Descriptive Statistics

3.1 SUMMATION: Summation is the operation of adding a sequence of numbers; the result is their **sum** or *total*. If numbers are added sequentially from left to right, any intermediate result is a partial sum, prefix sum, or running total of the summation. The numbers to be summed (called **addends**, or sometimes **summands**) may be integers, rational numbers, real numbers, or complex numbers. Besides numbers, other types of values can be added as well: vectors, matrices, polynomials and, in general, elements of any additive group (or even monoid). For finite sequences of such elements, summation always produces a well-defined sum (possibly by virtue of the convention for empty sums).

The summation of an infinite sequence of values is called a series. A value of such a series may often be defined, by means of a limit (although sometimes the value may be infinite, and often no value results at all). Another notion involving limits of finite sums is integration. The term summation has a special meaning related to extrapolation in the context of divergent series.

The summation of the sequence [1, 2, 4, 2] is an expression whose value is the sum of each of the members of the sequence. In the example, 1 + 2 + 4 + 2 = 9. Since addition is associative the value does not depend on how the additions are grouped, for instance (1 + 2) + (4 + 2) and 1 + ((2 + 4) + 2) both have the value 9; therefore, parentheses are usually omitted in repeated additions. Addition is also commutative, so permuting the terms of a finite sequence does not change its sum (for infinite summations this property may fail; see absolute convergence for conditions under which it still holds).

There is no special notation for the summation of such explicit sequences, as the corresponding repeated addition expression will do. There is only a slight difficulty if the sequence has fewer than two elements: the summation of a sequence of one term involves no plus sign (it is indistinguishable from the term itself) and the summation of the empty sequence cannot even be written down (but one can write its value "0" in its place). If, however, the terms of the sequence are given by a regular pattern, possibly of variable length, then a summation operator may be useful or even essential. For the summation of the sequence of consecutive integers from 1 to 100 one could use an addition expression involving an ellipsis to indicate the missing terms: 1 + 2 + 3 + 4 + ... + 99 + 100.

3.2 GROUPED DATA: Data

Data can be defined as groups of information that represent the qualitative or quantitative attributes of a variable or set of variables, which is the same as saying that data can be any set of information that describes a given entity. Data in statistics can be classified into grouped data and ungrouped data.

Any data that you first gather is ungrouped data. Ungrouped data is data in the raw. An example of ungrouped data is a any list of numbers that you can think of.

Grouped Data

Grouped data is data that has been organized into groups known as classes. Grouped data has been 'classified' and thus some level of data analysis has taken place, which means that the data is no longer raw.

A data class is group of data which is related by some user defined property. For example, if you were collecting the ages of the people you met as you walked down the street, you could group them into classes as those in their teens, twenties, thirties, forties and so on. Each of those groups is called a class.

Each of those classes is of a certain width and this is referred to as the **Class Interval** or **Class Size**. This class interval is very important when it comes to drawing Histograms and Frequency diagrams. All the classes may have the same class size or they may have different classes sizes depending on how you group your data. The class interval is always a whole number.

Calculating Class Interval

Given a set of raw or ungrouped data, how would you group that data into suitable classes that are easy to work with and at the same time meaningful?

The first step is to determine how many classes you want to have. Next, you subtract the lowest value in the data set from the highest value in the data set and then you divide by the number of classes that you want to have:

$$\text{Class Interval} = \frac{\text{Highest Value - Lowest Value}}{\text{number of classes you want to have}}$$

3.2.1 Measures of Central Tendency

The measures of central tendency are different ways of determining or indicating which value from the information is the central value. The different measures of central tendency are:

Arithmetic mean

The mean is the average value of the distribution.

Median

The median is the score of the scale that separates the upper half of the distribution from the lower, that is to say, it divides the series of data in two equal parts.

Mode

The mode is the most repeated value in a distribution.

Measures of Position

Measures of position are different techniques that divide a set of data into equal groups. To determine the measurement of position, the data must be sorted from lowest to highest. The different measures of position are:

Quartiles

The **quartiles** divide the data set into **four equal parts**.

Deciles

The **deciles** divide the data set into **ten equal parts**.

Percentiles

Percentiles divide the data set into **one hundred equal parts**.

3.2.2 Measures of Dispersion

The measures of dispersion report on how far the values of the distribution are from the center. The measures of dispersion are:

Range

The range is the difference between the highest and lowest data of a statistical distribution.

Average Deviation

The average deviation is the arithmetic mean of the absolute values of the deviations from the mean.

Variance

The variance is the arithmetic mean of the squared deviations from the mean.

Standard Deviation

The standard deviation is the square root of the variance.

