HQIP
Healthcare Quality
Improvement Partnership

An introduction to statistics for local clinical audit and improvement

December 2015

HQIP Healthcare Quality Improvement Partnership

Authors:

2015 production

Sally Fereday, Healthcare Quality Improvement Partnership

2010 production

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

Version history:

April 2010:	(first publication)
November 2011:	(revised: minor changes to wording)
December 2015:	(revised: previously "An Introduction to Statistics for Clinical Audit", updated, examples broadened, minor changes to wording)

Next review:
December 2016

Design: Pad Creative www.padcreative.co.uk

Contents

1 Overview

1.1 Background

To improve quality of care it is vital to establish if care provided meets best practice standards. Through quality improvement processes, current practice is measured against best practice, or desired practice, and any shortfalls in care are identified and addressed.

In measurement for quality improvement it is important that the right data are collected and analysed appropriately, to accurately reflect care provided. Results must be clearly presented to highlight any changes needed to practice or service delivery.

Careful statistical data analysis and presentation are essential to ensure conclusions drawn are correct and not misleading. Data must be presented in a balanced and transparent manner that enables scrutiny. Helpful summations are important in reporting key findings with clarity for all potential readers, whether patients, patient representatives, staff or the Board.

1.2 Purpose of this guide

This guide introduces the basics of statistical data analysis and presentation, to support those directly involved in local clinical audit and improvement projects, within health and social care organisations and their individual departments. It aims to:

- Explain how to use descriptive statistical methods to analyse and present quality improvement data
- Provide general principles on how to choose the most appropriate statistical methods
- Demonstrate how to present local statistical data clearly and concisely

The guide is broken down into sections, the first detailing the types of data that exist, followed by various methods of statistical analysis. Further sections address ways to present local data clearly, and guidance on sampling techniques.

1.3 Who is this guide for?

This guide has been designed for individuals working in the field of clinical audit and improvement within local health and social care organisations and their departments. It aims to help those who are new to or unfamiliar with statistical data analysis and presentation, whether patients, patient representatives, staff or Board members.

Statistics are an essential part of measurement in local clinical audit and improvement studies, and must be applied correctly and appropriately. Those working in quality improvement need to be confident in using, interpreting and presenting associated data in order to draw accurate conclusions.

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 must be interpreted to become information.

For example, if one of the standards you are measuring as part of a weekly clinical audit is "all service users admitted to the unit should have a falls risk assessment completed within 24 hours of admission", you will collect data on the number of service users admitted for that week, and the number who had a falls risk assessment completed. In your analysis you will ask the question "How many service users admitted to the unit had a falls risk assessment completed within 24 hours of admission?". If your answer is 47, it may be perfectly accurate data, but it is not information. A more complete and useful answer would be that 47 out of a total of 85 service users had a falls risk assessment on admission that week – this is information.

In many local clinical audit and improvement projects, **frequencies** (the number of service users meeting a standard) and **percentages** (the proportion of service users meeting a standard) are all the information a team will need. However, there are many ways in which information can be presented and explained. It is important to understand the type of data 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 methods you could lead a 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 quality improvement methods involve collecting qualitative data, such as service users' descriptions of how they feel about the treatment they have received. Qualitative data can provide additional insight into the way a service operates or the impact it has on service users. However, the analysis of qualitative data is not covered in this guide.

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

2.2 Continuous data

Continuous data run in a sequence and use real numbers. Theoretically there can 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. Examples of interval data are temperature or year. 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.

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. Examples of data useful in ratios include height, weight, and length of time in hours or minutes. 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.

