UNIT 9. STATISTICS

1. What is Statistics?

Every day we hear about data:

The percentage of unemployed people rose last month.

The most watched TV programme was "Aguila Roja".

The last September exam was passed by 65% of students.

Statistics studies data, or facts: it does surveys, it collects data, it draws graphs, it works out some values such as the mean and mode…

Solve the following exercises:

1. Write down three examples of a statistical datum.

Language corner

Data or Datum? /ˈdeɪtə ɔː ˈdeɪtəm/

The singular form is "datum", so we can say "That datum **is** very high".

"Data" is the plural, so we can say "the data **are** available", but it is also a **collection** of facts, so "the data **is** available" is OK, too.

2. Find some graphs of statistics that you have seen in a newspaper, cut and paste them here:

1.1 What are Data?

Data are a collection of facts, such as values or measurements. They can be numbers, words, measurements, observations or even just descriptions of things.

1.2. Qualitative versus Quantitative

Data can be qualitative or quantitative.

- **Qualitative data** is descriptive information (it *describes* something, like "high" or "fast").
- **Quantitative data**, is numerical information (numbers).

(To help you remember, think: "**Quant**itative is about **Quant**ity")

Quantitative data can be discrete or continuous:

- **Discrete data** can only take certain values (like whole numbers). These kinds of data are **counted**.

Example: *the number of students in a class (you can't have half a student).*

- **Continuous data** can take any value (within a range). These kinds of data are **measured**.

Examples:

- *A person's height: it could be any value (within the range of human heights), not just certain fixed heights,*
- *The time it takes to run a race; you could even measure it to fractions of a second,*
- *A dog's weight,*
- *The length of a leaf,*
- *And a lot more!*

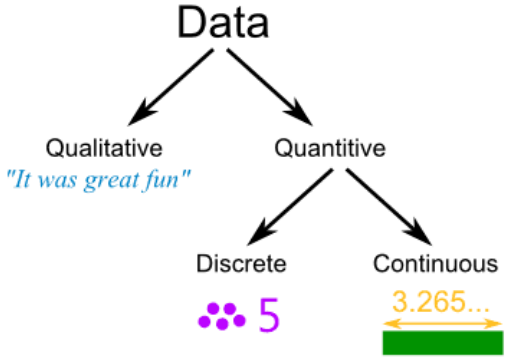
Example: *What can we know about this dog?*

***Qualitative**:*

- *He is brown and black*
- *He has long hair*
- *He has a lot of energy*

***Quantitative**:*

- *Discrete:*
  - *He has 4 legs*
  - *He has 2 brothers*
- *Continuous:*
  - *He weighs 25.5 kg*
  - *He is 565 mm tall*



Solve the following exercises:

3. Which of the following is a quantitative datum? (We are talking about a female dog)

a) She is black and white.
c) She has long hair.
b) She has two ears.
d) Her tail is quite long.

4. Which of the following is a continuous datum? (We are talking about a female cat)

a) She has two eyes.
c) She has five kittens.
b) She weighs 5.4 kg.
d) She has four paws.

5. Which of the following is a discrete datum? (We are talking about a female cat)

a) She is 54.4 cm long.
c) She is 22.3 cm high.
b) She has 30 teeth.
d) She weight 5.4 kgs.

2. Population, Sample or Census

Population is the set of individuals where we are going to carry out a statistical study.
A **Census** is the process of collecting data for **every** member of the group (the whole "**population**").
A **Sample** is the process of collectin data just for some **selected members** of the population.
Example:
There are 120 people in your local football club.
The population is all the members in your local football club, ie, 120 people.
You can ask everyone (all 120) what their age is. That is a census.
Or you can just ask the people that are in the club this afternoon. That is a sample.

A census is accurate, but hard to do. A sample is not as accurate, but it could be good enough, and it is much easier to do.

