From the 1980’s, a lot of focus has been made on improving the quality of production/ industry output in American industry. However, an industry miracle began in the middle of the 20th century in Japan where other countries failed, Japanese excelled in creating an atmosphere for high level production. This resulted because of statistical methods being followed by them.

Statistical methods are widely used across different industries, it has a wide spectrum usage in various sectors such as manufacturing, food production, software dev, energy, and pharmacy. All this requires inferential statistics as it equips us with tools to make informed judgments. In any process, there is variability. Statistical methods help in identifying areas of improvement. For instance, in pharmaceuticals, effectiveness of any drug is evaluated on patient-to-patient variability, hence here comes the need of statistical methods, as it helps us to take informed decisions.

Statisticians in order to draw inferences use fundamentals laws of probability. Also, information is gathered in the form of samples or collection of observations. Samples are collected from populations, which are collections of all individuals.

Probability and statistical inference are distinct, yet they are very interconnected. Probability is just a theory while statistical inference applies probability concepts to interpret real-life results. Statistical inference uses probability and sample data to draw conclusion about the characteristics of population. Probability is necessary because it helps to identify uncertainty in samples

Simple Random Sampling means that any sample of a given size has an equal chance of being chosen, regardless of elements. Sample size refers to the total number of elements included in the sample. Random number tables can be used to assist in selecting samples. The main benefit of simple random sampling is that it helps minimize the risk of the sample being unrepresentative of the broader population which indeed is very crucial

Measures of Location

Measures of Location: used to give analyst insights about the values of the data such as where is the center, or other location of data.

Sample mean is an important measure of location. It is denoted by x bar. Suppose there are n number of observation in a sample, then sample mean is equal to sum of observations divided by number of observations.

A mathematical equation with numbers and symbols

Description automatically generated

Fig 1: Sample Mean

Sample Median – used for central tendency. in simpler words it’s the middle value of an ordered sample set. To calculate median for any sample the first step is to arrange the sample in an increasing order, now we have to identify the number of observations in the sample are odd or even. Now the sample median is given by the image below

A math equations on a white background

Description automatically generated

Fig 2: sample median

Note: sample mean is considered as the **centroid of the data** in a sample.

Trimmed mean – it is another measure of location. It is calculated by trimming or eliminating a percentage of value from the smallest. eg 10% trimmed mean is calculating by elimination 10 percent of largest set and 10 % of smallest set.  
Note – trimmed mean is more insensitive to outliers than mean but not that insensitive as median

Measures of Variability

Sample Range – simplest one – calculated by identifying the max and min of a sample set. Sample Range = X max – X min, so basically, it’s the difference between maximum and minimum values in the sample set.

Sample Standard deviation – generally sample measure of spread, denoted by s

Sample variance denoted by s2 – is A mathematical equation with numbers and symbols

Description automatically generated

Fig 3: Sample variance

Sample Standard Deviation which is denoted by s is the square root of s2 i.e. Variance



Fig 4 : Sample Standard Deviation

Note – Larger the variability larger would-be sample variance.

Also, n-1, which is the denominator in variance, is known as degree of freedom associated with variance.

Variance measures the average squared deviation from the mean and using degrees of freedom. It has squared units, while sd (standard deviation) has linear units same as data.

Discrete vs Continuous data

For any observational study or experiment data gathered may be discrete or continuous depends on the application. In general, discrete data is like a countable outcome, for example of number of defective items while continuous data is measured in continuity like temperature.

Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

Stem and leaf plot

data can be very useful for studying the behavior of the distribution if it’s presented in tabular and graphical display which is called stem and leaf plot. In general, it’s a statistical tool that organizes numeric data into stems and leaves where stem represents significant digits and leaf represents the remaining one. It is an efficient way to summarize data

stem-and-leaf plot is a basis for constructing frequency distributions. Each stem defines a class interval and leaves within is the frequency of observations in that interval.

Histogram

When we divide each class frequency by number of observations we get the relative frequency distribution. It signifies proportion of observation within each class interval. Using this relative frequency distribution and midpoint of each interval we construct a relative frequency histogram. Many continuous frequency distributions can be represented graphically by the characteristic bell-shaped curve.

The distribution is symmetric if it can be folded along a vertical axis (two sides coincides)

A distribution that lacks symmetry with respect to vertical axis is said to be skewed.

Box Plot

Box plot, which is also known as box-and-whisker plot, is a graphical representation for summarizing a dataset by showing the central tendency, variability and skewness through quartiles. It gives a visual representation of the distribution of data by highlighting the median, quartiles and any outliers.

Components of a box plot - >

Min (Q0 or 0th percentile) – excluding outliers

First Quartile (Q1or 25th percentile) – median of the lower half of the dataset.

Median (Q2 or 50th percentile) – middle value of the dataset, that divides into two equal parts)

Third Quartile (Q3 or 75th percentile)- median of the upper half of the dataset.

Max (Q4 or 100th percentile) – excluding outliers

IQR, known as interquartile range, is the measure of 50 percent spread of the data. Whiskers extend from the box to the smallest and largest values within 1.5 times the IQR from the quartiles. Data points which are considered outliers and are beyond the whiskers and are plotted individually.

Box plot provided the viewer with which observations may be outliers. Outliers are basically those observations that are considered far from the sample data. It’s also called the rare event. The visual information gathered from box and whisker plot is not a formal test for outliers, it’s a diagnostic tool. A common procedure that can be used to identify outlier is to use a multiple of the interquartile range. Example if the distance exceeds 1.5 times IQR it can be considered that observation is outlier in the sample data set.

References

R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability & Statistics for Engineers & Scientists*, 9th ed. Prentice Hall