GETTING STARTED WITH STATISTICS

The purpose of this short guide is to provide MAX toolkit users with an overview of the key statistical and methodological terms that are used in the analysis guides and tools. **Tip**: users may find it helpful to use the navigation pane (which can be found in the view tab) to identify relevant terms.

TYPES OF DATA

Data are values or pieces of information that are often collected in a raw and unorganised form from surveys. Data can be quantitative or qualitative, and both kinds of data are collected by the Adult Social Care and Carers Surveys.[1]

Quantitative data

Numerical data (e.g. data collected from tick-box responses in the ASCS or PSS SACE).

Qualitative data

Non-numerical, text-based data (e.g. data collected from comments boxes in the ASCS or PSS SACE. In other words, respondents' comments).

DATASETS

Datasets are collections of data. Datasets can be drawn from a variety of sources and can be characterised by a number of properties, including: the types of variables in the dataset (e.g. categorical, continuous); the scales of measurements used to measure the variables (e.g. interval, ratio); and the kinds of statistics that can be used to analyse the dataset. Population and sample are two common forms of datasets referred to in statistics.

[1] The Adult Social Care Survey and Carers Survey will be respectively referred to in this document as the ASCS and PSS SACE.

Population

An entire group of units or people (e.g. all adult social-care service users and carers within a given local authority) to which we want to generalise a set of findings or fit a statistical model. Properties of populations (e.g. mean, standard deviation) are known as **parameters**.

Sample

A smaller group of units or people drawn from a population (e.g. all adult social care service users within a given local authority who completed the ASCS) that are used to make inferences about the population. Properties of samples (e.g. mean, standard deviation) are known as **statistics**.

STATISTICS

Descriptive statistics

Methods of **describing or summarising the dataset** in a meaningful way (e.g. by highlighting key features, trends or patterns) using quantitative measures. The most common types of descriptive statistics are **measures of central tendency** (e.g. mode, median and mean) and **measures of distribution or spread** (e.g. range, standard deviation). Descriptive statistics rely solely on the dataset – in other words, can only describe what has been measured – and **cannot be used to draw conclusions beyond the dataset**.

Inferential statistics

Methods that **allow generalisations about a population to be made from a sample of data drawn from the population**. The methods of inferential statistics are **estimation of parameter(s)** and **testing of statistical hypotheses**, and, in the case of ASCS and PSS SACE analysis, include independent t-tests, analysis of variance (ANOVA) and chi-square tests.

Statistical tests can be classified as **parametric** and **non-parametric**. See the section **Using samples of data to test theories about populations** for further information.

Note: descriptive and inferential statistics can be conducted on the same dataset (e.g. ASCS or PSS SACE data) but, while descriptive statistics treat the dataset as a population, inferential statistics treat the same dataset as a sample drawn from a larger population.

VARIABLES

A variable is any unit or case that can vary (e.g. between people and conditions or across time) and that can be measured, manipulated and controlled for in-data analysis. In the case of the ASCS and PSS SACE, each question can generate one or more variables.

Independent variable

The variable which is **manipulated**, rather than measured, in a statistical test. The value of an independent variable is not affected by any other variable (i.e. it is independent). Also known as a **predictor variable.**

Dependent variable

The variable which is thought to be affected by changes in an independent variable and is **measured**, rather than manipulated, in a statistical test. The value of a dependent variable may be affected by another variable (i.e. it is dependent). Also known as an **outcome variable**.

Confounding variable

A variable which is not controlled for or measured and which may affect the outcome variable.

Example

Using the question ***Do overall social care-related quality of life (SCRQoL) scores (ASCOF 1A) differ between men and women?*** **Gender** is the **independent variable** and is manipulated to assess its effect on SCRQoL (i.e. SCRQoL scores for men and women are allocated to separate groups prior to statistical testing). **SCRQoL** is the **dependent variable** and is measured and compared across both groups to assess the impact of gender. Potential **confounding variables** – in other words, variables other than gender that may affect SCRQoL – include age and health status.

