Departamento de Ciencia de la Computación (DC-UCSP)

Maestría en Ciencia de Datos

Descriptive Statistics

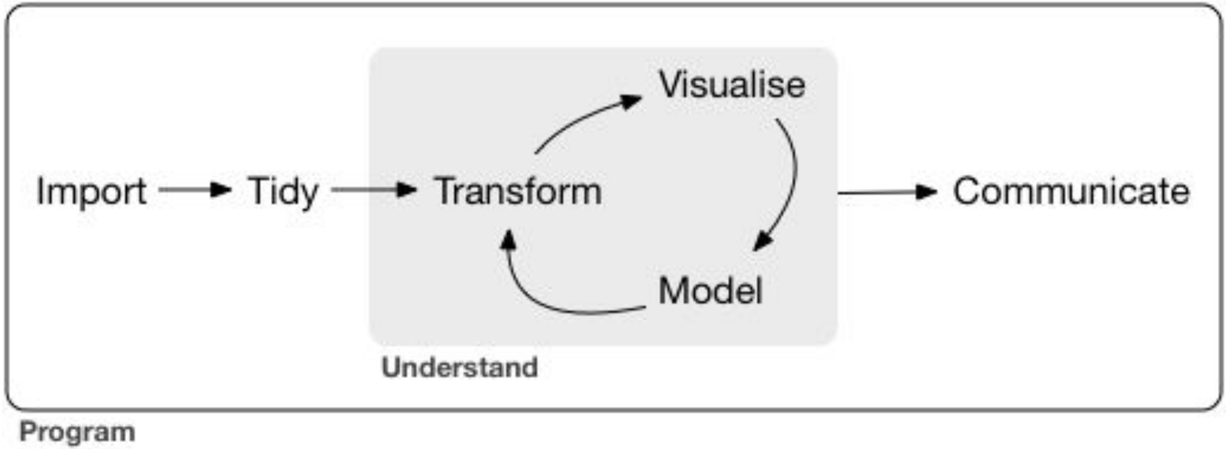
Daniel Alexis Gutierrez Pachas, PhD
dgutierrezp@ucsp.edu.pe

Universidad Católica San Pablo | Departamento de Ciencia de la Computación

- Statistics is the science of developing and applying methods for collecting, analyzing, and interpreting data.

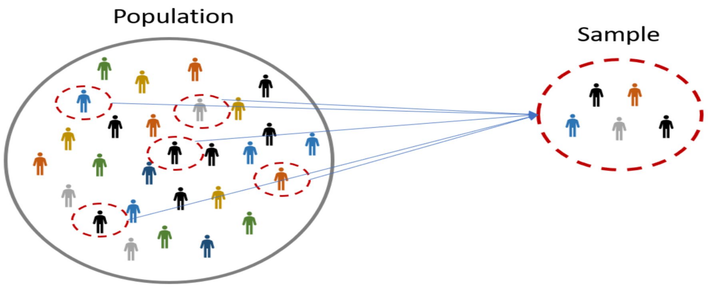
- Statistical methods incorporate reasoning using tools of probability. These methods enable us to deal with uncertainty and variability.

- Statistical science has three aspects: (a) Design, (b) Description, and (c) Inference.



The **units** on which we measure data—such as persons, cars, animals, or plants are called **observations**. These units/observations are represented by **ω**. The collection of all units is called **population** and is represented by Ω. A selection of observations are called **sample.**

The goal of most data analyses is to learn about populations. But it is almost always necessary, and more practical, to observe only samples from those populations.



To distinguish between a descriptive statistic calculated for a sample and the corresponding characteristic of the population, we use the term parameter for the population characteristic.

A variable is a characteristic that can vary in value among subjects in a sample or population

- Generally, a variable can be **Quantitative** or **Qualitative**.
- A quantitative variable can be **Discrete** or **Continuous**.
- A qualitative variable can be **Nominal or Ordinal.** A nominal variables cannot be ordered. However, a ordinal variable can be ordered.

