

# Probability (Math 323) winter 2019

Prof: David Wolfson  
Yimeng Hu

## Contents

|   |           |
|---|-----------|
| <b>1 Lec 01, Jan 08</b>                                       | <b>2</b>  |
| 1.1 Introduction . . . . .                                    | 2         |
| 1.2 Basic Set Algebra . . . . .                               | 2         |
| <b>2 Lec 02, Jan 10</b>                                       | <b>3</b>  |
| 2.1 Experiment . . . . .                                      | 3         |
| 2.2 Sample Space . . . . .                                    | 3         |
| 2.3 Kolmogorov Axioms . . . . .                               | 4         |
| <b>3 Lec 03, Jan 15</b>                                       | <b>5</b>  |
| 3.1 The 5 theorems . . . . .                                  | 5         |
| 3.2 Tools for calculating probability . . . . .               | 6         |
| <b>4 Lec 04, Jan 17</b>                                       | <b>8</b>  |
| 4.1 Counting Rule . . . . .                                   | 8         |
| 4.2 The Birthday Problem . . . . .                            | 8         |
| 4.3 The Fish in the Lake Problem . . . . .                    | 9         |
| 4.4 Capture & Recapture Problem . . . . .                     | 9         |
| <b>5 Lec 05, Jan 22</b>                                       | <b>11</b> |
| 5.1 Conditional Probability . . . . .                         | 11        |
| 5.2 Multiplication Rule for Conditional Probability . . . . . | 12        |
| <b>6 Lec 06, Jan 24</b>                                       | <b>13</b> |
| 6.1 Conditioning Backwards . . . . .                          | 13        |
| 6.2 The Law of Total Probability . . . . .                    | 13        |
| 6.3 Baye's Theorem . . . . .                                  | 14        |
| <b>7 Lec 07, Jan 29</b>                                       | <b>16</b> |
| 7.1 Statistical Independence* . . . . .                       | 16        |
| <b>8 Lec 08, Jan 31</b>                                       | <b>18</b> |
| 8.1 The Role of Independence . . . . .                        | 18        |
| <b>9 Lec 09, Feb 05</b>                                       | <b>21</b> |
| 9.1 Random Variable . . . . .                                 | 21        |
| 9.1.1 Specification of R.V . . . . .                          | 22        |

|  |           |
|--|-----------|
| <b>10 Lec 10, Feb 07</b>   | <b>23</b> |
| 10.1 Cumulative Distribution Function . . . . .  | 23        |
| 10.2 Discrete R.V and CDF . . . . .  | 24        |
| <b>11 Lec 11, Feb 12</b>   | <b>27</b> |
| 11.1 Some named discrete distribution . . . . .  | 27        |
| 11.1.1 The discrete uniform distribution . . . . .   | 27        |
| 11.1.2 The Bernouilli Distribution . . . . .   | 27        |
| 11.2 The Binomial Distribution . . . . .   | 27        |
| <b>12 Lec 12, Feb 14</b>   | <b>29</b> |
| 12.1 Binomial Setup . . . . .  | 29        |
| 12.2 <b>The geometric distribution</b> . . . . .   | 30        |
| 12.3 Negative Binomial Distribution . . . . .  | 31        |
| <b>13 Lec 13, Feb 19</b>   | <b>32</b> |
| 13.1 Poisson Distribution . . . . .  | 32        |
| 13.2 The Hypergeometric Distribution . . . . .   | 33        |
| 13.3 Mathematical Expectation & Variance . . . . .   | 34        |
| <b>14 Lec 14, Feb 26</b>   | <b>35</b> |
| 14.1 Expectation . . . . .   | 35        |
| 14.2 Variance . . . . .  | 36        |
| <b>15 Lec 15, Mar 01</b>   | <b>38</b> |
| <b>16 Lec 16, Mar 12</b>   | <b>39</b> |
| 16.1 Continuous Probability Distribution . . . . .   | 39        |
| <b>17 Lec 17, Mar 14</b>   | <b>43</b> |
| 17.1 Expected Value and Variance of Cont. R.V . . . . .                                    | 43        |
| 17.2 Some Named Continuous Distribution . . . . .  | 44        |
| 17.2.1 The Continuous Uniform Distribution . . . . .                                       | 44        |
| 17.2.2 The Gamma Distribution . . . . .  | 45        |
| <b>18 Lec 18, Mar 19</b>   | <b>48</b> |
| 18.1 Another special case of Gamma Distribution . . . . .                                  | 49        |
| 18.2 The Normal(Gaussian) Distribution . . . . .   | 49        |
| <b>19 Lec 19, Mar 21</b>   | <b>52</b> |
| 19.1 Transformations of Random Variables . . . . .   | 52        |
| <b>20 Lec 20, Mar 26</b>   | <b>55</b> |
| 20.1 Application to the simulation of random variables from a given distribution . . . . . | 56        |
| 20.2 A Final Important Comment in Transformation . . . . .                                 | 56        |
| 20.3 Moment Generating Function . . . . .  | 57        |
| <b>21 Lec 21, Mar 28</b>   | <b>58</b> |
| 21.1 Multivariate Distribution . . . . .   | 60        |

|   |           |
|---|-----------|
| <b>22 Lec 22, Apr 2</b>                   | <b>61</b> |
| 22.1 Joint p.d.f and p.m.f . . . . .      | 61        |
| 22.2 Expectation and Variance . . . . .   | 63        |
| 22.2.1 Application . . . . .              | 63        |
| 22.3 Conditional Distribution . . . . .   | 63        |
| <b>23 Lec 23, Apr 4</b>                   | <b>64</b> |
| 23.1 Conditional Distribution . . . . .   | 64        |
| 23.2 Covariance and Correlation . . . . . | 64        |

# 1 Lec 01, Jan 08

## 1.1 Introduction

Why study probability?

- as a discipline in its own right.
- as a part of mathematics/ applied mathematics
- Most importantly, as a tool for statistical inference

**The meaning of probability** (i.e when we say "the probability" of an event A is 2/3, what do we mean)

**Ex** a box has 6 Red and 4 green marbles. Draw a marble at random from the box. What is the probability that the marble is red?

1. If we say  $P(\text{red}) = \frac{6}{10}$  what do we mean by this statement?

**Sol.**

*We cannot simply define "the probability" of getting a red as  $\lim_{N \rightarrow \infty} \frac{\# \text{ of red}}{\# \text{ of trials}}$  since we do not know if this limit will exist and be unique for every sequence of trials. So instead in the 1930. The Great Russian mathematician A.N. Komolgorov proposed **3 axioms/assumptions** that probability should satisfies and then developed a theory of probability from these.*

**Note.** As a consequence of The law of large number, we interpret probability as a limiting relative frequency. However it has nothing to do with relative frequencies.

2. How did you arrive at this answer?

**Sol.**

*In order to arrive at the answer  $\frac{6}{10}$ , we need to use the Komolgorov axioms and any theorem that follows from that to prove that this is indeed correct.*

## 1.2 Basic Set Algebra

1.  $A = \{\omega : \omega \in A\}$  where A is an event (a set) consist of elementary outcomes w.
2.  $A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$       **note:** "and"  $\Rightarrow$  intersection
3.  $A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$       **note:** "or"  $\Rightarrow$  union.
4.  $A \subset B := \omega \in A \Rightarrow \omega \in B$
5. All discussion take place in the context of the **universal set**  $S$
6.  $A^c := \{\omega \in S : \omega \notin A\}$       **note:** It is sometimes easier to first find  $P(A^c)$
7.  $A \cap B = \emptyset \Rightarrow$  A and B disjoint or mutually exclusive
8. De Morgan's Law:
  - a)  $(A \cap B)^c = A^c \cup B^c$
  - b)  $(A \cup B)^c = A^c \cap B^c$

## 2 Lec 02, Jan 10

### 2.1 Experiment

**DEF 2.1.** An **experiment** is defined informally as the performance of some actions

**DEF 2.2.** An **Random Experiment** is one for which the outcome are not known in advance. i.e there is uncertainty in the outcome that will be observed. (Once thet experiment has been conducted though you may not know the outcome, there is nothing random about the outcome)

**Ex 2.1.** Toss a coin twice and observe the outcome The pre-experiment outcomes are random / uncertain

**Ex 2.2.** Take 60 subjects who will undergo surgery for a certain disorder before we observe their  $\underbrace{\text{time to recovery}}_{\text{outcome}}$ , are random/uncertain

**Ex 2.3.** Toss a coin until you observe the first head. Let the trial at which this happens be the outcome of interest. This outcome is random before you start tossing.

### 2.2 Sample Space

**DEF 2.3.** The set of all possible outcome of an experiment is called the **Sample Space** ( $S$ ) of the experiment. We denote each outcome as  $\omega$ , an elemetary outcome.

**Note.** a sample space mainly depend on how you define your outcomes.

**Ex 2.4.** Draw a marble at random from 6 Red and 4 Green.

1. If order **does not** matter, Let  $w_1 :=$  event which a red marble is drawn,  $w_2 :=$  event which a green marble is drawn. then the sample space is as following:  $S = \{w_1, w_2\}$

2. If order matters, number the marbles WLOG  $\underbrace{\{1, \dots, 6\}}_{\text{Red}}, \underbrace{\{7, \dots, 10\}}_{\text{Green}}$  Let  $w_i :=$  event which marble  $i$  is drawn for  $i = 1, \dots, 10$  Then the sample space is:  $S = \{w_1, \dots, w_{10}\}$

**Ex 2.5.** Suppose there are  $n$  people in a room, ask these people when their birthday are. Let  $S$  be the set of outcomes that we could get at the completion of our experiment. Then  $S$  could be defined as:

$$S = \underbrace{\{\{Jan1, \dots, Jan1\}, \dots, \{Jan1, \dots, Jan1\}\}}_{w_1} \cup \underbrace{\{\{Dec31, \dots, Dec31\}, \dots, \{Dec31, \dots, Dec31\}\}}_{w_n}$$

**Note.** Here the elementary outcomes are  $n$ -dim vectors

**Ex 2.6.** Toss a coin until you observe the 1<sup>st</sup> head. Let  $n$  be the trial number at which this occurs. Then  $S = \{w_1, \dots, w_n\} = \{1, \dots, n\}$ .  $S$  is an example of sample space with a countably many numbers of possible outcomes.

**Ex 2.7.** Suppose that you measure the height of a dam every July 1<sup>st</sup>. The set of possible heights might be  $S = \{[0, 20]\}$  where 20 is the height of the dam wall in meters. Here  $S$  is an uncountably infinite set. **Note:** In real life no such sample space exists.

## 2.3 Kolmogorov Axioms

**DEF** 2.4. Any subset  $E \subset S$  is defined as an **event**. The empty set  $\emptyset \subset S$  is also an event.

**DEF** 2.5. a function  $P()$  is a set function on the subset of  $S$  if  $P(A)$  is a real number for every subset of  $S$ . Let  $S$  be a sample space. Then  $P()$  is called a probability measure if  $P$  is a real valued set function on  $S$  s.t

1.  $\forall E \subset S, P(E) \geq 0$
2.  $P(S) = 1$
3. Let  $E_1, E_2, \dots$  be any countable connection of events such that they are mutually exclusive,  
i.e  $E_i \cap E_j = \emptyset \quad \forall i \neq j$ . Then  $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

**Note.** From these 3 axioms, we developpe the entire theory of probability including the Law of Large numbers which allows us to interpret probability as a limiting relative frequency. We shall state and prove 5 theorem that will be useful for solving word problems.

### 3 Lec 03, Jan 15

#### 3.1 The 5 theorems

**Thm 3.1.** 1. For any event  $A$ ,  $P(A^c) = 1 - P(A)$

PROOF:

$$A \cup A^c = S, \Rightarrow P(A \cup A^c) = P(S) = 1 \quad (\text{Ax 2})$$

$$A \cap A^c = \emptyset \Rightarrow P(A \cup A^c) = P(A) + P(A^c) \quad (\text{Ax 3})$$

$$\Rightarrow P(A^c) = 1 - P(A)$$

□

$$2. P(\emptyset) = 0$$

$$3. P(A \cap B^c) = P(A) - P(A \cap B)$$

PROOF: trick: Try to write unions as disjoint unions and apply Ax 3

$$\begin{aligned} A = (A \cap B) \cup (A \cap B^c) &\Rightarrow P(A) = P(A \cap B) + P(A \cap B^c) \\ &\Rightarrow P(A \cap B^c) = P(A) - P(A \cap B) \end{aligned} \quad (\text{Ax 3})$$

□

$$4. A \subset B \Rightarrow P(A) \leq P(B)$$

PROOF:

$$\begin{aligned} B = A \cup (B \cap A^c), A \cap (B \cap A^c) &= \emptyset \\ &\Rightarrow P(B) = P(A) + P(B \cap A^c) \\ &\geq P(A) \end{aligned} \quad (\text{Ax 1})$$

□

$$5. \text{ For any two events } A \text{ and } B, P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Note.** if  $A \cap B = \emptyset$ , then  $P(A \cup B) = P(A) + P(B)$

PROOF:

$$\begin{aligned} A \cup B &= \underbrace{(A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)}_{\text{mutually exclusive}} \\ \Rightarrow P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \quad (\text{Ax 3}) \\ &= P(A) - P(A \cap B) + P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) \quad (\text{Thm 3}) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

□

**Cor 3.1.1.** For any event  $A$ ,  $0 \leq P(A) \leq 1$

**Note.** Do not use tree diagram to present your answer. Start by defining the simplest possible event then construct more complicated event by using set operations.

"Either ... or...", "At least"  $\Rightarrow \cup$ , "and"  $\Rightarrow \cap$ , "Not"  $\Rightarrow$  complement.  
Then apply axiom or theorem.

**Ex 3.1.** Suppose that it's known that 20% of people smoke and that 1% of old people will develop lung cancer. Suppose that the probability of someone will either smoke or develop lung cancer is 0.205. Let  $A$ :=the event of someone smokes.

$B$ := the event of someone has cancer

Then  $P(A) = 20\%$ ,  $P(B) = 1\%$  and  $P(A \cup B) = 0.205$

1. Find the proportion of people who smoke and develop lung cancer.

**Sol.** WTS  $P(A \cap B)$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) && (\text{thm 5}) \\ \Rightarrow P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &= 0.2 + 0.01 - 0.205 \\ &= 0.005 \end{aligned}$$

2. What is the probability that someone does not smoke but have lung cancer.

**Sol.** WTS  $P(A^c \cap B)$

$$\begin{aligned} P(A^c \cap B) &= P(B) - P(A \cap B) && (\text{thm 3}) \\ &= 0.01 - 0.005 \\ &= 0.005 \end{aligned}$$

3. What is the probability that someone smokes but does not have lung cancer.

**Sol.** WTS  $P(A \cap B^c)$

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) && (\text{thm 3}) \\ &= 0.2 - 0.005 \\ &= 0.195 \end{aligned}$$

4. What is the probability that someone neither smoke nor have lung cancer.

**Sol.** WTS  $P(A^c \cap B^c)$

$$\begin{aligned} P(A^c \cap B^c) &= P((A \cup B)^c) && (\text{De Morgan's Law}) \\ &= 1 - P(A \cup B) && (\text{thm 1}) \\ &= 1 - 0.205 \\ &= 0.195 \end{aligned}$$

### 3.2 Tools for calculating probability

**Thm 3.2.** Let  $S$  be a finit sample space with  $N$  equally likely outcomes. Let  $E$  be any event in  $S$ . Then

$$P(E) = \frac{|E|}{N} = \frac{\# \text{ of outcomes in } E}{\text{Tot. possible outcomes}}$$

**Note.** The calculation of a probability can be then reduced to a counting problem

PROOF: write the event E as the union of the elementary outcomes i.e

$$\begin{aligned} E &= \bigcup_{w_i \in E} w_i \Rightarrow P(E) = P\left(\bigcup_{w_i \in E} w_i\right) = \sum_{w_i \in E} P(W_i) \\ P(S) &= \sum_{i=1}^N P(w_i) = 1 \Rightarrow P(w_i) = \frac{1}{N} \quad \forall i = 1, \dots, N \end{aligned} \quad (\text{Ax 2})$$

$$\begin{aligned} \text{Hence: } P(E) &= \sum_{i=w_i \in E} \frac{1}{N} \\ &= \frac{1}{N} \sum_{i=w_i \in E} 1 \\ &= \frac{|E|}{N} \end{aligned}$$

□

## 4 Lec 04, Jan 17

we want a sample space with equally likely outcomes.

**Recall.**  $S = \{1, 2, \dots, 10\}$ ,  $R = \{1, 2, 3, 4, 5, 6\}$ ,  $G = \{7, 8, 9, 10\}$

All of these outcomes are reasonably equally likely. There are  $N=10$  such outcomes.

Therefore by the above thm(last class),  $P(R) = \frac{\# \text{ of ways to get a red marble}}{\text{tot. no. of possible outcomes}} = \frac{6}{10}$ .

**Note.** Now although the above thm(last class) is easy to understand, the counting can sometimes be very difficult. it is useful to have some counting tools

### 4.1 Counting Rule

1. If you have a set of  $n$  distinct object, then the number of ways to order the objects in  $n!$
2. If you have a set of  $n$  distinct object, then the number of ways to draw  $r$  object from the set, and the order is unimportant, sampling **without** replacement is denoted as " $n$  choose  $r$ ".

