CHAPTER 2

ORGANIZING DATA

ORGANIZING DATA

- Sampling techniques
- Sampling and non-sampling errors
- Organizing and Graphing
  - Qualitative data
  - Quantitative data

Understanding Terms

- **Census**
  a study of every unit, everyone or everything, in a population. It is known as a complete enumeration, which means a complete count.

- **Sample**
  a subset of units in a population, selected to represent all units in a population of interest. It is a partial enumeration because it is a count from part of the population.

Concept of sampling

- A process of selecting units from a population
- A process of selecting a sample to determine certain characteristics of a population

“Why sample ?”

- Economy
- Timeliness
- The large size of many populations
- Inaccessibility of some of the population
- Destructiveness of the observation – accuracy

In most cases, **census** is unnecessary!

Sampling…

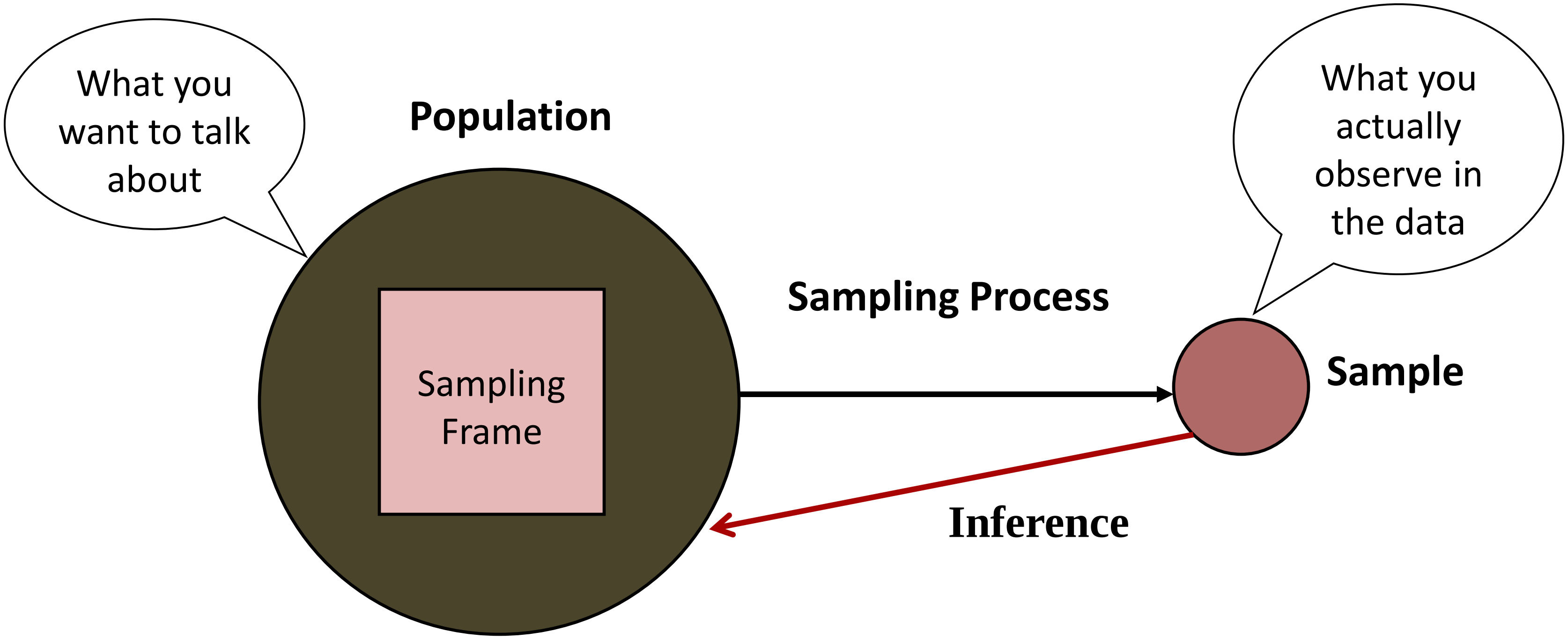
3 factors that influence sample representative-ness

- Sampling procedure
- Sample size
- Participation (response)

When might you sample the entire population?

