (1 - \lambda)g(b) \geq g(\lambda a + (1 - \lambda)b)$$

### Jensen's Inequality

For any random variable  $X$ , and any **convex function**  $g : \mathbb{R} \rightarrow \mathbb{R}$ , we have

$$\mathbf{E}[g(X)] \geq g(\mathbf{E}[X]).$$

If  $g$  is **strictly convex** and  $X$  is not constant, then the inequality is strict.

## Asymptotic Bias of the Second Estimator (non-examinable)

### Example 4c

$\mathbf{E}[h_2(X_1, \dots, X_n)] \xrightarrow{n \rightarrow \infty} e^{-\lambda}$  (hence it is **asymptotically unbiased**).

Answer

- Recall  $h_2(X_1, \dots, X_n) = e^{-\bar{X}_n}$ . For any  $0 \leq k \leq n$ ,

$$\mathbf{P}\left[h_2(X_1, \dots, X_n) = e^{-k/n}\right] = \mathbf{P}\left[\sum_{i=1}^n X_i = k\right] = \mathbf{P}[Z = k],$$

where  $Z \sim \text{Pois}(n \cdot \lambda)$  (since  $\text{Pois}(\lambda_1) + \text{Pois}(\lambda_2) = \text{Pois}(\lambda_1 + \lambda_2)$ )

$$\Rightarrow \mathbf{P}\left[h_2(X_1, \dots, X_n) = e^{-k/n}\right] = \frac{e^{-n\lambda} \cdot (n\lambda)^k}{k!}$$

$$\Rightarrow \mathbf{E}[h_2(X_1, \dots, X_n)] = \sum_{k=0}^{\infty} e^{-n\lambda} \cdot \frac{(n\lambda)^k}{k!} \cdot e^{-k/n}$$

By LOTUS

$$= e^{-n\lambda} \cdot e^{n\lambda e^{-1/n}} \sum_{k=0}^{\infty} e^{-n\lambda e^{-1/n}} \cdot \frac{(n\lambda e^{-1/n})^k}{k!}$$

$$= e^{-n\lambda \cdot (1 - e^{-1/n})} \cdot 1$$

since  $e^x = 1 + x + O(x^2)$  for small  $x$

$$\xrightarrow{n \rightarrow \infty} e^{-n\lambda \cdot (1 - 1/n + O(1/n^2))} = e^{-\lambda + O(\lambda/n)}.$$

Hence in the limit, the positive bias of  $h_2$  diminishes.

# Outline

---

Introduction

Defining and Analysing Estimators

More Examples

### Unbiased Estimators for Expectation and Variance

Let  $X_1, X_2, \dots, X_n$  be **identically distributed** samples from a distribution with finite expectation  $\mu$  and finite variance  $\sigma^2$ .

- Then

$$\bar{X}_n := \frac{X_1 + X_2 + \dots + X_n}{n}$$

is an **unbiased** estimator for  $\mu$ .

- Furthermore, for  $n \geq 2$ ,

$$S_n = S_n(X_1, \dots, X_n) := \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

is an **unbiased** estimator for  $\sigma^2$ .

## Example 5

We need to prove:  $\mathbf{E} [ S_n ] = \sigma^2$ .

Answer

Multiplying by  $n - 1$  yields:

$$\begin{aligned} (n - 1) \cdot S_n &= \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \sum_{i=1}^n (X_i - \mu + \mu - \bar{X}_n)^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n (\bar{X}_n - \mu)^2 - 2 \sum_{i=1}^n (X_i - \mu) (\bar{X}_n - \mu) \\ &= \sum_{i=1}^n (X_i - \mu)^2 + n (\bar{X}_n - \mu)^2 - 2 (\bar{X}_n - \mu) \cdot n \cdot (\bar{X}_n - \mu) \\ &= \sum_{i=1}^n (X_i - \mu)^2 - n (\bar{X}_n - \mu)^2. \end{aligned}$$

Let us now take **expectations**:

**By Lec. 8, Slide 21:**  $\mathbf{E} [ (\bar{X}_n - \mu)^2 ] = \mathbf{V} [ \bar{X}_n ] = \sigma^2/n$

$$\begin{aligned} (n - 1) \cdot \mathbf{E} [ S_n ] &= \sum_{i=1}^n \mathbf{E} [ (X_i - \mu)^2 ] - n \cdot \mathbf{E} [ (\bar{X}_n - \mu)^2 ] \\ &= n \cdot \sigma^2 - n \cdot \sigma^2/n \\ &= (n - 1) \cdot \sigma^2. \end{aligned}$$

## An Unbiased Estimator may not always exist

### Example 6

Suppose that we have one sample  $X \sim \text{Bin}(n, p)$ , where  $0 < p < 1$  is unknown but  $n$  is known. Prove there is **no unbiased estimator** for  $1/p$ .

Answer

- First a simpler proof which exploits that  $p$  might be arbitrarily small
- **Intuition:** By making  $p$  smaller and smaller, we force  $\max_{0 \leq k \leq n} T(k)$ ,  $k \in \{0, 1, \dots, n\}$  to become bigger and bigger
- **Formal Argument:**
  - Fix any estimator  $T(X)$
  - Define  $M := \max_{0 \leq k \leq n} T(k)$ . Then,

$$\begin{aligned} \mathbf{E}[T(X)] &= \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} \cdot T(k) \\ &\leq M \cdot \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = M. \end{aligned}$$

- Hence this estimator does not work for  $p < \frac{1}{M}$ , since then  $\mathbf{E}[T(X)] \leq M < \frac{1}{p}$  (negative bias!)
- The next proof will work even if  $p \in [a, b]$  for  $0 < a < b \leq 1$ .

