Biometry

Introduction

BA Part I

Bibha Verma

18 January 2022

Introduction to Biometry

Biostatistics (also known as **biometry**) 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 has historically been heavily used in the field of genetics. Mendel used statistics to to collect and understand the genetics of segregation pattern in family of peas. Population genetics used these concepts later.

The three leading figures in the establishment of population genetics and this synthesis all relied on statistics and developed its use in biology.

- Ronald Fisher developed several basic statistical methods in support of his work studying the crop experiments at Rothamsted Research, including in his books Statistical Methods for Research Workers (1925) end The Genetical Theory of Natural Selection (1930). He gave many contributions to genetics and statistics. Some of them include the ANOVA, p-value concepts, Fisher's exact test and Fisher's equation for population dynamics. He is credited for the sentence "Natural selection is a mechanism for generating an exceedingly high degree of improbability".[1]
- Sewall G. Wright developed F-statistics and methods of computing them and defined inbreeding coefficient.
- J. B. S. Haldane's book, *The Causes of Evolution*, reestablished natural selection as the premier mechanism of evolution by explaining it in terms of the mathematical consequences of Mendelian genetics. Also developed the theory of primordial soup.

Research planning answers scientific questions, and to have a high accuracy of results, statistics is integral. The correct definition of main hypothesis and the research plan will reduce errors while taking a decision and in understanding methods of data analysis, perspective and costs evolved. The three basic steps of : Randomisation, Replication and Local Control will be essential here.

Research Question: Defines the objective of the study. The research needs to be concise, and to define the way to ask a scientific question, literature review is necessary so that the research may add value to the scientific community.

Hypothesis definition: Once defined the possible answers can be proposed, transforming the question into a hypothesis. The main purpose of null hypothesis is to based on permanent knowledge or obvious occurrence of the phenomenon, sustained by deep literature review. It is the standard expected answer for data under situation in test. Ho assumes no association between treatments. Alternative hypothesis is denial of Ho.

Sampling: A study aims to understand an effect of a phenomenon over a population. It is generally not possible to take measures from all elements of a population. Sampling process is very important for statistical inference. Sample size is determined by several things including scope of the research and to resources.

Experimental Design: Sustains those basic principles of experimental statistics. Three kinds of experiment designs are:

1. Completely randomised design - for studying the effects of one primary factor without the need to take other nuisance variables into account.
2. Randomised block design - **blocking** is the arranging of experimental units in groups (blocks) that are similar to one another. Blocking can be used to tackle the problem of pseudoreplication.
3. Factorial design - is an experiment whose design consists of two or more factors, each with discrete possible values or "levels", and whose experimental units take on all possible combinations of these levels across all such factors. A full **factorial design** may also be called a **fully crossed design**.

Data collection: methods have to be considered in research planning, and highly influences the sample size and experimental design. Collection method also varies with the type of data. Qualitative data is collected by structured questionnaire or by observation. Quantitative data, collection is by measuring numerical information using instruments.

Analysis and Data Interpretation:

1. Descriptive tools: Representation through tables or graphs such as line chart, bar charts, histograms, scatter plots. Central tendency and variability are also useful to describe overview of data.
   - Frequency tables - number of occurrences or repetitions of data
   - Absolute - number of times a determined value appears
   - Relative - devision of the absolute frequency by total number
   - Line graph - variation over other metric, such as time
   - Bar chart - data in horizontal bars
   - Histograms - graphical representation of a dataset tabluated and divided into uniform and non-uniform
   - Scatter plot - mathematical diagram using cartesian coordinates
   - Mean - the average value (sum of all values/number of values)
   - Median - value at the middle of dataset
   - Mode - the value that appears most
   - Box plot - graphically deputising groups of numerical data
   - Correlation coefficients - data inferred by graphs, scatter plot and necessary to validate numerical informations.
   - Pearson correlation coefficient - association between two variables X and Y

