Chapter 1
The Where, Why, and How of Data Collection

Chapter Goals:
After completing this chapter, you should be able to:

- Describe key data collection methods.
- Know key definitions:
  - ♦Population vs. Sample
  - ♦Primary vs. Secondary data types
  - ♦Qualitative vs. Qualitative data
  - ♦Time Series vs. Cross-Sectional data
- Explain the difference between descriptive and inferential statistics.
- Describe different sampling methods.

1.1 : What is Business Statistics?

It is a collection of tools and techniques that are used to convert data into meaningful information in a business environment.

Tools of Business Statistics:

- **Descriptive statistics:**
  - Collecting, presenting using graphes, and describing data numerically.
- **Collect data**
  - e.g. Survey, Observation, Experiments
- **Present data**
  - e.g. Charts and graphs
- **Characterize data**
  - e.g. Sample mean = $\frac{\sum_{i=1}^{n} x_i}{n}$
- **Inferential statistics:**

To draw conclusions and/or to make decisions concerning a population based only on sample data.

Tools that allow a decision maker to reach a conclusion about a population of data based on a subset of data from the population, there are two catagories of statistical inference tools: estimation and hypothesis testing;

- Estimation
  Estimate the population mean weight using the sample mean weight.
- Hypothesis Testing
  Use sample evidence to test the claim that the population mean weight is 120 pounds.

1.2: Tools of collecting Data:

Objectives:

1. To define the methods of collecting data.
2. To define the problems of collecting data.

Data collection methods:

There are many methods to collecting data. the moest useful methods are:

1. Experiments:

any process that generate data as its outcome.

Experemental design: it is a plan for performing an experement in which variable of interest is defined, the variable can measured or observed.

2. Telephone surveys:

it must be short from 1 to 3 minutes. It is efficient and not expensive

3. Mail questionnaires and other written survey

It is most frequentlly used method and it is similar to telephone survey but not the same limit time. It can be low – cost, effective means of collecting data.

4. Direct observation and personal interviews: it is method to collecting data, any interview can be

a. Structured interview: the questions are scripted (written).
b. Unstructured interview: interviews with different questions based on the responses.

there is disadvantages for this method

a. it increase the cost.
b. it needs more time than other methods.
c. It is depend on the person who makes the interview.

Example: studying human behavior about a new product on the market

Survey Design Steps:

1. Define the issue: what are the purpose and objectives of the survey?
2. Define the population of interest.
3. Formulate survey questions.
   - make questions clear and unambiguous
   - use universally-accepted definitions
   - limit the number of questions
4. Pre-test the survey
   - pilot test with a small group of participants
   - assess clarity and length
5. Determine the sample size and sampling method
6. Select Sample and administer the survey

Types of Questions

A written survey can contain

2. Closed – end questions:
   questions that require the respondents to select from a short list of defined questions.

Example: What is your Major: __business__liberal arts __science __other

3. Open - end Questions:
   Respondents are free to respond with any value, words, or statement

Example: What did you like best about this course?

4. Demographic Questions:
   Questions about the respondents' personal characteristics.

Example: Gender: __Female __ Male

Note:

1. for any survey the budget must be considered
2. telephone survey can use open – end questions

Data Sources

1. Primary Data Collection: the data that obtained by
   a. observations,
   b. survey
   c. or by experimentations.
2. Secondary Data Compilation: any data from previous study (Print or Electronic data)

NOTES:

1. any collected data may contain errors.
2. any collected data may be biased, there are many sources of bias
   a. Response bias: the way that you make the interview
   b. Non response bias: the questions without answering
   c. Selection bias: bias can be interjected through the way subjects are selected for the data collection
3. measurments errors.
4. observer bias: some time the data are deferent from person to person.

1.3: Populations and Samples and sampling techniques:

Objectives:

1. To define the population, the sample, parameter and statistics .
2. To define the sampling techniques.

Population: is the set of all items or individuals of interest. i.e. A population is the entire collection of things under consideration.

A parameter is a summary measure computed to describe a characteristic of the population

Examples: All likely voters in the next election.
All parts produced today.
All sales receipts for November.
All students in kfupm.

Sample is a subset of the population. i.e. A **sample** is a portion of the population selected for analysis

A statistic is a summary measure computed to describe a characteristic of the sample

Examples: 1000 voters selected at random for interview
A few parts selected for destructive testing
Every 100^{th} receipt selected for audit

Census: an enumeration of the entire of measurements taken from the whole population.

