Lecture 5

Biostatistics

second stage 2022-2023

lecturer

Naba Mohammed Dhiaa ALashqar

AL-Noor University College

- **Descriptive Measures**

B. Measures of Dispersion (Measures of Variation)

The most used measures of Variation are:

1. Range 2. Variance (S^2) 3. Standard Deviation (S) 4. Quartile (IQR). 5. Percentile (IPR).

1. Range: The **range** of a data set is the difference between the maximum and minimum data entries in the set. To find the range, the data must be quantitati

$$\text{Range} = (\text{Maximum data entry}) - (\text{Minimum data entry})$$

a. For raw data. b. For frequency distribution. c. For grouped frequency distribution

Example: find the range for the following data sets?

a. (2,7,11,5,8,4,9,3)

Range = Max- Min = 11-2 = 9.

b.

X	1	2	3	4	5
Frequency	100	2	17	50	750

Range = Max- Min = 5-1 = 4

c. For grouped frequency distribution

Intervals	0-4	5-9	10-14	15-19
Frequency	3	8	7	2
U.R.B	4.5	9.5	14.5	19.5
L.R.B	-0.5	4.5	9.5	14.5

Range = U.R.B (Max value) - L.R.B (Min value)

For U.R.B $\rightarrow \frac{5-4}{2} = \mathbf{0.5}$ we add 0.5 to the upper terms.

For L.R.B $\rightarrow \frac{5-4}{2} = \mathbf{0.5}$ we subtract 0.5 from the lower terms.

Range = Max- Min = 19.5 – (-0.5) = 20.

Example: Consider the three data sets:

Set A: 0,2,6,10,12 Set B: 4,5,6,7,8 Set C: 6,6,6,6,6

Note: All sets in the previous example have the same mean, but the spread of the distribution in set A is greater than B and the spread in set C is zero.

Deviation: Distance of the data entry from the mean

Example: Find the range of data sets A, B, and C.

Note: The range is strongly affected by outliers

2. Variance, and Standard Deviation

Definition: The **deviation** of an entry x in a population is the difference between the entry and the mean.

$$\text{Deviation of } x = x - \mu$$

Example (1): Two corporations each hired 10 graduates. The starting salaries for each graduate are shown. Find the range of the starting salaries for Corporation A.

Starting Salaries for Corporation A (in thousands of dollars)

Salary	41	38	39	45	47	41	44	41	37	42

Starting Salaries for Corporation B (in thousands of dollars)

Salary	40	23	41	50	49	32	41	29	52	58

Deviations of Starting Salaries for Corporation A

Salary (in 1000s of dollars) x	Deviation (in 1000s of dollars) $x - \mu$
41	−0.5
38	−3.5
39	−2.5
45	3.5
47	5.5
41	−0.5
44	2.5
41	−0.5
37	−4.5
42	0.5
$\Sigma x = 415$	$\Sigma(x - \mu) = 0$

The sum of the deviations is 0.

Mean:

Sum of the deviations:

Note: the sum of deviations of any data set is always zero.

To overcome this issue, take the squares of each deviation and find the average.

Definition:

Variance

$$\text{Population variance} = \sigma^2 = \frac{\Sigma(x-\mu)^2}{N}.$$

$$\text{Sample variance} = s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

Standard deviation

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

$$\text{Sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

Symbols in Variance and Standard Deviation Formulas

	Population	Sample
Variance	σ^2	s^2
Standard deviation	σ	s
Mean	μ	$\bar{x}$
Number of entries	N	n
Deviation	$x-\mu$	$x-\bar{x}$
Sum of squares	$\Sigma(x-\mu)^2$	$\Sigma(x-\bar{x})^2$

Finding the Population Variance and Standard Deviation

In Words	In Symbols
1. Find the mean of the population data set.	$\mu = \frac{\Sigma x}{N}$
2. Find the deviation of each entry.	$x-\mu$
3. Square each deviation.	$(x-\mu)^2$
4. Add to get the sum of squares.	$SS_x = \Sigma(x-\mu)^2$
5. Divide by N to get the population variance.	$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$
6. Find the square root of the variance to get the population standard deviation.	$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

Finding the Sample Variance and Standard Deviation

In Words	In Symbols
1. Find the mean of the sample data set.	$\bar{x} = \frac{\Sigma x}{N}$
2. Find the deviation of each entry.	$x-\bar{x}$
3. Square each deviation.	$(x-\bar{x})^2$
4. Add to get the sum of squares.	$SS_x = \Sigma(x-\bar{x})^2$
5. Divide by $n-1$ to get the sample variance.	$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$
6. Find the square root of the variance to get the sample standard deviation.	$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

Note: We can use another formula to find the sample variance.

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

Steps to find the variance/ S.D:

- Find the mean
- Find the deviation of each entry
- Square each deviation
- Add to get the sum
- Divide by N or n-1
- Take the root to find the standard deviation

- Example (2): Find the population variance and standard deviation of the starting salaries for Corporation A listed in Example (1).

Solution:

Sum of Squares of Starting Salaries for Corporation A

Salary x	Deviation $x - \mu$	Squares $(x - \mu)^2$
41	−0.5	0.25
38	−3.5	12.25
39	−2.5	6.25
45	3.5	12.25
47	5.5	30.25
41	−0.5	0.25
44	2.5	6.25
41	−0.5	0.25
37	−4.5	20.25
42	0.5	0.25
$\Sigma x = 415$		$SS_x = 88.5$

For this data set, $N = 10$ and $\Sigma x = 415$. The mean is

$$\mu = \frac{415}{10} = 41.5.$$ Mean

The table at the left summarizes the steps used to find SS_x. Because

$$SS_x = 88.5$$ Sum of squares

you can find the variance and standard deviation as shown.

$$\sigma^2 = \frac{88.5}{10} \approx 8.9$$ Round to one more decimal place than the original data.

$$\sigma = \sqrt{\frac{88.5}{10}} \approx 3.0$$ Round to one more decimal place than the original data.

So, the population variance is about 8.9, and the population standard deviation is about 3.0, or \$3000.

Example (3):

Finding the Sample Variance and Standard Deviation

In a study of high school football players that suffered concussions, researchers placed the players in two groups. Players that recovered from their concussions in 14 days or less were placed in Group 1. Those that took more than 14 days were placed in Group 2. The recovery times (in days) for Group 1 are listed below. Find the sample variance and standard deviation of the recovery times. *(Adapted from The American Journal of Sports Medicine)*

Grope A 4 7 6 7 9 5 8 10 9 8 7 10

Grope B 43 57 18 45 47 33 49 24

Time x	Deviation $x - \bar{x}$	Squares $(x - \bar{x})^2$
4	−3.5	12.25
7	−0.5	0.25
6	−1.5	2.25
7	−0.5	0.25
9	1.5	2.25
5	−2.5	6.25
8	0.5	0.25
10	2.5	6.25
9	1.5	2.25
8	0.5	0.25
7	−0.5	0.25
10	2.5	6.25
$\Sigma x = 90$		$SS_x = 39$

For this data set, $n = 12$ and $\Sigma x = 90$. The mean is $\bar{x} = 90/12 = 7.5$. To calculate s^2 and s, note that $n - 1 = 12 - 1 = 11$.

$SS_x = 39$ Sum of squares (see table at left)

$s^2 = \frac{39}{11} \approx 3.5$ Sample variance (divide SS_x by $n - 1$)

$s = \sqrt{\frac{39}{11}} \approx 1.9$ Sample standard deviation

So, the sample variance is about 3.5, and the sample standard deviation is about 1.9 days.

Definition: **Variance for Grouped Data**

$$s^2 = \frac{\sum X_i^2.f - n\bar{X}^2}{n-1}, \qquad \text{where } \bar{X} = \frac{\sum X_i.f}{n}$$

Example: Calculate the variance and the S.D of the frequency table below.\

x	f	x.f	X^2f
0	10	0	0
1	19	19	19
2	7	14	28
3	7		63
4	2		32
5	1		25
6	4		144
	50	91	311

Mean=

S^2=

Note: When we classes, find the midpoints for each data entry.

Example: Given the following frequency table, find the variance and the S.D.

classes	f	X_m	$X_m.f$	$X_m^2.f$
6-10	1			
11-15	2			
16-20	3			
Total	6		88	1374

Properties of the Variance

1. The larger the variance is the more spread the data is.
2. $S^2 >= 0$
3. If the variance is zero, then all data entries are equal.

Definition: The **coefficient of variation (CV)** of a data set describes the standard deviation as a percent of the mean.

$$\text{Population: } CV = \frac{\sigma}{\mu} \cdot 100\% \qquad \text{Sample: } CV = \frac{s}{\bar{x}} \cdot 100\%$$

Remark: Why do we use the CV?
We use the variance or SD to compare data sets with the same units of measure and close means When the data sets have different units of measure or different means, use the CV.

Example: Find the coefficient of variation for the heights and the weights, then compare the results.

Heights and Weights of a Basketball Team

Heights	72	74	68	76	74	69	72	79	70	69	77	73
Weights	180	168	225	201	189	192	197	162	174	171	185	210

Solution:

The mean height is 72.8 (Check!) and the SD is 3.3 inches (Check!)

CV (height)=

The mean weight is 187.8 lb. and the SD is 17.7 lb. (Check!)

CV (weight)=

Lecture 3 and 4

Biostatistics

second stage 2022-2023

lecturer

Naba Mohammed Dhiaa ALashqar

AL-Noor University College

- **Descriptive Measures**

Measures of Central Tendency and Dispersion are common descriptive measures

for summarizing numerical data.

A. Measures of Central Tendency (mean, median, and mode)

Measures of central tendency are measures of the location of the middle or the center of a distribution. The most frequently used measures of central tendency are the mean, median, and mode.

1. The **mean** of a data set is the sum of the data entries divided by the size of the data set. It is denoted by $\bar{x}$ for the sample and μ (mew) for the population.

$$\text{Population Mean: } \mu = \frac{\Sigma x}{N}$$

$$\text{Sample Mean: } \bar{x} = \frac{\Sigma x}{n}$$

Example (1): The weights (in pounds) for a sample of adults before starting a weight-loss study are listed. What is the mean weight of the adults?

274 235 223 268 290 285 235

Solution: Find the sum and then, divide by the sample size.

Definition: The **mean of a frequency distribution** for a sample is estimated by

$$\bar{x} = \frac{\Sigma xf}{n}$$ Note that $n = \Sigma f$.

where x and f are the midpoint and frequency of each class, respectively.

Finding the Mean of a Frequency Distribution

1. Find the midpoint of each class.	$x = \frac{(\text{Lower limit}) + (\text{Upper limit})}{2}$
2. Find the sum of the products of the midpoints and the frequencies.	Σxf
3. Find the sum of the frequencies.	$n = \Sigma f$
4. Find the mean of the frequency distribution.	$\bar{x} = \frac{\Sigma xf}{n}$

Example (2): Complete the following table and find the mean.

class boundaries	f	X_m	$f.X_m$
5.5-10.5	1	8	
10.5-15.5	2		
15.5-20.5	3		54
20.5-25.5	5	23	115
25.5-30.5	4	28	112
30.5-35.5	3	33	99
35.5-40.5	2	38	76
Total	20		$\sum x_i f = 490$

$$\bar{x} = \frac{\sum xf}{n} = \frac{\qquad\qquad}{} =$$

2. The **median (Q_2)** of a data set is the value that lies in the middle of the data when it is ordered.

- When the data has an odd number of entries, the median is the middle data entry.
- When the data has an even number of entries, the median is the mean of the two middle ones.

Example (3): Find the median of the weights listed in Example (1).

Solution: To find the median weight, first order the data.

223 235 235 268 274 285 290

Because there are seven entries (an odd number), the median is the middle, or fourth, entry. So, the median weight is 268 pounds.

Example (4): The points scored by the winning teams in the Super Bowls for the National Football League's 2001 through 2016 seasons are listed. Find the median.

20 48 32 24 21 29 17 27
31 31 21 34 43 28 24 34

Solution: You have to order the data

- 3. The **mode** of a data set is the entry with the highest frequency.

There are three cases for the mode:

- **One mode;** there is one data entry that has the highest frequency (unimodal)
- **Two modes;** there are two entries with the greatest frequency (bimodal)
- **No mode;** no entry is repeated

Example (5): Find the mode in each of the following.

- {1,3,6,6,6,6,7,7,12,12,17}. ⟶
- {1,3,6,7,12,17} ⟶
- {1,1,3,3,6,6,6,7,7,12,7,12,17} ⟶

Definition: An **outlier** is a data entry that is far removed from the other entries in the data set.

Example (6): Find the mean, median, and mode of the ages. Are there any outliers? Which measure of central tendency best describes a typical entry of this data set?

Ages in a class						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

Mean: Mode:

Median: Outlier:

Interpretation:

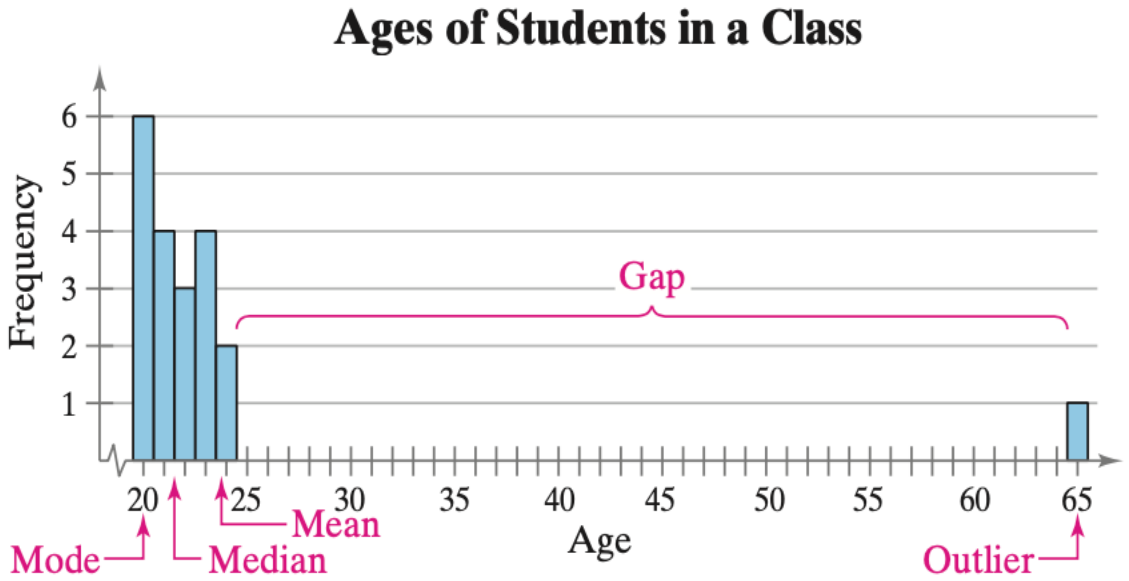
The mean takes every entry into account but is influenced by the outlier of 65.

The median also takes every entry into account, and it is not affected by the outlier.

In this case the mode exists, but it does not appear

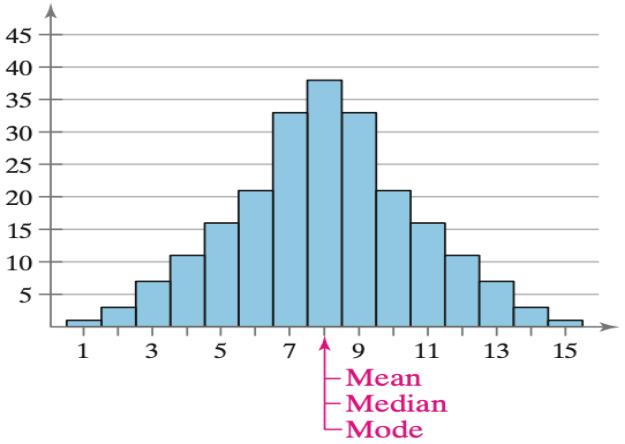
to represent a typical entry. Sometimes a graphical comparison can help you decide which measure of central tendency best represents a data set. The

histogram shows the distribution of the data and the locations of the mean, the median, and the mode. In this case, it appears that the median best describes the data set.

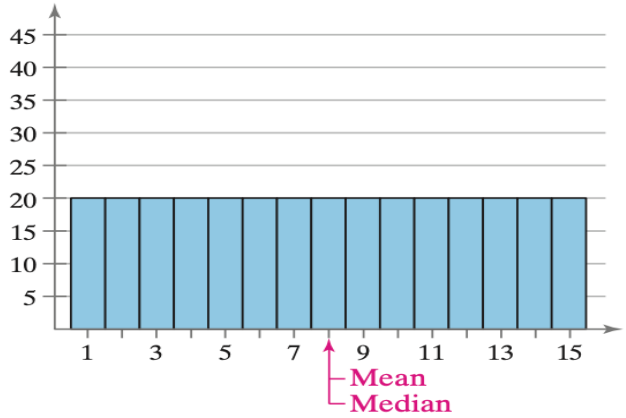


Shapes of Distributions

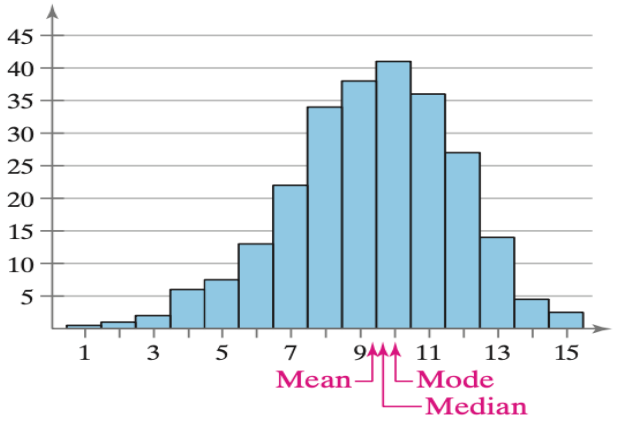
$\mathbf{mean}(\bar{x}) = \mathbf{medin}\,(Q_2) = \mathbf{Mode}$



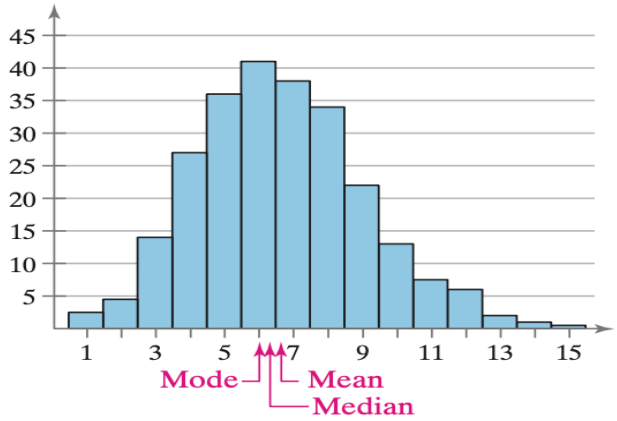
Symmetric Distribution



Uniform Distribution



Skewed Left Distribution



Skewed Right Distribution

$\mathbf{mean}(\bar{x}) < \mathbf{medin}\,(Q_2) < \mathbf{Mode}$

$\mathbf{mean}(\bar{x}) > \mathbf{medin}\,(Q_2) > \mathbf{Mode}$

- The mean will always fall in the direction in which the distribution is skewed. For instance, when a distribution is skewed left, the mean is to the left of the median.

Example (7): Describe the distribution shape if the median medin (Q_2) is 8, mean is 7, and mode is 8

Solution:

Example (8): on an exam given to 5 students, the mean grade is 78, the grades of them are 87, 81, 76 and 53. then the grade of the 5^{th} student is:

A) 65. **B) 93.** **C) 71.** **D) 85.** **E) 99.**

Solution:

Example (9): In a quiz, 3 students got 1, 5 students got 2 and 2 students got 5. the average score of these students in this quiz is:

A) 3.30 **B) 3.00** **C) 2.80** **D) 3.11** **E) 2.30.**

Solution:

Example (10): if the mean of 9 students is 15, a new student joined the class with mark 20, find the new sum, new number of students, and new mean.

