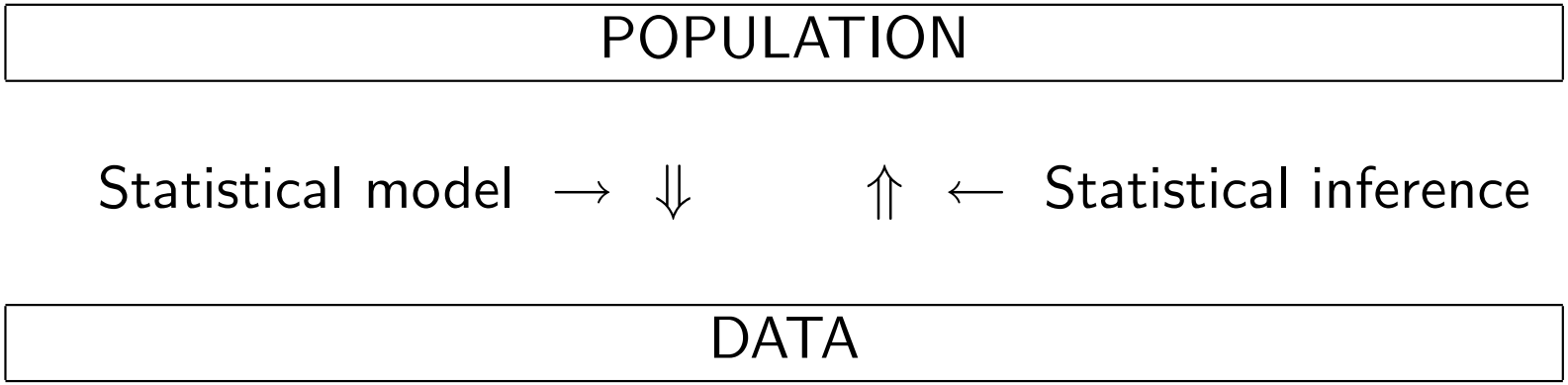
Descriptive Statistics

1. Introduction to Statistics

- Statistics is a mathematical science pertaining to the collection, analysis, interpretation, and presentation of data. (Wikipedia!!).
- Data description (17th century: first censuses in national states).
- Probability theory (18th century: games of chance).
- During the 19th century Statistics is applied to experimental sciences (Biology, Physics, Chemistry).
- During the 20th century Statistics benefits from the development of computers.
- During the 21th century Statistics benefits from the development of both big data analysis and artificial intelligence.

- **Statistical method:**



- **Descriptive statistics** studies how to analyze and present data.
- **Probability theory** explains how data are generated from a population. This is achieved by means of a statistical (or probability) model.
- **Statistical inference** allows us to say something about the population from the available data. This is achieved by "inverting" the statistical model.
- This chapter is about descriptive statistics.

2. Types of variables

- A **population** is a set of people or objects. The elements of a population are called **individuals.**
- A **sample** is a subset of a population.
- Population and sample are relative concepts.
- A **variable** is a characteristic of a population which can take different values (examples: weight, age, color of eyes, income, etc.)

- Two types of variables:
  1. **Qualitative (or categorical) variables**, which are the ones that cannot be measured numerically.
  2. **Quantitative variables**, which are the ones that can be measured numerically, i.e., through numerical values.

- There are two types of quantitative variables:
  1. **Discrete** or **countable**, which are the ones that can take values from a countable set of numbers.
  2. **Continuous**, which are the ones that can take values from an uncountable set of numbers (for instance, the set of real numbers).

- There are two types of discrete (or countable) variables:
  1. **Finite**, which are the ones that take values from a finite set of numbers.
  2. **Infinite**, which are the ones that take values from a countable infinite set of numbers (for instance, the set of rational numbers).

- A variable is represented by a capital letter: X, Y, Z, ...
- The different K values taken by a variable are represented by small letters: $x_1, x_2, ..., x_K$.
- If N is the number of individuals or objects of the population for which we study a characteristic, then $K \leq N$.

3. Univariate frequency distributions: absolute, relative, and cumulative frequencies

- Let us assume that we only care about one single variable X of the population.
- Let us assume that the number of different values taken by this variable is small ($K \leq 10$).
- The absolute frequency $n(x)$ of the value x is the number of times that the value x appears in the data.
- Note that $\sum_x n(x) = N$.
- The distribution of absolute frequencies is the function $n(\cdot)$ defined for all the possible values of the variable X.

- The relative frequency $f(x)$ of the value x is the fraction (or percentage) of times that the value x appears in the data,

$$f(x) = \frac{n(x)}{N}.$$

- Note that

$$\sum_x f(x) = \sum_x \frac{n(x)}{N} = \frac{\sum_x n(x)}{N} = 1.$$

- The cumulative absolute frequency $N(x)$ is the number of times that the variable X takes values smaller or equal than x,

$$N(x) = \sum_{y \leq x} n(y).$$

- The cumulative relative frequency $F(x)$ is the fraction of times that the variable X takes values smaller or equal than x,

$$F(x) = \frac{N(x)}{N} = \frac{\sum_{y \leq x} n(y)}{N} = \sum_{y \leq x} \frac{n(y)}{N} = \sum_{y \leq x} f(y).$$

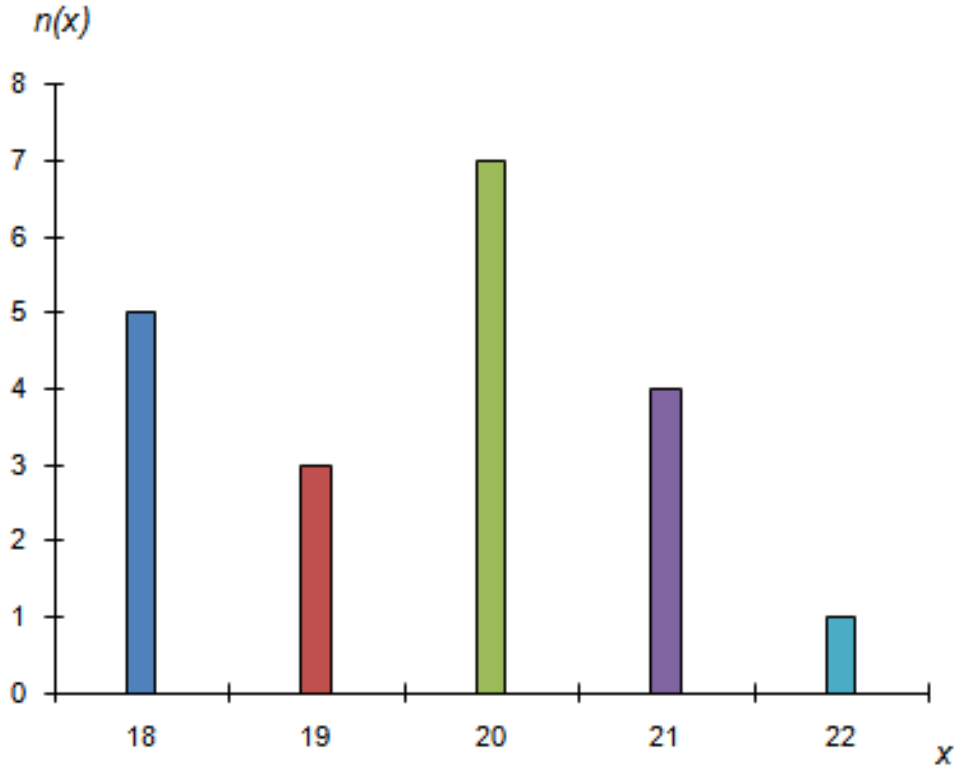
- Note that cumulative frequencies are only well-defined for quantitative variables.

The following table summarizes the distribution of absolute, relative, cumulative absolute and cumulative relative frequencies of the variable "age" for a population (or sample) of 20 students:

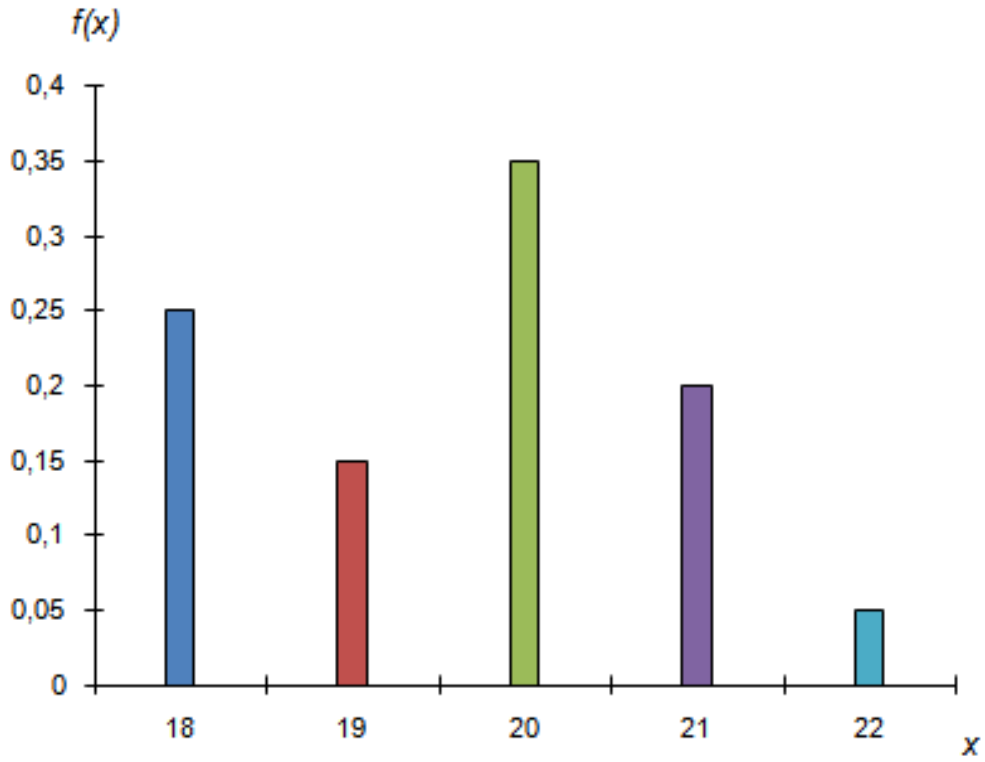
Age x	$n(x)$	$N(x)$	$f(x)$	$F(x)$
18	5	5	0.25	0.25
19	3	8	0.15	0.40
20	7	15	0.35	0.75
21	4	19	0.20	0.95
22	1	20	0.05	1
	$\sum_x n(x) = N = 20$		$\sum_x f(x) = 1$	

4. Graphic representation of frequency distributions: bar diagrams and histograms

Bar diagram of **absolute frequencies** of age**:**



Bar diagram of **relative frequencies** of age:



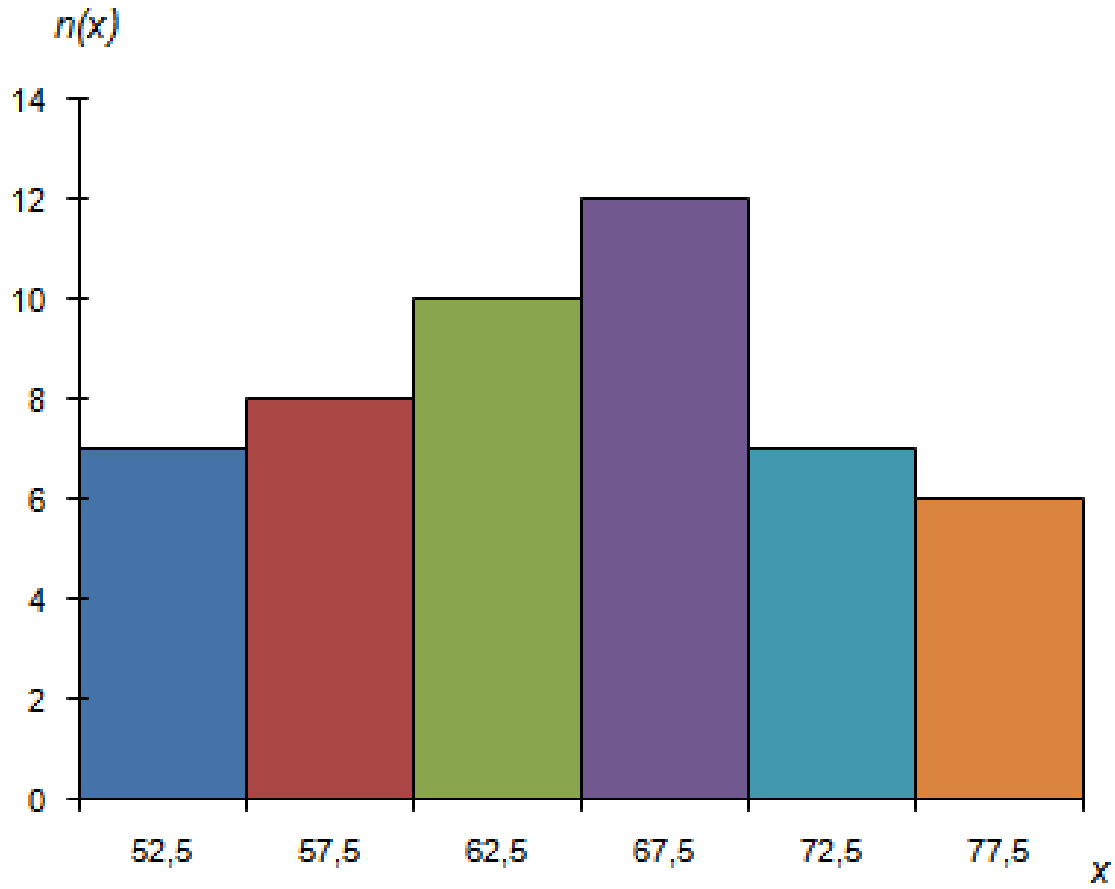
- If the number of values taken by a variable is large ($K > 10$), then we partition the set of values into classes, intervals, or bins. That is, we work with grouped data.
- Usually, we construct from 5 to 10 classes.