## An Unbiased Estimator may not always exist (cntd. - non-examinable)

### Example 6 (cntd.)

Suppose that we have one sample  $X \sim \text{Bin}(n, p)$ , where  $0 < p < 1$  is unknown but  $n$  is known. Prove there is **no unbiased estimator** for  $1/p$ .

Answer

- Suppose there exists an unbiased estimator with  $\mathbf{E}[T(X)] = 1/p$ .
- Then

$$\begin{aligned}1 &= p \cdot \mathbf{E}[T(X)] \\&= p \cdot \sum_{k=0}^n \mathbf{P}[X = k] \cdot T(k) \\&= p \cdot \sum_{k=0}^n \binom{n}{k} p^k \cdot (1-p)^{n-k} \cdot T(k)\end{aligned}$$

- Last term is a **polynomial of degree  $n + 1$**  with constant term zero  
 $\Rightarrow p \cdot \mathbf{E}[T(X)] - 1$  is a **(non-zero) polynomial of degree  $\leq n + 1$**   
 $\Rightarrow$  this polynomial has at most  $n + 1$  roots  
 $\Rightarrow \mathbf{E}[T(X)]$  can be equal to  $1/p$  for at most  $n + 1$  values of  $p$ , and thus cannot be an unbiased.

# Introduction to Probability

Lecture 11: Estimators (Part II)

Mateja Jamnik, [Thomas Sauerwald](#)

University of Cambridge, Department of Computer Science and Technology  
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

Easter 2026



Estimating Population Size (First Model)

Mean Squared Error

Estimating Population Size (Second Model)

## Estimating Population Size (First Model)

- Suppose we have a sample of a few serial numbers (IDs) of some product
- We assume IDs are running from 1 to an **unknown parameter**  $N$  (so  $N = \theta$ )
- Each of the IDs is drawn without replacement from the **discrete uniform distribution** over  $\{1, 2, \dots, N\}$
- This is also known as **Tank Estimation Problem** or **(Discrete) Taxi Problem**

7, 3, 10, 46, 14



Warning

- As before, we denote the samples  $X_1, X_2, \dots, X_n$
- Since sampling is without replacement:
  - they are **not independent!** (but identically distributed)
  - their number must satisfy  $n \leq N$

## First Estimator Based on Sample Mean

### Example 1

Construct an **unbiased estimator**  $T_1$  using the **sample mean**.

Answer

- The sample mean is

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

- Linearity of expectation applies (even for **dependent** random var.!):

$$\begin{aligned} \mathbf{E}[\bar{X}_n] &= \frac{n \cdot \mathbf{E}[X_1]}{n} = \mathbf{E}[X_1] \\ &= \sum_{i=1}^N i \cdot \frac{1}{N} = \frac{N+1}{2}. \end{aligned}$$

- Thus we obtain an **unbiased estimator** by

$$T_1 := 2 \cdot \bar{X}_n - 1.$$

## Example: Odd Behaviour of $T_1$

- Suppose  $n = 5$
- Let the sample be

7, 3, 10, 46, 14

- The estimator returns:

$$T_1 = 2 \cdot \bar{X}_n - 1 = 2 \cdot \frac{80}{5} - 1 = 31 \text{ ☹}$$

This estimator will often unnecessarily **underestimate** the true value  $N$ .

Challenging exercise: Find a lower bound on  $\mathbf{P} [ T_1 < \max(X_1, X_2, \dots, X_n) ]$

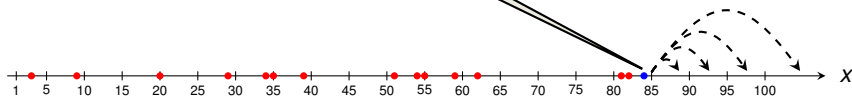
- Achieving **unbiasedness** alone is not a good strategy
- **Improvement:** find an estimator which always returns a value at least  $\max(X_1, X_2, \dots, X_n)$

## Intuition: Constructing an Estimator based on Maximum Sample

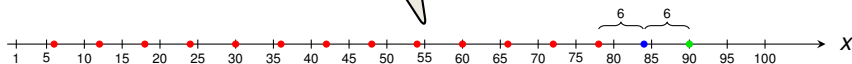
- Suppose  $n = 15$
- Our samples are:

9, 82, 39, 35, 20, 51, 54, 62, 81, 29, 84, 59, 3, 34, 55

How much should we add to the maximum?



Rearrange the other 14 points equi-spaced between 0 and 84.



$$\max(X_1, \dots, X_n) + \frac{\max(X_1, \dots, X_n)}{n-1}$$

This suggests  $84 + 6 = 90$  as the estimate!

