

Probability and Statistics

Lecture-1

What is probability?

- ▶ **Uncertainty** - condition when results/outcomes are not completely determined and the outcomes are affected by a number of factors or the outcomes may be by pure chance.
- ▶ **Simple examples of uncertainty**- buying a lottery ticket, turning a wheel of fortune, or tossing a coin to make a choice.
- ▶ **Uncertainty** appears in virtually all areas of computer science and software engineering.
- ▶ Installation of software requires **uncertain** time and often **uncertain** disk space
- ▶ A newly released software contains an **uncertain** number of defects/bugs

- ▶ When a computer program is executed, the amount of required memory may be **uncertain**.
- ▶ When a job is sent to a printer, it takes **uncertain** time to print as there is always a different number of jobs in a queue ahead of it.
- ▶ Electronic components fail at **uncertain** times, and the order of their failures cannot be predicted exactly.
- ▶ Viruses attack a system at **unpredictable** times and affect an **unpredictable** number of files and directories.

Everyday instances of uncertainty

- ▶ Stock markets - nobody would have ever lost a cent in stock trading had the market contained no uncertainty!
- ▶ Chance of rain on any particular day - forecasting weather precisely, with no error, is not a solvable problem, again, due to uncertainty!
- ▶ Bus pick-up times.
- ▶ The quantity of food that has to be served daily in the mess/canteen.
- ▶ And so on!

- ▶ Thus, most of the phenomena we come across are highly **uncertain**
- ▶ We have to understand and deal with it
- ▶ In fact, consciously or unconsciously, we are dealing with it in our day to day life
- ▶ Most of the time, we are forced to make decisions under **uncertainty**

- ▶ We have to deal with internet and e-mail knowing that we may not be protected against all kinds of viruses.
- ▶ New software has to be released even if its testing probably did not reveal all the defects.
- ▶ Some memory or disk quota has to be allocated for each customer by servers, internet service providers, etc., without knowing exactly what portion of users will be satisfied with these limitations.
- ▶ And so on.

- ▶ **Probability** is all about measuring the **uncertainty** and **randomness**
- ▶ In this course, we will learn
 - ▶ how to evaluate probabilities, or chances of different results (when the exact result is uncertain),
 - ▶ how to select a suitable model for a phenomenon containing uncertainty and use it in subsequent decision making,
 - ▶ how to evaluate performance characteristics and other important parameters,
 - ▶ how to make optimal decisions under uncertainty

⌚ What about Statistics?

- ▶ Statistics is the science of taking a decision or changing your mind under uncertainty!
- ▶ Let us look at a game (kind of gambling!).
- ▶ Suppose there is a huge warehouse which consists of a large (unknown) number of hats of only two colours - blue and orange.



- ▶ The warehouse is sealed and there is single see-through window with a button to press on its side.
- ▶ Whenever the button is pressed, a fully automated robot randomly picks a hat, shows that to you through the window and throws it back in to the pool of hats.
- ▶ You will be paid Rs. 100 if an orange hat comes up and you'll lose Rs. 80 if a blue hat comes up.
- ▶ It is also known that, before you, 50 persons have pressed the button and out of these 50 times, orange hat came up 20 times and blue hat came up 30 times.
- ▶ **Question:** Will you play the game?

- ▶ The question of winning or losing boils down to finding the proportion of orange hats in the warehouse
 - ▶ With the given information, how do we estimate this proportion?
 - ▶ Even though winning is uncertain in a single time, how many times do you have to play so that you do not end up losing any amount?
 - ▶ What would be "expected" value of your winnings?
- ☞ The 'statistics' part helps us in answering the above questions!

- ▶ The subject of statistics deals with **collection of data, organizing it** and **analysing it** for any possible conclusions
- ▶ The data we are interested may be people, computers, cars or any objects we have to analyse
- ▶ By **population**, we mean all the units/objects we are interested in. Population may not always mean people!
- ▶ If the population is too large, we sometimes do **sampling** to estimate the population characteristics

- ▶ For instance, consider the case where a baby food production company wishes to estimate the proportion of children below 5 years in a densely populated metropolitan city
- ▶ Instead of collecting data from each and every household, the company 'randomly' selects a certain (relatively small) number of houses and tries to 'estimate' the required proportion from this smaller data set
- ▶ Some serious concerns:
 - ▶ How random (or unbiased) was the sample?
 - ▶ What are the techniques used to 'estimate' the population proportion?
- ▶ The technique of sampling/estimation is very crucial and can lead to varying conclusions if not done ethically!

Do you know him?



Mr. X

- ▶ Communal riots broke out during India-Pakistan partition during 1946-47
- ▶ In Delhi, the capital city, an unknown number (say n) of refugees (majority of them are Muslims) who choose to remain in India arrived to seek shelter in the iconic red fort
- ▶ The government employed contractors to feed them.
- ▶ The contractors used to submit bills (of different commodities like rice, pulses, salt etc.) to the government to claim the money

- ▶ A secretary to the government of India suspected that the contractors were over quoting the commodities they purchased
- ▶ **Obvious thing to do** - send someone into fort and count the number of refugees
- ▶ **Problem** - the guards at the gate did not allow them to go inside as they were not members of the same community as that of the refugees
- ▶ So, the problem now is to “**estimate the unknown number, n** ”, of refugees “**without entering the fort**”
- ▶ How do we do that? Any guesses?

 Mr. X did the following:

- ▶ He, along with his team surveyed to estimate per capita/person consumption of rice, pulses and salt.
- ▶ They found out that these numbers are r , p and s respectively
- ▶ Let R, P, S be the amounts quoted by the contractor for rice, pulses and salt,
- ▶ Then the estimate for number of people is $\frac{R}{r}$ or $\frac{P}{p}$ or $\frac{S}{s}$
- ▶ If the contractor quoted the right amounts, these three numbers should be approximately equal
- ▶ It was found that

$$\frac{R}{r} = 30,253, \quad \frac{P}{p} = 21,122 \text{ and } \frac{S}{s} = 10,891$$

- ▶ Which of these numbers, 30,253,21122,10,891 is the correct estimate for number of refugees?
- ▶ Since salt was the least priced commodity, contractors usually do not over quote the salt quantity, and hence Mr. X chose the number $\frac{S}{s} = 10,891$
- ▶ When everything settled down and the refugees started moving out, it was found that the actual number was 10,887!

What went right?

- ▶ The estimates of per-capita consumption were pretty much accurate!
- ▶ Such accuracy can be achieved only by random (unbiased) sampling and using the right methods of estimation
- ▶ Wrong sampling methods might have resulted in wrong estimates!
- ▶ **Common sense/ethics** - While choosing the least number among $\frac{R}{r}$, $\frac{P}{p}$ and $\frac{S}{s}$

Moral of the story?

Though statistics is a powerful tool to make decisions, this tool should be wisely used and coupled with common sense, ethics etc. to get the accurate/unbiased results!

Coming back to Mr. X



Prasanta Chandra Mahalanobis

- ▶ One of the members of the **first planning commission of independent India**
- ▶ Founder of **Indian Statistical Institute**
- ▶ Considered as the father of modern statistics in India
- ▶ His birthday **June 29** is celebrated as **National Statistics day!**

Probability and Statistics

Lecture-2

Recall: The game from last class

- ▶ Huge warehouse which consists of a large (unknown) number of hats of only two colours - blue and orange.



- ▶ Whenever the button is pressed, a fully automated robot randomly picks a hat, shows that to you through the window and throws it back in to the pool of hats.
- ▶ You will be paid Rs. 100 if an orange hat comes up and you'll lose Rs. 80 if a blue hat comes up.

- ▶ Now, if we came to know that there are 1,00,000 hats in the warehouse of which 45000 are orange ones, what is the chance of your winning?
- ▶ 0.45 or 45% ?
- ▶ How did you get this 0.45?
- ▶ $0.45 = \frac{45000}{1,00,000}$
- ▶ Your chance of winning = $\frac{\text{number of ways picking a hat so that you win}}{\text{total number of ways of picking a hat}}$

Experiment: Toss two “fair” coins simultaneously.

 **Question-1:** What is the chance that we get (T, H) ?

- ▶ All possible outcomes = $\{(H, H), (T, H), (H, T), (T, T)\}$
- ▶ Our outcome is one of these 4 outcomes. So there is a $1/4$ or 25% chance that we get (T, H) .

 **Question-2:** What is the chance that exactly one of the two coins show up heads?

- ▶ Here our outcome set is $\{(T, H), (H, T)\}$. So the chance is $2/4 = 1/2$ or 50%

- ☞ In both the cases, the game and the example, we calculated the proportion of '**favourable outcomes**' to total number of possible outcomes of the experiment
- ☞ Important thing in both the cases: all possible outcomes have '**equal chance**' of showing up
- ☞ Under this assumption, let us guess the definition of probability

Guess: In an experiment, let all possible outcomes be given by a set S . Also assume that each outcome has equal chance of showing up. Then the chance of getting an outcome from a subset $E \subset S$ is

$$P(E) = \frac{|E|}{|S|}$$

where $|A|$ denotes the number of elements in the set A .

- ☞ Let us formalize this approach now!

 The first step is finding out all possible outcomes of the experiment

The set of all possible outcomes of an experiment is called **sample space** of the experiment and is denoted by S .

all possible outcomes \leftrightarrow sample space

Example-1: Tossing two coins simultaneously

- ▶ All possible outcomes are $\{(H, H), (T, H), (H, T), (T, T)\}$
- ▶ Hence sample space is $S = \{(H, H), (T, H), (H, T), (T, T)\}$

Example-2: Measuring (in hours) the lifetime of an electronic device

- ▶ The device may fail as soon as it is turned on or it may work for lifetime!
- ▶ Hence, the sample space consists of all non-negative real numbers, i.e., $S = \{x : 0 \leq x < \infty\}$

Example-3: Suppose we have 7 horses numbered 1, 2, 3, 4, 5, 6, and 7. In a race, the experiment is to determine the order of finish

- ▶ Here the horses can finish the race in any order
- ▶ So all possible orders are all possible permutations of $(1, 2, 3, 4, 5, 6, 7)$
- ▶ Thus, sample space is

$$S = \{\text{all } 7! \text{ permutations of } (1, 2, 3, 4, 5, 6, 7)\}$$

Example-4: Drawing 3 balls randomly from a bowl containing 6 white and 5 black balls

- ▶ The first ball can be chosen in $6 + 5 = 11$ ways and second ball can be chosen in $11 - 1 = 10$ ways and finally the third ball can be chosen in 9 ways.
- ▶ Thus, the sample space consists of $11 \cdot 10 \cdot 9 = 990$ outcomes!
- ▶ Observe that for our computation, the colour of the balls is irrelevant!

☞ Once we determine the sample space, the next step is to determine 'favourable' outcomes for which the probability has to be computed

Any subset E of the sample space is known as an **event**. If the outcome of the experiment is contained in E , we say that E has occurred.

favourable outcomes \leftrightarrow elements of event E

Example-1: Tossing two coins simultaneously

- ▶ Sample space $S = \{(H, H), (T, H), (H, T), (T, T)\}$
- ▶ Let E_1 denote the event of getting (T, H)
- ▶ Then, $E_1 = \{(T, H)\}$
- ▶ Let E_2 denote the event of getting exactly one head

$$\implies E_2 = \{(T, H), (H, T)\}$$

Example-2: Measuring (in hours) the lifetime of an electronic device

- ▶ Sample space $S = \{x : 0 \leq x < \infty\}$
- ▶ Let E be the event that the device does not last longer than 5 hours
- ▶ Then, $E = \{x : 0 \leq x < 5\}$

Example-3: Determining the order of finish of 7 numbered horses.

- ▶ Sample space $S = \{\text{all } 7! \text{ permutations of } (1, 2, 3, 4, 5, 6, 7)\}$
- ▶ Let E be the event that the horse which is numbered 4 ends up in the first place
- ▶ Then E consists of permutations in which 4 is in the first place
 $\implies E = \{\text{all outcomes in } S \text{ starting with a 4}\}$
 $\implies |E| = 6!$

Example-4: Drawing 3 balls randomly from a bowl containing 6 white and 5 black balls

- ▶ We saw that the sample space consists of $11 \cdot 10 \cdot 9 = 990$ outcomes
- ▶ Let E be the event that one of the balls is white and the other two are black
- ▶ Possible order of the balls is WBB, BWB and BBW .
- ▶ WBB is possible in $6 \cdot 5 \cdot 4 = 120$ outcomes
- ▶ BWB is possible in $5 \cdot 6 \cdot 4 = 120$ outcomes
- ▶ BBW is possible in $5 \cdot 4 \cdot 6 = 120$ outcomes
- ▶ Thus $|E| = 120 + 120 + 120 = 360$

Short recap

- ☞ Given an experiment,
 - ▶ first step is to determine the **sample space** S , i.e., the set of all possible outcomes of the experiment
 - ▶ determine the **event space** E , which is the set of favourable outcomes.
- ☞ E is always a subset of S
- ☞ In both the cases, determining the exact number of outcomes is very crucial
- ☞ This is where the knowledge of permutations and combinations comes into play!

☞ Most of the times, the experiment may not be simple and may be combination of two or more experiments

☞ Examples:

- ▶ Tossing a coin and rolling a dice simultaneously
- ▶ Rolling a dice and picking a card from a shuffle of cards

The basic principle of counting

Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of the m possible outcomes and if, for each outcome of the experiment 1, there are n possible outcomes of the experiment 2, then together there are mn possible outcomes of the two experiments

Example-1: Tossing a coin and rolling a dice simultaneously

- ▶ Tossing the coin is the first experiment and rolling the dice the second experiment
- ▶ The first experiment has 2 possible outcomes and the second one has 6 possible outcomes
- ▶ Thus, by basic principle of counting, total number of outcomes = $2 \times 6 = 12$

Example-2: Rolling a dice and picking a card from a shuffle of cards

- ▶ Here the first experiment has 6 possible outcomes and the second has 52 possible outcomes
- ▶ Hence total number of outcomes = $6 \times 52 = 312$

Example-3: A small community consists of 10 women, each of whom has 3 children. If one woman and one of her children are to be chosen as mother and child of the year, how many different choices are possible?

- ▶ Regard choice of woman as outcome of the first experiment and the subsequent choice of her children as the outcome of the second experiment.
- ▶ We have a choice of 10 woman. So outcomes of the first experiment are 10 in number and once a woman is chosen, we have a choice of her 3 children.
- ▶ Thus, by basic principle of counting, there are $10 \times 3 = 30$ possible choices

Example-4:

- (a) How many different 7-place license plates are possible if the first 2 places are for alphabets and the other 5 are for numbers?
- (b) Repeat part-(a) under the assumption that no letter or number can be repeated in a single license plate
- (a)
- ▶ First experiment is to place alphabets in first 2 places and the second experiment is to place numbers in the remaining 5 places.
 - ▶ There are $26 \times 26 = 26^2$ outcomes for the first experiment and $10 \times 10 \times 10 \times 10 \times 10 = 10^5$ outcomes for the second experiment
 - ▶ Thus, by basic principle of counting, total number of possible license plates are $26^2 \times 10^5$

(b)

- ▶ Even in this case, we have two experiments
- ▶ First experiment is to place alphabets in first 2 places and the second experiment is to place numbers in the remaining 5 places
- ▶ For the first place we have a choice of 26 alphabets and since repetition is not allowed, for the second place we only have a choice of 25 alphabets
- ▶ Thus, number of outcomes of the first experiment is 26×25
- ▶ Similarly, for second experiment, the number of outcomes is $10 \times 9 \times 8 \times 7 \times 6$
- ▶ Thus, total number of possible number plates without allowing repetitions is $26 \times 25 \times 10 \times 9 \times 8 \times 7 \times 6$

Summary

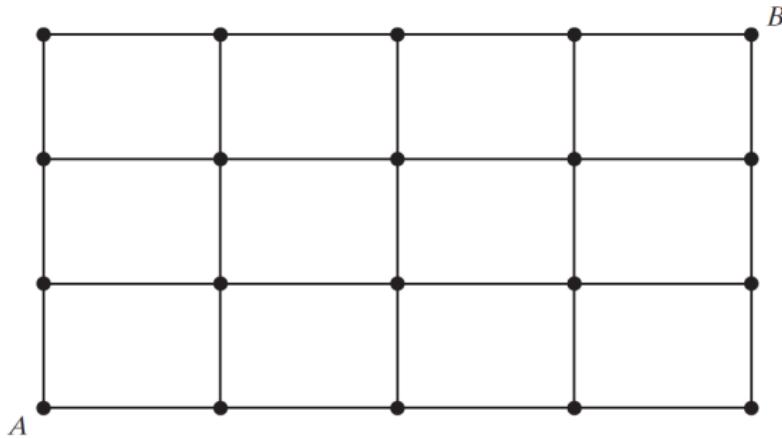
- ▶ Guessed the definition of probability
- ▶ To calculate probabilities, the first step is to list out all possible outcomes of the experiment, which we defined to be **sample space**
- ▶ The second step is to identify the set of favourable outcomes which is a subset of sample space.

all possible outcomes \leftrightarrow sample space

favourable outcomes \leftrightarrow elements of events

- ▶ In order to determine sample space or events, we need to know how to count different possibilities
- ▶ Basic principle of counting

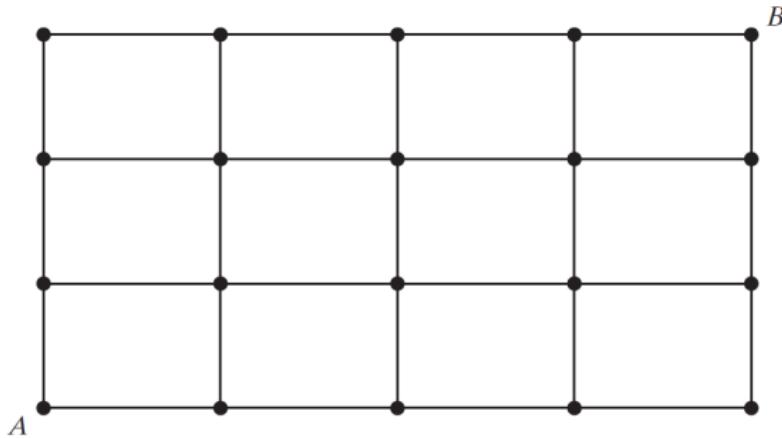
Exercise: Consider a grid of points shown below. Suppose that, starting at the point A , we can go one step to the right or one step up at each move. This procedure is continued until the point B is reached. How many different paths from A to B are possible?

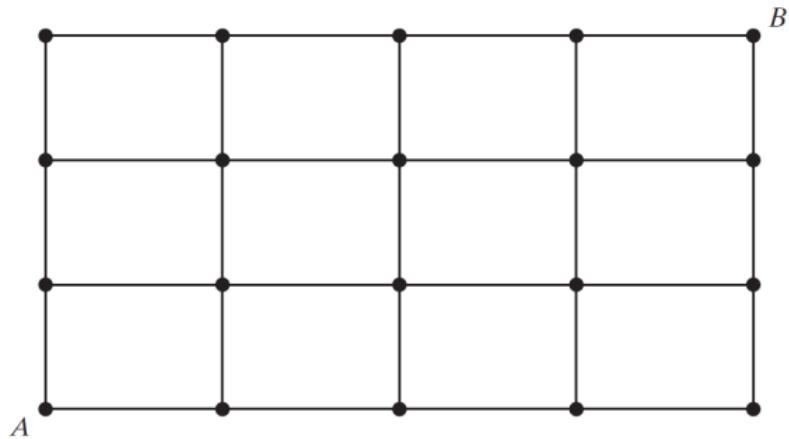


Probability and Statistics

Lecture-3

Exercise: Consider a grid of points shown below. Suppose that, starting at the point A , we can go one step to the right or one step up at each move. This procedure is continued until the point B is reached. How many different paths from A to B are possible?





☞ We shall discuss this during the class!

Question: A police department in a small city consists of 10 officers. If the department policy is to have 5 of the officers patrolling the streets, 2 of the officers working full time at the station, and 3 of the officers on reserve at the station, how many different divisions of the 10 officers into the 3 groups are possible?

 We shall discuss this during the class!

 The above two examples can be generalized as “grouping” problems

Problem: A set of n distinct items is to be divided into r distinct groups of respective sizes n_1, n_2, \dots, n_r , where $n_1 + n_2 + \dots + n_r = n$. How many different divisions are possible?

Solution:

- ▶ There are $\binom{n}{n_1}$ possible choices for the first group
- ▶ For each choice of the first group, there are $\binom{n-n_1}{n_2}$ possible choices for the second group
- ▶ For each choice of the second group, there are $\binom{n-n_1-n_2}{n_3}$ possible choices for the third group and so on.
- ▶ Thus, by basic principle of counting, the number of possible divisions is

$$\binom{n}{n_1} \binom{n-n_1}{n_2} \cdots \binom{n-n_1-n_2-\cdots-n_{r-1}}{n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

Notation: If $n_1 + n_2 + \cdots + n_r = n$, we define

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

The numbers $\binom{n}{n_1, n_2, \dots, n_r}$ are known as **multinomial coefficients**

☞ $\binom{n}{n_1, n_2, \dots, n_r}$ represents the number of possible divisions of n distinct objects into r distinct groups of respective sizes n_1, n_2, \dots, n_r .

☞ When the number of groups, r , is 2 we get the **binomial coefficients!**

Back to probability

Recall:

- ▶ **Sample space** S - set of all outcomes of the experiment
- ▶ **Event** - any subset of S
- ▶ We call an event E to be a **simple event** if E is singleton, i.e., E consists of only one outcome/element

Example - Rolling pair of dice simultaneously.

Recall E_1 and E_2 from the last class!

Then E_2 is a simple event and E_1 is **NOT** a simple event

Union of events

- ▶ For any two events E and F (of the same experiment!), we define the new event $E \cup F$ to consists of all the outcomes that are either in E or F
- ▶ Set theoretically, $E \cup F$ is the union of sets E and F
- ▶ The event $E \cup F$ will occur if **either** E or F occurs
- ▶ We call the event $E \cup F$ to be the **union** of the event E and the event F

Example: Tossing two coins simultaneously

- ▶ Here, $S = \{(H, H), (T, H), (H, T), (T, T)\}$
- ▶ Consider the events $E = \{(H, H), (H, T)\}$ and $F = \{(T, H), (H, H)\}$
- ▶ E is the event that first coin lands heads and F is the event that the second coin lands heads
- ▶ Then, $E \cup F = \{(H, H), (H, T), (T, H)\}$ is the event that at least one of the coins lands heads, i.e., either the first coin is a head or the second coin is a head.

Intersection of events

- ▶ For any two events E and F , **intersection** of E and F , $E \cap F$ is the event that consists of all the outcomes that are both in E and F
- ▶ Thus, the event $E \cap F$ will occur only if both E and F occur

Example: Tossing two coins simultaneously

- ▶ Here, $S = \{(H, H), (T, H), (H, T), (T, T)\}$
- ▶ Consider the events $E = \{(H, H), (H, T), (T, H)\}$ and $F = \{(H, T), (T, H), (T, T)\}$
- ▶ E is the event that at least one head occurs and F is the event that at least one tail occurs
- ▶ Then $E \cap F = \{(H, T), (T, H)\}$ is the event that exactly one head and one tail occur

- ▶ The empty set \emptyset is also a subset of the sample space
- ▶ We call this event \emptyset to be the **Null event**
- ▶ Since there is always an outcome for every experiment, the null event **never occurs!**
- ▶ Two events E and F are said to be **mutually exclusive** if $E \cap F = \emptyset$ i.e., both the events cannot happen at the same time!

Example: Rolling a pair of dice

- ▶ $E = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ - sum of two dice is 7
- ▶ $F = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$ - sum of two dice is 6
- ▶ E and F are mutually exclusive

- ▶ For any event E , the event E^c , the **complement** of E , is the event that consists of all outcomes in the sample space S that are not in E
- ▶ E^c will occur if and only if E does not occur

Example-1: Tossing a coin

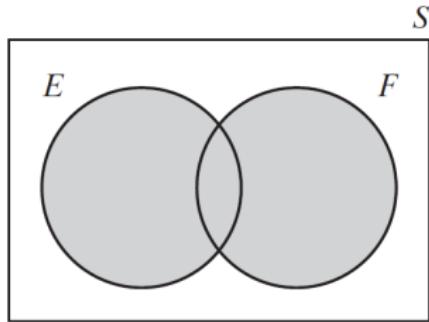
If E is the event that the coin lands heads, then describe E^c

Example-2: Rolling a pair of dice

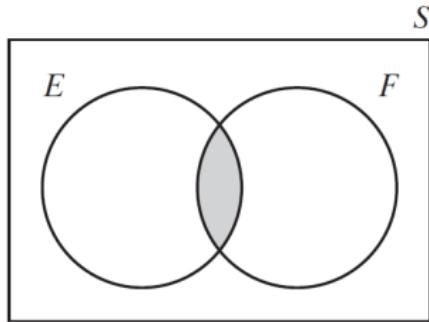
If $E = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ - sum of two dice is 7, then what will be E^c ?

- ★ For any event E and E^c are always mutually exclusive!

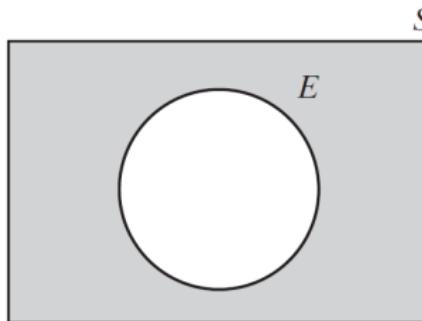
Graphical representation



$E \cup F$



$E \cap F$



E^c

Some useful identities:

- ▶ **Commutative laws** - $E \cup F = F \cup E$ & $E \cap F = F \cap E$
- ▶ **Associative laws** - $(E \cup F) \cup G = E \cup (F \cup G)$ &
 $(E \cap F) \cap G = E \cap (F \cap G)$
- ▶ **Distributive laws** - $(E \cup F) \cap G = (E \cap F) \cup (F \cap G)$ &
 $(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$
- ▶ **DeMorgan's laws** - For any sequence of events E_1, E_2, \dots, E_n

$$\left(\bigcup_{i=1}^n E_i \right)^c = \bigcap_{i=1}^n E_i^c$$

$$\left(\bigcap_{i=1}^n E_i \right)^c = \bigcup_{i=1}^n E_i^c$$

Example: A, B and C take turns flipping a coin. Assume that A flips first, then B , then C , then A and so on. The first one to get a head wins.

- (a) Describe the sample space S
- (b) Let A denote the event of A winning and B denote the event of B winning
 - (i) Describe A and B as subsets of S
 - (ii) Find $(A \cup B)^c$

 We shall discuss this during the class!

Exercise: A system is composed of 5 components, each of which is either working or failed. Consider an experiment that consists of observing the status of each component, and let the outcome of the experiment be given by the vector $(x_1, x_2, x_3, x_4, x_5)$, where x_i is equal to 1 if the i^{th} component is working and is equal to 0 if the i^{th} component is failed

1. How many outcomes are in the sample space of this experiment?
2. Suppose that the system will work if the components 1 and 2 are both working, or if components 3 and 4 are both working, or if components 1, 3 and 5 are all working. Let W be the event that the system will work. Specify all the outcomes in W .
3. Let A be the event that components 4 and 5 are both failed. How many outcomes are contained in the event A ?
4. Write out all the outcomes in the event $A \cap W$.

(S Ross book (ninth edition), page - 48, Problem-5)

Probability and Statistics

Lecture-4

Definition of probability

- 1.** The classical definition
- 2.** The relative frequency definition
- 3.** The subjective probability definition

Recall from lecture-2:

Experiment: Tossing two “fair” coins simultaneously

Sample space $S = \{(H, H), (T, H), (H, T), (T, T)\}$

Events $E_1 = \{(T, H)\}$ and $E_2 = \{(T, H), (H, T)\}$

Guess: $P(E_1) = \frac{|E_1|}{|S|} = \frac{1}{4}$ and $P(E_2) = \frac{|E_2|}{|S|} = \frac{2}{4} = \frac{1}{2}$

1. The Classical definition of probability:

- ▶ In this definition we assume that every outcome in the sample space has an equal chance of occurrence.
- ▶ We term this as **all outcomes in the sample space are equally likely to occur**
- ▶ Under this assumption, for any event E ,

$$P(E) = \frac{|E|}{|S|}$$

☞ **Some examples of sample spaces with equal likely outcomes:**

1. Tossing a fair coin
2. Rolling a fair dice
3. “Random draw” of balls

★ If the outcomes are not equally likely, then we cannot use the classical definition to get the probability of an event.

- ▶ Suppose we have a coin which is **not fair** or **biased**, then how do we find probability of getting a head?
- ▶ Toss the coin, say, 10 times and count the number of times heads has shown up, say, $n(H)$
- ▶ Then, an estimate for $P(\{H\})$ is $\frac{n(H)}{10}$
- ▶ To get a more precise estimate, we increase the number of tosses
- ▶ Thus, if m is the number of tosses and $n(H)$ is the number times heads has turned up in these m tosses, then

$$P(\{H\}) = \lim_{m \rightarrow \infty} \frac{n(H)}{m}$$

- ▶ This is called the **relative frequency definition of probability**

2. The relative frequency definition

- ▶ We suppose that an experiment, whose sample space is S , is repeatedly performed under exactly the same conditions
- ▶ For each event E of the sample space S , we define $n(E)$ to be the number of times in the first m repetitions of the experiment that the event E occurs
- ▶ Then $P(E)$, the probability of the event E , is defined as

$$P(E) = \lim_{m \rightarrow \infty} \frac{n(E)}{m}$$

- ▶ That is, $P(E)$ is defined as the (limiting) proportion of times that E occurs
- ▶ In other words, it is the limiting relative frequency of E

3. The subjective probability definition

- ▶ The probability of an event is a measure of how sure the person making the statement is that the event will happen
- ▶ For instance, after considering all available data, a weather forecaster might say that the probability of rain today is 30% or 0.3
- ▶ This definition gives no rational basis for people to agree on a right answer
- ▶ There is some controversy about when, if ever, to use subjective probability except for personal decision-making.

Summary

- ▶ **Classical definition** - outcomes must be equally likely and we have a sure shot and simple formula for probability
- ▶ **Relative frequency definition** - experiment is repeated large number of times and the frequency of occurrences of events are recorded
- ▶ **Subjective definition** - here probability is one's belief and is too personal/subjective and can only be used for personal decision-making

★ From now on, we assume that we are given a number $P(E)$ for each event E which satisfies the following three axioms

☞ The three axioms of probability

Axiom 1. $0 \leq P(E) \leq 1$ for any event E

Axiom 2. $P(S) = 1$

Axiom 3. For any sequence of mutually exclusive events E_1, E_2, \dots (that is, events for which $E_i \cap E_j = \emptyset$ whenever $i \neq j$),

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

★ We refer to $P(E)$ as the **probability of the event E** .

☞ Using these three axioms, we will prove several facts.

Fact 1. $P(\emptyset) = 0$

Proof. We will use Axiom 3.

Let $E_1 = S$ and $E_i = \emptyset$ for every $i = 2, 3, \dots$

We have $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$

$$\implies P(S) = \sum_{i=1}^{\infty} P(E_i) = P(S) + \sum_{i=2}^{\infty} P(\emptyset)$$

Thus, $P(\emptyset) = 0$



Fact 2. For any event E , $P(E^c) = 1 - P(E)$

Proof. We know that $S = E \cup E^c$ and $E \cap E^c = \emptyset$

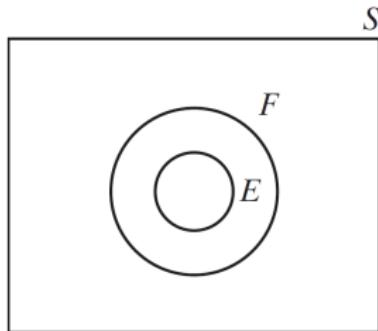
By Axiom 3, $P(S) = P(E) + P(E^c)$ and by Axiom 1, $P(S) = 1$

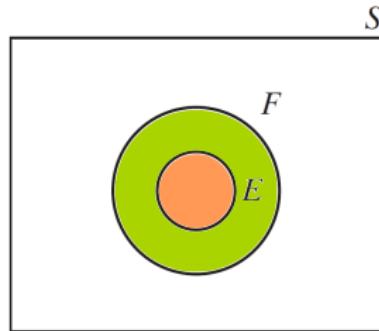
Thus, we get $P(E^c) = 1 - P(E)$.

□

Fact 3. For any $E \subset F$, $P(E) \leq P(F)$

Proof.





$$F = E \cup (E^c \cap F)$$

Further, $E \cap (E^c \cap F) = \emptyset$

By Axiom 3, $P(F) = P(E \cup (E^c \cap F)) = P(E) + P(E^c \cap F)$

By Axiom 1, $P(E^c \cap F) \geq 0$

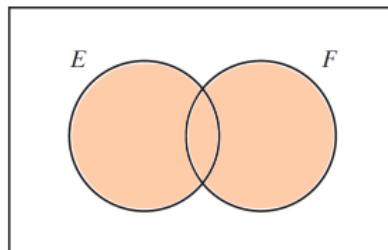
Thus, $P(E) \leq P(F)$.

□

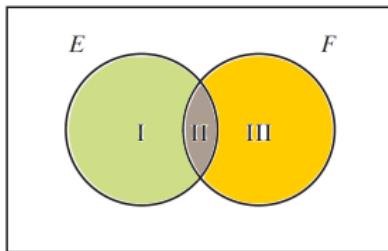
Fact 4. For any two events E and F ,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Proof.



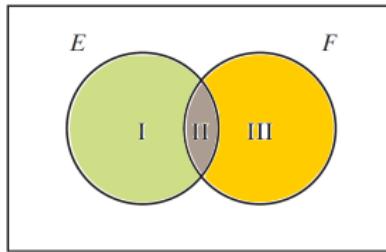
$$E \cup F$$



$$E \cup F = E \cup (E^c \cap F)$$

Since $E \cap (E^c \cap F) = \emptyset$, by Axiom 3,

$$P(E \cup F) = P(E \cup (E^c \cap F)) = P(E) + P(E^c \cap F) \quad \longrightarrow (*)$$



$$F = (E \cap F) \cup (E^c \cap F)$$

Again, by Axiom 3, $P(F) = P(E \cap F) + P(E^c \cap F)$
 $\implies P(E^c \cap F) = P(F) - P(E \cap F)$

Substituting in (*), we get,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

□

Example: J is taking two books along on her holiday vacation. With probability 0.5, she will like the first book; with probability 0.4, she will like the second book; and with probability 0.3, she will like both books. What is the probability that she likes neither of the books?

Let A be the event that J likes the first book and B be the event that J likes the second book

Given that $P(A) = 0.5$, $P(B) = 0.4$ and $P(A \cap B) = 0.3$

We need to find $P(A^c \cap B^c)$

Thus,

$$\begin{aligned} P(A^c \cap B^c) &= P((A \cup B)^c) && \text{(by DeMorgan's law)} \\ &= 1 - P(A \cup B) && \text{(by Fact 2)} \\ &= 1 - [P(A) + P(B) - P(A \cap B)] && \text{(by Fact 4)} \\ &= 1 - 0.5 - 0.4 + 0.3 \end{aligned}$$

$$\implies P(A^c \cap B^c) = 0.4$$

Fact 5:

(a) For any three events E, F and G ,

$$\begin{aligned} P(E \cup F \cup G) = & P(E) + P(F) + P(G) - P(E \cap F) - P(F \cap G) - P(E \cap G) \\ & + P(E \cap F \cap G) \end{aligned}$$

(b) For any set of n events, E_1, E_2, \dots, E_n ,

$$\begin{aligned} P(E_1 \cup E_2 \cup \dots \cup E_n) = & \sum_{i=1}^n P(E_i) - \sum_{i_1 < i_2} P(E_{i_1} \cap E_{i_2}) + \dots \\ & + (-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P(E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_r}) \\ & + \dots + (-1)^{n+1} P(E_1 \cap E_2 \cap \dots \cap E_n) \end{aligned}$$

★ (b) is known as **inclusion-exclusion identity**

Problem: A total of 36 members of a club play tennis, 28 play squash, and 18 play badminton. Furthermore, 22 of the members play both tennis and squash, 12 play both tennis and badminton, 9 play both squash and badminton, and 4 play all three sports. How many members of this club play at least one of three sports?

Solution:

- ▶ Let N be the total number of players in the club
- ▶ The experiment is to randomly select a member of this club
- ▶ Let T be the set of members who play tennis, S be the set of members who play squash, and B be the set of members who play badminton.
- ▶ We have $P(T) = \frac{36}{N}$, $P(S) = \frac{28}{N}$, $P(B) = \frac{18}{N}$, $P(T \cap S) = \frac{22}{N}$, $P(T \cap B) = \frac{12}{N}$, $P(S \cap B) = \frac{9}{N}$, and $P(T \cap S \cap B) = \frac{4}{N}$
- ▶ We need to find $P(T \cup S \cup B)$

$$\begin{aligned}P(T \cup S \cup B) &= P(T) + P(S) + P(B) - P(T \cap S) - P(S \cap B) - P(B \cap T) \\&\quad + P(T \cap S \cap B) \\&= \frac{36 + 28 + 18 - 22 - 12 - 9 + 4}{N} \\&= \frac{43}{N}\end{aligned}$$

⇒ Thus 43 members play at least one of the sports.



Probability and Statistics

Lecture-5

The birthday problem

Problem: If n people are present in a room, what is the probability that at least two of them celebrate their birthday on the same day of the year? How large need n be so that this probability is greater than $\frac{1}{2}$?

Solution:

- ▶ Each person can celebrate his/her birthday on one of the 365 days, there are $(365)^n$ possible outcomes
- ▶ Thus, the sample space S has $(365)^n$ points
- ▶ Let E be the event that at least two persons have the same birthday
- ▶ Then, E^c is the event that no two persons have the same birthday

- ▶ Then the number of outcomes in E^c is $365 \times 364 \times \dots \times (365 - (n - 1))$
- ▶ Thus the probability that no two persons share the same birthday is

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

Now, check that $\frac{365 \times 364 \times \dots \times (365 - (n - 1))}{(365)^n} \leq \frac{1}{2} \iff n \geq 23$

- ▶ That is, $P(E^c) \leq \frac{1}{2} \iff n \geq 23$
- ▶ Which implies, $P(E) \geq \frac{1}{2} \iff n \geq 23$
- ▶ That is, just 23 people are enough in a room for the probability of atleast two people sharing the same birthday to exceed $\frac{1}{2}$!

Experiment: Rolling a pair of “fair” dice

- ▶ Sample space $S = \{(i,j) : i,j = 1, 2, 3, 4, 5, 6\}$ and $|S| = 36$
- ▶ What is the probability of getting a sum of 8?
- ▶ Possible outcomes are $\{(2,6), (3,5), (4,4), (5,3), (6,2)\}$ and hence probability is $\frac{5}{36}$
- ▶ Suppose that we got some additional information that the first die landed on side 3
- ▶ With this information, what is the probability that sum of the two dice equals 8? is it the same?

- ▶ Since we know that the first die landed on 3, all such possible outcomes are $\{(3,1), (3,2), (3,3), (3,5), (3,6)\}$ and every other outcome has zero probability
 - ▶ Each of these outcomes are equally likely and hence $P((3,j)) = \frac{1}{6}$ for each $j = 1, 2, 3, 4, 5, 6$
 - ▶ Our desired outcome is $(3,5)$ and hence the probability is $\frac{1}{6}$
 - ▶ If the condition was not given, then the probability of getting a sum of 8 was $\frac{5}{36}$
 - ▶ Thus, the prior information/condition has changed the probability of getting a sum of 8!
 - ▶ Such probability is referred to as **conditional probability**.
- ☞ Suppose if we denote by F the event of getting 3 on the first die, and E the event of getting a sum of 8, then $\frac{1}{6} = \frac{P(E \cap F)}{P(F)}$

We define $P(E|F) = \frac{P(E \cap F)}{P(F)}$ and call it the **conditional probability of E given that F has occurred**

☞ Observe that, since F has already occurred $P(F) > 0$ and the definition makes perfect sense.

Example: A bin contains 5 defective (that immediately fail when put in use), 10 partially defective (that fail after a couple of hours of use), and 25 acceptable transistors.

A transistor is chosen at random from the bin and put into use. If it does not immediately fail, what is the probability it is acceptable?

Solution:

- ▶ E be the event that the transistor chosen is defective
- ▶ F be the event that the transistor chosen is partially defective
- ▶ G be the event that the transistor chosen is acceptable

- ▶ We are given that the transistor chosen did not fail immediately.
- ▶ That is, we are given the information that $(F \cup G)$ has happened.
- ▶ Given this information we need find probability that the transistor is acceptable.
- ▶ That is, we need to find $P(G|(F \cup G))$

$$\begin{aligned} P(G|(F \cup G)) &= \frac{P(G \cap (F \cup G))}{P(F \cup G)} \\ &= \frac{P(G)}{P(F \cup G)} \end{aligned}$$

- ▶ We have $P(E) = \frac{5}{40}$, $P(F) = \frac{10}{40}$ and $P(G) = \frac{25}{40}$
- ▶ $P(F \cap G) = 0$ as a transistor cannot be both partially defective and acceptable at the same time

- ▶ Thus, $P(F \cup G) = P(F) + P(G) = \frac{35}{40}$
- ▶ Hence, $P(G|(F \cup G)) = \frac{25/40}{35/40} = \frac{5}{7}$

Example: Joe is 80% certain that his missing key is in one of the two pockets of his hanging jacket, being 40% certain it is in the left-hand pocket and 40% certain it is in the right-hand pocket. If a search of the left-hand pocket does not find the key, what is the conditional probability that it is in the other pocket?

Solution:

- ▶ L be the event that the key is in the left-hand pocket of the jacket and R be the event that it is in the right-hand pocket

We are given the information $P(R) = 0.4$ and $P(L) = 0.4$.

The desired probability is $P(R|L^c)$

- ▶ We have $P(R|L^c) = \frac{P(R \cap L^c)}{P(L^c)}$
- ▶ Observe that $R \cap L = \emptyset$ and hence $R \cap L^c = R$
- ▶ Thus, $P(R|L^c) = \frac{P(R)}{1-P(L)} = \frac{0.4}{1-0.4} = \frac{2}{3}$

Example: Consider an urn containing 12 balls, of which 8 are white. A sample size of 4 is to be drawn randomly with replacement.

- (a) What is the conditional probability that the first and third balls drawn will be white given that the sample drawn contains exactly 3 white balls?
- (b) What will be the probability if the balls are drawn without replacement?

Solution: Let us assume that the remaining 4 balls are red.

Define the sample space

$$S = \{(x_1 x_2 x_3 x_4) : x_i = W \text{ or } R \text{ whether the } i^{\text{th}} \text{ ball is white or red}\}$$

Since the draw is random, each outcome of the sample space is equally likely

E be the event that the first and third balls are white

F be the event that there are exactly 3 white balls in the sample

We need to find $P(E|F) = \frac{P(E \cap F)}{P(F)}$

(a) In this case, clearly, $|S| = 12^4$

Now, $WWWR, WWRW, WRWW, RWWW$ are possible sequence of balls for the outcomes to be in F

Since each of this sequence is possible in $8^3 \times 4$ ways,

$$|F| = 4 \times 8^3 \times 4 = 8^3 \times 4^2$$

$$\implies P(F) = \frac{8^3 \times 4^2}{12^4}$$

$WWWR, WRWW$ are the only possible sequences for the outcomes $E \cap F$

Thus, $|E \cap F| = 2 \times 8^3 \times 4$ and $P(E \cap F) = \frac{8^3 \times 4^2 \times 2}{12^4}$

$$\implies P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{1}{2}$$

(b) In this case of drawing balls without replacement,

$$|S| = 12 \times 11 \times 10 \times 9$$

Here again, $WWWR$, $WWRW$, $WRWW$, $RWWW$ will be the possible sequence of balls for the outcomes to be in F

Since each of this sequence is possible in $8 \times 7 \times 6 \times 4$ ways,

$$|F| = 4 \times 8 \times 7 \times 6 \times 4$$

$$\implies P(F) = \frac{8 \times 7 \times 6 \times 4^2}{12 \times 11 \times 10 \times 9}$$

$WWWR$, $WRWW$ are the only possible sequences for the outcomes $E \cap F$

Thus, $|E \cap F| = 2 \times 8 \times 7 \times 6 \times 4$ and $P(E \cap F) = \frac{2 \times 8 \times 7 \times 6 \times 4}{12 \times 11 \times 10 \times 9}$

$$\implies P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{1}{2}$$



Problem: A total of 500 married working couples were polled about their annual salaries, with the following information resulting.

		Husband	
Wife		Less than \$50,000	More than \$50,000
Less than \$50,000	Less than \$50,000	212	198
	More than \$50,000	36	54

Thus, for instance, in 36 of the couples the wife earned more and the husband earned less than \$50,000. If one of the couples is randomly chosen, what is

- (a) the probability that the husband earns less than \$50,000;
- (b) the conditional probability that the wife earns more than \$50,000 given that the husband earns more than this amount;
- (c) the conditional probability that the wife earns more than \$50,000 given that the husband earns less than this amount?

Solution:

Let E be the event that the husband in the couple chosen earns less than \$50,000 and F be the event that the wife in the couple chosen earns less than \$50,000

- (a) Out of 500 couples, $212 + 36 = 248$ couples are such that the husband earns less than \$50,000

Thus, $P(E) = \frac{248}{500}$

- (b) Here, we need to find $P(F^c|E^c)$

Now, $E^c \cap F^c$ is the event that both wife and husband of the chosen couple earn more than \$50,000

From the given data, $P(E^c \cap F^c) = \frac{54}{500}$

Further, $P(E^c) = \frac{198+54}{500} = \frac{252}{500}$

Thus, $P(F^c|E^c) = \frac{54}{252}$

(c) We need to find $P(F^c|E)$

We have, $P(E \cap F^c) = \frac{36}{500}$ and $P(E) = \frac{248}{500}$

Thus, $P(F^c|E) = \frac{36}{248}$



Conditional probability of E given F is given by,

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

☞ Conditional probabilities satisfy all of the properties of ordinary probabilities!