- When your population is very small
- When you have extensive resources
- When you don’t expect a very high response

What is Sampling?


What you want to talk about
Population
Sampling Frame
Sampling Process
What you actually observe in the data
Sample
Inference

A **sampling frame** is the source material or device from which a sample is drawn. It is a list of all those within a population who can be sampled, and may include individuals, households or institutions.

Sampling Techniques

Probability sampling

Each member of the population has a certain probability to be selected into the sample. Types of probability sampling;

- Simple random sampling
- Systematic sampling
- Stratified sampling

Non-probability sampling

Members selected not according to logic of probability (or mathematical rules), but by other means;

- Convenience sampling
- Purposive sampling

Probability sampling…

Simple Random Sampling

Every possible sample of a given size have same chance of selection;

- ✓ Establish a sampling frame (a list, e.g. of all the company’s customers, or all UCT students)
- ✓ Assign a single number to each element in the list
- ✓ Use random numbers to select the elements

Probability sampling…

Systematic sampling

- This is random sampling with system. From the sampling frame, a starting point is chosen at random, and thereafter at regular interval
- Usually more efficient than Simple Random Sampling (SRS);
  - ➢ Establish a sampling frame
  - ➢ Select the first element at random
  - ➢ Then select every nth element in the list, until you have the required number of respondents
    - e.g. with a population of 300, if we want a sample of 10, choose every 30th element
    - Keep an eye out for peculiar arrangements in the sampling frame

Probability sampling…

Stratified sampling

- Modifies random sampling and systematic sampling, to obtain a greater degree of representativeness
- Organize the population into homogeneous subsets, then sample randomly within each one
  - e.g. for university students, stratify according to seniority and gender
- Stratification ensures equal proportions of people having the relevant characteristics are selected into your sample
- Depends on what variables are available to stratify on

Non - Probability sampling…

i. **Convenience sampling**

- Rely on available respondents
- Most convenient method
- Risky; exercise caution

ii. **Purposive sampling**

- *Select the sample on the basis of knowledge of the population: your own knowledge, or use expert judges to identify candidates to select*
- *Typically used for very rare populations, such as deviant cases*

Errors…

Two major types of error can arise when a sample of observations is taken from a population:

i. sampling error

ii. non - sampling error.

Sampling Error…

Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample.

- Sampling error occurs because researchers draw different subjects from the same population but still, the subjects have individual differences.

Increasing the sample size **will** reduce this type of error.

Non - sampling Error…

Nonsampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly. Three types of nonsampling errors:

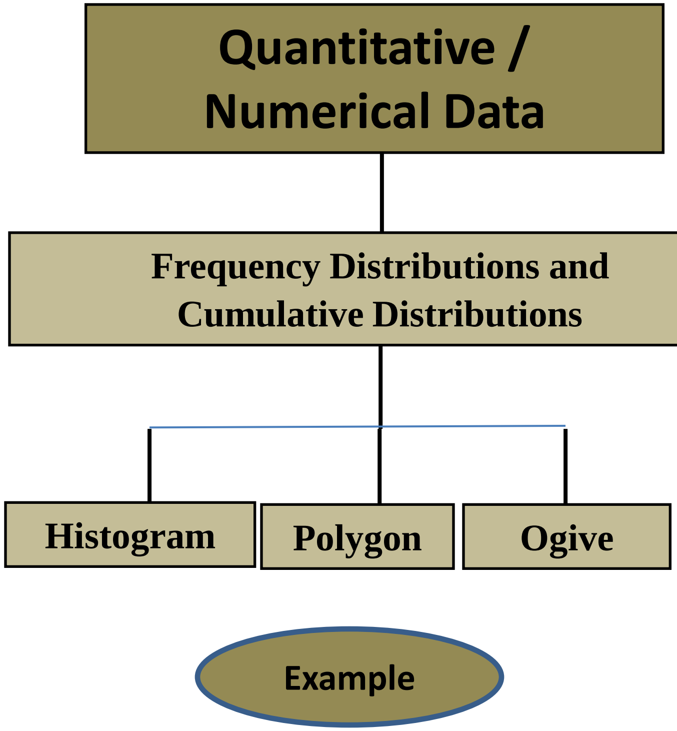
i. Errors in data acquisition,

ii. Nonresponse errors, and

iii. Selection bias.

