Chapter 4: Random Sampling and data Description

CLO4	Explain random sampling and data description.

1. Definition of Statistics

- Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions.
- In any conclusion, statistics provides measures of confidence (trust or faith).

2. The process of statistics

- **Identify the research objective:**
  - Identify the group to be studied and the questions to be answered.
  - The group is called the population.
  - A member of the population is called an individual.
  - A subset of the group to be studied is called a sample.

- **Collect the data needed to answer the questions:**
  - Usually, access to the entire population is difficult or impossible.
  - In conducting research, we typically use a sample, which is a subset of the population.

- **Organize and summarize the collected data:**
  - This is called descriptive statistics (quantitative description).
  - The information collected is described using numerical, measurements, charts, and tables.

- **Draw conclusions:**
  - The information collected from the sample is generalized to the population.
  - The methods that take a result from a sample, extend it to the population, and measure the reliability of the results are called inferential statistics (process of deducing properties of an underlying distribution by analysis of data).

Example 1:

A poll was conducted on 4-7 October 2007 about American gun-control laws.

- Identify the research objective:

  Determine the percentage of people aged 18 or older who favor more strict laws. Therefore, population is people 18 or older.

- Collect the needed data:

  A sample of 1010 people aged 18 or older was surveyed. 515 stated they were in favor of more strict laws.

- Organize and summarize the data:

  51% are in favor of more strict laws. This is a descriptive statistic.

- Draw conclusion:

  The researchers are 95% certain that the percentage of all people 18 or older in favor of more strict laws is between 48% and 54% (3% error).

3. Variables

- Once a research objective is stated, a list of the information the researcher desires about the individual must be created.
- Variables are the characteristics of the individual within the population.
- Variables can be classified into two groups: quantitative and qualitative.
- **Qualitative or categorical variable:** allow for classification of individuals based on some attribute or characteristic (generally not be measured with a numerical result).
- **Quantitative variables:** provide numerical measures of individuals. Arithmetic operations such as addition can be performed on the values of quantitative variable and provide meaningful results.
- Quantitative variables can be continuous or discrete.

- **A discrete variable**: is a quantitative variable that has either a finite number of possible values or a countable number of possible values. Countable means the values result from counting such as 0, 1, 2, …
- **A continuous variable**: is a quantitative variable that has an infinite number of possible values that are not countable.

Example 2

Determine whether the following variables are qualitative or quantitative.

(a) Gender (b) Temperature

(c) Number of red cars (d) Zip code

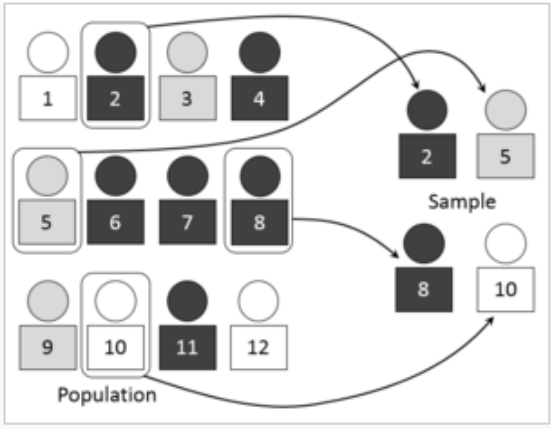
---

4 Observational Studies versus Design Experiments:

- Ones the research question is developed, we must develop methods for obtaining the data to answer the questions.
- There are two methods for collecting data: **observational studies** and **designed experiments.**
- **An observational study** measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, the researcher observes the behavior of the individuals without trying to influence the outcome of the study.
- **In a designed experiment (controlled experiments):** a researcher assigns the individuals to a certain group, intentionally changes the value of the explanatory variable, and then records the value of the response for each group.

5 Simple Random Sampling

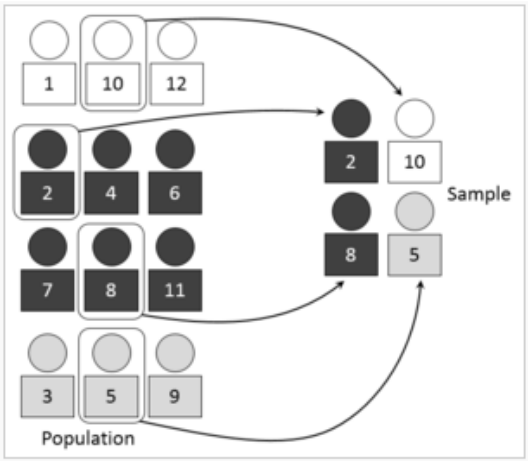
- A sample of size n from a population of size *N* is obtained through random sampling if every possible sample of size *n* has the same chance of occurring. The sample is then called a simple random sample.
- A list of all individuals within a population is called a frame.



