

# An Introduction to Statistics for Clinical Audit

**Jane Moore, Healthcare Quality Improvement Partnership**  
**Mandy Smith, Healthcare Quality Improvement Partnership**  
**Mary Barwick, Berkshire East Primary Care Trust**

*Clinical audit tool to promote quality for better health services*



**Revised -  
minor changes to wording -  
November 2011**

**Previous versions:**

April 2010 (first publication)

## Contents

---

<b>1</b>	<b>Overview</b>	<b>1</b>
<b>2</b>	<b>Types of data</b>	<b>2</b>
2.1	Data versus information	2
2.2	Continuous data	2
2.3	Discrete data	3
<b>3</b>	<b>Descriptive statistics</b>	<b>4</b>
3.1	Distributions of data	4
3.2	Averages	5
3.3	What type of average should you use	5
3.3.1	Mean	5
3.3.2	Standard deviation	6
3.3.3	Variance	6
3.3.4	Median	6
3.3.5	Interquartile range	7
3.3.6	Quantiles	7
3.3.7	Mode	7
3.3.8	Range	8
3.3.9	Percentages	8
<b>4</b>	<b>Ways of presenting clinical audit findings</b>	<b>9</b>
4.1	Tabular representation of data	9
4.2	Frequency categorisation	10
4.3	Graphical representation of data	10
4.4	Bar chart	10
4.4.1	Simple bar chart	11
4.4.2	Multiple bar chart	12
4.5	Histogram	13
4.6	Line graph	13
4.7	Pie chart	14
4.8	Summary	15
<b>5</b>	<b>Populations and sampling</b>	<b>16</b>
5.1	Population	16
5.2	Sampling	16
5.3	Representative or random sampling techniques	16
5.3.1	Simple random sampling	17
5.3.2	Systematic random sampling	17
5.3.3	Stratified random sampling	18
5.3.4	Cluster sampling	19
5.4	Non-representative sampling techniques	19
5.4.1	Purposive sampling	20
5.4.2	Convenience sampling	20
5.5	Summary	20
<b>6</b>	<b>Glossary of terms</b>	<b>21</b>

<b>7</b>	<b>Further reading</b>	<b>23</b>
	<b>Appendix 1. How to select a random sample using Excel</b>	<b>24</b>

# 1 Overview

---

To improve the quality of patient care, it is vital to establish if care provided currently meets best practice. Through the clinical audit process, current practice is measured against best practice and any gaps in care are identified and addressed. Therefore, it is imperative that the right data are collected and that data are analysed appropriately to obtain an accurate reflection of care provided and to easily identify any needed changes in clinical practice and service delivery.

This introductory guide has been designed for individuals who are new to clinical audit. It aims to:

- explain how descriptive statistics are used to present and analyse clinical audit data
- provide general principles on how to use the right statistic and how to present statistics clearly and concisely.

Appropriate statistics need to be applied correctly and people working in clinical audit need to be confident in using descriptive statistics.

## 2 Types of data

---

### 2.1 Data versus information

The term data is used to describe a collection of facts from which conclusions may be drawn. Data on their own have no meaning. Data have to be interpreted to become information.

For example, if one of the standards you are measuring is 'all patients admitted to the ward should have a pressure ulcer assessment completed within 24 hours of admission', you need to collect data on the number of patients admitted to the ward and the number of patients who had a pressure ulcer assessment completed. In your analysis of data, you will answer the question 'How many patients admitted to the ward had a pressure ulcer assessment completed within the time specified?' If your answer is 47, it may be a perfectly accurate bit of data, but it is not information. A more complete and useful answer would be 47 out of a total of 85 patients, which is 55% of the patients on the ward during the time over which data were collected.

In many clinical audit projects, **frequencies** (the number of patients meeting a standard) and **percentages** (the proportion of patients meeting a standard) are all the information a clinical team will need. However, there are many ways in which the information can be presented and explained. It is important to understand the type of data that you are collecting before you can decide on the statistical techniques to use and how best to present the information. If you use the wrong techniques or presentation, you could lead a clinical team to draw the wrong conclusions.

There are two types of data: **quantitative** and **qualitative**.

This guide is about the statistics you can use to analyse quantitative data, which is data expressed in numbers. Some clinical audits also can involve collecting qualitative data, such as patients' descriptions of how they feel about the treatment they received. The qualitative data can provide additional insight into the way a service operates or the impact it has on patients. However, the analysis of qualitative data is not covered in this guide.

There are two types of **quantitative data**: continuous data and discrete data.

### 2.2 Continuous data

Continuous data run in a continuous sequence and use real numbers. There can theoretically be an infinite number of values between any two points in the sequence. For example, between one centimetre (cm) and two cms, you can have tenths of cms, hundredths, thousandths and so on.

There are two types of continuous data: **interval** and **ratio**.

**Interval** — Interval data (also called integer) are measured along a scale in which each point is an equal distance from the next; however, the **zero point** (or point of beginning) is arbitrary. For example, data on temperature may be collected using a zero point of 35 degrees Celsius. Examples of interval data are temperature or year.

**Ratio** — Ratio data represent quantities in terms of equal intervals but have an **absolute zero point** of origin, thereby allowing a proportional relationship between two different numbers or quantities. For example, someone who has waited in a hospital’s emergency department for two hours can be said to have waited twice as long as someone who has waited one hour. Examples of ratio data are height, weight or length of time in hours or minutes.

### 2.3 Discrete data

Discrete data result when names or numbers are assigned to different mutually exclusive categories and the number of observations in each category is determined.

There are two types of discrete data: **ordinal** and **nominal**.

**Ordinal data** — Ordinal data are individual values that can be ordered or assigned a specific rank on a scale. For example, in a satisfaction survey, you could use a four–point response scale such as “very satisfied, satisfied, dissatisfied, very dissatisfied.” These responses can be placed in order of satisfaction; however, you could not say that one person is twice as satisfied as another person. Examples of ordinal data are level of satisfaction or grade of pressure sore.

**Nominal data** — Nominal data result from using categories that represent qualities rather than quantities and have no reference to a linear scale. For example, when reviewing who is accessing a particular health promotion service, it may be important to have information on people’s marital status, for example, single, married, divorced or co-habiting. One person cannot be more single than another and there is no progression or sequence among the categories. Examples of nominal data are gender, marital status or blood group.

Types of quantitative data and examples of each are in the box.

