

Statistics Summary

(prepared by Xuan (Tappy) He)

Statistics is the practice of collecting and analyzing data. The analysis of statistics is important for decision making in events where there are uncertainties.

Example 1: Consider the following table which records the marks of students on a test out of 6.

Name	Mohammad	Mary	Mamadou	Binta	Alieu	Lamin
Mark	5	3	6	3	4	3

Measures of Central Tendency:

- A good way to begin analyzing data is to summarize the data into a single representative value.
- The three most common measures of central tendency are **mean**, **median** and **mode**.

In order to analyze a set of data using measures of central tendency, we must know:

- The number of entries in the data set and,
- The value of each entry.

Mean:

- Definition: Another word for average.

$$\text{Mean} = \frac{\text{Sum of Value of Entries}}{\text{Number of Entries}}$$

- Formula:
- In Example 1, the mean is $\frac{5 + 3 + 6 + 3 + 4 + 3}{6} = \frac{24}{6} = 4$

Median:

- Definition: The median is the value that separates the higher half of a sample from the lower half.
- Formula: Arrange all of the values from lowest to highest. If there are an odd number of entries, the median is the middle value. If there are an even number of entries, the median is the mean of the two middle entries.
- In Example 1, the numbers can be reordered as 3, 3, 3, 4, 5, 6, and the median is $\frac{3+4}{2} = \frac{7}{2} = 3.5$

Mode:

- Definition: The mode is the most frequently occurring value in the data set.
- In a data set where each value occurs exactly once, there is no mode.
- In Example 1, since 3 occurred more frequently than any other value, the mode is 3.
-

Measures of Dispersion:

- Dispersion gives information about how spread out the values are in the data set.
- Common measures of dispersion are **range**, **standard deviation**, **mean deviation**, and **interquartile range**.

Range:

- Definition: The range in a data set measures the difference between the smallest entry value and the largest entry value.
- Formula: $\text{Range} = (\text{largest entry value} - \text{smallest entry value})$
- This is the simplest measure of dispersion.
- In Example 1, the range is $6-3=3$.

Mean Deviation:

- Definition: The mean deviation calculates the average difference of each entry from the mean. It is calculated by taking the absolute value of the difference between each entry and the mean of the data set, adding them together, then divided by the number of entries in the data set.
- Formula: Mean Deviation = $\frac{\text{Sum of } | \text{value of entry} - \text{mean of data set} |}{\text{Number of Entries}}$. (The vertical bars denote the absolute value).

- In Example 1, we have that Mohammad's mark is $|5-4|=1$ from the mean, Mary's mark is $|3-4|=1$ from the mean, 2 for Mamadou, 1 for Bintu, 0 for Alieu and 1 for Lamin. Thus the mean deviation

$$\text{is } \frac{1+1+2+1+0+1}{6} = \frac{6}{6} = 1$$

Standard Deviation:

- Definition: Standard deviation measures the variation or dispersion exists from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread over a large range of values.
- Formula: Standard Deviation = $\sqrt{\frac{\text{Sum of (value of entry - mean of data set)}^2}{\text{Number of Entries}}}$.
- In Example 1, the square of the difference between Mohammad's mark and the mean is $(5-4)^2=1$. For Mary, it is $(3-4)^2=(-1)^2=1$. For Mamadou, Bintu, Alieu and Lamin, they are 4, 1, 0, and 1 respectively. The standard deviation is $\sqrt{\frac{(5-4)^2 + (3-4)^2 + (6-4)^2 + (3-4)^2 + (4-4)^2 + (3-4)^2}{6}} = \sqrt{\frac{8}{6}} \approx 1.15$

Data presented in frequency tables

Large quantities of data are often presented in frequency tables. To do this we must have

- A common criteria,
- The common criteria divided into classes,
- The number of entries in each class.

For example, the heights of students in an elementary school class are presented in the following table. The common criteria is height, the classes are the different groups of heights, and the value of each class is the number of pupils in the group.

