CHAPTER 1

BASIC CONCEPTS

Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected on any sphere of inquiry. Knowledge on statistical concepts is desired for the development, production, evaluation and marketing of various pharmaceutical dosage forms. Statistics depends on the measurement of continuous and discrete variables. Discontinuous or discrete variable is lacking continuity and it may take several values. For example, the number of students in several classes in a school may vary from 50 to 100, taking several discrete integer values. A continuous variable takes on values which can be measured to any depth. Statistical analysis requires the generation of data and subsequent statistical treatment. During the collection of data, there is a possible presence of either determinate errors or indeterminate errors. Determinant errors are constant errors; unsuspected and corrected once they are identified. They are present in each case. For example, presence of impurities in active pharmaceutical active ingredient is inevitable and depends upon the raw materials used and the conditions employed for the synthesis. The affect of these impurities on the quality, performance and safety of the finished dosage form may vary. Such errors must be eliminated before studying the significance of indeterminate errors. Indeterminate errors occur by accident or chance and vary from one experiment to another experiment. They cannot be corrected due to the fluctuations that occur in all measurements.

Determinate and indeterminate errors can be explained by taking an experiment "determination of the strength of the bulk density of the sample." If the errors in density measurement are due to the transfer of a sample in to the container or in adequate working of the bulk density apparatus, then those errors are designated as determinate errors. However the errors may also due to the fluctuations in temperature during the repeated experimentation and hence they are termed as indeterminate errors.

The outcome of statistical decision depends upon the selected sample. Statistics involves the examination of the relatively small sample and making decision about the population. The sample should be carefully chosen, subsequently modified (if necessary) so that the sample has the same characteristics as the bulk material. During the dissolution

of tablet dosage form, if the samples are collected from different positions of the dissolution rate testing apparatus at different time intervals then the statistician cannot find the source of variation in amount of drug dissolved because it may be due to the differences in sample position or due to the time. In this experiment, the data is graphically represented by plotting a graph in between time (X-axis) and amount of drug dissolved (Y-axis). The validity or reproducibility of the parameter represented on Y-axis depends upon the sample collected at regular intervals of dissolution study.

1.1 Types of Statistics

The statistical procedures are divided into descriptive and inferential. Descriptive statistics describes the observed/collected data. It represents the procedure used to organize, categorize and summarize the data collected from the experiment. Such statistics accurately, efficiently and effectively represent the observed outcome in a clear and understandable manner. Examples include average of outcome, pie charts, bar graphs or other visual representations.

Inferential statistics make predictions about a large population based on a sample from those populations. It uses the statistical tests such as ANOVA, t-test, chi-square test.

Inferential statistics involves the following steps:

1. Identification/ establishment of a research problem
2. Establishment of a hypothesis
3. Selection of suitable statistical test
4. Selection of data
5. Collection of required data
6. Performing the statistical tests
7. Making the decision based on the test results.

1.2 Parameters and Statistics

Statistical data usually involves a relatively small portion of data. This data is numerically manipulated to take decisions about the population. Parameters are characteristics of population and statistics are characteristics of samples. Parameters are usually represented by Greek symbols (μ, σ, ψ) and statistics are denoted by letters ($\overline{X}$, S^2, and r).

Example:

A pharmaceutical company produces 20,000 Paracetamol tablets per batch. To meet pharmacopoeial standards, these tablets are subjected to various quality control tests such as weight variation, drug content (assay), hardness, disintegration time, friability and dissolution tests. To determine population parameter, all the formulated tablets should be weighed and from this either average weight or range of weights are calculated. However

it is very tedious. If all the tablets are subjected to disintegration tests to determine the average disintegration time of 20,000 tablets (whole batch), it destroys all tablets (destructive testing) which is not desirable. Thus calculation of population parameters may be either impractical or impossible. So performance of a statistical analysis is preferred to make a statement about population.

Statistical analysis of the average weight can be explained as follows:

20 tablets should be collected by following appropriate sampling procedure. The average weight of 20 tablets is calculated. With some statistical manipulation of this data, we can get the actual average weight of entire batch of 20,000 tablets at the selected confidence interval (usually 95%). If the sample is carefully and accurately collected then the observed/claim (descriptive statistic) is 100% accurate. Similarly disintegration and dissolution tests can be performed on 6 randomly collected tablets.

