UNIT III

Topic 3.1 : Basic concept in Bio-statistics (sampling design, data collection, scaling technique, parametric and non-parametric statistics)

What is Statistics ?

The field of statistics. The study and use of theory and methods for the analysis of data arising from random processes or phenomena. The study of how we make sense of data.

-The field of statistics provides some of the most fundamental tools and techniques of the scientific method.

-forming hypotheses.

-designing experiments and observational studies.

-gathering data.

-summarizing data.

-drawing inferences from data .e.g.. testing hypotheses.

The field of statistics. also refers to a numerical quantity computed from sample data (.e.g.. the mean, the media, the maximum).

What is Bio-statistics?

Biostatistics are the development and application of statistical methods to a wide range of topics in biology. It encompasses the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.

Biostatistics is sometimes distinguished from the field of biometry based upon whether applications are in the health sciences (bio statistics) or in broader biology (biometry, e.g., agriculture, ecology, wildlife biology).

Biostatistics covers applications and contributions not only from health, medicines and, nutrition but also from fields such as genetics, biology, epidemiology, and many others. Biostatistics mainly consists of various steps like generation of hypothesis, collection of data, and application of

statistical analysis. To begin with, readers should know about the data obtained during the experiment, its distribution, and its analysis to draw a valid conclusion from the experiment.

Sampling Design

What are sampling methods?

In a statistical study, sampling methods refer to how we select members from the population to be in the study. If a sample isn't randomly selected, it will probably be biased in some way and the data may not be representative of the population.

Bad ways to Sample:

Convenience sample: The researcher chooses a sample that is readily available in some non-random way.

Example—A researcher polls people as they walk by on the street.

Probably biased: The location and time of day and other factors may produce a biased sample of people.

Voluntary response sample: The researcher puts out a request for members of a population to join the sample, and people decide whether or not to be in the sample.

Example—A TV show host asks his viewers to visit his website and respond to an online poll.

Probably biased: People who take the time to respond tend to have similarly strong opinions compared to the rest of the population.

Good ways to sample:

Simple random sample: Every member and set of members has an equal chance of being included in the sample. Technology, random number generators, or some other sort of chance process is needed to get a simple random sample.

Example—A teachers puts students' names in a hat and chooses without looking to get a sample of students.

Why it's good: Random samples are usually fairly representative since they don't favor certain members.

Stratified random sample: The population is first split into groups. The overall sample consists of some members from every group. The members from each group are chosen randomly.

Example—A student council surveys 100100100 students by getting random samples of 252525 freshmen, 252525 sophomores, 252525 juniors, and 252525 seniors.

Why it's good: A stratified sample guarantees that members from each group will be represented in the sample, so this sampling method is good when we want some members from every group.

Cluster random sample: The population is first split into groups. The overall sample consists of every member from some of the groups. The groups are selected at random.

Example—An airline company wants to survey its customers one day, so they randomly select 555 flights that day and survey every passenger on those flights.

Why it's good: A cluster sample gets every member from some of the groups, so it's good when each group reflects the population as a whole.

Systematic random sample: Members of the population are put in some order. A starting point is selected at random, and every n(th) member is selected to be in the sample.

Example—A principal takes an alphabetized list of student names and picks a random starting point.

Sampling Design

Sampling design is a mathematical function that gives the probability of any given sample being drawn.

Since sampling is the foundation of nearly every research project, the study of sampling design is a crucial part of statistics, and is often a one or two semester course. It involves not only learning how to derive the probability functions which describe a given sampling method but also understanding how to design a best-fit sampling method for a real life situation.

Examples of Sampling Design

Sampling design can be very simple or very complex. In the simplest, one stage sample design where there is no explicit stratification and a member of the population is chosen at random, each unit has the probability n/N of being in the sample, where:

n is the total number of units to be sampled,

N is number of units in the total population.