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}, \quad \text{note: } 0! = 1 \text{ by def}$$

3. If we have a set of  $n$  distinct objects. The number of ways to draw  $r$  objects from these  $n$  and the order does matter, sampling **without** replacement is denoted by " $n$  permutation  $r$ "

$$P(n, r) = \frac{n!}{(n-r)!}$$

4. **Multiplication Rule** (Sausage Rule)

Suppose that you have  $k$  set of  $n_1, n_2, \dots, n_k$  distinct objects respectively. The number of ways to form a set by selecting one object from each set is given by

$$n_1 * n_2 * \dots * n_k$$

### 4.2 The Birthday Problem

Suppose there are  $n$  people in a room. What is the probability that at least two have the same birthday?

PROOF:

Suppose that there are 356 possbile birthday

Let  $E :=$  event that at least two people have the same birthday.

It is easier to compute  $P(E^c) = P(\text{no two have the same birthday})$ , then

$$P(E) = 1 - P(E^c) \quad (\text{by Thm 1})$$

The sample space here is  $S = \{(Jan 1, \dots, Jan 1), \dots, (Dec 31, \dots, Dec 31)\}$

First we assume that all of these outcomes are equally likely. There are finitely many of them. Therefore

$$\begin{aligned} P(E^c) &= \frac{\# \text{ of ways that } E^c \text{ can occur}}{\text{Tot.no.of outcomes in } S} \\ &= \frac{P(365,n)}{365^n} \\ &= \frac{365 * 364 * \dots * (365 - n + 1)}{365^n} \end{aligned}$$

Hence

$$P(E) = 1 - \frac{365 * 364 * \dots * (365 - n + 1)}{365^n}$$

□

### 4.3 The Fish in the Lake Problem

Suppose a lake has  $N$  fish in it, of which  $a$  are tagged and  $N-a$  are untagged. If you draw a sample  $n$  fish from the lake, sampling **without** replacement, What is the probability of getting  $x$  tagged fish in my sample?

PROOF:

We want a sample space with equally likely outcomes.

Start by numbering the fish from 1 to  $N$ .

We will suppose that the fish with numbers  $1, \dots, a$  correspond to those with tags and the remaining  $N-a$  numbers to the untagged fish.

An outcome for our experiment is defined to a set of  $n$  integers selected from the integers  $1, \dots, N$

The order is considered unimportant and assume that all sets of  $n$  numbers are equally likely.

Hence we can use our thm to solve the problem.

Let  $E$ :=event that there are  $x$  tagged in sample

$$P(E) = \frac{\text{number of ways to get } x \text{ tagged}}{\text{total number of poss. outcomes}}$$

We have

Tot. number of possible outcomes = number of ways to draw  $n$  integers from a set of  $N$  distinct integers

$$= \binom{N}{n} \quad (\text{counting rule 2.})$$

Now use **Multiplication Rule**

each sub-sausage contains  $x \leq a$  integers selected from integer  $1, \dots, a$

each sub-sausage contains  $n-x$  integers selected from integers  $(a+1), \dots, N$ . So we must count the number of objects in each of these two sausages,  $n_1, n_2$ . say for the number of ways to get  $x$  tagged fish is  $= n_1 * x * n_2$  we have

$$n_1 = \binom{a}{x}, \quad n_2 = \binom{N-a}{n-x}$$

finally

$$P(x \text{ tagged fish out of } n) = \frac{\binom{a}{x} * \binom{N-a}{n-x}}{\binom{N}{n}}$$

□

### 4.4 Capture & Recapture Problem

Have  $N$  fish in the lake **Capture Phase**

1. Remove and tag  $a$  fish.
2. Return the fish to the lake

**Recapture Phase**

1. capture  $n$  fish
2. Count how many tagged fish in the recaptured sample

$$P(X = x) = \frac{\binom{a}{x} * \binom{N-a}{n-x}}{\binom{N}{n}}$$

N is unknown, But

$$\begin{aligned}\frac{a}{N} &\approx \frac{x}{n} \\ N &= \frac{a * n}{x}\end{aligned}$$

However, in real world, the captured face tends to be harder to be recaptured

## 5 Lec 05, Jan 22

### 5.1 Conditional Probability

**Idea:** Sometimes, knowing that an event A has occurred influences the probability that the event B will occur.

**Ex 5.1.** In our marble problem, The probability of getting a red marble on the second of two draws (**without replacement**) knowing that we got a red on the first, is different from simply the probability of getting a red on the second draw. Argument

$$\begin{aligned} P(R_2) &= P[(R_2 \cap G_1) \cup (R_2 \cap R_1)] \\ &= P[R_2 \cap G_1] + P[R_2 \cap R_1] \end{aligned}$$

which is easy to see,  $P(R_2)$  is different from  $P(R_2 \text{ knowing } R_1)$

We therefore feel justified in formally defining the notion of "Conditional Probability"

**DEF 5.1.** Let A and B be two events such that  $P(A) \neq 0$ , then we define the probability of B given A as follows:

$$P[B \text{ given } A] := P[B | A] = \frac{P[A \cap B]}{P(A)}$$

**Note:** the RHS is the ratio of two probability and we have defined probability (The 3 Axioms)

**Note.**

1. we need to check whether conditional probability satisfies the 3 Axioms.

$$(a) P(B | A) \geq 0$$

PROOF:

$$P[B | A] = \frac{\underbrace{P[A \cap B]}_{\geq 0}}{\underbrace{P(A)}_{\geq 0}}$$

Hence true □

$$(b) P[S | A] = 1$$

PROOF:

$$P[S | A] = \frac{P[A \cap S]}{P(A)} = \frac{P(A)}{P(A)} = 1$$

Hence true □

$$(c) P[\bigcup B_i | A] = \sum_{i=1}^{\infty} P[B_i | A] \quad \text{where } B_i \cap B_j = \emptyset \quad \forall i \neq j$$

PROOF: (exercise) □

It then follows that the 5 theorem also go through for conditional Probability

$$2. P(A | B) = \frac{P(A \cap B)}{P(B)} \quad \text{if } P(B) \neq 0$$

The definition of conditional probability leads to a fundamental theorem that allows us to sometimes find the probability of an intersection

## 5.2 Multiplication Rule for Conditional Probability

Follows immediately from the definition of conditional probability

$$\begin{aligned} P(B \cap A) &= P(B | A) * P(A) \\ &= P(A | B) * P(B) \end{aligned}$$

The hope is that when you are required to find  $P[A \cap B]$ , you know either  $P(A)$  or  $P(B)$  and one of the conditional probability.

In word problems,

”of those that”  $\Rightarrow$  Conditional Probability

Do not confuse ”and” with ”given that”

**Ex 5.2.** We have two inspectors for items coming off an assembly. The proportion of items that are declared non-defective by the first inspector is 0.90. Of those items that are declared non-defective by the first inspector, 0.95 are declared non-defective by inspector 2. What is the probability that an item is declared non-defective by both inspectors.

**Sol.**

Let  $ND_i$  ( $i=1,2$ ) := event non-defective for each of the inspectors resp.

**WTS:**  $P[ND_1 \cap ND_2]$

Given:  $P[ND_1] = 0.90$  and  $P[ND_2 | ND_1] = 0.95$

$$\begin{aligned} P[ND_2 \cap ND_1] &= P[ND_2 | ND_1] * P[ND_1] \\ &= 0.90 * 0.95 \end{aligned}$$

□

**\*Extension:** Let  $A_1, \dots, A_n$  be any sequence of events. Then

$$\begin{aligned} P[A_1 \cap \dots \cap A_n] &= P[A_n | A_1 \cap \dots \cap A_{n-1}] * P[A_1 \cap \dots \cap A_{n-1}] \\ &= P[A_n | A_1 \cap \dots \cap A_{n-1}] * P[A_{n-1} | A_1 \cap \dots \cap A_{n-2}] * P[A_1 \cap \dots \cap A_{n-2}] \\ &= \dots \\ &= \prod_{i=1}^n P[A_i | A_1 \cap \dots \cap A_{i-1}] * P(A_0) \quad (\text{i}=1, \dots, n) \end{aligned}$$

**Ex 5.3.**

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_3 | A_1 \cap A_2)P(A_1 \cap A_2) \\ &= P(A_3 | A_1 \cap A_2)P(A_2 | A_1)P(A_1) \end{aligned}$$

The process of repeatedly conditioning starting with the last event is called the process of **conditioning backwards**.

When you are required to find the probability of the intersection of several events, think of conditioning backwards

## 6 Lec 06, Jan 24

### 6.1 Conditioning Backwards

**DEF** 6.1. *The process of repeatedly conditioning starting with the last event is called the process of **conditioning backwards**.*

When you are required to find the probability of the intersection of several events, think of conditioning backwards

**Ex** 6.1. *The Marble Problem*

- Suppose that you draw 2 marbles **without** replacement. What is the probability that the second marble drawn is green?

**Sol.** Conditioning Backwards.

$$G_2 = (R_1 \cap G_2) \cup (G_1 \cap G_2)$$

Implies

$$\begin{aligned} P(G_2) &= P((R_1 \cap G_2)) + P((G_1 \cap G_2)) \\ &= P(G_2|R_1)P(R_1) + P(G_2|G_1)P(G_1) \\ &= \frac{4}{9} * \frac{6}{10} + \frac{3}{9} * \frac{4}{10} \end{aligned} \tag{Ax 3}$$

- Suppose that you draw 5 marbles. What is the probability that you will get the sequence  $R_1, R_2, G_3, G_4, R_5$ .

**Sol.** Conditioning Backwards

$$\begin{aligned} P(R_1 \cap R_2 \cap G_3 \cap G_4 \cap R_5) &= P(R_5|R_1 \cap R_2 \cap G_3 \cap G_4) * P(G_4|R_1 \cap R_2 \cap G_3) \\ &\quad * (P(G_3|R_1 \cap R_2) * (P(R_2|R_1) * P(R_1)) \end{aligned}$$

Hence

$$P(R_1 \cap R_2 \cap G_3 \cap G_4 \cap R_5) = \frac{4}{6} * \frac{3}{7} * \frac{4}{8} * \frac{5}{9} * \frac{6}{10}$$

The following theorem on conditional probability are fundamental

### 6.2 The Law of Total Probability

**Thm** 6.1.

Let  $A$  be any event, let  $B_1, B_2, \dots$  be  $m$  events that satisfy the following

- $B_i \cap B_j = \emptyset \quad \forall i \neq j$
- $\bigcup_{i=1}^m B_i = S$  we call  $\{B_1, B_2, \dots\}$  a partition of  $S$

Then  $P(A) = \sum_{i=1}^m P(A|B_i)P(B_i)$

PROOF: (Of theorem)

known  $A = \bigcup_{i=1}^m \underbrace{(A \cap B_i)}_{\text{all disjoint}}$  Hence

$$\begin{aligned} P(A) &= \sum_{i=1}^m P(A \cap B_i) \\ &= \sum_{i=1}^m P(A|B_i)P(B_i) \end{aligned} \tag{Ax 3}$$

□

**Note.** Maybe  $A$  is complicated and it is difficult to find its probability directly or the given information does not provide  $P(A)$  directly.

The hope is that we can find  $P(A|B_i)$  easily or that they come with the provided information and that we know  $P(B_i)$ .

In word problem, the clue to use the Law of Total Probability is that you are given a bunch of conditional probability and the probability  $P(B_i)$  and you are asked to find  $P(A)$ .

### 6.3 Baye's Theorem

**Thm 6.2.** Let  $A$  and  $B_1, B_2, \dots$  be defined exactly as in the Law of Probability. Then we can write

$$P(B_k|A) = \frac{P(A|B_k) * P(B_k)}{\sum_{i=1}^m P(A|B_i)P(B_i)} \quad (k = 1, 2, \dots, m)$$

PROOF:

$$\begin{aligned} P(B_k|A) &= \frac{P(B_k \cap A)}{P(A)} \\ &= \frac{\overbrace{P(A|B_k)P(B_k)}^{\text{mult. rule}}}{\underbrace{\sum_{i=1}^m P(A|B_i)P(B_i)}_{\text{Law of tot. Prob.}}} \end{aligned}$$

□

**Note.** Mathematically, Baye's Theorem allows you to reverse one or more given conditional probability. In word problem, the clue to use the Baye's thm is that you are required to reverse one or more conditional probability statement.

**Ex 6.2.** Suppose that there is a diagnostic test for breast cancer and that in a certain population  $\frac{5}{1000}$  women have breast cancer. Known that the test has the following properties:

1. if a woman has breast cancer, the test will be positive 95% of the time.
2. if a woman does not have breast cancer, the test will be negative 95% of the time.

Question is

- i) What proportion of women will test positive?

**Sol.**

Let  $Pos :=$  event that a test is positive

Let  $Neg :=$  event that a test is negative

Let  $Bc :=$  event that a woman has breast cancer

Let  $Bc^c :=$  event that a woman does not have breast cancer

Known

$$P(Pos | Bc) = 95\%$$

$$P(Pos | Bc^c) = 1 - 95\% = 0.05$$

$$P(Neg | Bc^c) = 95\%$$

$$P(BC) = 0.005$$

$$P(BC^c) = 1 - 0.005$$

We have that  $BC$  and  $BC^c$  are disjoint and  $BC \cup BC^c = S$ .

Therefore by the Law of Tot. Prob.

$$\begin{aligned} P(\text{Pos}) &= P(Pos | Bc) * P(BC) + P(Pos | Bc^c) * P(BC^c) \\ &= 0.95 * 0.005 + 0.05 * (1 - 0.005) \\ &= 0.054 \end{aligned}$$

- ii) If a woman tests positive, what is the probability that she has breast cancer?

**Sol.** WTS  $P(Bc | Pos)$ .

**note:** we are required to reverse the  $P(Pos | Bc)$  (Baye's)

$$\begin{aligned} P(Bc | Pos) &= \frac{P(Pos | Bc) * P(Bc)}{P(Pos | Bc) * P(Bc) + P(Pos | Bc^c) * P(Bc^c)} \\ &= \frac{0.95 * 0.005}{\underbrace{0.054}_{\text{from part i}}} \\ &= 0.087 \end{aligned}$$

Some comments on Baye's Theorem and diagnostic tests:

In the world of diagnostic tests  $P(Pos | Disease)$  is called the sensitivity of the test  $P(neg | Disease^c)$  is called the specificity.  $P(Disease)$  is called the disease prevalence and  $P(Disease | Pos)$  is called the positive predictive value of the test. Note that pos predictive value depends on the sensitivity, specificity and prevalence of the disease.

## 7 Lec 07, Jan 29

Some comments on Baye's Theorem and diagnostic tests:

In the world of diagnostic tests  $P(\text{Pos} \mid \text{Disease})$ := the sensitivity of the test

$P(\text{neg} \mid \text{Disease}^c)$ := the specificity.

$P(\text{Disease})$ := the disease prevalence and

$P(\text{Disease} \mid \text{Pos})$ := the positive predictive value of the test.

Note that pos predictive value depends on the sensitivity, specificity and pre. value of the disease.

### 7.1 Statistical Independence\*

**Idea:** sometimes the occurrence of an event A does **NOT** influence the probability that an event B will occur.

**Ex 7.1.** In the marble problem, suppose that you draw two marbles **with** replacement. What is the probability that the second marble is drawn is red given the first is green. (i.e what is  $P[R_2|G_1]$ ).

Clearly, we obtained a green on the first draw has no impact on the probability of a red on the second draw, since the box was returned to its original composition. This idea leads to a definition of the independence of two events A and B

#### DEF 7.1.

1. We say that the events A and B are independent  $\iff P[B|A] = P(B)$

**Note:** This definition is while intuitive, it is not easily extended to the notion of independence of more than two events.

2. (non-intuitive but extendable) The events A and B are said to be independent,  
 $\iff P(A \cap B) = P(A) * P(B)$ .

**Thm 7.1.** A and B are independent according to DEF 1  $\iff$  they are independent according to DEF 2.  
(i.e The above two definition are equivalent.)

PROOF:

( $\Rightarrow$ ) Assume A and B are independent according to DEF 1, WTS they are also independent according to DEF2.

Then  $P(B \mid A) = P(B)$ .

Hence by the Multiplication rule

$$\begin{aligned} P(A \cap B) &= P(B|A) * P(A) \\ &= P(B) * P(A) \end{aligned}$$

( $\Leftarrow$ ) Trivial

□

**DEF 7.2.** The events  $A_1, A_2, \dots, A_n$  are said to be mutually independent if  $P[A_{i_1} \cap \dots \cap A_{i_k}] = P[A_{i_1}] * \dots * P[A_{i_k}]$  for all subset of  $A_{i_1} \dots A_{i_k}$  selected from  $A_1$  to  $A_n$ .

#### Ex 7.2.

1.  $P(A \cap B \cap C) = P(A)P(B)P(C)$
2.  $P(A \cap B) = P(A)P(B)$  etc...