Solution:

Example (11): **Consider the following data**

I	3-5	6-8	9-11	12-14
F	5	2	2	1

Then the mode is:

A) 13 **B)** 10 **C)** 4 **D)** 7 **E)** 8

Solution:

Example (12): **consider the following grouped sample data**

I	0-4	5-9	10-14	15-19
F	2	3	5	2

Then the mean is:

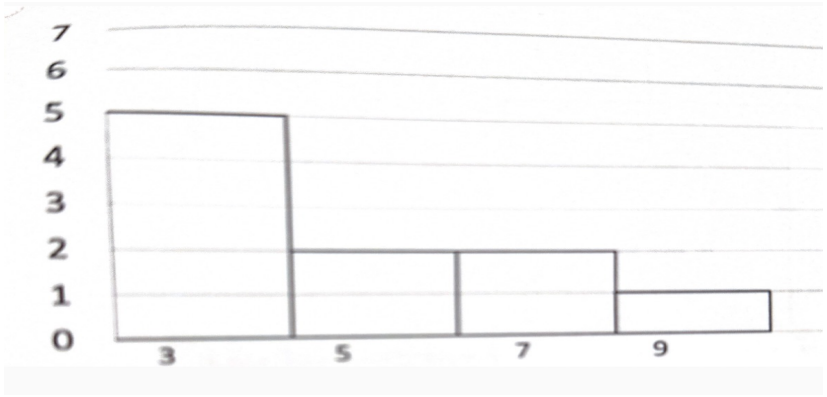
A) 8.1 **B)** 5.7 **C)** 7.2 **D)** 4.9 **E)** 9.92

Solution:

Example (13): **If the mean of 15, X, 2X+3, is 33, then the value of X is:**

A) 18 **B)** 16 **C)** 25 **D)** 22 **E)** 27

Example (14): the following is the histogram of a grouped sample data. The mean equals:

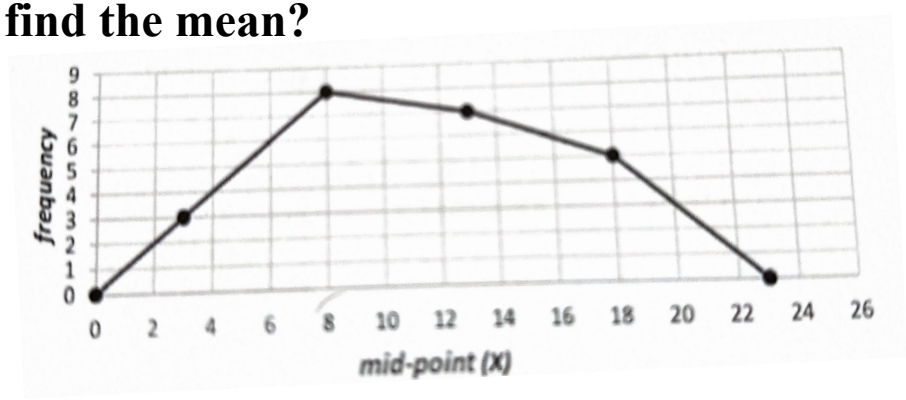


A) 2.2 **B)** 3.4 **C)** 3.2 **D)** 3 **E)** 4.8

Solution:

Example (15): for the following polygon, find the mean?

Solution:



Biostatistics

Second Stage 2023-2024

Lecturer: Naba Mohammed Dhiaa ALashqar

AL-Noor University College

Lecture 1

Chapter 1

Introduction of Statistics

In this chapter, we will talk about some basic concepts that would help you to understand all concepts of this course.

Statistics: is the science of collecting, organizing, and interpreting (analysis) data in order to make decisions.

Biostatistics

- Biostatistics can be defined as the application of the mathematical tools used in statistics to the fields of biological sciences and medicine.
- Biostatistics is a growing field with applications in many areas of biology including epidemiology, medical sciences, health sciences, educational research, and environmental sciences.
- Biostatistics is concerned with the collection, organization, summarization, and analysis of data.
- To draw inferences about data when only a part of the data is observed.

Data: consists of information coming from observations, counts, measurements, or responses.

- Biostatistics is concerned with the interpretation (analysis) of the data and the communication of information about the data.
- The measurements obtained in a research study are called the data.
- The goal of statistics is to help researchers to organize and interpret the data.
- Raw data: is data collected as they receive.
- Organize data is data organized either in ascending, descending or grouped data (frequency distribution table).

Sources of data

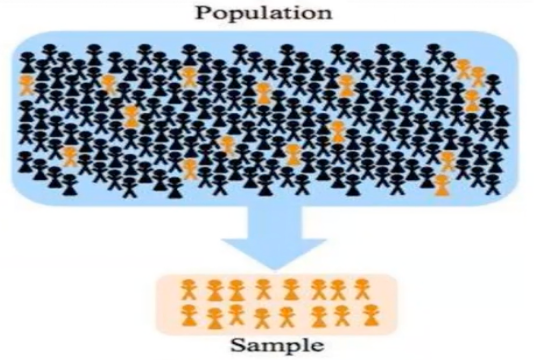
- Data: It's a raw material of statistics for examination.

Data are obtained from;

- Analysis of records
- Surveys
- Counting
- Experiments
- Reports

Data Sets: Populations and samples

- A **population** is the set of all individuals under Study for which we make observations with which we are concerned.
- A **population** is the collection of all outcomes, responses, or measurements.
- The size or number of observations of a population is denoted by N.
- The elements of the population possess common characteristics.
- A **sample** is a subset or part of a population, it is usually randomly taken.
- The size of the sample is denoted by n.



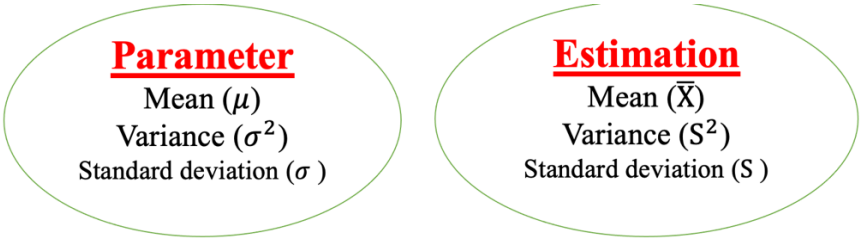
Two important terms are used within this course are parameters and estimators.

A **parameter** is a numerical description of a population characteristic.

هو معامل عددي نحن مهتمين نعرفه عن مجتمع معين، مثلاً متوسط دخل الفرد، نسبة التدخين، نسبة الإصابة بفيروس معين و غيرها

A **statistic** is a numerical description of a sample characteristic.

هو مقدر عددي ويتم حسابه من خلال عينة، ويمثل الشئ الذي استطعنا حسابه بشكل فعلي

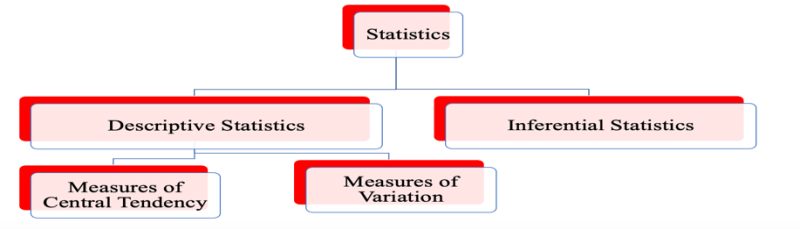


We must note that a sample statistic may differ by changing the sample taken from the population, but a population parameter is constant for the targeted population.

المقدر (Estimator) قد تختلف قيمته بتغير العينة المأخوذة من المجتمع، بينما المعامل (Parameter) يبقى كما هو ثابت لمجتمع الدراسة

Branches of Statistics

The study of statistics has two major branches



1. **Descriptive statistics (الإحصاء الوصفي)** : it involves organizing and summarizing, and displaying data to make them more understandable.

For example, tables or graphs are used to organize data, and descriptive values such as the mean score and the median value are used to summarize data.

- A descriptive value for a population is called a parameter
- While a descriptive value for a sample is called a statistic (Estimation).

2. **Inferential statistics (الإحصاء الاستدلالي)**: using a sample to draw conclusions about a population

- **True or False? In Exercises 1-5, determine whether the statement is true or false. If it is false, rewrite it as a true statement.**

1. A statistic is a numerical description of a population characteristic.
2. A sample is a subset of a population.
3. Inferential statistics involves using a population to draw a conclusion about a corresponding sample.
4. A population is the collection of some outcomes, responses, measurements, or counts that are of interest.
5. A sample statistic will not change from sample to sample.

Types of Data

Data may be classified into two classes:

1. **Qualitative (**Categorical): where we can put our data in groups or categories.

توصف بانها بيانات نوعية وفي هذا النوع يتم تصنيف البيانات الى مجموعات او فئات.

Example:

Gender: (male, female), **Marital Status**: (single, married, widow, divorced), **Color**: (white, red, blue), **Car Type**: Toyota, BMW, and **Blood Type**: A, AB, O.

2. **Quantitative data(Numeric):** where the values of the data are numbers, and we have two types: Discrete and Continuous.

توصف بانها بيانات كمية ويعبر عنها بارقام وتكون اما متصلة او منفصلة

1. **Discrete quantitative data**: If it is countable. توصف انها بيانات نستطيع عدها.

Example: The Number of patients, The number of children in the family, The number of deaths from a certain disease, The number of car accidents on certain roads on different days.

2. **Continuous quantitative data**: If it is measurable. توصف باانها بيانات تحتاج الى جهاز لقياسها

Example: Temperature, Height, Weight, length..., etc.

Example: If we wish to classify the student's major (accounting, economics, management, marketing) then it will be:

A) Qualitative. **B)** Quantitative **C)** Continuous. **D)** None

Solution: Since we are talking about majors, then it is categorical (A).

Example: One of the following is not a type of qualitative data :

A) Eye color

B) The number of emails received by a professor before the midterm exam.

C) Blood type. **D)** Gender

Solution: The answer is **…….**

Example: One of the following is a type of qualitative data:

A) Annual Income. **B)** Weight **C)** number of siblings. **D)** Marital Status

Solution: The answer is **…….**

Simple random sample

- A scientific sampling of the population is necessary to make a valid inference about the population.
- A simple random sample of size n is a sample that is chosen in such a way that every element in the population has an equal chance (probability) of being selected.
- The main difference between population and sample is:

1- the population includes all the units from a set of data.

2- the sample includes a small group of units selected from the population.

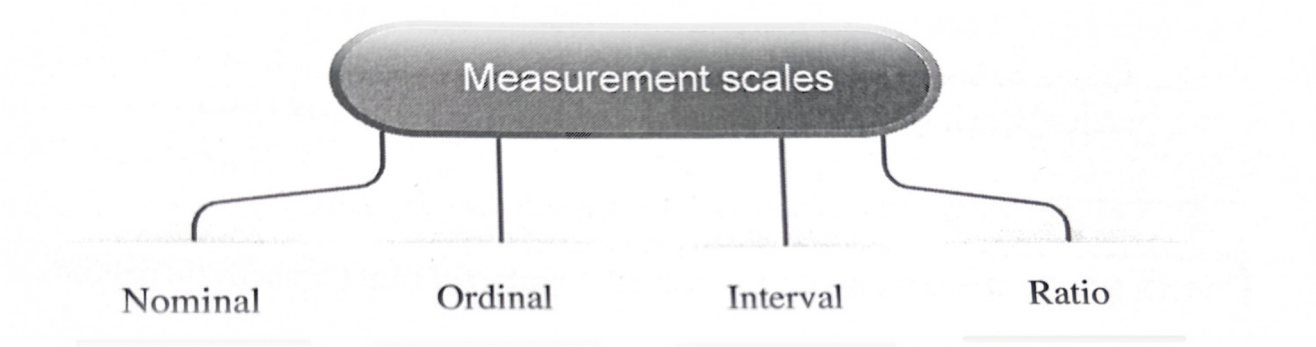
For example, the population of all patients who have eye problems and a sample of patients with glaucoma.

Reasons for sampling:

1. Necessity: sometimes it is simply not possible to study the whole population due to its size or inaccessibility.
2. Practicality: it is easier and more efficient to collect data from a sample.
3. Cost-effectiveness: there are fewer participant, laboratory, and researcher costs involved.

Classifying data by levels of measurement

Measurement scales: The four levels of measurement, in order from lowest to highest, are nominal, ordinal, interval, and ratio.



1. **Nominal Scale:** Classifies data into mutually exclusive categories in which no order or ranking can be imposed on the data, used for qualitative data only.(the data can only be categorized)

Example: 1. (Male-Female). 2. (Well - Sick)

2. **Ordinal Scale:** Classifies data into categories that can be ranked, but differences between data entries are not meaningful. Used for both qualitative and quantitative data. (the data can be categorized and ranked)

Example: 1. Rating scale (Poor, Good, Excellent). 2. Low, Medium, High.

Note: we can't determine if differences between categories are equal or not.

3. **Interval Scale:** can be ordered, and meaningful differences between data entries can be calculated. Also, there is no meaningful zero (0 ≠ nothing). (Interval: the data can be categorized, ranked, and evenly spaced)

Example: the temperature in Fahrenheit, where the difference between 10 and 20 degrees Fahrenheit is exactly the same as the difference between, say, 50 and 60 degrees Fahrenheit.

4. **Ratio Scale:** The ratio scale is exactly the same as the interval scale, with one difference: The ratio scale has what's known as a "true zero." A good example of ratio data is weight in kilograms. If something weighs zero kilograms, it truly weighs nothing compared to temperature (interval data), where a value of zero degrees doesn't mean there is "no temperature," it simply means it's extremely cold! You'll find a full guide to ratio data here. (the data can be categorized, ranked, evenly spaced, and has a natural zero.)

Example: 1. Height (h =0 means no height). 2. Number of cars (0 means there is no cars).

Identify the data type and classify it by the level of measurement.

1. Types of shows (Comedy, Drama, Sports...)
   The nominal level of measurement is most appropriate because the data cannot be ordered

2. The total numbers of branch locations of different banks
   Ratio

3. Average monthly temperatures (Jan: 30.7, Feb: 32.5...)
   Interval
4. Top 3 universities in Iraq
   Ordinal
5. 

The following table summarizes all the information you need about the four levels of measurement scales.

Level of measurement	Put data in categories	Arrange data in order	Subtract data entries	Determine whether one data entry is a multiple of another
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

The Data comes in one of the following forms:

1. **Raw Data** (البيانات العشوائية)

Ex:1,4,7,8,10,12,1,4

2. Frequency Distribution (التوزيع التكراري)
Ex:

X	1	3	4	5	6
F	50	70	300	70	30

3. Grouped Frequency Distribution (التوزيع التكراري بمجموعات)
Ex:

I	0-4	5-9	10-14	15-19	20-24
F	3	6	8	5	4

- **The following data represents the Marital Status for a sample of patients**

Single	**Married**	**widowed**	**divorced**	**single**	**widowed**	**Married**	**single**	**Married**	**single**
Divorced	**Single**	**divorced**	**single**	**Married**	**widowed**	**Married**	**Married**	**Married**	**Married**

Present the data in the Frequency table and construct the relative Frequency

- **Sum=** $\sum_{i=1}^{4} f_i$

$= f_1 + f_2 + f_3 + f_4$

=20

- **relative Frequency =**

$= \frac{f_i}{n} = \frac{f_i}{\sum f_i}$

Marital Status	Frequency (f_i)	Percent (f_i^*)
single	6	0.30
Married	8	0.40
widowed	3	0.15
divorced	3	0.15
Sum	20	1.00

Biostatistics

Second Stage 2023-2024

Lecturer: Naba Mohammed Dhiaa ALashqar

AL-Noor University College

Lecture 2

Chapter 2

Descriptive Statistics

❖ **Definitions:**

- **Frequency Distribution** is a table that shows classes or intervals of data entries with a count of the number of entries in each class.

- The difference between the maximum and minimum data entries is called the **Range**

- The **Relative Frequency** of a class is the percentage of the data that falls into that class

$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Sample size}} = \frac{f}{n} \qquad \text{Note that } n = \Sigma f.$$

- A **Midpoint** of a class is the sum of the lower- and upper-class limits divided by two.

$$\text{Midpoint} = \frac{(\text{Lower class limit}) + (\text{Upper class limit})}{2}$$

- The **Cumulative Frequency** is the sum of the frequency of all previous classes

- To Find the **Class boundaries**. Because the data entries are integers, subtract 0.5 from each lower limit to find the lower-class boundaries and add 0.5 to each upper limit to find the upper-class boundaries.

❖ Example: The following table consists of IQ test scores for randomly chosen 10-year-olds Construct a frequency table with 7 classes including the midpoints, R.f, and C.f.

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	80	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

Max: Min:

Range:

length:

Solution:

Class	Frequency	Class boundaries	midpoint	Relative frequency	Cumulative frequency
80-88	2	79.5-88.5	84	2/60	
89-97	5		93		
98-106	10				
107-115	15				
116-124	12				
125-133	8				
134-142	7				
143-151	1				

❖ Example: The following data represents the marks of 70 students in the final test of the biostatistics subject.

56	65	70	65	55	60	66	70	75	56
60	70	61	67	61	71	67	62	71	60
68	72	57	68	72	69	57	71	69	75
72	62	67	73	58	63	66	73	63	65
58	73	74	76	74	80	81	60	74	58
76	82	77	83	77	85	91	78	94	72
79	64	57	79	55	87	64	88	78	62

Compute the followings:

1- Construct the frequency distribution of marks.

2- Construct the relative frequency distribution

3- What is the proportion of students having marks between 70 and less than 80

4- what is the proportion of students having marks less than 70.

5- what is the proportion of students having mark 80 or more.

Solution:

1- Constructing the frequency distribution

Step 1: Calculate the range R, let X=marks

$$\mathbf{R = X_{Maximum} - X_{Minimum} = X_{Max} - X_{Min}}$$

$X_{Max} = 94$ and $X_{Min} = 55$

$$R = 94 - 55 = 39$$

Step 2: Identify the number of classes C

The number C depends on the researcher, the objective of experiment (research) and the size of the data.

C is usually taken between 5 and 15

i.e $5 \leq C \leq 15$

Here we assume C =8

Step 3: Class length (L) is computed by the formula $\mathbf{L = \frac{R}{C} = \frac{Range}{No.of\ classes}}$

$$\mathbf{L = \frac{39}{8} = 4.875 \cong 5}$$

Where as L is always rounded to the nearest integer (e.g. L=4.5 ≅ 5 , L=4.1 ≅ 5)

Step 4: Identify the classes: the class starts with value named the lower limit and end with value named the upper limit

- the lower limit of the first class takes minimum observation, **i.e lower limit =55**

the upper limit of the first class = lower limit + L = 55 + 5 = 60

Therefore the first class is written as (55-60) and it is read (55 to less than 60)

- The lower limit of the second class= the upper limit of the first class = 60

the upper limit of the second class = lower limit + L = 60 + 5 = 65

Therefore the second class is 6o to less and written as 60 – 65.

Similarly, We write the: limits of the remaining classes, and these are follows

Third class. 65-70

Fourth class 70-75

Fifth class 75-80

Six class 80-85

Seventh class 85-90

eight class 90-95

Construct frequency table

2- Relative frequency distribution

$\boldsymbol{f_i^* = \frac{f_i}{n}}$

for the class $i = 1,\ 2,\ \ldots,8,$ as shown in the third column of table

3-The proportion of students having marks $70 \leq X < 80$

Let P be proportion, $P = 0.229 + 0.143 + 0.372$ of students got

Marks between 70 and less than 80. This means 37.2% (p*100)

of the students got marks between [70,80)

4- $P(X < 70) = \boldsymbol{f_1^* + f_2^* + f_3^*} = 0.143 + 0.171 + 0.186 = 0.5.$

This means 50% have marks less than 70.