2. Inferential statistics - It is used to make inferences[14] about an unknown population, by estimation and/or hypothesis testing. In other words, it is desirable to obtain parameters to describe the population of interest, but since the data is limited, it is necessary to make use of a representative sample in order to estimate them. With that, it is

possible to test previously defined hypotheses and apply the conclusions to the entire population. The standard error of the mean is a measure of variability that is crucial to do inferences.[4]

Hypothesis testing

Hypothesis testing is essential to make inferences about populations aiming to answer research questions, as settled in "Research planning" section. Authors defined four steps to be set:[4]

1. *The hypothesis to be tested*: as stated earlier, we have to work with the definition of a null hypothesis (H_0), that is going to be tested, and an alternative hypothesis. But they must be defined before the experiment implementation.
2. *Significance level and decision rule*: A decision rule depends on the level of significance, or in other words, the acceptable error rate (α). It is easier to think that we define a *critical value* that determines the statistical significance when a test statistic is compared with it. So, α also has to be predefined before the experiment.
3. *Experiment and statistical analysis*: This is when the experiment is really implemented following the appropriate experimental design, data is collected and the more suitable statistical tests are evaluated.
4. *Inference*: Is made when the null hypothesis is rejected or not rejected, based on the evidence that the comparison of p-values and α brings. It is pointed that the failure to reject H_0 just means that there is not enough evidence to support its rejection, but not that this hypothesis is true.

Confidence Interval

A confidence interval is a range of values that can contain the true real parameter value in given a certain level of confidence. The first step is to estimate the best-unbiased estimate of the population parameter. The upper value of the interval is obtained by the sum of this estimate with the multiplication between the standard error of the mean and the confidence level. The calculation of lower value is similar, but instead of a sum, a subtraction must be applied.

Developments and Big Data

Recent developments have made a large impact on biostatistics. Two important changes have been the ability to collect data on a high-throughput scale, and the ability to perform much more complex analysis using computational techniques. This comes from the development in areas as sequencing technologies, Bioinformatics and Machine learning (Machine learning in bioinformatics).

Applications

Public health

Public health, including epidemiology, health services research, nutrition, environmental health and health care policy & management. In these medicine contents, it's important to consider the design and analysis of the clinical trials. As one example, there is the assessment of severity state of a patient with a prognosis of an outcome of a disease.

With new technologies and genetics knowledge, biostatistics are now also used for Systems medicine, which consists in a more personalized medicine. For this, is made an integration of data from different sources, including conventional patient data, clinico-pathological parameters, molecular and genetic data as well as data generated by additional new-omics technologies.

Qualitative genetics

The study of Population genetics and Statistical genetics in order to link variation in genotype with a variation in phenotype. In other words, it is desirable to discover the genetic basis of a measurable trait, a quantitative trait, that is under polygenic control. A genome region that is responsible for a continuous trait is called Quantitative trait locus (QTL). The study of QTLs become feasible by using molecular markers and measuring traits in populations, but their mapping needs the obtaining of a population from an experimental crossing, like an F2 or Recombinant inbred strains/lines (RILs). To scan for QTLs regions in a genome, a gene map based on linkage have to be built. Some of the best-known QTL mapping algorithms are

Interval Mapping, Composite Interval Mapping, and Multiple Interval Mapping.

Expression data

Studies for differential expression of genes from RNA-Seq data, as for RT-qPCR and microarrays, demands comparison of conditions. The goal is to identify genes which have a significant change in abundance between different conditions. Then, experiments are designed appropriately, with replicates for each condition/treatment, randomization and blocking, when necessary. In RNA-Seq, the quantification of expression uses the information of mapped reads that are summarized in some genetic unit, as exons that are part of a gene sequence. As microarray results can be approximated by a normal distribution, RNA-Seq counts data are better explained by other distributions. The first used distribution was the Poisson one, but it underestimate the sample error, leading to false positives. Currently, biological variation is considered by methods that estimate a dispersion parameter of a negative binomial distribution. Generalized linear models are used to perform the tests for statistical significance and as the number of genes is high, multiple tests correction have to be considered. Some examples of other analysis on genomics data comes from microarray or proteomics experiments. Often concerning diseases or disease stages.

Source: Wikipedia