STATISTICS

Module 2 *of*

MATHEMATICS AND STATISTICS

Marina Santacroce

Department of Decision Sciences, Bocconi

L1: Introduction to Statistics

Descriptive, inferential statistics & probability.
Descriptive statistics: Frequency tables & graphs

Outline

Introduction

Why Statistics?

Descriptive Statistics
- Frequency tables and graphs
- Grouped data, histograms, ogives and stem and leaf plots

General Information

- Teaching team:
  myself (*Instructor*) and Pieralberto Guarniero (*Instructor & T.A.*)

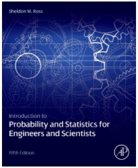
- Schedule:

Schedule		
L1-L2	05-06 feb	Descriptive statistics
L3	07 feb	Descriptive statistics Intro to R
L4-L9	12-21 feb	Elements of Probability
L10-L13	26-28 feb, 04 mar	Random variables
L14	05 mar	Random variables
L15-L18	06-13 mar	Random vectors
Partial Exam	20 mar	
L19	25 mar	Sampling statistics
L20	26 mar	Sampling statistics
L21-L22	27 mar, 01 apr	Sampling statistics
L23	02 apr	Sampling statistics
L24-L25	03-08 apr	Parameter estimation
L26	09 apr	Parameter estimation
L27	22 apr	Parameter interval estimation
L28	23 apr	Parameter interval estimation
L29-L30	24-28 apr	Parameter interval estimation
L31	29 apr	Hypothesis testing
L32	30 apr	Hypothesis testing
L33	06 mag	Hypothesis testing
L34-L35	07-08 mag	Exercises
L35	08 mag	Exercises

PIER

General Information

- My Office Hours:
  when: TUESDAYS 4 p.m.-5 p.m.
  where: Via Roentgen n.1- 3-rd floor D2-06
- Reference Book:



Sheldon M. Ross,
Introduction to Probability and Statistics for Engineers and Scientists

- Prior knowledge: Elementary calculus

Exam rules

You may choose between the following two options:

- Two partial written exams that contribute to the final grade with a 50% weight each
  - mid-term exam on March 20th
  - final exam on May 18th or June 12th
- A single general written exam (after the end of the course) that counts for 100% of the final mark.

- The tests consist of exercises. They aim at ascertaining students' mastery of concepts and results discussed during lectures as well as an adequate knowledge of R.
- In each test the maximum grade is 31 and questions related to R count 4 points (of the 31).

The assessment method is the same for both attending and non-attending students.

Students who take the mid-term exam may still take the written general exam instead of taking the final exam.

Access to the final (or second partial) exam follows the rules indicated in *Section 7.6 of the Guide to the University*

Why Statistics?

Sh. M. Ross "*It is become accepted in today's world that in order to learn about something, you must first collect data. Statistics is the art of learning from data*."

⇒ Statistics is concerned with data collection, data description, data analysis and often leads to draw conclusions or forecast.

- Descriptive statistics: description and summarization of data
- Inferential statistics: all techniques concerned to infer/draw conclusions from data

Example: We flip a coin 10 times and it comes up heads 7 times.
Is the coin fair?
what if it comes up heads 48 times out of 50?

$$\Downarrow$$

we need to make some assumption about the chances/probabilities...

- Probability: The *basis* of statistical thinking is *probability*. We need it in order to deal with uncertainty and, in particular, we need probability models to describe the data. Only doing so inference is possible.

Summing up

This is an elementary but rigorous *introductory course in*

Probability and Statistics:

⇒ Descriptive statistics:

- we describe data sets (tables and graphs);
- we summarize data sets through statistics regarding the “center” (mean, median, mode) and the “spread” of the values in the data set (variance, standard deviation).

- Probability: we study probability theory from its foundations with a particular attention to the fundamentals concepts which are the basis of statistical inference.
- Inferential Statistics: we study techniques to infer characteristics of the population from the observation of a sample (parameter estimation and hypothesis testing).

