Probability (Math 323) winter 2019

Prof: David Wolfson
Yimeng Hu

Contents

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(red) = $\frac{6}{10}$ what do we mean by this statement?

   Sol.
   We cannot simply define "the probability" of getting a red as $\lim_{N\to\infty} \frac{\#\ of\ red}{\#\ of\ trials}$ *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 developped 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 theorm 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. $A \cap B = \phi \Rightarrow$ 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* $\underbrace{\text{time to recovery}}_{outcome}$*, 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 ω, an elemenatry 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* $\{\underbrace{1, ..., 6}_{Red}, \underbrace{7, ..., 10}_{Green}\}$ *Let w_i := event which marble i is drawn for i = 1,...,10 Then the sample space is:* $S = \{w_1, ..., w_{10}\}$

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:*
$S = \{\underbrace{\{Jan1, ..., Jan1\}}_{w_1}, ..., \underbrace{\{Dec31, ..., Dec31\}}_{w_n}\}$

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

Ex 2.6. *Toss a coin until you observe the 1^{st} head. Let n be the trial number at which this occurs. Then* $S = \{w_1, ..., w_n\}$ = {1,...,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^st. 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 lift no such sample space exits.*

2.3 Kolmogorov Axioms

DEF **2.4.** *Any subset $E \subset S$ is defined as an **event**. The empty set $\phi \subset S$ 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 $E_1, E_2, ...$ be any countable connection of events such that they are mutually exclusive, i.e $E_i \cap E_j = \phi \qquad \forall i \neq j$. 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:.

$$\begin{aligned} A \cup A^c = S, \Rightarrow P(A \cup A^c) = P(S) = 1 \quad &\text{(Ax 2)} \\ A \cap A^c = \phi \Rightarrow P(A \cup A^c) = P(A) + P(A^c) \quad &\text{(Ax 3)} \\ \Rightarrow P(A^c) = 1 - P(A) \end{aligned}$$

□

2. $P(\phi) = 0$

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

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

□

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

PROOF:.

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

□

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

Note. *if $A \cap B = \phi$, 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" $\Rightarrow \bigcup$, *"and"* $\Rightarrow \bigcap$, *"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 developpe lung cancer. Suppose that the probability of someone will either smoke or developpe 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 developpe 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*

$$P(E) = \frac{|\ E\ |}{N} = \frac{\#\ of\ outcoms\ in\ E}{Tot.possibleoutcomes}$$

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

$$E = \bigcup_{w_i \in E} w_i \Rightarrow P(E) = P(\bigcup_{w_i \in E} w_i) = \sum_{w_i \in E} P(W_i)$$

$$P(S) = \sum_{i=1}^{N} P(w_i) = 1 \Rightarrow P(w_i) = \frac{1}{N} \qquad \forall i = 1, ..., N \qquad \text{(Ax 2)}$$

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

□

4 Lec 04, Jan 17

we want a sample space with equally likely outcomes.

Recall. *S = {1,2, ... , 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), $P(R) = \frac{\#\ of\ ways\ to\ get\ a\ red\ marble}{tot.\ no.\ of\ possible\ outcomes} = \frac{6}{10}$.

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".
$$\binom{n}{r} = \frac{n!}{(n-r)!r!}, \qquad \text{note: } 0! = 1 \text{ by def}$$
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 $n_1, n_2, ..., n_k$ distinct objects respectively. The number of ways to form a set by selecting one object from each set is given by
$$n_1 * n_2 * ... * n_k$$

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(no two have the same birthday), then
$$P(E) = 1 - P(E^c) \qquad \text{(by Thm 1)}$$
The sample space here is S = {(Jan 1, ..., Jan 1), ... , (Dec 31, ..., Dec 31)}
First we assume that all of these outcomes are equally likely. There are finitely many of them. Therefore
$$\begin{aligned} P(E^c) &= \frac{\#\text{ of ways that } E^c\text{can occur}}{Tot.no.of outcomesins} \\ &= \frac{\text{P(365,n)}}{365^n} \\ &= \frac{365 * 364 * ... * (365-n+1)}{365^n} \end{aligned}$$
Hence
$$P(E) = 1 - \frac{365 * 364 * ... * (365-n+1)}{365^n}$$
□

