Descriptive Statistics

Univariate Descriptive Statistics

Benchikh Tawfik

Faculty of Medicine

1st Year Medicine

October 5, 2025

Plan du cours

DISCRETE QUANTITATIVE VARIABLE: DEFINITION

- A **discrete quantitative variable** is a numerical variable that can only take specific, separate values, most often integers (and occasionally decimals). Mathematically, its values belong to a *countable set* (e.g., $\mathbb{N}$).
- The number of distinct values (also called **categories** or **modalities**) is typically limited, often fewer than twenty in practical applications.

DISCRETE QUANTITATIVE VARIABLE: EXAMPLES

- Number of children in a family;
- Number of years of study after high school (e.g. $0, 1, 2, 3, \ldots$);
- Heart rate measured in beats per minute (e.g. $60, 72, 85$);
- Number of vaccine doses received (e.g. $1, 2, 3$ injections).

Discrete Quantitative Variable: Remark

- Suppose we record the age (rounded to the nearest year) of 26 employees in a company. The raw data are: $43, 29, 57, 45, 50, 29, 37, 59, 46, 31, 47, 24, 33, 38, 49, 31, 62, 60, 52, 36, 38, 26, 41, 27, 52, 60$.
- Here, the variable **age** is technically discrete.
- However, the number of distinct values observed is 21, which is relatively large compared to the sample size.
- When a discrete variable takes on many different values (with small gaps between them), it often behaves like a **continuous variable**.
- This illustrates a **general rule:** *Discrete variables with a large number of possible values are often analyzed as if they were continuous.*

Absolute, Relative, and Cumulative Frequencies: Example

- In a survey of $n = 69$ families, we record the number of children per family. The raw data may look as follows: $1, 3, 1, 4, 1, 2, 2, 0, 3, 0, 0, 4, 2,$ $0, 2, 2, 3, 3, 1, 3, 2, 0, 3, 1, 0, 0, 2, 1, 1, 2, 3, 1, 1, 4, 4, 0, 0, 3, 2, 1, 3, 2, 2, 2, 1, 2, 1, 1,$ $0, 0, 1, 3, 2, 2, 0, 1, 1, 3, 3, 0, 1, 3, 3, 0, 2, 0, 0, 1, 2.$
- This sequence is unreadable in its raw form.
- The best way to summarize and interpret these data is to construct a **statistical table** showing the absolute, relative, and cumulative frequencies.

Statistical Table: Ordering the Data

- A statistical frequency table is built by listing the r distinct values of the variable X, sorted in increasing order (without repetition). We denote them by $\{x_j\,;\, j = 1, \ldots, r\}$.
- For each value x_j, we record:
  - its **absolute frequency:** n_j (number of occurrences),
  - its **relative frequency:** $f_j = \dfrac{n_j}{n} \times 100$ (percentage),
  - its **cumulative absolute frequency:** N_j, and
  - its **cumulative relative frequency:** F_j.

Value x_i	Absolute freq. n_i	Relative freq. f_i	Cumulative abs. N_i	Cumulative rel. F_i

Absolute, Relative, and Cumulative Frequencies: 2

- Where, for a discrete variable X with r **modalities**, we have:
  - **Absolute frequency** (or **count**) n_i associated with a value x_i: the number of times x_i appears in the dataset (sample or population).
  - **Relative frequency** f_i associated with the value x_i:

$$f_i = \frac{n_i}{n}$$

where n_i is the absolute frequency of x_i and n is the total number of observations (sample size).

Absolute, Relative, and Cumulative Frequencies: 3

- **Cumulative absolute frequency** associated with x_i: the number of individuals whose measurement is less than or equal to x_i: $N_i = \sum_{k=1}^{i} n_k$
- **Cumulative relative frequency**: $F_i = \sum_{k=1}^{i} f_k$
- Note that: $N_r = \sum_{j=1}^{r} n_j = n$ and $F_r = \sum_{j=1}^{r} f_j = 1.$

Remarks

- Relative frequencies and cumulative relative frequencies are especially useful for comparing two or more populations.

Frequency Table: Example 1

- Survey of $n = 69$ families.
- Variable studied: **number of children per family**.
- Observed distribution:
  - 16 families have 0 children $\Rightarrow$ $n_1 = 16$, $f_1 = \frac{16}{69} = 0.23$.
  - 18 families have 1 child $\Rightarrow$ $n_2 = 18$, $f_2 = \frac{18}{69} = 0.26$.
  - 17 families have 2 children $\Rightarrow$ $n_3 = 17$, $f_3 = \frac{17}{69} = 0.25$.
  - 14 families have 3 children $\Rightarrow$ $n_4 = 14$, $f_4 = \frac{14}{69} = 0.20$.
  - 4 families have 4 children $\Rightarrow$ $n_5 = 4$, $f_5 = \frac{4}{69} = 0.06$.

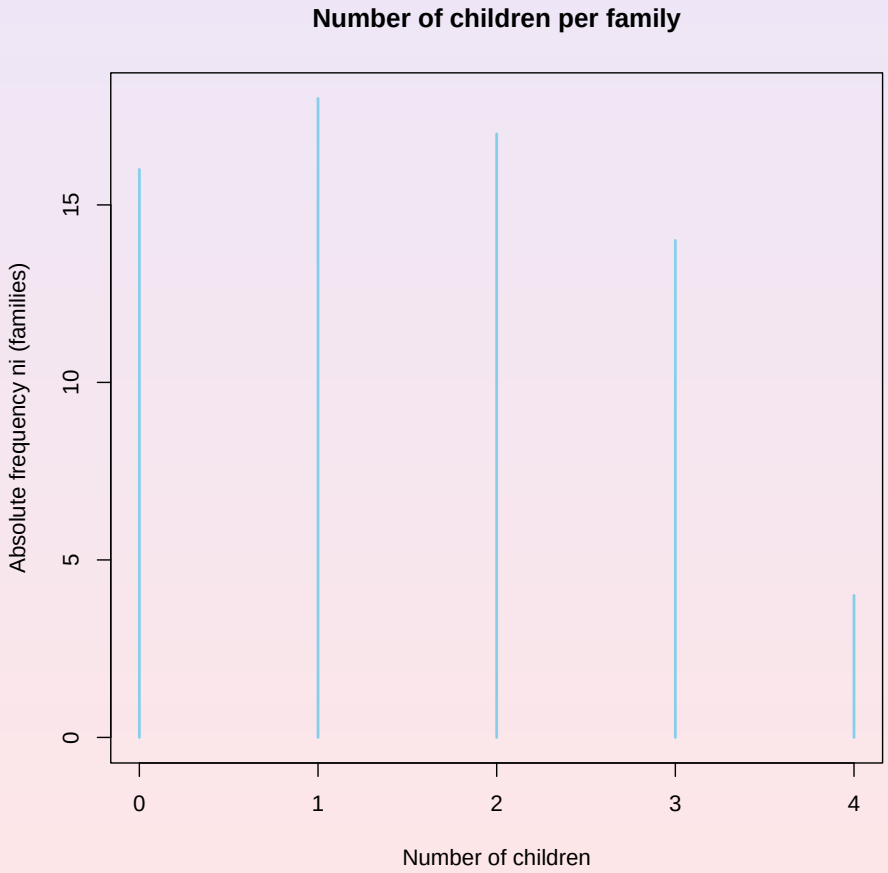
Frequency Table: Example 1 (continued)

