Probability and Statistics

Raluca Muresan

West University of Timisoara

raluca.muresan@e-uvt.ro

Overview

Statistics I

"Probability theory is the vehicle of Statistics"

If it weren't for the laws of probability, statistics wouldn't be possible.

To illustrate the **difference** between probabilities and statistics, let us consider two boxes: a probabilistic and a statistical one. For the probabilistic box we know that it contains 5 white, 5 black and 5 red balls; the probabilistic problem is that if we take a ball, what is the chance that it is white? For a statistical box we do not know the combination of balls in the box. We consider a sample and from this sample we conjecture what we think the box contains.

Statistics II

What have we learned so far:

- to analyze problems and systems involving uncertainty,
- to find probabilities, expectations, and other characteristics for a variety of situations,
- to produce forecasts that may lead to important decisions.

What we need to know for this: the **distribution and its parameters**.

In practice, the parameters or even the distribution are usually not known.

Solution: **collect data (descriptive statistics)** and try to deduce the parameters using different statistical methods **(statistical inference)**.

The most popular statistical softwares are: R, Minitab, SAS (Statistical analysis system), SPSS (statistics package for social sciences), etc.

Statistics III

Descriptive statistics

- deals with the collection, classification and presentation of numerical data.

Inferential statistics

- deals with the interpretation of the data and their use in formulating conclusions and taking decisions.

What happens next?

What will we learn:

- to visualize data, understand the patterns
- to characterize this behavior in quantities
- to estimate the distribution parameters
- to test statements about parameters and the entire system
- to understand relations among variables
- to fit suitable models and use them to make forecasts

Preliminary notions I

Definitions

A **population** consists of all units/items of interest.

Example: the students at Computer Science, 2nd year; all possible hands at poker; all the stars in our galaxy; all voters in Timiș county; all Romanians that play tennis; all the computers produced by a company; daily maximum temperatures in the month of July

We are interested in a specific **characteristic** of the population.

Example: the height/weight/grade at PS for the students at Computer Science; the preference for candidate A; the level of education of voters that prefer candidate A; the age of players/ how many times a month they play tennis; memory space for the computers.

Preliminary notions II

A **sample** consists of observed units collected from the population; it is a subset of the population.

Example: Students in a group; voters in Timișoara; tennis players of ages 18-35; computers assembled in a specific day;

