Introductory Statistics

Tatiana Tyukina
tt51@leicester.ac.uk

Topic 1 - Descriptive Statistics

- **Topic 0: Introduction**
- Topic 1: Descriptive Statistics
- Topic 2: Estimators: Point estimation, Interval estimation
- Topic 3: Hypothesis Testing: Test for single population mean, test for two population mean
- Topic 4: Inference based on Normal Distribution
- Topic 5: Goodness of Fit: The χ^2 test.

DEFINITION

Population *is the collection of the individuals or items under consideration in a statistical study.*

DEFINITION

Population *is the collection of the individuals or items under consideration in a statistical study.*

DEFINITION

Sample *is that part of the population from which information is obtained.*

DEFINITION

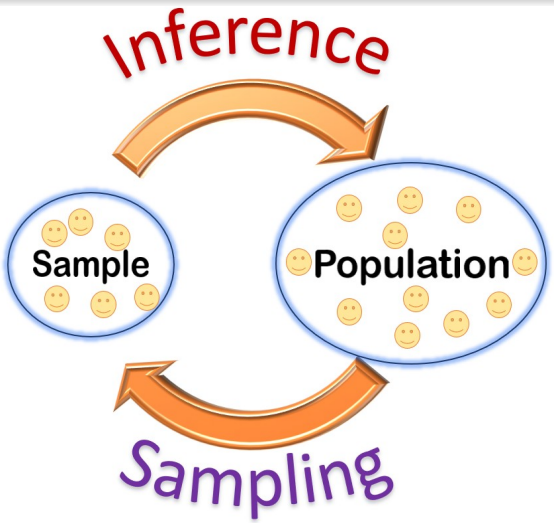
Population *is the collection of the individuals or items under consideration in a statistical study.*

DEFINITION

Sample *is that part of the population from which information is obtained.*

DEFINITION

Sample frame *is the list of individuals from which sample is actually selected*



Simple Random Sampling - A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained:

Simple Random Sampling - A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained:

- picking slips of paper out of the box;

Simple Random Sampling - A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained:

- picking slips of paper out of the box;
- using random-number tables;

73735	45963	78134	63873
02965	58303	90708	20025
98859	23851	27965	62394
33666	62570	64775	78428
81666	26440	20422	05720
15838	47174	76866	14330
89793	34378	08730	56522
78155	22466	81978	57323
16381	66207	11698	99314
75002	80827	53867	37797
99982	27601	62686	44711
84543	87442	50033	14021
77757	54043	46176	42391
80871	32792	87989	72248
30500	28220	12444	71840

another example `https://teorica.fis.ucm.es/ft8/tablern2.pdf`

Simple Random Sampling - A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained:

- picking slips of paper out of the box;
- using random-number tables;
- using random-number generators.

Simple Random Sampling - A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained:

- picking slips of paper out of the box;
- using random-number tables;
- using random-number generators.

Systematics Random Sampling - every Kth element in the sampling frame is selected after a suitable random start for the first element.

Simple Random Sampling - A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained:

- picking slips of paper out of the box;
- using random-number tables;
- using random-number generators.

Systematics Random Sampling - every Kth element in the sampling frame is selected after a suitable random start for the first element.

Stratified Sampling - a population with N elements is divided into s sub-populations (non-overlapping). A sample is drawn from each sub-population independently.

Simple Random Sampling - A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained:

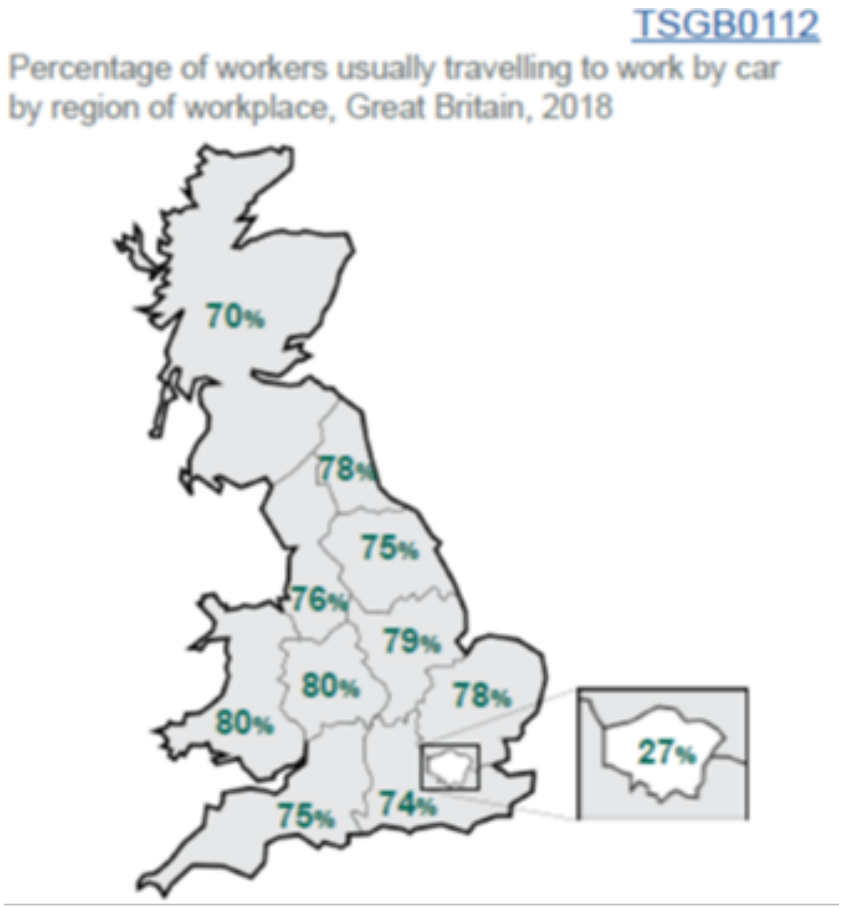
- picking slips of paper out of the box;
- using random-number tables;
- using random-number generators.

Systematics Random Sampling - every Kth element in the sampling frame is selected after a suitable random start for the first element.

