UNIT II

Statistical Modelling

Definition:

Statistical modelling is the use of mathematical models and statistical assumptions to generate sample data and make predictions about the real world. A statistical model is a collection of probability distributions on a set of all possible outcomes of an experiment.

What is Statistical Modelling?

Statistical modelling refers to the process of applying statistical analysis to datasets.

A statistical model is a mathematical relationship between one or more random variables and other non-random variables.

The application of statistical modelling to raw data helps data scientists approach data analysis in a strategic manner, providing intuitive visualizations that aid in identifying relationships between variables and making predictions.

Common data sets for statistical analysis include Internet of Things (IoT) sensors, census data, public health data, social media data, imagery data, and other public sector data that benefit from real-world predictions.

Statistical Modelling Techniques

The first step in developing a statistical model is gathering data, which may be sourced from spreadsheets, databases, data lakes, or the cloud. The most common statistical modelling methods for analyzing this data are categorized as either supervised learning or unsupervised learning.

Some popular statistical models include:

- Supervised
    - Regression model
    - Classification model:
- Un Supervised

Clustering

Reinforcement learning

Supervised learning techniques:

- Regression model: a type of predictive statistical model that analyzes the relationship between a dependent and an independent variable. Common regression models include logistic, polynomial, and linear regression models. Applications include forecasting, time series modelling, and discovering the causal effect relationship between variables.

- Classification model: a type of machine learning in which an algorithm analyzes an existing, large and complex set of known data points as a means of understanding and then appropriately classifying the data; common models are decision trees, Naive Bayes, nearest neighbour, random forests and neural networking models, which are typically used in Artificial Intelligence.

Unsupervised learning techniques:

- Clustering: aggregates a specified number of data points into a specific number of groupings based on certain similarities.

- Reinforcement learning: an area of deep learning that concerns models iterating over many attempts, rewarding moves that produce favourable outcomes and penalizing steps that produce undesired outcomes, therefore training the algorithm to learn the optimal process.

There are three main types of statistical models: parametric, nonparametric, and semiparametric:

- Parametric: a family of probability distributions that has a finite number of parameters.

- Nonparametric: models in which the number and nature of the parameters are flexible and not fixed in advance.

- Semiparametric: the parameter has both a finite-dimensional component (parametric) and an infinite-dimensional component (nonparametric).

How to Build Statistical Models?

- Start with univariate descriptives and graphs. Visualizing the data helps with identifying errors, understanding the variables you're working with, how they look, how they are behaving and why.
- Build predictors in theoretically distinct sets first in order to observe how related variables work together, and then the outcome once the sets are combined.
- Then, run bivariate descriptives with graphs in order to visualize and understand how each potential predictor relates individually to every other predictor and to the outcome.
- Frequently record, compare and interpret results from models
- Eliminate non-significant interactions first; any variable involved in a significant interaction must be included in the model by itself.

Machine Learning vs Statistical Modelling

Machine learning is a subfield of computer science and artificial intelligence that involves building systems that can learn from data rather than explicitly programmed instructions.

Machine learning models seek out patterns hidden in data independent of all assumptions, therefore predictive power is typically very strong.

Machine learning requires little human input and does well with large numbers of attributes and observations.

Statistical modelling is a subfield of mathematics that seeks out relationships between variables in order to predict an outcome.

Statistical models are based on coefficient estimation and are typically applied to smaller sets of data with fewer attributes, and require the human designer to understand the relationships between variables before inputting.

Random variables, Probability mass/density functions

A random variable is a numerical description of the outcome of a statistical experiment.

Discrete random variable:

A random variable that may assume only a finite number or an infinite sequence of values is said to be discrete

Continuous random variable:

A random variable that may assume any value in some interval on the real number line is said to be continuous.

Ex: A random variable representing the number of automobiles sold at a particular showroom on one day would be discrete.

A random variable representing the weight of a person in kilograms is continuous.

The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable.

For a discrete random variable, x, the probability distribution is defined by a probability mass function, denoted by f(x). This function provides the probability for each value of the random variable.