## Deriving the Estimator Based on Maximum Sample

### Example 2

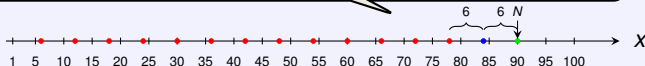
Construct an **unbiased estimator**  $T_2$  using  $\max(X_1, \dots, X_n)$

Answer

- Calculate expectation of the maximum (for details see Dekking et al.)

$$\mathbf{E}[\max(X_1, \dots, X_n)] = \dots = \frac{n}{n+1} \cdot N + \frac{n}{n+1} = \frac{n}{n+1} \cdot (N+1).$$

Equi-spaced configuration would suggest  $\max(X_1, \dots, X_n) \approx \frac{n-1}{n} \cdot N$

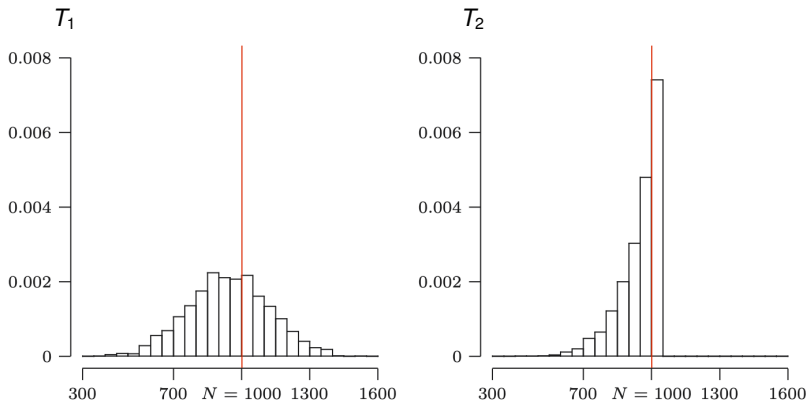


- Hence we obtain an **unbiased estimator** by

$$T_2 := \frac{n+1}{n} \cdot \max(X_1, \dots, X_n) - 1.$$

- For our samples before, we get  $t_2 = \frac{16}{15} \cdot 84 - 1 = 88.6$ .

## Empirical Analysis of the two Estimators



Source: Modern Introduction to Statistics

Figure: Histogram of 2000 values for  $T_1$  and  $T_2$ , when  $N = 1000$  and  $n = 10$ .

Can we find a quantity that captures the superiority of  $T_2$  over  $T_1$ ?

# Outline

---

Estimating Population Size (First Model)

Mean Squared Error

Estimating Population Size (Second Model)

## Mean Squared Error

### Mean Squared Error Definition

Let  $T$  be an estimator for a parameter  $\theta$ . The **mean squared error** of  $T$  is

$$\mathbf{MSE} [ T ] = \mathbf{E} [ ( T - \theta )^2 ].$$

- According to this, estimator  $T_1$  **better** than  $T_2$  if  $\mathbf{MSE} [ T_1 ] < \mathbf{MSE} [ T_2 ]$ .

### Bias-Variance Decomposition

The **mean squared error** can be decomposed into:

$$\mathbf{MSE} [ T ] = \underbrace{(\mathbf{E} [ T ] - \theta)^2}_{= \text{Bias}^2} + \underbrace{\mathbf{V} [ T ]}_{= \text{Variance}}$$

- If  $T_1$  and  $T_2$  are both **unbiased**,  $T_1$  is **better** than  $T_2$  iff  $\mathbf{V} [ T_1 ] < \mathbf{V} [ T_2 ]$ .

## Bias-Variance Decomposition: Derivation

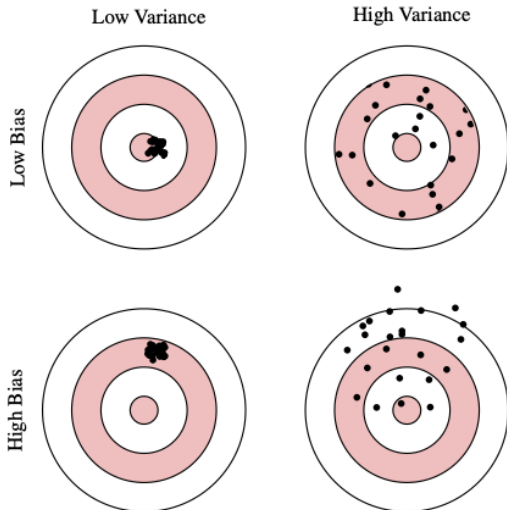
### Example 3

We need to prove:  $\mathbf{MSE}[T] = (\mathbf{E}[T] - \theta)^2 + \mathbf{V}[T]$ .

Answer

$$\begin{aligned}\mathbf{MSE}[T] &= \mathbf{E}[(T - \theta)^2] \\ &= \mathbf{E}[T^2 - 2T\theta + \theta^2] \\ &= \mathbf{E}[T^2] - 2 \cdot \mathbf{E}[T] \cdot \theta + \theta^2 + \mathbf{E}[T^2] - \mathbf{E}[T]^2 \\ &= (\mathbf{E}[T] - \theta)^2 + \mathbf{V}[T].\end{aligned}$$

# Bias-Variance Decomposition: Illustration



Source: Edwin Leuven (Point Estimation)

## Example 4

It holds that  $\mathbf{MSE} [ T_1 ] = \Theta \left( \frac{N^2}{n} \right)$ , where  $T_1 = 2 \cdot \bar{X}_n - 1$ .

