# 1 Probability

## Probability and Possibility

We all are familiar with the words, possibility and probability. Though these words seem to convey similar meanings, in reality they do not. Imagine, your greatest ambition is to climb Mount Everest. But you do not know the basics of mountaineering and have not climbed even a hill before. It may still be *possible* for you to climb Mount Everest, if you learn mountaineering techniques and undergo strenuous training in mountaineering. But the *probability* of accomplishing your ambition of climbing Mount Everest is remote. Possibility is the event that can happen in life, whereas the probability is the chance of that happening. In statistical terminology, an event is collection of results or outcomes of a procedure. Probability is the basic of statistics.

Mathematicians developed the 'principle of indifference' over 300 years ago to elucidate the 'science of gaming' (Murphy, 1985). According to Keynes (1921), the 'principle of indifference' asserts that "if there is no known reason for predicating of our subject one rather than another of several alternatives, then relatively to such knowledge the assertions of each of these alternatives have an equal probability." In other words, if you have no reason to believe the performance of drug A is better than B, then you should not believe that drug A is better than B.

The two approaches to probability are classical approach and relative frequency approach. In classical approach, the number of successful outcomes is divided by the total number of equally likely outcomes. Relative frequency is the frequency of an event occurring in large number of trials. For example, you flip a coin 1000 times and the number of occurrences of *head up* is 520. The probability of *head up* is 520/1000=0.52.

Both the classical and frequency approaches have some drawbacks. Because of these drawbacks, an axiomatic approach to probability has been suggested by mathematicians (Spiegel *et al.*, 2002).

However, in pharmacology and toxicology experiments, relative frequency approach proposed by Mises and Reichenbach (Carnap, 1995) works well.

We shall understand probability a bit more in detail by working out examples.

## Probability—Examples

Let us try to define a probability with regard to frequency approach. The probability of an occurrence for an event labeled A is defined as the ratio of the number of events where event A occurs to the total number of possible events that could occur (Selvin, 2004).

Let us understand some basic notations of probability:

*P* denotes probability.

If you toss a coin, only two events can occur, either a *head up* or a *tail up*.

*P(H)* denotes probability of event head is up. You can calculate the probability of head coming up using the formula:

$$P(H) = \frac{\text{Number of times head is up}}{\text{(Number of times head is up+Number of times tail is up)}}$$

Remember, a *head up* and a *tail up* have equal chance of occurring. Ideally you will get a value very close to 50% for *P(H),* if you toss the coin several times.

You roll an unbiased six-sided dice. The total number of outcomes is six, which are equally likely. This means the likelihood of 'any number' coming up is same as 'any other number'. The probability of any number coming up is 1/6. The probability of any two numbers coming up is 2/6.

Let us come back to our example of tossing a coin. The probability of a *head up* is ½ (0.5 or 50%). Now you flip the coin twice. The probability of a *head up* both times is ½ x ½ = ¼.

### *Mutually exclusive events*

While you toss a coin either a *head up* or a *tail up* occurs. When the event *head up* occurs, the event *tail up* cannot occur and *vice versa*; one event precludes the occurrence of the other. In this example, *head up* or *tail up* that occurs while tossing a coin is a mutually exclusive event.

### *Equally likely events*

Occurrence of *head up* or *tail up* is an equally likely event when you toss a fair coin. This means $P(H) = P(T)$, where $P(H)$ denotes probability of event *head up* and $P(T)$ denotes probability of event *tail up*.

## Probability Distribution

Let us try to understand probability distribution with the help of an example. You flip a coin twice. In this example the variable, $H$ is number of heads that results from flipping the coin. There are only 3 possibilities:

$H = 0$

$H = 1$

$H = 2$

Let us calculate the probabilities of the above occurrences of *head up*.

The probability of not occurring a *head up* in both the times ($H$=0) =0.25

The probability of occurring a *head up* in one time ($H$=1) = 0.5

The probability of occurring a *head up* in both times ($H$=2) = 0.25

0.25, 0.5 and 0.25 are the probability distribution of $H$.