5- $P(X \geq 80) = \boldsymbol{f_6^* + f_7^* + f_8^*}$

$= 0.057 + 0.043 + 0.028 = 0.128$

This means 12.8 % have $X \geq 80$.

Marks Classes	No.of students (frequencies) f_i	Relative frequencies f_i^*
55-60	10	0.143
61 – 65	12	0.171
66-70	13	0.186
71-75	16	0.229
76-80	10	0.143
81-85	4	0.057
86-90	3	0.043
91-95	2	0.028
Sum	70	1.00

❖ Example: Homework

Constructing a Frequency Distribution from a Data Set

The data set lists the out-of-pocket prescription medicine expenses (in dollars) for 30 U.S. adults in a recent year. Construct a frequency distribution that has seven classes. *(Adapted from: Health, United States, 2015)*

200	239	155	252	384	165	296	405	303	400
307	241	256	315	330	317	352	266	276	345
238	306	290	271	345	312	293	195	168	342

SOLUTION

Number of classes: The number of classes (7) is stated in the problem.

Min.= Max.=

Width:

Sample size=

Class	Frequency (f)	Midpoint	Relative frequency	Cumulative frequency
	$\sum f = 30$		$\sum \frac{f}{n} = 1$	

Graphical Representation of Data (Data Representation)

Graphical representation of Quantitative data is more cases graphical representation of data is more easy and quick to understand the phenomena under study.

- The most used graphs are Histogram and pie chart

A. <u>Histogram</u> (Quantitative -Continuous data): uses bars to represent the frequency distribution of a data set. A histogram has the following properties

1. The horizontal scale (X-axis) is quantitative and measures the data entries (represent the classes).
2. The vertical scale (Y-axis) measures the frequencies of the classes.
3. Consecutive bars must touch.

Because consecutive bars of a histogram must touch, bars must begin and end at class boundaries instead of class limits. **Class boundaries** are the numbers that separate classes without forming gaps between them. For data that are integers, subtract 0.5 from each lower limit to find the lower class boundaries. To find the upper-class boundaries, add 0.5 to each upper limit. The upper boundary of a class will equal the lower boundary of the next higher class.

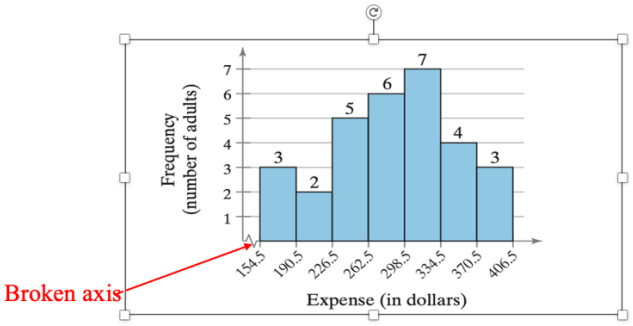
<u>Example:</u> The data set lists the out-of-pocket prescription medicine expenses (in dollars) for 30 U.S. adults in a recent year. *(Adapted from: Health, United States, 2015)*

- Draw a frequency histogram for the frequency distribution

200	239	155	252	384	165	296	405	303	400
307	241	256	315	330	317	352	266	276	345
238	306	290	271	345	312	293	195	168	342

<u>Solution:</u>

Class	Class boundaries	Frequency, f
155–190	154.5–190.5	3
191–226	190.5–226.5	2
227–262	226.5–262.5	5
263–298	262.5–298.5	6
299–334	298.5–334.5	7
335–370	334.5–370.5	4
371–406	370.5–406.5	3



Interpretation histogram, you can see that two-thirds of the adults are paying more than $262.50 for out-of-pocket prescription

B. Frequency Polygon (Quantitative -Continuous data)

Def: A frequency polygon is a graphical form of representation of data. It is used to depict the shape of the data and to depict trends. It is usually drawn with the help of a histogram but can be drawn without it as well

Steps to draw a frequency polygon

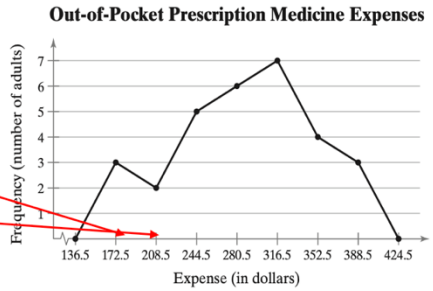
1. Mark the **Midpoint** (class mark) for each class of the horizontal axis.
2. Corresponding to each class mark, plot the frequency as given to you, on the vertical axis.
3. Join the potted points using line segments. The resulting curve is called the frequency polygon.

$$\text{Midpoint} = \frac{(\text{Lower class limit}) + (\text{Upper class limit})}{2}$$

$$= \frac{154.5 + 190.5}{2} = \frac{345}{2} = 172.5$$

$$= \frac{190.5 + 226.5}{2} = \frac{417}{2} = 208.5$$

It is usually, the frequency polygon is closed from both ends (left end and right end) by the classes.



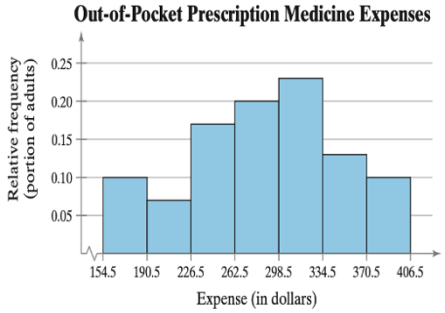
Frequency by the Relative Frequency Histogram

Expenses

Number of adults

Portion of adults

Class	Frequency, f	Midpoint	Relative frequency	Cumulative frequency
155–190	3	172.5	0.1	3
191–226	2	208.5	0.07	5
227–262	5	244.5	0.17	10
263–298	6	280.5	0.2	16
299–334	7	316.5	0.23	23
335–370	4	352.5	0.13	27
371–406	3	388.5	0.1	30
	$\Sigma f = 30$		$\Sigma \frac{f}{n} = 1$	



- Notice that the shape of the histogram is the same as the shape of the frequency histogram on previous page. The only difference is that the vertical scale measures the relative frequencies.
- Interpretation From this graph, you can quickly see that 0.2, or 20%, of the adults, have expenses between $262.50 and $298.50, which is not immediately obvious from the frequency histogram on page 10.

C. Cumulative Frequency Curve

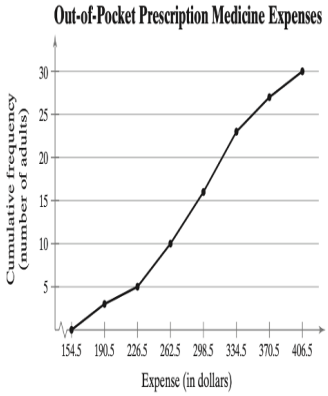
Constructing Cumulative Frequency Graph

1. Construct a frequency distribution that includes **cumulative frequencies** as one of the columns.
2. Specify the horizontal and vertical scales. The horizontal scale consists of upper-class boundaries, and the vertical scale measures cumulative frequencies.
3. Plot points that represent the upper-class boundaries and their corresponding cumulative frequencies.
4. Connect the points in order from left to right with line segments.
5. The graph should start at the lower boundary of the first class (cumulative frequency is 0) and should end at the upper boundary of the last class (cumulative frequency is equal to the sample size).

Example: draw the Frequency curve of the same example

Solution: Using the cumulative frequencies, you can construct a Cumulative Frequency Graph. The upper-class boundaries, frequencies, and cumulative frequencies are shown in the table. Notice that the graph starts at 154.5, where the cumulative frequency is 0, and the graph ends at 406.5, where the cumulative frequency is 30.

Upper class boundary	f	Cumulative frequency
190.5	3	3
226.5	2	5
262.5	5	10
298.5	6	16
334.5	7	23
370.5	4	27
406.5	3	30



D. Pie Chart (Qualitative Data)

Begin by finding the relative frequency, or percent, of each category. Then construct the pie chart using the central angle that corresponds to each category. To find the central angle, multiply 360° by the category's relative frequency.

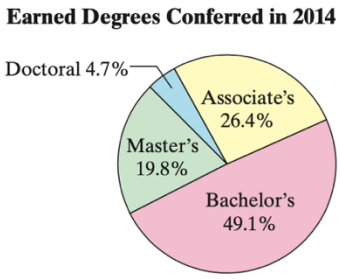
For instance, the central angle for Associate's degrees is $360°(0.264) \approx 95°$.

Earned Degrees Conferred in 2014

Type of degree	Number (in thousands)
Associate's	1003
Bachelor's	1870
Master's	754
Doctoral	178

$$\theta = \frac{f}{\sum f} * 360°$$

Type of degree	f	Relative frequency	Angle
Associate's	1003	0.264	95°
Bachelor's	1870	0.491	177°
Master's	754	0.198	71°
Doctoral	178	0.047	17°



E. Pareto Chart or Par Graph (Qualitative data)

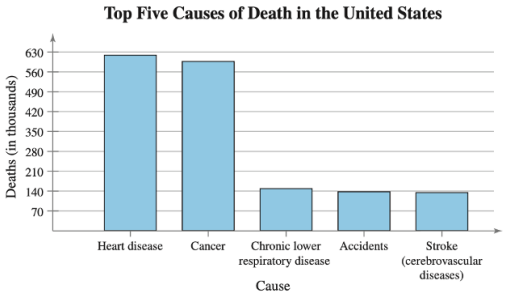
In 2014, these were the leading causes of death in the United States.

Accidents: 136,053
Cancer: 591,699
Chronic lower respiratory disease: 147,101
Heart disease: 614,348
Stroke (cerebrovascular diseases): 133,103

Use a Pareto chart to organize the data. What was the leading cause of death in the United States in 2014? (Source: Health, United States, 2015, Table 19)

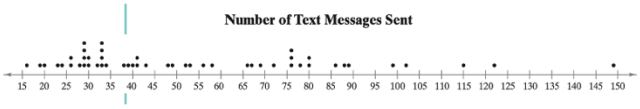
Solution:

Interpretation From the Pareto chart, you can see that the leading cause of death in the United States in 2014 was from heart disease.
Also, heart disease and cancer caused more deaths than the other three causes.



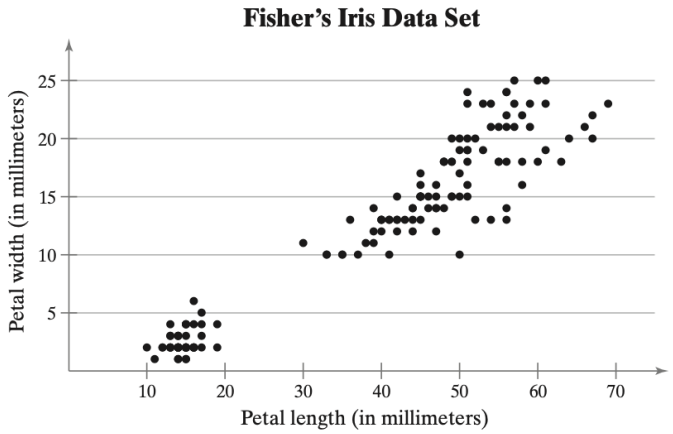
F. Dot Plot (Quantitative) :

Number of Text Messages Sent									
76	49	102	58	88	122	76	89	67	80
66	80	78	69	56	76	115	99	72	19
41	86	48	52	28	26	29	33	26	20
33	24	43	16	39	29	32	29	29	40
23	33	30	41	33	38	34	53	30	149



Interpretation From the dot plot, you can see that most entries occur between 20 and 80 and only 4 people sent more than 100 text messages. You can also see that 149 is an unusual data entry.

G. Scatter Plot (for paired data sets) :



Lecture 6

Biostatistics

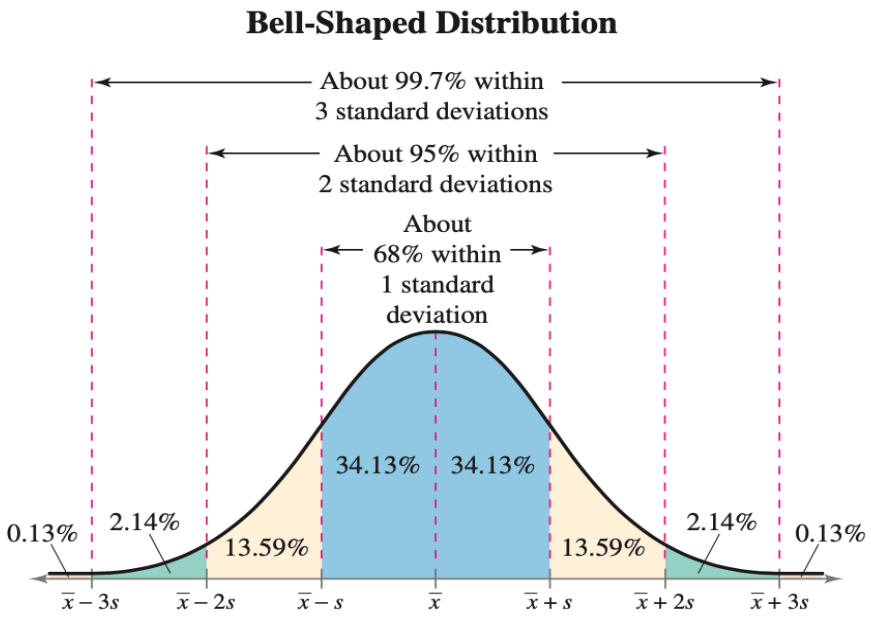
lecturer: Naba Mohammed Dhiaa ALashqar

Empirical Rule

Notes:

- Data that lie more than 2sd from the mean are considered unusual.
- Data that lie more than 3sd from the mean are very unusual.

The above-mentioned data points have a great influence on the SD than the ones closer to the mean



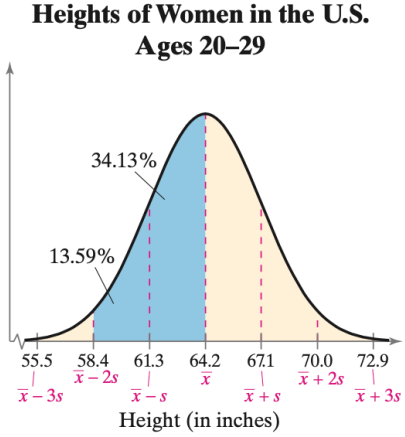
Example: The mean value of homes on a street is 125 thousand with a standard deviation of 5 thousand. The data set has a bell-shaped distribution.
Estimate the percentage of homes between120 and130 thousand.

Sol:

Example: 68% of the marks in a test are between 51 and 64. Assuming the data is bell-shaped distribution, What are the mean and SD?

Sol:

Example: In a survey conducted by the National Center for Health Statistics, the sample **mean** height of women in the United States (ages 20–29) was **64.2** inches, with a sample **standard deviation of 2.9** inches. Estimate the percent of women whose heights are between 58.4 inches and 64.2 inches. (Adapted from National Center for Health Statistics). **Using the Empirical Rule**



Sol:

$$\bar{x} - 2s = 64.2 - 2(2.9) = 58.4.$$

Because 58.4 is 2 standard deviations below the mean height, the percent of

the heights between 58.4 and 64.2 inches is about

$$13.59\% + 34.13\% = 47.72\%.$$

So, about 47.72% of women are between 58.4 and 64.2 inches tall.

***Example:** You are applying for jobs at two companies. Company A salaries with μ= \$30,000 and σ = \$4,000. Company B offers starting salaries with μ = \$30,000 and σ= \$2,000. From which company are you more likely to get an offer of \$36,000 or more?

SOL:

Example: You are applying for jobs at two companies. Company C offers starting salaries with μ = \$75,000 and S= \$2,500. Company D offers starting salaries with μ = \$75,000 and S= \$5,000. From which company are you more likely to get an offer of \$85,000 or more? H.W

Chebyshev's Theorem

When the shape of the distribution is not known, use Chebyshev's Theorem.

1. **(At least / within / minimum / between)** $1 - \frac{1}{K^2}$ **,** (At least = round up)
2. **(At most / outside / maximum).** $\frac{1}{K^2}$**.** **(**At most = round down**) , where K>1.**

The portion of any data set lying within k standard deviations $(k > 1)$ of the mean is at least

$$1 - \frac{1}{k^2}.$$

- $k = 2$: In any data set, at least $1 - \frac{1}{2^2} = \frac{3}{4}$, or 75%, of the data lie within 2 standard deviations of the mean.
- $k = 3$: In any data set, at least $1 - \frac{1}{3^2} = \frac{8}{9}$, or about 88.9%, of the data lie within 3 standard deviations of the mean.

Example: You are conducting a survey on the number of people per house in your region. From a sample with n = 60, the mean number of people per house is 3 and the standard deviation is 1 person. Using Chebyshev's Theorem, determine **at least** how many of the households have 0 to 6 people.

Sol: $\mu = 3$, $S = 1$

We need to determine at least how many of the households have 0 to 6 people.

$\overline{X} - K * S = 0.$ OR $\overline{X} + K * S = 6$

$3 - K * 1 = 0$ $3 + K * 1 = 6$

$K = 3$ $K = 3$

At least $\mathbf{1 - \frac{1}{K^2} = 1 - \frac{1}{(3)^2} = 1 - \frac{1}{9} = 0.888}$

$\mathbf{0.888 * 100 = 88.8\ \%}$

The corresponding number of households is the percentage multiplied by sample size $\mathbf{88.8\ \% * 60 = 53.28 = 53}$ (At least = round up)

At least 53 of the households have 0 to 6 people.

Example: Consider the following data $\bar{X} = 50$, and $S = 5$, Find the interval that contains at least 75% of the data?

SOL: At least $\mathbf{1 - \frac{1}{K^2} = 0.75}$ → $\mathbf{K^2 = 4}$ → **K=2**

Interval $\mathbf{(\bar{X} - K * S, \bar{X} + K * S)}$ → $\mathbf{(50 - 5 * 2, 50 - 5 * 2)} =$ **(40,60)**

Example: The grades of 200 students have mean 40 & standard deviation 6, then number of students whose grade **outside** the interval (31, 49) is:

A) At most 88 C) At least 111

B) At most 89 D) At most 111

SOL:

$\mathbf{\bar{X} - K * S = 31}$ **OR** $\mathbf{\bar{X} + K * S = 49}$

$\mathbf{40 - K * 6 = 31}$ $\mathbf{40 + K * 6 = 49}$

$\mathbf{K = 1.5}$ $\mathbf{K = 1.5}$

Outside → $\mathbf{\frac{1}{K^2} = \frac{1}{1.5^2} = 0.444}$

Number of students $\mathbf{= 0.444 * 200 + 88.8.}$ (Outside= round down=88) → **A.**

Example: The mean & standard deviation of a sample of size 100 are 12 and 1 ,the smallest possible number of observations that are between 10 & 14 is:

A)65 B) 84 C) 75 D) 80

SOL: **(H.W)**

Measures of Position (Quartiles, Percentiles). (مقاييس الموقع (الربعية، النسب المئوية

1. Quartiles

Definition: The three quartiles Q1, Q2, and Q3 divide an ordered data set into four equal parts.