Answer

- Since  $T_1$  is unbiased,  $\mathbf{MSE} [ T_1 ] = (\mathbf{E} [ T_1 ] - \theta)^2 + \mathbf{V} [ T_1 ] = \mathbf{V} [ T_1 ]$ , and

$$\mathbf{V} [ T_1 ] = \mathbf{V} [ 2 \cdot \bar{X}_n - 1 ] = 4 \cdot \mathbf{V} [ \bar{X}_n ] = \frac{4}{n^2} \cdot \mathbf{V} [ X_1 + \dots + X_n ]$$

- Note:** The  $X_i$ 's are **not independent!**
- Use generalisation of  $\mathbf{V} [ X_1 + X_2 ] = \mathbf{V} [ X_1 ] + \mathbf{V} [ X_2 ] + 2 \cdot \mathbf{Cov} [ X_1, X_2 ]$  (Exercise Sheet) to  $n$  r.v.'s, and then that the  $X_i$ 's are **identically distributed**, and also the  $(X_i, X_j)$ ,  $i \neq j$ :

$$\begin{aligned} \mathbf{V} [ X_1 + \dots + X_n ] &= \sum_{i=1}^n \mathbf{V} [ X_i ] + 2 \sum_{i=1}^n \sum_{j=i+1}^n \mathbf{Cov} [ X_i, X_j ] \\ &= n \cdot \mathbf{V} [ X_1 ] + 2 \binom{n}{2} \cdot \mathbf{Cov} [ X_1, X_2 ]. \end{aligned}$$

- By definition of the discrete uniform distribution,  $\mathbf{V} [ X_1 ] = \frac{(N+1)(N-1)}{12}$
- Intuitively,  $X_1$  and  $X_2$  are negatively correlated, which would be sufficient to complete the proof. For a more rigorous and precise derivation (see Dekking et al.):

$$\mathbf{Cov} [ X_1, X_2 ] = -\frac{1}{12} (N+1).$$

- Rearranging and simplifying gives

$$\mathbf{V} [ T_1 ] = \frac{(N+1)(N-n)}{3n}.$$

## Analysis of the MSE for $T_2$ (non-examinable)

### Example 5

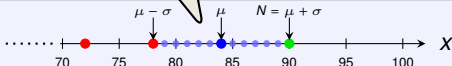
It holds that  $\mathbf{MSE} [ T_2 ] = \Theta \left( \frac{N^2}{n^2} \right)$ , where  $T_2 = \frac{n+1}{n} \cdot \max(X_1, \dots, X_n) - 1$ .

Answer

- $T_2$  is unbiased  $\Rightarrow$  need  $\mathbf{V} [ T_2 ]$  which reduces to  $\mathbf{V} [ \max(X_1, \dots, X_n) ]$
- One can prove: For details see Dekking et al.

$$\mathbf{V} [ \max(X_1, \dots, X_n) ] = \dots = \frac{n(N+1)(N-n)}{(n+2)(n+1)^2} = \Theta \left( \frac{N^2}{n^2} \right)$$

Equi-spaced (idealised) configuration suggests a standard deviation of  $\sigma \approx \frac{N}{n}$



Maximum could have equally likely taken any value between 79 and 90

- $\mathbf{MSE} [ T_2 ]$  is much lower than  $\mathbf{MSE} [ T_1 ] = \Theta \left( \frac{N^2}{n} \right)$ , i.e.,  $\frac{\mathbf{MSE} [ T_1 ]}{\mathbf{MSE} [ T_2 ]} = \frac{n+2}{3}$
- $\Rightarrow$  confirms **simulations** suggesting that  $T_2$  is better than  $T_1$ !
- can be shown  $T_2$  is the **best unbiased estimator**, i.e., it minimises MSE.

# Outline

---

Estimating Population Size (First Model)

Mean Squared Error

Estimating Population Size (Second Model)

## A New Estimation Problem

### Previous Model

- Population/ID space  $S = \{1, 2, \dots, N\}$
- We take **uniform** samples from  $S$  without replacement
- Goal:** Find estimator for  $N$

Similar idea applies to situations where elements are not labelled before we see them first time (**Mark & Recapture Method**)

### New Model

- Population/ID space of size  $|S| = N$
- We take **uniform** samples from  $S$  with replacement
- Goal:** Find estimator for  $N$

- Suppose  $n = 6$ ,  $N = 11$ ,  $S = \{3, 4, 7, 8, 10, 15.83356, 20, 21, 56, 81, 10000\}$
- Let the sample be

10, **81**, 20, 3, **81**, 10000

Let us call this a **collision**

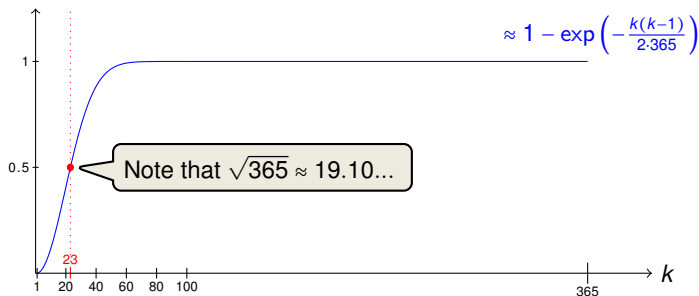
As we do not know  $S$ , our only clue are elements that **were sampled twice.**

## Birthday Problem

**Birthday Problem:** Given a set of  $k$  people

- What is the **probability** of having two with the same birthday (i.e., having at least one collision)?
- What is the **expected number** of people one needs to ask until the first collision occurs?

$P$  [ collision ]



## Estimation via Collision: The Algorithm

Recall: As we do not know  $S$ , our only information are **collisions**.

