



Session 6.5: Review of Statistical Methods

Topics

- 1 Prior to Engaging with Statistics
- 2 General Overview of Statistics
- 3 Descriptive and Inferential Statistics
- 4 Data Distribution, Sampling Issues and Sources of Bias
- 5 Hypothesis Testing

- 1 Prior to enegaging with Statistics
- 2 General Overview of Statistics
- 3 Descriptive and Inferential Statistics
- 4 Data Distribution, Sampling Issues and Sources of Bias
- 5 Hypothesis Testing

1. Prior to Engaging with Statistics

- A. What are the research questions, objectives and hypothesis?
- B. What will be the data requirement?
- C. Is there is a bias in the process?
- D. Arrive at a judgement of what is feasible

1 Prior to enegaging with Statistics

2 General Overview of Statistics

3 Descriptive and Inferential Statistics

4 Data Distribution, Sampling Issues and Sources of Bias

5 Hypothesis Testing

2. General Overview of Statistics

- Statistics could be described as:
 - designing or choosing appropriate ways of collecting data and extracting information from them;
 - exploring, analysing and summarizing data;
 - constructing and testing models which can be used as a basis to make inferences and drawing conclusions;
- A large part of statistics is concerned with statistical inference:
 - *Making assertions about population(s) from sample(s)*

2. General Overview of Statistics

- Statistics is not just the dry collection of facts; it is the science of making sense of the facts we have
- Statistics aim to cope with uncertainty not in producing certainty
- Precision is often unlikely
- **Statistic** refers to a quantity (such as mean or median) calculated from a set of data.

1 Prior to engaging with Statistics

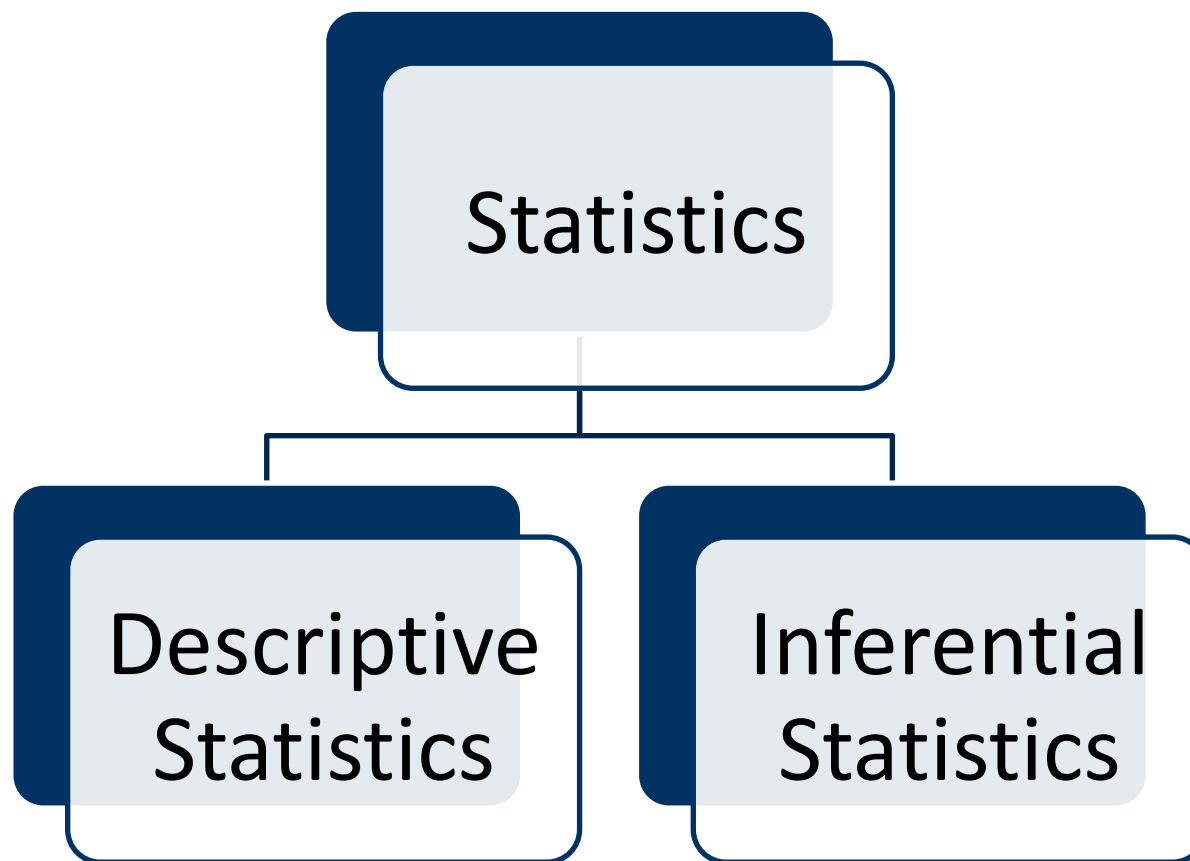
2 General Overview of Statistics

3 Descriptive and Inferential Statistics

4 Data Distribution, Sampling Issues and Sources of Bias

5 Hypothesis Testing

Descriptive and Inferential Statistics



Descriptive and Inferential Statistics

- Descriptive Statistics uses raw numbers, tables and graphs to reveal patterns and trends in a data set. The patterns and trends summarizes the information and presents it a manner that is easily interpretable.
- Some examples of descriptive statistics are the measures of location and spread and some of the graphs are histograms, scatter plot and so forth.

Descriptive and Inferential Statistics

- Inferential Statistics uses sample data to make estimates, decisions, predictions or other generalizations about a larger set of data.
- Some examples of inferential statistics are analysis of variance, chi-square, ordinary least squares and so forth.
- Core issue - degree of reliability (extent of uncertainty associated with the statistical inference)



1: Measures of location (averages) and spread

Datasets are characterised by their:

- Location/level - Size of the measured values

Example: What is the “average” reaction time?