2.3 Discrete data

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

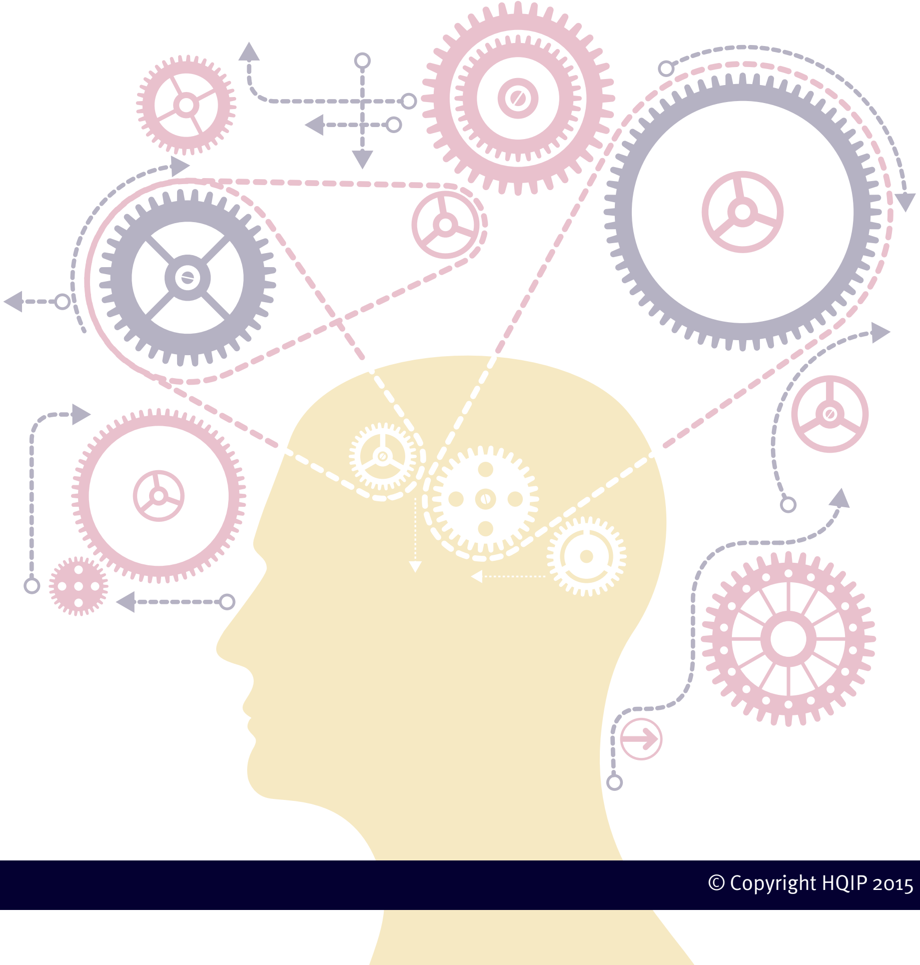
There are two types of discrete data: **ordinal** and **nominal**.

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, such as 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 set out in the table below:

Continuous:	Discrete:
Interval: Temperature, year	**Ordinal:** Level of satisfaction, grade of pressure sore
Ratio: Height, weight, length of time in minutes, age, blood haemoglobin levels	**Nominal:** Gender, marital status, social class, ethnic group, ward, GP practice, blood group



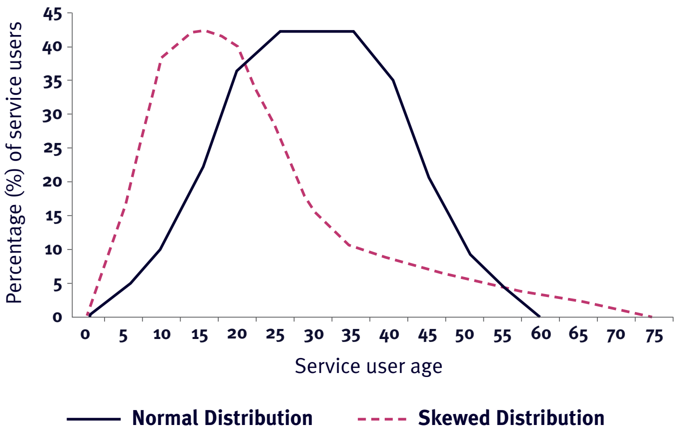
3 Descriptive statistics

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 to help describe a population or sample of individuals, events or circumstances that have been measured or observed.

3.1 Distributions of data

The graph below shows data on the age distribution of two groups of service users:

Age distribution of service users



There are two types of distribution 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 service user 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, where 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

An **average** is simply any single number that represents many numbers. When analysing data it can be difficult to understand significance in large sets of numbers, and therefore it is sometimes more convenient to describe a set of numbers using a single number. Calculating a single number is one of the most frequently used methods of presenting data, for example, providing a single “average”.

3.2.1 Which type of average should you use?

There are three types of averages: **mean**, **mode** and **median**. It is important to understand the type of average to use, which can depend on the distribution of the data.

3.2.2 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 table below - the mean blood haemoglobin level is 366/30 = 12.2g/100ml:

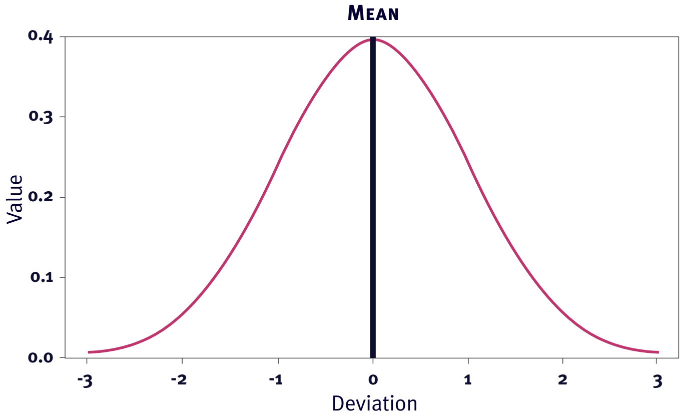
Blood haemoglobin levels (g/100ml) by patient

