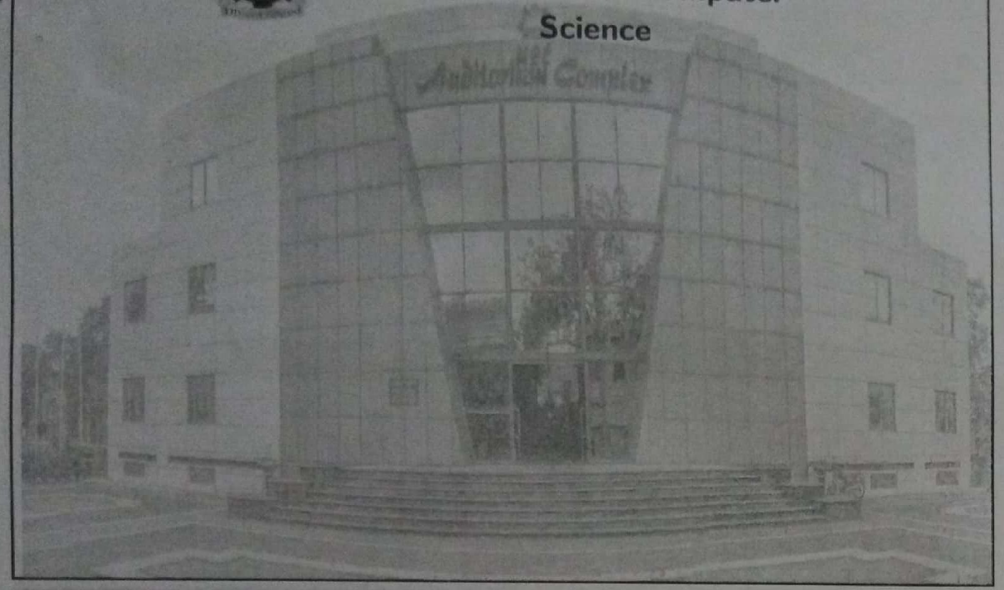
University of Engineering and Technology, Lahore
Department of Computer Science



An introduction to Statistics

What is Statistics?

Statistics is a branch of mathematics dealing with the collection, organization, analysis, interpretation, and presentation of data.

- The word Statistics is derived from the Latin word Status. It has various applications in almost every field of life. The use of Statistical methods in manufacturing, development of food products, Computer software, Pharmaceuticals and many other areas involve to gather information or scientific data.
- For example, Suppose we have selected a random sample of 100 students from a school out of 1000 students, The average height of the sampled students would be an example of a Statistic. Likewise, result grade of students, education rate, religion, colour etc are also some examples of Statistic.

Use of Statistical Information

Statistical Information can be used for variety of purposes. Some are listed as below:

- to inform general public;
- to explain things that have happened;
- to justify a claim;
- to provide general comparison;
- to predict the decisions regarding future outcomes;
- to estimate the unknown quantities;
- to establish relationship between factors;

Basic Definitions

Statistics

Statistics is the science that involves methods which are used to collect, process and analyze numerical data and draw conclusions in the face of uncertainty.

Statistics can be divided into two main areas:

1. **Descriptive Statistics**
2. **Inferential Statistics**

Descriptive Statistics

Descriptive Statistics consists of the collection, organization, summarization, and presentation of data.

In descriptive statistics the statistician tries to describe a situation. Consider the national census conducted by the government every 10 years. Results of this census give the average age, income, and other characteristics of the population.

Scanned with CamScanner

Basic Definitions

To obtain this information, the Census Bureau must have some means to collect relevant data. Once data are collected, the bureau must organize and summarize them. Finally, the bureau needs a means of presenting the data in some meaningful form, such as charts, graphs, or tables.

Inferential Statistics

Inferential Statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

The statistician tries to make inferences from samples to populations. Inferential statistics uses probability, i.e., the chance of an event occurring. If you play cards, dice, bingo, or lotteries, you win or lose according to the laws of probability. Probability theory is also used in the insurance industry and other areas.

Basic Definitions

It is important to distinguish between a sample and a population.

Population

A population consists of all subjects (human or otherwise) that are being studied. The number of observations in a population is called population size and it is denoted by **N**.