Choose the correct option:

6. A census collects information about:

a) All members of the population.
c) A large sample of the population.
b) All adult members of the population.
d) a small sample of the population.

7. A sample collects information about:

a) All member of the population.
c) All adult members of the population
b) None of the population.
d) Some, but not all, of the population.

3. Frequency distribution table.

After collecting the data we have to order them on a table.

3.1 Frequency, n_i

Frequency, n_i: is how often something occurs. Sometimes it is called *absolute frequency*.

Example: *Sam played football on*

- *Saturday morning,*
- *Saturday afternoon*
- *Thursday afternoon*

The frequency was twice on Saturday, once on Thursday and 3 times for the whole week.

By counting frequencies we can make a **Frequency distribution** table.

Example: *Goals*

Sam's team has scored the following number of goals in recent games:

2, 3, 1, 2, 1, 3, 2, 3, 4, 5, 4, 2, 2, 3

Sam put the numbers in order, then added up:

- *how often 1 occurs (twice),*
- *how often 2 occurs (5 times),*
- *etc,*

and wrote them down as a ***Frequency distribution*** *table:*

Scores:
1,1,2,2,2,2,2,3,3,3,3,4,4,5

Score	Frequency
1	2
2	5
3	4
4	2
5	1

From the table we can quickly see some interesting things, such as

- *scoring 2 goals happens most frequently*
- *they scored 5 goals only once*

This is the definition:

Frequency distribution: it consists in a list of values and their frequency (how often each value occurs).

Here is another example:

Example: *Newspapers*

These are the numbers of newspapers sold at a local shop over the last 10 days:

22, 20, 18, 23, 20, 25, 22, 20, 18, 20

Let us count how many of each number there are:

Papers Sold	***Frequency, n_i***
18	*2*
19	*0*
20	*4*
21	*0*
22	*2*
23	*1*
24	*0*
25	*1*

*It is also possible to **group** the values. Here they are grouped in 5s:*

Papers Sold	***Frequency, n_i***
15-19	*2*
20-24	*7*
25-29	*1*

The values are always in groups when the data are continuous.

Solve the following exercises:

8. Pi[1] is a special number that is used to find the area of a circle. The following number gives the first 100 digits of number pi:

π = 3.141 592 653 589 793 238 462 643 383 279 502 884 197 169 399 375 105 820 974 944 592 307 816 406 286 208 998 628 034 825 342 117 067

a) Which of the digits from 0 to 9 occurs most frequently in this number?

b) Make a frequency distribution table.

[1] Remember: pi is pronounced /paɪ/

3.2 Cumulative frequency, N_i.

Cumulative means "how much so far".

Think of the word "accumulate" which means to gather together.

To have cumulative frequency totals, just **add up the values as you go**.

Example: *Jamie has earned this much in the last 6 months:*

Month	***Earned***
March	*$120*
April	*$50*
May	*$110*
June	*$100*
July	*$50*
August	*$20*

To work out the cumulative frequency totals, just add up as you go.

The first line is easy, the total earned so far is the same as Jamie earned that month:

Month	***Earned***	***Cumulative***
March	*$120*	***$120***

But for April, the total earned so far is $120 + $50 = ***$170*** *:*

Month	***Earned***	***Cumulative***
March	*$120*	*$120*
April	*$50*	***$170***

And for May we continue to add up: $170 + $110 = ***$280***

Month	***Earned***	***Cumulative***
March	*$120*	*$120*
April	*$50*	*$170*
May	*$110*	***$280***

Do you see how we add the previous month's cumulative total to this month's earnings?

Here is the calculation for the rest:

- *June is $280 + $100 =* ***$380***
- *July is $380 + $50 =* ***$430***
- *August is $430 + $20 =* ***$450***

And this is the result

Month	***Earned***	***Cumulative***
March	*$120*	*$120*
April	*$50*	*$170*
May	*$110*	*$280*
June	*$100*	*$380*

July	*$50*	*$430*
August	*$20*	*$450*

The last cumulative total should match the total of all earnings:

***$450** is the last cumulative total ...*
... and it is also the total of all earnings:

*$120+$50+$110+$100+$50+$20 = **$450***

So we got it right.