1.3 Sampling and Independent Observations

Random sampling from a population is essential for any inferential test. In random sampling, each individual member/observation has an equal chance of being selected for the sample. For example, in a batch of 20,000 tablets each tablet is having theoretically an equal chance for selection. The second requirement of inferential testing is that the observations are independent of each other. The sample should not affect the outcome of any other sample. For example, the disintegration time of the 6^{th} tablet is not influenced by the disintegration time of other 5 tablets which are subjected to disintegration test. An independent result must represent an outcome not dependent on the result of any other observation.

1.4 Variables and Variation

A variable is any measurable property that can vary from one observation to another. These variables are identified and controlled in a well designed process/experiment. The variables shall be controlled to formulate a product having consistent quality. During the developmental stage, all probable variables which may directly/indirectly influence the quality, safety and effectiveness of a drug/drug delivery system should be identified. In process optimization studies, the process variables such as temperature, flow rate, speed of coating pan etc, and formulation variables like concentration of excipients should be optimized. The assay of a drug is influenced by the variables such as the method employed (gravimetric/volumetric/spectrophotometric), instrument used, analyst, sample, unidentified, uncontrollable errors (noise). In validation of a process, the process is splitted into different steps and the variables affecting each step is clearly defined.

The observation could have an infinite number of variables. For example, the hardness (observation) of a tablet is influenced by the variables like binder; moisture content of granules, granule size and size distribution, compression force … etc. The variables should be controlled to achieve reproducible results.

The number of possible variables is limited only on imagination. Variables may be discrete or continuous. The variable should be thoroughly determined before selecting the appropriate statistical test.

1.4.1 Discrete Variables

A discrete variable is characterized by jumps and gaps between one value and the next. These variables are placed in specific, finite number of categories or classifications. Variables that can only take on a finite number of values are called "discrete variables. All qualitative variables are discrete. Some quantitative variables are discrete, such as performance rated as 1, 2, 3, 4, or 5, or temperature rounded to the nearest degree. A type of variable, also called a categorical or nominal variable, which has a finite number of possible values that do not have an inherent order. For example, hair color would be a discrete variable, because it can only have a limited number of possibilities, such as red, brown, and black, that does not occur in any particular order.

The solid dosage forms can be categorized as powders, tablets and capsules. The tablets can be classified based on passage or failure of a specific assay or dissolution requirement. These variables are also referred as qualitative, category or nominal. They can be differentiated as above and below the midpoint of distribution. Levels of discrete variable are exhaustive when the variable accounts for all possible outcomes. For example, capsule size 0-5 are exhaustive for humans; whereas capsule no 2 is not exhaustive for the human use because there are capsules other than the size 2. The levels of discrete variable may be set as mutually exclusive when the categories are not having the same number in common with each other. For example, the weight variation tolerances for uncoated tablets are categorized as average weight < 130 mg, 130-324 mg and more than 324 mg. Such average weights are not mutually exclusive because the average weights 130 and 324 mg are included in two of the discrete groups. To represent a mutually exclusive and exhaustive set of categories, the average weight group should be as follows: less than 130 mg, 131-324 mg and greater than 325 mg.

1.4.2 Continuous Variables

Continuous variables also referred as quantitative variables. Discrete variables are usually based on counting, whereas continuous variables are involving measurements. Examples include viscosity, surface tension, blood glucose levels, systolic blood pressure. Theoretically a difference can be found in continuous variable value depending upon the sensitivity of the instrument used. For example, the plasma concentration of a sample may be analyzed by UV and HPLC techniques. The concentration of a drug detected from UV method may be 24 μg/mL whereas the concentration observed from HPLC method may be 24.6 μg/mL. Similarly in analytical balance of 10 mg sensitivity, the weight of the sample is 70 mg and the same sample weight may be 72, 72.4, 72.46 mg respectively from the balances having a sensitivity of 1 mg, 0.1 mg and 0.01 mg respectively.

Therefore, any measurement result for a continuous variable actually represents an interval from half a unit below or above the value.

Occasionally continuous variables cannot be easily measured but can be ranked in order of magnitude. The experimental results may be expressed as above or below a certain value (like midpoint). For example the temperature is usually represented as high, medium or low. Similarly, the pharmaceutical dosage forms are evaluated for a particular quality control test and the results are mentioned as passes/fails. In analgesic studies, pain severity is assigned with the scores like no pain = 0, slight pain = 1, moderate pain = 2 and severe pain = 3. Similarly the scores can be allotted to the taste of the drug.

