Unit 1

Statistical analysis

Dr. Shashi Kiran Misra
School Of Pharmaceutical Science

Statistical analysis

It's the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends.

Statistics are applied every day – in research, industry and government – to become more scientific about decisions that need to be made.

For example:
Manufacturers use statistics to weave quality into beautiful fabrics, to bring lift to the airline industry and to help guitarists make beautiful music.

Researchers keep children healthy by using statistics to analyze data from the production of viral vaccines, which ensures consistency and safety.

Communication companies use statistics to optimize network resources, improve service and reduce customer churn by gaining greater insight into subscriber requirements.

Government agencies around the world rely on statistics for a clear understanding of their countries, their businesses and their people.

What Is Confidence Interval?

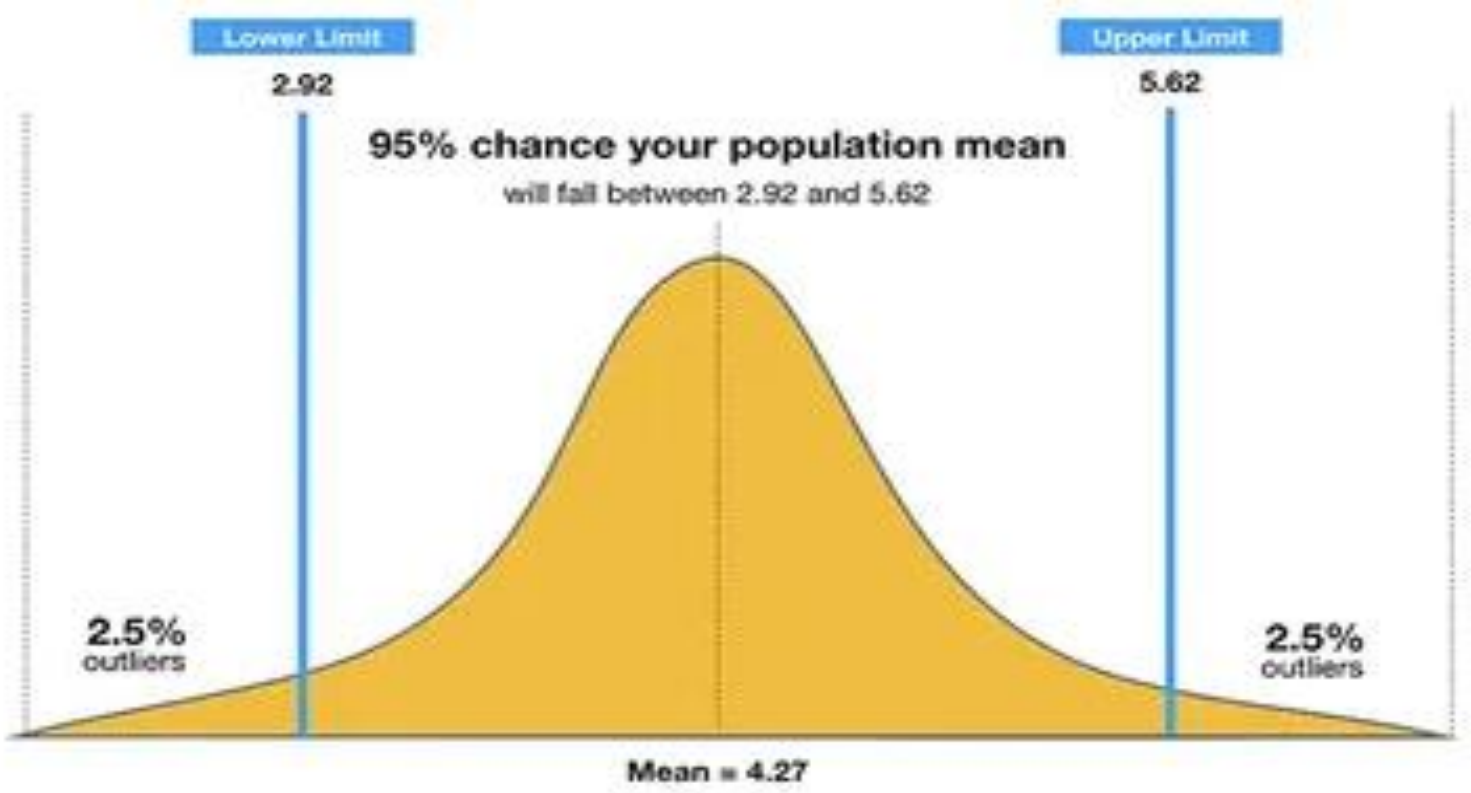
A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

A confidence interval displays the probability that a parameter will fall between a pair of values around the mean.

Confidence intervals measure the degree of uncertainty or certainty in a sampling method.