We also define independence for an infinite sequence of events  $A_1, A_2, \dots, A_3$

**DEF 7.3.** *The event  $A_1, A_2, \dots$  are independent if and only if every finite set of  $A_i$  is independent according to the previous definition of independence.*

**Note.**

1. Events  $A_1, A_2, \dots, A_n$  are said to be pairwise independent if  $P(A_i \cap A_j) = P(A_i)P(A_j)$  for all  $i \neq j$ . It can be shown that pairwise independence does **NOT** imply mutual independence.

2. we write  $A \perp\!\!\!\perp B$  to denote  $A$  is independent of  $B$

3. If  $A \perp\!\!\!\perp B$  then  $B \perp\!\!\!\perp A$  and vice-versa.

Further, we have if  $A, B$  and  $C$  are mutually independent then

- (a)  $A^c \perp\!\!\!\perp B$
- (b)  $(A \cup B)^c \perp\!\!\!\perp C$
- (c)  $(A^c \cap B^c) \perp\!\!\!\perp C$
- (d)  $(A \cup C)^c \perp\!\!\!\perp B$
- (e)  $A^c \perp\!\!\!\perp B^c$  etc...

PROOF: WTS  $P(A^c \cap B^c) = P(A^c)P(B^c)$

$$\begin{aligned} P(A^c \cap B^c) &= P((A \cup B)^c) && (\text{De Morgan}) \\ &= 1 - P(A \cup B) && (\text{Thm 1}) \\ &= 1 - [P(A) + P(B) - P(A \cap B)] && (\text{Thm 5}) \\ &= 1 - P(A) - P(B) + P(A \cap B) \end{aligned}$$

Now notice

$$\begin{aligned} P(A^c)P(B^c) &= (1 - P(A))(1 - P(B)) \\ &= 1 - P(A) - P(B) + P(A)P(B) \end{aligned}$$

Hence

$$P(A^c \cap B^c) = P(A^c)P(B^c)$$

□

4. In fact the following is true

Suppose that you have two sets  $A_1, A_2, \dots, A_n$  and  $B_1, B_2, \dots, B_n$ . We say that the set of  $A$  is independent of set  $B$  if the probability of the intersection of every set of  $A$  with the intersection of every set of  $B$  is the product of the intersection of the set of  $A$  and the set of  $B$ .

$$\text{E.g } P[(A_3 \cap A_5 \cap A_6) \cap (B_1 \cap B_2)] = P[A_3 \cap A_5 \cap A_6] * P[B_1 \cap B_2]$$

## 8 Lec 08, Jan 31

### 8.1 The Role of Independence

1. The most important Role of Independence in probability is the following:

If you can assume independence base on your knowledge of the substantive area and/or the way the experiment was carried out. Then subsequent probability calculations often become a lot easier than if you cannot make this assumptions.

This is so since the probability of intersection. becomes product of probabilities rather than product of conditional probabilities requiring knowledge of these conditional probabilities.

2. Second, we may want to decide whether events are independent (i.e This may be the goal.)

For instance, one may wish to know, whether recovery time from abdominal surgery is independent of the temperature of the operating room.

3. The relation between disjointness and independence:

It turns out that these notions are completely different. Disjointness is entirely a set property whereas independence depends on how probabilities are assigned to these events. The following theorem says it all:

**Thm 8.1.** Suppose that  $A$  and  $B$  are disjoint, then  $A$  and  $B$  are independent only if either  $P(A) = 0$ , or  $P(B) = 0$ .

PROOF:

$$\begin{aligned} A, B \text{ disjoint} &\Rightarrow A \cap B = \emptyset \\ &\Rightarrow P(A \cap B) = P(\emptyset) = 0 \end{aligned}$$

Now if  $A \perp\!\!\!\perp B$ , then we must have:

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \Rightarrow P(A)P(B) = 0 \\ &\Rightarrow P(A) = 0, \text{ or } P(B) = 0 \end{aligned}$$

□

**Note.** sometimes you will be required to find  $P(A \cup B)$ , you have from theorem 5, that  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ,

To deal with  $P(A \cap B)$ :

- (a)  $A \cap B = \emptyset \Rightarrow P(A \cap B) = 0$
- (b)  $A \perp\!\!\!\perp B \Rightarrow P(A \cap B) = P(A)P(B)$
- (c)  $A$  and  $B$  are dependent, then  $P(A \cap B) = P(B | A)P(A)$  (note, this is always true)

**Ex 8.1.** (on how independence can be used)**Note.** Preliminary note

- when sampling **without** replacement, the outcomes in the sequence of draws are dependent.

Thus if you have a box of 10 items of which 4 are defective and you remove 2 without replacement, whether or not you observe a defective on the second draw will depend on what was removed on the first draw. However, if "the box" from which we sample is very large, relative to the size of the sample, we may regard the outcomes of our draws, as being roughly independent.

Suppose that in a very large city, 20% of people have a certain genetic mutation. If 10 people are examined

Clearly we are sampling without replacement (by design).

Let the outcome on trial  $i$ , be  $M_i :=$  there is a mutation for subjects  $i$ .

$M_i^c :=$  there is no mutation for subjects  $i$ .

Let  $X :=$  the number of mutations in these ten trials. (random variable.)

- What is the probability that exactly 2 will have the mutation?

**Sol.** WTS  $P(X = 2)$ . Proceed as following

- Find the probability of a specific configuration of mutations and non-mutations in 10 trials that result in 2 mutations out of 10.

Consider the configuration

$$\begin{aligned} (M_1, M_2, M_3^c, \dots, M_{10}^c) &= M_1 \cap M_2 \cap M_3^c \cap \dots \cap M_{10}^c \\ P(M_1, M_2, M_3^c, \dots, M_{10}^c) &= P(M_1)P(M_2)P(M_3^c)\dots P(M_{10}^c) \\ &\quad (\text{Rough independenc (large 'city' - small sample)}) \\ &= 0.2 * 0.2 * 0.8 * \dots * 0.8 \end{aligned}$$

Now notice, all configurations which result in exactly two mutations will have probability  $(0.2)^2(0.8)^8$

- Sum up the probability of all such configurations

$$\begin{aligned} P[X = 2] &= P\left[\bigcup_{k=1}^{\text{all config.}} \text{configuration } k \text{ with 2 mutations}\right] \\ &= \sum_{k=1}^{\text{all config.}} P\left[\text{configuration } k \text{ with 2 mutations}\right] \quad (\text{Ax 3. (config. disjoint)}) \\ &= (0.2)^2(0.8)^8 \sum_{k=1}^{\text{all config.}} 1 \\ &= (0.2)^2(0.8)^8 \binom{10}{2} \end{aligned}$$

2. What is the probability that at least 2 will have the mutation?

**Sol.** WTS  $P[x \geq 2]$

$$\begin{aligned} P[x \geq 2] &= \sum_{k=2}^{10} P[x = k] \\ &= \sum_{k=2}^{10} (0.2)^k (1 - 0.2)^{10-k} \binom{10}{k} \end{aligned}$$

or computationally simpler:

$$\begin{aligned} &= 1 - P[x < 2] \\ &= 1 - [P[x = 0] + P[x = 1]] \\ &= 1 - \left[ \binom{10}{0} (0.2)^0 (0.8)^{10-0} + \binom{10}{1} (0.2)^1 (0.8)^{10-1} \right] \end{aligned}$$

**Note.** If the city has 10 million people with 20% having the mutation then the exact answer will be

$$P[x = 2] = \frac{\binom{2\text{million}}{2} \binom{8\text{million}}{8}}{\binom{10\text{million}}{10}}$$

If you are not giving the actual size of the city, but told it's large. You have to use the independent approximation.

## 9 Lec 09, Feb 05

### 9.1 Random Variable

**Idea:** Often not interested in the outcome of a random experiment but rather in real numbers that we can associate with each of these outcomes.

**Ex 9.1.** A pathologist looking a grid on slide, where each rectangle contains either a red blood cell or a white blood cell. The pathologist is not really interested in the SEQ  $R_1, R_2, R_3, W_4, R_5, W_6\dots$  rather she is interested in the number of red/white cells on the slide. This idea lead to the definition of the random variables that assigns a real number to each possible outcome of the experiment.

**DEF 9.1.** We call the function  $X(w)$  that assigns a real number of  $X(w)$  to every elementary outcome  $\omega \in E$  a real value random variable. That is  $X : w \rightarrow \mathbb{R}$

**Note.**

1.  $X$  is just a function. However, the argument and random outcomes of an experiment.  
i.e The outcomes are uncertain or random BEFORE the experiment is carried out. Hence, the value of the function  $X$  are uncertain or random prior to the experiment. This is why we called the function,  $X$ , a random variable.
2. Always denote random variable by capital letters (e.g  $X, Y, Z$  etc)
3. Once the experiment has been carried out and an  $\omega \in S$  has been observed  $X(w)$  is a real number and it's no longer random. i.e a realized or post experiment value of a random value is not random.

**DEF 9.2.** Random variable are broadly classified into one of 2 types

#### 1. Discrete R.V

By definition, a random variable is said to discrete if it can assume at most a countably infinite number of distinct value.

**Ex 9.2.**

- (a) The number of times we get heads in 3 tosses of a coin (here could assume the values of  $X$  to be 0, 1, 2, 3)
- (b) The times between arrivals of a bus we assume to the nearest minute.
- (c) The number of exploratory wells that are drilled before oil is first struck (here, could assume the value of  $X$  to be 1, 2, 3...)

#### 2. Continuous R.V

There are random variables that can be assumed any real value in some interval.

**Ex 9.3.**

- (a) The exact time between the arrival of buses at a bus stop
- (b) The exact time to recovery following surgery.
- (c) The exact blood pressure of a patient.

### 9.1.1 Specification of R.V

1. In classical mathematics we specify a function by giving its domain and the values of the function at these value of X. Since a R.V X has a inherent uncertainties (prior to the experiment) when specify X, we need to specify the probability that X will assume its various values(the rough idea)
2. The following is one way to specify a R.V in the case where R.V X is either discrete or continuous

**DEF** 9.3. Let  $X$  be a random variable. Let the real value function denoted by  $F_x$  is defined as follows:

$$F_x(X) \stackrel{\text{def.}}{=} P[X \leq x] \quad \forall -\infty < x < +\infty$$

$F$  := the cumulative distribution function (cdf) or distribution function (df) of the random variable  $X$ .

**Note.**

1. when we write  $P[X \leq x]$  we mean  $P[\omega : X(\omega) \leq x] = P[X^{-1}(-\infty, x)]$   
i.e  $P[X \leq x] = P$  (all  $w$  in  $S$  that are mapped by  $X$  into the interval  $(-\infty, x]$ )
2. When we have specified the cdf of a random variable  $X$  it turns out that we have (in theory) uniquely specified  $P[x \in A]$  for any  $A$  a subset in the real line. i.e we can uniquely specify the probability of all events associated with  $X$  once we know the cdf  $F_x$  that gives us only the probability of special types of sets,  $(-\infty, x]$
3. We call the set of probability  $\{P[x \in A] : A \subset \mathbb{R} \text{ 7y7is an event}\}$  The probability distribution of Random variable  $X$ .

## 10 Lec 10, Feb 07

### 10.1 Cumulative Distribution Function

**Summary:**

1. The probability distribution of a random variable is the specification of the probability that  $x \in A$  for every subset  $A$  of the real line.  
i.e this is  $\{P(x \in A) : A \text{ is a subset of } \mathbb{R}\}$
2. Amazingly, in order to uniquely specify the probability of all subsets, it's enough to specify the probability of just a small collection of subsets.  
i.e it's enough to specify the probability of intervals of the type  $(-\infty, x)$
3. The theorem telling us this = caratheodory extension theorem.
4. The real-value function of  $X$ , which gives  $P(X \leq x) = P[X \in (-\infty, x]]$  is called the cumulative distribution function (cdf) and is denoted by  $F_x()$  Thus  $F_x(X) \stackrel{\text{def}}{=} P(X \leq x)$
5. Thus the cumulative distribution function uniquely determines the probability distribution of a random variables.

**Note.**

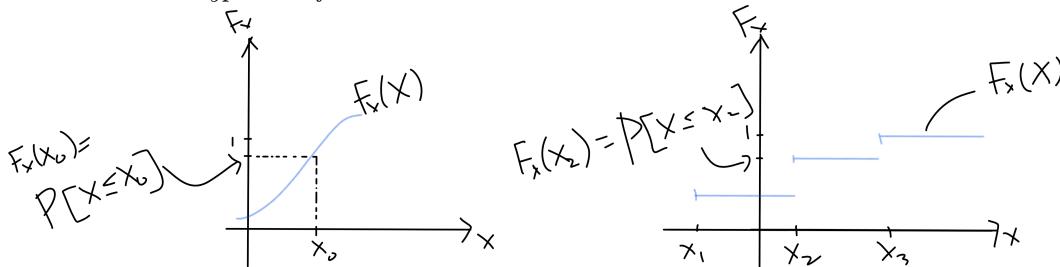
1. **The CDF's Properties:**

- (a)  $F_x(X)$  is non-decreasing of a function of  $X$ . i.e if  $x < y$  then  $F_x(X) \leq F_y(X)$
- (b)  $F_x(X)$  is continuous from the right i.e  $\lim_{y \neq x} F_x(Y) = F_x(X)$  (basically  $F_x(X)$  does not jump as you approach  $x$  from above)
- (c)  $\lim_{x \rightarrow \infty} F_x(X) = 1$  and  $\lim_{x \rightarrow -\infty} F_x(X) = 0$  in brief  $F_x(\infty) = 1$  and  $F_x(-\infty) = 0$

2. If  $F_x$  is continuous both from the left and right (i.e  $F$  is "plain" continuous, we say the random variable  $X$  is continuous)

3. We have  $P(a \leq X \leq b) = F_x(b) - F_x(a)$  for all  $a < b$

4. Here are some typical cdf's:



5. So far, our discussion has applied to random variables in general. We now examine **discrete** random variables and **continuous** random variables separately

- A random variable is discrete if its cdf is a step function with jumps of values  $X_1, X_2, \dots$  which can be assumed by the random variable  $X$ . Thus a discrete random variable has cdf with at most countably many jumps.

- The size of each jump at  $x_0$  is obtained as

$$F_X(x_0) - \lim_{y \rightarrow x_0^-} F_X(y) = F_X(x_0) - F_X(x_0^-)$$

Think roughly, that the size of the jump at  $x_0$  is  $F_X(x_0) - F_X(y)$  has a  $y$  very close to  $x_0^-$

- $P[X = x_0] = F_X(x_0) = F_X(x_0^-) = \text{size of the jump in } F_X \text{ at } x_0$

If  $x_0$  is not one of the value of  $X$ , then the size of the jump will be 0. This correspond to  $P[X = x_0] = 0$  for such  $x_0$

If  $x_0$  coincides with one of the  $x_i$  then the size of the jump at that  $x_i$  gives  $P[X = x_i]$

There is apart from the cdf another way to specify the probability distribution of a discrete random variable.

## 10.2 Discrete R.V and CDF

**DEF 10.1.** If  $X$  is a **discrete** random variable, then the real valued function of  $X$ , defined as follows, is called the **probability function** (probability mass function of the random variable)

- Let  $P_X(x) = P[X = x]$  for all  $x$  in range of  $X$

Then we will call  $P_X(x)$  the probability function of the random variable  $X$

i.e  $P_X(x)$  gives the probability of all  $x$ . we usually specify  $P_X(x)$  only for those  $x$  for which the probability is  $> 0$  and it's assumed that all other  $x$  have prob. = 0

Now it is easy to show that  $P_X()$  uniquely determines the probability distribution of a discrete random variable. The following theorem tells us there is 1-1 correspondence between the probability function  $P_X$  and the cdf  $F_X$  of a discrete random variable.

Hence  $P_X$  uniquely determines the probability distribution of  $X$

**Thm 10.1.**

1.  $P_X$  determines  $F_X$

2.  $F_X$  determines  $P_X$

PROOF: 1. Let  $P_X$  be given, then

$$\begin{aligned} F_X(x_0) &= P[X \leq x_0] \\ &= \sum_{\text{all } x \leq x_0} P[X = x] \\ &= \sum_{x \leq x_0} P_X(x) \end{aligned} \tag{Ax 3}$$

i.e  $P_X$  determines  $F_X(x_0)$  for all  $x_0$

2. suppose  $F_X$  is given

Let  $X_0$  be some arbitrary value in the range of  $X$ .

Then we have

$$\begin{aligned} P_X(x_0) &= F_X(x_0) - F_X(x_0^-) \\ &= F_X(x_0) - \lim_{y \rightarrow x_0^-} F_X(y) \end{aligned}$$

□

**Ex 10.1.** Suppose that if a flight is cancelled the airline loses 5000\$, if the flight leaves late (more than 1/2 hours late) it loses 2000\$, if it leaves on time, it makes a profit of 10000\$ if it leaves late but less than 1/2 hour, it still makes 10000\$

$$P[C] = 0.05$$

$$P[> 1/2 \text{ hour late}] = 0.1$$

$$P[\text{on time}] = 0.7$$

The questions are:

1. Find the probability function of the random variable that represent the gain on a flight
2. Find the cdf of this random variable.

