Samples and distributions

# Samples and sample sizes

# Inferential statistics and samples

A **sample** is a **smaller subset of the larger population** that shares the fundamental characteristics of the larger population from which inferences can be made.

**Inferential statistics** is a way of using data analysis to infer the characteristics of a population based on a sample of the population.

A **representative sample** is one that objectively and fairly represents or is comparable to the population.

# Example

Imagine we were tasked to study the **number of households that receive different types of social grants** from the South African government.

## We need...

A **data collection method** that enables us to collect accurate and reliable data to perform statistical analysis on and from which we can draw insights.

## The challenge

There are approximately **60.1 million people** in South Africa and an estimated **18 million households**. It would be too **costly** and **time-consuming** to collect data from each and every household.

# Generating samples

Things we need to consider when **generating a sample**

## 01. Sample size

**Example:**

**How many households** do we include in the sample?

## 02. Sampling technique

**Example:**

**Which households** should we use in the sample?

## 03. Sampling bias

**Example:**

Does the sample accurately **represent the entire population**?

# Sample size

A sample size is the **number** of people, units, or survey responses **included in the sample**.

## A larger sample size

**Pros:**

- Enables **more accurate estimates** to be made about the rest of the population.
- Increases the **statistical power** of a study, making it more likely for significant differences between groups or conditions to be detected.

**Cons:**

- Makes research **more expensive**, **time-consuming**, and **resource intensive**.

## A smaller sample size

**Pros:**

- More **cost-effective**, **time efficient**, and **manageable**.
- The **measurement errors** can easily be **identified** and **controlled**.

**Cons:**

- **Decreases accuracy** because sample results may not be generalisable to a broader population.

# The minimum sample size

We can calculate the **minimum sample size** required at a certain **margin of error** and **confidence level**.

When we use this formula, we assume that the **population is normally distributed.**

There are numerous internet resources that provide **sample size calculators**, so we typically don't need to manually calculate this minimum sample size.

However, we do need to **understand some of the variables that make up this formula** since determining the minimum sample size requires careful consideration of multiple factors.

$$n = \frac{(Z^2 \times p \times (1-p))/e^2}{(N-1) + (Z^2 \times p \times (1-p))/e^2}$$

Where:

**n** is the required sample size;
**N** is the population size;
**Z** is the Z-score associated with the desired confidence level;
**p** is the **sample proportion**, which is the **estimated** proportion of the sample who are **expected to share a specific trait**, e.g. the households receiving grants. Usually **set at 50%**; and
**e** is the margin of error.

# Z-score

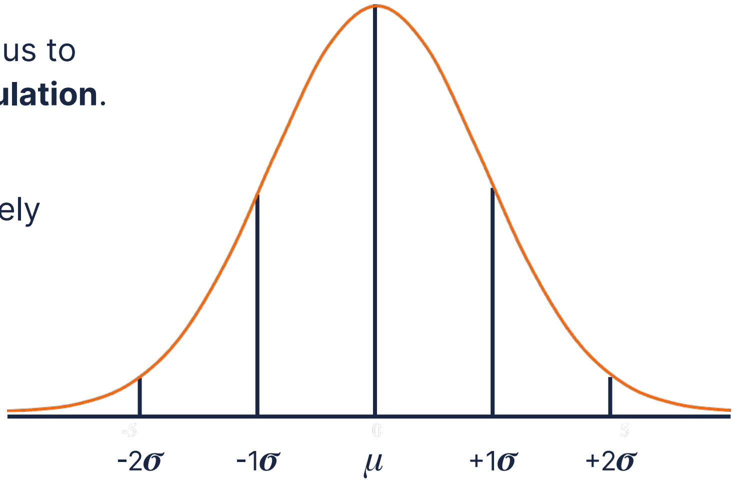
> The Z-score indicates the **number of standard deviations** ($\sigma$) between any **value in the sample** and the **mean of the population** ($\mu$). The higher the Z-score, the further from the population the sample can be considered to be.

## What is the purpose of this Z-score?

The Z-score is included in the sample size formula because it helps us to **set the sample size based** on the **variability that exists in the population**.

When there is greater variability within the population, the Z-score increases, indicating that a larger sample size is required to accurately represent the population.

We choose an appropriate Z-score based on the **desired level of confidence** we want to have in our results.



# Confidence level and margin of error

> The **confidence level** and **margin of error** indicate how **confident we are in the results** of our population statistical study and how much **error we are willing to accept**.

## Confidence level

The **degree of certainty** or **probability** that the results of the study will fall within the margin of error. If the confidence level is set at 95%, there is a 95% chance that the results of the study will fall within the margin of error. The higher the confidence level, the larger the sample size required.

## Margin of error

The **maximum amount of error that is acceptable** in the results of a study. It is typically expressed as a percentage and indicates the degree of precision required for the study. The smaller the margin of error, the larger the sample size required.

# Confidence level and margin of error

> The confidence level and margin of error required depend on a balance between the **research question** and the **resources available**.

## Research question

In a **clinical study** of the side effects of vaccines, you would want to have a **very low margin of error** and a **high degree of confidence** that the vaccine is unlikely to cause any deaths.

However, you would be completely content with a **lower confidence level and a higher margin of error** when studying the number of households receiving government grants since the **danger of getting it wrong will not have any serious repercussions.**

## Resources available

A **larger sample size** (which requires more resources) provides a **more accurate** representation of the population, which will result in a **higher confidence level** and **lower margin of error**.

However, if we don't have enough resources to collect the required samples, we might need to settle for a lower confidence level.

# Interpreting the confidence level and margin of error

## Vaccine example:

Let's say we manage to collect data from a **sample of 5000 patients** and our results show that 3,750 patients (**95%**) did not experience side effects from a vaccine.

If we set a **margin of error of 5%** and a **confidence level of 80%**, we can interpret the results by saying:

"We are **80% confident** that **90–100%** (that is 95% +/- the margin of error of 5%) **of patients** did not experience side effects from the vaccine, therefore, it's safe to use."

Based on these results, the **level of confidence is not high enough** to **warrant using this vaccination**. And since this has to do with the health and life of people, we **cannot take the risk**.

## Social grants example:

Since we are investigating the number of households receiving **social grants** from the government, we can afford to **accept this higher margin of error** and **low confidence level** because, if our estimates are off, there won't be any serious **repercussions** that cannot be undone or re-engineered.

# Sampling techniques

The second consideration we make is which households we should use in the sample.
Here are a few common techniques **used to select which individuals or units to include in a sample**.

## Simple random sampling

Samples are **randomly selected** and every member of the population has an equal chance of being selected for the sample.

## Stratified sampling

The population is **divided into homogeneous groups**, and a sample is selected from each group **in proportion to its size**.

## Systematic sampling

The population is **divided into groups**, and a **sample is selected from each group regardless of its size**.

## Cluster sampling

The population is **divided into clusters**, and a **sample of clusters is randomly selected**. Then a **sample is selected from each selected cluster.**

# Sampling biases

Sampling bias is the tendency for a **sample to systematically differ** from the population it is drawn from, leading to **inaccurate conclusions**. The most common sampling biases include:

## 01. Selection bias

Occurs when the sample is selected in a way that **does not represent the entire population**.

**For example:**
A researcher only samples from one city out of the whole country.

## 02. Measurement bias

Occurs when the **data collection method** or instrument used to measure a variable **is flawed or inaccurate**, leading to inaccurate conclusions.

**For example:**
There is a flaw in the survey used to collect population data.