☞ For an event F ,

- ▶ $0 \leq P(E|F) \leq 1$ for any event E
- ▶ $P(S|F) = 1$
- ▶ For E_1, E_2, \dots with $E_i \cap E_j = \emptyset$ whenever $i \neq j$,

$$P\left(\bigcup_{i=1}^{\infty} E_i | F\right) = \sum_{i=1}^{\infty} P(E_i | F)$$

 Hence, we get:

- ▶ $P(\emptyset|F) = 0$
- ▶ $P(E^c|F) = 1 - P(E|F)$
- ▶ For any $A \subset E$, $P(A|F) \leq P(E|F)$
- ▶ For any two events A and B ,
$$P((A \cup B)|F) = P(A|F) + P(B|F) - P((A \cap B)|F)$$
- ▶ For any set of n events, E_1, E_2, \dots, E_n ,

$$\begin{aligned} P((E_1 \cup E_2 \cup \dots \cup E_n)|F) &= \sum_{i=1}^n P(E_i|F) - \sum_{i_1 < i_2} P((E_{i_1} \cap E_{i_2})|F) + \dots \\ &\quad + (-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P((E_{i_1} \cap \dots \cap E_{i_r})|F) \\ &\quad + \dots + (-1)^{n+1} P((E_1 \cap \dots \cap E_n)|F) \end{aligned}$$

Probability and Statistics

Lecture-6

- ▶ $P(E|F)$ is in general not equal to $P(E)$
- ▶ The occurrence of F may affect the occurrence of E
- ▶ When does $P(E|F) = P(E)$?
- ▶ When does occurrence of F has no effect on occurrence of E ?
- ▶ We say that the events E and F are independent if
 $P(E|F) = P(E)$

$$P(E|F) = P(E) \iff \frac{P(E \cap F)}{P(F)} = P(E) \iff P(E \cap F) = P(E)P(F)$$

Two events E and F are said to be **independent** if $P(E \cap F) = P(E)P(F)$

☞ We say that E and F are **dependent** if they are not independent.

Example: A card is selected at random from an ordinary deck of 52 playing cards. If A is the event that the selected card is an ace and H is the event that it is a heart.

Are A and H independent?

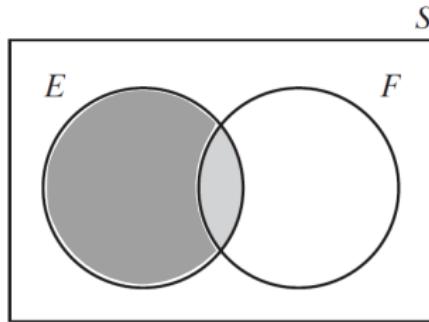
Solution: We have $P(A) = \frac{4}{52}$, $P(H) = \frac{13}{52}$ and $P(A \cap H) = \frac{1}{52}$

$$\implies P(A)P(H) = P(A \cap H)$$

Thus, A and H are independent.

Fact 6. If E and F are independent, then so are E and F^c

Proof: We need to show that $P(E \cap F^c) = P(E)P(F^c)$



- ▶ $E = (E \cap F) \cup (E \cap F^c)$
- ▶ $(E \cap F) \cap (E \cap F^c) = \emptyset$

$$\begin{aligned}\implies P(E) &= P(E \cap F) + P(E \cap F^c) \\ &= P(E)P(F) + P(E \cap F^c) \\ &\quad (\text{since } E \text{ and } F \text{ are independent})\end{aligned}$$

$$\implies P(E \cap F^c) = P(E)(1 - P(F)) = P(E)P(F^c)$$

That is, E and F^c are independent!



- ▶ Suppose now that E is independent of F and E is also independent of G .
- ▶ Is E then necessarily independent of $F \cap G$?
- ▶ The answer is **no!**

Example: Two fair dice are thrown. Let E denote the event that the sum of the dice is 7. Let F denote the event that the first die equals 4 and G denote the event that the second die equals 3.

- ▶ $P(E) = \frac{1}{6}, P(F) = \frac{1}{6}$ and $P(G) = \frac{1}{6}$
- ▶ $P(E \cap F) = \frac{1}{36}, P(E \cap G) = \frac{1}{36}, P(F \cap G) = \frac{1}{36}$
- ▶ $P(E \cap (F \cap G)) = \frac{1}{36}$
- ▶ $P(E \cap F) = P(E)P(F), P(E \cap G) = P(E)P(G)$
- ▶ But $P(E \cap (F \cap G)) \neq P(E)P(F \cap G)$
- ▶ E is independent of F and is also independent of G but E is **not independent** of $F \cap G$

☞ Three events E, F, G are said to be independent if **all** of the following are satisfied:

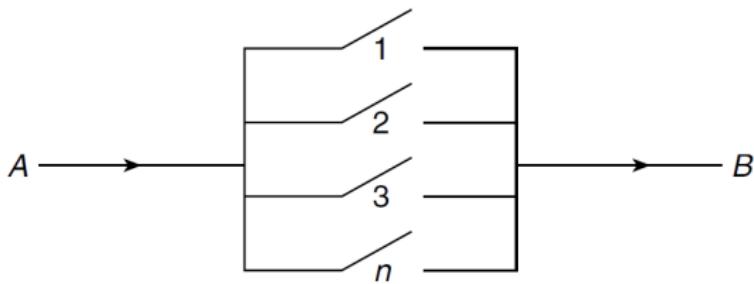
- ▶ $P(E \cap F \cap G) = P(E)P(F)P(G)$
- ▶ $P(E \cap F) = P(E)P(F)$
- ▶ $P(E \cap G) = P(E)P(G)$
- ▶ $P(F \cap G) = P(F)P(G)$

☞ The events E_1, E_2, \dots, E_n are said to be independent if

$$P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_r}) = P(E_{i_1})P(E_{i_2})\cdots P(E_{i_r})$$

for any $1 \leq r \leq n$ and $i_1, i_2, \dots, i_r \in \{1, 2, 3, \dots, n\}$

Problem: A system composed of n separate components is said to be a *parallel system* if it functions when at least one of the components functions. For such a system, if component i , independent of other components, functions with probability p_i , $i = 1, \dots, n$, what is the probability the system functions?



Solution: Let A_i denote the event that component i functions.

$$\begin{aligned}P(\text{system functions}) &= 1 - P(\text{system does not function}) \\&= 1 - P(\text{all components do not function}) \\&= 1 - P(A_1^c \cap A_2^c \cap \dots \cap A_n^c) \\&= 1 - P(A_1^c)P(A_2^c)\dots P(A_n^c) \quad (\text{independence}) \\&= 1 - (1 - p_1)(1 - p_2)\dots(1 - p_n)\end{aligned}$$

- ▶ Experiment under consideration may consist of performing a sequence of sub-experiments
- ▶ Examples - tossing a coin 10 times, tossing 3 coins and rolling a pair of dice etc.
- ▶ We may think of each toss of the coin or roll of the die is a sub-experiment
- ▶ In many cases, we may assume that the sub-experiments are independent of each other
- ▶ Two consecutive tosses can be assumed to be independent of each other
- ▶ Tossing a coin and rolling a die can be assumed to be independent of each other

- ▶ If each sub-experiment has the same set of possible outcomes, then the sub-experiments are often called **trials**
- ▶ For example, in the experiment of tossing a coin 10 times, each coin toss is a trial. Here, they are **independent trials!**
- ▶ If the coin tosses are independent of each other then the following two are equivalent:
 - ▶ Tossing a single coin n number of times
 - ▶ Tossing n identical coins at the same time

Problem: An infinite sequence of independent trials is to be performed. Each trial results in a success with probability p and a failure with probability $1 - p$. What is the probability that

- (a) at least 1 success occurs in the first n trials;
- (b) exactly k successes occur in the first n trials?

Solution: (a) We need to determine the probability of at least 1 success in the first n trials.

It is easy to compute first the probability of the complementary event

That is, we will first determine the probability of no successes in the first n trials

If we let E_i denote the event of a failure on the i^{th} trial, then the probability of no successes is

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2) \cdots P(E_n) = (1 - p)^n$$

Thus, the probability of atleast one success in n trials is $1 - (1 - p)^n$

(b) Consider any particular sequence of the first n outcomes containing k successes and $n - k$ failures

Each one of these sequences will occur with probability $p^k(1 - p)^{n-k}$ by independence

Further, there are $\binom{n}{k}$ such sequences

Thus, the desired probability is $\binom{n}{k}p^k(1 - p)^k$



- ▶ We have $P(E|F) = \frac{P(E \cap F)}{P(F)}$
- ▶ Since $P(F) \neq 0$, multiplying $P(F)$ on both the sides gives,

$$P(E \cap F) = P(F)P(E|F)$$

- ▶ This equation is quite useful in computing probability of intersection of events

Example: Rachel is undecided as to whether to take a French course or a chemistry course. She estimates that her probability of receiving an A grade would be $\frac{1}{2}$ in a French course and $\frac{2}{3}$ in a chemistry course. If Rachel decides to base her decision on the flip of a fair coin, what is the probability that she gets an A in chemistry?

Solution: Let E be the event that Rachel takes chemistry and F be the event that she receives an A grade.

We need to find $P(E \cap F)$

Since she chooses Chemistry or French by a flip of fair coin,

$$P(E) = \frac{1}{2} \text{ and given that } P(F|E) = \frac{2}{3}$$

$$\text{We have } P(E \cap F) = P(E)P(F|E) \text{ and hence } P(E \cap F) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$$

 **The multiplication rule:** For any events E_1, E_2, \dots, E_n ,

$$P(E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2) \dots P(E_n|E_1 \dots E_{n-1})$$

Proof. Expand the right hand side using definition of conditional probability and simplify

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

Solution:

	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													

Define events $E_i, i = 1, 2, 3, 4$, as follows:

$E_1 = \{\text{the ace of spades is in any one of the piles}\}$

$E_2 =$

$\{\text{the ace of spades and the ace of hearts are in different piles}\}$

$E_3 =$

$\{\text{the aces of spades, hearts, and diamonds are all in different piles}\}$

$E_4 = \{\text{all 4 aces are in different piles}\}$

The desired probability is $P(E_4) = P(E_1 \cap E_2 \cap E_3 \cap E_4)$

By multiplication rule,

$$P(E_1 \cap E_2 \cap E_3 \cap E_4) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2)P(E_4|E_1 \cap E_2 \cap E_3)$$

- ▶ Observe that E_1 is the sample space itself and hence $P(E_1) = 1$
- ▶ The event $E_2|E_1$ is “given that the ace of spades is in one of the piles, the ace of the hearts is in a different pile”
- ▶ Its compliment $E_2^c|E_1$ is the event “given that the ace of spades is in one of the piles, the ace of hearts is in the same pile”
- ▶ In the pile in which the ace of spades is present, there are 12 remaining cards and hence $P(E_2^c|E_1) = \frac{12}{51}$
 $\implies P(E_2|E_1) = 1 - \frac{12}{51} = \frac{39}{51}$

- ▶ Now, $E_3|E_1 \cap E_2$ is the event “given that ace of spades and ace of hearts are in different piles, the ace of diamonds is in the third pile different from these two piles”
- ▶ By similar argument as above, $P(E_3^c|E_1 \cap E_2) = \frac{24}{50}$ and hence $P(E_3|E_1 \cap E_2) = 1 - \frac{24}{50} = \frac{26}{50}$
- ▶ Again, using a similar argument, we get
 $P(E_4|E_1 \cap E_2 \cap E_3) = 1 - \frac{36}{49} = \frac{13}{49}$

Thus, $P(E_1 \cap E_2 \cap E_3 \cap E_4) = \frac{39 \cdot 26 \cdot 13}{51 \cdot 50 \cdot 49} \approx 0.105$

- ▶ Thus, there is approximately 10.5% chance that each pile will contain an ace.

Probability and Statistics

Lecture-7

- ▶ **Multiplication rule:** For two events, E and F ,

$$P(E \cap F) = P(E|F)P(F)$$

Now, consider $P(F|E) = \frac{P(E \cap F)}{P(E)}$

$$= \frac{P(E|F)P(F)}{P(E)} \quad (\text{multiplication rule})$$

Bayes' theorem/formula/rule:

For any two events E and F with $P(E) \neq 0$,

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

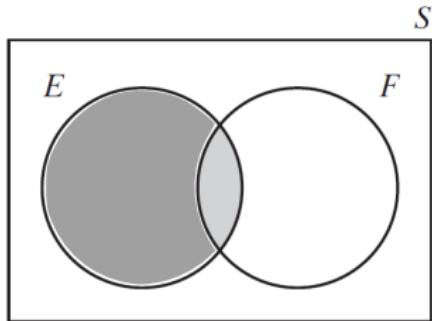
☞ Conditional probability + multiplication rule \leftrightarrow Bayes' theorem

Example: 1% of people have a certain genetic defect. 90% of tests for the gene detect the defect (true positives). Irrespective of the defect, 10.4% of tests show positive. If a person gets a positive test result, what is the probability that they actually have the genetic defect?

- ▶ Let D denote the event that a person has the genetic defect
- ▶ T^+ denote the event that the test shows positive
- ▶ We are given $P(T^+|D) = 0.9$, $P(D) = 0.01$ and $P(T^+) = 0.104$
- ▶ We need to find $P(D|T^+)$

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+)} \quad (\text{Bayes' Theorem})$$
$$= 0.0865$$





$$E = (E \cap F) \cup (E \cap F^c) \text{ and } (E \cap F) \cap (E \cap F^c) = \emptyset$$

$$\implies P(E) = P(E \cap F) + P(E \cap F^c)$$

$$\implies P(E) = P(E|F)P(F) + P(E|F^c)P(F^c)$$

(Multiplication rule)

Problem: You ask your neighbour to water a sickly plant while you are on vacation. Without water it will die with probability 0.8; with water it will die with probability 0.15. You are 90 percent certain that your neighbour will remember to water the plant.

- (a) What is the probability that the plant will be alive when you return?
- (b) If it is dead, what is the probability your neighbor forgot to water it?

Solution: Let A be the event that the plant is alive after return and W be the event that the plant receives water during the vacation

We are given $P(A^c|W^c) = 0.8$, $P(A^c|W) = 0.15$ and $P(W) = 0.9$

$$P(A|W^c) = 1 - P(A^c|W^c) = 0.2,$$

$$P(A|W) = 1 - P(A^c|W) = 0.85$$

$$P(W^c) = 1 - P(W) = 0.1$$

☞ For (a), we need to find $P(A)$

$$\begin{aligned}P(A) &= P(A|W)P(W) + P(A|W^c)P(W^c) \\&\implies P(A) = 0.85 \times 0.9 + 0.2 \times 0.1 = 0.785\end{aligned}$$

☞ In (b), we need to find $P(W^c|A^c)$

$$\begin{aligned}P(W^c|A^c) &= \frac{P(W^c \cap A^c)}{P(A^c)} \\&= \frac{P(W^c)P(A^c|W^c)}{P(A^c)} \quad (\text{multiplication rule}) \\&= \frac{(1 - P(W))P(A^c|W^c)}{1 - P(A)} \\&= \frac{0.1 \times 0.8}{0.215} \\&= 0.372\end{aligned}$$

Example: There are two local factories that produce microwaves. Each microwave produced at factory A is defective with probability 0.05, whereas each one produced at factory B is defective with probability 0.01. Suppose you purchase two microwaves that were produced at the same factory, which is equally likely to have been either factory A or factory B . If the first microwave that you check is defective, what is the conditional probability that the other one is also defective?

Solution:

- ▶ D_i be the event that microwave i is defective ($i = 1, 2$)
- ▶ A and B be the events that the microwaves were produced at factory A and B respectively
- ▶ We need to find $P(D_2|D_1)$

$$\begin{aligned}
 P(D_2|D_1) &= \frac{P(D_1 \cap D_2)}{P(D_1)} \\
 &= \frac{P((D_1 \cap D_2)|A)P(A) + P((D_1 \cap D_2)|B)P(B)}{P(D_1|A)P(A) + P(D_1|B)P(B)} \\
 &= \frac{P(D_1|A)P(D_2|A)P(A) + P(D_1|B)P(D_2|B)P(B)}{P(D_1|A)P(A) + P(D_1|B)P(B)}
 \end{aligned}$$

- ▶ Since it is equally likely that the microwave is from factory A or factory B , $P(A) = P(B) = 0.5$
- ▶ It is given that, $P(D_i|A) = 0.05$ and $P(D_i|B) = 0.01$ for each $i = 1, 2$

$$\begin{aligned}
 \implies P(D_2|D_1) &= \frac{(0.05)^2(0.5) + (0.01)^2(0.05)}{(0.05)(0.5) + (0.01)(0.5)} \\
 &= 0.043
 \end{aligned}$$

- ▶ In both the above examples, we used the following (in some form):

$$P(E) = P(E|F)P(F) + P(E|F^c)P(F^c) \text{ for any two events } E \text{ and } F$$

- ▶ Let $F_1 = F$ and $F_2 = F^c$
- ▶ Then, $F_1 \cup F_2 = S$ (sample space), $F_1 \cap F_2 = \emptyset$ and

$$P(E) = P(E|F_1)P(F_2) + P(E|F_2)P(F_2)$$

- ▶ We can generalise this!

- Let F_1, F_2, \dots, F_n are mutually exclusive and exhaustive events, that is,

$$F_i \cap F_j = \emptyset \text{ for } i \neq j \text{ and } \bigcup_{i=1}^n F_i = S$$

$$\implies E \cap \left(\bigcup_{i=1}^n F_i \right) = E \cap S$$

$$\implies \bigcup_{i=1}^n (E \cap F_i) = E$$

- Since the events $E \cap F_i, i = 1, 2, 3, \dots, n$ are mutually exclusive,

$$\begin{aligned} P(E) &= P\left(\bigcup_{i=1}^n (E \cap F_i)\right) = \sum_{i=1}^n P(E \cap F_i) \\ &= \sum_{i=1}^n P(E|F_i)P(F_i) \quad (\textbf{multiplication rule}) \end{aligned}$$

For events F_1, F_2, \dots, F_n such that

$$F_i \cap F_j = \emptyset \text{ for } i \neq j \text{ and } \bigcup_{i=1}^n F_i = S,$$

the equation

$$P(E) = \sum_{i=1}^n P(E|F_i)P(F_i)$$

is called the **law of total probability**

Theorem (Generalised Bayes' theorem)

Suppose that F_1, F_2, \dots, F_n are mutually exclusive events such that

$$\bigcup_{i=1}^n F_i = S \text{ (mutually exhaustive)}$$

$$\begin{aligned} \text{Then, for any event } E, P(F_j|E) &= \frac{P(E \cap F_j)}{P(E)} \\ &= \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^n P(E|F_i)P(F_i)} \end{aligned}$$

Example: A particular test for whether someone has been using cannabis gives 90% true "positive" results (meaning, "Yes he used cannabis") for cannabis users and 80% true negative results for non-users. Assuming 5% of people actually do use cannabis, what is the probability that a random person who tests positive is really a cannabis user?

- ▶ Let U denote the event of a person being a user of cannabis and $T+$ (resp. $T-$) be the event of test showing positive (resp. negative).
- ▶ We are given $P(T+|U) = 0.9$, $P(T-|U^c) = 0.8$ and $P(U) = 0.05$
- ▶ We need to find $P(U|T+)$

- ▶ By Bayes' theorem,

$$\begin{aligned} P(U|T+) &= \frac{P(T+|U)P(U)}{P(T+)} \\ &= \frac{P(T+|U)P(U)}{P(T+|U)P(U) + P(T+|U^c)P(U^c)} \\ &= \frac{0.9 \times 0.05}{0.9 \times 0.05 + (1 - 0.8) \times (1 - 0.05)} \\ &\approx 0.19 \end{aligned}$$

□

Problem: Suppose that we have 3 cards that are identical in form, except that both sides of the first card are colored red, both sides of the second card are colored black, and one side of the third card is colored red and the other side black.

The 3 cards are mixed up in a hat, and 1 card is randomly selected and put down on the ground. If the upper side of the chosen card is colored red, what is the probability that the other side is colored black?

Solution: Let RR , BB and RB denote, respectively, the events that the chosen card is all red, all black, or the red-black card.

Further, let R be the event that the upturned side of the chosen card is red.

Then, we need to find $P(RB|R)$

$$\begin{aligned}
P(RB|R) &= \frac{P(RB \cap R)}{P(R)} \\
&= \frac{P(R|RB)P(RB)}{P(R|RR)P(RR) + P(R|RB)P(RB) + P(R|BB)P(BB)} \\
&= \frac{\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)}{\left(1\right)\left(\frac{1}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + 0\left(\frac{1}{3}\right)} \\
&= \frac{1}{3}
\end{aligned}$$

□

Problem: A plane is missing and it is presumed that it was equally likely to have gone down in any of three possible regions. Let $1 - \alpha_i$ denote the probability the plane will be found upon a search of the i^{th} region when the plane is, in fact, in that region, $i = 1, 2, 3$.

What is the conditional probability that the plane is in the i^{th} region, given that a search of region 1 is unsuccessful, $i = 1, 2, 3$?

Solution: Let $R_i, i = 1, 2, 3$, be the event that the plane is in region i ; and let E be the event that a search of region 1 is unsuccessful.

Given data - $P(E^c|R_1) = 1 - \alpha_1$, $P(R_i) = \frac{1}{3}, i = 1, 2, 3$.

We need to find $P(R_i|E)$ for each $i = 1, 2, 3$

Given that the plane is found in region 2 or 3, then there is no way it can be found in region 1.

Thus, $P(E|R_2) = 1 = P(E|R_3)$.

Thus, by Bayes' formula,

$$\begin{aligned} P(R_1|E) &= \frac{P(E|R_1)P(R_1)}{P(E|R_1)P(R_1) + P(E|R_2)P(R_2) + P(E|R_3)P(R_3)} \\ &= \frac{\alpha_1 \left(\frac{1}{3}\right)}{\alpha_1 \left(\frac{1}{3}\right) + (1) \left(\frac{1}{3}\right) + (1) \left(\frac{1}{3}\right)} \\ &= \frac{\alpha_1}{\alpha_1 + 2} \end{aligned}$$

Now, for $j = 2, 3$

$$\begin{aligned} P(R_j|E) &= \frac{P(E|R_j)P(R_j)}{P(E|R_1)P(R_1) + P(E|R_2)P(R_2) + P(E|R_3)P(R_3)} \\ &= \frac{(1) \left(\frac{1}{3}\right)}{\alpha_1 \left(\frac{1}{3}\right) + (1) \left(\frac{1}{3}\right) + (1) \left(\frac{1}{3}\right)} \\ &= \frac{1}{\alpha_1 + 2} \end{aligned}$$

Probability and Statistics

Lecture-8

Look at these examples

- (1) In the experiment of rolling two dice, consider the event E that the sum of dice is 7
- (2) In the experiment of tossing 4 coins, consider the event F that the outcome has at least 2 heads
- (3) Four balls are randomly selected, without replacement, from an urn containing 20 balls numbered 1 through 20. Consider the event G that out of the 4 selected balls, the largest number is 10

(1) In the experiment of rolling two dice, consider the event E that the sum of dice is 7

☞ Here, we are interested in the “sum” of the dice and **not** in the actual outcomes of individual dice

(2) In the experiment of tossing 4 coins, consider the event F that the outcome has at least 2 heads

☞ Here, we are interested in the “number of total heads” and **not** in the actual outcome of individual toss

(3) Four balls are randomly selected, without replacement, from an urn containing 20 balls numbered 1 through 20. Consider the event G that out of the 4 selected balls, the largest number is 10

☞ Here, we are interested in the “largest numbered ball” and **not** in the actual sample of 4 balls

(1) In the experiment of rolling two dice, consider the event E that the sum of dice is 7

☞ Here, we are interested in the “sum” of the dice and **not** in the actual outcomes of individual dice

Once the two dice are rolled, let the **variable** X denote the sum of dice

Then, $P(E) = P\{X = 7\}$

(2) In the experiment of tossing 4 coins, consider the event F that the outcome has at least 2 heads

☞ Here, we are interested in the “number of total heads” and **not** in the actual outcome of individual toss

If we denote by the variable Y , the number of heads in the 4 tosses, then

$$P(F) = P\{Y \geq 2\}$$

(3) Four balls are randomly selected, without replacement, from an urn containing 20 balls numbered 1 through 20. Consider the event G that out of the 4 selected balls, the largest number is 10

☞ Here, we are interested in the “largest numbered ball” and **not** in the actual sample of 4 balls

Define the variable Z to be the largest number among the four selected balls

Then, $P(G) = P\{Z = 10\}$

- ▶ In all the examples, we were interested in a “variable” whose value is dependent on the outcome of the experiment and we did not care about what the actual outcomes were!
- ▶ These quantities of interest are known as **random variables**

Given an experiment whose sample space is S , a **random variable** X is real-valued function defined on the sample space S .

$$\text{i.e., } X : S \rightarrow \mathbb{R}$$

Example: Suppose that our experiment consists of tossing 3 fair coins. If we let Y denote the number of heads that appear in the three tosses.

What are the values Y can take?

$$Y = 0, 1, 2, 3$$

$$P\{Y = 0\} = P\{(T, T, T)\} = \frac{1}{8}$$

$$P\{Y = 1\} = P\{(T, T, H), (T, H, T), (H, T, T)\} = \frac{3}{8}$$

$$P\{Y = 2\} = P\{(T, H, H), (H, T, H), (H, H, T)\} = \frac{3}{8}$$

$$P\{Y = 3\} = P\{(H, H, H)\} = \frac{1}{8}$$

Example: A telecom company wishes to analyse length of phone calls. Let X be the length of a randomly selected telephone call.

What are the values X can take?

Example: Let Y be the height of a randomly selected person in a city.

What values can Y take?

Example: Let Z denote the time taken to execute a particular program.

What values can Z take?

Example: Four balls are randomly selected, without replacement, from an urn containing 20 balls numbered 1 through 20.

If X is the largest numbered ball among the selected balls, then X is a random variable

What values can X take?

$$X = 4, 5, \dots, 20$$

Question: For $i \geq 4$, $P\{X = i\} = ?$

$X = i \iff$ the ball numbered i is one of the 4 selected balls and we have to choose remaining 3 balls from the balls numbered $1, 2, \dots, (i-1)$.

$$\text{Thus, } P\{X = i\} = \frac{\binom{i-1}{3}}{\binom{20}{4}}$$

Question: $P\{X > 13\} = ?$

We shall calculate $P\{X > 13\}$ in two ways.

Method-1:

We have $\{X > 13\} = \{X = 14\} \cup \{X = 15\} \cup \dots \cup \{X = 20\}$ and
 $\{X = i\} \cap \{X = j\} = \emptyset$ for $i \neq j$

$$\text{Thus, } P\{X > 13\} = \sum_{i=14}^{20} P\{X = i\} = \sum_{i=14}^{20} \frac{\binom{i-1}{3}}{\binom{20}{4}} \quad (\text{by Axiom 3})$$

Method-2:

$$\text{Now, } P\{X > 13\} = 1 - P\{X \leq 13\}$$

In the event $\{X \leq 13\}$, we have to choose 4 balls from the balls numbered 1, 2, ..., 13 only. This can be done in $\frac{\binom{13}{4}}{\binom{20}{4}}$

$$\text{Thus, } P\{X > 13\} = 1 - \frac{\binom{13}{4}}{\binom{20}{4}}$$

For a random variable X , the function F defined by

$$F(x) = P\{X \leq x\}, \quad -\infty < x < \infty$$

is called the **cumulative distribution function** or, simply, the **distribution function** of X .

- ▶ Suppose $a \leq b$

Then, $\{X \leq a\} \subset \{X \leq b\}$

$$\implies P\{X \leq a\} \leq P\{X \leq b\} \quad (\text{by Fact-3})$$

$$\implies F(a) \leq F(b)$$

☞ This means that, $F(x)$ is a **non-decreasing function** of x

Now consider two experiments

Experiment 1: Flipping a coin infinite number of times

Experiment 2: Measuring lifetime of an electronic device

- ▶ Let X be the number of heads in experiment 1 and Y be the lifetime in hours in experiment 2
- ▶ X and Y are random variables. What are the values that X and Y can take?
- ▶ $X = 0, 1, 2, \dots$ and $0 \leq Y \leq \infty$
- ▶ What is the difference between X and Y ?
- ▶ X is taking only countably many values and Y is not so

☞ Such random variables which can take at most a countable number of values is said to be **discrete**

- ▶ A random variable that takes only countable number of values is said to be a **discrete random variable**.
- ▶ For a discrete random variable X , we define the **probability mass function (p.m.f)**, $p(\cdot)$, of X by

$$p(a) = P\{X = a\} \text{ for every real number } a$$

Some properties of p.m.f:

- ▶ Note that, as a function, $p : \mathbb{R} \rightarrow [0, 1]$
- ▶ For a given $b \in \mathbb{R}$, if the random variable does not take the value b , then $p(b) = P\{X = b\} = 0$

- ▶ Thus, if X can take only the values x_1, x_2, \dots , then

$$p(x_i) \geq 0 \text{ for } i = 1, 2, \dots$$

$$p(x) = 0 \text{ for all other values of } x$$

- ▶ Since X must take one of the values x_i , we have

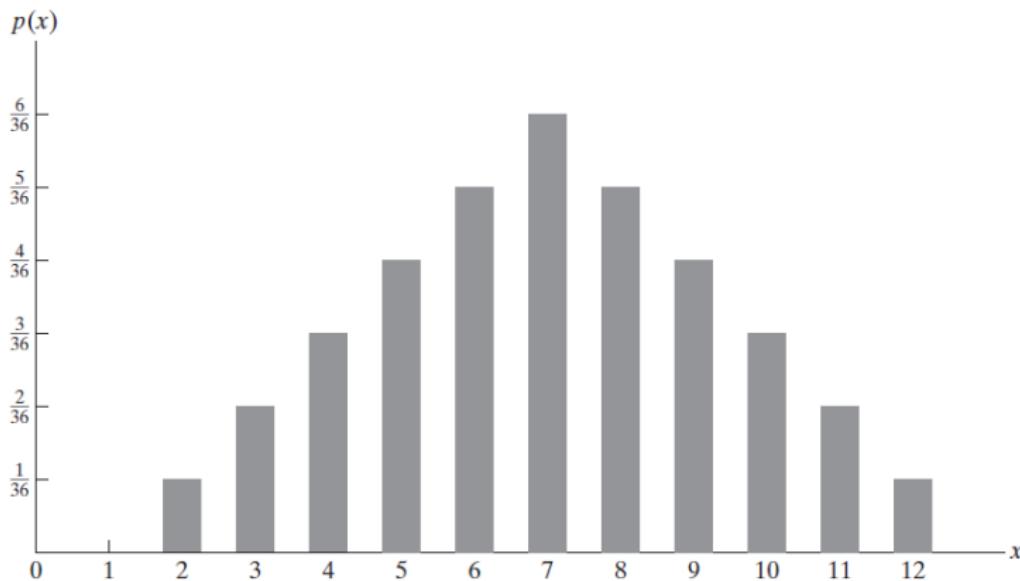
$$\sum_{i=1}^{\infty} p(x_i) = 1$$

- ▶ We often present p.m.f in a graphical format by plotting $p(x_i)$ on the y -axis against x_i on the x -axis

Example: Consider the experiment of rolling a pair of dice and the random variable X be the sum of the dice

We have the following

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Example: Five men and five women are ranked according to their scores on an examination. Assume that no two scores are alike and all $10!$ possible ranking are equally likely. Let X denote the highest ranking achieved by a woman.

Find the probability mass function (p.m.f) of X .

Solution: Step-1: Identify the values X can take

- ▶ First of all, note that the lowest possible position is 6, which means that all five women score worse than the five men.
- ▶ That is, X can take values $1, 2, 3, 4, 5, 6$

Step-2: Find $p(a) = P(X = a)$ for each value a that X can take

- ▶ For $X = 6$, we first need to find the number of different ways to arrange the 10 people such that the women all scored lower than the men.
- ▶ There are $5!$ ways to arrange the women and $5!$ ways to arrange the men, so $P\{X = 6\} = \frac{5! \cdot 5!}{10!}$

- ▶ Now consider the top woman scoring 5th on the exam.



- ▶ There are 5 possible positions for the lower scoring women, and we have 4 women that must be assigned to these ranks
- ▶ This can be accomplished in $\binom{5}{4}$ different ways
- ▶ Additionally, these 5 women can be arranged in 5! ways, and the men can be arranged in 5! ways. Thus,

$$P\{X = 5\} = \frac{\binom{5}{4} \cdot 5! \cdot 5!}{10!}$$

- ▶ Similarly for $P\{X = 4\}$, there are 6 positions for the 4 remaining women.
- ▶ The women can be arranged in $5!$ ways and the men can be arranged in $5!$ ways.

$$\implies P\{X = 4\} = \frac{\binom{6}{4} \cdot 5! \cdot 5!}{10!}$$

- ▶ By exactly same argument, we get

$$P\{X = 3\} = \frac{\binom{7}{4} \cdot 5! \cdot 5!}{10!}$$

$$P\{X = 2\} = \frac{\binom{8}{4} \cdot 5! \cdot 5!}{10!}$$

$$P\{X = 1\} = \frac{\binom{9}{4} \cdot 5! \cdot 5!}{10!}$$

Thus, the p.m.f of X is given by

$$p(a) = \begin{cases} \frac{\binom{(10-a)}{4} \cdot 5! \cdot 5!}{10!}, & \text{for } a = 1, 2, 3, 4, 5, 6, \\ 0, & \text{else.} \end{cases}$$

Probability and Statistics

Lecture-9

- ▶ A random variable that takes only countable number of values is said to be a **discrete random variable**.
- ▶ For a discrete random variable X , we define the **probability mass function (p.m.f)**, $p(\cdot)$, of X by

$$p(a) = P\{X = a\} \text{ for every real number } a$$

Some properties of p.m.f:

- ▶ Note that, as a function, $p : \mathbb{R} \rightarrow [0, 1]$
- ▶ For a given $b \in \mathbb{R}$, if the random variable does not take the value b , then $p(b) = P\{X = b\} = 0$

- ▶ Thus, if X can take only the values x_1, x_2, \dots , then

$$p(x_i) \geq 0 \text{ for } i = 1, 2, \dots$$

$$p(x) = 0 \text{ for all other values of } x$$

- ▶ Since X must take one of the values x_i , we have

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

Example: The p.m.f of a random variable X is given by

$$p(i) = \begin{cases} \frac{c\lambda^i}{i!}, & \text{for } i = 0, 1, 2, \dots, \\ 0, & \text{else.} \end{cases}$$

where λ is some positive number.

Find the value of c and calculate (a) $P\{X = 0\}$ (b) $P\{X > 2\}$.

Solution:

- ▶ By the properties of p.m.f,

$$\sum_{i=1}^{\infty} p(i) = 1$$

$$\implies c \sum_{i=1}^{\infty} \frac{\lambda^i}{i!} = 1$$

since $e^x = \sum_{i=1}^{\infty} \frac{x^i}{i!}$, we get $ce^\lambda = 1$ or $c = e^{-\lambda}$

- ▶ Thus, (a) $P\{X = 0\} = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-\lambda}$

- ▶ (b) We have,

$$P\{X > 2\} = 1 - P\{X \leq 2\} = 1 - P\{X = 0\} - P\{X = 1\} - P\{X = 2\}$$

$$\implies P\{X > 2\} = 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2 e^{-\lambda}}{2}$$

□

Recall: The cumulative distribution function of X is given by $F(a) = P\{X \leq a\}$ for every real number a .

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

☞ This gives the relation between the p.m.f and cumulative distribution function

Example: Suppose X is a discrete random variable and has a probability mass function given by

$$p(1) = \frac{1}{4}, \quad p(2) = \frac{1}{2}, \quad p(3) = \frac{1}{8}, \quad p(4) = \frac{1}{8}$$

and $p(a) = 0$ for every other value of a . Compute the distribution function of X .

- ▶ X takes values 1,2,3 and 4 only
- ▶ Consider the least of them, which is 1
- ▶ For any $x < 1$, we have $p(x) = 0$ and hence

$$F(a) = \sum_{\text{all } x \leq a} p(x) = 0 \text{ for any } a < 1$$

- ▶ Now, for $1 \leq a < 2$,

$$p(a) = \begin{cases} \frac{1}{4}, & \text{if } a = 1, \\ 0, & \text{else.} \end{cases}$$

- ▶ Thus, for $1 \leq a < 2$,

$$F(a) = \sum_{\text{all } x \leq a} p(x) = p(1) = \frac{1}{4} \text{ for } 1 \leq a < 2$$

- ▶ For, $2 \leq a < 3$

$$p(a) = \begin{cases} \frac{1}{2}, & \text{if } a = 2, \\ 0, & \text{else.} \end{cases}$$

- ▶ Thus, for $2 \leq a < 3$,

$$F(a) = \sum_{\text{all } x \leq a} p(x) = p(1) + p(2) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \text{ for } 2 \leq a < 3$$

- ▶ Similarly, for $3 \leq a < 4$,

$$F(a) = \sum_{\text{all } x \leq a} p(x) = p(1) + p(2) + p(3) = \frac{7}{8}$$

- ▶ And, for $a \geq 4$,

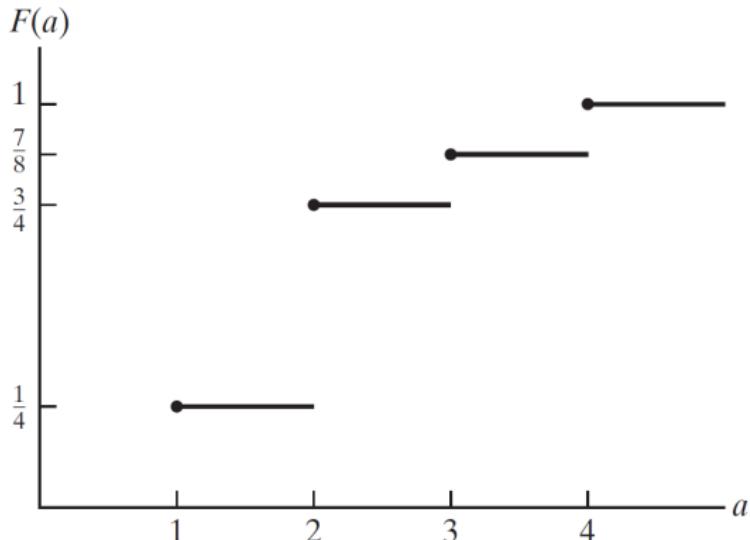
$$F(a) = \sum_{\text{all } x \leq a} p(x) = p(1) + p(2) + p(3) + p(4) = 1$$

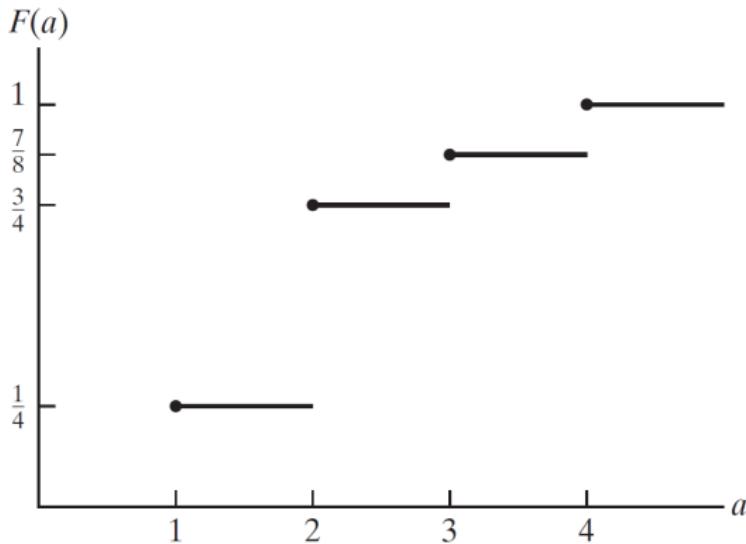
- ▶ Putting everything together, the distribution function of X is

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

- ▶ Let us look at graph of $F(x)$

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





- ▶ The function $F(x)$ is a discontinuous function and the graph of $F(x)$ has jumps of magnitude $p(a)$ at each point $x = a$ of discontinuity
- ▶ For example, the jump at 2 is $p(2) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}$

- ▶ The cumulative distribution function of X is given by $F(a) = P\{X \leq a\}$ for every real number a
- ▶ If X is discrete, F is discontinuous precisely at the values which X takes
- ▶ The jump at each discontinuity $X = a$ is given by the p.m.f $p(a)$
- ▶ $P(a < X \leq b) = \sum_{x:a < x \leq b} p(x) = F(b) - F(a)$
- ▶ $P(a \leq X < b) = \sum_{x:a \leq x < b} p(x) = F(b) - F(a) - p(b) + p(a)$
- ▶ $P(a \leq X \leq b) = \sum_{x:a \leq x \leq b} p(x) = F(b) - F(a) + p(a)$

Example: The distribution function of the random variable X is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \leq x < 1 \\ \frac{2}{3} & 1 \leq x < 2 \\ \frac{11}{12} & 2 \leq x < 3 \\ 1 & 3 \leq x \end{cases}$$

- (a) Find its probability mass function
- (b) $P\{X < 3\}$
- (c) $P\{X > \frac{1}{2}\}$
- (d) $P\{2 < X \leq 4\}$

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \leq x < 1 \\ \frac{2}{3} & 1 \leq x < 2 \\ \frac{11}{12} & 2 \leq x < 3 \\ 1 & 3 \leq x \end{cases}$$

(a)

- ▶ The values X takes are precisely the points at which F is discontinuous
- $$\implies X = 0, 1, 2, 3$$
- ▶ We need to find $p(i) = P\{X = i\}$ for each $i = 0, 1, 2, 3$
 - ▶ Now, for $0 \leq a < 1$,

$$\frac{1}{2} = F(a) = P\{X \leq a\} = \sum_{x \leq a} p(x) = p(0)$$

$$\implies p(0) = P\{X = 0\} = \frac{1}{2}$$

- ▶ Now, for $1 \leq a < 2$,

$$\frac{2}{3} = F(a) = P\{X \leq a\} = \sum_{x \leq a} p(x) = p(0) + p(1)$$

$$\implies p(1) = P\{X = 1\} = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$$

- ▶ For $2 \leq a < 3$,

$$\frac{11}{12} = F(a) = P\{X \leq a\} = \sum_{x \leq a} p(x) = p(0) + p(1) + p(2) = \frac{1}{2} + \frac{1}{6} + p(2)$$

$$\implies p(2) = P\{X = 2\} = F(a) - p(0) - p(1) = \frac{11}{12} - \frac{1}{2} - \frac{1}{6} = \frac{1}{4}$$

- ▶ Since $p(0) + p(1) + p(2) + p(3) = 1$, we have

$$p(3) = P\{X = 3\} = 1 - \frac{1}{2} - \frac{1}{6} - \frac{1}{4} = \frac{1}{12}$$

► Thus the p.m.f of X is

$$p(a) = \begin{cases} \frac{1}{2}, & \text{if } a = 0, \\ \frac{1}{6}, & \text{if } a = 1, \\ \frac{1}{4}, & \text{if } a = 2, \\ \frac{1}{12}, & \text{if } a = 3, \\ 0, & \text{else.} \end{cases}$$

(b) $P\{X < 3\} = p(0) + p(1) + p(2) = 1 - p(3) = \frac{11}{12}$

(c) $P\{X > \frac{1}{2}\} = 1 - P\{X \leq \frac{1}{2}\} = 1 - F\left(\frac{1}{2}\right) = \frac{3}{4}$

(d) $P\{2 < X \leq 4\} = F(4) - F(2) = \frac{1}{12}$

Average winnings: Suppose in a game of chance, we have a chance of winning x_i rupees with the probability $p(x_i)$, for $i = 1, 2, \dots, n$.

What will be our average winnings per game?

Answer: $\sum_{i=1}^n x_i p(x_i)$

So $\sum_{i=1}^n x_i p(x_i)$ is the **expected value** of the amount we win!

If X is a discrete random variable having a p.m.f $p(x)$, then the **expectation**, or the **expected value**, of X , denoted by $E[X]$, is defined by

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

Problem: Find $E[X]$, where X is the outcome when we roll a fair die

Solution: We have $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = \frac{1}{6}$

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



Example: The distribution function of the random variable X is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \leq x < 1 \\ \frac{2}{3} & 1 \leq x < 2 \\ \frac{11}{12} & 2 \leq x < 3 \\ 1 & 3 \leq x \end{cases}$$

Find $E[X]$.

- We computed the p.m.f earlier:

$$p(a) = \begin{cases} \frac{1}{2}, & \text{if } a = 0, \\ \frac{1}{6}, & \text{if } a = 1, \\ \frac{1}{4}, & \text{if } a = 2, \\ \frac{1}{12}, & \text{if } a = 3, \\ 0, & \text{else.} \end{cases}$$

- Thus,

$$\begin{aligned} E[X] &= 0 \cdot p(0) + 1 \cdot p(1) + 2 \cdot p(2) + 3 \cdot p(3) \\ &= \frac{1}{6} + \frac{1}{2} + \frac{1}{4} \\ &= \frac{11}{12} \end{aligned}$$

Probability and Statistics

Lecture-10

If X is a discrete random variable having a p.m.f $p(x)$, then the **expectation**, or the **expected value**, of X , denoted by $E[X]$, is defined by

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

- ☞ It is the average value that a random variable will take if we only repeat our experiment often enough!
- ☞ In fact, we may never observe the expected value!

- ▶ Let X be a discrete random variable with p.m.f p_X and $Y = X^2$
- ▶ Then Y is also a discrete random variable.
- ▶ What will be $E[Y]$?
- ▶ **Fact:** $E[Y] = \sum_{x:p_X(x)>0} x^2 p_X(x)$

☞ More generally,

Fact 7. If X is a discrete random variable that takes on one of the values $x_i, i \geq 1$, with respective probabilities $p(x_i)$, then, for any real-valued function g ,

$$E[g(X)] = \sum_i g(x_i)p(x_i)$$

Example: Let X denote a random variable that takes on any of the values $-1, 0$, and 1 with respective probabilities

$$P\{X = -1\} = 0.2, \quad P\{X = 0\} = 0.5, \quad P\{X = 1\} = 0.3$$

► $E[X^2] = (-1)^2 p(-1) + 0^2 p(0) + 1^2 p(1) = 0.5$

★ Observe that, $0.5 = E[X^2] \neq (E[X])^2 = 0.01$

☞ In general, $E[g(X)] \neq g(E[X])!$

Fact 8. If a and b are constants, then

$$E[aX + b] = aE[X] + b$$

Proof.