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,...,**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,...,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

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

We have

$$\begin{aligned}\text{Tot. number of possible outcomes} &= \text{number of ways to draw n integers from a set of N distinct integers} \\ &= \binom{N}{n} \qquad \text{(counting rule 2.)}\end{aligned}$$

Now use **Multiplication Rule**
each sub-sausage contains $x \leq a$ integers selected from integer 1,...,a
each sub-sausage contains n-x integers selected from integers (a+1),...,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

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

finally

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

□

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

$$\frac{a}{N} \approx \frac{x}{n}$$
$$N = \frac{a * n}{x}$$

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, knowning that an event A has occured influences the probability that the event B will occure.

Ex 5.1. *In our marble problem, The probability of getting a red marble on the second of two draws (**without** replacement) knowning 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 $P(R_2$ knowning $R_1)$

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:*

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

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) $P(B \mid A) \geq 0$

   PROOF:.

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

   Hence true □

   (b) $P[S \mid A] = 1$

   PROOF:.

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

   Hence true □

   (c) $P\big[\bigcup B_i \mid A\big] = \sum_{i=1}^{\infty} P[B_i \mid A] \qquad where B_i \cap B_j = \phi \qquad \forall i \neq j$

   PROOF:. (exercise) □

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

2. $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad$ *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

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

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: $P\big[ND_1 \cap ND_2\big]$
Given: $P[ND_1] = 0.90$ *and* $P\big[ND_2 \mid ND_1\big] = 0.95$

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

□

***Extension:** LEt $A_1, ..., A_n$ be any sequence of events. Then

$$\begin{aligned} P[A_1 \cap ... \cap A_n] &= P[A_n \mid A_1 \cap ... \cap A_{n-1}] * P[A_1 \cap ... \cap A_{n-1}] \\ &= P[A_n \mid A_1 \cap ... \cap A_{n-1}] * P[A_{n-1} \mid A_1 \cap ... \cap A_{n-2}] * P[A_1 \cap ... \cap A_{n-2}] \\ &= ... \\ &= \Pi_{i=1}^{n} P[A_i \mid A_1 \cap ... \cap A_{i-1}] * P(A_0) \qquad \text{(i=1,...,n)} \end{aligned}$$

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 stating with the last event is called the process of ***conditioning backwards***.
When you are required to find the probability of the intersection of several evetns, think of conditioning backwards

6 Lec 06, Jan 24

6.1 Conditioning Backwards

DEF 6.1. *The process of repeatedly conditioning stating with the last event is called the process of* ***conditioning backwards****.*
When you are required to find the probability of the intersection of several evetns, think of conditioning backwards

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

1. *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

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

2. *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*

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

   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* $B_1, B_2, ...$ *be* m *events that satisfy the following*

1. $B_i \cap B_j = \phi \qquad \forall i \neq j$
2. $\bigcup_{i=1}^{m} B_i = S$ *we call* $\{B_1, B_2, ...\}$ *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

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

□

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 $B_1, B_2, ...$ be defined exactly as in the Law of Probability. Then we can write*

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

PROOF:.

$$\begin{aligned} P(B_k|A) &= \frac{P(B_k \cap A)}{P(A)} \\ &= \frac{\overbrace{P(A|B_k)P(B_k)}^{\text{mult. rule}}}{\underbrace{\sum_{i=1}^{m} P(A|B_1)P(B_i)}_{\text{Law of tot.Prob.}}} \end{aligned}$$

□

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.

$$\begin{aligned} \textit{Let Pos} &:= \textit{event that a test is positive} \\ \textit{Let Neg} &:= \textit{event that a test is negative} \\ \textit{Let Bc} &:= \textit{event that a woman has breast cancer} \\ \textit{Let } Bc^c &:= \textit{event that a woman does not have breast cancer} \end{aligned}$$

Known

$$\begin{aligned} P(Pos\,|Bc) &= 95\% \\ P(Pos\,|Bc^c) &= 1 - 95\% = 0.05 \\ P(Neg\,|Bc^c) &= 95\% \\ P(BC) &= 0.005 \\ P(BC^c) &= 1 - 0.005 \end{aligned}$$

We have that BC *and* BC^c *are disjoint and* $BC \cup BC^c = S$.
Therefore by the Law of Tot. Prob.