Frequency tables and bar/line graphs

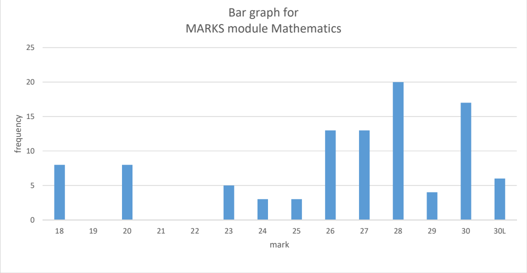
- Once collected, data have to be presented in a clear and concise way

$\Rightarrow$ we introduce tables and graphs

- Good to reveal essential features, such as range, degree of concentration and symmetry...

Example: Consider the marks of 100 students, who passed the exam in Mathematics. A possible way to represent them is through the following *frequency table*:

Mark	Frequency
18	8
20	8
23	5
24	3
25	3
26	13
27	13
28	20
29	4
30	17
30L	6

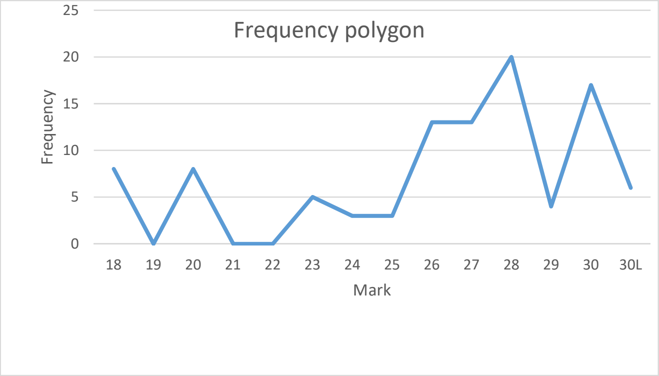


We graphically represent the data by a *bar graph* where the distinct data (marks) are plotted on the horizontal axis and the corresponding frequencies are represented by the heights of the related vertical lines.

If the lines are not thick the graph is called *line graph*.

Alternative graph: Frequency polygon

Referring to the same frequency table, *the frequency polygon* is obtained by plotting the frequencies on the vertical axis and connecting points with straight lines.

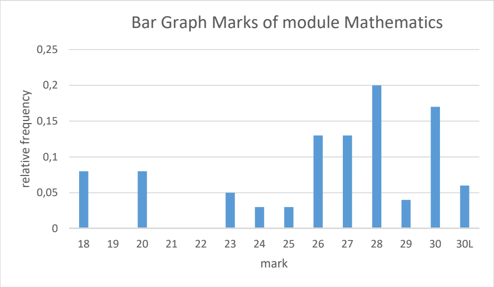


Relative frequency tables and bar/line graphs

- Consider a data set consisting of n observations, which presents k different values ($k \leq n$):
  in the example $n = 100$ number of students and $k = 14$.
- the *(absolute) frequency* f_i is the number of times the i-th value is observed; whereas the ratio $\frac{f_i}{n}$ is the *relative frequency* that is the proportion of the data which present that value.

Note: $\sum_{i:1}^{k} f_i = f_1 + \cdots + f_k = n$ and $\sum_{i:1}^{k} \frac{f_i}{n} = 1$.

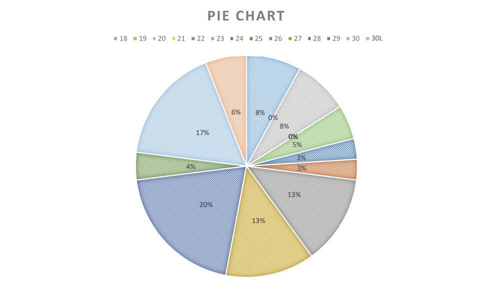
Mark	relative frequency
18	8/100=0.08
20	0.08
23	0.05
24	0.03
25	0.03
26	0.13
27	0.13
28	0.2
29	0.04
30	0.17
30L	0.06



Relative frequency tables and pie chart

- *Pie chart* are often used to represent relative frequencies when data are not numerical
- the circle is sliced into different sectors (one for each distinct data value)
- each sector area is proportional to its related relative frequency.

