Chapters 1. Introduction

January 11, 2011

1 What is Statistics

Definition 1: everything dealing with the collection, processing, analysis, and interpretation of numerical data about a **population** or a **sample** belongs to the domain of statistics.

Definition 2: Statistics is the science of acquiring & understanding data and making inferences about the population in the face of variability and uncertainty

The uncertainty includes:

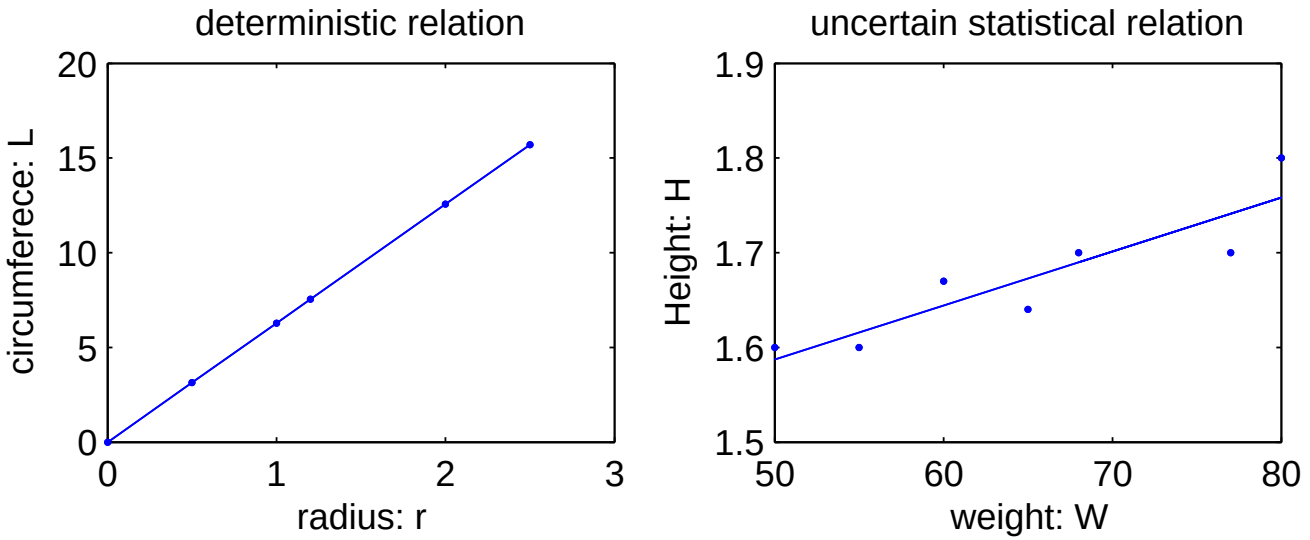
- sampling uncertainty (called random sample, observation random error)
- Dependence uncertainty (E.g. people's height and weight; today's stock market index and tomorrow's ...)

  Example Deterministic relationship: circumference of a circle (L) and its radius (r)

$$L = 2\pi r$$

  Statistical relationship: a person's height (H) and his/her weight (W)

$$H \stackrel{?}{=} \text{a funciton } f(W) \quad \not\checkmark$$



The uncertainty shakes the foundation of the deterministic sciences (e.g. mathematics), because only probabilistic statements can be made in describing a phenomenon with uncertainty.

Statistics may refer to

1. data

2. statistics science: methodology in making inference and prediction

3. formulas used in making inference.

Statistics in daily life

We are constantly being bombarded with statistics and statistical information.

E.g. Customer Surveys; Medical News; Economic Predictions; Marketing Information; traffic accidents; body fitness

- How can we make sense out of all this data?
- How do we differentiate valid from flawed claims?

Example After visiting a country or a city, we usually like to make comments (inferences) about the country and city based on what we observed, e.g. "people there are tall", ".... nice", "things are cheap", etc. These are statistical inferences. But we dont think about how reliable our inferences are!

Statistics in scientific research

Applying statistical methods to different disciplines, new research and applications areas are developed:

Econometrics	Technometrics
Biometrics	Bioinformetrics
Environmetrics	psychometrics
Actuarial science	Epidemiology
Statistical physics	.....

2 Statistical Population

A **(statistical) population** is defined as all the people or items with a characteristic we wish to understand. Each person or item in the population is called a **unit** or an **individual**.

