Medical Databases

- Data: is the formalized representation of concepts or facts that are appropriate for processing (automated or by humans).

Medical Databases

- Data: is the formalized representation of concepts or facts that are appropriate for processing (automated or by humans).
- File: represents an organized set of data.

Medical Databases

- Data: is the formalized representation of concepts or facts that are appropriate for processing (automated or by humans).
- File: represents an organized set of data.
  Data File Structure:
  - Records
  - Fields: Name, Type (numeric, char, bool etc.), Size

Medical Databases

Data Base: structured set of data, comprising both data and relations between data.

Medical Databases

Data Base: structured set of data, comprising both data and relations between data.

Minimalistic *structure* of a database:

- Files, restrictions: at least one (1) common field for all the files. Usually the common field is ID

Medical Databases

Data Base: structured set of data, comprising both data and relations between data.

Minimalistic *structure* of a database:

- Files, restrictions: at least one (1) common field for all the files. Usually the common field is ID
- Relations: represents how records and/or data is connected.

Medical Databases

Data Base: structured set of data, comprising both data and relations between data.

Creating a database, steps:

1. Collecting Data: record structure of DB; code data; fill data;

Medical Databases

Data Base: structured set of data, comprising both data and relations between data.

Creating a database, steps:

1. Collecting Data: record structure of DB; code data; fill data;
2. Validating Data: validate by field type; all possible relations.

Database Management Systems (DBMS)

A set of software tools that are used in order to:

- Design and build a DB

Database Management Systems (DBMS)

A set of software tools that are used in order to:

- Design and build a DB
- Control access to data

Database Management Systems (DBMS)

A set of software tools that are used in order to:

- Design and build a DB
- Control access to data
- Assure data security and integrity

Database Management Systems (DBMS)

- Functions

Database Management Systems (DBMS)

- Functions
- Description
  - data structure
  - relations
  - access conditions

Database Management Systems (DBMS)

- Functions
- Description
  - data structure
  - relations
  - access conditions
- Data manipulation
  - Create, delete, update a record
  - Search, sort, edit virtual records

Database Management Systems (DBMS)

- Functions
- Description
  - data structure
  - relations
  - access conditions
- Data manipulation
  - Create, delete, update a record
  - Search, sort, edit virtual records
- User functions (allows facile User-DB interaction)

Database Management Systems (DBMS)

- Functions
- Description
  - data structure
  - relations
  - access conditions
- Data manipulation
  - Create, delete, update a record
  - Search, sort, edit virtual records
- User functions (allows facile User-DB interaction)
- DBMD languages: MySQL, MS Access, Oracle, FoxPro etc.

Biostatistics

February 12, 2020

Public Health

Steps in the Paradigm of Public Health:

- Define the problem
- Measure its magnitude
- Understand the key determinants
- Develop intervention/prevention strategies
- Set the policy and priorities
- Implement and evaluate the results

Public Health

Biostatistics and epidemiology in public health:

- Use of quantitative methods which combine biostatistics and epidemiology
- Information is collected in order to investigate a question
- Methods and tools from biostatistics are used in order to analyse data $\rightarrow$ aid decision making.

Biostatistics

Definition **Biostatistics** is the science that studies characteristics of populations. Or in other words is the application of statistics to a wide range of topics in biology.

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. In other words is a method to *learn* from data.

Biostatisticians

A **biostatistician** is a person who's work is driven by questions relating to the health of people, as individuals or members of population.

In a nutshell they can be classified as:

- "Data detectives", they are in charge of uncovering patterns and clues through data description and exploration

Biostatisticians

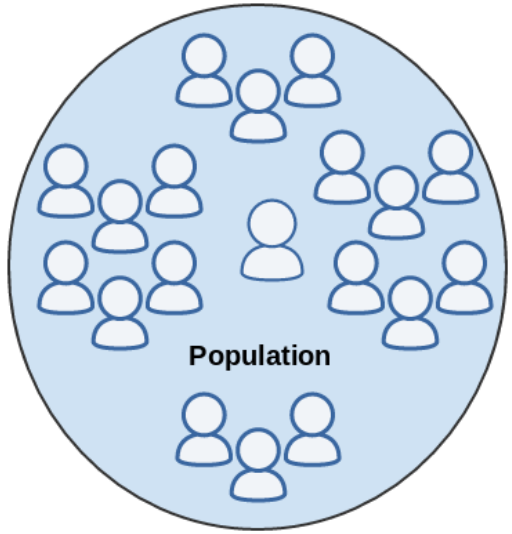
A **biostatistician** is a person who's work is driven by questions relating to the health of people, as individuals or members of population.

In a nutshell they can be classified as:

- "Data detectives"
- "Data judges", they are in charge of confirming and making decisions using *inferential methods*

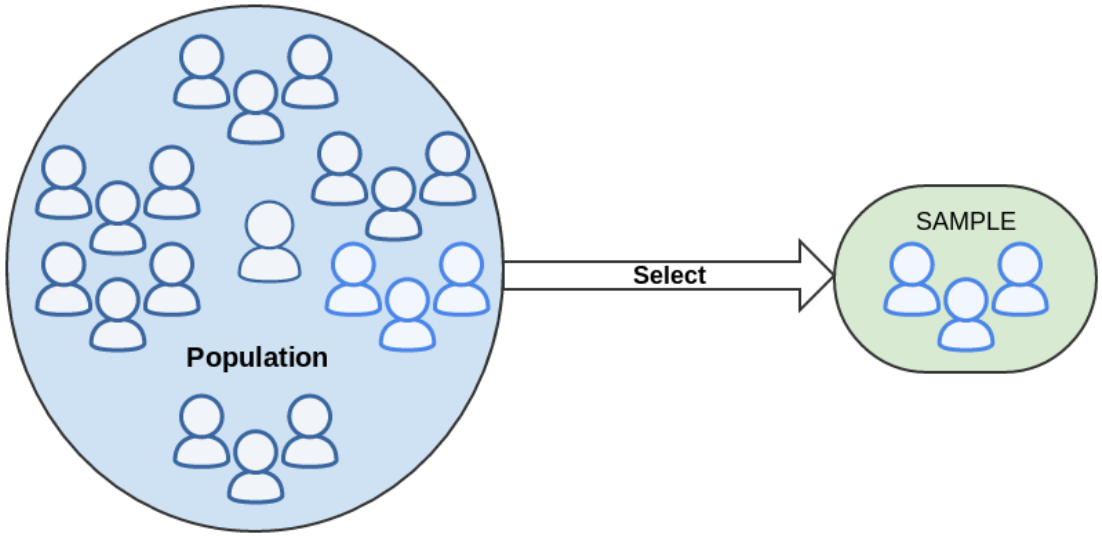
Basic notions

- **Population** represents the total of individuals having a common set of characteristics