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

ii) If a woman tests postive, 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)*

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

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 prevelance and P(Disease | Pos) is called the postive predictive value of the test. Note that pos predictive value depends on the sensitivity, specitivity and prevalue of the diseasea.*

7 Lec 07, Jan 29

Some comments on Baye's Theorem and diagnostic tests:
In the world of diagnostic tests **P(Pos | Disease)**:= the sensitivity of the test
P(neg |Disease^c):= the specificity.
P(Disease):= the disease prevelance and
P(Disease | Pos):= the postive predictive value of the test.
Note that pos predictive value depends on the sensitivity, specitivity and pre. value of the diseasea.

7.1 Statistical Independence*

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

Ex 7.1. *In the marble problem, supose 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 event 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 | 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 $A_1, A_2, ..., A_n$ are sait to be mutually independent if $P\big[A_{i_1} \cap ... \cap A_{i_k}\big] = P[A_{i_1}\big] * ... * P[A_{i_k}]$ for all subset of $A_{i_1}...A_{i_k}$ 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 $A_1, A_2, ...A_3$

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

Note.

1. *Events $A_1, A_2, ..., A_n$ 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 independece.*

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 $\mathrm{P}(A^c \cap B^c) = \mathrm{P}(A^c)\mathrm{P}(B^c)$

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

   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 $A_1, A_2, ...A_n$ and $B_1, B_2, ..., B_n$. 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.
   E.g $P\Big[(A_3 \cap A_5 \cap A_6) \cap (B_1 \cap B_2)\Big] = P\Big[A_3 \cap A_5 \cap A_6\Big] * P\Big[B_1 \cap B_2\Big]$

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 independece 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 evetns 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 independece 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:.

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

   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) $A \cap B = \phi \Rightarrow P(A \cap B) = 0$

   (b) $A \perp\!\!\!\perp B \Rightarrow P(A \cap B) = P(A)P(B)$

   (c) A and B are dependent, then $P(A \cap B) = P(B \mid A)P(A)$ *(note, this is always true)*

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

Note. *Preliminary note*

1. *when sampling **without** replacement, the outcomes in the sequence of draws are dependent.*
   Thus if you have a box of 10 items if 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. (randome variable.)

1. *What is the probability that exactly 2 will have the mutation?*

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

   (a) *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

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

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

   (b) *Sum up the probability of all such configurations*

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

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

Sol. *WTS P[x ≥ 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 completationally simpler:

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

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

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

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 $R_1, R_2, R_3, W_4, R_5, W_6$... 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 $X : w \to \Re$*

Note.

1. *X is just a function. However, the argument and random outcomes of an experiment.*
   i.e The outcomesare 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 explorality wells that are drilled before oil is first struck (here, could assume the value of X to be 1,2,3...)*

2. ***Countinous 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 continous

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

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

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 $P[\omega : X(\omega) \leq x] = P\big[X^{-1}(-\infty, x]\big]$*
   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 subsect 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 $\big\{P[x \in A] : A \subset \Re$?y?is an event$\big\}$ The probability distribution of Random variable X.*