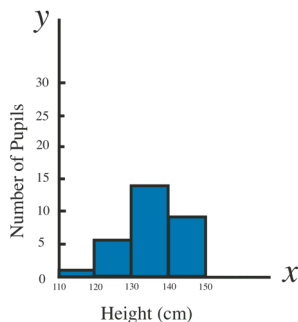
Example 2: Height of Pupils

Height	110cm-120cm	120cm-130cm	130cm-140cm	140cm-150cm
Number of pupils	1	6	14	9

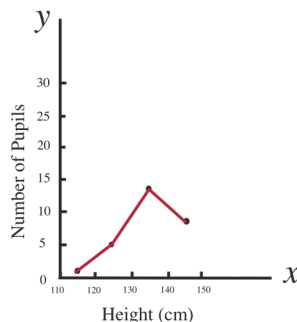
Histograms and Frequency Polygons:

- Graphical representations of the frequency distribution of a data set.
- A **histogram** is a bar graph whose x-axis labels the classes for the common criteria and y-axis labels the number of entries that satisfy each criteria.
- A **frequency polygon** is a line graph whose axis labels are the same as a histogram. However, after we mark down the points corresponding to our given data on the graph we connect the two points next to each other on the graph by a straight line. This usually results in a mountain-looking polygon.

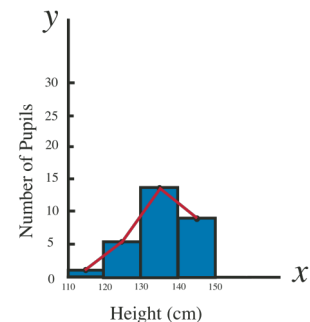
Histogram for the above example:



Frequency polygon for the above example:



A combined look:



Cumulative Frequency:

- Definition: The number of entries in a given class in a frequency table, plus all entries in earlier classes.
- In order to plot a cumulative frequency chart, we add another row or column to the frequency table that lists the cumulative frequency. See Example 3.
- We may draw a **Cumulative Frequency Curve** to represent the cumulative frequency information in the table.

Example 3:

Height	110cm-120cm	120cm-130cm	130cm-140cm	140cm-150cm
Number of pupils	1	6	14	9
Height	Under 120cm	Under 130cm	Under 140cm	Under 150cm
Cumulative number of pupils	1	7	21	30

Quartiles, Percentiles, Interquartile Range

Quartile:

- Definition: A quartile is each of the three points that divide a range of data into four equal groups.

Lower Quartile (Q1)

- Definition: The value such that 25% of all data entries have values less than this value.
- Formula: The entry at Q1 is calculated by $\frac{1}{4}(\text{total number of entries} + 1)$. Q1=Value of this entry.
- In Example 3, the height of the 7.75th person is the cut off height of the lower quartile, i.e. 135cm (assuming the height of pupils are arranged in ascending order)

Upper Quartile (Q3)

- Definition: The value such that 75% of all data entries are less than this value.
- Formula: The entry at Q3 is calculated by $\frac{3}{4}(\text{total number of entries} + 1)$. Q3=Value of this entry.
- In Example 3, the 23.25th person's height is the minimum height for the upper quartile, i.e. 145cm (again assuming the height of pupils are arranged in ascending order)

Median (Q2)

- Definition: The middle point that divides the middle two quartiles. It is the value such that 50% of all entries are less than this value. Recall that the median is the value of the middle entry out of all entries.
- Formula: refer to above measure of central tendency.
- In Example 3, the median is the average height of the 15th and 16th person, i.e. 135cm.