LEVELS OF MEASUREMENT

The level of measurement used to describe a variable depends on the underlying properties of the variable being explored. Broadly speaking, variables can be classified as **categorical** or **continuous.** With the exception of social care-related quality of life (SCRQoL), carer social care-related quality of life (Carer SCRQoL) and age, which are continuous variables, all the variables in the ASCS and PSS SACE are categorical.

Categorical variables

A categorical variable is a factor where each person (case) in the dataset falls into only one of two or more categories. For example, gender in the ASCS or PSS SACE is classified is as male, female or missing. Categorical variables can be further classified as **dichotomous**, **nominal** and **ordinal**.

Dichotomous variables

A variable that has only two categories or levels (e.g. yes/no responses). Also known as **binary variables**.

Nominal variables

Variables that have two or more categories but which do not have any intrinsic order. Examples of nominal variables in the ASCS and PSS SACE include primary support reason, ethnicity and gender.

Ordinal variables

Variables that have two or more categories and can be ordered or ranked. Examples of ordinal variables in the ASCS and PSS SACE include the responses to control over daily life, satisfaction with services, and ease of finding information.

Continuous variables

A continuous variable is a factor where each person (case) in the dataset can take any value on the measurement scale being used. For example, social care-related quality of life (SCRQoL) and carer social care-related quality of life (Carer SCRQoL) in the ASCS or PSS SACE are classified as continuous variables, and a person can generate a score anywhere between 1 and 24. Continuous variables can be further classified as **interval** or **ratio**.

Interval

Variables that can take on any value in a given scale and where the differences between the intervals of that measurement scale represent equal differences in the property being measured (i.e. the intervals are meaningful). The scale, however, does not have a true zero point.

Ratio

Variables that share the same properties as interval variables but where the scale does have a true zero point. Thus, the ratios and the intervals on the measurement scale are meaningful. Age is a good example of a ratio variable.

Variables are measured constructs that vary across units or individual people in a sample. **Parameters,** in contrast, are estimated (rather than measured) from the data.

DISPLAYING DATA

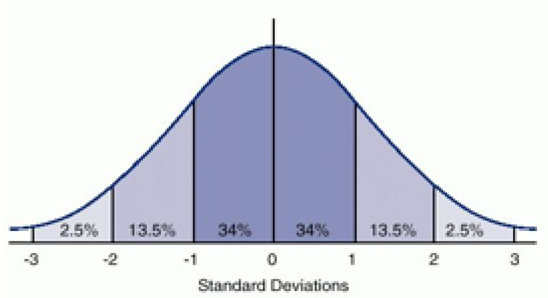
Frequency distribution

A frequency distribution is a graphical representation which plots how many times each score in the dataset occurs. Values are plotted along the horizontal (x) axis and frequencies are represented by columns and are plotted against the vertical (y) axis. Social care-related quality of life (SCRQoL) scores (ASCS) and carer social care-related quality of life (carer SCRQoL) scores are often displayed in reports as frequency distributions.

Normal distribution

The normal distribution is an important class of statistical distribution that is used in inferential statistical tests to represent the frequency distribution of independent, randomly generated variables whose distribution is not known. The normal distribution has two parameters – the mean (the centre of all scores or average, which is valued at 0 in the standard normal distribution) and the standard deviation (the amount of dispersion of scores away from the mean, which is valued at 1 in the standard normal distribution) – and is based on the assumption (known as

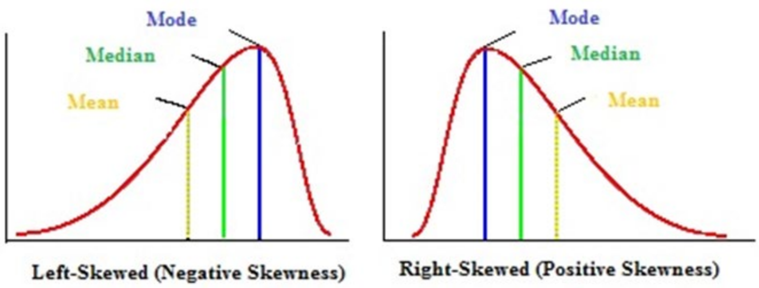
central limit theorem)[2] that a distribution will fall symmetrically around the mean to produce a **bell-shaped curve** if a sufficient sample is drawn. This means that most values will be grouped near the centre of the distribution and the remaining values will tail off away from the mean in equal measures.