- Sometimes, a population is obviously defined. For examples, (1) people in a country (or a place); (2) all students currently enrolled in a school; (3) daily temperature in 2011 in Singapore. In all these examples, the total number of individuals are finite, denoted by N.

- In other cases, the 'population' may be less tangible. For example, (1) in the study of the stability of a person in measuring the length of a robe, the population is outcomes of all repeated measurements/experiments. Thus, the population is 'super'. (2) the products from a production line in a factory. In these examples, $N = \infty$.
- Even less tangible situations where experiment cannot be done repeatedly. For example, we are interested in the weather on 01/01/2020? (to be discussed later)

Any characteristic of interest for each unit in the population, qualitative or quantitative, is called a **variable** or **random variable**[1]. E.g., for people in a country, the variables of interest might be height (x), weight (y), income (z), gender (w). For the products, we might be interested in its quality (x), size (y). For students in a school, student grades; For potatoes from a farmer; weight of a potato, For a coin; the chance of heads-up in a flip, etc. For individual i, we can further denote its variable value as x_i, y_i, ...

Data are the **observed values** of a random variable. E.g. student marks: 67, 74, 71, 83, 93, 55, 48

[1] because we do not know its value until we observe it

Example A. (All) 80 male adults in a village with their weights below (in kg)

12 76.1	24 75.7	36 76	48 76.1	60 75.8	72 76.3	
11 75.6	23 76	35 76.2	47 76.2	59 76	71 76.1	
10 76	22 76	34 75.7	46 75.7	58 75.7	70 75.9	
9 76	21 75.8	33 76	45 75.6	57 75.7	69 76	
8 75.8	20 75.9	32 75.8	44 75.9	56 76.1	68 76	80 75.9
7 76.1	19 76	31 75.7	43 76.1	55 76	67 76.1	79 76
6 75.9	18 76.2	30 76	42 76.4	54 76.1	66 75.6	78 75.8
5 75.9	17 75.6	29 75.9	41 75.9	53 76.1	65 75.6	77 75.7
4 75.9	16 75.8	28 76	40 76	52 75.9	64 76.2	76 76.3
3 76.2	15 76.3	27 76	39 76.2	51 75.8	63 76.1	75 76
2 76.1	14 75.9	26 75.9	38 75.7	50 76.3	62 75.9	74 76.3
1 76.4	13 76	25 76.2	37 75.6	49 75.9	61 76.2	73 75.8

In this example, $N = 80$, $x_1 = 76.4, x_2 = 76.1, ..., x_{80} = 75.9$.

We also call all the data { 76.4, 76.1, ..., 75.9 } the population, or generally we call $\{x_1, ..., x_N\}$ the population.

Types of variables

- Numerical/Quantitative variable [real numbers]:

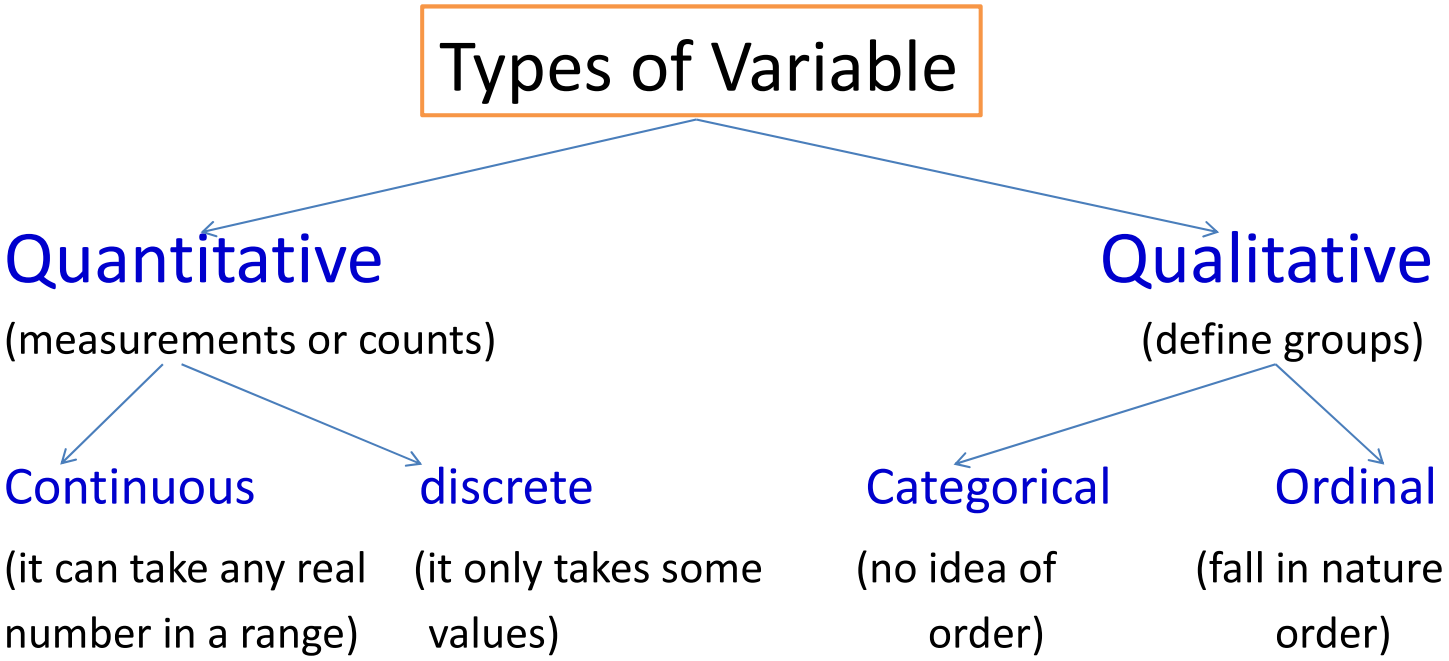
  E.g. (1) height of a randomly selected student (2) temperature of coffee bought from Macdonald (3) etc.

- Qualitative/Categorical variables [labels rather than numbers]:

  E.g. (1) Marital status: single, married, divorced, widowed. (2) favorite color (3) the part of a new automobile that breaks first (4) the reason a person gets mad at his/her spouse (5) etc.

Quantitative variables are further broken down into

- Continuous variable — variable can be any real number within a range. E.g. Normally measurement data [weight, time, volume, etc]
- Discrete variable — Data can only take very specific values which can be listed. E.g. normally count data [number of traffic accidents in a city, number of typos on a page, etc]


Types of Variable
Quantitative
(measurements or counts)
Qualitative
(define groups)
Continuous
(it can take any real number in a range)
discrete
(it only takes some values)
Categorical
(no idea of order)
Ordinal
(fall in nature order)

Distribution of the data

A frequency distribution is a display of the number/frequency (denoted by n_i or f_i) of occurrences of each value in a data set. Suppose there are k sets/groups,then