Stratified Sampling - a population with N elements is divided into s sub-populations (non-overlapping). A sample is drawn from each sub-population independently.

Cluster Sampling - the sampling unit contains groups of elements called clusters instead of individual elements of the population. A cluster is an intact group naturally available in the field.

Example of Cluster Sampling



Source:
Transport Statistics Great Britain 2019 [Link 1]

A sample should reflect all the characteristics (of importance) of the population - be *representative* - reflect as closely as possible the relevant characteristics of the population under consideration.
A sample that is not representative of the population characteristics is called a *biased* sample.

DEFINITION

Sampling error *occur when the sample is not representative, its characteristics differ from the population.*

DEFINITION

Sampling error *occur when the sample is not representative, its characteristics differ from the population.*

DEFINITION

The ***sample size*** *is an important feature of any empirical study*

Sample size depends:

on the population size;

on the required reliability;

available resources (money, time, etc.)

DEFINITION

Sampling error *occur when the sample is not representative, its characteristics differ from the population.*

DEFINITION

The ***sample size*** *is an important feature of any empirical study*

DEFINITION

Nonsampling errors *occur in the collection, recording, and processing of sample data.*

Sources:

poorly designed survey;

missing values;

incorrect measurements or responses.

DEFINITION

The values of a random variable for one or more people or things is called ***data****.*

DEFINITION

The values of a random variable for one or more people or things is called ***data****.*

DEFINITION

Each individual piece of data is called an ***observation****.*

DEFINITION

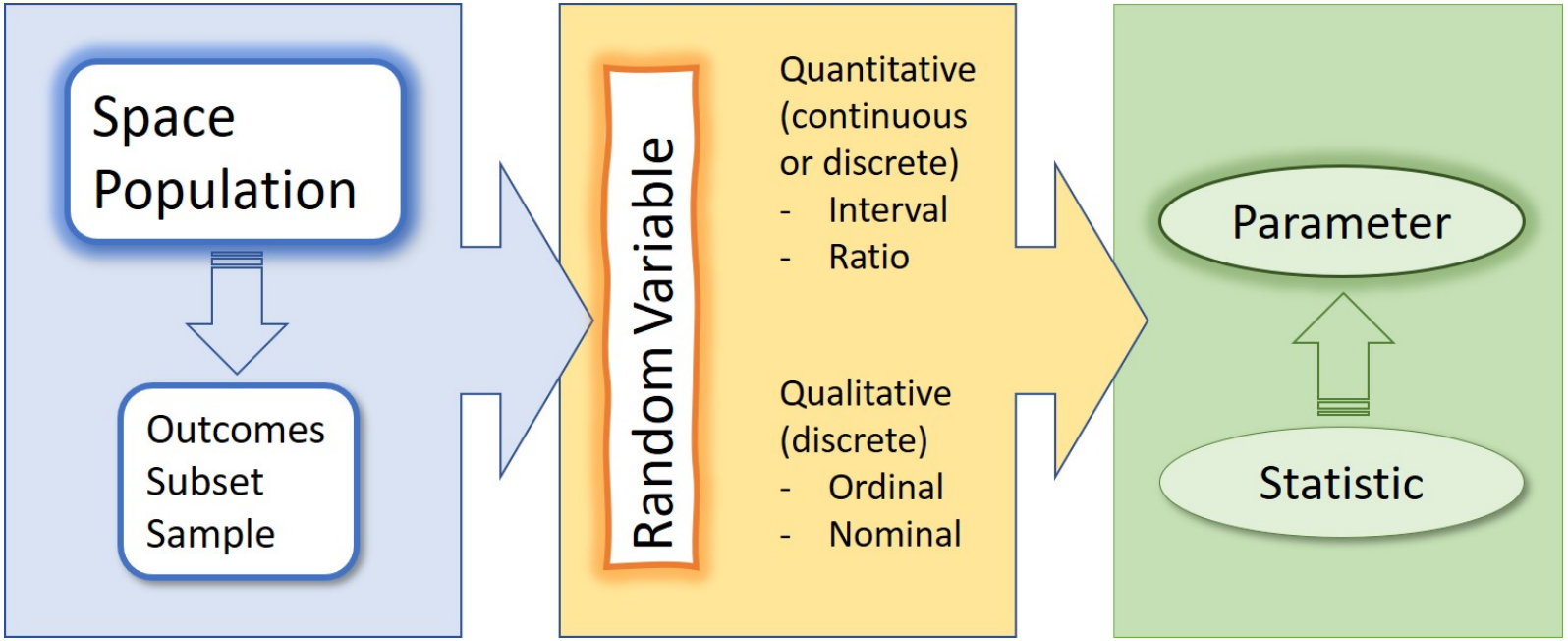
The values of a random variable for one or more people or things is called ***data****.*

DEFINITION

Each individual piece of data is called an ***observation****.*

DEFINITION

The collection of all observations for a particular variable is called a ***data set****.*


Space
Population
Outcomes
Subset
Sample
Random Variable
Quantitative
(continuous
or discrete)
- Interval
- Ratio
Qualitative
(discrete)
- Ordinal
- Nominal
Parameter
Statistic

The most common tabular data structures:

The most common tabular data structures:

Cross-Sectional Data

HoursW	Edu	Income	Sex	Age	State
40	13	44000	M	45	Illinois
35	12	12000	F	35	Wisconsin
40	16	48200	M	38	Ohio
35	8	25000	F	42	Michigan
.	.	.	.	.	.