- 25% of the data lies below Q1
- 50% of the data lies below Q2 (the second quartile is the same as the median of the data set).
- 75% of the data lies below Q3

To compute Q1, Q2, and Q3 data should be arranged in ascending order

Q1 has order $\frac{n}{4}$, Q2 has order $\frac{n}{2}$, Q3 has order $\frac{3}{4}n$

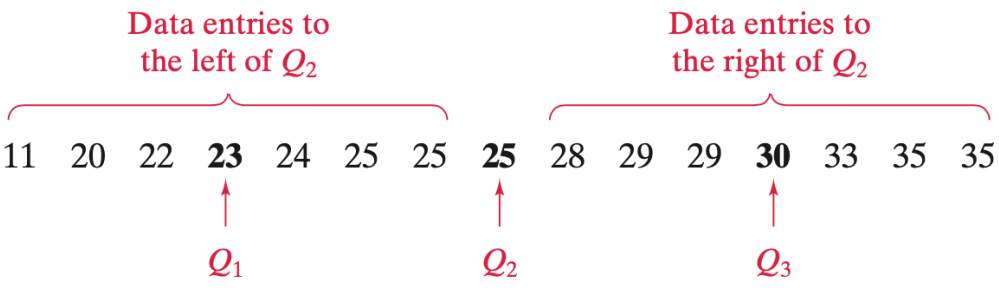
1-For Row data we calculate a by the following steps:

1. Arrange data in ascending order
2. Find. the median Q2 of the data
3. Q2 divides the data into two equal parts. find the median of the first part Q1 (the part before Q2)
4. Find the median of the second part Q3 (The part after Q2)

Example: Find Q1, Q2, and Q3 of the given data sets.

- **20,30,29,22,25,29,25,24,35,23,25,11,33,28,35**

Solution: First, order the data set and find the median Q2. The first quartile Q1 is the median of the data entries to the left of Q2. The third quartile Q3 is the median of the data entries to the right of Q2.



Example: Find Q1, Q2, and Q3 of the given data sets.

- **1,3,6,7,8,9,12,13,13,15**

Solution: n=10. (even number)

$$\mathbf{Q2} = \frac{\mathbf{X}_{(\frac{n}{2})} + \mathbf{X}_{(\frac{n}{2}+1)}}{2} = \frac{\mathbf{X}_{(5)} + \mathbf{X}_{(6)}}{2}$$

$\mathbf{X}_{(5)} = 8$, $\mathbf{X}_{(6)} = 9$.

$= \frac{8+9}{2} = \frac{17}{2} = 8.5$

The location of the median Q2 is 6^{th}

1,3,6,7,8,8.5,9,12,13,13,15

Q2=9

Left: 1,3,6,7,8 → Q1=6. $(\frac{n}{4})$

Right: 9,12,13,13,15 → Q3=13. $(\frac{3}{4}n)$

Definition: The **interquartile range (IQR)** is the range of the middle portion of the data.

IQR= Q3-Q1

Example: Find the interquartile range for the following data.

(2, 2, 4, 5, 7, 7, 8, 8, 11). (H.W)

2-For frequency distribution

Example: Find the interquartile range for the following data.

X	1	2	3	4	5
Frequency	3	7	8	2	5
C.F	3	10	18	20	25
Intervals	0-3	4-10	11-18	19-20	21-25

Q1= $\frac{n}{4} = \frac{25}{4}$ = 6.25 (fraction) → 7th value. ضمن الفترة الثانية

Q1= 2

Q3= $\frac{3\ n}{4} = \frac{3*\ 25}{4} = \frac{75}{4}$ = 18.7 (fraction) → 19th value ضمن الفترة الرابعة

Q3= 4

→IQR= **Q3 - Q1 = 4-2=2**

3- For grouped frequency distribution

Example: Find the inter-quartile-range for the following data.

Intervals	**0-4**	**5-9**	**10-14**	**15-19**
Frequency	**3**	**2**	**7**	**8**
C.F	**3**	**5**	**12**	**20**
U.R.B	**4.5**	**9.5**	**14.5**	**19.5**

Q1= $\frac{n}{4} = \frac{20}{4}$ = 5. (we don't care if its whole number) → 5th value. C.F موجوده بشكل مباشر بصف

→ **Q1**= 9.5

Q3= $\frac{3\ n}{4} = \frac{3*\ 20}{4} = \frac{60}{4}$ = 15 → 15th value. لايوجد بشكل مباشر لهذا السبب سوف نستخدم

(inter potation method)

C.F	**3**	**5**	**12**	**15**	**20**
U.R.B	**4.5**	**9.5**	**14.5**	**Q3**	**19.5**

12	15	20

$$\frac{15\ -\ 12}{20\ -\ 12} = \frac{Q3\ -\ 14.5}{19.5\ -\ 14.5} \rightarrow \mathbf{Q3 = 16.3}$$

14.5	**Q3**	**19.5**

→ IQR= **Q3 – Q1 = 16.3 – 9.5 = 6.8**

2. Percentiles

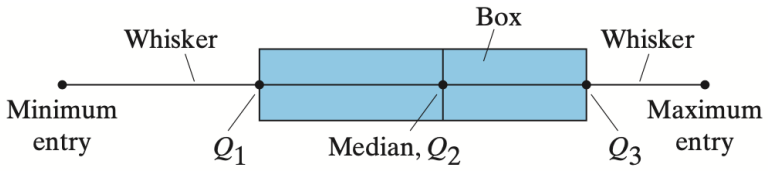
Percentiles P1, P2, ..., and P99 divide the ordered data into 100 equal parts.

- 1% of the data is below P1
- 2% of the data is below P2 •:
- 99% of the data is below P99

Note: P25 = Q1, P50 = Q2 = median, P75 = Q3

The five-number **summary** of a data set are:

- The minimum
- The first quartile Q1
- The Median (Q2)
- The third quartile Q3
- The maximum



(Box-plot)

Lecture 6+7

Biostatistics

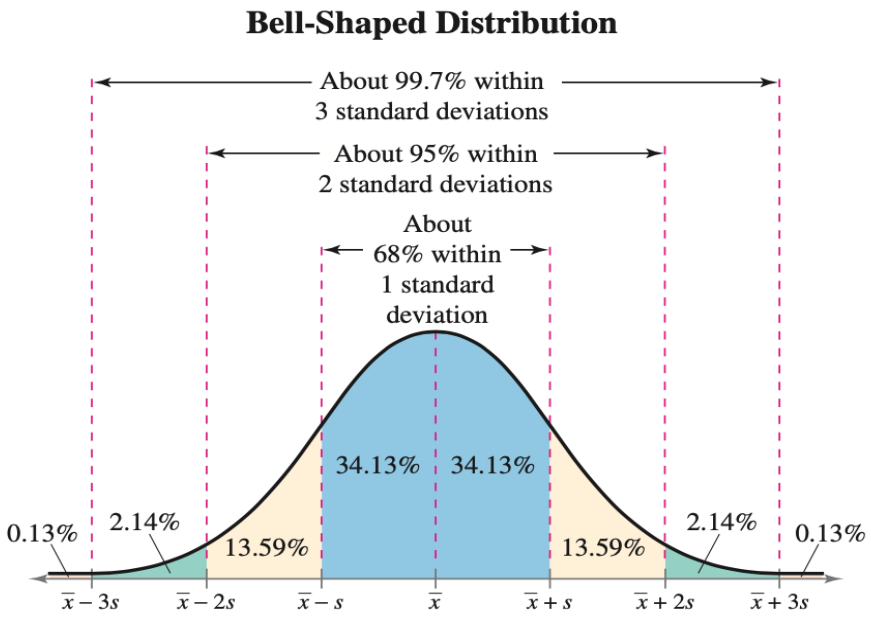
lecturer: Naba Mohammed Dhiaa ALashqar

Empirical Rule

Notes:

- Data that lie more than 2sd from the mean are considered unusual.
- Data that lie more than 3sd from the mean are very unusual.

The above-mentioned data points have a great influence on the SD than the ones closer to the mean



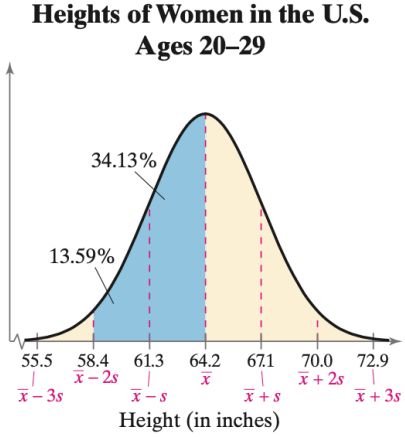
Example: The mean value of homes on a street is 125 thousand with a standard deviation of 5 thousand. The data set has a bell-shaped distribution.
Estimate the percentage of homes between120 and130 thousand.

Sol:

Example: 68% of the marks in a test are between 51 and 64. Assuming the data is bell-shaped distribution, What are the mean and SD?

Sol:

Example: In a survey conducted by the National Center for Health Statistics, the sample **mean** height of women in the United States (ages 20–29) was **64.2** inches, with a sample **standard deviation of 2.9** inches. Estimate the percent of women whose heights are between 58.4 inches and 64.2 inches. (Adapted from National Center for Health Statistics). **Using the Empirical Rule**



Sol:

$$\bar{x} - 2s = 64.2 - 2(2.9) = 58.4.$$

Because 58.4 is 2 standard deviations below the mean height, the percent of

the heights between 58.4 and 64.2 inches is about

$$13.59\% + 34.13\% = 47.72\%.$$

So, about 47.72% of women are between 58.4 and 64.2 inches tall.

***Example:** You are applying for jobs at two companies. Company A salaries with μ= \$30,000 and σ = \$4,000. Company B offers starting salaries with μ = \$30,000 and σ= \$2,000. From which company are you more likely to get an offer of \$36,000 or more?

SOL:

Example: You are applying for jobs at two companies. Company C offers starting salaries with μ = \$75,000 and S= \$2,500. Company D offers starting salaries with μ = \$75,000 and S= \$5,000. From which company are you more likely to get an offer of \$85,000 or more? H.W

Chebyshev's Theorem

When the shape of the distribution is not known, use Chebyshev's Theorem.

1. **(At least / within / minimum / between)** $\mathbf{1 - \frac{1}{K^2}}$ **,** (At least = round up)
2. **(At most / outside / maximum).** $\mathbf{\frac{1}{K^2}}$**.** **(**At most = round down**) , where K>1.**

The portion of any data set lying within k standard deviations $(k > 1)$ of the mean is at least

$$1 - \frac{1}{k^2}.$$

- $k = 2$: In any data set, at least $1 - \frac{1}{2^2} = \frac{3}{4}$, or 75%, of the data lie within 2 standard deviations of the mean.
- $k = 3$: In any data set, at least $1 - \frac{1}{3^2} = \frac{8}{9}$, or about 88.9%, of the data lie within 3 standard deviations of the mean.

Example: You are conducting a survey on the number of people per house in your region. From a sample with n = 60, the mean number of people per house is 3 and the standard deviation is 1 person. Using Chebyshev's Theorem, determine **at least** how many of the households have 0 to 6 people.

Sol: $\mu = 3$, $S = 1$

We need to determine at least how many of the households have 0 to 6 people.

$\overline{X} - K * S = 0.$ OR $\overline{X} + K * S = 6$

$3 - K * 1 = 0$ $3 + K * 1 = 6$

$K = 3$ $K = 3$

At least $\mathbf{1 - \frac{1}{K^2} = 1 - \frac{1}{(3)^2} = 1 - \frac{1}{9} = 0.888}$

$\mathbf{0.888 * 100 = 88.8\ \%}$

The corresponding number of households is the percentage multiplied by sample size $\mathbf{88.8\ \% * 60 = 53.28 = 53}$ (At least = round up)

At least 53 of the households have 0 to 6 people.

Example: Consider the following data $\bar{X} = 50$, and $S = 5$, Find the interval that contains at least 75% of the data?

SOL: At least $\mathbf{1 - \frac{1}{K^2} = 0.75}$ → $\mathbf{K^2 = 4}$ → **K=2**

Interval $\mathbf{(\bar{X} - K * S, \bar{X} + K * S)}$ → $\mathbf{(50 - 5 * 2, 50 - 5 * 2)} =$ **(40,60)**

Example: The grades of 200 students have mean 40 & standard deviation 6, then number of students whose grade **outside** the interval (31, 49) is:

A) At most 88 C) At least 111

B) At most 89 D) At most 111

SOL:

$\mathbf{\bar{X} - K * S = 31}$ **OR** $\mathbf{\bar{X} + K * S = 49}$

$\mathbf{40 - K * 6 = 31}$ $\mathbf{40 + K * 6 = 49}$

$\mathbf{K = 1.5}$ $\mathbf{K = 1.5}$

Outside → $\mathbf{\frac{1}{K^2} = \frac{1}{1.5^2} = 0.444}$

Number of students $\mathbf{= 0.444 * 200 + 88.8.}$ (Outside= round down=88) → **A.**

Example: The mean & standard deviation of a sample of size 100 are 12 and 1 ,the smallest possible number of observations that are between 10 & 14 is:

A)65 B) 84 C) 75 D) 80

SOL: **(H.W)**

Measures of Position (Quartiles, Percentiles). (مقاييس الموقع (الربعية، النسب المئوية

1. Quartiles

Definition: The three quartiles Q1, Q2, and Q3 divide an ordered data set into four equal parts.

- 25% of the data lies below Q1
- 50% of the data lies below Q2 (the second quartile is the same as the median of the data set).
- 75% of the data lies below Q3

To compute Q1, Q2, and Q3 data should be arranged in ascending order

Q1 has order $\frac{n}{4}$, Q2 has order $\frac{n}{2}$, Q3 has order $\frac{3}{4}n$

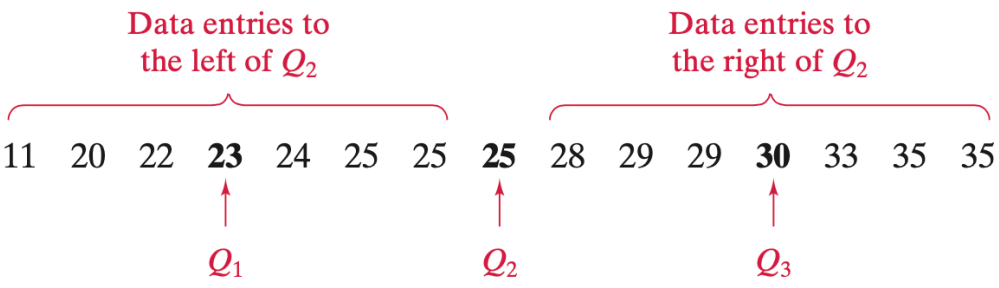
1-For Row data we calculate a by the following steps:

1. Arrange data in ascending order
2. Find. the median Q2 of the data
3. Q2 divides the data into two equal parts. find the median of the first part Q1 (the part before Q2)
4. Find the median of the second part Q3 (The part after Q2)

Example: Find Q1, Q2, and Q3 of the given data sets.

- **20,30,29,22,25,29,25,24,35,23,25,11,33,28,35**

Solution: First, order the data set and find the median Q2. The first quartile Q1 is the median of the data entries to the left of Q2. The third quartile Q3 is the median of the data entries to the right of Q2.



Example: Find Q1, Q2, and Q3 of the given data sets.

- **1,3,6,7,8,9,12,13,13,15**

Solution: n=10. (even number)

$$\mathbf{Q2} = \frac{\mathbf{X}_{(\frac{n}{2})} + \mathbf{X}_{(\frac{n}{2}+1)}}{2} = \frac{\mathbf{X}_{(5)} + \mathbf{X}_{(6)}}{2}$$

$$\mathbf{X}_{(5)} = 8 \ , \qquad \mathbf{X}_{(6)} = 9.$$

$$= \frac{8+9}{2} = \frac{17}{2} = 8.5$$

The location of the median Q2 is 6^{th}

1,3,6,7,8,8.5,9,12,13,13,15

Q2=9

Left: 1,3,6,7,8 → Q1=6. $(\frac{n}{4})$

Right: 9,12,13,13,15 → Q3=13. $(\frac{3}{4}n)$

Definition: The **interquartile range (IQR)** is the range of the middle portion of the data.

IQR= Q3-Q1

Example: Find the interquartile range for the following data.

(2, 2, 4, 5, 7, 7, 8, 8, 11). (H.W)

2-For frequency distribution

Example: Find the interquartile range for the following data.

X	1	2	3	4	5
Frequency	3	7	8	2	5
C.F	3	10	18	20	25
Intervals	0-3	4-10	11-18	19-20	21-25

Q1= $\frac{n}{4} = \frac{25}{4}$ = 6.25 (fraction) → 7th value. ضمن الفترة الثانية

Q1= 2

Q3= $\frac{3\ n}{4} = \frac{3*\ 25}{4} = \frac{75}{4}$ = 18.7 (fraction) → 19th value ضمن الفترة الرابعة

Q3= 4

→IQR= **Q3 - Q1 = 4-2=2**

3- For grouped frequency distribution

Example: Find the inter-quartile-range for the following data.

Intervals	**0-4**	**5-9**	**10-14**	**15-19**
Frequency	**3**	**2**	**7**	**8**
C.F	**3**	**5**	**12**	**20**
U.R.B	**4.5**	**9.5**	**14.5**	**19.5**

Q1= $\frac{n}{4} = \frac{20}{4}$ = 5. (we don't care if its whole number) → 5th value. C.F موجوده بشكل مباشر بصف

→ **Q1**= 9.5

Q3= $\frac{3\ n}{4} = \frac{3*\ 20}{4} = \frac{60}{4}$ = 15 → 15th value. لايوجد بشكل مباشر لهذا السبب سوف نستخدم

(inter potation method)

C.F	**3**	**5**	**12**	**15**	**20**
U.R.B	**4.5**	**9.5**	**14.5**	**Q3**	**19.5**

12	15	20

$$\frac{15\ -\ 12}{20\ -\ 12} = \frac{Q3\ -\ 14.5}{19.5\ -\ 14.5} \rightarrow \mathbf{Q3 = 16.3}$$

14.5	**Q3**	**19.5**

→ IQR= **Q3 – Q1 = 16.3 – 9.5 = 6.8**

The mid-quartile:

$$\text{mid-quartile} = \frac{Q3 - Q1}{2}.$$

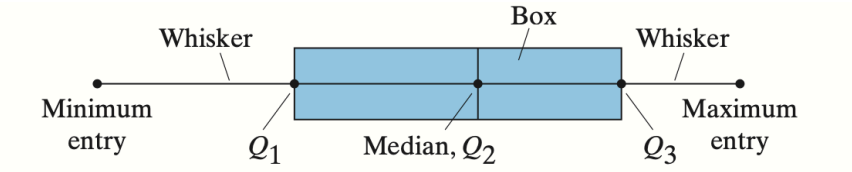
Inter-percentile-range(IPR) المدى النسبى

هي قيمه تحصر البيانات من نسبه الى أخرى، حيث كنا في حساب كنا نحصر البيانات من ، بينما هو الحالة العامه حيث سنحصر البيانات من نسبة الى أخرى

علينا ان نتعلم أولا كيفيه إيجاد النسب وبعدها سنناقش موضوع ال

The five-number summary of a data set are: (Box-plot)

- The minimum
- The first quartile Q1
- The Median (Q2)
- The third quartile Q3
- The maximum



(Box-plot)

Using the Interquartile Range to Identify Outliers

1. Find the first Q1 and third Q3 quartiles of the data set.
2. Find the interquartile range: IQR = Q3 - Q1.
3. Multiply IQR by 1.5: 1.5*(IQR).
4. Subtract 1.5(IQR2 from Q1. Any data entry **less than Q1 - 1.5*(IQR)** is an outlier.
5. Add 1.5(IQR) to Q3. Any data entry **greater than(more than) Q3 + 1.5*(IQR)** is an outlier.

Example: Find the interquartile range of the data set from (example P5) 11,20,22,23,24,25,25,25,28,29,29,30,33,35,35 Are there any outliers?

from (example P5) you know that Q1 = 23 and Q3= 30. So, the interquartile range is

IQR= Q3- Q1=30-23=7

To identify any outliers, first, note that 1.5(IQR) = 1.5(7) = 10.5. There is a data entry, 11, that is less than

$$Q_1 - 1.5(\text{IQR}) = 23 - 10.5 = 12.5$$

Subtract 1.5(IQR) from Q_1.

A data entry less than 12.5 is an outlier.

but there are no data entries greater than

$$Q_3 + 1.5(\text{IQR}) = 30 + 10.5 = 40.5.$$

Add 1.5(IQR) from Q_3.