Why Sample?

- Less time consuming than a census.
- Less costly to administer than a census.
- It is possible to obtain statistical results of a sufficiently high precision based on samples
- Less error than census.

Note:

1. Ay numarical measures (average, variance …) that are computed from an entire population are called parameters.
2. Ay numarical measures (average, variance …) that are computed from a sample are called statistics.

Sampling Techniques:

It falls into two caragories

1. non statistical sampling technique (non propability sample) this tipe of sample used Judgement or Convenience.
2. statistical sampling technique (propbablity sample): Items of the sample are chosen based on known or calculable probabilities

Propbablity Sample

1. Simple Random Sample:

a simple random sample from a population is a sample chosen randomly, so that each possible sample has the same probability of being chosen, also every item in the sample has the same chance to being selected.

Selection may be with replacement or without replacement (small populations)

Samples can be obtained from a table of random numbers or computer random number generators

Advantages

1. It is free of classification error,
2. It requires minimum advance knowledge of the population.
3. It best situations where the population is fairly homogeneous and not much information is available about the population.
4. If these conditions are not true, stratified sampling may be a better choice.

Drawing Simple Random Samples using a Table of Random Numbers

An easy way to select a SRS is to use a random number table, which is a table of digits 0,1,…,9, each digit having equal chance of being selected at each draw. To use this table in drawing a random sample of size n from a population of size N, we do the following:

1. Label the units in the population from 0 to N –1.
2. Find ***r***, the number of digits in N –1 . For example; if N = 100, then r = 2.
3. Read ***r*** digits at a time across the columns or rows of a random number table.
4. If the number in (3) corresponds to a number in (1), the corresponding unit of the population is included in the sample, otherwise the number is discarded and the next one is read.
5. Continue until ***n*** units have been selected.

If the same unit in the population is selected more than once in the above process of selection, then the resulting sample is called a **SRS with replacement**; otherwise it is called a **SRS without replacement**. The observations in the sample are the enumeration or readings of the units selected.

Example: To draw a SRS, consider the data below as our population. In a study of wrap breakage during the weaving of fabric, 98 pieces of yarn were tested. The number of cycles of strain to breakage was recorded for each yarn and the resulting data are given in the following table.

86	175	282	38	211	497	246	393	198	292	61	246	244	193
146	176	224	337	180	182	185	396	264	131	121	279	20	188
251	76	149	66	93	423	188	203	105	169	90	81	284	77
653	264	180	151	315	185	568	829	203	262	229	186	194	121

98	15	325	341	353	229	55	239	124	88	166	290	277	280
249	364	250	40	571	400	55	236	137	264	135	398	143	400
400	195	196	40	124	338	61	286	135	20	597	71	350	280

Here we have a population of size N = 98. To draw a simple random of size n =10 *without* replacement, we proceed as follows:

1. Label the units in the population from 00 to 97.
2. Find ***r***, the number of digits in N-1. For example, if N =98, then ***r*** = 2.
3. Read 2 digits at a time across the columns of a random number table.

Part of a **Random Number Table**

8571	7683	5118	7669	6126	3663	3059	7807	9219	4383	9021	7013	0233	3348	4077
0864	5055	8631	5770	0505	0386	9792	1690	4874	3084	0228	8539	9375	5046	8635
4753	1992	8182	2658	2914	4005	1577	1714	7862	7009	0252	3070	1563	3008	3716
1267	1063	4415	8496	6779	1563	7833	5351	2278	0674	1252	6813	4016	3961	6890
9497	0105	5626	0529	0602	4573	1499	7772	7759	9405	9502	3408	6931	7946	4655
6823	7365	6140	0357	7069	7715	9083	6180	1131	7059	9808	9803	7883	5943	6649
6532	4048	3044	8035	1045	8349	5422	0315	7470	7679	1726	1390	4997	5632	9033
8184	8336	5684	5846	7056	2847	4715	2869	2576	5373	8175	0384	5348	8232	8186
5605	0939	9380	1647	7307	5893	7569	7092	4437	2722	7807	5908	5425	9679	2348
4926	1561	7299	2195	5374	3664	8269	5241	4436	5265	7571	8299	6006	2142	2273
0933	6131	2406	0715	5069	1663	8015	9120	0667	4884	8601	3370	3449	7158	8950
7413	9526	9670	3075	8321	8295	6327	5475	5650	9061	7687	3849	2207	6910	4166

Suppose we read the first two digits of the first two columns of the above random number table to get the following numbers

85 71 76 83 51 18 76 69 61 26 36

4. Since the random digit 85 corresponds to a unit in (1), we select unit 85 of the population in the sample. If any random digit in (3) exceeds 99, the random digit is discarded and the next one is read. After selecting 6 random numbers of two digits, we find a random number 76 which is discarded for SRS without replacement as it appeared before.