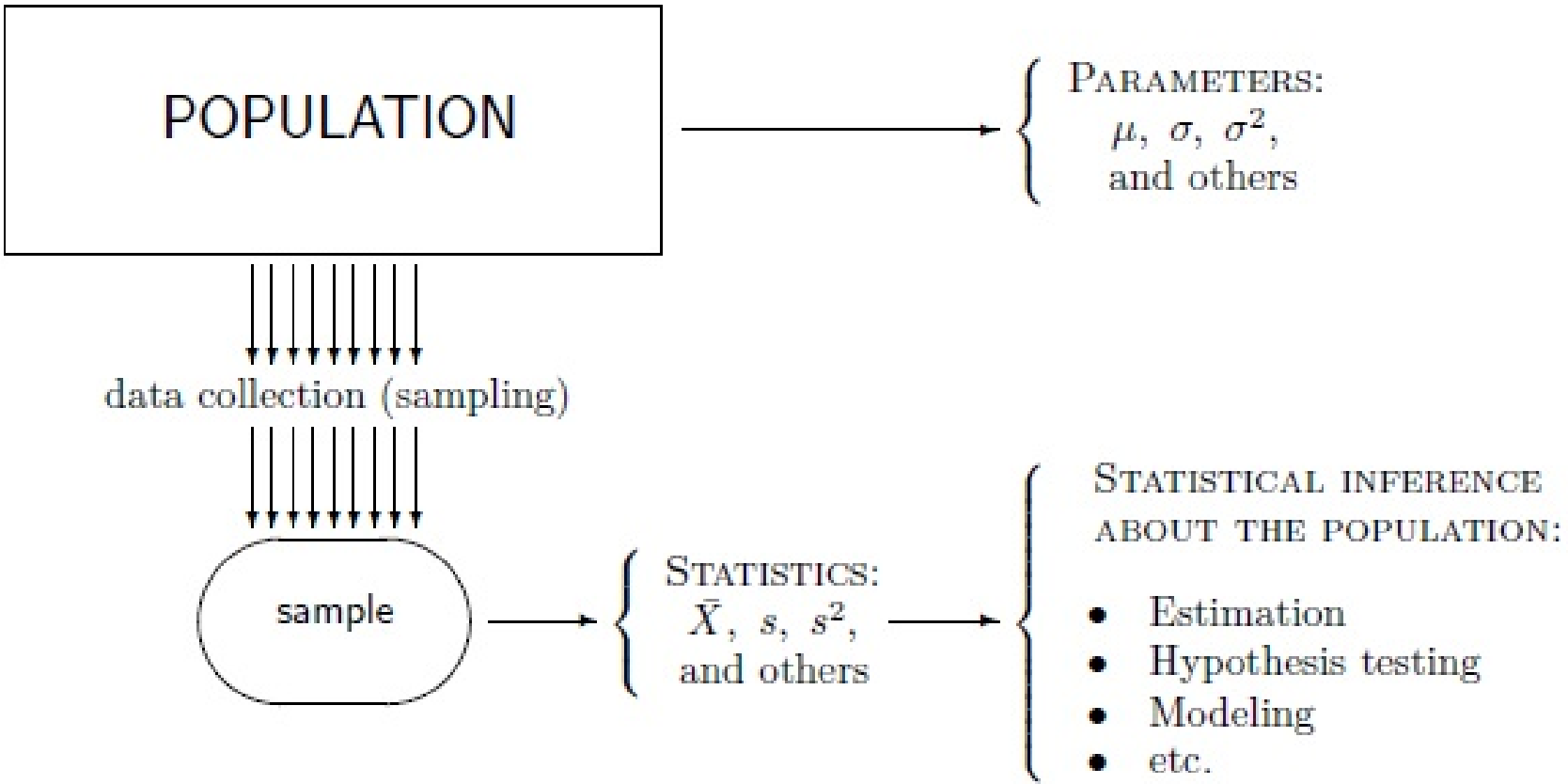
Any function of a sample is called **statistic**.

Example: the average height/weight/grade of the students; the percentage of students that pass the PS exam; the average age of the voters for candidate A; the dispersion of ages of tennis players.

Census vs sampling - what is a census and can we always perform such a census?

Preliminary notions III

The parameters of a population are denoted using Greek letters: μ, σ^2, σ. The estimators of these parameters (determined using the sample and usually being statistics) are denoted using Roman letters: m, s^2, s.



Descriptive Statistics I

Suppose a random sample has been collected:

$$S = (X_1, X_2, \ldots, X_n).$$

Example 1. To evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time for n = 30 randomly chosen jobs (in seconds)

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

What information do we get from this collection of numbers?

- population: processors considered (produced by a company)
- characteristic of interest: CPU time for a task

Descriptive Statistics II

We assign a random variable X to this characteristic.

What is the distribution of X and its parameters?

Objective 1: Compute descriptive statistics measuring the location, spread, position and other characteristics like

- **mean** - measures the average value of a sample;
- **median** - measures the central value;
- **quantiles and quartiles** - show where certain portions of a sample are located;
- **variance, standard deviation, and interquartile range** - measure variability and spread of data.

Objective 2: Determine graphically the distribution of X.

Characteristics of central tendency I

These are the **mean, median, mode**.

The **sample mean** $\bar{X}$ estimates the population mean $\mu = E(X)$.

Definition

Sample mean $\bar{X}$ is the arithmetic average of the values in the sample:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Properties: the sample mean is **unbiased, consistent, and asymptotically Normal**.

This is true if the population has finite mean and variance.

Characteristics of central tendency II

An estimator $\hat{\theta}$ is **unbiased** for a parameter θ if its expectation equals the parameter

$$E(\hat{\theta}) = \theta$$

for all possible values of θ.

Prove that the sample mean $\bar{X}$ is unbiased.
The **bias** of $\hat{\theta}$ is $E(\hat{\theta} - \theta)$.