FIND-FIRST-COLLISION( $S$ )

- 1:  $C = \emptyset$
- 2: **For**  $i = 1, 2, \dots$
- 3:     Take next i.i.d. sample  $X_i$  from  $S$
- 4:     **If**  $X_i \notin C$  **then**  $C \leftarrow C \cup \{X_i\}$
- 5:     **else return**  $T(i)$
- 6: **End For**

$T(i)$  will be the value of the estimator if algo returns after  $i$  rounds. (We want  $T$  unbiased)

- **Running Time:** The expected time until the algorithm stops is:  
= the expected number of samples until a **collision**...

Same as the birthday problem, but now with  $|S| = N$  days... ☺

Expected Running Time (Knuth, Ramanujan)

$$\sqrt{\frac{\pi N}{2}} - \frac{1}{3} + O\left(\frac{1}{\sqrt{N}}\right).$$

**Exercise:** Prove a bound of  $\leq 2 \cdot \sqrt{N}$

## Estimation via Collision: Getting the Estimator Unbiased

### Example 6

One can define  $T(i)$ ,  $i \in \mathbb{N}$ , such that  $\mathbf{E}[T] = |S|$  for any finite, non-empty set  $S$ .

Answer

- We outline a construction by **induction**.
- **Case  $|S| = 1$** : Algo always stops after  $i = 2$  rounds and returns  $T(2)$ . We want

$$1 = \mathbf{E}[T] = T(2) \quad \Rightarrow \quad T(2) = 1.$$

- **Case  $|S| = 2$** : Algo stops after 2 or 3 rounds (w.p. 1/2 each). We want

$$2 = \mathbf{E}[T] = \frac{1}{2} \cdot T(2) + \frac{1}{2} \cdot T(3) \quad \Rightarrow \quad T(3) = 3.$$

- **Case  $|S| = 3$** : gives  $3 = \mathbf{E}[T] = \frac{1}{3} \cdot T(2) + \frac{4}{9} \cdot T(3) + \frac{2}{9} \cdot T(4)$   
 $\Rightarrow T(4) = 6$ , similarly,  $T(5) = 10$  etc.
- can continue to define  $T(i)$  inductively in this way (note  $T$  is **unique**)  
(a proof that  $T(i) = \binom{i}{2}$  is harder)

## Extra Slide on the Collision Algorithm (non-examinable)

- + The algorithm runs in (expected) **sublinear time**  $O(\sqrt{N})$ , where  $N := |S|$
- The algorithm does not take a **pre-specified** and **fixed** number of samples

What can we do with a **fixed number of samples**  $n$ ?

- We cannot find an **unbiased** estimator that works for any  $N$  (similar to Lecture 10, Slide 22)
- Could use **hypothesis testing**: For a fixed sample  $(x_1, x_2, \dots, x_n)$  with  $c$  collisions, what is the probability to have  $c$  collisions under hypothesis that  $N \geq x$  (or  $N = x$ ) for some value  $x$ ?
- **Bayesian Approach**: Assume unknown parameter  $N$  comes from a (known) probability distribution (called **prior distribution**). For a fixed sample  $(x_1, x_2, \dots, x_n)$ , update the probability distribution (called **posterior distribution**)

$$N \sim \text{Exp}(1/1000) \xrightarrow{X_1 = x_1, \dots, X_n = x_n} N \sim \left( \text{Exp}(1/1000) \mid X_1 = x_1, \dots, X_n = x_n \right)$$

# Introduction to Probability

Lecture 12: Online Algorithms

Mateja Jamnik, [Thomas Sauerwald](#)

University of Cambridge, Department of Computer Science and Technology  
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

Easter 2026



UNIVERSITY OF  
CAMBRIDGE

# Outline

---

Stopping Problem 1: Dice Game

Stopping Problem 2: The Secretary Problem

A Generalisation: The Odds Algorithm (non-examinable)

The End...

## Introduction: Dice Game



### Dice Game

- We throw a fair, six-sided dice  $n$  times
- After each throw, you can either STOP or CONTINUE
- You win if you STOP at the last 6 within the  $n$  throws

What is the optimal strategy for maximising the probability of winning?

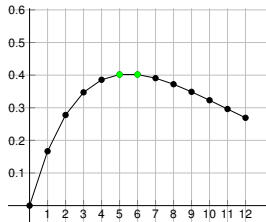
Example ( $n = 10$ )

- 3, 5, 6, 4, 2, 3, 1, 2, 6, 5  $\Rightarrow$  LOSE!  
STOP
- 3, 5, 6, 4, 2, 3, 1, 2, 6, 5  $\Rightarrow$  LOSE!  
STOP
- 3, 5, 6, 4, 2, 3, 1, 2, 6, 5  $\Rightarrow$  WIN!  
STOP

This boils down to finding a threshold from which we STOP as soon as a 6 is thrown.

## Dice Game (Solution)

$$P[\text{obtain exactly one 6 in last } k \text{ throws}] = \binom{k}{1} \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{k-1} = \frac{k}{6} \cdot \left(\frac{5}{6}\right)^{k-1}$$

