

Module 1

Basic Concepts

Objective: In this module you will learn mainly about classification of medical/clinical data and methods of drawing random sample from a large number of patients.

1.1 Introduction

The tools of statistics are employed in many fields such as medicine, pharmacy, biology, biochemistry, business, education, agriculture, psychology and economics, to mention a few. When the data analysed are derived from the biological sciences and medicine, we use the term “biostatistics” to distinguish this particular application of statistical tools from the rest. The application of biostatistics is the subject of this unit.

In this module we discuss some of the basic words and phrases that are encountered in the study of biostatistics. We should develop a clear understanding of these very preliminary concepts in statistics which will be used extensively in the remainder of our discussion. By the end of this module we will learn about:

- Observations & Variables
- Measurement Scales or Types of Data
 - Categorical (qualitative) Data:
 - Nominal data
 - Ordinal data
 - Numerical Data:
 - Discrete data
 - Continuous data
- Populations & Samples
- Random variables

1.2 Observations & Variables:

Observations:

In medicine, data is usually obtained by collecting information from a group of subjects/patients participating in a study. Information from a patient/subject is called observation. An observation may represent single or multiple pieces of information about the patients. For example, age, sex, height, weight, blood pressure level, cholesterol level, etc. of a patient.

Variables:

A variable is any attribute or event that can have different values. A variable may have different values when observed at different times for the same patient or a variable may have different values for different patients. Some examples of variables include age, weight, height, blood pressure, heart rate, blood sugar level, cholesterol level, etc. of patients under study.

1.3 Measurement Scale or Types of Data

In statistics, we mainly deal with data so it is sensible to start with a brief discussion of various types of data that may be encountered in medical research. The nature of the data is of major importance in relation to the choice of correct statistical methods of data analysis. Broadly data can be classified in the following two groups.

- Categorical data and
- Numerical data

However within these broad classifications there are various different types of data which we will discuss below.

1.3.1 Categorical Data

The simplest type of observation on a patient is the allocation of that patient to one of multiple possible categories. For example:

Smoker	yes/no
Sex	male/female
Marital status	married/single
Stress	low/medium/high
Pain	none/mild/moderate/severe
Blood group	A/B/AB/O

All of the above variables have been classified into two or more categories; data of this type is known as categorical data. For statistical analyses purposes categorical data must be coded by numbers, for examples, smoking status can be coded by “0” for smoker and “1” for non-smoker; stress level can be coded by “0” for low, “1” for medium and “2” for high. Categorical data can mainly be classified into two groups: nominal data and ordinal data, a brief discussion of them is given below.

Nominal data

When observations are classified into separate categories, with the categories having no ranking, the data is said to be nominal. The categories can be arranged in any order. Examples of nominal data are:

- Marital status: married, widowed, divorced, never married
- Sex: male-female
- Blood group: A, B, AB, O

For nominal data, if there are only two categories, the data is known as dichotomous or binary. For example, sex (male or female) and current smoking status (smoker, non-smoker) of a patient – each of these variables have only two categories. Similarly, data with more than two categories is called multinomial data, e.g., blood group has four categories of A, B, AB and O.

Ordinal Data

Whenever observations are not only different from category to category but can be ordered/ranked according to some criteria, they are said to be measured on an ordinal scale. For example:

- A cancer patient may classify his/her degree of pain as: 0 for no pain, 1 for mild pain, 2 for moderate pain and 3 for severe pain.
- Similarly, according to the number of cigarette consumption a patient can be classified as non-smoker, light smoker and heavy smoker.

The above examples show obvious natural order of categories. For ordinal data, although ranking exists among categories, any mathematical operations (e.g., differences, additions, divisions, multiplications) between numbers may lack meaning. To illustrate, consider the scores for degree of pain. The difference between scores of “2” and “3” is probably not the same clinical magnitude as the difference between scores of “0” and “1”. However, in psychological research, mathematical manipulation of ordinal data may be a common practice.