*So that's how to do it, **add up as you go down the list** and you will have cumulative totals.*

You could also call it a "Running Total"

Solve the following exercises:

9. The table shows the daily sales made by a mobile ice cream salesman over one week.

a) Work out the cumulative frequency for every day.

Day	Sales ($)
Monday	51
Tuesday	63
Wednesday	37
Thursday	54
Friday	81
Saturday	123
Sunday	157

b) What was the value of the cumulative sales for Friday?

3.3 Relative Frequency, f_i

Relative frequency consists in how often something happens divided by all the outcomes.

Example: *if your team has won 9 matches out of a total of 12 matches played:*

- the *Frequency* of winning is 9
- the *Relative frequency* of winning is 9/12 = 0,75 (the maximum relative frequency is 1)

The relative frequency, f_i, can be easily changed into a percentage, you just need to multiply it by 100:

- 9/12 = 0,75 = 75% (the maximum relative frequency is 100%)

All the Relative frequencies add up to 1 (except when there are any rounding errors).

Example: *Travel Survey*

92 people were asked how they got to work:

- *35 used a car*
- *42 took public transport*
- *8 rode a bicycle*
- *7 walked*

The Relative frequencies (to 2 decimal places) are:

- *Car: 35/92 =* ***0.38***
- *Public Transport: 42/92 =* ***0.46***
- *Bicycle: 8/92 =* ***0.09***
- *Walking: 7/92 =* ***0.08***

$$0.38+0.46+0.09+0.08 = \mathbf{1.01}$$

(It would be exactly 1 if we had not rounded the relative frequencies to two decimal points),

3.4. Whole frequency distribution table.

The absolute frequency, n_i, the cumulate frequency, N_i, the relative frecuency, f_i, and the percentage can be written in the same table, and this is called a *whole frequency distribution table*.

Solve the following exercises:

9. 60 students took test. The frequency distribution is shown in the following table:

Mark	Frequency
0	1
1	3
2	6
3	9
4	8
5	11
6	8
7	7
8	4
9	1
10	2

a) Write down the whole frequency distribution table:

Mark, x_i	Frequency, n_i	Cumulative, N_i	Relative frequency, f_i	Percent, %
0				
1				
2				
3				
4				
5				
6				

7				
8				
9				
10				
Total				

b) How many students scored greater than or equal to 4, but less than or equal to 7?

c) Which percentage of student got greater than 8?

3.5 Frequency distribution on a continuous data.

Grouped Frequency Distribution

We have just seen in the previous example how we can group frequencies. This is very useful when the data have many different values.

Example: Leaves

Alex measured the lengths of a sample of leaves from an oak tree, and got these values (to the nearest cm):

9,16,13,7,8,4,18,10,17,18,9,12,5,9,9,16,1,8,17,1,10,5,9,11,15,6,14,9,1,12,5,16,4,16,8,15,14,17

How can we group these values?

Trying to work out the groups takes some practice, but here is a guide:

To start with, put the numbers in order, then find the smallest and largest values in your data, and ***calculate the range*** *(range = largest minus smallest). In order the lengths are:*

1,1,1,4,4,5,5,5,6,7,8,8,8,9,9,9,9,9,9,10,10,11,12,12,13,14,14,15,15,16,16,16,16,17,17,17,18,18

The smallest value (the "minimum") is 1 cm

The largest value (the "maximum") is 18 cm

The range is 18-1 = 17 cm

Now ***calculate*** *an approximate* ***group size****, by dividing the range by how many groups you would like.*

Then round that group size up to some simple value (like 2 instead of 1.83 or 5 instead of 4.26).

Let us say we want about 5 groups.