Increasing the sample size **will** reduce this type of error.

Tables and Charts



Quantitative / Numerical Data

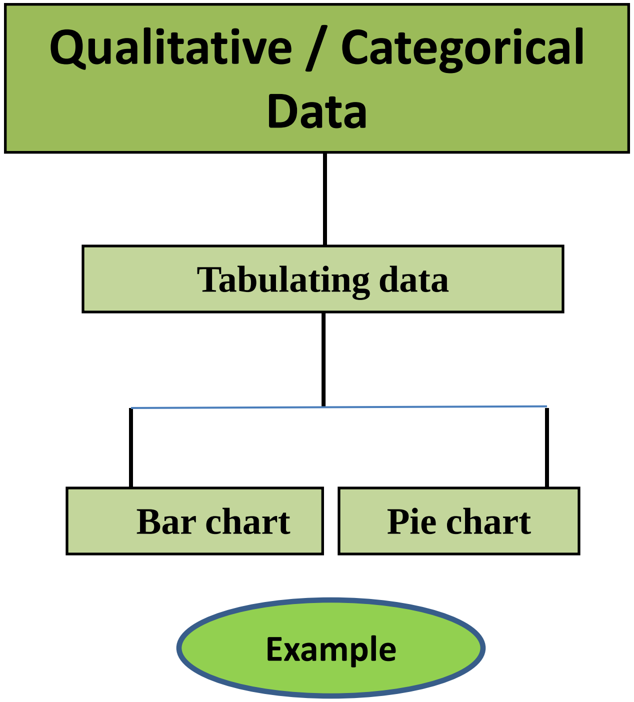
Frequency Distributions and Cumulative Distributions

Histogram

Polygon

Ogive

Example



Qualitative / Categorical Data

Tabulating data

Bar chart

Pie chart

Example

Organizing Data: Frequency Distribution

- A *frequency distribution* for quantitative data lists all the classes and the number of values that belong to each class.
- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered classes.
- In general, a frequency distribution should have at least 5 but no more than 15 classes.
- To determine the **width of a class interval,** divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

Frequency Distributions

***Class Boundary*:**

The class boundary is given by the midpoint of the upper limit of one class and the lower limit of the next class.

Finding Class Width:

$$\boxed{\text{Class size} = \text{Upper boundary} - \text{Lower boundary}}$$

Calculating Class Midpoint:

$$\boxed{\text{Class midpont} = \left(\text{Upper} \lim it + \text{Lower} \lim it\right)/2}$$

Relative Frequency and Percentage Distribution

The *relative frequency* shows what fractional part or proportion of the total frequency belongs to the corresponding category.

$$\boxed{\text{Relative frequency of a class} = \frac{\text{Frequency of a class}}{\text{Sum of all frequencies}}}$$

A *percentage distribution* lists the percentages for all categories.

$$\boxed{\text{Percentage} = (\text{Relative frequency}) \times 100}$$

Cumulative Frequency Distributions

- A *cumulative frequency distribution* gives the total number of values that fall below the upper boundary of each class
- A cumulative frequency distribution is constructed for *quantitative data* only.
- In cumulative frequency distribution table, each class has the *same lower limit* but a *different upper limit*.

Cumulative Relative Frequency and Cumulative Percentage

- Cumulative relative frequency

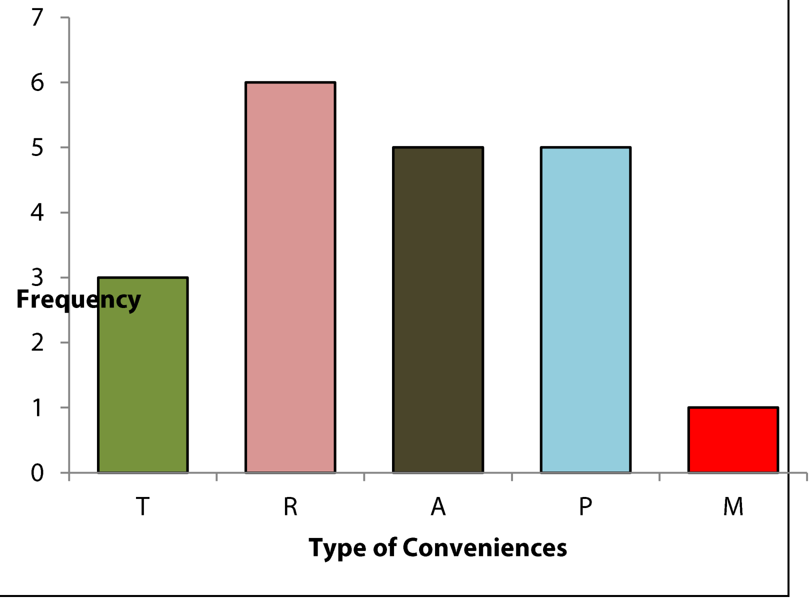
$$\boxed{= \frac{\text{Cumulative frequency}}{\text{Total observations in the data set}}}$$