## Cumulative Probability

A cumulative probability is a sum of probabilities. It refers to the probability that the value of a random variable falls within a specified range.

You toss a dice. What is the probability that the dice will land on a number that is smaller than 4? The possible 6 outcomes, when a dice is tossed are 1, 2, 3, 4, 5 and 6.

The probability that the dice will land on a number smaller than 4:

$$\mathrm{P}(X < 4\ ) = \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3) = 1/6 + 1/6 + 1/6 = 1/2$$

The probability that the dice will land on a number 4 or smaller than 4:

$$P(X \leq 4) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 2/3$$

Cumulative probability is commonly used in the analysis of data obtained from pharmacological (Kuo *et al.*, 2009; Rajasekaran *et al.*, 2009) and toxicological experiments.

## Probability and Randomization

In order to evaluate the efficacy of an anti-diabetic drug in rats, twenty rats are administered streptozotocin to induce diabetes. The blood sugar of individual rats is measured to confirm induction of diabetes. You find that 13 rats have blood sugar >250 mg/dl and remaining 7 rats have blood sugar <200 mg/dl. The 20 rats are then distributed randomly in two equal groups (Group 1 and Group 2). You want to treat the Group 1 (control group) with the vehicle alone and the Group 2 (treatment group) with the drug.

Initiate randomization by picking up a rat without any bias and place it in Group 1.

The probability of picking up a rat having blood sugar >250 mg/dl = 13/20 = 65%

The probability of picking up a rat having blood sugar <200 mg/dl = 7/20 = 35%

Assign 10 rats to Group 1 and then the remaining to Group 2. It is most likely that you will have more rats with blood sugar >250 mg/dl in Group 1.

Remember that both the groups are physiologically and metabolically different, because it is most likely that more number of rats in Group 1 will have blood sugar >250 mg/dl and more number of rats in Group 2 will have blood sugar <200 mg/dl. It is unlikely that the experiment with these groups will yield a fruitful result. Randomization is very important in animal studies. We shall be discussing more on randomization of animals in pharmacological studies in later chapters.

## References

Carnap, R. (1995): Introduction to the Philosophy of Science. Dover Publications, Inc., New York, USA.

Keynes, J.M. (1921): A Treatise on Probability. Macmillan, London, UK.

Kuo, S.P., Bradley, L.A. and Trussell, L.O. (2009): Heterogeneous kinetics and pharmacology of synaptic inhibition in the chick auditory brainstem. J. Neurosci., 29 (30), 9625–9634.

Rajasekaran, K., Sun, C. and Bertram, E.H. (2009): Altered pharmacology and GABA-A receptor subunit expression in dorsal midline thalamic neurons in limbic epilepsy. Neurobiol. Res., 33(1), 119–132.

Murphy, E.A. (1985): A Companion to Medical Statistics. Johns Hopkins University Press, Baltimore, USA.

Selvin, S. (2004): Biostatistics—How It Works. Pearson Education (Singapore) Pte. Ltd., India Branch, Delhi, India.

Spiegel, M.R., Schiller, J.J., Srinivsan, R.A. and LeVan, M. (2002): Probability and Statistics. The McGraw Hill Companies, Inc., USA.

# 2
# Distribution

## History

The most commonly used probability distribution is the normal distribution. The history of normal distribution goes way back to 1700s. Abraham DeMoivre, a French-born mathematician introduced the normal distribution in 1733. Another French astronomer and mathematician, Pierre-Simon Laplace dealt with normal distribution in 1778, when he derived 'central limit theorem'. In 1809 Johann Carl Friedrich Gauss (1777–1855), a German physicist and mathematician, studied normal distribution extensively and used it for analysing astronomical data. Normal distribution curve is also called as Gaussian distribution after Johann Carl Friedrich Gauss, who recognized that the errors of repeated measurements of an object are normally distributed (Black, 2009).