Divide the range by 5:

17/5 = 3.4

Then round that up to 4

Then ***choose a starting value*** *that is less than or equal to the smallest value. Try to make it a multiple of the group size if you can. In our case a start value of 0 makes sense.*

Next, ***calculate the list of groups****. (You must continue up to or past the largest value)*

Starting at 0 and with a group size of 4 we get: 0, 4, 8, 12, 16

Write down the groups, including the end value of each group (which must be less than the next group):

Length (cm)	Frequency
0-3	
4-7	
8-11	
12-15	
16-19	

The last group goes up to 19, which is greater than the largest value, so that is good.

(Note: If you don't like the groups, then go back and change the group size or starting value and try again.)

*After that**, include the Upper and Lower Values** for each Group. Even though Alex only measured in whole numbers, the data is continuous, so "4 cm" means that the actual value could have been anywhere from 3.5 cm to 4.5 cm. Alex just rounded the numbers to whole centimetres.*

Here are the groups with the Lower and Upper limits shown:

Length	Lower/Upper	Frequency
0-3 cm	0-3.5	
4-7 cm	3.5-7.5	
8-11 cm	7.5-11.5	
12-15 cm	11.5-15.5	
16-19 cm	15.5-19.5	

Finally, **tally and total**: tally (=make a table with) the results to find the frequencies, and do a total as well.

1,1,1,4,4,5,5,5,6,7,8,8,8,9,9,9,9,9,9,10,10,11,12,12,13,14,14,15,15,16,16,16,16,17,17,17,18,18:

Length	Lower/Upper	Frequency
0-3 cm	0-3.5	3
4-7 cm	3.5-7.5	7
8-11 cm	7.5-11.5	12
12-15 cm	11.5-15.5	7
16-19 cm	15.5-19.5	9
	Total:	38 leaves

Done!

Exercise.

10. Fill in the table above to get a whole frequency distribution table.

11. Make a whole frequency distribution table with the height of the students in your class.

4. Graphs.

4.1 Graphs for qualitative or quantitative discrete data.

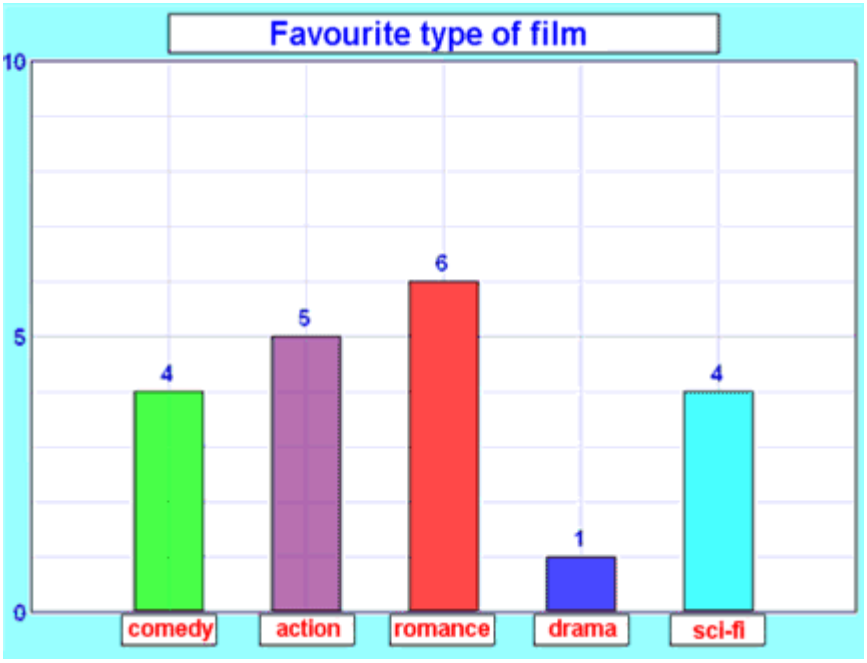
4.1.1 Bar Graphs.