Patient No.	Hb Level	Patient No.	Hb Level	Patient No.	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					

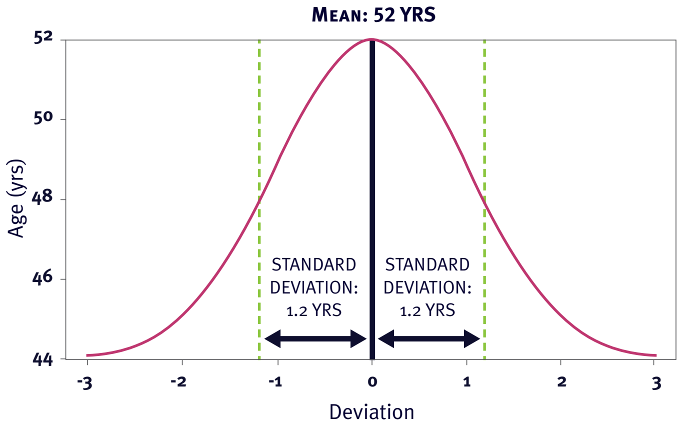
You should only use the mean if you have a normal data distribution. If data are skewed, the mean will be affected by the extreme values and will not give a true representation of what is typical for the sample. Also, only use the mean if there are a reasonable number of observations in your data set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

Standard deviation

If you use the mean for a set of data, you can also use the standard deviation. The standard deviation (SD) gives information about the spread of data - the deviation - around the mean. The value of the standard deviation is relative to the mean. A large standard deviation, when compared to the mean, implies that the data are widely spread around the mean, whereas a small standard deviation implies that the data are mainly concentrated around the mean. The graph below illustrates the normal distribution of data around the mean, known as the “bell curve”. When data are distributed normally, 95.44% of values lie between plus or minus two standard deviations (±2SD) of the mean, as illustrated below and overleaf.



A worked example is set out overleaf.



In a group of patients, the mean age is 52 years and the standard deviation is 1.2 years. Roughly 95% of the patients' ages are between 52 ± 2 (1.2).

52 ± 2SD = 52 ± (2 x 1.2) = 52 ± 2.4 = 49.6 to 54.4 years

This means that roughly 95% of patients are 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 investigated further.

Variance

Variance is the square (a number multiplied by itself) 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:

Variance = s^2 where s^2 is the square of the standard deviation.

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

Variance = $(1.2)^2$ = 1.44

When to use standard deviation or variance

As statistical measures of dispersion of data, both **standard deviation** and **variance** represent how much variation there is from the mean, i.e. to what extent values **deviate** or **vary** from the mean. As we have seen, variance is the square of the standard deviation (difference in values from the mean), and the standard deviation is the square root of that variance. Standard deviation therefore is used to identify outliers in the data, whereas variance is used to examine the spread of the data.

3.2.3 Mode

The **mode** is the most frequently occurring value in a set of data. There can be more than one mode if two or more values are equally frequent. The mode can be relevant if there is a distribution with a double peak or if the data are skewed. Examples are set out below:

Six patients who had a heart attack were admitted to a ward. The number of days they each took to recover was: 8, 10, 12, 12, 14 and 15. The mode is 12 as it occurs twice and is the most frequently occurring value in the data set.

Of a group of children admitted to hospital, the mean length of stay is 10 days, the median (see 3.2.4) is 5 days and the mode is 1 day. The mode is a better indication of practice because both the mean and the median are affected by a few children who have long stays.

3.2.4 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 set out below:

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

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

↑

Median

Range

In analysing and presenting data, it is always important to display the **range** of observations for context, that is, the highest and lowest values in a data set. An example is set out below:

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.

Interquartile range

If you use the median, you should also use the **interquartile range** (IQR). As with the standard deviation, the IQR gives more information about the distribution of values around the median. The IQR includes the middle 50% of the data. Calculating the IQR means dividing each half of the data into further halves. An example is set out below:

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

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

↑ ↑

Interquartile range

IQR = 6 - 1 = 5

The median is less affected than the mean by extreme values; however, the effect of diverse values is reflected in the change in the IQR, as shown in the example below:

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

0, 1, 1, 2, 2, 2, 2, 2, 4, 5, 9, 10, 11, 13, 20

↑ ↑

Interquartile range

IQR = 10 - 2 = 8

Quantiles

Quantiles are a set of cut off points that divide data into groups containing, as far as possible, equal numbers of observations. Examples of quantiles are as follows:

- **Deciles** divide data into ten equal groups, and are the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th and 100th percentiles
- **Quintiles** divide data into five equal groups, and are the 20th, 40th, 60th, and 80th percentiles
- **Quartiles** divide data into four equal groups, and are the 25th, 50th, and 75th percentiles
- The **median** is the 50th percentile

3.2.5 When to use mean, mode and median

The table below summarises when to use mean, mode and median:

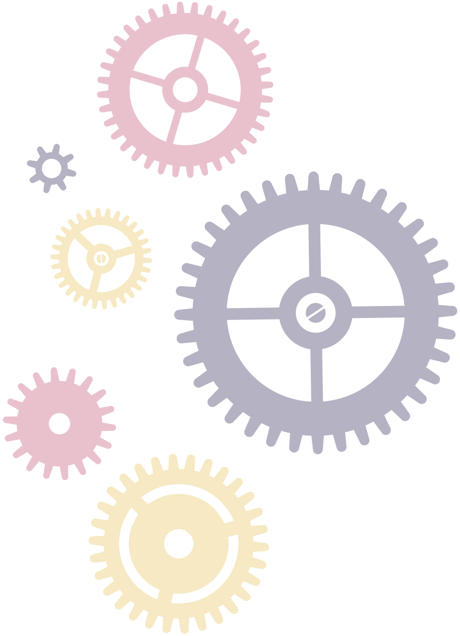
When to use mean, mode and median		
	When to use:	**What to use:**
Mean	• Normal distribution • Reasonable number of observations (30+)	• Standard deviation • Variance • Range
Mode	• Distribution with a double peak • Skewed distribution	• Range
Median	• Skewed distribution	• Interquartile range • Range

3.2.6 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 set out below:

> 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:
>
> **Health Visitor A = 50/145 x 100 = 34.5%**
>
> **Health Visitor B = 80/145 x 100 = 55.2%**
>
> **Paediatrician = 15/145 x 100 = 10.3%**

Conclusions can be drawn, for example, in the above scenario, for Health Visitor A, we might write 34.5% (n145).



4 Ways to present findings

When data has been collected and analysed, findings need to be presented in ways that enable accurate interpretation that is quick and easy.

When you have decided which type of data you are analysing you can choose the most appropriate way to present it. Using the wrong type of presentation for your data can lead to incorrect assumptions.

4.1 Tabular representation of data

The simplest way to present data is in a **table** of **frequencies** and **percentages**. An example is in the table below:

Frequency of method of delivery		
Method of delivery:	**Number:**	**Percentage (%):**
Normal	420	70.0%
Forceps	150	25.0%
Caesarean section	30	5.0%
Total:	600	100.0%

Percentages can be helpful but they can also 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 follows:

You are auditing the completion of falls risk assessments, and two of the standards are that:

- Every service user has a falls risk assessment
- The falls risk assessment is signed by the assessor

You check 100 care records and you find the following:

- 95 of 100 service users have falls risk assessments
- Of the 95 assessments, 85 have been signed

In your report, you should present the following information:

- In five cases, no falls risk assessment was found; therefore, compliance with the first standard is 95/100 (95.0%)
- In 85/95 cases in which a falls risk assessment was found, the falls risk assessment was signed by the assessor

You need to decide how to present the findings for the second standard. If the second standard implies that all service users should have a signed assessment, compliance would be 85/100 (85%). If the second standard does not assume that there should be a signed assessment for every service user, you could subtract the number of signed assessments from the number of assessments completed and report compliance as 85/95 (89%).

4.2 Frequency categorisation

With continuous data, or discrete data that have 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 it into categories. An example is in the table below:

Frequency categorisation of blood haemoglobin levels in women

Haemoglobin (g/100ml):	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%
Total:	70	100.0%

You need to make sure that observations can fall into only one category, for example, 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, otherwise your table will be misleading.

4.3 Graphical representation of data

Graphical representation of statistical data is more interesting, and more easily interpreted and understood than raw statistical data. There are a number of ways in which statistical data can be represented, but the basic methods include:

- 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, i.e. for ordinal or nominal data. The bars are of equal width and there is equal distance between the bars.

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

4.4.1 Simple bar chart

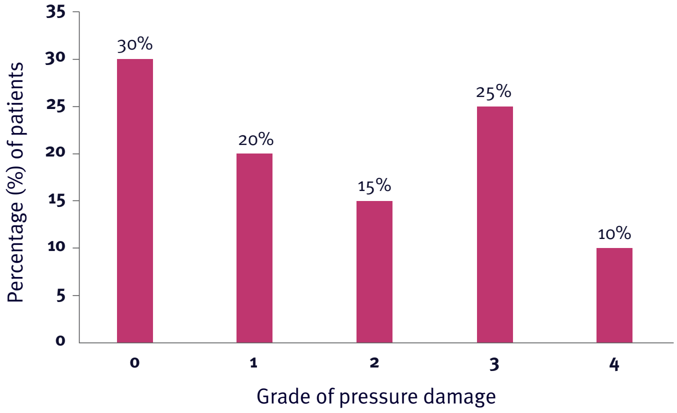
A **simple bar chart** is one in which the bars represent **one** quantity or variable only. The length of the bar indicates the number of people or items in that category. The bottom and side of the chart (known as "x" and "y" axes) must have clear titles. An example is set out below:

Number of patients with pressure damage by grade of damage

Grade of damage:	Number of patients with pressure damage:	Percentage (%) of patients:
0	6	30%
1	4	20%
2	3	15%
3	5	25%
4	2	10%
Total:	20	100%

The data from the table are presented in the bar chart below:

Bar chart of percentage of patients with pressure damage, by grade of damage



From the bar chart, we can see, for example, that 30% of patients had no pressure damage and 25% had grade 3 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 separate bar charts of quantities, which would make immediate comparisons difficult. The multiple bar chart is frequently used in quality improvement studies to demonstrate how data can be compared using one round of measurement and the next. An example of such data is set out in the table below:

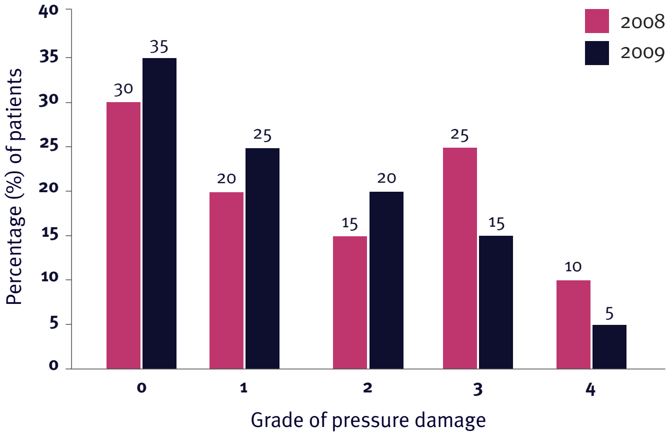
You carried out an audit of pressure damage in 2008 and repeated the audit in 2009. Now you want to display the data in order to compare the findings between one year and the next. Your results are displayed in the table below.

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

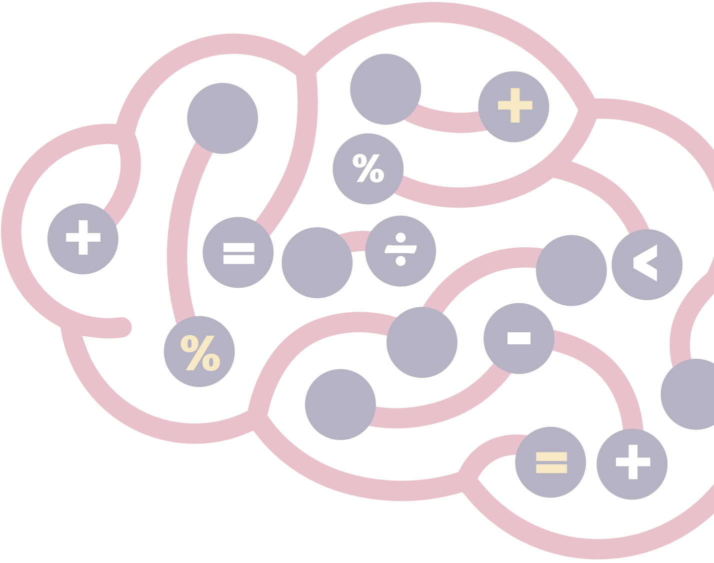
Grade of pressure damage:	2008		2009	
	No. of patients	% of patients	No. of patients	% of patients
0	6	30%	7	35%
1	4	20%	5	25%
2	3	15%	4	20%
3	5	25%	3	15%
4	2	10%	1	5%
Total:	20	100%	20	100%

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

Bar chart of pressure damage, by grade of damage



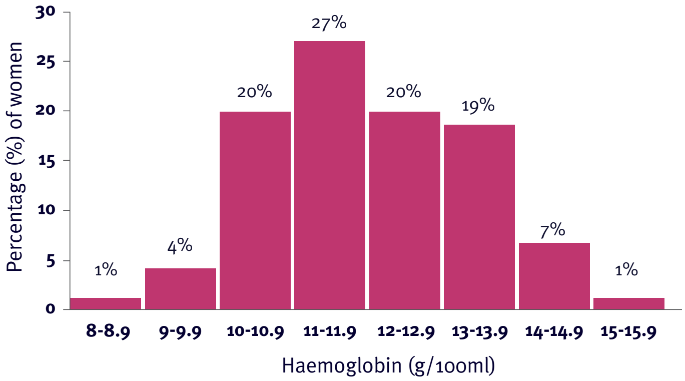
Bar charts can be presented vertically or horizontally; however, it is good practice to be consistent throughout a data analysis report to avoid confusing the reader. It is more common to have the categories across the bottom of the chart (x axis) and the number or percentage value along the side (y axis), as in the example above. 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 set out below:

Histogram of blood haemoglobin levels in women

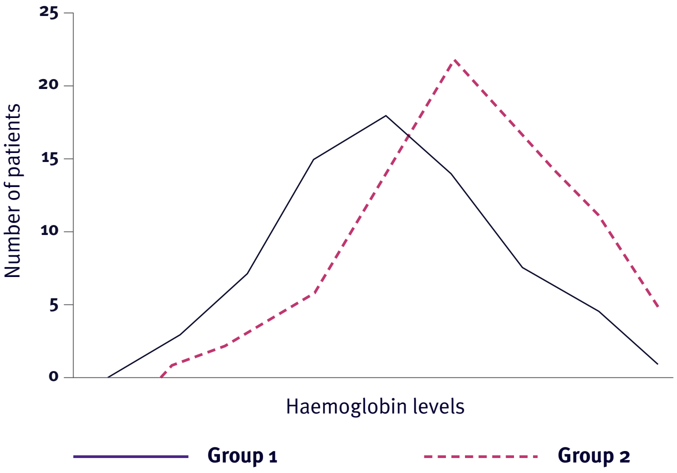


Histograms should not be used for multiple sets of data.

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 that would fall on the line, and with discrete data there are no intermediate values between the categories. An example is set out below:

Graph of Haemoglobin levels in two groups of patients



4.7 Pie chart

Any type of data can also be presented in a circular pie graph or pie chart; however, the chart is usually used 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. The size of each slice is worked out manually by calculating angles; however, the charts are normally constructed using computer software such as Excel, and a link to how to do this is available in the Further reading section of this guide.

Pie charts should only be used when there are three or more categories. An example is set out below:

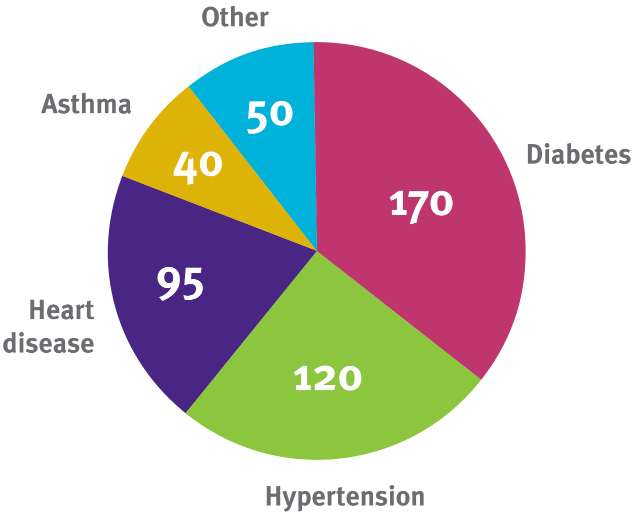
You have been asked to look at the number of people seen in a GP practice who are diagnosed with various diseases. You have put the results into the table below:

Number of patients diagnosed with different diseases:

Disease:	Number:
Diabetes	170
Hypertension	120
Heart disease	95
Asthma	40
Other	50
Total:	475

The results are presented in the following pie chart:

Pie chart of number of patients with each chronic disease



4.8 Summary

It is important to use the right graphical representation method for your data, clearly labelled with a title, and axes labelled completely and accurately. A summary of when to use each method is provided in the table below:

When to use each method

Method:	Use for:
Bar chart	Discrete data — nominal or ordinal
Histogram	Continuous data — interval or ratio
Line graph	Continuous data — interval or ratio
Pie chart	Discrete data — nominal or ordinal

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 often used to refer to all the people, things, items or cases that you are dealing with. An example is set out below:

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

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

5.2 Sampling

Sampling techniques are important in terms of robust quality improvement processes. It is important to understand them in order to choose the most appropriate technique for your study.

A population sometimes includes many individuals, making it inconvenient to include all of them in a quality improvement study. 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. To reflect a whole population we must draw a sample of a smaller number of individuals 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 a quality improvement study. An example follows:

The population is all 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, without electronic records from which to generate data, this would be time-consuming. 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 is drawn.

5.3 Sample size

If generalised conclusions need to be drawn from a quality improvement study sample in relation to a population, the sample has to be **representative**. It must therefore be drawn using a sampling technique to reflect the population, patients, events or situations under study, and **sufficiently large** to enable confidence in the reliability of any assumptions drawn.