A visual representation of the sampling process.

6 Other Effective Sampling Methods

- **Stratified Sample:** is obtained by separating the population into non-overlapping groups called strata and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogeneous (or similar) in some way.

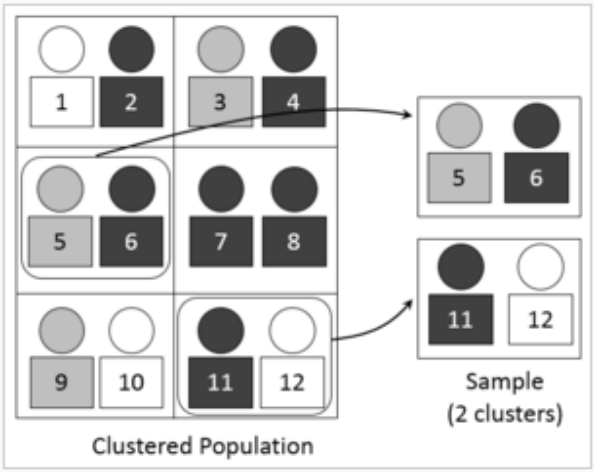


A visual representation of selecting a random sample using the stratified sampling technique

- **Systmatic Sample:** is obtained by selecting every *k*th individual from the population. The first individual selected corresponds to a random number between 1 and k.

A visual representation of selecting a random sample using the systematic sampling technique

- **A cluster sample:** is obtained by selecting all individuals within a randomly selected group of individuals.



A visual representation of selecting a random sample using the cluster sampling technique

- **A convenience sample:** is a sample in which the individuals are easily obtained and not based on randomness (close to hand).
- There are many types of convenience samples, but the most popular are those in which the individuals in the sample are self-selected. These are voluntary response samples.
- Note that results obtained from convenience sample may not be accurate because the individuals who decide to participate generally have strong opinions about the topic.

7 Bias in sampling

Survey results are typically subject to some error. If the results of the sample are not representative of the population, then the sample has ***bias*** (error). There are three sources of bias in sampling:

7.1 Sampling Bias:

- This technique used to obtain the individuals in the sample tends to favor one part of the population over another.
- Causes of sampling bias:
    - Any convenience sample because individuals are not chosen randomly.
    - Under coverage which occurs when the proportion of one segment in the sample is less than it is in the population.

7.2 Nonresponse Bias:

- Exists when individuals selected in the sample who don't respond to the survey have different opinions from those who do (Failure to obtain complete data from all selected individuals).

  Causes:
    - Individual selected in the sample do not respond.
    - The interviewer was unable to contact the individuals.

7.3 Response Bias:

- Exists when the answers on survey don't reflect the true opinion of the individuals.

Causes: (when respondents misunderstand a question, or find it difficult to answer)

- Interviewer error.
- Poorly-worded questions
- Incorrect data entry.

8 The Design of Experiments

- In the previous sections we studied obtaining data through surveys. Now, we discuss obtaining data through designed experiments.
- **A designed experiment:** is a controlled study to determine the effect of varying one or more explanatory variables (or factors) on a response variable, which represents the variable of interest.
- Any combination of the values of the factors is called **treatment**.
- The key ingredients of a well-designed experiment are: control, manipulation, randomization, and replication.
- **The experiment unit** (or subject) is a person, object, or some other well-defined item upon which a treatment is applied. The subject is analogous to the individual in a survey.

8.1 Steps in conducting an experiment:

- Identify the problem to be solved:
- The statement of the problem should be as explicit as possible, and must identify the response variable and the population to be studied. The statement is referred to as the claim.
- Determine the factors that affect the response variable.
- Determine the number of experimental units.
- Determine the level of each factor:
- There are 3 levels to deal with the factors:
    - **Control:** fix the variables whose effect on the response variable is not of interest.
    - **Manipulate:** manipulate the variables whose effect on the response variable is of interest.
    - **Randomize:** randomize the experimental units to various treatment groups so that the effect of factors whose levels cannot be controlled is averaged out (or minimized).
- Conduct the experiment: Collect and process the data.

- Test the claim: Inferential statistics is a process in which generalizations about the population are made on the basis of results obtained from a sample. The level of confidence in the generalization is also provided.