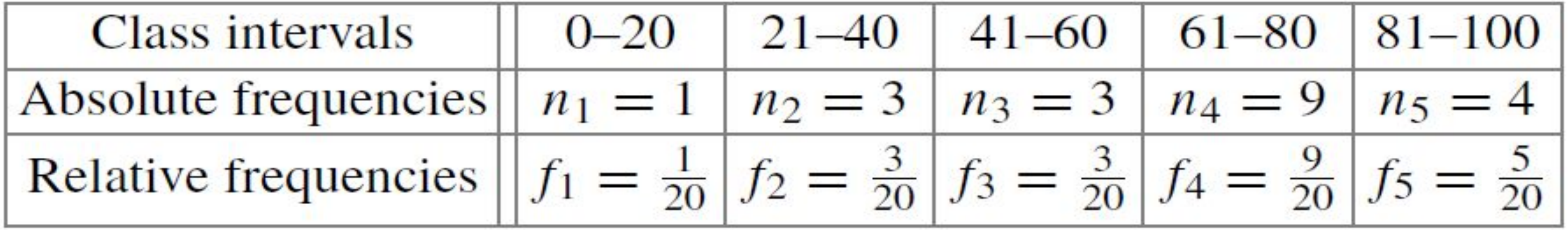
We can be interested in many different features. Each of them collected in a different variable. Each observation ω takes a particular value for X. We store in a **data matrix**.

$$\begin{array}{ccccc} \omega & \text{Variable}\,1 & \text{Variable}\,2 & \cdots & \text{Variable}\,p \\ \left(\begin{array}{c} 1 \\ 2 \\ \vdots \\ n \end{array}\right. & \begin{array}{c} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{array} & \begin{array}{c} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{array} & \begin{array}{c} \cdots \\ \cdots \\ \\ \cdots \end{array} & \left.\begin{array}{c} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{array}\right) \end{array}$$

Frequency tables and empirical **cumulative distribution** functions help provide a numerical summary of a variable. Also, **Graphs are an alternative** too**.** Generally, an adequate graph depends on the variable type.

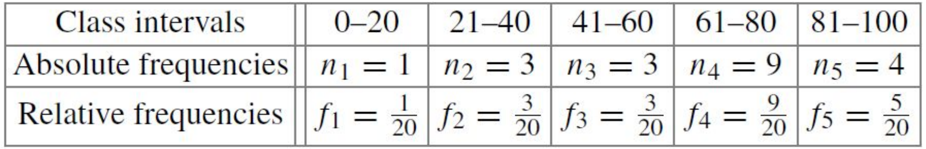
Considering 20 results of the written part of a driving licence examination (maximum value is 100)

28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.

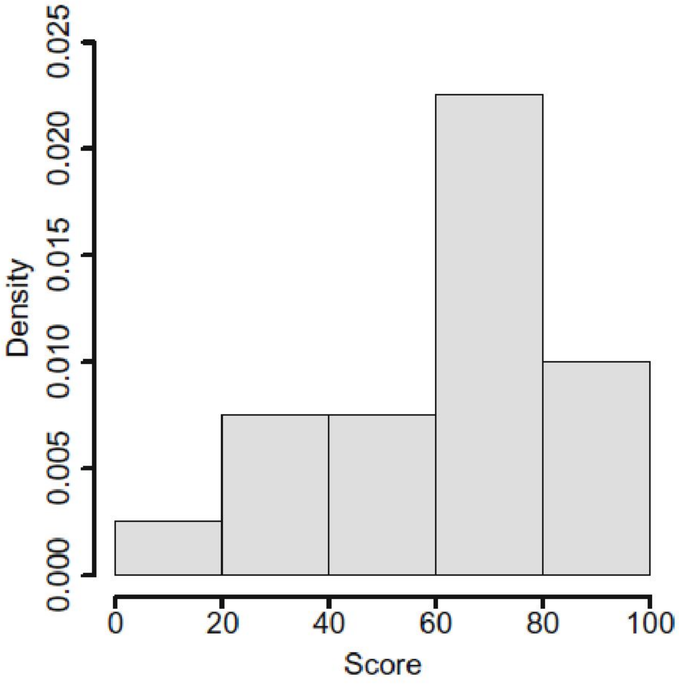


Class intervals	0–20	21–40	41–60	61–80	81–100
Absolute frequencies	$n_1 = 1$	$n_2 = 3$	$n_3 = 3$	$n_4 = 9$	$n_5 = 4$
Relative frequencies	$f_1 = \frac{1}{20}$	$f_2 = \frac{3}{20}$	$f_3 = \frac{3}{20}$	$f_4 = \frac{9}{20}$	$f_5 = \frac{5}{20}$

