

# Mathematical Statistics

MAS 713

Chapter 1.2

## Previous subchapter

- 1 What is statistics ?
- 2 The statistical process
- 3 Population and samples
- 4 Random sampling

**Questions?**

# This subchapter

## **Descriptive Statistics**

- 1.2.1 Introduction
- 1.2.2 Types of variables
- 1.2.3 Graphical representations
- 1.2.4 Descriptive measures

# Introduction

As we said, **statistics is the art of learning from data**

However, statistical data, obtained from surveys, experiments, or any series of measurements, are often so numerous that they are **virtually useless** unless they are condensed

~> **data should be presented in ways that facilitate their interpretation and subsequent analysis**

Naturally enough, the aspect of statistics which deals with **organising, describing** and **summarising data** is called **descriptive statistics**

# Types of variables

# Types of variables

The range of available descriptive tools for a given data set depends on the **type** of the considered variable

There are essentially two types of variables :

- 1 **categorical** (or qualitative) variables : take a value that is one of several possible categories (no numerical meaning)

**Ex.** : gender, hair color, field of study, political affiliation, status, ...

- 2 **numerical** (or quantitative) variables : naturally measured as a number for which meaningful arithmetic operations make sense

**Ex.** : height, age, temperature, pressure, salary, area, ...

## Types of variables

**Attention :** sometimes categorical variables are disguised as quantitative variables

For example, one might record gender information coded as

$$0 = \textit{Male}, \quad 1 = \textit{Female}.$$

It remains a categorical variable, it is not naturally measured as a number

## Types of variables

Within these two main variable types, there are also sub-types :

- categorical variables

- ▶ **ordinal** : there is a clear ordering of the categories

**Ex.:** salary class (low, medium, high), opinion (disagree, neutral, agree), etc.

- ▶ **nominal** : there is no intrinsic ordering to the categories

**Ex.:** gender, hair colour, etc.

- numerical variables

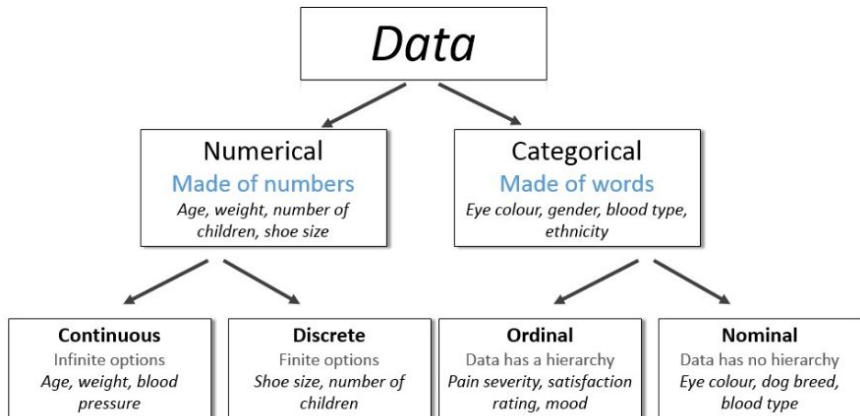
- ▶ **discrete** : the variable can only take a finite (or countable) number of distinct values

**Ex.:** number of courses you are enrolled in, number of persons in a household, etc.

- ▶ **continuous** : the variable can take any value in an entire interval on the real line (uncountable)

**Ex.:** height, weight, temperature, time to complete a task, etc.





# Descriptive statistics tools

Essentially, descriptive statistics tools consist of

- 1 graphical methods and
- 2 descriptive measures

# Graphical representations

# Graphical representations

*A picture is worth a thousand words*

~> graphical representations are often the most effective way to quickly obtain a feel for the essential characteristics of the data

## Fact

Any good statistical analysis of data should always begin with **plotting the data**

# Graphical representations

Plots often reveal useful information and opens paths of inquiry

They might also highlight the presence of irregularities or unusual observations (“outliers”)