- **A completely randomized design** is one in which each experimental unit is randomly assigned to a treatment.
- **A matched pair design:** in which the experimental units are paired up. The pairs are matched up so that they are somehow related (for example, a person before and after a treatment, twins, husband and wife, and so on). There are only two levels of treatment.
- **Blocking:** is grouping similar (homogeneous) experimental units together and then randomizing the experimental units within each group to a treatment. Each group is called a block.
- **A randomized block design:** is used when the experimental units are divided into homogeneous groups called blocks. Within each block, the experimental units are randomly assigned to treatment. This design is useful to solve the issue of confounding, which occurs when the effect of two factors on response variable cannot be distinguished.

9. Organizing Data

- In the Chapter 6, we discussed how to identify the research objective and collect data from either observational studies or designed experiments.
- Thus the obtained data are called: **raw data**.
- Raw data must be organized into a meaningful form, so we can recognize what the data are telling us.
- Tables and graphs allow quick overview of the collected information.
- **Frequency distribution:** lists each category of data and the number of occurrences for each category.

- **Relative frequency**: is the proportion (or percent) of observations within a category. It is found using the formula:

$$\textbf{relative frequency} = \frac{\textbf{frequency}}{\textbf{sum of all frequencies}}$$

- **Relative frequency distribution:** lists each category of data to gather with the relative frequency (sum is one).

Example 3:

A physical therapist wants to get idea of the types of rehabilitation required by his patients. Below, a simple random sample of 30 patients:

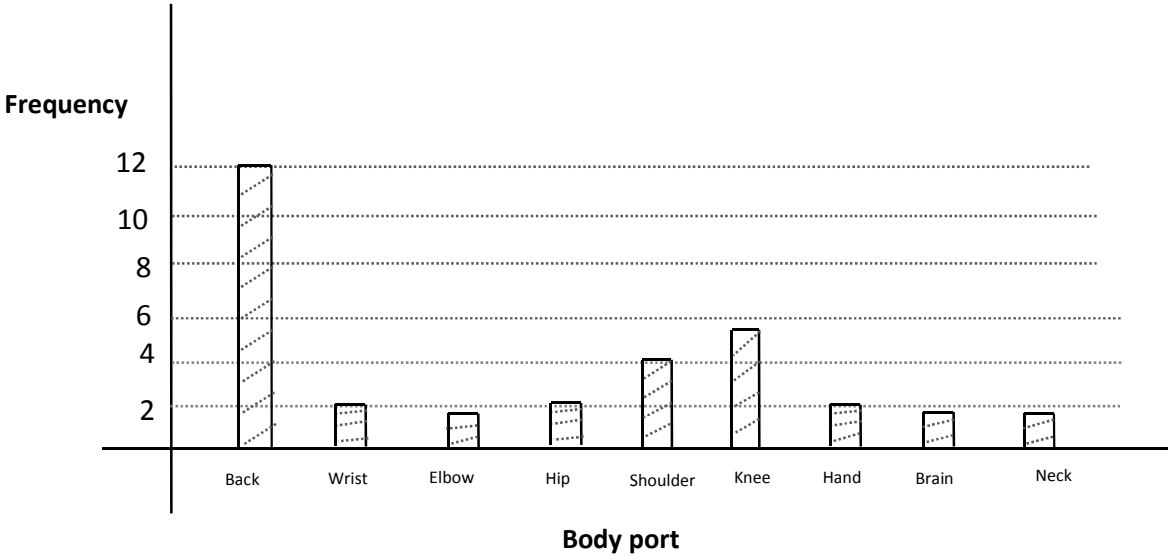
Back	Back	Hand	Neck	Knee	Knee
Wrist	Back	Brain	Shoulder	Shoulder	Back
Elbow	Back	Back	Back	Back	Back
Back	Shoulder	Shoulder	Knee	Knee	Back
Hip	Knee	Hip	Hand	Back	Wrist

1. Construct the frequency distribution of injury location and bar graph.
2. Construct the relative frequency distribution and bar graph.

Solution:

Body Part	count	frequency	relative frequency
Back	𝍸 𝍸 ||	12	12/30 =0.4
Wrist	||	2	2/30 ≈0.067
Elbow	|	1	1/30 ≈ 0.033
Hip	||	2	2/30 ≈0.067

Shoulder	||||	4	4/30 ≈0.133
Knee	|||||	5	5/30 ≈0.167
Hand	||	2	2/30 ≈0.067
Brain			1	1/30 ≈ 0.033
Neck			1	1/30 ≈ 0.033
Total			30	1



Similar graph can be obtained for relative frequency (exercise).

10 Summarizing Data

- Once data are organized into tables or graphs, many characteristics of the distribution can be obtained such as shape, center, and spread.