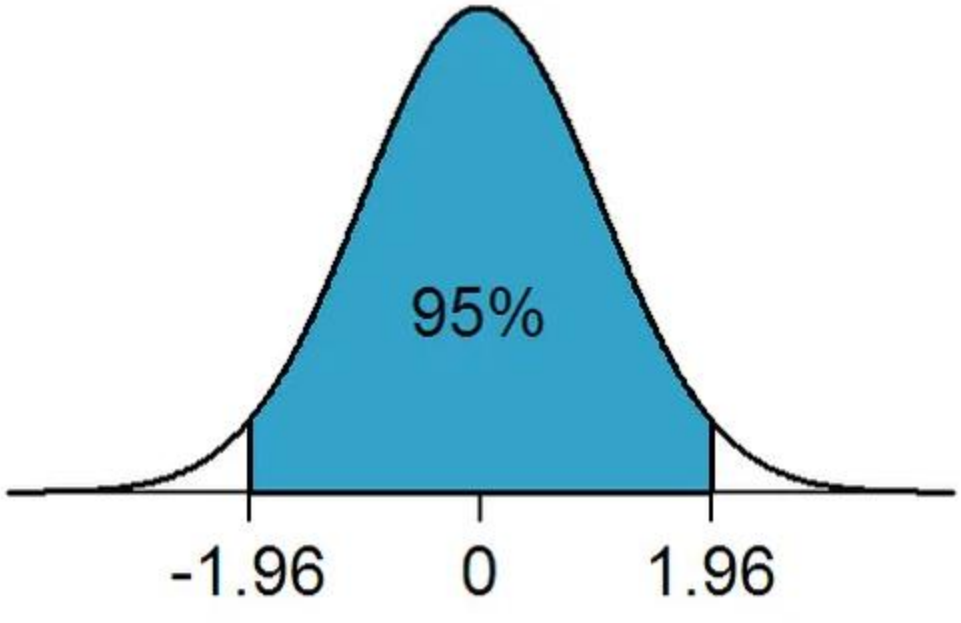
They are most often constructed using confidence levels of 95% or 99%.

The 95% confidence interval is a range of values that you can be 95% confident **contains the true mean of the population**. Due to natural sampling variability, the sample mean (center of the CI) will vary from sample to sample.



The confidence is in the method, not in a particular CI. If we repeated the sampling method many times, approximately 95% of the intervals constructed would capture the true population mean.

Therefore, as the sample size increases, the range of interval values will narrow, meaning that you know that mean with much more accuracy compared with a smaller sample. We can visualize this using a normal distribution.



For example, the probability of the population mean value being between -1.96 and +1.96 standard deviations (z-scores) from the sample mean is 95%.

Accordingly, there is a 5% chance that the population mean lies outside of the upper and lower confidence interval (as illustrated by the 2.5% of outliers on either side of the 1.96 z-scores).

Confidence intervals measure the degree of uncertainty or certainty in a sampling method. They can take any number of probability limits, with the most common being a 95% or 99% confidence level. Confidence intervals are conducted using statistical methods, such as a t-test.

What Is a T-Test

Confidence intervals are conducted using statistical methods, such as a t-test. A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related to certain features.

Calculating a t-test requires three key data values.
They include the difference between the mean values from each data set (called the mean difference),

the standard deviation of each group, and

the number of data values of each group.

Suppose for example, that an analyst is tasked with measuring the pH of ten samples taken from a batch of liquid product. The measurements yield the following pH data:

Sample #	pH
1	6.39
2	6.51
3	6.54
4	6.42
5	6.52
6	6.47
7	6.69
8	6.37
9	6.63
10	6.62

The average pH, in this example is 6.52; the sample standard deviation, is 0.11. The analyst wants to report the estimated mean pH along with a statement of the uncertainty of this estimate.

Using a = 0.05 (confidence level = 95%), the analyst calculates a two-sided limit. For a two-sided limit, the critical value for the *t* distribution is found using one-half of the value for a; that is, 0.05/2 = 0.025.

This value along with the degrees of freedom (*n-1= 9* in this case) will be used to look up the critical value of the *t* distribution.

The critical value in this example is 2.262. Entering this value into the two sided confidence interval formula gives:

$$\bar{x} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}} = 6.52 \pm 2.262 \frac{0.11}{\sqrt{10}} = 6.52 \pm 0.08$$

The lower interval bound in this example is 6.52 – 0.08 = 6.44; the upper bound is 6.52 + 0.08 = 6.60. The interval range is 0.16.

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

Sampling is the selection of a subset of the population of interest in a research study. ... Sampling from the population is often more practical and allows data to be collected faster and at a lower cost than attempting to reach every member of the population.