# Dotplot

# Dotplot

A **dotplot** is an attractive summary of **numerical** data when the data set is reasonably small (few distinct data values)

Each observation is represented by a dot above the corresponding location on a horizontal measurement scale

When a value occurs more than one time, there is a dot for each occurrence and these dots are stacked vertically

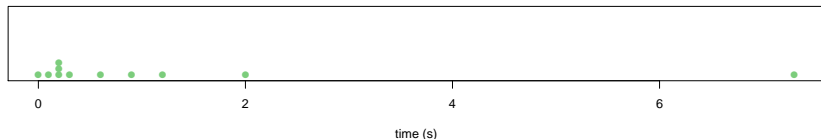
## Dotplot : example

### Example

In 1987, for the first time, physicists observed neutrinos from a supernova that occurred outside of our solar system. At a site in Kamiokande, Japan, the following times (in seconds) between neutrinos were recorded :

0.107; 0.196; 0.021; 0.283; 0.196; 0.854; 0.58; 0.196; 7.3; 1.18; 2

Draw a dotplot :



Note that the largest observation is extremely different to the others. Such an observation is called an **outlier**

Usually, outliers merit further attention (recording error? missed observations? real observation?) Important questions may hinge on the correct interpretation of an outlier



# Stem-and-leaf plot

## Stem-and-leaf plot

The dotplot is useful for small samples, up to (say) about 20 observations. However, when the number of observations is moderately large, other graphical display may be more useful

A **stem-and-leaf plot** (or just **stemplot**) is another effective way to organise **numerical** data without much effort

Idea : separate each observation into a **stem** (all but last digit) and a **leaf** (final digit)

Example :      24 := 2|4      139 := 13|9      5 := 0|5

Write all unique stems in vertical column with the smallest at the top, and draw a vertical line at the right of this column

Write each leaf in the row to the right of its stem, in increasing order out from the stem

## Stem-and-leaf plot : example

### Example 1.5

Study of use of alcohol by university students. We are interested in a variable  $X$ , the **percentage of undergraduate students who are binge drinkers**. We observe  $X$  on 140 campuses across the US

The sample is :

26	57	66	66	41	46	65	35	46	38	44	29	43				
14	11	68	37	27	18	46	30	32	35	59	39	32	31	39	21	58
65	50	44	29	53	27	38	52	29	58	45	34	36	56	47	22	59
46	24	51	26	39	23	55	50	42	18	48	64	44	46	66	33	61
38	35	22	57	42	42	26	47	67	37	39	58	26	41	61	51	61
56	48	53	13	28	52	36	62	31	38	42	64	51	54	33	19	25
42	37	36	55	37	56	43	28	56	49	39	57	48	52	60	17	49
61	44	18	67	36	58	47	16	33	27	29	48	45	34	57	56	48
46	49	15	52	04	41	64	37									

(Source : "Health and Behavioral Consequences of Binge Drinking in College", J. Amer. Med. Assoc., 1994, 1672-1677)

## Stem-and-leaf plot : example

~> you separate the tens digit (“stem”) from the ones digit (“leaf”)

0		4
1		1345678889
2		1223456666777889999
3		011223334455566667777888899999
4		1112222233444455666666777888889999
5		00111222233455666667777888899
6		01111244455666778

~> this stemplot suggests that

- a **typical value** is in the stem 4 row (probably in mid-40% range)
- there is a single peak, but the shape is **not perfectly symmetric**
- there is no observations unusually far from the bulk of the data (**no outliers**)
- at most universities in the sample, **at least one-quarter** of the students are binge drinkers

## Stem-and-leaf plot

The stemplot conveys information about the following aspects of the data :

- identification of a typical value
- extent of spread about the typical values
- presence of any gaps in the data
- extent of symmetry in the distribution values
- number and location of peaks
- presence of any outlying values

## Stem-and-leaf plot : variations

There are many ways in which stem-and-leaf plots can be modified to meet particular needs :