**Sol.** The elementary outcomes are:

$$\omega_1 = \text{cancelled}.$$

$$\omega_2 = 1/2 \text{ hour late}.$$

$$\omega_3 = \text{on time}.$$

$$\omega_4 = \text{less than } 1/2 \text{ hour late}.$$

Let  $X = \text{the gain for a flight}$ .

We have:

$$X(\omega_1) = -5000$$

$$X(\omega_2) = -2000$$

$$X(\omega_3) = 10000$$

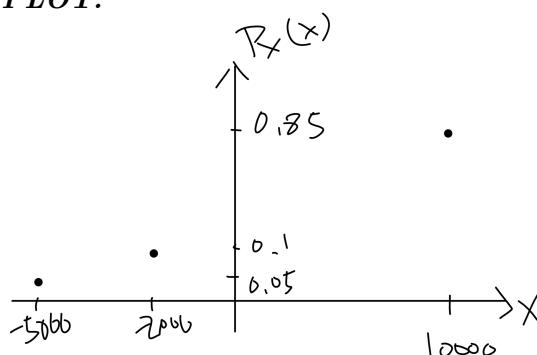
$$X(\omega_4) = 10000$$

1. We want  $P_X(x) = P[X = x]$  for all  $x$  that  $X$  can assume

$$\begin{aligned} P[X = -5000] &= P[\omega : X(\omega) = -5000] \\ &= P[C] = 0.05 \\ P[X = -2000] &= P[\omega : X(\omega) = -2000] \\ &= P[> 1/2 \text{ late}] \\ &= 0.1 \\ P[X = 10000] &= P[\omega : X(\omega) = 10000] \\ &= P[\text{on time}] + P[\leq 1/2 \text{ late}] \\ &= 0.7 + 1 - P[> 1/2 \text{ late}] - P[\text{on time}] - P[C] \\ &= 0.7 + 1 - 0.05 - 0.1 - 0.7 = 0.85 \end{aligned}$$

Hence  $P_X(x)$  is specified and

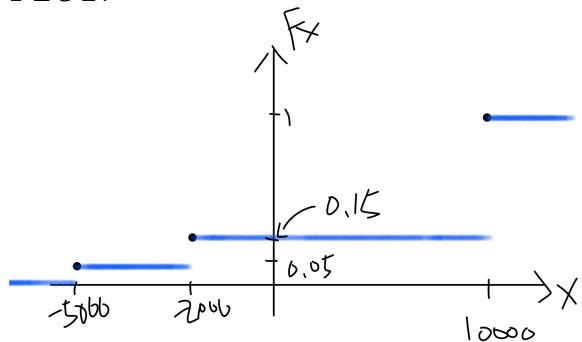
**PLOT:**



2. To find the cdf  $F_X(x)$ , we need to give  $P[X \leq x]$  for all  $-\infty < x < \infty$ .

$$\begin{aligned}
 F_X(x) &= 0 && (\text{for } x < -5000) \\
 F_X(x) &= P[X = -5000] = 0.05 && (\text{for } -5000 \leq x \leq -2000) \\
 F_X(x) &= P[-\infty < X < -5000] + P[-5000 \leq X < -2000] \\
 &\quad + P[-2000 \leq X < 10000] \\
 &= 0 + 0.05 + 0.1 \\
 &= 0.15 && (\text{for } -2000 \leq x \leq 10000) \\
 F_X(x) &= 0 + 0.05 + 0.15 + 0.85 = 1 && (\text{for } 10000 \leq x < \infty)
 \end{aligned}$$

**PLOT:**



## 11 Lec 11, Feb 12

### 11.1 Some named discrete distribution

Some discrete distribution arise so frequently that they are given special names. Here are some of them:

#### 11.1.1 The discrete uniform distribution

**DEF** 11.1. *The random variable  $X$  is said to be a discrete uniform distribution on the real numbers  $a_1, a_2, \dots, a_N$  if  $P_X[x] = P[X = x] = 1/N$  for all  $x = a_1, a_2, \dots, a_N$  (i.e  $P_X(a_i) = 1/N$  for all  $i = 1, 2, \dots, N$ )*

**Note.**

1. A probability function  $P_X(x)$  must satisfy the following 2 conditions:

$$(a) P_X(x) \geq 0 \quad \forall x$$

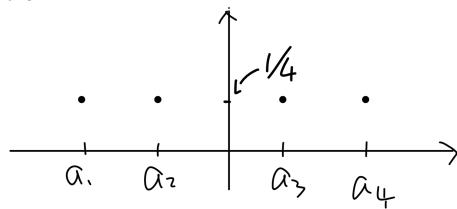
$$(b) \sum_{\text{all } x} P_X(x) = 1 \quad (P(S) = 1)$$

2. Reason for the name, "uniform":

All possible values of  $X$  are equally probable, with probability  $1/N$

3. The  $a_i$  need not be positive or integer valued.

i.e



4. The discrete uniform distribution is used to model what we understand to be "complete randomness"

#### 11.1.2 The Bernoulli Distribution

**DEF** 11.2. *The random variable  $X$  is said to have a Bernoulli distribution with parameter  $p$ , if*

1.  $P_X(x) = p$  when  $x = 1$
2.  $P_X(x) = 1 - p = q$  when  $x = 0$

**Note.**

1. We can write the probability function compactly as  $P_X(x) = p^x(1 - p)^{1-x}$  for  $x = 0, 1$
2. The Bernoullie Distribution is mostly used as a building block for random variables that can be regarded as sums.

## 11.2 The Binomial Distribution

**DEF** 11.3. *A random variable  $X$  is said to have a binomial distribution with parameters  $n$  and  $p$  if  $P_X(x) = P[X = x] = \binom{n}{x} p^x (1 - p)^{n-x}$  for  $x = 0, 1, \dots, n$  and  $0 < p < 1$*

**Note.**

1. Clearly  $P_X(x) \geq 0$  but  $\sum_{\text{all } x} P_X(x) \stackrel{?}{=} 1$

PROOF:

consider

$$\sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = [p + (1-p)]^n \quad (\text{by the binomial theorem})$$

(to be continued...) □

## 12 Lec 12, Feb 14

Midterm include materials up until and include this class How does the binomial distribution most often arrive?

### 12.1 Binomial Setup

1. We assume that we have  $n$  independent Bernoulli trials  
(A Bernoulli trial is the one that can result in exactly one of two possible outcome. By convention, we call these outcomes "success" and "Failure")
2. We assume that if  $S_i$  is the event success at trial  $i$  and  $F_i$  as failure, then  $P[S_i] = p$  for all  $i = 1, \dots, n$  and  $P[F_i] = 1 - P[S_i] = 1-p$ .

Note that  $p$  is assumed to be constant from trial to trial

Let  $X$  = the number of observed successes in these  $n$  trials

**Thm 12.1.**  $P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$  for  $x = 0, 1, \dots, n$

PROOF:

- Done already in the context of the gene mutation problem when sampling from a very population (small sample)
- **Idea:** Find the probability of a specific configuration with  $x$  successes and  $n-x$  failures. This is

$$p^x (1-p)^{n-x}$$

Every such configuration was this probability  $P[X=x] = \text{sum of the probabilities of all such configurations}$

THere are  $\binom{n}{x}$  terms in this sum

- Hence in a word problem if
  - (a) you have trials that result in one of two possible outcomes with equal probability
  - (b) you can assume that these trials are independent, then  $X$  = number of successes in these  $n$ -trials has a binomial distribution

□

3. Be careful of immediately invoking the binomial distribution because you see 2 types of outcomes on each trial. The trials may be dependent.

**Ex 12.1.** Suppose that patients undergoing a certain treatment can survive  $> 5$  years or  $\leq 5$  years. Suppose proportion of patients that survive more than 5 years is 0.80

If 30 patients are to receive this treatment.

What is the probability that at least 3 will survive more than 5 years.

**Sol.**

- Let  $X$  = number of patients who survive more than 5 years  
Assume that patients survive independent of each other  
Let  $S_i$ := event that patient  $i$  survives more than 5 years  $i = 1, 2, \dots, 30$   
We can assume the binomial setup with  $n = 30$  and  $P[S_i] = 0.80$

- Hence

$$\begin{aligned} P[X \geq 3] &= \sum_{x=3}^{30} P[X = x] \\ &= \sum_{x=3}^{30} \binom{30}{x} (0.8)^x (1 - 0.8)^{30-x} \end{aligned}$$

OR

$$\begin{aligned} P[X \geq 3] &= 1 - P[X < 3] \\ &= 1 - P[X \leq 2] \\ &= 1 - \sum_{x=0}^2 \binom{30}{x} (0.8)^x (0.2)^{30-x} \end{aligned}$$

**Q.E.D**

## 12.2 The geometric distribution

**DEF** **12.1.** A random variables  $X$  is said to have geometric distribution with parameter  $0 < p < 1$  if  $P_X(x) = P[X = x] = (1 - p)^{x-1}p$   $x = 1, 2, \dots$

How does the geometric distribution arrive ?

**Answer:**

- Imagine a sequence of independent Bernoulli trial, with  $P[S_i] = p$  for  $i=1,2,\dots$
- Let  $x =$  the trial at which the first success is observed

**Claim:**  $P[X = x] = (1 - p)^{x-1}p$  for  $x = 1, 2, \dots$

(i.e the geometric distribution is the distribution of the "time" to success in a sequence of independent Bernoulli Trials with constant probability of success)

PROOF:

- The event  $\{X = x\}$  is equivalent to the event  $\{F_1 \cap F_2 \cap \dots \cap F_{x-1} \cap S_x\}$

$$\begin{aligned} P[X = x] &= P[F_1 \cap F_2 \cap \dots \cap F_{x-1} \cap S_x] \\ &= P[F_1] * P[F_2] * \dots * P[F_{x-1}] * P[S_x] \quad (\text{By the assumption that the trials are indep.}) \\ &= (1 - p)^{x-1}p \quad (\text{for all } x = 1, 2, \dots) \end{aligned}$$

□

**Note.**

- The Binomial gives the number of successes in a fixed number of trials  $n$ .
- The geometric gives the trial at which the first success occurs.
- The number of trials is not fixed

### 12.3 Negative Binomial Distribution

**DEF** 12.2.

- A random variable  $X$  has a negative binomial distribution with parameter  $K$  and  $p$  if  $X$  describes the trial at which the  $k$ th success occurs in a sequence by independent Bernoulli trials with constant probability of success  $p$ .

Let's find  $P_X(x) = P[X = x]$  for  $x = k, k+1, \dots$

$$\begin{aligned}
 P[X = x] &= P[\text{getting } k-1 \text{ successes in } x-1 \text{ trials} \cap \text{getting a success on trial } x] \\
 &= P[A \cap B] \\
 &= P[B | A] * P(A) \\
 &= P(B) * P(A) && (\text{by independence}) \\
 &= \underbrace{P(B)}_{p} \binom{x-1}{k-1} p^{k-1} (1-p)^{x-1-(k-1)} && (\text{for } x = k, k+1, \dots) \\
 &= \binom{x-1}{k-1} p^k (1-p)^{x-k} && (\text{for } x = k, k+1, \dots)
 \end{aligned}$$

## 13 Lec 13, Feb 19

### 13.1 Poisson Distribution

**DEF** 13.1. *The random variable  $X$  is said to have a Poisson distribution with parameter  $\lambda > 0$  if and only if*

$$P_X(x) = P[X = x] = \frac{\lambda^x e^{-\lambda}}{x!} \quad \forall x = 0, 1, \dots$$

Notice

$$\begin{aligned} \sum_{x=0}^{\infty} P_X(x) &= \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \underbrace{\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}}_{e^\lambda} \\ &= e^{-\lambda} e^\lambda \\ &= 1 \end{aligned}$$

**Note.** The Poisson Distribution arises as an approximation to the Binomial distribution if "n is large and p is small"

**Thm** 13.1. (*Poisson approximation to the Binomial*)

Let  $X$  have a Binomial distribution with parameters  $n$  and  $p$ . Then

$$P[X = x] \rightarrow \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{as } n \rightarrow \infty, p \rightarrow 0 \quad \forall x = 0, 1, \dots$$

such that  $np = \lambda$  is constant

PROOF:

$$\begin{aligned} P[X = x] &= \binom{n}{x} p^x (1-p)^{n-x} && (\text{for } x = 0, 1, \dots) \\ &= \frac{n!}{x!(n-x)!} * \frac{p^x (1-p)^n}{(1-p)^x} \\ &= \frac{n * (n-1) * \dots * 2 * 1}{x!(n-x) * (n-x-1) * \dots * 2 * 1} * \left(\frac{\lambda}{n}\right)^x * \left(1 - \frac{\lambda}{n}\right)^{n-x} && (\text{note } \lambda = np \Rightarrow p = \lambda/n) \\ &= \frac{\lambda^x}{x!} * \overbrace{\frac{n * (n-1) * \dots * 2 * 1}{n * n * \dots * n}}^{x \text{ terms}} * \left(1 - \frac{\lambda}{n}\right)^{n-x} * \frac{1}{(1 - \frac{\lambda}{n})^x} \\ &= \frac{\lambda^x}{x!} * \frac{n * (n-1) * \dots * 2 * 1}{n * n * \dots * n} * \left(1 - \frac{\lambda}{n}\right)^{n-x} * \frac{1}{(1 - \frac{\lambda}{n})^x} \end{aligned}$$

Now let  $n \rightarrow \infty$

The first term on the R.H.S  $\rightarrow 1$  as  $n \rightarrow \infty$  and each "ratio"  $\rightarrow 1$  (e.g.  $\frac{n-1}{n} \rightarrow 1$  as  $n \rightarrow \infty$ )

The second term on the R.H.S  $\rightarrow 1$  as  $n \rightarrow \infty$  since  $(1 - \lambda/n \rightarrow 1)$  as  $n \rightarrow \infty$

Finally  $(1 - \lambda/n)^n \rightarrow e^{-\lambda}$  as  $n \rightarrow \infty$  (another definition of  $e^{-\lambda}$ )

Hence

$$P[X = x] \rightarrow \frac{\lambda^x e^{-\lambda}}{x!} \quad (\text{for } x = 0, 1, \dots)$$

□

**Ex 13.1.** Suppose that sections of textile of length 1cm have a flaw in them with probability .01. If 1000 such sections are examined. What is the approximate probability that at least 50 will have a flaw?

**Sol.**

- Let  $X$  be the number of 1cm length sections that have a flaw
- Assume the binomial setup so that  $X$  has binomial distribution with  $n = 1000$  and  $p = 0.01$
- Hence since  $n$  is "large" and  $p$  is "small", we can use the Poisson approximation to the Binomial distribution
- Exact answer(Binomial)

$$P[X \geq 50] = \sum_{x=50}^{1000} \binom{1000}{x} (0.01)^x (1 - 0.01)^{1000-x}$$

- Poisson Approximation:

- Let  $\lambda = np = 1000 * 0.01 = 10$  Then

$$P[X \geq 50] = \sum_{x=50}^{\infty} \frac{10^x e^{-10}}{x!}$$

## 13.2 The Hypergeometric Distribution

**DEF 13.2.** A random variable  $X$  is said to have a Hypergeometric distribution with parameter  $N$ ,  $a$  and  $n$  if and only if

$$P[X = x] = \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}} \quad \forall x = 0, 1, \dots$$

**Note.**

- We see immediately that this is just the distribution of the number of tagged fish in a sample of size  $n$  drawn without replacement from a lake with  $N$  fish of which  $a$  are tagged (Fish in the lake problem)
- In fact, the Hypergeometric distribution describes the number of Type 1 objects observed in a sample of size  $n$  drawn without replacement from a "box" with  $N$  objects ( $a$  of Type 1 and  $N-a$  of Type 2) (Justification was done earlier)

**Notation:** We have notation for describing the distribution of a named random variable

- $X \sim \text{Ber}(p)$ :  $X$  has Bernoulli distribution with parameter  $p$
- $X \sim \text{Bin}(n, p)$ :
- $X \sim \text{Poisson}(\lambda)$
- $X \sim \text{Geometric}(p)$
- $X \sim \text{NegBin}(n, p, k)$

### 13.3 Mathematical Expectation & Variance

**Idea:** The probability distribution of a discrete random variable  $X$ , tells the whole story about  $X$  i.e How its value are distributed, with different probability.

However, maybe you want one of two summaries of  $P_X(x)$ , that captures the main features of this distribution.