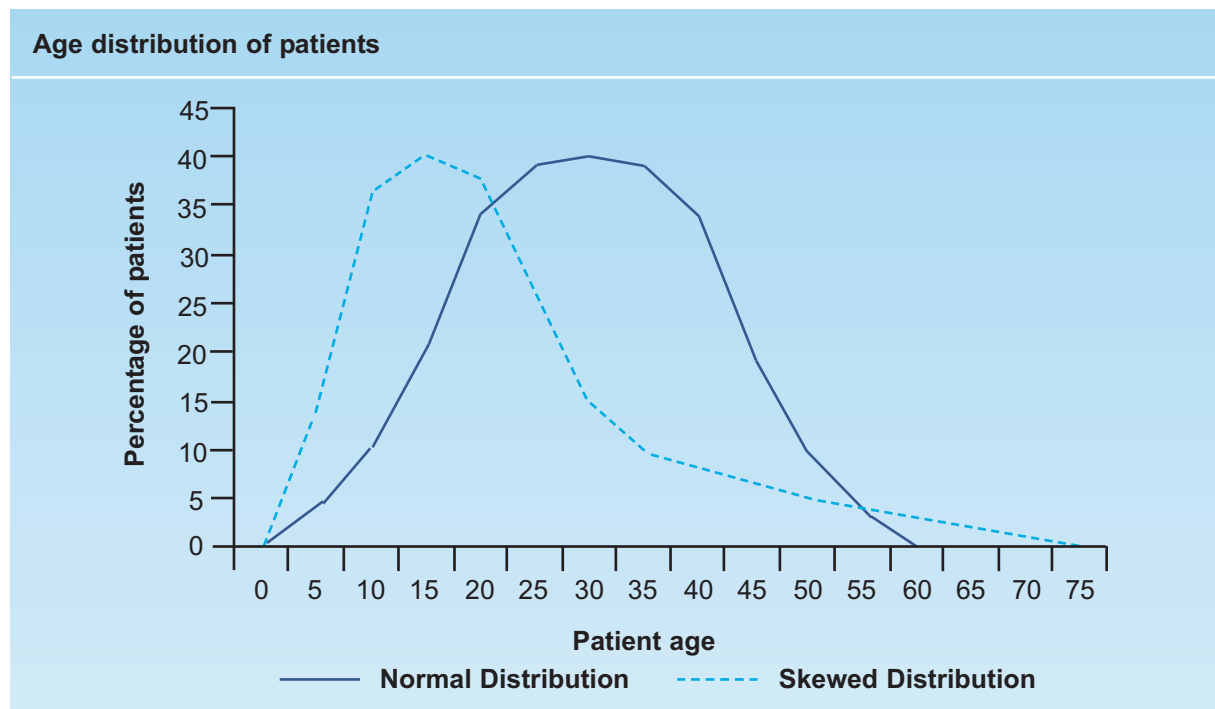
Types of quantitative data and examples			
Quantitative data			
Continuous		Discrete	
<b>Interval</b>	Temperature, year	<b>Ordinal</b>	Level of satisfaction, grade of pressure sore
<b>Ratio</b>	Height, weight, length of time in minutes, age, blood haemoglobin levels	<b>Nominal</b>	Gender, marital status, social class, ethnic group, ward, GP practice, blood group

### 3 Descriptive statistics

Descriptive statistics are used to describe the main features of a collection of data in quantitative terms. They involve summarising, tabulating, organising and displaying data for the purpose of describing a population or sample of individuals, events or circumstances that have been measured or observed.

#### 3.1 Distributions of data

The graph shows data on the age distribution of two groups of patients.



There are two types of distributions of age in this graph. The **solid line** represents what is called a **normal distribution**. It is symmetrical around a centre point or is bell shaped. Most of the patient ages are close to the average and relatively few are at one extreme or the other. You might see this type of distribution if you collected data on the heights of all children of a particular age.

The **dotted line** represents what is called a **skewed distribution**, that is, it is not normal. (It is also sometimes called non-normal.) A skewed distribution is common in clinical audits. For example, most patients are discharged from hospital in a few days but a few patients may stay in hospital much longer.

Other non-normal distributions of data can occur where there is more than one peak in the data.

## 3.2 Averages

Many healthcare staff have difficulty understanding the significance of large sets of numbers, and therefore, numbers are often described in tabular or pictorial form. However, it can be more convenient to describe a set of numbers by using a single number. In clinical audit, calculating a single number is one of the most frequent methods of presenting data, for example, providing a single 'average' to represent a set of numbers. An **average** is simply any single number that represents many numbers.

## 3.3 What type of average should you use

There are three types of averages, and it is important to understand the type of average to use, which can depend on the distribution of the data.

### 3.3.1 Mean

Many people use the term average when they are actually referring to the mean. The mean is dependent on all the observed values in a data set. To calculate the mean, add all the observed values and divide by the total number of values.

For example, data collected on blood haemoglobin levels (g/100ml) for a sample of patients are in the box. The mean blood haemoglobin level is  $366/30 = 12.2\text{g}/100\text{ml}$ .

Example of data on blood haemoglobin levels					
Patient	Hb Level	Patient	Hb Level	Patient	Hb Level
1	10.2	11	13.7	21	10.1
2	13.3	12	12.9	22	11.2
3	10.6	13	10.5	23	12.9
4	12.1	14	12.9	24	13.6
5	9.3	15	13.5	25	9.2
6	12.0	16	12.9	26	10.3
7	13.4	17	12.1	27	11.6
8	11.9	18	11.4	28	12.8
9	11.2	19	15.1	29	14.3
10	14.6	20	11.1	30	15.3
Total	Hb Level				366

**You should use the mean only if you have a normal distribution.** If data are skewed, the mean will be affected by the extreme values and will not give a true representation of what is typical for the sample. Also, use the mean if there is a reasonable number of observations in your data set (30 is the generally accepted minimum number). If there are fewer observations, you cannot be certain that your data really are normally distributed.

### 3.3.2 Standard deviation

If you use the mean for a set of data, you also can use the standard deviation. The standard deviation (SD) gives information about the spread of data around the mean. The value of the standard deviation should be compared relative to the mean. A large standard deviation, when compared to the mean, implies that the data are widely spread, whereas a small standard deviation implies that the data are mainly concentrated around the mean. When data are distributed normally, 95.44% of values lie between  $\pm 2SD$  of the mean. An example is in the box.

In a group of patients, the mean age is 52 years and the standard deviation is 1.2. Roughly 95% of the patients' ages will occur between  $52 \pm 2(1.2)$ .

$$52 \pm 2SD = 52 \pm (2 \times 1.2) = 52 \pm 2.4 = 49.6 \text{ to } 54.4 \text{ years}$$

This means that roughly 95% of patients will be between 49.6 and 54.4 years of age.

