

Statistics

Statistics is a group of methods that are used to collect, analyze, present, and interpret data and make decisions.

We will encounter the notions of population and sample for much of this course. It is important that we are clear on what both terms mean

A **population** consists of all elements - individuals, items or objects - whose characteristics are being studied.

If we choose a portion of the population to study we call this portion a **sample**.

Sampling Theory

1. An Irish opinion poll typically might use ~ 1000 voters to estimate how an electorate of $\sim 1,500,000$ would vote.
2. A purchaser might inspect a few dozen items of a product for defects and on this basis, estimate the total percentage of defects in a batch of several thousand.
3. A TV station may use the viewing habits of a few hundred families to compile the popularity ratings for programs.

The common idea between these situations is that a relatively small number of units from a large collection is used to predict characteristics of the collection in total.

Recall that a **sample** of a population is a (typically much smaller) subset of the population. The idea is to choose a sample that approximately represents the population.

Example 1

If in number 2 above, it is the case that 8% of all items in the population are defective then we would hope that about 8% of items in the sample are defective. Such a sample is said to be unbiased sample.

If we are careful in selecting a sample then we can make very accurate predictions for the population on the basis of a sample.

Example 2

In number 1 above, if about 1,000 people are sampled and 56% say they would vote for party X then we could deduce (with 95% certainty) that the true support for party X in the electorate of 1,500,000 would be $56\% \pm 3\%$.

Census: In a census, information is collected from every member of the population.

Example 3

The population census is collected every 7 years.

Example 4

A county vehicle registration office could provide a census for all cars registered in that county.

Sampling Frame

This is a list of all units in a population.

Once compiled, it may be used as a sampling frame for another user.

Random Number A random number is a number chosen from a set of numbers in a manner which ensures that all numbers in the set have the same chance of being selected.

Example 5

Choose a random number between 1 and 6.

Possible solution

Roll a die and use the result to choose.

Example 6

Choose a random number between 1 and 12.

Possible solution

Write the numbers 1 to 12 on a piece of paper, place them in a hat and choose one.

Exercise 1

Could rolling two dice together be used as a solution?

Exercise 2

Could rolling one die twice be used as a solution?

Arbitrary number: An arbitrary selected number is a number selected without conditions.

N.B. An arbitrary number is not the same as a random number!!

Example 7 Choose a number between 5 and 12.

Note Many people when asked this question pick 7!

In general it is very difficult to generate random numbers by requiring somebody (or yourself) to pick them.

Simple Random Sampling (SRS)

This method of sampling assigns each member of the population a number. To select a sample of size n, the numbers assigned are randomly selected (by say drawing the numbers from a hat) and the members of the population corresponding to these numbers are used as the sample.

Note The two examples above used SRS.

Problem

Use SRS to select a sample of size 8 from a group of 50 people.

Note

Rolling a die or selecting numbers from a hat is a tedious way to generate a random number.

Alternatives to this include

1. using tables of random numbers (used a lot before computer age)
2. using a random number generator (RNG) on a calculator

Exercise 3

On your calculator, find the RNG.

Typically it will generate a number between 0 and 1 quoting this number to 3 places of decimal.

Solution

Assign each member of the group a number between 1 and 50.

Generate a random number on your calculator and use the first two digits to determine which group member (if any!) to select.

My calculator generated the following numbers but yours will generate different numbers.

<u>Random number</u>	<u>Number</u>	<u>Comment</u>
0.417	41	#1
0.599	59	Too big!
0.786	78	Too big!
0.100	10	#2
0.317	31	#3
0.305	30	#4
0.021	2	#5
0.486	48	#6
0.312	31	Used before!
0.185	18	#7
0.261	26	#8

Thus our sample from the group consists of

2 10 18 26 30 31 41 48

Exercise 4

Generate a sample of size 10 from a population of 105

Note When using SRS, it is not necessary to assign the numbers 1 up to n (sample size). Sometimes a number is already assigned for another reason.

Example 8

To select a sample of students from WIT, we could use students identification number (i.e. 99024572) and write a computer program to generate suitable random numbers.

Exercise 5

How might a SRS be selected from the population of

1. All workers who pay PRSI?
2. All motor vehicles registered in Waterford since Jan 87?

Note SRS is often not very practical.

Example 9

If in the example above 10 motor vehicles are randomly selected, it would be difficult and expensive to locate all of them.

This problem is even more pronounced when dealing with polls.