1.4.3 Measurement of Variables

The variables are measured in terms of scales. In case of **nominal scale**, observations are classified qualitatively based on a characteristic. They differ in nature and cannot be arranged in a meaningful order. E.g.: medicines in a shop are arranged into tablets, capsules, liquids etc.

In **ordinal scale,** the items in a population are divided into two or more categories and the categories are arranged on a scale by some attribute, that is a ordinal scale. The difference between the two observations is not considered but their relative position is considered based on ranking with numbers. E.g. medicines in a shop are arranged in order of their acceptance by patients: very well accepted, accepted, tolerated, rejected. Such scale is not precise but extremely important in non parametric tests. Nominal and ordinal scales are also referred as non-metric scales.

The third type of scale is **interval scale** where the difference between each level of scale is equal. For example, the particle size is determined by optical microscopy and the observed data is represented as the number of particles belongs to a particular interval.

The fourth type of scale is **ratio scale** where the numerical value of one observation is divided with the other observation value. E.g.: the AUC value of formulation A is 100 μg ml/hr and the second formulation AUC is 50 μg mL/hr. Then the data can be expressed as the AUC value observed from the first formulation is twice the AUC value of second formulation. The interval and ratio scales are referred as metric scales.

1.4.4 Independent and Dependent Variables

Independent variables are qualitative (either continuous or discrete). The independent variables should be identified before conducting an experiment (predictor variable) and controlled during experiment. The variables which are measured against the independent variable are called dependent variables. During research work, the dependent variables are measured and compared based on different levels or categories of independent variables. For example the hardness of a tablet is the independent variable, if all the variables influencing the hardness are properly controlled. The hardness of such tablets may be tested with three different hardness testers (Monsanto, Pfizer and Strong Cobb).

The variations in hardness observed from these testers are dependent variable as it is influenced by the device. The results are different even though the hardness of the tablet is same. Such extraneous factors which may influence the dependent variable are known as confounding or nuisance variable.

1.5 Error Analysis

No measurement of a physical quantity can be entirely accurate. It is important to know, therefore, just how much the measured value is likely to deviate from the unknown, true, value of the quantity. The art of estimating these deviations should probably be called uncertainty analysis, but for historical reasons is referred to as error analysis. Even under constant experimental conditions (same operator, same tools, and same laboratory), repeated measurements of series of identical samples always lead to results which differ among themselves and from the true value of the sample. Therefore, quantitative measurements cannot be reproduced with absolute reliability.

Absolute and Relative Errors

The absolute error in a measured quantity is the uncertainty in the quantity and has the same units as the quantity itself. For example if you know a weight is 0.628 mg ± 0.003 mg, the 0.003 mg is an absolute error. The relative error (also called the fractional error) is obtained by dividing the absolute error in the quantity by the quantity itself. The relative error is usually more significant than the absolute error. For example a 1 mg error in the weight of a potent medicaments such as digitoxin is probably more serious than a 1 mg error in a drug having high therapeutic index like paracetamol. Note that relative errors are dimensionless. When reporting relative errors it is usual to multiply the fractional error by 100 and report it as a percentage.

Random Errors

Random errors are the components of measurement errors that vary in an unpredictable manner in replicated measurements. They reflect the distribution of the results around the mean value of the sample which are randomly distributed to lower and higher values. Random errors characterize the reproducibility of measurements, and, therefore, their precision. They are caused by effects such as measuring techniques (e.g. noise), sample properties (e.g. in homogeneities), and chemical effects (e.g. equilibrium). Even under carefully controlled conditions random errors cannot, in principle, be avoided, they can only be minimized and evaluated with statistical methods.

