# Unit 5: Statistical Distribution Functions

Contents

## 1.0 Introduction

Although simulation can be a valuable tool for better understanding the underlying mechanisms that control the behaviour of a system, using simulation to make ***predictions*** of the future behaviour of a system can be difficult. This is because, for most real-world systems, at least some of the controlling parameters, processes and events are often stochastic, uncertain and/or poorly understood. The objective of many simulations is to identify and quantify the risks associated with a particular option, plan or design. Simulating a system in the face of such uncertainty and computing such risks requires that the uncertainties be quantitatively included in the calculations. To do this we collect data about the system parameters and subject them to statistical analysis.

## 2.0 Intended Learning Outcomes (ILOs)

After studying this unit the reader should be able to

- Define Statistics
- Explain Statistical Distributions

- Compute measures of Central Tendency and Variations
- Explain the Components of Statistical Distributions
    - Normal Distributions,
    - z-score
    - percentile,
    - Skewed Distributions
    - Ways to transform data to Graphs

## 3.0 Main Content

### 3.1 What is Statistics?

The field of statistics is concerned with the collection, description, and interpretation of data (data are numbers obtained through measurement). In the field of statistics, the term “statistic” denotes a measurement taken on a sample (as opposed to a population). In general conversation, “statistics” also refers to facts and figures.

### 3.2 What is a Statistical Distribution?

A statistical distribution describes the numbers of times each possible outcome occurs in a sample. If you have 10 test scores with 5 possible outcomes of A, B, C, D, or F, a statistical distribution describes the relative number of times an A,B,C,D or F occurs. For example, 2 A’s, 4 B’s, 4 C’s, 0 D’s, 0 F’s.

### 3.3 Measures of Central Tendency

Suppose we have a sample with the following 4 observations: 4, 1, 4, 3.

**Mean** - the sum of a set of numbers divided by the number of observations.

$$\text{Mean} = \frac{4+1+4+3}{4} = \frac{12}{4} = 3$$

**Median** - the middle point of a set of numbers (for odd numbered samples). the mean of the middle two points (for even samples).

$$\text{Median} = 1,\underline{3},\underline{4},4 \text{ or } \frac{3+4}{2} = \frac{7}{2} = 3.5$$

**Mode -** the most frequently occurring number.

Mode = 4 (4 occurs most).

The mean, median and mode are called measures of central tendency.

### 3.4 Measures of Variation

**Range** - the maximum value minus the minimum value in a set of numbers. Range = 4-1 = 3.

**Standard Deviation -** the average distance a data point is away from the mean.

$$\text{standard deviation} = \frac{|4-3|+|1-3|+|4-3|+|3-3|}{4} = \frac{1+2+1+0}{4} = \frac{4}{4} = 1$$

Standard deviation computes the difference between each data point and the mean. Take the absolute value of each difference. Sum the absolute values. Divide this sum by the number of data points. Median: first arrange data points in increasing order.

Mean, Median, Mode, Range, and Standard Deviations are measurements in a sample (statistics) and can also be used to make inferences on a population.

### 3.5 Showing Data Distribution in Graphs

- **Bar graphs** use bars to compare frequencies of possible data values (see Fig a).
- **Double bar graphs** use two sets of bars to compare frequencies of data values between two levels of data (e.g. boys and girls) (see fig b).
- **Histograms** use bars to show how frequently data occur within equal spaces within an interval (see fig c & d).
- **Pie Charts** use portion of a circle to show contributions of data values (see fig c & d).

### 3.6 The Difference between a Continuous and a Discrete Distribution

**Continuous distributions** describe an infinite number of possible data values (as shown by the curve). For example someone's height could be 1.7m, 1.705m, 1.71m, ...

**Discrete distributions** describe a finite number of possible values. (shown by the bars)

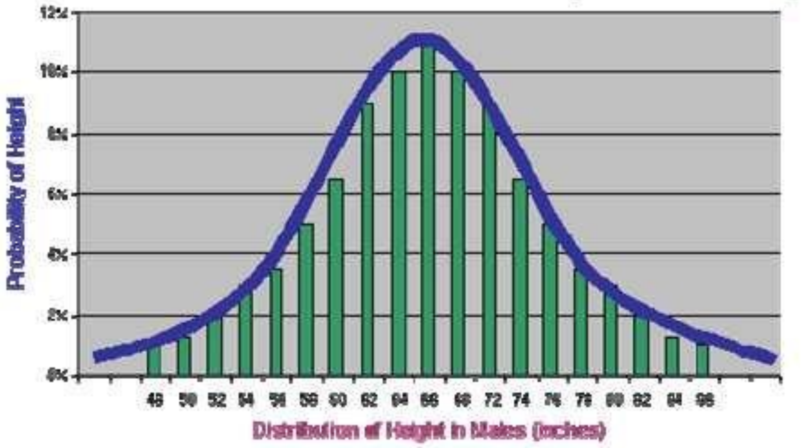

Fig 2: Distribution of Height in Males

### 3.7 Normal Distribution

**A normal distribution** is a continuous distribution that is "bell-shaped". Data are often assumed to be normal. Normal distributions can estimate probabilities over a continuous interval of data values.