Note A second difficulty with SRS is the problem of assigning numbers to population members and generating random numbers.

This difficulty is eased with the following method.

Systematic Sampling

Example 10.

Select say a sample of 100 invoices from a large file of invoices in a company.

The first invoice might be selected at random (SRS) and then say every 25th invoice afterwards is selected for the sample.

N.B. It is important to ensure that there are no cyclic patterns in the data if this sampling method is to be used.

Example 11.

Suppose a group of 4 quality controllers on a factory's production line each take every fourth product to examine for defects.

If a systematic sample of products is taken from this production line with every 8th product being selected after the initial one, then all of the products will have been checked by the same controller.

Exercise 6

The residents of a large street in an urban area are being surveyed to determine if they are content with the width of the footpath in front of their home.

Using house numbers, a starting house is selected and every 6th house afterwards is included in the sample.

Discuss this sampling method.

Stratified sampling

Note With SRS, we could expect on average to obtain a sample which has the same characteristics of the population.

Example 12.

If 28% of the Irish population is unemployed and 51% are female then on average we would get the same proportions in our sample, i.e. if we were to repeat the sample selection procedure a large number of times, the average % of unemployed would be 28% and the average % of females would be 51%.

The problem is that in any particular situation we are likely to take only one sample which as a once off may have say 23% unemployed and 57% female.

Stratified sampling addresses this problem by guaranteeing that the important characteristics of the population are precisely represented in the sample.

Example 13.

An opinion poll is being conducted in advance of a referendum. It is believed by the pollsters that the attitudes of the electorate depends on the factors work status (employed/unemployed) and sex (male/female).

Suppose that the following table describes the population with respect to these characteristics.

employed	80%	unemployed	20%
male	50%	female	50%

A sample of size 400 might be constructed as follows.

200 (50%) men would be selected and of these 200, 160 (80%) would be employed. The other 40 (20%) would be unemployed.

200 (50%) women would be selected and of these 200, 160 (80%) would be employed. The other 40 (20%) would be unemployed.

Note

- 1 Sampling in this way is called stratified sampling. In practice, a pollster is likely to consider a number of other strata such as age (e.g. under 30/over 30), area of residence (rural/urban), social class (middle/lower), religion (more likely in Northern Ireland), etc.
2. In our example above we assume that the unemployment rates for men and women are the same. This may not be the case.

Example 14.

If it is thought that 25% of men and 15% of women are unemployed, redo the sample selection problem above.

3. This type of sampling is better than SRS as it guarantees that each group is proportionally represented while SRS can only guarantee that on average each group is proportional represented. However this does not guarantee that the sample characteristics will be the same as the population characteristics.

Exercise 7.

A survey is conducted amongst 2nd level students to determine their attitude to having Irish as a compulsory subject in schools.

The surveyors know that 10% of all students attend schools which teach through Irish. Students are thought to be divided equally in to males and females. Design a stratified sample of size 280.

Exercise 8.

Is the stratification according to sex necessary?

Multi-stage sampling

One problem with all of our methods so far is as follows.

If a sample is selected from the population without considering where the members selected are located within the population, it may be very difficult to reach these members.

Example 15.

A sample of 1000 people from the Irish population would likely give 1000 people spread all over the country.

A pollster would then have to spend a great deal of time (and money) traveling from place to place rather than collecting data.

Example 16.

A sample of 600 fuses selected from all fuses manufactured in a factory might require the sampler to open a large number of cartons which would clearly be undesirable.

A possible solution (say to the problem in example 1) might be to travel to 20 different locations and interview 50 people in each location.

The locations would have to be carefully selected.

i.e. If the population of the country has 75% living in towns or cities then we may wish to take 75% of our locations (i.e. 15) to be in towns or cities.

And similarly for other considerations.