The two features that immediately comes to mind are:

1. The "centre" of the distribution or "average value" of  $X$
2. The "variability" or "spread" of the values of  $X$

**DEF 13.3.** (*Expectation*) Let  $X$  be a discrete random variable, Let  $E(X) = \sum_{\text{all } x} xP[X = x]$ . Then call  $E(X)$  the expected value of  $X$  (or the expectation of  $X$ )

**Note.**

1. Often we denote  $E(X)$  by  $\mu_X$
2. we also call  $E(X)$  the population mean of  $X$
3.  $E(X)$  can be thought of as the weighted average of  $X$ , with weights  $P[X=x]$

## 14 Lec 14, Feb 26

### 14.1 Expectation

**DEF** 14.1. (*Expectation*) Let  $X$  be a discrete random variable, Let  $E(X) = \sum_{\text{all } x} xP[X=x]$ . Then call  $E(X)$  the expected value of  $X$  (or the expectation of  $X$ )

**Note.** cont'd with last class

1. Often we denote  $E(X)$  by  $\mu_X$
2. we also call  $E(X)$  the population mean of  $X$
3.  $E(X)$  can be thought of as the weighted average of  $X$ , with weights  $P[X=x]$
4.  $E(X)$  can also be thought of what you would get of observed infinitely many  $X_i$  all with the same distribution as  $X$ , and then take their average (followed from the law of Large numbers)
5. Two important properties of expectation, if  $c$  is a constant then

$$(a) E(cX) = cE(x) \Rightarrow \sum_{\text{all } x} cxP[X=x] = c \sum_{\text{all } x} xP[X=x]$$

$$(b) E(\sum_{i=1}^n x_i) = \sum_{i=1}^n E(X_i)$$

This does not require the random variables to be "independent"

6. The expected value does not provide a value of  $X$  that "you would expect to see" In fact, most often  $E(X)$  would not be one of the possible values of  $X$ . It is an average.
7. We say that the expectation of a random variable  $X$  exists if  $E(|X|) < \infty$  (i.e its finite)

**Ex** 14.1. if a random variable does not have a finite expectation

$$(a) \text{ Let } X = n \text{ with probability } \frac{c}{n^2} \quad \forall n = 1, 2, \dots$$

(b) Notice we must have

$$\sum_{x=1}^{\infty} P[X=x] = 1 \Rightarrow \sum_{n=1}^{\infty} \frac{c}{n^2} = 1 \Rightarrow c = 1 / \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

(c) Now

$$E(x) = \mu_x = \sum_{x=1}^{\infty} n \frac{c}{n^2} = \sum_{\text{all } x} xP(X=x) \quad (\text{which diverges})$$

(d) Hence  $E(x)$  does not exist

**Ex 14.2.** Calculation of insurance premiums.

Suppose that you wish to insure your computer against theft for 1000\$. The insurance company knows that 5% of such computers are stolen every year, and 95% are not. What premium should the company charge in order that its expected gain per customer is 0\$

**Sol.**

1. Let  $X = \text{gain of the company}$ , and  $c = \text{desired premium}$ . We first need  $P_X(x) = P[X = x]$

2. We have

$$\begin{aligned} P[X = c] &= 0.95 \\ P[X = -1000 + c] &= 0.05 \end{aligned}$$

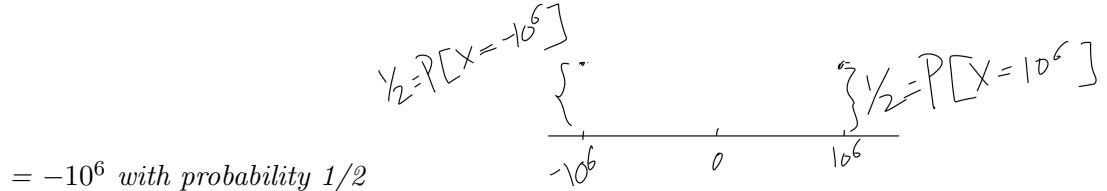
Hence

$$E(X) = 0.95c + 0.05(c - 1000)$$

3. set  $E(X) = 0$  we get  $c = 50\$$

Now it's clear that  $\mu_x$  alone does not provide an adequate picture (summary) of distribution. We need at least to describe the variability of our random variable about its expected value.

**Ex 14.3.** Let  $X = 10^6$  with probability 1/2



Here  $E(X) = -10^6 * 1/2 + 10^6 * 1/2 = 0$  which alone hardly tells us about the random variable  $X$ . This leads to the definition of a measure of spread of variation

## 14.2 Variance

**DEF 14.2.** Let  $X$  be a random variable. Let  $\text{Var}(X) = \sigma_x^2 = E[(X - \mu)^2]$ . Then we call  $\text{Var}(X)$  the variance of  $X$

**Note.**

1.  $\sigma_x^2$  is often used as the var of  $X$
2.  $\sigma_x^2$  is a constant parameter and a characteristic of the distribution of  $X$  (in a parameter setting) it is also called the population variance
3. The units of  $\text{Var}(X)$  are the same as the units of  $X^2$ , which renders the interpretation of  $\text{Var}(X)$  a bit awkward. Hence we often use  $\sqrt{\text{Var}(X)} = \sqrt{\sigma_x^2} = \sigma_x$  as a measure of spread. We call  $\sigma_x$  the standard deviation of  $X$ .  $\sigma_x^2, \sigma_x$  are mathematically equivalent.  $\sigma_x$  is in the same units as  $X$
4. There is an alternative form for  $\sigma_x^2$ , we have by definition that

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

This immediately tells that

$$E(X^2) \geq E^2(X) \quad (\text{since } E[(X - \mu)^2] > 0)$$

5.  $\text{Var}(cX) = c^2 \text{Var}(X)$

6. so if  $X$  is discrete

$$\begin{aligned} \sigma_x^2 &= \sum_{\text{all } X} (X - \mu_x)^2 P[X = x] \\ &= \underbrace{\sum_{\text{all } X} X^2 P[X = x]}_{E(X^2)} - \underbrace{\left[ \sum_{\text{all } X} X P[X = x] \right]^2}_{E^2(X)} \end{aligned}$$

**Ex 14.4.** Suppose that botanist knows the leaf length of certain plant has the following distribution (in cm)  
Let  $X$  be a random leaf length. Known:

$$P(X = 1.2) = 0.3$$

$$P(X = 2.2) = 0.2$$

$$P(X = 3.7) = 0.5$$

Find (i)  $E(X)$  and (ii)  $\sigma_x$

**Sol.**

i  $E(X) = \mu_x = 1.2 * 0.3 + 2.2 * 0.2 + 3.7 * 0.5 = 2.65$

ii  $\text{Var}(X) = E(X^2) - \mu_x^2 = 1.2^2 * 0.3 + 2.2^2 * 0.2 + 3.7^2 * 0.5 - 2.65^2 = 1.2225$

or  $\sigma_x^2 = (1.2 - 2.65)^2 * 0.3 + (2.2 - 2.65)^2 * 0.2 + (3.7 - 2.65)^2 * 0.5$

**15 Lec 15, Mar 01**

## 16 Lec 16, Mar 12

### 16.1 Continuous Probability Distribution

**DEF 16.1.**

- We say that the random variable  $X$  is continuous if its c.d.f  $F_X(x) (= P[X \leq x])$  is a continuous function of  $x$ .
- If  $X$  is continuous, then it turns out that  $P[X = x] = 0$  for every real number  $x$ .

we have that

$$\begin{aligned} P[X = x] &= P[X \leq x] - P[X < x] \\ &= F_X(x) - \lim_{y \rightarrow x^+} F_X(y) \\ &= F_X(x) - F_X(x) \quad (\text{since } F_X \text{ is continuous from below as well}) \end{aligned}$$

**Question:** Does this mean that since for any  $A = \{x: x \in A\}$  we have  $P[A] = 0$ ?

**Sol.** NO, since  $A$  is not necessarily the countable union of the  $x$  that make up  $A$ . If it were, then certainly by Ax 3, we would have

$$P[A] = \sum_{x_i \in A} P[X = x_i] = 0$$

**Question:** How do we then specify the probability distribution of a continuous random variable?

**Sol.** We can do this through  $F_X(x)$  as stated before, but unlike with discrete random variables, we can alternatively specify the distribution through  $P[X = x]$  for all  $x$ . These are equal to zero.

Instead we specify a function, called the **probability density function (pdf)** that takes the place of the probability mass function of discrete random variables.

**DEF 16.2.** Let  $X$  be continuous random variable with c.d.f  $F_X(x)$ . Then we call the real valued function of  $x$ , the **probability density function** of  $X$  if  $f_X(x)$  has the following property:

$$F_X(x) = \int_{-\infty}^x f_X(y) dy \quad \forall -\infty < x < \infty$$

i.e  $f_X$  is a p.d.f if it has the property that when integrated from  $-\infty$  to  $x$ , it gives  $P[X \leq x]$

**Note.**

1.  $f_X(x)$  can be shown to have the property that for any event  $A$ ,

$$P[X \in A] = \int_A f_X(y) dy$$

2. There is a 1-1 correspondence between a p.d.f and its c.d.f

**Reason:**

- we know that given  $f_X$  we can recover  $F_X$  by integration (definition), and conversely, given  $F_X$  we can recover  $f_X$  by using the fundamental theorem of calculus:

we have

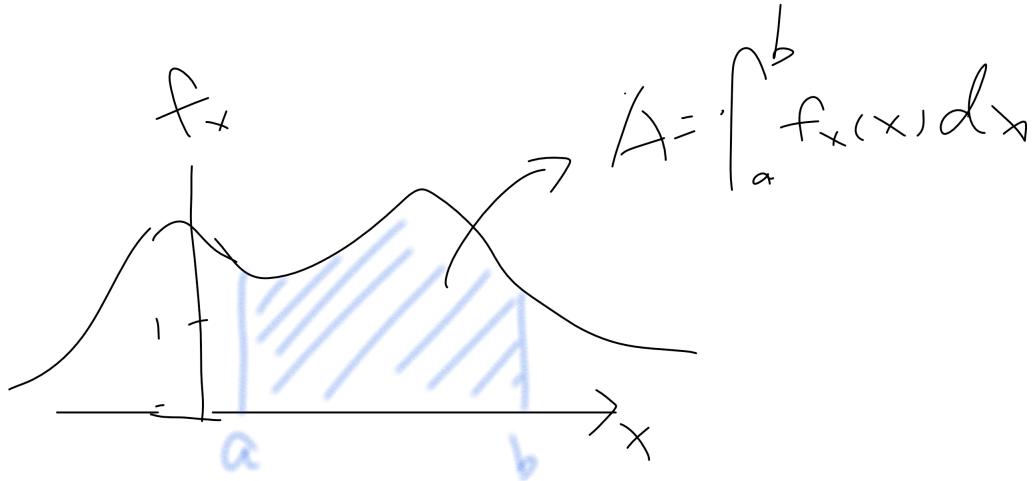
$$F'_X(x) = \frac{d}{dx} \int_{-\infty}^x f_X(y) dy = f_X(x)$$

3. It follows immediately that

$$\int_{-\infty}^{\infty} f_X(y) dy = 1 \quad (= P[X \leq \infty])$$

4. Unlike a p.m.f (which gives the probabilities and therefore has to lie in the interval  $[0,1]$ ), a p.d.f must be  $\geq 0$  but does not have to be  $\leq 1$ . The requirement is that when integrated over a set, it gives a probability

5. Pictorially we have,



Note that  $f_X(x) > 1$  for some  $x$

6. since  $F_X(x)$  is continuous, we have

$$P[a < X < b] = P[a \leq X \leq b] = P[a < X \leq b] = P[a \leq X < b]$$

7. Interpretation of the p.d.f

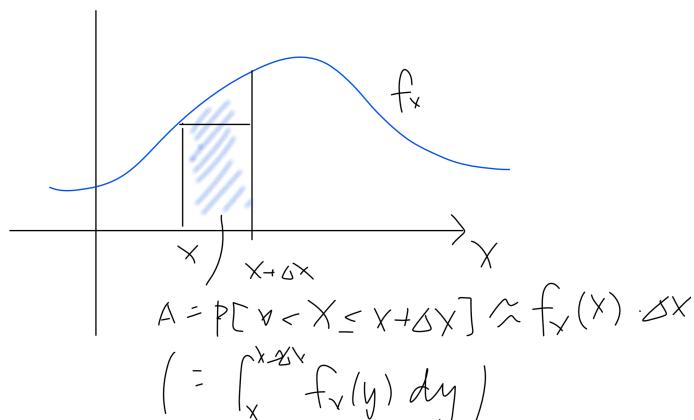
- It does not give a probability itself.
- However, we have that

$$\frac{P[x < X \leq x + \Delta x]}{\Delta x} \approx F'_X(x) \quad (\text{for small } \Delta x)$$

$$\frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} \approx f_X(x)$$

Hence we get

$$f_X(x)\Delta x \approx P[x < X \leq x + \Delta x]$$



8. Any non-negative function  $g$  that has the property that  $\int_{-\infty}^{\infty} g(x)dx = c < \infty$  for some constant  $c$  can be converted into a p.d.f by normalizing  $g$  to  $f$  where  $f = g/c$ .

Now  $f \geq 0$  and  $\int_{-\infty}^{\infty} f(x)dx = 1$

i.e  $g/\int_{-\infty}^{\infty} g(x)dx = f$

**Ex 16.1.** Suppose that  $X$  has a p.d.f

$$f_X(x) = \begin{cases} cx^2, & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

1. find  $c$
2. find  $P[\frac{1}{4} \leq X < \frac{1}{2}]$
3. find  $F_X(x)$  for all  $-\infty < x < \infty$

**Sol.**

1. since

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

we must have

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)dx &= \int_{-\infty}^0 0dx + \int_0^1 cx^2dx + \int_1^{\infty} 0dx \\ &= 1 \quad (\text{if } \int_0^1 cx^2dx = 1) \end{aligned}$$

to get that we must have

$$\begin{aligned} c &= \frac{cx^3}{3} \Big|_0^1 = 1 \\ c &= 3 \end{aligned}$$

2.

$$P\left[\frac{1}{4} \leq X < \frac{1}{2}\right] = \int_{1/4}^{1/2} 3x^2dx = \frac{7}{64}$$

3.

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f_X(y)dy && (\text{for } x \leq 0) \\ &= \int_{-\infty}^x 0dy = 0 \\ &= \int_{-\infty}^0 f_X(y)dy + \int_0^x f_X(y)dy && (\text{for } 0 < x < 1) \\ &= 0 + \int_0^x 3y^2dy \\ &= x^3 \end{aligned}$$

finally

$$F_X(x) = 1 \quad (\text{for } x \geq 1)$$

or

$$\int_{-\infty}^0 0 dy + \int_0^1 3y^2 dy + \int_1^\infty 0 dy = 1$$

**DEF 16.3.** If  $X$  is continuous with p.d.f  $f_X$ , then we define the **expected value** of  $X$  as follows:

$$E(X) = \mu = \int_{-\infty}^{\infty} f_X(x) dx =$$

We say that the expected value exists if  $E | X | < \infty$ .

More generally, we define the  $k$ th moment of  $X$  (if it exists) to be

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

Hence if  $X$  is continuous, then the variance of  $X$

$$\text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \underbrace{\mu^2}_{\left[ \int_{-\infty}^{\infty} x f_X(x) dx \right]^2}$$

We call  $\sqrt{\sigma_X^2} = \sigma_X$  the standard deviation as before

## 17 Lec 17, Mar 14

### 17.1 Expected Value and Variance of Cont. R.V

**DEF** 17.1. If  $X$  is continuous with p.d.f  $f_X$ , then we define the **expected value** of  $X$  as follows:

$$E(X) = \mu = \int_{-\infty}^{\infty} x f_X(x) dx =$$

We say that the expected value exists if  $E | X | < \infty$ .

More generally, we define the  $k$ th moment of  $X$  (if it exists) to be

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

Hence if  $X$  is continuous, then the variance of  $X$

$$\text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \underbrace{\mu^2}_{\left[ \int_{-\infty}^{\infty} x f_X(x) dx \right]^2}$$

We call  $\sqrt{\sigma_X^2} = \sigma_X$  the standard deviation as before

**Ex** 17.1. Recall: Suppose that  $X$  has a p.d.f

$$f_X(x) = \begin{cases} 3x^2, & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

1.  $E(X) = \mu$
2.  $\text{Var}(X) = \sigma_X^2$

**Sol.**

1.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^0 x * 0 dx + \int_0^1 x * 3x^2 dx + \int_1^{\infty} x * 0 dx \\ &= \frac{3}{4} \end{aligned}$$

2.

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 * 3x^2 dx \\ &= \frac{3}{5} \end{aligned}$$

Hence

$$\begin{aligned} \text{Var}(X) &= E(X^2) - \mu^2 \\ &= \frac{3}{5} - \left( \frac{3}{4} \right)^2 \end{aligned}$$

## 17.2 Some Named Continuous Distribution

The following distribution has occurred so often that they have acquired names

### 17.2.1 The Continuous Uniform Distribution

**DEF** 17.2.

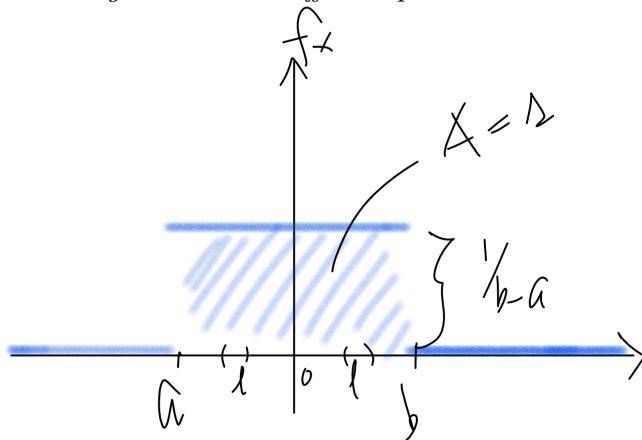
- The random variable  $X$  is said to have a uniform distribution on the interval  $(a, b)$  if

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{elsewhere} \end{cases}$$

- We write  $X \sim U(a, b)$
- if  $a = 0$ , and  $b = 1$ , i.e  $X \sim U(0, 1)$ , we say that  $X$  has a standard uniform distribution

**Note.**

1. Reason for the name: The p.d.f is constant on  $(a, b)$  i.e there are no intervals between  $a$  and  $b$  of the same length that have different probabilities



both interval  $l$  have the same probabilities.