Example. Population (or sample) size: 50 students. Variable: weight.

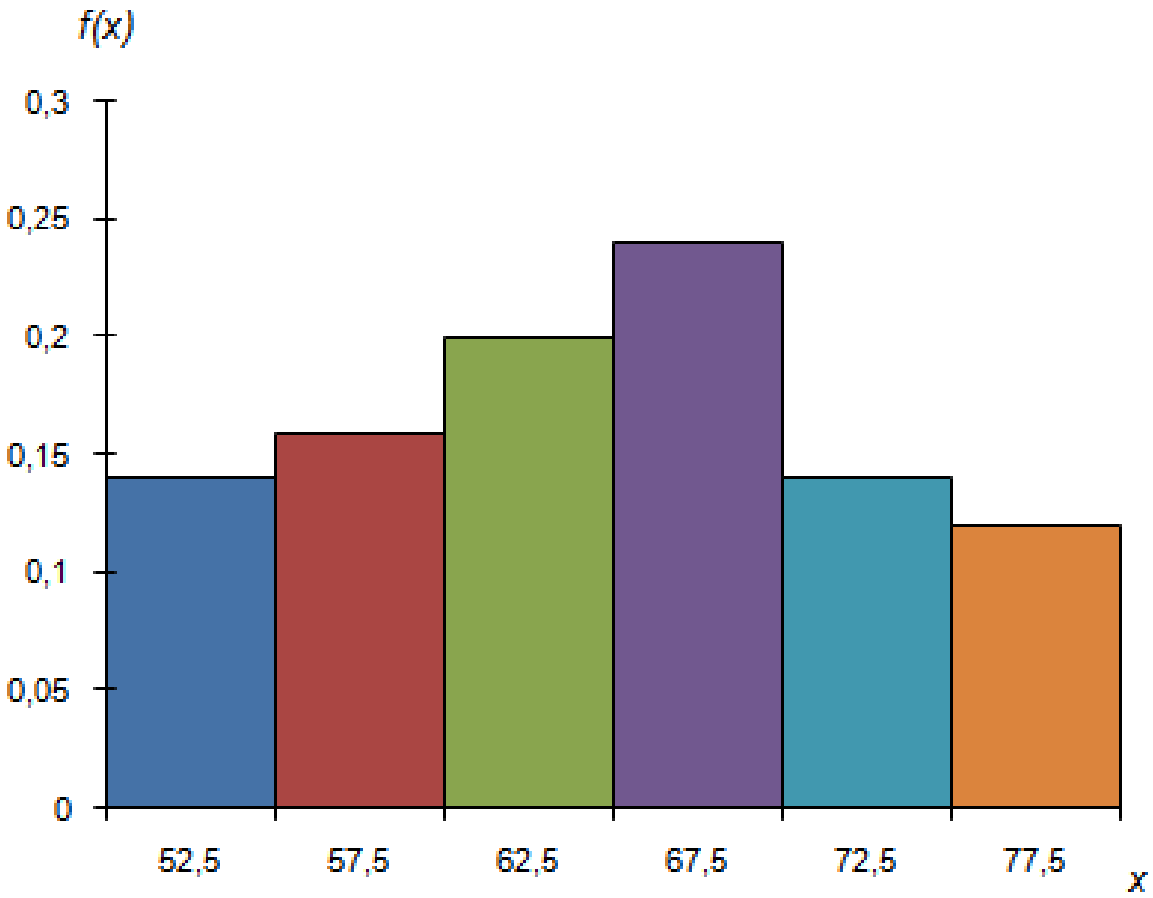
Interval	Midpoint x	Absolute freq. $n(x)$	Cumulat. absolute freq. $N(x)$	Relative freq. $f(x)$	Cumulat. relative freq. $F(x)$
$(50, 55]$	52.5	7	7	0.14	0.14
$(55, 60]$	55.5	8	15	0.16	0.30
$(60, 65]$	62.5	10	25	0.20	0.50
$(65, 70]$	67.5	12	37	0.24	0.74
$(70, 75]$	72.5	7	44	0.14	0.88
$(75, 80]$	77.5	6	50	0.12	1
		$N = 50$		$\sum_x f(x) = 1$	

- Note that the midpoint is the value that represents the interval. Therefore, we are committing an approximation error since we treat all the values of an interval as if they were equal to the value of the midpoint.

- When we work with grouped data we use histograms for the corresponding graphical representation.
- Histogram of absolute frequencies:



- Histogram of relative frequencies:



5. Measures of central tendency

- **Mean (or average value or arithmetic mean) of the variable X :**

$$\overline{X} = \frac{\sum_i x_i}{N} = \sum_x \frac{x \cdot n(x)}{N} = \sum_x x \cdot f(x),$$

where x_i is the value of the variable X for individual i.

- Note that in the sum $\sum_i$ we sum over individuals so that this sum has N terms.
- However, in the sum $\sum_x$ we sum over the different values taken by the variable X so that this sum has K terms.

- **Properties of the mean.**

1. $\sum_i \left(x_i - \overline{X}\right) = 0$ or $\sum_x \left(x - \overline{X}\right) n(x) = 0$ or $\sum_x \left(x - \overline{X}\right) f(x) = 0$.

   Proof.

$$\sum_i \left(x_i - \overline{X}\right) = \sum_i x_i - \sum_i \overline{X} = N\overline{X} - N\overline{X} = 0. \qquad Q.E.D.$$

2. $\overline{kX} = k\overline{X}$, where k is a constant (or scalar).

   Proof. Note that the values taken by the variable kX for the N individuals are $kx_1, kx_2, ..., kx_N$. Therefore,

$$\overline{kX} = \frac{\sum_i kx_i}{N} = k \cdot \left(\frac{\sum_i x_i}{N}\right) = k\overline{X}. \qquad Q.E.D.$$

- **Median:** We order all the values of the variable X taken by the N individuals from the smallest to the largest. We re-index the values accordingly so that

$$x_1 \leq x_2 \leq x_3 \leq ... \leq x_{N-1} \leq x_N.$$

Then, the median of X is

$$Median(X) = \begin{cases} x_{\frac{N}{2}+\frac{1}{2}} & \text{if } N \text{ is odd} \\ \dfrac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} & \text{if } N \text{ is even.} \end{cases}$$

- *Examples:* (a) $N = 7$ and the ordered values are $1, 1, 3, 6, 8, 8, 12$. The median is $x_4 = 6$.
- (b) $N = 8$ and the ordered values are $1, 1, 3, 6, 8, 8, 12, 3472$. The median is

$$\frac{x_4 + x_5}{2} = \frac{6 + 8}{2} = 7.$$

- The mean uses all the values of all individuals but it is more sensitive to outliers and errors.
- The median does not use all the values but, in general, is less sensitive to outliers and errors.

- **Mode:** It is the value that appears a larger number of times, i.e., is the value with a larger absolute (or relative) frequency.
- A variable may have more than one mode.
- If we work with grouped data, then we use the midpoint values on the previous formulae. We are thus committing an approximation error.
- Finally, note that the mean and the median (and the next measures) are defined only for quantitative variables.

6. Measures of variability

- **Variance:** It measures the average of the square of the deviations from the mean,

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{S}_X^2 = \mathrm{S}^2 = \frac{\sum_i \left(x_i - \overline{X}\right)^2}{N} = \sum_x \frac{\left(x - \overline{X}\right)^2 \cdot n(x)}{N} \\ &= \sum_x \left(x - \overline{X}\right)^2 \cdot f(x). \end{aligned}$$

- The subindex X can be omitted when we work with a single variable.
- Obviously, the variance is always non-negative.
- Note that, if we suppress the square, the previous formula is meaningless since, for every variable X, we have

$$\frac{\sum_i \left(x_i - \overline{X}\right)}{N} = 0.$$

- **Properties of the variance**

1. $S_X^2 = \overline{X^2} - \overline{X}^2$.

In words: the variance is equal to the mean of the square minus the square of the mean.

Proof:

$$S_X^2 = \frac{\sum_i \left(x_i - \overline{X}\right)^2}{N} = \frac{1}{N}\sum_i \left[x_i^2 - 2x_i\overline{X} + \overline{X}^2\right]$$

$$= \frac{1}{N}\left[\sum_i x_i^2 - 2\overline{X}\sum_i x_i + N\overline{X}^2\right] = \frac{\sum_i x_i^2}{N} - 2\overline{X}\left(\frac{\sum_i x_i}{N}\right) + \overline{X}^2$$

$$= \overline{X^2} - 2\cdot\overline{X}\cdot\overline{X} + \overline{X}^2 = \overline{X^2} - \overline{X}^2 \ . \qquad Q.E.D.$$

- Thus,

$$S_X^2 = \overline{X^2} - \overline{X}^2 = \frac{\sum_i x_i^2}{N} - \overline{X}^2 = \sum_x \frac{x^2 \cdot n(x)}{N} - \overline{X}^2 = \sum_x x^2 \cdot f(x) - \overline{X}^2.$$

2. $S_{kX}^2 = k^2 S_X^2$ or $\text{Var}(kX) = k^2 \text{Var}(X)$, where k is a scalar.

Proof:

$$S_{kX}^2 = \frac{\sum_i \left(kx_i - \overline{kX}\right)^2}{N} = \frac{\sum_i \left(kx_i - k\overline{X}\right)^2}{N} = \frac{k^2 \sum_i \left(x_i - \overline{X}\right)^2}{N} = k^2 S_X^2 \ .$$

Q.E.D.

- **Standard deviation:**

$$S_X = S = \left(S_X^2\right)^{1/2} \equiv +\sqrt{\text{Var}(X)}.$$

- Note that

$$S_{kX} = kS_X \quad \text{if} \quad k \geq 0.$$

- **Average absolute deviation:**

$$\frac{\sum_i \left|x_i - \overline{X}\right|}{N} = \sum_x \frac{\left|x - \overline{X}\right| \cdot n(x)}{N} = \sum_x \left|x - \overline{X}\right| \cdot f(x).$$

- The average absolute deviation is not differentiable with respect to the value x_i when $x_i = \overline{X}$.

- **Coefficient of variation (or variation coefficient):**

$$\mathrm{CV}_X = \mathrm{CV} = \frac{S_X}{\left|\overline{X}\right|}, \quad \text{when } \left|\overline{X}\right| \neq 0.$$

- Note that

$$\mathrm{CV}_{kX} = \frac{S_{kX}}{\left|\overline{kX}\right|} = \frac{kS_X}{\left|k\overline{X}\right|} = \frac{kS_X}{k\left|\overline{X}\right|} = \frac{S_X}{\left|\overline{X}\right|} = \mathrm{CV}_X \quad \text{if } k > 0,$$

  so that the coefficient of variation is immune to the units of measurement.

- The **range** of the variable X is the difference between the largest and the smallest value taken by the variable.

- The **p% percentile:** We order all the values of the variable X taken by the N individuals from the smallest to the largest. We re-index the values accordingly so that

$$x_1 \leq x_2 \leq x_3 \leq ... \leq x_{N-1} \leq x_N.$$

We compute the integer j (between 1 and N) nearest to the value

$$\frac{p \cdot N}{100} + \frac{1}{2}. \qquad (\star)$$

Then, the $p\%$ percentile is

$$x_{p\%} = x_j,$$

with $p \in [0, 100]$. That is, $x_{p\%}$ is the value of the variable for which a $p\%$ of the individuals exhibits a lower value of the variable.

- If there are two integers j and $j+1$ that are at the same distance from $(\star)$, then the $p\%$ percentile is

$$x_{p\%} = \frac{x_j + x_{j+1}}{2}.$$

- Note that the range equals to $x_{100\%} - x_{0\%}$.
- The **interquartile range** is equal to $x_{75\%} - x_{25\%}$.
- The median is equal to $x_{50\%}$.