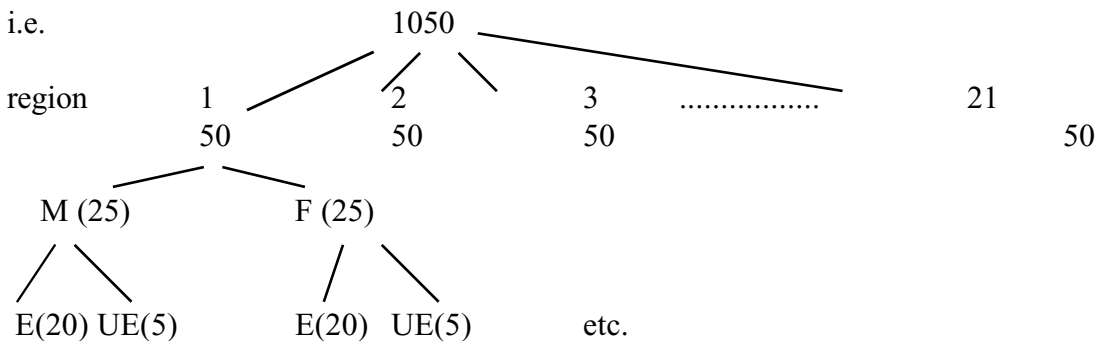
Once each location is selected, stratification methods can be applied to this location to improve the representativeness of the sample, i.e. in our example where we had

employed	80%	unemployed	20%
male	50%	female	50%

we might select a sample of 1050 as follows

1. Divide the total sample size in to say 21 parts of size 50 and choose 21 regions in the country to go to.
2. Within each region stratify as necessary.

i.e.



Note

1. This is called multi-stage sampling.
i.e. The first stage is to divide the sample to be selected in to regions and then divide the allocation for each region in to strata.
2. This is the procedure usually used in professional opinion polls.

The fact that the sample is taken in (21) groups of size (50) tends to make it less representative of the population. However the fact that it is stratified tends to make it more representative and so these factors tend to cancel one another out.

3. In practice, we may not always get integer values. For example if 26% of the population were unemployed, we would need to select 26% of 25 = 6.75 unemployed men and women from each region.

One solution here might be to round upwards for 75% of the regions (because of the 0.75) and downwards for the other 25%.

Example 20. (Welfare)

Q1. Do you feel that the government is spending too much, too little or about the right amount on welfare?

A: Too little 22%

Q2. Do you feel that the government is spending too much, too little or about the right amount on assistance to the poor?

A: Too little 61%

Some opinion polls in Ireland have also produced “odd” results.

Example 21.

In an Irish opinion poll questioning people about the funding of political parties, a majority said that they did not agree with funding being provided by the state.

But in the same poll, a majority said that they did not agree with large donations from the business community.

Example 22.

In opinion polls in the Irish Republic concerning the future of Ireland a majority say that they favour a united Ireland.

But in the same poll, a majority say that they would not be prepared to accept the tax rises that would result.

Data Analysis.

Types of Data.

Note We later consider two criteria which can be used to classify data.

1. Raw data v Grouped data
2. Discrete data v Continuous data

There is another important criteria for classifying data and that is according to type.

Nominal, Ordinal and Cardinal Data:

Nominal Data (or categorical data).

Example 1. A furniture manufacturer produced 6 types of chairs. An inventory of all chairs produced would typically list each chair type along with the number of this type produced. However there would be no natural way to order the chair types.

Example 2. A EU report on crime for each of the unions member countries would typically list each country along with some statistic which measures crime in that country. However, again there is no natural ordering of EU countries.

(The countries could be ordered by their crime levels but this is an ordering of the statistics (crime levels) rather than the categories)

Thus nominal data types are typified by the lack of any ordering of the various categories.

Ordinal Data

Example 3

A customer comment card in a restaurant requests customers to rate the service provided by checking one box beside the entries excellent, good, average, poor.

The possible choices can clearly be ordered in the obvious way but they cannot be quantified.

i.e. good is better than average but by how much?

Example 4

A survey consists of a number of statements to which each person surveyed can respond with one of the following.

Strongly agree, agree, no opinion, disagree, strongly disagree.

1.4.3. Cardinal Data

Example 5 The scores obtained by students in an examination.

As with ordinal data, the data here can be ordered but unlike ordinal data the ordering can be quantified.

e.g. If A gets 40% and B gets 60%, we can say that B has scored higher than A and by 20%.

Example 6 A soft drink sold in a take away restaurant is classified as small, medium or large.

At first glance this looks like ordinal data but we could make a case for classifying it as cardinal data.

i.e. Small may be 200 ml, medium 300 ml and large 450 ml.

Our classification is sometimes context dependent.

N.B. Most quantitative statistical techniques use cardinal data.

Surveys of the type above (Ordinal data, example 2) frequently are processed by “cardinalising” the data.

i.e. Strongly agree is coded as 5 and Strongly disagree as 1 etc.

Care must be taken when applying quantitative techniques here.

Example 7 Suppose the survey was designed to gauge the public's intolerance of crime and includes the following two questions.