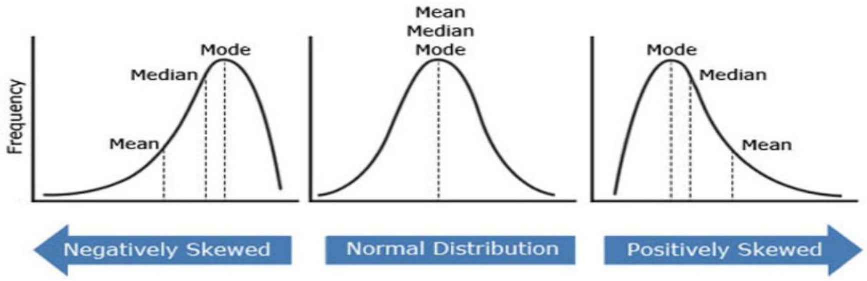
A **histogram** is the appropriate choice to represent the distribution of values of continuous variables. A disadvantage of histograms is that continuous data is categorized artificially. The choice of the class intervals is crucial for the final look of the graph.



Class intervals	0–20	21–40	41–60	61–80	81–100
Absolute frequencies	$n_1 = 1$	$n_2 = 3$	$n_3 = 3$	$n_4 = 9$	$n_5 = 4$
Relative frequencies	$f_1 = \frac{1}{20}$	$f_2 = \frac{3}{20}$	$f_3 = \frac{3}{20}$	$f_4 = \frac{9}{20}$	$f_5 = \frac{5}{20}$



A natural human tendency is to make comparisons with the “average”. We call statistical functions named measures of central tendency, as **Mean** and **Median**. The **Mode** is the value that appears most frequently in a data set.



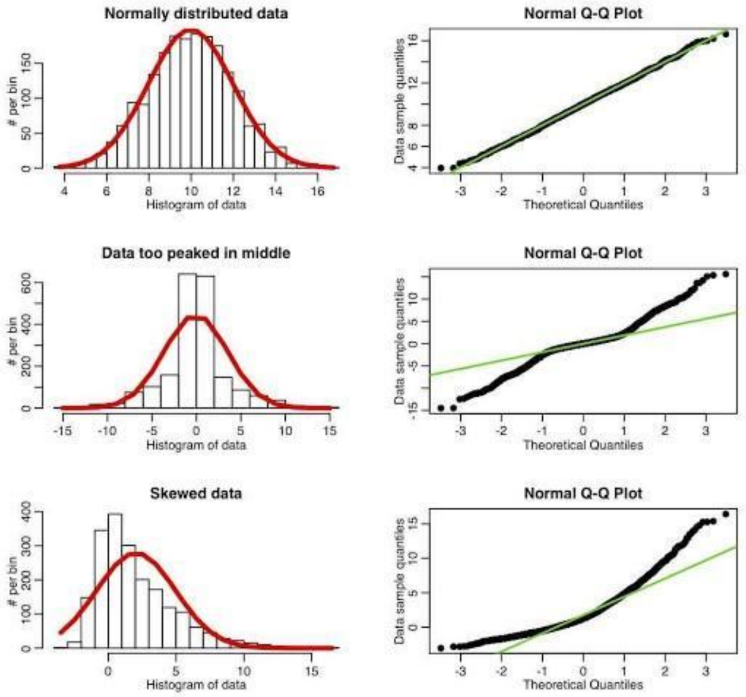
- Example 1: Given the values 1,3,5,5,6,7,8,10,12. Compute the mean and median.

- Example 2: Given the values 1,3,5,5,6,7,8,9,10,12. Compute the mean and median.

- Example 3: Given the values 1,3,5,5,6,7,8,9,10,12,100. Compute the mean and median.

Besides center and variability ,another way to describe a distribution is with a measure of position. The pth percentile is the point such that p% of the observations fall below or at that point and (100 − p)% fall above it. The 50th percentile is the median. Quantiles are percentiles expressed in proportion form. For example, the 95th percentile is also called the 0.95 quantile.