- Cumulative percentage

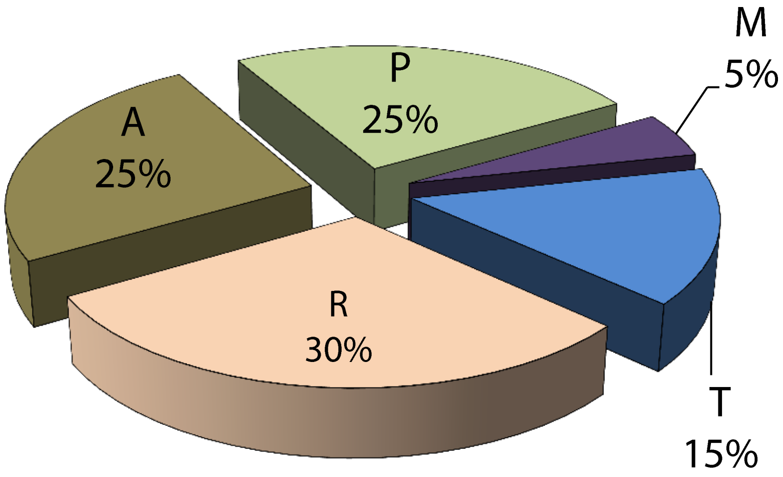
$$\boxed{= \left(\text{Cumulative relative frequency}\right) \times 100}$$

Bar Graphs and Pie Chart

A graph made of bars whose heights represent the frequencies of respective categories is called a *bar graph*.



A circle divided into portions that represent the relative frequencies or percentages of a population or a sample belonging to different categories.

