Data collection, Analysis and Inference

Subject Code: CPE-RPE,

May 2021,
SRM Univeristy-AP, Andhrapradesh

Lecture- 1: Principle of Counting, Probability and conditional probability

- Aim: To be able to compute probabilities and distinguish independent events.

- **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.

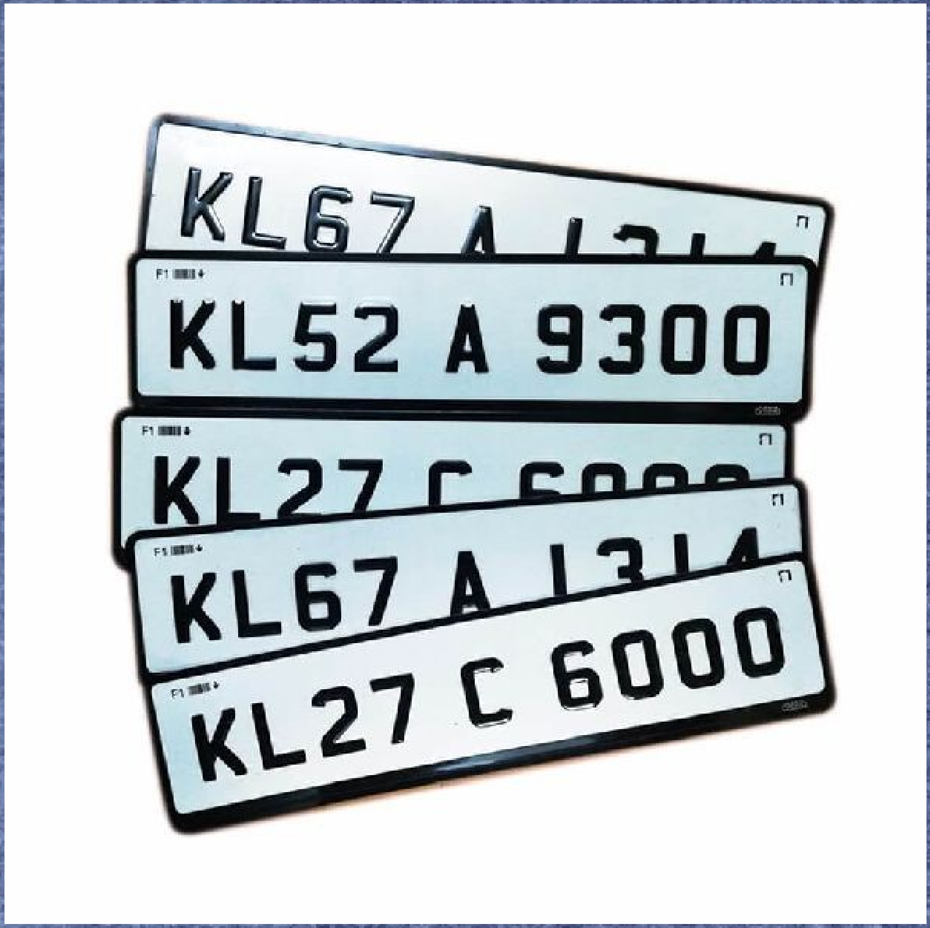
- Example: A small community consists of 10 women, each of whom has 3 children. If one woman and one of her children are to be chosen as mother and child of the year, how many different choices are possible?
- Regard the choice of woman as outcome of the first experiment and the subsequent choice of 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:(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 put alphabets in first 2 places and the second experiment is to put 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 $26 \times 25 = (26\ p\ 2)$
- Similarly, for second experiment, the number of outcomes is $10 \times 9 \times 8 \times 7 \times 6 = (?\ p\ ?)$
- Thus, total number of possible number plates without allowing repetitions is $26 \times 25 \times 10 \times 9 \times 8 \times 7 \times 6 = (26\ p\ 2) \times (?\ p\ ?)$

Count all the 4-digit lucky numbers ?



A number is considered lucky if its iterative sum terminates with single digit 9.
Ex: 9693 which leads to 9+6+9+3=27 and further again it turns into 2+7=9.

Sample Space

- The set of all possible outcomes of an experiment is called sample space of the experiment and is denoted by S.
- all possible outcomes ↔ sample space(S).
- To calculate probabilities, the first step is to list out all possible outcomes of the experiment.
- Given an experiment, determine the event space E , which is the set of favourable outcomes.
- E is always a subset of S

- Example - Rolling pair of dice simultaneously.
- Sample Space $S = \{(i, j) : i, j = 1, 2, 3, 4, 5, 6\}$
- E1 be the event that sum of dice equals 7 and E2 be the event that the outcome is (1, 5)
- That is, $E1 = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ and $E2 = \{(1, 5)\}$
- Then E2 is a simple event and E1 is NOT a simple event

- For any two events E and F (of the same experiment), we define the new event $E \cup F$ to consist of all the outcomes that are either in E or F.
- 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.

- 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 = {(T , H), (H, H), (H, T)}
- Then, E ∪ 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.
- Then E ∩ F = {(H,T), (T , H)} is the event that exactly one head and one tail occur

- **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)

Definition of probability

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

The Classical definition of probability

- In this definition we assume that every outcome in the sample space has an equal chance of occurrence. (equally likely)

- Under this assumption, for any event E ,

$$P(E) = |E| / |S|$$

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

The relative frequency definition of probability

- 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 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

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

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.**

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 $E_1, E_2, \ldots$

  (that is, events for which $E_i \cap E_j = \varnothing$ whenever $i \neq j$),
-
$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Some results

- For any event E , P(E c) = 1 − P(E)
- P(∅) = 0
- For any two events E and F ,

  P(E ∪ F) = P(E) + P(F) − P(E ∩ F)
- **Inclusion-exclusion identity:**

For any set of n events, $E_1, E_2, \ldots E_n$,

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

- **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?

Conditional Probability

- 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 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)) = 16 for each j = 1, 2, 3, 4, 5, 6
- Our desired outcome is (3, 5) and hence the probability is 1/6
- If the condition was not given, then the probability of getting a sum of 8 was 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.**

Definition

- 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**

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	212	198
More than \$50,000	36	54

- 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?

- 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
- Given the information $P(R) = 0.4$ and $P(L) = 0.4$.
- The desired probability is $P(R \mid L^c)$

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, $P(R|L^c) == \frac{P(R)}{1-P(L)} = \frac{0.4}{1-0.4} = \frac{2}{3}$

Independent events

- $P(E \mid F)$ is in general not equal to $P(E)$
- The occurrence of F may affect the occurrence of E

- When does $P(E \mid F) = P(E)$?
- When does occurrence of F has no effect on occurrence of E ?
- We say that the events E and F are independent if

$$\mathbf{P(E \mid F) = P(E)}$$

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.

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

Fact :: **If E and F are independent, then so are E and F c**

- 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!

- Counter 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.
- Hint: Compute $P(E \cap (F \cap G))$ and $P(E)P(F \cap G)$

Bayes's Rule to compute reverse conditional probability

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

- END