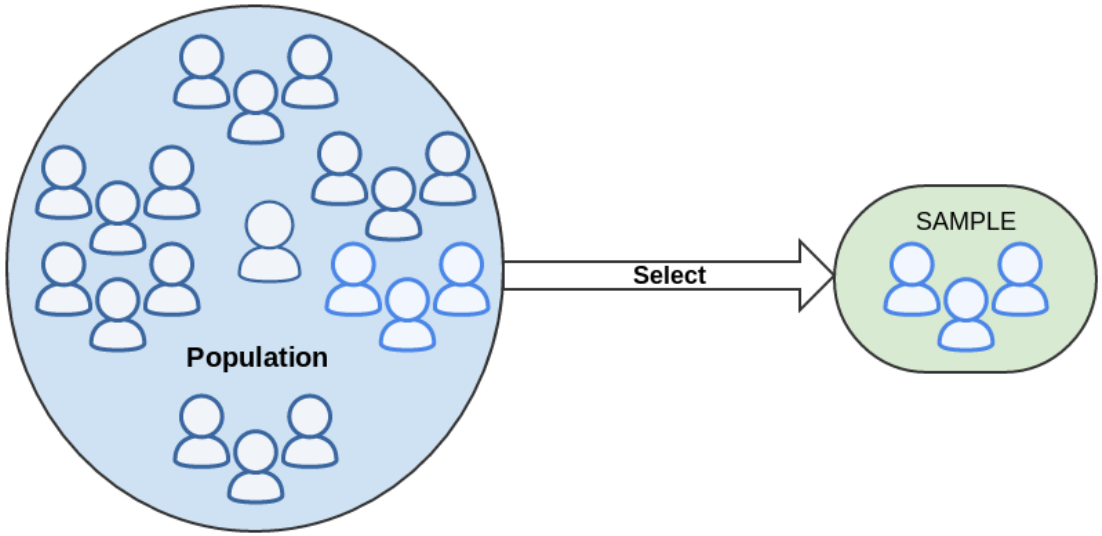
Basic notions

- **Population** represents the total of individuals having a common set of characteristics
- The **Sample** represents a subgroup of the population that is analysed



Basic notions

- **Population** represents the total of individuals having a common set of characteristics
- The **Sample** represents a subgroup of the population that is analysed



- Representative sample?

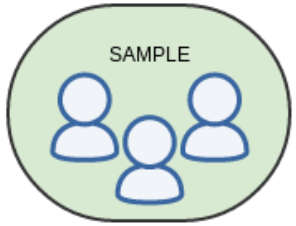
Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.

Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.

Starting with a *Sample*,



Statistical inference

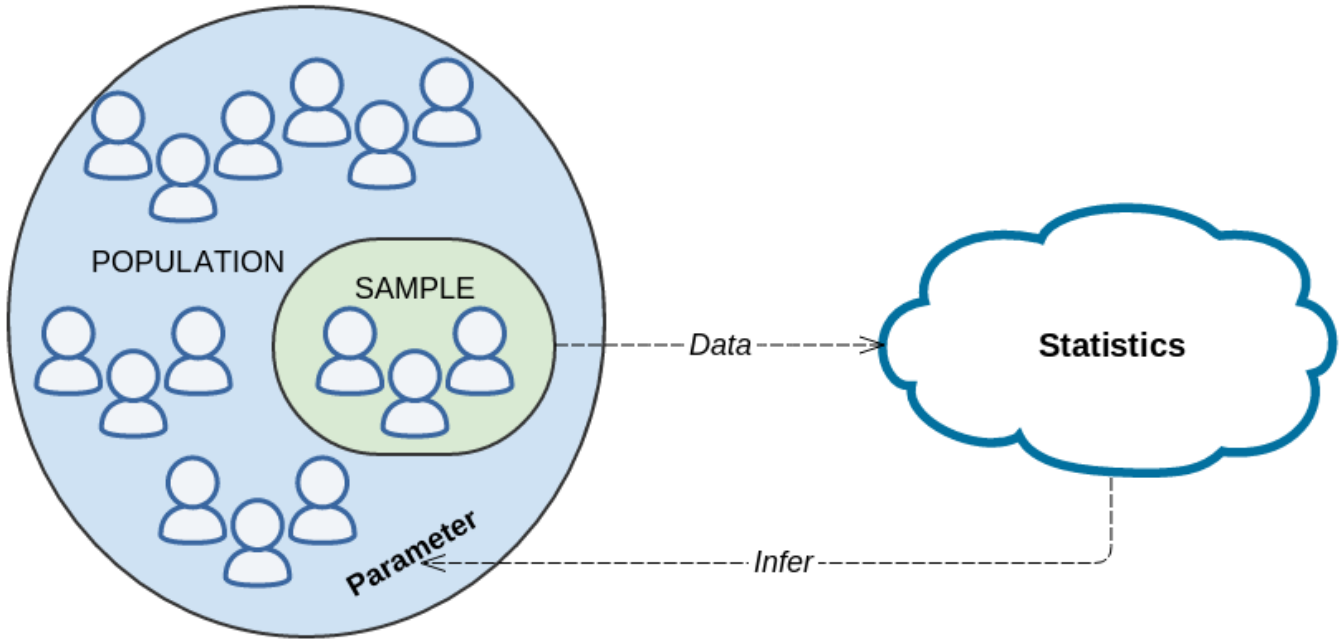
Statistical inference is the process that generalizes characteristics of a sample to the entire population.

Starting with a *Sample*, we are interested in finding out parameters about a *Population*

Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.

Starting with a *Sample*, we are interested in finding out parameters about a *Population* . . . but we can only compute statistics on the sample.



Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.

- Measurement: assigning numbers (codes) according to prior-set rules

Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.

- Measurement: assigning numbers (codes) according to prior-set rules
- Variable: what is measured (i.e. age, color of the eyes etc.)

Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.

- Measurement: assigning numbers (codes) according to prior-set rules
- Variable: what is measured (i.e. age, color of the eyes etc.)
- Value: the result of a measurement (i.e. 21 y., blue etc.)