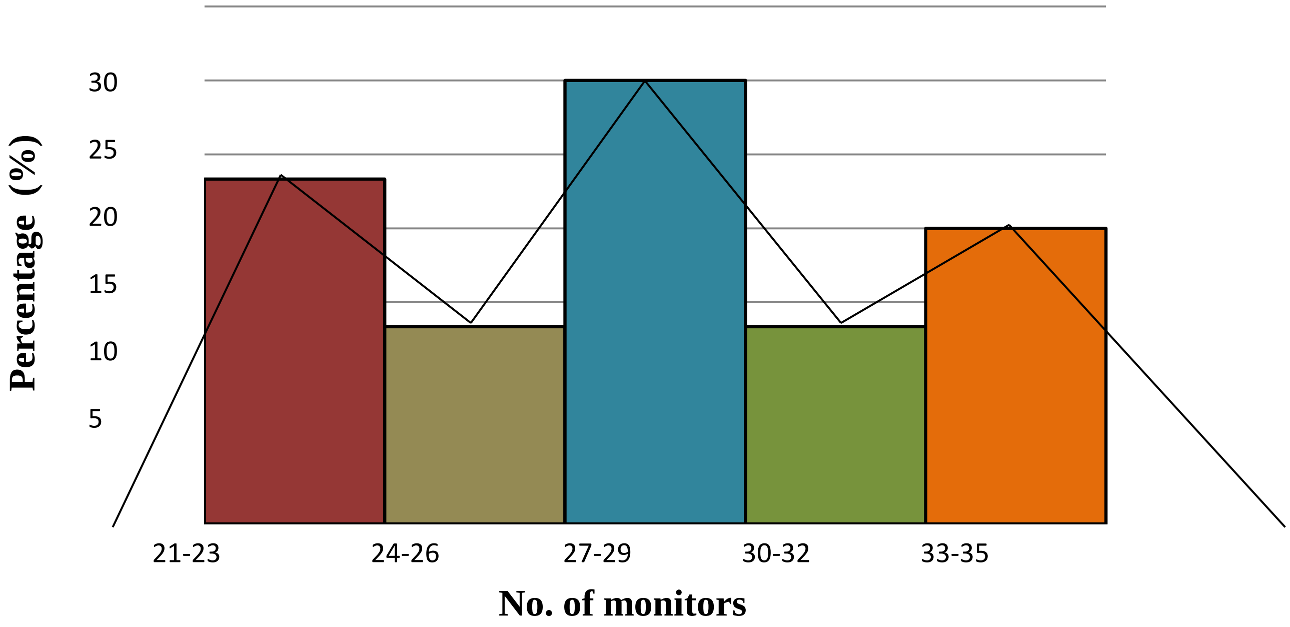


Histogram and Poligons

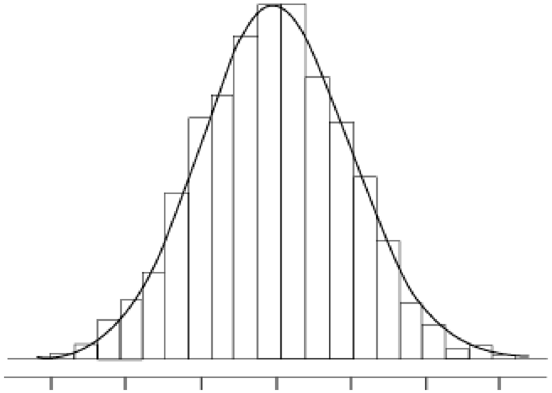
A vertical bar chart of the data in a frequency distribution is called a **histogram.**

- In a histogram there are no gaps between adjacent bars.
- In a percentage histogram the vertical axis would be defined to show the percentage of observations per class

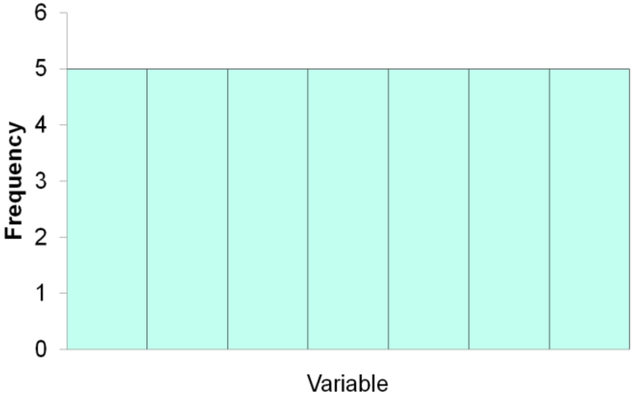
A graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines is called a polygon.