1. Offenders should be required to compensate crime victims.
2. Offenders should be put in a swimming pool full of piranhas

In both cases “strongly agree” would contribute 5 to the total measure of intolerance while is clearly not appropriate.

Presentation of Data.

When data is first collected, it is usually not in a form that conveys much information. It may consist of a list or a table of values, and it will often need some refinement before one can draw any conclusions from it.

Section 1. Discrete Data.

A discrete variable contains a list of separated, or *discrete* values. This is a variable that is found by counting something, such as the number of days in a month that an employee is absent or the number of broken watches in batches of 100. These two things can only take on certain separated values.

1.1. Raw Data.

Usually statistical analysis involves looking at a large collection of numbers and making sense of them. Occasionally one may be able to work with this raw data, without having to resort to grouping it. There are three main aspects in the presentation of data:

1. the use of tables to clarify the numbers e.g. a frequency distribution.
2. the use of graphs to pictorially represent the data e.g. a histogram
3. the use of a single figure which is representative of the whole data set e.g.

Frequency Distributions.

Example 1

In order to monitor the efficiency of his work-force, a manager checks the number of days which his employees have been absent over the past month. This search reveals the following information:

1	5	3	3	2
3	0	4	1	4
3	3	2	1	2
1	1	0	3	6
5	0	3	4	2

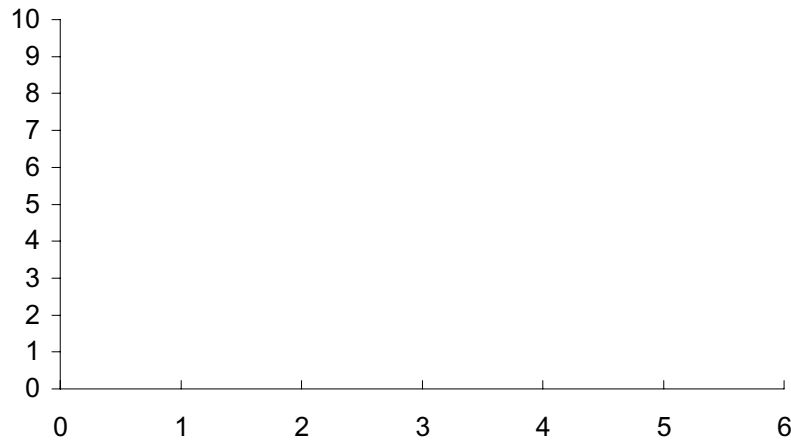
All of the values in the previous example lie between zero and six inclusive. It might be useful to find out how often each value in the range occurs in the table. This can be done by tallying, or counting, and will give the following result:

Number of days Absent	Frequency
0	
1	
2	
3	
4	
5	
6	
Total	

The totals in the above column are called frequencies and the table is called a frequency distribution. Thus, the frequency of 3 days absent is 8, and so on.. Note that there is a total of 25 people working in the factory. This should be clear from the raw data.

Histogram.

A histogram is a bar graph of a frequency distribution. The x-axis is the variable being measured and the y-axis is the corresponding frequency. The frequency is represented by the height of a block and the class being covered is represented by the base of the block. For discrete variables, the ‘blocks’ become vertical lines.



Cumulative Frequency Distribution.

A cumulative frequency distribution lists the number of values **up to and including** a certain point. It can easily be computed as a running total from the corresponding frequency distribution.

For example, the frequency distribution of the number of days absent can be used to obtain:

Number of days Absent (less than or equal)		Cumulative Frequency		% Cumulative Frequency
0				
1				
2				
3				
4				
5				
6				

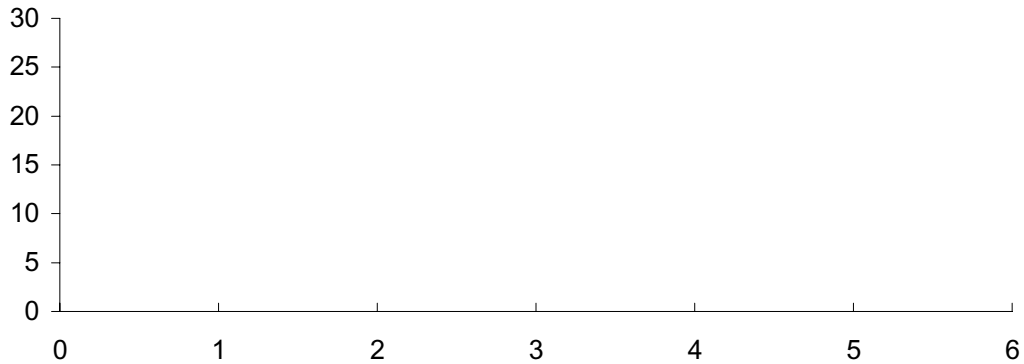
Suppose we are asked how often the employees are more than four days absent. The answer can be found from the above table as follows:

25 - 22 = 3 occasions out of 25: i.e. 12%

Ogives.