In the pharmaceutical industry, a sampling plan is a method for deciding whether to accept or reject a batch of materials. The sampling plan sets out what will be tested, when and how it will be tested, and the measurements for determining whether the sample that is being tested will be accepted or rejected

Statistical modeling

An introduction to statistical modeling is pivotal for any data analyst to make sense of the data and make scientific predictions.

In its essence, statistical modeling is a process using statistical models to analyze a set of data. Statistical models are mathematical representations of the observed data.

Statistical modeling methods are a powerful tool in understanding the consolidated data and making generalized predictions using this data.

A statistical model could be in the form of a mathematical equation or a visual representation of the information.

Types of Statistical Models

The different types of statistical models are essentially the statistical methods used for computation. These are the mathematical equations and visual representations that make statistical modeling possible.

Some of them are:

Linear regression
Logistic regression
Cluster analysis
Factor analysis
Analysis of variation (ANOVA)
Chi-squared test
Correlation
Decision trees
Time series
Experimental design
Bayesian theory – Naïve Bayes classifier
Pearson's r
SamplingAssociation rules
Matrix operations
K-nearest neighbor algorithm (k-NN)

Statistical Modeling in Pharma, R, and Excel

Statistical modeling holds an important place in all types of data analysis, making it relevant to various fields of science and industry. This especially holds in the data analytics field, where analysts rely heavily on statistical methods and techniques to interpret and draw conclusions from any given dataset.

Statistical modeling in pharmaceutical research and development

Statistical models are being introduced into the pharmaceutical industry to determine the efficacy of drugs for particular individuals, ensuring that individuals are given the right drugs for optimal response. Statistical techniques are used to filter biomarkers from the data, using which models are developed to predict the groups in which the drugs are most effective.

Statistical modelling in R

Owing to the extensive usage of statistical modeling in data science, convenient tools embedded within the R programming language. R allows analysts to run various statistical models and is built specifically for statistical analysis and data mining. It can also enable the analyst to create software and applications that allow for reliable statistical analysis. Its graphical interface is also beneficial for data clustering, time-series, lineal modeling, etc.

Statistical modelling in Excel

Excel can be used conveniently for statistical analysis of basic data. It may not be ideal for huge sets of data, where R and Python work seamlessly. Microsoft Excel provides several add-in tools under the Data tab. Enabling the Data Analysis tool on Excel opens a wide range of convenient statistical analysis options, including descriptive analysis, ANOVA, moving average, regression, and sampling.

Descriptive model

In this type of model, the purpose is to provide a reasonable description of the data in some appropriate way without any attempt at understanding the underlying phenomenon , that is the data-generating mechanism , then the family of models is selected based on its adequacy to represents the data structure

In this instance, the order of the model is chosen based on its competence to describe the data arrangement.

This type of model is very useful for discriminating between alternating hypothesis but are useless for capturing the fundamental characteristics of a mechanism.

Mechanistic model

In the mechanistic model, the importance rests in the knowledge of the device of development, it is important to be able to score on a powerful collaboration among scientists, specialists in the field and statisticians or mathematicians.

The former must provide updated , rich and reliable information about the problem.

whereas the latter are trained for translating scientific information in mathematical models.

Purpose of model

To translate the known properties about as well as some new hypothesis into a mathematical representation.

The family of models is selected depends on the main purpose of the exercise.

If the purpose is just to provide a reasonable description of the data without any attempt at understanding the underlying phenomenon, that is, the **data-generating mechanism.**

Then the family of models is selected based on its adequacy to represent the data structure.

Animal tumor growth data are used for the representation of the different concepts encountered during the development of a model and its after-identification use.