If the standard deviation was 10.5 years, for example, the age range of the patients would be more widely spread. 95% of the patient ages would be between 31.0 and 73.0 years.

If the amount of spread is unexpected or unlikely, it may indicate an error in data collection or sampling that should be further investigated.

### 3.3.3 Variance

**Variance** is the square of the standard deviation. Variance is another way of calculating the degree to which data are dispersed or spread out from the mean. The formula to calculate variance is  $\text{variance} = s^2$  where  $s^2$  is the square of the standard deviation.

In the example above, if the standard deviation is 1.2 years, the variance would be:

$$\text{Variance} = (1.2)^2 = 1.44$$

### 3.3.4 Median

If your data are skewed (and in most clinical audits, data collected will be skewed), you should use the median, not the mean. The **median** is the middle value of the data set when all the numbers are arranged in order. The median divides a data set into equal parts. An example is in the box.

Waiting times (in days) for a CT scan for 15 patients

0, 0, 1, 1, 1, 2, 2, 2, 4, 5, 5, 6, 8, 9, 10

↑  
Median



### 3.3.8 Range

Whether you use a mean and standard deviation or variance, or a median and interquartile range, you should also use the **range** of observations, that is the highest and lowest values in a data set. Examples are in the box.

The mean height of children in the class on the day of data collection was 125 cm, the standard deviation was 2.7 cm, and the range was 117 to 133 cm.

The median length of stay on ward X for patients in the audit sample was 10 days, the IQR was 4–14 days, and the range was 3–22 days.

#### Summary of when to use mean, median and mode

	When to use	What to use
<b>Mean</b>	Normal distribution Reasonable number of observations (30+)	Standard deviation Variance Range
<b>Median</b>	Skewed distribution	Interquartile range Range
<b>Mode</b>	Distribution with a double peak Skewed distribution	Range

### 3.3.9 Percentages

Percentages translate as 'per hundred'. They are useful when comparing groups of different sizes. A percentage is calculated by dividing the number of observations by the total in the sample and multiplying by 100. An example is in the box.

145 children were seen in a paediatric clinic in one week. 50 children were seen by Health Visitor A, 80 were seen by Health Visitor B and 15 were seen by the Paediatrician. The percentages of children seen by each professional are as follows:

$$\text{Health Visitor A} = 50/145 \times 100 = 34.5\%$$

$$\text{Health Visitor B} = 80/145 \times 100 = 55.2\%$$

$$\text{Paediatrician} = 15/145 \times 100 = 10.3\%$$

In clinical audit reports, it is good practice to include numbers and percentages together so that the data can be easily interpreted and accurate conclusions can be drawn.

## 4 Ways of presenting clinical audit findings

After clinical audit data are collected and analysed, the findings need to be presented in ways that enable interpretation quickly and easily.

When you have decided what type of data you are analysing, you can choose the most appropriate way of presenting the data. Using the wrong type of presentation for your data can lead to incorrect assumptions being made about the meaning.

### 4.1 Tabular representation of data

The simplest way of presenting data is in a **table of frequencies and percentages**.

Frequency of method of delivery			
Method of delivery	Number of births		
	Number		Percentage
Normal	420		70.0%
Forceps	150		25.0%
Caesarean section	30		5.0%
<b>Total</b>	<b>600</b>		<b>100.0%</b>

Percentages can be helpful but they also can be misleading. You should not use percentages if your sample size is small (for example, less than 20). Also, you should make sure that you only use percentages for cases for which you have valid data. An example is in the box.

**Say you are auditing the completion of falls assessments, and two of the standards are that:**

- Every patient has a falls assessment.
- The falls assessment is signed by the assessor.

**You check 100 patient records and you find the following:**

- 95 of 100 patients have falls assessments.
- Of the 95 assessments, 85 have been signed.

**In your report, you should present the following information:**

- In 5 cases, no falls assessment was found. Therefore, compliance with the first standard is 95/100 (95.0%).
- In 85/95 cases in which a falls assessment was done, the falls assessment was signed by the assessor.

You need to decide how to present the findings for the second standard. If the second standard implied that all patients would have an assessment, compliance would be 85/100 (85%). If the second standard does not make the assumption that there is a falls assessment for every patient, you could subtract from the denominator the patients for whom an assessment was not done and report compliance as 85/95 (89%).

## 4.2 Frequency categorisation

With continuous data, or discrete data with a broad span of values, you can end up with many observations all with different values. To make it easier to work with the data, you can use a frequency categorisation table to group the data into categories.

Frequency categorisation of blood haemoglobin levels in women		
Haemoglobin (g/100 ml)	Number of women	Percentage
8 – 8.9	1	1.4%
9 – 9.9	3	4.3%
10 – 10.9	14	20.0%
11 – 11.9	19	27.1%
12 – 12.9	14	20.0%
13 – 13.9	13	18.6%
14 – 14.9	5	7.1%
15 – 15.9	1	1.4%
<b>Total</b>	<b>70</b>	<b>100.0%</b>

You need to make sure that observations can fall into only one category. Therefore, you should have 8–8.9 and 9–9.9, NOT 8–9 and 9–10. Also, you should make sure your categories are all the same size as otherwise your table will be misleading.

## 4.3 Graphical representation of data

Graphical representation of data makes the data appear more interesting and easy to understand. There are a number of ways in which data can be represented graphically but the basic graphic representations of data are:

- bar charts
- histograms
- graphs
- pie charts.

## 4.4 Bar chart

A **bar chart** is used to show the distribution of any type of discrete data, ie, for ordinal or nominal data. The bars are of equal width and have equal distance between the bars.