The most common tabular data structures:

Cross-Sectional Data

Time Series Data

Year	AvgMinWage	AvgCovRate	UnempRate	GNP,mln$
1950	0.20	20.1	15.4	878.7
1951	0.21	20.7	16.0	925.0
1952	0.23	22.06	14.8	1015.9
.	.	.	.	.
1987	3.35	58.2	16.8	4496.7

The most common tabular data structures:

Cross-Sectional Data

Time Series Data

Pooled Cross Sections

Year	Hprice,$	PropTax	sqrft	Bedrms
1993	85500	42	1600	3
1993	67300	36	1440	3
.	.	.	.	.
1993	243600	41	2600	4
1995	65000	16	1250	2
1995	182400	20	2200	4
.	.	.	.	.
1995	97500	15	1540	3

The most common tabular data structures:

- Cross-Sectional Data
- Time Series Data
- Pooled Cross Sections
- Panel or Longitudinal Data

City	Year	Murders	Population	Unem	Police
1	1986	5	350000	8.7	440
1	1990	8	359200	7.2	471
2	1986	2	64300	5.4	75
2	1990	1	65100	5.5	75
.	.	.	.	.	.
150	1986	25	543000	4.3	520
150	1990	32	546200	5.2	493

Graphical representation

- frequency distribution (Qualitative Data) or histogram (Quantitative Data);

Graphical representation

- frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- relative frequency distribution (Qualitative Data) or histogram (Quantitative Data);

Graphical representation

- frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- relative frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- Pareto Chart;

Graphical representation

- frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- relative frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- Pareto Chart;
- Pie chart (Qualitative Data or Quantitative Data - for Range values);

Graphical representation

- frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- relative frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- Pareto Chart;
- Pie chart (Qualitative Data or Quantitative Data - for Range values);
- Horizontal Bar Chart (Nominal Qualitative Data);

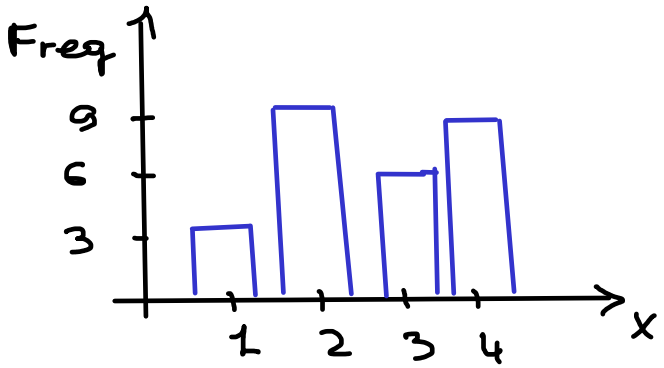
Graphical representation

- frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- relative frequency distribution (Qualitative Data) or histogram (Quantitative Data);
- Pareto Chart;
- Pie chart (Qualitative Data or Quantitative Data - for Range values);
- Horizontal Bar Chart (Nominal Qualitative Data);
- Stem-Leaves Diagram (Quantitative Data);

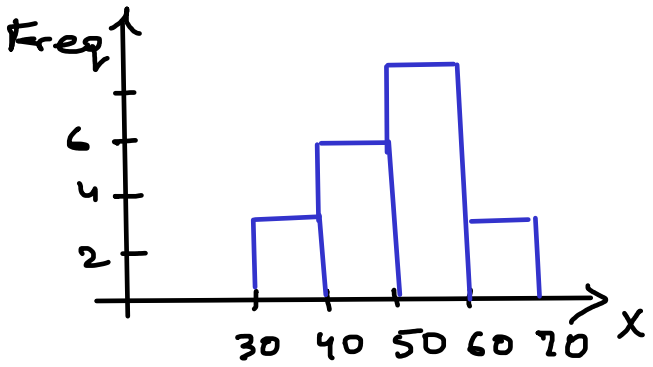
- **frequency distribution** (Qualitative Data) or **histogram** (Quantitative Data);
- **relative frequency distribution** (Qualitative Data) or **histogram** (Quantitative Data);
- Pareto Chart;
- **Pie chart** (Qualitative Data or Quantitative Data - for Range values);
- Horizontal Bar Chart (Nominal Qualitative Data);
- Stem-Leaves Diagram (Quantitative Data);

