Theory Session | Statistical Inference

Tamojit Maiti
Masters in Applied Statistics and Operations Research, ISI Kolkata
Data Scientist at Sixt R&D, previously at Rapido & AB InBev

Agenda

- Probability
  - Definitions
  - Exercises
- Random Variables
  - Definitions
  - Exercises
- Probability Distributions
  - Uniform
  - Binomial
  - Poisson
  - Exponential
  - Normal
  - Exercises
- Central Limit Theorem
  - Exercises

Probability

Probability | Definitions

- An experiment is something that be infinitely repeated and has a well-defined set of possible outcomes, called the sample space.

- Event is a subset of the sample space.
  - Experiment: Rolling a die
  - Sample Space, S = {1, 2, 3, 4, 5, 6}
  - Event: Getting an odd number, A = {1, 3, 5}

- Probability is a numerical way of describing how likely (or not) an event is to happen.
  - If each of the elements in the sample space S are equally likely, then we can define the probability of event A as $P(A) = \frac{n(A)}{n(S)}$
  - The probability of any event is between 0 and 1 inclusive

Probability | Axioms

- The probability of a certain event is 1
- The probability of an impossible event is zero, and probabilities cannot be negative
- For any two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- A conditional probability is the probability of an event, given some other event has already occurred.
  - It is denoted by $P(A|B)$ and read as probability of A given B
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- For independent events A and B, $P(A \cap B) = P(A) \times P(B)$

Probability | Exercises

- In a simultaneous toss of two fair coins, what is the probability that you get two heads?

- In a simultaneous roll of two fair dice, what is the probability of getting a sum of greater than or equal to 9?

- In a group of 25 people, 18 have a mortgage, 13 own some shares and 2 people have neither a mortgage nor any shares. You select a person at random. What is the probability that they have both mortgage and shares?

Random Variables

Random Variables | Definitions

- A random variable
  - is a variable whose possible values are numerical outcomes of a random phenomenon
  - is a mapping of each element in the sample space to the set of real numbers

- Example: Discrete
  - Experiment: Tossing a coin
  - Sample space: S = {H, T}.
  - Random Variable: X is the outcome of the coin toss
  - X(H) = 1 and X(T) = 0
- Example: Continuous
  - Experiment: Measuring the temperature during the day
  - Sample space: S = $[0, \infty]$ K
  - Random Variable: T is the temperature in Kelvins

Random Variables | Definitions

- An event is when a random variable equals one/more values in the sample space

- Example: Rolling a fair dice
  - Random Variable: Number on the dice face, denoted by X
  - Event: The dice face shows 3
  - Probability of the event P(X = 3) = 1/6

- The probability distribution function is a mapping from the set of values of the random variable to the set [0,1]. It describes how the probabilities are 'distributed' across the values of the random variables

- It is defined as $f(x) = P(X = x)$ and has the following properties
  - $f(x) \geq 0 \; \forall \, x \in X$
  - $\sum_X f(x) = 1$

Random Variables | Definitions

- The probability distribution function of a discrete random variable is a mapping from the set of values of the random variable to the set [0,1]. It describes how the probabilities are 'distributed' across the values of the random variables

- It is defined as $f(x) = P(X = x)$ and has the following properties
  - $f(x) \geq 0 \ \forall \, x \in X$
  - $\sum_X f(x) = 1$

- The cumulative distribution function of a discrete random variable is also a mapping from the set of values of the random variable to the set $[0,1]$. However, it describes the probability of X not exceeding some value x

- It is defined as $F(x) = P(X \leq x)$

Random Variables | Exercises

- Define suitable random variables and write their corresponding probability distribution function and the cumulative distribution function for the following experiments:

    - Tossing a fair coin

    - Rolling a fair dice

Random Variables | Definitions

- The probability distribution function for a continuous random variable is always zero

$$P(X = x) = 0 \; \forall \; x \in X$$

- The probability density function of a continuous random variable is defined as

$$P(a < X < b) = \int_a^b f_X(x)dx$$

- The cumulative density function for a continuous random variable is defined as

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

Random Variables | Exercises

- Given a PDF for a random variable W., perform the following exercises

$$f_W(w) = 12w^2(1-w) \forall w \in [0,1]$$

  - Check that the PDF is non-negative for all values of w

  - Calculate the CDF

Random Variables | Definitions

- The mean or expected value of a random variable X refers to the central location of the probability distribution function of X