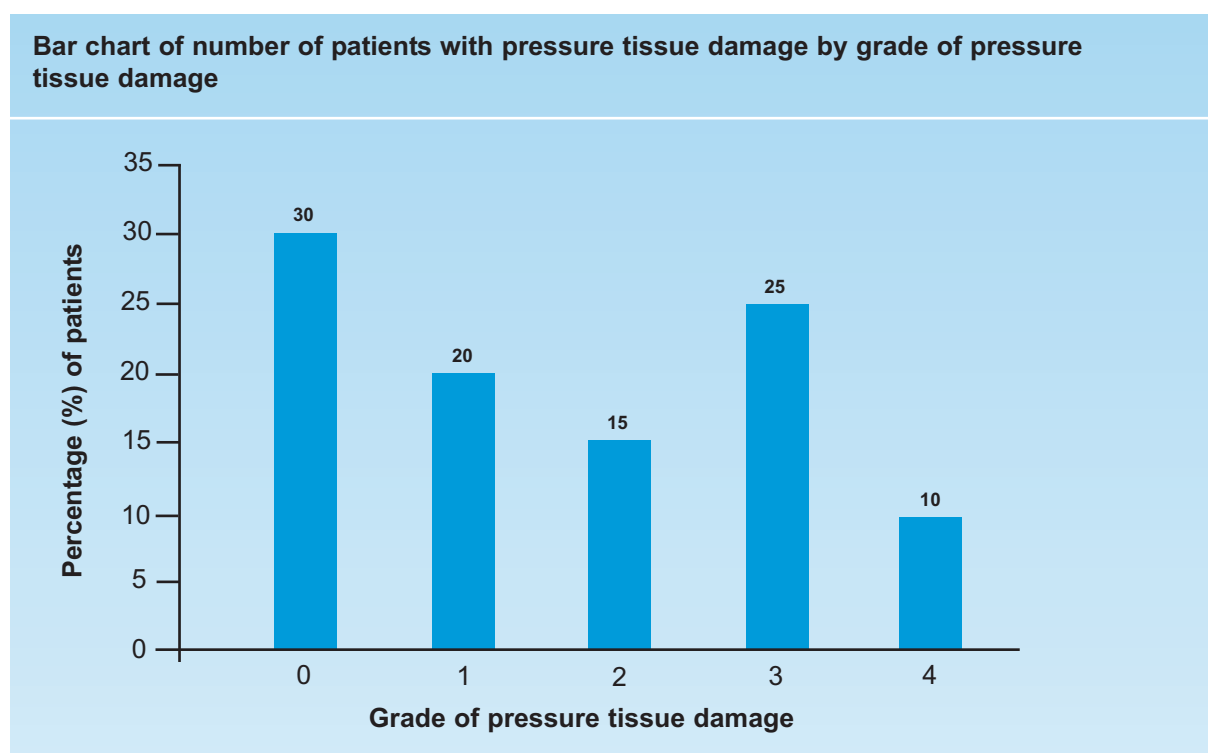
There are two types of bar charts—**simple** bar chart and **multiple** bar chart.

#### 4.4.1 Simple bar chart

A **simple bar chart** is one in which the bars represent **one** quantity or variable only. The length of the bar indicates the number of people or items in that category. The bottom and side of the chart must have clear titles. An example is in the box.

Number of patients with pressure tissue damage by grade of pressure tissue damage		
Grade of pressure tissue damage	Number of patients with pressure tissue damage	Percentage of patients
0	6	30%
1	4	20%
2	3	15%
3	5	25%
4	2	10%
<b>Total</b>	<b>20</b>	<b>100%</b>

The data from the table are presented in the bar chart.



From the bar chart, we can see, for example, that 30% of patients had no pressure tissue damage, and 25% had grade 3 pressure tissue damage.

#### 4.4.2 Multiple bar chart

A **multiple bar chart** is one in which the bars are displayed side by side, often in pairs or triples, in order to show comparisons. This is usually preferable to drawing two separate bar charts of the quantities where immediate comparisons would be difficult. The multiple bar chart is frequently used in clinical audit to demonstrate how data can be compared between one round of measurement in an audit and the next. An example is in the box.

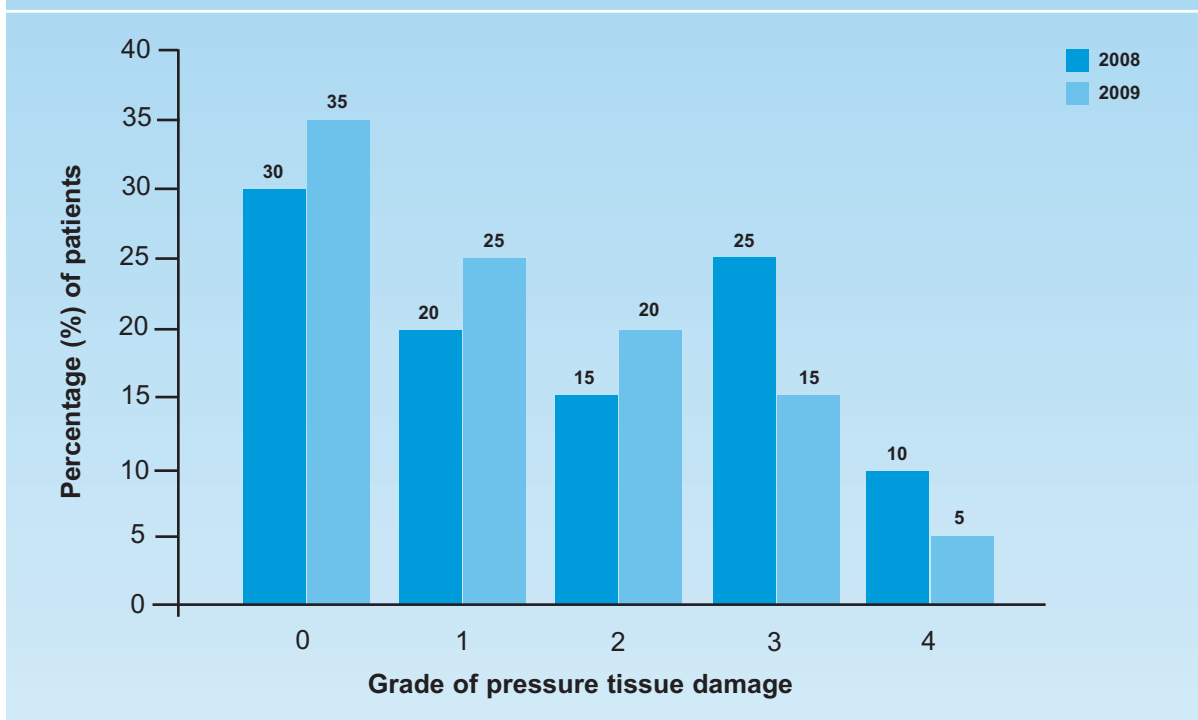
You carried out an audit of pressure tissue damage in 2008 and repeated the audit in 2009. Now you want to display the data in order to compare the findings between one year and the next. Your results are displayed in the table.