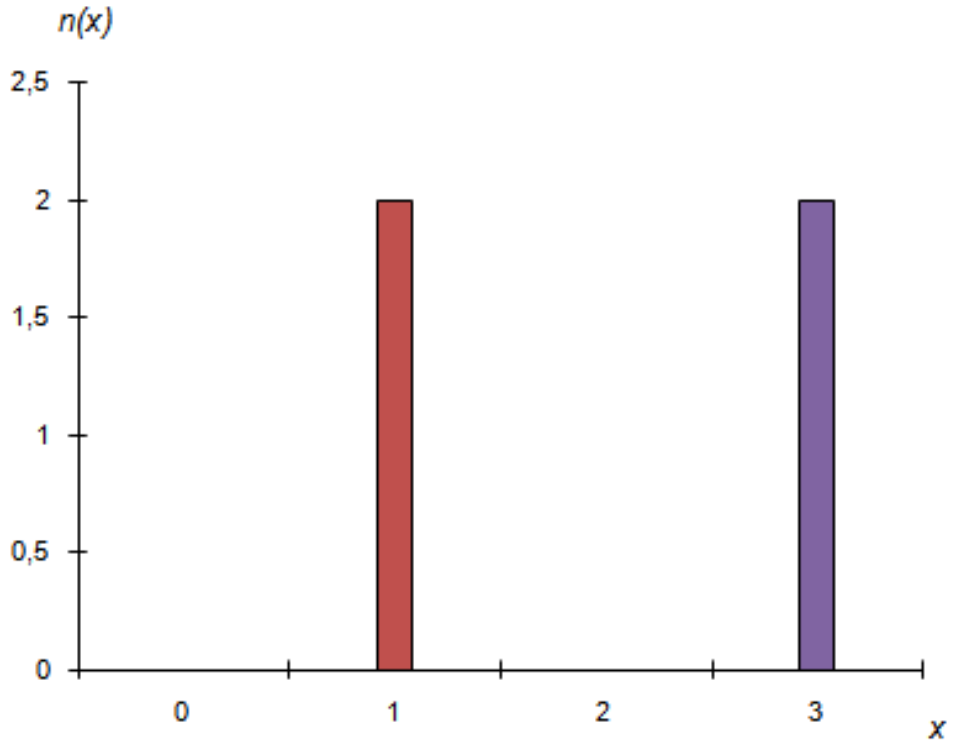
7. Other measures summarizing the shape of a distribution

- **Coefficient of Asymmetry:**

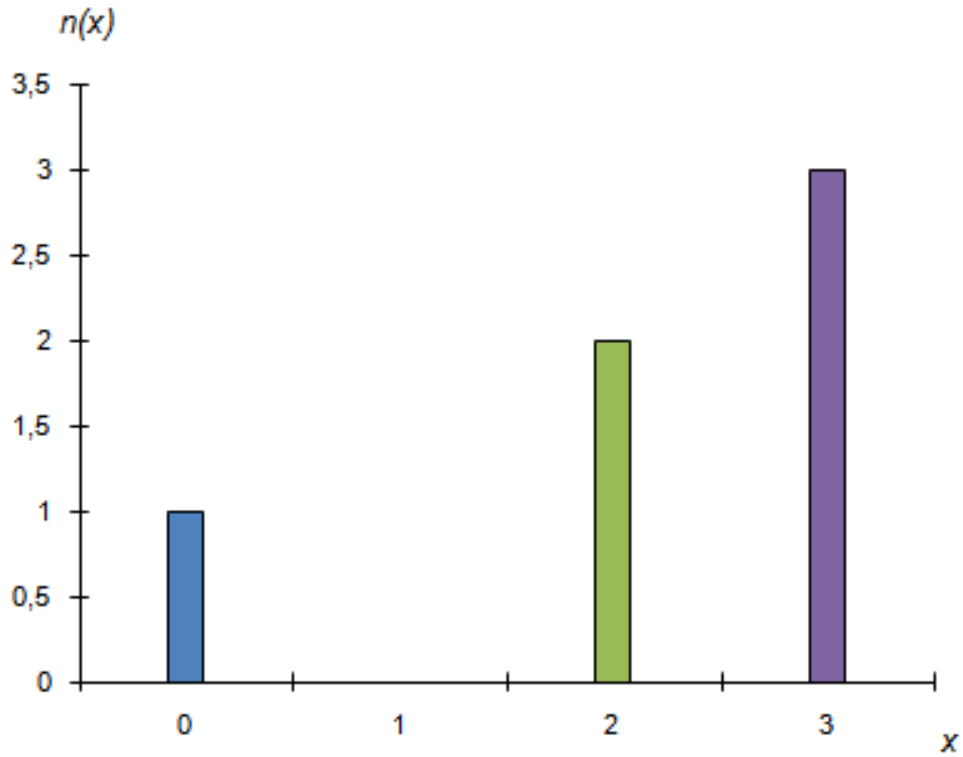
$$\text{CA}_X = \frac{\sum_i \left(x_i - \overline{X}\right)^3}{N \cdot S_X^3} = \frac{\sum_x \left(x - \overline{X}\right)^3 \cdot n(x)}{N \cdot S_X^3} = \frac{\sum_x \left(x - \overline{X}\right)^3 \cdot f(x)}{S_X^3},$$

for $S_X^3 \neq 0$.

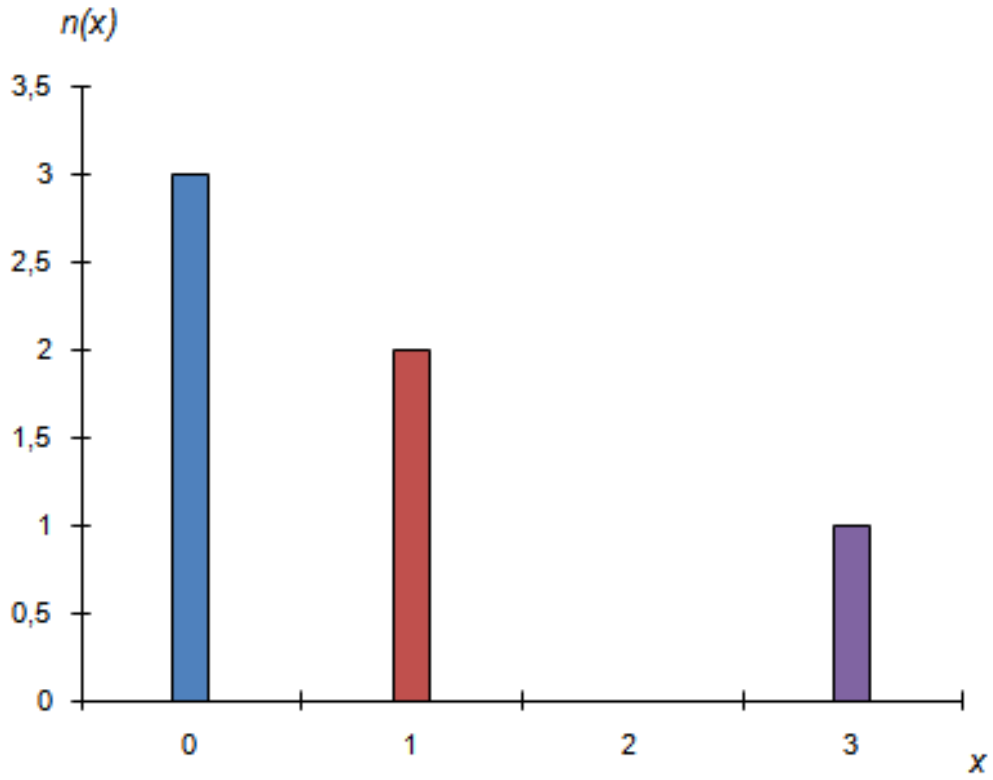
- *Examples:* $\mathrm{CA}_X = 0$. The distribution of x is symmetric.



- $\mathrm{CA}_X < 0$. The left tail is the longest.



- $\mathrm{CA}_X > 0$. The right tail is the longest

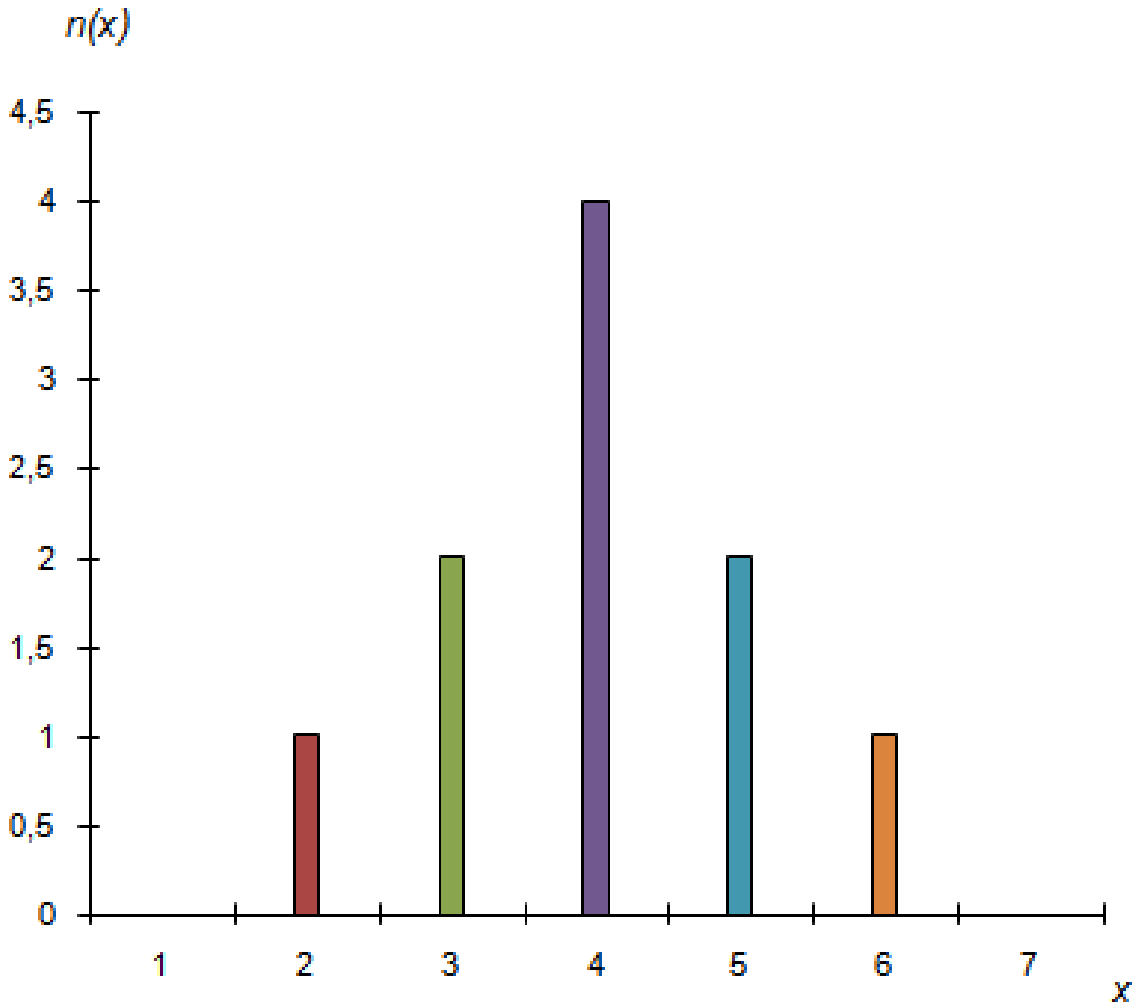


- The **coefficient of Kurtosis** measures the thickness of the tails of the distribution and is given by