- Spread/dispersion – How much do the measured values vary

Example: How spread out are reaction times?

Measures of location Averages

Mean

- Uses all of the data
- Suitable for further analysis
- **Affected by a few extreme values**

Median

- Not affected by a few extreme values
- **Does not use all of the data**
- **Not particularly suitable for further analysis**

Mode

- Simple
- **Bimodal and multimodal data is common**
- **Not particularly suitable for further analysis**

Each average has a different use

- **Standard deviation** (represented by the symbol σ) shows how much variation or "dispersion" exists from the average (mean, or expected value).
- A low standard deviation indicates that the data points tend to be very close to the mean; a high standard deviation indicates that the data points are spread out over a large range of values.
- **Variance** is a measure of how far a set of numbers is spread out
- Variance of X is given by:

$Var(X) = E[(X - \mu)^2]$ if a random variable X has the expected value (mean) $\mu = E[X]$

Range	<ul style="list-style-type: none">▪ Easy to understand▪ Does not use all of the data▪ Not suitable for further analysis▪ Affected by extreme values
Inter-quartile range (IQR)	<ul style="list-style-type: none">▪ Difference between the upper and lower quartiles $IQR = Q_3 - Q_1$▪ Not affected by a few extreme values▪ Does not use all of the data▪ Not particularly suitable for further analysis
Variance	<ul style="list-style-type: none">▪ Uses every value▪ Suitable for further analysis▪ Affected by a few extreme observations

Coefficient of variation (CV) measures the **spread** of a set of data as a proportion of its mean.

It is the **ratio** of the sample **standard deviation** to the sample **mean**

Shape of the Distribution

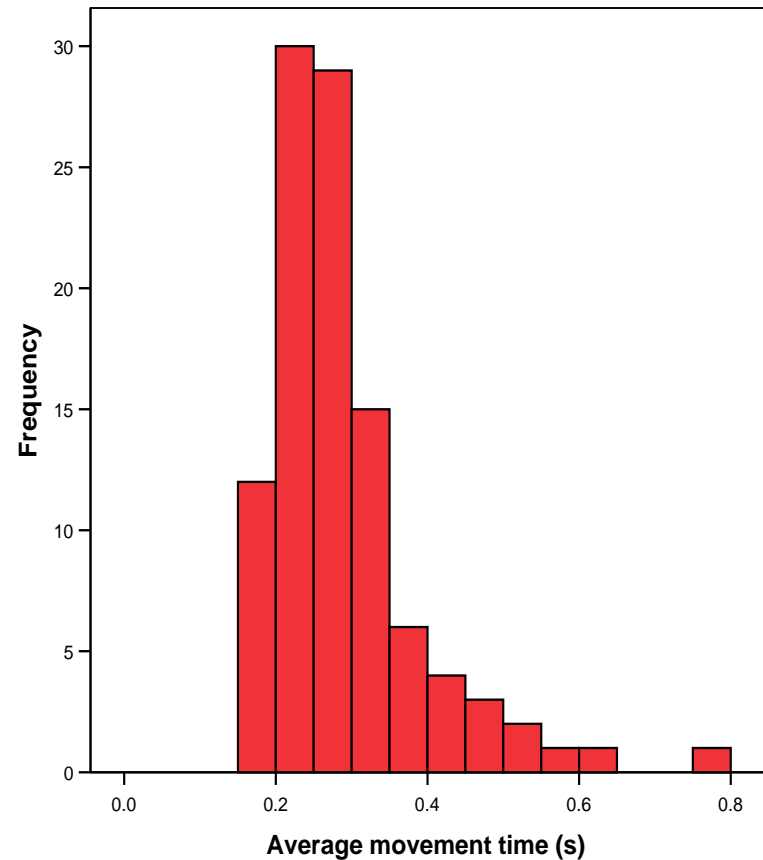
- Histogram
 - Summarizes the distribution of a dataset
- Skewness
 - Measures the degree of asymmetry exhibited by the data
 - If skewness equals zero, the histogram is symmetric about the mean
 - Positive and negative skewness
- Kurtosis
 - Measures the peakedness of the histogram
 - The kurtosis of a normal distribution is 0



Recommendations on use of summary statistics

Recommendations

- Roughly symmetrical with a single peak:
 - Use mean and standard deviation
- Markedly skewed with a single peak:
 - Use median and IQR