Patients with pressure tissue damage by grade of pressure tissue damage in 2008 and 2009

Grade of pressure tissue damage	2008		2009	
	Number of patients with pressure tissue damage	Percentage of patients	Number of patients with pressure tissue damage	Percentage of patients
0	6	30%	7	35%
1	4	20%	5	25%
2	3	15%	4	20%
3	5	25%	3	15%
4	2	10%	1	5%
<b>Total</b>	<b>20</b>	<b>100%</b>	<b>20</b>	<b>100%</b>

The findings of both audits can be displayed in a multiple bar chart as follows.

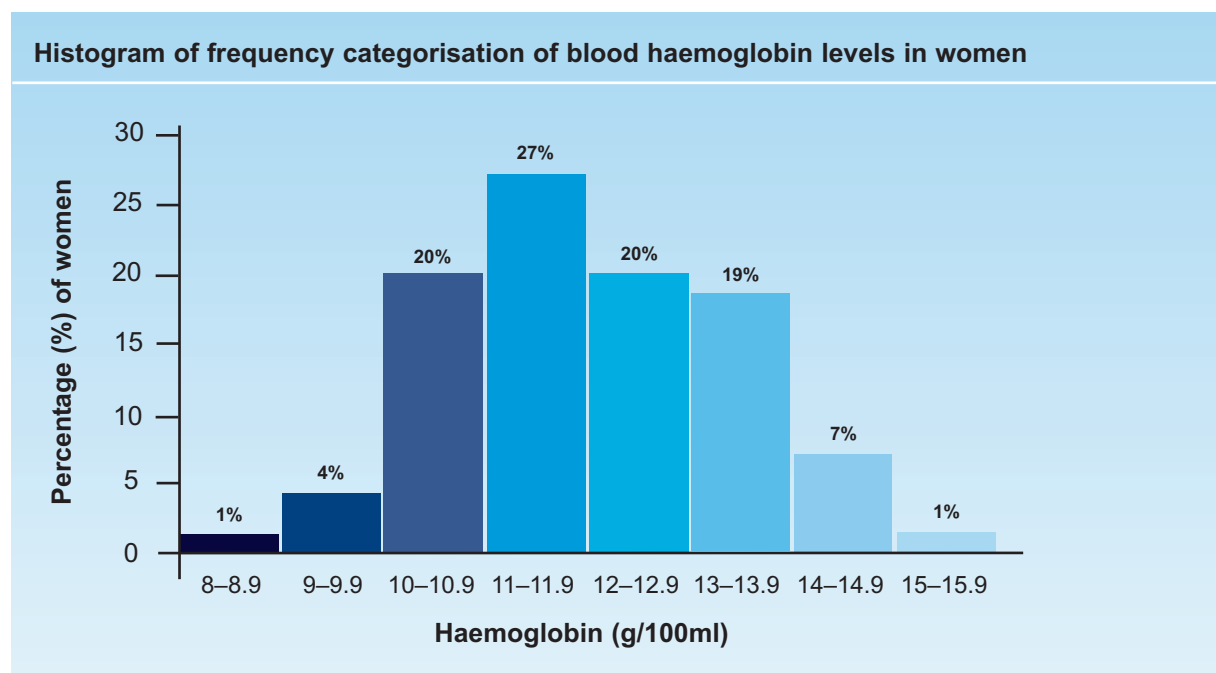
Bar chart of pressure tissue damage by grade of pressure tissue damage



Bar charts can be presented vertically or horizontally; however, it is good practice to be consistent throughout your audit report rather than changing the presentations, which may be confusing to the reader. It is more common to have the categories across the bottom of the chart (x axis) and the number or percentage along the side (y axis), as in the example. Most readers prefer a simple uncluttered chart.

#### 4.5 Histogram

A **histogram** is used to show the distribution of a continuous variable (interval or ratio data). Thus, there are no gaps between the bars as there are in a bar chart, because the data are continuous. The area of each column represents the number of observations in each category. An example is in the box.



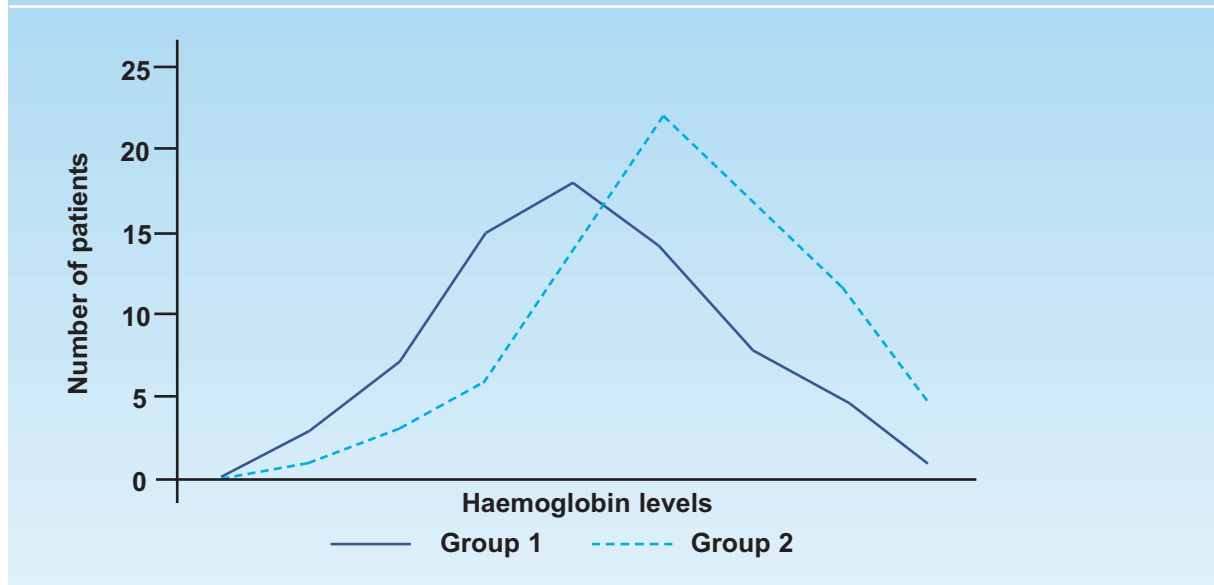
Histograms should not be used for multiple sets of data.

#### 4.6 Line graph

Graphs can be used to display multiple sets of data, for example, year on year comparisons.

You should not use a line graph to display discrete data (ordinal or nominal), because by joining up the points you are implying that there are continuous and intermediate values which would fall on the line, and with discrete data there are no intermediate values between the categories. An example is in the box on the next page.