- **rounding** or **truncating** the numbers to a few digits before making a stemplot to avoid too much irrelevant detail in the stems
- **splitting each stem** to give greater detail in distribution
- **back-to-back** stemplots with common stems to compare two related distributions

## Stem-and-leaf plot : splitting each stem

A more informative display can be created by repeating each stem value twice, once for the low leaves 0, 1, 2, 3, 4 and again for the high leaves 5, 6, 7, 8, 9

For the binge-drinking data this yields (compare Slide 20)

0	4
0	
1	1 3 4
1	5 6 7 8 8 8 9
2	1 2 2 3 4
2	5 6 6 6 6 7 7 7 8 8 9 9 9 9
3	0 1 1 2 2 3 3 3 4 4
3	5 5 5 6 6 6 6 7 7 7 7 7 8 8 8 8 8 9 9 9 9 9 9
4	1 1 1 2 2 2 2 2 3 3 4 4 4 4
4	5 5 6 6 6 6 6 6 7 7 7 7 8 8 8 8 8 8 9 9 9
5	0 0 1 1 1 2 2 2 2 3 3 4
5	5 5 6 6 6 6 6 6 7 7 7 7 8 8 8 8 8 9 9
6	0 1 1 1 1 2 4 4 4
6	5 5 6 6 6 7 7 8

## Stem-and-leaf plot : back-to-back plots

Suppose you have two data sets, each consisting of observations on the same variable

For instance, exam scores for two different classes

→ in what ways are the two data sets similar ? how do they differ ?

→ comparative stem-and-leaf plot, or back-to-back stemplots

The leaves for one data set are listed to the right of the stems and the leaves for the other to the left

	2588		5		9
	2234578		6		01445
<b>Example:</b> exam scores	0225556689		7		1223567
	4479		8		01334578
			9		156688

→ two sides are similar, except that the right side appears to be shifted up one row from the other side



# Frequency distribution, bar charts and histograms

## Frequency distribution, bar charts and histograms

A natural continuation of the stemplot is to count the number of observations (that is, the **frequency**) in each row

↪ this can be done each time that proper “categories” are available

The frequency of observations in each category then forms the **frequency distribution**

Which categories ?

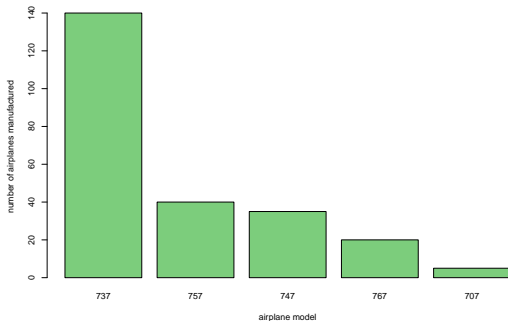
- for a categorical variable, the categories are obviously defined
- ↪ count the frequency of observations in each category, mark each category on a horizontal axis and draw rectangles whose heights are the corresponding frequencies.
- for a discrete numerical variable, the categories are given by the distinct values taken by the variable
- ↪ then, proceed as above. This is however called a **histogram**

## Bar chart : example

### Example

In 1985 the Boeing Company published the figures for its production of transport aircraft. That year, they produced 5 Boeing's 707, 140 Boeing's 737, 35 Boeing's 747, 40 Boeing's 757 and 20 Boeing's 767

The frequency distribution for this production is presented in the following bar chart



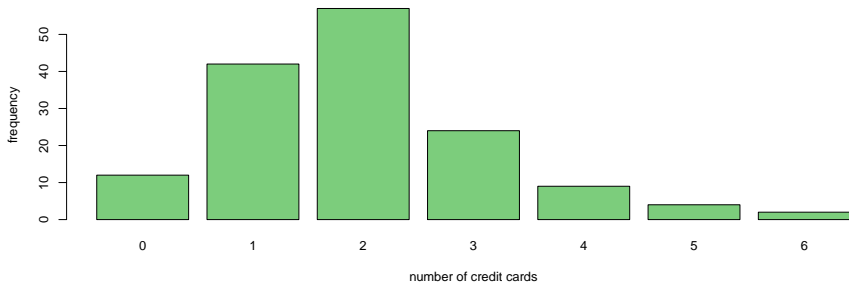
# Histogram : example

## Example

A sample of students from a statistics class were asked how many credit cards they carry. For 150 students, the frequency distribution is given below