1 Prior to engaging with Statistics

2 General Overview of Statistics

3 Descriptive and Inferential Statistics

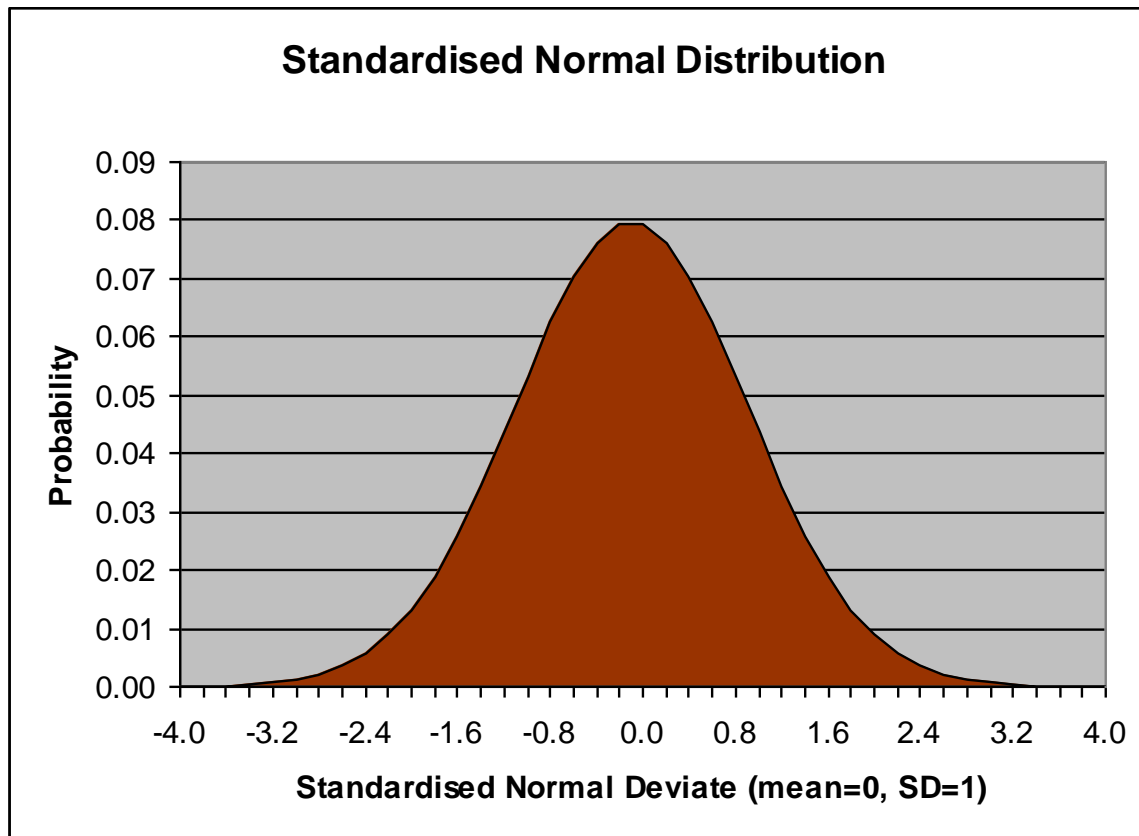
4 Data Distribution, Sampling Issues and Sources of Bias

5 Hypothesis Testing

- *Most common underlying probability distribution for continuous variables*
- *The most important distribution for statistical analysis purposes*
- A plot of the normal distribution roughly follows a bell-shaped curve

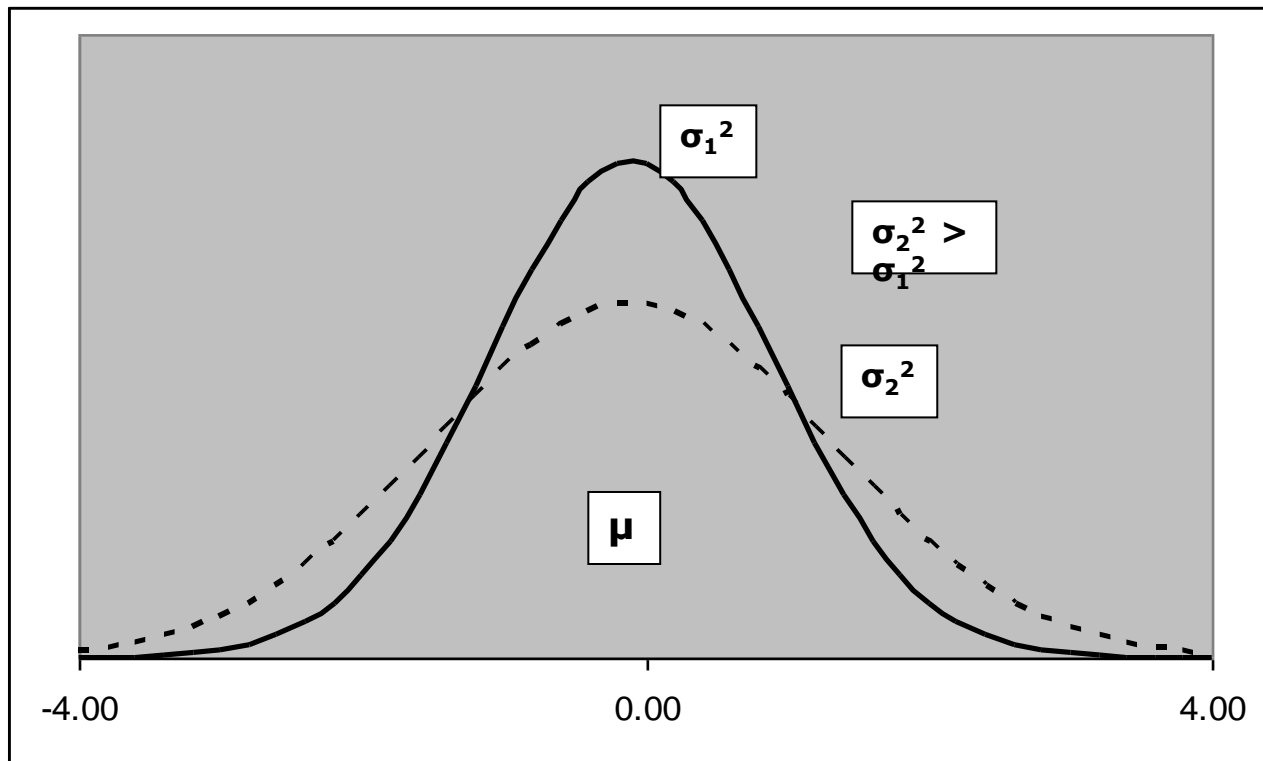
The normal distribution is characterised by:

- Mean: μ
- Variance: σ^2



The normal distribution is written as $N(\mu, \sigma^2)$

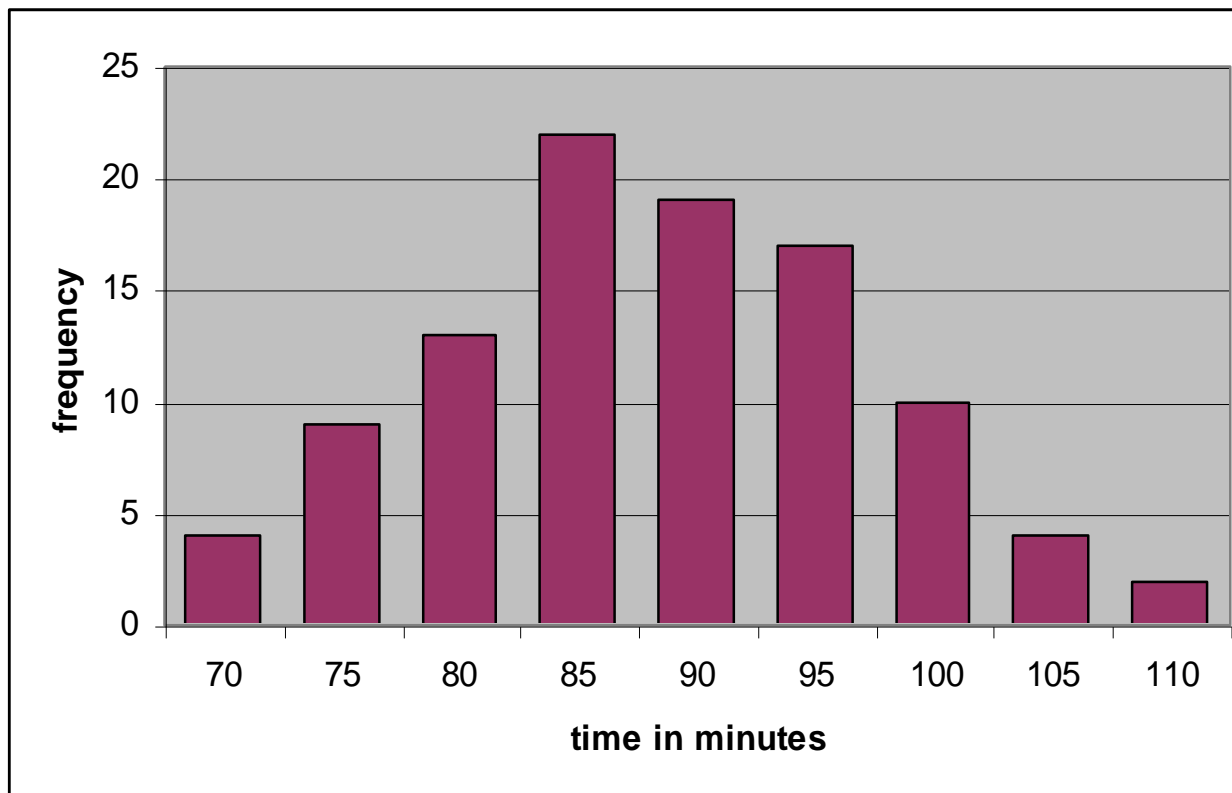
The normal distribution is symmetric about μ and the shape depends on σ^2



Plot of the journey time from Dar es Salaam to Morogoro

X = journey time in minutes

n = the total number of observations = 100



How to check:

- *Plot the histogram and check for symmetry*
- *The mean and median should be similar*
- Check the sample skewness (measure of symmetry)
- Check the sample kurtosis (measure of “peakedness” i.e. shape of distribution)
- Do a statistical test (e.g. Kolmogorov-Smirnov test to compare two samples)

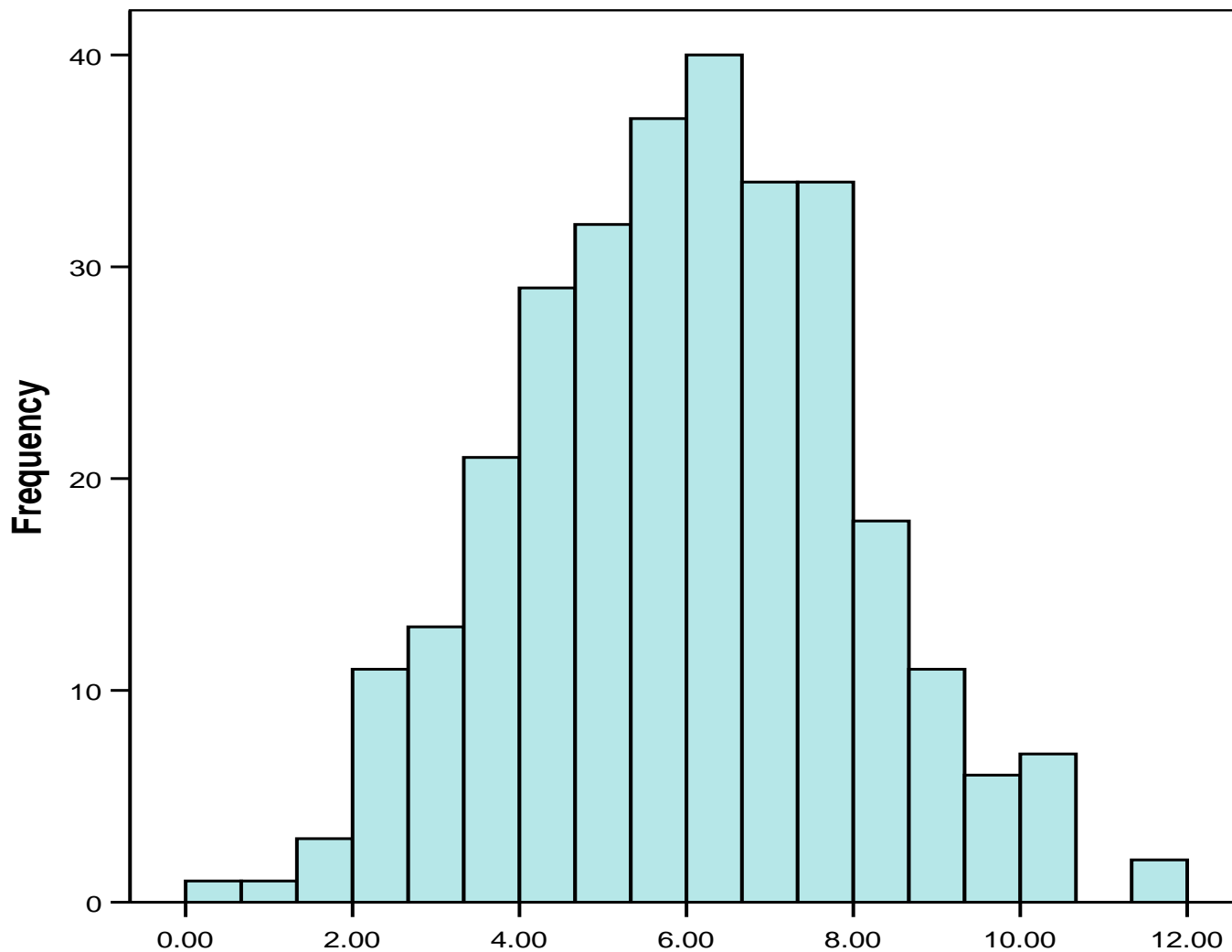
A man driving from Dar es Salaam to Morogoro recorded the following journey times in minutes on 10 occasions