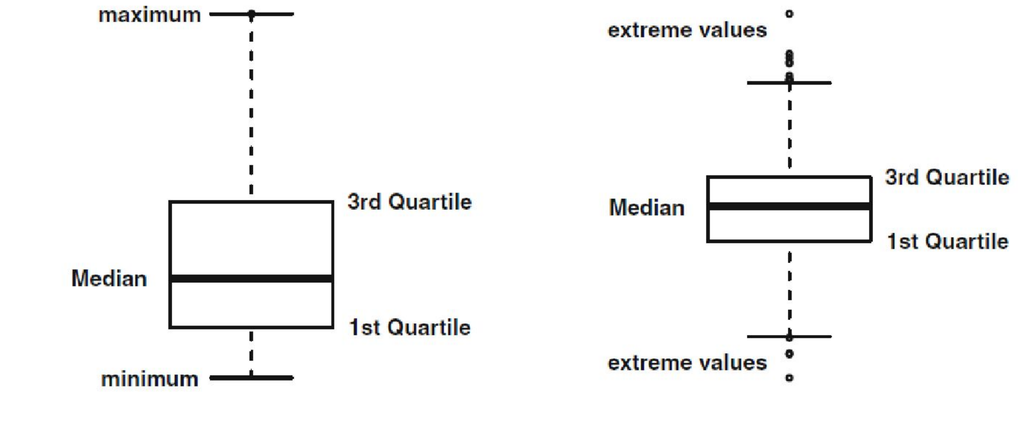
Two especially useful percentiles are the 25th percentile, called the lower quartile (Q1), and the 75th percentile, called the upper quartile (Q3). The interquartile range (IQR), is IQR=Q3-Q1. We illustrate the behavior of two variables employing a **Quantile–Quantile plot (QQ-plot)**.



So far we have described various measures of central tendency and dispersion. It can be tedious to list those measures in summary tables.

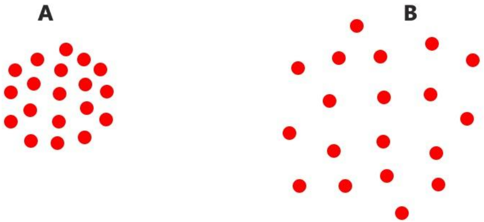
A simple and powerful graph is the **box plot** which summarizes the distribution of a continuous variable by using its median, quartiles, minimum, maximum, and extreme values. The box plot is one of many methods of exploratory data analysis proposed by John Tukey (1977).

An observation is identified as an outlier if it falls more than 1.5 (IQR) below Q1 or above Q3.