10 Lec 10, Feb 07

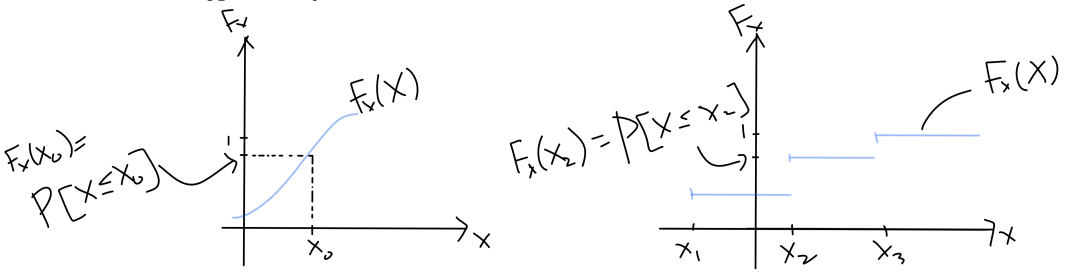
10.1 Cumulative Distribution Function

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 ∈ A): A is a subset of $\mathbb{R}$ }
2. Amazingly, in order to uniquely specify the probability of of all subsets, it's enough to specify the probabilty of just a small collection of subsets.
   i.e it's enough to specify the probability of intervals of the type (-∞, 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 ∈ $(-\infty, x]$] is called the cumulative distribution function (cdf) and is denoted by $F_x()$ Thus $F_x(X) \overset{\text{def}}{=} P(X \leq x)$
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 $F_x(X) \leq F_x(Y)$*
   (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) *$\lim_{x \to \infty} F_x(X) = 1$ and $\lim_{x \to -\infty} F_x(X) = 0$ 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 hasa 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 $X_1, X_2, ...$ which can be assumed by the random vairable X. Thus a discrete random vairable has cdf with at most countably many jumps.*

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

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

*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^-*

- *$P[X = x_0] = F_X(x_0) = F_X(x_0^-)$ = size of the jum in F_X 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 **probablity 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 probablity 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 wich the probablity 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 probablity 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

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

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

2. suupose F_X is given
Let X_0 be some arbitrary value in the range of X.
Then we have

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

□

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 hour late] = 0.1
P[on time] = 0.7
The questions are:

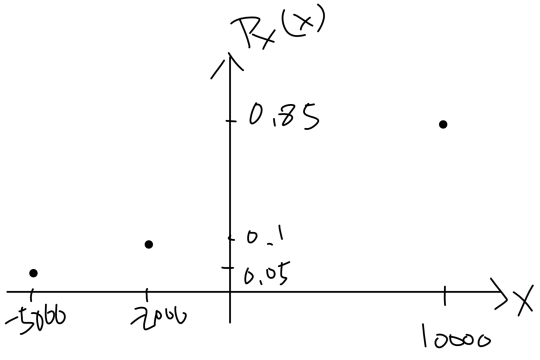
1. *Find the probability function of the random variable that represent the gain on a flight*
2. *Find the cdf of this random variable.*

Sol. *The elementary outcomes are:*
$\omega_1 =$ *cancelled.*
$\omega_2 =$ *1/2 hour late.*
$\omega_3 =$ *on time.*
$\omega_4 =$ *less than 1/2 hour late.*
Let X = 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*

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

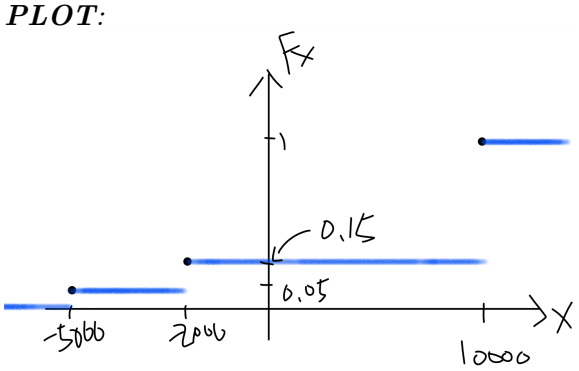
Hence $P_X(x)$ *is specified and*
PLOT:



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

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

***PLOT**:*



11 Lec 11, Feb 12

11.1 Some named discrete distribution

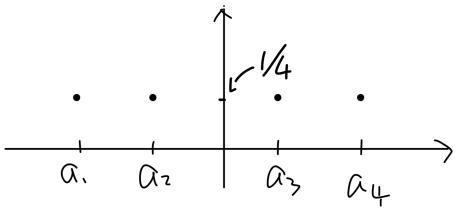
Some discrete distribution arise so frequently that they are given special names. Here are some of them:

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* $a_1, a_2, ..., a_N$ *if* $P_X[x] = P[X = x] = 1/N$ *for all* $x = a_1, a_2, ..., a_N$ *(i.e* $P_X(a_i) = 1/N$ *for all i = 1, 2, ..., N)*

Note.