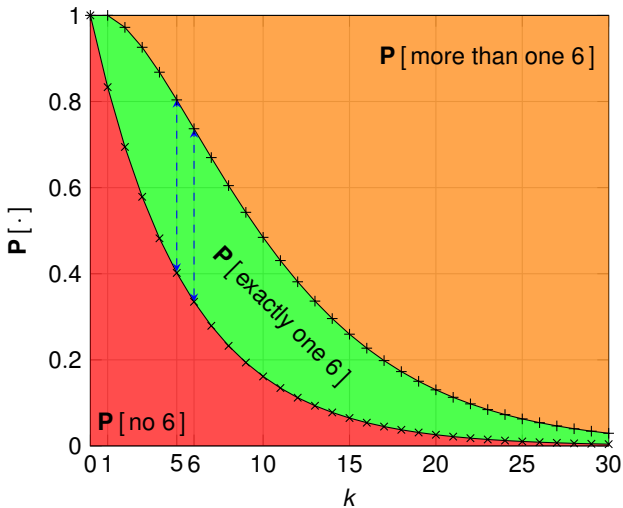


We obtain a **unimodal** distribution

- This is **maximised** for  $k = 6$  (or  $k = 5$ )  $\Rightarrow$  **best strategy**: wait until we have 6 (5) throws left, and then STOP at the first 6
- **Probability of success** is:

$$\left(\frac{5}{6}\right)^5 \approx 0.40.$$

## Illustration of the Three Possibilities



# Outline

---

Stopping Problem 1: Dice Game

Stopping Problem 2: The Secretary Problem

A Generalisation: The Odds Algorithm (non-examinable)

The End...

## The Secretary Problem

### The Problem

- We are interviewing  $n$  candidates for **one job** in a sequential, **random** order
- A candidate must be accepted (STOP) or rejected **immediately** after the interview and cannot be recalled
- **Goal:** **maximise** the probability of hiring the **best** candidate

also known as **marriage problem** (Kepler 1613),  
**hiring problem** or **best choice problem**.

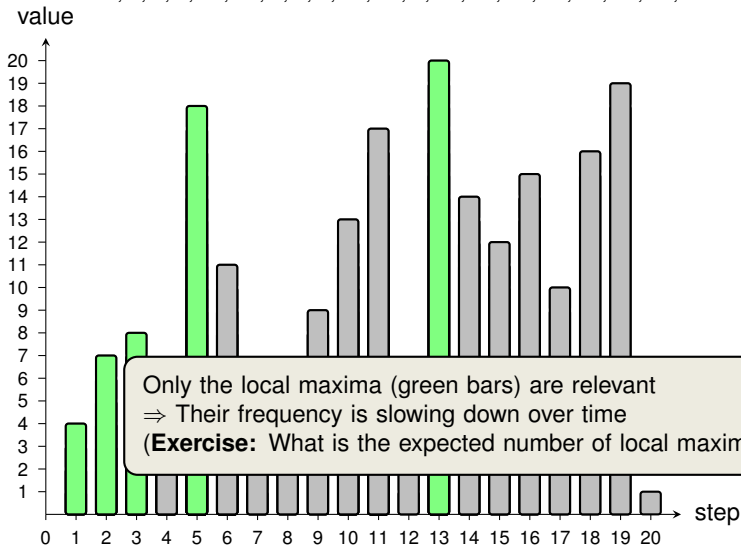
### Further Remarks

- After seeing candidate  $i$ , we only know the **relative order** among the first  $i$  candidates.
- ⇒ For our problem we may as well assume that the only information we have when interviewing candidate  $i$  is whether that candidate is best among  $\{1, \dots, i\}$  or not.

## Illustration ( $n = 20$ )

unknown permutation:

4, 7, 8, 6, 18, 11, 3, 5, 9, 13, 17, 2, 20, 14, 12, 15, 10, 16, 19, 1.



## Two Basic Strategies

### Naive Approach

- Always pick the **first** (or any other) candidate
- Probability for success is:

$$P[\text{hire best candidate}] = \frac{1}{n}.$$

A typical **exploration-exploitation** based strategy.

### Smarter Approach

- Reject the first  $n/2$  candidates, then take the first candidate that is better than the first  $n/2$  (if none is taken before, take last candidate)

How good is this approach?

### Example 1

Find a lower bound on the success probability of the refined approach (picking the first candidate better than the first  $n/2$ ).

Answer

- Probability for success is:

$$\mathbf{P} [ \text{hire best candidate} ]$$

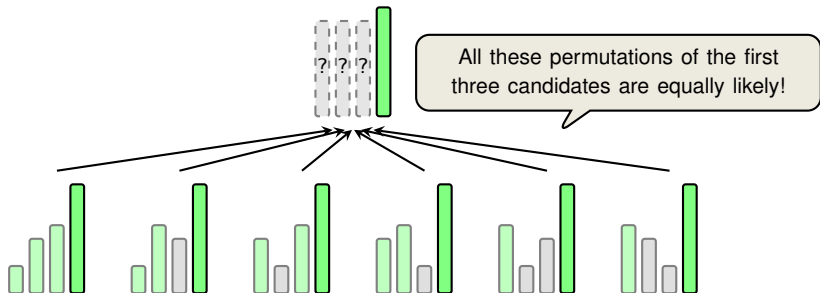
$$\geq \mathbf{P} [ \text{best in 2nd half} \cap \text{second best in 1st half} ]$$

$$= \mathbf{P} [ \text{best in 2nd half} ] \cdot \mathbf{P} [ \text{second best in 1st half} \mid \text{best in 2nd half} ]$$

$$= \frac{n/2}{n} \cdot \frac{n/2}{n-1} > \frac{1}{4}.$$

## Finding the Optimal Strategy (1/2)

- **Observation 1:** At interview  $i$ , it only matters if current candidate is best so far (i.e., no benefit in counting how many “best-so-far” candidates we had).