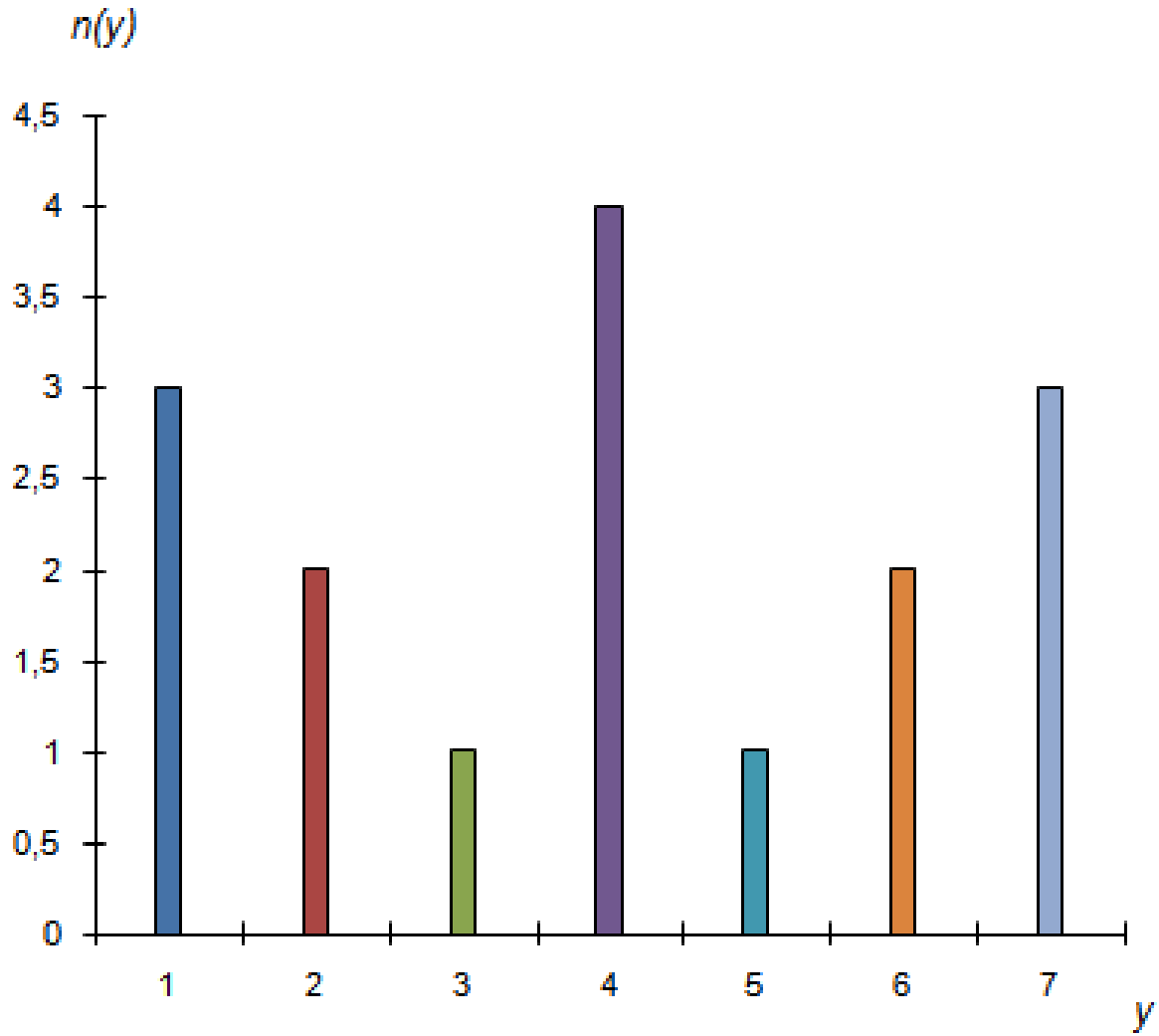
$$\mathrm{CK}_X = \frac{\sum_i \left(x_i - \overline{X}\right)^4}{N \cdot \mathrm{S}_X^4} = \frac{\sum_x \left(x - \overline{X}\right)^4 \cdot n(x)}{N \cdot \mathrm{S}_X^4} = \frac{\sum_x \left(x - \overline{X}\right)^4 \cdot f(x)}{\mathrm{S}_X^4},$$

for $\mathrm{S}_X^4 > 0$.

- *Examples:* Distribution of X



- Distribution of Y



- Then, $\mathrm{CK}_X < \mathrm{CK}_Y$ since the distribution of Y has thicker tails than the distribution of X. In other words, the distribution of X has thinner tails than the distribution of Y.

8. Multivariate frequency distributions

- The multivariate frequency distribution (or joint distribution) gives us the distribution of several variables.
- For instance, the joint distribution of absolute frequencies of two variables X and Y gives us the number of times that each pair of values (x, y) corresponding to the pair of variables (X, Y) appears in the data.
- *Example:* Consider a population consisting of 100 father/son pairs and consider two qualitative variables: $X =$ color of the parent's eyes, $Y =$ color of the son's eyes. This table summarizes the joint distribution of absolute frequencies of these two variables:

		Father's color		
		Light	Dark	
Son's color	Light	25	8	33
	Dark	12	55	67
		37	63	N =100

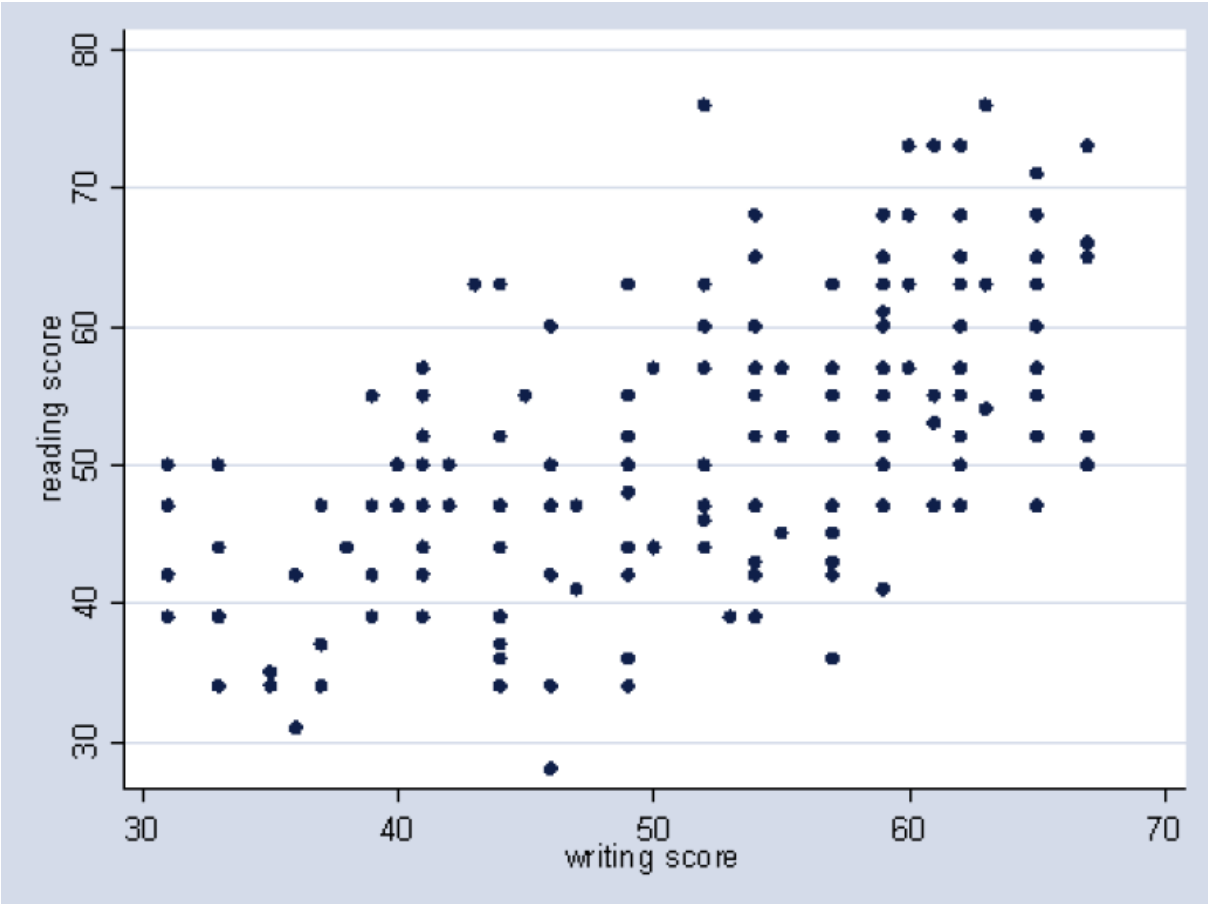
- *Another example:* Consider the quantitative variables X and Y. We have N observations for the values of these two variables.

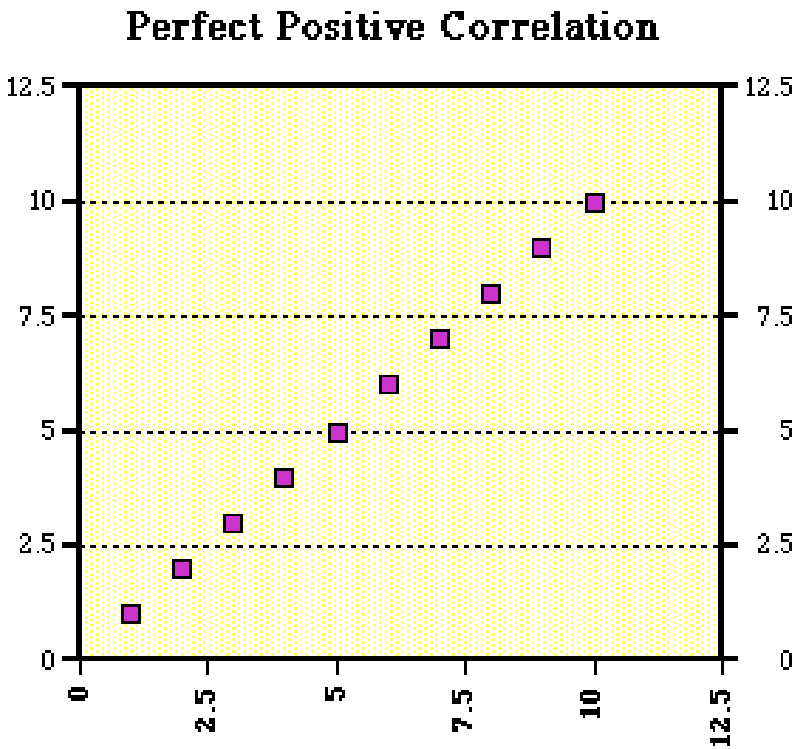
$$X : x_1, x_2, x_3, ..., x_N$$

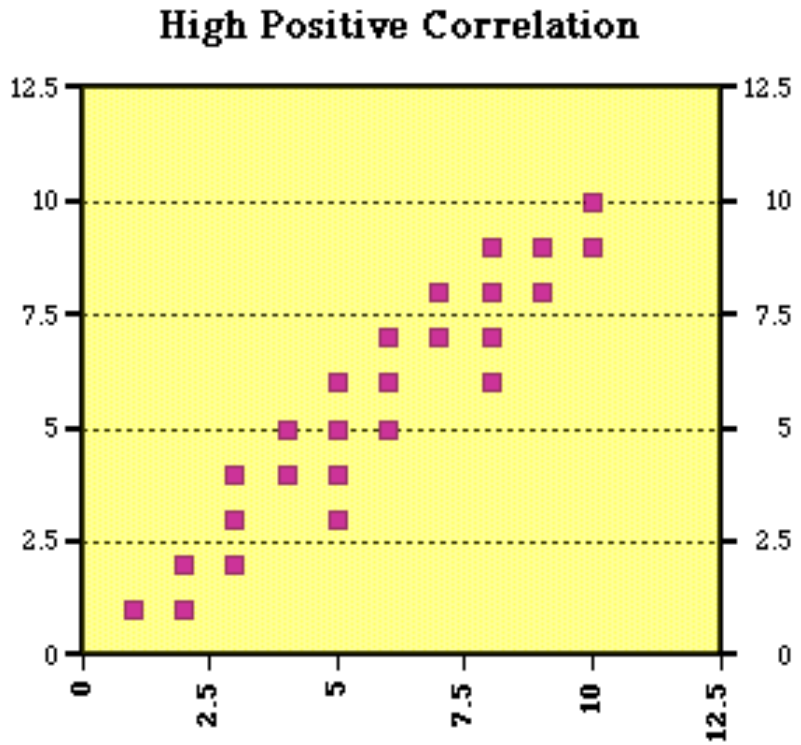
$$Y : y_1, y_2, y_3, ..., y_N$$

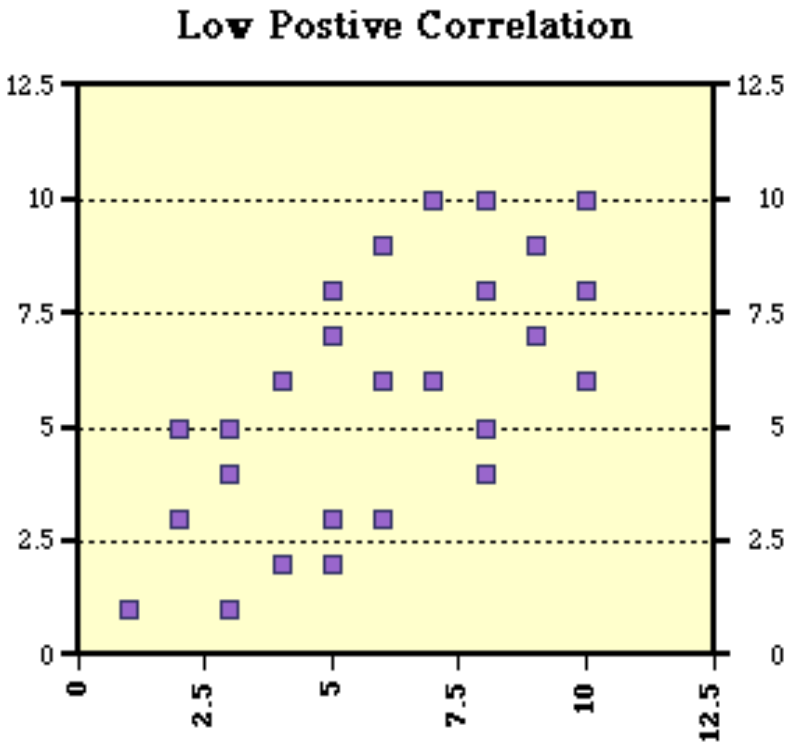
- The pair (x_i, y_i) gives the values of the variables X and Y for individual i.