Number of cards	0	1	2	3	4	5	6
Frequency	12	42	57	24	9	4	2

Represent the histogram



## Histogram for continuous numerical variable

If the variable is numerical continuous, there are no obvious categories

→ we have to create some categories, called **classes**, which will be intervals that **do not overlap** and **accommodate all observations**

Once the classes have been defined, we can proceed similarly to above, that is :

**Determine the frequency** of observations in each class,  
**mark the class boundaries** on a horizontal axis and  
**draw rectangles** whose heights are the corresponding frequencies

However, important practical questions arise, like how many classes to use and what are the limits for each class

## Histogram for continuous numerical variable

Generally speaking, this will depend on the total number of observations and the range of the data

Actually, you should find a trade-off between

1. choosing too few classes at a cost of losing information about actual values
2. choosing too many classes, which will result in the frequencies of each class to be too small for a pattern to be discernible

An empirical rule is

$$\text{number of classes} \simeq \sqrt{\text{number of observations}}$$

↪ between 5 and 20 classes will be satisfactory for most data sets

Besides, it is common, although not essential, to choose classes of equal width

## Histogram : example 1.8

### Example

Power companies need information about customer usage to obtain accurate forecasts of demand. Here we consider the energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes

The sample is

```
10.04 13.47 13.43 9.07 11.43 12.31 4.00 9.84 10.28 8.29
6.94 10.35 12.91 10.49 9.52 12.62 11.09 6.85 15.24 18.26
11.21 11.12 10.28 8.37 7.15 9.37 9.82 9.76 8.00 10.21
6.62 12.69 13.38 7.23 6.35 5.56 5.98 6.78 7.73 9.43 9.27
8.67 15.12 11.70 5.94 11.29 7.69 10.64 12.71 9.96 13.60
16.06 7.62 2.97 11.70 13.96 8.81 12.92 12.19 16.90 9.60
9.83 8.26 8.69 6.80 9.58 8.54 7.87 9.83 10.30 8.61 7.93
13.11 7.62 10.95 13.42 6.72 10.36 12.16 10.40 5.20 10.50
8.58 14.24 14.35 8.47 7.29 12.28 11.62 7.16
```

## Histogram : example

The data set contains 90 observations, and since  $\sqrt{90} \simeq 9.48$ , we suspect that about nine classes will provide a satisfactory frequency distribution

The smallest and largest data values are 2.97 and 18.26, so the classes must cover a range of at least 15.29 BTU

→ as  $15.29/9 \simeq 1.7$ , we take the common class width equal to 2 (for simplicity), and we start at 2 (again for simplicity)

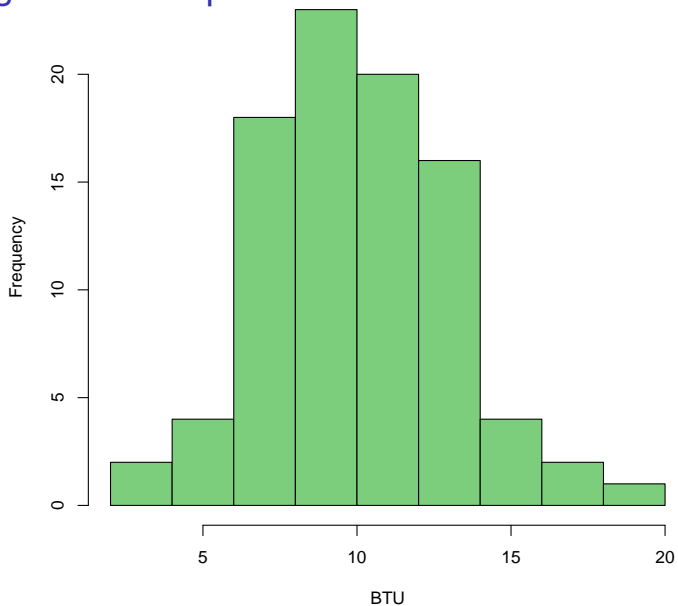
Counting the frequencies in the so-defined classes, we get