Measurement and Variable

- Categorical (Nominal) : qualitative or quantitative; eg. color of a ball or breed of dog etc.

Measurement and Variable

- Categorical (Nominal) : qualitative or quantitative; eg. color of a ball or breed of dog etc.
- Ordinal (Rank) : often arbitrary and has the purpose to provide a ranking on data eg. ranking of students, pain levels etc.

Measurement and Variable

- Categorical (Nominal) : qualitative or quantitative; eg. color of a ball or breed of dog etc.
- Ordinal (Rank) : often arbitrary and has the purpose to provide a ranking on data eg. ranking of students, pain levels etc.
- Quantitative (Scale) : measured on a scale; eg. A country's population, person shoe size, car speed etc.

Nominal variables

Nominal variables are classifying observations in named categories.
Properties:

- Cannot be ordered
- Individuals that belong to the same category are regarded as equivalent
- Examples:
  - color of the eyes
  - taste
  - gender
  - blood group (0, A, B, AB).

Ordinal variables

Categories can be arranged in a rank order.
Examples:

- Stages of cancer: I, II, III, IV
- Opinion: strongly agree, agree, neutral, disagree, strongly disagree
- Position in an ordered set of data

Quantitative variables

- *Numerical* values having *equal spacing* between them

Quantitative variables

- *Numerical* values having *equal spacing* between them
- Sales are *proportional*

Quantitative variables

- *Numerical* values having *equal spacing* between them
- Sales are *proportional*
- Usually reported to an *Unit of Measurement*

Quantitative variables

- *Numerical* values having *equal spacing* between them
- Sales are *proportional*
- Usually reported to an *Unit of Measurement*

Example Height (um: cm or m), Blood pressure (millimetres of mercury mmHg) etc.

Visual description of samples

- Frequency chart.

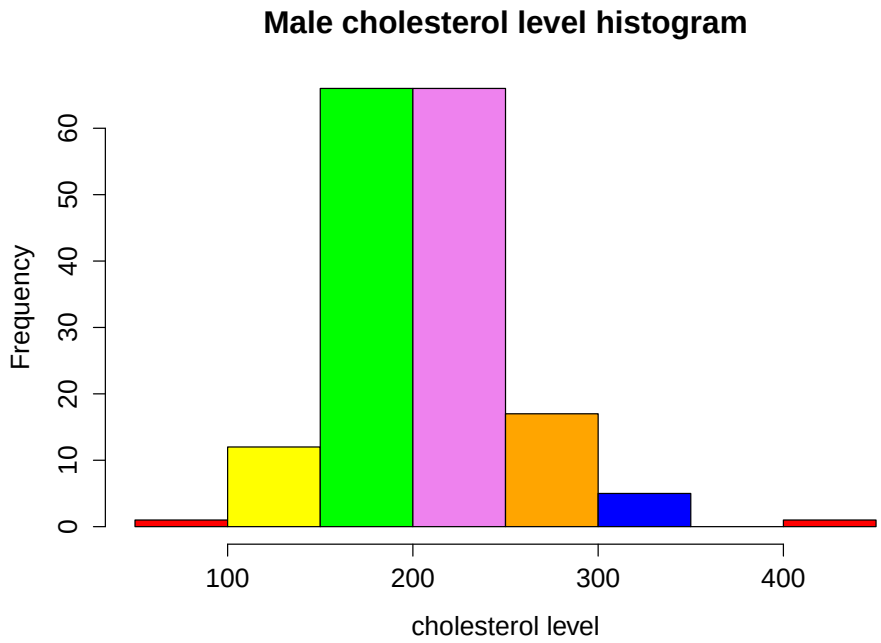
make

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	AMC	3	11.5	11.5	11.5
	Audi	2	7.7	7.7	19.2
	BMW	1	3.8	3.8	23.1
	Buick	7	26.9	26.9	50.0
	Cad.	3	11.5	11.5	61.5
	Chev.	6	23.1	23.1	84.6
	Datsun	4	15.4	15.4	100.0
	Total	26	100.0	100.0	

Visual description of samples

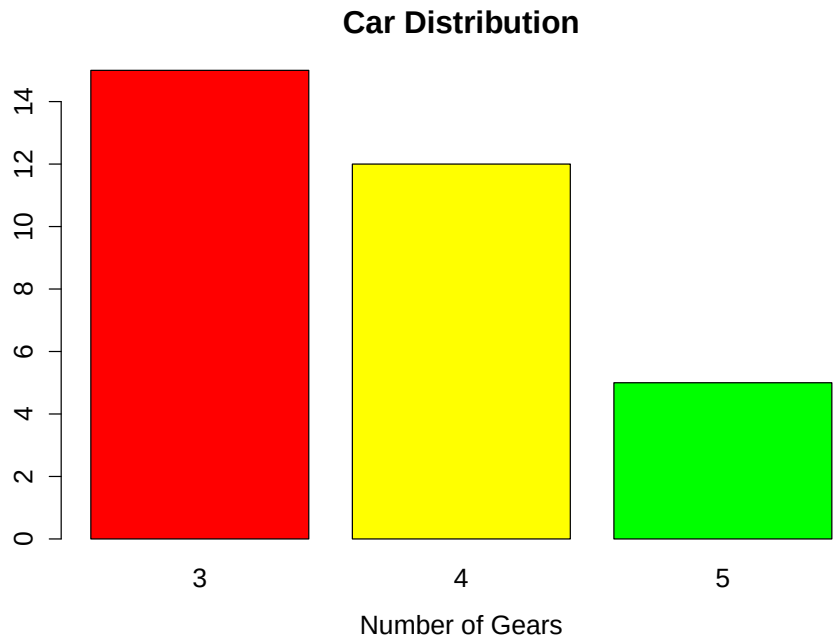
- Frequency chart.
- Histogram. Note that this type of representation is valid only for quantitative data.

```
 78 249 248 195 227 177 191 230 194 186 234 232
184 158 214 292 218 244 283 186 194 174 225 268
197 205 174 192 205 143 169 174 188 215 207 179
234 191 181 293 198 135 212 162 255 404 239 204
307 157 235 202 194 227 337 255 289 209 214 144
```