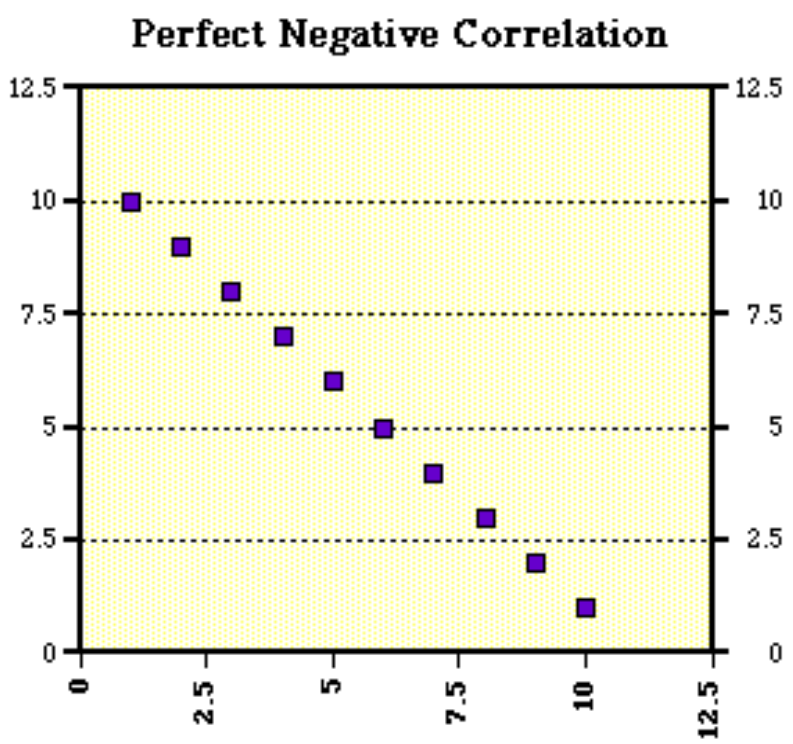
We can summarize these values in a **scatter plot,** which gives us an idea of the type of association (or correlation) between two quantitative variables:

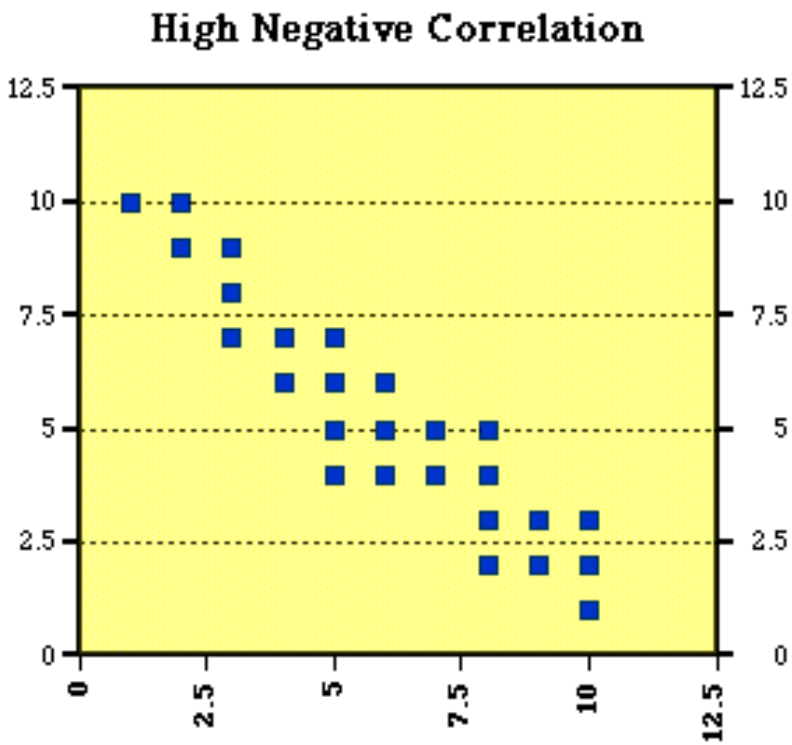


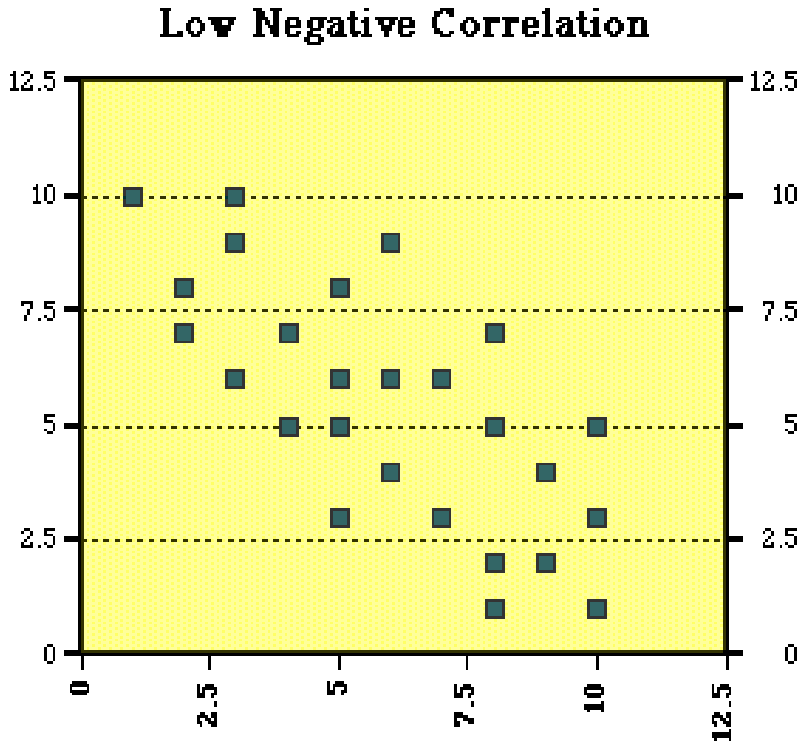

Perfect Positive Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5

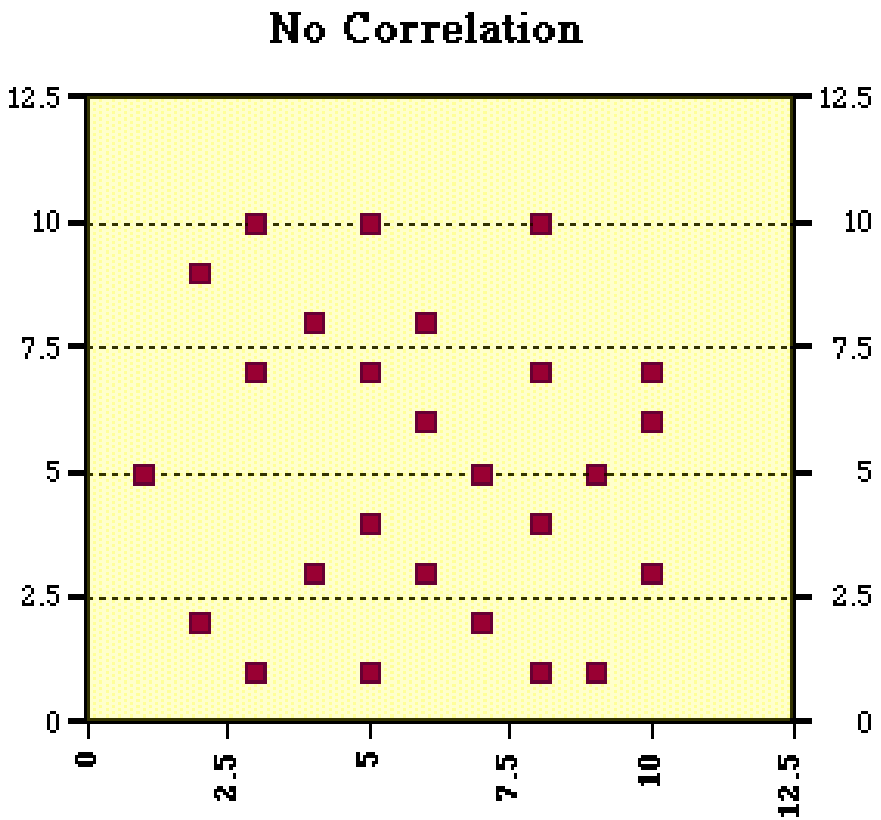

High Positive Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Low Postive Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Perfect Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


High Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Low Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


No Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5

- *Another example:* Consider the variables
  $X =$ average monthly temperature in $^{\circ}C$ of a machine.
  $Y =$ number of breakdowns of the same machine in a month.

- We have 100 observations (months) for the values of these two variables.

$$X : x_1, x_2, x_3, ..., x_{100}$$

$$Y : y_1, y_2, y_3, ..., y_{100}$$

- For the variable X (which is continuous) we work with 3 intervals or classes. The variable Y is discrete and, in the data, only takes on the values 2, 3, 4, 5.

- The following table summarizes the distribution of absolute frequencies:

$Y \setminus X$	$(110-130]$ $120°$	$(130-150]$ $140°$	$(150-170]$ $160°$	
2	20	15	10	45
3	12	7	5	24
4	4	10	2	16
5	0	5	10	15
	36	37	27	100

- The absolute frequency $n_{X,Y}(x,y)$ of the pair (x,y) is the number of times that this pair appears in the data. Example: $n_{X,Y}(140,4) = 10$.
- Recall that the joint distribution of absolute frequencies of the variables X and Y is the function $n_{X,Y}(\cdot,\cdot)$ defined for the Cartesian product of the set of values taken by each of these variables.
- Note that $\sum_x \sum_y n_{X,Y}(x,y) = N$.

- The relative frequency $f_{X,Y}(x,y)$ of the pair (x,y) is the fraction (or percentage) of times that this pair appears in the data:

$$f_{X,Y}(x,y) = \frac{n_{X,Y}(x,y)}{N}.$$

- Example: $f_{X,Y}(140,4) = 0.1$.
- The following table summarizes the distribution of relative frequencies:

$Y \setminus X$	$(110-130]$ 120°	$(130-150]$ 140°	$(150-170]$ 160°	
2	0.20	0.15	0.10	0.45
3	0.12	0.07	0.05	0.24
4	0.04	0.10	0.02	0.16
5	0	0.05	0.10	0.15
	0.36	0.37	0.27	1

- Note that $\sum_x \sum_y f_{X,Y}(x,y) = 1$.

9. Marginal and conditional frequencies

- The **distribution of (absolute / relative) marginal frequencies of the variable** X is the frequency distribution of this variable with independency of the values taken by the other variables.
- The absolute marginal frequency $n_X(x)$ of the value x taken by the variable X is

$$n_X(x) = \sum_y n_{X,Y}(x,y).$$

- The relative marginal frequency $f_X(x)$ of the value x taken by the variable X is

$$f_X(x) = \frac{n_X(x)}{N} = \frac{\sum_y n_{X,Y}(x,y)}{N} = \sum_y \frac{n_{X,Y}(x,y)}{N} = \sum_y f_{X,Y}(x,y).$$

- Example: $n_Y(3) = 24$ and $f_Y(3) = 0.24$.

- Note that

$$\overline{X} = \frac{\sum_i x_i}{N} = \sum_x \frac{x \cdot n_X(x)}{N} = \sum_x \frac{x \cdot \sum_y n_{X,Y}(x,y)}{N}$$

$$= \sum_x \frac{\sum_y x \cdot n_{X,Y}(x,y)}{N} = \frac{\sum_x \sum_y x \cdot n_{X,Y}(x,y)}{N}.$$

or

$$\overline{X} = \sum_x x \cdot f_X(x) = \sum_x x \cdot \sum_y f_{X,Y}(x,y) = \sum_x \sum_y x \cdot f_{X,Y}(x,y),$$

and similarly for the formulae for the variance and other measures.

- The **distribution of conditional frequencies of the variable X given $Y = y$** is the relative frequency distribution of the variable X for all the observations where $Y = y$.
- The conditional frequency $f_{X|Y}(x|y)$ of the value x taken by the variable X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{n_{X,Y}(x,y)}{n_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

- Note that

$$f_{Y|X}(y|x) = \frac{n_{X,Y}(x,y)}{n_X(x)} = \frac{f_{X,Y}(x,y)}{f_X(x)} \neq \frac{f_{X,Y}(x,y)}{f_Y(y)} = f_{X|Y}(x|y).$$

- Example:

$$f_{X|Y}(160|3) = \frac{n_{X,Y}(160,3)}{n_Y(3)} = \frac{5}{24} = \frac{f_{X,Y}(160,3)}{f_Y(3)} = \frac{0.05}{0.24} = 0.2083,$$

whereas

$$f_{Y|X}(3|160) = \frac{n_{X,Y}(160,3)}{n_X(160)} = \frac{5}{27} = \frac{f_{X,Y}(160,3)}{f_X(160)} = \frac{0.05}{0.27} = 0.1852.$$

10. Covariance and correlation coefficient

- Consider the quantitative variables X and Y. We have N observations for the values of these two variables,

$$X : x_1, x_2, x_3, ..., x_N$$

$$Y : y_1, y_2, y_3, ..., y_N$$

- The **covariance** between X and Y measures the strength of the association between these two variables and is given by

$$\begin{aligned} S_{X,Y} = \text{Cov}(X, Y) &= \frac{\sum_i (x_i - \overline{X}) \cdot (y_i - \overline{Y})}{N} \\ &= \frac{\sum_x \sum_y (x - \overline{X}) \cdot (y - \overline{Y}) \cdot n_{X,Y}(x, y)}{N} \\ &= \sum_x \sum_y (x - \overline{X}) \cdot (y - \overline{Y}) \cdot f_{X,Y}(x, y). \end{aligned}$$

- Note that the covariance can be positive, negative or equal to zero.
- **Properties of the covariance**

  1. $S_{X,X} = S_X^2$.