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



Example: A discrete random variable has the following p.m.f

$$p(a) = \begin{cases} \frac{1}{2}, & \text{if } a = 0, \\ \frac{1}{6}, & \text{if } a = 1, \\ \frac{1}{4}, & \text{if } a = 2, \\ \frac{1}{12}, & \text{if } a = 3, \\ 0, & \text{else.} \end{cases}$$

Compute $E[X^2 + 2X + 3]$

- We have $E[X^2 + 2X + 3] = E[X^2] + 2E[X] + 3$

$$\begin{aligned} E[X] &= 0 \cdot p(0) + 1 \cdot p(1) + 2 \cdot p(2) + 3 \cdot p(3) \\ &= \frac{1}{6} + \frac{1}{2} + \frac{1}{4} \\ &= \frac{11}{12} \end{aligned}$$

$$\begin{aligned}E[X^2] &= 0^2 \cdot p(0) + 1^2 \cdot p(1) + 2^2 \cdot p(2) + 3^2 \cdot p(3) \\&= \frac{1}{6} + 1 + \frac{3}{4} \\&= \frac{23}{12}\end{aligned}$$

$$\implies E[X^2 + 2X + 3] = \frac{23}{12} + 2\left(\frac{11}{12}\right) + 3 = \frac{27}{4}$$

□

- ▶ Given a random variable X , its expected value, $E[X]$, is also referred to as the **mean** or the **first moment of X** .
- ▶ The quantity $E[X^n]$, $n \geq 1$, is called the n^{th} **moment of X**

Example: Find the expected value of the sum obtained when n fair dice are rolled

- ▶ X be the random variable which denotes the sum.
- ▶ X can take values $n, n+1, \dots, 6n$.
- ▶ *Usual approach:* calculate $E[X]$ by finding probabilities $p(i)$ for each of these values $i = n, n+1, \dots, 6n$ and find the sum
$$\sum_{i=n}^{6n} ip(i)$$
- ▶ *Problem:* For large values of n , this approach is not practically feasible!
- ▶ Let X_i denote the random variable which is the upturned value on the die i . Thus, $X = X_1 + X_2 + \dots + X_n$
- ▶ X_i takes values $1, 2, 3, 4, 5, 6$ with equal probability $\frac{1}{6}$ and hence $E[X_i] = \frac{7}{2}$
- ▶ If $E[X] = E[X_1] + E[X_2] + \dots + E[X_n]$, then $E[X] = \frac{7n}{2}$

☞ Is $E[X] = E[X_1] + E[X_2] + \cdots + E[X_n]$ true?

Fact 9. For random variables X_1, X_2, \dots, X_n ,

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

Example: Suppose that 10 balls are put into 5 boxes, with each ball independently being put in box i with probability p_i , $\sum_{i=1}^5 p_i = 1$.

- (a) Find the expected number of boxes that do not have any balls.
- (b) Find the expected number of boxes that have exactly 1 ball.

(a) X denote the number of boxes that do not have any balls

- ▶ We need to find $E[X]$

- ▶ Define random variables $X_i = 1$ if the i^{th} box is empty and 0, else, for each $i = 1, 2, 3, 4, 5$
- ▶ Then, $X = X_1 + X_2 + X_3 + X_4 + X_5$
- ▶ Now, let us concentrate on X_i .
- ▶ $X_i = 1$, i.e, the i^{th} box is empty, if all the 10 balls go into the other 4 boxes
- ▶ Thus, $P\{X_i = 1\} = (1 - p_i)^{10}$ and hence

$$E[X_i] = (0)(1 - (1 - p_i)^{10}) + (1)(1 - p_i)^{10} = (1 - p_i)^{10} \text{ for each } i = 1, 2, 3$$

$$\implies E[X] = \sum_{i=1}^5 (1 - p_i)^{10}$$

(b) Now, let Y be the number of boxes that have exactly 1 ball

- ▶ Again, let Y_i be the random variable which takes value 1 if the i^{th} box has exactly 1 ball and 0, else
- ▶ Then, $Y = Y_1 + Y_2 + Y_3 + Y_4 + Y_5$
- ▶ For Y_i to be 1, one of the 10 balls should go into the i^{th} box and all others into the remaining 4 boxes
- ▶ Hence $P\{Y_i = 1\} = \binom{10}{1} p_i (1 - p_i)^9$ for each $i = 1, 2, 3, 4, 5$

$$\implies E[Y_i] = \binom{10}{1} p_i (1 - p_i)^9, i = 1, 2, 3, 4, 5$$

$$\implies E[Y] = \sum_{i=1}^5 \binom{10}{1} p_i (1 - p_i)^9$$

☞ Mean or expected value of a random variable X does not tell anything about spread or variation of the values of X

Example: Consider the following three random variables

$W = 0$ with probability 1

$$Y = \begin{cases} -1 & \text{with probability } \frac{1}{2} \\ +1 & \text{with probability } \frac{1}{2} \end{cases}$$

$$Z = \begin{cases} -100 & \text{with probability } \frac{1}{2} \\ +100 & \text{with probability } \frac{1}{2} \end{cases}$$

- ▶ We get $E[W] = E[Y] = E[Z] = 0$
- ▶ But the values of Z are **more spread from 0** compared to that of Y
- ▶ Values of Y in turn are **more spread from 0** compared to the values of W .

- ▶ How do we measure this spread?
- ▶ We will try to measure the spread of random variable X from its mean $E[X] = \mu$ (say), i.e., $|X - \mu|$
- ▶ Since X can take many values, computing $|X - \mu|$ for each value of X may be cumbersome
- ▶ So we will look at the **average or expectation** of this spread from the mean
- ▶ That is, we will look at $E[|X - \mu|]$
- ▶ Mathematically, $|\cdot|$ is not so convenient to handle, so we will consider a similar quantity $(X - \mu)^2$
- ▶ Thus, $E[(X - \mu)^2]$ is the measure of the spread of the values of random variable X from its mean μ

If X is a random variable with mean $\mu = E[X]$, then the **variance** of X , denoted by $\text{Var}(X)$, is defined by

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

Fact 9. $\text{Var}(X) = E[X^2] - (E[X])^2$

Proof.

$$\begin{aligned}\text{Var}(X) &= E[(X - \mu)^2] = \sum_x (x - \mu)^2 p(x) \\ &= \sum_x (x^2 - 2\mu x + \mu^2) p(x) \\ &= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\ &= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2\end{aligned}$$

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

Example: Calculate $\text{Var}(X)$ if X represents the outcome when a fair die is rolled.

We already saw that $E[X] = \frac{7}{2}$

Now, $E[X^2] = ??$

$$E[X^2] = 1^2 \left(\frac{1}{6}\right) + 2^2 \left(\frac{1}{6}\right) + 3^2 \left(\frac{1}{6}\right) + 4^2 \left(\frac{1}{6}\right) + 5^2 \left(\frac{1}{6}\right) + 6^2 \left(\frac{1}{6}\right) = \frac{91}{6}$$

$$\text{Thus, } \text{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

Example: A box contains 5 red and 5 blue marbles. Two marbles are drawn randomly. If they are the same color, then you win \$1.10; if they are different colors, then you win -\$1.00. (That is, you lose \$1.00). Calculate

- the expected value of the amount you win;
- the variance of the amount you win.

Solution: Let X be the amount of winnings.

X can take values 1.10 and -1.00

The probability of $X = 1.10$ is the probability of drawing 2 balls of the same color, which is $\frac{\binom{2}{1} \cdot \binom{5}{2}}{\binom{10}{2}} = \frac{4}{9}$

(a) Thus, $P\{X = -1\} = 1 - \frac{4}{9} = \frac{5}{9}$

Therefore, $E[X] = (1.10)\left(\frac{4}{9}\right) - (1)\left(\frac{5}{9}\right) = -\frac{1}{15}$

So you are expected to **lose money!**

(b) We will use the formula $\text{Var}(X) = E[X^2] - (E[X])^2$

$$\text{Now, } E[X^2] = (1.10)^2 \left(\frac{4}{9}\right) + (1)^2 \left(\frac{5}{9}\right)$$

$$\text{Hence, } \text{Var}(X) = (1.10)^2 \left(\frac{4}{9}\right) + (1)^2 \left(\frac{5}{9}\right) - \left(\frac{1}{15}\right)^2 \approx 1.089$$

□

Fact 10. For any constants a and b ,

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

Proof. Let $\mu = E[X]$ and by Fact 8, $E[aX + b] = a\mu + b$.

Therefore,

$$\begin{aligned}\text{Var}(aX + b) &= E[(aX + b - a\mu - b)^2] \\ &= E[a^2(X - \mu)^2] \\ &= a^2E[(X - \mu)^2] \\ &= a^2\text{Var}(X)\end{aligned}$$

Example: Let X be a discrete random variable. If $E[X] = 1$ and $\text{Var}(X) = 5$, find

- (a) $E[(2 + X)^2]$
- (b) $\text{Var}(4 + 3X)$

Solution: (a)

$$E[(2 + X)^2] = E[4 + 4X + X^2] = 4 + 4E[X] + E[X^2]$$

$$\text{Now, } E[X]^2 = \text{Var}(X) + (E[X])^2 = 5 + 1^2 = 6 \quad (\text{by Fact 9})$$

$$\implies E[(2 + X)^2] = 4 + (4 \times 1) + 6 = 14$$

(b) $\text{Var}(4 + 3X) = 3^2 \text{Var}(X) = 45 \quad (\text{by Fact 10})$



The square root of the $\text{Var}(X)$ is called the **Standard deviation** of X , and we denote it by $SD(X)$. That is,

$$SD(X) = \sqrt{\text{Var}(X)}$$

Probability and Statistics

Lecture-11

- ▶ Suppose that a trial, or an experiment, whose outcome can be classified as either a **success** or **failure** is performed
- ▶ Example - tossing a coin (Heads - success, Tails - failure)
- ▶ X be the random variable which takes 1 when the outcome is success and 0 when the outcome is a failure
- ▶ The probability mass function of X is given by

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

$$p(1) = P\{X = 1\} = p$$

where p , $0 \leq p \leq 1$ is the probability that the trial is a success.

A random variable X is said to be a **Bernoulli random variable** if it takes only two values 0,1 and its probability mass function is given by

$$p(a) = \begin{cases} 1 - p, & \text{if } a = 0, \\ p, & \text{if } a = 1, \\ 0, & \text{else,} \end{cases}$$

for some $p \in (0, 1)$.

- ☞ Suppose now that we have n independent trials, each of which results in a success with probability p or in a failure with probability $1 - p$, are to be performed.
- ☞ If X denotes the number of successes that occur in the n trials, then X is said to be a **binomial random variable** with parameters (n, p) .

If X is a binomial random variable with parameters (n, p) , then the p.m.f of X is given by

$$p(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & k = 0, 1, 2, \dots, n, \\ 0, & \text{else.} \end{cases}$$

☞ By binomial theorem,

$$\sum_{k=0}^{\infty} p(k) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1$$

☞ Thus, a Bernoulli random variable is just a binomial random variable with parameters $(1, p)$

Example: Five fair coins are flipped. If the outcomes are assumed independent, find the probability mass function of the number of heads obtained.

- ▶ X be the number of heads (successes) that appear.
- ▶ X is a binomial random variable with parameters $(n = 5, p = \frac{1}{2})$.
- ▶ The p.m.f of X is given by

$$p(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & k = 0, 1, 2, \dots, 5, \\ 0, & \text{else.} \end{cases}$$

- ▶ If you wish to calculate, say, $P\{X = 3\}$, just substitute 3 in place of k

Properties of Binomial Random Variables

Let $X \sim \text{Bin}(n,p)$, i.e., X is a binomial random variable.

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

 **Exercise:** $i \binom{n}{i} = n \binom{n-1}{i-1}$ for $i \geq 1$

$$\begin{aligned} E[X^k] &= np \sum_{i=1}^n i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} \\ &= np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \quad (\text{by setting } j = i-1) \\ &= npE[(Y+1)^{k-1}] \end{aligned}$$

where $Y \sim \text{Bin}(n-1, p)$

$$\implies E[X^k] = npE[(Y+1)^{k-1}] \text{ for } k \geq 1 \text{ and } Y \sim \text{Bin}(n-1, p)$$

In particular, $k = 1$, $E[X] = npE[1] = np$

Further, when $k = 2$, $E[X^2] = npE[Y+1]$ where $Y \sim \text{Bin}(n-1, p)$

$$\implies E[X^2] = np[E[Y] + 1] = np[(n-1)p + 1]$$

$$\begin{aligned} \text{Thus, } \text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= np[(n-1)p + 1] - (np)^2 \\ &= np(1-p) \end{aligned}$$

For a binomial random variable X with parameters (n, p) , we have

$$E[X] = np$$

$$\text{Var}(X) = np(1-p)$$

$$E[X^k] = npE[(Y+1)^{k-1}] \text{ where } Y \sim \text{Bin}(n-1, p)$$

A random variable X that takes on one of the values $0, 1, 2, \dots$ is said to be a **Poisson random variable** with parameter λ if, for some $\lambda > 0$,

$$p(i) = P\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \dots$$

We have $\sum_{i=0}^{\infty} p(i) = e^{-\lambda} \left(\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} \right) = e^{-\lambda} e^{\lambda} = 1$

★ Some examples of random variables which obey Poisson probability law:

- ▶ The number of misprints on a page (or a group of pages) of a book
- ▶ The number of people in a community who survive to age 100

Some examples of random variables which obey Poisson probability law:

- ▶ The number of wrong telephone numbers that are dialled in a day
- ▶ The number of packages of biscuits sold in a particular store each day
- ▶ The number of customers entering a post office on a given day
- ▶ The number of vacancies occurring during a year in any of the government departments
- ▶ The number of α -particles discharged in a fixed period of time from some radioactive material

 Observe that in each case the number of objects is very large and the probability is very small!

Expectation of Poisson random variable:

$$\begin{aligned} E[X] &= \sum_{i=0}^{\infty} i \left(\frac{e^{-\lambda} \lambda^i}{i!} \right) \\ &= \sum_{i=1}^{\infty} i \left(\frac{e^{-\lambda} \lambda^i}{i!} \right) \\ &= \lambda \sum_{i=1}^{\infty} i \left(\frac{e^{-\lambda} \lambda^{i-1}}{(i-1)!} \right) \\ &= \lambda e^{-\lambda} \left(\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \right) \quad (\text{by letting } j=i-1) \\ &= \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda \end{aligned}$$

$$\begin{aligned}
E[X^2] &= \sum_{i=0}^{\infty} i^2 \left(\frac{e^{-\lambda} \lambda^i}{i!} \right) \\
&= \sum_{i=1}^{\infty} i^2 \left(\frac{e^{-\lambda} \lambda^i}{i!} \right) \\
&= \lambda \sum_{i=1}^{\infty} i \left(\frac{e^{-\lambda} \lambda^{i-1}}{(i-1)!} \right) \\
&= \lambda \left(\sum_{j=0}^{\infty} \frac{(j+1)e^{-\lambda} \lambda^j}{j!} \right) \quad (\text{by letting } j=i-1) \\
&= \lambda \left[\sum_{j=0}^{\infty} j \left(\frac{e^{-\lambda} \lambda^j}{j!} \right) + \sum_{i=0}^{\infty} \left(\frac{e^{-\lambda} \lambda^i}{i!} \right) \right] \\
&= \lambda \left[E[X] + \sum_{i=0}^{\infty} p(i) \right] \\
&= \lambda(\lambda + 1)
\end{aligned}$$

$$\text{Thus, } \text{Var}(X) = \lambda(\lambda + 1) - \lambda^2 = \lambda$$

Summing up,

For a Poisson random variable X with parameter λ , we have

$$p(i) = \begin{cases} e^{-\lambda} \frac{\lambda^i}{i!}, & i = 0, 1, 2, \dots \\ 0, & \text{else.} \end{cases}$$

$$E[X] = \lambda$$

$$\text{Var}(X) = \lambda$$

A Poisson experiment is a statistical experiment that has the following properties:

- ▶ The experiment results in outcomes that can be classified as successes or failures
- ▶ The average number of successes (λ) that occurs in a specified region is known
- ▶ The probability that a success will occur is proportional to the size of the region
- ▶ The probability that a success will occur in an extremely small region is virtually zero

Example: Vehicles pass through a junction on a busy road at an average rate of 300 per hour.

- (a) Find the probability that none passes in a given minute.
 - (b) What is the expected number passing in two minutes?
 - (c) Find the probability that this expected number actually pass through in a given two-minute period.
- ▶ X denote the number of vehicles that pass through the junction per minute
 - ▶ X follows Poisson distribution
 - ▶ The average number of cars per minute is $\lambda = \frac{300}{60} = 5$
 - ▶ Thus, $p(i) = P\{X = i\} = e^{-5} \frac{5^i}{i!}, \quad i = 0, 1, 2, \dots$

(a) $P\{X = 0\} = e^{-5}$

(b) Per minute average is 5 \implies per 2-minute average is $\frac{300}{30} = 10$

- ▶ Thus, if Y denoted the number of vehicles passing through the junction per 2 minutes, then Y is also a Poisson random variable with parameter 10

$$\implies E[Y] = 10$$

(c) We have $P\{Y = i\} = e^{-10} \frac{10^i}{i!}, \quad i = 0, 1, 2, \dots$

- ▶ We need to find $P\{Y = 10\}$

$$\implies P\{Y = 10\} = e^{-10} \frac{10^{10}}{10!} \approx 0.12511$$



Summary

Bernoulli random variable:

X takes on one of the values 0,1 and its probability mass function is given by

$$p(0) = P\{X = 0\} = 1 - p,$$

$$p(1) = P\{X = 1\} = p,$$

$$p(a) = 0 \text{ for } a \neq 0, 1,$$

for some $p \in (0, 1)$.

 Here p is the parameter.

Binomial random variable:

X takes the values $0, 1, 2, \dots, n$ and its probability mass function is given by

$$p(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & k = 0, 1, 2, \dots, n, \\ 0, & \text{else.} \end{cases}$$

☞ Here (n, p) are the parameters and we have

$$E[X] = np$$

$$\text{Var}(X) = np(1 - p)$$

☞ Bernoulli random variable is a particular case of binomial random variable

Poisson random variable:

X takes on one of the values $0, 1, 2, \dots$ and its probability mass function is given by

$$p(i) = \begin{cases} e^{-\lambda} \frac{\lambda^i}{i!}, & i = 0, 1, 2, \dots \\ 0, & \text{else.} \end{cases}$$

☞ Here λ is the only parameter and we found that

$$E[X] = \text{Var}(X) = \lambda$$

Probability and Statistics

Lecture-12

- ▶ **Experiment:** flip a coin until heads occur
- ▶ Let X equal the number of flips required
- ▶ What are the values X can take?
- ▶ $X = 1, 2, \dots$
- ▶ Let p be the probability of heads in a single flip
- ▶ Then, $P\{X = n\} = (1 - p)^{n-1}p, \quad n = 1, 2, 3, \dots$
- ▶ $\sum_{n=1}^{\infty} P\{X = n\} = p \sum_{n=1}^{\infty} (1 - p)^{n-1} = \frac{p}{1-(1-p)} = 1$
 $\implies p(i) = P\{X = i\}, i = 1, 2, \dots$ is a probability mass function of X

A random variable X which takes on values $1, 2, 3, \dots$ and whose probability mass function is given by

$$p(i) = \begin{cases} (1 - p)^{i-1} p, & \text{if } i = 1, 2, 3, \dots, \\ 0, & \text{else,} \end{cases}$$

for some $p \in (0, 1)$, is called a **geometric random variable** with the parameter p .

Example: An urn contains N white and M black balls. Balls are randomly selected, one at a time, until a black one is obtained. If we assume that each ball selected is replaced before the next one is drawn, what is the probability that

- (a) exactly n draws are needed?
- (b) at least k draws are needed?

- ▶ X be the number of draws needed for a black ball to turn up
- ▶ The probability of black ball being selected is $p = \frac{M}{M+N}$
- ▶ That is, probability of “success” is $p = \frac{M}{M+N}$
- ▶ We are interested in number of draws for the first success
- ▶ Hence X is a geometric random variable with parameter p

- ▶ p.m.f of X is given by

$$p(i) = \begin{cases} (1-p)^{i-1}p, & \text{if } i = 1, 2, 3, \dots, \\ 0, & \text{else,} \end{cases}$$

- (a) Exactly n draws $\implies X = n$

Now,

$$p(n) = P\{X = n\} = p((1-p)^{n-1}p) = \left(\frac{N}{M+N}\right)^{n-1} \frac{M}{M+N} = \frac{MN^{n-1}}{(M+N)^n}$$

- (b) at least k draws $\implies X \geq k$

$$\begin{aligned}
P\{X \geq k\} &== \sum_{n=k}^{\infty} p(n) \\
&= \sum_{n=k}^{\infty} P\{X = n\} \\
&= \sum_{n=k}^{\infty} (1-p)^{n-1} p \\
&= \sum_{n=k}^{\infty} \frac{MN^{n-1}}{(M+N)^n} \\
&= \frac{M}{M+N} \sum_{n=k}^{\infty} \left(\frac{N}{M+N}\right)^{n-1} \\
&= \left(\frac{M}{M+N}\right) \left(\frac{N}{M+N}\right)^{k-1} \Bigg/ \left[1 - \frac{N}{M+N}\right] \\
&= \left(\frac{N}{M+N}\right)^{k-1}
\end{aligned}$$

Example: Consider a roulette wheel consisting of 38 numbers - 1 through 36, 0 and double 0. If Smith always bets that the outcome will be one of the numbers 1 through 12, what is the probability that

- (a) Smith will lose his first 5 bets;
 - (b) his first win will occur on his fourth bet?
- ▶ Smith always bets on the numbers 1 through 12, which occupy 12 spaces on the wheel
 - ▶ The probability of success (winning) is $p = \frac{12}{38} \approx 0.316$
- (a) The first five bets form a finite set of $n = 5$ trials and each spin of the roulette wheel is independent, and the probability of success p is constant
- ▶ X be the number of bets won by smith in 5 trials
 - ▶ $\implies X \sim \text{Bin}(5, 0.316)$
 - ▶ Losing all 5 bets $\implies X = 0$
 - ▶ Hence $P\{X = 0\} = \binom{5}{0}(0.316)^0(1 - 0.316)^5 \approx 0.15$

(b) Now, let Y denote the number of bets for his first win

- ▶ The question talks about the number of trials for the first win (success)
- ▶ Hence Y follows **geometric distribution** with parameter $p = 0.316$
- ▶ We have $P\{Y = i\} = (1 - p)^{i-1}p, i = 1, 2, 3, \dots$
- ▶ We wish to find probability for the first win to occur on fourth bet
- ▶ Hence required probability is

$$P\{Y = 4\} = (1 - 0.316)^3(0.316) \approx 0.1012$$



Example: A programmer has a 90% chance of finding a bug every time he compiles his code, and it takes him two hours to rewrite his code every time he discovers a bug. What is the probability that he will finish his program by the end of his workday? (Assume that a workday is 8 hours and that the programmer compiles his code immediately at the beginning of the day.)

- ▶ Here success is a bug-free compilation and failure is the discovery of a bug.
- ▶ Hence the probability of success is $p = 0.1 (= 1 - 0.9)$
- ▶ Program should be finished in 8 hours means that the number of compilations of the program should be at most 3
- ▶ X denote the number of compilations
 $\implies X$ is a geometric random variable with parameter $p = 0.1$

- ▶ probability mass function is given by

$$p(i) = \begin{cases} (1-p)^{i-1}p, & \text{if } i = 1, 2, 3, \dots, \\ 0, & \text{else,} \end{cases}$$

with $p = 0.1$

- ▶ Thus, the required probability is

$$\begin{aligned} P\{X \leq 3\} &= p(0) + p(1) + p(2) + p(3) \\ &= P\{X = 0\} + P\{X = 1\} + P\{X = 2\} + P\{X = 3\} \\ &= (0.9)^0(0.1) + (0.9)(0.1) + (0.9)^2(0.1) + (0.9)^3(0.1) \\ &\approx 0.344 \end{aligned}$$



- ▶ Geometric random variable is the number of trials required for the first success
- ▶ What if, we need to have r successes ?
- ▶ Let X be the number of coin flips required to get r heads
- ▶ What are the values X can take?
- ▶ $X = r, r + 1, \dots$
- ▶ For some $n \geq r$, $P\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r}$
- ▶ $\sum_{n=r}^{\infty} P\{X = n\} = \sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (1-p)^{n-r} = 1$ by **negative binomial expansion**

A random variable X which takes on values $r, r+1, r+2, \dots$ and whose probability mass function is given by

$$p(i) = \begin{cases} \binom{i-1}{r-1} p^r (1-p)^{i-r}, & \text{if } i = r, r+1, \dots, \\ 0, & \text{else,} \end{cases}$$

for some $p \in (0, 1)$, is called a **negative binomial random variable** with parameters (r, p)

☞ Geometric random variable is a particular case of negative binomial random variable when $r = 1$

Example: If independent trials, each resulting in a success with probability p , are performed, what is the probability of r successes occurring before m failures?

- ▶ X be the number of trials required to achieve r successes.
- ▶ Then X is a negative binomial random variable with parameters (r, p) and hence

$$P\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r}, n = r, r+1, \dots$$

- ▶ For exactly m failures to occur (as we must have exactly r successes as well), X must be $m+r$.
- ▶ Since the r^{th} success should occur before m failures, $X \leq r+m-1$ and we have $X \geq r$
- ▶ Thus, we need to find $P\{r \leq X \leq r+m-1\}$

$$P\{r \leq X \leq r+m-1\} = \sum_{n=r}^{n=r+m-1} \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

Banach's matchbox problem

At all times, a smoker carries 2 matchboxes; 1 in his left-hand pocket and 1 in his right-hand pocket. Each time he needs a match, he is equally likely to take it from either pocket.

Consider the moment when the smoker first discovers that one of his matchboxes is empty. If it is assumed that both matchboxes initially contained N matches, what is the probability that there are exactly k matches, $k = 0, 1, \dots, N$, in the other box?

- ▶ Let E denote the event that the smoker first discovers that the right-hand matchbox is empty and that there are k matches in the left-hand box at that time
- ▶ X be the number of times the smoker lights a cigarette
- ▶ Here success is that he chooses the right-hand match box.
Thus, $p = \frac{1}{2}$.

- ▶ For the event E to happen, the right-hand match box should be empty
- ▶ Thus, there should be N “successes”
- ▶ Hence, X is a negative binomial random variable with parameters $(N, \frac{1}{2})$
- ▶ Now, the required event will occur \iff the right-hand matchbox is empty and the left-hand matchbox has k matches
 \iff the $(N + 1)^{\text{th}}$ choice of the right-hand matchbox is made at the $(N + 1 + N - k)^{\text{th}}$ trial.
- ▶ Thus, $P(E) = P\{X = 2N - k + 1\} = \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k+1}$
- ▶ Recall - our event E considers only the right-hand matchbox

- ▶ There is an equal probability that it is left-hand matchbox that is first discovered to be empty and there are k matches in the right-hand box at that time
- ▶ Hence the desired result is

$$2 \times P(E) = \binom{2N-k}{N} \left(\frac{1}{2}\right)^{2N-k}$$



Summary

A random variable X which takes on values $1, 2, 3, \dots$ and whose probability mass function is given by

$$p(i) = \begin{cases} (1-p)^{i-1} p, & \text{if } i = 1, 2, 3, \dots, \\ 0, & \text{else,} \end{cases}$$

is called a **geometric random variable** with parameter p .

☞ By using similar techniques as in the case of binomial and Poisson random variables, we get

$$E[X] = \frac{1}{p}$$

$$\text{Var}(X) = \frac{1-p}{p^2}$$

For a **negative binomial random variable** X with parameters (r, p) , we have

$$p(i) = \begin{cases} \binom{i-1}{r-1} p^r (1-p)^{i-r}, & \text{if } i = r, r+1, \dots, \\ 0, & \text{else,} \end{cases}$$

☞ It turns out that

$$E[X] = \frac{r}{p} \text{ and } \text{Var}(X) = \frac{r(1-p)}{p^2}$$

Probability and Statistics

Lecture-13

Recall..

- ▶ We defined sample space, events
- ▶ Looked at three definitions of probability
- ▶ Defined random variables and started to see events in terms of random variables
- ▶ Narrowed down our focus to “discrete” random variables and defined expectation of a (discrete) random variable
- ▶ Derived probability mass function using distribution function and vice-versa
- ▶ Looked at five types of random variables - Bernoulli, binomial, Poisson, geometric and negative binomial, and their probability mass functions, expectation and variance

- ▶ “Discrete” random variables - random variables which takes on finite or countably many values
- ▶ We can even have random variables which can take uncountably many values
- ▶ Typical examples -
 - ▶ various times like service time, installation time, download time, failure time, and
 - ▶ physical measurements like weight, height, distance, velocity, temperature, and connection speed etc.

If X is a continuous random variable, then there exists a non-negative function f , called **probability density function**, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers,

$$P\{X \in B\} = \int_B f(x) dx$$

- ▶ Total probability must be 1

$$\implies P\{-\infty < X < \infty\} = \int_{-\infty}^{\infty} f(x) dx = 1$$

- ▶ If we let $B = [a, b]$, then

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

- ▶ If we let $a = b$ in the above equation,

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

- ▶ That is, $p(a) = P\{X = a\} = 0$ for every real number a
- ▶ Hence the probability mass function **does not carry any information in the case of continuous random variables!**

- ▶ Since $P\{X = a\} = 0$,

$$P\{X \leq a\} = P\{X = a\} + P\{X < a\} = P\{X < a\}$$

$$P\{a \leq X \leq b\} = P\{a < X \leq b\} = P\{a \leq X < b\} = P\{a < X < b\}$$

- ▶ **Recall:** For any random variable X , the cumulative distribution function is defined as $F(a) = P\{X \leq a\}$
- ▶ Now, if X is continuous, we have

$$F(a) = P\{X \leq a\} = \int_{-\infty}^a f(x) dx$$

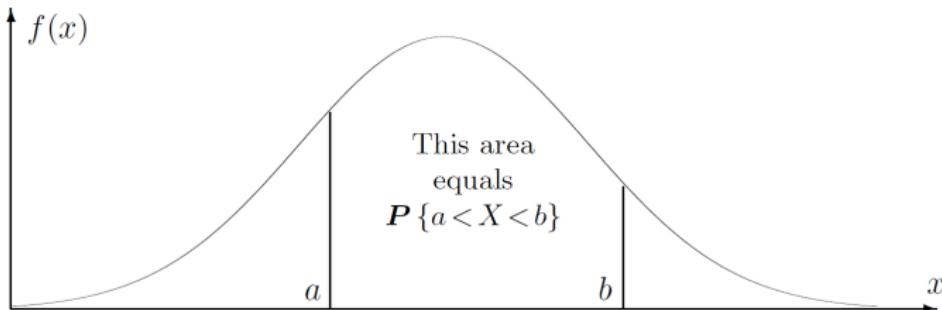
- ▶ Hence, $F(x)$ is a **continuous** non-decreasing function

- Whenever $F(x)$ is differentiable, we have

$$f(x) = \frac{d}{dx} F(x) = F'(x)$$

- By the **Fundamental Theorem of Calculus**,

$$\int_a^b f(x) dx = F(b) - F(a) = P\{a \leq X \leq b\}$$



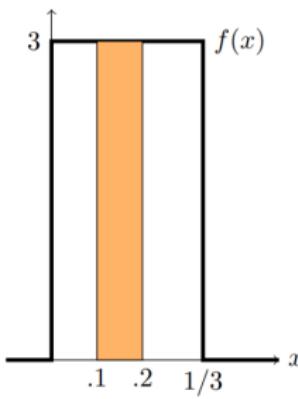
Summary

Distribution	Discrete	Continuous
We use	p.m.f $p(x) = P\{X = x\}$	p.d.f $f(x)$ (p.d.f)
Computing probabilities	$P\{X \in A\} = \sum_{x \in A} p(x)$	$P\{X \in A\} = \int_A f(x)dx$
Cumulative distribution function	$F(a) = \sum_{x \leq a} p(x)$	$F(a) = \int_{-\infty}^a f(x)dx$
Total probability	$\sum_x p(x) = 1$	$\int_{-\infty}^{\infty} f(x)dx = 1$

Example: Suppose a random variable X has density

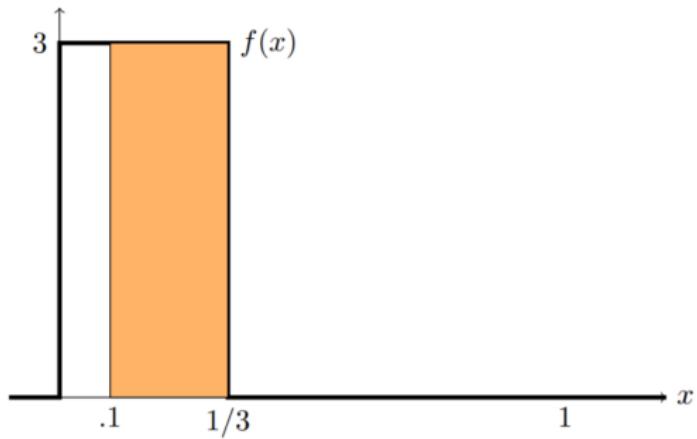
$$f(x) = \begin{cases} 3, & \text{if } x \in [0, \frac{1}{3}], \\ 0, & \text{else.} \end{cases}$$

Find $P\{0.1 \leq X \leq 0.2\}$, $P\{0.1 \leq X \leq 1\}$.



$$P\{0.1 \leq X \leq 0.2\}$$

$$\text{☞ } P\{0.1 \leq X \leq 0.2\} = \int_{0.1}^{0.2} f(x) dx = \int_{0.1}^{0.2} 3 dx = 0.3$$



$$P\{0.1 \leq X \leq 1\}$$

☞ $P\{0.1 \leq X \leq 1\} = \int_{0.1}^1 f(x) dx = \int_{0.1}^{1/3} 3 dx = 0.7$

□

Example: Let X be a random variable with p.d.f

$$f(x) = \begin{cases} Cx^2, & \text{if } x \in [0, 1], \\ 0, & \text{else.} \end{cases}$$

1. Find the value of C
 2. Find the cumulative distribution function
 3. Compute $P\left\{X < \frac{1}{2}\right\}$
- Since total probability should be 1,

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^1 Cx^2 dx = 1$$

$$\implies C \left[\frac{x^3}{3} \right]_{x=0}^1 = 1$$
$$\implies C = 3$$

- ▶ For $a < 0$, $f(a) = 0$

$$\implies F(a) = \int_{-\infty}^a f(x) dx = 0$$

- ▶ For $a \in [0, 1]$,

$$\begin{aligned} F(a) &= \int_{-\infty}^a f(x) dx \\ &= \int_0^a 3x^2 dx \\ &= a^3 \end{aligned}$$

- ▶ For $a > 1$,

$$\begin{aligned} F(a) &= \int_{-\infty}^a f(x) dx = \int_0^1 3x^2 dx \\ &= 1 \end{aligned}$$

- ▶ Thus, the cumulative distribution function of X is

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ x^3, & \text{if } x \in [0, 1], \\ 1, & \text{if } x > 1. \end{cases}$$

- ▶ $P\left\{X < \frac{1}{2}\right\}$ can be calculated in two ways:

☞ Using density function

$$P\left\{X < \frac{1}{2}\right\} = \int_{-\infty}^{1/2} f(x) dx = \int_0^{1/2} 3x^2 dx = \frac{1}{8}$$

☞ Using distribution function

$$P\left\{X < \frac{1}{2}\right\} = F\left(\frac{1}{2}\right) = \frac{1}{8}$$



Example: The lifetime, in years, of some electronic component is a continuous random variable with the density

$$f(x) = \begin{cases} \frac{k}{x^3} & x \geq 1, \\ 0 & x < 1 \end{cases}$$

Find k , draw a graph of the distribution function $F(x)$, and compute the probability for the lifetime to exceed 5 years.

- We must have $\int_{-\infty}^{\infty} f(x)dx = 1$

- Thus, $\int_1^{\infty} \frac{k}{x^3} dx = 1$

$$\implies \left[-\frac{k}{2x^2} \right]_{x=1}^{x=\infty} = 1$$

$$\implies \frac{k}{2} = 1$$

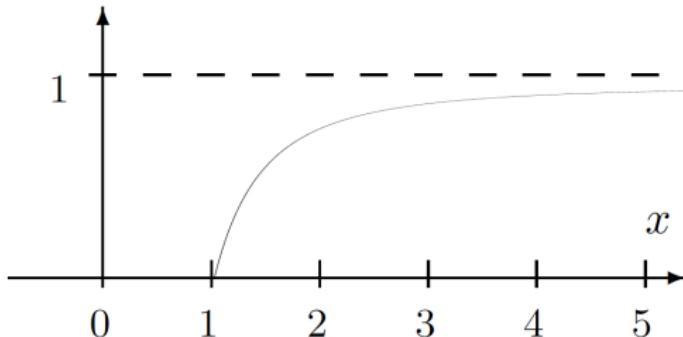
$$\implies k = 2$$

- ▶ Now,

$$\begin{aligned}F(x) &= \int_{-\infty}^x f(y) dy \\&= \int_1^x \frac{2}{y^3} dy \\&= \left[-\frac{1}{y^2} \right]_{y=1}^{y=x} \\&= 1 - \frac{1}{x^2}\end{aligned}$$

for $x \geq 1$

- ▶ Its graph looks like,



- ▶ Finally, using distribution function,

$$\begin{aligned}P\{X > 5\} &= 1 - P\{X \leq 5\} \\&= 1 - F(5) \\&= 1 - \left(1 - \frac{1}{5^2}\right) \\&= 0.04\end{aligned}$$

- ▶ Using density function,

$$\begin{aligned}P\{X > 5\} &= \int_5^{\infty} f(x) dx \\&= \int_5^{\infty} \frac{2}{x^2} dx \\&= \left[-\frac{1}{x^2} \right]_{x=5}^{\infty} \\&= \frac{1}{25} \\&= 0.04\end{aligned}$$

Example: The lifetime in hours of a certain kind of radio tube is a random variable having a probability density function given by

$$f(x) = \begin{cases} 0, & x \leq 100, \\ \frac{100}{x^2}, & x > 100 \end{cases}$$

What is the probability that exactly 2 of 5 such tubes in a radio set will have to be replaced within the first 150 hours of operation? Assume that the replacement of any two tubes are independent of each other.

- ▶ X denote the lifetime i hours of a radio tube
- ▶ Here “success” is replacement of a tube with in first 150 hours.

- ▶ Hence, probability of success is

$$\begin{aligned} p &= P\{X \leq 150\} \\ &= \int_{-\infty}^{150} f(x) dx \\ &= \int_{100}^{150} \frac{100}{x^2} dx \\ &= 100 \left[-\frac{1}{x} \right]_{x=100}^{150} \\ &= \frac{1}{3} \end{aligned}$$

- ▶ Y be the number of tubes to be replaced with in first 150 hrs
- ▶ We are looking at number of replacements out of 5 tubes and it is given that each replacement is independent of other replacements.
- ▶ Hence Y is a binomial random variable with parameters $(n = 5, p = \frac{1}{3})$

- ▶ Thus, the desired probability is

$$P\{Y = 2\} = \binom{5}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3 = \frac{80}{243}$$



Example: If X is continuous with distribution function F_X and density function f_X , find the density function of $Y = 2X$.

- ▶ We have $f_Y(a) = \frac{d}{da}F_Y(a)$

$$\text{Now, } F_Y(a) = P\{Y \leq a\} = P\{2X \leq a\}$$

$$= P\left\{X \leq \frac{a}{2}\right\} = F_X\left(\frac{a}{2}\right)$$

$$\begin{aligned}\implies f_Y(a) &= \frac{d}{da}F_Y(a) = \frac{d}{da}F_X\left(\frac{a}{2}\right) \\ &= \frac{1}{2}f_X\left(\frac{a}{2}\right)\end{aligned}$$

Probability and Statistics

Lecture-14

- ▶ For a continuous random variable X , the probability mass function is of no use as the point masses have zero probability
- ▶ The probability density function $f(x)$ plays the role of probability mass function in the continuous case
- ▶ $P\{X \in B\} = \int\limits_B f(x) dx$
- ▶ Distribution function $F(a) = \int\limits_{-\infty}^a f(x) dx$
- ▶ $P(a < X < b) = \int\limits_a^b f(x) dx = F(b) - F(a)$

Let X be any random variable. Then the expectation of X is defined in the following way

Discrete	Continuous
$E [X] = \sum_x xp(x)$ $p(x)$ - p.m.f	$E [X] = \int_{-\infty}^{\infty} xf(x)dx$ $f(x)$ - p.d.f

Example: Compute $E[X]$ if X has a density function given by

$$f(x) = \begin{cases} \frac{1}{4}xe^{-x/2}, & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Solution:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_0^{\infty} \frac{1}{4}x^2 e^{-x/2} dx \\ &= \frac{1}{4} \left[x^2 \int e^{-x/2} dx - \int (2x) \left(\int e^{-x/2} dx \right) dx \right]_{x=0}^{\infty} \\ &= \frac{1}{4} \left[-2x^2 e^{-x/2} + 4 \int xe^{-x/2} dx \right]_{x=0}^{\infty} \\ &= \frac{1}{4} \left[-2x^2 e^{-x/2} \right]_{x=0}^{\infty} + \left[\int xe^{-x/2} dx \right]_{x=0}^{\infty} \end{aligned}$$

$$\begin{aligned}
 E[X] &= \left[\int xe^{-x/2} dx \right]_{x=0}^{\infty} \\
 &= \left[-2xe^{-x/2} + 2 \int e^{-x/2} dx \right]_{x=0}^{\infty} \\
 &= \left[-2xe^{-x/2} - 4e^{-x/2} \right]_{x=0}^{\infty} \\
 &= 4
 \end{aligned}$$

Thus, $E[X] = 4$



Example: The density function of X is given by

$$f(x) = \begin{cases} a + bx^2, & 0 \leq x \leq 1, \\ 0, & \text{otherwise} \end{cases}$$

If $E[X] = \frac{3}{5}$, find a and b .

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

$$\implies E[X] = \int_0^1 (ax + bx^3)dx = \left[a\left(\frac{x^2}{2}\right) + b\left(\frac{x^4}{4}\right) \right]_{x=0}^{x=1} = \frac{a}{2} + \frac{b}{4}$$

$$\implies \frac{a}{2} + \frac{b}{4} = \frac{3}{5}$$

$$\implies 10a + 5b = 12 \longrightarrow ①$$

- Two unknowns, one equation!
- How do we get one more equation?

- ▶ **Total probability:** $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \int_0^1 (a + bx^2)dx \\&= \left[ax + b\left(\frac{x^3}{3}\right) \right]_{x=0}^{x=1} \\&= a + \frac{b}{3}\end{aligned}$$

- ▶ Thus, $a + \frac{b}{3} = 1 \implies 3a + b = 3 \rightarrow ②$
- ▶ We already have (by ①), $10a + 5b = 12$
- ▶ On solving, we get $a = \frac{3}{5}$ and $b = \frac{6}{5}$

□

Recall - Fact 7. If X is a discrete random variable that takes on one of the values $x_i, i \geq 1$, with respective probabilities $p(x_i)$, then, for any real-valued function g ,

$$E[g(X)] = \sum_i g(x_i)p(x_i)$$

We similarly have:

Fact 7(a). If X is a continuous random variable with probability density function $f(x)$, then, for any real-valued function g ,

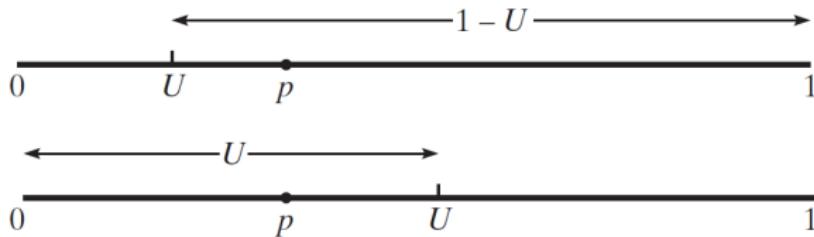
$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Recall - Fact 8. If a and b are constants, then

$$E[aX + b] = aE[X] + b$$

Example: A stick of length 1 is split at a point U having density function $f(u) = 1, 0 < u < 1$. Determine the expected length of the piece that contains a point $P, 0 \leq P \leq 1$

Solution:



If $L_P(U)$ denotes the length of the substick that contains the point P , then,

$$L_P(U) = \begin{cases} 1 - U, & \text{if } U < P, \\ U, & \text{if } U > P \end{cases}$$

By Fact 7(a),

$$\begin{aligned} E[L_P(U)] &= \int_{\infty}^{\infty} L_P(u)f(u)du \\ &= \int_0^1 L_P(u)du \\ &= \int_0^P (1-u)du + \int_P^1 udu \\ &= \frac{1}{2} + P(1-P) \end{aligned}$$



Problem: Suppose that if you are s minutes early for an appointment, then you incur the cost cs , and if you are s minutes late, then you incur the cost ks . Suppose also that the travel time from where you presently are to the location of your appointment is a continuous random variable having probability density function f . Determine the time at which you should depart if you want to minimize your expected cost.