A statistical formula (see Appendix 1) can be used to determine what a sample size should be. When the study findings are collated, a statement is added regarding the level of certainty that the true population value falls within a confidence interval. An example is set out below:

For a finding of 84% compliance with a clinical audit standard, using a sample size sufficient for a 95% level of confidence and a 5% range of accuracy, you could say, "I am 95% sure that the true value is 84%±5%, or that the true value lies between 79% and 89%." In other words, you can say that you are 95% confident that the compliance with the audit standard in the entire population would be between 79% and 89%.

See Appendix 1 for guidance on the number of cases for a study sample. The table shown takes into consideration the need for reasonable confidence to allow findings to be generalised from a sample to a population. The formulas for determining the number of cases needed for different confidence levels for any population are also included. It should be noted that this formula is only one of the factors that should be taken into account in determining sample size. Detailed information on the practical considerations that must be taken into account in determining the sample size for a clinical audit project can be found in *HQIP's Guide to Ensuring Data Quality in Clinical Audit*[1].

5.4 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 (or random) sampling techniques, which are described in the sections that follow.

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 a quality improvement study. Four useful common random sampling techniques are:

- Simple random sampling
- Systematic sampling
- Stratified sampling
- Cluster sampling

5.4.1 Simple random sampling

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

Each service user attending a rehabilitation centre over a two-year period could have a number allocated instead of having their 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 papers and 200 numbers could be picked out of a hat. These numbers could then be matched to names on the cardiac outpatient list, thereby providing a random list of 200 people.

[1] http://www.hqip.org.uk/public/cms/253/625/19/191/HQIP-Guide-to-Ensuring-Data-Quality-in-CA-Reviewed%202011.pdf?realName=Zmh8bI.pdf

See Appendix 2 for how to obtain a simple random sample using Excel.

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

5.4.2 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 often used 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 individual is chosen. The technique could also be used to select a fixed-size sample. An example is set out below:

We want to carry out 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 of 100 patients. 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 nth patient on the list (e.g. fifteenth/15th in the above example), 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.

5.4.3 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 healthcare 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 quality improvement study.

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. Examples are set out below:

If you want to audit hand-washing among nursing staff and you are interested in looking at the hand-washing technique among different groups of staff, the following three groups could be chosen: healthcare 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, as a mini reproduction of the population. For example, 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 set out below:

Using the example on hand-washing, 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

5.4.4 Cluster sampling

It is sometimes expensive to spread a sample across the population as a whole. For example, travel can become costly 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 set out below:

You are going to undertake an audit on healthcare in care homes in a county (nursing, residential, and respite). There are far too many care homes to include in the audit. Therefore, a cluster sample in one locality could be selected that would include the different types of care homes.

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.

5.5 Non-representative sampling techniques

There are other sampling techniques that do not follow the random sampling model but are very useful in quality improvement studies.

5.5.1 Purposive sampling

A **purposive sample** is a non-representative sample in which the sample is selected for a specific purpose. It may be used when a population cannot 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 set out below:

A doctor wishes to study the implementation of a new intervention on patients across a hospital. The sample might be selected from the wards that have been using the new intervention for the longest time, as staff members working on the ward will have the most experience.

5.5.2 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 set out below:

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 if they would like to complete your questionnaire. Those who agree to complete the questionnaire comprise your sample. The sample is made up of people who are simply available in a convenient way to the auditor. It is not random, and the likelihood of bias is high.

5.6 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. See *HQIP's Guide to ensuring data quality in clinical audit*[2] for further information on reducing bias in samples.

[2] http://www.hqip.org.uk/public/cms/253/625/19/191/HQIP-Guide-to-Ensuring-Data-Quality-in-CA-Reviewed%20 2011.pdf?realName=Zmh8bI.pdf

6 Glossary of useful terms

Average: See **mean**.

Bias: How far a statistic lies from the parameter it is estimating, i.e., 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.

Chi-square test: When looking at categorical data, statistically significant data (results not occurring by random chance) are found using a chi-square test, which is a statistical test used to compare expected data with that collected.

Class intervals: One of several convenient intervals into which data values 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 which reflects the simultaneous change in value of pairs of numerical values over time.

A negative correlation suggests that if one variable's value increases, another variable's value decreases.

A positive correlation suggests that a variable increases in value with an increase in value of another variable.

A correlation coefficient very close to 0.00 means two variables have no correlation, indicating that their statistical relationship is completely random.

Confidence interval: The range within which the true size of an 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 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 another 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 are also called summative scales, as scores can be added up for groups. The scales produce ordinal data but could provide interval data. An even-numbered set of responses also removes the neutral option—forcing either a positive or negative response, and is called the forced choice method.

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 test 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 forms a bell-shaped curve; 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 which 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 often used 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 middle (2nd) quartile is the median.