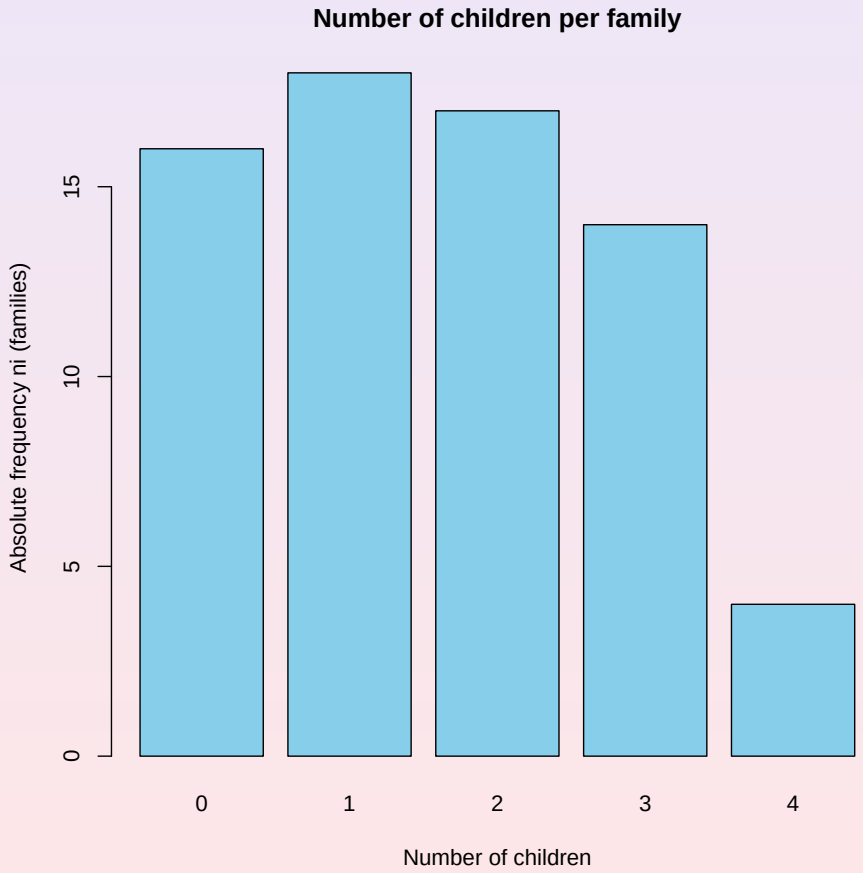
numb. of Children (x_i)	Count n_i	Relative f_i	Cumulative Count N_i	Cumulative Rel. F_i
0	16	0.23	16	0.23
1	18	0.26	34	0.49
2	17	0.25	51	0.74
3	14	0.20	65	0.94
4	4	0.06	69	1.00

- Example: $N_2 = n_1 + n_2 = 16 + 18 = 34$; $N_4 = n_1 + n_2 + n_3 + n_4 = 65$.

Bar Chart

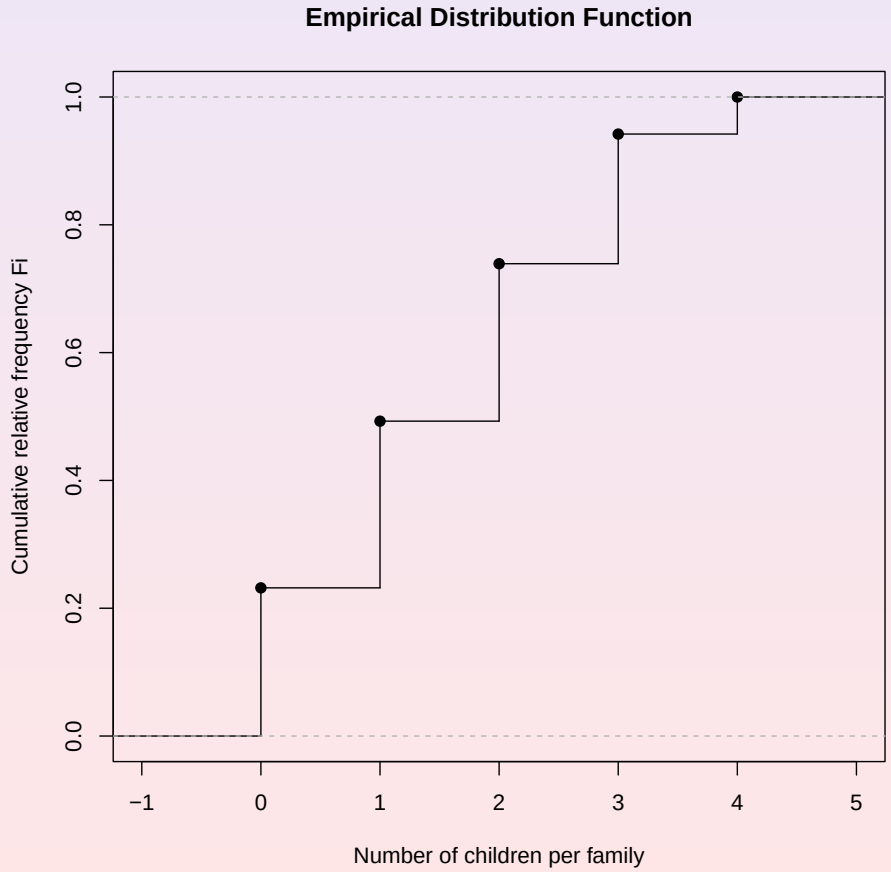
- Provides an overall view of the observed data.
- Example: number of children in 69 families.





THE CUMULATIVE DIAGRAM

- It allows us to determine the number (or proportion) of observations that are less than or equal to a given value in the dataset.
- Example: number of children in 69 families.



- The numerical characteristics (or summaries) introduced here serve to condense a statistical dataset into a small number of numerical values, classified into two main categories:
  - **measures of central tendency**,
  - **measures of dispersion**.

Measures of Central Tendency (or Position)

- Their objective is to provide a representative value for the dataset, i.e., to locate its center or midpoint.
- The main measures of central tendency are: the **arithmetic mean**, the **median**, the **mode**, and the **quantiles**.

Measure of Position: The Mean

- The **mean** of a dataset is defined as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{r} n_i x_i \,,$$

  where r is the number of modalities of the variable X.

- **Properties:**
  - The arithmetic mean summarizes the entire dataset with a single number.
  - It takes into account all observed values and is easy to compute.
  - It is highly sensitive to extreme values (outliers). In some cases, removing such outliers is necessary.
  - The quantity $e_i = x_i - \overline{X}$ is the deviation of x_i from the mean.

Mean: Examples

- Example "Number of children per family": $\overline{X} = 1.59 \backsimeq 2$ children per family.
- For the dataset: 1,1,1,39,39,39, we obtain $\overline{X} = 20$ (which is not meaningful).

Measure of Position: The Mode

- The **mode** of a dataset is the value that occurs most frequently, i.e., the one with the highest frequency.
- **Properties:**
  - The mode may not exist, and if it does, it may not be unique.
  - A dataset with only one mode is called unimodal.

Mode: Examples

1. Dataset: number of children per family $\Rightarrow$ mode $= 1$.
2. Dataset: 3,5,8,10,12,15,16 $\Rightarrow$ no mode.
3. Dataset: 2,3,4,4,4,5,5,7,7,7,9 $\Rightarrow$ two modes (4 and 7), hence bimodal.

Measure of Position: The Median

- If the values of a dataset are ordered, the **median** is the central value, i.e., the smallest value m_e such that at least half of the observations are less than or equal to it, and at least half are greater or equal.

Median: Computation

- Two cases:
  - If the dataset has an odd number of values ($N = 2k + 1$), the median is the $(k + 1)^{st}$ observation.
  - If the dataset has an even number of values ($N = 2k$), the median lies between the two central values, often taken as their arithmetic mean (which may not correspond to an actual observation).