Other types of design include:

Systematic sample: all members of a population are listed in order and samples are chosen at defined intervals

Stratified sample: the population is first divided into strata and then samples are randomly selected from the strata (for example, divide a population between men and women, then randomly select a given number of men and a given number of women)

Cluster strata: a population is divided into clusters and first clusters are randomly selected, then random members of the selected clusters are sampled. (for instance, first randomly select a number of classes, then, from the class lists of those classes, randomly sample a number of students)

Each of these have their own sampling design function. The sampling method chosen will depend on the situation and priorities of the researcher. Sometimes, non-probability sampling methods will be chosen; for example, convenience sampling, where the sample is simply those easily reached and observed. Unlike systematic, stratified, or cluster sampling, these types of sampling cannot be easily described by a function.

Data Collection:

When faced with a research problem, you need to collect, analyze and interpret data to answer your research questions. Examples of research questions that could require you to gather data include how many people will vote for a candidate, what is the best product mix to use and how useful is a drug in curing a disease. The research problem you explore informs the type of data you'll collect and the data collection method you'll use. In this section, we will explore various types of data, methods of data collection and advantages and disadvantages of each.

Types of Data

Quantitative Data

Data that is expressed in numbers and summarized using statistics to give meaningful information is referred to as **quantitative data**. Examples of quantitative data we could collect are heights, weights, or ages of students. If we obtain the mean of each set of measurements, we have meaningful information about the average value for each of those student characteristics.

Qualitative Data

When we use data for description without measurement, we call it **qualitative data**. Examples of qualitative data are student attitudes towards school, attitudes towards exam cheating and friendliness of students to teachers. Such data cannot be easily summarized using statistics.

Primary Data

When we obtain data directly from individuals, objects or processes, we refer to it as **primary data**. Quantitative or qualitative data can be collected using this approach. Such data is usually collected solely for the research problem to you will study. Primary data has several advantages. First, we tailor it to our specific research question, so there are no customizations needed to make the data usable. Second, primary data is reliable because you control how the data is collected and can monitor its quality. Third, by collecting primary data, you spend your resources in collecting only required data. Finally, primary data is proprietary, so you enjoy advantages over those who cannot access the data.

Despite its advantages, primary data also has disadvantages of which you need to be aware. The first problem with primary data is that it is costlier to acquire as compared to secondary data. Obtaining primary data also requires more time as compared to gathering secondary data.

Secondary Data

When you collect data after another researcher or agency that initially gathered it makes it available, you are gathering **secondary data**. Examples of secondary data are census data published by the US Census Bureau, stock prices data published by CNN and salaries data published by the Bureau of Labor Statistics.

One advantage to using secondary data is that it will save you time and money, although some data sets require you to pay for access. A second advantage is the relative ease with which you can obtain it. You can easily access secondary data from publications, government agencies, data aggregation websites and blogs. A third advantage is that it eliminates effort duplication since you can identify existing data that matches your needs instead of gather new data.

Despite the benefits it offers, secondary data has its shortcomings. One limitation is that secondary data may not be complete. For it to meet your research needs, you may need to enrich it with data from other sources. A second shortcoming is that you cannot verify the accuracy of secondary data, or the data may be outdated. A third challenge you face when using secondary data is that documentation may be incomplete or missing. Therefore, you may not be aware of any problems that happened in data collection which would otherwise influence its interpretation. Another challenge you may face when you decide to use secondary data is that there may be copyright restrictions.

Now that we've explained the various types of data you can collect when conducting research, we will proceed to look at methods used to collect primary and secondary data.

Methods Employed in Primary Data Collection

When you decide to conduct original research, the data you gather can be quantitative or qualitative. Generally, you collect quantitative data through sample surveys, experiments and observational studies. You obtain qualitative data through focus groups, in-depth interviews and case studies. We will discuss each of these data collection methods below and examine their advantages and disadvantages.

Sample Surveys

A **survey** is a data collection method where you select a sample of respondents from a large population in order to gather information about that population. The process of identifying individuals from the population who you will interview is known as **sampling**.