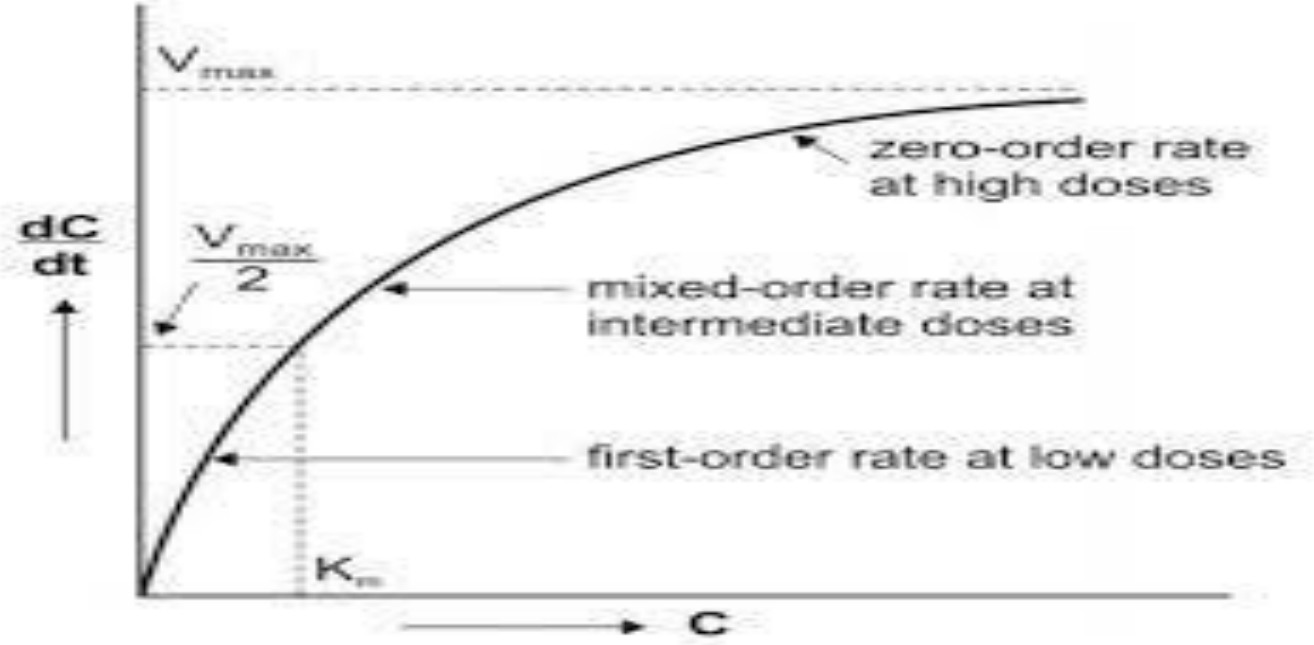
The data represent the tumor growth in rats over a period of 80 days.

We are interested in modeling the growth of experimental tumors subcutaneously implanted in rats to be able to differentiate between treatment regimens.

Mathematical Models	Statistical Models
Study of concepts (in space and time) • quantity (discrete and continuous) • structure (geometric figures) • Patterns	To decide on suitable course of action, it deals with (in space and time) • collection and analysis of data • Extracting information
Mechanistic or perspective models (primarily causation) of how system changes	Descriptive or phenomenological models (primarily correlation; traditional approach)
Some precision (limited data are used)	Precise (exact data are used)
Realism: Explicitly considering the process that produce given observation or changes in the system	Little realism: Make no claim about the nature of the underlying mechanisms that produce the behavior of the system
Study dynamics (iterative interactions over time and space) of interacting populations using deterministic or probabilistic models	Attempts at estimating probabilistic future behavior of a system based on its past behavior

Nonlinearity at the maximum

Nonlinearity is a term used in statistics to describe a situation where there is not a straight-line or direct relationship between an independent variable and a dependent variable. In a nonlinear relationship, changes in the output do not change in direct proportion to changes in any of the inputs.



Linear Pharmacokinetics	Non-linear Pharmacokinetics
Pharmacokinetic parameters for a drug would not change with change in dose	Pharmacokinetic parameters for a drug can change with change in dose.
Dose Independent	Dose dependent
First Order kinetics	Also called as Mixed order, saturated kinetics, capacity limited
All semilog plots of C vs t for diff. doses are superimposable.	Not superimposable

Causes of Nonlinearity

Cause	Drug
Drug Absorption	
Absorption involves carrier-mediated transport system (F, K_a, C_{max}, AUC) (↓)	Riboflavin, ascorbic acid, Cyanocobalamine, gebapentin, L-dopa, baclofen, ceftibuten
Drugs with low solubility in GI but relatively high dose (F, K_a, C_{max}, AUC) (↓)	Chorothiazide, griseofulvin, danazol
Presystemic hepatic metabolism attains saturation (F, K_a, C_{max}, AUC) (↑)	Propranolol

Sensitivity Analysis

Sensitivity analysis is the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be divided and allocated to different sources of uncertainty in its inputs.

A related practice is uncertainty analysis, which has a greater focus on uncertainty quantification and propagation of uncertainty; ideally, uncertainty and sensitivity analysis should be run in tandem.

The process of recalculating outcomes under alternative assumptions to determine the impact of a variable under sensitivity analysis can be useful for a range of purposes, including:

- Testing the robustness of the results of a model or system in the presence of uncertainty.
- Increased understanding of the relationships between input and output variables in a system or model.

METHODS

There are a large number of approaches to performing a sensitivity analysis, many of which have been developed to address one or more of the constraints discussed above.[2] They are also distinguished by the type of sensitivity measure, be it based on (for example) variance decompositions, partial derivatives or elementary effects.

- **Derivative-based local method**
- **Regression analysis**
- **Variance-based methods**
- **Variogram analysis of response surfaces (VARS)**
- **Screening**

References

***The Pharmaceutical Journal*, PJ, December 2016, Vol 297, No 7896;297(7896):DOI:10.1211/PJ.2016.20202033**