## Variable

We need to understand a terminology very commonly used in statistics, *i.e.,* 'variable'. Variable is the fundamental element of statistical analysis. Variables are broadly classified into categorical (attribute) and quantitative variables. Categorical and quantitative variables are further classified into two subgroups each—Categorical variables into nominal and ordinal, and Quantitative variables into discrete and continuous.

*Nominal variable*: The key feature of nominal variables is that the observation is not a number but a word (example—male or female, blood types). Nominal variables cannot be ordered. It makes no difference if you write the blood types in the order A, B, O, AB or AB, O, B, A.

*Ordinal variable*: Here the variable can be ordered (ranked); the data can be arranged in a logical manner. For example, intensity of pain can be ordered as—mild, moderate and severe.

*Discrete variable*: Discrete variable results from counting. It can be 0 or a positive integer value. For example, the number of leucocytes in a µl of blood.

*Continuous variable*: Continuous variable results from measuring. For example, alkaline phosphatase activity in a dl of serum.

The variables can be independent and dependent. In a 90 day repeated dose administration study you measure body weight of rats at weekly intervals. In this situation week is the independent variable and the body weight of the rats is the dependent variable.

## Stem-and-Leaf Plot

Stem- and Leaf-Plot (Tukey, 1977) is an elegant way of describing the data (Belle *et al.*, 2004). Let us construct a stem-and-leaf plot of the body weight of rats given in Table 2.1.

**Table 2.1.** Body weight of rats

| Body weight (g) |
|---|
| 132, 139, 134, 141, 145, 141, 140, 166, 154, 165, 145, 158, 162, 148, 154, 146, 154, 148, 140, 153, 154 |

Now arrange the data in an ascending order as given in Table 2.2:

**Table 2.2.** Body weight of rats arranged in an ascending order

| Body weight (g) |
|---|
| 132, 134, 139, 140, 140, 141, 141, 145, 145, 146, 148, 148, 153, 154, 154, 154, 154, 158, 162, 165, 166 |

Stem-and-leaf plot of the above data is drawn in Figure 2.1:

| Stem | Leaf |
|---|---|
| 13 | 2 4 9 |
| 14 | 0 0 1 1 5 5 6 8 8 |
| 15 | 3 4 4 4 4 8 |
| 16 | 2 5 6 |

**Figure 2.1.** Stem- and- Leaf plot

Each data is split into a "leaf" (last digit) and a "stem" (the first two digits). For example, 132 is split into 13, which forms the 'stem' and 2, which forms the 'leaf'. The stem values are listed down (in this example 13, 14, 15 and 16) and the leaf values are listed on the right side of the stem values.

The Stem-and-leaf plot provides valuable information on the distribution of the data. For example, the plot indicates that more number of the animals is having body weight in the 140 g range, followed by the 150 g range.

## Box-and-Whisker Plot

Another way of describing the data is by constructing a box-and-whisker plot. The usefulness of box-and-whisker plot is better understood by learning how to construct it. For this purpose we shall use the same body weight data given in Table 2.1. As we have done for plotting the stem-and-leaf plot, arrange the data in an ascending order (Table 2.2). The first step in constructing a box-and-whisker plot is to find the median. You will learn more about median in Chapter 3.

The median of the data given in Table 2.2 is the 11th value, *i.e.,* 148 (see Table 2.3).

**Table 2.3.** Median value of the body weight data

**Median** ⇩

132, 134, 139, 140, 140, 141, 141, 145, 145, 146, **148**, 148, 153, 154, 154, 154, 154, 158, 162, 165, 166

The median divides the data into 2 halves (a lower and an upper half). The lower half consists of a range of values from 132 to 146 and the upper half consists of a range of values from 148 to 166 (see Table 2.4).