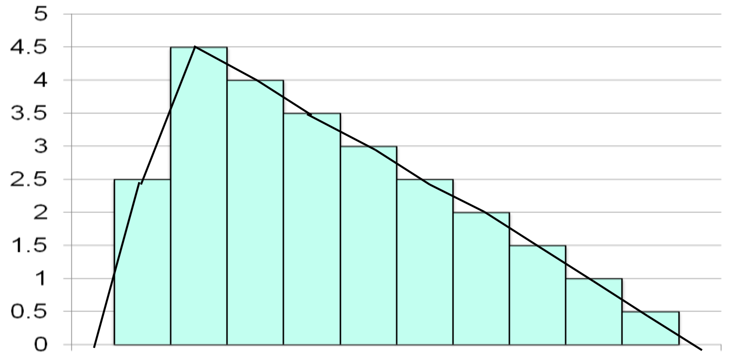
Shapes Of Histograms



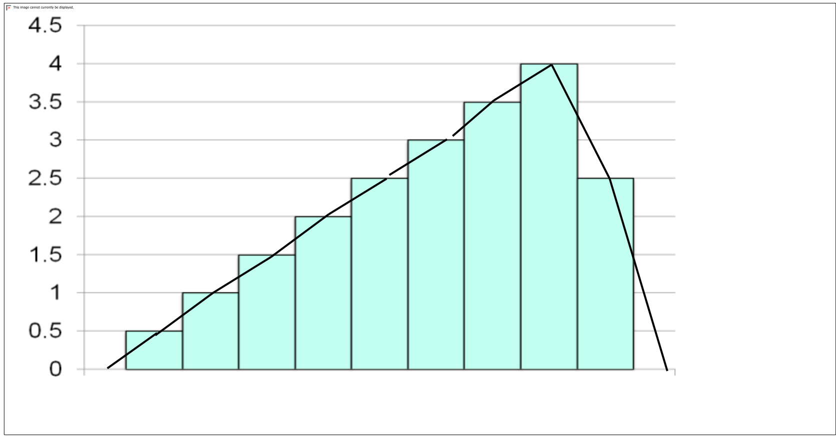
Symmetric Histogram



Uniform / Rectangular Histogram



Skewed to the right

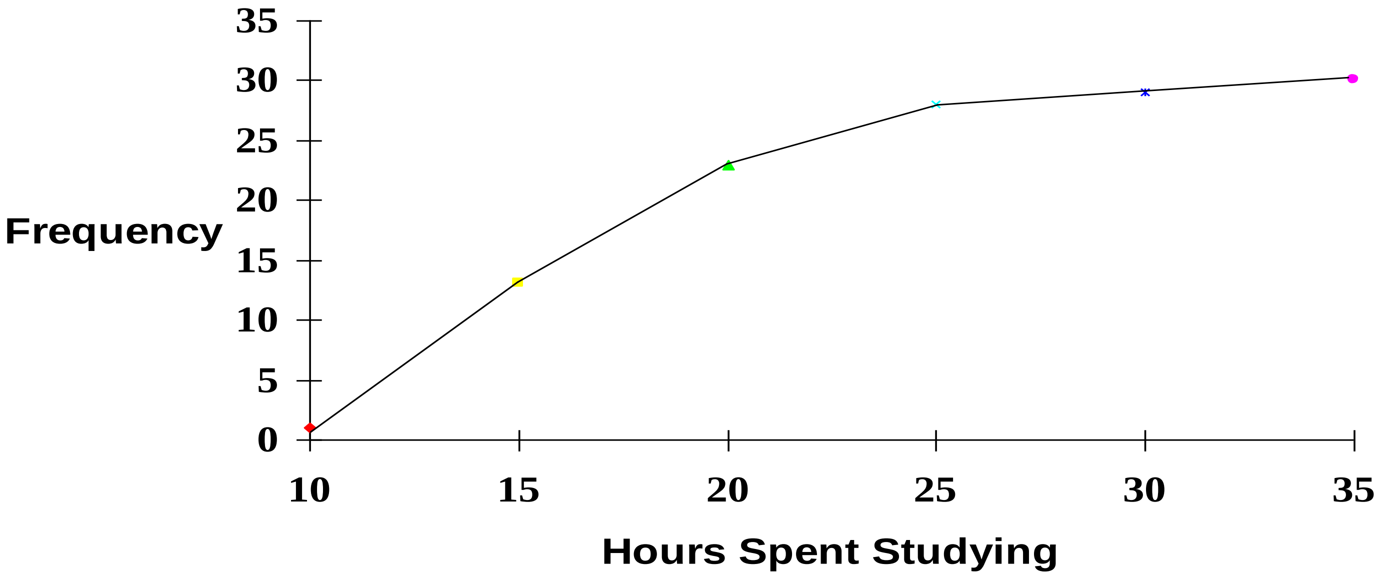


Skewed to the left

Ogives

An **Ogive** (a cumulative line graph) is best used when you want to display the total at any given time.

Meaning Ogive: An cumulative frequency distribution by joining with straight lines the dots marked above the upper boundaries of classes at heights equal to the cumulative frequencies of respective classes.



Example:

A distribution of the number of hours that boat batteries lasted is:

Number of Hours	Frequency
24-30	3
31-37	1
38-44	5
45-51	9
52-58	6
59-65	1

a) Find the class boundaries and the class mid-point

b) Do all classes have the same width? If so, what is this width?

c) Prepare the relative frequency and the percentage distribution columns.

d) What the percentage of these boat posses 38 or more hours batteries?