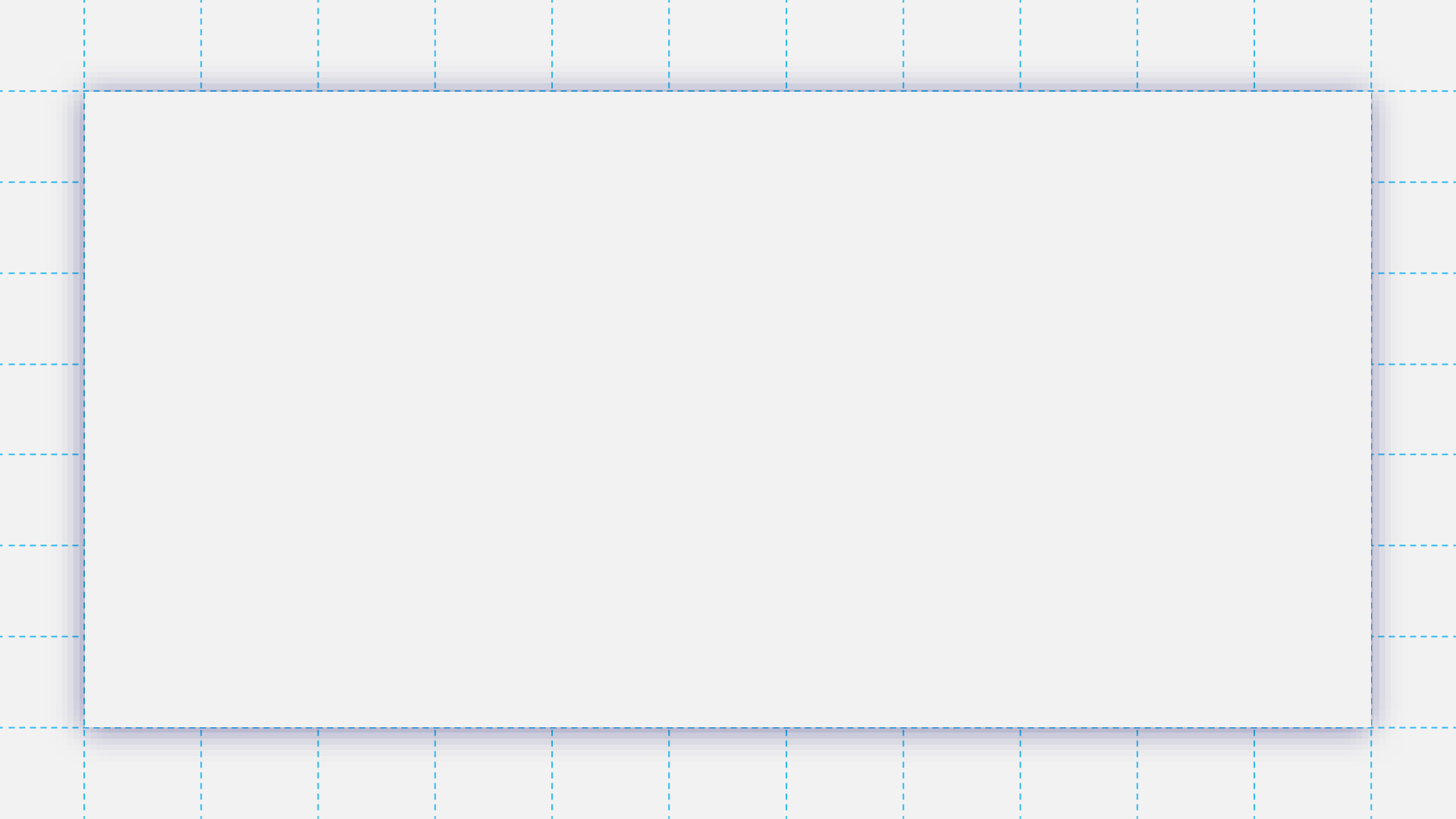


Statistics

Statistics is the science that is involved with the collection, presentation, analysis and interpretation of data



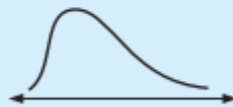


Key terms

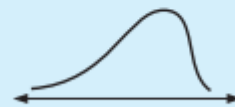
- **Population** - an entire collection of individuals about which we want to draw conclusions
- **Census** - the collection of information from the **whole population**
- **Sample** - a subset of the population which should be chosen at **random** to avoid **bias** in the results
- **Survey** - the collection of information from a **sample**
- **Data** - information about individuals in a population
- **Categorical variable** - describes a particular quality or characteristic which can be divided into categories
- **Numerical variable** - describes a characteristic which has a numerical value that can be counted or measured
- **Parameter** - a numerical quantity measuring some aspect of a population
- **Statistic** - a quantity calculated from data gathered from a sample, usually used to estimate a population parameter
- **Distribution** - the pattern of variation of data, which may be described as:



symmetrical



positively skewed



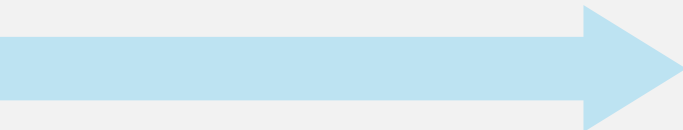
negatively skewed

- **Outliers** - data values that are either much larger or much smaller than the general body of data; they should be included in analysis *unless* they are the result of human or other known error



SAMPLING

process of obtaining subjects who are representative of population events, behaviours or other elements with which to conduct study



Sampling techniques

Simple random sampling: every member of the population is equally likely to be chosen. For example, allocate each member of the population a number. Then use random numbers to choose a sample.