1.3.2 Numerical Data

Observations for which mathematical operations (eg, the differences) between numbers have meaning on a numerical scale are called numerical data or quantitative data. For example, number of children in a family, age of diabetic patients seen in a clinic, the weight of preschool children, etc.

There are two types of numerical data, namely:

- Discrete data and
- Continuous data

Discrete Data

Discrete data are measurements where the possible values are clearly separated from each other or can only take values equal to whole numbers or integers, e.g., number of children, family size, number of visits to the general practitioner in a year, number of children admitted into Monash Medical Centre in 2006. None of these data can take fractional values, this means, one can not say for example the number of members in

his/her family is 4.5. Mathematical operations such as addition, subtraction, division and multiplication are possible for discrete data.

Continuous Data

For continuous data, no matter what two values you mention, it is always possible to imagine more possible values in between them. Common examples include height, weight, age, serum cholesterol level, blood glucose level, etc. A patient may have morning and evening blood glucose levels 7 mmol/L and 7.5 mmol/L respectively, however, he/she may have been an infinite number of glucose levels in between morning and evening (e.g. 7.1, 7.2, 7.01, 7.52 mmol/L, etc.).

1.4 Populations & Samples

Populations

The average person thinks of a population as a collection of entities/items, usually people. A population or collection of entities may, however, consist of animals, machines, places etc. So, a population can be defined as follows: A population is the largest collection of entities for which we have an interest at a particular time. The term “target population” is often used in scientific literature to define the population of interest. Consider the following examples:

- Cholesterol level (CL) of diabetic patients in Victoria.
- Stress levels of all first year Monash students.
- Job satisfaction levels of all employees at Australian universities.
- Pulse rates of all patients seen in a clinic on a particular day (see Table 1.1 in the appendix).
- Length of intensive care unit (ICU) stay of patients after having coronary artery bypass graft (CABG) surgery in Australian hospitals.

Samples

A sample may be defined simply as a representative part of a population. Consider the population of pulse rate data. If we collect for analysis the pulse rate of only a fraction of these patients, for example 5 patients, we have only a part of our population that is, we have a sample.

Thus the sample is a small part of the population however this small part must represent the whole population as much as possible. Consider the age of five students from the first year biomedical science course in 2007 – 18, 18.5, 19, 65 and 18.5 years. There might be only a few students in the population who have higher age like 65 years. However, a student of age 65 years in a small sample of five students is unusual and hence this sample does not represent the age of students in the population – observations in a sample like this is known as outliers or extreme values or unusual values.

1.5: Selection of Sample

There are many ways to draw a sample from a population. The simplest way is simple random sampling (SRS). In SRS, every entity in the population has an equal chance to be included in the sample. Using SRS, samples can be drawn by:

- Lottery System and
- Random number System

A common example of lottery system is drawing of lotto lottery. Using a random number system, samples can be drawn manually or using computer – details about how to draw samples manually is discussed as follows.

Table 1.1 lists the pulse rates together with a “dummy” patient ID numbers. We will select a random sample from this pulse rate population. Consider Table 1.2, which contains 5-digit random numbers. Suppose we want to draw a random sample of size 4 pulse rates from the pulse rate data. Pick a starting spot on the table (you may wish to close your eyes), and then choose 4 random numbers either down or across the table. Suppose that the starting point is the 2nd row and 3rd column in the random number table (Table 1.2) which is 40311. Thus, starting at 40311, the chosen random numbers from the 3rd column of the random number table are 40311, 93540, 05085 and 80110. Since the population size is 99 (a two digit number), each of the selected numbers should not have more than two digits, so we may consider only the first two digits from each of them. Hence, the selected random numbers are 40, 93, 05 and 80. These numbers then correspond to the ID#s in Table 1.1 – select the pulse rates corresponding to these selected ID#s. Thus the pulse rates selected in the sample are 80, 66, 79 and 94. Similarly, you can draw random samples of any sizes (sample size must be less than the population size) from your population of interest.