Most of the time, due to the expense, time, size of population, medical concerns, etc., it is not possible to use the entire population for a statistical study; therefore, researchers use samples.

Sample

A sample is a group of subjects selected from a population. The number of observations in a sample is called sample size and it is denoted by **n**.

If the subjects of a sample are properly selected, most of the time they should possess the same or similar characteristics as the subjects in the population.

Variables and Classification of Variables

statisticians gain information about a particular situation by collecting data for random variables. This section will explore in detail the nature of variables and types of data.

Variable

A variable is a characteristics that varies over time or for different individuals or objects under consideration. e.g., income, heights, age, colour, religion etc.

Variables can be classified as quantitative or qualitative:

Quantitative variables

A variable is called quantitative when a characteristic can be expressed numerically and can be ordered or ranked. For example, income, heights, age, body temperatures, weights etc.

Variables and Classification of Variables

Quantitative variables can be further classified into two groups:

1. **Discrete Variable**
2. **Continuous Variable**

Discrete Variable A discrete variable can assume only a finite or countable values such as number of children in a family, the number of students in a classroom, number of rooms in a house.

Continuous Variable A continuous variable can assume infinite number of values in an interval for example, heights, age, body temperatures, weights etc.

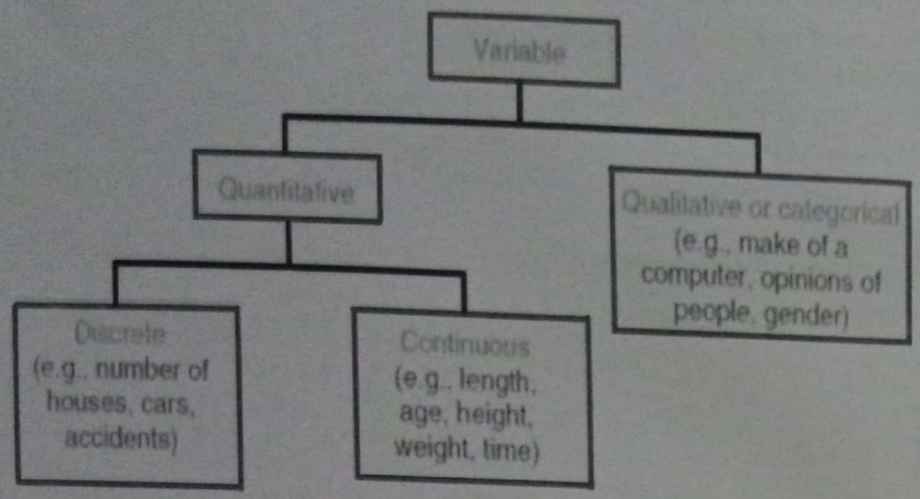
Qualitative Variable

A variable is qualitative when a characteristic is non-numerical such as colour, religion, education, property etc. A qualitative variable is also called an attribute.

Scanned with CamScanner

Variables and Classification of Variables

The classification of variables can be summarized as follows:



Data and Types of Data

Data

Data is a set of values or information of quantitative or qualitative variables about one or more person or objectives.

There are two types of data:

1. **Primary Data**
2. **Secondary Data**

Primary Data

Data that have been originally collected and have not undergone through any sort of statistical treatment is called primary data.

Secondary Data

Data that have undergone through some sort of statistical treatment by statistical method i.e., organized, tabulated or presented in some form for a certain purpose is called secondary data.

Data

DATA Facts or figures, which are numerical or otherwise, collected with a definite purpose are called data.

Types Of Data

Quantitative Data	Qualitative Data
These represent numerical value.	These represent some characteristics or attributes.
These can be numerically computed.	These depict descriptions that may be observed but cannot be computed.

Primary Data	Secondary Data
Data collected for first time.	Data that is sourced by someone other than the user.

Discrete Data	Continuous Data
These are the data that can take only specific value.	These are the data that can take values from a given range.