[2, 4)	[4, 6)	[6, 8)	[8, 10)	[10, 12)	[12, 14)	[14, 16)	[16, 18)	[18, 20)
2	4	18	23	20	16	4	2	1

**Note :** we adopt the left-end inclusion convention, i.e. a class contains its left-end but not its right-end boundary point (interval  $[a, b)$ , left-close right-open)



# Histogram : example



## Histogram : comments

Unlike the histogram for discrete numerical variables, the histogram for continuous numerical variables consists of **adjacent rectangles**. This precisely renders the **continuity** of the underlying variable

Like the stemplot, the histogram (discrete or continuous) provides a **visual impression of the shape of the distribution** of the observations, as well as information about the **central tendency** and **dispersion** in the data, that are not immediately apparent from the data themselves

# Histogram : comments

## Typical words/phrases used to describe histograms :

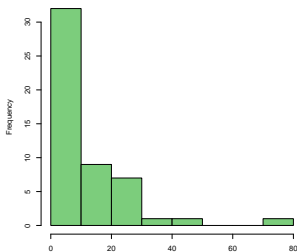
- **symmetric**, or **skewed** to the right/left;
- **unimodal**, or bimodal/multimodal;
- **bell-shaped** (if symmetric & unimodal);
- there are possible **outliers** around..., or there are no outliers;
- **typical value** of the data is..., the **range** of the data is...

## Unequal class width

Sometimes, equal-width classes may not be a sensible choice

For instance, if the data have several extreme observations or outliers, nearly all observations will fall in just a few of the classes

Consider the following histogram :



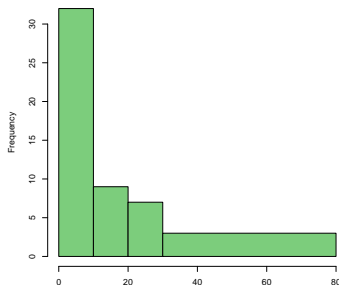
→ the last 5 classes together have only 3 observations!

→ it is preferable to regroup the data in the right-hand tail into a wider class

## Unequal class width

However, **wider classes are likely to include more observations than narrower ones!**

~> when class widths are unequal, the frequencies will give a distorted representation of the reality



~> the **rectangles areas** (not their heights) should be proportional to the frequencies

~> this is what a **density histogram** achieves

## Density histogram

The **relative frequency** of a class is the proportion of observations in that class, that is, the frequency of the class divided by the total number of observations

We call the **density** of a class the relative frequency of the class divided by the class width

A **density histogram** is a histogram whose rectangle heights are given by the densities of each class

So we have

$$\begin{aligned}\text{relative frequency} &= \text{density} \times \text{class width} \\ &= \text{rectangle height} \times \text{rectangle width} \\ &= \text{rectangle area}\end{aligned}$$

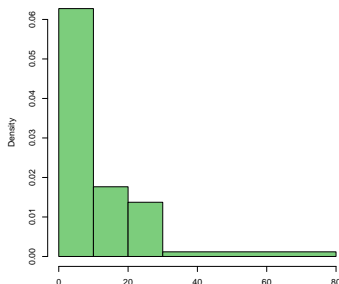
## Density histogram

Besides, the **total rectangles area must be equal to 1**, as the sum of all relative frequencies must be 1

This property will play an important role in the sequel, so that **it is always preferable to represent a density histogram** instead of a (frequency) histogram, even when the classes have equal width.