2. Note that  $f_X$  is discontinuous at  $a$  and  $b$  which does not matter, since we only demanded that the c.d.f to be continuous for a continuous random variable
3. The c.d.f  $F_X$  if  $X \sim U(0, 1)$  looks like this Note that

$$F_X(x) = \begin{cases} 0, & x \leq 0 (\text{trivial}) \\ \int_0^x \frac{1}{1-0} dx = x, & 0 < x < 1 \\ \int_0^1 \frac{1}{1-0} dx + \int_1^x 0 dx = 1, & x \geq 1 \end{cases}$$

4. The continuous uniform is used to model the idea of complete randomness in an interval

5.

$$E(X) = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2} \quad (\text{as anticipated})$$

it is similarly, easy to show that

$$\sigma_X^2 = \text{Var}(X) = \frac{(b-a)^2}{12} \quad (\text{start with } E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx)$$

### 17.2.2 The Gamma Distribution

Some Background:

- A creature called the **Gamma Function** plays a role

**DEF 17.3.** Let  $\alpha > 0$ , then we call the function of  $\alpha$

$$\int_0^\infty x^{\alpha-1} e^{-x} dx$$

the gamma function and denoted by  $\Gamma(\alpha)$

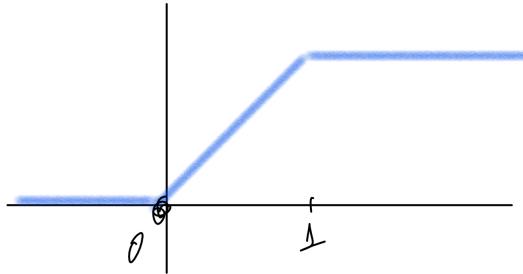
- There are three important properties of  $\Gamma(\alpha)$

1.  $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
2. if  $n$  is a positive integer, then  $\Gamma(n) = (n - 1)!$
3.  $\Gamma(1/2) = \sqrt{\pi}$

**DEF 17.4. (Gamma Distribution)**

we say that  $X$  has a Gamma distribution with parameters  $\alpha$  and  $\beta$ , if its p.d.f

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx, & x \geq 0 \\ 0, & x < 0 \end{cases}$$



**Note.**

1. Does this define a p.d.f?

- First, clearly  $f_X(x) \geq 0$  for all  $x$
- Second, does it integrate 1? i.e.  $\int_{-\infty}^{\infty} f_X(x) dx \stackrel{?}{=} 1$ . Consider

$$\int_0^\infty \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx = \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx$$

Idea: try to set the integral up so that it looks like a gamma function

Therefore, make a change of variable and set  $y = x/\beta$  and we get  $dx = \beta dy$  and  $x = y\beta$ , and the range of integration is still 0 to  $\infty$

$$\begin{aligned} &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \int_0^\infty \beta^{\alpha-1} y^{\alpha-1} e^{-y} \beta dy \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \beta^\alpha \int_0^\infty y^{\alpha-1} e^{-y} dy \\ &= \frac{1}{\Gamma(\alpha)} \Gamma(\alpha) \\ &= 1 \end{aligned}$$

2.  $E(X)$  and  $\text{Var}(X)$

- 

$$\begin{aligned} E(X) &= \int_0^\infty x \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \int_0^\infty x^{\alpha-1+1} e^{-x/\beta} dx \end{aligned}$$

Set  $y = x / \beta$  and get

$$\begin{aligned} &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \int_0^\infty y^{\alpha+1-1} \beta^\alpha e^{-y} \beta dy \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \beta^\alpha * \beta \int_0^\infty y^{(\alpha+1)-1} e^{-y} dy \\ &= \frac{1}{\Gamma(\alpha)} \beta \Gamma(\alpha+1) \\ &= \frac{1}{\Gamma(\alpha)} \beta \alpha \Gamma(\alpha) \quad (\text{prop. of Gamma func.}) \\ &= \alpha \beta \end{aligned}$$

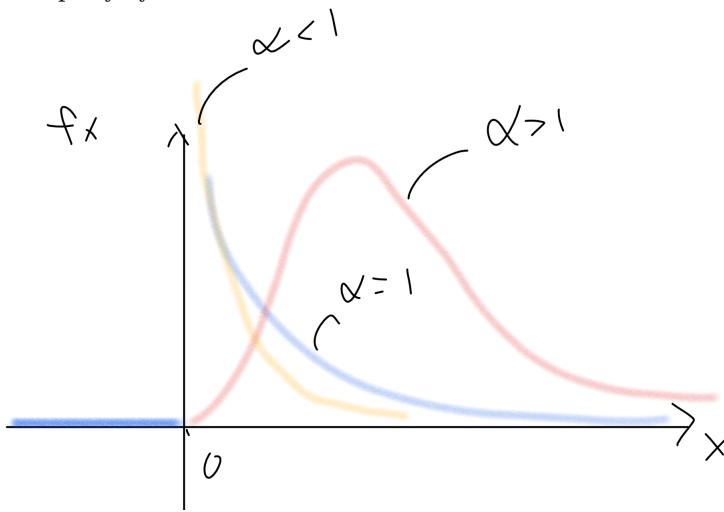
- Similarly, one can find  $E(X^2)$  and hence show that

$$\text{Var}(X) = \alpha \beta^2 \quad (\text{exercise})$$

- (Later we will see an easier way to find the moment of  $X$ )

3. We call  $\alpha$  the shape parameter and  $\beta$  the scale parameter of the Gamma Distribution and We denote it as  $X \sim \text{Gamma}(\alpha, \beta)$ .

4. The p.d.f of the Gamma Distribution looks like this:



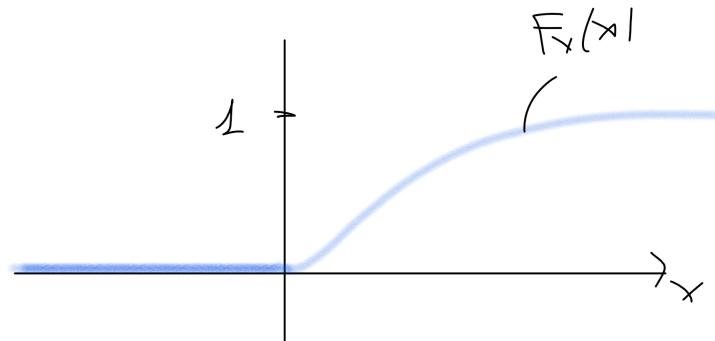
5. There are two very important distribution that are special cases of the Gamma Distribution

(a) The Exponential distribution with parameter  $\beta$

- Set  $\alpha = 1$  in the Gamma p.d.f and we get

$$f_X(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

- the p.d.f of the exponential distribution looks like this:
- The c.d.f looks like this:



we have

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \int_0^x \frac{1}{\beta} e^{-y/\beta}, & x \geq 0 \\ 1 - e^{-x/\beta}, & x < 0 \end{cases}$$

## 18 Lec 18, Mar 19

**Note.**

1. *The memoryless property of Experimental Distribution:*

- If  $X \sim \exp(\beta)$ , then

$$P[a \leq X < a+x \mid X \geq a] = P[0 \leq X \leq x]$$

Given that  $X \geq a$ , this historical plays no role in determining the probability that  $X$  will fill in the interval  $[a, a+x]$ . Hence "memoryless"

PROOF:

$$\begin{aligned} P[a \leq X < a+x \mid X \geq a] &= \frac{P[a \leq X < a+x \cap X \geq a]}{P[X \geq a]} \\ &= \frac{P[a \leq X < a+x]}{P[X \geq a]} \\ &= [F_X(a+x) - F_X(a)] * [1 - F_X(a)] \\ &= \frac{(1 - e^{-(a+x)/\beta}) - (1 - e^{a/\beta})}{1 - (1 - e^{-x/\beta})} \\ &= \frac{e^{-a/\beta}(1 - e^{-x/\beta})}{e^{-a/\beta}} \quad (\text{for } a \geq 0) \\ &= 1 - e^{-x/\beta} \\ &= P[0 \leq X \leq x] \end{aligned}$$

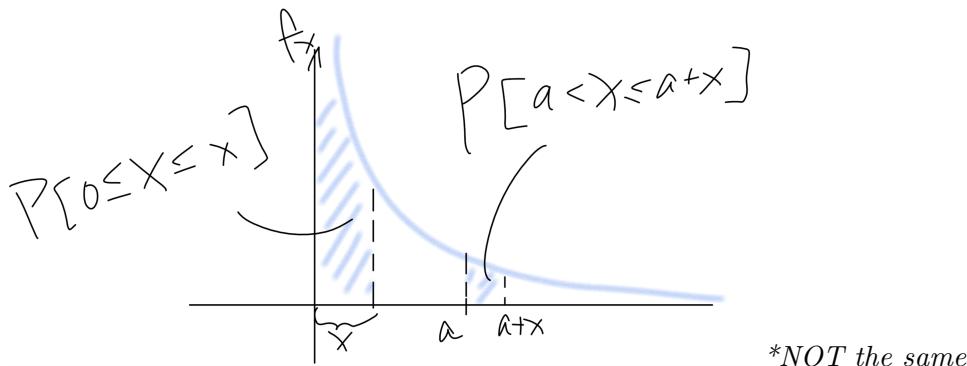
□

Amazingly, it can be shown that the only continuous distribution with the memoryless property is the exponential distribution.

**Note.**

- (a) It is not asserted that

$$P[a < X \leq a+x] = P[0 \leq X \leq x]$$



- (b) Some people write(parametrise) the exponential distribution as

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

with this parametrization, we have

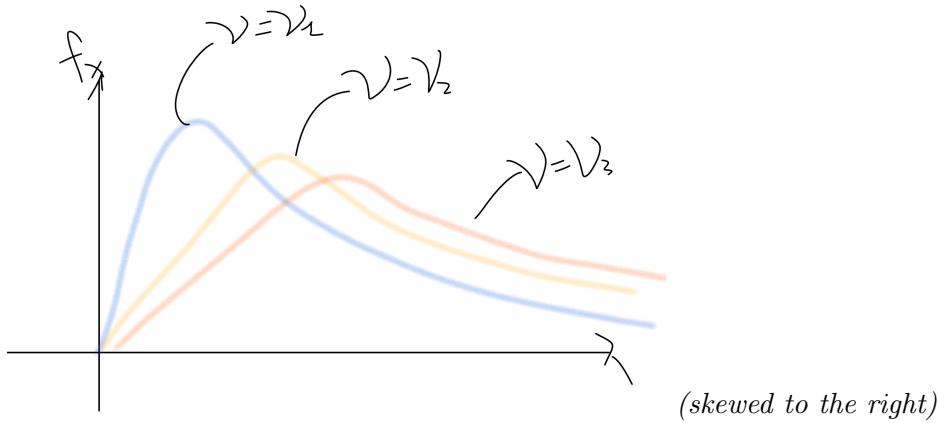
$$E(X) = \frac{1}{\lambda} \text{ and } \text{Var}(X) = \frac{1}{\lambda^2}$$

**Recall.** if  $X \sim \text{Gamma}(\alpha, \beta)$  then  $E(X) = \alpha\beta$ , and  $\text{Var}(X) = \alpha\beta^2$  since  $X \sim \exp(\beta) = \beta$  and  $\text{Var}(X) = \beta^2[\alpha = 1]$

## 18.1 Another special case of Gamma Distribution

**DEF** 18.1.

1. In the Gamma p.d.f, if we set  $\alpha = \nu/2$  and  $\beta = z$ . We get what is called a Chi-square distribution with parameter  $\nu$
2. We call  $\nu$  The degree of freedom of the chi-square distribution and denote as  $X \sim \chi_{\nu}^2$
3. Its p.d.f (depending on  $\nu$ )



4. We have immediately that

$$\begin{aligned} E(X) &= \nu & (\alpha\beta \text{ with } \alpha = \nu/2, \beta = 2) \\ \text{Var}(X) &= 2\nu \end{aligned}$$

5. The p.d.f

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\nu/2)} * \frac{1}{2^{\nu/2}} * x^{\nu/2-1} e^{-x/2} & X \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

## 18.2 The Normal(Gaussian) Distribution

**DEF** 18.2.

- We say that the random variable  $X$  has a normal distribution with parameter  $\mu$  and  $\sigma^2$  if its p.d.f is given by

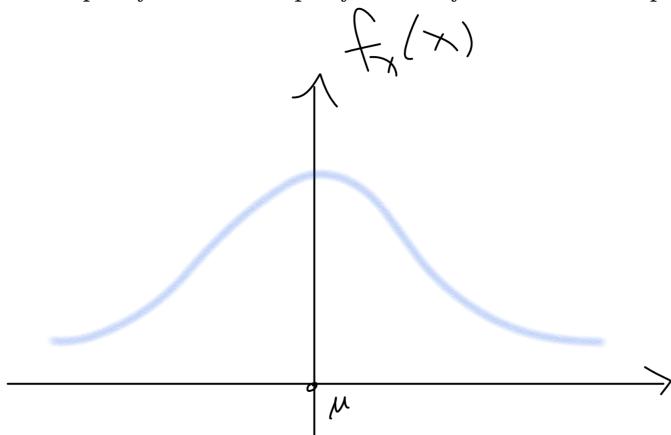
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (-\infty < x < \infty)$$

**Note.**

1. It is not too difficult to show that

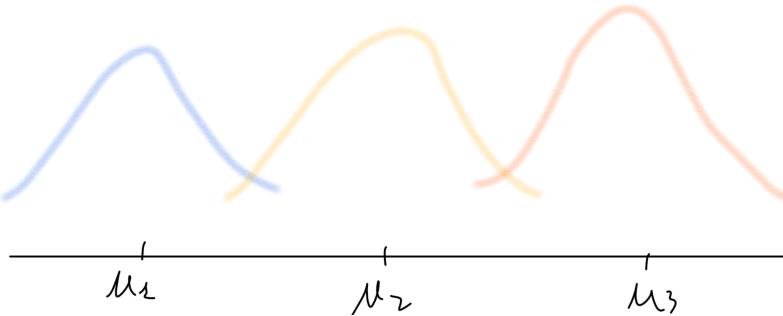
$$\int_{-\infty}^{\infty} f_X(x) dx = 1 \quad (\text{omit it})$$

2. The shape of a normal p.d.f is the famous bell-shape. It is NOT the only p.d.f with a bell-shape

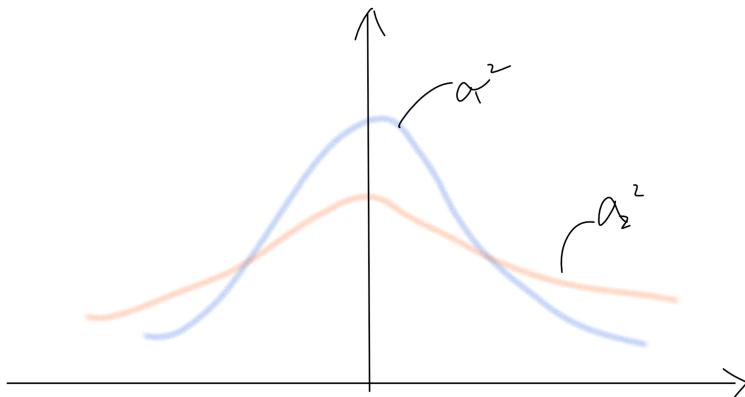


3. Not surprisingly the parameter  $\mu$  and  $\sigma^2$  are just the mean and variance of  $X$ . (Will show later using moment generating function)

4. If we keep  $\sigma^2$  fixed and change  $\mu$ . we shift the p.d.f along the x-axis



5. Notice  $f_X$  is symmetric along its mean  $\mu$ . If we keep  $\mu$  constant and change  $\sigma^2$ , we get  $\sigma_1^2 < \sigma_2^2$



6. when  $\mu = 1, \sigma^2 = 1$ , we call the resulting Normal Distribution a **standard Normal** and denote as  $Z \sim N(0, 1)$

7. Result: if  $X \sim N(\mu, \sigma^2)$  then

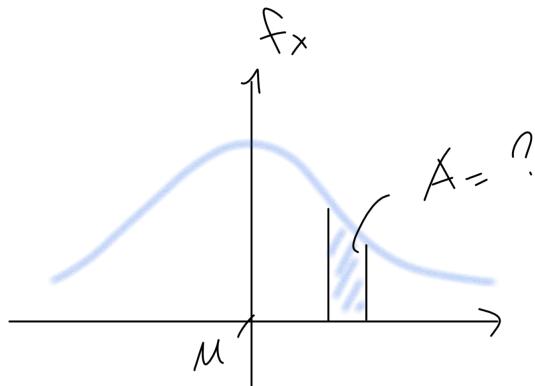
$$Z = \underbrace{\frac{X - \mu}{\sigma}}_{\text{standardization}} \sim N(0, 1) \quad (\text{Proof will be given later})$$

(a) No matter what the distribution if  $X$  standardization yields a new random variable with mean 0 and variance 1

(b) Easy to show by the property of expected value

(c) However standardization may NOT give a random variable of the same type

8. To calculate Normal Probability, we will have to use the Normal Table.



The trick is to use only one side of the table to find any  $\mu$  and  $\sigma^2$

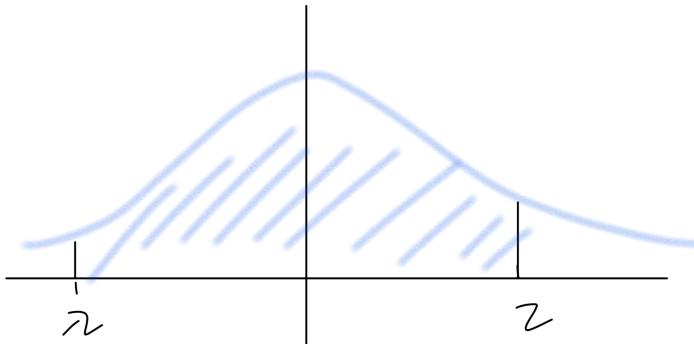
## 19 Lec 19, Mar 21

So, the idea to start with your probability statement about  $X$  and reduce to a probability statement about  $Z \sim N(0, 1)$ , by standardizing throughout.