Classification and Presentation of Data

1. Quantitative Data
2. Qualitative Data

Quantitative Data

Here are some methods to arrange quantitative data:

- Stem and Leaf Display (Stem Plot)
- Dot Plots
- Frequency Distribution
- Cumulative Frequency

(a) Stem and Leaf Display (Stem Plot)

In this method the data is divided in to two groups i.e., a Stem and a Leaf. The stem is a leading digit and a leaf is a trailing digit.

Scanned with CamScanner

Stem Plot

Example

The ages of 30 patients admitted to a certain hospital during a particular week were as follows:

48, 31, 54, 37, 18, 64, 61, 43, 40, 71, 51, 12, 52, 65, 53, 42, 39, 62, 74, 48, 29, 67, 30, 68, 35,57, 26, 27, 58, 49

Stem	Leaf					
1	2	8				
2	6	7	9			
3	0	1	5	7	9	
4	0	2	3	8	8	9
5	1	2	3	4	7	8
6	1	2	4	5	7	8
7	1	4				

Back to Back Stem Plot

Stem and leaf display can be to compare two sets of data showing together on back to back stem plot.

Example

The following data are the marks of students in an examination.

Mathematics

75, 69, 58, 46, 44, 32, 50, 53, 47, 58, 78, 81, 61, 45, 31, 44, 53, 66, 57

Statistics

52, 58, 68, 77, 38, 85, 43, 44, 63, 69, 56, 65, 79, 44, 71, 84, 72, 72, 79

Leaf(Maths)						Stem	Leaf(Stats)					
				2	1	3	8					
	7	6	5	4	4	4	3	4	4			
8	8	7	3	3	0	5	2	6	8			
		9	6	1	1	6	3	5	5	8	9	
				8	5	7	1	2	2	7	9	9
					1	8	4	5				

Stem-and-Leaf Plot

Example

To illustrate the construction of a stem-and-leaf plot, consider the data of Table given below which specifies the "life" of 40 similar car batteries recorded to the nearest tenth of a year.

Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Stem-and-Leaf Plot of Battery Life

Stem	Leaf	Frequency
1	69	2
2	25669	5
3	00111122233344455677788 99	25
4	11234577	8

(b) Dot Plot

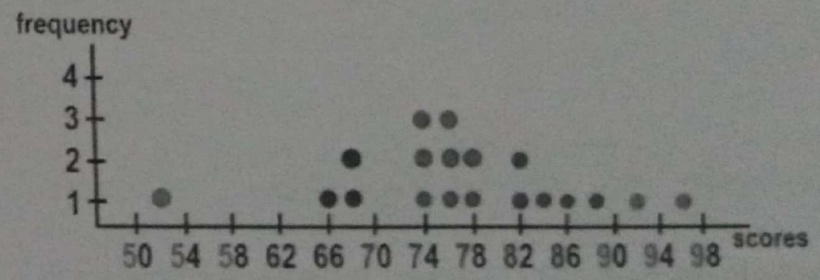
Example

The following data are the marks of students in an examination.

The Dot Plot

19 Exam Score

76 86 74 84 82 68 96 78 66
74 76 92 78 82 74 76 52 88 68



Scanned with CamScanner

Grouped and Ungrouped Data

Ungrouped Data

Ungrouped data which is also known as raw data is data that has not been placed in any group or category after collection. Data is categorized in numbers or characteristics therefore, the data which has not been put in any of the categories is ungrouped.

Grouped Data

When the number of observations in the data are very large then the data has been placed in a group or category after collection is called Grouped Data

Grouped Data verses Ungrouped Data

Grouped Data
VERSUS
Ungrouped Data

Differences	Grouped	Ungrouped
Classification	Organized into classes	No form of organization
Preference	Preferred when analyzing data	Preferred when collecting data
Accuracy	Has higher accuracy levels when calculating mean and median	Less accurate in determining mean and median
Presentation	Frequency tables are mostly used	Lists are used in the data type
Summary	Summarized in frequency distribution	No form of summarization

Summary of Grouped Data verses Ungrouped Data

- In statistics, the term data is used to refer to information that has been collected and recorded for the purpose of specific projects and it could be either qualitative or quantitative.
- Both grouped and ungrouped data are types of data however, grouped data has been classified into categories based on similar characteristics whereas ungrouped data is raw data.
- Both types of data can be represented by frequency tables. However, for grouped data, there are no class limits thus the use of tally marks. Grouped data in a frequency table has limits and that is the upper class limit and lower class limit.
- Both types of data can be used to calculate the mean, mode and median of samples of population therefore they are useful.