**Table 2.4.** Median value of the lower and upper quartiles

Median

Lower half ← ⇩ → Upper half

132, 134, 139, 140, 140, 141, 141, 145, 145, 146, **148**, 148, 153, 154, 154, 154, 154, 158, 162, 165, 166

Next step is to find the median of lower half and upper half:

| | |
|---|---|
| Median of the lower half | = (140+141)/2 = 140.5 |
| Median of the upper half | = (154+154)/2 = 154.0 |

Median of the lower half is also called as 'lower hinge' or ' lower quartile' and the median of the upper half as 'upper hinge' or ' upper quartile'. The term, quartile was introduced by Galton in 1882 (Crow, 1993). About 25% of the data are at or below the 'lower hinge', about 50% of the data are at or below the median and about 75% of the data are at or below the 'upper hinge'.

Next step is calculation of 'hinge spread', the range between lower and upper quartiles:

Hinge spread = 154.0–140.5 = 13.5

Hinge spread is also called as inter-quartile range (IQR).

Now, we need to determine 'inner fence'. The limits of 'inner fence' are determined as given below:

Lower limit of 'inner fence' = Lower hinge–1.5 x hinge spread
= 140.5–(1.5x13.5)= 120.25
Upper limit of 'inner fence' = Upper hinge+1.5 x hinge spread
= 154.0+(1.5x13.5)= 174.25

We now have all the required information to construct the 'whiskers'. The lowest body weight data observed (see Table 2.4) between 140.5 g and 120.25 g is 132 g and the highest body weight data observed between 154.0 g and 174.25 g is 166 g. Hence, the whiskers are extended from the lower quartile to 132 g and from the upper quartile to 166 g.

Box-and-whisker plot of the data (Table 2.1) is given in Figure 2.2.

The box-and-whisker plot is based on five numbers: the least value, the lower quartile, the median, the upper quartile and the greater value in a data set.

If the data are normally distributed:

1. the median line will be in the centre of the box dividing the box into two equal halves
2. the whiskers will have similar lengths
3. observed values will scarcely be outside the 'inner fence'.

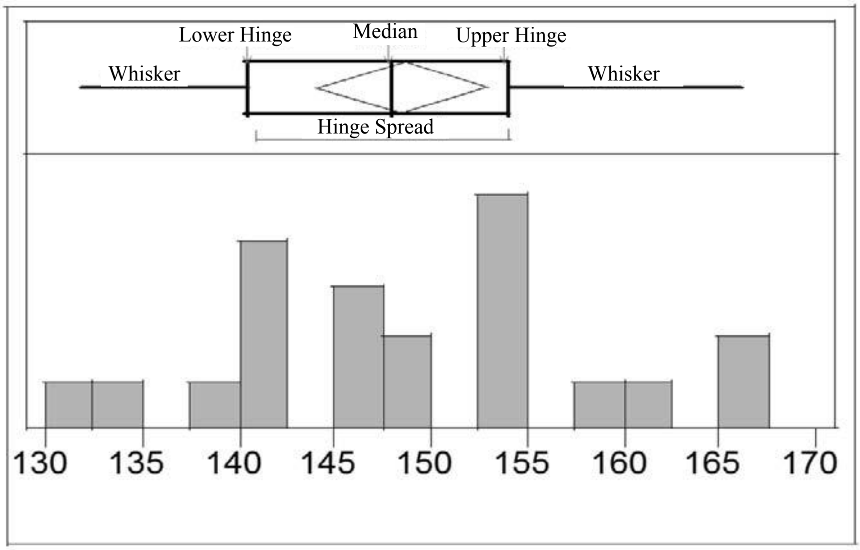

**Figure 2.2.** Box-and-whisker plot of the data

It is important to examine whether the data are normally distributed before applying a statistical tool. We shall learn more about this in later chapters.

## References

Belle, G,V., Fisher, L.D., Heagerty, P.J. and Lumley, T. (2004): Biostatistics-A Method for the Health Sciences. 2nd Edition, Wiley Interscience, New Jersey, USA.

Black, K. (2009): Business Statistics: Contemporary Decision Making. 6th Edition. John Wiley and Sons, Inc., USA.

Crow, J.F. (1993): Francis Galton: Count and measure, measure and count. Genetics, 135, 1–4.