  2. $S_{X,Y} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y}$.

- *In words:* the covariance is equal to the mean of the product minus the product of the means.

Proof:

$$S_{X,Y} = \frac{\sum_i \left(x_i - \overline{X}\right) \cdot \left(y_i - \overline{Y}\right)}{N} = \frac{1}{N} \sum_i \left[x_i \cdot y_i - x_i \cdot \overline{Y} - y_i \cdot \overline{X} + \overline{X} \cdot \overline{Y}\right]$$

$$= \frac{1}{N} \left[\sum_i x_i \cdot y_i - \overline{Y} \sum_i x_i - \overline{X} \sum_i y_i + N \cdot \overline{X} \cdot \overline{Y}\right]$$

$$= \frac{\sum_i x_i \cdot y_i}{N} - \overline{Y} \left(\frac{\sum_i x_i}{N}\right) - \overline{X} \left(\frac{\sum_i y_i}{N}\right) + \overline{X} \cdot \overline{Y}$$

$$= \overline{X \cdot Y} - \overline{Y} \cdot \overline{X} - \overline{X} \cdot \overline{Y} + \overline{X} \cdot \overline{Y} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y} \; . \qquad Q.E.D.$$

- Thus,

$$S_{X,Y} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y} = \frac{\sum_i x_i \cdot y_i}{N} - \overline{X} \cdot \overline{Y}$$

$$= \sum_x \sum_y \frac{x \cdot y \cdot n_{X,Y}(x,y)}{N} - \overline{X} \cdot \overline{Y} = \sum_x \sum_y x \cdot y \cdot f_{X,Y}(x,y) - \overline{X} \cdot \overline{Y}.$$

3. $S_{\alpha X,\beta Y} = \alpha \cdot \beta \cdot S_{X,Y}$, where α and β are scalars.

Proof:

$$S_{\alpha X,\beta Y} = \frac{\sum_i \left(\alpha x_i - \overline{\alpha X}\right)\left(\beta y_i - \overline{\beta Y}\right)}{N} = \frac{\sum_i \left(\alpha x_i - \alpha \overline{X}\right)\left(\beta y_i - \beta \overline{Y}\right)}{N}$$

$$= \frac{\alpha \cdot \beta \cdot \sum_i \left(x_i - \overline{X}\right) \cdot \left(y_i - \overline{Y}\right)}{N} = \alpha \cdot \beta \cdot S_{X,Y} \ . \qquad Q.E.D.$$

4. $S_{X,Y} = S_{Y,X}$.

- The **coefficient of correlation** $r_{X,Y}$ between the variables X and Y is given by

$$r_{X,Y} = \frac{S_{X,Y}}{S_X \cdot S_Y} \quad \text{when } S_X > 0 \text{ and } S_Y > 0.$$

- Note that $r_{\alpha X,\beta Y} = r_{X,Y}$ if $\alpha > 0$ and $\beta > 0$ so that the coefficient of correlation is immune to the units of measurement.

- Another interesting property of the coefficient of correlation is that $|r_{X,Y}| \leq 1$, that is, $-1 \leq r_{X,Y} \leq 1$ (see the proof in the handout).

- In the temperature-breakdowns example we can compute (please do it!) the following:

$$\overline{X} = 138.2, \quad \overline{Y} = 3.01,$$

$$S_X = 15.77 > S_Y = 1.1, \text{ and } CV_X = 0.11 < CV_Y = 0.37,$$

$$S_{X,Y} = 5.62 > 0, \; r_{X,Y} = 0.32 > 0.$$

11. Mean and variance of linear combinations of variables

- Let us consider the variables X_1 and X_2,

$$X_1 : x_{11}, x_{12}, ..., x_{1N}$$

$$X_2 : x_{21}, x_{22}, ..., x_{2N}$$

and consider the scalars α_1 and α_2. Then, the values taken by the variable

$$Z = \alpha_1 X_1 + \alpha_2 X_2$$

are

$$Z : \underbrace{\alpha_1 x_{11} + \alpha_2 x_{21}}_{z_1}, \underbrace{\alpha_1 x_{12} + \alpha_2 x_{22}}_{z_2}, ..., \underbrace{\alpha_1 x_{1N} + \alpha_2 x_{2N}}_{z_N}$$

- The mean of the variable Z is

$$\overline{Z} = \frac{\sum_i z_i}{N} = \frac{\sum_i (\alpha_1 x_{1i} + \alpha_2 x_{2i})}{N} = \frac{\sum_i \alpha_1 x_{1i} + \sum_i \alpha_2 x_{2i}}{N}$$

$$= \frac{\alpha_1 \sum_i x_{1i} + \alpha_2 \sum_i x_{2i}}{N} = \frac{\alpha_1 \sum_i x_{1i}}{N} + \frac{\alpha_2 \sum_i x_{2i}}{N}$$

$$= \alpha_1 \cdot \left(\frac{\sum_i x_{1i}}{N} \right) + \alpha_2 \cdot \left(\frac{\sum_i x_{2i}}{N} \right) = \alpha_1 \overline{X}_1 + \alpha_2 \overline{X}_2 \ .$$

- *In words:* the mean of a linear combination of variables is equal to the linear combination of means.
- In particular, if $Z = X + Y$, then

$$\overline{Z} = \overline{X} + \overline{Y} \ .$$

The variance of the variable Z is

$$S_Z^2 = \frac{\sum_i \left(z_i - \overline{Z}\right)^2}{N} = \frac{\sum_i \left[\alpha_1 x_{1i} + \alpha_2 x_{2i} - \left(\alpha_1 \overline{X}_1 + \alpha_2 \overline{X}_2\right)\right]^2}{N}$$

$$= \frac{\sum_i \left[\alpha_1 \left(x_{1i} - \overline{X}_1\right) + \alpha_2 \left(x_{2i} - \overline{X}_2\right)\right]^2}{N}$$

$$= \frac{\sum_i \left[\alpha_1^2 \left(x_{1i} - \overline{X}_1\right)^2 + \alpha_2^2 \left(x_{2i} - \overline{X}_2\right)^2 + 2\alpha_1\alpha_2 \left(x_{1i} - \overline{X}_1\right)\left(x_{2i} - \overline{X}_2\right)\right]}{N}$$

$$= \alpha_1^2 \cdot \left[\frac{\sum_i \left(x_{1i} - \overline{X}_1\right)^2}{N}\right] + \alpha_2^2 \cdot \left[\frac{\sum_i \left(x_{2i} - \overline{X}_2\right)^2}{N}\right]$$

$$+2\alpha_1\alpha_2 \left[\frac{\sum_i \left(x_{1i} - \overline{X}_1\right)\left(x_{2i} - \overline{X}_2\right)}{N}\right] = \alpha_1^2 \cdot S_{X_1}^2 + \alpha_2^2 \cdot S_{X_2}^2 + 2\alpha_1\alpha_2 S_{X_1,X_2}.$$

- In particular, if $Z = X + Y$, then

$$S_Z^2 = S_X^2 + S_Y^2 + 2S_{X,Y} \ .$$

12. The mean vector and the variance covariance matrix

- Consider the following (column) vector of M variables:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{pmatrix}.$$

- The mean vector (or vector of means) of the vector of variables X is

$$\overline{X} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_M \end{pmatrix},$$

where $\overline{X}_j$ is the mean of the variable X_j.

- The variance covariance matrix (or covariance matrix) of the vector X of variables is the following $M \times M$ matrix

$$\mathrm{S} = \begin{pmatrix} \mathrm{S}_1^2 & \mathrm{S}_{12} & \cdots & \mathrm{S}_{1M} \\ \mathrm{S}_{21} & \mathrm{S}_2^2 & \cdots & \mathrm{S}_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{S}_{M1} & \mathrm{S}_{M2} & & \mathrm{S}_M^2 \end{pmatrix},$$

  where $\mathrm{S}_{jq} = \mathrm{S}_{X_j,X_q}$ and $\mathrm{S}_j^2 = \mathrm{S}_{X_j}^2 = \mathrm{S}_{X_j,X_j} = \mathrm{S}_{jj}$.

- Since $\mathrm{S}_{jq} = \mathrm{S}_{qj}$, for all pairs (j, q), the matrix S is symmetric.

- Define the vector of scalars

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{pmatrix}.$$

- Define the variable Z as a linear combination of the variables appearing in X,

$$Z = \sum_j \alpha_j X_j = \alpha^\intercal X,$$

where $^\intercal$ denotes the transpose.

- Then,

$$\overline{Z} = \sum_j \alpha_j \overline{X}_j = \alpha^\intercal \overline{X}$$

and

$$S_Z^2 = \sum_j \alpha_j^2 S_j^2 + 2 \sum_{\substack{q \\ j<q}} \sum_j \alpha_j \alpha_q S_{jq} = \sum_j \alpha_j^2 S_j^2 + \sum_{\substack{q \\ j \neq q}} \sum_j \alpha_j \alpha_q S_{jq}$$

$$= \sum_j \sum_q \alpha_j \alpha_q S_{jq} = \alpha^\intercal S \alpha, \quad \text{where}$$

- $\sum\limits_{\substack{q \\ j<q}} \sum\limits_j$ means $\sum\limits_{q=j+1}^{M} \sum\limits_{j=1}^{M-1}$,
- $\sum\limits_{\substack{q \\ j \neq q}} \sum\limits_j$ means $\sum\limits_{\substack{q=1 \\ q \neq j}}^{M} \sum\limits_{j=1}^{M}$,
- $\sum\limits_j \sum\limits_q$ means $\sum\limits_{j=1}^{M} \sum\limits_{q=1}^{M}$.

- Note that $S_Z^2 = \alpha^\intercal S \alpha \geq 0$ for all vector of scalars α. Therefore, the covariance matrix is (symmetric) positive semi-definite. This implies that the determinant of the matrix S is non-negative.

The Correlation Coefficient $\mathrm{r}_{X,Y}$

The correlation coefficient $\mathrm{r}_{X,Y}$ between two variables X and Y satisfies $-1 \leq \mathrm{r}_{X,Y} \leq 1$ or, equivalently, $|\mathrm{r}_{X,Y}| \leq 1$.

Proof. Consider the symmetric 2×2 variance covariance matrix of the vector of variables (X, Y),

$$\mathrm{S} = \begin{pmatrix} \mathrm{S}_X^2 & \mathrm{S}_{X,Y} \\ \\ \mathrm{S}_{X,Y} & \mathrm{S}_y^2 \end{pmatrix}.$$

Since for every vector of scalars,

$$\alpha = \begin{pmatrix} \alpha_1 \\ \\ \alpha_2 \end{pmatrix},$$

it is true that $\mathrm{Var}(\alpha_1 X + \alpha_2 Y) = \alpha^{\intercal} \mathrm{S} \alpha \geq 0$, it follows that the matrix S is symmetric positive semi-definite. This means that the determinants of all the main minors of S are non-negative, that is, $\mathrm{S}_X^2 \geq 0$, $\mathrm{S}_y^2 \geq 0$, and $\det(\mathrm{S}) \geq 0$. Note that

$$\det(\mathrm{S}) = \mathrm{S}_X^2 \mathrm{S}_y^2 - (\mathrm{S}_{X,Y})^2 = \mathrm{S}_X^2 \mathrm{S}_y^2 \left[1 - \left(\frac{\mathrm{S}_{X,Y}}{\mathrm{S}_X \mathrm{S}_y} \right)^2 \right] = \mathrm{S}_X^2 \mathrm{S}_y^2 \left[1 - (\mathrm{r}_{X,Y})^2 \right] \geq 0.$$

Since $\mathrm{S}_X^2 \mathrm{S}_y^2 > 0$, we have that $\det(\mathrm{S}) \geq 0$ if and only if $(\mathrm{r}_{X,Y})^2 \leq 1$. That is, it must hold that $|\mathrm{r}_{X,Y}| \leq 1$ or, equivalently $-1 \leq \mathrm{r}_{X,Y} \leq 1$. *Q.E.D.*

Exercises. Probability and Statistics. IDEA.
Descriptive Statistics

1. (a) Compute the mean, the median and the standard deviation of the following data in cm: 28, 22, 35, 42, 44, 53, 58, 41, 40, 32, 31, 38, 37, 61, 25, 35.

   (b) Classify the previous data into 5 classes (or intervals) with a length of 10 cm each ($(20-30]$; $(30-40;]$; etc.) and compute the previous characteristic measures using the formulae for grouped data.