For example, if 20% of students in a school were in Grade 7, then you would choose 20% of your sample from Grade 7.

Systematic sampling: find a sample of size n from a population of size N by selecting every k th member where $k = \frac{N}{n}$ to the nearest whole number.

For example, asking people at a shopping centre to fill out a survey until you have 50 responses.

Stratified sampling: is selecting a random sample where numbers in certain categories are proportional to the numbers in the population.

For example, the school canteen is considering introducing a new lunch menu and would like feedback from the students. The school has 250 boys and 300 girls and so the canteen manager decides to interview 25 boys and 30 girls to find out their opinion of the new menu. He stands at the entrance to the canteen and interviews the first 25 boys and 30 girls that come into the canteen.

Quota sampling: decide how many members of each group you want to sample and take samples from the population until you have a large enough sample for each group.

Convenience sampling: take samples from the members of the population that you have access to until you have a sample of the desired size.

For example, choosing every 15th student from the school register to find a sample of 100 from a population of 1500.

Examples

A chocolate factory produces 80 000 blocks of chocolate per day. Today, the factory operator wants to sample 2% of the blocks for quality testing. He uses a systematic sample, starting from the 17th block.

- a List the first five blocks to be sampled.
- b Find the total size of the sample.

An annual dog show averages 3540 visitors. The catering manager is conducting a survey to investigate the proportion of visitors who will spend more than €20 on food and drinks at the show. He decides to survey the first 40 people through the gate.

- a Identify the sampling method used.
- b Discuss any problems with the sampling method.
- c Suggest a better sampling method that includes a suitable sample size and which better represents the population.

Examples

A sporting club wants to ask its members some questions about the clubhouse. The club has 80 tennis members, 60 lawn bowls members, and 20 croquet members.

- a How many members does the club have in total?
- b The club decides to use a sample of 40. How many members of each sport should be sampled?

Examples

Mona wants to gauge the opinions of her peers on the design of the school's yearbook. She uses her own home room class as her sample.

- a** Explain why Mona's sample is a convenience sample.
- b** In what ways will Mona's sample be biased?
- c** Suggest a more appropriate sampling method that Mona should use.

Types of Data

Classify each variable as categorical, discrete, or continuous.

If the variable is categorical, list some possible categories.

If the variable is quantitative, suggest possible values or a range of values the variable may take.

- a** The number of brothers a person has.
- b** The colours of lollies in a packet.
- c** The time children spend brushing their teeth each day.
- d** The heights of the trees in a garden.
- e** The brand of car a person drives.
- f** The number of petrol pumps at a service station.
- g** The most popular holiday destinations.
- h** The scores out of 10 in a diving competition.
- i** The amount of water a person drinks each day.
- j** The number of hours spent per week at work.
- k** The average temperatures of various cities.
- l** The items students ate for breakfast before coming to school.
- m** The number of televisions in each house.

Errors in Sampling

Sampling error occurs when a characteristic of a sample differs from that of the whole population. This error is random, and will occur even for samples which are well-chosen to avoid bias.

Coverage errors occur when a sample does not truly reflect the population we are trying to find information about.

Non-response errors occur when a large number of people selected for a survey choose not to respond to it. For example:

- An online survey is less likely to be completed by elderly people who are unfamiliar with technology. This means that elderly people will be under-represented in the survey.
- In surveys on customer satisfaction, people are more likely to respond if they are dissatisfied.

Measurement error refers to inaccuracies in measurement at the data collection stage. For example:

- When we record a person's height to the nearest centimetre, the recorded height is slightly different from the person's *exact* height.
- If the questions in a survey are not well worded, they may be misunderstood and produce answers which are not relevant to the question. Survey construction is discussed more thoroughly in

Errors in Sampling

A new drug called Cobrasyl has been developed for the treatment of high blood pressure in humans. A derivative of cobra venom, it is able to reduce blood pressure to an acceptable level. Before its release, a research team treated 7 high blood pressure patients with the drug, and in 5 cases it reduced their blood pressure to an acceptable level.

Do you think this sample can be used to draw reliable conclusions about the drug's effectiveness for all patients? Explain your answer.

A polling agency is employed to investigate the voting intention of residents in a particular electorate. From the data collected, they want to predict the election result for that electorate in the next election. Explain why each of the following situations may produce a biased sample:

- a** A random selection of people in the local large shopping complex is surveyed between 1 pm and 3 pm on a weekday.
- b** The members of the local golf club are surveyed.
- c** A random sample of people at the local train station between 7 am and 9 am are surveyed.
- d** A door to door visit is undertaken, surveying every voter in a particular street.