An estimator $\hat{\theta}$ is **consistent** for a parameter θ if the probability of its sampling error of any magnitude converges to 0 as the sample size increases to infinity,

$$P(|\hat{\theta} - \theta| > \varepsilon) \to 0, n \to \infty, \forall \varepsilon > 0.$$

The sample mean $\bar{X}$ is consistent (prove it using Chebyshev's Inequality).

Characteristics of central tendency III

The **asymptotic Normality** of the sample mean $\bar{X}$ is deduced using the **Central Limit Theorem**.

Therefore the statistic

$$Z = \frac{\bar{X} - E(\bar{X})}{Std(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

converges to Standard Normal as $n \to \infty$.

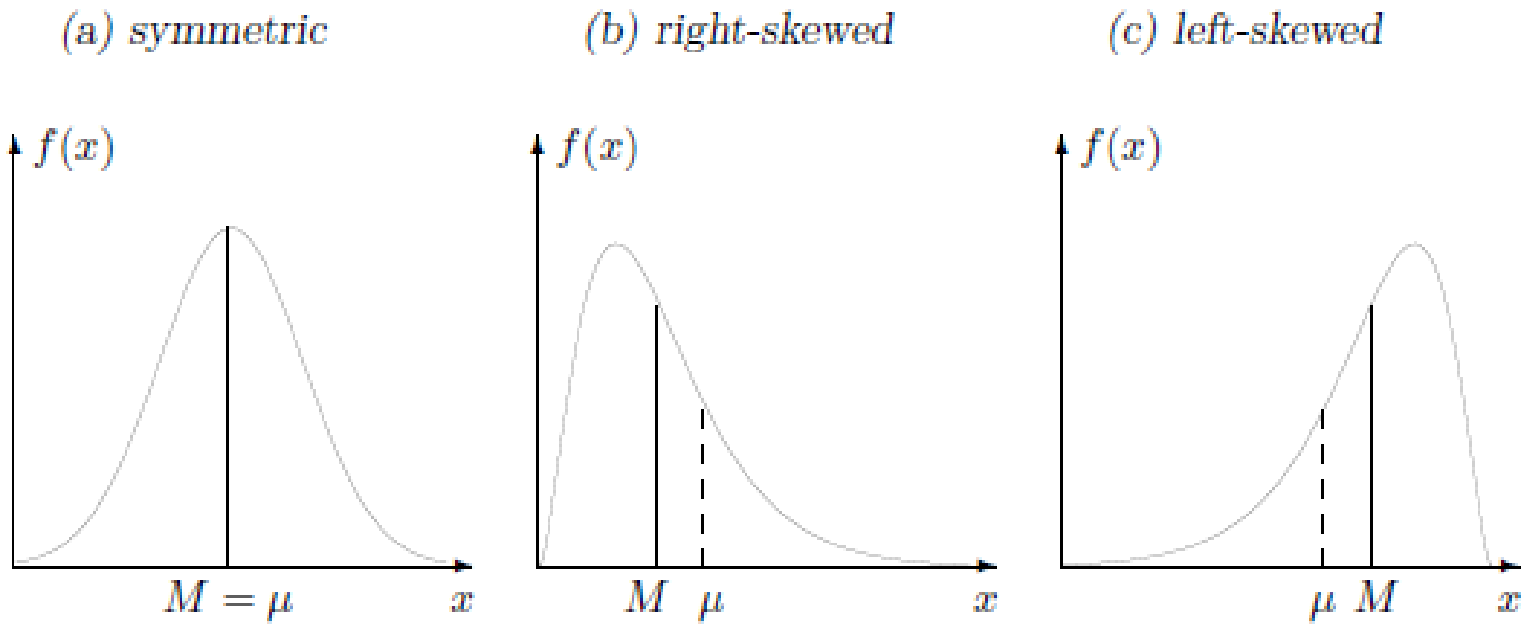
Example 2. Compute the sample mean for the CPU data in Example 1.

Characteristics of central tendency IV

One disadvantage of a sample mean is its **sensitivity to extreme observations**.

Another simple measure of location is a **sample median**, which estimates the population median.

It is much less sensitive than the sample mean.