In the development of the probability function for a discrete random variable, two conditions must be satisfied:

(1) f(x) must be nonnegative for each value of the random variable

(2) the sum of the probabilities for each value of the random variable must equal one.

A continuous random variable may assume any value in an interval on the real number line or in a collection of intervals. Since there is an infinite number of

values in any interval, it is not meaningful to talk about the probability that the random variable will take on a specific value; instead, the probability that a continuous random variable will lie within a given interval is considered.

In the continuous case, the counterpart of the probability mass function is the probability density function, also denoted by f(x).

For a continuous random variable, the probability density function provides the height or value of the function at any particular value of x; it does not directly give the probability of the random variable taking on a specific value. However, the area under the graph of f(x) corresponding to some interval, obtained by computing the integral of f(x) over that interval, provides the probability that the variable will take on a value within that interval.

A probability density function must satisfy two requirements:

(1) f(x) must be nonnegative for each value of the random variable

(2) The integral over all values of the random variable must equal one.

The expected value, or mean, of a random variable—denoted by E(x) or μ—is a weighted average of the values the random variable may assume.

In the discrete case the weights are given by the probability mass function, and in the continuous case the weights are given by the probability density function.

The formulas for computing the expected values of discrete and continuous random variables respectively are given by following equations.

$E(x) = \Sigma x\, f(x)$

$E(x) = \int x\, f(x)dx$

The variance of a random variable, denoted by Var(x) or $\sigma 2$, is a weighted average of the squared deviations from the mean.

In the discrete case the weights are given by the probability mass function, and in the continuous case the weights are given by the probability density function.

The formulas for computing the variances of discrete and continuous random variables are given by following equations respectively.

$Var(x) = \sigma 2 = \Sigma\, (x - \mu)2f(x)$

$Var(x) = \sigma 2 = \int (x - \mu)2f(x)dx$

The standard deviation, denoted σ, is the positive square root of the variance. Since the standard deviation is measured in the same units as the random variable and the variance is measured in squared units, the standard deviation is often the preferred measure.

Probability distributions

Two of the most widely used discrete probability distributions are the binomial and Poisson.

1.The binomial distribution

The binomial probability mass function provides the probability that x successes will occur in n trials of a binomial experiment.

Below is the equation of Binomial distribution.

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

A binomial experiment has four properties:

(1) it consists of a sequence of n identical trials

(2) two outcomes, success or failure, are possible on each trial

(3) the probability of success on any trial, denoted p, does not change from trial to trial

(4) the trials are independent.

2.The Poisson distribution

The Poisson probability distribution is often used as a model of the number of arrivals at a facility within a given period of time.

For example, a random variable might be defined as the number of telephone calls coming into an airline reservation system during a period of 15 minutes. If the mean number of arrivals during a 15-minute interval is known, the Poisson

probability mass function given by following equation can be used to compute the probability of x arrivals.

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

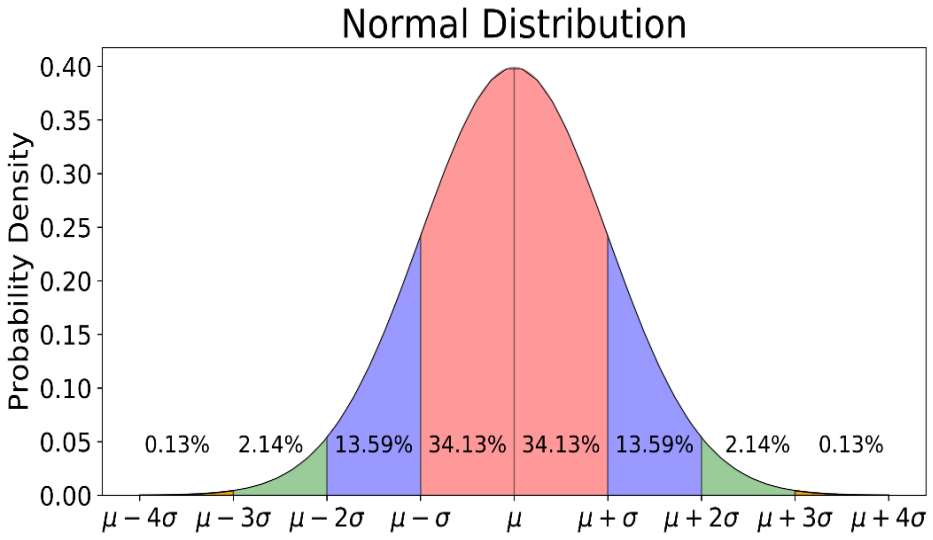
For example, suppose that the mean number of calls arriving in a 15-minute period is 10. To compute the probability that 5 calls come in within the next 15 minutes, μ = 10 and x = 5 are substituted in equation 7, giving a probability of 0.0378.