Median: Properties

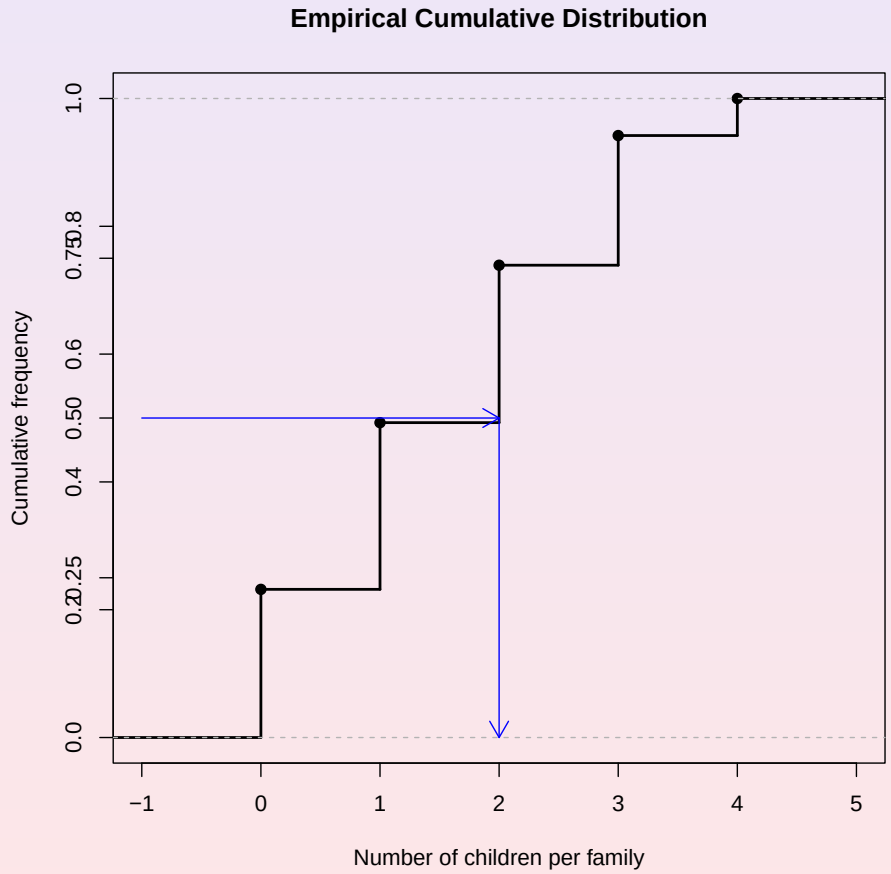
- The median is less sensitive than the mean to outliers.
- It depends on the number of data points but not their magnitude, and therefore cannot be used in estimation theory.
- For discrete variables, the median may not correspond to an actual observed value.
- The median is the intersection point of the increasing and decreasing cumulative curves.

Median: Examples

- Dataset: 1,2,4,4,4,5,6,7,8,8,9,9,10,11,12. $n = 15$, so $M_e = x_8 = 7$.
- Dataset: 4,5,8,8,9,11,12,14,17,19. $N = 10$, so $M_e = \frac{x_5 + x_6}{2} = 10$ (not an observed value).
- Dataset: 2,2,2,3,3,3,4,4,4,5,5. $N = 11$, so $M_e = x_6 = 3$.
- Example "Number of children per family": $m_e = 2$.

Graphical Determination of the Median

- The median can also be obtained graphically using the cumulative distribution curve:



Measure of Position: Quantiles

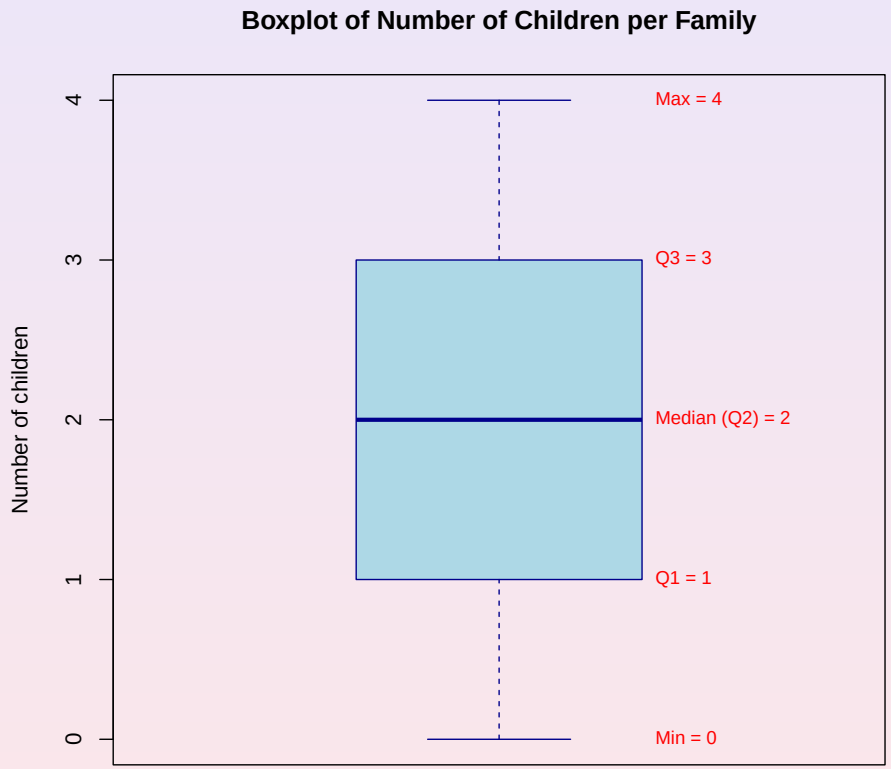
- **Quantiles** divide an ordered dataset into k equal parts:
  1. For $k = 4$, the quantiles are the **quartiles** Q_1, Q_2 (median), and Q_3. The interquartile range $Q_3 - Q_1$ contains 50% of the data.
  2. For $k = 10$, the quantiles are **deciles** $D_1, \ldots, D_9$, each corresponding to 10% of the data.
  3. For $k = 100$, the quantiles are **percentiles**, with 99 values, each representing 1% of the data.

Quartiles: Boxplot (Tukey)

- The **boxplot** is a schematic representation of a distribution based on quartiles. The interquartile range defines the central box, the median is shown inside, and whiskers extend to the minimum and maximum values.

Boxplot: Example

- Example: "Number of children per family"



Measures of Dispersion

- These measures quantify the spread of the values around the center. The main ones are: the variance (or its square root, the standard deviation), the coefficient of variation, and the range.

Measure of Dispersion: Range

- The range is defined as:

$$E = x_{max} - x_{min}.$$

- **Properties**
  - Easy to compute.
  - Depends only on extreme values, ignoring intermediate ones.
  - Used in industrial quality control with small samples (typically $n \leq 5$).
  - For larger samples ($n \geq 15$), the standard deviation is preferred.

Range: Example

- Dataset: 0,10,11,12,13,14,14. $E = 14 - 0 = 14$, which suggests wide variability, even though most values are close together.

Measure of Dispersion: Interquartile Range

- The interquartile range $[Q_1, Q_3]$ provides a robust measure of variability.

Measure of Dispersion: Variance and Standard Deviation

- The **variance** of a dataset, denoted $s^2 = Var(X)$, is also called the empirical variance.

$$s^2 = Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{r} n_i\, x_i^2 - (\overline{X})^2.$$

- Its square root is the **standard deviation:** $s = \sqrt{s^2}$, which measures the average deviation from the mean:

VARIANCE: PROPERTIES

- The mean $\overline{X}$ and standard deviation s are expressed in the same units as the data.
- A small s indicates that data are tightly clustered around $\overline{X}$ (greater homogeneity).
- Variance incorporates all data points, making it the most widely used measure of dispersion.

Coefficient of Variation: definition

- The Coefficient of Variation (CV) is a standardized measure of dispersion of a dataset, expressed as a percentage of the mean:

$$CV = \frac{s}{\overline{X}} \times 100,$$

- The CV measures relative variability: how large the standard deviation is compared to the mean.
- Allows comparison between datasets with different units or scales.
- It is unit-free, which makes it useful for comparing variability across datasets with different units or scales.

Coefficient of Variation: interpretation / Guidelines (in practice):

- A $CV < 15\%$ Low variability $\rightarrow$ The mean is representative of the data (homogeneous dataset).
- CV between 15% and 30% $\rightarrow$ Moderate variability $\rightarrow$ The mean is less reliable as a central measure.
- If $CV > 30\%$, High variability $\rightarrow$ The mean is not representative; the median is often a better measure.

Variance: Example

- Example: Number of children per family:
  - $\sum x_i = 110$, $\sum x_i^2 = 276$.
  - Variance $s^2 = 1.458$
  - Standard deviation $s = 1.207$
  - $CV = 75.72\% > 30\%$, hence the mean is not representative; the median is more appropriate.