A data entry greater than 40.5 is an outlier.

So, 11 is an outlier.

Example: the following is the frequency table of a sample data. The first quartile Q1=7 and the third quartile Q3=17, the number of outliers in this sample is:

X	F
3	2
7	6
17	11
18	2
33	3

A) 3. **B)2.** **C)0.** **D)5.** **E)4.**

Solution: IQR=Q3-Q1= 17-7=10

Less than Q1 - 1.5*IQR. → 7 - 1.5 (10) = -8

More than Q3 - 1.5*IQR. → 17 - 1.5 (10) = 32

33 is an outlier, and since it was repeated 3 times, then the number of outliers is 3 → A.

Example: A set of data has the following five number summary.

Minimum	First quartile	median	Third quartile	maximum
17	37	40	49	90

Which of the following contains all the outliers in the distribution

A) 83,85,90,95. B) 17, 81, 80,85, 90. C)64, 80, 85. D)2, 3, 85, 90.

E) 0, 80, 84, 89. (H.W)

Example: **Draw a box plot for the following data:**

(2, 7, 5, 11, 3, 8, 6, 10, 10)

Sol: we must arrange them first: (2, 3, 5, 6, 7, 8, 10, 10, 11)

Q1= $\frac{n}{4} = \frac{9}{4} = 2.25$ **(fraction)** $\rightarrow$ **3rd value** $\rightarrow$ **Q1= 5**

Q2 = $\frac{n}{2} = \frac{9}{2} = 4.5$ **(fraction)** $\rightarrow$ **5rd value Q2 =7**

Q3 = $\frac{3n}{4} = \frac{3*9}{4} = 6.75$ **(fraction)** $\rightarrow$ **7rd value Q3 =10**

Max = 11. **Min = 2.** **(We don't have outliers) check!**

2. Percentiles $(P_K)^{th}$

Percentiles P_1, P_2, ..., and P_{99} divide the ordered data into 100 equal parts.

- 1% of the data is below P_1
- 2% of the data is below P_2
- 99% of the data is below P_{99}

Note: P_{25} **= Q1,** P_{50}**= Q2 = median,** P_{75} **= Q3.**

هنا سنجد قيمة محددة ينحصر تحتها نسبة معينه من البيانات

حتى نوجد P_K **نستخدم القانون التالي:**

$$P_K = \frac{K}{100} * n \begin{cases} \text{Fraction (كسر)} \\ \text{whole number (صحيح عدد)} \end{cases}$$

→Fraction: we take the next integer **Ex: 9.5 → 10**

→ whole number: we take $\frac{K^{th} + (K+1)^{th}}{2}$ **Ex: 10 →** $\frac{10 + 11}{2}$

1. For raw data:

→ we must order the data firstly

Example: Find the 65% for the following data (or P_{65}**) ?**

(2, 8, 5, 11, 3, 6, 2, 1, 10, 12)

→ ordered: (1, 2, 2, 3, 5, 6, 8, 10, 11, 12)

$P_{65} = \frac{K}{100} * n = \frac{65}{100} * 10 = 6.5$ **(fraction)** $\rightarrow$ **7th value** $\rightarrow P_{65} = 8$.

2. For frequency distribution:

Example: find 60% for the following data or or (P_{60}) ?

X	1	2	3	4	5
F	3	2	7	8	5
C.F	3	5	12	20	25
Intervals	0-3	4-5	6-12	13-20	21-25

$P_{60} = \frac{K}{100} * n = \frac{60}{100} * 25 = 15$ → (whole number) $\frac{15^{th} + 16^{th}}{2} = \frac{4+4}{2} = 4$

$P_{60} = 4.$

3. For grouped frequency distribution:

Example : find 70% for the following data (or P_{70})?

Intervals	0-6	6-12	12-18	18-24
Frequency	3	8	7	2
C.F	3	11	18	20
U.R.B	6	12	18	24

$P_{70} = \frac{K}{100} * n = \frac{70}{100} * 20 = 14^{th}$ value غير موجودة بشكل مباشر رح نستخدم

Inter polation method

C.F	3	11	14	18	20
U.R.B	6	12	P_{70}	18	24

11	14	18

$$\frac{14 - 11}{18 - 11} = \frac{P_{70} - 12}{18 - 12} \rightarrow P_{70} = 14.5$$

12	P_{70}	18

***To find the percentile that corresponds to a specific data entry x, use the following formula:**

$$\text{Percentile of } x = \frac{\text{number of data entries less than } x}{\text{total number of data entries}} \cdot 100$$

Example: **The age distribution for a random sample of 20 school students is**

Age (X)	10	12	15	18
F	2	6	7	5

1- The percentage of students having age above 15 years is:

A) 20%. **B)24%** **C) 25%** **D) 30%** **E) 40%**

Solution:

C.F	2	8	15	20
I	0-2	3-8	9-15	16-20

$$P_{age\ above\ 15\ years} = \frac{number\ of\ students\ above\ 15\ years}{total\ number\ of\ data\ entries} * 100 = \frac{5}{20} * 100$$

$$= 0.25 * 100 = 25\% \rightarrow C$$

2- The 50^{th} percentile of this sample. **(H.W)**

A) 11 **B)15** **C) 13.5** **D) 25** **E) 12.5.**

Example (H.W): **The IQ (Inter Quotient) for people has a Bell-shaped distribution with mean 100 and standard deviation of 15. The percentage of people with IQ≤ 115 is?**

A) 16%. **B)68%** **C) 95%** **D) 84%** **E) 52%**

The Inter-percentile-range(IPR) ($P_n - P_m$):

Example: find 35% to 65% inter- percentile-range (IPR) for the following data :

(2, 2, 3, 5, 8, 8, 9, 10, 12, 13)

$P_{35} = \frac{K}{100} * n = \frac{35}{100} * 10 = 3.5$ (fraction) $\rightarrow 4^{th}$ value $\rightarrow P_{35} = 5$.

$P_{65} = \frac{K}{100} * n = \frac{65}{100} * 10 = 6.5$ (fraction) $\rightarrow 7^{th}$ value $\rightarrow P_{65} = 9$.

$IPR = P_{65} - P_{35} = 9 - 5 = 4$.

Example (H.W) : In a test for a class of 30 student, scores obtain by 11 students were: 4, 4, 5, 5, 6, 7, 8, 9, 9, 10, 10.

1. **Find the 5 number summary.**
2. **Find the percentile rank(x) in P_x for a score of 8 in this test.**
3. **Find the mean and standard deviation.**

Comparison two collection — المقارنة بين مجموعتين او اكثر

We can compare two collection by:

I. **Z-score or the standard score** — الدرجة المعيارية او القياسية.

II. **Coefficient of variation C.V** — معامل التشتت

1. Z-score or the standard score:

هنا سنقيس مدى جودة العينة وذلك بإيجاد الفرق بين قيمة من البيانات و المتوسط الحسابي وقسمته على الانحراف المعياري

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

S **: Sample standard deviation**

- Z>0, means that the corresponding x-value is greater than the mean.
- Z<0, means that the corresponding x-value is less than the mean.
- Z=0, means that the corresponding x-value is the mean.

Example: consider the following data for Physics and Math , compute Z-score for both? Which one is the best?

	Math	Physics
X	87	72
$\bar{X}$	80	65
S	6	4

Sol: **Z-score** $= \frac{X-\bar{X}}{S}$ $\begin{cases} \textbf{for math} = \frac{87-80}{6} = 1.167 \\ \textbf{for physics} = \frac{72-65}{4} = 1.75 \end{cases}$

Z-score for physics > Z-score for math, it is means that 72 in physics is better than 87 in math.

2. Coefficient of variation C.V معامل التشتت.

سنقوم بقياس معامل التشتت في عينتين وبناء عليه التي لها معامل تشتت اعلى ، يكون التشتت في العينة اكبر

C.V$=\frac{S}{\bar{X}} * 100\%$. , $\bar{X}$: **Sample mean**

S **: Sample standard deviation**

Example: compute the coefficient of variation for the following data? Which data set has more variability?

	Class(1)	Class(2)
$\bar{X}$	60	70
S	4.5	5

Sol: **C.V**$=\frac{S}{\bar{X}} * 100\%$ $\begin{cases} \textbf{for Class(1)} = \frac{4.5}{60} * 100\% = 75\% \\ \textbf{for Class(2)} = \frac{5}{70} * 100\% = 7.14\% \end{cases}$

The variability of Class(1) is higher than it for Class(2)

Example: The mean speed of vehicles along a highway is 56 mi/h with a SD of 4 mi/h.

Find the z-scores of three cars traveling at 62mi/h, 47mi/h, and 56mi/h

SOLUTION The z-score that corresponds to each speed is calculated below.

$x = 62$ mph $\quad z = \frac{62-56}{4} = 1.5$

$x = 47$ mph $\quad z = \frac{47-56}{4} = -2.25$

$x = 56$ mph $\quad z = \frac{56-56}{4} = 0$

Interpretation From the z-scores, you can conclude that a speed of 62 miles per hour is 1.5 standard deviations above the mean; a speed of 47 miles per hour is 2.25 standard deviations below the mean; and a speed of 56 miles per hour is equal to the mean. The car traveling 47 miles per hour is said to be traveling unusually slow, because its speed corresponds to a z-score of -2.25.

*Standard Error:

$$S.E = \frac{S}{\sqrt{n}}$$ **, S : Sample standard deviation**

n : is a sample size

Example: Two random samples A and B 13 are drawn from the same population for which we have the following information

n_A= 15, S_A= 8.5 and n_B= 10 , S_B= 7

which sample is the more representative to the population

SOL:

For sample A $$S.E = \frac{S}{\sqrt{n}} = \frac{8.5}{\sqrt{15}} = 2.195$$

For sample B $$S.E = \frac{S}{\sqrt{n}} = \frac{7}{\sqrt{10}} = \frac{7}{3.162} = 2.214$$

Sample A is better than sample B, because its size is more than that of B.
Therefor A is more representative to the population than B.

Lecture 8

Biostatistics

Lecturer: Naba Mohammed Dhiaa Alashqar

Chapter 3
Elements of Probability

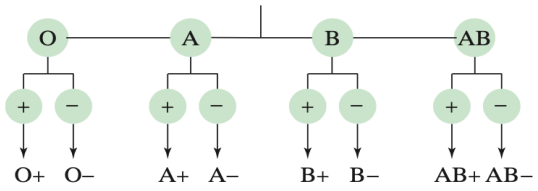
3.1: Basic concepts of probability and counting

- A **probability experiment** is a trial where results are obtained.
- An **outcome** is a single trial in the experiment.
- A **sample space(s)** is the set of all possible outcomes.
- An **event** is a subset of the sample space.
- A **simple event** consists of only one outcome.
- **N(S):** number of elements in the sample space.

Example(1): A survey consists of asking people for their blood type (O, A, and AB), including whether they are RH-positive or Rh-negative. List all the outcomes of the sample space.

Sol: There are four blood types: O, A, B, and AB. For each person, they are either Rh-positive or Rh-negative. A tree diagram gives a visual display of the outcomes of a probability experiment by using branches that originate from a starting point. It can be used to find the number of possible outcomes in a sample space as well as individual outcomes.

Tree Diagram for Blood Types



From the tree diagram, you can see that the sample space has eight possible outcomes, which are listed below.

{O+, O-, A+, A-, B+, B-, AB+, AB-}. Sample space

Example(2): find the sample space (S) for the following random experiments:

1. Tossing a coin one time
   S={H, T} → N(S)= 2.
2. Tossing a coin two times
   S={HH,TH, HT, TT} → $N(S) = 2^2 = 4$.
3. Tossing a coin three times
   S=

Note: When tossing a coin K-times → $N(S) = 2^k$.

Question: if you toss a coin 6 times, then the number of elements in the sample space is:

4. Tossing a die one time
   S={1, 2, 3, 4, 5, 6} → $N(S) = 6$

5. Tossing a die two times

S=

	1	2	3	4	5	6
1	**1,1**	**1,2**	**1,3**	**1,4**	**1,5**	**1,6**
2	**2,1**	**2,2**	**2,3**	**2,4**	**2,5**	**2,6**
3	**3,1**	**3,2**	**3,3**	**3,4**	**3,5**	**3,6**
4	**4,1**	**4,2**	**4,3**	**4,4**	**4,5**	**4,6**
5	**5,1**	**5,2**	**5,3**	**5,4**	**5,5**	**5,6**
6	**6,1**	**6,2**	**6,3**	**6,4**	**6,5**	**6,6**

$\rightarrow$ $N(S) = 6^2 = 36.$

Note: When tossing a die K-times $\rightarrow$ $N(S) = 6^k$.

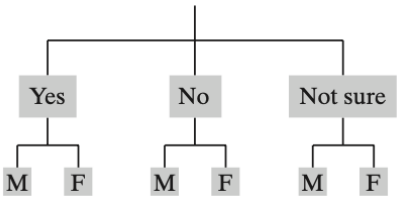
- An event is a well-defined subset of the sample space

For example, the event the outcome of the sum of the faces on the two dice equal six consists of the five outcomes (1,5), (2,4), (3,3), (4,2), (5,1).

Example(3): For each probability experiment, determine the number of outcomes and identify the sample space.

1. A probability experiment consists of recording a response to the survey statement and the gender of the respondent.
2. A probability experiment consists of recording a response to the survey statement at the left and the age (18–34, 35–49, 50 and older) of the respondent.
3. A probability experiment consists of recording a response to the survey statement at the left and the geographic location (Northeast, South, Midwest, West) of the respondent.

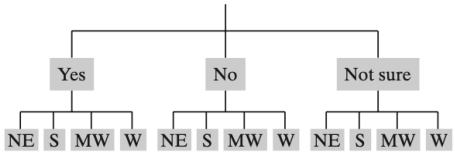
Sol: (1)



6 outcomes

Let Y= Yes, N= No, NS= Note sure, M= Male, F= Female.

Sample space (S)= {YM, YF, NM, NF, NSM, NSF}.

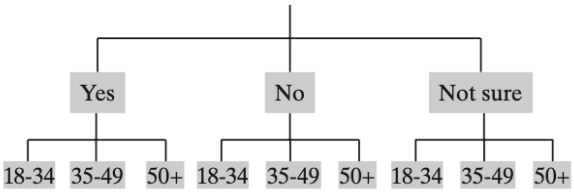


Sol: (2): 9 outcomes

Let Y= Yes, N= No, NS= Note sure, 50+=50 and older

Sample space (S)= {Y18-34, Y35-49, Y50+, N18-34, N35-49, N50+, NS18-34, NS35-49, N50+ }

(3) SOL:



Outcomes=

H.W

Example(4): Determine the number of outcomes of each event.

A. **Selecting a defective machine part.**
B. **Rolling at least four in a six-sided die.**

Sol:

Event A has only one outcome.

Event B has three outcomes: rolling a 4, a 5, or a 6.

The fundamental Counting Principle

The number of ways that events can occur in sequence is found by multiplying the number of ways one event can occur By the number of ways other events can occur.

Example(5): From the table below, Count the number of different ways you can select one manufacturer, one car size, and one color.

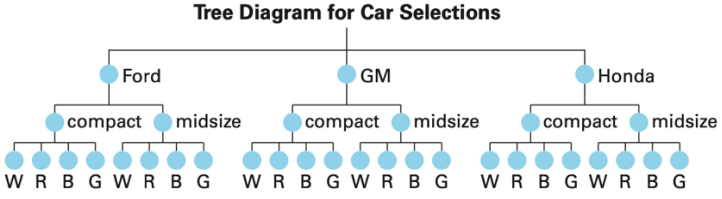
Manufacturer	Car size	Color
Ford	compact	white (W)
GM	midsize	red (R)
Honda		black (B)
		green (G)

Sol:

There are three choices of manufacturers, two choices of car sizes, and four choices of colors. **Using the Fundamental Counting Principle,** you can determine that the number of ways to select one manufacturer, one car size, and one color is

3*2*4 = 24 ways

Using a tree diagram, you can see why there are 24 options



Example(6): The access code for a car's security system consists of four digits. Each digit can be any number from 0-9.
How many access codes are there when

A. **Each digit can be used only once and not repeated.**

B. **Each digit can be repeated.**

C. **Each digit can be repeated but the first digit cannot be 0 or 1.**

Definition: **Classical probability** is used when all outcomes are equally likely to happen.
The classical probability for an event E is given by:

$$P(E) = \frac{\text{Number of outcomes in event } E}{\text{Total number of outcomes in sample space}}.$$

Example(7): You roll a six-sided die. Find the probability of

1. Event A: rolling a 4.
2. Event B: rolling a number less than 5.
3. Event C: rolling an odd number.
4. Event D: rolling a number greater than 3.
5. Event E: rolling a number less than 1.
6. Event F: rolling a number greater than zero.

SOL:

Space ={1, 2, 3, 4, 5, 6}. → N(S)= 6

1) A= {4} → simple event → P(A)= $\frac{1}{6}$
2) B={1, 2, 3, 4} → P(B)= $\frac{4}{6}$
3) C={1, 3, 5} → P(C)= $\frac{3}{6} = 2$
4) D={ } → P(D)=
5) E={} OR φ → P(E)= $\frac{0}{6} = 0$
6) F={ } → P(F)=

Remarks:

- The probability of the sample space is 1 (**Sure Event**)
- The probability of the **impossible** event is 0
- All events have a probability between 0 and 1

Definition: Empirical probability is based on observations from the experiment.
The empirical probability of an event E is the relative frequency

$$P(E) = \frac{\text{Frequency of event } E}{\text{Total frequency}}$$
$$= \frac{f}{n} \qquad \text{Note that } n = \Sigma f.$$

Example(8): In a sample of 50 people, 21 had type O blood, 22, had type A blood, 5 had type B blood, and 2 had type AB blood.
Set up a frequency distribution and find the following probabilities.

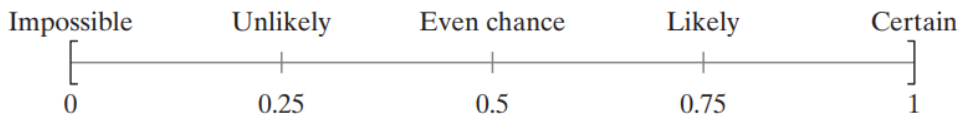
1) A person has type O blood.
2) A person has type AB blood.
3) A person does not have type A blood.

Definition: **Complement of event E** (E') is the set of all outcomes that are not in event E.

$$P(E) + P(E') = 1$$
$$P(E) = 1 - P(E')$$
$$P(E') = 1 - P(E)$$

- **Law of Large Numbers:** If an experiment is repeated over and over, the empirical probability approaches the Classical probability.
- **Subjective probability** is a result of an educated guess. (Probability of rain)

Range of probabilities: The probability of an event E is between 0 and 1, inclusive. That is
$\mathbf{0 \leq P(E) \leq 1}$



Example(9): The frequency distribution on the right shows the number of voting-age in American citizens by age.
Find the probability that a citizen chosen at random is in the age range

- 18-29 years old
- Less than 18 years old
- 65 years old and over
- Not 30-44 years old

Sol:

Ages	Frequency, f (in millions)
18 to 29	48.9
30 to 44	53.9
45 to 64	78.1
65 and over	46.0

Unions, intersections, and complements of sets

Use the following information to answer the questions below.
S= {1,2,3,4,5,6,7,8,9,10} (Sample Space)
A= {1,2,5,6,9,10} (Subset)
B= {3,4,7,8} (Subset)
C= {2,3,8,9,10} (Subset)

Find:

- $\overline{A}$
- $A \cup C$
- $A \cap B$

- $\bar{A} \cap C$
- $\overline{B \cup C}$
- $A \cap B \cap C$
- $\bar{\bar{A}}$

3.2 : Multiplication Rule, Addition Rule, and Conditional Probability

Definitions:

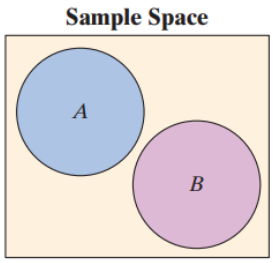
- Two events are **independent** when the occurrence of one of the events does not affect the probability of the occurrence of the other event. Two events A and B are independent when

$$P(B|A) = P(B) \qquad \text{Occurrence of } A \text{ does not affect probability of } B$$

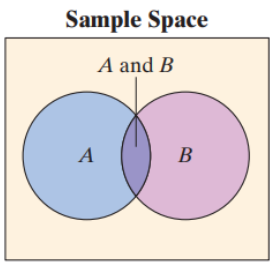
or when

$$P(A|B) = P(A). \qquad \text{Occurrence of } B \text{ does not affect probability of } A$$

- Events that are not **independent** are called dependent.
- Two events A and B are called mutually exclusive if they don't have any common outcomes



A and *B* are mutually exclusive.