Frequency distribution

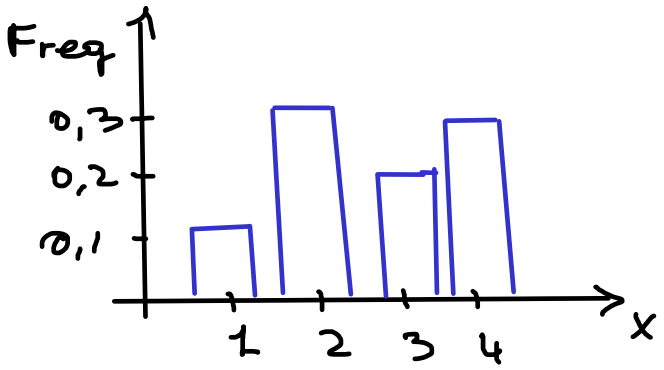
Bar chart

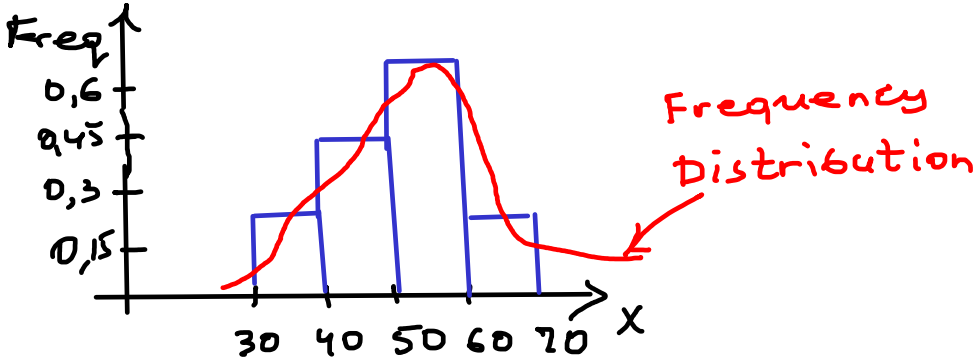


Histogram

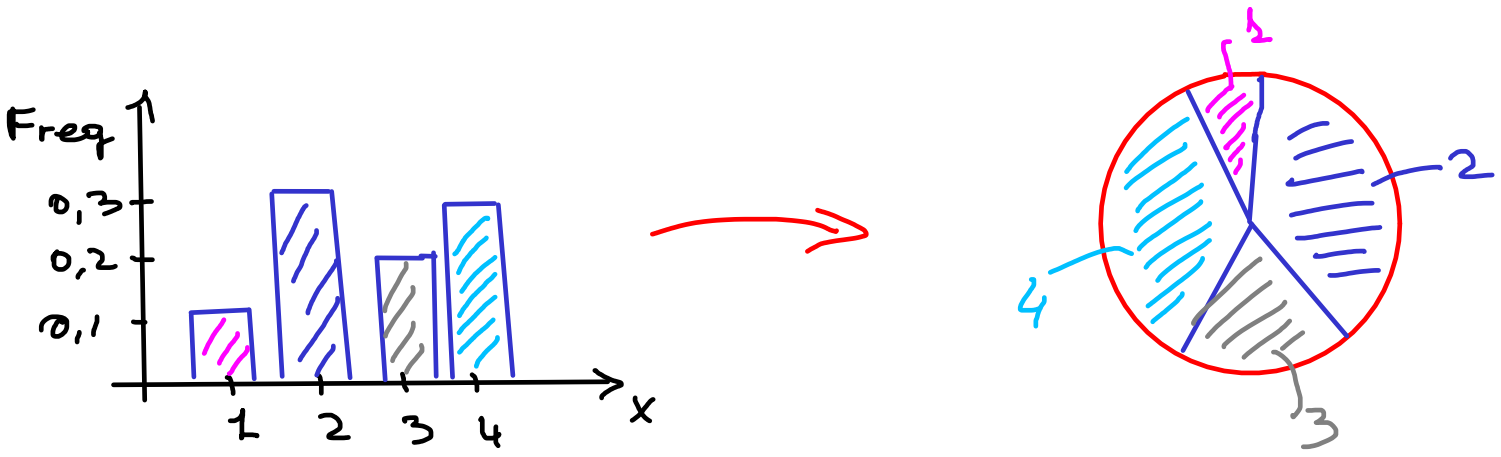


Relative frequency distribution

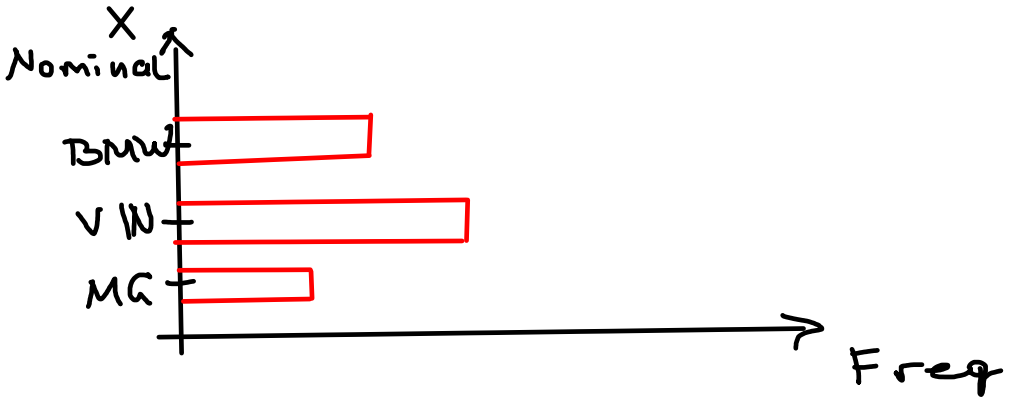




Pie chart

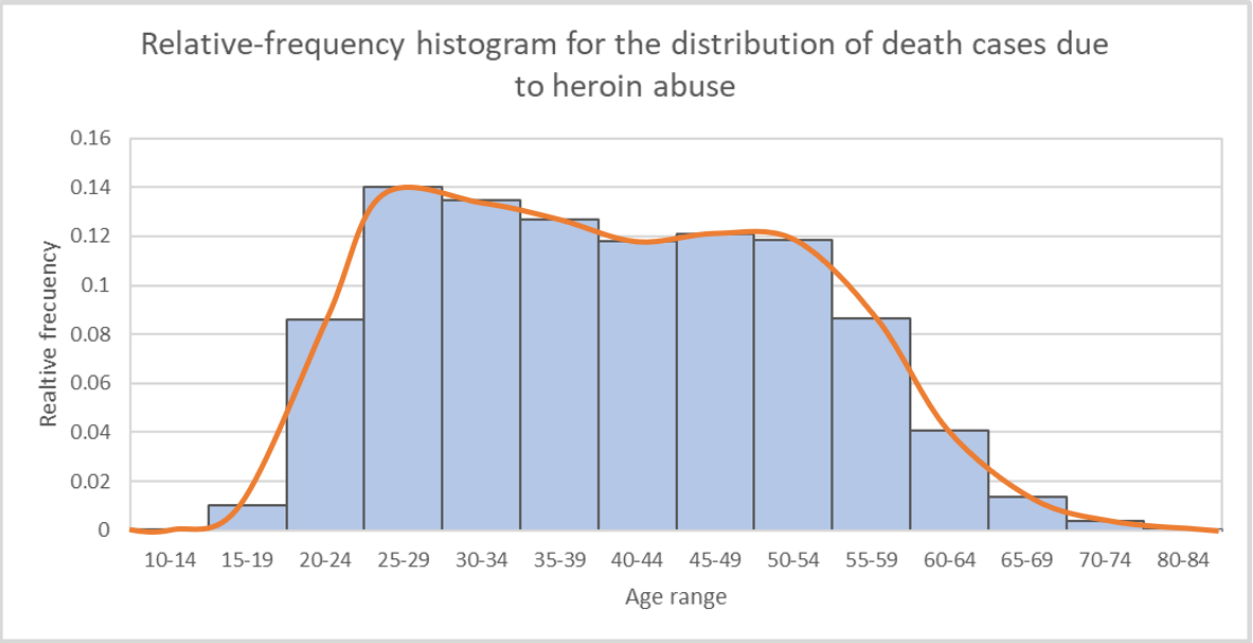


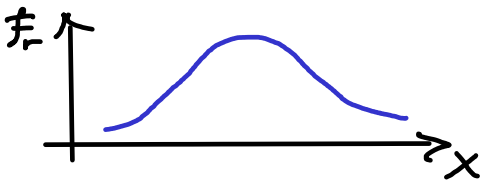
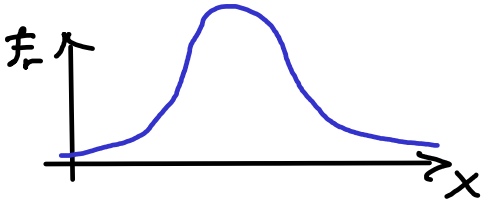
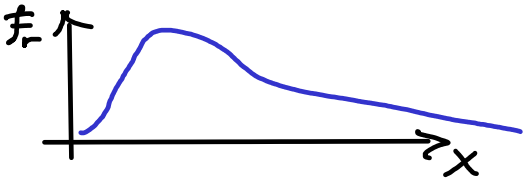
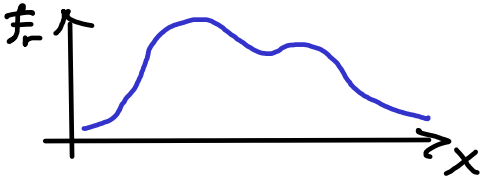
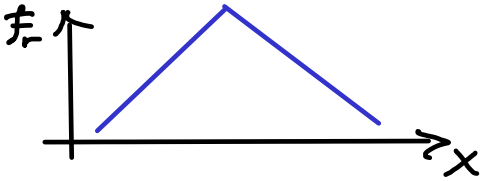
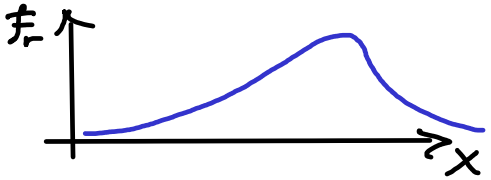
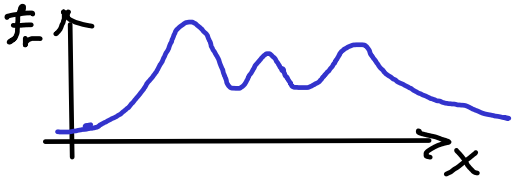
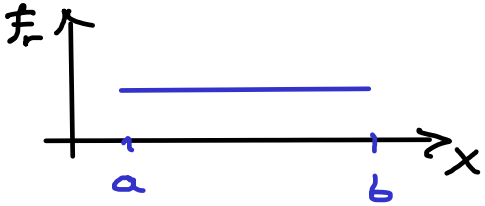
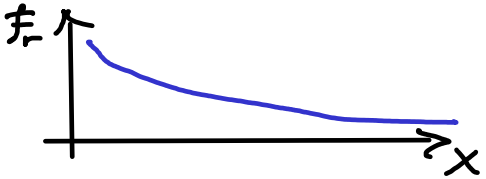
Horizontal Bar Chart



DEFINITION

*The **distribution of a data set** is a table, graph, or formula that provides the values of the observations and how often they occur.*



Modality	Symmetry	Skewness
 unimodal	 bell-shaped	 right-skewed
 bimodal	 triangular	 left-skewed
 multimodal	 uniform	 inverse J-shaped

Drug Classification This database contains information about certain drug types.

EXAMPLE

This data has been collected from a drug company which has labelled data set of the drugs and the parameters that effect it. Therefore, on the basis of the disease and the type of patient the drug is recommended. See drug200.xlsx file on Blackboard

`https://www.kaggle.com/prathamtripathi/drug-classification/metadata`

DEFINITION

Numbers that are used to describe data sets are called ***descriptive measures****.*

DEFINITION

Numbers that are used to describe data sets are called ***descriptive measures****.*

DEFINITION

Descriptive measures that indicate where the center or most typical value of a data set lies are called ***measures of central tendency*** *or, more simply,* ***measures of center****. (Measures of center are often called averages.)*

DEFINITION

Numbers that are used to describe data sets are called ***descriptive measures****.*

DEFINITION