CONTINUOUS QUANTITATIVE VARIABLE: DEFINITION

- A continuous quantitative variable can take an infinite number of possible values.
- Its domain is $\mathbb{R}$ or an interval of $\mathbb{R}$.

Continuous Quantitative Variable: Note

- For example, height can be measured in centimeters or even millimeters. In practice, continuous variables may sometimes be treated as discrete.
- However, for graphical representations and statistical tables, values must be grouped into classes.
- A class-based table is often called a **grouped distribution**.

Statistical Table

- Data are grouped into r classes.
- A **class** is defined by its boundaries e_{i-1}, e_i and its frequency n_i.
- **Absolute frequency**: number n_i of values of X such that $e_{i-1} \leq X < e_i$.
- **Class width (amplitude)**: $e_i - e_{i-1}$.
- **Class midpoint**: $\frac{e_i + e_{i-1}}{2}$.

Statistical Table

- N_i: cumulative frequency of class i.
- f_i: relative frequency of class i.
- F_i: cumulative relative frequency of class i, defined as

$$F_i = \sum_{j=1}^{i} f_j, \quad \text{with } F_1 = f_1.$$

It gives the proportion of observations such that $X < e_i$.

Classes	Absolute frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
$e_{i-1} \leq X < e_i$	n_i	f_i	N_i	F_i

Number of Classes: Practical Guidelines

- The number of classes should not be too small (loss of information), nor too large (unnecessary complexity, very small frequencies).
- Usually, the number of classes is between 5 and 15 (sometimes up to 20), depending on sample size n and data spread.
- e_0 is less than the smallest value; e_r is greater than the largest value. Each observation belongs to exactly one class.

Number of Classes: Practical Guidelines (cont.)

- Classes are generally of equal width, preferably chosen so that midpoints are integers (or with few decimals).
- Sometimes one or two classes may have very large frequencies. In such cases, unequal class widths can be used.
- Open classes (without lower or upper limits) should be avoided if possible.

Number of Classes: Rules of Thumb

- Several formulas can guide the choice of the number of classes r and class width:
  - Sturges' rule: $r = 1 + 3.3 \log_{10}(n)$
  - Yule's rule: $r = 2.5 \sqrt{\sqrt{n}}$
  - General rule: $r = \sqrt{n}$
- In each case, r is rounded to the nearest integer.
- Similarly, the class width can be rounded for convenience.

Class Amplitude

- If classes have equal width: $a = \frac{E}{r}$, where $E = x_{\max} - x_{\min}$.
- If the class width a is chosen first, it should not be too small or too large.
- Each class is represented by its midpoint.
- Grouping into classes sacrifices individual detail and fine structure of the distribution.
- Starting from the minimum value, class boundaries are obtained by adding successive class widths.
- Once grouped, a continuous variable can be analyzed in the same way as a discrete one.

Example

- To study child growth, the height (cm) of 50 children was measured:

152	152	152	153	153	154	154	154	155	155
156	156	156	156	156	157	157	157	158	158
159	159	160	160	160	160	160	161	161	162
162	162	163	164	164	164	164	165	166	167
168	168	168	169	169	170	171	171	171	171

Example: Step 2

- Number of distinct values = 20 $\Rightarrow$ group values into classes.
- Suppose we choose 5 classes starting at 151.5.
- Using Yule's, Sturges' or $\sqrt{n}$ rule gives about 7 classes.
- Class width:

$$a = \frac{\text{Range}}{\text{Number of classes}} = 4 \text{ cm.}$$

EXAMPLE: GROUPED TABLE WITH YULE'S METHOD

Class interval	n_i	f_i	N_i	F_i
$[152, 155[$	8	0.16	8	0.16
$[155, 158[$	10	0.20	18	0.36
$[158, 161[$	9	0.18	27	0.54
$[161, 164[$	6	0.12	33	0.66
$[164, 167[$	6	0.12	39	0.78
$[167, 170[$	6	0.12	45	0.90
$[170, 173[$	5	0.10	50	1.00
Total	50	1.00		

Example: Grouped Table 2

- Statistical table of class distribution with $r = 5$

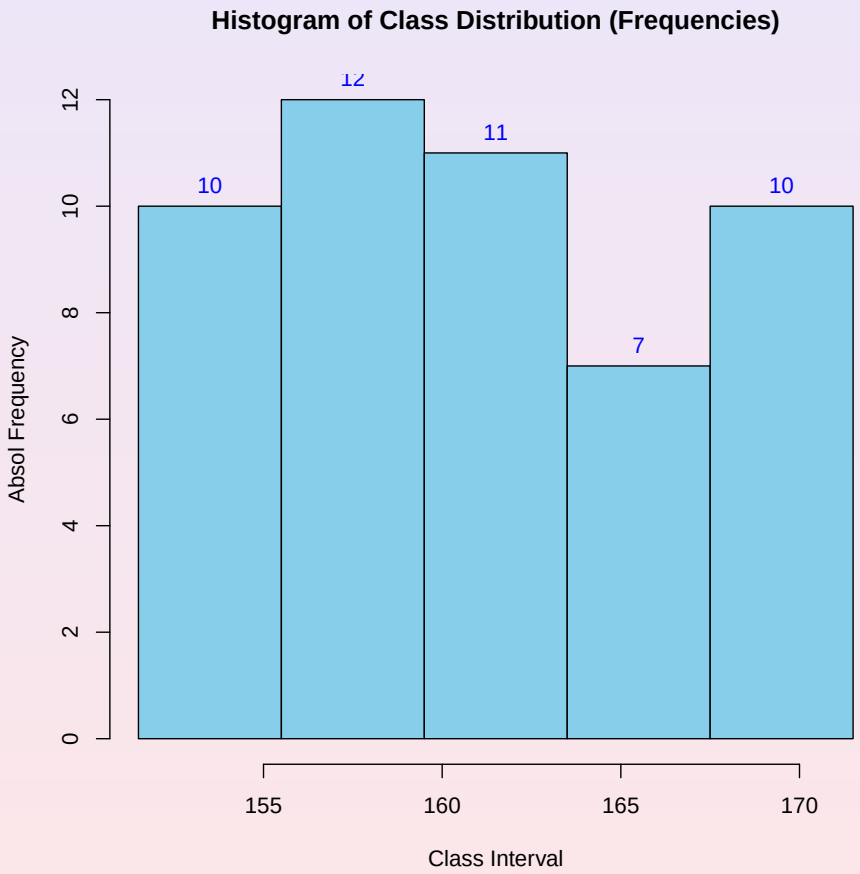
Class interval	n_i	f_i	N_i	F_i
$[151.5, 155.5[$	10	0.20	10	0.20
$[155.5, 159.5[$	12	0.24	22	0.44
$[159.5, 163.5[$	11	0.22	33	0.66
$[163.5, 167.5[$	7	0.14	40	0.80
$[167.5, 171.5[$	10	0.20	50	1.00
Total	50	1.00		