$$n_1 + n_2 + ... + n_k = n(or N)$$

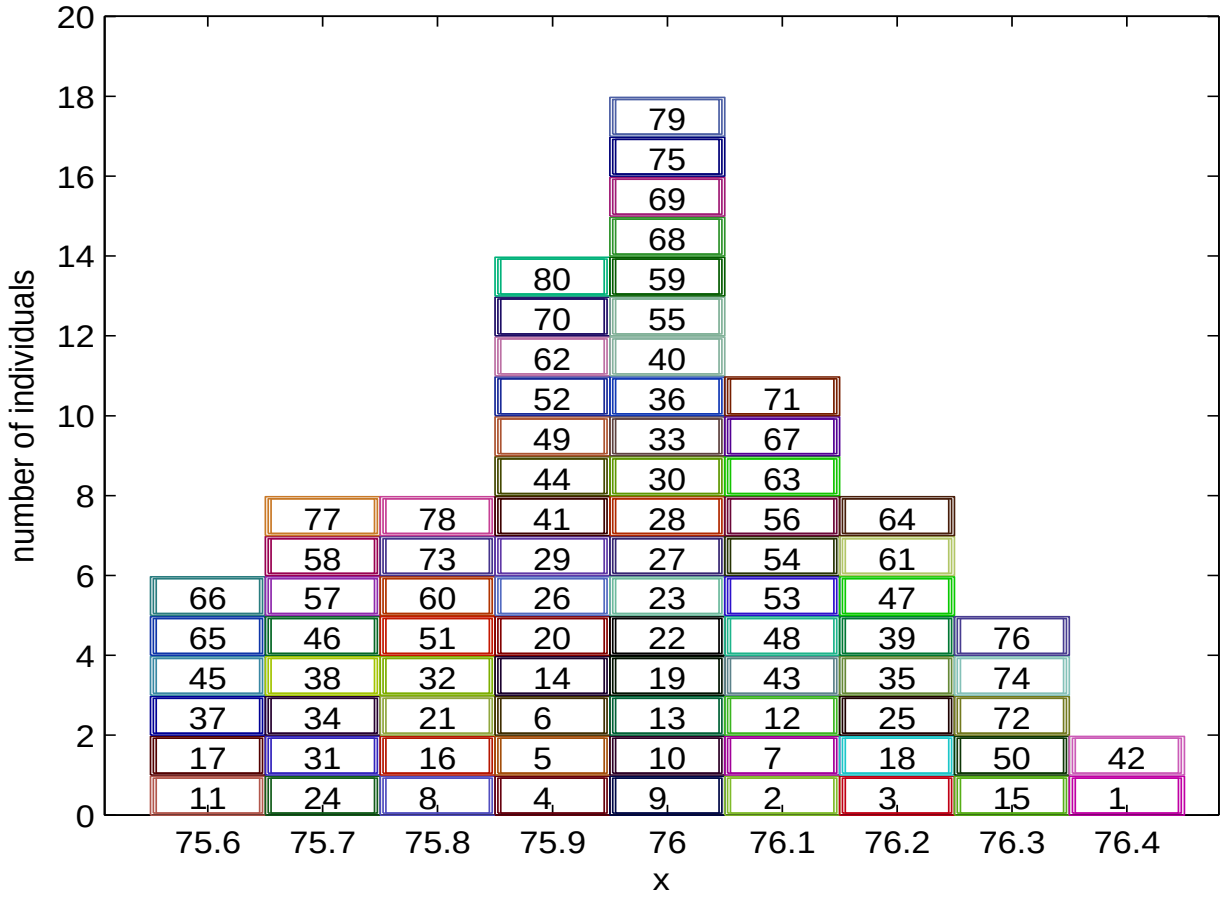
A relative frequency distribution is a display of the percentage of occurrences (denoted by p_i) of each value in a data set. Then

$$p_1 + ... + p_k = 1$$

Example A(continued).

value	frequency (n_i or f_i)	relative frquency (p_i)
75.6	6	0.075
75.7	8	0.100
75.8	8	0.100
75.9	14	0.175
76.0	18	0.225
76.1	11	0.1375
76.2	8	0.100
76.3	5	0.0625
76.4	2	0.025
total	80	1.00

For Example A, after grouping according to the values, we can stack the individuals as follows



The number of individuals with $x = x_i$ is n_i, denoted as $\#(x = x_i) = n_i$, e.g. $\#(x = 75.9) = 14$. The proportion for $x = x_i$ is $P(x = x_i) = p_i$, e.g. $P(x = 75.9) = 0.175$.

Example B. The ethnic groups in Singapore (categorical variable)

Ethnic	Relative Frequency
Chinese	75.2%
Malays	13.6%
Indians	8.8%
Eurasians and others	2.4%

Disadvantages of studying the entire population:

- the cost is too high (for example, **census** of population in a big country)
- Sometimes it is impossible (the life expectancies of lamps from a production line),
- the population is dynamic in that the individuals making up the population may change over time (e.g. the chance of head-up in tossing a coin).

3 Sample and Sampling

A statistical **sample** is a subset (of individuals or units) from a statistical population. The number of observations (individuals) in the sample is called sample size, usually denoted by n.

Example A (continued). For the above population, if individuals 11, 46, 32, 14, 33, 2, 7, 47, 72 are sampled, then we have a subset with their weights

$$\{75.6, 75.7, 75.8, 75.9, 76, 76.1, 76.1, 76.2, 76.3\}$$

we call this subset a sample, with $n = 9$.

- **Advantages of using samples**: the cost is lower (because $n << N$), data collection is faster.
- **Disadvantage of using samples:** The accuracy (or efficiency) of inference is a big concern. Usually, the bigger the sample is, the more accurate the inference will be.

Sampling is the statistical practice concerned with the selection of a sample within a population intended to yield some knowledge about the population.

Random numbers and **random sample**

The selection of a sample from a population must be done impartially and objectively.

- For finite population, each individual must have equal probability to be selected.
- For infinite population, each value is selected with the chance being equal to the relative frequency of the value (to be discussed later).