Mark	Relative frequency
18	8/100=0.08
20	0.08
23	0.05
24	0.03
25	0.03
26	0.13
27	0.13
28	0.2
29	0.04
30	0.17
30L	0.06



Pie chart

For plotting a pie chart we compute the angle related to each sector:

$$\alpha_i = 360^\circ \cdot \frac{f_i}{n}$$

Example: Consultant expertise (Business, Education, Healthcare, IT)

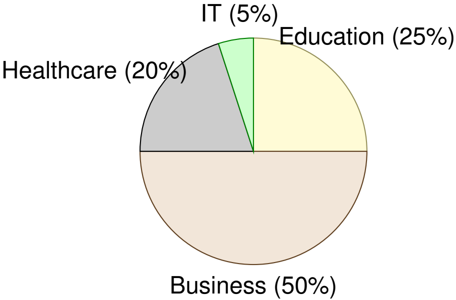
Consultant expertise	f_i
E	25
B	50
H	20
IT	5

$$\alpha_1 = 360^\circ \cdot \frac{25}{100} = 90^\circ$$

$$\alpha_2 = 360^\circ \cdot \frac{50}{100} = 180^\circ$$

$$\alpha_3 = 360^\circ \cdot \frac{20}{100} = 72^\circ$$

$$\alpha_4 = 360^\circ \cdot \frac{5}{100} = 18^\circ$$



Grouped data

- If the number of distinct data values is too large (or the values are all different) it is useful to *divide the data in suitable class intervals*

$\Rightarrow$ derive the *class frequency table* accordingly.

- The number of classes is subjective:
  too few classes $\rightarrow$ cost of loss of information
  too many $\rightarrow$ risk of small frequencies (not useful!)
- classes may be chosen of *equal length* or not
- *class boundaries* (ends points of class intervals) should belong to one and only one class; e.g., left end inclusion convention: 3-5 means [3,5)

Example: A sample of heights of 10 people is given by

$$172, 162, 178, 155, 175, 188, 195, 173, 181, 184$$

We could consider 5 classes of the same length

$$[150, 160), [160, 170), [170, 180), [180, 190), [190, 200)$$

or, alternatively, 3 classes of different length

$$[150, 170), [170, 180), [180, 200)$$

In any case, if data are grouped in classes, we use *histograms* to represent graphically the frequencies.

Histogram

The *histogram* is composed by k rectangles, one for each class:

- the *base* is represented by the *class interval*;
- the *area* is *proportional to the frequency* of the class.

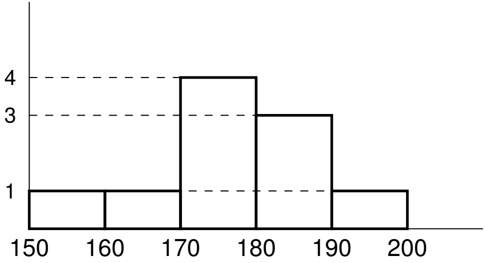
If the classes have equal length, also the heights will be proportional to the class frequencies.

Example: Sample of 10 people heights

$$172, 162, 178, 155, 175, 188, 195, 173, 181, 184$$

I consider 5 classes of length 10 cm from 150 to 200 cm.

	f_i
[150,160)	1
[160,170)	1
[170,180)	4
[180,190)	3
[190,200)	1



If the length of the classes were *not equal*, we would need to find the *correct height* of the rectangles.

Histograms with class intervals of different length

Consider k classes $[x_0, x_1), \ldots, [x_{k-1}, x_k)$ with frequencies $f_1, \ldots, f_k$.

Denote by a_i the length of the class $a_i = x_i - x_{i-1} \qquad i = 1, \ldots, k$.