Characteristics of central tendency V

Definition

Sample median $\hat{M}$ is a number that is exceeded by at most a half of observations and is preceded by at most a half of observations.

Computing the sample median $\hat{M}$:

- if n (the sample size) is odd, then $(\frac{n+1}{2})$-th smallest observation is a median;
- if n is even, then the average between the $(\frac{n}{2})$ and $(\frac{n}{2}+1)$-th smallest observations is the median.

Example 3. Compute the median for the CPU data in Example 1.

Characteristics of central tendency VI

Definition

The mode of a set of data values is the value that appears most often.

The mode is not necessarily unique.

In symmetric unimodal distributions, such as the normal distribution, the mean, median and mode all coincide.

Except for extremely small samples, the mode is insensitive to "outliers".

Quantiles I

Definition

A p-**quantile** of a population is such a number x that solves the inequations

$$P(X \leq x) \leq p$$

$$P(X \geq x) \leq 1 - p,$$

where $0 < p < 1$.

A **sample** p-**quantile** is any number that exceeds at most $100p\%$ of the sample, and is exceeded by at most $100(1-p)\%$ of the sample.

A γ-**percentile** is (0.01γ)-quantile.

Quantiles II

First, second, and third quartiles are the 25th, 50th, and 75th percentiles. They split a population or a sample into four equal parts.

The **median** is at the same time a 0.5-quantile, 50th percentile, and 2nd quartile.

NOTATION

q_p	=	population p-quantile
$\hat{q}_p$	=	sample p-quantile, estimator of q_p
π_γ	=	population γ-percentile
$\hat{\pi}_\gamma$	=	sample γ-percentile, estimator of π_γ
$Q_1,\ Q_2,\ Q_3$	=	population quartiles
$\hat{Q}_1,\ \hat{Q}_2,\ \hat{Q}_3$	=	sample quartiles, estimators of $Q_1,\ Q_2,$ and Q_3
M	=	population median
$\hat{M}$	=	sample median, estimator of M

Quantiles, quartiles, and percentiles are related as follows.

Quantiles III

Quantiles, quartiles, percentiles

$$q_p = \pi_{100p}$$

$$Q_1 = \pi_{25} = q_{1/4} \qquad Q_3 = \pi_{75} = q_{3/4}$$

$$M = Q_2 = \pi_{50} = q_{1/2}$$

Example 4. Compute the 1st and 3rd quartiles of CPU data, as well as the 95% percentile.

Example 5. Compute the 90%, 95% and 99% percentiles for the Standard Normal distribution.

Characteristics of dispersion I

Next we are going to measure variability of our variable (data), i.e. how much the actual value can differ from its expectation.

Definition

For a sample $S = (X_1, X_2, \ldots, X_n)$, a **sample variance** is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

It is an estimate for the population variance $\sigma^2 = Var(X)$.

Characteristics of dispersion II

The **sample standard deviation** is a square root of the sample variance,

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

It estimates the population standard deviation $\sigma = Std(X)$.

The sample variance can be computed as follows

$$s^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)$$

Example 6. Compute the sample variance and standard deviation for the CPU data.

Characteristics of dispersion III

Sample mean, variance, and standard deviation are sensitive to outliers.

To detect and identify outliers, we need measures of variability that are not very sensitive to them.

Such a measure is the **interquartile range**.

Definition

An **interquartile range** is defined as the difference between the first and the third quartiles,

$$IQR = Q_3 - Q_1.$$

IQR is estimated by the **sample interquartile range**

$$\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1.$$

Characteristics of dispersion IV

A "rule of thumb" for identifying outliers is the **rule of** $1.5 \times IQR$. The values in the sample that are smaller than $\hat{Q}_1 - 1.5\widehat{IQR}$ or larger than $\hat{Q}_3 + 1.5\widehat{IQR}$ are viewed as outliers.

Example 7. Compute the IQR for the CPU data and find out if there are possible outliers in the sample.

Other measures of dispersion: **range, coefficient of variation**.

The **range** of a sample is the difference between the maximum and the minimum value.

Characteristics of dispersion V

The **coefficient of variation** of a sample is the ratio between the sample mean and the standard deviation,

$$\hat{c} = \frac{s}{|\bar{X}|} \cdot 100.$$

It represents a standardized measure of dispersion.