- ▶ X denote the travel time
- ▶ If we leave t minutes before the appointment, then our cost, say, $C_t(X)$, is given by

$$C_t(X) = \begin{cases} c(t - X), & \text{if } X \leq t, \\ k(X - t), & \text{if } X \geq t \end{cases}$$

- ▶ We need to calculate $E[C_t(X)]$

$$\begin{aligned}
 E[C_t(X)] &= \int_0^{\infty} C_t(x) f(x) dx \\
 &= \int_0^t c(t-x) f(x) dx + \int_t^{\infty} k(x-t) f(x) dx \\
 &= ct \int_0^t f(x) dx - c \int_0^t xf(x) dx + k \int_t^{\infty} xf(x) dx \\
 &\quad - kt \int_t^{\infty} f(x) dx
 \end{aligned}$$

- ▶ We want to minimize this function $E[C_t(X)]$ (function of t)
- ▶ We differentiate it and equate it to 0

By calculus,

$$\begin{aligned}
 \frac{d}{dt} E[C_t(X)] &= ctf(t) + cF(t) - ctf(t) - ktf(t) + ktf(t) - k[1 - F(t)] \\
 &= (k + c)F(t) - k
 \end{aligned}$$

- ▶ $\frac{d}{dt} E[C_t(X)] = 0 \implies F(t) = \frac{k}{k+c}$
- ▶ Thus, to minimize the expected cost, we need to leave before t minutes before the appointment, where t satisfies $F(t) = \frac{k}{k+c}$. □

- ▶ $E[X^n] = \int_{-\infty}^{\infty} x^n f(x) dx$
- ▶ If X is a random variable with expected value μ , then the variance of X is defined (for any type of random variable) by

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

- ▶ The alternative formula,

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

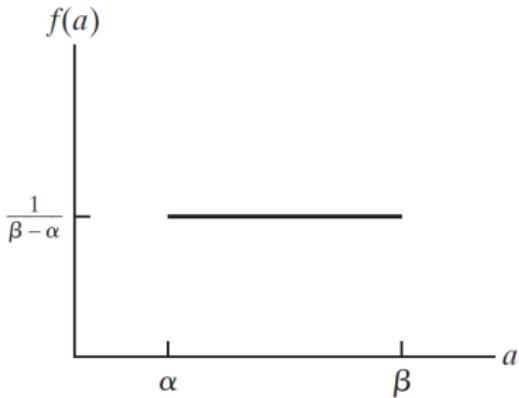
can be proved in a similar way as in discrete case

Exercise: Compute $\text{Var}(X)$ if X has a density function given by

$$f(x) = \begin{cases} \frac{1}{4}xe^{-x/2}, & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

We say that X is a **uniform random variable on the interval** (α, β) if the probability density function of X is given by

$$f(a) = \begin{cases} \frac{1}{\beta - \alpha}, & \alpha < a < \beta, \\ 0, & \text{else} \end{cases}$$

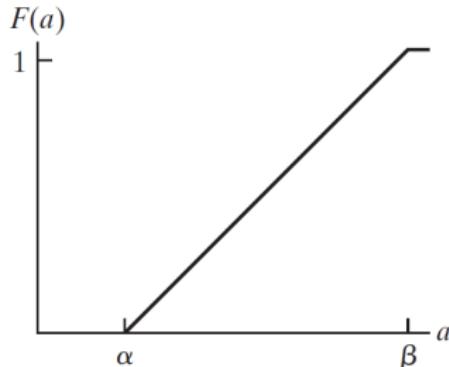


- ▶ Clearly, $\int_{-\infty}^{\infty} f(x)dx = \int_{\alpha}^{\beta} \frac{1}{\beta - \alpha} dx = 1$

- For $\alpha < a < b < \beta$, $P\{a \leq X \leq b\} = \int_a^b \frac{1}{\beta-\alpha} dx = \frac{b-a}{\beta-\alpha}$
- Using the relation $F(a) = \int_{-\infty}^a f(x)dx$, we can see that the distribution function of X is

$$F(a) = \begin{cases} 0, & a \leq \alpha \\ \frac{a-\alpha}{\beta-\alpha}, & \alpha < a < \beta, \\ 1, & a \geq \beta \end{cases}$$

Graph of F



☞ If X is uniformly distributed over (α, β) , then

- ▶ X takes values in the interval (α, β)
- ▶ Its density is given by

$$f(a) = \begin{cases} \frac{1}{\beta - \alpha}, & \alpha < a < \beta, \\ 0, & \text{else} \end{cases}$$

- ▶ The distribution function of X is

$$F(a) = \begin{cases} 0, & a \leq \alpha \\ \frac{a - \alpha}{\beta - \alpha}, & \alpha < a < \beta, \\ 1, & a \geq \beta \end{cases}$$

$$\blacktriangleright E[X] = \frac{\beta + \alpha}{2}$$

$$\blacktriangleright \text{Var}(X) = \frac{(\beta - \alpha)^2}{12}$$

Probability and Statistics

Lecture-15

☞ If X is uniformly distributed over (α, β) , then

- ▶ X takes values in the interval (α, β)
- ▶ Its density is given by

$$f(a) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha < a < \beta, \\ 0, & \text{else} \end{cases}$$

- ▶ The distribution function of X is

$$F(a) = \begin{cases} 0, & a \leq \alpha \\ \frac{a-\alpha}{\beta-\alpha}, & \alpha < a < \beta, \\ 1, & a \geq \beta \end{cases}$$

$$\text{▶ } E[X] = \frac{\beta+\alpha}{2}$$

$$\text{▶ } \text{Var}(X) = \frac{(\beta-\alpha)^2}{12}$$

Example: If X is uniformly distributed over $(0, 10)$, calculate the probability that (a) $X < 3$, (b) $X > 6$, and (c) $3 < X < 8$.

☞ We shall discuss this during the class!

Example: Trains headed for destination *A* arrive at the train station at 15-minute intervals starting at 7 a.m., whereas trains headed for destination *B* arrive at 15-minute intervals starting at 7 : 05 a.m.

- (a) If a certain passenger arrives at the station at a time uniformly distributed between 7 and 8 a.m. and then gets on the first train that arrives, what is the probability that he or she will end up at the destination *A*?
- (b) What if the passenger arrives at a time uniformly distributed between 7 : 10 and 8 : 10 a.m.?

 We shall discuss this during the class!

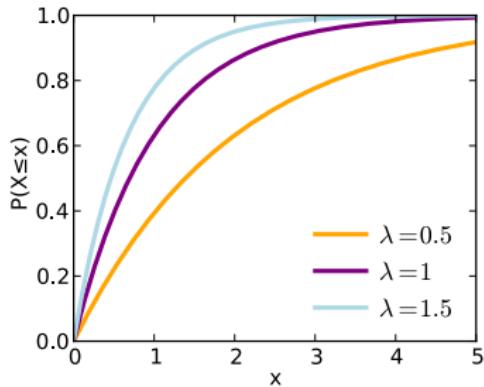
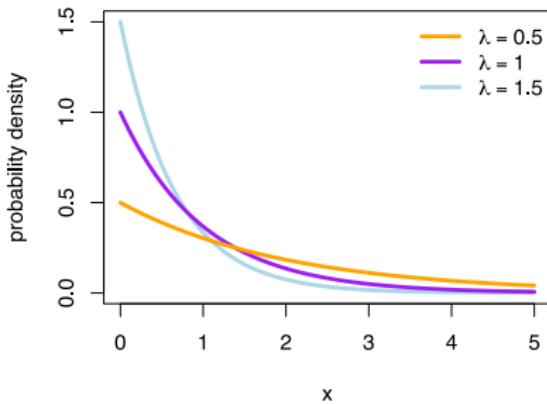
A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

$$f(a) = \begin{cases} \lambda e^{-\lambda a}, & \text{if } a \geq 0, \\ 0, & \text{else} \end{cases}$$

is said to be **an exponential random variable** (or, more simply, is said to be **exponentially distributed**) with parameter λ .

☞ Using the relation $F(a) = \int_{-\infty}^a f(x)dx$, we get the cumulative distribution function to be

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{else.} \end{cases}$$



 **Fact:** If X is an exponential random variable with parameter λ ,

$$E[X^n] = \frac{n}{\lambda} E[X^{n-1}]$$

 Using the above fact and $E[X^0] = 1$, we get

$$E[X] = \frac{1}{\lambda}$$

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

- ▶ In practice, the exponential distribution often arises as the distribution of the amount of time until some specific event occurs
- ▶ For instance, the amount of time (starting from now) until
 - ▶ an earthquake occurs, or
 - ▶ a new war breaks out, or
 - ▶ a telephone call you receive turns out to be a wrong number

Example: Suppose that the length of a phone call in minutes is an exponential random variable with parameter $\lambda = \frac{1}{10}$. If someone arrives immediately ahead of you at a public telephone booth, find the probability that you will have to wait

- (a) more than 10 minutes;
- (b) between 10 and 20 minutes.

☞ We shall discuss this during the class!

We say that a non-negative random variable X is **memoryless** if

$$P\{X > s + t | X > t\} = P\{X > s\} \text{ for all } s, t \geq 0$$

- ▶ Suppose that a random variable X represents waiting time
- ▶ Memoryless property means that the fact of having waited for t minutes gets “**forgotten**” and it does not affect the future waiting time
- ▶ Regardless of the event $X \geq t$, when the total waiting time exceeds t , the remaining waiting time still has same distribution of that of X
- ▶ The above condition can also be written as
$$P\{X > s + t\} = P\{X > s\}P\{X > t\}$$

- ▶ If X is an exponential random variable with parameter λ , then $P\{X > s + t\} = e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t} = P\{X > s\}P\{X > t\}$
- ▶ That is, exponential random variable has memoryless property!
- ▶ If X is a (discrete) random variable following geometric distribution with parameter p , then

$$P\{X > s+t\} = (1-p)^{(s+t)} = (1-p)^s(1-p)^t = P\{X > s\}P\{X > t\}$$

- ▶ Hence geometric random variable also has memoryless property!

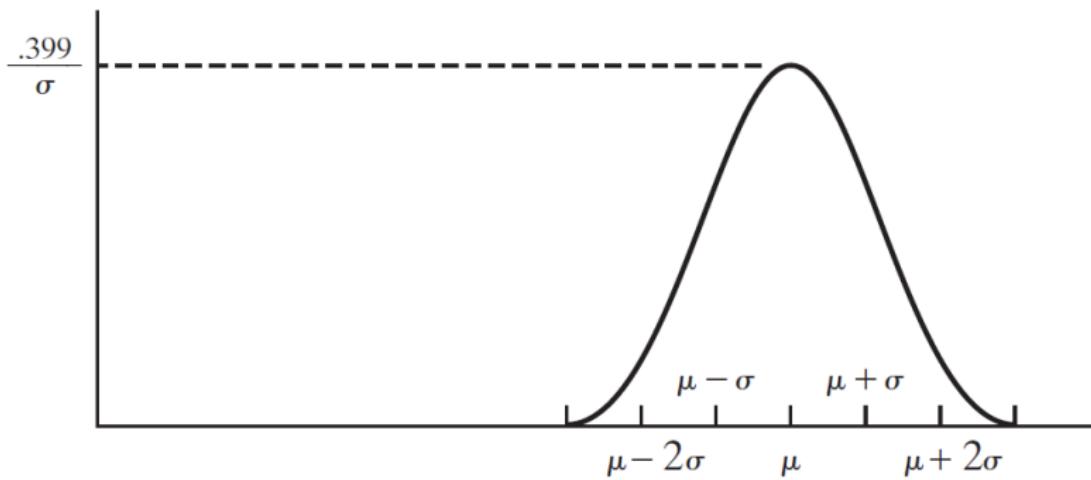
Example: The time (in hours) required to repair a machine is an exponentially distributed random variable with parameter $\lambda = \frac{1}{2}$. What is

- (a) the probability that a repair time exceeds 2 hours?
- (b) the conditional probability that a repair takes at least 10 hours, given that its duration exceeds 9 hours?

 We shall discuss this during the class!

We say that X is a **normal random variable**, or simply that X is normally distributed, with parameters μ and σ^2 if the density of X is given by

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



- ▶ To prove that $f(x)$ is indeed a probability density function, we need to show that

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx = 1$$

- ▶ By substituting $y = \frac{x-\mu}{\sigma}$, this boils down to showing that

$$\int_{-\infty}^{\infty} e^{-y^2/2} dy = \sqrt{2\pi}$$

- ▶ This is proved by using some “**convolution**” techniques!

- ☞ If X is a normal random variable with parameters μ and σ , then $Y = aX + b$ is also a normal random variable with parameters $a\mu + b$ and $a^2\sigma^2$
- ☞ In particular, if $Z = \frac{X-\mu}{\sigma}$, then Z is a normal random variable with parameters 0 and 1
- ☞ A normal random variable with parameters 0 and 1 is called the **standard normal random variable**.
- ☞ The alphabet Z is usually reserved for the standard normal random variable

For a normal random variable X with parameters μ and σ^2 ,

$$E[X] = \mu \text{ and } \text{Var}(X) = \sigma^2$$

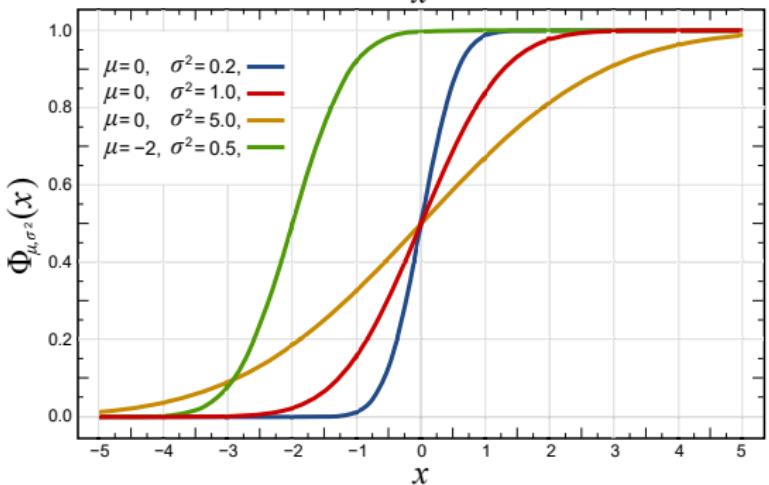
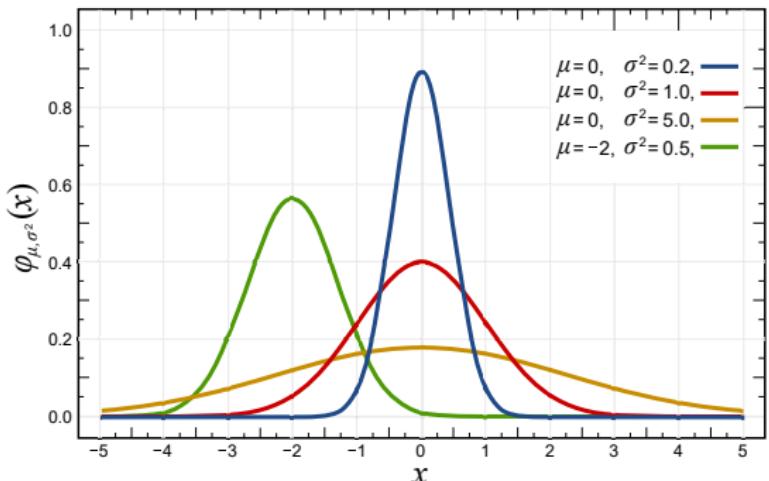
Probability and Statistics

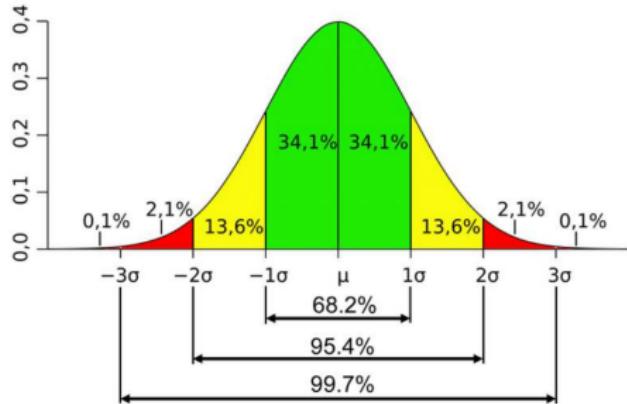
Lecture-16

- ▶ X is normally distributed, with parameters μ and σ^2 if the density of X is given by

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

- ▶ $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$
- ▶ $Y = aX + b$ is also a normal random variable with parameters $a\mu + b$ and $a^2\sigma^2$
- ▶ In particular, if $Z = \frac{X-\mu}{\sigma}$, then Z is a normal random variable with parameters 0 and 1
- ▶ A normal random variable with parameters 0 and 1 is called the **standard normal random variable**
- ▶ The alphabet Z is usually reserved for the standard normal random variable





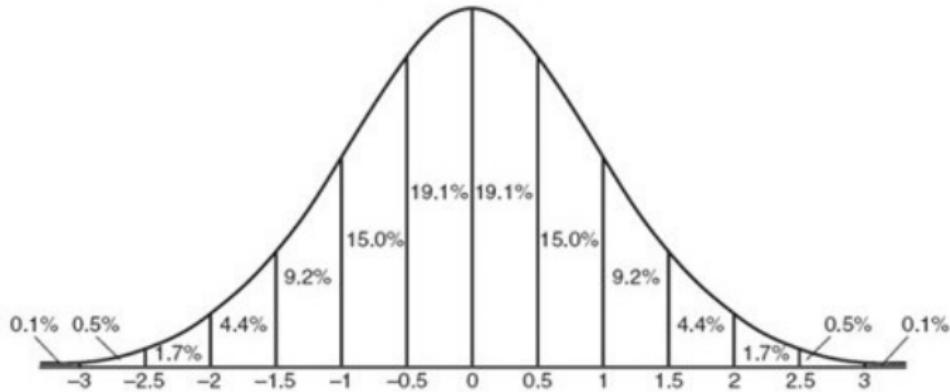
- ▶ $P\{\mu - \sigma < X < \mu + \sigma\} = 0.682$
 - ▶ $P\{\mu - 2\sigma < X < \mu + 2\sigma\} = 0.954$
 - ▶ $P\{\mu - 3\sigma < X < \mu + 3\sigma\} = 0.997$

- ▶ **Recall:** The probability density function (p.d.f) of standard normal variable Z is given by

$$\phi_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

- ▶ We denote the distribution function of Z **customarily** by $\Phi(z)$. That is,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx$$



- ▶ $P\{-1 < Z < 1\} = 0.682$
- ▶ $P\{-2 < Z < 2\} = 0.954$
- ▶ $P\{-3 < Z < 3\} = 0.997$
- ▶ The curve is 'symmetric' about Y-axis

That is, $P\{Z > z\} = P\{Z < -z\}$ for any z

$$\implies P\{Z < z\} = 1 - P\{Z < -z\} \text{ for any } z$$

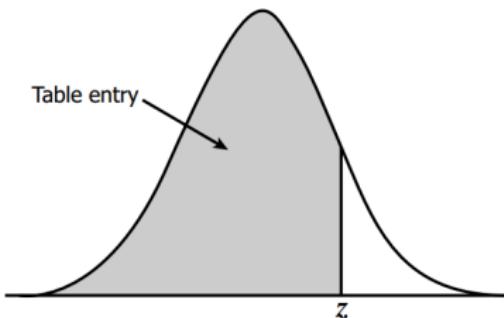
- ▶ **Question:** How do we compute $P\{a < Z < b\}$ for any a and b ?
- ▶ For instance, $P\{-1.75 < Z < 0.62\} = ?$
- ▶ We use the distribution function Φ
- ▶ $P\{-1.75 < Z < 0.62\} = \Phi(0.62) - \Phi(-1.75)$
- ▶ Now,

$$\Phi(0.62) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0.62} e^{-x^2/2} dx$$

$$\Phi(-1.75) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1.75} e^{-x^2/2} dx$$

- ▶ **Big question:** Is it easy to compute the integrals?
- ▶ **Answer:** No!
- ▶ We will instead use '**standard normal table**'

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998



- ▶ Table entries give $\Phi(z)$ for **positive values** of z

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790

- ▶ $\Phi(0.62) = 0.7324$

- ▶ What about negative values of z ? $\Phi(-1.75)$?
- ▶ **Recall:** $P\{Z < z\} = 1 - P(Z < -z)$ for any z . That is,
 $\Phi(z) = 1 - \Phi(-z)$.
- ▶ Thus, $\Phi(-1.75) = 1 - \Phi(1.75)$

z	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686

- ▶ $\Phi(-1.75) = 1 - \Phi(1.75) = 1 - 0.9599 = 0.0401$

☞ Thus,

$$P\{-1.75 < Z < 0.62\} = \Phi(0.62) - \Phi(-1.75) = 0.7324 - 0.0401 = 0.6923$$

☞ That was all about 'standard' normal variable

☞ For any normal variable X with parameters μ and σ^2 , how do we find $P\{c \leq X \leq d\}$?

☞ **Trick:** Transform it to standard form!

$$P\{c \leq X \leq d\} = P\left\{\frac{c-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{d-\mu}{\sigma}\right\} = P\{a \leq Z \leq b\}$$

where $a = \frac{c-\mu}{\sigma}$ and $b = \frac{d-\mu}{\sigma}$

☞ We know how to compute $P\{a \leq Z \leq b\}$!

Example: If X is a normal random variable with parameters $\mu = 3$ and $\sigma^2 = 9$, find

- (a) $P\{2 < X < 5\}$
- (b) $P\{X > 0\}$
- (c) $P\{|X - 3| > 6\}$

Solution: $Z = \frac{X-3}{3}$

(a)

$$\begin{aligned}P\{2 < X < 5\} &= P\left\{\frac{2-3}{3} < \frac{X-3}{3} < \frac{5-3}{3}\right\} \\&= P\left\{-\frac{1}{3} < Z < \frac{2}{3}\right\} \\&= \Phi\left(\frac{2}{3}\right) - \Phi\left(-\frac{1}{3}\right) \\&= \Phi\left(\frac{2}{3}\right) - \left[1 - \Phi\left(\frac{1}{3}\right)\right] \quad (\text{since } \Phi(-x) = 1 - \Phi(x)) \\&\approx 0.7486 - (1 - 0.6293) \quad (\textbf{standard normal table}) \\&= 0.3779\end{aligned}$$

$$\begin{aligned}(b) \ P\{X > 0\} &= P\left\{\frac{X - 3}{3} > \frac{0 - 3}{3}\right\} = P\{Z > -1\} \\&= 1 - P\{Z < -1\} = 1 - \Phi(-1) \\&= 1 - (1 - \Phi(1)) = \Phi(1) \approx 0.8413\end{aligned}$$

$$\begin{aligned}(c) \ P\{|X - 3| > 6\} &= P\{X - 3 > 6\} + P\{X - 3 < -6\} \\&= P\{X > 9\} + P\{X < -3\} \\&= P\{Z > 2\} + P\{Z < -2\} \\&= 1 - \Phi(2) + \Phi(-2) = 2(1 - \Phi(2)) \\&\approx 0.0456\end{aligned}$$



Example: Let X be a normal random variable with mean 12 and variance 4. Find the value of c such that $P\{X > c\} = 0.1$.

► Now, $X > c \implies \frac{X-12}{2} > \frac{c-12}{2}$

► Thus,

$$P\{X > c\} = P\left\{Z > \frac{c-12}{2}\right\} = 1 - P\left\{Z < \frac{c-12}{2}\right\} = 1 - \Phi\left(\frac{c-12}{2}\right)$$

► We are given that $1 - \Phi\left(\frac{c-12}{2}\right) = 0.1$

$$\implies \Phi\left(\frac{c-12}{2}\right) = 0.9$$

► From the standard normal table,

$$\frac{c-12}{2} \approx 1.28 \implies c \approx 14.56$$



Example: The systolic blood pressure in the population is usually modelled by a normal distribution with mean 120 mmHg (millimetres of mercury) and standard deviation 8 mmHg.

- (a) Below which blood pressure do we find one third of the population?
- (b) Above which blood pressure do we find 5% of the population?

Solution: Let X be the systolic blood pressure of a randomly selected individual.

Given that, X is a normal random variable with parameters 120 and 8^2

Let $Z = \frac{X-120}{8}$. Then Z is the standard normal random variable.

(a) Below which blood pressure do we find one third of the population?

We need to find c such that $P\{X < c\} = \frac{1}{3}$

$$\implies P\left\{Z < \frac{c-120}{8}\right\} = \frac{1}{3}$$

$$\implies \Phi\left(\frac{c-120}{8}\right) = \frac{1}{3}$$

$$\implies 1 - \Phi\left(-\frac{c-120}{8}\right) = \frac{1}{3} \quad (\text{as } \Phi(x) = 1 - \Phi(-x))$$

$$\implies \Phi\left(-\frac{c-120}{8}\right) = \frac{2}{3}$$

$$\implies -\frac{c-120}{8} \approx 0.43 \quad (\text{from the standard normal table})$$

$$\implies c \approx 116.56$$

In words, one third of the population has a blood pressure below 116.56 mmHg.

(b) Above which blood pressure do we find 5% of the population?

We need to find c such that $P\{X > c\} = 0.05$

$$\implies P\left\{Z > \frac{c-120}{8}\right\} = 0.05$$

$$\implies 1 - \Phi\left(\frac{c-120}{8}\right) = 0.05$$

$$\implies \Phi\left(\frac{c-120}{8}\right) = 0.95$$

$$\implies \frac{c-120}{8} \approx 1.645 \text{ (**from the standard normal table**)}$$

$$\implies c \approx 133.16$$

In words, 5% of the population have a blood pressure above 133.16 mmHg. □

Summary

☞ For any normal variable X with parameters μ and σ^2 , to find $P\{c \leq X \leq d\}$:

- ▶ **Step-1:** Transform the variable to standard variable

$$P\{c \leq X \leq d\} = P\{a \leq Z \leq b\}$$

where $a = \frac{c-\mu}{\sigma}$ and $b = \frac{d-\mu}{\sigma}$

- ▶ **Step-2:** $P\{a \leq Z \leq b\} = \Phi(b) - \Phi(a)$
- ▶ **Step-3:** Look for the values of $\Phi(a)$ and $\Phi(b)$ from the standard normal table. You may need use the relation $\Phi(z) = 1 - \Phi(-z)$ for negative values of z
- ▶ **Step-4:** Substitute these values to get the required probability

$$P\{c \leq X \leq d\} = \Phi(b) - \Phi(a)$$

Probability and Statistics

Lecture-17

- ▶ **Recall:** A random variable X is simply a function from the sample space S to the set of real numbers \mathbb{R} , i.e., $X : S \rightarrow \mathbb{R}$
- ▶ For example, in the experiment of rolling a pair of dice, X can be the sum of dice
- ▶ X can take on any value $2, 3, \dots, 12$.
- ▶ Here X is discrete
- ▶ If the random variable X is the blood pressure of a randomly selected person from a population, then X is a continuous random variable
- ▶ These random variables are functions of **one variable!**
- ▶ That is, $X \rightarrow \mathbb{R}$, the co-domain \mathbb{R} is just one-dimensional

- ▶ We can have functions of two/more variables on the sample space

$$(X, Y) : S \rightarrow \mathbb{R}^2$$

- ▶ For instance, X be the sum as above and Y be the value of the first die minus the second
- ▶ Thus, we can make sense of a pair of random variables and we can talk about its "**joint distribution**"
- ▶ That is, we can talk about the probabilities $P\{X \leq a, Y \leq b\}$

For any two random variables X and Y , the **joint cumulative probability distribution function of X and Y** is defined by

$$F(a, b) = P\{X \leq a, Y \leq b\}, \quad -\infty < a, b < \infty$$

- Now, looking only at X , the distribution function of X is

$$F_X(a) = P\{X \leq a\} = P\{X \leq a, Y < \infty\} = F(a, \infty) \text{ for any } a \in \mathbb{R}$$

- Similarly,

$$F_Y(b) = P\{Y \leq b\} = P\{X < \infty, Y \leq b\} = F(\infty, b) \text{ for any } b \in \mathbb{R}$$

- F_X and F_Y are referred to as the **marginal distributions** of X and Y

- ▶ All joint probability statements about X and Y can be answered by their joint distribution function

Example:

$$\begin{aligned} P\{X > a, Y > b\} &= 1 - P(\{X > a, Y > b\}^c) \\ &= 1 - P(\{\{X > a\} \cap \{Y > b\}\}^c) \\ &= 1 - P(\{\{X > a\}^c \cup \{Y > b\}^c\}) \\ &= 1 - P(\{\{X \leq a\} \cup \{Y \leq b\}\}) \\ &= 1 - [P\{X \leq a\} + P\{Y \leq b\} - P(\{\{X \leq a\} \cap \{Y \leq b\}\})] \\ &= 1 - P\{X \leq a\} - P\{Y \leq b\} + P\{X \leq a, Y \leq b\} \\ &= 1 - F_X(a) - F_Y(b) + F(a, b) \end{aligned}$$

$$\implies P\{X > a, Y > b\} = 1 - F_X(a) - F_Y(b) + F(a, b)$$

$$\begin{aligned}P\{a_1 < X \leq a_2, b_1 < Y \leq b_2\} &= F(a_2, b_2) + F(a_1, b_1) \\&\quad - F(a_1, b_2) - F(a_2, b_1)\end{aligned}$$

whenever $a_1 < a_2$ and $b_1 < b_2$.

- In the case of X and Y being discrete, we define the **joint probability mass function of X and Y** by

$$p(x, y) = P\{X = x, Y = y\}$$

- Using the joint probability mass function, one can obtain the **marginal** probability mass functions of X and Y respectively as follows:

$$p_X(x) = P\{X = x\} = \sum_{y:p(x,y)>0} p(x, y)$$

$$p_Y(y) = P\{Y = y\} = \sum_{x:p(x,y)>0} p(x, y)$$

Example: Suppose that 3 balls are randomly selected from an urn containing 3 red, 4 white, and 5 blue balls. If we let X and Y denote, respectively, the number of red and white balls chosen, calculate the joint probability mass function of X and Y

- ▶ X can take values 0, 1, 2, 3 and Y can take values 0, 1, 2, 3
- ▶ we need to find $p(i,j) = P\{X = i, Y = j\}$ for each $0 \leq i, j \leq 3$
- ▶ Instead of listing out all the values, we **always** represent the joint probability mass function in the form of a table!

Joint p.m.f of X and Y

$X = i \backslash Y = j$	0	1	2	3	$P\{X = i\}$
0					
1					
2					
3					
$P\{Y = j\}$					

- $(ij)^{\text{th}}$ cell of the table corresponds to $p(i,j) = P\{X = i, Y = j\}$
- Column sum gives the p.m.f of Y , $p_Y(j) = P\{Y = j\}$
- Row sum gives the p.m.f of X , $p_X(i) = P\{X = i\}$

Joint p.m.f of X and Y

$X = i \backslash Y = j$	0	1	2	3	$P\{X = i\}$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
$P\{Y = j\}$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	1

★ You can cross-check your calculations by checking that sum of the last column is 1 and sum of the last row is 1

Problem: Suppose that 3 balls are chosen without replacement from an urn consisting of 5 white and 8 red balls. Let X_i equal 1 if the i^{th} ball selected is white, and let it equal 0 otherwise. Give the joint probability mass function of X_1, X_2

Solution: Both X_1 and X_2 can take on values 0, 1

Joint p.m.f of X_1 and X_2

$X_1 = i \backslash X_2 = j$	0	1	$P\{X_1 = i\}$
0	$\frac{56}{156}$	$\frac{40}{156}$	$\frac{96}{156}$
1	$\frac{40}{156}$	$\frac{20}{156}$	$\frac{60}{156}$
$P\{X_2 = j\}$	$\frac{96}{156}$	$\frac{60}{156}$	1

Summary

- ▶ For any two random variables X and Y , the **joint cumulative probability distribution function of X and Y** is defined by

$$F(a, b) = P\{X \leq a, Y \leq b\}, \quad -\infty < a, b < \infty$$

- ▶ The **marginal distribution functions** of X and Y can be deduced in the following way:

$$F_X(a) = F(a, \infty) \text{ and } F_Y(b) = F(\infty, b) \text{ for any } -\infty < a, b < \infty$$

- ▶ All joint probability statements about X and Y can, in theory, be answered using their joint distribution function
- ▶ $P\{X > a, Y > b\} = 1 - F_X(a) - F_Y(b) + F(a, b)$

$$\begin{aligned}P\{a_1 < X \leq a_2, b_1 < Y \leq b_2\} &= F(a_2, b_2) + F(a_1, b_1) \\&\quad - F(a_1, b_2) - F(a_2, b_1)\end{aligned}$$

whenever $a_1 < a_2$ and $b_1 < b_2$.

Summary

- When X and Y are discrete, we define the **joint probability mass function of X and Y** by

$$p(x, y) = P\{X = x, Y = y\}$$

- Using the joint probability mass function, one can obtain the **marginal** probability mass functions of X and Y respectively as follows:

$$p_X(x) = P\{X = x\} = \sum_{y:p(x,y)>0} p(x, y)$$

$$p_Y(y) = P\{Y = y\} = \sum_{x:p(x,y)>0} p(x, y)$$

This was about discrete random variables

- ▶ Now suppose that X and Y are continuous
- ▶ We will have the **joint probability density function** $f(x, y)$ for $(x, y) \in \mathbb{R}^2$
- ▶ The **marginal density functions** can be computed by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

- ▶ For any $C \subset \mathbb{R}^2$,

$$P\{(X, Y) \in C\} = \iint_C f(x, y) dx dy$$

- ▶ If $C = \{(x, y) : x \in A, y \in B\}$, $A, B \subset \mathbb{R}$ then

$$P\{X \in A, Y \in B\} = \int_A \int_B f(x, y) dy dx$$

- ▶ The joint distribution function can be calculated using the joint density function as below

$$F(a, b) = \int_{-\infty}^a \int_{-\infty}^b f(x, y) dy dx$$

Example: The joint density function of X and Y is given by

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

Compute

- $P\{X > 1, Y < 1\},$
- $P\{X < Y\},$
- $P\{X < a\},$ and
- the joint distribution function

(a)

$$\begin{aligned}
 P\{X > 1, Y < 1\} &= \int_1^\infty \int_{-\infty}^1 f(x, y) dy dx \\
 &= \int_1^\infty \int_{-\infty}^1 2e^{-x} e^{-2y} dy dx \\
 &= 2 \int_1^\infty e^{-x} \int_{-\infty}^1 e^{-2y} dy dx \\
 &= e^{-1}(1 - e^{-2}) \quad (\text{check!})
 \end{aligned}$$

(b)

$$\begin{aligned}
 P\{X < Y\} &= \iint_{(x,y):x < y} 2e^{-x} e^{-2y} dx dy \\
 &= \int_0^\infty \int_0^y 2e^{-x} e^{-2y} dx dy \\
 &= \frac{1}{3} \quad (\text{check!})
 \end{aligned}$$

(c)

$$P\{X < a\} = \int_0^a \int_0^\infty 2e^{-x} e^{-2y} dy dx = 1 - e^{-a} \quad (\text{check!})$$

☞ Whenever $x \leq 0$ or $y \leq 0$, we have $f(x, y) = 0$

$$\implies F(a, b) = 0 \text{ when } a \leq 0 \text{ or } b \leq 0$$

☞ Now, for $a > 0$ and $b > 0$

$$\begin{aligned} F(a, b) &= \int_{-\infty}^a \int_{-\infty}^b f(x, y) \, dy \, dx \\ &= \int_0^a \int_0^b 2e^{-x-2y} \, dy \, dx \\ &= (1 - e^{-a}) (1 - e^{-2b}) \quad (\text{check!}) \end{aligned}$$



Example: The joint density of X and Y is given by

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

Find the density function of the random variable $\frac{X}{Y}$.

- ▶ The first step is to calculate the distribution function of $W = \frac{X}{Y}$
- ▶ Now, for $a \in \mathbb{R}$,

$$\begin{aligned} F_W(a) &= P\left\{\frac{X}{Y} \leq a\right\} \\ &= \iint_C f(x, y) \, dx \, dy \end{aligned}$$

where $C = \{(x, y) : \frac{x}{y} \leq a\}$ is a region in \mathbb{R}^2

- ▶ Whenever $a \leq 0$, for $\frac{x}{y} \leq a$ to be true, either $x \leq 0$ or $y \leq 0$
- ▶ Thus, for $a \leq 0$, $f(x, y) = 0 \implies F_W(a) = 0$ if $a \leq 0$

- ▶ For $a > 0$,

$$\begin{aligned}F_W(a) &= \iint_C f(x, y) dx dy \\&= \iint_C e^{-(x+y)} dx dy \\&= \int_0^\infty \int_0^{ay} e^{-(x+y)} dx dy = \int_0^\infty (1 - e^{-ay}) e^{-y} dy \\&= \left[-e^{-y} + \frac{e^{-(a+1)y}}{a+1} \right]_{y=0}^\infty = 1 - \frac{1}{a+1}\end{aligned}$$

- ▶ Thus,

$$F_W(a) = \begin{cases} 1 - \frac{1}{a+1}, & \text{if } a > 0, \\ 0, & \text{else.} \end{cases}$$

- ▶ We have $f_W(a) = \frac{d}{da} F_W(a)$
- ▶ Hence the probability density function of $W = \frac{X}{Y}$ is

$$f_W(a) = \begin{cases} \frac{1}{(a+1)^2}, & \text{if } a > 0, \\ 0, & \text{else.} \end{cases}$$



Probability and Statistics

Lecture-18

For any two sets A and B , if

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}$$

then we say that the random variables X and Y are **independent**.

Equivalent conditions:

- ▶ $P\{X \leq a, Y \leq b\} = P\{X \leq a\}P\{Y \leq b\}$ for any two numbers a and b
- ▶ $F(a, b) = F_X(a)F_Y(b)$ for any two numbers a and b
- ▶ If X and Y are discrete, $p(a, b) = p_X(a)p_Y(b)$ for any two numbers a and b
- ▶ If X and Y are continuous, $f(a, b) = f_X(a)f_Y(b)$ for any two numbers a and b

If X and Y are not independent, we call them to be **dependent**

Example: The joint probability mass function of two discrete random variables X and Y is as given below

$X = i \backslash Y = j$	0	1	2	3
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0
3	$\frac{1}{220}$	0	0	0

Are X and Y independent?

- ▶ Computing the sums of rows and columns we will get the marginal p.m.f

$X = i \backslash Y = j$	0	1	2	3	$P\{X = i\}$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
$P\{Y = j\}$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	1

$$P\{X = 0\} = \frac{84}{220}, P\{X = 1\} = \frac{108}{220}, P\{X = 2\} = \frac{27}{220}, P\{X = 3\} = \frac{1}{220}$$

$$P\{Y = 0\} = \frac{56}{220}, P\{Y = 1\} = \frac{112}{220}, P\{Y = 2\} = \frac{48}{220}, P\{Y = 3\} = \frac{4}{220}$$

☞ Now, $P\{X = 3\} P\{Y = 1\} = \frac{112}{(220)^2} \neq 0 = P\{X = 3, Y = 1\}$

⇒ X and Y are **NOT** independent!



Example: The joint density of X and Y is given by

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

Are X and Y independent?

- ▶ First step is calculate the marginal densities
- ▶ Now, for $x > 0$

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy = \int_0^{\infty} e^{-(x+y)} dy \\ &= e^{-x} \int_0^{\infty} e^{-y} dy = e^{-x} \end{aligned}$$

- ▶ Hence,

$$f_X(x) = \begin{cases} e^{-x}, & \text{for } x > 0 \\ 0, & \text{else.} \end{cases}$$

- ▶ Similarly,

$$f_Y(y) = \begin{cases} e^{-y}, & \text{for } y > 0 \\ 0, & \text{else.} \end{cases}$$

- ▶ Clearly,

$$f(x, y) = f_X(x)f_Y(y) \text{ for every } 0 < x, y < \infty$$

$\implies X$ and Y independent!



Example: Suppose that $n + m$ independent trials having a common probability of success p are performed. Let X be the number of successes in the first n trials, Y be the number of successes in the final m trials, and Z be the number of successes in the $n + m$ trials.

(a) Are X and Y independent?

(b) Are X and Z independent?

- ▶ We have $X \sim \text{Bin}(n,p)$, $Y \sim \text{Bin}(m,p)$ and $Z \sim \text{Bin}(n+m,p)$
- ▶ $\{X = i, Y = j\}$ is the event of having i successes in n trials and j successes in m trials

$$\begin{aligned}\implies P\{X = i, Y = j\} &= \binom{n}{i} \binom{m}{j} p^{i+j} (1-p)^{n+m-i-j} \\ &= \left\{ \binom{n}{i} p^i (1-p)^{n-i} \right\} \cdot \left\{ \binom{m}{j} p^j (1-p)^{m-j} \right\} \\ &= P\{X = i\} P\{Y = j\}\end{aligned}$$

for any $1 \leq i \leq n$, $1 \leq j \leq m$

$\implies X$ and Y are independent

- ▶ Now, observe that $Z = X + Y$

$$\begin{aligned} P\{X = i, Z = j\} &= P\{X = i, Y = j - i\} \\ &= \binom{n}{i} \binom{m}{j-i} p^j (1-p)^{n+m-j} \end{aligned}$$

for $1 \leq i \leq n, i \leq j \leq n + m$

- ▶ Whereas,

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i} \text{ and } P\{Z = j\} = \binom{n+m}{j} p^j (1-p)^{n+m-j}$$

- ▶ Thus, $P\{X = i, Z = j\} \neq P\{X = i\} P\{Z = j\}$
- ▶ For instance, $P\{X = 1, Z = 1\} \neq P\{X = 1\} P\{Z = 1\}$
- ▶ Hence X and Z are **not** independent!

□

Example: Let X, Y, Z be independent and uniformly distributed over $(0, 1)$. Compute $P\{X \geq YZ\}$.

Solution: Since X, Y, Z are independent, we get

$$f_{X,Y,Z}(x,y,z) = f_X(x)f_Y(y)f_Z(z) = 1, \quad 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1$$

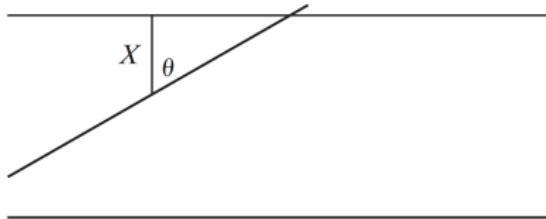
$$\begin{aligned} \text{Now, } P\{X \geq YZ\} &= \iiint_{x \geq yz} dx dy dz \\ &= \int_0^1 \int_0^1 \int_{yz}^1 dx dy dz \\ &= \int_0^1 \int_0^1 (1 - yz) dy dz \\ &= \int_0^1 \left(1 - \frac{z}{2}\right) dz \\ &= \frac{3}{4} \end{aligned}$$



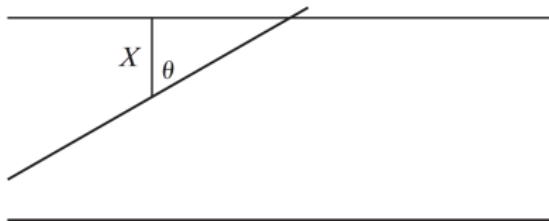
Problem (Buffon's needle problem, 18th century)

A table is ruled with equidistant parallel lines a distance D apart. A needle of length L , where $L \leq D$, is randomly thrown on the table. What is the probability that the needle will intersect one of the lines (the other possibility being that the needle will be completely contained in the strip between two lines)?

Solution:



- X = the distance from the middle point of the needle to the nearest parallel line
- θ = the angle between the needle and the projected line of length X



- If the length of the hypotenuse of the right-angled triangle thus formed is H , then $X = H \cos \theta$
- The needle will intersect a line if the hypotenuse H of the right triangle is less than $\frac{L}{2}$, i.e.,

$$H < \frac{L}{2} \iff \frac{X}{\cos \theta} < \frac{L}{2} \text{ or } X < \frac{L \cos \theta}{2}$$

- Now, since the needle is randomly thrown, it is reasonable to assume that X and θ are independent uniform random variables on $(0, \frac{D}{2})$ and $(0, \frac{\pi}{2})$

- We have, X uniform on $(0, \frac{D}{2})$, θ uniform on $(0, \frac{\pi}{2})$ and we want to find

$$P\left\{X < \frac{L \cos \theta}{2}\right\}$$

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

- ▶ **Question:** For two discrete **independent** random variables X and Y with p.m.fs p_X and p_Y , what can we say about the p.m.f of $W = X + Y$?
- ▶ Since X and Y are independent, their joint p.m.f is just the product of marginal p.m.fs

$$p(x, y) = p_X(x)p_Y(y) \text{ for every } x, y$$

- ▶ Now, the p.m.f of W is $p_W(w) = P\{X + Y = w\}$
- ▶ Observe that

$$\{X + Y = w\} = \bigcup_x \{X = x, Y = w - x\}$$

$$\begin{aligned} \text{Hence } p_W(w) &= P\{X + Y = w\} \\ &= \sum_x P\{X = x, Y = w - x\} \\ &= \sum_x P\{X = x\}P\{Y = w - x\} \quad (\text{independence}) \\ &= \sum_x p_X(x)p_Y(w - x) \\ \implies p_{X+Y}(w) &= \sum_x p_X(x)p_Y(w - x) \end{aligned}$$

☞ The process

$$\sum_x p_X(x)p_Y(w - x) = p_X * p_Y(w)$$

is called the **discrete convolution** of the p.m.fs p_X and p_Y

☞ That is, $p_{X+Y} = p_X * p_Y$

Question: Let X and Y be independent binomial random variables with respective parameters (n, p) and (m, p) .

Is $X + Y$ binomial?

- By the discrete convolution, the p.m.f of $X + Y$ is

$$\begin{aligned} p_{X+Y}(k) &= \sum_{i=0}^k p_X(i)p_Y(k-i) \\ &= \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \binom{m}{k-i} p^{k-i} (1-p)^{m-k+i} \\ &= p^k (1-p)^{n+m-k} \sum_{i=0}^k \binom{n}{i} \binom{m}{k-i} \end{aligned}$$

Exercise: $\binom{n+m}{k} = \sum_{i=0}^k \binom{n}{i} \binom{m}{k-i}$

$$\implies p_{X+Y}(k) = P\{X+Y = k\} = \binom{n+m}{k} p^k (1-p)^{n+m-k}, 0 \leq k \leq n+m$$

- ▶ If $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$ are independent, then

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

☞ By similar computations one can prove the following:

- ▶ If X_1, X_2, \dots, X_n are independent Bernoulli random variables with the same “success” parameter p , then

$$X_1 + X_2 + \dots + X_n \sim \text{Bin}(n,p)$$

- ▶ If X_1, X_2, \dots, X_n are independent Poisson random variables with parameters $\lambda_1, \lambda_2, \dots$, then

$$X_1 + X_2 + \dots + X_n \sim \text{Poisson}(\lambda_1 + \lambda_2 + \dots + \lambda_n)$$

- ▶ If X_1, X_2, \dots, X_n are independent geometric random variables with the same “success” parameter p , then

$$X_1 + X_2 + \dots + X_n \sim \text{NB}(n, p)$$

Probability and Statistics

Lecture-19

Recall

☞ Discrete convolution of p.m.fs

$$p_{X+Y} = \sum_w p_X(x)p_Y(w-x) = p_X * p_Y$$

☞ If X_1, X_2, \dots, X_n are independent random variables which are

- ▶ Bernoulli with the same “success” parameter p , then

$$X_1 + X_2 + \dots + X_n \sim \text{Bin}(n, p)$$

- ▶ binomial with parameters $(k_1, p), (k_2, p), \dots, (k_n, p)$, then

$$X_1 + X_2 + \dots + X_n \sim \text{Bin}(k_1 + k_2 + \dots + k_n, p)$$

- ▶ Poisson with parameters $\lambda_1, \lambda_2, \dots$, then

$$X_1 + X_2 + \dots + X_n \sim \text{Poisson}(\lambda_1 + \lambda_2 + \dots + \lambda_n)$$

- ▶ geometric with the same “success” parameter p , then

$$X_1 + X_2 + \dots + X_n \sim \text{NB}(n, p)$$

- ▶ What can we say about sum of independent '**continuous**' random variables?
- ▶ We have **continuous** convolution of p.d.fs
- ▶ If X and Y are independent continuous random variables with probability density functions f_X and f_Y , then the density of $X + Y$ is given by the convolution

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(x)f_Y(a-x) dx$$

- ▶ **Question:** If X and Y are independent uniform random variables on $(0, 1)$, is $X + Y$ again a uniform random variable?
- ▶ **Answer:** No!