A and *B* are not mutually exclusive.

- Conditional Probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, P(A) > 0$$

- De Morgan's laws

$$P(\overline{A \cap B}) = P(\bar{A} \cup \bar{B})$$
$$P(\overline{A \cup B}) = P(\bar{A} \cap \bar{B})$$

- If A and B are **mutually exclusive (disjoint)**

  $P(A \cap B) = 0$

  $P(A \cup B) = P(A) + P(B)$

  $P(A|B) = 0 \; and \; P(B|A) = 0$

- Two events A and B are **independent** if any one of the following holds.

  $P(A \cap B) = P(A)P(B)$

  $P(A|B) = P(A)$

$P(B|A) = P(B)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap \bar{B}) = P(A) - P(A \cap B)$
- $P(B \cap \bar{A}) = P(B) - P(A \cap B)$

Type of probability and probability rules	In words	In symbols
Classical Probability	The number of outcomes in the sample space is known and each outcome is equally likely to occur.	$P(E) = \frac{\text{Number of outcomes in event } E}{\text{Number of outcomes in sample space}}$
Empirical Probability	The frequency of each outcome in the sample space is estimated from experimentation.	$P(E) = \frac{\text{Frequency of event } E}{\text{Total frequency}} = \frac{f}{n}$
Range of Probabilities Rule	The probability of an event is between 0 and 1, inclusive.	$0 \le P(E) \le 1$
Complementary Events	The complement of event E is the set of all outcomes in a sample space that are not included in E, and is denoted by E'.	$P(E') = 1 - P(E)$
Multiplication Rule	The Multiplication Rule is used to find the probability of two events occurring in sequence.	$P(A \text{ and } B) = P(A) \cdot P(B\|A)$ Dependent events $P(A \text{ and } B) = P(A) \cdot P(B)$ Independent events
Addition Rule	The Addition Rule is used to find the probability of at least one of two events occurring.	$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ $P(A \text{ or } B) = P(A) + P(B)$ Mutually exclusive events

Example(10): If P(A) =0.3, P(B)= 0.6, and P(A and B) =0.2. Find P(A or B).

Sol: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.6 - 0.2 = 0.7$

Example(11): The probability that a student will pass calculus is 0.8, that he will pass physics is 0.65, and that he will pass Both is 0.6. Find the probability that

1. He will pass at least one of the two.
2. He will fail both classes.
3. If he passes physics, what is the probability he will pass calculus?

SOL: 1. $P(C \cup P) = P(C) + P(P) - P(C \cap P) = 0.8 + 0.65 - 0.6 = 0.85.$
2. $1 - p(C \cup P) = 1 - 0.85 = 0.15$
3. $p(C \backslash P) = \frac{p(C \cap P)}{p(P)} = \frac{0.6}{0.65} = 0.9231.$

Lecture 9

Biostatistics

Lecturer: Naba Mohammed Dhiaa Alashqar

(Multiplication Rule, Addition Rule, and Conditional Probability)

Rules of Probability:

1. $P(\bar{A})=1-P(A)$

2. i) $P(A \cap \bar{B}) = P(A) - P(A \cap B)$
   ii) $P(\bar{A} \cap B) = P(B) - P(A \cap B)$

3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4. De Morgan's laws
   i) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$
   ii) $P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B)$

5. A and B are said to be mutually exclusive (disjoint) if
   i) $P(A \cap B) = 0$. Or. ii) $P(A \cup B) = P(A) + P(B)$.

6. A and B are two independent if they don't influence each other

Mathematically,

$$P(A \cap B) = P(A)P(B)$$

7. Conditional Probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Note: If A&B are independent, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

$$\rightarrow P(A|B) = P(A)$$

Example: If $P(A) = 0.8, P(B) = 0.7$ and $P(A \cap B) = 0.6$. Find:

1) $P(\bar{A})$ 2) $P(\bar{B})$ 3) $P(A \cap \bar{B})$ 4) $P(\bar{A} \cup \bar{B})$ 5) $P(A \cup B)$ 6) $P(\bar{A} \cap \bar{B})$ 7) $P(\bar{A} \cup \bar{B})$

8) $P(\bar{A} \cap B)$ 9) $P(A \cup \bar{B})$ 10) $P(A|B)$ 11) $P(\bar{A}|B)$ 12) $P(A|\bar{B})$ 13) $P(\bar{A} \backslash \bar{B})$ 14) $\overline{P(A|B)}$.

Sol:

Example: **If $P(A) = 0.6, P(B) = 0.5$ & $P(A \cup B) = 0.8$, Are A&B Mutually exclusive? Independent? Or neither**

Sol:

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$0.8 = 0.6 + 0.5 - P(A \cap B)$

$P(A \cap B) = 0.6 + 0.5 - 0.8$

$P(A \cap B) = 0.3 \neq 0$

∴ A and B are not Mutually exclusive

- $P(A \cap B) = P(A)P(B)$

Now $P(A) * P(B) = 0.6 * 0.5 = 0.3$

- Since $P(A \cap B) = P(A) * P(B)$, then A and B are independent.

Example: **If A and B be two independent events such that $P(A) = 2P(B)$ and $P(A \cup B) = 0.8$. Find $P(A)$?**

Sol:

let $x = P(B) \rightarrow 2x = P(A) \rightarrow 2x^2 = P(A \cap B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$0.8 = 2x + x - 2x^2 \quad \rightarrow \quad 2x^2 - 3x + 0.8 = 0$

$x = \frac{-b \pm \sqrt{(b)^2 - 4ac}}{2a} \quad \rightarrow = \frac{3 \pm \sqrt{(-3)^2 - 4(2)(0.8)}}{2(2)} \quad \rightarrow \quad x = 1.15$ or $x = 0.347$.

Since $0 \leq P \leq 1$. $\rightarrow P(B) = 0.347$

$\rightarrow P(A) = 2x \rightarrow 2(0.347)$

$\rightarrow P(A) = 0.694$.

Example(12): **Let A and B be two independent events in the sample space such that $P(A) = 0.1$, and $P(B) = 0.2$, find $P(\overline{A \cup B})$**

A)0.28. **B) 0.18** **C)0.4** **D)0.72** **E) 0.3**

Sol:

A and B independent $\rightarrow P(A \cap B) = P(A) * P(B) = 0.1 * 0.2$

$\rightarrow P(A \cap B) = 0.02$

$P(\overline{A \cup B}) = 1 - P(A \cup B)$

$= 1 - [P(A) + P(B) - P(A \cap B)] = 1 - [0.1 + 0.2 - 0.02] = 0.72 \rightarrow D$.

Example(13): **Let A and B be two events in a given sample space, $P(B) = 0.8$, $P(A \cup B) = 0.7$, $P(\overline{A} \cap B) = 0.5$, then P(A)=?**

A)0.2. **B) 0.4** **C)0.5** **D)0.8** **E) 0.31**

Sol:

$P(\overline{A} \cap B) = P(B) - P(A \cap B) \rightarrow 0.5 = 0.8 - P(A \cap B) \rightarrow P(A \cap B) = 0.3$

$P(A \cup B) = P(A) + P(B) - P(A \cap B) \rightarrow 0.7 = P(A) + 0.8 - 0.3 \rightarrow P(A) = O.2 \rightarrow A$

Example(14): **A coin is flipped and a dice is rolled. Find the probability of getting a head and an even number.**

Sol:

Example(15): **suppose P(A)=0.28, and P(B)=0.52. if P(A\B)=0.14, Then P(B\A)=?**

A)0.26. **B) 0.0728** **C)0.1456** **D)0.0392** **E) 0.0754.**

Sol:

Example(16): Let A and B be two independent events such that P(A)>0 and $P(A \cap B) = 2 * P(A \cap \bar{B})$. Find P(B)

Sol: A and B be two independent→ $P(A \cap B) = P(A) * P(B)$

→ We know that $P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A) * P(B)$

→ $P(A \cap B) = 2 * P(A \cap \bar{B})$

→ $P(A) * P(B) = 2 * (P(A) - P(A) * P(B))$

$= 2P(A) - 2(P(A) * P(B))$

→ $P(A) * P(B) = P(A) * (2 - 2P(B))$

→ $P(B) = 2 - 2P(B)$

→ $P(B) = \frac{2}{3} = 0.667.$

Example(17): **If the probabilities that Jane, Tom, and Mary will be chosen as a chairperson of the board are 0.5, 0.3, and 0.2 respectively. Find P (Tom or Mary)**

A)0.4 **B) 0.7** **C) 0.8** **D) 0.5** **E) 0.6**

Sol:

Example(17): **If A and B are independent, $P(A) = 0.6$ and $P(B) = 0.3$, then find the following:**

1) $P(A \cap \bar{B})$ **2)** $P(\bar{A} \cap B)$. **3)** $P(A \setminus \bar{B})$

Sol:

H.W

Tree diagram or Bayes' theorem (الرسم كشجرة):

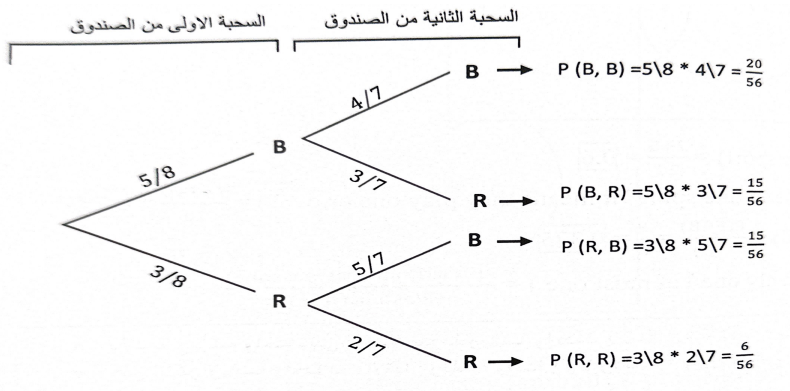
- تستخدم هذه الطريقة عندما نختار عنصرين او اكثر من مجموعة عناصر
- او تستخدم عندما تكون لدينا تجزئة بالسؤال وهو درس ال (Bayes' theorem)

Example(18): Two balls are to be selected at random from a box contains five black balls & three red balls. Find the probability of getting the following: "without replacement"

1. Two black balls
2. Two balls of same color
3. Two balls of different colors
4. At least one ball is black
5. The first is black given that the second is red

6. The second ball is black given that the two balls of the same color.

Sol: B: Black R:Red B,B=B∩ B



5. $P(1^{st}\ B \backslash 2^{nd}\ R) = \frac{P(1^{st}\ B \cap 2^{nd}\ R)}{P(2^{nd}\ R)} = \frac{P(B,R)}{P(B,R)+P(R,R)} = \frac{\frac{15}{56}}{\frac{15}{56}+\frac{6}{56}} = \frac{5}{7}$

6. $P(2^{nd}\ B \backslash two\ of\ same\ color) = \frac{P(B,B)}{P(B,B)+P(R,R)} = \frac{\frac{20}{56}}{\frac{20}{56}+\frac{6}{56}} = \frac{20}{26}$.

Example:

	smoking	Not smoking	Total
Male	8	2	10
Female	6	4	10
Total	14	6	20

a) a student is selected at random, what is the prob. That student:
i) a female smoking
ii) a female or smoking
iii) a female if the student is a smoker

b) 2 students are selected at random, what is the prob. That exactly 1 of them is a smoker
sol:

a) i) $P(F \cap S) = \frac{6}{20} = \frac{3}{10}$

ii) $P(F \cup S) = P(F) + P(S) - P(F \cap S)$

$$= \frac{10}{20} + \frac{14}{20} + \frac{6}{20} = \frac{18}{20}$$

iii) $P(F \backslash S) = \frac{P(F \cap S)}{P(S)} = \frac{\frac{6}{20}}{\frac{14}{20}} = \frac{6}{14} = \frac{3}{7}$

b)

Example: The following table represents the distribution of blood type by gender in a group of 1,000 individuals.

Blood type	Male	Female	Total
O	80	370	450
A	150	250	400
B	50	50	100
AB	20	30	50
Total	300	700	1,000

1. **What is the probability of selecting a person with blood type AB?**

$$P(AB) = 1 - P(O) + P(A) + P(B)$$

$$= 1 - \left[\frac{450}{1000} + \frac{400}{1000} + \frac{100}{1000}\right] = 1 - \frac{950}{1000} = 1 - 0.95 \rightarrow P(AB) = 0.05$$

2. **What is the probability of selecting a male and blood type A?**

$$\boldsymbol{P(M \cap A) = \frac{150}{1000} = 0.15}$$

3. **What is the probability of selecting a female or blood type O?**

$$\boldsymbol{P(F \cup O) = P(F) + P(O) - P(F \cap O) = \frac{700}{1000} + \frac{450}{1000} - \frac{370}{1000} = \frac{780}{1000} = 0.78}$$

4. **What is the probability that the person selected has blood type B if the person is a male? (Given he is a male)**

$$\boldsymbol{P(B \backslash M) = \frac{P(B \cap M)}{P(M)} = \frac{\frac{50}{1000}}{\frac{300}{1000}} = \frac{0.05}{0.3} = 0.166}$$

5. **What is the probability that the person selected is a male if the blood type is B? (Given that the blood type is B)**

$$\boldsymbol{P(M \backslash B) = \frac{P(M \cap B)}{P(B)} = \frac{\frac{50}{1000}}{\frac{100}{1000}} = \frac{0.05}{0.1} = 0.5}$$

Example: The following table represents the distribution of blood type by gender in a group of 65 individuals. **(H.W)**

	A	B	C(OR)	Total
Male	18	19	3	40
Female	12	4	9	25
Total	30	23	12	65

1. **Find the probability that the person was female OR got "C"**

2. **Find the probability that the person did NOT get "B"**

Example: The probability that a head will appear when we toss a two-sided coin is 0.6. if the coin is tossed three times, the probability of getting the same face in all three tosses is: (H.W)

A) $\frac{1}{4}$ B) $\frac{1}{8}$ C) $(0.6)^3 + (0.4)^3$ D) $(0.6)^3$ E) $2\,(0.4)^3\,(0.6)^3$.

Example: Determine whether the statement is true or false. If it is false, rewrite it as a true statement. (H.W)

1. If two events are independent, then P(A\B) = P(B).
2. If events A and B are dependent, then P(A and B) = P(A) * P(B).
3. If events A and B are mutually exclusive, then P(A or B) = P(A) + P(B).

Example: The table shows the numbers of male and female students in the U.S. who received bachelor's degrees in business in a recent year. A student is selected at random. Find the probability of each event. (H.W)

	Business degrees	Nonbusiness degrees	Total
Male	191.310	621,359	812,669
Female	172,489	909,776	1,082,265
Total	363,799	1,531,135	1,894,934

1. The student is male or received a business degree.
2. The student is female or received a nonbusiness degree.
3. The student is not female or received a nonbusiness degree.

Lecture 10

Biostatistics

Lecturer: Naba Mohammed Dhiaa Alashqar

3.4 Permutations, Combinations, and Applications of Counting Principles

Definition:

- **Fundamental Counting Principle**: if one event can occur in m ways and a second event can occur in n ways, then the number of ways the two # events can occur in sequence is $\boldsymbol{m * n}$.

- A **permutation** is an ordered arrangement of objects where the order is important. The number of different permutations of n distinct objects is **n!.**

- The expression **n!** is read as **n factorial**. If **n** is a **positive integer**, then n! is defined as follows.

$$n! = n(n-1)(n-2)(n-3)\ldots\ldots(3)(2)(1)$$

As a special case, 0! = 1

Example: find the following

1. **5!** = 5*4*3*2*1=120
2. **10!= 10*9*8……..2*1=3628800**
3. $\frac{5!}{7!} = \frac{5!}{7*6*5!} = \frac{1}{7*6} = \frac{1}{42}$
4. Let $\frac{n!}{(n-2)!} = 30$, **then find the value of n?**
   $\frac{n(n-1)(n-2)!}{(n-2)!} = 30 \rightarrow n*(n-1) = 30 \rightarrow n^2 - n - 30 = 0$
   $\rightarrow (n-6)*(n+5) = 0 \rightarrow$ **n=6 "accepted"** n=-5 "rejected".

Example: In how many ways can 6 people be seated in a row?

- The number of **permutations** of n **distinct** objects taken r at a time is given by the formula:

$$_nP_r = \frac{n!}{(n-r)!}, \text{ where } r \leq n.$$

Example: How many ways can we award a 1st, 2nd, and 3d place prize winners among 8 contestants?
Sol:

$$_nP_r = \frac{n!}{(n-r)!} \rightarrow \qquad {}_8P_3 = \frac{8!}{(8-3)!} = \frac{8*7*6*5!}{5!} = 8*7*6 = 336$$

Example: **let** $_nP_2 = 6$, find the value of (n)?

$$_nP_2 = \frac{n!}{(n-2)!} = 6 \rightarrow \frac{n*(n-1)*(n-2)!}{(n-2)!} = 6 \rightarrow n*(n-1) = 6 \rightarrow n^2 - n - 6 = 0$$

$\rightarrow (n-3)(n+2) = 0 \rightarrow n = 3$ **"accepted"** $\quad n = -2$ **"rejected"**

Note:

1. if r = n then ${}^{n}P_{n} = n!$
2. ${}^{n}P_{n-1} = n!$
3. ${}^{n}P_{0} = 1$
4. ${}^{n}P_{1} = n$

Example: find the following:

1. ${}^{10}P_{10} = \frac{10!}{(10-10)!} = \frac{10!}{0!} = 10!$
2. ${}^{10}P_{9} = \frac{10!}{(10-9)!} = 10!$
3. ${}^{10}P_{0} =$
4. ${}^{10}P_{1} =$

Definition: The number of distinguishable permutations of n objects, where n1 are of one type, n2 are of another type… is

$$\frac{n!}{n_1! \cdot n_2! \cdot n_3! \cdots n_k!}$$

where

$$n_1 + n_2 + n_3 + \cdots + n_k = n.$$

Example: 1. How many distinguishable permutations are there to order the letters AAAABBC?

2. How many different words can be formed from the letters of the word MISSISSIPPI?

Definition: The number of combinations of r objects selected from a group of n objects without regard to order is:

$${}^{n}C_{r} = \binom{n}{r} = \frac{n!}{r!\,(n-r)!} = \frac{{}^{n}P_{r}}{r!}$$

Where r ≤ n.

Example:

- In how many ways can a committee of 3 be chosen from a group of 12 people?
  Sol:

$${}^{12}C_{3} = \binom{12}{3} = \frac{12!}{3!\,(12-3)!} = \frac{12 * 11 * 10 * 9!}{3! * 9!} = \frac{12 * 11 * 10}{3!} = \frac{1320}{6} 220$$

- A committee of 16 people, 7 women and 9 men, is forming an 8-member subcommittee that must consist of 3 women and 5 men. In how many ways can the subcommittee be formed?