MEASURES OF DISPERSION: Measures of dispersion express quantitatively the degree of variation or dispersion of values in apopulation or in a sample . Along with measures of central tendency , measures of dispersion are widely used in practice as descriptive statistics . Some measures of dispersion are the standard deviation , theaverage deviation , the range , the interquartile range .
For example, the dispersion in the sample of 5 values (98,99,100,101,102) is smaller than the dispersion in the sample (80,90,100,110,120), although both samples have the same central location - "100", as measured by, say, the mean or the median . Most measures of dispersion would be 10 times greater for the second sample than for the first one (although the values themselves may be different for different measures of dispersion).

It is important from a practical standpoint that measures of dispersion are normally constructed to beshift invariant and scale invariant . If a measure is not scale invariant, for example, then the value of dispersion might depend on the units of measurement. For example, say the value of dispersion of prices of a particular CD-player model across a country is $10. If the measure of dispersion is scale-invariant and you convert all the prices from dollars to cents by multiplying them by 100, then the measure of dispersion will change from 10 (dollars) to 1000 (cents).

3.3 Sample space and events: In probability theory, the **sample space** of an experiment or random trial is the set of all possible outcomes or results of that experiment. A sample space is usually denoted using set notation, and the possible outcomes are listed as elements in the set.

For example, if the experiment is tossing a coin, the sample space is typically the set {head, tail}. For tossing two coins, the corresponding sample space would be {(head,head), (head,tail), (tail,head), (tail,tail)}. For tossing a single six-sideddie,

the typical sample space is {1, 2, 3, 4, 5, 6} (in which the result of interest is the number of pips facing up)

A well-defined sample space is one of three basic elements in a probabilistic model (a probability space); the other two are a well-defined set of possible events (a sigma-algebra) and a probability assigned to each event (a probability measure function).

Multiple sample spaces

For many experiments, there may be more than one plausible sample space available, depending on what result is of interest to the experimenter. For example, when drawing a card from a standard deck of fifty-two playing cards, one possibility for the sample space could be the various ranks (Ace through King), while another could be the suits (clubs, diamonds, hearts, or spades). A more complete description of outcomes, however, could specify both the denomination and the suit, and a sample space describing each individual card can be constructed as the Cartesian product of the two sample spaces noted above (this space would contain fifty-two equally likely outcomes). Still other sample spaces are possible, such as {right-side up, up-side down} if some cards have been flipped in shuffling.

Equally likely outcomes

Flipping a coin leads to a sample space composed of two outcomes that are almost equally likely. Up or down? Flipping a brass tack leads to a sample space composed of two outcomes that are not equally likely.

In some sample spaces, it is reasonable to estimate or assume that all outcomes in the space are equally likely (that they occur with equal probability). For example, when tossing an ordinary coin, one typically assumes that the outcomes "head" and "tail" are equally likely to occur. An implicit assumption that all outcomes in the sample space are equally likely underpins most randomization tools used in common games of chance (e.g. rolling dice, shuffling cards, spinning tops or wheels, drawing lots, etc.). Of course, players in such games can try to cheat by subtly introducing systematic deviations from equal likelihood (e.g. with marked cards, loaded or shaved dice, and other methods).

Some treatments of probability assume that the various outcomes of an experiment are always defined so as to be equally likely. However, there are experiments that are not easily described by a sample space of equally likely outcomes— for example, if one were to toss a thumb tack many times and observe whether it landed with its point upward or downward, there is no symmetry to suggest that the two outcomes should be equally likely.

Though most random phenomena do not have equally likely outcomes, it can be helpful to define a sample space in such a way that outcomes are at least approximately equally likely, since this condition significantly simplifies the computation of probabilities for events within the sample space. If each individual outcome occurs with the same probability, then the probability of any event becomes simply:

$$P(event) = \frac{\text{number of outcomes in event}}{\text{number of outcomes in sample space}}$$

Simple random sample

In statistics, inferences are made about characteristics of a population by studying a sample of that population's individuals. In order to arrive at a sample that presents an unbiased estimate of the true characteristics of the population, statisticians often seek to study a simple random sample— that is, a sample in which every individual in the population is equally likely to be included.

The result of this is that every possible combination of individuals who could be chosen for the sample is also equally likely (that is, the space of simple random samples of a given size from a given population is composed of equally likely outcomes).

Infinitely large sample spaces

In an elementary approach to probability, any subset of the sample space is usually called an event. However, this gives rise to problems when the sample space is infinite, so that a more precise definition of an event is necessary. Under this definition only measurable subsets of the sample space, constituting a σ-algebra over the sample space itself, are considered events. However, this has essentially only theoretical significance, since in general the σ-algebra can always be defined to include all subsets of interest in applications.