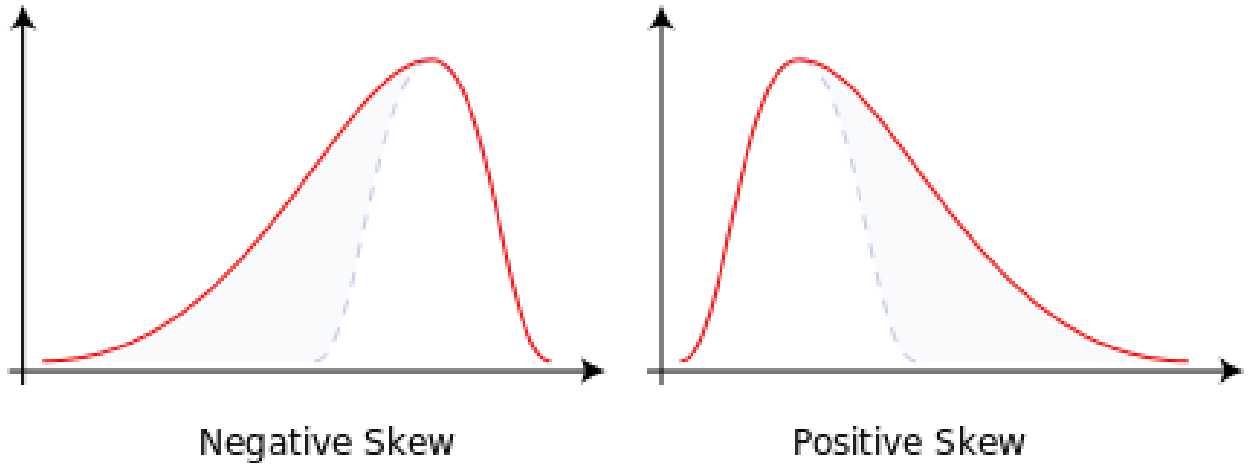
Skewness

The **skewness** is a measure of asymmetry of the distribution of the sample. It can be positive or negative, or undefined.

Definition

The **sample skewness** is defined as the ratio

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\sqrt{\frac{1}{n-1}(X_i - \bar{X})^2}^3}$$



Kurtosis I

Kurtosis is a measure of the "tailedness" of the probability distribution.

For this measure, higher kurtosis is the result of infrequent extreme deviations (or outliers), as opposed to frequent modestly sized deviations.

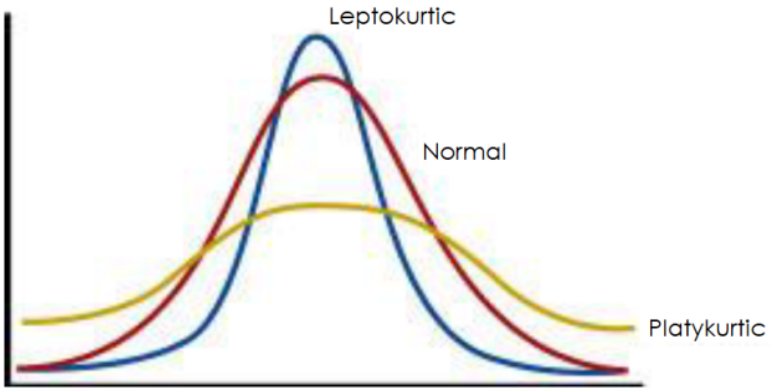
The kurtosis of any univariate normal distribution is 3. Distributions with kurtosis less than 3 are said to be **platykurtic**, which means the distribution produces fewer and less extreme outliers than does the normal distribution. Distributions with kurtosis greater than 3 are said to be **leptokurtic**.

Kurtosis II

Definition

The **sample kurtosis** is defined as the ratio

$$k = \frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{\sqrt{\frac{1}{n-1}(X_i - \bar{X})^2}^4}$$



Example 8. Compute the skewness and kurtosis for the CPU data.

Exercises

Exercise 1. Solve Ex. 8.9 page 235 from the textbook.