- It is defined as
  - $E(X) = \sum_X x \times f_X(x)$ for discrete random variables
  - $E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$ for continuous random variables

- The variance of a random variable X refers to the spread of the probability distribution function of X

- It is defined as
  - $Var(X) = E[X^2] - (E[X])^2$

Probability Distributions

Probability Distributions | Uniform

- Probability All outcomes are equally likely
- Random variable $X = \{1,2,3, \dots k\}$
- The PDF

$$f_X(x) = \frac{1}{k} \; \forall \; x \in \{1,2,3,\cdots,k\}$$

- The expected value

$$E(X) = \sum_{x=1}^{k} x\left(\frac{1}{k}\right) = \frac{1}{k}\sum_{x=1}^{k} x = \frac{1}{k} \times \frac{(k)(k+1)}{2} = \frac{k+1}{2}$$

- The variance

$$var(X) = \sum_{x=1}^{k} x^2\left(\frac{1}{k}\right) = \frac{1}{k}\sum_{x=1}^{k} x^2 = \frac{1}{k} \times \frac{(k)(k+1)(2k+1)}{6} = \frac{(k+1)(2k+1)}{6}$$

Probability Distributions | Bernoulli

- Probability Two possible outcomes, either a success or a failure

- Random variable $X = \{0,1\}$

- The PDF

$$f_X(x) = p^x(1-p)^{1-x} \; \forall \, x \in \{0,1\}, p \in [0,1]$$

- The expected value

$$E(X) = p$$

- The variance

$$var(X) = p(1-p)$$

Probability Distributions | Binomial

- Probability Outcome of the number of successes out of a sequence of N independent and identically distributed Bernoulli trials, each of which as a chance of success p

- Random variable $X = \{0,1,2,3,\cdots,N\}$

- The PDF

$$f_X(x) = \binom{N}{x} p^x (1-p)^{N-x} \;\forall\, x \in \{0,1,2,\cdots,N\}, p \in [0,1]$$

- The expected value

$$E(X) = Np$$

- The variance

$$var(X) = Np(1-p)$$

Probability Distributions | Exponential

- Probability Outcome of the time to failure of a certain equipment

- Random variable $X = [0, \infty)$

- The PDF

$$f_X(x) = \lambda e^{-\lambda x} \; \forall \; x \in [0, \infty), \lambda \in [0, \infty)$$

- The expected value

$$E(X) = 1/\lambda$$

- The variance

$$var(X) = 1/\lambda$$

- Probability Outcome of many natural processes
- Random variable $X = (-\infty, \infty)$
- The PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \; \forall\, x \in (-\infty, \infty), \mu \in (-\infty, \infty), \sigma^2 \in [0, \infty)$$

- The expected value

$$E(X) = \mu$$

- The variance

$$var(X) = \sigma^2$$

Probability Distributions | Exercises

- Calculate the expected value and the variance of the random variable described by the number shown when a fair dice is thrown

- Calculate the expected value and the variance of the random variable whose PDF is given by the following expression

$$f_X(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

The Central Limit Theorem

Central Limit Theorem | Definitions

- A population refers to the entire group that we want to draw conclusions about
- But often the population size is very large, so we cannot collect data about the entire population
- In that case, we collect data from a small part of the population
- A sample refers to the subset of the group from which we will collect our data from
- A parameter is a number describing some attribute of the entire population
- A statistic is a number describing some attribute of the sample in question
- The basic aim of statistical inference is to conclude about population parameters using sample statistics

Central Limit Theorem | Statement

- No matter the distribution from which a bunch of random variables are drawn
  - If we sample randomly from the bunch of random variables
  - And calculate the mean of the sample
  - And repeat the above two steps many many times, keeping the sample size same in each draw

- We will see that the sample means form a normal distribution with
  - Expected value equal to the population mean
  - Variance equal to the population variance divided by the sample size

- It means that under certain circumstances, most distributions can be transformed and approximated by a normal distribution
  - $\mathrm{Binomial}(n, p) \sim \mathrm{N}(np, np(1-p))$ when $np > 5$ and $n(1-p) > 5$
  - $\mathrm{Poisson}\ (\lambda) \sim \mathrm{N}(\lambda, \lambda)$ if λ is large

Q & A