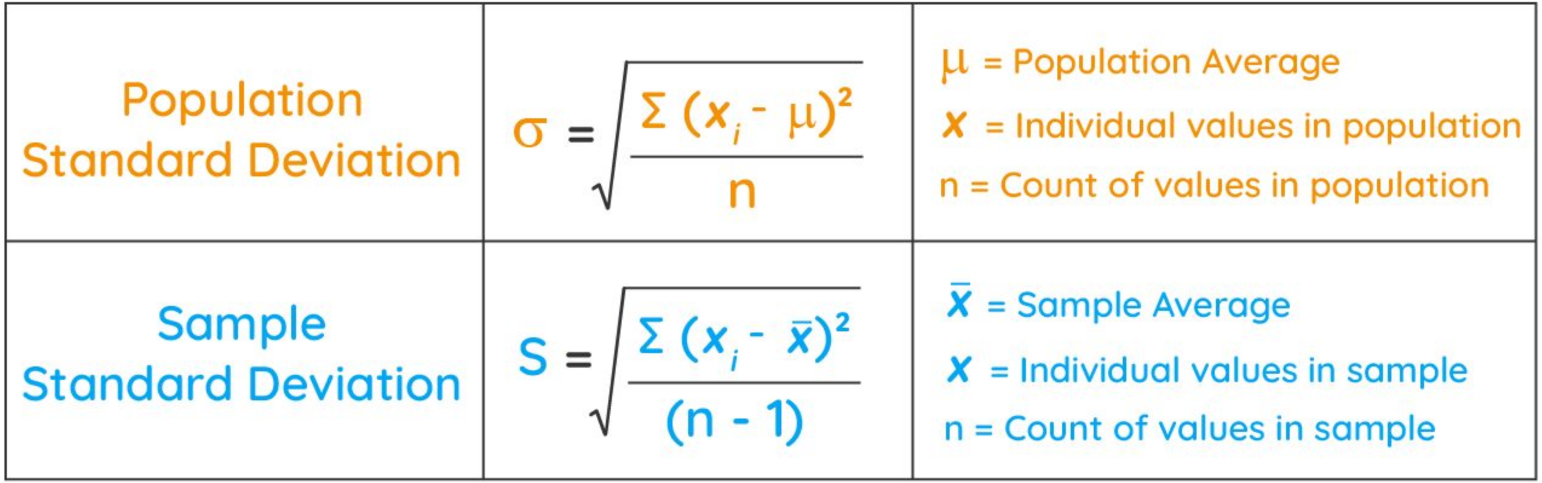


(a) Box plot without extreme values

(b) Box plot with extreme values



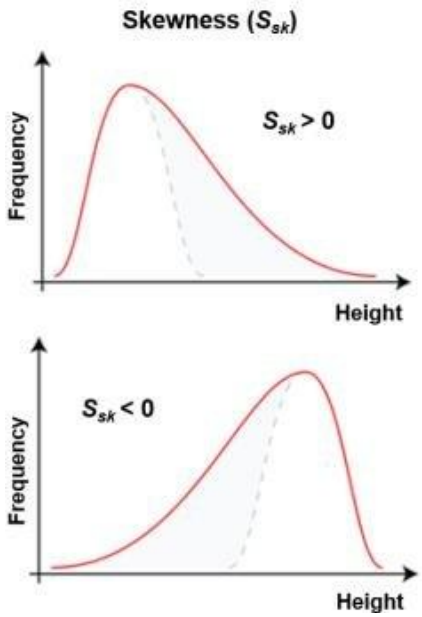
The concentration or dispersion of observations around any particular value is another property which characterizes the data and its distribution. We now introduce statistical measures named **Standard Deviation** and **Variance**.

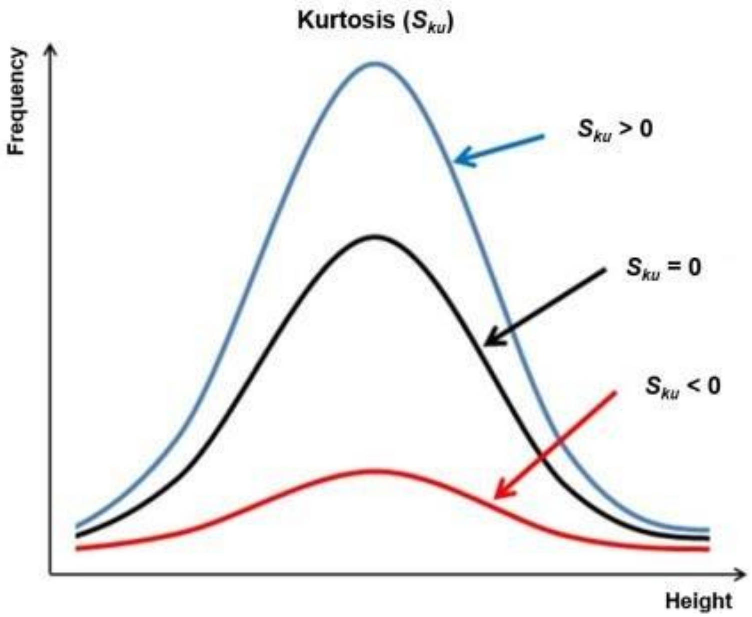


Population Standard Deviation	$\sigma = \sqrt{\dfrac{\Sigma (x_i - \mu)^2}{n}}$	μ = Population Average x = Individual values in population n = Count of values in population
Sample Standard Deviation	$S = \sqrt{\dfrac{\Sigma (x_i - \bar{x})^2}{(n - 1)}}$	$\bar{x}$ = Sample Average x = Individual values in sample n = Count of values in sample