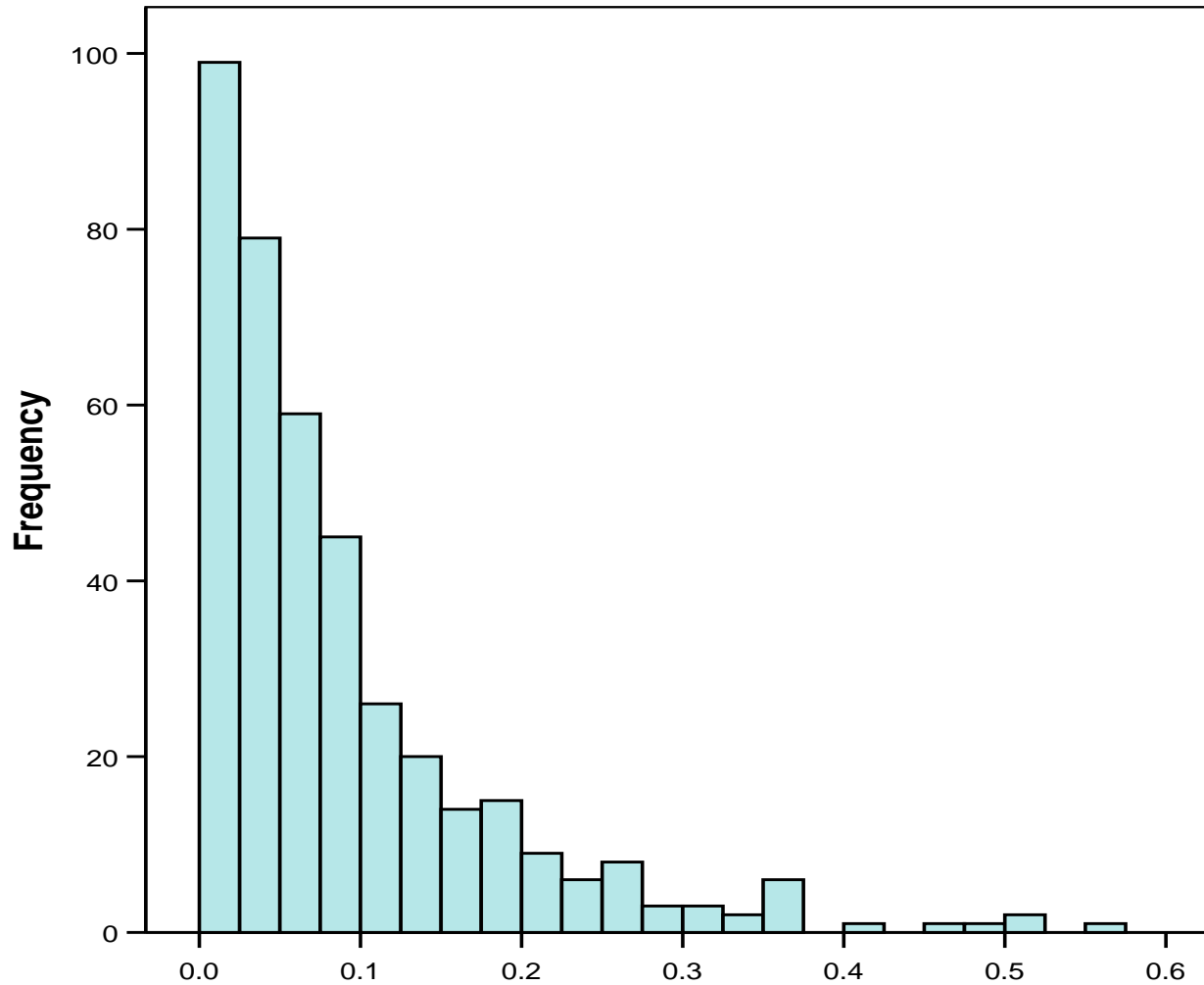
116 121 117 108 122 113 122 114 112 116

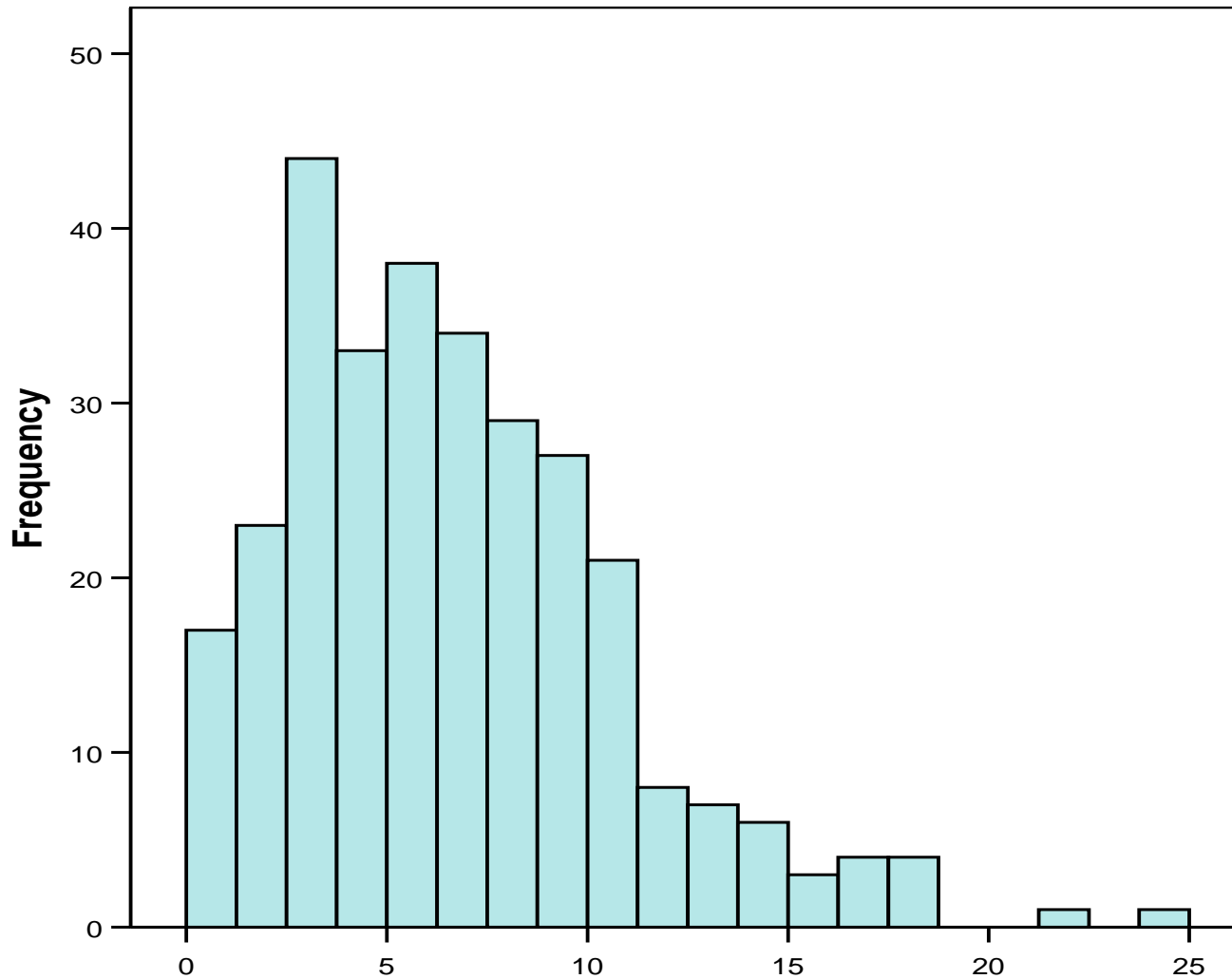
- Does this data follow a normal distribution (only based on 10 observations)?

Use journey time to decide.

- Use **Histogram** (data analysis) – does the distribution look ‘bell-shaped’?
- Use **Descriptive Statistics** – Check the mean, median, skewness and kurtosis







10 Observed journey times from Dar es Salaam to Morogoro

116 121 117 108 122 113 122 114 112 116

Calculate the mean, standard deviation and standard error in Excel:

- **AVERAGE**(data range) =
- **STDEV**(data range) =
- **COUNT**(data range) =
- SE for mean = {Standard deviation / **SQRT**(sample size)}

10 Observed journey times from Dar es Salaam to Morogoro

116 121 117 108 122 113 122 114 112 116

Calculate the mean, standard deviation and standard error in Excel:

- **AVERAGE**(data range) = 116.1mins
 - **STDEV**(data range) = 4.6055 mins
 - **COUNT**(data range) = 10
 - SE for mean = {Standard deviation / **SQRT**(sample size)}
- Try **Descriptive Statistics** (data-data analysis)

Most common statistic

- Use the standard error:
 - $s.e. = \sigma / \sqrt{n}$
- 95% confidence interval of the mean:
 - $\bar{x} \pm 1.96 * \sigma / \sqrt{n}$
- Estimate it using:
 - $\bar{x} \pm t_{n-1} * s / \sqrt{n}$
- t-distribution similar to normal distribution but with fatter tails

Degrees of freedom	t-value
5	2.57
10	2.23
20	2.09
50	2.01
Infinity	1.96

Sampling Issues

A characteristic or measure obtained from a population is known as a **parameter** and it is usually denoted by N , μ , σ

Sample:

It is a portion or a subgroup of an entire group (population) from which an **estimate** can be obtained to make **generalizations/inferences** about the population

A characteristic or measure obtained from a sample is known as a **statistic** and it is usually denoted by n , x , s

Sampling Issues

Sampling Frame:

This is a list of all elements in the population based on different characteristics/peculiarity identifiable with the population — from this the researcher selects units to create the study sample

Sampling:

This is the process of selecting observations (a sample) to provide credible and reliable description and inference about a population

Sampling bias:

Consistent error that arises due to the sample selection.

Sampling Issues

Question:

What about non-sampling error?

Why Sample?:

- Populations are usually large
- It is often impossible to get data for every object that we are studying
- Cost implications
- Time constraints
- Unlimited dynamics within population

Sampling Issues

Expected outcome for engaging with a sample?:

- Mimic the population
- The sample must contain essentially the same variation as the population.

Need to understand degree of homogeneity (heterogeneity) of a population

Key concern - how representative is a sample?

However, engaging with a sample can result in more accurate data.

- Managing a representative sample to obtain a high and accurate response rate may yield more accurate information than surveying everyone and having response bias.