Continue until n = 10 units have been selected. Thus we have the sample units:
85 71 76 83 51 18 69 61 26 36
so that the sample observations are:
195 364 55 400 262 180 280 229 20 105

A SRS with replacement in the above example would be:
195 364 55 400 262 180 55 280 229 20.

2. Stratified Sampling:

Stratification is the process of grouping members of the population into relatively homogeneous subgroups before sampling. The population is divided into subgroups (called strata) according to some common characteristic such that every element in the population must be assigned to one (and only one) stratum.

To develop strata, we look at the characteristic of interest for which items are quite homogeneous.

A SRS is selected from each stratum, then the samples from all strata are combined into one sample.

The main objective is to increase precision.

Advantages

1. Focuses on important subpopulations but ignores irrelevant ones.
2. Improves the accuracy of estimation.
3. Efficient.

Disadvantages

1. Can be difficult to select relevant stratification variables.
2. Not useful when there are no homogeneous subgroups.
3. Can be expensive.
4. Requires accurate information about the population, otherwise it introduces bias.
5. looks randomly within specific sub headings.

Example: In general the size of the sample in each stratum is taken in proportion to the size of the stratum. This is called *proportional allocation*. Suppose that in a company there are the following staff:

- male, full time: 90
- male, part time: 18
- female, full time: 9
- female, part time: 63

then the total is 180
and we are asked to take a sample of 40 staff, stratified according to the above categories. The first step is to find the total number of staff (180) and calculate the percentage in each group.

- % male, full time $= \left(90/180\right) \times 100\% = 0.5 \times 100\% = 50\%$
- % male, part time $= \left(18/180\right) \times 100\% = 0.1 \times 100\% = 10\%$
- % female, full time $= \left(9/180\right) \times 100\% = 0.05 \times 100\% = 5\%$
- % female, part time $= \left(63/180\right) \times 100\% = 0.35 \times 100\% = 35\%$

This tell us that of our sample of 40:
50% should be male, ful time, 50% of 40 = 20
10% should be male, part time, 10% of 40 = 4
5% should be female, full time, 5% of 40 = 2
35% should be female, part time, 35% of 40 = 14

Select SRS from each strata, then the final sample is the sum of all sub - samples.

3. *Systematic Sampling:*

Systematic sampling is the selection of every $\boldsymbol{k}^{th}$ element from a sampling frame, where $\boldsymbol{k}$, is the *sampling interval*, is calculated as:

$$k = \frac{Size\ of\ \ the\ \ population}{Size\ \ of\ \ the\ \ sample} = \frac{N}{n}$$

Using this procedure each element in the population has a known and equal probability of selection. This makes systematic sampling functionally similar to simple random sampling. It is however, much more efficient and much less expensive to do.
The researcher must ensure that the chosen sampling interval does not have a pattern. Any pattern makes the sample NOT random. A random starting point must also be selected.

The method of selecting a systematic random sample

1. Decide on sample size: n.
2. Divide frame of N individuals into groups of $\boldsymbol{k}$ individuals: $k = \frac{N}{n}$.
3. *Randomly* select one individual from the 1^{st} group.
4. Select every $\boldsymbol{k}^{th}$ individual thereafter.

Example: Let N = 64, n = 8,Then $k = \frac{64}{8} = 8$

Chose any nomber between 1and 8 randomly, then take it as the first observation in the sample, then add 8 to the number that you select, the result will be the second observation in the sample and so on.
Asume that the first is number 3 then the systematic random sample is:
3, 11, 19, 27, 35, …., etc

4. *Cluster Sampling:*

Cluster sampling is a sampling technique used when "natural" groupings are evident in the population. The total population is divided into these groups (or clusters) (usually on a geographical basis), and a sample of the groups is selected. Then the required information is collected from the elements within each *selected* group. This may be done for every element in these groups, or a subsample of elements may be selected within each of these groups.

All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique

The main objective of cluster sampling is to reduce costs by increasing sampling efficiency.