For example, if you were to measure the amount of drug decomposed from a sample at same time interval but different trials, you would find that your measurements were not always the same. The main source of these fluctuations would probably be the difficulty of judging exactly when the sample should be collected, and time interval maintained for analysis of the sample. Since you would not get the same value each time that you try to

measure it, your result is obviously uncertain. There are several common sources of such random uncertainties in the type of experiments that you are likely to perform:

- Uncontrollable fluctuations in initial conditions in the measurements. Such fluctuations are the main reason why, no matter how skilled the technician, no individual can create a formulation with same quality.
- Limitations imposed by the precision of your measuring apparatus, and the uncertainty in interpolating between the smallest divisions. The precision simply means the smallest amount that can be measured directly. A typical burette is subdivided into 0.1 milliliters and its precision is thus 0.1 milliliter.
- Lack of precise definition of the quantity being measured. The length of a table in the laboratory is not well defined after it has suffered years of use. You would find different lengths if you measured at different points on the table. Another possibility is that the quantity being measured also depends on an uncontrolled variable. (The temperature of the object for example).
- Sometimes the quantity you measure is well defined but is subject to inherent random fluctuations. Such fluctuations may be of a quantum nature or arise from the fact that the values of the quantity being measured are determined by the statistical behavior of a large number of particles.

Estimating Random Errors

There are several ways to make a reasonable estimate of the random error in a particular measurement. The best way is to make a series of measurements of a given quantity (say, x) and calculate the mean x, and the standard deviation σx from this data. We become more certain that $\overline{x}$, is an accurate representation of the true value of the quantity x, the more we repeat the measurement. A useful quantity is therefore the standard deviation of the mean σx defined as $\sigma x \equiv \sigma x / N$. The quantity σx is a good estimate of our uncertainty in $\overline{x}$. Notice that the measurement precision increases in proportion to N as we increase the number of measurements. Not only have you made a more accurate determination of the value, you also have a set of data that will allow you to estimate the uncertainty in your measurement. The calculation of standard deviation, relative standard deviation are explained in next chapter.

Systematic Errors

Systematic errors arise from a flaw in the measurement scheme which is repeated each time a measurement is made. If you do the same thing wrong each time you make the measurement, your measurement will differ systematically (that is, in the same direction each time) from the correct result. Some sources of systematic error are:

- Errors in the calibration of the measuring instruments.
- Incorrect measuring technique: For example, one might make an incorrect scale reading because of parallax error.

- Bias of the experimenter. The experimenter might consistently read an instrument incorrectly, or might let knowledge of the expected value of a result influence the measurements.

It is clear that systematic errors do not average to zero if you average many measurements. If a systematic error is discovered, a correction can be made to the data for this error. If you measure a absorbance of a drug substance from a spectrophotometer that later turns out to have a 0.08 offset, you can correct the originally determined absorbance by this amount and eliminate the error. Although random errors can be handled more or less routinely, there is no prescribed way to find systematic errors. One must simply sit down and think about all of the possible sources of error in a given measurement, and then do small experiments to see if these sources are active. The goal of a good experiment is to reduce the systematic errors to a value smaller than the random errors. For example a pipette should have been manufactured such that the milliliter markings are positioned much more accurately than one milliliter.

Systematic deviations (errors) displace the results of analytical measurements to one side, to higher or lower values which lead to false results. Such an effect is described by the performance characteristic trueness, which is defined as "the closeness of agreement between the expectation of a test result or measurement result and a true value". Measurement trueness is not a quantity and cannot be expressed numerically, but measures for closeness of agreement can be given. Thus the trueness can be quantified as bias which is defined as the difference between the average of several measurements on the same sample $\bar{x}$ and its (conventionally) true value μ:

$$\text{Bias } (x) = \bar{x} - \mu$$

or if expressed as a percentage Bias

$$\text{Bias } \% = \frac{\left(\bar{x} - \mu\right)}{\mu} \times 100$$

or as the recovery ratio

$$\text{Rr } \% = \frac{\bar{x}}{\mu} \times 100$$

Systematic errors are always combined with random errors.

1.6 Applications of Statistics in Pharmacy

- Analytical statistics such as correlation, regression is used to establish the functional relation between variables like time and amount of drug dissolved.
- Descriptive statistical parameters such as mean and standard deviation are used to express the data generated from the replicated experiments to confirm the reproducibility.

- Inferential statistical tests such as t test, ANOVA are used in bioavailability and bioequivalence studies.
- Applied statistics is used quality control.
- Statistical treatment of data is required for validation of a process.
- Statistical quality control charts are used to determine the quality of the product.
- Statistical knowledge is essential for validation of analytical technique.
- A useful tool in the design of experiments like bioavailability studies.
- Useful in interpretation of experimental results.
- Statistical usage is recommended by the drug regulatory agencies.
- Collection of true representative sample from the bulk needs the concept of statistics.