Descriptive measures that indicate where the center or most typical value of a data set lies are called ***measures of central tendency*** *or, more simply,* ***measures of center****. (Measures of center are often called averages.)*

DEFINITION

Descriptive measures that indicates the amount of variation, or spread, in a data set are referred to as ***measures of variation*** *or* ***measures of spread****.*

- The **Mean (Sample Mean)** - the sum of observations divided by the number of observations.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where $x_1, x_2, \ldots, x_n$ are observations, n is the sample size;

- The **Mean (Sample Mean)** - the sum of observations divided by the number of observations.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i,$$

  where $x_1, x_2, \ldots, x_n$ are observations, n is the sample size;
- The **Median** is the point that divides the sample in half.

- The **Mean (Sample Mean)** - the sum of observations divided by the number of observations.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i,$$

  where $x_1, x_2, \ldots, x_n$ are observations, n is the sample size;
- The **Median** is the point that divides the sample in half.

DEFINITION

*The **order statistics** of $X_1, \ldots, X_n$ are their values put in increasing order ($x_1 \leq x_2 \leq \ldots \leq x_n$), which we denote $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.*
$x_{(i)}$ is called i-th order statistic.

- The **Mean (Sample Mean)** - the sum of observations divided by the number of observations.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where $x_1, x_2, \ldots, x_n$ are observations, n is the sample size;

- The **Median** is the point that divides the sample in half.

$$\tilde{x} := \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & \text{if } n \text{ is even.} \end{cases}$$

DEFINITION

*The **order statistics** of $X_1, \ldots, X_n$ are their values put in increasing order ($x_1 \leq x_2 \leq \ldots \leq x_n$), which we denote $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.*
$x_{(i)}$ is called i-th order statistic.