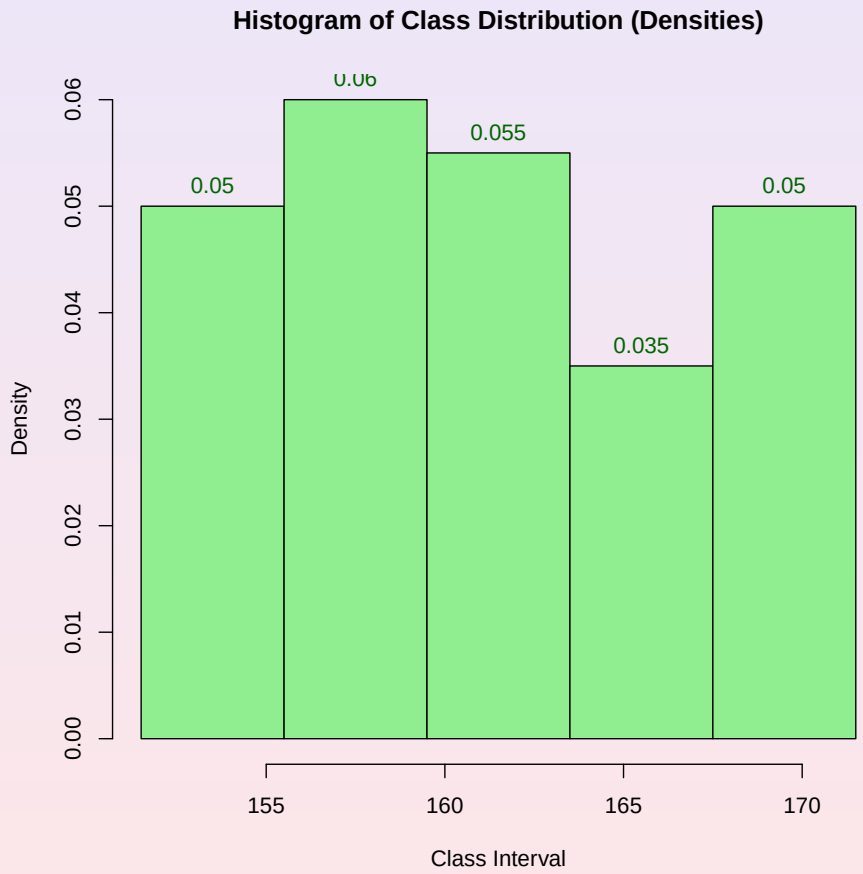
Graphical Representation: Histogram

- A histogram represents the counts (or relative frequencies) of the classes using adjacent rectangles, where the **area** (not the height) of each rectangle corresponds to the count (or relative frequency).

HISTOGRAM: EXAMPLE

- Children's height:





Histogram: Remark 1

- The histogram is an easy-to-use statistical tool that quickly provides an overview of the behavior of an industrial process and the overall shape of the distribution. It illustrates the spread of the data, gives information about dispersion and extreme values, and may help detect potential outliers.

Histogram: Remark 2

- For a histogram of counts, the height of the rectangle corresponding to class j is given by:

$$h_j = \frac{n_j}{a_j},$$

where n_j is the count of class j and a_j its width.

  - h_j is called the **count density**.
  - The total area of the histogram equals the total sample size n, since the area of each rectangle equals the count of class j: $a_j \times h_j = n_j$.

HISTOGRAM: REMARK 3

- For a histogram of relative frequencies, we have:

$$d_j = \frac{f_j}{a_j},$$

  where f_j is the relative frequency of class j.

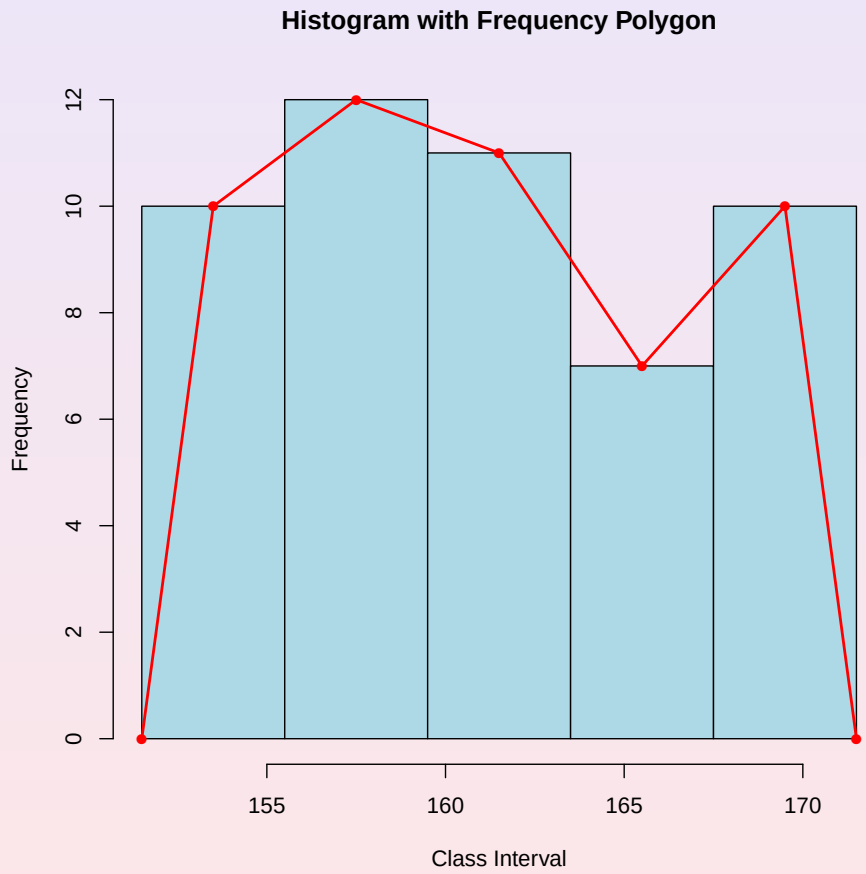
- d_j is called the **frequency density**.
- The total area of the histogram equals 1, since the area of each rectangle equals the frequency of class j: $a_j \times d_j = f_j$.

Graphical Representation: Frequency Polygon

- A frequency polygon represents the distribution of absolute or relative frequencies as a curve.
- It is constructed by joining, with straight lines, the midpoints of the upper sides of each histogram rectangle.
- To close the polygon, an additional class with zero frequency is added at each end.

Frequency Polygon: Example

- Children's height:



The Cumulative Curve

- **Increasing cumulative curve:** obtained by joining points whose abscissas are the upper boundaries of the classes and whose ordinates are the corresponding increasing cumulative frequencies (for the first point, the value is set to 0). It shows the number of observations less than or equal to a given value in the dataset.
- **Decreasing cumulative curve:** constructed in the same way, but using the lower boundaries of the classes as abscissas and the decreasing cumulative frequencies as ordinates (for the last point, the value is 0). It shows the number of observations greater than a given value in the dataset.

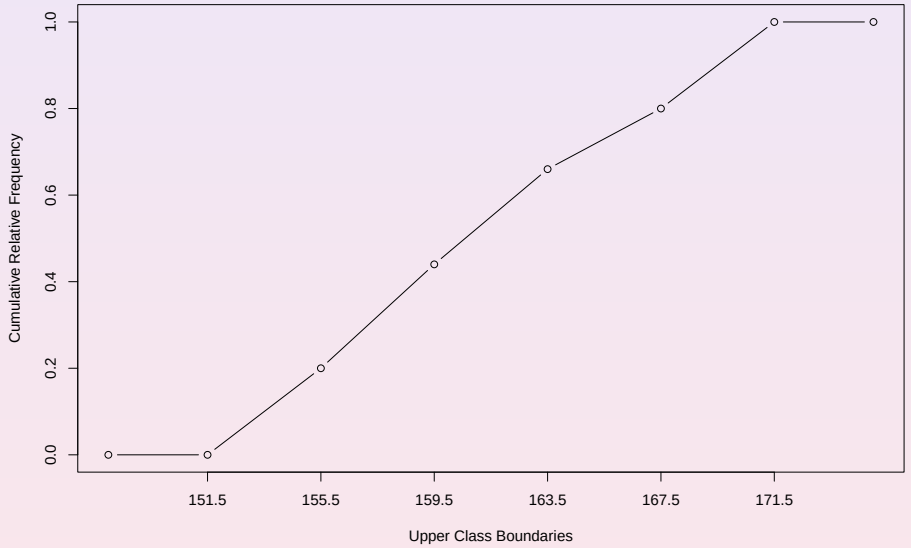
The Distribution Function: Definition

- The distribution function $F(x)$ (the graph of the cumulative curve) is a function from $\mathbb{R}$ to $[0, 1]$, defined as:

$$F(x) = \begin{cases} 0 & \text{if } x < e_1 \\ F_{i-1} + \dfrac{f_i}{e_i - e_{i-1}}(x - e_{i-1}) & \text{if } x \in [e_{i-1}, e_i) \\ 1 & \text{if } x \geq e_r \end{cases}$$

CUMULATIVE CURVE: EXAMPLE

- Children's height:



Statistical Parameters

- **Modal class:** the class with the highest frequency; the mode is taken as the midpoint of this class.
- **Median class:** the class that splits the sample into two groups of equal size; the median is the midpoint of this class.
- **Quantiles:** the quantiles of a continuous variable can be directly determined from the cumulative curve. In practice, one first identifies the class containing the desired quantile, then determines its value within that class by linear interpolation.