- **Observation 2:** If at interview  $i$ , the best strategy is to accept the candidate (if it is “best-so-far”), then the same holds for interview  $i + 1$

Optimal Strategy

- **Explore** but reject the first  $x - 1$  candidates
- **Accept** first candidate  $i \geq x$  which is better than **all candidates before**

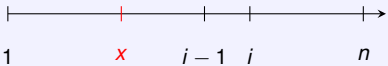
## Example 2

Find  $x$  which maximises the probability of hiring the best candidate.

Answer

- First compute **success probability** for any  $x \in \{1, \dots, n\}$ , and then **optimise**:

**P** [hire best candidate]

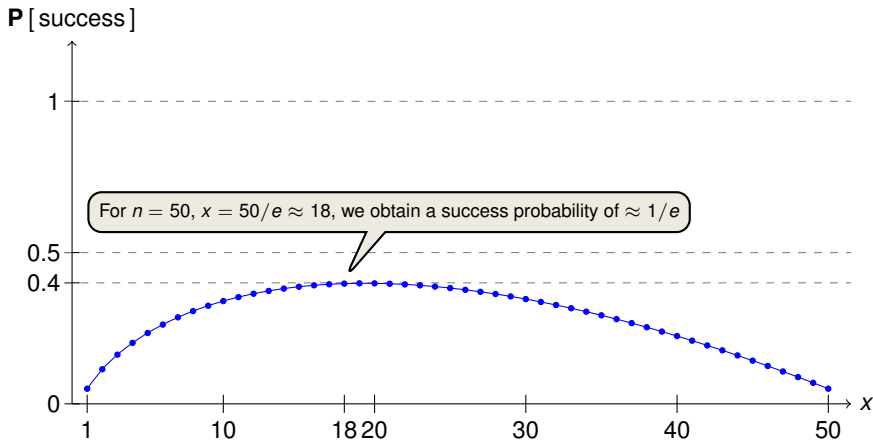
$$\begin{aligned}
 &= \sum_{i=1}^n \mathbf{P}[\text{hire candidate } i \cap \text{candidate } i \text{ is best}] \\
 &= \sum_{i=x}^n \mathbf{P}[\text{hire candidate } i \cap \text{candidate } i \text{ is best}] \\
 &= \sum_{i=x}^n \mathbf{P}[\text{hire candidate } i \mid \text{candidate } i \text{ is best}] \cdot \mathbf{P}[\text{candidate } i \text{ is best}] \\
 &= \frac{1}{n} \cdot \sum_{i=x}^n \mathbf{P}[\text{second best of first } i \text{ candidates is in the first } x-1 \mid \text{candidate } i \text{ is best}] \\
 &= \frac{1}{n} \cdot \sum_{i=x}^n \frac{x-1}{i-1} = \frac{x-1}{n} \cdot \sum_{i=x}^n \frac{1}{i-1}
 \end{aligned}$$


The diagram shows a horizontal line representing the range from 1 to n. There are vertical tick marks at 1, x, i-1, i, and n. The tick mark at x is highlighted in red, and the label 'x' is written in red below it. The labels 'i-1' and 'i' are placed below their respective tick marks.

$$\Rightarrow \sum_{i=x}^n \frac{1}{i-1} \approx \ln(n/x) \Rightarrow \text{maximum success probability for } x = \frac{1}{e} \cdot n.$$

## Probability for Success (Illustration)

Suppose  $n = 50$ :



## Another Variant of the Secretary Problem

“The Postdoc Variant of the Secretary Problem” (Vanderbei’80)

- same setup as in the secretary problem before
- **difference**: we want to pick the **second-best** (“the best [postdoc] is going to Harvard”)
- Success probability of the optimal strategy is:

$$\frac{0.25n^2}{n(n-1)} \xrightarrow{n \rightarrow \infty} \frac{1}{4}$$

- Thus it is **easier** to pick the best than the second-best(!)

# Outline

---

Stopping Problem 1: Dice Game

Stopping Problem 2: The Secretary Problem

A Generalisation: The Odds Algorithm (non-examinable)

The End...

## Details of the Odds Algorithm

- Let  $I_1, I_2, \dots, I_n$  be a sequence of **independent** indicators and let  $p_j = \mathbf{E}[I_j]$
- Let  $r_j := \frac{p_j}{1-p_j}$  (**the odds**) and  $p_j \in (0, 1)$  for all  $j = 1, 2, \dots, n$

### Example 3

What is the probability that after trial  $k$ , there is exactly one success?

Answer

$$\mathbf{P} \left[ \sum_{j=k}^n I_j = 1 \right] = \sum_{j=k}^n p_j \cdot \prod_{k \leq j \leq n, j \neq i} (1 - p_i) = \sum_{j=k}^n r_j \cdot \left( \prod_{i=k}^n (1 - p_i) \right)$$

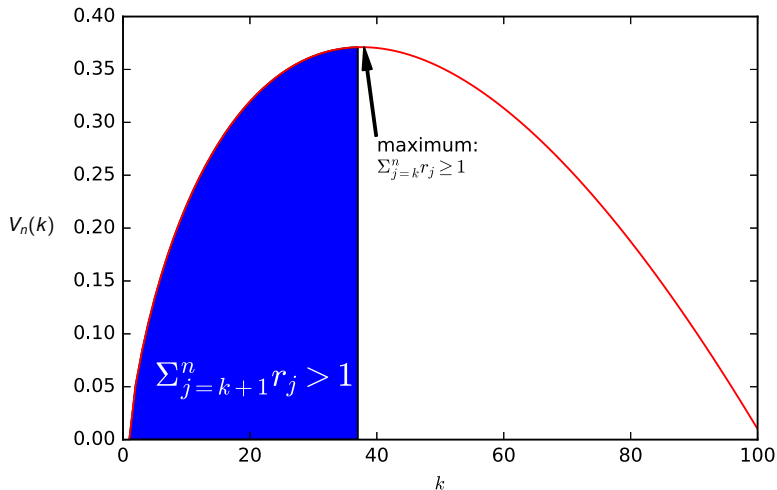
- One can prove that  $\mathbf{P} \left[ \sum_{j=k}^n I_j = 1 \right]$  is **unimodal** in  $k \Rightarrow$  there is an **ideal point** from which on we should **STOP at the first success!**