A sample of $n = 6$:

$x_1 = 25$, $x_2 = 15$, $x_3 = 13$, $x_4 = 21$,
$x_5 = 18$, $x_6 = 23$

The order statistics

$x_{(1)} = 13$, $x_{(2)} = 15$, $x_{(3)} = 18$, $x_{(4)} = 21$,
$x_{(5)} = 23$, $x_{(6)} = 25$

$$\text{Median} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

$$\text{median} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{18 + 21}{2} = 19.5$$

- A α**-trimmed Mean** - the mean of the remaining middle $100(1-2\alpha)\%$ of the observations.

$$\bar{x}_\alpha := \frac{x_{([n\alpha]+1)} + x_{([n\alpha]+2)} + \ldots x_{(n-[n\alpha])}}{n-2[n\alpha]} = \frac{\sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}}{n-2[n\alpha]},$$

where $[s]$ is the greatest integer $\leq s$.

$n = 20 \qquad \alpha = 0{,}05 \qquad 2\alpha = 0{,}1 \qquad n\alpha = 1$

$\underbrace{x_{(1)}} \ldots x_{(19)}, \underbrace{x_{(20)}}$

$x_{(2)}, x_{(3)} \ldots, x_{(19)}$

- A α**-trimmed Mean** - the mean of the remaining middle $100(1-2\alpha)\%$ of the observations.

$$\bar{x}_\alpha := \frac{x_{([n\alpha]+1)} + x_{([n\alpha]+2)} + \ldots x_{(n-[n\alpha])}}{n-2[n\alpha]} = \frac{\sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}}{n-2[n\alpha]},$$

  where $[s]$ is the greatest integer $\leq s$.
- The **Mode** is most frequently occurring value of the data.

- A α**-trimmed Mean** - the mean of the remaining middle $100(1-2\alpha)\%$ of the observations.

$$\bar{x}_\alpha := \frac{x_{([n\alpha]+1)} + x_{([n\alpha]+2)} + \ldots x_{(n-[n\alpha])}}{n-2[n\alpha]} = \frac{\sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}}{n-2[n\alpha]},$$

  where $[s]$ is the greatest integer $\leq s$.
- The **Mode** is most frequently occurring value of the data.
  1) Find the frequency of each value in the data;

- A α**-trimmed Mean** - the mean of the remaining middle $100(1-2\alpha)\%$ of the observations.

$$\bar{x}_\alpha := \frac{x_{([n\alpha]+1)} + x_{([n\alpha]+2)} + \ldots x_{(n-[n\alpha])}}{n-2[n\alpha]} = \frac{\sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}}{n-2[n\alpha]},$$

  where $[s]$ is the greatest integer $\leq s$.

- The **Mode** is most frequently occurring value of the data.
  1) Find the frequency of each value in the data;
  2) If no value occurs more than once, then the data set has *no mode*;

- A α**-trimmed Mean** - the mean of the remaining middle $100(1-2\alpha)\%$ of the observations.

$$\bar{x}_\alpha := \frac{x_{([n\alpha]+1)} + x_{([n\alpha]+2)} + \ldots x_{(n-[n\alpha])}}{n-2[n\alpha]} = \frac{\sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}}{n-2[n\alpha]},$$

  where $[s]$ is the greatest integer $\leq s$.

- The **Mode** is most frequently occurring value of the data.
  1) Find the frequency of each value in the data;
  2) If no value occurs more than once, then the data set has *no mode*;
  3) Otherwise, any value that occurs with the greatest frequency is a mode of the data.

- A α**-trimmed Mean** - the mean of the remaining middle $100(1-2\alpha)\%$ of the observations.

$$\bar{x}_\alpha := \frac{x_{([n\alpha]+1)} + x_{([n\alpha]+2)} + \ldots x_{(n-[n\alpha])}}{n-2[n\alpha]} = \frac{\sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}}{n-2[n\alpha]},$$

  where $[s]$ is the greatest integer $\leq s$.
- The **Mode** is most frequently occurring value of the data.
  1) Find the frequency of each value in the data;
  2) If no value occurs more than once, then the data set has *no mode*;
  3) Otherwise, any value that occurs with the greatest frequency is a mode of the data.

The mode is the only measure of center that can be used for qualitative data!

- A **Grouped-Data Mean**

Grouped data are data formed by aggregating individual observations of a variable into groups, so that a frequency distribution of these groups serves as a convenient means of summarizing or analyzing the data.

$$\bar{x}_g := \frac{\sum_{j=1}^{m} x_j f_j}{\sum_{j=1}^{m} f_j} = \frac{\sum_{j=1}^{m} x_j f_j}{n},$$

where x_j denotes either a class mark or a midpoint, f_j - a class frequency, m is the number of classes or intervals, and $n = \sum_{j=1}^{m} f_j$) is the sample size.

- A **Grouped-Data Mean**

$$\bar{x}_g := \frac{\sum_{j=1}^{m} x_j f_j}{\sum_{j=1}^{m} f_j} = \frac{\sum_{j=1}^{m} x_j f_j}{n},$$

  where x_j denotes either a class mark or a midpoint, f_j - a class frequency, m is the number of classes or intervals, and $n = \sum_{j=1}^{m} f_j$) is the sample size.

- The **Weighted Mean** is used when some data points contribute more "weight" than others.

- A **Grouped-Data Mean**

$$\bar{x}_g := \frac{\sum_{j=1}^{m} x_j f_j}{\sum_{j=1}^{m} f_j} = \frac{\sum_{j=1}^{m} x_j f_j}{n},$$

where x_j denotes either a class mark or a midpoint, f_j - a class frequency, m is the number of classes or intervals, and $n = \sum_{j=1}^{m} f_j$) is the sample size.

- The **Weighted Mean** is used when some data points contribute more "weight" than others.

$$\bar{x}_w := \frac{\sum_{j=1}^{n} x_j w_j}{\sum_{j=1}^{n} w_j},$$

where x_j denotes some data, w_j - corresponding non-negative weights, n is the sample size (or a number of data points).