3.The normal distribution

The most widely used continuous probability distribution in statistics is the normal probability distribution.

The graph corresponding to normal distribution is a bell-shaped curve. Probabilities for the normal probability distribution can be computed using statistical tables for the standard normal probability distribution, which is a normal probability distribution with a mean of zero and a standard deviation of one.

A simple mathematical formula is used to convert any value from a normal probability distribution with mean μ and a standard deviation σ into a corresponding value for a standard normal distribution. The tables for the standard normal distribution are then used to compute the appropriate probabilities.



There are many other discrete and continuous probability distributions.

Other widely used discrete distributions include the geometric, the hypergeometric, and the negative binomial.

Other commonly used continuous distributions include the uniform, exponential, gamma, chi-square, beta, t, and F.

Hypothesis testing

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

Initially, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H0.

An alternative hypothesis (denoted Ha), which is the opposite of what is stated in the null hypothesis, is then defined.

The hypothesis-testing procedure involves using sample data to determine whether or not H0 can be rejected. If H0 is rejected, the statistical conclusion is that the alternative hypothesis Ha is true.

For example, assume that a radio station selects the music it plays based on the assumption that the average age of its listening audience is 30 years.

To determine whether this assumption is valid, a hypothesis test could be conducted with the null hypothesis given as H0: $\mu = 30$ and the alternative hypothesis given as Ha: $\mu \neq 30$.

Based on a sample of individuals from the listening audience, the sample mean age, $\bar{x}$, can be computed and used to determine whether there is sufficient statistical evidence to reject H0.

Conceptually, a value of the sample mean that is “close” to 30 is consistent with the null hypothesis, while a value of the sample mean that is “not close” to 30 provides support for the alternative hypothesis.

What is considered “close” and “not close” is determined by using the sampling distribution of $\bar{x}$.

Ideally, the hypothesis-testing procedure leads to the acceptance of H0 when H0 is true and the rejection of H0 when H0 is false.

Unfortunately, since hypothesis tests are based on sample information, the possibility of errors must be considered. A type-I error corresponds to rejecting H0 when H0 is actually true, and a type-II error corresponds to accepting H0 when H0 is false.

The probability of making a type-I error is denoted by α, and the probability of making a type-II error is denoted by β.

In using the hypothesis-testing procedure to determine if the null hypothesis should be rejected, the person conducting the hypothesis test specifies the maximum allowable probability of making a type-I error, called the level of significance for the test.

Common choices for the level of significance are $\alpha = 0.05$ and $\alpha = 0.01$.

Although most applications of hypothesis testing control the probability of making a type-I error, they do not always control the probability of making a type-II error.

A graph known as an operating-characteristic curve can be constructed to show how changes in the sample size affect the probability of making a type-II error.

A concept known as the p-value provides a convenient basis for drawing conclusions in hypothesis-testing applications. The p-value is a measure of how likely the sample results are, assuming the null hypothesis is true.

The smaller the p-value, the less likely the sample results. If the p-value is less than α, the null hypothesis can be rejected; otherwise, the null hypothesis cannot be rejected. The p-value is often called the observed level of significance for the test.

A hypothesis test can be performed on parameters of one or more populations as well as in a variety of other situations.

In each instance, the process begins with the formulation of null and alternative hypotheses about the population.