Exercise 2. Solve Ex. 8.2 page 234 from the textbook.

Preliminaries I

Before you do anything with a data set, look at it!

A quick look at a sample may clearly suggest

- a family of distributions to be used;
- statistical methods suitable for the given data;
- presence or absence of outliers;
- presence or absence of heterogeneity;
- existence of time trends and other patterns;
- relation between two or several variables.

Preliminaries II

Types of statistical plots:

- histograms
- stem-and-leaf plots
- boxplots
- time plots
- scatter plots

Histogram I

A **histogram** shows the shape of a pmf or a pdf of data, checks for homogeneity, and suggests possible outliers.

To construct a histogram, we split the range of data into equal intervals, "bins," and count how many observations fall into each bin.

There are two types of histograms:

- frequency histogram - consists of columns, one for each bin, whose height is determined by the number of observations in the bin.
- relative frequency histogram - column heights represent the proportion of all data that appeared in each bin.

Histogram II

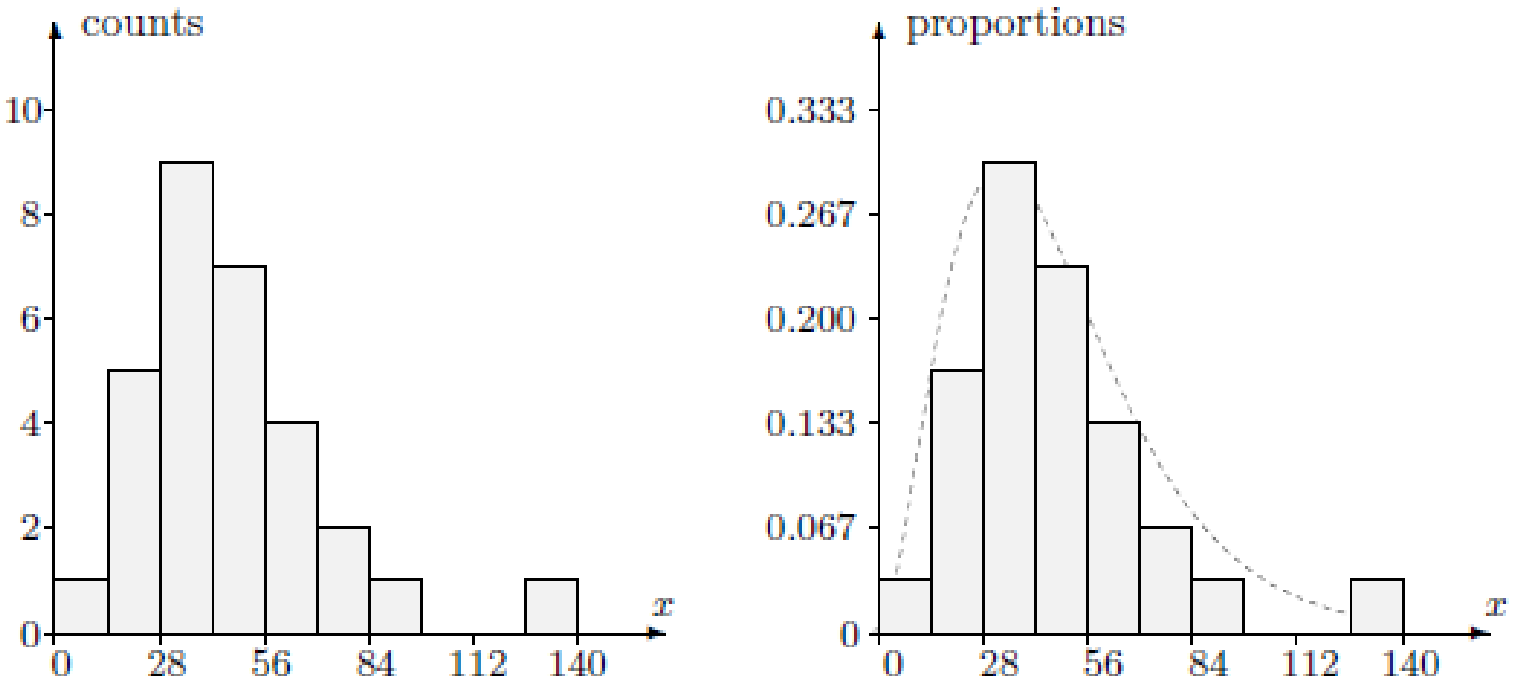
Example 1. Construct the histogram for the CPU data given at Example 1 in the previous lecture. Use bins of length 14.
Example 2. Construct the relative frequency histogram for the same data keeping the same bins.