To gather data through a survey, you construct a questionnaire to prompt information from selected respondents. When creating a questionnaire, you should keep in mind several key considerations. First, make sure the questions and choices are unambiguous. Second, make sure the questionnaire will be completed within a reasonable amount of time. Finally, make sure there are no typographical errors. To check if there are any problems with your questionnaire, use it to interview a few people before administering it to all respondents in your sample. We refer to this process as pretesting.Using a survey to collect data offers you several advantages. The main benefit is time and cost savings because you only interview a sample, not the large population. Another benefit is that when you select your sample correctly, you will obtain information of acceptable accuracy. Additionally, surveys are adaptable and can be used to collect data for governments, health care institutions, businesses and any other environment where data is needed.

A major shortcoming of surveys occurs when you fail to select a sample correctly; without an appropriate sample, the results will not accurately generalize the population.

Ways of Interviewing Respondents

Once you have selected your sample and developed your questionnaire, there are several ways you can interview participants. Each approach has its advantages and disadvantages.

In-person Interviewing

When you use this method, you meet with the respondents face to face and ask questions. In-person interviewing offers several advantages. This technique has excellent response rates and enables you to conduct interviews that take a longer amount of time. Another benefit is you can ask follow-up questions to responses that are not clear.

In-person interviews do have disadvantages of which you need to be aware. First, this method is expensive and takes more time because of interviewer training, transport, and remuneration. A

second disadvantage is that some areas of a population, such as neighborhoods prone to crime, cannot be accessed which may result in bias.

Telephone Interviewing

Using this technique, you call respondents over the phone and interview them. This method offers the advantage of quickly collecting data, especially when used with computer-assisted telephone interviewing. Another advantage is that collecting data via telephone is cheaper than in-person interviewing.

One of the main limitations with telephone interviewing it's hard to gain the trust of respondents. Due to this reason, you may not get responses or may introduce bias. Since phone interviews are generally kept short to reduce the possibility of upsetting respondents, this method may also limit the amount of data you can collect.

Online Interviewing

With online interviewing, you send an email inviting respondents to participate in an online survey. This technique is used widely because it is a low-cost way of interviewing many respondents. Another benefit is anonymity; you can get sensitive responses that participants would not feel comfortable providing with in-person interviewing.

When you use online interviewing, you face the disadvantage of not getting a representative sample. You also cannot seek clarification on responses that are unclear.

Mailed Questionnaire

When you use this interviewing method, you send a printed questionnaire to the postal address of the respondent. The participants fill in the questionnaire and mail it back. This interviewing method gives you the advantage of obtaining information that respondents may be unwilling to give when interviewing in person.

The main limitation with mailed questionnaires is you are likely to get a low response rate. Keep in mind that inaccuracy in mailing address, delays or loss of mail could also affect the response

rate. Additionally, mailed questionnaires cannot be used to interview respondents with low literacy, and you cannot seek clarifications on responses.

Focus Groups

When you use a focus group as a data collection method, you identify a group of 6 to 10 people with similar characteristics. A moderator then guides a discussion to identify attitudes and experiences of the group. The responses are captured by video recording, voice recording or writing—this is the data you will analyze to answer your research questions. Focus groups have the advantage of requiring fewer resources and time as compared to interviewing individuals. Another advantage is that you can request clarifications to unclear responses.

One disadvantage you face when using focus groups is that the sample selected may not represent the population accurately. Furthermore, dominant participants can influence the responses of others.

Observational Data Collection Methods

In an observational data collection method, you acquire data by observing any relationships that may be present in the phenomenon you are studying. There are four types of observational methods that are available to you as a researcher: cross-sectional, case-control, cohort and ecological.

In a **cross-sectional** study, you only collect data on observed relationships once. This method has the advantage of being cheaper and taking less time as compared to case-control and cohort. However, cross-sectional studies can miss relationships that may arise over time.

Using a **case-control** method, you create cases and controls and then observe them. A case has been exposed to a phenomenon of interest while a control has not. After identifying the cases and controls, you move back in time to observe how your event of interest occurs in the two groups. This is why case-control studies are referred to as retrospective. For example, suppose a medical researcher suspects a certain type of cosmetic is causing skin cancer. You recruit people who have used a cosmetic, the cases, and those who have not used the cosmetic, the controls. You request participants to remember the type of cosmetic and the frequency of its use. This method is cheaper and requires less time as compared to the cohort method. However, this approach has limitations

when individuals you are observing cannot accurately recall information. We refer to this as recall bias because you rely on the ability of participants to remember information. In the cosmetic example, recall bias would occur if participants cannot accurately remember the type of cosmetic and number of times used.