- The sample mean
- The median
- The α-trimmed mean
- The mode
- The grouped-data mean
- The weighted mean

EXAMPLE

Some quantitative data were collected:

$$79,\ 61,\ 77,\ 74,\ 54,\ 70,\ 62,\ 83,\ 93,\ 66,\ 70,\ 89$$

$$80,\ 98,\ 54,\ 90,\ 83,\ 50,\ 72,\ 82,\ 60,\ 83,\ 51,\ 86,\ 61$$

Question:
Using the limit grouping with a first class of $50 - 59$ and a class width of 10. Find the sample mean, the median, the α-trimmed mean ($\alpha = 5\%$), the mode and the grouped mean.

EXAMPLE

Some quantitative data were collected:

$$79,\ 61,\ 77,\ 74,\ 54,\ 70,\ 62,\ 83,\ 93,\ 66,\ 70,\ 89$$

$$80,\ 98,\ 54,\ 90,\ 83,\ 50,\ 72,\ 82,\ 60,\ 83,\ 51,\ 86,\ 61$$

Question:
Using the limit grouping with a first class of $50 - 59$ and a class width of 10. Find the sample mean, the median, the α-trimmed mean ($\alpha = 5\%$), the mode and the grouped mean.

$$50, 51, 54, 54, 60, 61, 61, 62, 66, 70, 70, 72, 74, 77, 79,$$

$$80, 82, 83, 83, 83, 86, 89, 90, 93, 98$$

Frequency distribution table

interval	Frequency
50-59	4
60-69	5
70-79	6
80-89	7
90-99	3

Grouped mean

$$\text{mean}_g = \frac{\left(\frac{50+59}{2}\right)\cdot 4 + \left(\frac{60+69}{2}\right)\cdot 5 + \left(\frac{70+79}{2}\right)\cdot 6 + \left(\frac{80+89}{2}\right)\cdot 7 + \left(\frac{90+99}{2}\right)\cdot 3}{25} = \underline{74.5}$$

$$\text{Mean} = \frac{50+51+54+54+\ldots+93+98}{25} = \underline{73.12}$$

Median:

order all values – see previous slide

$n = 25$ – sample size $\Rightarrow$ median $= x_{\left(\left[\frac{25+1}{2}\right]\right)} = x_{(13)} = \underline{74}$

5%-trimmed mean

$n\alpha = 25 \cdot 0{,}05 = 1{,}25 \Rightarrow$ we should exclude $[1{,}25] = 2$ observations from lower and upper side

$$\text{mean}_{5\%} = \frac{54+54+60+\ldots+89+90}{21} = \underline{73.14}$$

The **Sample Variance** is defined as the expectation of the squared difference of data points from the mean of the data set:

The **Sample Variance** is defined as the expectation of the squared difference of data points from the mean of the data set:

$$\tilde{s}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n},$$

The **Sample Variance** is defined as the expectation of the squared difference of data points from the mean of the data set:

$$\tilde{s}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n},$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

The **Sample Standard Deviation** measures variation by indicating how far, on average, the observations are from the mean:

The **Sample Standard Deviation** measures variation by indicating how far, on average, the observations are from the mean:

$$\tilde{s}_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}},$$

The **Sample Standard Deviation** measures variation by indicating how far, on average, the observations are from the mean:

$$\tilde{s}_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}},$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

The **Sample Variance for grouped data**:

$$s^2_{x,grouped} = \frac{\sum_{j=1}^{m}(x_j - \bar{x})^2 f_j}{n-1},$$

where x_j denotes either a class mark or a midpoint, f_j - a class frequency, m is the number of classes or intervals, and $n = \sum_{j=1}^{m} f_j$ is the sample size.

The **Sample Variance for grouped data**:

$$s^2_{x,grouped} = \frac{\sum_{j=1}^{m}(x_j - \bar{x})^2 f_j}{n-1},$$

where x_j denotes either a class mark or a midpoint, f_j - a class frequency, m is the number of classes or intervals, and $n = \sum_{j=1}^{m} f_j$ is the sample size.

The **Sample Standard Deviation for grouped data**:

The **Sample Variance for grouped data**:

$$s^2_{x,grouped} = \frac{\sum_{j=1}^{m}(x_j - \bar{x})^2 f_j}{n-1},$$

where x_j denotes either a class mark or a midpoint, f_j - a class frequency, m is the number of classes or intervals, and $n = \sum_{j=1}^{m} f_j$ is the sample size.

The **Sample Standard Deviation for grouped data**:

$$s_{x,grouped} = \sqrt{\frac{\sum_{j=1}^{m}(x_j - \bar{x})^2 f_j}{n-1}},$$

The **Range** is the difference between the largest and smallest values of the sample

$$range = x_{(max)} - x_{(min)} = x_{(n)} - x_{(1)};$$

The **Range** is the difference between the largest and smallest values of the sample

$$range = x_{(max)} - x_{(min)} = x_{(n)} - x_{(1)};$$

The **Midrange** is average of the largest and the smallest values of the sample

$$midrange = \frac{x_{(max)} + x_{(min)}}{2}.$$

The **Interquartile Range (IQR)** is the difference between the upper and lower quartiles of the sample.

The **Interquartile Range (IQR)** is the difference between the upper and lower quartiles of the sample.

Quantile is a nonparametric estimator of the population distribution based on the set of independent observations:

$$Q(p) = F^{-1}(p) = \inf\{x : \ F(x) \geq p\}, \quad 0 < p < 1,$$

where $F(x)$ is the distribution function. Percentile is the value below which a percentage of data falls. Percentiles vary between 0 and 100. Quartiles are the values that split the data into quarters, denoted as $Q_0, \ Q_1, \ Q_2, \ Q_3, \ Q_4$.
The range is $r = Q_4 - Q_0$.

https://www.researchgate.net/publication/222105754_Sample_Quantiles_in_Statistical_Packages

The **Interquartile Range (IQR)** is the difference between the upper and lower quartiles of the sample.

$$Q_1 = \frac{x_{(\lfloor 1+(n-1)/4 \rfloor)} + x_{(\lceil 1+(n-1)/4 \rceil)}}{2},$$

$$Q_3 = \frac{x_{(\lfloor 1+3(n-1)/4 \rfloor)} + x_{(\lceil 1+3(n-1)/4 \rceil)}}{2},$$

$$\text{IQR} = Q_3 - Q_1.$$

A **five-number summary** consists of

A **five-number summary** consists of

the *smallest data value*,

A **five-number summary** consists of

the *smallest data value*,

the *first quartile*,

A **five-number summary** consists of

the *smallest data value*,
the *first quartile*,
the *median*,

A **five-number summary** consists of

the *smallest data value*,
the *first quartile*,
the *median*,
the *third quartile*,

A **five-number summary** consists of

the *smallest data value*,
the *first quartile*,
the *median*,
the *third quartile*,
and the *largest data value*.