The **normal distribution** which produces a bell-shaped curve and plots the percentage of the data-set that should fall within a given range.

Skew

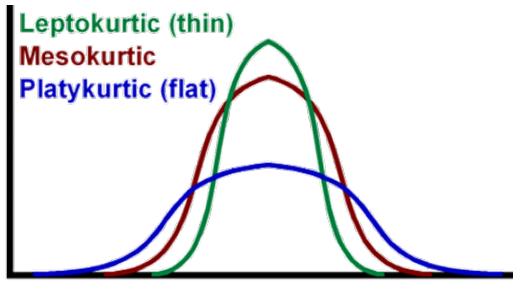
Measures the **symmetry** of a frequency distribution (i.e. the degree to which the scores in a dataset cluster equally around the centre). A frequency distribution with perfect symmetry has a skew of 0. Most distributions, however, are not perfect, and the most frequently occurring scores cluster at the lower end/left side of the distribution (**positively skewed distributions**) or at the higher end/right side of the distribution (**negatively skewed distributions**).



Kurtosis

Measures the **pointiness** of a frequency distribution (i.e. the degree to which scores in a dataset cluster in the tails of the frequency distribution). A frequency distribution with perfect pointiness has a kurtosis of 0. Most distributions, however, are not perfect and have too many scores in the tails of the distribution (which produces a **peaked or leptokurtic curve**) or too few scores (which produces a **flatter or platykurtic curve**).

[2] **Central limit theorem** states that the sampling distribution of large samples (30+) will form the shape of a normal distribution irrespective of the population from which the sample was drawn.



DESCRIBING DATA

Datasets are most commonly described using the centre and dispersion of the distribution.

The centre of the distribution

The centre of the distribution (or central position within a dataset) are described using measures of central tendency such as the **mean**, **median** and **mode**. These measures are classed as summary statistics and the most appropriate description depends on the underlying properties of the data being described.

Mode

The **most frequently occurring score** in a dataset. Represents the tallest bar(s) in a frequency distribution.[3] Best used to describe nominal data.

Median

The **middle score** when scores in a dataset are ranked in order of magnitude. The median is relatively unaffected by extreme scores and skewed distributions, and can be used with ordinal, interval and ratio data. Best used with ordinal data and skewed interval and ratio data.

Mean

The **average score** in a dataset. The mean is the most commonly reported measure of central tendency but is affected by extreme scores and skewed distributions and can only

[3] Some datasets will have more than one mode. Datasets with two modes are known as **bi-modal** while datasets with more than two modes are known as **multi-modal**.

be used with interval or ratio data. Unlike the mode and median, however, the mean uses all the scores in a dataset. Best used with interval and ratio data that is not skewed.

The dispersion in a distribution

The dispersion of the distribution refers to the spread of data around the centre of the dataset and are most commonly described using the **range** and **standard deviation**.

Range

The range measures the spread of scores from the lowest to the highest score. A range of scores can also be divided into equal portions known as **quantiles**. The most common quantiles are **quartiles**, which divide the data into four portions, and are used to deal with extreme scores.

Standard deviation

A measure of the spread or variance of the scores in a dataset around the mean. Standard deviation (SD) provides a measure of how well the mean represents the observed (sample) data: a small SD indicates that most of the scores in a dataset are close to the mean and that the mean therefore is a good representation of the sample. A larger SD, in contrast, indicates that the scores are spread widely around the mean and that the mean therefore is not such a good representation of the sample.

USING SAMPLES OF DATA TO TEST THEORIES ABOUT POPULATIONS

Samples of data, such as those collected by the ASCS and PSS SACE (which are based on a sample of adult social care service users or carers), can be used to test theories about the population to which they belong (in the case of the ASCS and PSS SACE, the entire population of adult social care service users or carers within a given local authority).

