Georgian

Statistics and Probability part 1

Garima Malik

January 10, 2022

Outline

Central Tendencies

Probability and its axioms

Random Variables

Probability Distributions

Central Tendencies

Central Tendencies

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.
- The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.
- The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

Mean

- The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data.
- Sample Mean :

$$\bar{x} = \frac{x_1 + x_2 ..... + x_n}{n} \tag{1}$$

- To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lower case letter "mu", denoted as
- population Mean :

$$\mu = \frac{\sum x}{n} \tag{2}$$

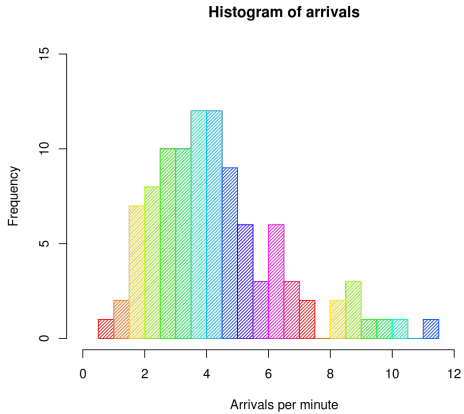
- The mean has one main disadvantage: it is particularly susceptible to the influence of outliers.

Median

- The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data.

Mode

- The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option.



Summary - Central tendencies

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Probability

- Empirical Definition : Given an event A,

$$Pr(A) = \lim_{n \to \infty} \frac{n_A}{n} \tag{3}$$

where n_A is no of times event A is observed and n is the number of trials.

- Classical Definition : Given an event A,

$$Pr(A) = \frac{N_A}{N} \tag{4}$$

where N_A Total number of outcomes that are favorable to A, N is the total number of all possible outcomes that are equally likely

Probability

- Example : Estimate the prob. of rolling 2 dice where the sum = 7 using definitions of probability.

Axioms of Probability

- For any event A, P(A) $\geq$ 0.
- Probability of the sample space S is P(S)=1.
- if $A_1, A_2, A_3 ..... A_n$ are disjoint events then

$$p(A_1 \cup A_2 \cup A_3 \cup A_4 ...) = P(A_1) + P(A_2) + P(A_3) + ... + P(A_n) \quad (5)$$

- P(A $\cap$ B) = P(A and B) = P(A,B)
- P(A $\cup$ B) = P(A or B)

Example

- In a presidential election, there are four candidates. Call them A, B, C, and D. Based on our polling analysis, we estimate that A has a 20 percent chance of winning the election, while B has a 40 percent chance of winning. What is the probability that A or B win the election?

Conditional Probability

- If A and B are two events in a sample space S, then the conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{6}$$

where P(B) should be greater than 0.

Example

- I roll a fair die twice and obtain two numbers X1= result of the first roll and X2= result of the second roll. Given that I know X1+X2=7, what is the probability that X1=4 or X2=4?

Random Variables

- A random variable is a real-valued variable whose value is determined by an underlying random experiment.
- A random variable X is a function from the sample space to the real numbers

$$X : S \rightarrow R \tag{7}$$

Example

- I toss a coin five times. This is a random experiment and the sample space can be written as S=TTTTT,TTTTH,...,HHHHH.
- Note that here the sample space S has 32 elements. Suppose that in this experiment, we are interested in the number of heads. We can define a random variable X whose value is the number of observed heads. The value of X will be one of 0,1,2,3,4 or 5 depending on the outcome of the random experiment.

Random Variables

- Discrete Random Variable : Range is countable
- Continuous Random Variable : Range is not countable

Discrete Random Variables

- Concept of PMF (Probability Mass Function):

Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, \dots\}$ (finite or countably infinite). The function

$$P_X(x_k) = P(X = x_k), \text{ for } k = 1, 2, 3, \dots,$$

is called the *probability mass function (PMF)* of X.

- Thus, the PMF is a probability measure that gives us probabilities of the possible values for a random variable. While the above notation is the standard notation for the PMF of X, it might look confusing at first. The subscript X here indicates that this is the PMF of the random variable X. Thus, for example, $P_X(1)$ shows the probability that X=1.

PMF

- For discrete random variables, the PMF is also called the probability distribution.
- I toss a fair coin twice, and let X be defined as the number of heads I observe. Find the range of X, R_X, as well as its probability mass function P_X.

Bernoulli RV

A random variable X is said to be a *Bernoulli* random variable with *parameter* p, shown as $X \sim Bernoulli(p)$, if its PMF is given by

$$P_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Geometric RV

A random variable X is said to be a *geometric* random variable with *parameter* p, shown as $X \sim Geometric(p)$, if its PMF is given by

$$P_X(k) = \begin{cases} p(1-p)^{k-1} & \text{for } k = 1, 2, 3, \dots \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Example

- I roll a fair die repeatedly until a number larger than 4 is observed. If N is the total number of times that I roll the die, find P(N=k), for k=1,2,3,....

Binomial RV

A random variable X is said to be a *binomial* random variable with parameters n and p, shown as $X \sim Binomial(n, p)$, if its PMF is given by

$$P_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, 2, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

Example

- What is the probability of finding at most one defective part in picking 3 parts at a time from a box containing 500 parts. The probability of finding the defective part is 0.25?

Poisson RV

A random variable X is said to be a *Poisson* random variable with parameter λ, shown as $X \sim Poisson(\lambda)$, if its range is $R_X = \{0, 1, 2, 3, \dots\}$, and its PMF is given by

$$P_X(k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{for } k \in R_X \\ 0 & \text{otherwise} \end{cases}$$

Example

- The number of emails that I get in a weekday can be modeled by a Poisson distribution with an average of 0.2 emails per minute.
  - What is the probability that I get no emails in an interval of length 5 minutes?
  - What is the probability that I get more than 3 emails in an interval of length 10 minutes?

Thank You