In addition to the population mean, hypothesis-testing procedures are available for population parameters such as proportions, variances, standard deviations, and medians.

Hypothesis tests are also conducted in regression and correlation analysis to determine if the regression relationship and the correlation coefficient are statistically significant.

A goodness-of-fit test refers to a hypothesis test in which the null hypothesis is that the population has a specific probability distribution, such as a normal probability distribution.

Nonparametric statistical methods also involve a variety of hypothesis-testing procedures.

A sample statistic is a piece of information you get from a fraction of a population.

When your statistical information (like an average, median, or some other kind of statistic) comes from a fraction of data or part of a population, it's called a sample statistic.

Difference between Descriptive and Inferential Statistics

Descriptive and inferential statistics are two broad categories in the field of statistics. Some of the statistical measures are similar, but the goals and methodologies are very different.

Descriptive Statistics

Both descriptive and inferential statistics help make sense out of row after row of data!

Use descriptive statistics to summarize and graph the data for a group that you choose. This process allows you to understand that specific set of observations.

Descriptive statistics describe a sample. You simply take a group that you're interested in, record data about the group members, and then use summary statistics and graphs to present the group properties. With descriptive statistics, there is no uncertainty because you are describing only the people or items that you actually measure. You're not trying to infer properties about a larger population.

The process involves taking a potentially large number of data points in the sample and reducing them down to a few meaningful summary values and graphs. This procedure allows us to gain more insights and visualize the data than simply pouring through row upon row of raw numbers!

Common tools of descriptive statistics

Descriptive statistics frequently use the following statistical measures to describe groups:

Central tendency: Use the mean or the median to locate the center of the dataset. This measure tells you where most values fall.

Dispersion: How far out from the center do the data extend? You can use the range or standard deviation to measure the dispersion. A low dispersion indicates that the values cluster more tightly around the center. Higher dispersion signifies that data points fall further away from the center. We can also graph the frequency distribution.

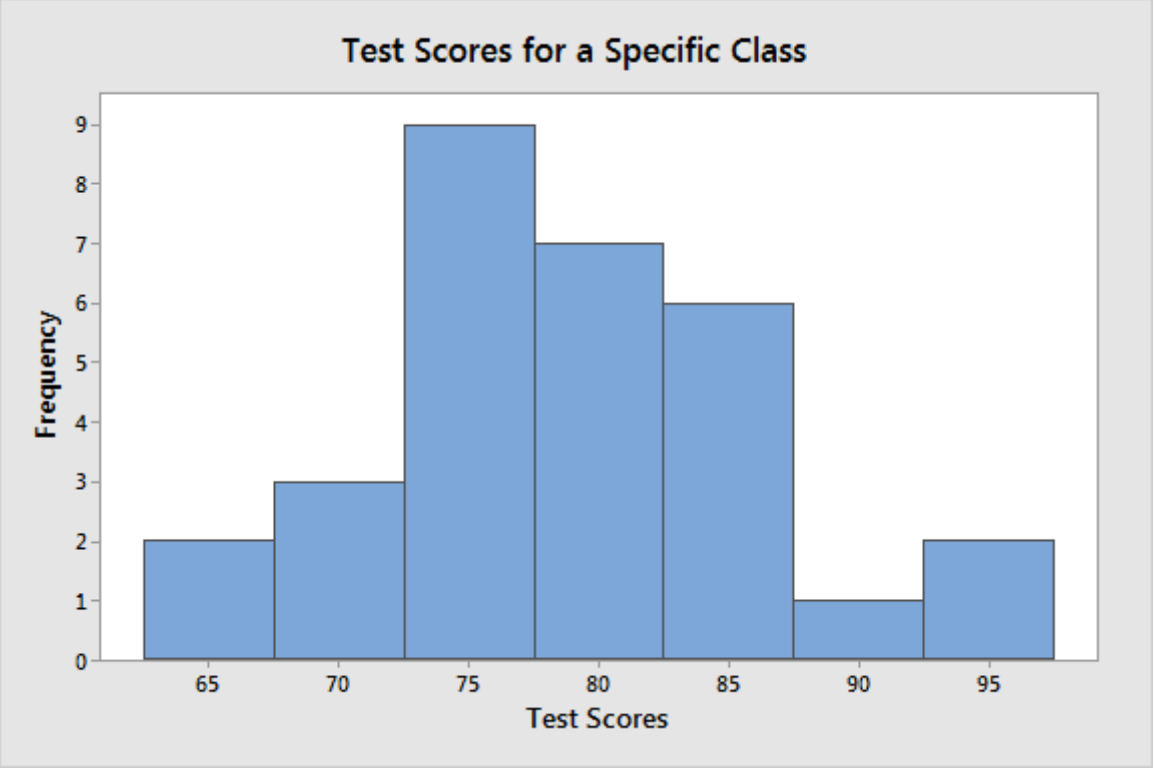
Skewness: The measure tells you whether the distribution of values is symmetric or skewed.