Statistical Parameters: Mean and Standard Deviation

- The **mean**, **variance**, and **standard deviation** of a continuous variable are computed in the same way as in the discrete case. In the formulas, the class midpoints c_i are used in place of the raw observations.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} c_i = \frac{1}{n}\sum_{i=1}^{r} n_i c_i \,,$$

$$s^2 = Var(X) = \frac{1}{n}\sum_{i=1}^{n}(c_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} c_i^2 - (\overline{X})^2$$

Standard deviation: $s = \sqrt{s^2}$.

NOMINAL QUALITATIVE VARIABLE

- A nominal qualitative variable has distinct values that cannot be ordered. We denote by r the number of distinct values or categories.
  - The distinct values are denoted $x_1, \ldots, x_i, \ldots, x_r$.
  - The **count** (effectif) of a category or a distinct value is the number of times this category appears. It is denoted by n_i for the category x_i.
  - The **frequency** of a category is the count divided by the total number of observational units:

$$f_i = \frac{n_i}{n}, \quad i = 1, \ldots, r.$$

Nominal Qualitative Variable: Example

- With the following example: "Marital status", we obtain the statistical table

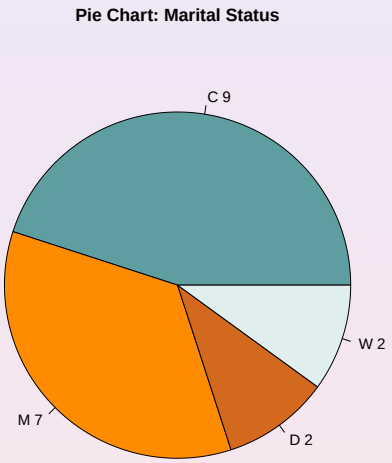
x_i	n_i	f_i
C	9	0.45
M	7	0.35
D	2	0.10
W	2	0.10
	$n = 20$	1

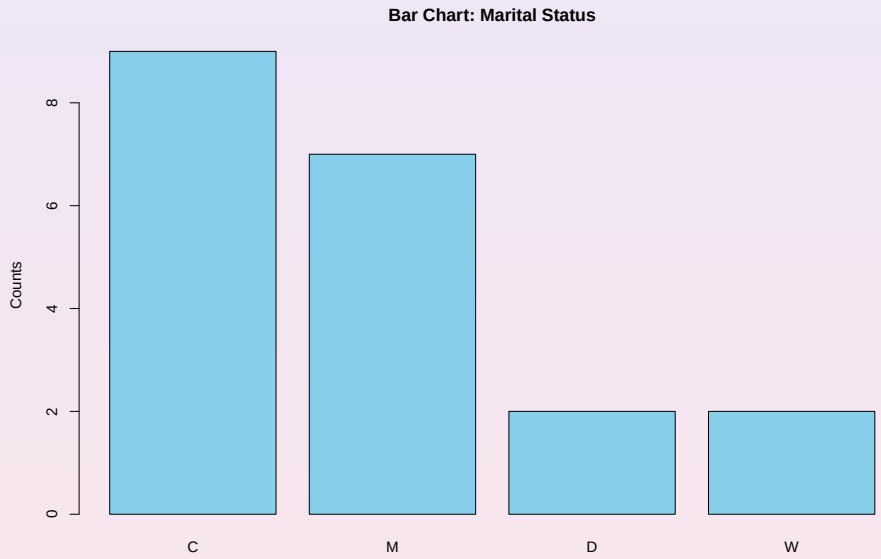
Graphical Representation

- The statistical table of a nominal qualitative variable can be represented in two types of charts: counts are represented by a bar chart, and frequencies by a pie chart.

PIE CHART,BAR CHART

- Example: "Marital status":





Ordinal Qualitative Variable

- The distinct values of an ordinal variable can be ordered, which we write as:

$$x_1 < x_2 < \ldots < x_i < \ldots < x_r \, .$$

- The notation $x_1 < x_2$ means that x_1 precedes x_2.

Statistical Table

- For an ordinal variable, we can compute the cumulative frequencies:

$$N_j = \sum_{i=1}^{j} n_i \,, \quad j = 1, \ldots, r \,.$$

Example

- We survey 50 individuals about their highest educational qualification (variable Y). The coding is given in the following table:

Highest qualification	x_i
No diploma	Sd
Primary	P
Secondary	Se
Non-university tertiary	Su
University	U

EXAMPLE

- The collected data are:

Sd	Sd	Sd	Sd	P	P	P	P	P	P	P	P	P	P
Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Su	Su
Su	Su	Su	Su	U	U	U	U	U	U	U	U	U	U

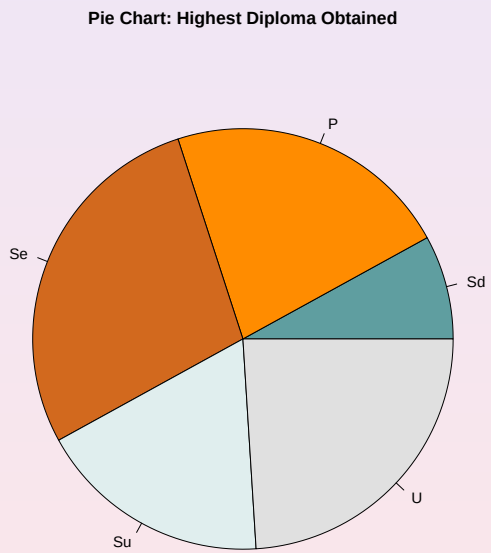
Example

- Finally, we obtain the complete statistical table:

x_i	n_i	N_i	f_i	F_i
Sd	4	4	0.08	0.08
P	11	15	0.22	0.30
Se	14	29	0.28	0.58
Su	9	38	0.18	0.76
U	12	50	0.24	1.00
Total	50		1.00	

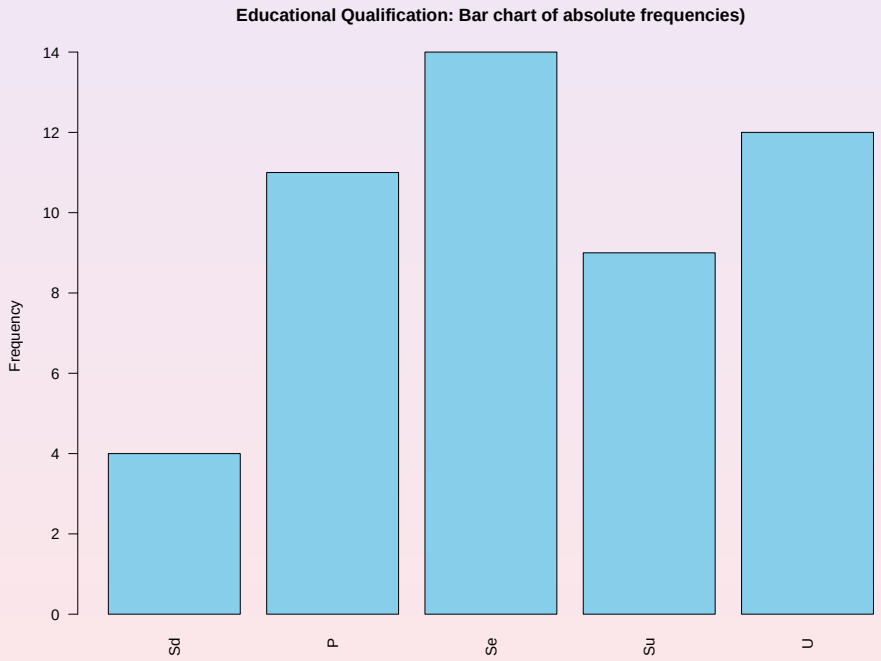
Graphical Representation: Pie Chart

- The frequencies of an ordinal qualitative variable can be represented using a pie chart. Example: "Highest diploma obtained".