   (c) Draw the histogram for these grouped data.

2. Find the value of a minimizing $\sum_{i=1}^{n}(x_i - a)^2$. Discuss.

3. Prove that the arithmetic mean of the variable Z obtained by adding the data of two variables, X and Y, equals the sum of the arithmetic means of these variables.

4. Prove that, if we construct a variable Z by mixing n_1 values of X and n_2 values of Y, the mean of Z is

$$\bar{Z} = \left(\frac{n_1}{n_1+n_2}\right)\bar{X} + \left(\frac{n_2}{n_1+n_2}\right)\bar{Y},$$

   where $\bar{X}$ and $\bar{Y}$ are the means of the two initial variables.

5. If $Z = X + Y$, prove that the variance of Z could be greater or smaller than the sum of the variances of the summands.

6. In 1879 Michelson obtained the following values for the speed of the light in the air (we provide the results subtracting 299.000 from the original data, in km/sec, to make easier the computations): 850, 740, 900, 1070, 930, 850, 950, 980, 980, 880, 1000, 980, 930, 650, 760.

   In 1882 Newcomb, using another procedure, obtained (subtracting again 299.000): 883, 816, 778, 796, 682, 711, 611, 599, 1051, 781, 578, 796, 774, 820, 772.

   (a) Compute the mean and the standard deviations.

   (b) What conclusions could you infer from (a)?

7. The geometric mean of the variable X is defined as

$$G = (x_1 \cdot x_2 \cdot .... \cdot x_n)^{1/n}$$

   and the harmonic mean of X as

$$H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}.$$

You are asked

(a) to find the relationship between the geometric mean and the mean of the logarithm of the original variable X;

(b) the same between H and the mean of Y, where $Y = X^{-1}$.

8. Prove that, if there exists a exact linear relationship between two variables, $Y = a + bX$, with $b > 0$, then their correlation coefficient is equal to one.

9. Compute the variance and the correlation coefficient of the following data:

$$X: \quad 2 \quad 1 \quad 2 \quad 1 \quad 4$$

$$Y: \quad 8 \quad 9 \quad 8 \quad 5 \quad 10$$

10. Consider the following joint distribution of absolute frequencies $n(x, y)$ of the variables $X =$ age and $Y =$ weight of a group of 42 secondary school boys. There are only two ages: 13 and 14 years. The weights (in kilograms) are grouped into three intervals (or classes): (35,45], (45,55] and (55,65], which are parametrized by their corresponding midpoints (40, 50 and 60).

		$Y =$ weight		
		40	50	60
$X =$ age	13	4	7	7
	14	3	8	13

You are asked to

(a) compute the distributions of marginal relative frequencies of age, $f_X(x)$ for $x = 13, 14,$ and weight, $f_Y(y)$ for $y = 40, 50, 60$;

(b) compute the means of age $\bar{X}$ and weight $\bar{Y}$;

(c) compute the variances of age S_X^2 and weight S_Y^2;

(d) compute the covariance between age and weight S_{XY} and the corresponding correlation coefficient r_{XY};

(e) compute the distribution of conditional (relative) frequencies of age given that the weight is 60, $f_{X|Y}\,(x\,|60)$ for $x = 13, 14$.

11. Let the variable X be the number of theater attendances and Y the number of cinema attendances. Consider the following table of monthly relative frequencies of theater and cinema attendances for a group of retired people:

		X		
		0	1	2
Y	1	0.41	0.05	0
	2	0.19	0.06	0.02
	3	0.10	0.05	0.02
	4	0.02	0.07	0.01

You are asked to

(a) find the distribution of marginal relative frequencies of the number of cinema attendances;

(b) find the distribution of conditional (relative) frequencies of the number of cinema attendances for the retired people who have not gone to the theater;

(c) compute the means and the standard deviations of the variables X and Y.

(d) compute the covariance between the variables X and Y.

12. The distribution of the number of accidents caused by 705 bus drivers in the last 4 years has been the following:

Number of accidents	n
0	114
1	157
2	158
3	115
4	78
5	44
6	21
7	7
8	6
9	1
10	3
11	1

(a) What type of variable are we dealing with? (discrete or continuous, qualitative or quantitative).

(b) Complete the frequency table by adding the relative frequencies f, the absolute cumulative frequencies N, and the relative cumulative frequencies F.

(c) Draw the bar diagram (or chart) of absolute frequencies.

(d) Compute the mode, the mean and the median.

(e) Find the 50%, 75% and 40% percentiles; $x_{50\%}$, $x_{75\%}$ and $x_{40\%}$.

13. An exporting firm sells its product to three countries (China, Japan and Korea). The variable X represents the number of commercial representatives in each country and the variable Y represents the revenues from each country (measured in units of one hundred thousand euros). The following table summarizes the information:

	X	Y
China	3	7
Japan	4	9
Korea	5	7

(a) Compute the variances of X and Y, $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$.

(b) Compute the covariance between X and Y, $\mathrm{Cov}(X, Y)$.

(c) Find the conditional (relative) frequencies of X given $Y = 7$, $f_{X|Y}(x\,|7)$ for $x = 3, 4, 5$, and the conditional (relative) frequencies of X given $Y = 9$, $f_{X|Y}(x\,|9)$ for $x = 3, 4, 5$.

14. The following observations correspond to the market value of two given shares of stock during 10 days:

    Share of stock A: 1.25, 1.34, 1.02, 1.01, 0.98, 1.12, 1.40, 1.23 , 1.10, 1.02
    Share of stock B: 1.40, 0.98, 1.32, 1.20, 0.97, 1.24, 1.10, 0.89, 1.36, 1.01

    (a) What is the mean value of each share of stock during the observed period?

    (b) What is the median of the values of each share of stock during the observed period?

    (c) From the point of view of a potential investor, and according to these observations, which of the two shares of stock is less risky (that is, which one has smaller variance)?

15. The variable X represents the number of years that a former student of Economics and Business of the UAB remained unemployed after completing her degree. The variable Y represents the number of questions of the final exam of Statistics that the same former student answered correctly. Consider the following table of relative frequencies of X and Y from the population of former students:

		X = 0	X = 1
Y	0	0,01	0,28
	1	0,07	0,17
	2	0,13	0,08
	3	0,10	0,01
	4	0,15	0,00

    You are asked to:

    (a) Find the distribution of marginal relative frequencies of the number of correct answers in the exam.

    (b) Find the distribution of conditional frequencies of the number of correct answers for the former students who remained unemployed during one year.

    (c) Compute the mean and the standard deviation of the variables X and Y.

    (d) Compute the covariance and the correlation coefficient between X and Y. What conclusion do you reach about the relationship between the number of correct answers and the number of years unemployed?

Probability and Statistics. IDEA. Answers.

Descriptive Statistics

1. (a) The mean is

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{622}{16} = 38.875$$

The median is:

1. the central value, if it is unique, after ordering the observations according to their magnitude, or

2. if there are two central values, then it is the mean of these two values.

Since $n = 16$, the central values are $x_8 = 37$ and $x_9 = 38$. Thus, the median is:

$$Median = x_{50\%} = \frac{x_8 + x_9}{2} = 37.5.$$

$$\mathrm{S}_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{X}^2} = 10.7114.$$

(b)

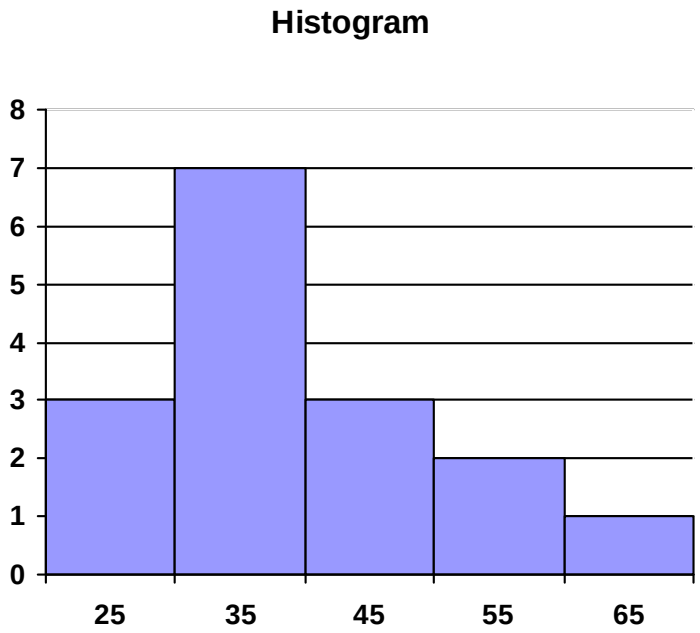
classes	midpoint	observations	$n(x_{i)}$
$(20, 30]$	25	$22, 25, 28$	3
$(30, 40]$	35	$31, 32, 35, 35, 37, 38, 40$	7
$(40, 50]$	45	$41, 42, 44$	3
$(50, 60]$	55	$53, 58$	2
$(60, 70]$	65	61	1

$$\bar{X}_c = \sum_x x f(x) = 25\frac{3}{16} + 35\frac{7}{16} + 45\frac{3}{16} + 55\frac{2}{16} + 65\frac{1}{16} = 39.375.$$

$$Median = 35.$$

$$\mathrm{S}_x^c = \sqrt{\sum_x x^2 f(x) - \bar{X}_c^{\,2}} = \sqrt{124.61} = 11.163.$$

(c)



– – – – – – – – – – – –

2.

$$\frac{d(\sum_{i=1}^n (x_i - a)^2)}{da} = \sum_{i=1}^n 2(x_i - a) = 0.$$

Therefore $\sum_i x_i = na$ and, hence,

$$a = \frac{\sum_i x_i}{n} = \bar{X}.$$

This means that, if we want to make the smallest sum of squared deviations from a value, this value must be the mean of the observations. Note that the formula for the variance uses this value.

– – – – – – – – – – –

3. Let $X : x_1, \ldots, x_n$ and $Y : y_1, \ldots, y_n$, and $Z = X + Y : x_1 + y_1, \ldots, x_n + y_n$. Then

$$\bar{Z} = \frac{\sum_i z_i}{n} = \frac{\sum_i (x_i + y_i)}{n} = \frac{\sum_i x_i}{n} + \frac{\sum_i y_i}{n} = \bar{X} + \bar{Y}.$$

– – – – – – – – – – –

4. Let $X : x_1, \ldots, x_{n_1}$ and $Y : y_1, \ldots, y_{n_2}$, and $Z = X, Y : x_1 \ldots, x_{n_1}, y_1, \ldots, y_{n_2} = z_1, \ldots, z_{n_1+n_2}$. Then

$$\bar{Z} = \frac{\sum_{i=1}^{n_1+n_2} z_i}{n_1+n_2} = \frac{\sum_{i=1}^{n_1} z_i}{n_1+n_2} + \frac{\sum_{i=n_1+1}^{n_1+n_2} z_i}{n_1+n_2} = \frac{\sum_i x_i}{n_1+n_2} + \frac{\sum_i y_i}{n_1+n_2} = \frac{n_1}{n_1+n_2}\bar{X} + \frac{n_2}{n_1+n_2}\bar{Y}.$$

– – – – – – – – – – –

5. $Z = X + Y$

$$\mathrm{Var}(\alpha X + \beta Y) = \alpha^2 \mathrm{Var}(X) + \beta^2 \mathrm{Var}(Y) + 2\alpha\beta \mathrm{Cov}(X, Y),$$

so that

$$\mathrm{Var}(Z) = \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).$$

Since $\mathrm{Cov}(X,Y) \gtreqless 0$, then $\mathrm{Var}(Z) \gtreqless \mathrm{Var}(X) + \mathrm{Var}(Y)$.

– – – – – – – – – – – –

6. (a) Let the Michelson's observations be the variable X, and those of Newcomb be Y. Then $\bar{X} = 896.667$ and $\bar{Y} = 763.2$, while $\mathrm{S}_X = 108.115$ and $\mathrm{S}_Y = 115.418$.

(b) The Michelson data's mean is closer to the theoretical velocity of the light. Moreover, it has smaller standard deviation. Summing up, Michelson made a better job.

– – – – – – – – – – – –

7. (a) Let $X : x_1, \ldots, x_n$ and let $\ln X : \ln x_1, \ldots, \ln x_n$. Note that we must have $x_i > 0$. Then, the mean of $\ln X$:

$$\overline{\ln X} = \frac{1}{n}\sum_i \ln x_i = \frac{1}{n}\ln\left(\prod_i x_i\right) = \ln\left(\prod_i x_i\right)^{1/n} = \ln G.$$

(b) Now let $Y = X^{-1} : \frac{1}{x_1}, \ldots, \frac{1}{x_n}$. Then

$$\overline{Y} = \overline{X^{-1}} = \frac{1}{n}\sum_i \frac{1}{x_i} = H^{-1}.$$

– – – – – – – – – – – –

8. $Y = a + bX$ and $b > 0$, then $\mathrm{Var}(Y) = b^2\mathrm{Var}(X)$ and $\mathrm{Cov}(X,Y) = b\mathrm{Var}(X)$. Therefore,

$$\mathrm{r}_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{S}_X\mathrm{S}_Y} = \frac{b\mathrm{S}_X^2}{|b|\mathrm{S}_X^2} = \frac{b}{|b|} = 1$$

if $b > 0$.