The main difference between cluster sampling and stratified sampling:
Question 34. page 19:

1. In cluster sampling, the idea is to break the population into *heterogeneous* groups called clusters (usually on a geographical basis) such that each cluster looks as much like the original population as possible. Then clusters are randomly selected (usually a SRS) and from the cluster, individual items can be selected using a statistical sampling method.

   In stratified random sampling, the population is divided into *homogeneous* groups called strata. The idea is to make all items in a stratum as much alike as possible with respect to the variable of interest thereby reducing the number of items that will need to be sampled from each stratum.

2. In cluster sampling: the cluster is treated as the sample unit so analysis is done on a population of clusters.

   In stratified sampling, the analysis is done on elements within strata.

3. In cluster sampling only the selected clusters are studied.

   In stratified sampling, a random sample is drawn from *each* of the strata.

4. The main objective of cluster sampling is to reduce cost by increasing sampling efficiency.

This contrasts with stratified sampling where the main objective is to *increase precision*.

1.4: Data Types and Data Measurement Levels:

Objectives:

1. To present two general types of data.
2. To classify data measurements into one of four levels.

Data Types

1. **Qualitative data** (Categorical data): Defined categories.
   Marital Status, Political Party, Eye Color,…
2. **Quantitative data** (Numerical data).
   It is two parts
   a. Discrete: Counted items
      number of children, Defects per hour,…
   b. Continuous: *Measured* characteristics
      Weight, Voltage,…
3. **Time Series Data** (Ordered data values observed over time).
4. **Cross Sectional Data** (Data values observed at a fixed point in time).

Example:

Sales (in $1000's)

Time Series Data

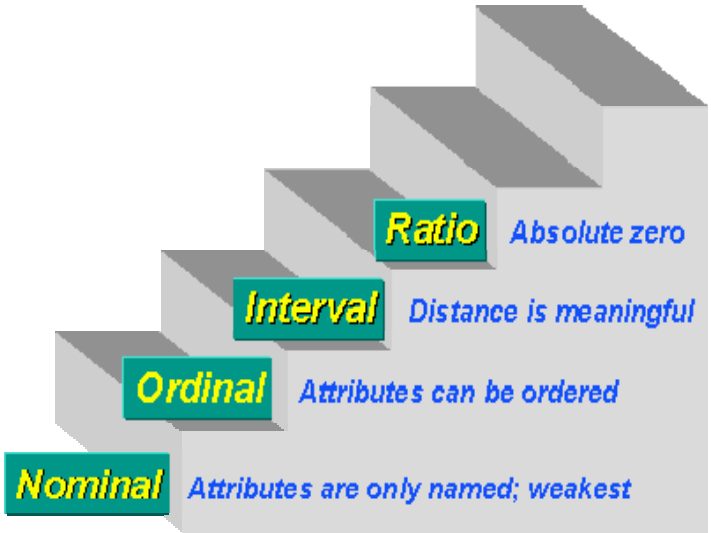
	2003	2004	2005	2006
Atlanta	435	460	475	490
Boston	320	345	375	395
Cleveland	405	390	410	395
Denver	260	270	285	280

Cross Sectional Data

Data Measurement Levels:

There are typically *four* levels of measurement that are defined:

1. **Nominal**: Lowest Level
   We can do basic Aanalysis
   ategorical Codes
   ID Numbers
   Category Names
2. **Ordinal**: Higher Level
   We can do Mid-level Analysis
   Rankings
   Ordered Categories
3. **Interval**



4. **Ratio**: Highest Level
   We can do complete cnalysis
   Measurements

In nominal measurement the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. For example, the numbers of the players in basketball are measures at the nominal level. A player with number 30 is not more of anything than a player with number 15, and is certainly not twice whatever number 15 is.

In ordinal measurement the attributes can be rank-ordered. Here, distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as 0=less than H.S.; 1=some H.S.; 2=H.S. degree; 3=some college; 4=college degree; 5=post college. In this measure, higher numbers mean more education. But is distance from 0 to 1 same as 3 to 4? Of course not. The interval between values is not interpretable in an ordinal measure.

In interval measurement the distance between attributes does have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values has meaning. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).

Finally, in ratio measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most "count" variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that "...we had twice as many clients in the past six months as we did in the previous six months."

Notes:

1. The interval data has ordinal properties (> or < or =).
2. An interval data does not have a true zero.
3. The ratio data has all charasteristics of interval data and also have a true zero meaning (zero means nothing)

Question 54. page 23:

a. Ratio data.
b. Nominal data.
c. Nominal data.
d. Ratio data.

e. Interval data.
f. Nominal or ordinal data.