What information can we draw from these histograms?

Histograms have a shape similar to the pmf or pdf of data, especially in large samples.

What other information can be deduced: outliers, symmetry, homogenity.

Histogram III

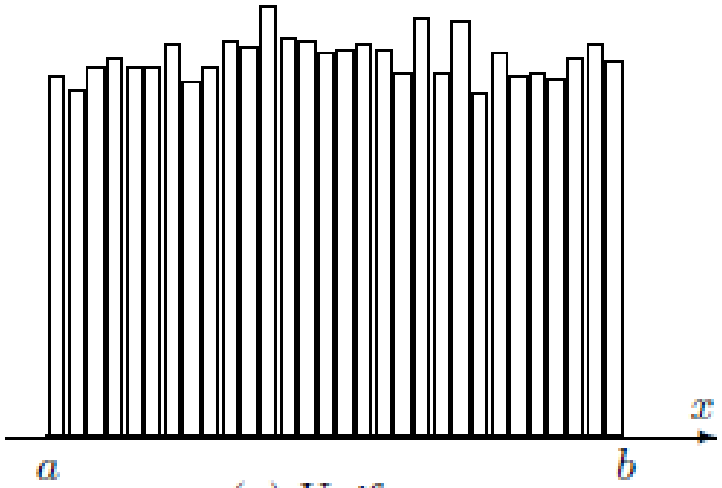


(a) Frequency histogram

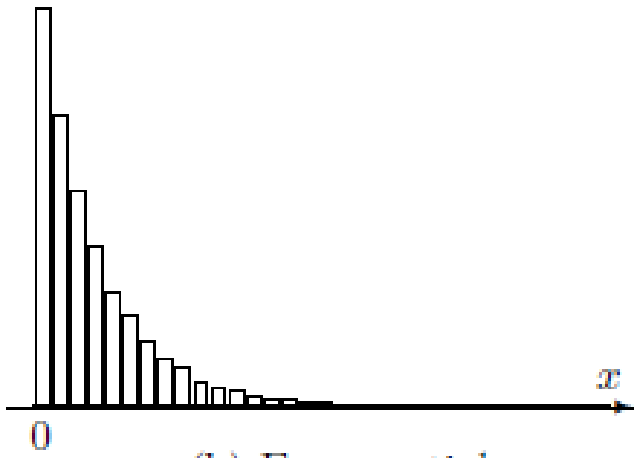
(b) Relative frequency histogram

FIGURE 8.6: *Histograms of CPU data.*

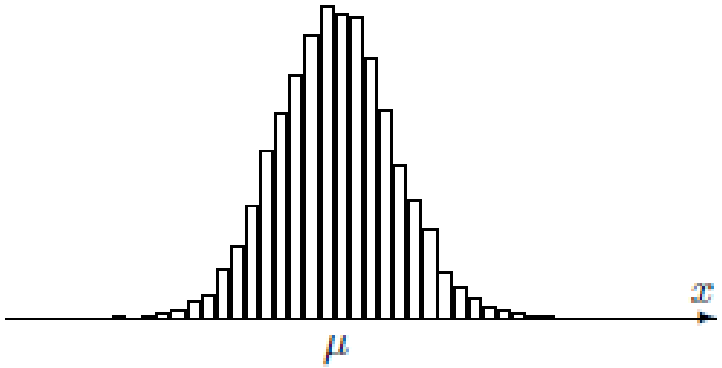
Histogram IV



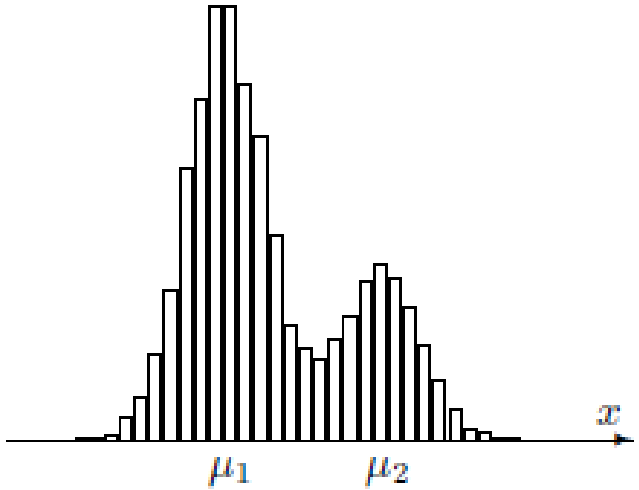
(a) Uniform



(b) Exponential



(c) Normal



(d) Mixture

The choice of bins

Experimenting with histograms, you can notice that their shape may depend on the choice of bins.
Rules of thumb about a good choice of bins, but in general,

- there should not be too few or too many bins;
- their number may increase with a sample size;
- they should be chosen to make the histogram informative, so that we can see shapes, outliers, etc.

Stem-and-leaf plot I

Stem-and-leaf plots are similar to histograms although they carry more information. Namely, they also show how the data are distributed within columns.

The stem: first one or several digits of the values

The leaf: the next digit form the leaf

Example: 239 can be written as 23|9 (23 is the stem and 9 is the leaf) or as 2|3 (2 is the stem and 3 is the leaf, 9 is dropped here).

In the first case, the leaf unit equals 1 while in the second case, the leaf unit is 10, showing that the (rounded) number is not 23 but 230.

Stem-and-leaf plot II

Example 2. Construct the stem-and-leaf plot for the CPU data such that the last digit is the leaf.