In a **cohort** method, you follow people with similar characteristics over a period. This method is advantageous when you are collecting data on occurrences that happen over a long period. It has the disadvantage of being costly and requiring more time. It is also not suitable for occurrences that happen rarely.

The three methods we have discussed previously collect data on individuals. When you are interested in studying a population instead of individuals, you use an **ecological** method. For example, say you are interested in lung cancer rates in Iowa and North Dakota. You obtain number of cancer cases per 1000 people for each state from the National Cancer Institute and compare them. You can then hypothesize possible causes of differences between the two states. When you use the ecological method, you save time and money because data is already available. However the data collected may lead you to infer population relationships that do not exist.

Experiments

An experiment is a data collection method where you as a researcher change some variables and observe their effect on other variables. The variables that you manipulate are referred to as **independent** while the variables that change as a result of manipulation are **dependent** variables. Imagine a manufacturer is testing the effect of drug strength on number of bacteria in the body. The company decides to test drug strength at 10mg, 20mg and 40mg. In this example, drug strength is the independent variable while number of bacteria is the dependent variable. The drug administered is the treatment, while 10mg, 20mg and 40mg are the levels of the treatment.

The greatest advantage of using an experiment is that you can explore causal relationships that an observational study cannot. Additionally, experimental research can be adapted to different fields like medical research, agriculture, sociology, and psychology. Nevertheless, experiments have the disadvantage of being expensive and requiring a lot of time.

Parametric and Non Parametric Statistics

Parametric statistics is a branch of statistics which assumes that sample data come from a population that can be adequately modeled by a probability distribution that has a fixed set of parameters. Conversely a non-parametric model differs precisely in that the parameter set (or feature set in machine learning) is not fixed and can increase, or even decrease, if new relevant information is collected.

TableI. Parametric vs Non-Parametric tests.	
Parametric	Non-Parametric
1 Sample T-test	Sign Test/Wilcoxon Signed Rank test
Paired T-test	Sign Test/Wilcoxon Signed Rank test
2 Sample T-test	Mann Whitney U test/Wilcoxon Sum Rank test
ANOVA	Kruskal Wallis test
We shall look at various examples to understand when each test is being used.	

Parametric Statistics

A parameter in statistics refers to an aspect of a population, as opposed to a statistic, which refers to an aspect about a sample. For example, the population mean is a parameter, while the sample mean is a statistic. A parametric statistical test makes an assumption about the population parameters and the distributions that the data came from. These types of test includes Student's T tests and ANOVA tests, which assume data is from a normal distribution.

The opposite is a nonparametric test, which doesn't assume anything about the population parameters. Nonparametric tests include chi-square, Fisher's exact test and the Mann-Whitney test.

Every parametric test has a nonparametric equivalent. For example, if you have parametric data from two independent groups, you can run a 2 sample t test to compare means. If you have nonparametric data, you can run a Wilcoxon rank-sum test to compare means.

Parametric Data Definition

Data that is assumed to have been drawn from a particular distribution, and that is used in a parametric test.

Parametric Equations

Parametric equations are used in calculus to deal with the problems that arise when trying to find functions that describe curves. These equations are beyond the scope of this site, but you can find an excellent rundown of how to use these types of equations here.

Non Parametric Statistics

What is a Non Parametric Test?

A non parametric test (sometimes called a distribution free test) does not assume anything about the underlying distribution (for example, that the data comes from a normal distribution). That's compared to parametric test, which makes assumptions about a population's parameters (for example, the mean or standard deviation); When the word "non parametric" is used in stats, it doesn't quite mean that you know nothing about the population. It usually means that you know the population data does not have a normal distribution.

For example, one assumption for the one way ANOVA is that the data comes from a normal distribution. If your data isn't normally distributed, you can't run an ANOVA, but you can run the nonparametric alternative—the Kruskal-Wallis test.

