

## **Introduction to Biostatistics**

Data – numbers resulting from counting or measurement. An individual number is a datum.

Statistics – a field of study concerned with

- (1) the collection, organization, summarization, and analysis of data (descriptive statistics);
- (2) the drawing of inferences about a body of data (population) when only part of the data (sample) is observed (inferential statistics).

Sources of Data

- (1) Routinely kept records
- (2) Surveys
- (3) Experiments
- (4) External Sources

Biostatistics – the application of statistical tools and concepts to data derived from the biological sciences or medicine.

Variable – a characteristic that takes on different values in different persons, places, or things.

- (1) Quantitative variable – one measured in the usual sense and conveys information regarding amount
- (2) Qualitative or categorical variable – measuring consists of categorizing and the measurements convey information regarding attribute. Frequencies (or counts) are the numbers we manipulate when our analysis involves qualitative variables.

- (3) Random variables – values arise as a result of chance factors, and cannot be predicted in advance. Values resulting from measurement procedures are referred to as observations or measurements.
- (4) Discrete random variable – characterized by gaps or interruptions in the values it can assume; you can count out possible values
- (5) Continuous Random Variable – can assume any value within a specified relevant interval of values assumed by the variable

Population – the largest collection of entities for which we have an interest at a particular time. A population of values is the largest collection of values of a random variable for which we have an interest at a particular time. Populations may be finite or infinite.

Sample – part of a population.

Measurement – the assignment of numbers to objects or events according to a set of rules

- (1) Nominal Scale – classifying into mutually exclusive and exhaustive categories.

Examples are medical diagnoses and age groups.

- (2) Ordinal Scale - ranking among categories, where the distance between categories need not be equal.

Examples are below average, average, above average, and the pain scale, where nurses ask you to rate your pain on a scale of 1 to 10.

- (3) Interval Scale – has a unit distance and a zero point, so there is equality of intervals; this is a truly quantitative scale.

An example is temperature. For  $F^{\circ}$  and  $C^{\circ}$ , we have arbitrary 0's. The distance from  $30^{\circ}$  to  $40^{\circ}$  represents the same heat gain as the distance from  $70^{\circ}$  to  $80^{\circ}$ ; but  $20^{\circ}$  is not twice as hot as  $10^{\circ}$ .

- (4) Ratio scale – equality of intervals and ratios may be determined; there is a true 0 point.

Example are height, weight, and annual income. Doubling weight will take a 50 pounder to 100 pounds;  $\frac{100}{50} = \frac{160}{80} = 2$ .

### Presenting Data

- (1) Identify the source and the individuals. How many are in the set?
- (2) Identify variables and type.
- (3) Identify units of measurement.
- (4) Label everything.

Statistical Inference – the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample that has been drawn from that population.

Simple Random Sample (SRS) – a sample of size  $n$  from a population of size  $N$  in such a way that every possible sample of size  $n$  has the same chance of being selected.

**NOTE.** As a rule, sampling is done without replacement.

**EXAMPLE.** We have 10 individuals numbered 0 thru 9, and want to choose an SRS of 3. For our procedure, we note that there are 50 rows of random numbers in Table A on page A-2. Without looking at the table, we choose a row from 1 to 50. Take the person with the same number as the first digit in that row along with the two succeeding people. Consider 0 the successor of 9. Suppose we pick row 36, where the first digit is 6, so we take for our sample persons 6, 7, 8. Each person has a 30% chance of being chosen. Is this an SRS?

**NO** – group {6, 7, 8} has a 10% chance of being chosen, but group {3, 6, 8} has 0% chance of being chosen.

**EXAMPLE.** Pick an SRS from the population of 169 individuals on pages 8-9 with  $n = 4$ . Procedure:

- (1) number the individuals from 001-169, i.e. use 3 digits to represent individuals.
- (2) Pick a random starting point in the table, say, row 23, column 6 (from the left).
- (3) read off sequences of 3 successive digits until you get 4 distinct (no replacement) numbers from 001 thru 169

(413)(721)(083)(766)(992)(931)(835)(692)(046)(479)(320)(728)(008)  
 (363)(868)(709)(308)(965)(405)(359)(471)(961)(245)(238)(234)(598)  
 (479)(719)(755)(147)

An alternate, and easier, way to get the numbers is to take your TI-89 and choose **2nd/Math/7:Probability/4:Rand** to get **Rand(** and then complete the command to **Rand(169)**. Then each time you hit **ENTER** you get a random integer from 1 to 169.

There are 32,795,126 possible samples of size 4, each equally likely.

Research Study – a scientific study of a phenomenon of interest. Research studies involve designing sampling protocols, collecting and analyzing data, and providing valid conclusions based on the results of the analysis.

Experiments – a special type of research study in which observations are made after specific manipulations of conditions have been carried out; they provide the foundation for scientific research.

## Two Additional Sampling Techniques in the Health Sciences:

- (1) Systematic Sampling - often used with a set of files of medical records. With the files numbered and ordered, a random starting point  $x$  is chosen along with an interval  $k$ . Then the records used are  $x$ ,  $x + k$ ,  $x + 2k$ , etc.
- (2) Stratified Random Sampling – a population of interest is partitioned into groups, or strata, in which the sample units within a particular stratum are more similar to each other than they are to the sample units that compose the other strata. Then an SRS is taken from each stratum.

Scientific Method – a process by which scientific information is collected, analyzed, and reported in order to produce unbiased and replicable results in an effort to produce an accurate representation of observable phenomena.

- (1) Making an observation – leads to the formulation of questions or uncertainties that can be answered in a scientifically rigorous way.
- (2) Formulating a hypothesis – to explain the observation and to make quantitative predictions of new observations. Hypotheses may be stated as either research hypotheses or statistical hypotheses.
- (3) Designing an experiment – that will yield the data necessary to validly test an appropriate statistical hypothesis.
  - (a) Accuracy - the correctness of a measurement (validity).
  - (b) Precision – the consistency of a measurement (reliability).
- (4) Conclusion – based on the degree of confidence about the hypotheses that were posed as part of the design. But results need to be replicated, often a large number of times, before scientific credence is granted them.