(c) Frequency Distribution

A frequency distribution is the organization of raw data in table form, with number of times each value occurs.

Example 1: Frequency Distribution for Ungrouped Data

The following data given below is the number of children born to 50 families

Note: When the data are in original form, they are called raw data and are listed next.

2	4	5	1	7	6	3	3	4	5	1	3	3	2
6	5	0	6	4	5	4	5	3	4	3	3	2	3
4	2	3	1	2	6	3	8	4	2	8	9	3	3
7	10	2	1	3	3	7	2						

Scanned with CamScanner

Frequency Distribution

No. of Children	Tally Marks	No. of Families
0	I	1
1	IIII	4
2	~~IIII~~ III	8
3	~~IIII~~ ~~IIII~~ IIII	14
4	~~IIII~~ II	7
5	~~IIII~~	5
6	IIII	4
7	III	3
8	II	2
9	I	1
10	I	1
		Total = 50

Grouped Frequency Distribution

Class Limits

The class limits are the numbers which describes the classes; the smallest number is called lower class limit(LCL) and the larger number is called upper class limit(UCL). e.g.,

10-14, 15-19, 20-24 etc

In 10-14 lower class limit(LCL) is 10 and upper class limit(UCL) is 14.
In 15-19 lower class limit(LCL) is 15 and upper class limit(UCL) is 19.

Class Width

The class width is the difference between class boundaries and it denoted by **h**.

h = UCB-LCB
= 14.5-9.5
= 5

Grouped Frequency Distribution

Class Boundaries

The class boundaries are the precise numbers which separate one class from another.e.g., 9.5-14.5, 14.5-19.5, 19.5-24.5 etc

In 9.5-14.5 lower class boundary(LCB) is 9.5 and upper class boundary(UCB) is 14.5

In 14.5-19.5 lower class boundary(LCB) is 14.5 and upper class boundary(UCB) is 19.5

Note:

9.5– < 14.5	or	$9.5 \leq x < 14.5$
14.5– < 19.5		$14.5 \leq x < 19.5$
19.5– < 24.5 etc		$19.5 \leq x < 24.5$ etc

Grouped Frequency Distribution

$$LCB = LCL - \frac{D}{2}$$

In 44-48

$$\frac{D}{2} = \frac{1}{2} = 0.5$$

LCB=44-0.5=43.5

and

$$UCB = UCL + \frac{D}{2}$$

$$\frac{D}{2} = \frac{1}{2} = 0.5$$

LCB=48+0.5= 48.5

where D is the difference between the LCL of the next class interval and the UCL of the given class interval.

Class limit	Class boundary
44 - 48	43.5 - 48.5
D=1 49 - 53	48.5 - 53.5
D=1 54 - 58	53.5 - 58.5
59 - 63	58.5 - 63.5
64 - 68	63.5 - 68.5
69 - 73	68.5 - 73.5

Scanned with CamScanner

Class Marks

A class mark is number which devides each class in to two equal parts. It is the mid point of the class interval or class boundaries. It is denoted by x.

$$X_m = \frac{\text{lower boundary} + \text{upper boundary}}{2} \qquad \frac{43.5 + 48.5}{2} = 46$$

or

$$X_m = \frac{\text{lower limit} + \text{upper limit}}{2} \qquad \frac{44+48}{2} = 46$$

Steps to construct Frequency Distribution

Step 1

Devide the number of classes

There is no hard and fast rules concerning the number of classes. Generally the larger number of observations in a data set, the more classes should be used.

- In general between 5 and 20 classes are used.
- No. of classes $\approx \sqrt{n}$ n=No. of observations.
- No. of classes $\approx 1 + 3.3 logn$

Steps to construct Frequency Distribution

Step 2

Range of variation

Find the range of data.

Range = Max. Value- Min. Value

Step 3

Determine width of class

The approximate width of the class is obtained by the formula:

Class width $= \frac{Range}{No.ofclasses}$

Steps to construct Frequency Distribution

Step 4

Determine the class limits

The first class starts with Min. value or any number less than this Min. value. Add class width repeatdly to find remaining class limits.