## Density histogram

For the previous case, the density histogram would be



which is much more faithful to reality than histogram on Slide 26

We can check that the area of the first rectangle is  $10 \times 0.063 = 0.63$ , that is the first class  $[10, 20)$  includes 63% of the observations

We could calculate the areas of the other rectangles the same way, and check that their sum is equal to 1



# Descriptive measures

# Descriptive measures

Dotplots, stem-and-leaf plots, histograms and density histograms summarise a data set pictorially so we can **visually** discern the overall pattern of variation

On the other hand, we might also find useful to **numerically** describe the data set

~> summary measures that tell:

- where a sample is centred (**measures of center or location**), and
- what is its extent of spread around that center (**measures of variability**)

# Descriptive measures

To proceed, we introduce the notation

$$X_1, X_2, \dots, X_n$$

for a generic sample consisting of  $n$  observations of the variable  $X$

**Note :** except when indicated otherwise, we assume that  $X$  is a numerical variable

# Measure of center

## Measure of center : the sample mean

The most frequently used **measure of center** of a sample is simply the arithmetic mean (or average) of the  $n$  observations

It is usually denoted  $\bar{x}$  and is given by

### Sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Note : the unit of  $\bar{x}$  is the same as that of  $X$

## Measure of center : the sample mean

### Example

Metabolic rate is the rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are :

1792; 1666; 1362; 1614; 1460; 1867; 1439

$$\leadsto \bar{x} = \frac{1}{7}(1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439) = 1600 \text{ (calories/24h)}$$

## Measure of center : the sample median

The **median** is another descriptive measure of the center of the sample

### Sample median

The median, usually denoted  $m$ , is the value which **divides the data into two equal parts**, half below the median and half above

↪ the sample median  $m$  is the “middlemost” value of the sample

Denote the ordered sample

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}$$

If  $n$  is odd, the median is the middle observation in the ordered data series, that is,  $m = x_{(\frac{n+1}{2})}$

If  $n$  is even, the median is *defined as* the average of the middle two observations in the ordered data series, that is

$$m = \frac{1}{2} \left( x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right)$$

## Measure of center : the sample median

### Example (ctd.)

Metabolic rate is rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are :

1792; 1666; 1362; 1614; 1460; 1867; 1439

Find the median

First, order the sample

$$1362 \leq 1439 \leq 1460 \leq 1614 \leq 1666 \leq 1792 \leq 1867$$

Here,  $n = 7$  (odd), so the median is the middle value

$$x_{(7+1)/2} = x_{(4)} = 1614 \text{ (calories/24h)}$$

**Note :** as only the order of the observations matters, the median can also be defined for an ordinal categorical variable



## Measure of center : the sample median

Sometimes, it is preferable to use the median instead of the mean, as it is resistant to outliers

### Example

A small company employs four young engineers, who each earn \$70,000, and the owner (also an engineer), who gets \$160,000. The latter claims that on the average, the company pays \$88,000 to its engineers and, hence, is a good place to work

The mean of the five salaries is indeed

$$\bar{x} = \frac{1}{5}(4 \times 70,000 + 160,000) = 88,000 \$$$

but this hardly describes the situation

On the other hand, the median is the middle observation in

$$70,000 = 70,000 = 70,000 = 70,000 < 160,000$$

↪ here the median= \$70,000

↪ most representative of what a young engineer earns with the firm

# Quartiles and Percentiles

## Quartiles and Percentiles

We can also divide the sample into more than two parts

~> when a sample is divided into four equal parts, the division points are called sample **quartiles**

The **first** or **lower quartile**  $q_1$  is the value that has 25% of the observations below (or equal to) it and 75% of the observations above (or equal to) it

The **third** or **upper quartile**  $q_3$  is the value that has 75% of the observations below (or equal to) it and 25% of the observations above (or equal to) it

Obviously, the second quartile  $q_2$  would split the sample into two equal halves (50% below - 50% above)

# Quartiles and Percentiles

More generally, the sample  $(100 \times p)$ th **percentile** is the value such that  $100 \times p\%$  of the observations are below this value, and the other  $100 \times (1 - p)\%$  are above this value (or equal to it)

$\leadsto q_1$  is the 25th, the median is the 50th and  $q_3$  is the 75th percentile.