**Ex 19.1.** Let the lengths of human thigh bones be approximately Normally distributed with mean to be 20cm and variance to be  $4 \text{ cm}^2$ . What proportion of thigh bones will have length between 16 cm and 24 cm.

**Sol.**

- Let  $X = \text{length of a thigh bone}$ . Known  $X \sim N(20, 4)$
- We want  $P(16 < X < 24)$ , reduce to a  $N(0, 1)$  statement.



- We have

$$\begin{aligned} P(16 < X < 24) &= P\left(\frac{16 - 20}{\sqrt{4}} < \frac{X - 20}{\sqrt{4}} < \frac{24 - 20}{\sqrt{4}}\right) \\ &= P(-z < Z < z) \\ &= P(Z \geq -2) - P(Z > 2) \\ &= \underbrace{(1 - 0.228)}_{\text{by symmetry of the } N(0, 1)} - 0.228 \\ &= 0.9544 \end{aligned}$$

### 19.1 Transformations of Random Variables

**Idea:** Although we may start with random variable  $X$ , our interest may really be in some random variable  $Y = g(X)$ .

**Ex 19.2.** We observe  $V$  and know its distribution, where  $V$  is the velocity of a particle A. We want the distribution of its kinetic energy,  $KE = \frac{1}{2}mv^2$ .

- If  $X$  is discrete then making such a transformation is usually simple and direct.

**Ex 19.3.** Let  $X \sim \text{Bin}(6, 0.4)$ . Find the distribution of  $g(X) = Y = \frac{x-2}{3}$

**Sol.**

$$\begin{aligned} P(Y = y) &= P\left(\frac{x-2}{3} = y\right) && \text{(for all } y \text{ in the range of } Y) \\ &= P(X = 3y + 2) \\ &= \binom{6}{3y+2} * (0.4)^{3y+2} * (1-0.4)^{6-(3y+2)} \\ &&& \text{(for } 3y+2 = 0, 1, 2, \dots, 6 \text{ (i.e for } y = -2/3, -1/3, 0, \dots, 4/3)) \end{aligned}$$

- Continuous random variable are more interesting :)

**Ex 19.4.** • Let  $X$  have c.d.f  $F_X$  and p.d.f  $f_X$

- Let  $Y = g(X)$ , for some function  $g$  while it is strictly increasing
- Find the p.d.f of  $Y$

**Sol.**

- Find  $F_Y(y) = P(Y \leq y)$  in terms of  $F_X$

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \quad (\text{g is strictly increasing hence the inverse must exist}) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

2.

$$\begin{aligned} f_Y(y) &= F'_Y(y) \\ &= \frac{d}{dy} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) \quad (\text{chain rule}) \\ &= f_X(g^{-1}(y)) \frac{dx}{dy} \\ &= f_X(g^{-1}(y)) \frac{1}{dy/dx} \end{aligned}$$

- Now if  $g$  is strictly decreasing, so that the inverse still exist.

- Then we use the fact that

$$g(X) \leq y \iff X \geq g^{-1}(y)$$

- Using this we get that (Do this)

$$f_Y(y) = f_X(g^{-1}(y)) * \frac{-1}{dy/dx}$$

Now we can combine these two formulas to give

$$f_Y(y) = f_X(g^{-1}(y)) \frac{1}{|dy/dx|} \quad (\text{for } y \text{ in the range of Y})$$

**Ex 19.5.** • Let  $X \sim N(\mu, \sigma^2)$

- Find the distribution of  $Z = \frac{X-\mu}{\sigma}$

**Sol.**

1.

$$\begin{aligned} F_Z(z) &= P(Z \leq z) \\ &= P\left(\frac{X-\mu}{\sigma} \leq z\right) \\ &= P(X \leq \sigma z + \mu) \\ &= F_X(\sigma z + \mu) \quad (-\infty < z < \infty) \end{aligned}$$

2.

$$\begin{aligned}
f_Z(z) &= F'_Z(z) \\
&= \frac{d}{dz} F_X(\sigma z + \mu) \\
&= f_X(\sigma z + \mu) \sigma && \text{(chain rule)} \\
&= \sigma \frac{1}{2\pi} \frac{1}{\sigma} e^{-1/2((\sigma z + \mu)/\sigma)^2} \\
&= \frac{1}{2\pi} e^{-1/2z^2} && \text{(for } -\infty < Z < \infty\text{)}
\end{aligned}$$

which, we recognize, is the p.d.f of a  $N(0, 1)$  r.v

**Ex 19.6.** The following example is of great importance in probability and statistic.

- Let  $Z \sim N(0, 1)$
- Then  $Y = Z^2 \sim \chi_1^2$

**Sol.**

1.

$$\begin{aligned}
F_Y(y) &= P(Z^2 \leq y) \\
&= P(-\sqrt{y} \leq Z \leq \sqrt{y}) \\
&= F_Z(\sqrt{y}) - F_Z(-\sqrt{y})
\end{aligned}$$

2.

$$\begin{aligned}
f_Y(y) &= F'_Y(y) \\
&= \frac{d}{dy} [F_Z(\sqrt{y}) - F_Z(-\sqrt{y})] \\
&= f_Z(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_Z(-\sqrt{y}) \frac{1}{2\sqrt{y}} && \text{(chain rule)} \\
&= f_Z(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_Z(\sqrt{y}) \frac{1}{2\sqrt{y}} && \text{(by sysmmetry of } N(0,1)\text{)} \\
&= \frac{1}{\sqrt{y}} f_Z(\sqrt{y}) \\
&= \frac{1}{\sqrt{y}} \frac{1}{2\pi} e^{-t/2} && \text{(for } y > 0\text{)} \\
&= \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{2}} y^{1/2-1} e^{-y/2} && \text{(for } y > 0\text{)}
\end{aligned}$$

Trivially since  $F_Y(y) = 0$  for  $y \leq 0$  and  $f_Y(y) = 0$  for  $y \leq 0$ . Now a  $\text{Gamma}(\alpha = 1/2, \beta = 2)$  distribution is a  $\chi_1^2$  distribution. we get

$$f_X(x) = \underbrace{\frac{1}{\Gamma(1/2)}}_{\sqrt{\pi}} \frac{1}{\sqrt{2}} x^{1/2-1} e^{-x/2} \quad (x > 0)$$

Which coincide with the p.d.f of  $Y = Z^2$

## 20 Lec 20, Mar 26

**Thm 20.1.** Let  $X$  be a continuous random variable with strictly increasing c.d.f  $F_x$ . Let  $U = F_x(X)$ . Then  $U \sim U(0, 1)$

**Note.**

1.  $F_x(X)$  is just a function of the real numbers  $x$
2. Think of  $F_x(X)$  as  $g(X)$ , a transformation of  $X$  using  $F_x(X)$  as your  $g$ .
3. Note that  $F_x(X) \neq P(X \leq x)$  since  $F_x(X)$  only equals to  $P(X \leq x)$  by the definition of  $x$  when it's a real number. If the argument is random, then  $F_x(X) \neq P(X \leq x) = 1$
- 4.

**Ex 20.1.**

- (a) Let  $X \sim \exp(\beta)$ . We have seen that

$$F_x(X) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-x/\beta} & x > 0 \end{cases}$$

The theorem says that if  $U = F_x(X)$  then  $U \sim U(0, 1)$

Notice

$$F_u(U) = \begin{cases} P[U \leq u] = 0 & u < 0 \\ P[\underbrace{1 - e^{-x/\beta}}_U \leq u] = u & u > 0 \end{cases} \quad (\text{simple transformation})$$

i.e

$$F_u(U) = \begin{cases} F'_u(U) = 1 & 0 < u < 1 \\ 0 & \text{elsewhere} \end{cases}$$

i.e  $U \sim U(0, 1)$

- (b) If  $X$  is normal  $(\mu, \sigma^2)$ , then its c.d.f is given by

$$F_X(x) = \int_{-\infty}^x \frac{1}{2\pi\sigma} e^{-\frac{1}{2}(\frac{y-u}{\sigma})^2} dy$$

The theorem says

$$U = F_X(x) = \int_{-\infty}^x \frac{1}{2\pi\sigma} e^{-\frac{1}{2}(\frac{y-u}{\sigma})^2} dy \sim U(0, 1)$$

PROOF:

i.

$$F_u(U) = P[U \leq u] = P[F_x(X) \leq u] = \begin{cases} 0 & u \leq 0 \\ P[X \leq F_x^{-1}(U)] = F_x[F_x^{-1}(U)] = U & u > 0 \end{cases}$$

ii.

$$f_u(U) = F'_u(U) = \frac{d}{du} U = \begin{cases} 1 & 0 < u < 1 \\ 0 & \text{elsewhere} \end{cases}$$

i.e  $U \sim U(0, 1)$

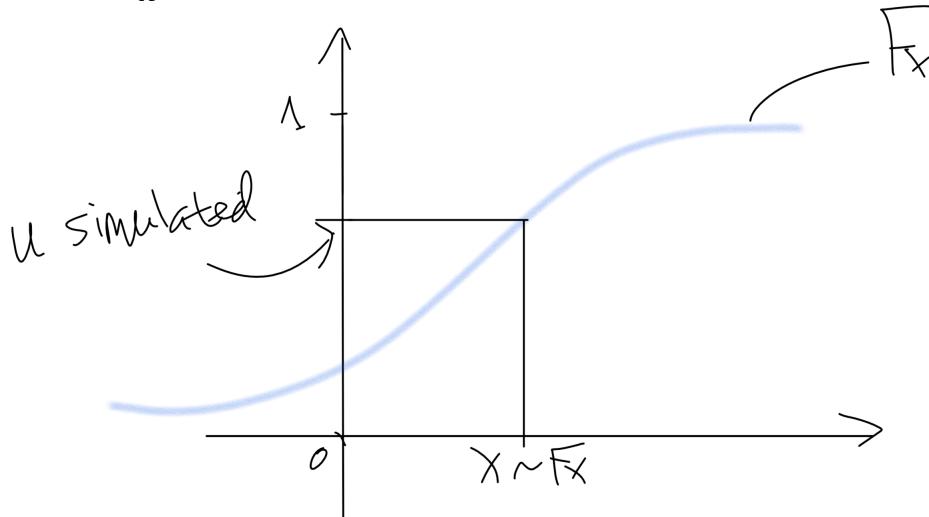
□

## 20.1 Application to the simulation of random variables from a given distribution

**Ex 20.2.** Suppose that we have some given distribution i.e we know its c.d.f  $F_X$ . We would like to simulate random variables from that distribution. It will be assumed that we can simulate random variables from a  $U(0,1)$  distribution. (Using one of several algorithms)

**Sol.** We can simulate from  $F_x(X)$  when  $X$  is not  $U(0,1)$  (e.g Normal Gamma, etc).

1. Simulate a  $U \sim U(0,1)$
2. We know by our theorem that  $U$  has the same distribution as  $F_x(X)$  since the latter is also  $U(0,1)$ . We write  $F_x(X) \stackrel{D}{=} U$  (Equal in distribution)
3. Now  $X \stackrel{D}{=} F_X^{-1}(U)$  we have simulated an  $X \sim F_X$



## 20.2 A Final Important Comment in Transformation

**Recall.**

We defined

$$E[g(X)] = \sum_X g(X)P[X = x] = \sum_X g(x)P_X(x) \quad (\text{X discrete})$$

Also

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

However if you think of  $g(X)$  as  $Y$ , then finding  $E[g(X)]$  is just finding  $E[Y]$  W.R.T distribution  $Y$ . i.e

$$E[g(X)] = E(Y) = \sum_y yPY = y = \sum_y yPY(y) \quad (\text{Y discrete})$$

and

$$E[g(X)] = \int_{-\infty}^{\infty} yFY(y)dy \quad (\text{Y continuous})$$

- There is a theorem that says that both way of thinking of  $E(g(X))$  give the same answer.
- The beauty of the first way is that you do not need change of variable to find the distribution of  $y = g(x)$ . You can use the original distribution of  $X$ .

### 20.3 Moment Generating Function

There is a function of the real numbers  $t$  that has great use in probability and statistics.

**DEF** **20.1.** Let  $X$  be the discrete (continuous) random variable.

Let

$$M_X(t) = \sum_{\text{all } x} e^{tx} P[X = x] = \begin{cases} \sum_{\text{all } x} e^{tx} P_X(x) & \text{discrete} \\ \int_{-\infty}^{\infty} e^{tx} F_X(x) dx & \text{continuous} \end{cases}$$

**Note.**

1.  $M_X(t) = E[e^{tx}]$  (obviously)
2.  $M_X(t)$  is a function of  $t$ . It may not exist for some distribution. We say that  $M_X(t)$  exists if it exists in some interval around  $t=0$
3. We call  $M_X(t)$  the moment generating function of  $X$

**Thm** **20.2.** (The Correspondence Theorem)

There is a 1-1 correspondence between the moment generating function of a random variable and its probability distribution.

i.e If you give me  $P_X$  or  $f_X$ , I can get  $M_X(t)$  (by definition)

If you give me  $M_X(t)$ , I can retrieve  $P_X$  and  $F_X$  uniquely

**Note.**

4. Reason for the name:  $M_X(t)$  can be used to find the moments  $E[X^k]$  of a distribution, as follows:  
Assume  $X$  is continuous:

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} F_X(x) dx \\ M'_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} F_X(x) dx \\ &= \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} F_X(x) dx \\ &= \int_{-\infty}^{\infty} x e^{tx} F_X(x) dx \end{aligned}$$

Hence

$$M'_X(t) = \int_{-\infty}^{\infty} x e^{tx} F_X(x) dx = E(x)$$

It's now easy to see that

$$M_X^{(k)}(0) = E[X^k]$$

## 21 Lec 21, Mar 28

Here are some important common moment generating function

1. Binomial:

$$\begin{aligned} M_X(t) &= E[e^{tx}] \\ &= \sum_{x=0}^n \underbrace{(e^t p)^x}_a \binom{n}{x} \underbrace{(1-p)^{n-x}}_b \\ &= (e^t p + 1 - p)^n \end{aligned} \quad ((a+b)^n)$$

Now

$$\begin{aligned} M'_X(0) &= n(e^t p + 1 - p)^{n-1} p e^t \Big|_{t=0} \\ &= np \end{aligned}$$

Next

$$M''_x(0) = np(1-p) + n^2 p^2$$

Hence

$$Var(X) = np(1-p) \quad (E(X^2) - \mu^2)$$

2. Poisson

$$\begin{aligned} E[e^{tx}] &= \sum_{x=0}^{\infty} \frac{e^{tx} \lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t \lambda)^x}{x!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t - 1)} \end{aligned}$$

Which limits for all t

It is easy to see that

$$M'_X(0) = \lambda,$$

and

$$M''_X(0) = E(X^2) = \lambda^2 + \lambda$$

Hence

$$Var(X) = E(X^2) - E^2(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

3. Let  $X \sim Gamma(X, \beta)$

$$\begin{aligned} M_X(t) &= E[e^{tx}] \\ &= \int_0^{\infty} e^{tx} \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^{\alpha}} x^{\alpha-1} e^{-\alpha/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^{\alpha}} \int_0^{\infty} x^{\alpha-1} e^{-x(1/\beta-t)} dx \end{aligned}$$