If at all possible, you should us parametric tests, as they tend to be more accurate. Parametric tests have greater statistical power, which means they are likely to find a true significant effect. Use nonparametric tests only if you have to (i.e. you know that assumptions like normality are being violated). Nonparametric tests can perform well with non-normal continuous data if you have a sufficiently large sample size (generally 15-20 items in each group).

When to use it

Non parametric tests are used when your data isn't normal. Therefore the key is to figure out if you have normally distributed data. For example, you could look at the distribution of your data. If your data is approximately normal, then you can use parametric statistical tests.

Q. If you don't have a graph, how do you figure out if your data is normally distributed?

A. Check the skewness and Kurtosis of the distribution using software like Excel. A normal distribution has no skew. Basically, it's a centered and symmetrical in shape. Kurtosis refers to how much of the data is in the tails and the center. The skewness and kurtosis for a normal distribution is about non parametric, Negative kurtosis (left) and positive kurtosis (right). If your distribution is not normal (in other words, the skewness and kurtosis deviate a lot from 1.0), you should use a non parametric test like chi-square test. Otherwise you run the risk that your results will be meaningless.

Data Types Does your data allow for a parametric test, or do you have to use a non parametric test like chi-square?

The rule of thumb is: For nominal scales or ordinal scales, use non parametric statistics. For interval scales or ratio scales use parametric statistics, use nonparametric tests. A skewed distribution is one reason to run a nonparametric test. Other reasons to run nonparametric tests:
One or more assumptions of a parametric test have been violated. Your sample size is too small to run a parametric test. Your data has outliers that cannot be removed. You want to test for the median rather than the mean (you might want to do this if you have a very skewed distribution).

Types of Nonparametric Tests

When the word "parametric" is used in stats, it usually means tests like ANOVA or a t test. Those tests both assume that the population data has a normal distribution. Non parametric do not assume that the data is normally distributed. The only non parametric test you are likely to come across in elementary stats is the chi-square test. However, there are several others. For example: the Kruskal Willis test is the non parametric alternative to the One way ANOVA and the Mann Whitney is the non parametric alternative to the two sample t test.

The main nonparametric tests are:

(1) 1-sample sign test. Use this test to estimate the median of a population and compare it to a reference value or target value.
(2) 1-sample Wilcoxon signed rank test. With this test, you also estimate the population median and compare it to a reference/target value. However, the test assumes your data comes from a symmetric distribution (like the Cauchy distribution or uniform distribution).

(3) Friedman test. This test is used to test for differences between groups with ordinal dependent variables. It can also be used for continuous data if the one-way ANOVA with repeated measures is inappropriate (i.e. some assumption has been violated).
(4) Goodman Kruska's Gamma: a test of association for ranked variables.
(5) Kruskal-Wallis test. Use this test instead of a one-way ANOVA to find out if two or more medians are different. Ranks of the data points are used for the calculations, rather than the data points themselves.
(6) The Mann-Kendall Trend Test looks for trends in time-series data.
(7) Mann-Whitney test. Use this test to compare differences between two independent groups when dependent variables are either ordinal or continuous.
(8) Mood's Median test. Use this test instead of the sign test when you have two independent samples.
(9) Spearman Rank Correlation.Use when you want to find a correlation between two sets of data.

Advantages and Disadvantages:

Compared to parametric tests, nonparametric tests have several advantages, including:

(1) More statistical power when assumptions for the parametric tests have been violated. When assumptions haven't been violated, they can be almost as powerful.
(2) Fewer assumptions (i.e. the assumption of normality doesn't apply).
(3) Small sample sizes are acceptable.
(4) They can be used for all data types, including nominal variables, interval variables, or data that has outliers or that has been measured imprecisely.

However, they do have their disadvantages. The most notable ones are:

(1) Less powerful than parametric tests if assumptions haven't been violated.
(2) More labor-intensive to calculate by hand (for computer calculations, this isn't an issue).
(3) Critical value tables for many tests aren't included in many computer software packages.

** Internet and public domain resources have been used to collate the material. Websites – wiki pedia;statisticshowto.com; Quantitative Data – Parametric & Non-parametric Tests (Y H Chan)*