The random sample can be obtained by using random number table or other random number generators

In R if we want to get a random sample with size n from population with size N, we can use

sample(N, n)

Example A(continued) If we want to draw 5 individuals,

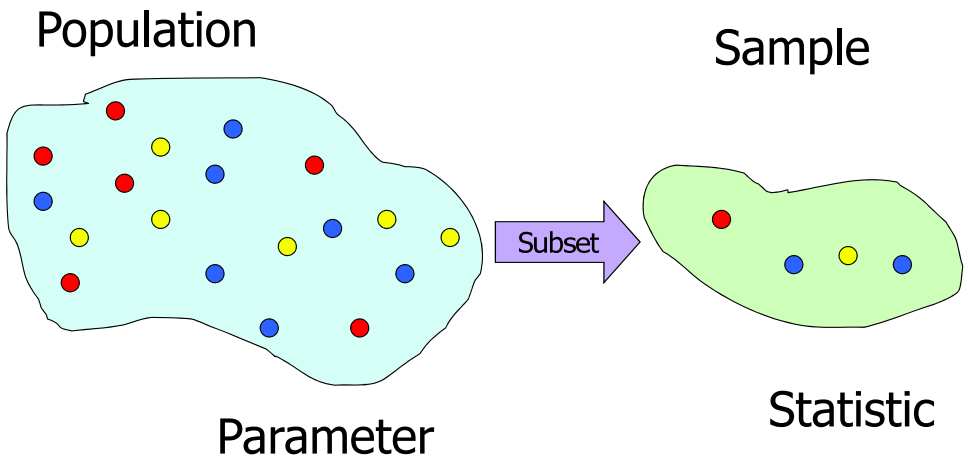
> sample(80, 5)

[1] 9 50 73 60 53

i.e. individuals 9, 50, 73, 60, 53 should be drown to form a sample

(every time, you may get different samples)

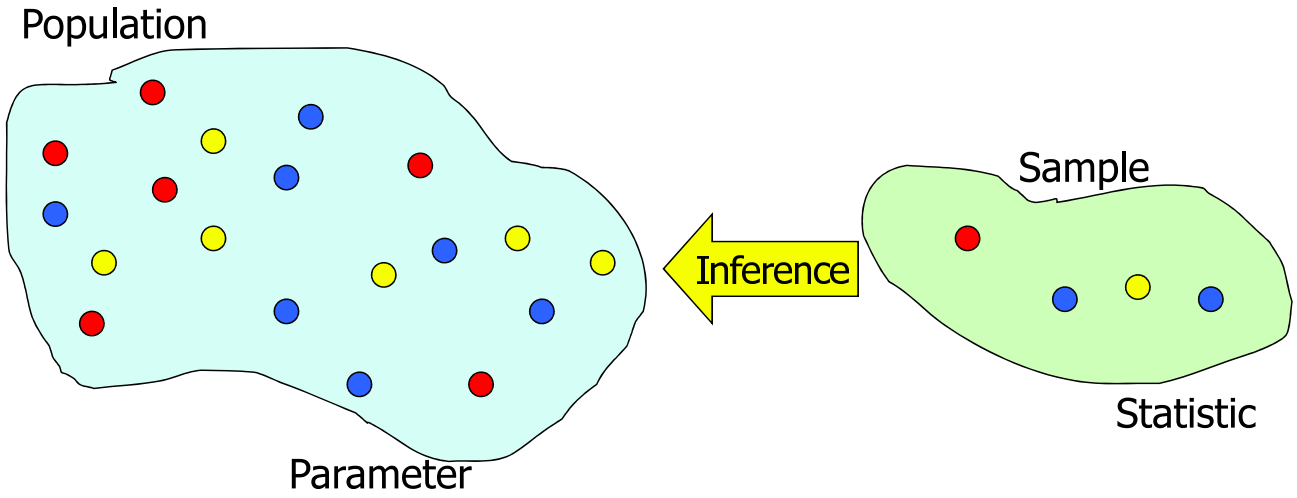
The relationship between population and sample



Populations have Parameters,

Samples have Statistics.

Statistical inference is the ***process*** of making an estimate, prediction, or decision about a population based on a sample.



What can we ***infer*** about a Population's Parameters based on a Sample's Statistics?