- ▶ Since X and Y are uniform on $(0, 1)$, we have

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{else} \end{cases}$$

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & \text{else} \end{cases}$$

- ▶ By the convolution formula,

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(x)f_Y(a-x) dx \text{ for any } a \in \mathbb{R}$$

- ▶ Since f_X and f_Y are zero outside the interval $(0, 1)$, we need to be extra careful about the limits of integration

- ▶ $f_X(x)f_Y(a-x) \neq 0 \iff f_X(x) \neq 0$ and $f_Y(a-x) \neq 0$
- ▶ $f_X(x) = 0$ whenever $x \notin (0, 1)$
- ▶ Hence $0 < x < 1$
- ▶ When $0 \leq a \leq 1$,
 - ▶ we have $-x \leq a - x \leq 1 - x < 1$
 - ▶ Since $f_Y(a-x) = 0$ whenever $a-x \leq 0$, we will only consider $0 < a-x \leq 1-x < 1$
 - ▶ In particular, $0 < a-x$. That is, $x < a$
 - ▶ Hence the limits of integration will be 0 to a when $0 \leq a \leq 1$

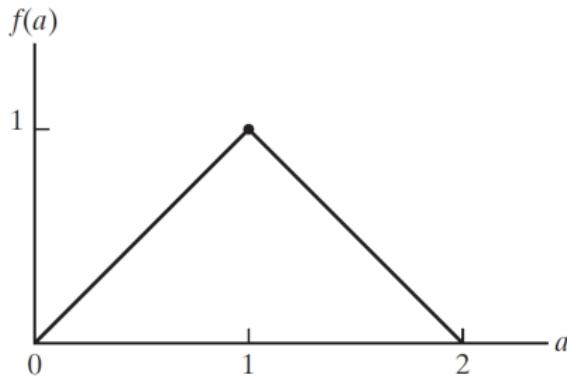
$$\implies f_{X+Y}(a) = \int_0^a dx = a \text{ when } 0 \leq a \leq 1$$

- ▶ Now, when $1 < a < 2$,
 - ▶ we have $0 < 1 - x < a - x < 2 - x$
 - ▶ Since $2 - x > 1$ (as $0 < x < 1$), we need to consider the limits only up to 1
 - ▶ $\implies 0 < 1 - x < a - x < 1$
 - ▶ In particular, $a - x < 1 \implies x > a - 1$
 - ▶ Hence the limits of integration will be $a - 1$ to 1 when $1 < a < 2$

$$\implies f_{X+Y}(a) = \int_{a-1}^1 dx = 2 - a \text{ when } 1 < a < 2$$

- ▶ For every other value of a , we have $f_Y(a - x) = 0$

$$\text{Thus, } f_{X+Y}(a) = \begin{cases} a, & 0 \leq a \leq 1, \\ 2 - a, & 1 < a < 2 \\ 0, & \text{otherwise} \end{cases}$$



- Hence, sum of two uniform random variables is **NOT** uniform!

- ☞ By a similar computation, we can see that sum of two independent exponential random variable **will not** be exponential
- ☞ This is not the case with independent normal variables!

If X_1, X_2, \dots, X_n are independent normal random variables with parameters $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \dots, (\mu_n, \sigma_n^2)$, then

$$X_1 + X_2 + X_3 + \dots + X_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

- ☞ In particular, if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent, then

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Example: You wait for an elevator, whose capacity is 2000 lbs.

The elevator comes with ten adult passengers. Suppose your own weight is 150 lbs, and you heard that human weights are normally distributed with the mean of 165 lbs and the standard deviation of 20 lbs. Would you board this elevator or wait for the next one? In other words, is overload likely?

- ▶ Let X_i denote the weight of i^{th} person in the lift,
 $i = 1, 2, \dots, 10$
- ▶ Since the elevator already has 10 passengers, we are interested in sum of their weights

$$Y_{10} = X_1 + X_2 + \dots + X_{10}$$

- ▶ X_1, X_2, \dots, X_{10} are normally distributed with parameters $(165, 20^2)$ and are independent of each other
- ▶ Hence their sum Y_{10} is also normally distributed with mean $10 \times 165 = 1650$ and variance $10 \times 20^2 = 4000$

- ▶ Overload happens if $Y_{10} + 150 > 2000$
- ▶ Hence we need to find $P \{ Y_{10} + 150 > 2000 \}$

$$\begin{aligned}
 P \{ Y_{10} + 150 > 2000 \} &= P \{ Y_{10} > 1850 \} \\
 &= P \left\{ \frac{Y_{10} - 165(10)}{20(\sqrt{10})} > \frac{1850 - 165(10)}{20(\sqrt{10})} \right\} \\
 &= P \{ Z > 3.16 \} \\
 &= 1 - P \{ Z \leq 3.16 \} \\
 &\approx 1 - \Phi(3.16) \quad (\textbf{standard normal table}) \\
 &\approx 1 - 0.9992 = 0.0008
 \end{aligned}$$

☞ Thus, there is only 0.08% chance of overload.

☞ Hence it is safe to take the elevator!



- ▶ Random variables X_1, X_2, \dots are said to be **identically** distributed if each of X_i has the exactly same distribution. That is,

$$F_{X_i}(x) = F_{X_j}(x) \text{ for every } x \text{ and } i \geq 1, j \geq 1$$

- ▶ In addition, if they are independent of each other, then we call them **independent and identically distributed** (abbreviated as i.i.d.) random variables

☞ Some examples:

- ▶ A sequence of outcomes of spins of a fair or unfair roulette wheel
- ▶ Any kind of random noise such as the hiss on a telephone line
- ▶ The ages of people queuing up at a bus-stand

- If X_1, X_2, \dots, X_n are i.i.d. normal random variables with mean μ and variance σ^2 , we know that $X_1 + X_2 + \dots + X_n$ is also normal with mean $\sum_{i=1}^n \mu = n\mu$ and variance $\sum_{i=1}^n \sigma^2 = n\sigma^2$
- What can we say when X_1, X_2, \dots, X_n are i.i.d. **but not normal?**

Central limit theorem (CLT)

If X_1, X_2, \dots is a sequence of i.i.d. random variables, each having mean μ and variance σ^2 . Then, for large values of n , $X_1 + X_2 + \dots + X_n$ is normally distribution with mean $n\mu$ and variance $n\sigma^2$.

- In such a case, for large n , by “standardizing”,

$$\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \sim N(0, 1)$$

☞ How do we apply CLT?

- ▶ In several instances, we come across i.i.d. random variables X_1, X_2, \dots
- ▶ We might have information about their mean μ and variance σ^2 but we may not be knowing the exact distribution of each X_i
- ▶ In such a situation, how do we calculate probabilities involving any (finite) sum of these random variables?
- ▶ That is, if $Y_n = X_1 + X_2 + \dots + X_n$, how do compute probabilities like $P\{c < Y_n < d\}$?

Idea: CLT says that, for large values of n , $Y_n \sim N(n\mu, n\sigma^2)$

- ▶ Then, by standardizing, $Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}} \sim N(0, 1)$

- ▶ Thus,

$$P\left\{c < Y_n < d\right\} = P\left\{\frac{c-n\mu}{\sigma\sqrt{n}} < Z_n < \frac{d-n\mu}{\sigma\sqrt{n}}\right\} = \Phi\left(\frac{d-n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{c-n\mu}{\sigma\sqrt{n}}\right)$$

(Recall: Φ is the distribution function of the standard normal variable Z)

- ▶ And then we use the standard normal table!

☞ **Rule of thumb:** We apply CLT only when $n > 30$!

Example: A disk has free space of 330 megabytes. Is it likely to be sufficient for 300 independent images, if each image has expected size of 1 megabyte with a standard deviation of 0.5 megabytes?

- ▶ Let X_i denote the size of i^{th} image in megabytes
- ▶ X_1, X_2, \dots are i.i.d. with mean 1 and variance $(0.5)^2 = 0.25$
- ▶ We are interested in the combined size of 300 images
- ▶ Let $Y_{300} = X_1 + X_2 + \dots + X_{300}$ be the combined size
- ▶ We wish to compute $P\{Y_{300} \leq 330\}$

- ▶ Since the $300 > 30$, we use the central limit theorem

$$\begin{aligned} P\{Y_{300} \leq 330\} &= P\left\{\frac{Y_{300} - 300(1)}{(0.5)\sqrt{300}} \leq \frac{330 - 300(1)}{(0.5)\sqrt{300}}\right\} \\ &= P\{Z_{300} \leq 3.46\} \\ &\approx \Phi(3.46) \quad (\textbf{central limit theorem}) \\ &\approx 0.9997 \quad (\textbf{standard normal table}) \end{aligned}$$

- ☞ Since the probability is very high, the available disk space is very likely to be sufficient for 300 images! □

Normal approximation to binomial distribution

- ▶ Suppose X_1, X_2, \dots be independent Bernoulli variables with the same 'success' probability p
- ▶ **Recall:** $E[X_i] = p = \mu$ (say) and $\text{Var}(X_i) = p(1 - p) = \sigma^2$ (say)
- ▶ We know that, for any n , $Y_n = X_1 + X_2 + \dots + X_n \sim \text{Bin}(n, p)$
- ▶ On the other hand, the CLT says that, for large values of n (> 30),

$$Y_n = X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2)$$

- ▶ **Moral:** We can assume binomial = normal for large values of n and certain values of p
- ▶ **Thumb rule:** We approximate binomial with normal if $np > 5$ and $n(1 - p) > 5$

Example: A new computer virus attacks a folder consisting of 200 files. Each file gets damaged with probability 0.2 independently of other files. What is the probability that fewer than 50 files get damaged?

- ▶ Let 'success' be the event of a file getting infected by virus
- ▶ Given that probability of success is $p = 0.2$
- ▶ Let X denote the number of infected files
- ▶ $X \sim \text{Bin}(200, p)$
- ▶ We have $n = 200$
 $\implies np = 40$ and $n(1 - p) = 160$
- ▶ Since both $np > 5$ and $n(1 - p) > 5$, we will assume that

$$X \sim N(np, np(1 - p))$$

- ▶ We have $X \sim N(\mu = 40, \sigma^2 = 32)$
- ▶ Required probability is $P\{X < 50\}$

$$\begin{aligned}P\{X < 50\} &= P\left\{\frac{X - 40}{\sqrt{32}} < \frac{50 - 40}{\sqrt{32}}\right\} \\&= P\{Z < 1.78\} \\&= \Phi(1.78) \\&\approx 0.9625 \quad (\text{standard normal table})\end{aligned}$$



- ☞ If we did not assume that X follows normal distribution, calculating $P\{X < 50\}$ would have been cumbersome!
- ☞ Thus, normal approximation will help us in reducing our calculations!

Probability and Statistics

Lecture-20

- ▶ **Recall:** If X_1, X_2, \dots, X_n are i.i.d. normal random variables with mean μ and variance σ^2 , then

$$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$$

- ▶ Let $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$
- ▶ $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \implies \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$
- ▶ Let $S = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})$
- ▶ Then S is also a random variable and in practice, we will be needing the distribution of

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

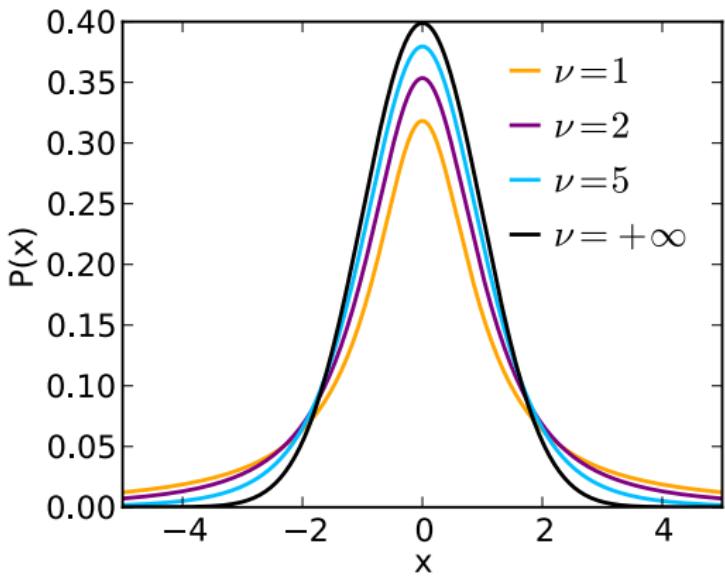
- ▶ It turns out that T follows ***Students' t-distribution*** with $(n - 1)$ degrees of freedom

We say that a random variable T is said to follow ***Student's t-distribution with ν degrees of freedom*** if it has the probability density function given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \text{ for } -\infty < t < \infty,$$

where $\nu \in \mathbb{N}$ is the number of degrees of freedom (df) and Γ is the gamma function.

- ☞ To compute probabilities we will use tables as we did in the normal case
- ☞ Student's t-distribution is a 'family' of distributions
- ☞ For different values of the parameter, df, ν , we get different distributions
- ☞ For any value of ν , the mean, mode and median of Student's t-distributed random variable will be 0



- ☞ As $\nu \rightarrow \infty$, t-distribution converges to standard normal distribution (black curve in the above picture)
- ☞ It is similar to having a 'kind' of standard normal distribution for each value of ν
- ☞ Thus, we will be having different tables for different values of ν

☞ Since we have to maintain large of number of tables, we instead use online calculators for obtaining probabilities

☞ We will be using the calculator given at

<https://stattrek.com/online-calculator/t-distribution.aspx>

Example: If T is a random variable following t-distribution with df $\nu = 13$, then compute $P \{-0.12 < T < 1\}$

- ▶ We know that $P \{-0.12 < T < 1\} = F_{13}(1) - F_{13}(-0.12)$, where F_{13} is the distribution function for df $\nu = 13$
- ▶ In the above calculator, we select 't-score' for the random variable
- ▶ We then enter 'Degrees of freedom' to be '13'
- ▶ Finally, to know the value of $F_{13}(-0.12)$, we enter '-0.12' against 't-score' and **leave the last field blank**

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Random variable ▾

Degrees of freedom

t score

Probability: $P(T \leq t)$

Calculate

☞ On clicking 'calculate', we get,

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Random variable	<input type="text" value="t score"/>
Degrees of freedom	<input type="text" value="13"/>
t score	<input type="text" value="-0.12"/>
Probability: $P(T \leq -0.12)$	<input type="text" value="0.4532"/>

Calculate

$$\implies F_{13}(-0.12) = 0.4532$$

☞ Similarly,

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Random variable

Degrees of freedom

t score

Probability: $P(T \leq 1)$

Calculate

$$\implies F_{13}(1) = 0.8322$$

☞ Thus, $P\{-0.12 < T < 1\} = F_{13}(1) - F_{13}(-0.12) = 0.8322 - 0.4532 = 0.379$



☞ We can also use the calculator to compute the t-scores for a given probability

Example: Find the value of c in $P\{T > c\} = 0.5672$, where T follows student's t-distribution with df $\nu = 19$

☞ $P\{T > c\} = 1 - P\{T < c\} = 1 - F_{19}(c)$

☞ Thus, $1 - F_{19}(c) = 0.5672 \implies F_{19}(c) = 0.4328$

☞ Now, in the calculator, we will leave the 't-score' field blank and enter *0.4328* in the 'Probability' field

☞ We get the t-score to be *-0.172*

☞ Hence $c = -0.172$



Tail probabilities/areas

☞ The probability that a random variable deviates by a given amount from its mean/expectation is referred to as a **tail probability/area**

☞ For a random variable X with mean μ , the statement

Probability that X deviates more than 3 from its mean

is equivalent to

$$P \{ \{X < \mu - 3\} \cup \{X > \mu + 3\} \} = P \{X < \mu - 3\} + P \{X > \mu + 3\}$$

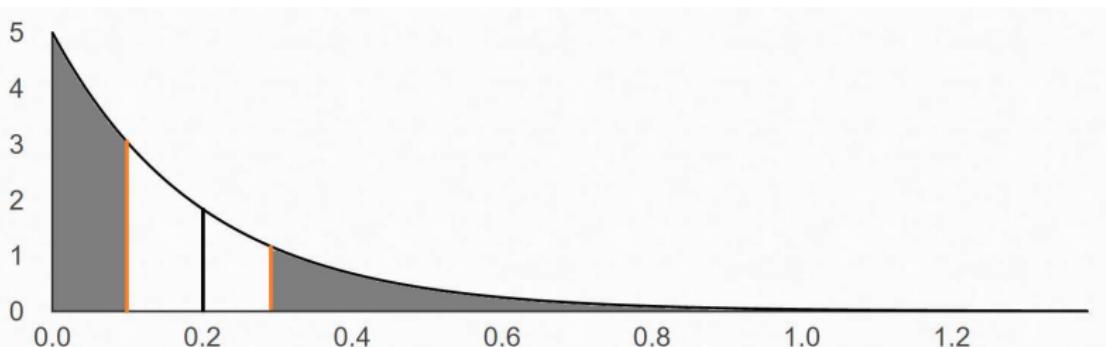
☞ Further, if the standard deviation is σ , the statement

Probability that X deviates more than 2 standard deviations from its mean

is equivalent to

$$P \{ \{X < \mu - 2\sigma\} \cup \{X > \mu + 2\sigma\} \} = P \{X < \mu - 2\sigma\} + P \{X > \mu + 2\sigma\}$$

- ☞ The above probabilities are **two-tailed** as we are considering deviations in both the directions
- ☞ For example, let $X \sim \exp(5)$
- ☞ Recall that $\mu = \frac{1}{\lambda} = 0.2 = \sigma$



$$P \{ \{X < \mu - 0.5\sigma\} \cup \{X > \mu + 0.5\sigma\} \} = P \{X < 0.1\} + P \{X > 0.3\}$$

☞ We could have **one-tailed** probabilities

☞ The statement,

Probability that X is 2 standard deviations above the mean

is equivalent to

$$P\{X > \mu + 2\sigma\}$$

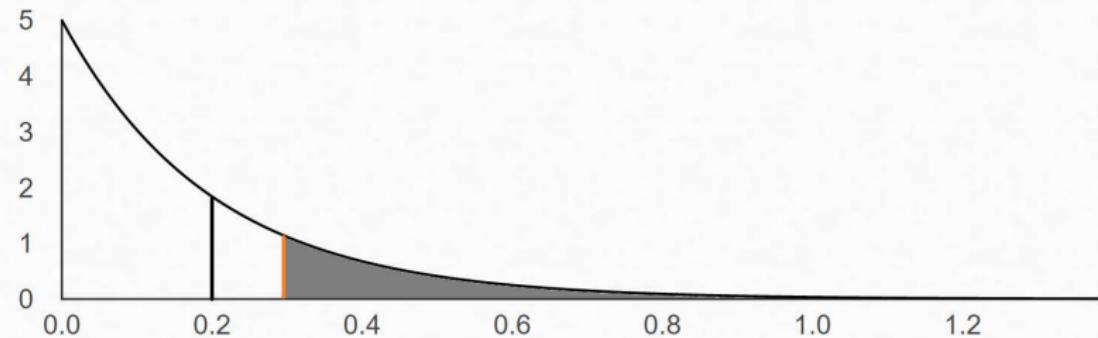
☞ Also, the statement,

Probability that X is 2 standard deviations below the mean

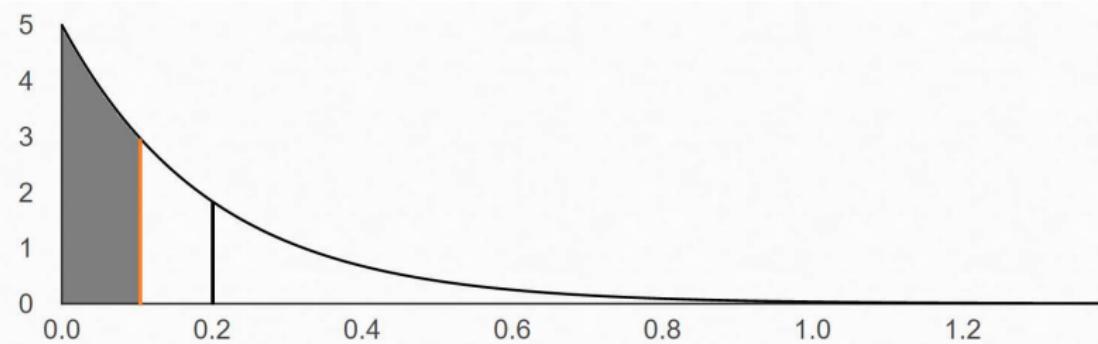
is equivalent to

$$P\{X < \mu - 2\sigma\}$$

☞ In our example, $X \sim \exp(5)$,

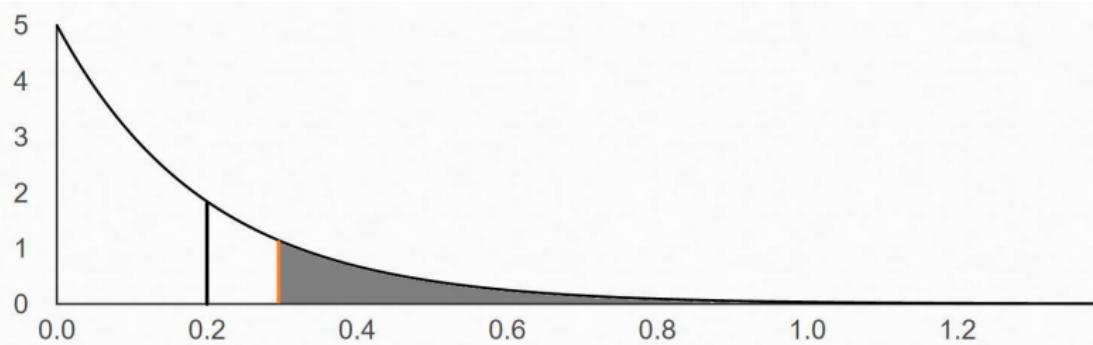


$$P\{X > \mu + 0.5\sigma\} = P\{X > 0.3\}$$

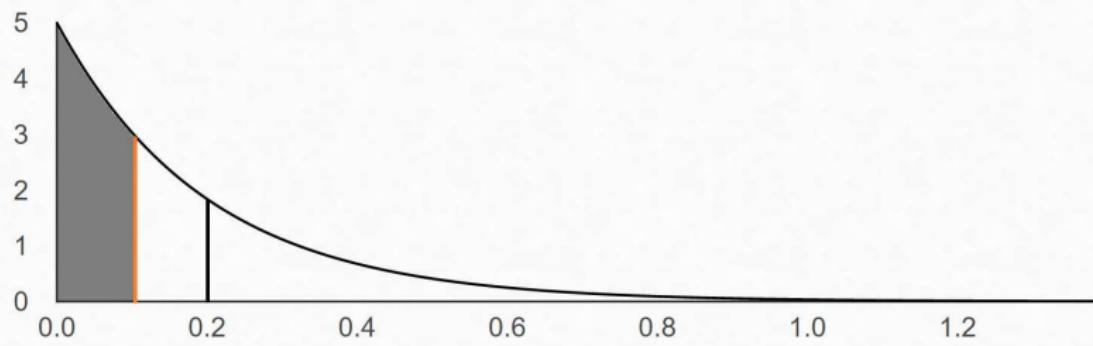


$$P\{X < \mu - 0.5\sigma\} = P\{X < 0.1\}$$

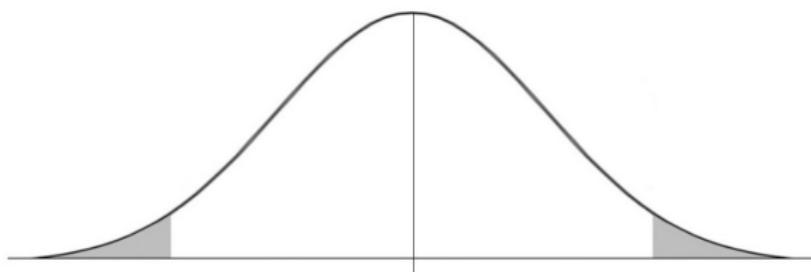
☞ The probabilities involving “above” mean statements (that is $P\{X > \mu + a\sigma\}$) are called **right-tailed** probabilities



☞ The probabilities involving “below” mean statements (that is $P\{X < \mu - a\sigma\}$) are called **left-tailed** probabilities



☞ In symmetric distributions like *normal* and *student's t*-distributions, the right-tail probabilities and left-tail probabilities (at the same level!) will be equal



☞ Thus, if X follows normal or student's t-distribution, then

$$\begin{aligned} P \{ \{X < \mu - a\sigma\} \cup \{X > \mu + a\sigma\} \} &= P \{X < \mu - a\sigma\} + P \{X > \mu + a\sigma\} \\ &= 2P \{X < \mu - a\sigma\} \end{aligned}$$

☞ If X follows normal or student's t-distribution, then the two-tailed probability of some deviation will be twice the one-tailed (left or right) probability of the same deviation

☞ In particular, if $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} P\{\{X < \mu - a\sigma\} \cup \{X > \mu + a\sigma\}\} &= 2P\{X < \mu - a\sigma\} \\ &= 2P\{Z < -a\} \\ &= 2\Phi(-a) \end{aligned}$$

$$\begin{aligned} P\{X > \mu + a\sigma\} &= P\{X < \mu - a\sigma\} \\ &= P\{Z < -a\} \\ &= \Phi(-a) \end{aligned}$$

Example: Suppose $X \sim N(3, 9)$

- (i) Compute the probability that X is 1.4 standard deviations above mean
- (ii) Find the probability that X deviates more than 1.8 standard deviations from mean

Solution: We have $X \sim N(3, 9)$. That is, $\mu = 3$ and $\sigma = 3$.

- (i) We need to compute $P\{X > \mu + (1.4)\sigma\}$

$$\begin{aligned} P\{X > \mu + (1.4)\sigma\} &= \Phi(-1.4) \\ &= 1 - \Phi(1.4) \\ &= 1 - 0.9192 = 0.0808 \end{aligned}$$

- (ii) We need to compute $P\{\{X < \mu - (1.8)\sigma\} \cup \{X > \mu + (1.8)\sigma\}\}$

$$\begin{aligned} P\{\{X < \mu - (1.8)\sigma\} \cup \{X > \mu + (1.8)\sigma\}\} &= 2\Phi(-1.8) \\ &= 2(1 - \Phi(1.8)) \\ &= 2(1 - 0.9641) = 0.0718 \end{aligned}$$

Probability and Statistics

Lecture-21

 Recall:

- ▶ For a random variable X , the expectation of X is defined as

$$E[X] = \begin{cases} \sum_x x p_X(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}$$

- ▶ If $h : \mathbb{R} \rightarrow \mathbb{R}$ is any function, then

$$E[h(X)] = \begin{cases} \sum_x h(x) p_X(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} h(x) f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}$$

- ▶ In particular,

$$E[X^n] = \begin{cases} \sum_x x^n p_X(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f_X(x) dx, & \text{if } X \text{ is continuous} \end{cases}$$

☞ Now, for two random variables X and Y ,

$$E[XY] = \begin{cases} \sum_x \sum_y xy p(x,y), & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x,y) dx dy, & \text{if } X \text{ and } Y \\ & \text{are continuous} \end{cases}$$

☞ For any two functions $h : \mathbb{R} \rightarrow \mathbb{R}$, $g : \mathbb{R} \rightarrow \mathbb{R}$

$$E[h(X)g(Y)] = \begin{cases} \sum_x \sum_y h(x)g(y) p(x,y), & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)g(y) f(x,y) dx dy, & \text{if } X \text{ and } Y \\ & \text{are continuous} \end{cases}$$

Example-1: In a study of response times (Y) versus the number of bars of signal strength (X) of a particular network, the joint p.m.f is given below.

$y = \text{Response Time}$ (nearest second)	$x = \text{Number of Bars of Signal Strength}$		
	1	2	3
4	0.15	0.1	0.05
3	0.02	0.1	0.05
2	0.02	0.03	0.2
1	0.01	0.02	0.25

Compute the 'joint' expectation $E[XY]$

☞ Since $E[XY] = \sum_{y=1}^4 \sum_{x=1}^3 xyp(x,y)$, we first compute $xyp(x,y)$ for each $x = 1, 2, 3$ and $y = 1, 2, 3, 4$

We tabulate the values of $xyp(x, y)$ as below:

$X = i$	1	2	3
$Y = j$			
4	0.6	0.8	0.6
3	0.06	0.6	0.45
2	0.04	0.12	1.2
1	0.01	0.04	0.75

Summing up all the entries, we get $E[XY] = 5.27$



Example-2: The random variables X and Y have a joint density function given by

$$f(x,y) = \begin{cases} 24xy, & 0 < x < 1, 0 < y < 1, 0 < x + y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Find $E[XY]$.

$$\begin{aligned} \text{We have } E[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y) dx dy \\ &= \int_0^1 \int_0^{1-y} 24x^2y^2 dx dy \\ &= \int_0^1 24y^2 \left(\int_0^{1-y} x^2 dx \right) dy \\ &= \int_0^1 8y^2(1-y)^3 dy \\ &= 8 \left[\frac{y^3}{3} - \frac{y^6}{6} + \frac{3y^5}{5} - \frac{3y^4}{4} \right]_{y=0}^{y=1} \\ &= \frac{2}{15} \end{aligned}$$

Fact: If X and Y are independent, then, for any functions h and g ,

$$E[h(X)g(Y)] = E[h(X)]E[g(Y)]$$

☞ In particular, for independent random variables X and Y ,

$$E[XY] = E[X]E[Y]$$

The **covariance** between any two random variables X and Y , denoted by $\text{Cov}(X, Y)$, is defined by

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

Fact: $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$

Proof: Let $a = E[X]$ and $b = E[Y]$

$$\begin{aligned}\text{Now, } \text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[(X - a)(Y - b)] \\ &= E[XY - bX - aY + ab] \\ &= E[XY] - bE[X] - aE[Y] + ab \\ &= E[XY] - ab \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

★ If X and Y are independent, we have

$$E[XY] = E[X]E[Y] \implies \text{Cov}(X, Y) = 0$$

★ **Caution:** $\text{Cov}(X, Y) = 0$ does not mean that X and Y are independent!

Example: In a study of response times (Y) versus the number of bars of signal strength (X) of a particular network, the joint p.m.f is given below.

$y = \text{Response Time}$ (nearest second)	$x = \text{Number of Bars of Signal Strength}$		
	1	2	3
4	0.15	0.1	0.05
3	0.02	0.1	0.05
2	0.02	0.03	0.2
1	0.01	0.02	0.25

Compute the $\text{Cov}(X, Y)$.

☞ In Example-1 we computed $E[XY] = 5.27$

☞ We have $E[X] = \sum_x x p_X(x)$ and $E[Y] = \sum_y y p_Y(y)$

☞ First step is to compute the marginal p.m.fs - p_X and p_Y

$y = \text{Response Time}$ (nearest second)	$x = \text{Number of Bars of Signal Strength}$			Marginal Probability Distribution of Y
	1	2	3	
4	0.15	0.1	0.05	0.3
3	0.02	0.1	0.05	0.17
2	0.02	0.03	0.2	0.25
1	0.01	0.02	0.25	0.28
	0.2	0.25	0.55	
Marginal Probability Distribution of X				

$$p_X(X=1) = 0.2, p_X(X=2) = 0.25, p_X(X=3) = 0.55$$

$$p_Y(Y=1) = 0.28, p_Y(Y=2) = 0.25, p_Y(Y=3) = 0.17, p_Y(Y=4) = 0.3$$

 Hence,

$$E[X] = 1(0.2) + 2(0.25) + 3(0.55) = 2.35$$

$$E[Y] = 1(0.28) + 2(0.25) + 3(0.17) + 4(0.3) = 2.49$$

 Thus,

$$\begin{aligned}\text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= 3.558 - (2.35)(2.49) \\ &= -0.5815\end{aligned}$$

Example: The random variables X and Y have a joint density function given by

$$f(x,y) = \begin{cases} 24xy, & 0 < x < 1, \ 0 < y < 1, \ 0 < x + y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Compute the covariance between X and Y .

- ☞ In Example-1 we computed $E[XY] = \frac{2}{15}$
- ☞ We have $E[X] = \int_{-\infty}^{\infty} xf_X(x) dx$ and $E[Y] = \int_{-\infty}^{\infty} yf_Y(y) dy$
- ☞ First step is to compute the marginal p.d.fs - f_X and f_Y

For $0 < x < 1$,

$$\begin{aligned}f_X(x) &= \int_0^{1-x} 24xy \, dy \\&= 24x \int_0^{1-x} y \, dy \\&= 12x(1-x)^2\end{aligned}$$

For $0 < y < 1$,

$$\begin{aligned}f_Y(y) &= \int_0^{1-y} 24xy \, dx \\&= 24y \int_0^{1-y} x \, dx \\&= 12y(1-y)^2\end{aligned}$$

Next we compute the expectations $E[X]$ and $E[Y]$

$$\begin{aligned}E[X] &= \int_{-\infty}^{\infty} xf_X(x) \, dx \\&= 12 \int_0^1 x^2(1-x)^2 \, dx \\&= 12 \left[\frac{x^3}{3} + \frac{x^5}{5} - \frac{x^4}{2} \right]_{x=0}^{x=1} \\&= \frac{2}{5}\end{aligned}$$

$$\begin{aligned}
 E[Y] &= \int_{-\infty}^{\infty} y f_Y(y) dy \\
 &= 12 \int_0^1 y^2 (1-y)^2 dy \\
 &= 12 \left[\frac{y^3}{3} + \frac{y^5}{5} - \frac{y^4}{2} \right]_{y=0}^{y=1} \\
 &= \frac{2}{5}
 \end{aligned}$$

Finally,

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{2}{15} - \frac{4}{25} = -\frac{2}{75}$$



Summary

☞ For any two random variables X and Y , given their joint p.m.f p (p.d.f f if X and Y are continuous), to compute $\text{Cov}(X, Y)$:

- ▶ **Step-1:** Compute the marginal p.m.fs (or p.d.fs) p_X and p_Y (f_X and f_Y)
- ▶ **Step-2:** Compute the expectations $E[X]$ and $E[Y]$
- ▶ **Step-3:** Compute the joint expectation $E[XY]$
- ▶ **Step-4:** $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$

Properties:

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2. $\text{Cov}(X, X) = \text{Var}(X)$
3. $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$ for any constant a
4. $\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$

Fact: $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i < j} \text{Cov}(X_i, X_j)$

Proof. $\text{Var}\left(\sum_{i=1}^n X_i\right) = \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right)$ (**Property-2**)
 $= \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, X_j)$ (**Property-4**)

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^n X_i\right) &= \sum_{i=j} \sum \text{Cov}(X_i, X_j) + \sum_{i \neq j} \sum \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \sum \text{Cov}(X_i, X_j) \quad (\textbf{Property-2})\end{aligned}$$



★ If X_1, X_2, \dots, X_n are pairwise independent, then

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$$

Probability and Statistics

Lecture-22

 **Recall:** The **covariance** between any two random variables X and Y , denoted by $\text{Cov}(X, Y)$, is defined by

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

 The **(Pearson's) correlation coefficient** of two random variables X and Y , denoted by $\rho(X, Y)$, is defined, by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

Fact: $-1 \leq \rho(X, Y) \leq 1$

Example-1: In a study of response times (Y) versus the number of bars of signal strength (X) of a particular network, the joint p.m.f is given below.

$y = \text{Response Time}$ (nearest second)	$x = \text{Number of Bars of Signal Strength}$		
	1	2	3
4	0.15	0.1	0.05
3	0.02	0.1	0.05
2	0.02	0.03	0.2
1	0.01	0.02	0.25

Compute $\rho(X, Y)$.

☞ In the last class, we computed the marginal p.m.fs:

$$p_X(X = 1) = 0.2, p_X(X = 2) = 0.25, p_X(X = 3) = 0.55$$

$$p_Y(Y = 1) = 0.28, p_Y(Y = 2) = 0.25, p_Y(Y = 3) = 0.17, p_Y(Y = 4) = 0.3$$

We also computed

$$E[X] = 2.35, E[Y] = 2.49 \text{ and } \text{Cov}(X, Y) = -0.5815$$

We have $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$

Now,

$$E[X^2] = 1^2(0.2) + 2^2(0.25) + 3^2(0.55) = 6.15$$

$$E[Y^2] = 1^2(0.28) + 2^2(0.25) + 3^2(0.17) + 4^2(0.3) = 0.28 + 1 + 1.53 + 4.8 = 7.61$$

Hence

$$\text{Var}(X) = E[X^2] - (E[X])^2 = 6.15 - 5.5225 = 0.6275$$

$$\text{Var}(Y) = E[Y^2] - (E[Y])^2 = 7.61 - 6.2001 = 1.4099$$

Thus, $\rho(X, Y) = \frac{-0.5815}{\sqrt{0.6275}\sqrt{1.4099}} = -0.62$



Example: The random variables X and Y have a joint density function given by

$$f(x, y) = \begin{cases} 24xy, & 0 < x < 1, 0 < y < 1, 0 < x + y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Compute the correlation coefficient between X and Y .

We computed the following in the last class:

$$f_X(x) = \begin{cases} 12x(1-x)^2, & 0 < x < 1, \\ 0, & \text{else.} \end{cases}$$

$$E[X] = \frac{2}{5}$$

$$f_Y(y) = \begin{cases} 12y(1-y)^2, & 0 < y < 1, \\ 0, & \text{else.} \end{cases}$$

$$E[Y] = \frac{2}{5}$$

$$\text{Cov}(X, Y) = \frac{-2}{75}$$

We have $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$

Now,

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ &= 12 \int_0^1 x^3 (1-x)^2 dx \\ &= 12 \left[\frac{x^4}{4} + \frac{x^6}{6} - \frac{2x^5}{5} \right]_{x=0}^{x=1} \\ &= \frac{1}{5} \end{aligned}$$

$$\begin{aligned} E[Y^2] &= \int_{-\infty}^{\infty} y^2 f_Y(y) dy \\ &= 12 \int_0^1 y^3 (1-y)^2 dy \\ &= 12 \left[\frac{y^4}{4} + \frac{y^6}{6} - \frac{2y^5}{5} \right]_{y=0}^{y=1} \\ &= \frac{1}{5} \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= \frac{1}{25} \end{aligned}$$

$$\begin{aligned} \text{Var}(Y) &= E[Y^2] - (E[Y])^2 \\ &= \frac{1}{25} \end{aligned}$$

$$\implies \rho(X, Y) = \frac{-2/75}{(1/5)(1/5)} = -\frac{2}{3}$$



Summary

- ▶ **Step-1:** Calculate the marginal density or mass functions
- ▶ **Step-2:** Calculate the expectations $E[X], E[X^2], E[Y], E[Y^2]$ using marginal functions
- ▶ **Step-3:** Calculate the joint expectation $E[XY]$
- ▶ **Step-4:** Calculate $\text{Cov}(X, Y)$, $\text{Var}(X)$ and $\text{Var}(Y)$

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

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

- ▶ **Step-5:** Finally,

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

Check: $-1 \leq \rho(X, Y) \leq 1$

- ▶ If $\rho(X, Y) = 0$, then X and Y are said to be **uncorrelated**
- ▶ **Recall:** If X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

$$\implies \text{Cov}(X, Y) = 0$$

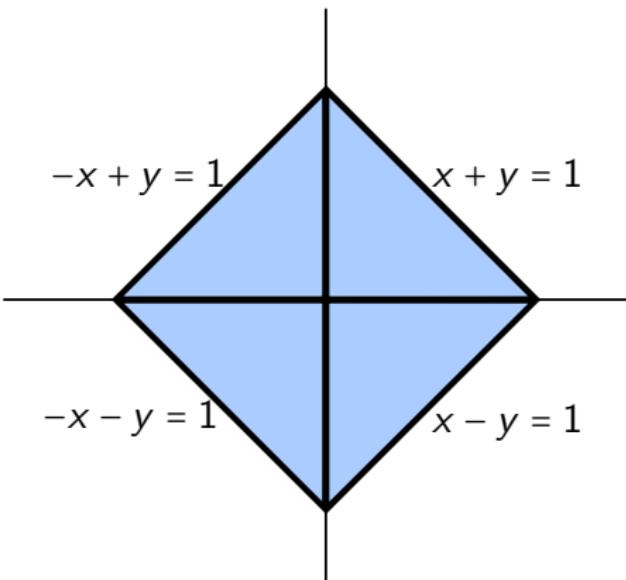
$$\implies \rho(X, Y) = 0$$

- ▶ Thus, independent variables are uncorrelated
- ▶ Converse is **not true!**
- ▶ That is even if $\rho(X, Y) = 0$, X and Y may not be independent!

Example: Let X and Y be continuous random variable with joint p.d.f

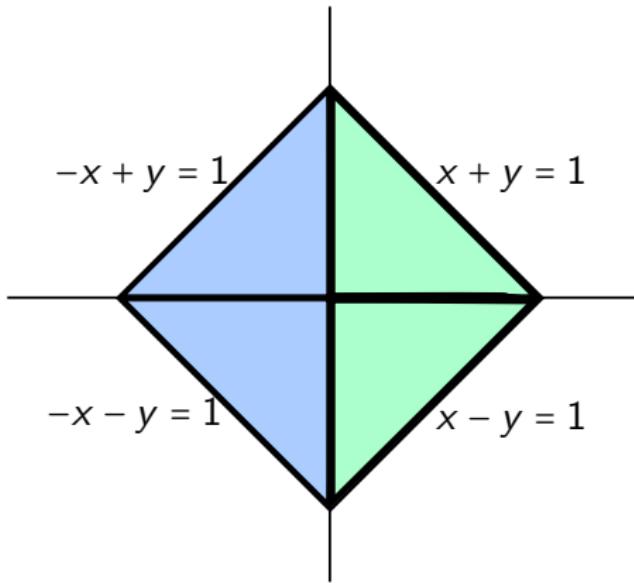
$$f(x, y) = \begin{cases} \frac{1}{2}, & \text{if } |x| + |y| < 1, \\ 0, & \text{else.} \end{cases}$$

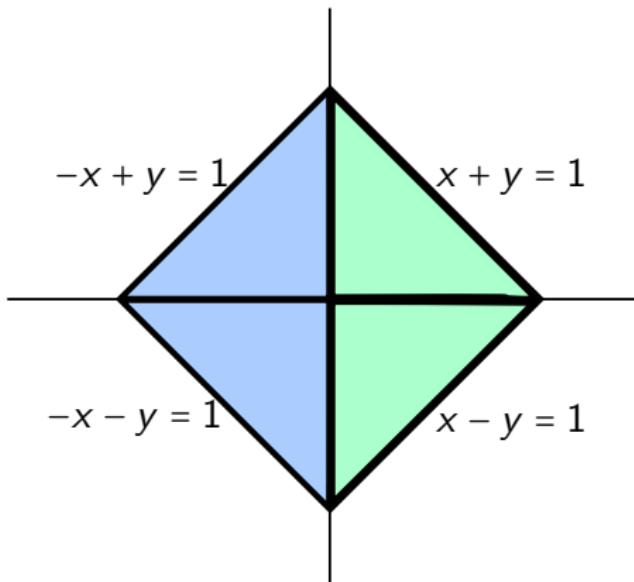
- Let us compute $\text{Cov}(X, Y)$



$$A = \{(x, y) : |x| + |y| < 1\}$$

- ▶ We first need to find the marginal densities f_X and f_Y
- ▶ Now, $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$
- ▶ We need to find limits of integration for y in the given region A
- ▶ For this, we will divide the region into two parts



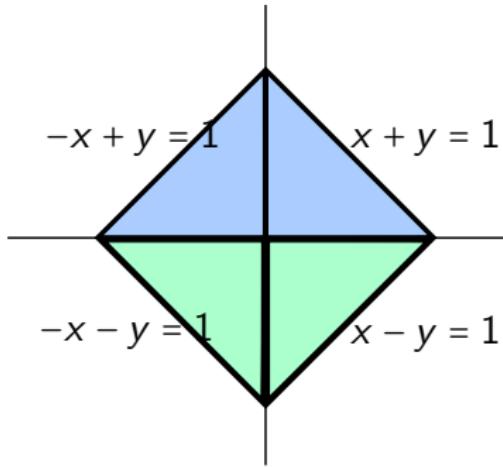


- The blue region is bounded below by $y = -1 - x$ and above by $y = 1 + x$
- The green region is bounded below by $y = x - 1$ and above by $y = 1 - x$

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-1-x}^{1+x} \frac{1}{2} dy + \int_{x-1}^{1-x} \frac{1}{2} dy = 2 \quad (\text{check!})$$

Thus,

$$f_X(x) = \begin{cases} 2, & \text{if } |x| < 1, \\ 0, & \text{else.} \end{cases}$$



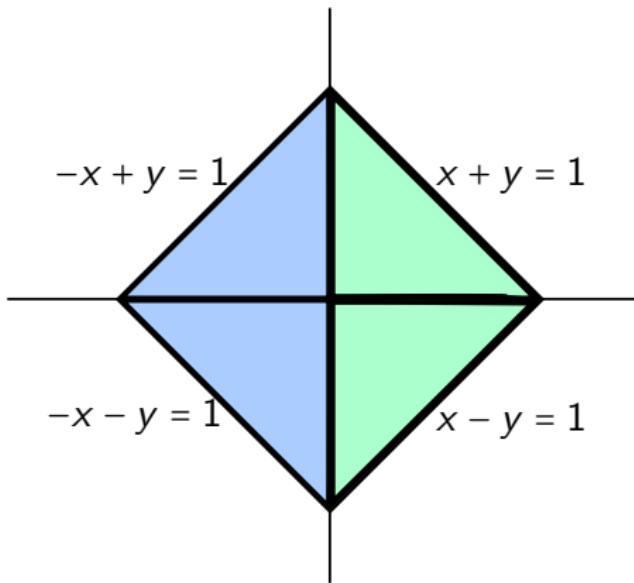
- ▶ The blue region is bounded in the left by $x = y - 1$ and right by $x = 1 - y$
- ▶ The green region is bounded in the left by $x = -1 - y$ and above by $x = 1 + y$

$$\implies f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{y-1}^{1-y} \frac{1}{2} dx + \int_{-1-y}^{1+y} \frac{1}{2} dx = 2 \quad (\text{check!})$$

Thus,

$$f_Y(y) = \begin{cases} 2, & \text{if } |y| < 1, \\ 0, & \text{else.} \end{cases}$$

- ▶ $E[X] = \int_{-\infty}^{\infty} f_X(x) dx = \int_{-1}^1 2 dx = 0$
- ▶ Similarly, $E[Y] = 0$
- ▶ Now, $E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy$
- ▶ We again look at the picture



► We get

$$\begin{aligned}
 E[XY] &= \int_{-1}^0 \int_{-x-1}^{1+x} \frac{1}{2} dy dx + \int_0^1 \int_{x-1}^{1-x} \frac{1}{2} dy dx \\
 &= \left(-1 + \frac{1}{2}\right) + \left(1 - \frac{1}{2}\right) \\
 &= 0
 \end{aligned}$$

- ▶ Thus $\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0$
- ▶ Hence $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = 0$
- ▶ Clearly $f(x, y) \neq f_X(x)f_Y(y)$ for any x, y such that $|x| + |y| < 1$
 $\implies X$ and Y are uncorrelated but **not** independent!