**Odds Algorithm** ("Sum the Odds to One and Stop", F. Thomas Bruss, 2000)

- Let  $k^*$  be the largest  $k$  such that  $\sum_{j=k}^n r_j \geq 1$
- Ignore** everything before the  $k^*$ -th trial, then **STOP** at the **first** success.

- The **success probability** is  $\sum_{j=k^*}^n r_j \cdot \left( \prod_{i=k^*}^n (1 - p_i) \right)$ .
- This algorithm always executes the **optimal strategy!**

## Illustration of the probability of having the last success ( $n = 100$ )



Source: Group Fibonacci

Use the **Odds Algorithm** to analyse the **Secretary Problem**.

Answer

- Let  $I_j = 1$  if and only if secretary  $j$  is the best secretary so far.
- The  $I_j$ 's are **independent** (this is an exercise on the exercise sheet)
- Then:

$$p_j = \mathbf{P} [ I_j = 1 ] = \frac{1}{j}$$

$$r_j = \frac{p_j}{1 - p_j} = \frac{1/j}{(j-1)/j} = \frac{1}{j-1}$$

- Largest  $k$  for which  $\sum_{j=k}^n \frac{1}{j-1} \geq 1$  is  $k = 1/e \cdot n$
- Probability for success:

$$\mathbf{P} \left[ \sum_{j=k}^n I_j = 1 \right] = \sum_{j=k}^n r_j \cdot \left( \prod_{i=k}^n (1 - p_i) \right)$$

$$= \sum_{j=k}^n \frac{1}{j-1} \cdot \left( \prod_{i=k}^n \frac{i-1}{i} \right)$$

$$= \sum_{j=k}^n \frac{1}{j-1} \cdot \frac{k-1}{n} \approx \frac{1}{e}.$$

We re-derived the solution of the **secretary problem** as a special case!

## Extra Slide: Why the Odd's Algorithm works

Define **two quantities**: For any  $1 \leq k \leq n$ ,

$$Q_k := \prod_{i=k}^n (1 - p_i), \quad \text{so} \quad Q_k = Q_{k+1} \cdot (1 - p_k),$$
$$R_k := \sum_{i=k}^n r_i = \sum_{i=k}^n \frac{p_i}{1 - p_i}, \quad \text{so} \quad R_k = R_{k+1} + \frac{p_k}{1 - p_k}.$$

The **success probability** (a.k.a. value or reward) for stopping after  $k$  rounds is

$$V_k := R_k \cdot Q_k.$$

The **difference** between two **consecutive values** is

$$\begin{aligned} V_{k+1} - V_k &= R_{k+1} Q_{k+1} - R_k Q_k \\ &= R_{k+1} Q_{k+1} - \left( R_{k+1} + \frac{p_k}{1 - p_k} \right) \cdot Q_{k+1} (1 - p_k) \\ &= R_{k+1} Q_{k+1} - ((1 - p_k) R_{k+1} + p_k) \cdot Q_{k+1} \\ &= Q_{k+1} \cdot p_k \cdot (R_{k+1} - 1), \end{aligned}$$

as the other two factors are  $\geq 0$ , the sign of  $V_{k+1} - V_k$  is determined by  $R_{k+1}$  (which is non-increasing in  $k$ )  $\Rightarrow$   **$V_k$  is unimodal between  $1 \leq k \leq n$ .**

# Outline

---

Stopping Problem 1: Dice Game

Stopping Problem 2: The Secretary Problem

A Generalisation: The Odds Algorithm (non-examinable)

The End...

- **Part I: Introduction to Probability**

- **Lecture 1:** Conditional probabilities and Bayes' theorem

- **Part II: Random Variables**

- **Lecture 2:** Random variables, probability mass function, expectation
- **Lecture 3:** Expectation properties, variance, discrete distributions
- **Lecture 4:** More discrete distributions: Poisson, Geometric, Negative Binomial
- **Lecture 5:** Continuous random variables
- **Lecture 6:** Marginals and Joint Distributions
- **Lecture 7:** Independence, Covariance and Correlation

- **Part III: Moments and Limit Theorems**

- **Lecture 8:** Basic Inequalities and Law of Large Numbers
- **Lecture 9:** Central Limit Theorem

- **Part IV: Applications and Statistics**

- **Lecture 10:** Estimators (Part I)
- **Lecture 11:** Estimators (Part II)
- **Lecture 12:** Online Algorithms

## List of Distributions

---

### Very Important:

- Bernoulli, Binomial, Poisson
- (Continuous) Uniform, Normal, Exponential

### (Somewhat Less) Important:

- Geometric, Negative Binomial, Hypergeometric, Discrete Uniform

### Not used or not defined in this course (and thus not examinable):

- Cauchy, Gamma, bivariate Normal
- Beta

Thank you and Best Wishes for the Exam!