The **normal distribution** refers to a family of continuous probability distributions described by the normal equation.

In a normal distribution, data are most likely to be at the mean. Data are less likely to be farther away from the mean.

The normal distribution is defined by the following equation:

$$Y = [\ 1/\sigma * \mathrm{sqrt}(2\pi)\ ] * e^{-(x\ -\ \mu)2/2\sigma2}$$

where *X* is a normal random variable, μ is the mean, σ is the standard deviation, π is approximately 3.14159, and *e* is approximately 2.71828.

The random variable *X* in the normal equation is called the **normal random variable**. The normal equation is the probability density function for the normal distribution.

The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve, as shown in figure 3a and 3b.

(a) (b)

Fig. 3: Graph of Normal Distribution Based on Size of Mean and Standard Deviation

The curve on the left is shorter and wider than the curve on the right, because the curve on the left has a bigger standard deviation.

### 3.7.1 Standard Normal Distribution

The **standard normal distribution** is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.

The normal random variable of a standard normal distribution is called a **standard score** or a **z-score**. Every normal random variable $X$ can be transformed into a $z$ score via the following equation:

$$z = (X - \mu) / \sigma$$

where $X$ is a normal random variable, μ is the mean of $X$, and σ is the standard deviation of $X$.

### 3.7.2 The Normal Distribution as a Model for Measurements

Often, phenomena in the real world follow a normal (or near-normal) distribution. This allows researchers to use the normal distribution as a model for assessing probabilities associated with real-world phenomena. Typically, the analysis involves two steps.

- Transform raw data. Usually, the raw data are not in the form of z-scores. They need to be transformed into z-scores, using the transformation equation presented earlier: $z = (X - \mu) / \sigma$.
- Find the probability. Once the data have been transformed into z-scores, you can use standard normal distribution tables, online calculators (e.g., Stat Trek's free normal distribution calculator) to find probabilities associated with the z-scores.

The problem in the next section demonstrates the use of the normal distribution as a model for measurement.

**Example 1 - Ada** earned a score of 940 on a national achievement test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a higher score than Ada? (Assume that test scores are normally distributed.)

**Solution -** As part of the solution to this problem, we assume that test scores are normally distributed. In this way, we use the normal distribution as a model for measurement. Given an assumption of normality, the solution involves three steps.

- First, we transform Ada's test score into a z-score, using the z-score transformation equation.
  $$z = (X - \mu) / \sigma = (940 - 850) / 100 = 0.90$$
- Then, using a standard normal distribution table, we find the cumulative probability associated with the z-score. In this case, we find $P(Z < 0.90) = 0.8159$.
- Therefore, the $P(Z > 0.90) = 1 - P(Z < 0.90) = 1 - 0.8159 = 0.1841$.

Thus, we estimate that 18.41 percent of the students tested had a higher score than Ada.

**Example 2 -** An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

*Solution:* Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following:

- The value of the normal random variable is 365 days.
- The mean is equal to 300 days.
- The standard deviation is equal to 50 days.

We enter these values into the formula and compute the cumulative probability. The answer is: P( X $\leq$ 365) = 0.90. Hence, there is a 90% chance that a light bulb will burn out within 365 days.

### 3.7.3 Conversion to a Standard Normal Distribution

The values for points in a standard normal distribution are **z-scores**. We can use a standard normal table to find the probability of getting at or below a z-score. (a percentile).

- Subtract the mean from each observation in your normal distribution, the new mean=0.
- Divide each observation by the standard deviation, the new standard deviation=1.

### 3.7.4 Skewed Distributions µ

Skewness is the degree of asymmetry or departure from symmetry, of a distribution. Skewed distributions are not symmetric. If the frequency curve of a distribution has a longer tail to the right of the right of the central maximum than to the left, the distribution is said to be skewed to the right, or have a positive skewness. If the reverse is the case, it is said to be skewed to the left or negative skewness.

For skewed distributions, the mean tend to lie on the same side of the mode as the longer tail. Thus a measure of the asymmetry is supplied by the difference:

Mean – mode. This can be made dimensionless if we divide it by a measure of dispersion, such as the standard deviation, leading to the definition:

$$\text{Skewness} = \frac{mean - \text{mod}\,e}{SD} = \frac{\mu - \text{mod}\,e}{s} \quad (1)$$

To avoid using mode, we can use the empirical formula:

$$\text{Skewness} = \frac{3(mean - \text{median})}{SD} = \frac{3(\mu - \text{median})}{s} \quad (2)$$

Equations (1) and (2) are called; Pearson’s first and second coefficients of skewness.

### 3.8 What is a Percentile?

**A percentile** (or **cumulative probability**) is the proportion of data in a distribution less than or equal to a data point. If you scored a 90 on a math test and 80% of the class had scores of 90 or lower; your percentile is 80. In the figure 4, b=90 and P(Z<b)=80.