Qualitative data: Non–numerical data, for example, that 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 can be drawn.

7 Further reading

Other materials from the Healthcare Quality Improvement Partnership:

A guide to quality improvement methods:

http://www.hqip.org.uk/public/cms/253/625/19/38/Guide-to-quality-improvement-methods-2015-7-1.pdf?realName=3hoUed.pdf

A guide to ensuring data quality in clinical audits:

http://www.hqip.org.uk/public/cms/253/625/19/191/HQIP-Guide-to-Ensuring-Data-Quality-in-CA-Reviewed%202011.pdf?realName=Zmh8bI.pdf

Using root cause analysis techniques in clinical audit

Published in January 2016

Statistics Notes: British Medical Journal (BMJ) series of papers:

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

http://www.bmj.com/specialties/statistics-notes

Statistics at Square One:

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

http://bmj.bmjjournals.com/collections/statsbk/index.dtl

Sample size calculator:

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

Randomisation:

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

http://www.random.org

Excel:

Introduction to basic statistics in Excel:

http://www.real-statistics.com/

How to create charts in Excel:

http://www.ehow.com/how_4441127_make-charts-excel.html

More complex statistics:

More difficult statistical concepts with practical examples of how they work:

http://onlinestatbook.com/stat_sim/

A full 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:

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

A problem-based text for medical students, medical researchers, and others in medical fields who need to use statistics but have no special mathematics background:

Douglas G. Altman, 2011. Practical Statistics for Medical Research, Second Edition. Chapman & Hall

Appendix 1. Selecting and calculating study sample sizes

NB: The numbers in the table below, intended to help determine study sample size, assume an expected incidence of 50% for the thing(s) being measured, that is, you assume that the patient care you are studying happens about half the time. Also, the numbers assume that the data being collected are binomial, i.e. two discrete categories, such as yes or no, or present or absent.

Recommended sample sizes for ±5% accuracy (when you expect the care you are measuring to happen about 50% of the time) and data are binomial			
Population:	**90% confidence ±5% accuracy**	**95% confidence ±5% accuracy**	**99% confidence ±5% accuracy**
<30	all	all	all
30	27	28	29
50	42	44	47
100	73	79	87
150	97	108	122
200	115	132	154
250	130	151	182
300	142	168	207
350	153	183	229
400	161	196	250
450	169	207	268
500	176	217	286
600	186	234	316
700	195	248	341
800	202	260	363
900	208	269	383
1000	213	278	400
2000	238	322	499
3000	248	341	545
4000	253	350	571
5000	257	357	587

If there is more than one measure in a study, the sample sizes could vary for each measure. For example, if there are two standards under examination, and one standard relates to an occurrence likely to happen 50% of the time, whilst the other relates to an occurrence likely to happen 80% of the time, the recommended sample sizes for the same confidence level will vary. In this situation, the larger of the recommended sample sizes may be used, because the larger sample will "cover" both standards under examination, and make it easier to carry out the study.

The formulae in the box below can also be used to calculate sample size:

How to calculate sample size

For 90% confidence level and ±5% accuracy and data are binomial

$$\text{sample size} = \frac{1.645^2 \text{ x N x p(1-p)}}{(0.05^2 \text{ x N}) + (1.645^2 \text{ x p(1-p))}}$$

For 95% confidence level and ±5% accuracy and data are binomial

$$\text{sample size} = \frac{1.96^2 \text{ x N x p(1-p)}}{(0.05^2 \text{ x N}) + (1.96^2 \text{ x p(1-p))}}$$

For 99% confidence level and ±5% accuracy and data are binomial

$$\text{sample size} = \frac{2.58^2 \text{ x N x p(1-p)}}{(0.05^2 \text{ x N}) + (2.58^2 \text{ x p(1-p))}}$$

1.645 = constant for a 90% confidence level

1.96 = constant for a 95% confidence level

2.58 = constant for a 99% confidence level

N = the number in the population

0.05 = the required range of accuracy

p is the percentage of cases for which you estimate the measure of quality will be present (or absent)

Appendix 2. 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.

See:

http://www.ehow.com/how_2272795_get-random-sample-excel.html

Also see:

http://www.ehow.com/how_5901348_randomly-select-rows-excel-spreadsheet.html

HQIP Healthcare Quality Improvement Partnership

Further information is available at: www.hqip.org.uk

ISBN NO 978-1-907561-07-8

6th Floor, Tenter House, 45 Moorfields, London, EC2Y 9AE

T 0207 997 7370 F 0207 997 7398
E communications@hqip.org.uk

www.hqip.org.uk

Registered Office: 70 Wimpole Street, London W1G 8AX
Registration No. 6498947
Registered Charity Number: 1127049

December 2015