1. *A probability function* $P_X(x)$ *must satisfy the following 2 conditions:*
   (a) $P_X(x) \geq 0 \qquad \forall x$
   (b) $\sum_{all\ x} P_X(x) = 1 \qquad$ *(P(S) = 1)*
2. *Reason for the name, "uniform":*
   All possible values of X are equally probable, with probabilty 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 Bernouilli Distribution

DEF 11.2. *The random variable X is said to have a Bernouilli 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,...,n and* $0 < p < 1$

Note.

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

PROOF:.

consider

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

(to be continued...) □

12 Lec 12, Feb 14

Midterm include materials up untill 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 Bernouilli 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, ..., 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 , ..., n*

PROOF:.

- Done already in the constant 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] = 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 ≤ 5 years.*
Suppose proportion of patients that survive more than 5 years is 0.80
If 30 patients are to recieve 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,..., 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, ...*

How does the geometric distribution arrive ?
Answer:

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

Claim: $P[X = x] = (1-p)^{x-1}p$ for x =1,2,...
(i.e the geometric distribution is the distribution of the "time" to success in a sequence of independent Bernouilli Trials with constant probability of success)

PROOF:.

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

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

□

Note.

- *The Binomial gives the number of successes in a fixed number of trials n.*
- *The geometric gives the trial at which th 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 kth success occurs in a sequence by independent Bernouilli trials with constant probability of success p.*

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

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

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*

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

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

$$P[X = x] \to \frac{\lambda^x e^{-\lambda}}{x!} \qquad \textbf{\textit{as}}\ n \to \infty, p \to 0 \qquad \forall x = 0, 1...$$

such that $np = \lambda$ is constant

PROOF:.

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

Now let $n \to \infty$

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

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

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

Hence

$$P[X=x] \to \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{(for x = 0,1,...)}$$

□

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 hvae a Hypergeometric distribution with parameter N, a and n if and only if*

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

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) (Justitication was done earlier)*

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

- X $\sim$ Ber(p): X has bernoulli distribution with paramenter p
- X $\sim$ Bin(n, p):
- X $\sim$ Poisson(λ)
- X $\sim$ Geometric(p)
- X $\sim$ 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* $E(X) = \sum_{all\ x} xP[X = x]$. *Then call E(X) the expected value of X (or the expectation of X)*

Note.

1. *Often we denote E(X) by* μ_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* $E(X) = \sum_{all\ x} xP[X = x]$. *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* μ_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 man* 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) $E(cX) = cE(x) \Rightarrow \sum_{all\ x} cxP[X = x] = c\sum_{all\ x} P[X = x]$
   (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* $E(\mid X \mid) < \infty$ *(i.e its finite)*

   Ex 14.1. *if a random variable does not have a finite expectation*
   (a) *Let* $X = n$ *with probability* $\frac{c}{n^2}$ $\quad \forall n = 1, 2, ...$
   (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*

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

   (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 = gain of the company, and c = desired premium. We first need* $P_X(x) = P[X = x]$
2. *We have*

$$P[X = c] = 0.95$$
$$P[X = -1000 + c] = 0.05$$

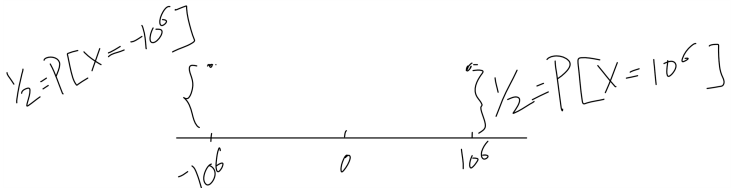
Hence

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

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

Now it's clear that μ_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*



$= -10^6$ *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 Var(X) =* $\sigma_x^2 = E[(X-\mu)^2]$. *Then we cal Var(X) the variance of X*

Note.

1. σ_x^2 *is often used as the var of X*
2. σ_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 Var(X) are the same as the units of* X^2, *which renders the interpretation of Var(X) a bit awkward. Hence we often use* $\sqrt{Var(X)} = \sqrt{\sigma_x^2} = \sigma_x$ *as a measure of spread. We call* σ_x *the standard deviation of X.* σ_x^2, σ_x *are mathematically equivalent.* σ_x *is in the same units as X*
4. *There is an alternative form for* σ_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

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

5. *Var(cX) = c^2 Var(X)*
6. *so if X is discrete*

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

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) σ_x

Sol.