Tukey, J.W. (1977): Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts, USA.

# 3

# Mean, Mode, Median

## Average and Mean

Average and mean are interchangeably used in everyday life. Average is the synonym for the central tendency. There are various types of central tendencies, such as mean, mode and median.

## Mean

The procedure for calculating mean is very simple; sum of all individual observations divided by the sum of number of observations. There are several types of means, such as arithmetic mean, geometric mean and harmonic mean. Let us work out the example given in Table 3.1 to familiarise the reader with the calculation procedure of these means.

**Table 3.1.** Calculation of arithmetic mean of body weight of rats

| Body weight (g) | Sum |
|---|---|
| 132, 139, 134, 141, 145, 141, 140, 166, 186, 183 | 1507 |

In statistics, the number of observations is denoted by the letter *N* or *n* (both cases). Number of observations is also called the sample size. The Greek letter Σ (uppercase only) is used to denote sum. Mean is denoted as $\overline{X}$ (X bar).

Mean body weight of above example:

$$\overline{X} = \frac{\sum X}{N} = \frac{1507}{10} = 150.7 \text{ g}$$

Mean in the above example is called the arithmetic mean. Arithmetic mean is sensitive to extreme values in data set. There is a condition for calculating arithmetic mean—the data should fit a normal distribution.

## Geometric Mean

Mathematically geometric mean is defined as the n$^{th}$ root of the product of n numbers. An easy way to calculate the geometric mean is to find the mean of logarithmic values of the data and then to find the antilog of the mean. Steps involved in the calculation of the geometric mean of the body weight data of the rats (Table 3.1) are given in Table 3.2.

**Table 3.2.** Calculation of geometric mean of body weight of rats

| Body weight (g) | | Σ | *N* | $\overline{X}$ |
|---|---|---|---|---|
| Linear scale | 132, 139, 134, 141, 145, 141, 140, 166, 186, 183 | 1507 | 10 | 150.7 |
| Log scale | 2.12, 2.14, 2.13, 2.15, 2.16, 2.15, 2.15, 2.22, 2.27, 2.26 | 21.7 | 10 | 2.17 |

Geometric mean is the antilog of 2.17 = 147.9

If any observed value is 0 or negative, geometric mean cannot be calculated. Geometric mean is very rarely used in pharmacology. However, use of it is witnessed in some pharmacokinetic studies (Schuirmann, 1987).

## Harmonic Mean

Harmonic mean is calculated by finding the mean of the reciprocals of the values and then finding the reciprocal of the mean.

Calculation procedure of the harmonic mean of the data given in Table 3.1 is described in Table 3.3:

**Table 3.3.** Calculation of harmonic mean of body weight of rats

| Body weight (g) | | Σ | *N* | $\overline{X}$ |
|---|---|---|---|---|
| Linear scale | 132, 139, 134, 141, 145, 141, 140, 166, 186, 183 | 1507 | 10 | 150.7 |
| Reciprocal | 0.0076, 0.0072, 0.0075, 0.0071, 0.0069, 0.0071, 0.0071, 0.0060, 0.0054, 0.0055 | 0.0673 | 10 | 0.0067 |

Harmonic mean = 1/0.0067 = 148.5 g

Unlike arithmetic mean, the harmonic mean is not influenced by the extreme values. Harmonic mean has limited use in pharmacology. A pharmacokinetic study carried out with cyclosporine-A revealed that there was little use of harmonic mean to describe the central tendency (Lum *et al.*, 1992). However, Iwamoto *et al.* (2008) used harmonic mean to evaluate the central tendency of pharmacokinetics in a clinical study conducted with Raltegravir in healthy subjects.

## Weighted Mean

In an experiment designed to administer a drug to rats, 15 rats were randomly assigned to 3 cages (Cage 1, Cage 2 and Cage 3), each cage consisting of 5 rats. At the end of 2 weeks of the drug administration in Cages 1 and 2, two rats each survived, whereas in Cage 3 all the five rats survived. The body weight of the survived rats is given in Table 3.4.