Turning this plot by 90 degrees counterclockwise, we get a histogram with 10-unit bins (because each stem unit equals 10).

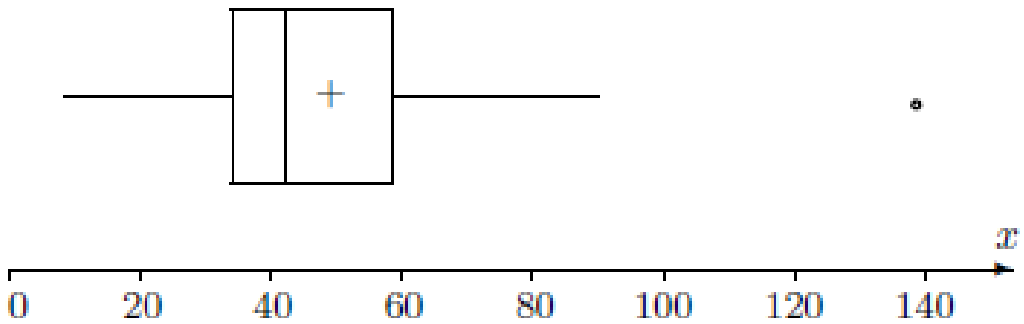
Boxplot I

The main descriptive statistics of a sample can be represented graphically by a **boxplot**.

To construct a boxplot, we need the **five-point summary**

$$\text{five point summary} = (minX_i, Q_1, M, Q_3, maxX_i)$$

Example 3. Construct the boxplot of the CPU data. Determine the five-point summary first.



Scatter plot I

Scatter plots are used to see and understand a relationship between two variables.

These can be temperature and humidity, experience and salary, age of a network and its speed, number of servers and the expected response time, and so on.

Example 4 (Antivirus maintenance). Protection of a personal computer largely depends on the frequency of running antivirus software on it. One can set to run it every day, once a week, once a month, etc.
During a scheduled maintenance of computer facilities, a computer manager records the number of times the antivirus software was launched on each computer during 1 month (variable X) and the number of detected worms (variable Y). The data for 30 computers are in the table.

Scatter plot II

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

Is there a connection between the frequency of running antivirus software and the number of worms in the system? Draw a scatter plot of the data.

When we study **time trends** and development of variables over time, we use **time plots**. These are scatter plots with x-variable representing time.

Exercise 8.1 on page 233.

The End