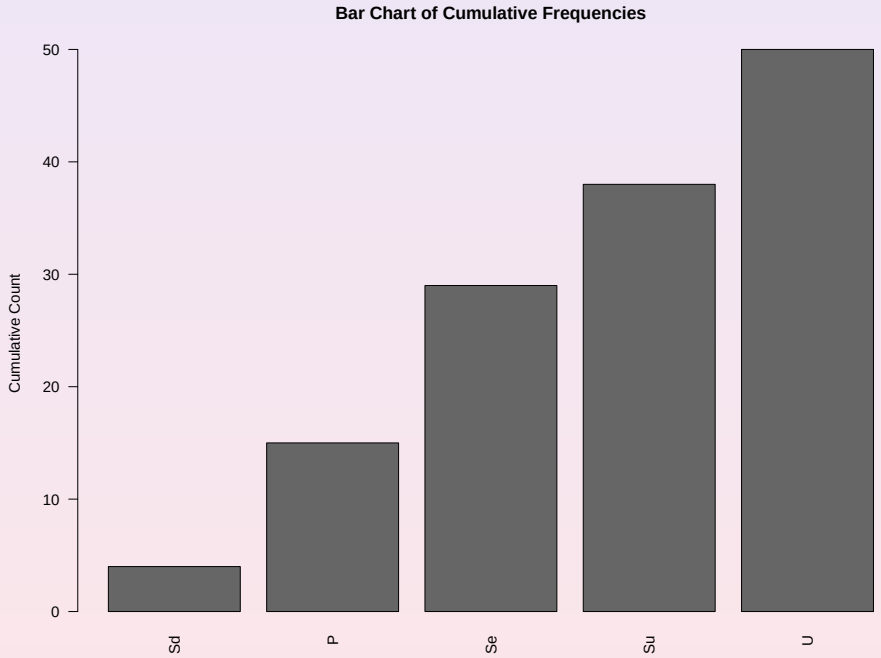
Bar Chart of Counts

- The counts (absolute frequencies) of an ordinal qualitative variable can be represented with a bar chart. Example: "Highest diploma obtained".



Bar Chart of Cumulative Counts

- The cumulative counts of an ordinal qualitative variable can be represented with a bar chart. Example: "Highest diploma obtained".



EXERCISE

- In order to establish a possible relationship between age and leisure activities, a psychosociologist surveyed a population of 20 people and obtained the following information:

EXERCISE

Subject	1	2	3	4	5	6	7	8	9	10
Age x	12	14	40	35	26	30	30	50	75	50
y	S	S	C	C	S	T	T	R	R	R

Subject	11	12	13	14	15	16	17	18	19	20
Age x	30	45	25	55	28	25	50	40	25	35
Leisure	T	C	C	C	S	R	R	C	T	T

Notations: S: Sport, C: Cinema, T: Theatre, R: Reading.

EXERCISE

1. What does the first row of this table represent? How many subjects were surveyed?
2. How many variables are there? What type is each variable?
3. For each variable, construct the complete statistical table.
4. Represent the frequencies of the variables using bar charts. What is the mode?

EXERCISE

5. Plot the cumulative distribution function of x. Can we plot that of y?

   Deduce the value of the median.

6. What percentage of subjects are:

   (A) younger than 30?

   (B) not preferring reading?

7. Compute the empirical mean, variance, and standard deviation of x.

   Briefly comment on these results. ($\sum x_i = 720$; $\sum x_i^2 = 30200$).

Exercise: Solution 1

1. The first row represents the subjects (individuals) of the sample.

   The sample size: $n = 20$ individuals.

2. We have 2 variables: Age X and Leisure Y.

   - X is a discrete quantitative variable (since the values are integers and the number of modalities equals 12 ≤ 15).
   - Y is a qualitative nominal variable (the modalities are not ordered).

3.1 Frequency distribution table: distribution of age (X)

x_j	n_j	*relative* f_j	*cumulative* N_j	*cumulative* F_j
12	1	0.05	1	0.05
14	1	0.05	2	0.10
25	3	0.15	5	0.25
26	1	0.05	6	0.30
28	1	0.05	7	0.35
30	3	0.15	10	0.50
35	2	0.10	12	0.60
40	2	0.10	14	0.70
45	1	0.05	15	0.75
50	3	0.15	18	0.90
55	1	0.05	19	0.95
75	1	0.05	20	1.00

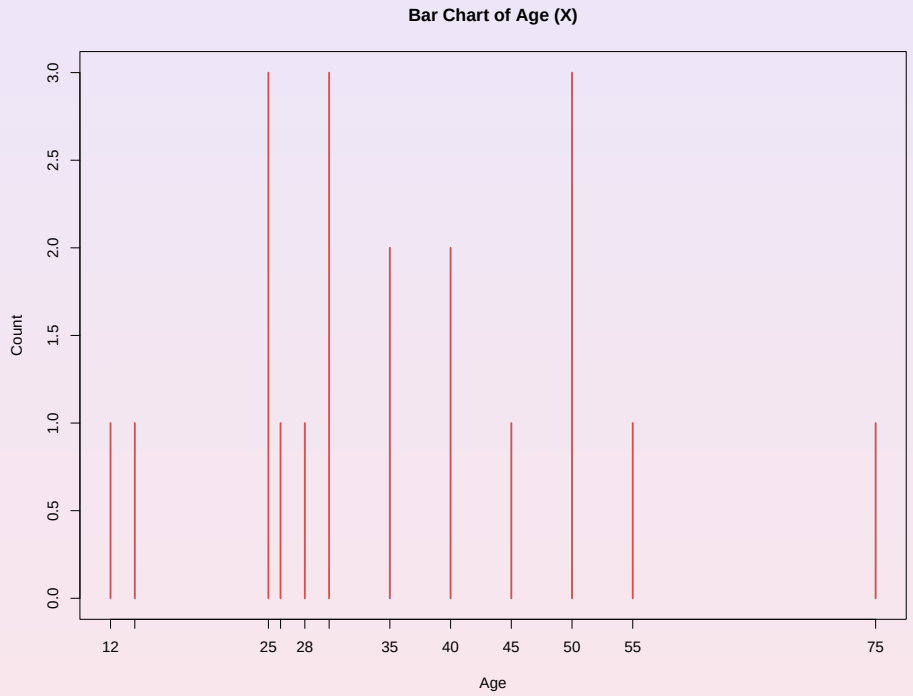
EXERCISE: SOLUTION 3

3.2 Frequency distribution table: distribution of 20 individuals according to their leisure activity (Y).

y_i	n_i	f_i
C	6	0.30
R	5	0.25
S	4	0.20
T	5	0.25
	$n = 20$	1

EXERCISE: SOLUTION 4

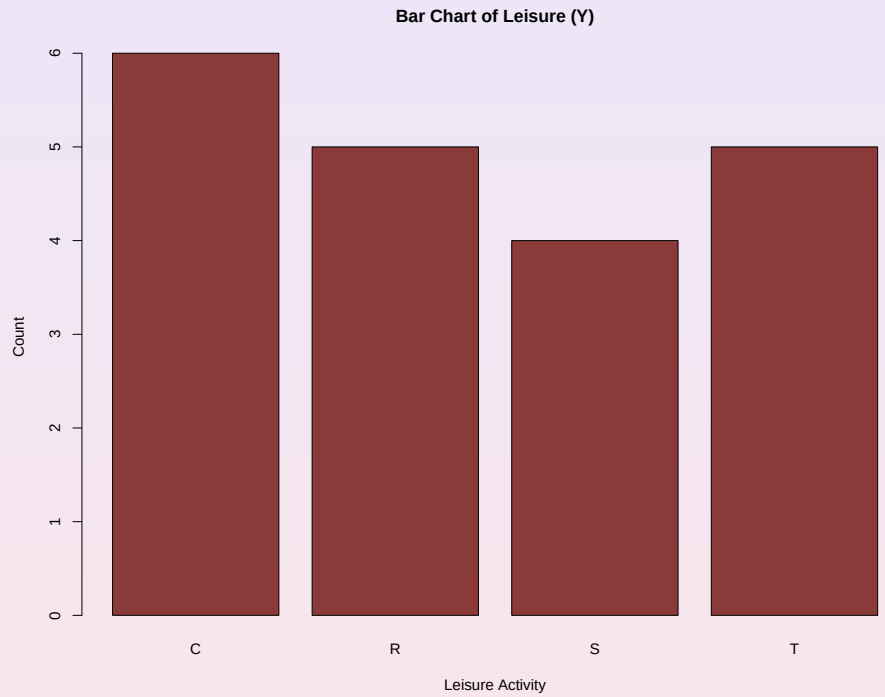
4.1 Bar chart of counts: Age X.