Visual description of samples

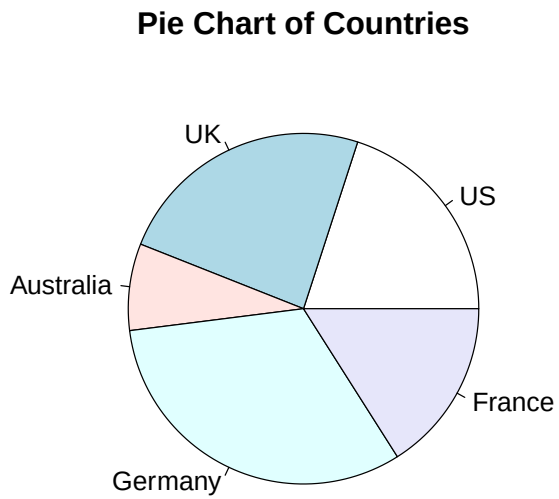
- Frequency chart.
- Histogram.
- Barchart. Note that this representation is valid for qualitative and quantitative measurements



Visual description of samples

- Frequency chart.
- Histogram.
- Barchart.
- Piechart. Note that this type of representation is valid only for qualitative measurements.

Country	Count
US	10
UK	12
Australia	4
Germany	16
France	8



Central tendency

Average: represents the arithmetic mean of the values from a sample.

$$Y_m = \sum_{i=1}^{n} \frac{Y_i}{n}$$

Properties:

- appropriate for numerical values
  e.g: sample $Y = \{2, 4, 6, 8, 10\}$

Central tendency

Average: represents the arithmetic mean of the values from a sample.

$$Y_m = \sum_{i=1}^{n} \frac{Y_i}{n}$$

Properties:

- appropriate for numerical values
  e.g: assume we have the following sample $Y = \{2, 4, 6, 8, 10\}$ then

$$Y_m = \sum_{i=1}^{5} \frac{Y_i}{5} = \frac{2+4+6+8+10}{5} = 6$$

Central tendency

Average: represents the arithmetic mean of the values from a sample.

$$Y_m = \sum_{i=1}^{n} \frac{Y_i}{n}$$

Properties:

- appropriate for numerical values
  e.g: sample $Y = \{2, 4, 6, 8, 10\}$ and $Y_m = 6$
- strongly influenced by extreme values
  e.g: $Y = \{2, 4, 6, 8, 10, 100\}$ then

$$Y_m = \sum_{i=1}^{6} \frac{Y_i}{6} = \frac{2 + 4 + 6 + 8 + 10 + 100}{6} = 21.66$$

Central tendency

Average: represents the arithmetic mean of the values from a sample.

$$Y_m = \sum_{i=1}^{n} \frac{Y_i}{n}$$

Properties:

- appropriate for numerical values
  e.g: sample $Y = \{2, 4, 6, 8, 10\}$ and $Y_m = 6$
- strongly influenced by extreme values
  e.g: $Y = \{2, 4, 6, 8, 10, 100\}$ and $Y_m = 21.66$
- "gravitational center"

Central tendency

Median: represents the value that splits an *ordered* array into two equal parts.

Properties:

- Appropriate for ordinal variables

$$Y = \{2, 4, 6, 8, 10\} \Rightarrow Me = 6$$

- Appropriate for quantitative variables
- **NOT** influenced by extremes

$$Y = \{2, 4, 6, 8, 10, 100\} \Rightarrow Me = \frac{6+8}{2} = 7$$

Central tendency

Mode: represents the most frequent value in a sample (array).

- Appropriate for nominal variables
- Appropriate for ordinal variables
- Appropriate for quantitative variables
- Not influenced by extremes

$$Y = \{2, 3, 4, 2, 3, 2, 10, 3, 2, 2, 100\} \Rightarrow Mo = 2$$

Indicators for Spread

Range

Range represents the distance between the min and max in a sample.
Properties:

- easy to compute
- influenced only by *min* and *max*
- not the most appropriate

Example:

$$Y = \{2, 3, 4, 2, 3, 2, 10, 3, 2, 2, 100\} \Rightarrow Range = 100 - 2 = 98$$

Standard Deviation

Deviation represents the distance from arithmetic mean in a sample.

Standard Deviation

Assume we have Y a sample and $x_i \in Y$ then Standard Deviation is computed as follows:

Standard Deviation

Assume we have Y a sample and $x_i \in Y$ then Standard Deviation is computed as follows:

- Compute sum of squares (SS)

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

where $\bar{x}$ is the arithmetic mean of the sample

Standard Deviation

Assume we have Y a sample and $x_i \in Y$ then Standard Deviation is computed as follows:

- Compute sum of squares (SS)
- Compute the variance (s^2)

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

Standard Deviation

Assume we have Y a sample and $x_i \in Y$ then Standard Deviation is computed as follows:

- Compute sum of squares (SS)
- Compute the variance (s^2)
- Compute the standard deviation (SD)

$$SD = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$

Standard error of mean

Describes the spread of the samples mean around population mean.

$$S_{\bar{x}} = \frac{SD}{\sqrt{N}}$$

Spread in case of ordinal values

Quantiles: split the array in "n" values.

- Quartile $n = 4$
- Decile $n = 10$
- Centile $n = 100$
- Promile $n = 1000$

Choosing the right method

- Nominal data
  - Frequency of observations
  - Mode
- Ordinal data
  - Frequency of observations
  - Median
  - Mode
  - Quantiles
- Numerical data
  - All types of descriptions are possible

Probabilities

$$P(E) = \frac{number\ of\ occurences}{number\ of\ all\ possible\ occurences}$$

Values taken by $P(E) \in [0, 1]$

Key property: sum of all probabilities for a given sample is 1.

$$\sum_{i=1}^{n} P_i(E) = 1$$

Probabilities

Q: How to quantify uncertainty?

Probabilities

Q: How to quantify uncertainty?
A: Probabilities allow us to quantify levels of belief

Probability	Expression
0.00	Never
0.005	Seldom
0.20	Infrequent
0.50	As often as not
0.80	Very frequent
0.95	Highly likely
1.00	Always

Distributions

Distributions describe the probability to encounter a value in the studied population.