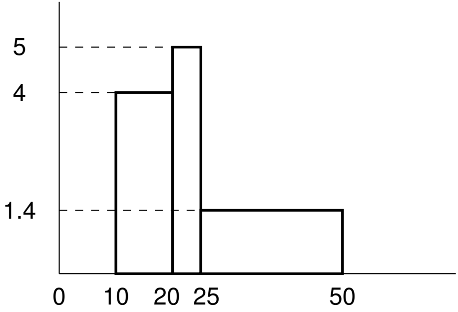
recall the area of the rectangle should be proportional to the frequency of the class

$\Rightarrow$ thus we can find the *height* of the rectangles

$$f_i = a_i \times h_i \quad \Rightarrow \quad h_i = \frac{f_i}{a_i} \qquad i = 1, \ldots, k.$$

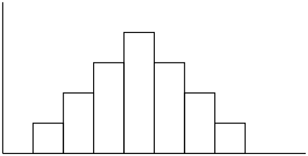
Example:

	f_i	a_i	h_i
[10,20)	40	10	$\frac{40}{10} = 4$
[20,25)	25	5	$\frac{25}{5} = 5$
[25,50)	35	25	$\frac{35}{25} = 1.4$

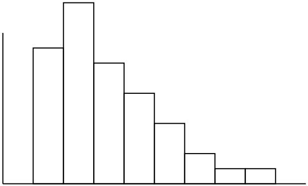


Symmetry

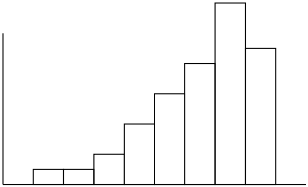
The histogram is *symmetric* if the observations are approximately regularly distributed around the center of the histogram.



Symmetry implies that very small and very big values, with respect to the center, are equally likely.
If the frequencies decrease slower on a side, there will be asymmetry: positive if there is a heavier tail on the right and negative if the heavier tail extends on the left.



positive asymmetry



negative asymmetry

Ogive

The *ogive* represents the plot of the *cumulative frequencies* (or cumulative relative frequencies).
A point on the horizontal axis represents a possible data value and the vertical plot gives the number (or proportion) of data which are less or equal to it.

Example: The following data show the distribution of salaries (millions dollars, per team) for thirty-one NFL teams.
Plot the corresponding ogive.

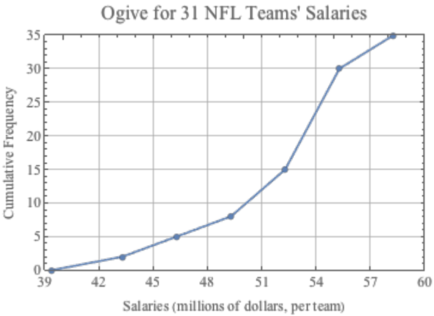
	f_i
[39.5,43.5)	2
[43.5,46.5)	3
[46.6,49.5)	3
[49.5,52.5)	7
[52.5,55.5)	15
[55.5,58.5)	1

Ogive

The *ogive* represents the plot of the *cumulative frequencies* (or cumulative relative frequencies).
A point on the horizontal axis represents a possible data value and the vertical plot gives the number (or proportion) of data which are less or equal to it.

Example: The following ogive shows the distribution of salaries for thirty-one NFL teams.
This ogive is useful for finding how many teams (out of the 31) pay salaries less or equal to a specific value.

	f_i	$F_i = f_1 + \dots f_i$
[39.5,43.5)	2	2
[43.5,46.5)	3	5
[46.6,49.5)	3	8
[49.5,52.5)	7	15
[52.5,55.5)	15	30
[55.5,58.5)	1	31



Stem and leaf plot

- The *stem and leaf* plot is an *efficient way to represent a data set*
- It is a method for showing the frequency with which certain classes of values occur.
- Typically the *stem* is the left hand column containing the tens digits and the *leaf* is in the right column and contains the ones.
  (Typically if the data have decimals, decimals will be the *leaf*).

Example: This example shows the heights of 200 adults.

Table 1

Stem-and-Leaf Plot Showing Distribution of Adult Heights (N=200)

Inches	.1 inches
55	3
56	9
57	7
58	57
59	25
60	228899
61	1225667
62	033556899
63	011223558888999
64	012333344555688
65	112334555555667888889
66	01122344556688889999
67	00012222345555556666666667899
68	1122233333444556666799999
69	112335555589
70	1244566779
71	234455789
72	11223344666778
73	06678
74	8
75	18

The left (tens column) contains the number of whole inches and the right column (ones) shows the decimal of that inch. When a number is repeated in the ones column, more than one person is that height.