A **five-number summary** consists of

the *smallest data value*,
the *first quartile*,
the *median*,
the *third quartile*,
and the *largest data value*.

The five-number summary usually presented as a box-plot.

Consider the previous example

$$\text{Range} = 98 - 50 = 48$$

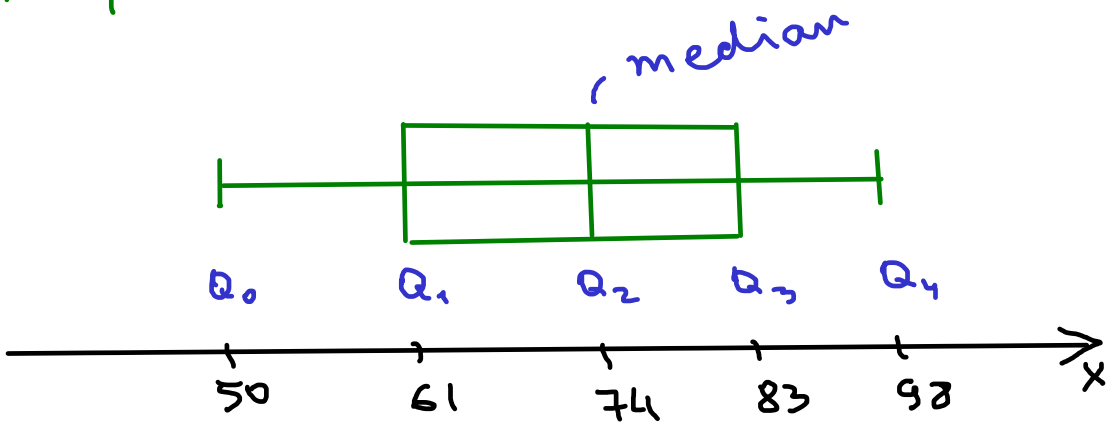
$$\text{Midrange} = \frac{98+50}{2} = 74$$

$$Q_1 = \frac{x_{(\lfloor 1 + \frac{25-1}{4} \rfloor)} + x_{(\lceil 1 + \frac{25-1}{4} \rceil)}}{2} = \frac{x_{(7)} + x_{(7)}}{2} = 61$$

$$Q_3 = \frac{x_{(\lfloor 1 + 3 \cdot \frac{25-1}{4} \rfloor)} + x_{(\lceil 1 + 3 \cdot \frac{25-1}{4} \rceil)}}{2} = \frac{x_{(19)} + x_{(19)}}{2} = 83$$

$$IQR = 83 - 61 = 22$$

Box-plot



- the sample variance or sample standard deviation;
- the sample range;
- the sample midrange
- the interquartile range (IQR);
- the five-point summary.

Questions to take home

1 which of the following (if any) can be represented by ordinal data:
   - Height
   - Weight
   - Age
   - Number of siblings
   - Religion
   - Place of birth
   - High school class rank

Questions to take home

1. which of the following (if any) can be represented by ordinal data:
   - Height
   - Weight
   - Age
   - Number of siblings
   - Religion
   - Place of birth
   - High school class rank
2. Answer true or false to each of the statements:

Questions to take home

1 which of the following (if any) can be represented by ordinal data:
   - Height
   - Weight
   - Age
   - Number of siblings
   - Religion
   - Place of birth
   - High school class rank

2 Answer true or false to each of the statements:
   a. Two data sets that have identical frequency distributions have identical relative-frequency distributions.

Questions to take home

1 which of the following (if any) can be represented by ordinal data:
   - Height
   - Weight
   - Age
   - Number of siblings
   - Religion
   - Place of birth
   - High school class rank

2 Answer true or false to each of the statements:
   a. Two data sets that have identical frequency distributions have identical relative-frequency distributions.
   b. Two data sets that have identical relative-frequency distributions have identical frequency distributions.

Questions to take home

1 which of the following (if any) can be represented by ordinal data:
   - Height
   - Weight
   - Age
   - Number of siblings
   - Religion
   - Place of birth
   - High school class rank

2 Answer true or false to each of the statements:
   a. Two data sets that have identical frequency distributions have identical relative-frequency distributions.
   b. Two data sets that have identical relative-frequency distributions have identical frequency distributions.

   Use your answers to parts (a) and (b) to explain why relative frequency distributions are better than frequency distributions for comparing two data sets.

Questions to take home

1 which of the following (if any) can be represented by ordinal data:
   - Height
   - Weight
   - Age
   - Number of siblings
   - Religion
   - Place of birth
   - High school class rank

2 Answer true or false to each of the statements:
   a. Two data sets that have identical frequency distributions have identical relative-frequency distributions.
   b. Two data sets that have identical relative-frequency distributions have identical frequency distributions.

   Use your answers to parts (a) and (b) to explain why relative frequency distributions are better than frequency distributions for comparing two data sets.

3 define main descriptive measures of central tendency;

Questions to take home

1. which of the following (if any) can be represented by ordinal data:
   - Height
   - Weight
   - Age
   - Number of siblings
   - Religion
   - Place of birth
   - High school class rank
2. Answer true or false to each of the statements:
   a. Two data sets that have identical frequency distributions have identical relative-frequency distributions.
   b. Two data sets that have identical relative-frequency distributions have identical frequency distributions.

   Use your answers to parts (a) and (b) to explain why relative frequency distributions are better than frequency distributions for comparing two data sets.
3. define main descriptive measures of central tendency;
4. define main descriptive measures of variation.

Link 1

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/870647/tsgb-2019.pdf