## Quartiles and Percentiles

Practically,

lower quartile = median of the lower half of the data

upper quartile = median of the upper half of the data

**Note** : if  $n$  is a odd number, include the median in each half

### Example (ctd.)

Metabolic rate is rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are :

1792; 1666; 1362; 1614; 1460; 1867; 1439

Find the quartiles

The median is 1614 (calories/24h). The lower half of the observations is thus

$$1362 \leq 1439 \leq 1460 \leq 1614,$$

whose median is  $q_1 = \frac{1}{2}(1439 + 1460) = 1449.5$  (calories/24h)

Similarly, the third quartile  $q_3 = 1729$  (calories/24h)

# Five number summary

## Five number summary

Quartiles give more detailed information about location of a data set

Often, the three quartiles (i.e., including the median) together with the minimum and maximum observation give a good insight into the data set

↪ this is known as the **five number summary**

### Five number summary

$$\{x_{(1)}, q_1, m, q_3, x_{(n)}\}$$

#### Note :

- $m = q_2$  is the 50th percentile
- $q_1$  is the 25th percentile
- $q_3$  is the 75th percentile
- $x_{(1)}$  is the 0th percentile
- $x_{(n)}$  is the 100th percentile

# Measures of variability



## Measures of variability

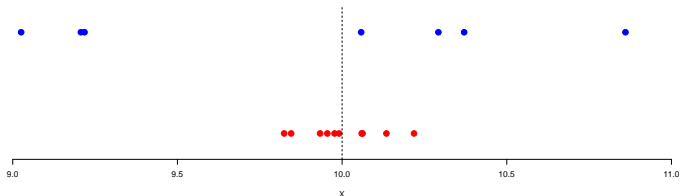
One of the most important characteristics of any data set is that the observations are not all alike

- ~> mean and median describe the location of a data set, but tell us nothing about the spread or variability of the observations
- ~> different samples may have identical measures of center, yet differ from one another in other important ways

## Measures of variability

- We observe that the dispersion of a set of observations
- is small if the values are closely bunched about their mean (red sample)
  - it is large if the values are scattered widely about their mean (blue sample)

**Notice:** both samples have mean 10



## Measures of variability

↪ it would seem reasonable to measure the variability in a data set in terms of the amounts by which the values deviate from their mean

Define the **deviations from the mean**

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$$

We might then think of using the average of those deviations as a measure of variability in the data set

Unfortunately, this will not do, because this average is always 0 :

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} = \bar{x} - \bar{x} = 0$$

↪ we need to remove the signs of the observations, so that positive and negative deviations do not counteract one another when combined

↪ **square** is a natural choice to do so

## Measure of variability : the sample variance

The most common measure of variability in a sample is the **sample variance**, usually denoted  $s^2$

The sample variance  $s^2$  is *essentially* the average of the **squared deviations** from the mean  $\bar{x}$

### Sample variance

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

See that the divisor for the sample variance is  $n - 1$ , not  $n$

↪ this is again due to the fact  $\sum_{i=1}^n (x_i - \bar{x}) \equiv 0$

↪ the value of any deviation  $(x_i - \bar{x})$  is entirely known from the other  $n - 1$  deviations, so that everything happens as if there were only  $n - 1$  independent “observations”  $(x_i - \bar{x})$

The number  $n - 1$  is the **number of degrees of freedom** for  $s^2$

## Measure of variability : the sample standard deviation

Notice that the unit of the variance is not that of the original observations :

$$\text{unit of } s^2 = (\text{unit of } X)^2$$

~> difficult to interpret

Hence, one defines the **sample standard deviation**  $s$

### Sample standard deviation

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

~> the unit of  $s$  is the same as the original unit of  $X$

~> ease of interpretation in measuring spread about the mean in the original scale