- The center and spread are numerical summaries of the data. The center of data is commonly called the average or mean.
- **The arithmetic mean** of a variable is the sum of all values of the variable divided by their number.
- **The population mean** μ is computed using all the individuals in a population.

  If $x_1, \dots \dots \dots x_N$ are the N observations of a variable from a population, then:

$$\mu = \frac{x_1 + x_2 + \cdots\cdots\cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- The sample mean $\bar{X}$ is computed using the sample data.

  If $x_1, \cdots \cdots x_n$ are n observations of variable from a sample, then:

$$\bar{X} = \frac{x_1 + x_2 + \cdots \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- The population mean μ is parameter whereas the sample mean $\bar{X}$ is a statistic.
- It is also useful to measure the amount of dispersion in the data, which is degree of spread out.
- **The population variance** of a variable is the sum of the squared deviations about the population mean divided by the number of observation in the population :

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N} = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- **The sample variance** is the sum of squared deviations about the sample mean divided by n-1:

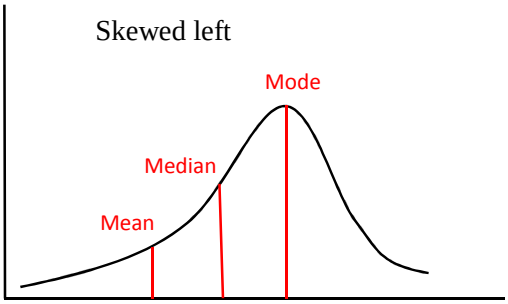
$$S^2 = \frac{\sum (x_i - \bar{X})^2}{n - 1}$$

- We note that S^2 is divided by n-1 because if we divide by n, S^2 would consistently underestimate σ^2 (biased). So we divide by n-1 to remove the bias.
- We call n-1 the **degrees of freedom** because the first n-1 observations have freedom to be whatever value, but the n^{th} value has no freedom as the sum of deviations about the mean always sums to zero:

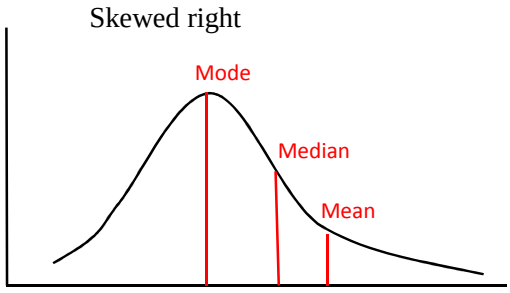
$$\sum_{i=1}^{n}(x_i - \bar{X}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{X} = \sum_{i=1}^{n} x_i - n\bar{X} = \sum x_i - n\frac{\sum x_i}{n} = 0$$

- **The range R** of a variable =largest value –smallest value.
- **The population standard deviation σ** is the square root of σ^2.
- **The sample standard deviation** *S* is the square root of S^2.
- **The median** M of a variable is the value that lies in the middle when data are arranged in ascending order. If the number of observations n is odd, the median is the observation in the $\frac{n+1}{2}$ position. If n is even, the median is the mean of the two observations in the $\frac{n}{2}$ and $\frac{n}{2}+1$ positions. For example, the median of {5, 7, 0, 100, 1} is 5 and the median of {5, 7, 0, 100, 1, 6} is 5.5.
- A numerical summary of data is said to be **resistant** if extreme values relative to the data don't affect its value substantially. The median is resistant while the mean is not.
- **The mode** of a variable is the most frequent observation of the variable that exists in the data set. For example, the mode of {5, 6, 0, 1, 5, 7, 5} is 5. In addition, {1, 7, 5, 6, 0} has no mode science each value occurs only once.
- **Relationship between distribution shape, mean and median:**

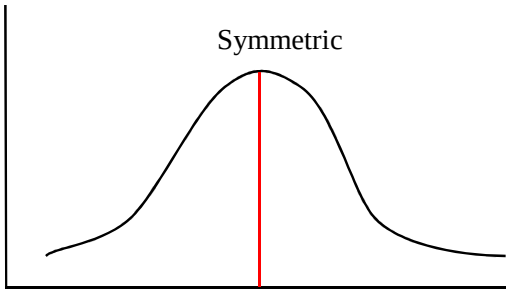
  When there are extreme values, they cause the data distribution to be skewed left or right with long tail and pull the mean in the direction of the tail.