  Sol:
  the 3 women can be chosen from 7 women in

$$\binom{7}{3} = \frac{7!}{3!\,(7-3)!} = \frac{7!}{3! * 4!} = \frac{7 * 6 * 5 * 4!}{3! * 4!} = \frac{7 * 6 * 5}{6} = 35\ ways$$

  Let $n_1 = 35$

  The 5 men can be chosen from 9 men in

$$\binom{9}{5} = \frac{9!}{5!\,(9-5)!} = \frac{9!}{5! * 4!} = \frac{9 * 8 * 7 * 6 * 5!}{5! * 4!} = \frac{9 * 8 * 7 * 6}{4!} = \frac{3024}{24} = 126\ \ ways$$

  Let $n_2 = 126$
  Hence the committee can be chosen by (Fundamental Counting Principle)

$$n_1 * n_2 = 35 * 126 = 4410$$

This can means that we can form 4410 different committees consisting of 3 women and 5 men from 7 women and 9 men.

Notes:

1. ${}^{n}\mathrm{C}_{n} = \binom{n}{n} = 1$ 2. ${}^{n}\mathrm{C}_{0} = \binom{n}{0} = 1$ 3. ${}^{n}\mathrm{C}_{n-1} = \binom{n}{n-1} = n$ 4. ${}^{n}\mathrm{C}_{1} = \binom{n}{1} = n.$

Example: If a person can select 3 presents from 10, how many different combinations are there?
Sol:

Example: a box contains five black balls, three white & seven green, if four balls are selected at random

A) Find the probability of getting the following:
   1. Two black & two white **2. Two black & two of the other colors**
   3. At least one ball is black

B) number of ways to choose three of them are black

Sol: **Hint** $P(A) = \frac{n(A)}{n(S)}$

A) 1. P(Two black, two white, zero green)= $\frac{\binom{5}{2}\binom{3}{2}\binom{7}{0}}{\binom{15}{4}} = 0.0219.$

2. P(Two black, two not black)= $\frac{\binom{5}{2}\binom{10}{2}}{\binom{15}{4}} = 0.329.$

3. P(At least one ball is black)=
P(1 black, 3 not black)+ P(2 black, 2 not black)+ P(3 black, 1 not black)+ P(4 black, 0 not black)= 1- P(4 not black)

$$1 - \frac{\binom{5}{0}\binom{10}{4}}{\binom{15}{4}} = 0.846$$

B) Number of ways = $\binom{5}{3}\binom{10}{1} = 100$ ways.

Example: **there are 5 students in class, what is the prob. That:**

a) No 2 students in the class have the same birthday
b) All the students in the class have the same birthday
c) Exactly 2 students in the class have the same birthday
d) Exactly 3 students in the class have the same birthday

Sol:

Example: **The number of ways of selecting 3 red and 2 white balls from a box that contains 5 red and 4 white balls? H.W**

A) 12. B)60. C)40. D)20

Lecture 11+12

Biostatistics

Lecturer: Naba Mohammed Dhiaa Alashqar

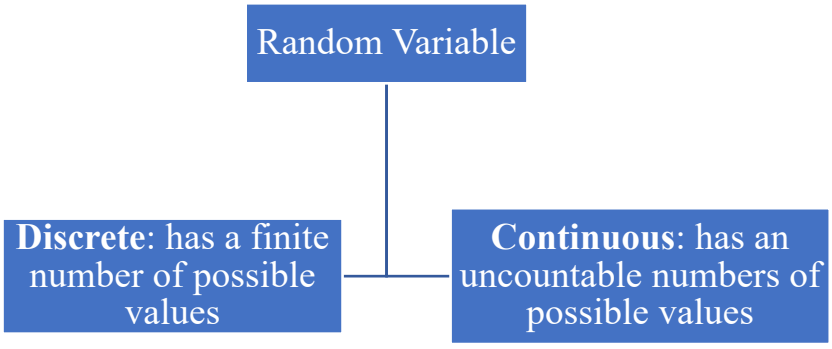
Chapter 4

Discrete Probability Distributions

4.1 Probability Distributions

Definition: A random variable x represents a single value associated with each outcome of a probability experiment.

Random variable



Example: Identify the following random variables:

1. The number of arrivals at an emergency room between midnight and 6:00$a.m.$ discrete
2. The number of applicants for a job. discrete
3. The temperature of a cup of coffee served at a restaurant. Continuous
4. The volume of gasoline in a 21-gallon tank. continuous, the amount can be any volume between 0 and 21 gallons
5. Number of calls you make in a day. discrete
6. The speed of a rocket. continuous
7. Number of days of rain for the next three days. discrete

Example: Toss a coin 2 times, define the random variable X as the number of heads in the tosses.

Difference between A discrete & continuous random variable

→ let X be a discrete variable with p.d.f, then

$P(2 \le x \le 4) = P(x = 2) + P(x = 3) + P(x = 4)$

→ let X be a continuous variable with p.d.f, then

$$P(2 \le x \le 4) = \int_2^4 f(x)dx$$

❖ **A probability Density Function (P.D.F):**

A function F(X)= P(X=x_i) is called (P.D.F) if

I. F(X)≥ 0, for all x

II. $\sum F(x_i) = 1$

Definition: A discrete probability distribution lists each possible value the random variable can assume, together with its probability. (Probability Density Function "PDF")

X	x_1	x_2	x_3	…	x_n
P(x)	P(x_1)	P(x_2)	P(x_3)	…	P(x_n)

- $0 \leq P(x_i) \leq 1, \forall i$
- $\sum_{i=1}^{n} P(x_i) = 1$

Example: Tell whether these tables represent a probability distribution function or not

Days of rain, *x*	0	1	2	3
Probability, *P*(*x*)	0.216	0.432	0.288	0.064

x	5	6	7	8
***P*(*x*)**	0.28	0.21	0.43	0.15

x	1	2	3	4
***P*(*x*)**	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{5}{4}$	−1

x	5	6	7	8
***P*(*x*)**	$\frac{1}{16}$	$\frac{5}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

x	1	2	3	4
***P*(*x*)**	0.09	0.36	0.49	0.10

Example: Complete the probability distribution using the table below.

Number of movies	0	1	2	3	4	5
Residents	25	35	40	64	88	148

X	**Frequency**	**Relative frequency = P(x)**
0	25	0.0625
1		35/400 = 0.0875
2	40	
3		
4	88	88/400=0.22
5	148	0.37

Example: Given the following probability distribution

X	0	1	2	3
P(x)	1/8	3/8	3/8	1/8

Find:

1. **P(x=1)**
   $\rightarrow P(x=1)=\frac{3}{8}=0.375.$
2. **P(0<x<1)= zero**
3. **P(1.5<x<2.5)**
   $\rightarrow P(x=2)=\frac{3}{8}=0.375$
4. **P(1.5<x<3.5)**
   →
5. **P(x=3\x≥1)**
   $$\rightarrow \frac{P(x=3)}{P(x\geq 1)}=\frac{P(x=3)}{P(x=1)+P(x=2)+P(x=3)}=\frac{\frac{1}{8}}{\frac{3}{8}+\frac{3}{8}+\frac{1}{8}}=\frac{0.125}{0.375+0.375+0.125}$$
   $$=\frac{0.125}{0.875}=0.1428$$
6. **P(x is odd\x≥1)**
   →

Definition: **The mean of a discrete random variable is given by:** $E(X)=\ \mu=\Sigma xP(x).$

The variance of a discrete random variable is

$$\sigma^2=\Sigma(x-\mu)^2P(x).$$

The **standard deviation** is

$$\sigma=\sqrt{\sigma^2}=\sqrt{\Sigma(x-\mu)^2P(x)}.$$

Variance(X) $=E(x-\mu)^2=E(x^2)-\mu^2$

$$\mathrm{Var(X)}=\mathrm{E(x^2)}-(\mathrm{E(X)})^2$$

Note: $\mathrm{E(x^2)}=\mathrm{Var(X)}+(\mathrm{E(X)})^2$

Example: **Find the mean and the standard deviation of the probability distribution given below.**

Score, *x*	1	2	3	4	5
Probability, *P*(*x*)	0.16	0.22	0.28	0.20	0.14

Sol:

$$E(X)=\sum X*P(X=x_i)=1*0.16+2*0.22+3*0.28+4*0.20+5*0.14$$
$$=\ 0.16+0.44+0.84+0.8+0.7=2.94\approx 2.9$$

standard deviation= $\sqrt{\text{variance}}$

variance (σ^2)= $E(x-\mu)^2=E(x^2)-\mu^2$

$$E(x^2) = \sum X^2 * P(X = x_i) = 1^2 * 0.16 + 2^2 * 0.22 + 3^2 * 0.28 + 4^2 * 0.20 + 5^2 * 0.14$$
$$= 0.16 + 0.88 + 2.52 + 3.2 + 3.5 = 10.26$$

$\mu^2 = (E(X))^2 = (2.94)^2 = 8.643$

Now Var(X) $= E(x^2) - \mu^2 = 10.26 - 8.643 = 1.617 \approx 1.6$
→ **standard deviation** $\boldsymbol{\sigma} = \sqrt{\textbf{variance}} = \sqrt{\mathbf{1.617}} = \mathbf{1.271} \approx 1.3$

Definition: The mean of a random variable represents what you would expect to happen (Expected Value)

$$\text{Expected Value} = E(x) = \mu = \Sigma x P(x)$$

It can be positive or negative.

Remarks: **Properties of the expected value and Properties of the variance**

1. E(c)= c, where c is a constant
2. E(aX)= aE(x)
3. E(X $\pm$ Y)= E(X)+ E(Y)
4. E(aX $\pm$ b)= a E(X)+ b
5. Var(a)=0
6. Var(aX)= a^2 Var(X)
7. Var (aX $\pm$ b) = a^2 Var(X)
8. Sd(aX $\pm$ b) = |a| Sd(X)

Example: consider $E(X) = 5$, then find the following:

1) E(7) =
2) E(E(X)) =
3) E(5 − 2x) =
4) $\mathrm{E}\left(1 + \frac{x}{2}\right) =$

Example: consider the following:

X	1	2	3	4
P(X=x_i)	0.4	0.3	0.2	0.1

Then find the following :

1) E(X) **2) E(X^2)** **3) E(e^x)** **4) E($\frac{1}{x}$)**

1. $E(X) = \Sigma X * P(X = x_i) = 1*0.4 + 2*0.3 + 3*0.2 + 4*0.1 = 2$
2. $E(X^2) = \Sigma X^2 * P(X = x_i) = 1^2*0.4 + 2^2*0.3 + 3^2*0.2 + 4^2*0.1 = 5$
3. $E(e^x) = \Sigma e^x * P(X = x_i) = e^1*0.4 + e^2*0.3 + e^3*0.2 + e^4*0.1 =$
4. $E(\frac{1}{x}) = \Sigma \frac{1}{x} * P(X = x_i) = \frac{1}{1} * 0.4 + \frac{1}{2} * 0.3 + \frac{1}{3} * 0.2 + \frac{1}{4} * 0.1 = 0.64$

Example: Consider the following is the probability of a random variable X:

X	1	2	4	5	7
$P(X = x_i)$	0.3	0.3	0.1	0.2	0.1

Find P(X is odd or X>2)?

a. 1 b. 0.8 c. 0.7 d. 0.6

Sol:

$$P(X \text{ is odd} \cup X > 2) = P(X \text{ odd}) + P(X > 2) - P(X \text{ odd} \cap X > 2)$$
$$= [P(x = 1) + P(x = 5) + P(x = 7)] + [P(x = 4) + P(x = 5) + P(x = 7)]$$
$$- [P(x = 5) + P(x = 7)] = (0.3 + 0.2 + 0.1) + (0.1 + 0.2 + 0.1) - (0.2 + 0.1)$$
$$= 0.7 \rightarrow C$$

Example: **Let X be a random variable, $E(X) = 10$ & $Var(X) = 9$, Then $E(3X^2 - 4X + 7) =$?**

Sol: (H.W)

Example: let X be a random variable. If X=1, 2, 3 and the p.d.f of X is $P(X=k)=\frac{1}{3}$ for all k=1, 2, 3. Then E(X)=?(H.W)

A)1. B) 2. C)2.5 D)1/3

Example: Suppose X is a random variable with possible values 3, -2 and -3 and with respective probabilities 0.15, 0.48 and0.37 then the mean and standard deviation of X, respectively, are:

A) -1.62 and 2.21 **B) -1.62 and 1.994**

C) -1.62 and 0.9533 **D) -0.667 and 0.9533**

Sol: (H.W)

Example: In a symmetric distribution, which is greater, the mean or the median? Explain.

Sol: Neither; in a symmetric distribution, the mean and median are equal.

4.2 Binomial Distributions

Definition: A binomial experiment is a probability experiment that satisfies the following conditions:

1. **The experiment has a fixed number of trials that are independent of each other ***
2. **The number of trials (n) should be more than or equal to 3 ($n \geq 3$)***
3. **There are only two possible outcomes for each trial (Success or Failure)***
4. **The probability of success is the same for each trial.***
5. **The random variable X counts the number of successes in each trial***

Notation for Binomial Experiments

Symbol	Description
n	The number of trials
p	The probability of success in a single trial
q	The probability of failure in a single trial ($q = 1 - p$)
x	The random variable represents a count of the number of successes in n trials: $x = 0, 1, 2, 3, \ldots, n$.

Example: **Tell whether the following experiments are binomial experiments or not and identify n, p, q, and x.**

- A certain surgical procedure has an 85% chance of success. A doctor performs the procedure on eight patients. The random variable represents the number of successful surgeries.

Sol: Is binomial

$n = 8$

$P = 0.85$

$q = 1 - P = 0.15$

$x = 0,1,2,\ldots,8$

- A jar contains five red marbles, nine blue marbles, and six green marbles. You randomly select three marbles from the jar, without replacement. The random variable represents the number of red marbles.

Sol: Not binomial

- You take a multiple-choice quiz that consists of 10 questions. Each question has four possible answers, only one of which is correct. To complete the quiz, you randomly guess the answer to each question. The random variable represents the number of correct answers.

Sol: Is binomial

$n = 10$

$P = 1/4$

$q = 1 - P = 3/4$

$x = 0,1,2,\ldots,10$

- **Finding the probability of a binomial experiment using the tree diagram and the binomial rule.**

Example: Assuming that having rain on one day is independent of having rain on another day, you have determined that there is a 40% probability of rain (and a 60% probability of no rain) on each of the three days. What is the probability that it will rain on 0, 1, 2, or 3 of the days?

1. **Is this a binomial experiment? (Determine n, p, q, and x)**

   Sol:

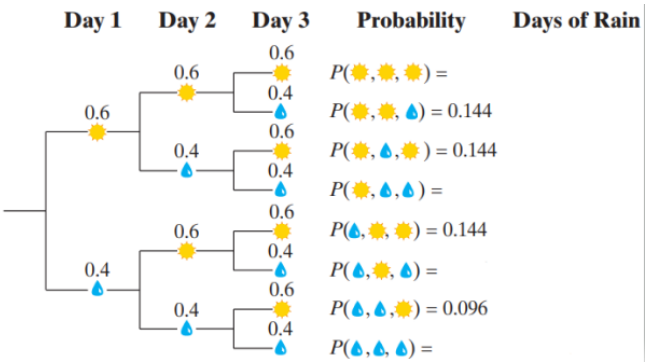
   $n = 3$

   $P = 0.4$

   $q = 1 - P = 0.6$

   $x = 0,1,2,3$

2. **Draw a tree diagram to find all the probabilities for all possible outcomes.**



3. Fill in the following probability distribution

Probability Distribution

Days of rain	Tally	Probability
0		
1		
2		
3		

4. Use the binomial probability formula to calculate the probabilities

Binomial Probability Formula

In a binomial experiment, the probability of exactly x successes in n trials is

$$P(x) = {}_nC_x p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}.$$

Note that the number of failures is $n - x$.

Population Parameters of a Binomial Distribution

Mean: $\mu = np$

Variance: $\sigma^2 = npq$

Standard deviation: $\sigma = \sqrt{npq}$

Example: **Finding Binomial Probabilities Using Formulas**

A survey found that 17% of U.S. adults say that Google News is a major source of news for them. You randomly select four adults and ask them whether Google News is a major source of news for them. Find the probability that (1) exactly two of them respond yes, (2) at least two of them respond yes, and (3) fewer than two of them respond yes. *(Source: Ipsos Public Affairs)*

SOLUTION

1. Using $n = 4$, $p = 0.17$, $q = 0.83$, and $x = 2$, the probability that exactly two adults will respond yes is

$$\begin{aligned} P(2) &= {}_4C_2(0.17)^2(0.83)^2 \\ &= 6(0.17)^2(0.83)^2 \\ &\approx 0.119. \end{aligned}$$

2. To find the probability that at least two adults will respond yes, find the sum of $P(2)$, $P(3)$, and $P(4)$. Begin by using the binomial probability formula to write an expression for each probability.

$$P(2) = {}_4C_2(0.17)^2(0.83)^2 = 6(0.17)^2(0.83)^2$$
$$P(3) = {}_4C_3(0.17)^3(0.83)^1 = 4(0.17)^3(0.83)^1$$
$$P(4) = {}_4C_4(0.17)^4(0.83)^0 = 1(0.17)^4(0.83)^0$$

So, the probability that at least two will respond yes is

$$\begin{aligned} P(x \geq 2) &= P(2) + P(3) + P(4) \\ &= 6(0.17)^2(0.83)^2 + 4(0.17)^3(0.83)^1 + (0.17)^4(0.83)^0 \\ &\approx 0.137. \end{aligned}$$

3. To find the probability that fewer than two adults will respond yes, find the sum of $P(0)$ and $P(1)$. Begin by using the binomial probability formula to write an expression for each probability.

$$P(0) = {}_4C_0(0.17)^0(0.83)^4 = 1(0.17)^0(0.83)^4$$
$$P(1) = {}_4C_1(0.17)^1(0.83)^3 = 4(0.17)^1(0.83)^3$$

So, the probability that fewer than two will respond yes is

$$\begin{aligned} P(x < 2) &= P(0) + P(1) \\ &= (0.17)^0(0.83)^4 + 4(0.17)^1(0.83)^3 \\ &\approx 0.863. \end{aligned}$$

Example: if a family has 7 children, what is the probability that three of them are boys?

Sol:

$X \sim Bin(n, P)$

$X \sim Bin(7, 0.5)$

$$P(X = 3) = \binom{7}{3} * (0.5)^3 * (0.5)^{7-3} = 0.273$$

Example (H.W): A coin is tossed 10 times (X: number of heads obtained),

A) Find the probability of getting:

1. Exactly 7 heads. 2. At least 2 heads. 3. At least one head given that at most two heads.

B) Find: 1. E(X) 2. Variance (X) 3. Standard deviation

Binomial Distributions

❖ **How to use binomial tables?**

Example: Let X be a discrete random variable following a binomial distribution Binomial (5, 0.1).

1. Find P(X=3)= 0.000
2. $P(X \leq 3) = 0.000 + 0.001 + 0.048 + 0.961 = 1$
3. P(X<3) = $0.001 + 0.057 + 0.941 = 0.999 = 1$
4. P(X>3) = $1 - P(X < 3) = 0$
5. $P(X \geq 3) = 0$
6. $P(1 \leq X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.048 + 0.001 + 0.000 = 0.049$
7. $P(1 < X \leq 3) = P(X = 2) + P(X = 3) = 0.001 + 0.000 = 0.001$
8. $P(1 < X < 3) = P(X = 2) = 0.001$
9. $P(X \geq 1 \backslash X \leq 3) = \frac{P(X \geq 1 \cap X \leq 3)}{P(X \leq 3)} = \frac{P(1 \leq X \leq 3)}{P(X \leq 3)} = \frac{P(X=1)+P(X=2)+P(X=3)}{P(X \leq 3)} = \frac{0.049}{1} = 0.049$
10. $P(X \leq 2 \backslash X \geq 4) = \frac{P(X \leq 2 \cap X \geq 4)}{P(X \geq 4)} = zero$
11. Find the mean, variance and standard deviation.