Hypothesis testing

Hypothesis testing involves using inferential statistical tests such as chi square or ANOVA to determine whether the patterns observed in a sample of data (e.g. the sample of respondents who completed the ASCS or PSS SACE) are likely to reflect those that would be found in the population (e.g. the entire population of adult social care service user or carers within a given LA).

A hypothesis test explores two hypotheses about a population - the **null hypothesis** and the **alternative hypothesis** – and determines which one should be accepted as true.

Null hypothesis

The null hypothesis states that there is **no effect or difference** between the variables of interest. In other words, the frequency counts observed in the data have arisen 'by chance'. For example, ***there is no difference in the proportion of men and women who say they have as much control over their daily life (ASCOF 1B) as they would like***.

Alternative hypothesis

The alternative hypothesis states that there **is an effect or difference** between the variables of interest. In other words, the frequency counts observed in the data have arisen 'by something other than chance' (e.g. an underlying factor). For example, ***there is a difference in the proportion of men and women who say they have as much control over their daily life (ASCOF 1B) as they would like.***

The statistical test used to test a hypothesis generates a **p-value** which enables you to determine whether or not to reject the null hypothesis. This p value is the probability that the null hypothesis is true. If the p value is less than a set significance level (usually 5% [0.05] or 1% [0.001]), it is said that the pattern is 'statistically significant'. This means that there is a low probability that the observed pattern arose by chance.

A word of caution

Hypothesis testing is not 100% certain so there is a risk of drawing an incorrect conclusion.

Type I errors: occur when the null hypothesis is rejected when, in fact, it is true (i.e. you conclude your data shows there is an effect or difference when there is none). The risk of generating a type I error can be reduced by using a lower significance value (e.g. 0.001 rather than 0.05)

Type II errors: occur when the null hypothesis is accepted when, in fact, it is false (i.e. you conclude your data shows no effect or difference when it actually does). The risk of generating a type II error can be reduced by ensuring your sample size is sufficient. Guidance notes are provided alongside the statistical tests explored in the MAX toolkit.

Two-tailed hypothesis: a hypothesis which predicts that there will be an effect or difference but which does not specify the direction of that effect or difference. For example, ***there is a difference in the proportion of men and women who say they have as much control over their daily life (ASCOF 1B) as they would like.***

One-tailed hypothesis: a hypothesis which predicts that there will be an effect or difference and which specifies the direction of that effect or difference. For example, ***more women will say they have as much control over their daily life (ASCOF 1B) as they would like.***

Statistical tests

Statistical tests are used to explore samples of data and can be classified as **parametric** and **non-parametric**.

Parametric

Parametric statistical tests make assumptions about the parameters (defining properties) of the population distribution(s) from which the sample of data are drawn (e.g. sampling distribution is normally distributed, observations are independent, etc.).
Examples of parametric tests: independent t-test; independent ANOVA.

Non-parametric

Non-parametric statistical tests make no assumptions about the parameters (defining properties) of the population distribution(s) from which the sample of data are drawn. Such tests use less information and are therefore less powerful than the parametric equivalents.
Examples of non-parametric tests: spearman; Mann-Whitney test.

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Assumed variance	Same/similar	Any
Type of data	Ratio or interval	Ordinal or nominal

	Parametric	Non-parametric
Tests		
Correlation	Pearson chi-square	Spearman
2 independent groups	Independent t-test	Mann-Whitney test
2+ independent groups	Independent ANOVA	Kruskal-Wallis test
2 repeated measures groups	Matched-pair t-test	Wilcoxon test
2+ repeated measures groups	Repeated measures ANOVA	Friedman's test

Comparison of parametric and non-parametric statistical tests

The analysis tools included in the MAX toolkit focus on independent group parametric tests (i.e. independent t-tests and ANOVA) and chi-square.

DISCLAIMER

The MAX toolkit and website are based on independent research commissioned and funded by the NIHR Policy Research Programme (Maximising the value of survey data in adult social care (MAX) project and the MAX toolkit implementation and impact project). The views expressed on the website and in publications are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care or its arm's length bodies or other government departments.