Let  $y = x(\frac{1}{\beta} - t) = x(\frac{1-\beta t}{\beta})$  then  $dx = \frac{\beta}{1-\beta t} dy$

$$\begin{aligned} M_X(t) &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \int_0^\infty \frac{\beta^{\alpha-1} y^{\alpha-1}}{(1-\beta t)^{\alpha-1}} e^{-y} \frac{\beta}{1-\beta t} dy \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^\alpha} \frac{\beta^\alpha}{(1-\beta t)^\alpha} \int_0^\infty y^{\alpha-1} e^{-y} dy \end{aligned}$$

which is valid for  $|t| < \frac{1}{\beta}$  (we want  $1 - \beta t > 0 \iff t < \frac{1}{\beta}$ )

$$\begin{aligned} &= \frac{1}{\Gamma(\alpha)} \frac{1}{(1-\beta t)^\alpha} \Gamma(\alpha) \\ &= \frac{1}{(1-\beta t)^\alpha} \quad (\text{for } |t| < \frac{1}{\beta}) \end{aligned}$$

Immediately, we have

$$M_X(t) = \frac{1}{(1-\beta t)}$$

If  $X \sim \chi_v^2$  then

$$M_X(t) = \frac{1}{(1-2t)^{v/2}} \quad (\alpha = v/2, \beta = 2)$$

It follows that

$$E(X) = M'_X(0) = \alpha\beta \quad \text{and} \quad E(X^2) = M''_X(0) = \alpha^2\beta^2 + \alpha\beta^2$$

Hence

$$Var(X) = \alpha^2\beta^2 + \alpha\beta^2 - (\alpha\beta)^2 = \alpha\beta^2$$

Therefore for the exponential distribution  $E(X) = \beta$ ,  $Var(X) = \beta^2$

And for the Chi-square distribution  $E(X) = v(\underbrace{\frac{v}{2}}_{\alpha} \underbrace{2}_{\beta}) = v$  and  $Var(X) = 2v$

#### 4. The $N(\mu, \sigma^2)$ distribution

side result about m.g.f:

Let  $y = ax + b$  where  $a$  and  $b$  are constants, then

$$M_Y(t) = M_{ax+b}(t) = e^{bt} M_X(at)$$

PROOF:

$$\begin{aligned} E[e^{yt}] &= M_Y(t) \\ &= E[e^{(ax+b)t}] \\ &= E[e^{atx} e^{bt}] \\ &= e^{bt} E[e^{(at)x}] \\ &= e^{bt} M_X(at) \end{aligned}$$

Return to the Normal distribution:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2}(\frac{(x-\mu)^2}{\sigma^2})} dx = e^{\mu t} + \frac{1}{2}\sigma^2 t^2 \quad (\text{for all } -\infty < t < \infty)$$

Now it follows immediately that

$$\begin{aligned} E(X) &= M'_X(0) = \mu \\ E(X^2) &= M''_X(0) = \mu^2 + \sigma^2 \\ Var(X) &= E(X^2) - E^2(X) = \mu^2 + \sigma^2 - \mu^2 = \sigma^2 \end{aligned}$$

□

*PROOF:* Another proof of  $X \sim N(\mu, \sigma^2)$  implies that  $Z = \frac{x-\mu}{\sigma} \sim N(0, 1)$

We have  $Z = \frac{1}{\sigma}x - \frac{\mu}{\sigma}$  (ax+b form where  $a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$ ). Therefore by theorem

$$\begin{aligned} M_Z(t) &= M_{\frac{x-\mu}{\sigma}}(t) \\ &= M_{\frac{1}{\sigma}x - \frac{\mu}{\sigma}}(t) \\ &= e^{-\frac{\mu}{\sigma}t} M_X\left(\frac{1}{\sigma}t\right) \\ &= e^{-\frac{\mu}{\sigma}t} e^{\mu\frac{1}{\sigma}t} + \frac{1}{2}\sigma^2 \frac{1}{\sigma^2} t^2 \\ &= e^{t^2/2} \end{aligned}$$

Now it is obvious that  $e^{t^2/2}$  is the moment generating function of  $N(\mu, \sigma^2)$  random variables with  $\mu = 0$  and  $\sigma^2 = 1$  (straight up substitution in the general formula for  $M_X(t)$ )

Therefore since  $M_Z(t) = e^{t^2/2}$  we have that  $Z \sim N(0, 1)$

by the uniqueness theorem i.e there cannot be any other random variable with the moment generating function  $e^{t^2/2}$  □

**Remark.** If  $Z \sim N(0, 1)$ , then  $x = \sigma z + \mu \sim N(\mu, \sigma^2)$

## 21.1 Multivariate Distribution

Idea : We wish to consider the simultaneous distribution of 2 or more random variables. For example  $x =$  tread depth on a car tire and  $y =$  the distance travelled. Maybe we want to know  $P[X \leq 2mm \cap X > 10000km]$  This is the topic under consideration.

Warning There is a lot of jargon.

### DEF 21.1.

- Let  $X$  and  $Y$  be any 2 random variables (discrete or continuous), We define the joint cumulative distribution function (c.d.f) of  $X$  and  $Y$  as follows:

$$F_{X,Y}(x, y) \stackrel{def}{=} P[X \leq x \cap Y \leq y] = P[X \leq x, Y \leq y]$$

## 22 Lec 22, Apr 2

**Note.**

1. We call  $F_X(x)$  and  $F_Y(y)$  the marginal c.d.f of  $X$  and  $Y$  or of  $F_{X,Y}(x,y)$ . We have by definition

$$F_X(x) = \lim_{y \rightarrow \infty^-} F_{X,Y}(x,y) \quad (F_{X,Y}(x, +\infty))$$

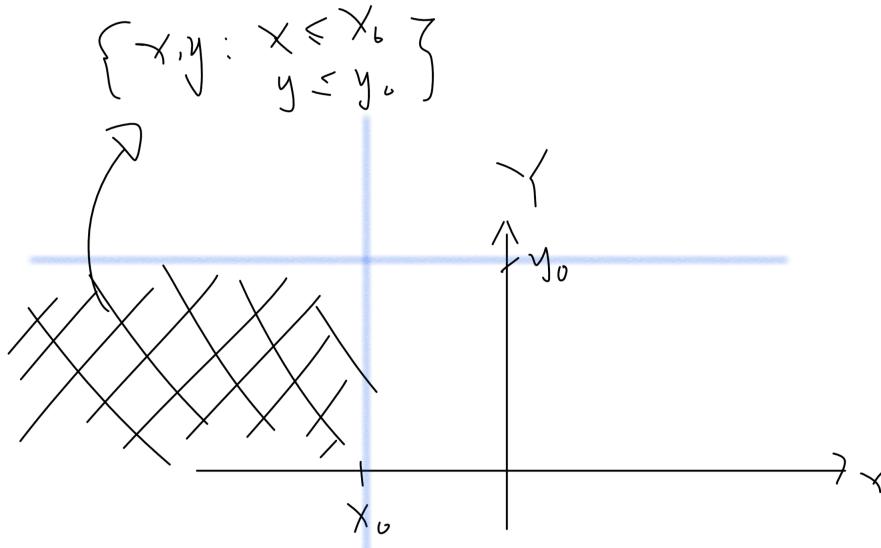
and

$$F_X(x) = \lim_{x \rightarrow \infty^-} F_{X,Y}(x,y) \quad (F_{X,Y}(+\infty, y))$$

2.

$$\begin{cases} \lim_{x,y \rightarrow \infty^-} F_{X,Y}(x,y) = 1 \\ \lim_{x,y \rightarrow \infty^+} F_{X,Y}(x,y) = 0 \end{cases}$$

3.  $F_{X,Y}(x,y) = P[X \leq x_0, Y \leq y_0]$  is the probability of the region below



### 22.1 Joint p.d.f and p.m.f

**DEF** 22.1.

*Continuous* If  $X$  and  $Y$  are jointly continuous random variables ( $F_{X,Y}(x,y)$  is continuous in  $X$  and  $Y$ ) Then we define the joint p.d.f  $f_{X,Y}(x,y)$  to be that non-neg function of  $x$  and  $y$  such that

$$F_{X,Y}(x,y) = \int_{-\infty}^{y_0} \int_{-\infty}^{x_0} f_{X,Y}(x,y) dx dy$$

- More generally

$$P[(X,Y) \in A] = \int \int_A f_{X,Y}(x,y) dx dy$$

*Discrete* If  $X$  and  $Y$  are jointly discrete random variables, then the joint probability mass function  $P_{X,Y}(x,y)$  is that function such that

$$P[X \leq x_0, Y \leq y_0] = F_{X,Y}(x_0, y_0) = \sum_{y \leq y_0} \sum_{x \leq x_0} P_{X,Y}(x,y)$$

In fact

$$P_{X,Y}(x,y) = P[X = x, Y = y]$$

**Note.**

1. We call  $f_X(x), f_Y(y), P_X(x), P_Y(y)$  the marginal p.d.f's and p.m.f's respectively of  $X$  and  $Y$  (i.e. of  $f_{X,Y}(x,y), P_{X,Y}(x,y)$ )

2. We have

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$$

$$\sum_{\text{all } y} \sum_{\text{all } x} P_{X,Y}(x,y) = 1$$

3. We have

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \quad , \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$$

and

$$P_X(x) = \sum_{\text{all } y} P_{X,Y}(x,y) \quad , \quad P_Y(y) = \sum_{\text{all } x} P_{X,Y}(x,y)$$

4. How do we recover the p.d.f (p.m.f) from  $F_{X,Y}(x,y)$

p.d.f We have

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = f_{X,Y}(x,y)$$

c.m.f By differencing the joint c.d.f (can be very messy, particularly in higher dimensions) Don't worry about this

5. If  $X$  and  $Y$  are jointly continuous, then

$$P[a_1 < X \leq a_2, b_1 \leq Y \leq b_2] = \int_{b_1}^{b_2} \int_{a_1}^{a_2} f_{X,Y}(x,y) dx dy = \sum_{(x,y): a_1 < X \leq a_2, b_1 \leq Y \leq b_2} P_{X,Y}(x,y)$$

**Ex 22.1.** see posted example on mycourses.

## 22.2 Expectation and Variance

**DEF** 22.2. Let  $g(x,y)$  be some function of  $X, Y$

- Then

$$E[g(x,y)] = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy & \text{continuous} \\ \sum_{\text{all } x} \sum_{\text{all } y} P_{X,Y}(x,y) & \text{discrete} \end{cases}$$

### 22.2.1 Application

For any two random variables with finite expectation

$$E[X + Y] = E(X) + E(Y)$$

PROOF:

continuous Think of  $X + Y$  as  $g(X, Y)$ , then

$$\begin{aligned} E(X + Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{X,Y}(x,y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) + y f_{X,Y}(x,y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) dx dy \\ &= \int_{-\infty}^{\infty} x dx \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy + \int_{-\infty}^{\infty} y dy \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx \\ &= \int_{-\infty}^{\infty} x f_X(x) dx + \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= E(X) + E(Y) \end{aligned}$$

Discrete similarly

□

## 22.3 Conditional Distribution

**DEF** 22.3. We define the conditional p.d.f and p.m.f of  $Y$  given  $X = x$  as follows

p.d.f

$$f_{Y|X=x} = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad (\text{for all } x * f_X(x) \neq 0)$$

p.m.f

$$P_{Y|X=x} = \frac{P[X=x, Y=y]}{P[X=x]} = \frac{P_{X,Y}(x,y)}{P_X(x)} \quad (P_X(x) \neq 0)$$

## 23 Lec 23, Apr 4

### 23.1 Conditional Distribution

**DEF** 23.1. We define the conditional p.d.f and p.m.f of  $Y$  given  $X = x$  as follows

p.d.f

$$f_{Y|X=x} = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad (\text{for all } x * f_X(x) \neq 0)$$

p.m.f

$$P_{Y|X=x} = \frac{P[X=x, Y=y]}{P[X=x]} = \frac{P_{X,Y}(x,y)}{P_X(x)} \quad (P_X(x) \neq 0)$$

**Note.**

1. The conditional p.d.f/p.m.f behaves just like a regular p.d.f/p.m.f as long as you do not mess with 'Conditioning Variable'. Thus let  $g(Y)$  be some function of some  $Y$ , then

$$E[g(Y) | X = x] = \int_{-\infty}^{\infty} g(y) f_{Y|X=x}(y | x) dy \quad (\text{continuous})$$

$$E[g(Y) | X = x] = \sum_g g(y) P_{Y|X=x}(y | x) \quad (\text{discrete})$$

In particular if  $g(Y) = y$ , then

$$E[Y | X = x] = \int_{-\infty}^{\infty} y f_{Y|X=x}(y | x) dy, \quad E[Y^2 | X = x] = \int_{-\infty}^{\infty} y^2 f_{Y|X=x}(y | x) dy \quad (\text{Continuous})$$

$$E[Y | X = x] = \sum_y y P_{Y|X=x}(y | x), \quad E[Y^2 | X = x] = \sum_y y^2 P_{Y|X=x}(y | x) \quad (\text{discrete})$$

Hence we have

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

2. We have

$$P[a < Y < b | X = x] = \int_a^b f_{Y|X=x}(y | x) dy \quad (\text{p.d.f})$$

$$P[a < Y < b | X = x] = \sum_{a < y < b} P_{Y|X=x}(y | x) \quad (\text{p.m.f})$$

**Ex** 23.1. (Posted on mycourses)

### 23.2 Covariance and Correlation

**Big Idea:** When considering invariate random variables, the main parameters of interest are the mean  $E(X) = \mu_x$  and  $\text{Var}(X) = \sigma_x^2$

If you have two random variables, say,  $X$  and  $Y$ , apart from the interest in  $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$ , you may also be interested in how these two random variables change W.R.T one another.

To this end, we introduce something called the **Covariance** between  $X$  and  $Y$

**DEF** 23.2. Let  $X$  and  $Y$  be two random variables (continuous / discrete), we define  $\text{Cov}(X, Y)$  as follows

•

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

**Note.**

1.  $\text{Cov}(X, Y)$  denote the covariance between  $X$  and  $Y$
2. The sign of  $\text{Cov}(X, Y)$  indicates whether  $X$  and  $Y$  vary in the same direction on average or in opposite directions. Thus if

$\text{Cov}(X, Y) > 0 \Rightarrow$  as  $X$  increases,  $Y$  increases (as  $X$  decreases,  $Y$  decreases)

$\text{Cov}(X, Y) < 0 \Rightarrow$  as  $X$  increases,  $Y$  decreases (as  $X$  decreases,  $Y$  increases)

3. Unfortunately  $\text{Cov}(X, Y)$  denoted by  $\sigma_{XY}$  is not scale invariant. That is

$$\text{Cov}(\alpha X, \beta Y) \neq \text{Cov}(X, Y)$$

4. Claim:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

PROOF:

$$\begin{aligned} \text{Cov}(X, Y) &\stackrel{\text{def}}{=} E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y \quad (\text{E}(X) = \mu_X, \text{E}(Y) = \mu_Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

□

Now

$$E(\alpha X \beta Y) - E(\alpha X)E(\beta Y) = \alpha \beta E(XY) - \alpha \beta E(X)E(Y) \neq E(XY) - E(X)E(Y) \quad (\text{unless } \alpha \beta = 1)$$

Hence we have shown that  $\text{Cov}(X, Y)$  is not scale invariant.

So, we would like a measure of covariation that has the same sign as  $\text{Cov}(X, Y)$ , But is also scale invariant. To this end, we have

**DEF 23.3.** Let  $X$  and  $Y$  be two random variables, we define the **Correlation** between  $X$  and  $Y$ , denoted by  $\text{Corr}(X, Y)$  ( $\rho(X, Y)$ ) as follows:

•

$$\text{Corr}(X, Y) \stackrel{\text{def}}{=} \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

**Note.**

1.  $\text{Corr}(X, Y)$  has the same sign as  $\text{Cov}(X, Y)$
2. It is easy to see that  $\text{Corr}(X, Y)$  is scale invariant in the sense that

$$|\text{Corr}(X, Y)| = |\text{Corr}(\alpha X, \beta Y)|$$

PROOF:

$$\begin{aligned}
 | \text{Corr}(\alpha X, \beta Y) | &= \frac{| \text{Cov}(\alpha X, \beta Y) |}{\sqrt{\text{Var}(\alpha X) \text{Var}(\beta Y)}} \\
 &= \frac{| \alpha \beta \text{Cov}(X, Y) |}{\sqrt{\alpha^2 \sigma_X^2} \sqrt{\beta^2 \sigma_Y^2}} \\
 &= \frac{| \alpha \beta | \text{Cov}(X, Y)}{| \alpha \beta | \sigma_X \sigma_Y}
 \end{aligned}$$

□

**Thm 23.1. (BIG THEOREM)**

$$| \rho(X, Y) | \leq 1 \quad (-1 \leq \text{Corr}(X, Y) \leq 1)$$

with the fact that  $| \rho(X, Y) | = 1$  iff  $Y$  is related to  $X$  in a linear fashion. (i.e  $Y$  can be written as  $Y = aX + b$  where  $a$  and  $b$  are constants)

Elaboration

- A correlation of  $\pm 1$  is an indication of a perfect linear relationship between  $X$  and  $Y$
- $\rho(X, Y) = 0$  implies that there is NO linear relationship between  $X$  and  $Y$ . (There may be a perfect curvilinear relationship.)
- Here are various scenarios that demonstrate different values of  $\rho(X, Y)$  between 0 and 1