An ogive is a graph of a cumulative frequency distribution. The x-axis is the variable being measured, and the y-axis is the corresponding cumulative frequency.

In the case of discrete variables, we do not join up the values, but rather draw vertical lines.



Measures of Average.

1. The Arithmetic Mean.

The arithmetic mean is obtained by dividing the sum of all the values in question by the number of values.

For a variable, x, this measure can be found as follows:

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

where \bar{x} = the arithmetic mean
n = the number of values

In our example above, if we add up all 25 values given to us in our initial data set then we obtain a value of 62. This figure means that there was a total of 62 days lost due to absenteeism. Since there are 25 employees this means that each employee was absent on average for $\frac{62}{25} = 2.48$ days. We can arrive at this figure in a more straightforward way by means of the frequency distribution table.

x	f	fx
0	3	
1	5	
2	4	
3	8	
4	3	
5	2	
6	1	
Totals	25	

The formula being used here is

$$\bar{x} = \frac{\sum f x}{\sum f}$$

2. The Median.

The median of a set of values is the value of that item which lies exactly halfway along the set. In order to find this value we need to look at the cumulative frequency table.

Number of days Absent (\leq)	Cumulative Frequency
0	3
1	8
2	12
3	19
4	22
5	24
6	25

Since there are 25 values in total we are looking for the value that has 12.5 values below it. Since there are 12 values equal to or below the value 2 and 19 values equal to or below 3, we find that the median is 3.

3. Mode.

The mode of a data set is the value that occurs most often, or equivalently has the largest frequency. When dealing with raw data, it is simply a *counting process* to determine the mode. In our above example it is easy to see from the histogram that the mode is 3.

Measures of Spread.

Having found a measure for the center of the data we now wish to see how the data is spread about that center. The simplest measure is that of the

1. Range.

Range=largest value-smallest value.

In our example above this amounts to 6-0=6. This is not a great measure since it can be very badly skewed by one extreme point. If, for example, there was one person absent for 30 days in a given month but everyone else was absent for no more than 6 days then a range of 30 is not representative of the data.

2. The Standard Deviation.

Spread is a measure of how much the data points deviate from a measure of central tendency such as the mean. If we were to calculate how much each data entry deviated from the arithmetic mean and then summed these deviations we would always end up with an answer of 0. This is because those values that are larger than the mean would end up canceling out those that are smaller than the mean.

One way around this problem is to square all the deviations. We can then sum these squared deviations and divide by the total number of values to obtain a measure of spread. Since we squared the deviations initially it is customary to take the square root of the sum. Therefore if a point is close to the mean it will contribute only a small amount to this number, if a point is far from the mean then it will contribute a large amount to this number.

We can use the frequency distribution table to calculate this figure. The standard deviation is given by

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$$

x	f	fx		
0	3	0		
1	5	5		
2	4	8		
3	7	21		
4	3	12		
5	2	10		
6	1	6		
	25	62		

3. Inter-Quartile Range.

An alternative measure of spread is the **Inter-Quartile Range**. This basically measures the middle 50%. If we let Q1 denote the value that has 25% of the data below it and so 75% of the data above it, and Q3 denote the value that has 75% of the data below it and 25% above it then

$$\mathbf{IQR = Q3 - Q1.}$$

To calculate this we look at the cumulative frequency table. We see this as follows.

Number of days Absent (\leq)	Cumulative Frequency
0	3
1	8
2	12
3	19
4	22
5	24
6	25

Q1 is the value that has 25/4 or 6.25 values below it. This corresponds to the value 1. Q3 is the value that has 19.75 values below it. This corresponds to the value 3. Therefore

$$\mathbf{IQR = 3 - 1 = 2.}$$

Essentially this means that the middle 50% of the values are spread between the values 1 to 3 and so has a range of 2.

The only problem with this is that it does not use all of the values in its calculation.