*i E(X) = μ_x = 1.2 * 0.3 + 2.2 * 0.2 + 3.7 * 0.5 = 2.65*

ii Var(X) = E(X^2) - μ_x^2 $= 1.2^2 * 0.3 + 2.2^2 * 0.2 + 3.7^2 * 0.5 - 2.05^2$ *= 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

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

Question: Does this mean that since for any A = {x: x ∈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:*

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

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

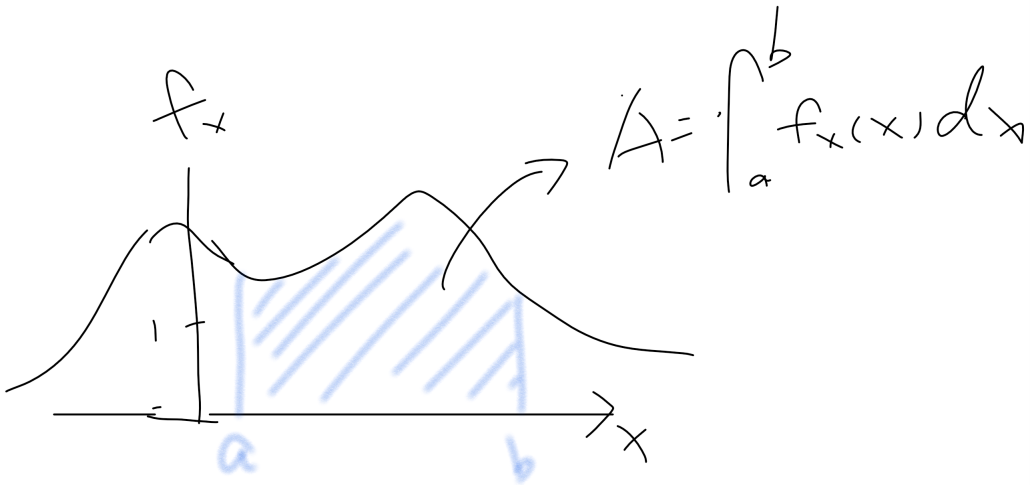
$$F_X'(x) = \frac{d}{dx} \int_{-\infty}^{x} f_X(y)dy = f_X(x)$$

3. *It follows immediately that*

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

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* ≥ 0 *but does not have to be* ≤ 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*

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

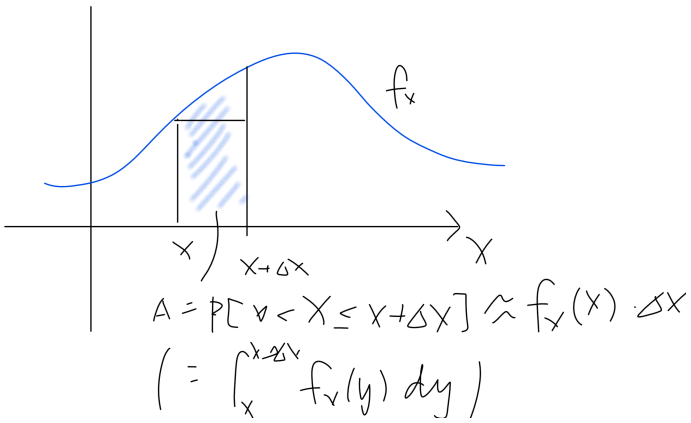
7. *Interpretation of the p.d.f*

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

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

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* $\int_{-\infty}^{\infty} g_X(y)dy = c < \infty$ *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* $\int_{-\infty}^{\infty} f_X(y)dy = 1$
i.e $g/\int_{-\infty}^{\infty} gdy = f$

Ex **16.1.** *Suppose that X has a p.d.f* $f_X(x) = cx^2$ *for* $0 < x < 1$ *and 0 elsewhere.*

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

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

to get that we must have

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

2.

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

3.

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

finally

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

or

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

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 $E \mid X \mid < \infty$.
More generally, we define the kth moment of X (if it exists) to be

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