BTW-Data Science Task 3

Bisma Fajar

June 2024

1 Statistics

Required to understand data science.there are two main types that serve distinct purposes:

1.1 Descriptive Statistics:

Descriptive statistics summarize and organize data from a sample. They provide a concise overview of the data without making inferences beyond what is observed. Common descriptive measures include: **Mean:** The average value of a set of data points.
Median: The middle value when data points are arranged in ascending or descending order.
Mode: The most frequently occurring value.
Measures of dispersion: These indicate how spread out the data is (e.g., standard deviation, range).
Descriptive statistics are useful for presenting and understanding data, but they don't involve making predictions or drawing conclusions about a larger population.

1.2 Inferential Statistics:

Inferential statistics draw conclusions about a population based on a sample. They allow us to make predictions and generalize findings beyond the observed data.

2 Probability

It is defined as a number of positive outcomes (that lead to the event), divided by total number of outcomes, given that all outcomes are equally probable.
For example when we roll a dice, the probability that we get an even number is $3/6 = 0.5$.

2.1 Random variables

the random variable that represents a number obtained when rolling a dice would take values from 1 to 6. Set of numbers from 1 to 6 is called **sample space**. The random variable in previous example is called **discrete**, because it has a countable sample space, i.e. there are separate values that can be enumerated. There are cases when sample space is a range of real numbers, or the whole set of real numbers. Such variables are called **continuous**. A good example is the time when the bus arrives.

2.2 Probability Distribution

The most well-known discrete distribution is uniform distribution, in which there is a sample space of N elements, with equal probability of 1/N for each of them.
Probability density function A continuous analog of uniform distribution is

$$P(t_1 \leq X < t_2) = \int_{t_1}^{t_2} p(x)dx$$

Figure 1: Probability density function.

called **continuous uniform**, which is defined on a finite interval. A probability that the value X falls into an interval of length l is proportional to l, and rises up to 1.

Another important distribution is **normal distribution**.The distribution of weights , and many measurements from real world follow the same type of distribution, but with different mean and variance. This distribution is called normal distribution.
Expectation
It can be demonstrated that for any discrete distribution with values x1, x2, ..., xN and corresponding probabilities p1, p2, ..., pN, the expectation would equal to E(X)=x1p1+x2p2+...+xNpN.
To identify how far the values are spread, we can compute the **variance**
$\sigma^2 = \frac{\sum(x_i-\mu)^2}{n}$
, where is the mean of the sequence. The value is called **standard deviation**, and 2 is called a variance.

Confidence Interval for the Mean

A confidence interval provides a range of values which is likely to contain the population parameter with a specified level of confidence.

For a known population standard deviation σ, the confidence interval for the population mean μ is calculated as:

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

Where:

- $\bar{x}$ is the sample mean.
- $z_{\alpha/2}$ is the critical value from the standard normal distribution for a given confidence level.
- σ is the population standard deviation.
- n is the sample size.

When the population standard deviation σ is unknown, the sample standard deviation s is used, and the t-distribution is employed:

$$\bar{x} \pm t_{\alpha/2,\, n-1} \left(\frac{s}{\sqrt{n}} \right)$$

Where:

- $t_{\alpha/2,\, n-1}$ is the critical value from the t-distribution with $n-1$ degrees of freedom.

Example

Suppose we have a sample with the following data:

- Sample mean ($\bar{x}$): 50
- Sample standard deviation (s): 10
- Sample size (n): 25
- Confidence level: 95

To find the 95
Thus, the confidence interval is:

$$50 \pm 2.064 \left(\frac{10}{\sqrt{25}} \right)$$

Calculating the margin of error:

$$2.064\left(\frac{10}{5}\right) = 2.064 \times 2 = 4.128$$

Therefore, the 95

$$50 \pm 4.128 = [45.872, 54.128]$$