- ☞ A positive value of $\rho(X, Y)$ indicates that
 - ▶ X and Y **positively correlated**, and
 - ▶ Y tends to **increase** when X increases.
- ☞ Whereas, a negative value of $\rho(X, Y)$ indicates
 - ▶ X and Y **negatively correlated**, and
 - ▶ Y tends to **decrease** when X increases.
- ☞ The correlation coefficient is a measure of the degree of **linear relationship** between X and Y
- ☞ Even when $\rho(X, Y) = 0$, X and Y may have a non-linear relationship
- ☞ For example, if $X \sim \text{Uniform}(-1, 1)$ and $Y = X^2$, then $\rho(X, Y) = 0$ despite the clear (non-linear) relation $Y = X^2$!

Probability and Statistics

Lecture-23

- ▶ Till now we learnt to analyze problems and systems involving uncertainty, to find probabilities, expectations, and other characteristics for a variety of situations, and to produce forecasts that may lead to important decisions
- ▶ What was given to us in all these problems?
- ▶ We were able to compute probabilities by considering outcomes of the experiment or **we were given the distribution and its parameters**
- ▶ Often the distribution may not be given, and we learned how to fit the suitable model, say, Binomial, Exponential, or Poisson, given the type of variables we deal with
- ▶ In any case, parameters of the fitted distribution had to be reported to us explicitly, or they had to follow directly from the problem

- ▶ This, however, is rarely the case in practice
- ▶ Then, how can one apply the knowledge of techniques we learnt so far and compute probabilities?
- ▶ The answer is simple: **we need to collect data**
- ▶ We collect data to get an idea of the distribution, estimate the distribution parameters and hence compute probabilities
- ▶ If the “population” is too large, we will collect a “good” sample and analyse it
- ▶ A properly collected sample of data can provide rather sufficient information about distribution and the parameters of the observed system

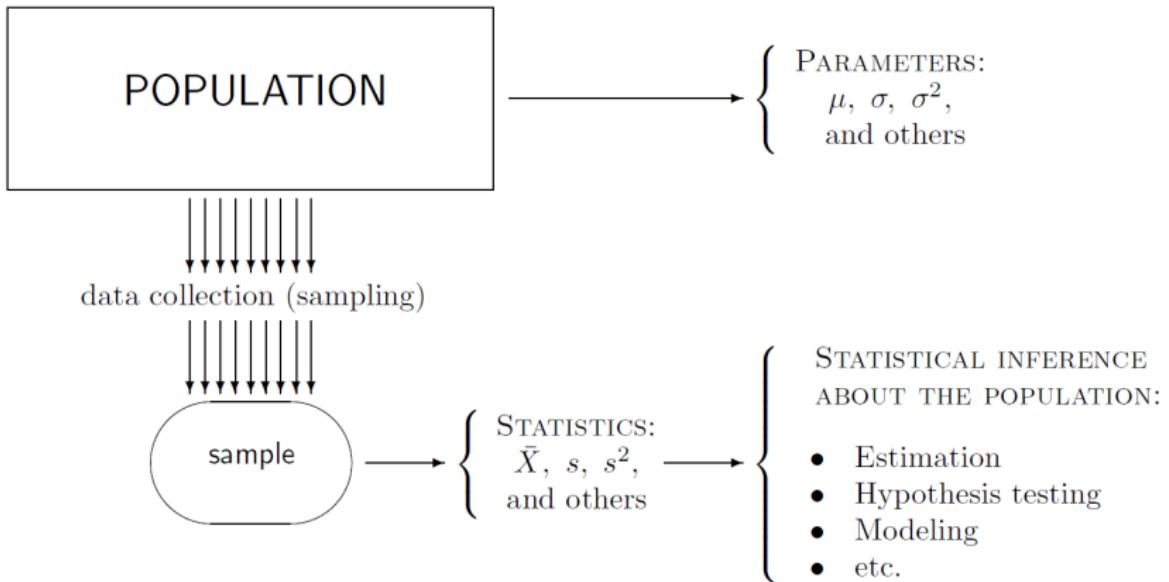
 In the remaining part of this course, we learn how to use this sample

- ▶ to visualize data, understand the patterns, and make quick statements about the system's behaviour;
- ▶ to characterize this behavior in simple terms and quantities;
- ▶ to estimate the distribution parameters;
- ▶ to assess reliability of our estimates;
- ▶ to test statements about parameters and the entire system;
- ▶ to understand relations among variables;
- ▶ to fit suitable models and use them to make forecasts.

- ☞ Our requirement from the data boils down to two things:
 - ▶ getting an idea of the distribution the population may follow
 - ▶ estimating the parameters of the distribution
- ☞ A **population** consists of all units of interest. Any numerical characteristic of a population is a **parameter**.
- ☞ Population may be people, computers, vehicles, etc. and parameters may be average height of people, maximum computing speed of computers, etc.
- ☞ A **sample** consists of observed units collected from the population. It is used to make statements about the population. Any function of a sample is called **statistic**.
- ☞ Each statistic is also a random variable because it is computed from random data. It has a so-called **sampling distribution**

- ▶ In real problems, we would like to make statements about the population
- ▶ To compute probabilities, expectations, and make optimal decisions under uncertainty, we need to know the population parameters
- ▶ However, the only way to know these parameters is to measure the entire population, i.e., to conduct a **census**
- ▶ Instead of a census, we may collect data in a form of a random sample from a population
- ▶ This is our data. We can measure them, perform calculations, and estimate the unknown parameters of the population up to a certain **measurable degree of accuracy**

NOTATION: $\left\{ \begin{array}{l} \theta \text{ -- population parameter} \\ \hat{\theta} \text{ -- its estimator, obtained from a sample} \end{array} \right.$



 Because of sampling there may be some discrepancy between the estimated values and actual values. These errors can be classified into two types:

- ▶ **Sampling errors** are caused by the mere fact that only a sample, a portion of a population, is observed. For most of reasonable statistical procedures, sampling errors decrease (and converge to zero) as the sample size increases.
- ▶ **Non-sampling errors** are caused by inappropriate sampling schemes.
 - ★ No statistical technique can reduce this kind of errors. Hence proper sampling practices must be followed.

Some examples of wrong sampling practices

Sampling from a wrong population

- ▶ To evaluate the work of a Windows help desk, a survey of social science students of some university is conducted
- ▶ This sample poorly represents the whole population of all Windows users
- ▶ For example, computer science students and especially computer professionals may have a totally different opinion about the Windows help desk

Some examples of wrong sampling practices

Dependent observations

- ▶ Comparing two brands of notebooks, a senior manager asks all employees of her group to state which notebook they like and generalizes the obtained responses to conclude which notebook is better
- ▶ Again, these employees are not randomly selected from the population of all users of these notebooks.
- ▶ Also, their opinions are likely to be dependent. Working together, these people often communicate, and their points of view affect each other
- ▶ Dependent observations do not necessarily cause non-sampling errors, if they are handled properly. The fact is, in such cases, we cannot assume independence

Some examples of wrong sampling practices

Not equally likely

- ▶ A survey among passengers of some airline is conducted in the following way
- ▶ A sample of random flights is selected from a list, and ten passengers on each of these flights are also randomly chosen
- ▶ Each sampled passenger is asked to fill a questionnaire. Is this a representative sample?
- ▶ Suppose Mr. X flies only once a year whereas Ms. Y has business trips twice a month
- ▶ Obviously, Ms. Y has a much higher chance to be sampled than Mr. X
- ▶ Unequal probabilities have to be taken into account, otherwise a non-sampling error will inevitably occur

☞ An effective way to avoid non-sampling errors is to use **simple random sampling**

☞ **Simple random sampling** is a sampling design where units are collected from the entire population independently of each other, all being equally likely to be sampled.

- ▶ Observations collected by means of a simple random sampling design are **iid** (independent, identically distributed) random variables
- ▶ Obtaining a good, representative random sample is rather important in Statistics
- ▶ Although we have only a portion of the population in our hands, a rigorous sampling design followed by a suitable statistical inference allows to estimate parameters and make statements with a certain measurable degree of confidence

☞ From now on, we assume that a good random sample

$$\mathcal{S} = (X_1, X_2, \dots, X_n)$$

has been collected

☞ An estimator $\hat{\theta}$ is **unbiased** for a parameter θ if its expectation equals the parameter,

$$E[\hat{\theta}] = \theta$$

for all possible values of θ .

- **Bias** of $\hat{\theta}$ is defined as $\text{Bias}(\hat{\theta}) = E[\hat{\theta} - \theta]$.

☞ An estimator $\hat{\theta}$ is **consistent** for a parameter θ if the probability of its sampling error of any magnitude converges to 0 as the sample size increases to infinity. Stating it rigorously,

$$P\{|\hat{\theta} - \theta| > \varepsilon\} \rightarrow 0 \text{ as } n \rightarrow \infty \text{ for any } \varepsilon > 0$$

☞ An estimator $\hat{\theta}$ for a parameter θ is said to be **asymptotically normal** if

$$\frac{\hat{\theta} - E[\hat{\theta}]}{SD(\hat{\theta})}$$

follows the standard normal distribution for considerably large sample size

Example: To evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time for $n = 30$ randomly chosen jobs (in seconds)

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

☞ What information do we get from this collection of numbers?

- ☞ **Variable of interest, say, X ?** - the time CPU takes to perform a random job
- ☞ X is a random variable, and its value does not have to be among the observed thirty
- ☞ We will use the collected data to describe the distribution of X
- ☞ To start with, we calculate the following **descriptive statistics** which measure the location, spread, variability, and other characteristics
 - ▶ **mean**, measuring the average value of a sample;
 - ▶ **median**, measuring the central value;
 - ▶ **quantiles and quartiles**, showing where certain portions of a sample are located;
 - ▶ **variance, standard deviation, and interquartile range**, measuring variability and spread of data.

- ☞ Each statistic is a random variable because it is computed from random data. It has a so-called **sampling distribution**
- ☞ Each statistic estimates the corresponding population parameter and adds certain information about the distribution of X , the variable of interest

☞ **Population mean** - $\mu = E [X]$

☞ **Sample mean** -

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

- ☞ Sample mean \bar{X} is an unbiased, consistent and asymptotically normal estimator for the population mean μ

Example: To evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time for $n = 30$ randomly chosen jobs (in seconds),

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

$$\text{Sample mean, } \bar{X} = \frac{\text{sum of observations}}{\text{number of observations}} = 48.23$$



- ▶ Though sample mean is a good descriptive measure, one disadvantage of a sample mean is its **sensitivity to extreme observations**
- ▶ For instance, consider the earlier data

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

- ▶ We found that the sample mean is 48.23
- ▶ Now replace the first value of 70 with some large value, say, 1800
- ▶ Check that the sample mean now is 105.9
- ▶ This single extremely large observation shifts the sample mean from 48.23 sec to 105.9 sec
- ▶ Can we call such an estimator “reliable”?

Probability and Statistics

Lecture-24

- ☞ We saw that extreme values affect mean adversely!
- ☞ Another simple measure of location is a **sample median**, which estimates the **population median**. It is much less sensitive than the sample mean

- ▶ **Median** means a “central” value
- ▶ **Sample median** \widehat{M} is a number that is exceeded by at most a half of observations and is preceded by at most a half of observations
- ▶ **Population median** M is a number that is exceeded with probability no greater than 0.5 and is preceded with probability no greater than 0.5. That is, M is such that

$$P\{X > M\} \leq 0.5 \text{ and } P\{X < M\} \leq 0.5$$

☞ Computing sample median

- ▶ If n is odd, then $(\frac{n+1}{2})$ -th smallest observation is a median
- ▶ If n is even, then any number between the $(\frac{n}{2})$ -th smallest and the $(\frac{n+2}{2})$ -th smallest observations is a median

Example: Let us compute the median of $n = 30$ CPU times from the data we saw earlier

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Step-1: Order the data

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

Step-2: Since $n = 30$ is even, find the $(\frac{n}{2})$ -th smallest and the $(\frac{n+2}{2})$ -th smallest observations

They are 42 and 43 respectively. Any number between them is a sample median. Typically we will take it as average of these two, which is 42.5



Exercise: Find the median for the following data

92, 90, 92, 74, 69, 80, 94, 98, 65, 96, 84, 69, 86, 91, 88

74, 97, 85, 88, 68, 77, 94, 88, 65, 76, 75, 60

69, 97, 92, 85, 70, 80, 93, 91, 68, 82, 78, 89

Step-1: Order the data

Step-2: Since the sample size $n = 39$ is odd, median will be the $(\frac{n+1}{2}) = 20^{\text{th}}$ smallest observation

Answer: Sample median $\hat{M} = 85$



☞ We learnt to compute the sample median

☞ How do we calculate population median?

- **Computation of population median for continuous distributions**

For continuous distributions, we compute population median by solving the equations:

$$\left. \begin{array}{l} P\{X > M\} = 1 - F(M) \leq 0.5 \\ P\{X < M\} = F(M) \leq 0.5 \end{array} \right\} \implies F(M) = 0.5$$

Example: Calculate median for a population which follows uniform distribution on (a, b)

We know that, the distribution function is given by

$$F(x) = \begin{cases} \frac{x-a}{b-a}, & \text{if } x \in (a, b) \\ 0, & \text{else} \end{cases}$$

We need to solve $F(M) = 0.5$ for M . **Check:** $M = \frac{a+b}{2}$



Example: Compute the population median for exponential population with parameter λ

We know that, the distribution function is given by

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x > 0 \\ 0, & \text{else} \end{cases}$$

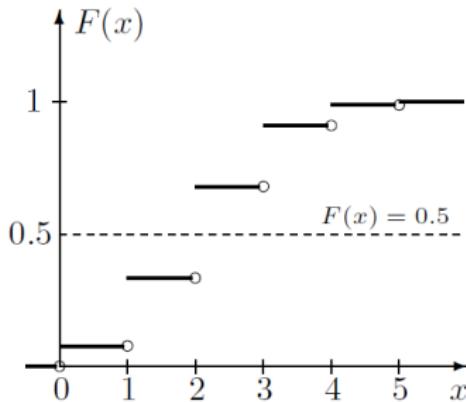
We need to solve $F(M) = 0.5$ for M

$$\text{That is, } 1 - e^{-\lambda M} = 0.5 \implies M = \frac{\ln(2)}{\lambda} = \frac{0.6931}{\lambda}$$

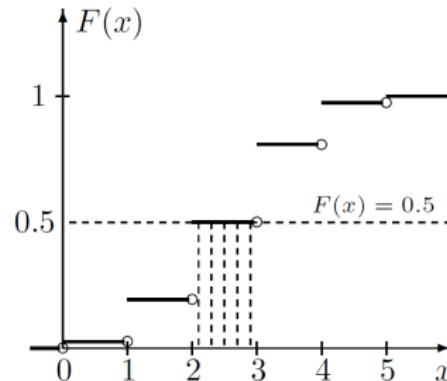
Exercise: Calculate the population median for normal population with mean μ and variance σ^2

Question: In any continuous distribution, does such an M always exist? That is, does the equation $F(M) = 0.5$ always have a solution?

- ☞ In the continuous random variable case, the distribution function $F(x)$ is a continuous function taking values in $[0,1]$ and hence, by **intermediate value theorem**, there is always a value M such that $F(M) = 0.5$
- ☞ But, in the discrete case, $F(M) = 0.5$ may have no solutions at all!

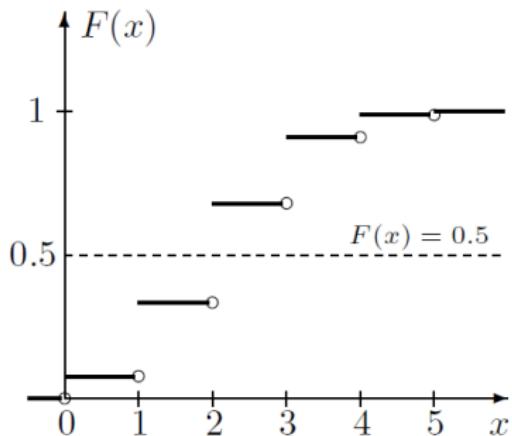


☞ It may also be the case that $F(M) = 0.5$ has infinitely many solutions



Computation of population median - discrete case

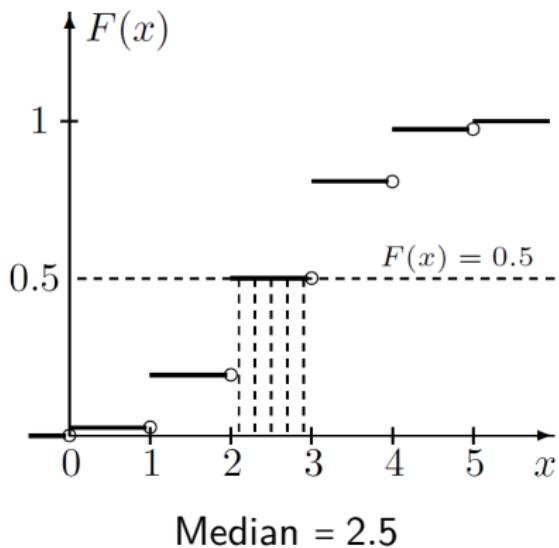
If the equation $F(M) = 0.5$ has no solution, then the smallest x with $F(x) \geq 0.5$ is the median



Median = 2

Computation of population median - discrete case

If the equation $F(M) = 0.5$ has an interval of solutions, then any number in this interval, excluding the ends, is a median. Notice that the median in this case is not unique. Often the middle of this interval is reported as the median.



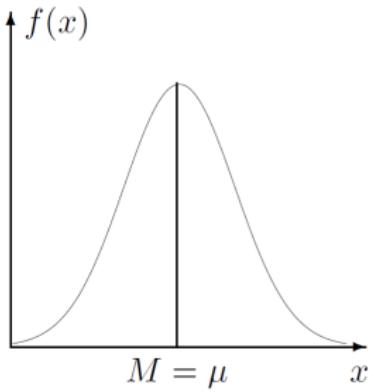
To sum up, to find population median in case of discrete distributions, we make use of the distribution function

- ▶ If the equation $F(M) = 0.5$ has no solution, then the smallest x with $F(x) \geq 0.5$ is the median
- ▶ If the equation $F(M) = 0.5$ has an interval of solutions, the middle of this interval is reported as the median

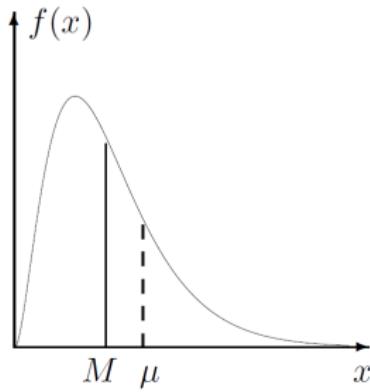
Suppose the random variable X has mean μ and median M , then X is said to have

- ▶ **Symmetric distribution** if $M = \mu$
- ▶ **Right-skewed distribution** if $M < \mu$
- ▶ **Left-skewed distribution** if $M > \mu$

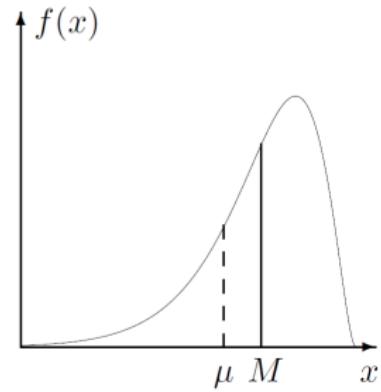
(a) symmetric



(b) right-skewed



(c) left-skewed



- ▶ Median is a number M for which $P\{X < M\} \leq 0.5$ and $P\{X > M\} \leq 0.5 = \frac{50}{100}$
- ▶ On similar grounds, we may extend this definition for any $0 < \gamma < 100$ as follows

☞ A **γ -percentile** of a population is such a number x that solves equations

$$P\{X < x\} \leq \frac{\gamma}{100} \text{ and } P\{X > x\} \leq 1 - \frac{\gamma}{100}$$

☞ A **sample γ -percentile** is any number that exceeds at most $\gamma\%$ of the sample, and is exceeded by at most $(100 - \gamma)\%$ of the sample.

☞ Observe that for $\gamma = 50$, the 50-percentile is nothing but the median!

- Computing sample percentiles

Let us go back to our data set of CPU times

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Say, we wish to calculate the 42-percentile

Step-1: Order the data

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

Computing sample percentiles

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

Step-2: Sample size $n = 30$ and $p = \frac{\gamma}{100} = 0.42$

$$np = 12.6 \text{ and } n(1 - p) = 17.4$$

Thus, the 42-percentile should have no more than 12.6 observations below it and no more than 17.4 observations above it

Such an observation is the 13th observation, which is 37

Thus, the 42-percentile for the given sample is 37



Computing population percentiles

- Recall that for calculating population median we solved the equation $F(M) = 0.5$ for M
- Along the similar lines, to calculate the population γ -percentile, we solve $F(x) = \frac{\gamma}{100}$ for x

NOTATION: $\left\{ \begin{array}{l} \pi_\gamma \text{ -- population } \gamma\text{-percentile} \\ \hat{\pi}_\gamma \text{ -- sample } \gamma\text{-percentile, estimator of } \pi_\gamma \\ M \text{ -- population median} \\ \bar{M} \text{ -- sample median, estimator of } M \end{array} \right.$

 $M = \pi_{50}$ and $\bar{M} = \hat{\pi}_{50}$

- ▶ For any data, there will be 99 percentiles. Of all these, three are considered important while computing the variability of the data

- ▶ **First quartile** - 25-percentile
- ▶ **Second quartile** - 50-percentile
- ▶ **Third quartile** - 75-percentile

- ▶ Quartiles split a population or a sample into four equal parts
- ▶ A median is at the same time a 50th percentile and 2nd quartile.

NOTATION: $\left\{ \begin{array}{l} Q_1, Q_2, Q_3 \text{ -- population quartiles} \\ \hat{Q}_1, \hat{Q}_2, \hat{Q}_3 \text{ -- sample quartiles, estimators of } Q_1, Q_2, \text{ and } Q_3 \end{array} \right.$

☞ By definition, $Q_1 = \pi_{25}$, $Q_2 = \pi_{50} = M$, and $Q_3 = \pi_{75}$

Probability and Statistics

Lecture-25

Recall

- ▶ **First quartile** - 25-percentile
- ▶ **Second quartile** - 50-percentile
- ▶ **Third quartile** - 75-percentile

- ▶ Quartiles split a population or a sample into four equal parts
- ▶ A median is at the same time a 50th percentile and 2nd quartile

NOTATION: $\left\{ \begin{array}{l} Q_1, Q_2, Q_3 \text{ -- population quartiles} \\ \hat{Q}_1, \hat{Q}_2, \hat{Q}_3 \text{ -- sample quartiles, estimators of } Q_1, Q_2, \text{ and } Q_3 \end{array} \right.$

☞ By definition, $Q_1 = \pi_{25}$, $Q_2 = \pi_{50} = M$, and $Q_3 = \pi_{75}$

How do we compute the sample quartiles?

Step-1: Order the data and note down the sample size n

Step-2: For \hat{Q}_1 , take $p = \frac{25}{100}$ and calculate np and $n(1 - p)$

\hat{Q}_1 will be that observation which has no more than np observations below it and $n(1 - p)$ observations above it

Step-3: \hat{Q}_2 is nothing but median which is the central value

Step-4: For \hat{Q}_3 , take $p = \frac{75}{100}$ and calculate np and $n(1 - p)$

\hat{Q}_3 will be that observation which has no more than np observations below it and $n(1 - p)$ observations above it

Example: A network provider investigates the load of its network. The number of concurrent users is recorded at fifty locations (thousands of people),

17.2	22.1	18.5	17.2	18.6	14.8	21.7	15.8	16.3	22.8
24.1	13.3	16.2	17.5	19.0	23.9	14.8	22.2	21.7	20.7
13.5	15.8	13.1	16.1	21.9	23.9	19.3	12.0	19.9	19.4
15.4	16.7	19.5	16.2	16.9	17.1	20.2	13.4	19.8	17.7
19.7	18.7	17.6	15.9	15.2	17.1	15.0	18.8	21.6	11.9

Compute the sample mean, sample median and the three sample quartiles.

☞ Sample mean = $\frac{\text{sum of observations}}{\text{number of observations}} = \frac{897.7}{50} = 17.954$

☞ To calculate median and quartiles, we need to order the data

☞ **Ordered data:**

11.9	12	13.1	13.3	13.4	13.5	14.8	14.8	15	15.2
15.4	15.8	15.8	15.9	16.1	16.2	16.2	16.3	16.7	16.9
17.1	17.1	17.2	17.2	17.5	17.6	17.7	18.5	18.6	18.7
18.8	19	19.3	19.4	19.5	19.7	19.8	19.9	20.2	20.7
21.6	21.7	21.7	21.9	22.1	22.2	22.8	23.9	23.9	24.1

☞ Since the number of observations $n = 50$ is even, we have two central values, the 24-th and 25-th observations

☞ They are 17.5 and 17.6. Thus, the sample median \bar{M} is the average of these two, which is 17.55

First quartile \hat{Q}_1

$\hat{Q}_1 = \hat{\pi}_{25}$, the 25-percentile

$$n = 50 \text{ and } p = \frac{25}{100} = 0.25$$

$$np = 12.5 \text{ and } n(1 - p) = 37.5$$

Thus, \hat{Q}_1 should have no more than 12.5 observations below it and 37.5 observations above it

Thus, \hat{Q}_1 is the 13-th observation. That is, $\hat{Q}_1 = 15.8$

Second quartile \hat{Q}_2

Since $\hat{Q}_2 = \bar{M}$, we have $\hat{Q}_2 = 17.55$

Third quartile \hat{Q}_3

$\hat{Q}_3 = \hat{\pi}_{75}$, the 75-percentile

$$n = 50 \text{ and } p = \frac{\gamma}{100} = 0.75$$

$$np = 37.5 \text{ and } n(1 - p) = 12.5$$

Thus, \hat{Q}_3 should have no more than 37.5 observations below it and 12.5 observations above it

Thus, \hat{Q}_3 is the 38-th observation. That is, $\hat{Q}_3 = 19.9$



- ▶ The descriptive measures defined so far (mean median, quartiles etc.) showed us where the average value and certain percentages of a population are located
- ▶ We now wish to measure *variability* of our variable, how unstable the variable can be, and how much the actual value can differ from its expectation
- ▶ With these additional measures we will be able to assess reliability of our estimates and accuracy of our forecasts

☞ For a sample (X_1, X_2, \dots, X_n) , a **sample variance** is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

☞ **Sample standard deviation** is a square root of a sample variance,

$$s = \sqrt{s^2}$$

☞ Both sample variance and sample standard deviation measure variability and they estimate the population variance $\sigma^2 = \text{Var}(X)$ and population standard deviation $\sigma = SD(X)$

Observe that

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\&= \sum_{i=1}^n X_i^2 - 2\left(\sum_{i=1}^n X_i\right)\bar{X} + \sum_{i=1}^n \bar{X}^2 \\&= \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\end{aligned}$$

$$\implies s^2 = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1}$$

- ☞ The coefficient $\frac{1}{n-1}$ ensures that s^2 is an unbiased estimator for the population variance σ^2
- ☞ In fact, it can be shown that under rather mild assumptions, sample variance and sample standard deviation are consistent and asymptotically normal

Let us calculate the sample variance for the CPU times

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

We calculated the sample mean earlier

$$\text{Sample mean, } \bar{X} = \frac{\text{sum of observations}}{\text{number of observations}} = 48.23$$

$$\text{Now, } s^2 = \frac{(70-48.23)^2 + (36-48.23)^2 + \dots + (19-48.23)^2}{30-1} = \frac{20391.37}{29} = 703.1507$$

- ▶ Just like sample mean, the sample variance and standard deviation are sensitive to extreme observations (**outliers**)
- ▶ If an extreme observation (an outlier) erroneously appears in our data set, it can rather significantly affect the values of \bar{X} and s^2
- ▶ To detect and identify outliers, we need measures of variability that are not very sensitive to them
- ▶ One such measure is the interquartile range

An **interquartile range** is defined as the difference between the first and the third quartiles,

$$IQR = Q_3 - Q_1$$

IQR is estimated by the **sample interquartile range**

$$\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1$$

Question: Given a data set, how do we detect outliers?

Answer: Any observation outside the interval

$$(\hat{Q}_1 - 1.5 (\widehat{\text{IQR}}), \hat{Q}_3 + 1.5 (\widehat{\text{IQR}}))$$

is an outlier!

Example: Detect the outliers (if any) in the data of CPU times.
Here we are looking at the ordered data

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

Step-1: Order the data - already done!

Step-2: Calculate the first and third sample quartiles \hat{Q}_1 and \hat{Q}_3

Step-3: Calculate the interquartile range $\widehat{\text{IQR}}$

Step-4: Set up the interval of “acceptance”

$$(\hat{Q}_1 - 1.5 (\widehat{\text{IQR}}), \hat{Q}_3 + 1.5 (\widehat{\text{IQR}}))$$

Step-5: Look for observations which do not lie in this interval

Step-2: Calculate the first and third sample quartiles \hat{Q}_1 and \hat{Q}_3

☞ \hat{Q}_1

$$n = 30, p = \frac{25}{100} = 0.25$$

$$np = 7.5 \text{ and } n(1 - p) = 22.5$$

Hence 8-th observation is the first quartile. That is, $\hat{Q}_1 = 34$

☞ \hat{Q}_3

Here we have $n = 30$ and $p = 0.75$

Check that $\hat{Q}_3 = 59$

Step-3: Calculate the interquartile range $\widehat{\text{IQR}}$

We get $\widehat{\text{IQR}} = 59 - 34 = 25$

Step-4: Set up the interval of “acceptance”

$$(\hat{Q}_1 - 1.5 (\widehat{\text{IQR}}), \hat{Q}_3 + 1.5 (\widehat{\text{IQR}}))$$

The interval here is $(34 - (1.5)25, 59 + (1.5)25) = (-3.5, 96.5)$

Step-5: Look for observations which do not lie in this interval

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

From the data, we see that 139 is the only outlier!



Question: If there are outliers in our data, what do we do with them?

- ▶ We may be tempted to delete them right away as they significantly affect sample mean and standard deviation and therefore spoil our statistical analysis
- ▶ However, deleting them immediately may not be the best idea
- ▶ It is rather important to track the history of outliers and understand the reason they appeared in the data set
- ▶ There may be a pattern that we would want to be aware of
- ▶ It may be a new trend that was not known before
- ▶ Or, it may be an observation from a very special part of the population
- ▶ Sometimes important phenomena are discovered by looking at outliers
- ▶ If it is confirmed that a suspected observation entered the data set by a mere mistake, it can be deleted

Probability and Statistics

Lecture-26

Recall

We learnt to compute the following descriptive statistics

- ▶ Sample mean \bar{X}
- ▶ Sample median \overline{M}
- ▶ Sample percentiles $\hat{\pi}_\gamma$
- ▶ Sample quartiles $\hat{Q}_1, \hat{Q}_2, \hat{Q}_3$
- ▶ Sample variance s^2
- ▶ Sample inter-quartile range \widehat{IQR}

☞ These statistics estimate respective population parameters and help in detecting outliers in the sample (if any)

☞ How do you guess the distribution of the population or draw quick inferences about the population?

**Before you do anything with a data set,
look at it!**

A quick look at a sample may clearly suggest

- ▶ a probability model, i.e., a family of distributions to be used;
- ▶ statistical methods suitable for the given data;
- ▶ presence or absence of outliers;
- ▶ presence or absence of heterogeneity;
- ▶ existence of time trends and other patterns;
- ▶ relation between two or several variables.

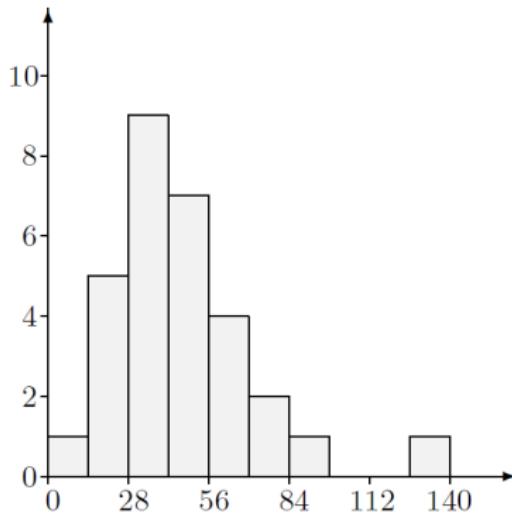
☞ We will be looking at the following ways of **visualizing data**

- ▶ histograms,
- ▶ scatter plots, and
- ▶ time plots.

☞ A **histogram**

- ▶ shows the shape of a pmf or a pdf of data,
- ▶ checks for homogeneity, and
- ▶ suggests possible outliers.

☞ A histogram looks like this:



- ▶ Given a data set, we will visualize it in the form of such vertical bars, whose width is fixed and height is varying
- ▶ The procedure of constructing the histogram involves determining width and heights of bars from the given data

To construct a histogram for a given data,

- ▶ the range of data into equal intervals, referred to as **bins**, and,
- ▶ count how many observations fall into each bin

Example: Let us look at the data of CPU times

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

Ordered data:

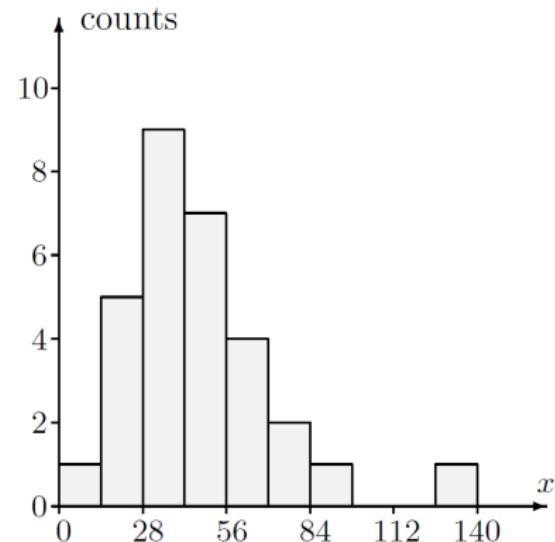
9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

- ☞ Since the whole data is (roughly) between 0 – 140, let us divide the range into 10 equal parts (bins), 0 – 14, 14 – 28, 28 – 42, 42 – 56, 56 – 70, 70 – 84, 84 – 98, 98 – 112, 112 – 126, 126 – 140
- ☞ We then count the number of observations in each of these bins

9	15	19	22	24	25	30	34	35	35
36	36	37	38	42	43	46	48	54	55
56	56	59	62	69	70	82	82	89	139

Bins	Frequency
0 – 14	1
14 – 28	5
28 – 42	9
42 – 56	7
56 – 70	4
70 – 84	2
84 – 98	1
98 – 112	0
112 – 126	0
126 – 140	1

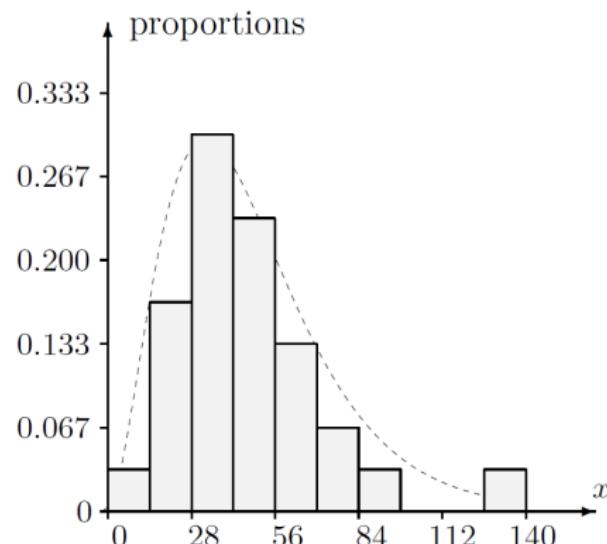
Bins	Frequency
0 – 14	1
14 – 28	5
28 – 42	9
42 – 56	7
56 – 70	4
70 – 84	2
84 – 98	1
98 – 112	0
112 – 126	0
126 – 140	1



Frequency histogram

$$\text{Relative frequency} = \frac{\text{frequency}(f)}{\text{sample size}(n)}$$

Bins	Freq.(f)	f/n
0 – 14	1	0.03
14 – 28	5	0.17
28 – 42	9	0.3
42 – 56	7	0.23
56 – 70	4	0.13
70 – 84	2	0.07
84 – 98	1	0.03
98 – 112	0	0
112 – 126	0	0
112 – 140	1	0.03



Relative frequency histogram

Example: A network provider investigates the load of its network. The number of concurrent users is recorded at fifty locations (thousands of people),

17.2	22.1	18.5	17.2	18.6	14.8	21.7	15.8	16.3	22.8
24.1	13.3	16.2	17.5	19.0	23.9	14.8	22.2	21.7	20.7
13.5	15.8	13.1	16.1	21.9	23.9	19.3	12.0	19.9	19.4
15.4	16.7	19.5	16.2	16.9	17.1	20.2	13.4	19.8	17.7
19.7	18.7	17.6	15.9	15.2	17.1	15.0	18.8	21.6	11.9

 **Ordered data:**

11.9	12	13.1	13.3	13.4	13.5	14.8	14.8	15	15.2
15.4	15.8	15.8	15.9	16.1	16.2	16.2	16.3	16.7	16.9
17.1	17.1	17.2	17.2	17.5	17.6	17.7	18.5	18.6	18.7
18.8	19	19.3	19.4	19.5	19.7	19.8	19.9	20.2	20.7
21.6	21.7	21.7	21.9	22.1	22.2	22.8	23.9	23.9	24.1

11.9	12	13.1	13.3	13.4	13.5	14.8	14.8	15	15.2
15.4	15.8	15.8	15.9	16.1	16.2	16.2	16.3	16.7	16.9
17.1	17.1	17.2	17.2	17.5	17.6	17.7	18.5	18.6	18.7
18.8	19	19.3	19.4	19.5	19.7	19.8	19.9	20.2	20.7
21.6	21.7	21.7	21.9	22.1	22.2	22.8	23.9	23.9	24.1

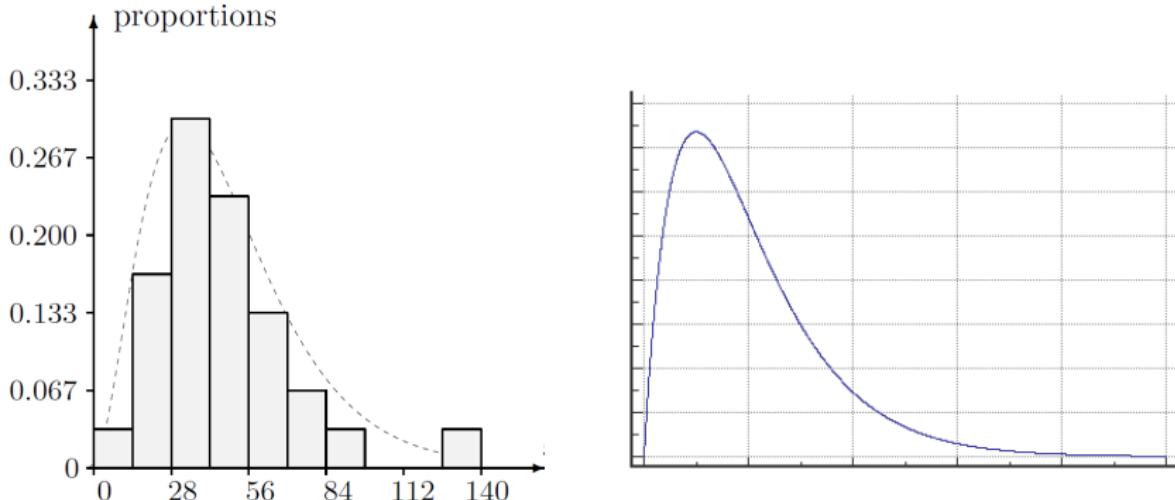
☞ Let us divide the range 10 – 25 into 6 equal parts. That is the bin size is 2.5 units

☞ Setup the frequency and relative frequency tables, and draw both the histograms

Now!

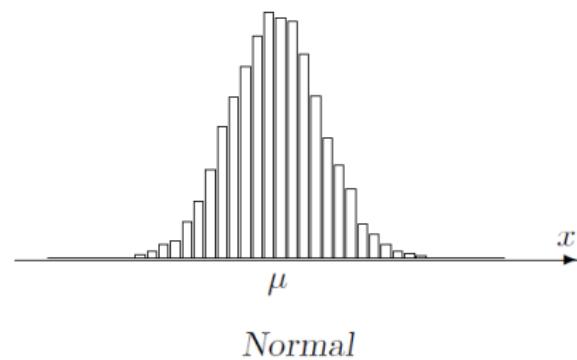
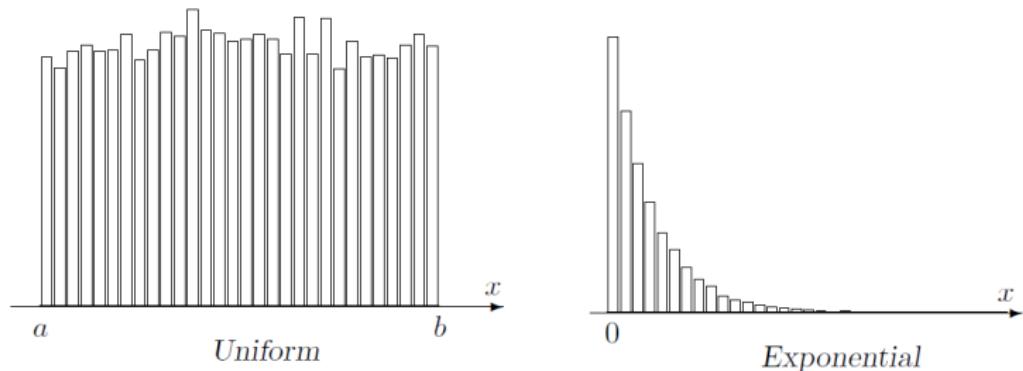
Histograms have a shape similar to the pmf or pdf of data, especially in large samples

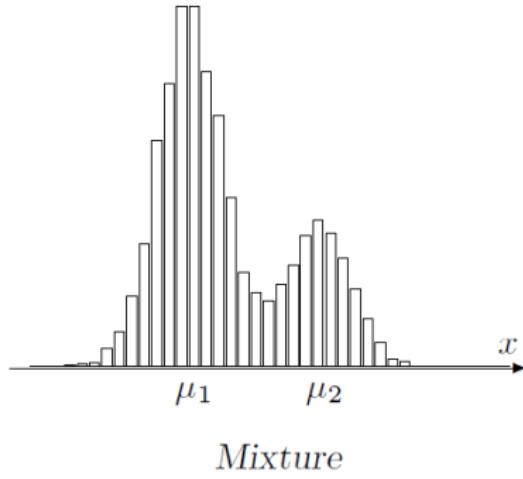
Let us look at the relative frequency histogram once again.



- ☞ The above curve is the density of **Gamma distribution** except for one outlier 139
- ☞ Thus, the gamma distribution seems to be a good fit for the CPU times

How else may histograms look like?





☞ Bell shapes of both humps in above figure suggest that the sample came from a mixture of two Normal distributions (with means around μ_1 and μ_2), with a higher probability of having mean μ_1 , since the left hump is bigger

Probability and Statistics

Lecture-27

Scatter plots

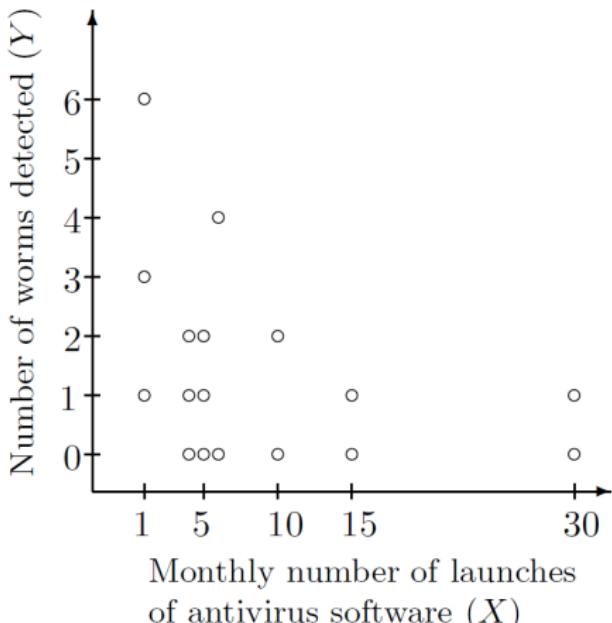
- ▶ Till now we had only one variable of interest and we were plotting their sample values
- ▶ Scatter plots are used to see and understand a relationship between two variables
- ▶ These can be
 - temperature and humidity,
 - experience and salary,
 - age of a network and its speed,
 - number of servers and the expected response time, and so on.
- ▶ To study the relationship, both variables are measured on each sampled item

Example: During a scheduled maintenance of computer facilities, a computer manager records the number of times the antivirus software was launched on each computer during 1 month (variable X) and the number of detected worms (variable Y). The data for 30 computers are in the following table

X	30	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	0	1	1	0
X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1	1

X	30	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1	1

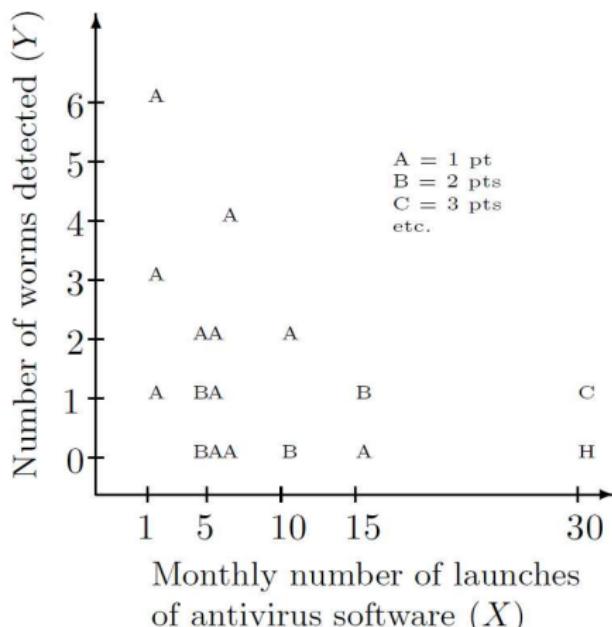


Plotting identical points

- ▶ In the scatter plot we have drawn, some portion of the data is hidden
- ▶ This is because there are some identical observations
- ▶ For example, no worms were detected on 8 computers where the antivirus software is used daily (30 times a month)
- ▶ **8 repeated observations (30,0) is plotted as only one point!**
- ▶ Hence the plot could be misleading
- ▶ When the data contain identical pairs of observations, the points on a scatter plot are often depicted with letters - **A** for 1 point, **B** for two identical points, **C** for three, etc.