Interquartile Range (IQR)

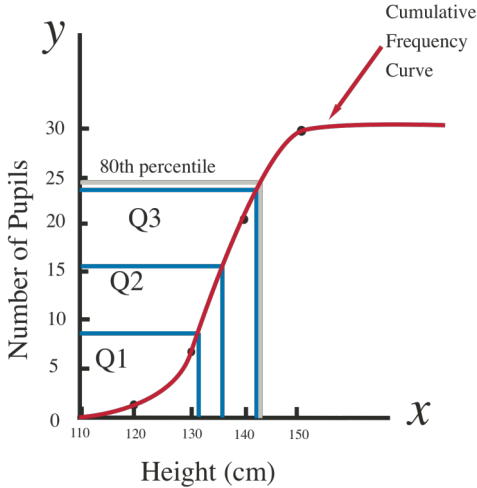
- Definition: A measure of dispersion. It is the difference between the lower quartile and upper quartile.
- Formula: $IQR = Q3 - Q1$
- In Example 3, we have that Q3 is 145cm (When each class of our criteria is given a range of values, we typically take the middle value of that class. In this example, 145 is the middle value of the 140-150cm height class); Q1 is 135cm, hence $IQR=145-135=10\text{cm}$.

Percentile

- Definition: The nth percentile of a data set is the value such that n% of all entries are less than this value.
- Formula: The nth percentile entry is calculated by: $n\%(\text{total number of entries} + 1)$. The value of this entry is the nth percentile.
- The percentile of a given value is determined by the percentage of the values that are smaller than that value.
- In Example 3, the 80th percentile is the height of the 24.8th person, approximately the 25th person when the heights are arranged in ascending order. That is, 145cm is the 80th percentile.

Note: We usually **round up** if the cut off for a certain quartile or percentile lands between two entries. (In the example above, we would round up the upper quartile, Q3, to the height of the 24th person.)

Estimating the median, quartiles and percentile from a Cumulative Frequency Curve



When we look at a cumulative frequency curve, the y-axis labels the number of entries for the set of data. The y-coordinate corresponding to the highest point on the curve is the total number of entries in the data set. Then if we take the middle of the y-axis between 0 and the value of the highest point, we have the median (Q2). Taking the value halfway between 0 and the median on the y-axis (i.e. the lower 25% of total number of entries) is the lower quartile (Q1). Taking the value halfway between the median and the value of the highest point on the y-axis is the upper quartile (Q3). To derive the n^{th} percentile, we use the same idea and take the value on the y-axis such that this value divided by the y-value of our highest point on the curve equals $n\%$.

Examples

1). This table shows the ages of students in a class.

Age (Years)	16	17	18
Number of Students	7	22	13

- How many students are in the class?
- What is the probability that a randomly chosen student in the class is 17 years old?
- What is the mean, median and modal age, and what is the range of the ages?

Solution:

a). Let n be the total number of students in the class. Then $n=7+22+13$. There are 42 students in the class.

b). There are 22 out of 42 total students who are aged 17. Hence, a randomly chosen student has the probability $\frac{22}{42} = \frac{11}{21}$, or approximately 52.3% of being 17 years old.

c).

(Mean) Remember that the mean is the sum of all values divided by the number of entries. We have 42 students as our number of entries (answer from part a).). Since there are 7 students of age 16, 22 students of age 17, 13 students of age 18, the sum of all values in our data set is $7 \times 16 + 22 \times 17 + 13 \times 18 = 112 + 374 + 234 = 720$. Therefore, the average age (i.e. mean) is $\frac{720}{42} \approx 17.15$ yrs old.

(Median) The median would be the age of the $\frac{(1 + 42)}{2} = 21.5^{\text{th}}$ student. (i.e. the average age between the 21st and 22nd student). Since there are 7 students aged 16, and 22 students aged 17, if we line these students up in ascending order of ages, the 21st student would be one of the students that is 17 years old, so would the 22nd student. Therefore, the median is just 17 years old.

(Mode) Out of the 3 possible ages, 17 is the age with the most number of people, therefore the mode is 17 years old.

(Range) Remember that the range is the highest entry value minus the lowest entry value. The highest entry value is 18 years old; the lowest entry value is 16 years old. Therefore the range is $18 - 16 = 2$.

2). Table shows the number of goals scored in a game of “football shootout” (each person gets 5 kicks at the goal) by students in a class.

Goal	0	1	2	3	4	5
Frequency	1	4	9	8	5	3

a). Calculate the mean and median.

b). A pie chart (or a circle graph) is a circular chart divided into sectors. Each of its sectors represents a class of the common criteria. The size of a sector illustrates the proportion of the number of entries of that class with respect to all entries in the data set. If we represent the information in a pie chart, what would be the sectorial angle for 4 goals? 2 goals?