Graph of haemoglobin levels in two groups of patients



#### 4.7 Pie chart

Any type of data also can be presented in a pie graph or pie chart; however, the chart is used normally for nominal and ordinal data. Each slice of the pie represents the number of observations in a category. The various slices of the pie are proportionally represented by parts of the circle. The size of each slice is calculated manually by calculating the angles of each slice; however, the charts are normally constructed using computer software.

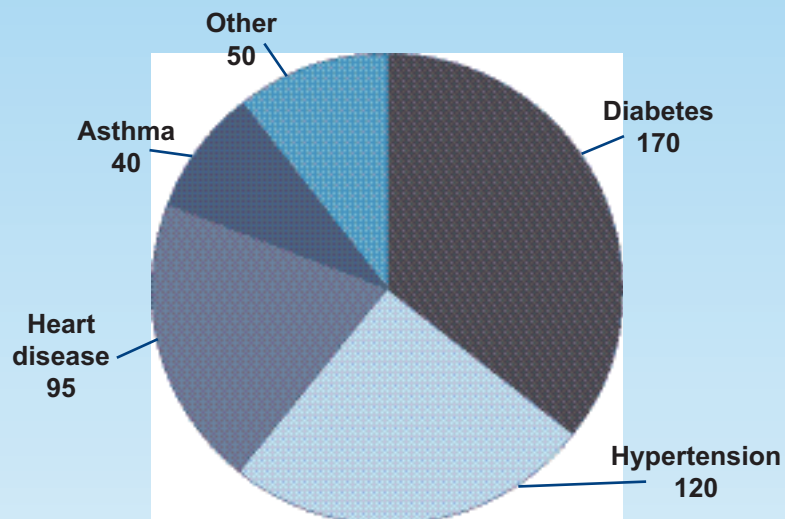
Pie charts should be used only when there are three or more categories. An example is in the box.

You have been asked to look at the number of people seen in a GP practice who are diagnosed with various diseases. You have put the results into a table.

Number of patients diagnosed with different diseases	
Disease	Number diagnosed
Diabetes	170
Hypertension	120
Heart disease	95
Asthma	40
Other	50
Total	475

The results are presented in a pie chart.

**Pie chart of number of patients with different diseases**



#### **4.8 Summary**

It is important to use the right type of chart for the right data. All charts always should be clearly labelled with a title, and axes always should be labelled completely and accurately.

##### **When to use what type of chart**

<b>Type of chart</b>	<b>Use for</b>
Bar chart	Discrete data — nominal or ordinal data
Histogram	Continuous data — interval or ratio data
Line graph	Continuous data — interval or ratio data
Pie chart	Discrete data — nominal or ordinal data

## 5 Populations and sampling

---

### 5.1 Population

In statistics, the term **population** means the entire collection of items that is the focus of concern. The term population is used often to refer to all the people, things, items or cases that you are dealing with. An example is in the box.

If a group of podiatrists was interested in carrying out an audit on adult patients over 18 years of age with type 2 diabetic foot ulcers attending the clinic over a 3-month period, the population would be all type 2 diabetic adults with a foot ulcer seen in the clinic over the 3-month period.

A population can be of any size. Patients or persons in a population need not be uniform; however, the items must share at least one measurable feature.

### 5.2 Sampling

Sampling techniques are an important tool in clinical audit. It is important to understand sampling techniques in order to identify and use the correct one.

A population sometimes includes many individuals, making it inconvenient to include all of them in a clinical audit. A **sample** is some, that is, a specific collection of the people, things, items or cases that are drawn from a population in which you are interested. Therefore, we often draw a sample from the population, the sample containing a smaller number of individuals selected from the population, but meeting all the population parameters. The reason for taking a sample rather than including the entire population is to reduce the resources needed to carry out an audit. An example is in the box.

The population is all the patients seen in a clinic in a year. We want to audit whether or not the date the patients are seen is documented in their records. If 2500 patients have been seen in the clinic in the year, we could check all 2500 records. We would be certain that we knew exactly how many cases had complied with the standard. However, we could check a sample of 100 records. If we find that in 94/100 (94%) cases in the sample, the dates have been documented, then it may be reasonable to assume that the dates have been documented in 94% (or 2350) of all 2500 cases.

When data for a population are collated, normally the purpose is to identify characteristics of the population. When data for a sample are used the purpose is to make inferences about the characteristics of the population from which the sample was drawn.

### 5.3 Representative or random sampling techniques

In order to get a representative sample of a population, you need to draw the sample in a systematic way so that each and every individual in a sample has an equal opportunity to be selected. There are a number of representative sampling techniques which are described in the following sections.

In random sampling, all individuals have an equal chance of being selected in the sample. Random sampling techniques ensure that bias is not introduced regarding who is included in the audit. Four common random sampling techniques that are useful in clinical audit are:

- simple random sampling
- systematic sampling
- stratified sampling
- cluster sampling.

### 5.3.1 Simple random sampling

With **simple random sampling**, each item in a population has an equal chance of inclusion in the sample. An example is in the box.

Each patient attending an outpatient cardiac clinic over a 2-year period could have a number allocated instead of having the patients' names on a list, such as 1, 2, 3, 4, 5, 6, 7... 50... 100... and so on. If the sample was supposed to include 200 patients, then 200 numbers could be randomly generated by a computer programme or all the numbers could be put on individual chits and 200 numbers could be picked out of a hat. These numbers then could be matched to names on the cardiac outpatient list, thereby providing a random list of 200 people.

See Appendix 1 for how to obtain a simple random sample using Excel.

The advantage of simple random sampling is that it is simple and easy to apply when small populations are involved. However, because every person or item in a population has to be assigned a number and listed before the random numbers can be selected and then the needed number of random numbers picked, this method is cumbersome to use for large populations.

### 5.3.2 Systematic random sampling

**Systematic random sampling** means that the individuals included in the sample are selected according to an interval between individuals on the population list. The interval remains fixed for the entire sample. This method is used often for large populations. We might decide to select every 20th individual in a population to be included in a sample. This technique requires the first individual to be selected at random as a starting point and thereafter, every 20th item is chosen. The technique could also be used to select a fixed size sample. An example is in the box.

We want to do a documentation audit on district nurses' patient records. There are 1500 patients on a district nurse's caseload and we want a sample of 100 patients for the audit. The sampling interval would be determined as follows:  $1500/100 = 15$ . Thus, we would need to include every 15th patient from a list of the 1500 patients to get a systematic random sample. All patients would be assigned a number in sequence. The first case would be selected at random, and after the first case, we would select every 15th patient until we get 100 patient records.