Step 5

Distribute the data into classes

Use tally marks to distribute the data into classes.

Step 6

Find frequency of classes

Add the tally marks.

Scanned with CamScanner

Grouped Frequency Distribution

When the range of the data is large, the data must be grouped into classes that are more than one unit in width, in what is called a grouped frequency distribution.

Example 2: Frequency Distribution for Grouped Data

Suppose a researcher wished to do a study on the ages of the top 50 wealthiest people in the world.

49 57 38 73 81 74 59 76 65 69 54 56 69 68 78 65
85 49 69 61 48 81 68 37 43 78 82 43 64 67 52 56
81 77 79 85 40 85 59 80 60 71 57 61 69 61 83 90
87 74

Construction of Frequency Distribution

Information

n=50, Max. value=90, Min. value=37

Step 1:

No. of classes $\approx \sqrt{n} = \sqrt{50} \approx 7.07$ or 7

Step 2

Range=Max. value- Min. value=90-37=53

Step 3

$$\text{Class width} = \frac{Range}{No.ofclasses}$$
$$= \frac{53}{7}$$
$$= 7.57 \approx 7$$

Frequency Distribution

Class limits	Tally	Frequency
35–41	///	3
42–48	///	3
49–55	////	4
56–62	~~////~~ ~~////~~	10
63–69	~~////~~ ~~////~~	10
70–76	~~////~~	5
77–83	~~////~~ ~~////~~	10
84–90	~~////~~	5
	Total	50

Frequency Distribution Table

Frequency Distribution Table A list, table or graph that displays the frequency of various outcomes in a sample of data.

Frequency Distribution Table

Ungrouped

It is used for small data set. For eg.

Marks Obtained	Frequency
16	3
17	4
18	8
19	10
20	12
21	6
22	3

Grouped

It is used for large data set. For eg.

Class Interval	Frequency
0-5	3
5-10	11
10-16	14
15-20	2

Scanned with CamScanner

Example 3:

These data represent the record high temperatures in degrees Fahrenheit (°F) for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Construction of Frequency Distribution

Information

n=50, Max. value=134, Min. value=100

Step 1:

No. of classes $\approx \sqrt{n} = \sqrt{50} \approx 7.07$ or 7

Step 2

Range=Max. value- Min. value=134-100=34

Step 3

$$\text{Class width} = \frac{Range}{No.ofclasses}$$

$$= \frac{34}{7}$$

$$= 4.85 \approx 5$$

Construction of Frequency Distribution

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	~~////~~ ///	8
110–114	109.5–114.5	~~////~~ ~~////~~ ~~////~~ ///	18
115–119	114.5–119.5	~~////~~ ~~////~~ ///	13
120–124	119.5–124.5	~~////~~ //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			$n = \Sigma f = 50$

Example 4:

The weights recorded to nearest grams of 60 apples picked out at random,

106	111	100	98	148	123	107	92
186	110	90	115	76	86	84	78
107	90	82	70	99	185	181	158
109	126	113	162	131	118	107	68
204	178	75	125	115	130	111	140
184	128	187	139	173	110	194	95
119	123	146	80	82	93	129	141
152	104	115	136				

Scanned with CamScanner

Construction of Frequency Distribution

Information

n=60, Max. value=204, Min. value=68

Step 1:

No. of classes $\approx \sqrt{n} = \sqrt{60} \approx 7.74$ or 7

Step 2

Range=Max. value- Min. value=204-68=136

Step 3

$$\text{Class width} = \frac{Range}{No.ofclasses}$$
$$= \frac{136}{7}$$
$$= 19.42 \approx 20$$

Construction of Frequency Distribution

Weights	Class Boundaries	Tally Marks	Frequency	Class Mark
65-84	64.5 - 84.5	𝍸 ||||	9	74.5
85-104	84.5 - 104.5	𝍸 𝍸	10	94.5
105-124	104.5 - 124.5	𝍸 𝍸 𝍸 ||	17	114.5
125-144	124.5 - 144.5	𝍸 𝍸	10	134.5
145-164	144.5 - 164.5	𝍸	5	154.5
165-184	164.5 - 184.5	||||	4	174.5
185-204	184.5 - 204.5	𝍸	5	194.5

(d) Cumulative Frequency Distribution

A cumulative frequency distribution is a distribution that shows the number of data values less than or equal to a specific value (usually an upper boundary). The values are found by adding the frequencies of the classes less than or equal to the upper class boundary of a specific class. This gives an ascending cumulative frequency.