You can present this summary information using both numbers and graphs. These are the standard descriptive statistics, but there are other descriptive analyses you can perform, such as assessing the relationships of paired data using correlation and scatterplots.

Example of descriptive statistics

Suppose we want to describe the test scores in a specific class of 30 students. We record all of the test scores and calculate the summary statistics and produce graphs.

Histogram of test score distribution for the descriptive statistics example.



Statistic	Class value
Mean	79.18
Range	66.21 – 96.53
Proportion >= 70	86.7%

These results indicate that the mean score of this class is 79.18. The scores range from 66.21 to 96.53, and the distribution is symmetrically centered around the mean. A score of at least 70 on the test is acceptable. The data show that 86.7% of the students have acceptable scores.

Collectively, this information gives us a pretty good picture of this specific class. There is no uncertainty surrounding these statistics because we gathered the scores for everyone in the class. However, we can't take these results and extrapolate to a larger population of students.

Inferential Statistics

Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn. Because the goal of inferential statistics is to draw conclusions from a sample and generalize them to a population, we need to have confidence that our sample accurately reflects the population. This requirement affects our process. At a broad level, we must do the following:

Define the population we are studying.

Draw a representative sample from that population.

Use analyses that incorporate the sampling error.

We don't get to pick a convenient group. Instead, random sampling allows us to have confidence that the sample represents the population. This process is a primary method for obtaining samples that mirrors the population on average. Random sampling produces statistics, such as the mean, that do not tend to be too high or too low. Using a random sample, we can generalize from the sample to the broader population. Unfortunately, gathering a truly random sample can be a complicated process.

You can use the following methods to collect a representative sample:

Simple random sampling

Stratified sampling

Cluster sampling

Systematic sampling

Pros and cons of working with samples

You gain tremendous benefits by working with a random sample drawn from a population. In most cases, it is simply impossible to measure the entire population to understand its properties. The alternative is to gather a random sample and then use the methodologies of inferential statistics to analyze the sample data.

While samples are much more practical and less expensive to work with, there are trade-offs. Typically, we learn about the population by drawing a relatively small sample from it. We are a very long way off from measuring all people or objects in that population. Consequently, when you estimate the properties of a population from a sample, the sample statistics are unlikely to equal the actual population value exactly.

For instance, your sample mean is unlikely to equal the population mean exactly. The difference between the sample statistic and the population value is the sampling error. Inferential statistics incorporate estimates of this error into the statistical results.

In contrast, summary values in descriptive statistics are straightforward. The average score in a specific class is a known value because we measure all individuals in that class. There is no uncertainty.

Standard analysis tools of inferential statistics

The most common methodologies in inferential statistics are hypothesis tests, confidence intervals, and regression analysis. Interestingly, these inferential

methods can produce similar summary values as descriptive statistics, such as the mean and standard deviation.

In descriptive statistics, we picked the specific class that we wanted to describe and recorded all of the test scores for that class. Nice and simple. For inferential statistics, we need to define the population and then draw a random sample from that population.

A study using descriptive statistics is simpler to perform. However, if you need evidence that an effect or relationship between variables exists in an entire population rather than only your sample, you need to use inferential statistics.