The advantage of systematic sampling is that it is simpler to select one random number and then every  $n$ th (eg, 15th in the above example) patient on the list, than to select a random number for every case in the sample. It also gives a good spread across the population. A disadvantage is that you will need a list of all patients to start with (as you would for any type of representative sample), to know your sample size and calculate your sampling interval.

### 5.3.3 Stratified random sampling

In **stratified random sampling**, the population is divided into groups called strata. A sample is then drawn from within these strata. Some examples of strata common to health care are age, gender, diagnosis, ethnic group or geographical area. Stratification is most useful when the stratifying variables are simple to work with, easy to observe and closely related to the topic of the audit.

An important aspect of stratification is that it can be used to select more from one group than another. You may decide how much of your sample comes from each strata if you feel that responses are more likely to vary in one group than another. So, if you know everyone in one group has the same value, you only need a small sample to get information for that group, whereas in another group, the values may differ widely and a bigger sample is required. An example is in the box.

**If you want to audit hand washing among nursing staff and you are interested in looking at the handwashing technique among different groups of staff, the following three groups could be chosen: health care assistants, nurses with basic training in infection control, and nurses with a certificate in infection control. These three groups would become your strata. You might decide that nurses who have a certificate will be able to provide you with more information. So you may decide to take 60% of your sample from that group, and 20% from healthcare assistants and 20% from nurses with basic training.**

**Alternatively, you can select your sample from each strata to reflect the makeup of the population, like a mini reproduction of the population.**

**Of the patients who visit your healthcare organisation, 45% of all patients live in borough A, 35% in borough B and 20% in borough C. You want to select a sample of 500 patients diagnosed with depression. Instead of selecting 500 people randomly from across the boroughs, you could select 45% of your sample from borough A, 35% from borough B and 20% from borough C. Using this method, your sample size will more accurately reflect the geographical distribution of your whole population.**

To calculate a stratified random sample, find out how many cases there are in each strata, then decide how many you need in your sample from each strata and how to select the cases. An example is in the box on the next page.

Using the example on hand washing, say in your hospital there are 1000 members of nursing staff, of which 300 are healthcare assistants, 500 are nurses with basic training in infection control and 200 are nurses with a certificate in infection control. You would follow these steps.

- Calculate the percentage of the total each group comprises (HCAs 30%, nurses with basic training 50% and nurses with a certificate 20%).
- Decide how many of the total population you wish to sample, for example, 200 members of the nursing staff from the total of 1000.
- Calculate the same percentages for each group from the total sample you are going to collect. (For HCAs, 30% of 200 = 60; for nurses with basic training, 50% of 200 = 100; and for nurses with a certificate 20% of 200 = 40.)
- Use simple random sampling to select the staff to be included in the sample for each strata.

#### 5.3.4 Cluster sampling

It is sometimes expensive to spread a sample across the population as a whole. For example, travel can become expensive if you are collecting data from across a wide area. To reduce costs, you could choose a cluster sampling technique.

**Cluster sampling** divides the population into groups or 'clusters'. A number of clusters are selected at random to represent the population and all units in the selected clusters are included in the sample. No units from non-selected clusters are included in the sample. This technique differs from stratified sampling, where some units are selected from each group. An example is in the box.

You are going to undertake an audit on health education in all schools in a county (primary, secondary and special needs). There are far too many schools to include in the audit. Therefore, a cluster sample in one locality could be selected that would include the different types of schools.

Cluster sampling has several advantages including reduced costs, simplified field work and convenient administration. Instead of having a sample scattered over the entire coverage area, the sample is more localised in relatively few areas. The disadvantage of cluster sampling is that less accurate results are obtained due to higher sampling error than for simple random sampling with the same sample size.

#### 5.4 Non-representative sampling techniques

There are also other sampling techniques, that do not follow the random sampling model but are very useful in clinical audit.

### 5.4.1 Purposive sampling

A **purposive sample** is a non-representative sample in which the sample is selected by the auditor for a specific purpose. It may be used when a population can not be specified or when it seems sensible to focus on a particular group even if the group is not selected using a random sampling method. An example is in the box.

A doctor wishes to audit the implementation of a new intervention on patients across a hospital. You might select your sample from the wards that have been using the new intervention for the longest, as staff members working on the ward will have the most experience.

### 5.4.2 Convenience sampling

A **convenience sample** is taking cases you can get. It is an accidental sample that is not randomly selected. Volunteers for participation in a project would constitute a convenience sample. An example is in the box.

You want to learn the views of patients attending a certain clinic. You can't interview everyone. You sit in the waiting room on one particular morning and ask everyone who attends the clinic that day to complete your questionnaire. Those who agree and complete the questionnaire comprise your sample. The sample is made up of people who are simply available in a convenient way to the auditor. There is no randomness and the likelihood of bias is high.

## 5.5 Summary

Non-representative samples are limited with regard to generalisation. Because they do not truly represent a population, we cannot make valid inferences about the larger group from which they are drawn. Validity can be increased by approximating random selection as much as possible and making every attempt to avoid introducing bias into sample selection.

## 6 Glossary of terms

---

**Average** — See **mean**.

**Bias** — How far the statistic lies from the parameter it is estimating, ie, the error that arises when estimating a quantity. Errors from chance will cancel each other out in the long run, those from bias will not.

**Class intervals** — One of several convenient intervals into which the values of the variate of a frequency distribution may be grouped. The set of limits by which data are classified, as in 0–4, 5–9, 10–14, and so on.

**Correlation** — A measure of the strength of linear association between two variables. Correlation will always be between  $-1.0$  and  $+1.0$ . If the correlation is positive, there is a positive relationship. If it is negative, the relationship is negative.

For example, for investigators, there may be two conditions such as age and cholesterol levels, age and blood pressure, or diet consumed and weight of a person. The relationship can be causal, complementary, parallel or reciprocal, and is stated as the correlation coefficient and always reflects the simultaneous change in value of the pairs of numerical values over time. A negative correlation suggests that if one of the variable's values increase, the other variable's values decrease. A positive correlation suggests that both the variables increase in value with an increase in any one of the variables. A correlation coefficient very close to  $0.00$  means the two variables have no correlation, indicating that their statistical relationship is completely random.

**Confidence interval** — The range within which the true size of effect (never exactly known) lies, with a given degree of assurance (95% or 99%)

**Descriptive statistics** — Techniques used to describe the basic features of data. They provide simple summaries about the sample and the measures. Together with simple graphic analysis, they form the basis of virtually every quantitative analysis of data.