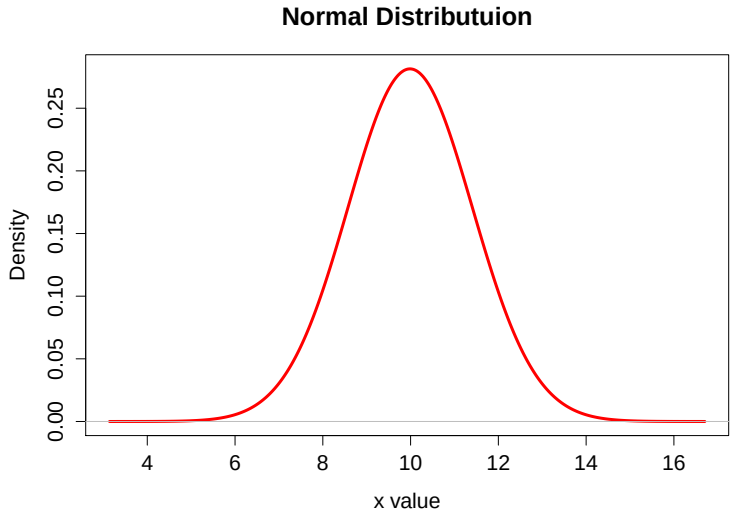


Figure: Symmetrical, bell-shaped

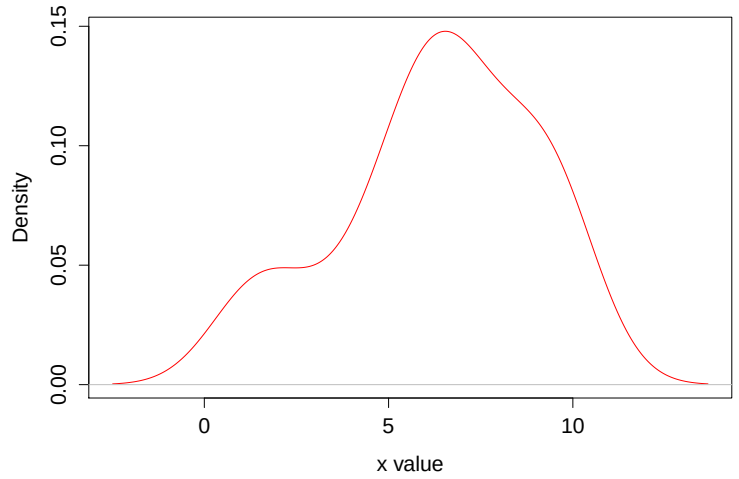


Figure: Non-symmetrical, not bell shaped

Distributions

Distributions describe the probability to encounter a value in the studied population.