Random Variable

Whenever we record age, height, weight, blood pressure, heart rate, blood sugar level of patients seen in a hospital, the result is frequently referred to as the value of the respective variable. When the values obtained arise as a result of chance factors, so that they cannot be exactly predicted in advance, the variable is called a random variable.

Alternatively, if in any study the patients are randomly selected, then the characteristics of interest (e.g., age, sex, height, weight, blood pressure, blood sugar level, etc) of the study patients are known as random variables.

Note 1: The choice of statistical methods to summarise, graph or analyse data depends critically on the scale of measurement i.e., nominal scale, ordinal scale, discrete scale and continuous scale.

Note 2: A numerical variable (discrete or continuous) can easily be converted to a categorical variable. For example, consider the pulse rate data in Table 1.1. The minimum and maximum pulse rates in this table are 56 and 95 beats respectively. We

can make categories for pulse rates like <60, 60-69, 70-79, 80-89 and 90+ beats and count the number of students which fall in each category. Thus, we can convert a numerical data to a categorical data by making different categories for the values of the variable of interests (see the following table for the pulse rate data in Table 1.1).

Pulse Rate Category	<60	60-69	70-79	80-89	90+	Total
#of students	3	17	42	33	5	100

Pulse rate in the above table is a categorical data with 5 different categories. Please note that if data is recoded on a categorical scale it can not be converted to a numerical scale. However, as was discussed above a numerical data can easily be converted to a categorical data.

Table 1.1: ID# and Pulse Rate for 100 patients seen in a CLINIC

ID#	Pulse	ID#	Pulse	ID#	Pulse	ID#	Pulse	ID#	Pulse
00	73	21	74	42	75	63	81	84	69
01	80	22	88	43	79	64	74	85	80
02	76	23	74	44	73	65	86	86	85
03	81	24	89	45	80	66	67	87	80
04	95	25	80	46	89	67	56	88	68
05	79	26	74	47	79	68	75	89	85
06	67	27	80	48	64	69	80	90	94
07	75	28	83	49	95	70	80	91	87
08	73	29	80	50	57	71	68	92	94
09	79	30	87	51	88	72	74	93	66
10	75	31	78	52	75	73	74	94	63
11	71	32	76	53	84	74	80	95	76
12	83	33	76	54	76	75	68	96	63
13	80	34	95	55	77	76	74	97	73
14	79	35	67	56	84	77	79	98	60
15	66	36	56	57	70	78	81	99	77
16	81	37	88	58	78	79	65		
17	64	38	70	59	71	80	76		
18	82	39	69	60	74	81	73		
19	60	40	80	61	76	82	79		
20	83	41	76	62	78	83	84		

Table 1.2: Random Numbers

75933	05250	79362	42350	37650	79788	25335	32049	63707
68531	69567	40311	16521	69648	35863	31181	46469	45242
93184	82616	93540	86013	57602	32260	44012	64961	65637
02302	22807	05085	56534	43573	47791	77031	46321	95628
43153	30553	80110	87607	17250	27264	78850	12048	35586
41066	30148	00860	62858	46809	31903	34738	77915	80790
28316	06672	38914	90497	95178	64608	38025	68181	29261
27989	94197	32122	88310	14709	69994	37726	24989	75495
31619	48676	64713	73498	50414	39009	30398	57971	57006
56726	27952	38589	04251	68956	42928	16710	30639	34116
28963	60880	28741	84568	64754	69143	74842	43250	73202
75555	44854	30396	42543	35374	29120	08167	27282	47122
58599	26917	72287	53123	45053	88809	16884	39602	73383
47450	23293	75429	11883	19237	43154	40181	34165	62547
55937	82563	73472	91301	91474	87244	51343	63042	70890

Module Summary:

- Variables and random variables
- Scale of Measurements
 - Categorical
 - Nominal
 - Ordinal
 - Numerical
 - Discrete
 - Continuous
- Sample and population
- Selection of sample using SRS (e.g., random number table)