**Frequency** — The number of times an observation occurs

**Graph** — A visual display of the relationship between variables

**Histogram** — A bar chart used to show the distribution of a continuous variable, thus, there are no gaps in the class intervals

**Independent events** — Events where the occurrence of one event does not affect the likelihood of the other event occurring

**Inferential statistics** — Techniques used to make judgements on the probability that an observed difference between groups is a dependable one or one that might have happened by chance. These statistics are used to test an inference drawn about a population from a random sample taken from it or, more generally, about a random process from its observed behaviour during a finite period of time.

**Likert scale** — A scale used frequently in patient satisfaction and experience questionnaires. It makes use of a set of ordered responses and it can range from 3 –10 point scales. An example of a three–point Likert scale is: 1. Disagree, 2. Neither agree nor disagree, 3. Agree.

Likert scales also are called summative scales, as scores can be added up for groups. The scales produce ordinal data but could provide interval data. A four–point scale in which the neutral—neither agree nor disagree—option has been removed forces either a positive or negative response and is called the forced choice method. Even numbered sets of responses also remove the neutral option.

Likert scales can induce bias as a result of respondents wanting to agree or be positive or disagree based on the phrasing of the question. Careful consideration needs to be given to phrasing of questions or statements to avoid bias. Data from Likert scales are sometimes reduced to the nominal level by combining all agree and disagree responses into two categories of “accept” and “reject”. The Chi-square can be used to analyse the combined data.

**Mean** — A measure of central tendency in which the sum of all observations is divided by the number of observations; also known as the average

**Normal distribution** — A symmetrical statistical distribution that is bell shaped; also known as a Gaussian distribution

**Population** — Any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in or that we wish to describe or draw conclusions about.

**Parameter** — Value, usually unknown, and therefore has to be estimated, used to represent a certain population characteristic. For example, the population mean is a parameter that is used often to indicate the average value of a quantity.

**Patient Reported Outcome Measures (PROMs)** — Measures of health status or health-related quality of life (HrQL) that are provided directly by patients. The term PROMs is a misnomer as the measures don’t attempt to determine the outcome or impact of a healthcare intervention; they assess a person’s health status or HrQL at a point in time. The impact of a healthcare intervention is determined by comparing the patient’s self reported health status at two points in time; for example, in surgery, the two points in time could be before and after an operation.

**Quartiles** — A measure that divides a distribution into four equal parts. The top 25% is cut off by the upper (3rd) quartile and the bottom 25% is cut off by the lower (1st) quartile; the 2nd quartile is the median.

**Qualitative data** — Non–numerical data. Examples of qualitative data are those that could be generated through interviews, written comments, video or photographic information.

**Quantitative data** — Numerical data that can be analysed using statistical methods

**Random sample** — A sample in which every member of the population has an equal chance of being selected

**Representative sample** — A small quantity of a targeted group such as customers, data, people or products, whose characteristics represent as accurately as possible the entire batch, lot, population or universe

**Statistics** — A branch of applied mathematics concerned with the collection and interpretation of quantitative data and the use of probability theory to estimate population parameters

**Sample** — A group of units selected from a larger group (the population). By studying the sample, it is hoped that valid conclusions about the larger group could be drawn.

## 7 Further reading

---

### **Statistical Notes: *British Medical Journal (BMJ)* series of papers**

[www.bmj.com/cgi/search?&titleabstract=%22statistics+notes%22&&journalcode=bmj&&hits=20](http://www.bmj.com/cgi/search?&titleabstract=%22statistics+notes%22&&journalcode=bmj&&hits=20)

A collection of 45 papers from the BMJ that explain many of the statistical methods and processes used in the medical literature

### ***Statistics at Square One***

<http://bmj.bmjournals.com/collections/statsbk/index.dtl>

The full text of the BMJ's bestselling medical statistics book

### **Sample size calculators**

<http://www.ginns.info/ssc.htm>

<http://www.raosoft.com/samplesize.html>

### **Randomisation**

<http://www.random.org>

Information on selecting random samples, including some useful tools such as a random number generator

### **Excel**

<http://www.bioss.ac.uk/smart/unix/mbasexc/slides/frames.htm>

Introduction to basic statistics in Excel

### **More complex statistics**

[http://onlinestatbook.com/stat\\_sim/](http://onlinestatbook.com/stat_sim/)

More difficult statistical concepts with practical examples of how they work

<http://davidmlane.com/hyperstat/index.html>

A full text statistical textbook, broken down into logical sections with links to sites that explore each concept in more detail and glossary links for all technical terms

## Appendix 1. How to select a random sample using Excel

---

1. Start Microsoft Excel 2007 and open an existing spreadsheet or workbook from your files that contains data you want to use to get a random sample, that is, your list of all cases in your population. Or you could create a new blank spreadsheet into which you want to generate random numbers within a range that you designate.
2. Verify that column A is empty, so you can use it to generate random numbers into. If you have data in column A, then you will need to add a number column so column A becomes empty and can be used to store the random numbers.
3. Click and drag to select the cells in column A that correspond with the records in the other cells. You want to select an empty cell for each row of information you have in the spreadsheet.
4. Type "=RAND()" (no quotes) in the "Formula" textbox near the top of the Excel screen. Press the "Enter" key on your keyboard if you selected one cell or the "CTRL+ENTER" if you have selected multiple cells to generate the random numbers into column A. You will now see random numbers have been generated in the range that you have specified.
5. Select all of the data in your spreadsheet along with the corresponding random numbers. Do not select any titles or headings.
6. Use the "Data" tab at the top of the screen and click the "Sort" button from the "Sort and Filter" group in the "Data" ribbon. The "Sort" dialog box will open onto the screen.
7. Choose "Column A" from the "Sort by Column" drop-down list and "Smallest to Largest" from the "Sort by Order" drop-down list. Click the "OK" button to close the dialog box and return to your spreadsheet. Choose the top number of rows to make up your random sample.

[http://www.ehow.com/how\\_2272795\\_get-random-sample-excel.htm](http://www.ehow.com/how_2272795_get-random-sample-excel.htm).

Accessed 19 March 2010.

Also see:

[http://www.ehow.com/how\\_2272796\\_generate-random-number-excel.html](http://www.ehow.com/how_2272796_generate-random-number-excel.html)

[http://www.ehow.com/how\\_2272797\\_generate-random-number-range-excel.html](http://www.ehow.com/how_2272797_generate-random-number-range-excel.html)

[http://www.ehow.com/how\\_2321632\\_numbers-microsoft-excel-move-control.html](http://www.ehow.com/how_2321632_numbers-microsoft-excel-move-control.html)