The standard deviation  $s$  has a *rough* interpretation as the average distance from an observation to the sample mean

## Measure of variability : example

### Example (ctd.)

Metabolic rate is rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are :

1792; 1666; 1362; 1614; 1460; 1867; 1439

Find the variance and the standard deviation

Here,  $n = 7$  and the sample mean is  $\bar{x} = 1600$  calories/24h. It follows

$$\begin{aligned}s^2 &= \frac{1}{6} ((1792 - 1600)^2 + \dots + (1439 - 1600)^2) \\ &= 35812 \text{ (calories/24h)}^2\end{aligned}$$

The standard deviation is  $s = \sqrt{35812} = 189.24$  calories/24h

## Measure of variability : example

**Note** : it is not difficult to see that

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

This is often useful for easing the computation of the variance

## Measure of variability : iqr

The sample variance  $s^2$  is a variability measure related to the central tendency measured by the sample mean  $\bar{x}$

→ the sample **Interquartile Range** (iqr) is a measure of variability related to the sample median and the quartiles

As the name suggests, iqr is given by the difference between the upper and the lower quartiles

### Sample Interquartile Range

$$\text{iqr} = q_3 - q_1$$

iqr describes the amount of variation in **the middle half of the observations**

It enjoys the properties of the quartiles, mainly the fact that it is **less sensitive to outliers than the sample variance**



# Detecting outliers from iqr

## Detecting outliers from iqr

We have defined an **outlier** as an observation which is too different from the bulk of the data

↪ how much different should be an observation to be an outlier?

An empirical rule is the following :

an outlier is an observation farther than  
 $1.5 \times \text{iqr}$  from the closest quartile

Besides, we say that

- an outlier is **extreme** if it is more than  $3 \times \text{iqr}$  from the nearest quartile
- it is a **mild** outlier otherwise

↪ it is important to examine the data for possible outliers as those abnormal observations may affect most of the statistical procedures

## Outliers : example

### Example

Consider the energy consumption data on Slide 31.  
Find possible outliers

**Exercise** : show that the five number summary is

$$\{2.97, 7.9475, 9.835, 12.045, 18.26\}$$

Specifically,  $q_1 = 7.9475$  (BTUs) and  $q_3 = 12.045$  (BTUs)

Hence,  $iqr = 12.045 - 7.9475 = 4.0975$  (BTUs)

The limits for not being an outlier are thus

$$[q_1 - 1.5 \times iqr, q_3 + 1.5 \times iqr] = [1.80125, 18.19125]$$

↪ only one observation (check in the data set!) falls outside that interval : the largest value 18.26 (**mild outlier**)

# Boxplots

A **boxplot** (also **box-and-whisker plot**) is a graphical display showing

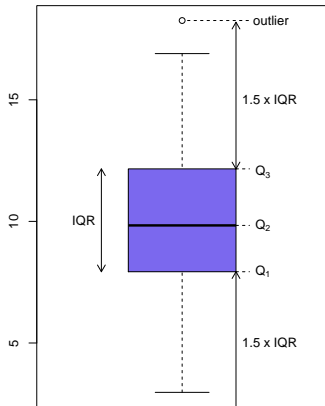
- the Five Number summary and
- any outlier value

A **central box** spans the quartiles

A line in the box marks the **median**

Lines extend from the box out to the smallest and largest observations which are not suspected outliers

Observations more than  $1.5 \times \text{IQR}$  outside the central box are plotted individually as outliers



Boxplots are very useful graphical comparisons among data sets, because they have **high visual impact** and are easy to understand

# Objectives

Now you should be able to :

- compute and interpret the sample mean, sample variance, sample standard deviation, sample median and sample quartiles
- construct and interpret visual data displays, including the stem-and-leaf plot, the histogram and the boxplot
- comment and assess the overall pattern of data from visual displays
- explain how to use the boxplots to visually compare two or more samples of data