Solution:

a). (Mean) Let n be the number of entries, then $n=1+4+9+8+5+3=30$. Remember that the mean is the sum of all values divided by total number of entries. Sum of all values is $1 \times 0 + 4 \times 1 + 9 \times 2 + 8 \times 3 + 5 \times 4 + 3 \times 5 = 81$. Therefore, the mean is $\frac{81}{30} = 2.7$.

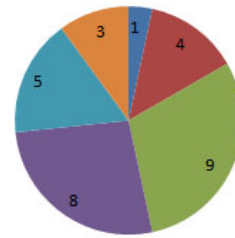
A student in the class scores on average 2.7 goals in the shootout.

(Median) Since there are 30 students, the median is the goal shot by the $\frac{(1+30)}{2} = 15.5^{\text{th}}$ student; i.e. the average of the 15th and 16th student. From the table, $1+4+9=15$, the 15th student is one of the 9 students that shot 2 goals, and the 16th student is one of 8 students that shot 3 goals. Hence the median is $\frac{2+3}{2} = 2.5$ goals.

b). (Approach 1) There are 5 students that scored 4 goals; there are 30 students in total. A full pie chart consists of all entries, and around a circle is 360° . Letting x be the sector angle, then $\frac{5}{30} = \frac{1}{6} = \frac{x}{360^\circ}$, so 60° of a full circle is the sectorial angle for 4 goals.

(Approach 2) We’ll take another approach for 2 goals. Since there are 30 students in total, and the pie is 360° . Then each student represents $\frac{360^\circ}{30} = 12^\circ$ of the pie’s sectorial angle. Since there are 9 students that scored 2 goals, we have $12^\circ \times 9 = 108^\circ$ as the sectorial angle for 2 goals.

Number of people per certain number of goals in football shootout



3). Sona, Karina, Omar, Mustafa and Amie earned scores of 6, 7, 3, 7, 2 on a standardized test respectively. Find the mean deviation and standard deviation of their scores.

Solution:

Mean deviation: We must first find the mean of the data set. The mean is $\frac{6+7+3+7+2}{5} = 5$. Then the mean deviation is calculated by $\frac{|6-5|+|7-5|+|3-5|+|7-5|+|2-5|}{5} = \frac{1+2+2+2+3}{5} = 2$. The mean deviation is 2.

Standard deviation: $\sqrt{\frac{(6-5)^2 + (7-5)^2 + (3-5)^2 + (7-5)^2 + (2-5)^2}{5}} = \sqrt{\frac{1+4+4+4+9}{5}} = \sqrt{\frac{22}{5}} \approx 2.1$

4). The following table contains the number of championships these clubs have won after the 2011 UEFA Champions League.

Club	Real Madrid	Liverpool	AC Milan	Manchester United	Internazionale	Bayern Munich	Barcelona
# of championships won	9	5	7	X	3	X+1	4

If these clubs won 5 championships over the years on average, how many championships did Bayern Munich win over the years?

Solution:

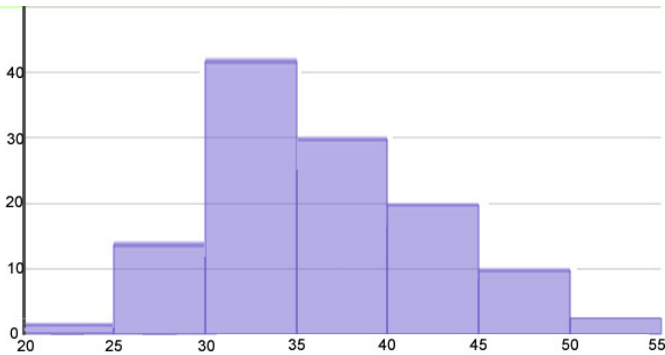
There are 7 clubs listed here, so the number of entries is 7. Now the average is 5 championships won per team over the years, then rearranging the formula for the mean yields: Average # of championships won \times 7 = Total # of championships won by all 7 teams. So the Total number of championships won by all 7 teams together is $5 \times 7 = 35$. Adding up the winnings from the other 6 teams, we've $9 + 5 + 7 + 3 + 4 = 28$. So Bayern Munich and Man.United combined must've won $35 - 28 = 7$ championships. We now have the equation $x + (x + 1) = 7$. Solving for x gives $x = 3$. So Man.United won 3 UEFA Championship league titles, then Bayern Munich won $x + 1 = 4$ tournaments over the years.