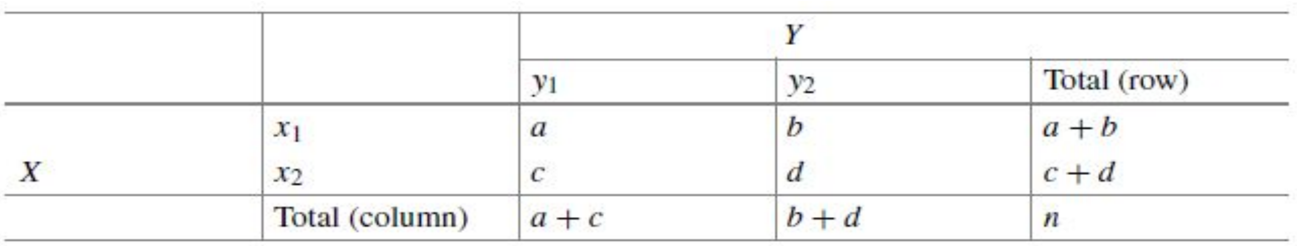
- The coefficient of **Skewness** is a measure for the degree of symmetry in the variable distribution.

- The coefficient of **Kurtosis** is a measure for the degree of peakedness/flatness in the variable distribution.





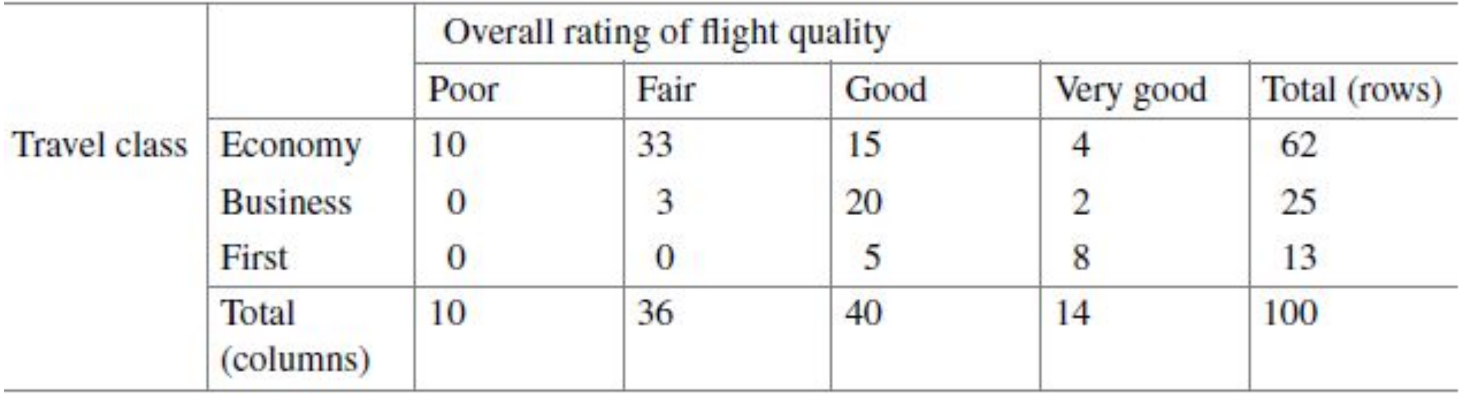
In many situations, we may be interested in the interdependence of two or more variables. For example, suppose we want to know whether male and female students in a college have any preference between the subjects mathematics and biology. We expect that if there is no association between the two variables “gender of student” (male or female) and “subject” (mathematics or biology), then an equal proportion of male and female students should choose the subjects biology and mathematics respectively. Any difference in the proportions may indicate a preference of males or females for a particular topic. Generally, the data can be summarized in a two-dimensional **contingency table.**



		Y		
		y_1	y_2	Total (row)
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
	Total (column)	$a+c$	$b+d$	n