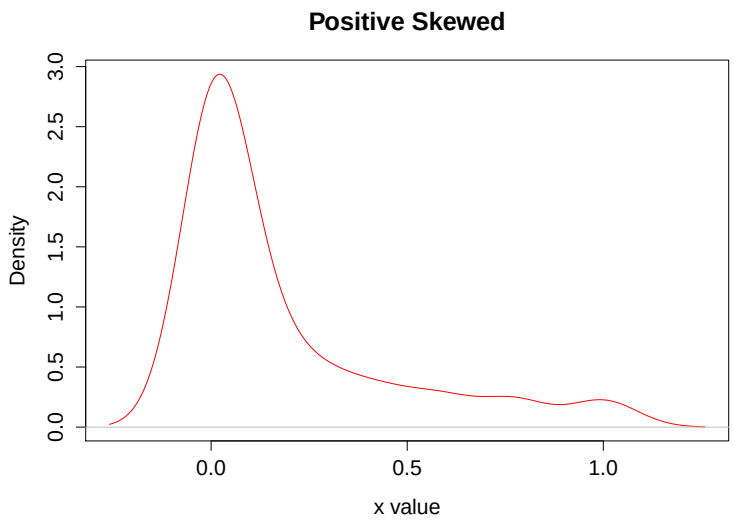


Figure: Positive Skewed

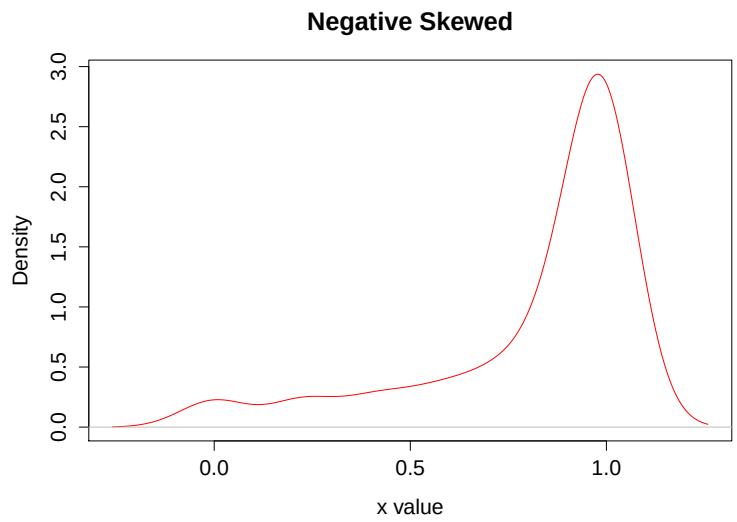
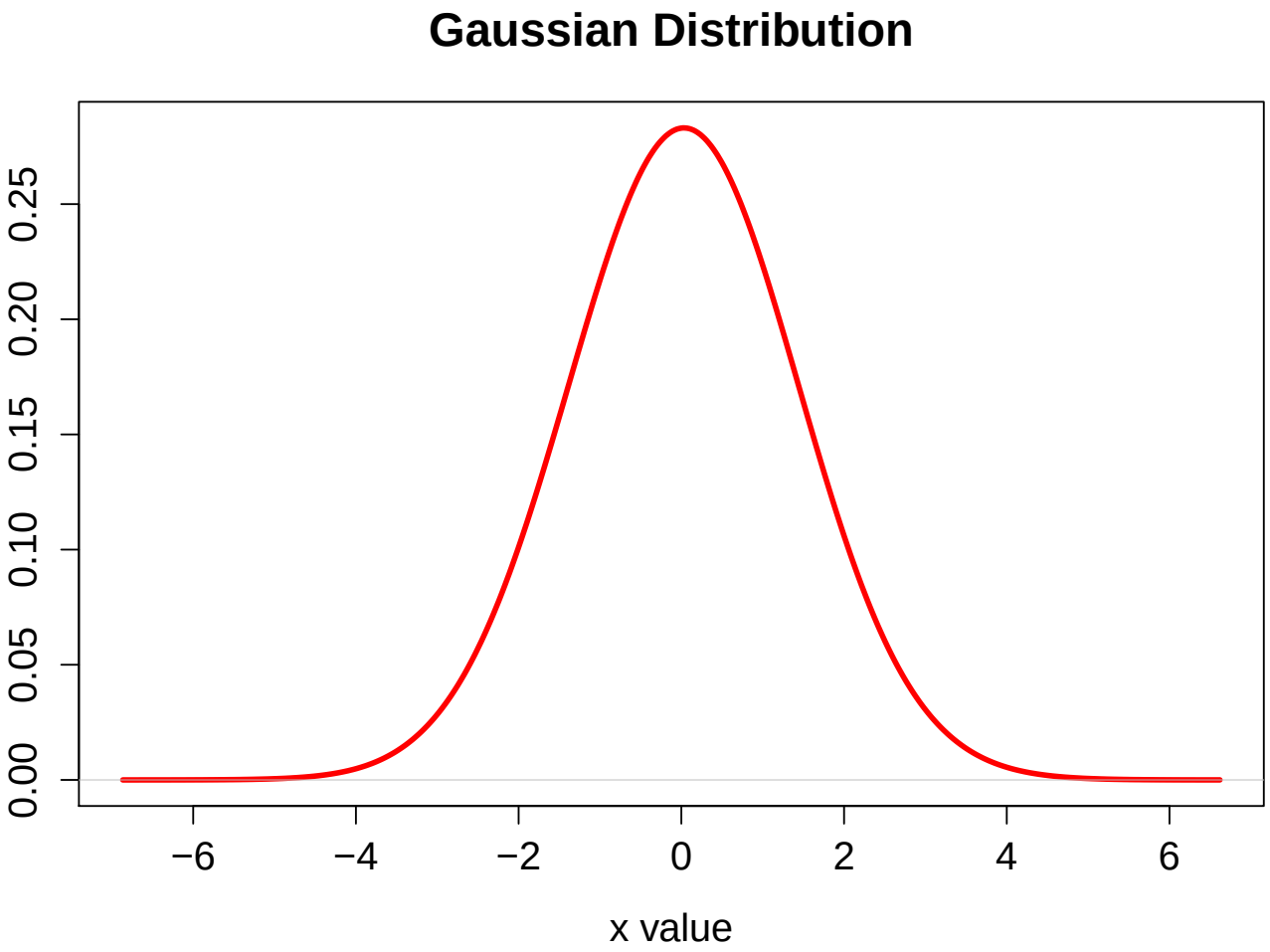
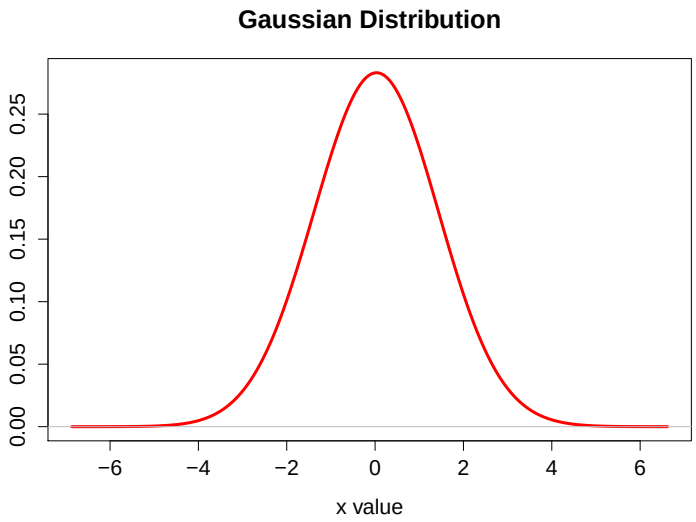


Figure: Negative Skewed

Gaussian distribution



Gaussian distribution

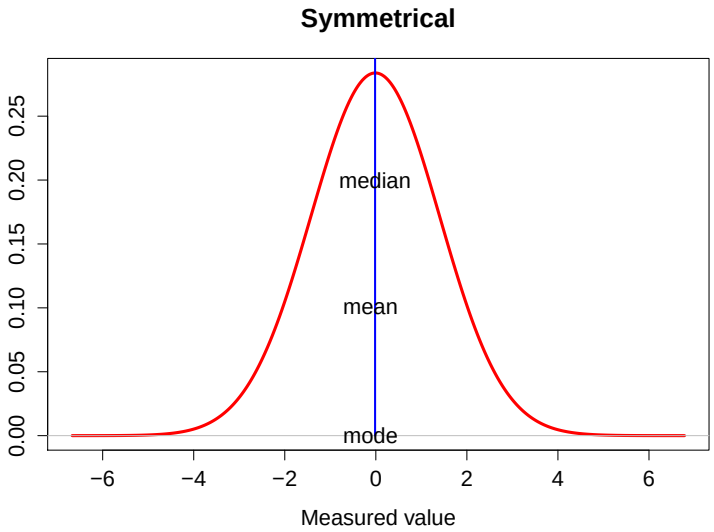


Important properties

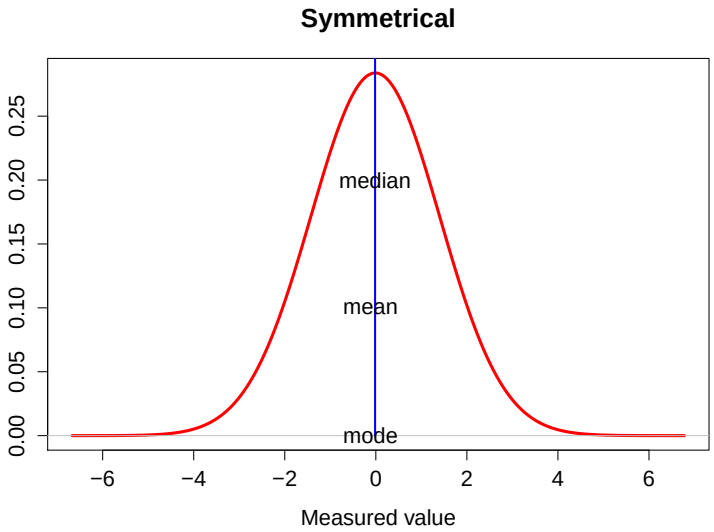
- Continuous distribution
- Maximum values at central tendency
- Symmetrical
- Lower values at extremes