X	30	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	0	1	1	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1	1



Time plots

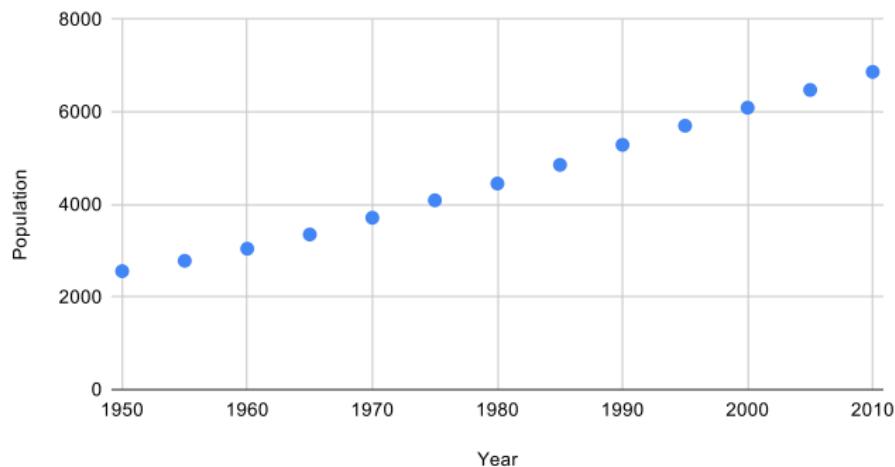
- When we study time trends and development of variables over time, we use **time plots**
- These are scatter plots with x-variable representing time

Example: According to the International Data Base of the U.S. Census Bureau, population of the world grows according to following table.

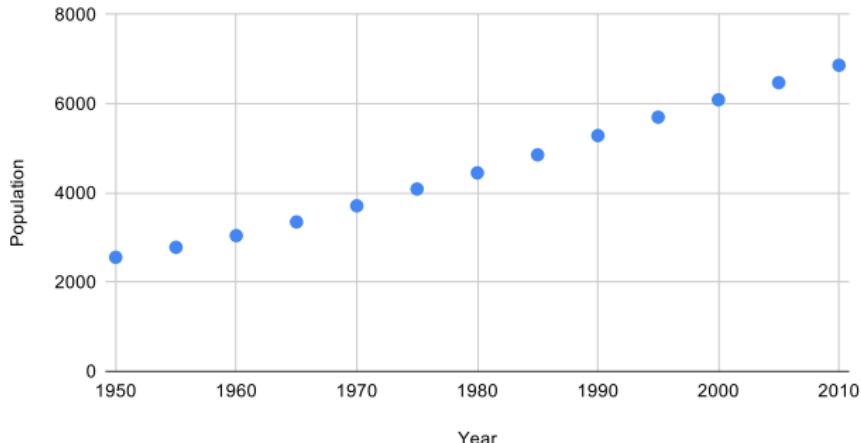
Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287		
1970	3712	1995	5700		

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287		
1970	3712	1995	5700		

Population vs. Year



Population vs. Year



☞ By looking at the above plot, we can see that there is some kind of **linear relationship** between the given two variables

☞ **Question:** How do we measure the extent of such linear relationships in the sample data?

☞ **Answer:** Correlation!

 **Recall**

- ▶ For any two random variables X and Y , the **covariance** between them, denoted by $\text{Cov}(X, Y)$, is defined by

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

- ▶ The **correlation coefficient** of two random variables X and Y , denoted by $\rho(X, Y)$, is defined by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

-  $\text{Cov}(X, Y)$ and $\rho(X, Y)$ are population parameters and to estimate them we need corresponding sample statistics

☞ For sample data consisting of pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$,

- ▶ the **sample covariance** between X and Y is given by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- ▶ the **sample correlation coefficient** between X and Y is given by

$$r = \frac{s_{xy}}{s_x s_y}$$

where $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ and $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ are the sample variances of X and Y respectively.

(Recall that s_x and s_y are sample standard deviations of X and Y !)

☞ By substituting s_x , s_y and s_{xy} in the above formula for r , the formula can be seen to be

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left(\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \left(\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \right)}$$

☞ We shall denote few things as below:

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = (n - 1)s_{xy}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = (n - 1)s_x^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = (n - 1)s_y^2$$

☞ With the above notations, we have

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Example: Let us compute the correlation coefficient for our population data

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287		
1970	3712	1995	5700		

☞ We are given pairs of data $(x_1, y_1), (x_2, y_2), \dots, (x_{13}, y_{13})$, where x_i represents year and y_i represents the population for the corresponding year

☞ The first step is to compute the sample means

$$\bar{x} = \frac{1}{13} \sum_{i=1}^{13} x_i = 1980 \text{ and } \bar{y} = \frac{1}{13} \sum_{i=1}^{13} y_i = 4558.1$$

Then, we setup and fill the following table

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1950	2558	-30	-2000.1	900	60003	4000400.01
1955	2782	-25	-1776.1	625	44402.5	3154531.21
1960	3043	-20	-1515.1	400	30302	2295528.01
1965	3350	-15	-1208.1	225	18121.5	1459505.61
1970	3712	-10	-846.1	100	8461	715885.21
1975	4089	-5	-469.1	25	2345.5	220054.81
1980	4451	0	-107.1	0	0	11470.41
1985	4855	5	296.9	25	1484.5	88149.61
1990	5287	10	728.9	100	7289	531295.21
1995	5700	15	1141.9	225	17128.5	1303935.61
2000	6090	20	1531.9	400	30638	2346717.61
2005	6474	25	1915.9	625	47897.5	3670672.81
2010	6864	30	2305.9	900	69177	5317174.81

- ▶ Taking sums of last three columns, we get,

$$S_{xx} = 4550, S_{xy} = 337250 \text{ and } S_{yy} = 25115320.93$$

- ▶ Hence the sample correlation coefficient is

$$r = \frac{337250}{\sqrt{4550}\sqrt{25115320.93}} = 0.9976$$

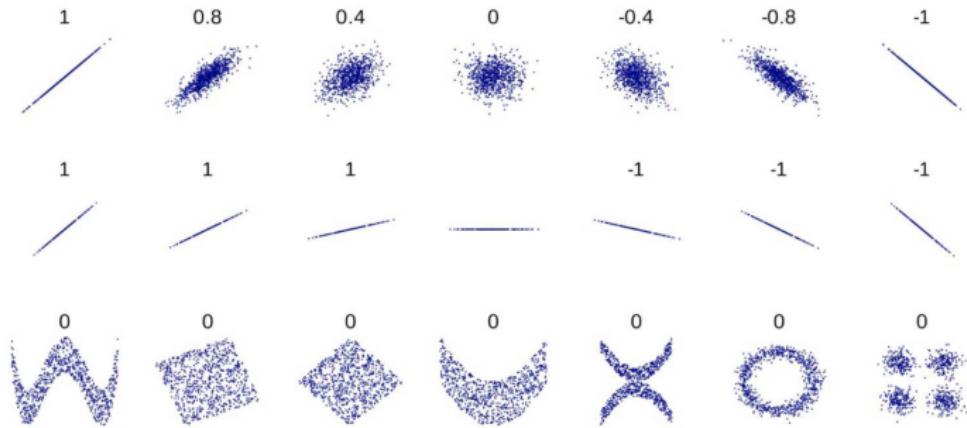


Some properties of the correlation coefficient:

- ▶ $-1 \leq r \leq 1$ (Recall $-1 \leq \rho(X, Y) \leq 1$)
- ▶ Correlation, most often quantified by Pearson's correlation coefficient (r), indicates the degree of interdependence of the values of two variables
- ▶ This observation can in some cases be used to estimate how likely two events or characteristics are to occur together
- ▶ Care must be taken to apply and interpret correlation appropriately
- ▶ Correlation does not indicate that one event or characteristic causes the other: both can be caused by a third event or characteristic or can be merely coincidental
- ▶ That is, correlation does not imply causation!
- ▶ It surely indicates an association between the variables, but the relationship might not be linear

- ▶ Pearson's r only tells us how close the data are to fitting on a line.
- ▶ A line of slope equal to 1 is only a single example of line that data could lie on and have that value of $r = 1$.
- ▶ Similarly, a line of slope equal to -1 is only a single example of line that data could lie on and have that value of $r = -1$
- ▶ The slope is only equal to r if the variables have been put into standard units.
- ▶ However, the slope and r always have the same sign, so we do know that the data have a positive or negative slope by looking at the sign of r
- ▶ $r = 0$ when
 - ▶ all the data points do not lie on any line, or
 - ▶ can best be fit with a line with slope = 0

☞ Some examples of scatter plots with their corresponding correlation coefficients



☞ Correlation was all about measure of linear relationship

☞ Suppose for some data we feel that there is a linear relationship, a natural question would be "**how do we find the equation of the line which best fits the given data?**"

Probability and Statistics

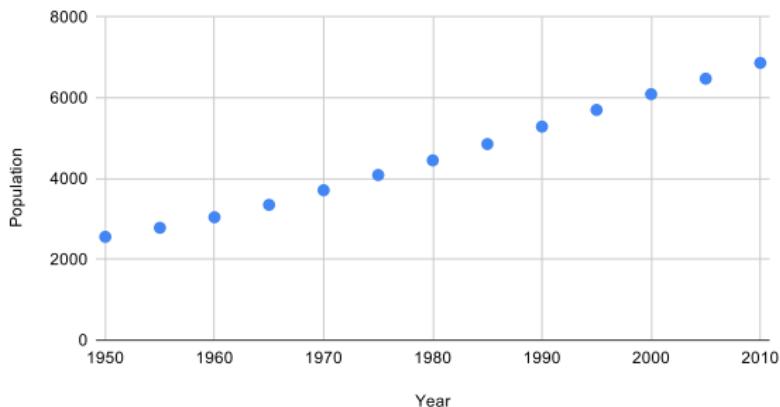
Lecture-28

Example: Let us get back to our population data

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287		
1970	3712	1995	5700		

⌚ We saw that the scatter plot will be

Population vs. Year



- ☞ We can clearly see that the data points lie on a straight line (approximately)
- ☞ **Question:** How do we find the equation of the line?
- ☞ If we could get the equation of this line which “**fits best**” to given set of points, then we can easily forecast or estimate population for any year
- ☞ That is, suppose $G(x) = b_0 + b_1x$ is the equation of “best fit” line, then $G(2015)$ gives the population forecast for the year 2015

☞ To sum up, suppose

- ▶ we have bi-variate data (variables be x, y), and
- ▶ the scatter plot of the data shows a linear trend,

then, to predict future values of y for particular values of x , our aim should be finding the equation of “**best fit**” line, say,

$$y = G(x) = b_0 + b_1 x$$

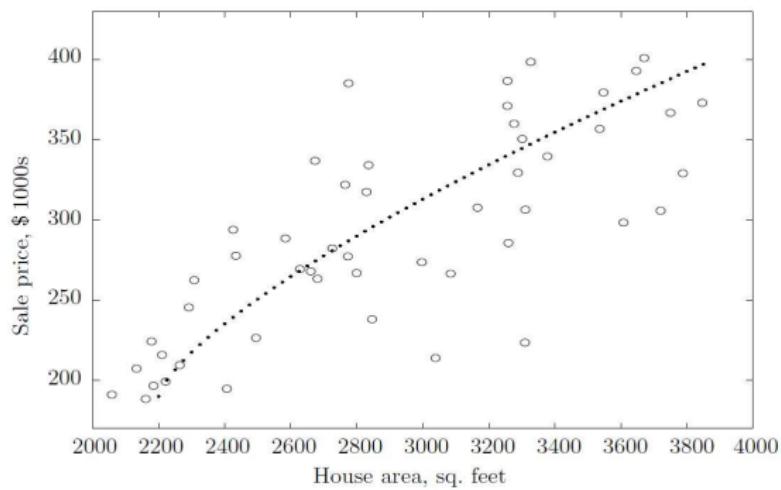
☞ The final step is to substitute values of x to get the corresponding “**estimates**” of y

☞ **Question:** Does all the data we come across show linear trend?
That is, does the scatter plot of every data show a linear trend?

Answer: Obviously, no!

☞ Let us look at an example!

Example: Seventy house sale prices in a certain county are depicted in following figure along with the house area



☞ First, we see a clear relation between these two variables, and in general, bigger houses are more expensive. However, the trend no longer seems linear

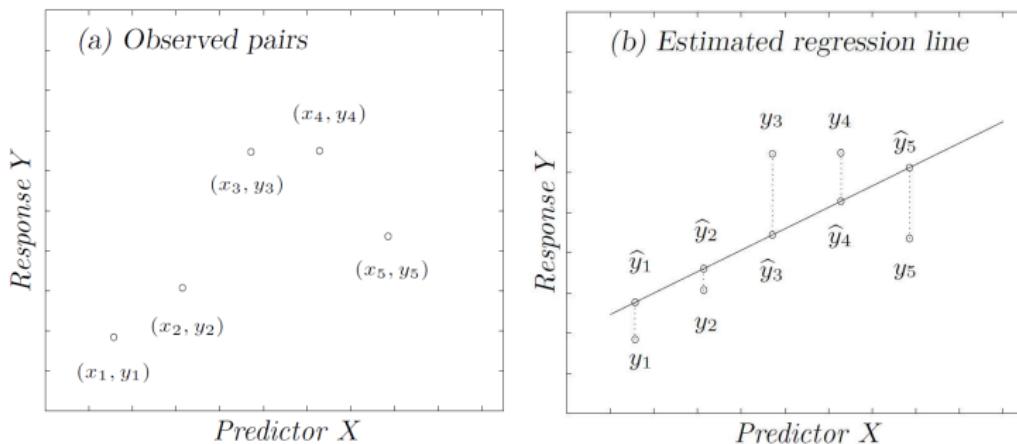
☞ Second, there is a large amount of variability around this trend.

- ☞ Then, in such cases, how can we estimate the price of a, say, 3200-square-foot house?
- ☞ One way is to still try to fit the data to a line (dotted). But, now, due to obviously high variability, our estimation will not be as accurate as in previous example
- ☞ A better but tougher method is trying to fit the data to a curve and finding the equation of the curve. This is called "**non-linear regression**". This process involves more advanced techniques.
- ☞ As of now, **we only consider models based on straight lines.**

- ☞ Coming back, given pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, we wish to find the equation $y = b_0 + b_1x$ of the line which fits best
- ☞ That is, we need to estimate the unknown coefficients b_0 and b_1 using the given data
- ☞ Here, using the equation, we are trying to predict or estimate the values of y given values of x
- ☞ We usually call x **the explanatory or predictor or independent variable**, and we call y **the response or dependent variable**.
- ☞ **Question:** How do we find the equation $y = b_0 + b_1x$?

Given pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, we wish to find the equation $y = b_0 + b_1 x$ of the line which fits best

What do we mean by “fitting best”?



$$\hat{y}_i = b_0 + b_1 x_i$$

The line will be the best fit if the difference between the estimates \hat{y}_i (given by the equation) and the actual values y_i is minimum for every $i = 1, 2, \dots, n$

Residual: difference between observed and expected

The residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit (\hat{y}_i) :

$$e_i = y_i - \hat{y}_i$$

We typically identify \hat{y}_i by plugging x_i into the model.

- ☞ We can say that the line $y = b_0 + b_1x$ is the best fit if all the residuals $e_i = y_i - \hat{y}_i$ are minimum
- ☞ Each observation will have a residual. i.e. the vertical distance from the observation to the line.
- ☞ If an observation is above the regression line then its residual is positive. Observations below the line have negative residuals.
- ☞ Since we talk about “distance” between the estimated and observed values, we need to minimize $|e_i| = |y_i - \hat{y}_i|$ for every i
- ☞ In one shot, we need to minimize the sum $|e_1| + |e_2| + \dots + |e_n|$

☞ Considering Mathematical technicalities, **the method of least squares** consists of finding the unknown coefficients b_0 and b_1 by minimizing the **sum of squares of residuals**

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

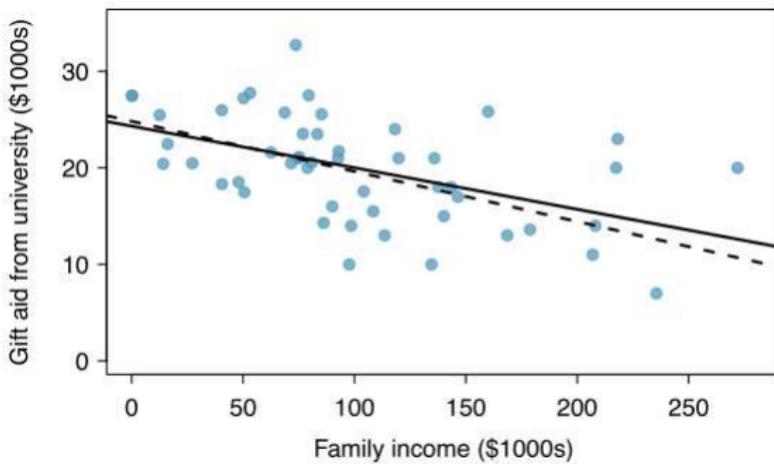
☞ This sum is denoted by SSE , which stands for "**sum of squared errors**"

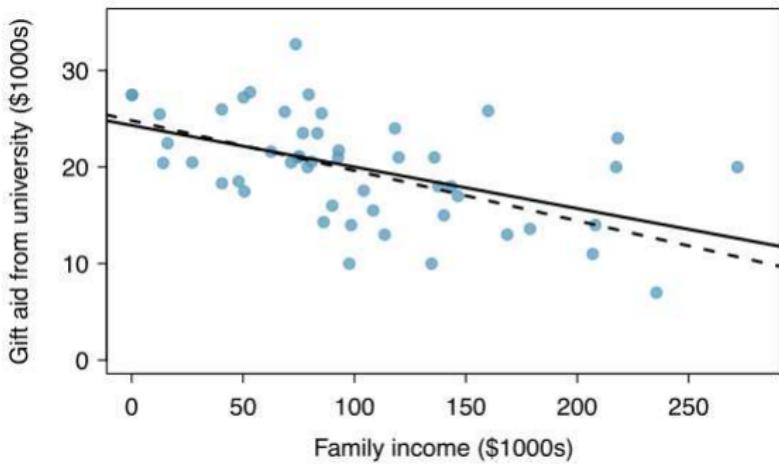
$$SSE = e_1^2 + e_2^2 + \cdots + e_n^2$$

☞ The best fit line obtained by the method of least squares is sometimes referred to as **the least squares line**

☞ The least squares method is the most common method and is easy to compute by hand or using statistical software

- ▶ Consider a scenario of **family income** and **gift aid** data from a random sample of 50 students in the 2011 freshman class of **Elmhurst College** in Illinois
- ▶ *Gift aid is financial aid that does not need to be paid back, as opposed to a loan*
- ▶ The scatterplot of the data is shown below along with two linear fits





- ▶ The lines follow a negative trend in the data; students who have higher family incomes tended to have lower gift aid from the university
- ▶ The **solid line being the least squares line** and the dotted line is of another uncommon approach which involves more computational efforts

When do we use the method of least squares?

- ▶ The data should show a linear trend
- ▶ Nearly normal residuals
- ▶ The variability of points around the least squares line remains roughly constant (constant Variability)
- ▶ **Independent observations** - Be cautious about applying regression to *time series data*, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis

When not to go for linear regression?

- ▶ The trend is not linear
- ▶ There are outliers
- ▶ The variability of the data around the line increases with larger values of x (i.e., the data is heteroscedastic)
- ▶ Time series data where successive observations are highly correlated

☞ The method of least squares estimates b_0 and b_1 by minimizing

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

☞ **Conclusions from the method:**

- ▶ $b_1 = \frac{s_y}{s_x} r$, which is the slope of the least squares line
- ▶ The line passes through the point (\bar{x}, \bar{y})

☞ The method gave us the slope of the line and a point through which the line passes

☞ How to find equation of the line with these two?

☞ **Point-slope form of a line!**

☞ The equation of a line which passes through the point (a, b) and which has slope m is given by $(y - b) = m(x - a)$

☞ Our least squares line passes through (\bar{x}, \bar{y}) and has slope

$$b_1 = \frac{s_y}{s_x} r$$

Given sample data consisting of pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, to find the equation of best-fit line,

Step-1: Compute sample means \bar{x} and \bar{y}

Step-2: Set up a table with columns $x_i, y_i, (x_i - \bar{x}), (y_i - \bar{y}), (x_i - \bar{x})^2, (x_i - \bar{x})(y_i - \bar{y}), (y_i - \bar{y})^2$ and hence compute

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

Step-3: Compute Pearson's correlation coefficient r by the formula,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Step-4: Compute the sample standard deviations

$$s_x = \sqrt{\frac{1}{n-1} S_{xx}} \text{ and } s_y = \sqrt{\frac{1}{n-1} S_{yy}}$$

Step-5: The best-fit line will pass through (\bar{x}, \bar{y}) and has slope $b_1 = \frac{s_y}{s_x} r$. That is, equation of the best-fit line is

$$y - \bar{y} = b_1(x - \bar{x})$$

☞ Let us work out an example!

☞ We are given the data

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287		
1970	3712	1995	5700		

☞ We wish to compute the equation of the best-fit line

We computed the following in the last class

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1950	2558	-30	-2000.1	900	60003	4000400.01
1955	2782	-25	-1776.1	625	44402.5	3154531.21
1960	3043	-20	-1515.1	400	30302	2295528.01
1965	3350	-15	-1208.1	225	18121.5	1459505.61
1970	3712	-10	-846.1	100	8461	715885.21
1975	4089	-5	-469.1	25	2345.5	220054.81
1980	4451	0	-107.1	0	0	11470.41
1985	4855	5	296.9	25	1484.5	88149.61
1990	5287	10	728.9	100	7289	531295.21
1995	5700	15	1141.9	225	17128.5	1303935.61
2000	6090	20	1531.9	400	30638	2346717.61
2005	6474	25	1915.9	625	47897.5	3670672.81
2010	6864	30	2305.9	900	69177	5317174.81

☞ We computed the following

$$S_{xx} = 4550, S_{xy} = 337250 \text{ and } S_{yy} = 25115320.93$$

☞ We found the sample correlation coefficient to be

$$r = \frac{337250}{\sqrt{4550}\sqrt{25115320.93}} = 0.9976$$

☞ Further,

$$s_x = \sqrt{\frac{1}{n-1}S_{xx}} = 19.47 \text{ and } s_y = \sqrt{\frac{1}{n-1}S_{yy}} = 1446.7$$

☞ We had $\bar{x} = 1980$ and $\bar{y} = 4558.1$

☞ Slope of the least squares line is $m = b_1 = \frac{s_y}{s_x} r = 74.1$ and the line passes through $(1980, 4558.1)$.

☞ Thus the equation of the line is

$$y - 4558.1 = 74.1(x - 1980)$$

Probability and Statistics

Lecture-29

Recall

Given sample data consisting of pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, to find the equation of best-fit line,

Step-1: Compute sample means \bar{x} and \bar{y}

Step-2: Set up a table with columns $x_i, y_i, (x_i - \bar{x}), (y_i - \bar{y}), (x_i - \bar{x})^2, (x_i - \bar{x})(y_i - \bar{y}), (y_i - \bar{y})^2$ and hence compute

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

Recall

Step-3: Compute Pearson's correlation coefficient r by the formula,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Step-4: Compute the sample standard deviations

$$s_x = \sqrt{\frac{1}{n-1}S_{xx}} \text{ and } s_y = \sqrt{\frac{1}{n-1}S_{yy}}$$

Step-5: The best-fit line will pass through (\bar{x}, \bar{y}) and has slope $b_1 = \frac{s_y}{s_x} r$. That is, equation of the best-fit line is

$$y - \bar{y} = b_1(x - \bar{x})$$

Multi-variate linear regression

- ▶ We learned how to predict (linearly!) a response variable Y from a predictor variable X
- ▶ We saw in several examples (as in the house sale prices) that including more information and using multiple predictors instead of one will enhance our prediction
- ▶ We now learn to connect a response Y with several predictors X_1, X_2, \dots, X_n in a **linear** fashion
- ▶ This is referred to as **multiple linear regression** or **multi-variate linear regression**

- That is, given a sample data consisting of $(n + 1)$ -tuples (one for Y and n for X_1, X_2, \dots, X_n),

y	x₁	x₂	...	x_n
y_1	x_{11}	x_{21}	...	x_{n1}
y_2	x_{12}	x_{22}	...	x_{n2}
\vdots	\vdots	\vdots	\vdots	\vdots
y_k	x_{1k}	x_{2k}	...	x_{nk}

★ x_{ij} – j^{th} observation of i^{th} variable

- We wish to find a linear equation

$$y = G(x_1, x_2, \dots, x_n) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

which **fits best** to the sample data

- We re-organize the data as follows:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{n1} \\ 1 & x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix}_{k \times (n+1)}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}_{k \times 1}$$

“Unknowns” β

$$= \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}_{(n+1) \times 1}$$

- We have to find the unknown matrix β in terms of \mathbf{X} and \mathbf{y}
- **Solution:** $\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
- Thus we get the equation
 $G(x_1, x_2, \dots, x_n) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$

Procedure - finding multi-variate linear regression equation

Step-1: Set up the matrices

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{n1} \\ 1 & x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}$$

“**Unknowns**” β =
$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

Step-2: Compute $(\mathbf{X}^T \mathbf{X})^{-1}$

Step-3: Compute $\mathbf{X}^T \mathbf{y}$ and hence $\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

Example: A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data and the database structure. Efficiency will be measured by the number of processed requests per hour. Applying the program to data sets of different sizes and different number of tables used to arrange the data in the database, she gets the following results,

Processed Requests y	Data size (gigabytes) x_1	Number of tables x_2
40	6	4
55	7	20
50	7	20
41	8	10
17	10	10
26	10	2
16	15	1

Processed Requests y	Data size (gigabytes) x_1	Number of tables x_2
40	6	4
55	7	20
50	7	20
41	8	10
17	10	10
26	10	2
16	15	1

- ▶ The response variable here is the number of processed requests (y)
- ▶ We attempt to predict it from the size of a data set (x_1) and the number of tables in the database (x_2)
- ▶ Let us compute the linear regression function for y in terms of x_1 and x_2

Step-1: Setting up the matrices

$$\mathbf{X} = \begin{pmatrix} 1 & 6 & 4 \\ 1 & 7 & 20 \\ 1 & 7 & 20 \\ 1 & 8 & 10 \\ 1 & 10 & 10 \\ 1 & 10 & 2 \\ 1 & 15 & 1 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 40 \\ 55 \\ 50 \\ 41 \\ 17 \\ 26 \\ 16 \end{pmatrix}$$

“Unknowns” β = $\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

Step-2: Calculate $(\mathbf{X}^T \mathbf{X})^{-1}$

$$\mathbf{X} = \begin{pmatrix} 1 & 6 & 4 \\ 1 & 7 & 20 \\ 1 & 7 & 20 \\ 1 & 8 & 10 \\ 1 & 10 & 10 \\ 1 & 10 & 2 \\ 1 & 15 & 1 \end{pmatrix} \implies \mathbf{X}^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 7 & 7 & 8 & 10 & 10 & 15 \\ 4 & 20 & 20 & 10 & 10 & 2 & 1 \end{pmatrix}$$

$$\implies \mathbf{X}^T \mathbf{X} = \begin{pmatrix} 7 & 63 & 67 \\ 63 & 623 & 519 \\ 67 & 519 & 1021 \end{pmatrix} \quad (\mathbf{X}^T \mathbf{X})^{-1} = ???$$

Finding inverse of a 3×3 matrix

☞ Let $M = \mathbf{X}^T \mathbf{X}$

Step-2(a): Compute $\det(M)$ and write down M^T

$$\det(M) = 99456 \text{ and } M^T = \begin{pmatrix} 7 & 63 & 67 \\ 63 & 623 & 519 \\ 67 & 519 & 1021 \end{pmatrix} = M$$

Step-2(b) Write down all the minors of M^T and find their determinants

$$M_{11} = \begin{pmatrix} 623 & 519 \\ 519 & 1021 \end{pmatrix}$$

$$M_{12} = \begin{pmatrix} 63 & 519 \\ 67 & 1021 \end{pmatrix}$$

$$M_{13} = \begin{pmatrix} 63 & 623 \\ 67 & 519 \end{pmatrix}$$

$$M_{21} = \begin{pmatrix} 63 & 67 \\ 519 & 1021 \end{pmatrix}$$

$$M_{22} = \begin{pmatrix} 7 & 67 \\ 67 & 1021 \end{pmatrix}$$

$$M_{23} = \begin{pmatrix} 7 & 63 \\ 67 & 519 \end{pmatrix}$$

$$M_{31} = \begin{pmatrix} 63 & 67 \\ 623 & 519 \end{pmatrix}$$

$$M_{32} = \begin{pmatrix} 7 & 67 \\ 63 & 519 \end{pmatrix}$$

$$M_{31} = \begin{pmatrix} 7 & 63 \\ 63 & 623 \end{pmatrix}$$

Finding inverse of a 3×3 matrix

$$|M_{11}| = 366722 \quad |M_{12}| = 29550 \quad |M_{13}| = -9044$$

$$|M_{21}| = 29550 \quad |M_{22}| = 2658 \quad |M_{23}| = -588$$

$$|M_{31}| = -9044 \quad |M_{32}| = -588 \quad |M_{33}| = 392$$

Step-2(c): Write down the **adjugate matrix** $\text{Adj}(M)$

$$\text{Adj}(M) = \begin{pmatrix} +|M_{11}| & -|M_{12}| & +|M_{13}| \\ -|M_{21}| & +|M_{22}| & -|M_{23}| \\ +|M_{31}| & -|M_{32}| & +|M_{33}| \end{pmatrix}$$

$$\implies \text{Adj}(M) = \begin{pmatrix} 366722 & -29550 & -9044 \\ -29550 & 2658 & 588 \\ -9044 & 588 & 392 \end{pmatrix}$$

Finding inverse of a 3×3 matrix

Step-2(d): $M^{-1} = \frac{1}{\det(M)} \times \text{Adj}(M)$

$$M^{-1} = \frac{1}{99456} \begin{pmatrix} 366722 & -29550 & -9044 \\ -29550 & 2658 & 588 \\ -9044 & 588 & 392 \end{pmatrix}$$

$$= \begin{pmatrix} 3.687 & -0.297 & -0.091 \\ -0.297 & 0.027 & 0.006 \\ -0.091 & 0.006 & 0.004 \end{pmatrix}$$

$$\implies \text{Step-2: } (\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} 3.687 & -0.297 & -0.091 \\ -0.297 & 0.027 & 0.006 \\ -0.091 & 0.006 & 0.004 \end{pmatrix}$$

Step-3: Compute $\mathbf{X}^T \mathbf{y}$ and hence $\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

$$\mathbf{X}^T \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 7 & 7 & 8 & 10 & 10 & 15 \\ 4 & 20 & 20 & 10 & 10 & 2 & 1 \end{pmatrix} \begin{pmatrix} 40 \\ 55 \\ 50 \\ 41 \\ 17 \\ 26 \\ 16 \end{pmatrix} = \begin{pmatrix} 245 \\ 1973 \\ 2908 \end{pmatrix}$$

$$\text{Now, } \beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \begin{pmatrix} 3.687 & -0.297 & -0.091 \\ -0.297 & 0.027 & 0.006 \\ -0.091 & 0.006 & 0.004 \end{pmatrix} \begin{pmatrix} 245 \\ 1973 \\ 2908 \end{pmatrix}$$

$$\text{Thus, } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 52.706 \\ -2.046 \\ 1.175 \end{pmatrix}$$

Hence, the regression equation is

$$y = G(x_1, x_2) = 52.706 - 2.046x_1 + 1.175x_2$$



- ☞ We learnt how to fit data to a linear equation
- ☞ How do you test the “**goodness**” of our fit?
- ☞ Let y_1, y_2, \dots, y_n be the observed values of the sample data and $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ be the estimated values obtained using the linear equation computed

- ▶ **Total sum of squares** is defined as

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = (n - 1)s_y^2 = S_{yy}$$

- ▶ **Regression sum of squares** is defined as

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

☞ **Recall:** Error sum of squares

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

☞ It turns out that

$$SST = SSR + SSE$$

☞ **R-square** or **coefficient of determination** is the proportion of SSR to SST . That is,

$$R^2 = \frac{SSR}{SST}$$

☞ In fact, $R^2 = r^2$

☞ R^2 is always between 0 and 1, with high values generally suggesting a good fit

Example: For the world population data,

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287		
1970	3712	1995	5700		

☞ We found that $S_{yy} = 25115320.93$ and the best-fit line is

$$G(x) = \beta_0 + \beta_1 x = -142201 + 74.1x$$

☞ Hence $SST = 25115320.93 \approx 2.512 \times 10^7$

☞ Substituting $x = 1950, 1955, \dots, 2010$, we get $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{13}$.

☞ **Check!** $SSR = \sum_{i=1}^{13} (\hat{y}_i - \bar{y})^2 \approx 2.5 \times 10^7$

☞ Thus, $R^2 = \frac{SSR}{SST} \approx 0.995$ which implies that the straight line we obtained is a very good fit for the given data!

Probability and Statistics

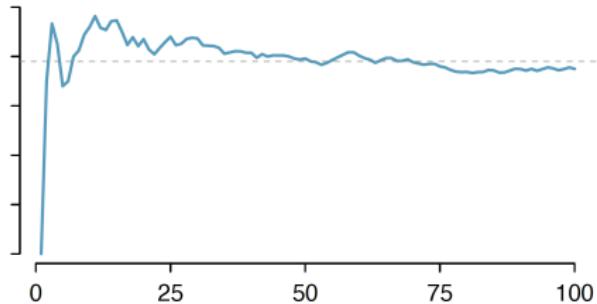
Lecture-30

- ☞ Suppose we wish to determine the following for students enrolled at a particular university:
 - ▶ μ = (unknown) **average** height of all the students in the university
 - ▶ p = (unknown) **proportion** of students whose height falls in the range 160-165 cm
- ☞ In either case, we can't possibly survey the entire population. That is, we can't survey all university college students.
- ☞ So, of course, we do what comes naturally and take a random sample from the population, and use the resulting data to make generalizations about unknown population
- ☞ This part of statistics is called **inferential statistics**
- ☞ The estimates obtained from the sample data are called **point estimates**

☞ Estimates generally vary from one sample to another, and this sampling variation suggests our estimate may be close, but it will not be exactly equal to the parameter

☞ Estimates better as more data become available. We can see this by plotting a **running mean**

☞ A **running mean** is a sequence of means, where each mean uses one more observation in its calculation than the mean directly before it in the sequence



☞ The running mean tends to approach the true population average (dotted line) as more data become available

- ☞ Sample point estimates only approximate the population parameter, and they vary from one sample to another
- ☞ If we take another simple random sample from the population, we would find that the estimate is little different
- ☞ It will be useful to quantify how variable an estimate is from one sample to another
- ☞ If this variability is small (i.e. the sample mean doesn't change much from one sample to another) then that estimate is probably very accurate
- ☞ If it varies widely from one sample to another, then we should not expect our estimate to be very good
- ☞ This variation is measured by **standard error (SE)** of the corresponding estimate/statistic

Sampling distribution of statistics

- ☞ Suppose we are interested in estimating the unknown parameter θ (may be mean, median, proportion etc.) of a population
- ☞ From the population, start picking up random samples of fixed size, say n . That is, we pick a random sample x_1, x_2, \dots, x_n from the population
- ☞ Calculate statistic which estimates θ . Let us denote it by $\hat{\theta}_1$
- ☞ Replace the sample and pick another random sample of same size and do the same calculation to get $\hat{\theta}_2$
- ☞ Repeating this procedure, we get the sampling distribution of the statistic $\hat{\theta}$

☞ **Standard error** of $\hat{\theta}$ is the standard deviation of the sampling distribution of $\hat{\theta}$

☞ Now, let us restrict ourselves to estimation of population mean μ and look at the corresponding sampling distribution of sample mean \bar{X} based on samples of size n

☞ Since sample mean is an unbiased estimate of population mean, we have $E[\bar{X}] = \mu$

☞ That is, the mean of the sampling distribution of sample mean is the population mean itself!

☞ What about standard deviation of the sampling distribution of sample mean?

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\ &= \frac{1}{n^2} [\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)] \quad (\text{i.i.d!}) \\ &= \frac{1}{n^2} (n \cdot \sigma^2) = \frac{\sigma^2}{n}\end{aligned}$$

where σ is the population standard deviation.

☞ The standard error of sample mean ($SE(\bar{X})$) is $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation and n is the sample size

☞ What does this tell us?

☞ If we are trying to estimate mean of a population with a sample of size n , then the magnitude of error we get in such estimation is $\frac{\sigma}{\sqrt{n}}$

☞ Clearly, given σ , as the sample size n increases, the error goes to zero and we get better estimates!

☞ On the other hand, for a fixed sample size, error is directly dependent on the population standard deviation σ

☞ We were able to compute the mean and standard deviation of sampling distribution of sample mean without actually knowing about the distribution

☞ In practice, we come across problems similar to the one given below

Example: It is found that average weight of people residing in a particular city is 62 Kg with a standard deviation of 5.2 Kg. If we choose a random sample of 75 people, what is probability that mean weight of the chosen sample is between 61 Kg and 64 Kg?

☞ To answer such questions we need to know the actual distribution of sample means!

☞ How do we get to know about the distribution?

☞ **Answer:** Central limit theorem!

- ☞ For each sample (X_1, X_2, \dots, X_n) , each of the observations X_i is a random variable
- ☞ Moreover, since the sample is chosen randomly, X_1, X_2, \dots, X_n are i.i.d. random variables!
- ☞ CLT says that if $n > 30$, $X_1 + X_2 + \dots + X_n$ will be normal with parameters $n\mu$ and $n\sigma^2$
- ☞ That is, for $n > 30$, sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ will be normal with parameters $\left(\mu, \frac{\sigma^2}{n}\right)$
- ☞ Hence, for $n > 30$, any probability involving \bar{X} can be computed using the standard normal table!

Caution: This approach is valid only if $n > 30$, that is, if the sample size is greater than 30!

Example: It is found that average weight of people residing in a particular city is 62 Kg with a standard deviation of 5.2 Kg. If we choose a random sample of 75 people, what is probability that mean weight of the chosen sample is between 61 Kg and 64 Kg?

- ☞ The variable of interest, X , is the weight of people residing in the city
- ☞ We are given that population mean and population SD are 62 Kg and 5.2 Kg respectively
- ☞ That is, $\mu = E[X] = 62$ and $\sigma = SD(X) = 5.2$
- ☞ The question asks for probability involving the sample mean \bar{X} computed from random samples of size $n = 75$
- ☞ Since $n > 30$, by CLT, we may assume that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- ☞ That is, $\bar{X} \sim N\left(62, \frac{(5.2)^2}{75}\right)$

Now,

$$\begin{aligned}P\{61 < \bar{X} < 64\} &= P\left\{\frac{61 - 62}{(5.2/\sqrt{75})} < Z < \frac{64 - 62}{(5.2/\sqrt{75})}\right\} \\&= P\{-1.67 < Z < 3.33\} \\&= \Phi(3.33) - \Phi(-1.67) \\&= \Phi(3.33) - (1 - \Phi(1.67)) \\&= 0.9996 - (1 - 0.9525) \\&= 0.9521\end{aligned}$$



☞ We started this lecture with the following examples

- ▶ μ = (unknown) **average** height of all the students in the university
- ▶ p = (unknown) **proportion** of students whose height falls in the range 160-165 cm

☞ Till now we were talking about sampling distribution of sample mean, its standard error etc., which mainly aims at the first example

☞ This, along with CLT, helped us to answer questions about probabilities involving sample mean

☞ What would be the standard error in estimation of population proportion with a sample of size n ? Sampling distribution of sample proportion?

Example: It is known that 45% of students have their heights in the range 160-165 cm. If random sample of 150 students is chosen, what is the probability of proportion of students with height in the range 160-165 cm is greater than 40%?

- ☞ Here the variable of interest is the “proportion” of students with height in the range 160-165 cm
- ☞ It is given that population proportion $p = 0.45$
- ☞ We are choosing random samples of 150 students and looking at the proportion \hat{p} of students with height in the range 160-165 cm in this sample
- ☞ To answer the question about the probability, we need to know the distribution of \hat{p}

- We have a population which consists of students with heights in the range 160-165 cm
- Let us term a student as “success” if his/her height falls in the range 160-165 cm
- In the sample of size 75, we are looking at the proportion of successes \hat{p}
- If the number of successes is S , then $\hat{p} = \frac{S}{75}$
- S is a random variable which records the number of successes in 75 trials! Thus, $S \sim \text{Bin}(75, p=0.45)$
- Mean of S is $np = 75 \times 0.45 = 33.75$ and variance is $np(1 - p) = 18.5625$

The standard error of sample proportion ($SE(\hat{p})$) is $\sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion and n is the sample size

- ☞ By CLT, S can be approximated by normal distribution (as both np and $n(1 - p)$ are greater than 5)
- ☞ With such an approximation $\hat{p} = \frac{S}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$
- ☞ Thus, $\hat{p} \sim N(0.45, 0.0033)$
- ☞ The question asks for probability of $\hat{p} > 0.4$

$$\begin{aligned}P\{\hat{p} > 0.4\} &= P\left\{\frac{\hat{p} - 0.45}{\sqrt{0.0033}} > \frac{0.4 - 0.45}{\sqrt{0.0033}}\right\} \\&= P\{Z > -0.87\} \\&= 1 - \Phi(-0.87) \\&= \Phi(0.87) = 0.8078\end{aligned}$$



Summary

☞ Sampling distribution of sample mean

- ▶ From a population with mean μ and variance σ^2 , we draw random samples of size n
- ▶ Since sample mean (\bar{X}) varies from sample to sample, we are interested in the distribution of sample mean, which we call as **sampling distribution** of \bar{X}
- ▶ The standard deviation of sampling distribution of \bar{X} is termed as **standard error (SE)** of \bar{X}
- ▶ The mean of sampling distribution of \bar{X} will be the population mean μ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- ▶ CLT \implies for $n > 30$, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Summary

☞ Sampling distribution of sample proportion

- ▶ We are interested in a proportion of population p and we draw random samples of size n to estimate it
- ▶ Different samples give different sample proportions (\hat{p}) and we looked at sampling distribution of \hat{p}
- ▶ The standard deviation of sampling distribution of \hat{p} is termed as **standard error (SE)** of \hat{p}
- ▶ The mean of sampling distribution of \hat{p} will be the population proportion p and $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$
- ▶ $n\hat{p} \sim \text{Bin}(n, p)$
- ▶ By CLT, for $np > 5$ and $n(1 - p) > 5$, $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

Probability and Statistics

Lecture-31

☞ Let us get back to earlier examples of determining the following for students enrolled at a particular university:

- ▶ μ = (unknown) **average** height of all the students in the university
- ▶ p = (unknown) **proportion** of students whose height falls in the range 160-165 cm

☞ We pick random samples from the population, and use the resulting data to estimate the value of these unknown population parameters

☞ Such kind of estimates are called **point estimates**

☞ A point estimate provides a single plausible value for a parameter

☞ However, a point estimate is rarely perfect; usually there is some error in the estimate

- ☞ Instead of supplying just a point estimate of a parameter, the next logical step would be to provide a plausible range of values for the parameter
- ☞ A plausible range of values for the population parameter is called a **confidence interval (CI)**
- ☞ Our point estimate is the most plausible value of the parameter, so it makes sense to build the confidence interval around the point estimate
- ☞ The standard error, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval

☞ Thus, computation of confidence interval(CI) for population parameter requires:

- ▶ the point estimate, i.e., the *computed* statistic $\hat{\theta}$,
- ▶ the investigator's desired **level of confidence** or **confidence level** α (most commonly 95, but any level between 0 – 100 can be selected),
- ▶ the sampling variability or the standard error(SE) of the point estimate, $SE(\hat{\theta})$

☞ A typical confidence interval looks like

$$(\hat{\theta} - ME, \hat{\theta} + ME)$$

where the **margin of error(ME)** is determined by the confidence level α and $SE(\hat{\theta})$

☞ What does it mean to say that (a, b) is the 90% confidence interval for the population parameter θ ?

- ▶ It **does not** mean that, if we take another sample of same size, there is a 90% chance that the estimate from the new sample will fall in the interval (a, b)
- ▶ It **means** that we are 90% confident that the population mean lies in the interval (a, b)

☞ What does it mean to say that we have considered 95% confidence level?

- ▶ It **means** that, if we repeat the process of sampling and calculate such intervals, then about 95% of those intervals will contain the population parameter
- ☞ Confidence levels tell us the long-term rate at which a certain type of confidence interval will successfully capture the parameter of interest

☞ For any confidence level α ($0 < \alpha < 100$), the corresponding confidence interval should satisfy

$$P\{\hat{\theta} - ME < \theta < \hat{\theta} + ME\} = \frac{\alpha}{100}$$

☞ This can be re-written as

$$P\{-ME < \hat{\theta} - \theta < ME\} = \frac{\alpha}{100}$$

$$\implies P\left\{-\frac{ME}{SE(\hat{\theta})} < \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} < \frac{ME}{SE(\hat{\theta})}\right\} = \frac{\alpha}{100}$$

☞ To get hold of such an ME, we need to know the distribution of $\frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$

☞ If F is the distribution function of $\frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$, then the above equality becomes

$$F\left(\frac{ME}{SE(\hat{\theta})}\right) - F\left(-\frac{ME}{SE(\hat{\theta})}\right) = \frac{\alpha}{100}$$

Confidence intervals for population mean

- Let us fix confidence level α with $0 < \alpha < 100$
- Given a sample of size n , to compute the $\alpha\%$ confidence interval for population mean, We need to find ME in the interval

$$(\bar{x} - ME, \bar{x} + ME)$$

satisfying

$$F\left(\frac{ME}{SE(\bar{x})}\right) - F\left(-\frac{ME}{SE(\bar{x})}\right) = \frac{\alpha}{100}$$

where F is the distribution function of $\frac{\bar{x}-\mu}{SE(\bar{x})}$

- Since $SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$, we should find ME satisfying

$$F\left(\frac{ME}{\sigma/\sqrt{n}}\right) - F\left(-\frac{ME}{\sigma/\sqrt{n}}\right) = \frac{\alpha}{100}$$

where σ is the population standard deviation which may be known or unknown

We divide our problem in to two cases:

Case-1: Population sd σ is known

Case-2: Population sd σ is unknown

Case-1: When σ is known, we assume that $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ and hence

$$\begin{aligned} F\left(\frac{ME}{\sigma/\sqrt{n}}\right) - F\left(-\frac{ME}{\sigma/\sqrt{n}}\right) &= \frac{\alpha}{100} \implies \Phi\left(\frac{ME}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{ME}{\sigma/\sqrt{n}}\right) = \frac{\alpha}{100} \\ &\implies \Phi\left(\frac{ME}{\sigma/\sqrt{n}}\right) = \frac{(\alpha/100) + 1}{2} \end{aligned}$$

NOTATION: z_α be such that $\Phi(z_\alpha) = \frac{(\alpha/100)+1}{2}$

Since $\frac{\sigma}{\sqrt{n}}$ is known, we have $ME = z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$ and hence we get the interval

$$(\bar{x} - ME, \bar{x} + ME)$$

$\alpha\%$ confidence intervals for population mean with known σ

☞ Given a sample of size n and population sd σ , to compute $\alpha\%$ confidence interval for the population mean,

- ▶ **Step-1:** Calculate sample mean \bar{x} and $\frac{\sigma}{\sqrt{n}}$
- ▶ **Step-2:** Using the standard normal table find the value of z_α satisfying

$$\Phi(z_\alpha) = \frac{(\alpha/100) + 1}{2}$$

- ▶ **Step-3:** Compute $ME = z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$
- ▶ **Step-4:** The $\alpha\%$ confidence interval for the population mean will be

$$(\bar{x} - ME, \bar{x} + ME)$$

Some standard values of z_α and the corresponding ME

$\alpha\%$	z_α	ME
90%	1.645	$1.645 \cdot \frac{\sigma}{\sqrt{n}}$
95%	1.96	$1.96 \cdot \frac{\sigma}{\sqrt{n}}$
99%	2.576	$2.576 \cdot \frac{\sigma}{\sqrt{n}}$

Example: You are testing chocolate chip cookies to estimate the mean number of chips per cookie. You sample 50 cookies and find that the sample mean of 10 chips per cookie. If it is known that there is a standard deviation of 2 per cookie in the entire population of cookies, set up a 95% confidence interval for the mean number of chocolate chips per cookie in the entire population.