(We can verify by checking that $\frac{9 + 5 + 7 + 3 + 3 + 4 + 4}{7} = 5$.)

5). The table below shows the distribution of the ages, in year, of 120 members of a local football club.

Age (Years)	21-25	26-30	31-35	36-40	41-45	46-50	51-55
Number of people	1	14	42	30	20	10	3

- Draw a histogram for the above distribution
- Calculate the mean, median and mode of the distribution.
- Find the standard deviation of the distribution, correct to 2 decimal places.

Solution:

a). Remember that when we draw a histogram, gaps are closed between the bars. Hence, to make up for the difference of 1 between each age group, we extend the bars by 0.5 on both sides. (Example: the bar representing the 21-25 age group would expand from 20.5 to 25.5, the next bar representing 26-30 age group would extend from 25.5-30.5, etc.)

b). **(Mean)** We note that when we are given the classes of our common criteria (our common criteria is age, the classes are the age groups) in a range of values, we generally take the middle value of each class. For example, for the class of 21-25 years old, we take 23 as the age of the people in this age class. Then we have 1 person 23 years old, 14 people 28 years old, 42 people 33 years old, etc. We can now calculate the mean as

$$\frac{1(23) + 14(28) + 42(33) + 30(38) + 20(43) + 10(48) + 3(53)}{1 + 14 + 42 + 30 + 20 + 10 + 3} = \frac{4440}{120} = 37$$

(Median) There are 120 members in the club, so the median is the average age of the 60th and 61st member. Since $1 + 14 + 42 = 57 < 60$ and $1 + 14 + 42 + 30 = 87 > 61$, the 60th and the 61st person are all from the age group 36-40. Hence the median age is 38.

(Mode) The most commonly occurring age is the age group 31-35, whose members' age we assumed to be the middle value of the age group: 33. The mode is 33.

c). **Standard deviation:** The mean we calculated is 37. The standard deviation is calculated as

$$\sqrt{\frac{1(23 - 37)^2 + 14(28 - 37)^2 + 42(33 - 37)^2 + 30(38 - 37)^2 + 20(43 - 37)^2 + 10(48 - 37)^2 + 3(53 - 37)^2}{1 + 14 + 42 + 30 + 20 + 10 + 3}} = \sqrt{\frac{4730}{120}} \approx \sqrt{39.42} \approx 6.28$$

6). The frequency distribution shows the marks of 100 students in a Math test.

Marks	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
No. of Pupils	2	4	9	13	18	32	13	5	3	1

- Fill in another row of cumulative frequency in the table.
- Draw a cumulative frequency curve for the distribution.
- Use your curve to estimate:

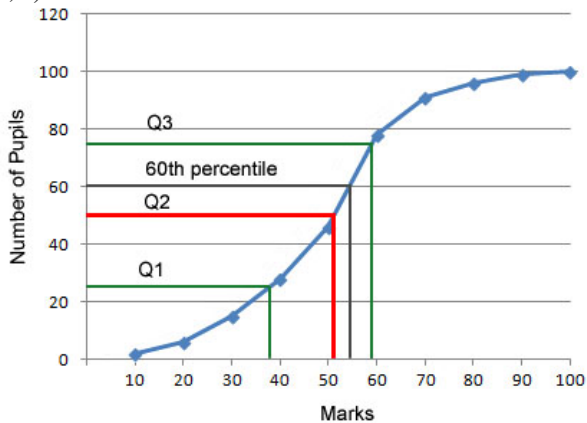
- i). The median;
- ii). Lower and upper quartile;
- iii). The 60th percentile;

Solutions:

a).