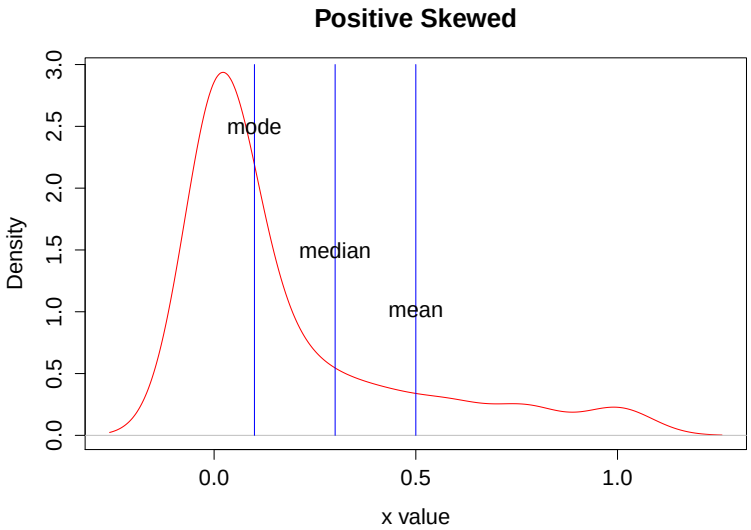
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Central tendency