Mean < Median



Mean > Median



Mean = Median

11 Measures of Position and Outliers:

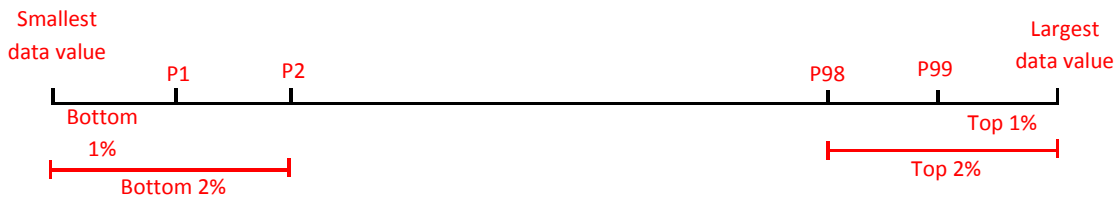
- Measures of position describe the relative position of certain data value within the entire dataset.
- **The Z-score** represents the distance between a data value and the mean in terms of the number of standard deviations.

  Population Z-score: $$Z = \frac{X-\mu}{\sigma}$$

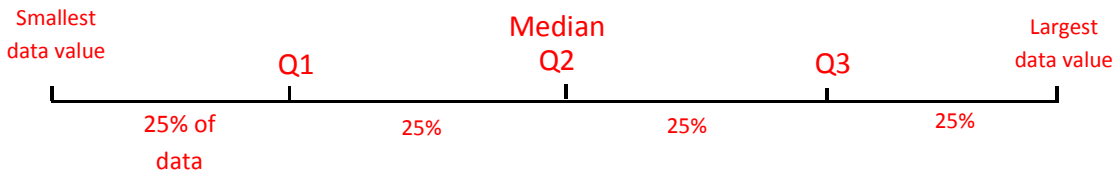
  Sample Z-score: $$Z = \frac{X-\bar{X}}{S}$$

- We note that the Z-score has zero mean and standard deviation one.
- **The K^{th} percentile P_K** of a set of data is a value such that k percent of the observations are less than or equal to the value.

  Note that P_{50} is the median.



- **Quartiles:** divide data into fourths or four equal parts. The first quartile $\boldsymbol{Q_1}$ divides the bottom 25% of the data from the top 75%. Therefore, $\boldsymbol{Q_1}$ is equivalent to $\boldsymbol{P_{25}}$.



- **Steps in finding quartiles:**
  - ✓ Arrange the data in ascending order.
  - ✓ Determine the median M or second quartile $\boldsymbol{Q_2}$.
  - ✓ Divide the data into two halves: below and above $\boldsymbol{Q_2}$. $\boldsymbol{Q_1}$ is the median of the bottom half and $\boldsymbol{Q_3}$ is the median of the top half.

Example 4:

A random sample of 18 collision claims are shown.

Find and interpret the first, second, and third quartiles.

6,751	9,908	3,461	2,336	21,147	2,332
189	1185	370	1414	4668	1953
10,034	735	802	618	180	1657

Solution:

Arrange the data in ascending order:

180 189 370 618 735 802 1185 1414 1657 1953 2332 2336 3461 4668 6751 9908 10034 21147

$\boldsymbol{Q_2}$ is the median which is the mean of data 9 and 10:

$$Q_2 = \frac{1657 + 1953}{2} = 1,805$$

Q_1 is the median of the bottom half =735

Q_3 is the median of the top half =4668

Interpretation: 25% of collision claims are less than or equal to 735, and 75% are less than or equal to 4668. 50% of the claims are less than or equal to 1805.

- **The interquartile range**, denoted IQR is the range of the middle 50% of the observations in the dataset. That is:

$$IQR=Q_3-Q_1$$

Example 5

Determine the inter-quartile range of the data in example 7.2.

Solution:

$$IQR = Q_3 - Q_1 = 4668 - 735 = 3933$$

- Extreme observations in the data set are referred to as outliers.
- Outliers occur because of error in the measurement, during data entry, or from errors in sampling.

- **Checking for outliers using quartiles:**
  - ✓ Determine the first and third quartiles $Q_1 and Q_3$.
  - ✓ Compute the inter-quartile range IQR.
  - ✓ Determine the fences, which serve as cutoff points for determining outliers:
    - **Lower fence** $= Q_1 - 1.5\ IQR$

- **Upper fence** $= Q_3 + 1.5\,\text{IQR}$

✓ If a data value is less than the lower fence, or greater than the upper fence , it is considered as an outliers.

✓

Example 6

Check the data in example 7.2 for outliers.

Solution:

$Q_1 = 735 \quad Q_3 = 4668$

2. $IQR = 3933$

3. $lower\ fence = Q_1 - 1.5\,IQR$

$= 735 - 1.5(3933) = -5164.5$

$upper\ fence = 4668 + 1.5(3993) = 10567.5$

4. There is an outliers above the upper fence =21147.