A **Bar Graph** (also called Bar Chart) is a graphical display of data using bars of different heights.

Imagine you have just done a survey of your friends to find which kinds of films they prefer:

Table: Favorite Type of Movie				
Comedy	**Action**	**Romance**	**Drama**	**SciFi**
4	5	6	1	4

You could show those data on a bar graph like this:

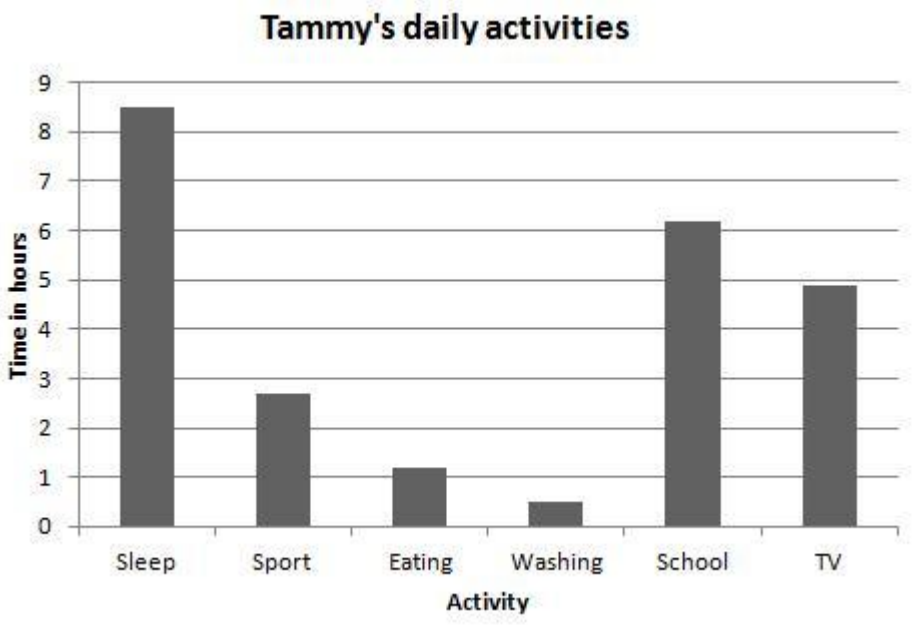


It is a really good way to show relative sizes: it is easy to see which types of film are most popular, and which are least popular, at a glance.

You can use bar graphs to show the relative sizes of many things, such as what type of car people have, how many customers a shop has on different days, and so on.

Solve the following exercises:

12. Tammy recorded the amount of time she spent on six activities over a twenty-four hour period and drew a bar graph, as follows:



a) Approximately how many more hours did she spend sleeping than watching TV?

b) Work out the approximate frequency distribution.

4.1.2 Pie Chart

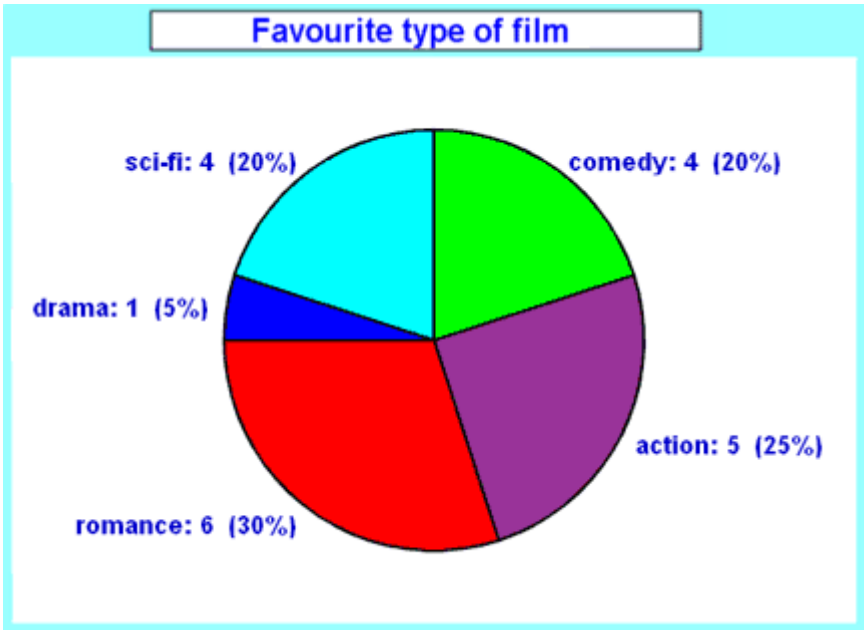
A **Pie Chart** is a special chart that uses "pie slices" to show relative sizes of data.