Sampling Issues

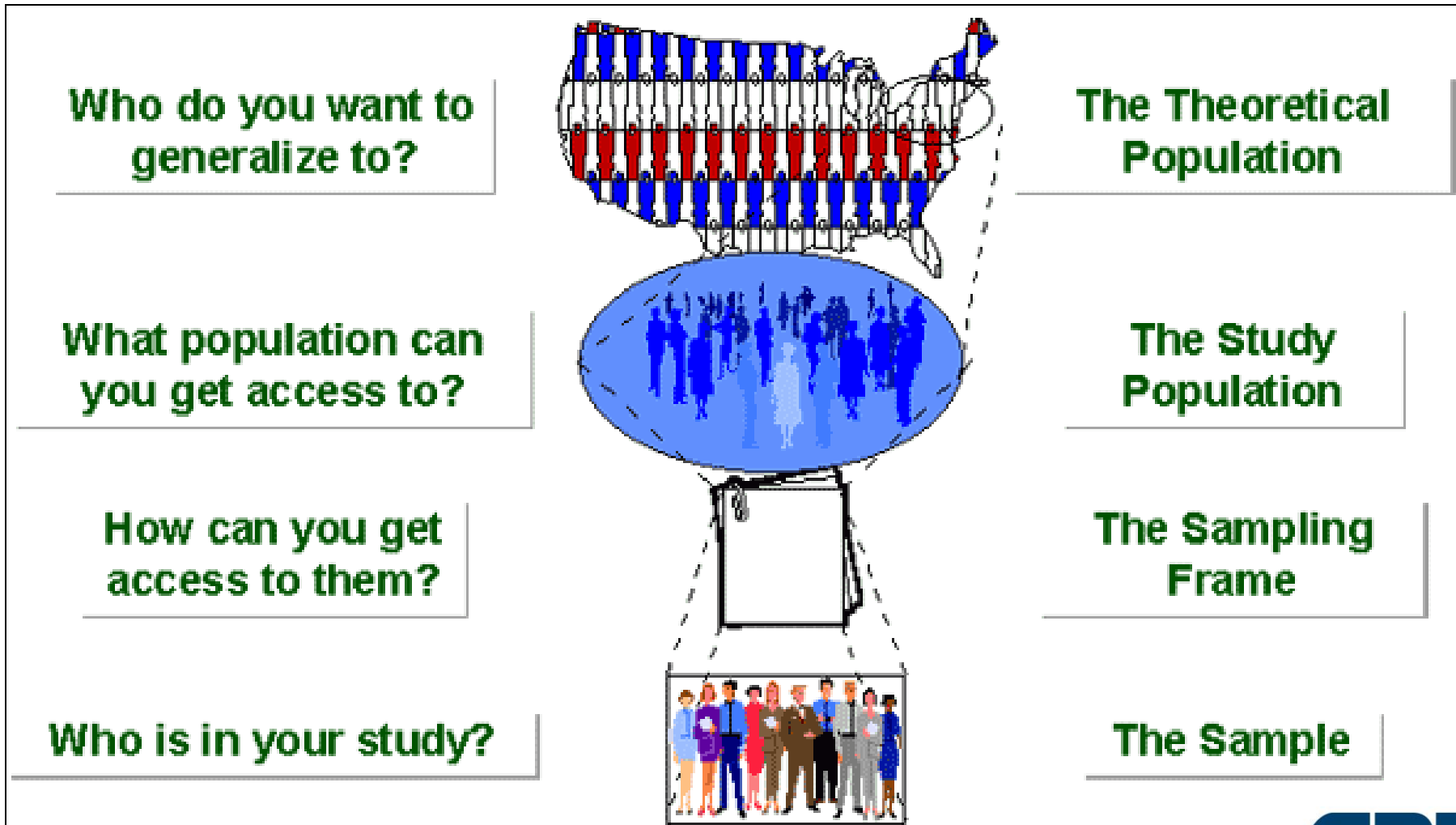
What do we mean by a representative sample?:

- Size - How large should the sample size be?
- Reflection of the variability among elements in the population

Factors to be considered for the computation of sample size?

- Size of population
- Prevalence rate
- Design effect
- Margin of error
- Confidence interval
- Response Rate

AFCAP Summary of Sampling definitions



Methods of Sampling

Probability

- Generalizability to population – element of equal chance and void of bias

Some examples:

- Simple random sample
- Stratified sample
- Cluster sample
- Systematic sample

Non-probability

- Non-generalizable but targeting the ‘right’ units for investigation – has an element of bias

Some examples:

- Quota sample
- “Purposeful” sample
- “Convenience” or “opportunity” sample

Summary

- Each unit has an equal chance of being selected
- Suitable for small sample with complete sampling frame

Advantages and disadvantages

Advantages:

- Simple

Disadvantages:

- Not statistically efficient
- Poor representation of subgroups possible

Procedure

1. Select population and sampling frame
2. Give each unit of population a unique number
3. Calculate/guess your sample size (n)
4. Retrieve n unique numbers from your random number table

Summary

- Randomly order the population
- Select every n th unit

Advantages and disadvantages

Advantages:

- Easy and efficient

Disadvantages:

- Vulnerable to periodicity (i.e. values repeat periodically)

Procedure

1. Select population and sampling frame
2. Give each unit of population a unique random number and sort on this number
3. Write down your sample size (n)
4. Calculate your sampling fraction $f = n/N$
5. Start at a random number in the list
6. Pick every f th value

Summary

- Organise your distinct categories
- Randomly sample within each category
- May want to over sample small categories to ensure good estimates

Advantages and disadvantages

Advantages:

- More economical

Disadvantages:

- May not equally represent all responses
- Lose variability information

Procedure

1. Select population and sampling frame
2. Choose categories of your population to sample
3. Assess whether you need to over sample from small categories
4. Calculate your sampling fraction for each category
5. Sample $n_1 * f_1$
6. Repeat for each category