Consider the Frequency Distribution of Example 4.

Upper Class Boundary(weights)	Cumulative Frequency(F)
Less than 84.5	9
Less than 104.5	$9 + 10 = 19$
Less than 124.5	$19 + 17 = 36$
Less than 144.5	$36 + 10 = 46$
Less than 164.5	$46 + 5 = 51$
Less than 184.5	$51 + 4 = 55$
Less than 204.5	$55 + 5 = 60$

The cumulative frequency of the last class is the sum of all frequencies in

Relative and percentage Frequencies

The relative frequency of a value is the fraction or proportion of times the value occurs.

$$\text{Relative Frequency} = \frac{No.\ of\ times\ the\ value\ occurs}{No.\ of\ observations\ in\ the\ data}$$

Consider the Frequency Distribution of Example 1.

No. of children	f	Relative Frequency	% Frequency
0	1	$\frac{1}{50} = 0.02$	$\frac{1}{50} \times 100 = 2\ \%$
1	4	$\frac{4}{50} = 0.08$	$\frac{4}{50} \times 100 = 8\%$
2	8	$\frac{8}{50} = 0.16$	$\frac{8}{50} \times 100 = 16\%$
3	14	$\frac{14}{50} = 0.28$	$\frac{14}{50} \times 100 = 28\%$
4	7	$\frac{7}{50} = 0.14$	$\frac{7}{50} \times 100 = 14\%$
5	5	$\frac{5}{50} = 0.1$	$\frac{5}{50} \times 100 = 10\%$
6	4	$\frac{4}{50} = 0.08$	$\frac{4}{50} \times 100 = 8\%$
7	3	$\frac{3}{50} = 0.06$	$\frac{3}{50} \times 100 = 6\%$
8	2	$\frac{2}{50} = 0.04$	$\frac{2}{50} \times 100 = 4\%$
9	1	$\frac{1}{50} = 0.02$	$\frac{1}{50} \times 100 = 2\%$
10	1	$\frac{1}{50} = 0.02$	$\frac{1}{50} \times 100 = 2\%$

Scanned with CamScanner

Relative and percentage Frequencies

The relative frequency of a value is the fraction or proportion of times the value occurs.

Relative Frequency $= \frac{\text{No. of times the value occurs}}{\text{No. of observations in the data}}$

Consider the Frequency Distribution of Example 1.

No. of children	f	Cumulative Frequency	% Frequency
0	1	1	$\frac{1}{50} \times 100 = 2\ \%$
1	4	5	$\frac{5}{50} \times 100 = 10\%$
2	8	13	$\frac{13}{50} \times 100 = 26\%$
3	14	27	$\frac{27}{50} \times 100 = 54\%$
4	7	34	$\frac{34}{50} \times 100 = 68\%$
5	5	39	$\frac{39}{50} \times 100 = 78\%$
6	4	43	$\frac{43}{50} \times 100 = 86\%$
7	3	46	$\frac{46}{50} \times 100 = 92\%$
8	2	48	$\frac{48}{50} \times 100 = 96\%$
9	1	49	$\frac{49}{50} \times 100 = 98\%$
10	1	50	$\frac{50}{50} \times 100 = 100\%$

Construction of Frequency Distribution

Construct a Frequency Distribution for the data given below. Indicate clearly the class limits, class boundaries, class marks, cumulative and relative frequencies.

41.78	29.32	31.47	35.35	32.82	39.42
61.65	28.31	44.63	22.78	44.44	48.12
81.71	33.47	50.35	29.19	51.26	50.32
26.84	18.95	48.19	43.72	43.89	47.15
60.20	44.43	41.17	37.50	22.35	29.17

Scanned with CamScanner