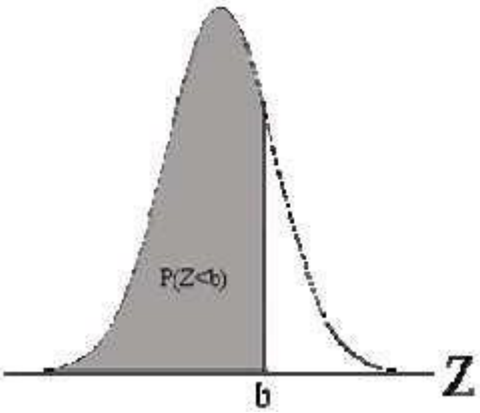

Fig. 4: illustration of Percentiles

### 3.9 Probabilities in Discrete Distributions

Suppose for your 10 tests you received 5 As, 2 Bs, 2 Cs, 1 D and want to find the probability of receiving an A or a B. Sum the frequencies for A and B and divide by the sample size. The probability of receiving an A or a B is (5+2)/10 = .7 (a 70% chance).

### 3.10 Probability and the Normal Curve

The normal distribution is a continuous probability distribution. This has several implications for probability.

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable *X* equals any particular value is 0.
- The probability that *X* is greater than b equals the area under the normal curve bounded by *b* and plus infinity (as indicated by the *non-shaded* area in the figure 4).
- The probability that *X* is less than *a* equals the area under the normal curve bounded by b and minus infinity (as indicated by the *shaded* area in the figure below).

### 4.0 Self-Assessment Exercise(s)

Answer the following questions:

1. Why Convert to a Standard Normal Distribution?
2. What is the difference between a Continuous and a Discrete Distribution?
3. Given the following: mean=279.76, median=279.06, mode=277.5 and SD=15.6, find the first and second coefficients of skewness
4. Find the mode, median and mean deviation of the following sets of data: (a) 3, 7, 9, 5 and (b) 8, 10, 9, 12, 4, 8, 2.

## 5.0 Conclusion

We use Statistical distributions to: investigate how a change in one variable relates to a change in a second variable, represent situations with numbers, tables, graphs, and verbal descriptions, understand measurable attributes of objects and their units, systems, and processes of measurement, identify relationships among attributes of entities or systems and their association.

## 6.0 Summary

In this unit:

- We defined Statistics as field of study that is concerned with the collection, description, and interpretation of data.
- We saw that Statistical Distributions describe the numbers of times each possible outcome occurs in a sample.
- We computed various measures of Central Tendency and Variations which can be used to make inferences.
- And explained the following components of Statistical Distributions:
    - Normal Distributions,
    - z-score
    - percentile,
    - Skewed Distributions
    - Ways to transform data to Graphs

## 7.0 Further Readings

- Devore, J. L. (2018). *Probability and statistics for engineering and the sciences*. Toronto, Ontario: Nelson.
- Georgii, H. (2013). *Stochastics: Introduction to probability and statistics*. Berlin: De Gruyter.
- Giri, N. C. (2019). *Introduction to probability and statistics*. London: Routledge.
- Johnson, R. A., Miller, I., & Freund, J. E. (2019). *Miller & Freunds probability and statistics for engineers*. Boston: Pearson Education.
- Laha, R. G., & Rohatgi, V. K. (2020). *Probability theory*. Mineola, NY: Dover Publications.
- Mathai, A. M., & Haubold, H. J. (2018). *Probability and statistics: A course for physicists and engineers*. Boston: De Gruyter.
- Pishro-Nik, H. (2014). *Introduction to probability, statistics, and random processes*. Blue Bell, PA: Kappa Research, LLC.
- Spiegel, M. R., Schiller, J. J., & Srinivasan, R. A. (2013). *Schaums outline of probability and statistics*. New York: McGraw-Hill.

## Unit 6: Common Probability Distributions

Contents

### 1.0 Introduction

In this section we look at the branch of statistics that deals with analysis of random events. Probability is the numerical assessment of likelihood on a scale from 0 (impossibility) to 1 (absolute certainty). Probability is usually expressed as the ratio between the number of ways an event can happen and the total number of things that can happen (e.g., there are 13 ways of picking a diamond from a deck of 52 cards, so the probability of picking a diamond is 13/52, or ¼). Probability theory grew out of attempts to understand card games and gambling. As science became more rigorous, analogies between certain biological, physical, and social phenomena and games of chance became more evident (e.g., the sexes of newborn infants follow sequences similar to those of coin tosses). As a result, probability became a fundamental tool of modern genetics and many other disciplines.

### 2.0 Intended Learning Outcomes (ILOs)

By the end of this unit, the reader should be able to:

- Explain the role of probability distribution functions in simulations
- Describe Probability theory
- Explain the fundamental concepts of Probability theory
- Explain Random Variable
- Explain Limiting theorems
- Describe Probability distributions in simulations