Sol:

1. $E(X) = \mu = n * p = 5 * 0.1 = 0.5$

2. Var(X)=$n * p * q = 5 * 0.1 * 0.9 = 0.45$

3. Std(X)= $\sqrt{\text{Var(X)}} = \sqrt{0.45} = 0.67$

Example: let $\boldsymbol{X \sim Bin(25, 0.2)}$ **Find** $\boldsymbol{P(\mu - \sigma < X \leq \mu)}$?

Sol:

$E(X) = \mu = n * p = 25 * 0.2 = 5$

Var(X)=$n * p * q = 25 * 0.2 * 0.8 = 4$

Std(X)= $\sigma = \sqrt{\text{Var(X)}} = \sqrt{4} = 2$

$\rightarrow P(\mu - \sigma < X \leq \mu) = P(5 - 2 < X \leq 5) = P(3 < X \leq 5)$

$= P(x \leq 5) - P(x \leq 3)$

$or = P(x = 5) + P(x = 4)$

$= 0.186681 + 0.196015 = 0.383$

Example: **if $X \sim Bin(n,p)$ with mean 4 and variance 2.4**, find $P(X \geq 2)$?

SOL:

$E(X) = \mu = n * p = 4 \ldots (1)$

Var(X)=$n * p * q = 2.4 \ldots (2)$

Substitute 1 in 2

$4 * q = 2.4 \rightarrow q = 0.6$ & $p = 1 - q = 1 - 0.6 = 0.4 \rightarrow p = 0.4$

Substitute $p = 0.4$ in 1

$n * 0.4 = 4 \rightarrow n = 10$

$\therefore$ $\boldsymbol{X \sim Bin(10, 0.4)}$

$p(X \geq 2) = 1 - p(X < 2) = 1 - p(X \leq 1)$

$= 1 - [(0.006) + (0.040)] = 1 - 0.046 = 0.954.$

NOTES:

1. $\mathbf{p(X < k) = (X < k - 1)}$
2. $\mathbf{(X > k) = 1 - p(X \leq k)}$
3. $\mathbf{p(X \geq k) = 1 - p(X < k) = 1 - p(X \leq k)}$

Example: **if $X \sim Bin(10, 0.20)$ then find $p(X \geq 1)$ (H.W)**

a) 0.107 b)0.915 c)0.791 d)0.893

Example: Let **$X \sim Bin(11, 0.50)$, then $p(3 < X < 8)$equals? (H.W)**

a) 0.774 b)0.85 c)0.61 d)0.89

Example: **A biased coin is weight such that the probability of obtaining a head is $\frac{4}{7}$. the coin tossed is 6 times and X denotes the number of heads observed. The value of the ratio $\frac{p(X=3)}{p(X=2)}$ =?**

a) 16 b) 9 c)16/9 d)9/16

Example: Let **$X \sim Bin(7, 0.7)$, find $p(X = 4)$ (using binomial distribution formula)**
Sol:

Example: In testing a new drug, researcher found that 20% of all patients using it will have a mild side effect. A random sample of 9 patients using this drug is selected. The probability that none will have this mild effect is:
Sol: (using binomial distribution formula)

Example: Let **$X \sim Bin\left(3, \frac{1}{3}\right)$, $then\ p(X = 3 \backslash X \geq 1)$ equals?** **(H.W +)**

4.3 Poisson Distribution

- In a binomial experiment, you are interested in the number of success in the given number of trials.
- In the Poisson distribution, you are interested in the number of specific events that occur randomly in a given time or interval.

For instance, the number of accidents per month, number of typos per page, number of patients per week.

Definition: The Poisson distribution is a discrete probability distribution of a random variable x that satisfies these conditions

1. The experiment consists of counting the number of times an event occurs in a given interval (time, area, volume)
2. The probability of the event is the same for each interval.
3. The number of occurrences in one interval is independent of the number of occurrences in other intervals.

The probability of exactly x occurrences in an interval is $P(x) = \frac{\mu^x e^{-\mu}}{x!}$

$X \sim poi(\mu)$

Where μ, is the expected value $E(X)$ or mean number of occurrences in the given interval.

e is the natural number = 2.71828…

The variance $(\sigma^2) = \mu$.

The standard deviation $\sigma = \sqrt{\mu}$

Note: A Poisson random variable can take only positive values. (In the Binomial distribution always has a finite upper limit).

Example: suppose that X is the number of deaths per month and if the average number per month is two, find the following:

1) $p(X = 0)$ 2) $p(X = 5)$ 3) $E(X)$ 4)Var(X) 5) Std(X)

Sol:

Average number per month → $X \sim poi(2)$

1. $p(X = 0) = \frac{e^{-2} * 2^0}{0!} = 0.135$
2. $p(X = 5) =$
3. $E(X) = 2$
4. Var$(X) =$
5. Std$(X) = \sqrt{\mu} =$

Example: The mean number of accidents per month at a certain intersection is 3. What is the probability that in any given month The number of the accidents is 4 at that intersection?

Sol: Average=3 → $X \sim poi(3)$

$$p(X=4) = \frac{e^{-3} * 3^4}{4!} = \frac{0.0496 * 81}{24} = \frac{4.0176}{24} = 0.167$$

❖ **Find the probability that more than 4 accidents will occur at that intersection.**

Sol:

$$p(X \geq 4) = 1 - p(X < 4) = 1 - [p(X=3) + p(X=2) + p(X=1) + p(X=0)]$$

▪ Using the Poisson Tables

Example: 1) If X follows a Poisson distribution with a mean value of 5, find P(x=4).

$$\mathbf{P(x=4) = 1755}$$

2) If X follows a Poisson distribution with a mean value of 6, find P(x=3).

$$\mathbf{P(x=3) = 0.0892}$$

3) If X follows a Poisson distribution and P(X=0)= 5 P(X=1).find P(X=0)

Sol: $P(X=0) 5 * P(X=1) \rightarrow \frac{e^{-\mu} * \mu^0}{0!} = 5 * \frac{e^{-\mu} * \mu^1}{1!} \rightarrow 1 = 5 * \mu \rightarrow \mu = 0.2$

$\rightarrow X \sim poi(0.2)$

$P(X=0) = \frac{e^{-0.2} * 0.2^0}{0!} = 0.8187$

Using the Poisson to approximate the Binomial

- Conditions of approximate:

1. (n) should be large $n \geq 30$, where n: the number of trials
2. P should be small → $X \sim Poi(\mu)$ where $\mu = n * p$

Example: let $X \sim Bin(100, 0.02)$, find the following

1. $p(x=3)$ 2. $p(2 \leq x \leq 6)$

Sol: $n = 100 \rightarrow n \geq 30$ & $p = 0.02 \rightarrow 0.1$

$X \sim Bin(100, 0.02) \rightarrow\rightarrow X \sim poi(100, 0.02) \rightarrow\rightarrow X \sim poi(2)$

1. $p(x=3) = 0.1804$
2. $p(2 \leq x \leq 6) = p(x=2) + p(x=3) + p(x=4) + p(x=5) + p(x=6)$

$$= 0.2707 + 0.1804 + 0.0902 + 0.0361 + 0.0120 = 0.5894$$

Lecture 13

Biostatistics

Lecturer: Naba Mohammed Dhiaa Alashqar

Chapter 5: Continuous probability distribution (will be discussed later after Chapter 6)

Chapter 6
Testing Statistical Hypothesis
(Level of significance, test statistic, and P-value)

In this chapter, we will learn how to test claims about a parameter by using a random sample.

Elements of hypothesis testing

Definition

- **A statistical hypothesis is a statement or a claim about a population parameter.**

- **A null hypothesis (H_0): is a statistical hypothesis that contains equality, such as ≤, = or ≥.**
- **Alternative hypothesis (H_a): is the complement of the null hypothesis (H_a is true if H_0 is false) and it contains a statement ≠, > or <.**

Verbal Statement H_0 *The mean is . . .*	Mathematical Statements	Verbal Statement H_a *The mean is . . .*
. . . greater than or equal to k. *. . . at least k.* *. . . not less than k.* *. . . not shorter than k.*		*. . . less than k.* *. . . below k.* *. . . fewer than k.* *. . . shorter than k.*
. . . less than or equal to k. *. . . at most k.* *. . . not more than k.* *. . . not longer than k.*		*. . . greater than k.* *. . . above k.* *. . . more than k.* *. . . longer than k.*
. . . equal to k. *. . . k.* *. . . exactly k.* *. . . the same as k.* *. . . not changed from k.*		*. . . not equal to k.* *. . . different from k.* *. . . not k.* *. . . different from k.* *. . . changed from k.*

Example: Write each claim as a mathematical statement. State the null and alternative hypotheses, and identify which represents the claim.

1. A school publicizes that the proportion of its students who are involved in at least one extracurricular activity is 61%.
2. A car dealership announces that the mean time for an oil change is less than 15 minutes.
3. A company advertises that the mean life of its furnaces is more than 18 years.

❖ When you make a hypothesis test, you make one of **two decisions**:

1. **Reject the null hypothesis (reject H_0)**

Or

2. **Fail to reject H_0 (Accepted H_0)**

Since your decision is based on random sample, there is a possibility for errors in your decision.

	H_0true/ H_1 false	**H_0 false / H_1 true**
Reject H_0	Type 1 error $\propto$: significance level	Correct decision or $1-\beta$ Power of the test
Accepted H_0	Correct decision	Type 2 error β

→ Type 1 error: it occurs when rejecting $\mathbf{H_0}$ while $\mathbf{H_0}$ is true.

- Probability of type (1) error is called $\propto$ which is called the **significance level**.
- **$\propto$ takes one small values, such as $\propto = 0.01$, $\propto = 0.05$, $\propto = 0.1$, the most often values $\propto = 0.05$ and $\propto = 0.01$.**

→ Type 2 error: it occurs when accepting $\mathbf{H_0}$ while $\mathbf{H_0}$ is false.

- Probability of type (1) error is called β
- The Power of the test $= 1-\beta$ (It represents the probability of rejecting the null hypothesis when it is false).
-

❖ **Statistical Hypothesis Test (Testing the Hypothesis)**

It is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis.

The tests used for testing hypothesis are Z-test and T-test

The steps for the test of hypothesis of one population mean μ

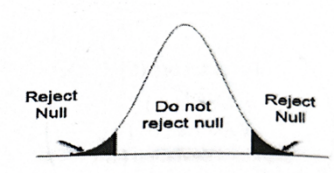
→ Hypothesis: population is equal to specified value when σ^2 (variance of population is known)

We have three cases of hypothesis

A. Two-tailed (two-sided) test. B. One-sided test $\mu > \mu_0$ **C.** One-sided test $\mu < \mu_0$

→

A. Two-tailed and the shaded region on the left and on the right represent the rejection region



1. State the experimental goal
2. State the hypothesis and alternatives:
   $\mathbf{H_0}$: $\boldsymbol{\mu = \mu_0}$
   $\mathbf{H_1}$: $\boldsymbol{\mu \neq \mu_0}$ $\boldsymbol{(\mu > \mu_0 \; or \; \mu < \mu_0)}$

3. **Decide on the level of significance (** probability of rejecting the hypothesis if it is true) $\boldsymbol{\propto = 0.05}$ **and** $\boldsymbol{\propto = 0.01}$ **.**
4. **Decide upon the statistic to be used for testing this hypothesis, in this case, (when** σ^2 is known), using Z-test statistic

$$\boldsymbol{z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}$$

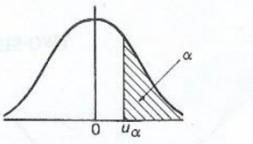
$\bar{X}$: mean of the sample
μ_0: population parameter
σ: standard deviation
n: sample size

5. **Find the sampling distribution of the statistic** Z-test **under the assumption that hypothesis is true**
   Note: that z is the value of the standard normal distribution variable Z which has mean=0 and variance = $\sigma^2 = 1$, i.e $Z \sim N(0,1)$.

6. **Decide a critical region, those values of the Statistic which would cause to reject the hypothesis. For this, we find in Table-1 the Values of $Z_{\propto/2}$. From the table, if $\propto = 0.05$, then the tabulated value**

Table-1 (PERCENTAGE POINTS OF THE STANDARDISED NORMAL DISTRIBUTION)

The $Z_\propto = U_\propto$ values tabulated, where Z ~ N(0,1)



α	u_α	α	u_α	α	u_α	α	u_α
0.50	0.00000	0.34	0.41246	0.18	0.91537	0.025	1.96000
0.49	0.02507	0.33	0.43991	0.17	0.95416	0.020	2.05375
0.48	0.05015	0.32	0.46770	0.16	0.99446	0.010	2.32635
0.47	0.07527	0.31	0.49585	0.15	1.03643	0.009	2.36562
0.46	0.10004	0.30	0.52440	0.14	1.08032	0.008	2.40891
0.45	0.12566	0.29	0.55338	0.13	1.12639	0.007	2.45726
0.44	0.15097	0.28	0.58284	0.12	1.17499	0.006	2.51214
0.43	0.17637	0.27	0.61281	0.11	1.22653	0.005	2.57583
0.42	0.20189	0.26	0.64335	0.10	1.28155	0.004	2.65207
0.41	0.22754	0.25	0.67449	0.09	1.34076	0.003	2.74778
0.40	0.25335	0.24	0.70630	0.08	1.40507	0.002	2.87816
0.39	0.27932	0.23	0.73885	0.07	1.47579	0.001	3.09023
0.38	0.30548	0.22	0.77219	0.06	1.55477	0.0005	3.29053
0.37	0.33185	0.21	0.80642	0.05	1.64485	0.0001	3.71902
0.36	0.35846	0.20	0.84162	0.04	1.75069	0.00005	3.89060
0.35	0.38532	0.19	0.87790	0.03	1.88079	0.00001	4.26489

if $\propto = 0.05 \rightarrow Z_{\propto/2} = Z_{0.05/2} = Z_{0.025} = 1.96$
and if $\propto = 0.01, \rightarrow Z_{\propto/2} = Z_{0.01/2} = Z_{0.005} = 2.58$

7. **Compute the value of the test statistic from the formula $z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$**
8. **Make a statement of acceptance or rejection of H_0 as follows**

   Reject H_0, if $|z| \geq Z_{\propto/2}$
   Accept H_0, if $|z| < Z_{\propto/2}$

9. **State the consequences of the experimental findings in light of the acceptance or rejection of the Statistical hypothesis.**

Example: A standard intelligence test. has been given for several years with an average score of 80 and standard deviation of 7. some modifications is made for this test which emphasizes on reading skill.
A group of 25 students is taught with the new change, the obtain a mean grade of 83 on the examination.
Is there reason to believe that the modification changes the results on the test?
Sol: we shall test the hypothesis that the population mean of students taught by the new method is 80
H_0: $\mu = 80$
H_1: $\mu \neq 80. (\mu > 80 \; or \; \mu < 80)$

Set **$\propto = 0.05$ (level of significance)**
Since the standard deviation of the population is known, $\sigma = 7$, then we use the Z-test

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}, \qquad \bar{X} = 83, \mu_0 = 80, \sigma = 7, n = 25$$

$$z = \frac{83 - 80}{7/\sqrt{25}} = \frac{3(5)}{7} = 2.14$$

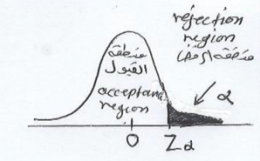
From Table 1, $Z_{\propto/2} = Z_{0.05/2} = Z_{0.025} = 1.96$
We note that $z > Z_{0.025}$
i.e the calculated z greater than the tabulated value $Z_{0.025}$, so we reject the hypothesis H_0 → higher significance

→ B. Hypothesis: population mean is equal to specified value when σ^2 is known
One-sided test (**right-sided**), alternative $\mu > \mu_0$
Oure procedure for testing a hypothesis of this sort as follows

1. State the experimental goal
2. State the hypothesis and alternatives:
   H_0: $\mu \leq \mu_0$
   H_1: $\mu > \mu_0$

3. Choose $\propto$
4. Use $\boldsymbol{z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}}$ **as the test statistic**
5. **Assume that the sampling distribution of Z is N(0,1) (normal with mean 0 and variance 1)**
6. **The critical region will be $z > Z_\propto$ (right-sided)**
7. Compute the value of $\boldsymbol{z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}}$ **and take the decision as**

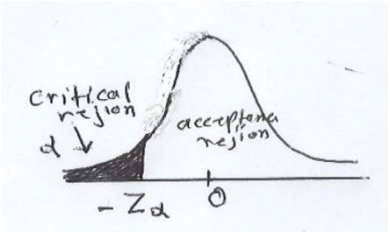
   Reject H_0 if $z \geq Z_\propto$
   Accept H_0 if $z < Z_\propto$



8. State the statistical conclusion

→ C. Hypothesis: population mean is equal to specified value when σ^2 is known
One-sided test(on the left side),
$\mathbf{H_0}: \mu \geq \boldsymbol{\mu_0}$
$\mathbf{H_1}: \mu < \boldsymbol{\mu_0}$

The procedure for testing this hypothesis is the same as in the previous case except that the critical region will be on the left side of the normal curve, see figure



The decision will be as follows
Reject H_0, if $|z| \geq Z_\propto$
Accept H_0, if $|z| < Z_\propto$

Example(1): In a hypothesis test for a proportion, the null hypothesis is $H_0: \mu \leq 0.7$ and he alternative hypothesis is $H_1: \mu > 0.7$ with $\propto = 0.05$. The test statistic is Z cal= 1.68, What is your decision in terms of H_0?

a. **Reject H_0** b. Accept **H_0** c. **Reject H_1** d. Accept both

Sol:
It is one-tailed test $\propto = 0.05 \rightarrow Z_{0.05} = 1.645$
Since 1.68>1.645, we reject **H_0** → **a**

Example(2): if $H_0: \mu = 1.5$ and $H_1: \mu > 1.5$, $\propto = 0.01$ and $\sigma = 0.2$. we conclude that

a. **Z-table= 1.645, H_0** should be rejected.
b. **Z-table= 1.645, H_0** should be **accepted.**
c. **Z-table= 2.326, H_0** should be rejected.
d. **Z-table= 2.326, H_0** should be **accepted**

sol:

$H_0: \mu = 1.5$ and $H_1: \mu > 1.5$

σ is known → use Z-test

$z = \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{1.53-1.5}{0.2/\sqrt{100}} = 1.5$

It is one tailed test (Right tailed)

$\propto = 0.01 \rightarrow Z_{0.01} = 2.33$

We can't reject $H_0 \rightarrow d$

Example(3): We want to test $H_0 : \mu = 60$ Vs $H_1: \mu \neq 60$ using a random sample of size 25, selected from a population of $\propto = 0.01$. In such a case we reject H_0 if the value of the test statistic is :

a. Between -2.57 and 2.57 b. Between -1.96 and 1.96

c. Smaller than -2.57 and Bigger than 2.57. d. Smaller than -1.96 and Bigger than 1.96

sol:

it is two-tailed test $\propto = 0.01 \rightarrow \frac{\propto}{2} = 0.005 \rightarrow Z_{0.005} = \pm 2.57 \rightarrow c$

Example(4): type II error occurs when (H.W)

A. We accept a true hypothesis. B. We reject a false hypothesis.

C. We accept a false hypothesis D. We reject a true hypothesis.

Example(5): in testing $H_0: \mu \geq \mu_0$ vs $H_1: \mu < \mu_0$, if the null hypothesis is not rejected when the alternative hypothesis is true

A. A Type 1 error is committed B. A Type 2 error is committed

C. A power of the test is committed D. A two-tailed test is made.

Example(6): Test the claim about the population mean μ at the level of significance $\propto$. Assume the population is normally distributed.

Claim $\mu \neq 40$; $\propto = 0.05$; $\sigma = 1.97$ Sample statistics $\bar{X} = 39.2$, n=25.

Sol: (H.W)