The mode: $Mo = 25, 30, 50$.

EXERCISE: SOLUTION 5

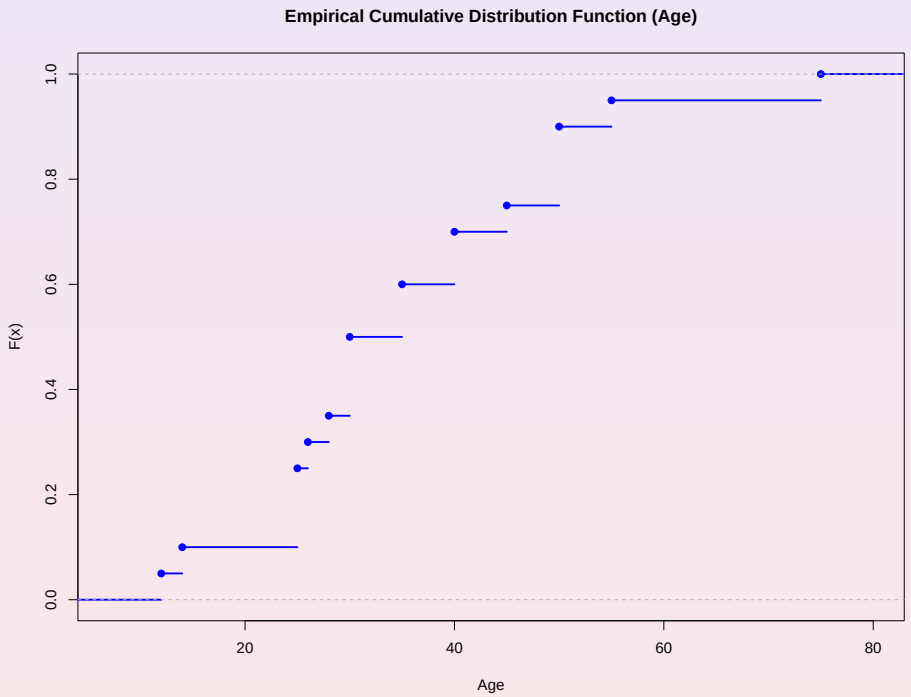
4.1 Bar chart of counts: Leisure Y.



The mode: $Mo =$ "cinema".

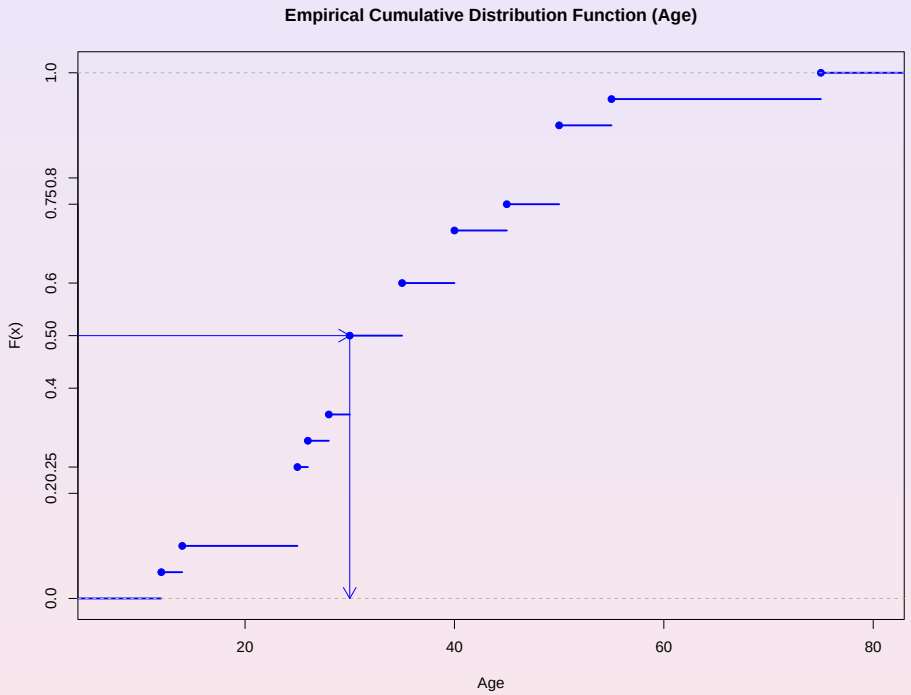
EXERCICE: SOLUTION 5

5.1 Diagramme cumulatif ou le graphe de la fonction de répartition: Âge X.



EXERCISE: SOLUTION 5

5.2 The median is equal to 32.5:



5.3 It is not possible to draw a cumulative diagram for a qualitative variable: Leisure Y.

EXERCISE: SOLUTION 5

7.1 The mean of X (age):

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{r} n_i x_i \; = 720/20 = 36.$$

7.2 The variance:

$$s^2 = Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{r} n_i\, x_i^2 - (\overline{X})^2 \; = 214$$

Standard deviation $s = 11.57$.

EXERCISE 2

- The blood glucose levels measured in 31 subjects are given below (in g/l):

0.85	0.87	0.90	0.90	0.93	0.94	0.95	0.97
0.97	0.98	0.99	1.00	1.01	1.03	1.03	1.03
1.04	1.06	1.07	1.08	1.08	1.10	1.10	1.11
1.13	1.14	1.14	1.15	1.17	1.19	1.20	

Exercise 2 (continued)

1. Define the studied variable and specify its type.
2. Group the data into classes of width 0.06 g/l (starting from the smallest value).
3. Determine the frequencies and cumulative frequencies.
4. Draw the histogram of frequencies. Deduce the mode. On the same graph, draw the frequency polygon.
5. Draw the cumulative frequency curve. Deduce the median graphically, then verify it by calculation.
6. Compute the mean, variance, and standard deviation of x.
7. Briefly comment on these results.

EXERCISE 3

A red blood cell count was performed using the 500 squares of a hemocytometer. The following results were obtained (unordered), where for each $i = 0, 1, \ldots, 10$, n_i is the number of squares containing i red blood cells:

i	10	1	8	3	6	5	4	7	2	9	0
n_i	1	42	8	111	46	66	100	21	91	2	12

$$\sum n_i x_i = 1802, \quad \sum n_i x_i^2 = 8114$$

1. In statistics, how many types of variables are there?

   $(A)\ 0 \quad (B)\ 1 \quad (C)\ 3 \quad (D)\ 4 \quad (E)\ > 4$

Exercise 3 (continued)

2. The statistical population is: (A) red blood cells (B) hemocytometer squares (C) number of hemocytometer squares (D) number of red blood cells (E) arbitrary individuals.
3. The statistical variable is: (A) red blood cells (B) hemocytometer squares (C) number of hemocytometer squares (D) number of red blood cells (E) number of red blood cells per square.
4. This dataset is: (A) unimodal with mode = 3 (B) unimodal with mode = 10 (C) unimodal with mode = 111 (D) bimodal (E) trimodal.

Exercise 3 (continued)

5. The median of this dataset is: (A) 3 (B) 4 (C) 111 (D) 250 (E) does not exist.
6. The interquartile range is: (A) 2 (B) 3 (C) 4 (D) 5 (E) 250
7. The mean of this dataset is: (A) 180.2 (B) 163.82 (C) 16.23 (D) 5 (E) 3.6