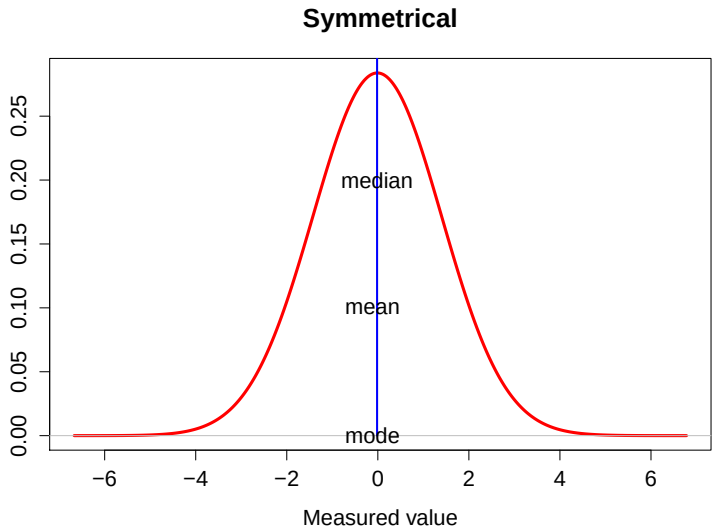


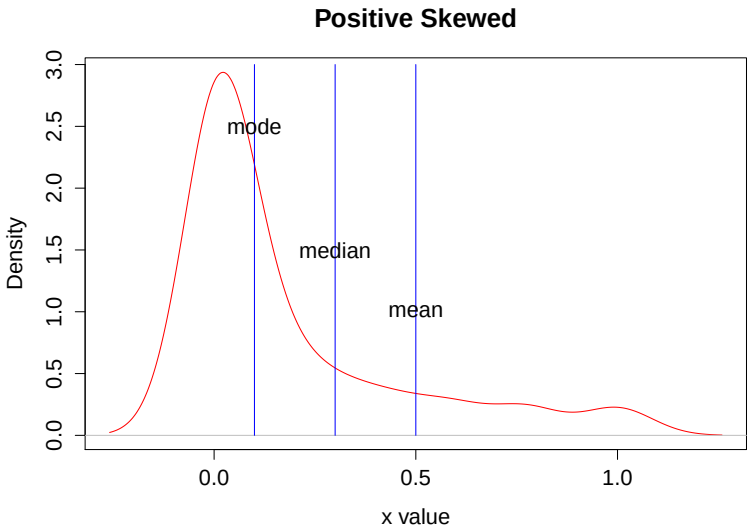
Central tendency

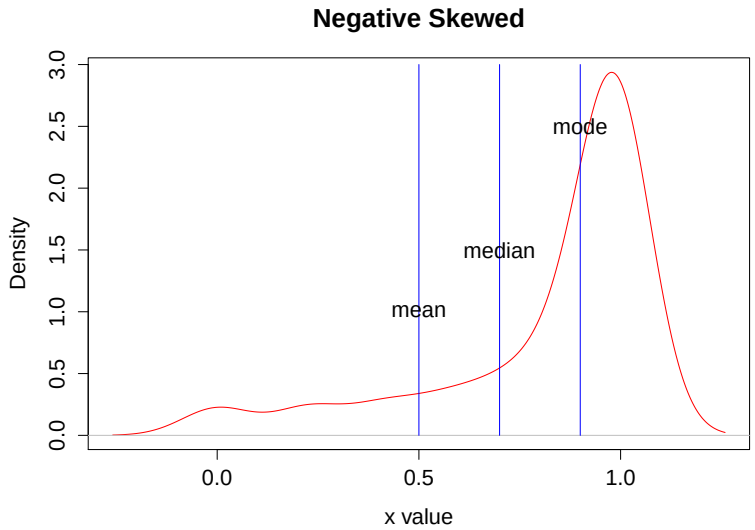




Central tendency





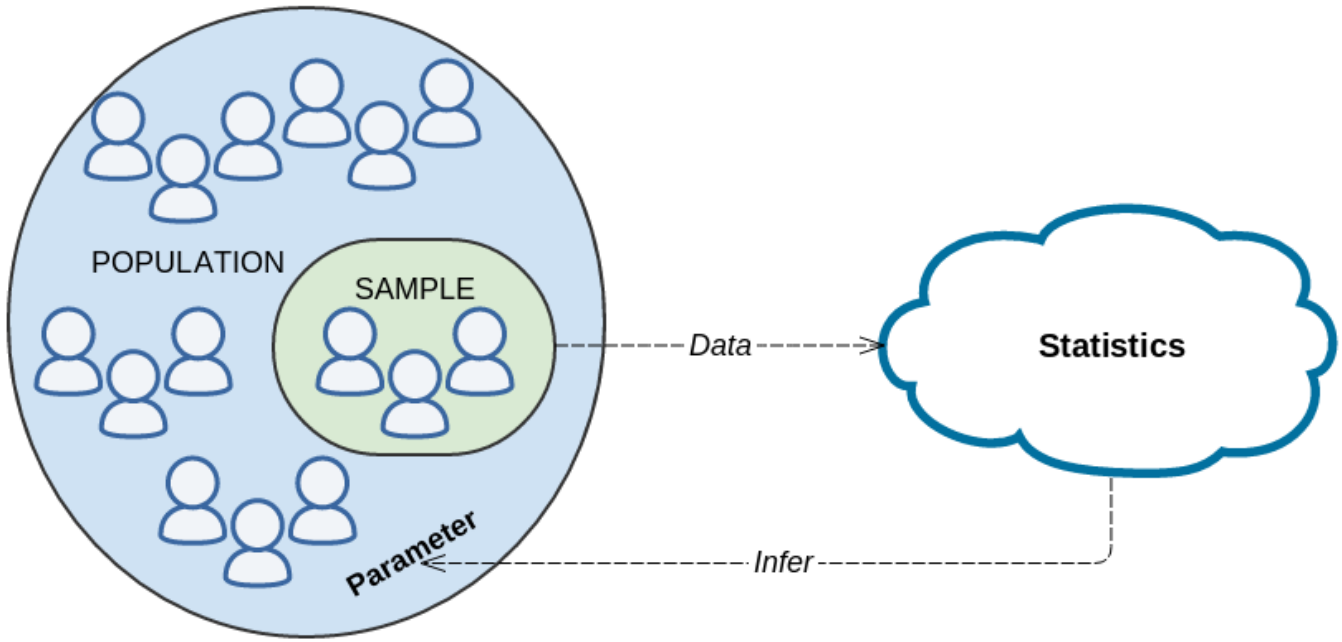


Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.

Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.
Starting with a *Sample*, we are interested in finding out parameters about a *Population* . . . but we can only compute statistics on the sample.



Statistical inference

Statistical inference is the process that generalizes characteristics of a sample to the entire population.

	Population Parameter The reality Greek letters	Sample Statistics Estimates the reality Latin letters
Mean	μ	X
Standard Deviation	σ	SD
Number of observations	N	n

Properties

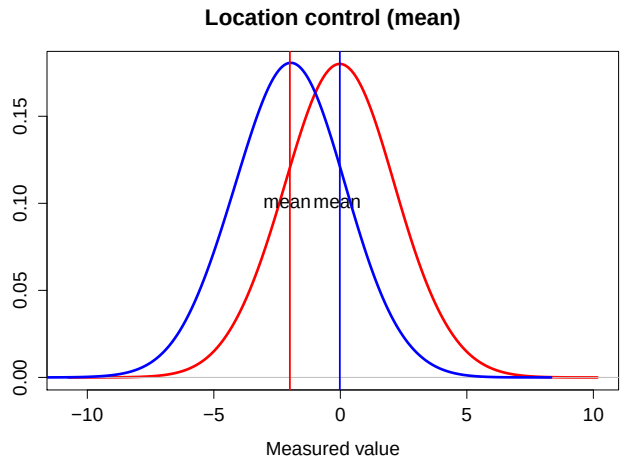


Figure: μ controls the location

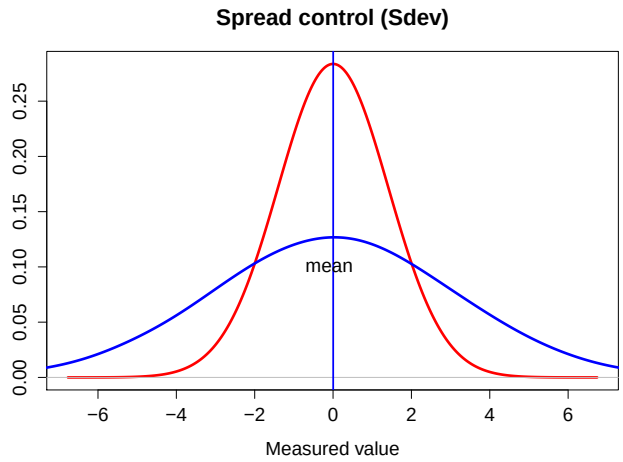


Figure: σ controls the location

Confidence Intervals

- We are unable to calculate exactly the parameters of a population

Confidence Intervals

- We are unable to calculate exactly the parameters of a population
- The value of a statistic is biased by the selection of the sample

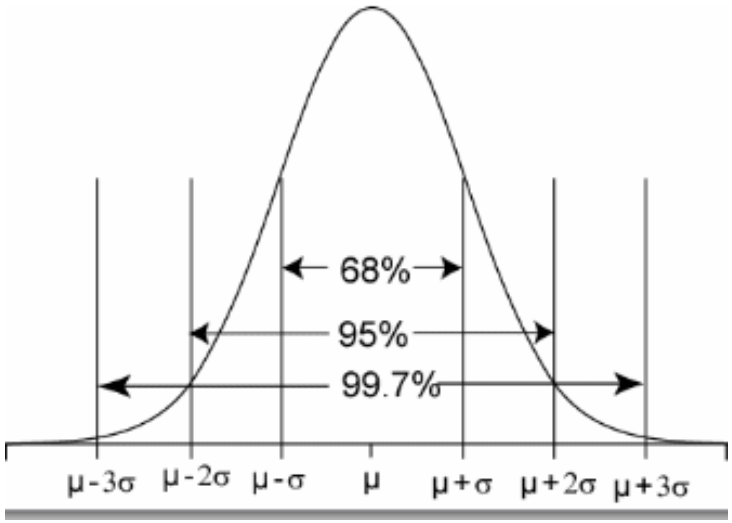
Confidence Intervals

- We are unable to calculate exactly the parameters of a population
- The value of a statistic is biased by the selection of the sample
- By using confidence intervals we are able to estimate a parameter of a population regarding the variations of statistics in samples

Confidence intervals

Definition: Confidence Interval defines an interval in which is probable to have the parameter of the population
Estimating a value interval:

- 68% of individuals can be found at $\pm 1 * \sigma$ around μ
- 95% of individuals can be found at $\pm 2 * \sigma$ around μ
- 99.7% of individuals can be found at $\pm 3 * \sigma$ around μ



Estimating the mean

$$SEM = \frac{SD}{\sqrt{N}}$$

- $(1 - \alpha) = 0.68 \Rightarrow X = \mu \pm 1 * SEM$
- $(1 - \alpha) = 0.95 \Rightarrow X = \mu \pm 2 * SEM$
- $(1 - \alpha) = 0.997 \Rightarrow X = \mu \pm 3 * SEM$