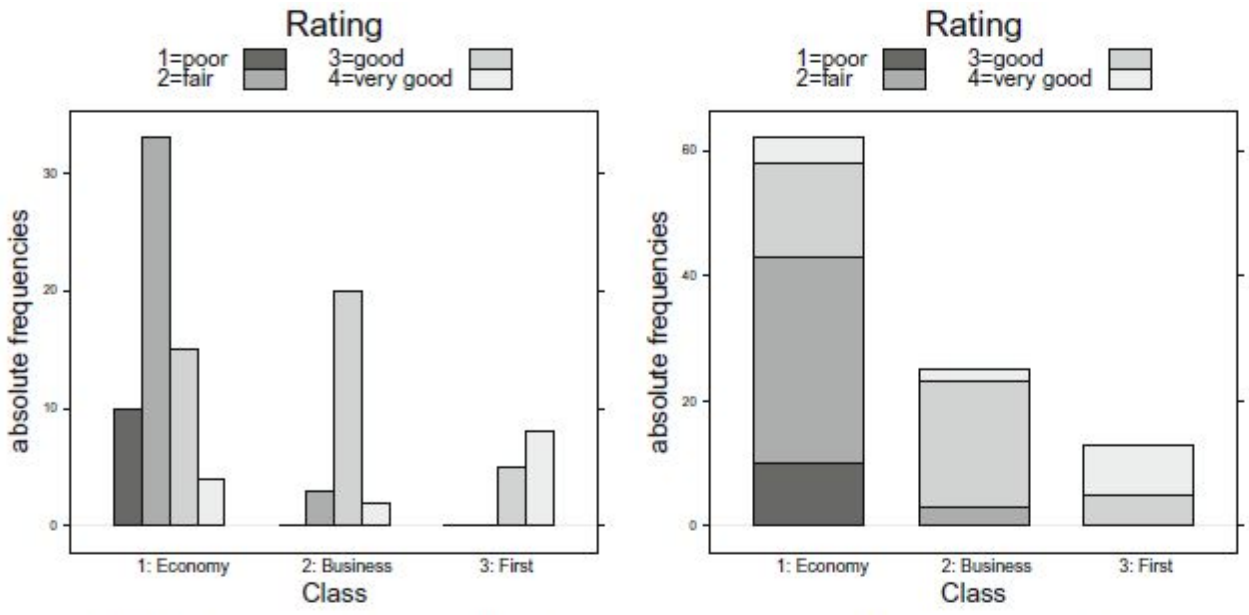
The variables X and Y are independent if $a/(a+c) = b/(b+d) = (a+b)/n$.

An airline conducts a customer satisfaction survey. The survey includes questions about travel class and satisfaction levels with respect to different categories such as seat comfort, in-flight service, meals, safety, and other indicators. A contingency table for 100 customers is:



		Overall rating of flight quality				
		Poor	Fair	Good	Very good	Total (rows)
Travel class	Economy	10	33	15	4	62
	Business	0	3	20	2	25
	First	0	0	5	8	13
	Total (columns)	10	36	40	14	100

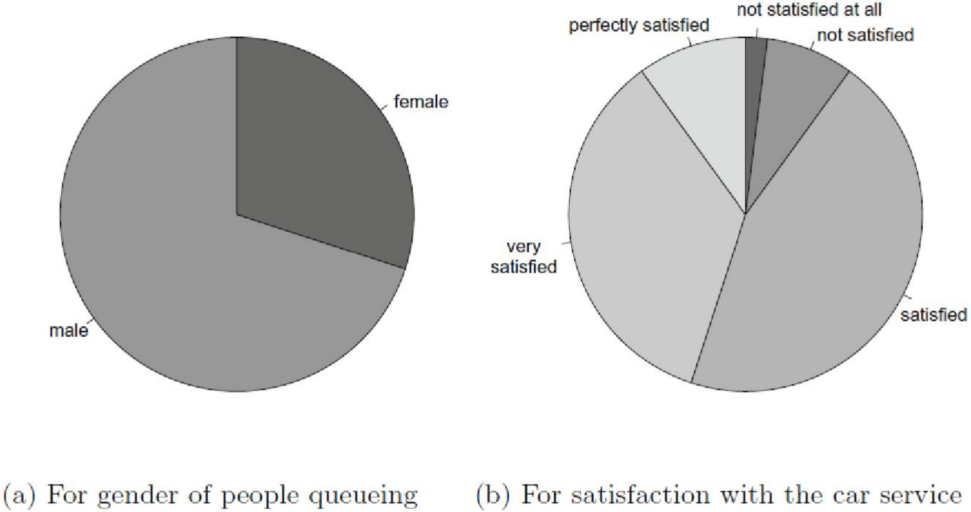
A **bar chart** can be used for nominal and ordinal variables, as long as the number of categories is not very large. It consists of one bar for each category.



(a) Categories next to each other

(b) Categories stacked

Pie chart is another option to visualize the absolute and relative frequencies of nominal and ordinal variables. This is a circle partitioned into segments, where each of the segments represents a category. The size of each segment depends upon the relative frequency and is determined by the angle.



(a) For gender of people queueing

(b) For satisfaction with the car service