☞ We are given the sample size $n = 50$, the sample mean $\bar{x} = 10$ and the population sd $\sigma = 2$

☞ Hence $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{50}} \approx 0.2828$

☞ Now, $\alpha = 95 \implies z_{\alpha} = 1.96$

☞ $ME = z_{\alpha} \times \frac{\sigma}{\sqrt{n}} \approx 0.5543$

☞ Thus, the 95% confidence interval for population mean is

$$(10 - 0.5543, 10 + 0.5543) = (9.4457, 10.5543)$$

☞ **Case-2:** Population sd σ is unknown

☞ We further divide this case into two sub-cases:

- ▶ **Case-2(a):** Sample size $n \geq 30$
- ▶ **Case-2(b):** Sample size $n < 30$

☞ **Case-2(a):** In this case, it turns out that the sample sd s is a pretty good estimate for σ and by CLT, again, $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim N(0, 1)$ and hence

$$ME = z_\alpha \times \frac{s}{\sqrt{n}} \text{ with } \Phi(z_\alpha) = \frac{(\alpha/100) + 1}{2}$$

☞ **Case-2(b):** Even here we estimate σ using s but since the estimate is not accurate, this error can be compensated by taking $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ to be following Student's t-distribution with df $\nu = n - 1$.

Hence $ME = t_\alpha \times \frac{s}{\sqrt{n}}$ with $F_{n-1}(t_\alpha) = \frac{(\alpha/100)+1}{2}$, where F_{n-1} is the distribution function of Student's t-distributed random variable with df $\nu = n - 1$

$\alpha\%$ confidence intervals for population mean with unknown σ

☞ Given a sample of size n to compute $\alpha\%$ confidence interval for population mean,

- ▶ **Step-1:** Compute the sample mean \bar{x}
- ▶ **Step-2:** Compute $\frac{s}{\sqrt{n}}$, where s is the sample standard deviation (to be computed using the sample!)
- ▶ **Step-3:** If $n \geq 30$ (else go to Step-4), find the value of z_α which satisfies $\Phi(z_\alpha) = \frac{(\alpha/100)+1}{2}$. Then compute

$$ME = z_\alpha \cdot \frac{s}{\sqrt{n}}$$

(to find z_α , we use the standard normal table)

- ▶ **Step-4:** If $n < 30$, find the value of t_α which satisfies $F_{n-1}(t_\alpha) = \frac{(\alpha/100)+1}{2}$. Then compute

$$ME = t_\alpha \cdot \frac{s}{\sqrt{n}}$$

(to find t_α , we use the calculator given at
<https://stattrek.com/online-calculator/t-distribution.aspx>)

- ▶ **Step-5:** Then the $\alpha\%$ confidence interval for population mean will be

$$(\bar{x} - ME, \bar{x} + ME)$$

Example: A researcher randomly selected 20 batteries from the production line of Duracell and tested these batteries. The tested batteries had a mean life span of 270 hours with a standard deviation of 45 hours. Set up a 90% confidence interval for the average life span of batteries that are manufactured by Duracell.

- ▶ We are given the sample size $n = 20$, sample mean $\bar{x} = 270$ and sample sd $s = 45$
- ▶ Since the population sd is unknown and sample size is less than 30, we need to compute t_{90} which satisfies

$$F_{19}(t_{90}) = \frac{(90/100) + 1}{2}$$

where F_{19} is the distribution function of Student's t-distribution with df $\nu = 19$

- ▶ Now, $\frac{(90/100)+1}{2} = 0.95$. By entering df as 19 and t-score as 0.95 in the t-calculator, we get $t_{90} = 0.8230$

- ▶ Thus, $ME = t_{90} \times \frac{s}{\sqrt{n}} \approx 8.2813$
- ▶ Hence the 90% confidence interval for the average life span of batteries that are manufactured by Duracell is

$$(270 - 8.2813, 270 + 8.2813) = (261.7187, 278.2813)$$



Example: A school nurse takes a random sample of 200 students and finds that the average height of her sample is 147 cm with a standard deviation of 7 cm. Construct a 99% confidence interval for the mean height of students in the school.

- ▶ We are given the sample size $n = 200$, sample mean $\bar{x} = 147$ and sample sd $s = 7$
- ▶ Since the population sd is unknown and sample size is greater than 30, we need to compute z_{99} which satisfies

$$\Phi(z_{99}) = \frac{(99/100) + 1}{2} = 0.995 \implies z_{99} = 2.576$$

- ▶ Thus, $ME = z_{99} \times \frac{s}{\sqrt{n}} \approx 1.2751$
- ▶ Hence the 99% confidence interval for the mean height of students in the school is

$$(147 - 1.2751, 147 + 1.2751) = (145.7249, 148.2751)$$



Probability and Statistics

Lecture-32

☞ For any confidence level α ($0 < \alpha < 100$), the corresponding confidence interval should satisfy

$$P\{\hat{\theta} - ME < \theta < \hat{\theta} + ME\} = \frac{\alpha}{100}$$

☞ This can be re-written as

$$P\{-ME < \hat{\theta} - \theta < ME\} = \frac{\alpha}{100}$$

$$\implies P\left\{-\frac{ME}{SE(\hat{\theta})} < \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} < \frac{ME}{SE(\hat{\theta})}\right\} = \frac{\alpha}{100}$$

$$\implies F\left(\frac{ME}{SE(\hat{\theta})}\right) - F\left(-\frac{ME}{SE(\hat{\theta})}\right) = \frac{\alpha}{100}$$

where F is the distribution function of $\frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$

☞ We computed confidence intervals for population mean in the last lecture

Confidence intervals for population proportion

- Let us fix confidence level α with $0 < \alpha < 100$
- Given a sample of size n , to compute the $\alpha\%$ confidence interval for population proportion p , we need to find ME in the interval

$$(\hat{p} - ME, \hat{p} + ME)$$

satisfying

$$F\left(\frac{ME}{SE(\hat{p})}\right) - F\left(-\frac{ME}{SE(\hat{p})}\right) = \frac{\alpha}{100}$$

where F is the distribution function of $\frac{\hat{p}-p}{SE(\hat{p})}$ with p being the population proportion which is unknown

Since $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$, we should find ME satisfying

$$F\left(\frac{ME}{\sqrt{p(1-p)/n}}\right) - F\left(-\frac{ME}{\sqrt{p(1-p)/n}}\right) = \frac{\alpha}{100}$$

★ We will be able to compute confidence intervals for population proportion only if $np > 5$ and $n(1 - p) > 5$

★ If these conditions are not met there are alternative procedures, called **exact methods**, which we will not be going into

☞ When $np > 5$ and $n(1 - p) > 5$, by CLT, $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$ and hence,

$$\begin{aligned} F\left(\frac{ME}{\sqrt{p(1-p)/n}}\right) - F\left(-\frac{ME}{\sqrt{p(1-p)/n}}\right) &= \frac{\alpha}{100} \\ \implies \Phi\left(\frac{ME}{\sqrt{p(1-p)/n}}\right) - \Phi\left(-\frac{ME}{\sqrt{p(1-p)/n}}\right) &= \frac{\alpha}{100} \\ \implies 2\Phi\left(\frac{ME}{\sqrt{p(1-p)/n}}\right) - 1 &= \frac{\alpha}{100} \\ \implies \Phi\left(\frac{ME}{\sqrt{p(1-p)/n}}\right) &= \frac{(\alpha/100 + 1)}{2} \end{aligned}$$

☞ Thus, $\frac{ME}{\sqrt{p(1-p)/n}} = z_\alpha \implies ME = z_\alpha \cdot \sqrt{\frac{p(1-p)}{n}}$

☞ Since the population proportion p is not known, we estimate it by the sample proportion \hat{p}

☞ Hence $ME = z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

☞ Thus we get the interval

$$(\hat{p} - ME, \hat{p} + ME)$$

$\alpha\%$ confidence intervals for population proportion

☞ Given a sample of size n , to compute $\alpha\%$ confidence interval for the population proportion,

- ▶ **Step-1:** Calculate sample proportion \hat{p} and check that $n\hat{p} > 5$ and $n(1 - \hat{p}) > 5$.

☞ Proceed to step-2 only if these conditions are satisfied.

- ▶ **Step-2:** Compute $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and using the standard normal table find the value of z_α satisfying

$$\Phi(z_\alpha) = \frac{(\alpha/100) + 1}{2}$$

- ▶ **Step-3:** Compute $ME = z_\alpha \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- ▶ **Step-4:** The $\alpha\%$ confidence interval for the population proportion will be

$$(\hat{p} - ME, \hat{p} + ME)$$

Example: Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people sampled, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

- ▶ We are given the sample size $n = 500$.
- ▶ Since 421 responded yes, sample proportion $\hat{p} = \frac{421}{500} = 0.842$
- ▶ Since $n\hat{p} = 421 > 5$ and $n(1 - \hat{p}) = 79 > 5$, we can compute the confidence interval
- ▶ $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.0163$
- ▶ Since $\alpha = 95$, $z_{95} = 1.96$

- ▶ $ME = z_{\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.032$
- ▶ Hence a 95% confidence interval estimate for the true proportion of adult residents of this city who have cell phones is

$$(0.842 - 0.032, 0.842 + 0.032) = (0.81, 0.874)$$



- ☞ One job of a statistician is to make statistical inferences about populations based on samples taken from the population
- ☞ Confidence intervals are one way to estimate a population parameter
- ☞ Another way to make a statistical inference is to make a decision about a parameter
- ☞ For instance,

- ▶ a car dealer advertises that its new small truck gets 35 miles per gallon, on average
- ▶ a bike company claims that 6% of employees in a city travel to their work place by riding their bikes
- ▶ a tutoring service claims that its method of tutoring helps 90% of its students get an A or a B

- ▶ a company says that women managers in their company earn an average of \$60,000 per year
 - ▶ a university claims that 30% of the students stay on campus
- ☞ A statistician will make a decision about these claims (called **null hypothesis**)
- ☞ This process is called **hypothesis testing**
- ☞ A hypothesis test involves collecting sample data and evaluating it
- ☞ Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis
- ☞ We are mainly interested in tests of hypothesis involving statements about population mean and population proportions

- ☞ The actual test begins by considering two hypotheses
- ☞ They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints
 - H_0 : **The null hypothesis**: It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt
 - H_a : **The alternative hypothesis**: It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0
- ☞ Since the null and alternative hypotheses are contradictory, we must examine evidence to decide if we have enough evidence to reject the null hypothesis or not
- ☞ The evidence is in the form of sample data

- ☞ After we have determined which hypothesis the sample supports, we make a decision
- ☞ There are two options for a decision
- ☞ They are “**reject H_0** ” if the sample information favors the alternative hypothesis or “**do not reject H_0** ” or “**decline to reject H_0** ” if the sample information is insufficient to reject the null hypothesis
- ☞ The null is not rejected unless the hypothesis test shows otherwise
- ☞ The null statement must always contain some form of equality ($=, \leq$ or \geq)
- ☞ Always write the alternative hypothesis, typically denoted with H_a , using less than, greater than, or not equals symbols, i.e., ($\neq, >$ or $<$).

Example-1: We want to test whether the mean GPA of students in colleges is different from 5.0 (out of 10.0). The null and alternative hypotheses are:

- ▶ $H_0: \mu = 5.0$
- ▶ $H_a: \mu \neq 5.0$

Example-2: A quality control expert at a factory that paints car parts. He knows that 20% of parts have an error in their painting. He recommended a change in the painting process and he wants to see if this error rate had changed. The null and alternative hypotheses will be:

- ▶ $H_0: p = 0.2$
- ▶ $H_a: p \neq 0.2$

Example-3: According to a very large poll in 2015, about 90% of homes in California had access to the internet. Market researchers want to test if that proportion is now higher. The null and alternative hypotheses in this case will be

- ▶ $H_0: p = 0.9$
- ▶ $H_a: p > 0.9$

Example-4: We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

- ▶ $H_0: \mu \geq 5$
- ▶ $H_a: \mu < 5$

- ☞ Hypothesis can be **one-tailed** or **two-tailed** depending on the alternative hypothesis H_a
- ☞ We call a hypothesis to be **one-tailed** if the alternative hypothesis states that a parameter is **larger or smaller** ($>$ or $<$) than the null hypothesis value
- ☞ It is **two-tailed** if it states that the parameter is different (\neq) from the null value

Example-1: We want to test whether the mean GPA of students in colleges is different from 5.0 (out of 10.0). The null and alternative hypotheses are:

- ▶ $H_0: \mu = 5.0$

- ▶ $H_a: \mu \neq 5.0$

- ☞ Since H_a has \neq , the hypothesis is two-tailed

Example-2: A quality control expert at a factory that paints car parts. He knows that 20% of parts have an error in their painting. He recommended a change in the painting process and he wants to see if this error rate had changed. The null and alternative hypotheses will be:

- ▶ $H_0: p = 0.2$
- ▶ $H_a: p \neq 0.2$

☞ Even here H_a has \neq . Hence the hypothesis is two-tailed

Example-3: According to a very large poll in 2015, about 90% of homes in California had access to the internet. Market researchers want to test if that proportion is now higher. The null and alternative hypotheses in this case will be

- ▶ $H_0: p = 0.9$
- ▶ $H_a: p > 0.9$

☞ Now H_a has $>$. So the hypothesis is one-tailed

Example-4: We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

- ▶ $H_0: \mu \geq 5$
- ▶ $H_a: \mu < 5$

☞ Since H_a has $<$, the hypothesis is one-tailed



☞ Setting up appropriate hypotheses is very crucial in deciding whether the hypothesis is a one-tailed or two-tailed one

☞ This identification will in turn help us in choosing the appropriate test to be conducted!

Probability and Statistics

Lecture-33

Recall

- ☞ A hypothesis test involves collecting sample data and evaluating it
 - ☞ Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis
 - ☞ The actual test begins by considering two hypotheses
 - ☞ They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints
- H_0 : **The null hypothesis:** It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt
- H_a : **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0

Recall

- ☞ Since the null and alternative hypotheses are contradictory, we must examine evidence to decide if we have enough evidence to reject the null hypothesis or not
- ☞ After we have determined which hypothesis the sample supports, we make a decision
- ☞ There are two options for a decision
- ☞ They are “**reject H_0** ” if the sample information favors the alternative hypothesis or “**do not reject H_0** ” or “**decline to reject H_0** ” if the sample information is insufficient to reject the null hypothesis
- ☞ The null is not rejected unless the hypothesis test shows otherwise

☞ One way of testing a null hypothesis is by using confidence intervals

Example: Colleges frequently provide estimates of student expenses such as housing. A consultant hired by a community college claimed that the average student housing expense was \$650 per month. The community college decides to collect data to evaluate the \$650 per month claim. They take a random sample of 175 students at their school and find that sample mean is \$616.91 and sample SD is \$128.65. Does the sample support the \$650 per month claim?

☞ **Step-1:** Set up the null and alternative hypothesis

- ▶ $H_0: \mu = 650$
- ▶ $H_a: \mu \neq 650$

where μ is the population mean

Step-2: Compute a confidence interval for the population mean

- ▶ Since no confidence level is mentioned, in such case, we always take the 95% confidence level
- ▶ We are given $\bar{x} = 616.91$, $s = 128.65$ and $n = 175$
- ▶ Since the sample size is large enough, we use the normal approximation (via CLT) to find the 95% confidence interval
- ▶ As σ is unknown, $\frac{s}{\sqrt{n}} = \frac{128.65}{\sqrt{175}} \approx 9.725$
- ▶ $\alpha = 95 \implies z_{\alpha} = 1.96$. Thus,

$$ME = 1.96 \times 9.725 \approx 19.061$$

- ▶ Thus, a 95% confidence interval for the population mean is
$$(597.849, 635.971)$$

Step-3: Take a decision

- ▶ Because the null value \$650 is not in the confidence interval, a true mean of \$650 is implausible and hence we reject H_0
- ▶ The data provide statistically significant evidence that the actual average housing expense is less than \$650 per month



Example: Abdullah is a quality control expert at a factory that paints car parts. He knows that 20% of parts have an error in their painting. He recommended a change in the painting process, and he wanted to see if this error rate had changed. He took a random sample of 400 parts painted in a month and found that 60 had an error. At 95% confidence level, does the sample give evidence of change in error rate?

☞ **Step-1:** Set up the null and alternative hypotheses

- ▶ $H_0 : p = 0.20$
- ▶ $H_a : p \neq 0.20$

where p is the proportion of parts painted at the factory that have error in their painting

☞ **Step-2:** Compute a confidence interval for the population proportion p

- ▶ We have $n = 400$, $\hat{p} = 0.15$ and hence $n\hat{p} > 5$ and $n(1 - \hat{p}) > 5$

- ▶ $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.0179$
- ▶ $\alpha = 95 \implies z_\alpha = 1.96$
- ▶ $ME = 1.96 \times 0.0179 = 0.035084$
- ▶ Thus, a 95% confidence interval for the population proportion p is

$$(0.1149, 0.1851)$$

☞ Step-3: Take a decision

- ▶ Since the null value 0.2 is not falling in the confidence interval, we reject the null hypothesis
- ▶ Thus, the sample gives sufficient statistical evidence in favour of change in the error rate □

- ☞ A better or popular method of testing hypothesis is by using '**P-values**'
- ☞ This way of testing hypothesis using P-value is usually termed as '**tests of significance**'
- ☞ There is a subtle difference between "**tests of hypothesis**" and "**tests of significance**" which we can ignore for our purposes
- ☞ Every significance test starts by fixing a '**significance level**' β in $(0, 1)$
- ☞ Significance level $\beta = 1 - \frac{\alpha}{100}$, α being the confidence level
- ☞ Default significance level is 0.05 (with the default confidence level 95%)
- ☞ The next step in a significance test is to determining the type of test - **one-tailed** or **two-tailed**
- ☞ We call a test to be **one-tailed** or **two-tailed** if the hypothesis is one-tailed or two-tailed respectively

Example-1: Carlos is designing a video game, and he's concerned that one of the levels is too difficult. He will redesign the level if he has convincing evidence that it takes the average player longer than 45 minutes to complete the level. He plans on recruiting players to play the level until they complete it, and he'll use their completion times to perform a significance test.

- ▶ $H_0: \mu \leq 45$
- ▶ $H_a: \mu > 45$

Then the test would be **one-tailed**.

Example-2: Amanda read a report saying that 49% of teachers in the United States were members of a labour union. She wants to test whether this holds true for teachers in her state, so she is going to take a random sample of these teachers and see what percent of them are members of a union. Let p represent the proportion of teachers in her state that are members of a union.

- ▶ $H_0: p = 0.49$
- ▶ $H_a: p \neq 0.49$

Then the test would be **two-tailed**.

- ☞ We primarily learn to conduct tests of significance to test hypothesis involving population mean and population proportion
- ☞ In particular, significance tests to test hypotheses with the statements pertaining to 'single' mean and 'single' proportion are termed as **tests of significance for single population mean** and **tests of significance for single population proportion**
- ☞ We shall start with an example of **test of significance for single population mean**

Example: A local pizza store knows the mean amount of time it takes them to deliver an order is 45 minutes after the order is placed. The manager has a new system for processing delivery orders, and they want to test if it changes the mean delivery time. They take a random sample of 15 delivery orders and find their mean delivery time is 48 minutes with a sample standard deviation of 10 minutes.

☞ First thing to note from the problem is that the manager wishes to test a claim about “**mean**” delivery time

☞ Once that is noted, we need to setup the hypotheses

- ▶ $H_0 : \mu = 45$
- ▶ $H_a : \mu \neq 45$

☞ We are given that for a sample of size 15, sample mean was $\bar{x} = 48$ and sample SD was $s = 10$

☞ P-value is “the probability that sample mean is more extreme to population mean than \bar{x} given that the null hypothesis is true”. That is,

$$\text{P-value} = P \left\{ \bar{X} \text{ is more extreme to } \mu \text{ than } \bar{x} \mid \mu = 45 \right\}$$

☞ What does “more extreme to μ than \bar{x} ” mean?

☞ We are assuming that H_0 is true. Hence the population mean is $\mu = 45$

☞ More extreme to μ than $\bar{x} = 48$ means that “our sample mean \bar{X} should be more farther from the population mean $\mu = 45$ than $\bar{x} = 48$ ”.

☞ Farther in which direction?

☞ This is where importance of one-tailed or two-tailed comes into play

☞ Since our alternative hypothesis is two-tailed, “farther” in our case can be on both the sides

$$\begin{aligned}\text{Thus, P-value} &= P \left\{ \left\{ \bar{X} - \mu > \bar{x} - \mu \right\} \cup \left\{ \bar{X} - \mu < -(\bar{x} - \mu) \right\} \right\} \\ &= P \left\{ \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} > \frac{\bar{x} - \mu}{SE(\bar{X})} \right\} \cup \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} < -\frac{\bar{x} - \mu}{SE(\bar{X})} \right\} \right\} \\ &= P \left\{ \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} > \frac{48 - 45}{SE(\bar{X})} \right\} \cup \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} < -\frac{48 - 45}{SE(\bar{X})} \right\} \right\} \\ &= P \left\{ \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} > \frac{3}{SE(\bar{X})} \right\} \cup \left\{ \frac{\bar{X} - \mu}{SE(\bar{X})} < -\frac{3}{SE(\bar{X})} \right\} \right\} \\ &= P \left\{ \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{3}{\sigma/\sqrt{n}} \right\} \cup \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -\frac{3}{\sigma/\sqrt{n}} \right\} \right\}\end{aligned}$$

as $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation

☞ The problem here is that we do not know the population SD σ

☞ An estimate for the population SD would be s

☞ We have $\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{3}{2.582} = 1.162$

☞ Hence P-value will be

$$P \left\{ \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} < -1.162 \right\} \cup \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} > 1.162 \right\} \right\}$$

☞ Since population sd is unknown, the distribution of $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ will be either $N(0, 1)$ or Student's t-distribution with $(n - 1)$ degrees of freedom depending on the sample size n

☞ It turns out that for large sample size (hence degrees of freedom), t-scores will be close to z-scores

☞ Hence, instead of going for different approaches for different sample sizes, we will be using t-distribution irrespective of sample size

☞ That is, $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ follows Student's t-distribution with df $\nu = 14$

☞ Thus,

$$\begin{aligned}\text{P-value} &= P\left\{\left\{\frac{\bar{X} - \mu}{s/\sqrt{n}} < -1.162\right\} \cup \left\{\frac{\bar{X} - \mu}{s/\sqrt{n}} > 1.162\right\}\right\} \\ &= 2F_{14}(-1.162) \\ &= 2 \times 0.1323 = 0.2646\end{aligned}$$

☞ **Decision:**

- ▶ if P-value $< \beta$ (significance level), we reject the null hypothesis in favour of alternative hypothesis
- ▶ If P-value $\geq \beta$, we do not reject null hypothesis or we say that sample do not provide enough evidence against the null hypothesis

- ☞ In our example, since no significance level is specified, we take $\alpha = 0.05$ and compare it with the P-value 0.2646
- ☞ Since $0.2646 > 0.05$, we do not reject the null hypothesis and conclude that the sample does not provide enough evidence to say that the average delivery time is different from 45 minutes □

Example: Fernanda runs a large bowling league. She suspects that the league average score is greater than 150 per game. She takes a random sample of 36 game scores from the league data. The scores in the sample have a mean of 156 and a standard deviation of 30. Conduct a test of significance with $\beta = 0.10$ to decide on Fernanda's suspicion.

☞ The hypotheses would be:

- ▶ $H_0 : \mu \leq 150$
- ▶ $H_a : \mu > 150$

☞ Here, since there is $>$ in the alternative hypothesis, the test is going to be a one-tailed

☞ The only difference in here is the calculation of P-value

☞ “more extreme” in this case would be in only one direction!

☞ Since we are interested in only $>$, we should look at extreme values above mean

- ☞ Further, recall that P-value is “the probability that sample mean is more extreme to population mean than \bar{x} given that the null hypothesis is true”
- ☞ Since the null and alternative hypothesis are contradictory to each other, and we are looking for extremity in the ‘direction’ of alternative hypothesis, in one-tailed tests, we assume μ to be the ‘border’ value from null hypothesis. That is, $\mu = 150$

- ☞ Hence the P-value in this case is

$$P \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} > \frac{\bar{x} - \mu}{s/\sqrt{n}} \right\} = P \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} > \frac{156 - 150}{30/\sqrt{36}} \right\} = P \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} > 1.2 \right\}$$

- ☞ Now we have $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ following Student’s t-distribution with df $\nu = 35$

☞ Thus,

$$\text{P-value} = P \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} > 1.2 \right\} = P \left\{ \frac{\bar{X} - \mu}{s/\sqrt{n}} < -1.2 \right\} = F_{35}(-1.2) = 0.1191$$

☞ As P-value $0.1191 > \beta = 0.10$, we do not reject the null hypothesis!

☞ That is at 0.10 significance level, the sample data does not provide enough evidence against the null hypothesis of the league average score being 150 per game



Test of significance for single population mean

☞ We will be given a sample of size n (hence its mean \bar{x} and SD s), and significance level β (default is 0.05)

Step-1: Setup the hypotheses and note down if the test is one-tailed or two-tailed

Step-2: Compute $\frac{s}{\sqrt{n}}$

Step-3: Note down the value of μ from null hypothesis and compute the '**test statistic**' $t = \left| \frac{\bar{x} - \mu}{s/\sqrt{n}} \right|$

Step-4: Compute P-value as below:

- ▶ For two-tailed test, P-value = $2F_{n-1}(-t)$
- ▶ For one-tailed test, P-value = $F_{n-1}(-t)$

where F_{n-1} is the distribution function of a Student's t-distributed random variable with $df = n - 1$

Step-5: Take a decision

- ▶ If $P\text{-value} < \beta$ (significance level), we reject the null hypothesis
- ▶ If $P\text{-value} \geq \beta$, we do not reject null hypothesis

★ Since we are using the Student's t-distribution as the '*reference distribution*' in the above test, the test is usually referred to as '*t-test for single population mean*'

Probability and Statistics

Lecture-34

Test of significance for single population mean

☞ We will be given a sample of size n (hence its mean \bar{x} and SD s), and significance level β (default is 0.05)

Step-1: Setup the hypotheses and note down if the test is one-tailed or two-tailed

Step-2: Compute $\frac{s}{\sqrt{n}}$

Step-3: Note down the value of μ from null hypothesis and compute the '**test statistic**' $t = \left| \frac{\bar{x} - \mu}{SE} \right|$

Step-4: Compute P-value as below:

- ▶ For two-tailed test, P-value = $2F_{n-1}(-t)$
- ▶ For one-tailed test, P-value = $F_{n-1}(-t)$

where F_{n-1} is the distribution function of a Student's t-distributed random variable with $df = n - 1$

Step-5: Take a decision

- ▶ If $P\text{-value} < \beta$ (significance level), we reject the null hypothesis
- ▶ If $P\text{-value} \geq \beta$, we do not reject null hypothesis

★ Since we are using the Student's t-distribution as the '*reference distribution*' in the above test, the test is usually referred to as '*t-test for single population mean*'

☞ We follow a similar approach if we wish to test hypothesis involving single population proportion

Example: Abdullah is a quality control expert at a factory that paints car parts. He knows that 20% of parts have an error in their painting. He recommended a change in the painting process, and he wanted to see if this error rate had changed. He took a random sample of 400 parts painted in a month and found that 60 had an error. At 0.05 significance, does the sample give evidence of change in error rate?

☞ Observe that, here we are interested in the “proportion” of parts that have an error in painting

☞ We are trying to test claim about this proportion

☞ Define p to be the proportion of parts with errors in the **population**

☞ The hypotheses would be

- ▶ $H_0 : p = 0.20$
- ▶ $H_a : p \neq 0.20$

☞ Since the alternative hypothesis has \neq , we are in the context of two-tailed test

☞ We are given the observed sample proportion $\hat{p} = \frac{60}{400} = 0.15$

☞ P-value is “the probability of \hat{P} (random variable) being more extreme to p than the observed sample proportion $\hat{p} = 0.15$ assuming $p = 0.20$ ”

☞ By exactly same approach as in the mean case, we end up in

$$\text{P-value} = P \left\{ \left\{ \frac{\hat{P} - p}{SE(\hat{P})} < -\frac{p - \hat{p}}{SE(\hat{P})} \right\} \cup \left\{ \frac{\hat{P} - p}{SE(\hat{P})} > \frac{p - \hat{p}}{SE(\hat{P})} \right\} \right\}$$

☞ **Recall:** $SE(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$

☞ Hence, in our example, $SE(\hat{P}) = \sqrt{\frac{0.2(1-0.2)}{400}} = 0.02$

☞ Thus,

$$\begin{aligned} \text{P-value} &= P \left\{ \left\{ \frac{\hat{P} - p}{SE(\hat{P})} < -\frac{0.20 - 15}{0.02} \right\} \cup \left\{ \frac{\hat{P} - p}{SE(\hat{P})} > \frac{0.20 - 15}{0.02} \right\} \right\} \\ &= P \left\{ \left\{ \frac{\hat{P} - p}{SE(\hat{P})} < -2.5 \right\} \cup \left\{ \frac{\hat{P} - p}{SE(\hat{P})} > 2.5 \right\} \right\} \end{aligned}$$

☞ **Recall:** If $np > 5$ and $n(1 - p) > 5$, then $\frac{\hat{P} - p}{SE(\hat{P})} \sim N(0, 1)$

☞ In our example, $n = 400$ and $p = 0.20$. Hence both the conditions are satisfied

☞ Thus,

$$\text{P-value} = P \{ \{Z < -2.5\} \cup \{Z > 2.5\} \} = 2\Phi(-2.5) = 0.0124$$

☞ **Decision:** Since P-value $0.0124 < 0.05$ (significance level), we reject the null hypothesis

☞ That is, the sample gives enough (statistical) evidence to say that error rate has changed □

Example: According to a very large poll in 2015, about 90% of homes in California had access to the internet. Market researchers want to test if that proportion is now higher, so they take a random sample of 100 homes in California and find that 96 of them have access to the internet. Conduct a significance test for this proportion with $\beta = 0.01$

☞ Let p be the proportion of homes in California which have internet access

☞ The hypotheses would be:

- ▶ $H_0 : p = 0.9$
- ▶ $H_a : p > 0.9$

☞ Here, since there is $>$ in the alternative hypothesis, the test is going to be a one-tailed

☞ The only difference, again, in here is the calculation of P-value

☞ “more extreme” in this case would be in only one direction!

☞ Since we are interested in only $>$, by following similar procedure as earlier, we get

$$\text{P-value} = P \left\{ \frac{\hat{P} - p}{SE(\hat{P})} > \frac{\hat{p} - p}{SE(\hat{P})} \right\}$$

☞ Since $p = 0.9$ and $n = 100$, $SE(\hat{P}) = \sqrt{\frac{p(1-p)}{n}} = 0.03$

$$\implies \text{P-value} = P \left\{ \frac{\hat{P} - p}{SE(\hat{P})} > 2 \right\}$$

☞ As $np > 5$ and $n(1 - p) > 5$, $\frac{\hat{P} - p}{SE(\hat{P})} \sim N(0, 1)$

☞ Thus, $\text{P-value} = P \{Z > 2\} = 1 - \Phi(2) = 0.0228$

☞ Since $\text{P-value} > 0.01$, we do not reject the null hypothesis □

Test of significance for single population proportion

☞ We will be given a sample of size n , a proportion \hat{p} computed from it, and significance level β (default is 0.05)

Step-1: Setup the hypotheses and note down if the test is one-tailed or two-tailed

Step-2: Note down the value of p from null hypothesis and compute $SE(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$

Step-3: Calculate the '**test statistic**' $z = \left| \frac{\hat{p}-p}{SE(\hat{P})} \right|$

Step-4: Check if $np > 5$ and $n(1 - p) > 5$. Only if both the conditions are satisfied, proceed to step-5

Step-5: Compute the P-value

- ▶ For two-tailed test, $P\text{-value} = 2\Phi(-z)$
- ▶ For one-tailed test, $P\text{-value} = \Phi(-z)$

where Φ is the distribution function of standard normal variable

Step-6: Take a decision

- ▶ If $P\text{-value} < \beta$ (significance level), we reject the null hypothesis
- ▶ If $P\text{-value} \geq \beta$, we do not reject null hypothesis

★ Since we are using the Standard normal distribution (z -distribution) as the '*reference distribution*' in the above test, the test is usually referred to as '*z-test for single population proportion*'

☞ One can see that the following is the common procedure in any significance test

- ▶ **Parameter of interest:** From the problem context, identify the parameter of interest
- ▶ **Null hypothesis H_0 :** State the null hypothesis, H_0
- ▶ **Alternative hypothesis H_a :** Specify an appropriate alternative hypothesis, H_a
- ▶ **Test statistic:** Determine an appropriate test statistic and its distribution
- ▶ **Computations:** Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute the P-value
- ▶ **Draw conclusions:** Decide whether or not H_0 should be rejected and report that in the problem context

☞ Hypothesis tests are not flawless as we can make a wrong decision in statistical hypothesis tests based on the data

☞ There are four possible scenarios, which are summarized in following table:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

☞ A **Type I Error** (false positive) is rejecting the null hypothesis when H_0 is actually true.

☞ A **Type 2 Error** (false negative) is failing to reject the null hypothesis when the alternative is actually true.

- ☞ There are tools in hypothesis testing which quantify these types of errors
- ☞ At a particular significance level β it turns out that

$$P(\text{Type I error}) = \beta$$

- ☞ The probability of **not committing** a Type II error is called **the power of a hypothesis test**
- ☞ That is,

$$\begin{aligned}\text{Power} &= 1 - P(\text{Type II error}) \\ &= 1 - P(\text{not rejecting } H_0 \mid H_0 \text{ is false})\end{aligned}$$

Factors that affect power

- ▶ **Sample size (n)** - Other things being equal, the greater the sample size, the greater the power of the test
- ▶ **Significance level (β)** - The power of the test is reduced when you reduce the significance level; and vice versa
- ▶ **The “true” value of the parameter being tested.** The greater the difference between the “true” value of a parameter and the value specified in the null hypothesis, the greater the power of the test

- ☞ In tests of significance, if the P-value is less than the significance level (typically 0.05), then we conclude that results are statistically significant (to reject the null hypothesis)
- ☞ Results are said to be **statistically significant** when the difference between the hypothesized population parameter and observed sample statistic is large enough to conclude (by P-value) that it is unlikely to have occurred by chance
- ☞ **Practical significance** refers to the magnitude of the difference, which is known as the **effect size**.
- ☞ Results are practically significant when the difference is large enough to be meaningful in real life. What is meaningful may be subjective and may depend on the context
- ☞ **This is especially important to research:** If we conduct a study, we want to focus on finding a meaningful result. We do not want to spend lots of money finding results that hold no practical value!

Probability and Statistics

Lecture-35

- ☞ Till now we were testing statements about single population mean or proportion
- ☞ Let us consider the following scenario
- ☞ The safety of drinking water is a serious public health issue
- ☞ An article reported on arsenic contamination in the water sampled from 10 communities in a metropolitan city area and 10 communities from a rural area in the same state
- ☞ The data showed dramatic differences in the arsenic concentration, ranging from 3 parts per billion (ppb) to 48 ppb
- ☞ Some natural questions arise from this article

- ☞ Is the difference in the arsenic concentrations in the metropolitan city and in the rural communities **real**?
- ☞ How large is this difference? Is it large enough to require action on the part of the public health service and other state agencies to correct the problem?
- ☞ Are the levels of reported arsenic concentration large enough to constitute a public health risk?
- ☞ Some of these questions can be answered by statistical methods
- ☞ Think of the metropolitan communities as one population and the rural communities as a second population
- ☞ Then we should determine whether a statistically significant difference in the mean arsenic concentration exists for the two populations by testing the hypothesis $\mu_1 \neq \mu_2$, where μ_1 and μ_2 are means of the two populations
- ☞ This is a relatively simple extension to two samples of the one-sample hypothesis testing procedures

★ We only go through tests of significance for difference in means of two populations under the assumption that **both the populations are normal and independent of each other**

☞ We are interested in testing hypothesis about difference in means of two populations

☞ Instead of testing statements like $\mu_1 = \mu_2$, $\mu_1 < \mu_2$ etc., we consider more general statements like $\mu_1 - \mu_2 = \Delta_0$, $\mu_1 - \mu_2 < \Delta_0$ etc., for a fixed number Δ_0

☞ As done in single mean context, we divide our problem into two cases:

Case-1: Population variances σ_1 and σ_2 are known

Case-2: Population variances σ_1 and σ_2 are unknown

Test of significance for difference of means - known population variances

☞ We will be having

- ▶ a significance level $\beta \in (0, 1)$,
- ▶ two samples x_1, x_2, \dots, x_{n_1} (size n_1) and y_1, y_2, \dots, y_{n_2} (size n_2) from two different populations
- ▶ population variances σ_1^2 and σ_2^2
- ▶ $H_0 : \mu_1 - \mu_2 = \Delta_0$ or \leq or $\geq \Delta_0$

☞ **Step-1:** Compute the sample means \bar{x} , \bar{y} and $\frac{\sigma_1^2}{n_1}$, $\frac{\sigma_2^2}{n_2}$

☞ **Step-2:** Compute the '**test statistic**'

$$z = \left| \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right|$$

Step-3: Compute the P-value

- ▶ For two-tailed test, $P\text{-value} = 2\Phi(-z)$
- ▶ For one-tailed test, $P\text{-value} = \Phi(-z)$

where Φ is the distribution function of standard normal variable

Step-4: Take a decision

- ▶ If $P\text{-value} < \beta$ (significance level), we reject the null hypothesis
- ▶ If $P\text{-value} \geq \beta$, we do not reject null hypothesis

Example:

- ▶ A product developer is interested in reducing the drying time of a primer paint
- ▶ Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time
- ▶ From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient
- ▶ Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order
- ▶ The two sample average drying times are $\bar{x} = 121$ minutes and $\bar{y} = 112$ minutes, respectively.

What conclusions can the product developer draw about the effectiveness of the new ingredient with a significance level of 5%?

- ▶ The quantity of interest is the difference in mean drying times, $\mu_1 - \mu_2$
- ▶ Setup null and alternative hypotheses
 - ▶ $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$ (i.e., $\Delta_0 = 0$)
 - ▶ $H_a : \mu_1 > \mu_2$
- ▶ We are given $n_1 = n_2 = 10$, $\bar{x} = 121$, $\bar{y} = 112$, $\sigma_1 = \sigma_2 = 8$ and $\beta = 0.05$
- ▶ $\frac{\sigma_1^2}{n_1} = \frac{\sigma_2^2}{n_2} = 6.4$
- ▶ ‘Test statistic’

$$z = \left| \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| = \left| \frac{121 - 112 - 0}{\sqrt{6.4 + 6.4}} \right| \approx 2.52$$

- ▶ Since the test is one-tailed, $P\text{-value} = \Phi(-2.52) = 0.0059$
- ▶ Since $P\text{-value} < \beta = 0.05$, we reject null hypothesis and conclude that adding the new ingredient to the paint significantly reduces the drying time



Test of significance for difference of means - unknown but equal population variances

☞ We will be having

- ▶ a significance level $\beta \in (0, 1)$,
- ▶ two samples x_1, x_2, \dots, x_{n_1} (size n_1) and y_1, y_2, \dots, y_{n_2} (size n_2) from two different populations
- ▶ $H_0 : \mu_1 - \mu_2 = \text{ or } \leq \text{ or } \geq \Delta_0$

☞ **Step-1:** Compute the sample means \bar{x} , \bar{y} , sample variances s_1^2 , s_2^2 and $\frac{s_1^2}{n_1}$, $\frac{s_2^2}{n_2}$

☞ **Step-2:** Compute the '**pooled estimator**' for $\sigma^2 = \sigma_1^2 = \sigma_2^2$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

☞ Step-3: Compute the 'test statistic'

$$t = \left| \frac{\bar{x} - \bar{y} - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|$$

☞ Step-4: Compute the P-value

- ▶ For two-tailed test, $P\text{-value} = 2F_{n_1+n_2-2}(-t)$
- ▶ For one-tailed test, $P\text{-value} = F_{n_1+n_2-2}(-t)$

where $F_{n_1+n_2-2}$ is the distribution function of Student's t-distributed random variable with n_1+n_2-2 degrees of freedom

☞ Step-5: Take a decision

- ▶ If $P\text{-value} < \beta$ (significance level), we reject the null hypothesis
- ▶ If $P\text{-value} \geq \beta$, we do not reject null hypothesis

Example:

- ▶ Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process
- ▶ Specifically, catalyst 1 is currently used; but catalyst 2 is acceptable
- ▶ Because catalyst 2 is cheaper, it should be adopted, if it does not change the process yield
- ▶ A test is run in the pilot plant and results in the data is shown below

Observation Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75

Is there any difference in the mean yields? Use $\beta = 0.05$, and assume equal variances.

- ▶ The parameters of interest are μ_1 and μ_2 , the mean process yield using catalysts 1 and 2, respectively, and we want to know if $\mu_1 - \mu_2 = 0$
- ▶ Setup null and alternative hypotheses
 - ▶ $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$ (i.e., $\Delta_0 = 0$)
 - ▶ $H_a : \mu_1 \neq \mu_2$
- ▶ We are given $n_1 = n_2 = 8$, two samples and $\beta = 0.05$

Observation Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75

- ▶ From the above data, $\bar{x} = 92.255$, $\bar{y} = 92.733$, $s_1 = 2.39$ and $s_2 = 2.98$

► 'Pooled estimate'

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 7.30$$

$$s_p = 2.7$$

► 'Test statistic'

$$t = \left| \frac{\bar{x} - \bar{y} - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| = \left| \frac{92.255 - 92.733 - 0}{2.7 \sqrt{1/8 + 1/8}} \right| \approx 0.35$$

- As the test is two-tailed,

$$\text{P-value} = 2F_{n_1+n_2-2}(-0.35) = 2F_{14}(-0.35) = 0.7316$$

- Since P-value > $\beta = 0.05$, we do not reject null hypothesis and conclude that we do not have strong evidence to conclude that catalyst 2 results in a mean yield that differs from the mean yield when catalyst 1 is used

Test of significance for difference of means - unknown and unequal population variances

☞ We will be having

- ▶ a significance level $\beta \in (0, 1)$,
- ▶ two samples x_1, x_2, \dots, x_{n_1} (size n_1) and y_1, y_2, \dots, y_{n_2} (size n_2) from two different populations
- ▶ $H_0 : \mu_1 - \mu_2 = \Delta_0$ or \leq or $\geq \Delta_0$

☞ **Step-1:** Compute the sample means \bar{x} , \bar{y} , sample variances s_1^2 , s_2^2 and $\frac{s_1^2}{n_1}$, $\frac{s_2^2}{n_2}$

☞ **Step-2:** Compute the '**test statistic**'

$$t = \left| \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right|$$

☞ Step-3: Compute the P-value

- ▶ For two-tailed test, $P\text{-value} = 2F_\nu(-t)$
- ▶ For one-tailed test, $P\text{-value} = F_\nu(-t)$

where F_ν is the distribution function of Student's t-distributed random variable with ν degrees of freedom with ν being the nearest integer to

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

☞ Step-4: Take a decision

- ▶ If $P\text{-value} < \beta$ (significance level), we reject the null hypothesis
- ▶ If $P\text{-value} \geq \beta$, we do not reject null hypothesis



Example: Arsenic concentration in public drinking water supplies is a potential health risk. An article reported drinking water arsenic concentrations in parts per billion (ppb) for 10 communities in Vijayawada and 10 communities in Guntur

Vijayawada	3	7	25	10	15	6	12	25	15	7
Guntur	48	44	40	38	33	21	20	12	1	18

We wish to determine whether any difference exists in mean arsenic concentrations for Vijayawada communities and for communities in Guntur.

- The parameters of interest are the mean arsenic concentrations for the two geographic regions, say, μ_1 and μ_2 , and we are interested in determining whether $\mu_1 - \mu_2 = 0$
- Setup null and alternative hypotheses
 - $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$ (i.e., $\Delta_0 = 0$)
 - $H_a : \mu_1 \neq \mu_2$

Vijayawada	3	7	25	10	15	6	12	25	15	7
Guntur	48	44	40	38	33	21	20	12	1	18

- From the above data, $\bar{x} = 12.5$, $\bar{y} = 27.5$, $s_1 = 7.63$ and $s_2 = 15.3$
- 'Test statistic'**

$$t = \left| \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right| = \left| \frac{12.5 - 27.5 - 0}{\sqrt{\frac{(7.63)^2}{10} + \frac{(15.3)^2}{10}}} \right| \approx 2.77$$

- Degrees of freedom ν is the nearest integer to

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} = \frac{\left(\frac{(7.63)^2}{10} + \frac{(15.3)^2}{10} \right)^2}{\frac{((7.63)^2/10)^2}{9} + \frac{((15.3)^2/10)^2}{9}} \approx 13.2$$

- Thus, $\nu = 13$
- As the test is two-tailed,

$$\text{P-value} = 2F_{13}(-2.77) = 0.0080$$

- Since no significance level is specified, we take $\beta = 0.05$
- As P-value $< \beta$, we reject null hypothesis and conclude that mean arsenic concentration in the drinking water in Guntur is different from the mean arsenic concentration Vijayawada drinking water