**Table 3.4.** Body weight (g) of rats in 3 Cages at the end of 2 weeks following a drug administration

| Cage | N | Mean (g) |
|---|---|---|
| 1 | 2 | 119 |
| 2 | 2 | 125 |
| 3 | 5 | 134 |

Let us calculate the grand mean:

(119+125+134)/3 = 126 g. But there is a problem with this grand mean. It is very close to the mean value of 2 animals of Cage 2, but does not seem to represent the body weight of animals in Cages 1 and 3. Hence calculating grand mean by the above method is not advisable. In such situation, calculating weighted mean value may be the right approach.

Weighted Mean = [(2x119)+(2x125)+(5x134)]/(2+2+5) =1158/9 = 128.7 g

## Mode

The mode is the value which appears the most in the data. It is usually calculated for discrete data (Belle *et al.*, 2004). There can be more than one mode, if there is more than one value which appears the most.

In the following data,

130, 140, 140, 150, 140, 160, 140, 110, 120

The mode is 140 (140 appears 4 times in the data).

In the following data,

130, 140, 140, 150, 140, 160, 140, 110, 120, 130, 130

There are two modes, 140 and 130 (140 appears 4 times in the data, whereas 130 appears 3 times).

## Median

To measure the central tendency, median is second in popularity to mean (Rosner, 2006). Median is also termed as 0.50 quantile. Another term for the median is the 50th percentile.

The first step to calculate the median is to rank the values from lowest to the highest. If the number of data values is odd, add 1 to the number of data values and divide that by 2. For example, if there are 9 sample values, divide (9+1) by 2. The median is the 5th ranked value. If the number of data values is even, again add 1 to the number of data values and divide that by 2. For example, if there are 10 sample values, divide (10+1) by 2 to get 5.5. Median is the mean of the 5th and 6th ranked values.

The second situation where the median is useful is when it is impractical to measure all of the values, such as when you are measuring the time until something happens. Survival time is a good example of this; in order to determine the mean survival time, you have to wait until every individual is dead, whereas to determine the median survival time you do not need to wait until every individual is dead; you need to wait only until half the individuals are dead.

Mean, mode and median are theoretically the same for the data collected from a symmetrical distribution (Lemma, 2008). Median and mode are not affected by the extreme values (outliers). One disadvantage of mode is that it does not include all the data for the analysis. Though mean and median are commonly used in statistical analysis of pharmacological and toxicological data, the use of mode is not very common.

## References

Belle, V.G., Fisher, L.D., Heagerty, P.J. and Lumley, T. (2004): Biostatistics—A Methodology for the Health Sciences. John Wiley & Sons, Inc., New Jersey, USA.

Iwamoto, M., Wenning, L.A., Petry, A.S., Laethem, M., De Smet, M., Kost, J.T., Merschman, S.A., Strohmaier, K.M., Ramael, S., Lasseter, K.C., Stone, J.A., Gottesdiener, K.M. and Wagner, J.A. (2008): Safety, tolerability, and pharmacokinetics of Raltegravir after aingle and multiple doses in healthy subjects. Clin. Pharmacol. Therapeutics, 83, 293–299.

Lemma, A. (2008): Introduction to the Practice of Psychoanalytic Psychotherapy. John Wiley & Sons Ltd., Chichester, UK.

Lum, B.L., Tam, J., Kaubisch, S. and Flechner, S.M. (1992): Arithmetic versus harmonic mean values for cyclosporin-A pharmacokinetic parameters. J. Clin. Pharmacol., 32 (10), 911–014.

Rosner, B. Fundamentals of Biostatistics. 6th Edition, Thomson Brooks/Cole, Belmont, USA.

Schuirmann, D.J. (1987): A comparison of the two one-sided tests procedure and the power approach for assessing the bioequivalence of average bioavailability. J. Pharmacokinetics Biopharm., 15, 657–680.