Marks	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
No. of Pupils	2	4	9	13	18	32	13	5	3	1
Cumulative Number of Pupils	2	6	15	28	46	78	91	96	99	100

b, c).



From the Cumulative Frequency Curve above, we see that the median (red line) lies approximately between 51 and 60. Hence we may conclude that (i) The median mark is approximately 55. The lower quartile (Q1, lower green line) lies approximately between 31 to 40, therefore (ii) The lower quartile mark is approximately 35 and the upper quartile, by observation is approximately 55. Lastly, we see that there are 100 students in this data set. So the 60th percentile is the mark that is above the lowest 60 students' marks, namely, the mark of the student with the 61st lowest mark. (iii) The 60th percentile mark, as illustrated by the gray line, is approximately 55.

Exercises

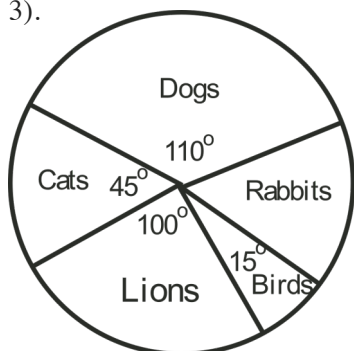
1). Which of the following is not a measure of central tendency? Mean, Median, Range, Mode.

2). The following table illustrates marks obtained by students in a test.

Marks	0	2	3	4	5
Frequency	2	2	8	X	2

- a). If the mean score of the class is 3, find the value of x.
- b). What is the median score? What is the modal score? What is the range of the scores?

3).



In a favorite animal survey of 120 students, each student is asked to identify his or her favorite animal. The data is displayed in the pie chart on the left.

- a). What percentage of students like rabbits as their favorite animal?
- b). How many students like cats as their favorite animal?
- c). How many more students like dogs as their favorite animal than rabbits as their favorite animal?

4). The average age of 5 women in a group is 27 years. If two other women aged 40 and 28 years join the group, find, in years, the new average age of the group of women.

5). The following table shows the frequency distribution of the weights of 25 children in a community.

Weight(kg)	14	15	16	17	18	19	20
# of Children	2	5	8	3	4	1	2

- Find the mode of the distribution
- Calculate the mean weight of the children
- What is the probability that a child selected at random weighs more than 17kg?
- Add a new row of cumulative frequency of children's weights.
- Use information in the question and your answer in part d). to plot a cumulative frequency curve.
- Using the cumulative frequency curve, determine the upper quartile weight of the children.

6). 10 different teams played football over the summer. After the summer, the top goal scorers from each team scored the following number of goals:

10 5 18 2 X 4 10 15 11 4

If the mean number of goals scored is 9, what is:

- The value of X?
- The mode?
- The median?
- The range?
- The mean deviation?
- The standard deviation?
- The 50th percentile?
- What is the percentile of the goal scorer with 11 goals scored?

7). The table below shows the marks scored by 30 students in a test.

Mark	5	6	7	8	9
Frequency	7	3	8	7	5

- Draw a frequency polygon to represent this information.
- A student is selected at random, what is the probability that he scored 7 or less points on this test?
- What is the median score?

8). In northern parts of the world, a 365-day year generally has temperature as follows:

Temperature (in degrees Celsius)	-39 to -30	-29 to -20	-19 to -10	-9 to 0	1 to 9	10 to 19	20 to 29	30 to 39
# of Days	15	45	60	40	45	60	65	35

- Draw a histogram to illustrate the above information.
- Add another row of cumulative frequency on the number of days of different temperature.
- Draw a cumulative frequency curve to illustrate the cumulative information.
- What percent of the year has temperature below 0 degrees?
- What percent of the year has temperature above 30 degrees?
- Using the cumulative frequency curve, determine the following:
 - The lower and upper quartile. Explain what they represent in context of the data.
 - The median temperature
 - The 30th percentile and 70th percentile. Explain what they represent in context of the data.