Imagine you have done a survey of your friends to find which kind of film they prefer.

Here are the results:

Table: Favorite Type of Movie				
Comedy	**Action**	**Romance**	**Drama**	**SciFi**
4	5	6	1	4

You could show that by means of this pie chart:



It is also a good way to show relative sizes: it is easy to see which film types are most liked, and which are least liked, at a glance.

How to Make Them Yourself

First, put your data into a table (like above), then add up all the values to get a total:

Comedy	**Action**	**Romance**	**Drama**	**SciFi**	**TOTAL**
4	5	6	1	4	**20**

Next, divide each value by the total and multiply by 100 to get a percent:

Comedy	**Action**	**Romance**	**Drama**	**SciFi**	**TOTAL**
4	5	6	1	4	**20**
4/20 = **20%**	5/20 = **25%**	6/20 = **30%**	1/20 = **5%**	4/20 = **20%**	**100%**

Now you need to figure out how many degrees you need for each "pie slice" (called a sector).

A full circle has 360 degrees, so we do this calculation:

Comedy	**Action**	**Romance**	**Drama**	**SciFi**	**TOTAL**
4	5	6	1	4	**20**

4/20 = **20%**	5/20 = **25%**	6/20 = **30%**	1/20 = **5%**	4/20 = **20%**	**100%**
4/20 × 360° = **72°**	5/20 × 360° = **90°**	6/20 × 360° = **108°**	1/20 × 360° = **18°**	4/20 × 360° = **72°**	**360°**

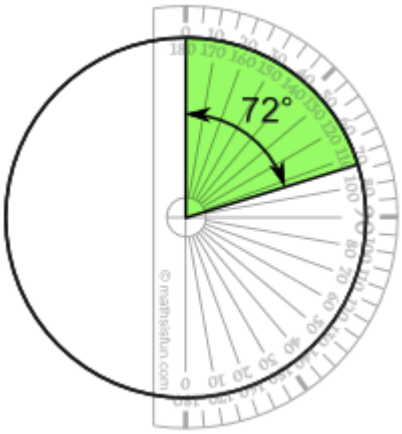
Now you are ready to start drawing!

Draw a circle.

Then use your protractor to measure the degrees of each sector.

Here is the first sector:

... you can do the rest!



Solve the following exercises:

13. Make a pie chart studying the colour of eyes of the students in your class.

4.1.3 Pictograms

4.2 Graph for continuous data.

5. Measures of central data

6. Measures of spread

7. Survey

EXERCISES

UNIT 9. STATISTICS.

Exercise 1: Which one of the following is NOT quantitative data?

a) The snake is 7 feet long.
b) The snake is green and yellow.
c) The snake has two eyes.
d) The snake has no legs.

Exercise 2: Which one of the following is discrete data?

a) Sam is 160 cm tall.
b) Sam weighs 60 kg.
c) Sam has two brothers and one sister.
d) Sam ran 100 meters in 10.2 seconds.

Exercise 3: Make a whole distribution frequency table of the number of siblings (brothers and sisters) of all the students in your class.