Summary

- Population dispersed over a large area? Reduce mileage by cluster analysis
- Split geographical region into areas and select a number p of them.
- Each unit of population within these p regions should be measured estimates

Procedure

1. Select population and sampling frame
2. Split your population into small regions
3. Write down your sample size p
4. Sample p regions using simple random sampling
5. Measure every unit within each sampled region

Advantages and disadvantages

Advantages:

- More economical

Disadvantages:

- May not equally represent all responses

Summary

- Generally used for street surveys
- Sample selection is made by an interviewer
- A quota is specified from a subset or set of subsets of the population

Advantages and disadvantages

Advantages:

- Quick and cheap to organise

Disadvantages:

- Not necessarily random
- Possibly biased samples
 - more approachable people

Procedure

1. Select population and sampling frame
2. Choose your subsets of population
3. Specify a quota for each subset
4. Interview or measure units until the quota is filled for a subset and continue until all other subset quotas are complete

Summary

- Combine the sampling strategies above
- Similar to cluster sampling, but sample within your chosen sample
- For example, sample n administration regions, sample m streets in those wards and sample every p th house in each street

Procedure

1. Select population and sampling frame
2. Split your population into small regions
3. Sample regions
4. Sample areas within regions
5. Sample subsets of areas within regions choosing the most appropriate sampling strategy each time

Advantages and disadvantages

Advantages:

- Convenience, economy and efficiency

Disadvantages:

- Lower accuracy due to higher sampling error

Sources of Bias

- A statistic is **biased** if it is calculated in such a way that it is systematically different from the population parameter of interest.
- Types of bias:
 - **Selection bias**, where individuals or groups are more likely to take part in a research project than others, resulting in biased samples.
 - **Spectrum bias** arises from evaluating diagnostic tests on biased samples, leading to an overestimate of the sensitivity and specificity of the test.
 - **Omitted-variable bias** appears in estimates of parameters in a regression analysis when the assumed specification is incorrect, in that it omits an independent variable that should be in the model.
 - **Funding bias** may lead to selection of outcomes, test samples, or test procedures that favour a study's financial sponsor.
 - **Reporting bias** involves a skew in the availability of data, such that observations of a certain kind may be more likely to be reported and consequently used in research.
 - **Data-snooping bias** comes from the misuse of data mining techniques.
 - **Analytical bias** arises due to the way that the results are evaluated.
 - **Exclusion bias** arises due to the systematic exclusion of certain individuals from the study.

Addressing bias

- Carefully design surveys which aim for a random sample of the population
- Recognise that bias is a risk not a certainty
- Sampling is inevitable – there is too much to count properly
- Undertake a pilot and adjust approach if necessary

- 1 Prior to enegaging with Statistics
- 2 General Overview of Statistics
- 3 Descriptive and Inferential Statistics
- 4 Data Distribution, Sampling Issues and Sources of Bias
- 5 Hypothesis Testing

The Concept Hypothesis Testing

- Statistical test
 - Aim is to test a hypothesis concerning the values of one or more population parameters
 - The null hypothesis is specified as H_0
 - The null hypothesis is never accepted
 - The null hypothesis is counter to what the researcher hopes for
 - The available data is the basis for proving that the null hypothesis is false

The Concept Hypothesis Testing

- Statistical test
 - The test requires alternative option hence the counterpart of the null hypothesis is termed as the alternative hypothesis
 - This is typically denoted by H_a or H_1
 - The alternative to the null hypothesis is that there is any difference (two-sided) or whether the difference has to be in a particular direction (one-sided)
 - Another name for the alternative hypothesis is research hypothesis
 - Typically, we start research with research/alternative hypothesis as that is what we hope for

Plausible Errors

- Type I Error
 - Failing to accept (Rejecting) a true null hypothesis
 - This type of error should really be marginal
 - As in the case of the court of law the jury committing this error implies conviction of an innocent person
 - The probability of a type I error is denoted by α and is often referred to as the significance “level of the test”.

- Type II Error
 - Failing to reject (Accepting) a false null hypothesis
 - The probability of a type II error is denoted by β
 - The likelihood of committing either a Type I or Type II error measures the goodness of a test.
 - Note that $\alpha \neq 1 - \beta$
 - Both α and β can only be reduced simultaneously as the sample size increases

Table: Four Possible Results of a Hypothesis Test

State of the World	Decision	
	H_0 Not Rejected	H_0 Rejected
If H_0 is true	Correct Decision Probability = $1 - \alpha$ = Confidence level	Type I error Probability = α = Level of test
If H_0 is false	Type II error Probability = β	Correct decision Probability = $1 - \beta$ = Power of test

AFCAP Steps in Hypothesis Testing



1. State research/alternative hypothesis
2. Decide on the level of significance for the test
3. Set up the null hypothesis
4. Identify/collect sample data
5. Calculate a test statistic from the sample data
6. Compare the test statistic to its sampling distribution under the null hypothesis and calculate the p-value (calculate the critical region)
7. Reject the null hypothesis if the p-value is less than the level of significance or the test statistic lies in the critical region



Critical values for one and two tailed tests at traditional levels of significance

Level of Significance (α)	Type of Test	
	One-Tailed	Two-Tailed
0.05	+1.645 or -1.645	± 1.96
0.01	+2.33 or -2.33	± 2.58
0.001	+3.09 or -3.09	± 3.30

- GETstats:

<http://www.getstats.org.uk/category/goodstats/>

- TDM Encyclopedia: Data Collection and Surveys:

<http://www.vtpi.org/tdm/tdm40.htm>

- Wikipedia statistics:

<http://en.wikipedia.org/wiki/Statistics>

- Statsoft: Statistica:

<http://www.statsoft.com>

- Royal Statistical Society

<http://www.rss.org.uk>

- BBC

<http://www.bbc.co.uk/bitesize/>



**Now read
Session 6.5
Notes!**