– – – – – – – – – – – –

9. $\bar{X} = 2$, $\bar{Y} = 8$, $\mathrm{Var}(X) = \mathrm{S}_X^2 = 1.2$, $\mathrm{Var}(Y) = \mathrm{S}_Y^2 = 2.8$.

$$\mathrm{Cov}(X,Y) = \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{n} = \frac{\sum_i x_i y_i}{n} - \bar{X}\bar{Y} = 17.2 - 16 = 1.2.$$

Thus,

$$\mathrm{r}_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{S}_X \mathrm{S}_Y} = \frac{1.2}{\sqrt{1.2}\sqrt{2.8}} = 0.65465.$$

– – – – – – – – – – – –

10. (a) $f_X(13) = \frac{18}{42} = \frac{3}{7}$, $f_X(14) = \frac{24}{42} = \frac{4}{7}$.

$f_Y(40) = \frac{7}{42} = \frac{1}{6}$, $f_Y(50) = \frac{15}{42} = \frac{5}{14}$, $f_Y(60) = \frac{20}{42} = \frac{10}{21}$.

(b) $\bar{X} = \sum_x x f_X(x) = 13.5714$, and $\bar{Y} = \sum_y y f_Y(y) = 53.0952$.

(c) $\mathrm{S}_X^2 = \sum_x x^2 f_X(x) - \bar{X}^2 = 184.4286 - 184.1837 = 0.2449$ and

$\mathrm{S}_Y^2 = \sum_y y^2 f_Y(y) - \bar{Y}^2 = 2873.8095 - 2819.1043 = 54.7052$.

(d)

$$\mathrm{S}_{X,Y} = \sum_x \sum_y xy f(x,y) - \bar{X}\bar{Y} = 721.1905 - 720.5782 = 0.6122.$$

$$\mathrm{r}_{X,Y} = \frac{\mathrm{S}_{X,Y}}{\mathrm{S}_X \mathrm{S}_Y} = 0.1673.$$

(e) $f_{X|Y}(13|60) = \frac{7}{20}$ and $f_{X|Y}(14|60) = \frac{13}{20}$.

– – – – – – – – – – – –

11. (a) $f_Y(1) = 0.46,\ f_Y(2) = 0.27,\ f_Y(3) = 0.17,\ f_Y(4) = 0.10.$

(b) Note that $f_X(0) = 0.72, f_X(1) = 0.23, f_X(2) = 0.05.$

$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \Rightarrow f_{Y|X}(y|0) = \frac{f(0,y)}{f_X(0)}$. Therefore, $f_{Y|X}(1|0) = \frac{0.41}{0.72} = 0.57,\ f_{Y|X}(2|0) = \frac{0.19}{0.72} = 0.26,\ f_{Y|X}(3|0) = \frac{0.10}{0.72} = 0.14,\ f_{Y|X}(4|0) = \frac{0.02}{0.72} = 0.03.$

(c) $\bar{X} = \sum_{x=0}^{2} x f_X(x_i) = 0.33,\ \bar{Y} = \sum_{y=1}^{4} y f_Y(y_i) = 1.91.$

$S_X^2 = \sum_{x=0}^{2} x^2 f_X(x) - \bar{X}^2 = 0.43 - 0.1089 = 0.3211.$ Thus, $S_X = 0.5667.$

$S_Y^2 = \sum_{y=1}^{4} y^2 f_Y(y) - \bar{Y}^2 = 4.67 - 3.6481 = 1.0219.$ Thus, $S_Y = 1.0109.$

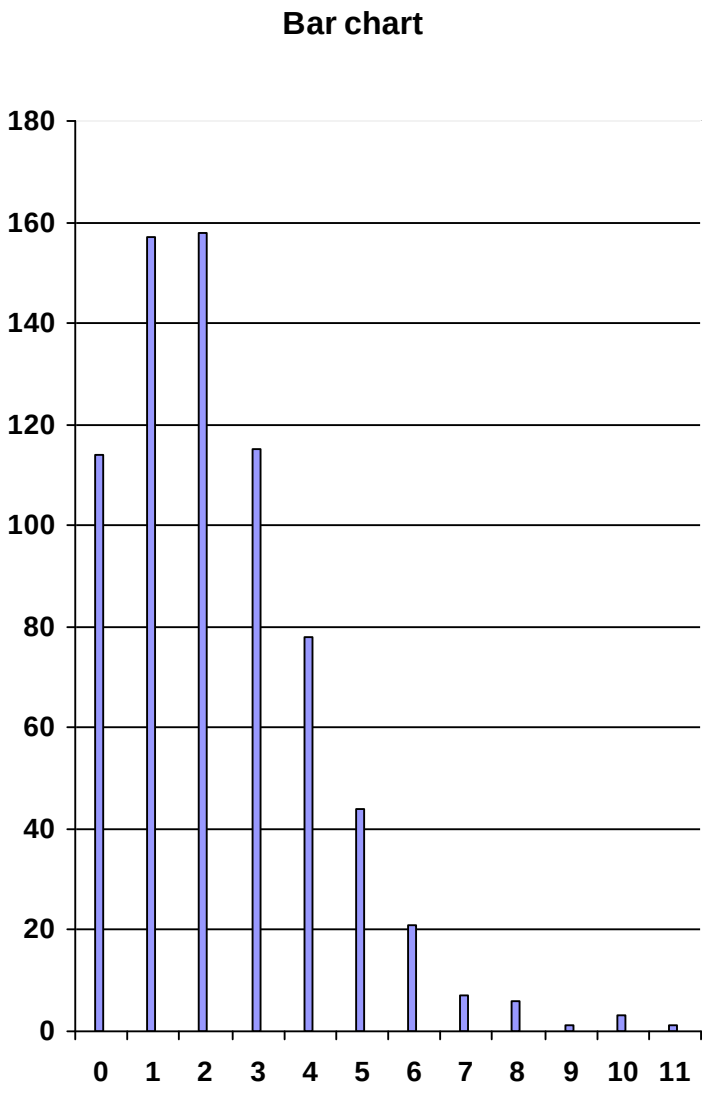
(d) $S_{X,Y} = \sum_{x=0}^{2} \sum_{y=1}^{4} xy f(x,y) - \bar{X}\bar{Y} = 0.88 - 0.6303 = 0.2497.$

– – – – – – – – – – –

12. (a) Discrete and quantitative.

(b) $f(x) = \frac{n(x)}{705},\ N(x) = \sum_{j=0}^{x} n(j),\ F(x) = \frac{N(x)}{705},\ x = 0, 1, ..., 11.$

(c)



(d) $Mode = 2, Mean = 2.3021, Median = 2.$

(e) $x_{50\%} = 2, x_{75\%} = 3, x_{40\%} = 2.$

– – – – – – – – – – –

13. Table of absolute frequencies, $n(x, y)$:

$Y \backslash X$	3	4	5	
7	1	0	1	2
9	0	1	0	1
	1	1	1	3

(a) $\overline{X} = \frac{3+4+5}{3} = 4$ and $\overline{Y} = \frac{7+9+7}{3} = \frac{23}{3}$.
$\text{Var}(X) = \frac{\sum_i (x_i)^2}{3} - \overline{X}^2 = \frac{9+16+25}{3} - (4)^2 = \frac{2}{3}$.
$\text{Var}(Y) = \frac{\sum_i (y_i)^2}{3} - \overline{Y}^2 = \frac{49+81+49}{3} - \left(\frac{23}{3}\right)^2 = \frac{8}{9}$.
(b) $\text{Cov}(X, Y) = \frac{\sum_i x_i y_i}{3} - \left(\overline{X} \cdot \overline{Y}\right) = \frac{3 \cdot 7 + 4 \cdot 9 + 5 \cdot 7}{3} - \left(4 \cdot \frac{23}{3}\right) = \frac{21+36+35}{3} - \frac{92}{3} = 0$.

(c) $$f_{X|Y}(x\,|7) = \frac{f(x,7)}{f_Y(7)} = \frac{n(x,7)}{n_Y(7)} = \frac{n(x,7)}{2} = \begin{cases} \dfrac{n(3,7)}{2} = \dfrac{1}{2} & \text{for } x = 3 \\ \dfrac{n(4,7)}{2} = 0 & \text{for } x = 4 \\ \dfrac{n(5,7)}{2} = \dfrac{1}{2} & \text{for } x = 5. \end{cases}$$

$$f_{X|Y}(x\,|9) = \frac{f(x,9)}{f_Y(9)} = \frac{n(x,9)}{n_Y(9)} = \frac{n(x,9)}{1} = \begin{cases} \dfrac{n(3,9)}{1} = 0 & \text{for } x = 3 \\ \dfrac{n(4,9)}{1} = \dfrac{1}{1} = 1 & \text{for } x = 4 \\ \dfrac{n(5,9)}{1} = 0 & \text{for } x = 5. \end{cases}$$

14. (a) $Mean = \frac{\sum_{i=1}^{10} x_i}{10}$. $Mean\ of\ A = Mean\ of\ B = 1.1470$.

(b) $Median\ of\ A = \frac{1.10+1.12}{2} = 1.11$, $Median\ of\ B = \frac{1.10+1.20}{2} = 1.15$.

(c) $Variance = \frac{\left(\sum_{i=1}^{10} x_i^2\right)}{10} - \left(\frac{\sum_{i=1}^{10} x_i}{10}\right)^2$. Then,

$Variance\ of\ A = 0.0201$, $Variance\ of\ B = 0.0297 \Longrightarrow$ A is less risky

than B.

– – – – – – – – – – – –

15. (a) $f_Y(0) = 0.29$, $f_Y(1) = 0.24$, $f_Y(2) = 0.21$, $f_Y(3) = 0.11$, $f_Y(4) = 0.15$.

(b)

$$f_{Y|X}(y\,|1) = \frac{f_{X,Y}(1,y)}{f_X(1)} = \frac{f_{X,Y}(1,y)}{0.54}, \quad \text{for } y = 0,1,2,3,4.$$

Then, $f_{Y|X}(0\,|1) = 0.5185$, $f_{Y|X}(1\,|1) = 0.3148$, $f_{Y|X}(2\,|1) = 0.1481$, $f_{Y|X}(3\,|1) = 0.0185$, $f_{Y|X}(4\,|1) = 0$.

(c) $\bar{X} = 0 \cdot f_X(0) + 1 \cdot f_X(1) = 0.54$,

$\bar{Y} = \sum_{y=0}^{4} y f_Y(y) = 0 \cdot 0.29 + 1 \cdot 0.24 + 2 \cdot 0.21 + 3 \cdot 0.11 + 4 \cdot 0.15 = 1.59.$

$\mathrm{S}_X = \sqrt{(0-0.54)^2 \cdot 0.46 + (1-0.54)^2 \cdot 0.54} = 0.4984$, or alternatively,

$\mathrm{S}_X = \sqrt{(0^2 \cdot 0.46 + 1^2 \cdot 0.54) - (0.54^2)} = 0.4984$

$$\mathrm{S}_Y = \sqrt{\begin{array}{c}(0-1.59)^2 \cdot 0.29 + (1-1.59)^2 \cdot 0.24 + (2-1.59)^2 \cdot 0.21 \\ + (3-1.59)^2 \cdot 0.11 + (4-1.59)^2 \cdot 0.15\end{array}} = 1.3935,$$

or alternatively,

$\mathrm{S}_Y = \sqrt{(0^2 \cdot 0.29 + 1^2 \cdot 0.24 + 2^2 \cdot 0.21 + 3^2 \cdot 0.11 + 4^2 \cdot 0.15) - (1.59^2)} = 1.3935.$

(d) $\mathrm{S}_{X,Y} = \sum_{x=0}^{1}\sum_{y=0}^{4} \big(x - \bar{X}\big) \cdot \big(y - \bar{Y}\big) \cdot f(x,y) = \left[\sum_{x=0}^{1}\sum_{y=0}^{4} x.y \cdot f(x,y)\right] - \bar{X}\bar{Y}$

$= \left[\sum_{y=1}^{4} 0 \cdot y \cdot f(0,y) + \sum_{y=0}^{4} 1 \cdot y \cdot f(1,y)\right] - \bar{X}\bar{Y} = \left[\sum_{y=0}^{4} 1 \cdot y \cdot f(1,y)\right] - \bar{X}\bar{Y}$

$= (0 + 1 \cdot 0.17 + 2 \cdot 0.08 + 3 \cdot 0.01 + 0) - (0.54 \cdot 1.59) = -0.4986.$

$\mathrm{r}_{X,Y} = \dfrac{\mathrm{S}_{X,Y}}{\mathrm{S}_X \mathrm{S}_y} = \dfrac{-0.4986}{0.4984 \cdot 1.3935} = -0.7179.$

Therefore, we see that the larger is the number of correct answers, the lower is the average number of years that the former students remain unemployed.

– – – – – – – – – – –