Introduction to Machine Learning

Data analysis using basic statistical methods

Descriptive statistics

Learning objectives: be able to

- describe data by means of statistical measures and graphical approaches
- interpret statistics and graphics to draw conclusions about data

Uni-variate analysis applied to qualitative variables

Frequency

Given a set of N instances, for the outcome i:

- n_i: Frequency (the number of times the outcome i appeared in the set)
- f_i: Relative frequency $f_i = \frac{n_i}{N}$
- N_i: Cumulative frequency $N_i = \sum_{k=1}^{i} n_k$
- F_i: Cumulative relative frequency $F_i = \sum_{k=1}^{i} f_k$

Uni-variate analysis applied to qualitative variables

Frequency

Exercise: can we compute the cumulative relative frequency (F_i) as follows? (either proof or provide a counterexample)

$$F_i = \frac{N_i}{N}$$

Uni-variate analysis applied to qualitative variables

Frequency

Exercise: dice rolling. Given the absolute frequency (n_i) of each outcome (i), compute the relative frequency (f_i) and the cumulative relative frequency (N_i).

i	n_i	f_i	N_i
1	12		
2	20		
3	15		
4	19		
5	10		
6	24		

Uni-variate analysis applied to qualitative variables

Frequency

Exercise: how was the coronavirus weekly death toll in your country during April and May? Which was the total death toll? How would you compare the data from two countries?

Week	n_i	f_i	N_i	F_i
1				
2				
3				
4				
5				
6				
7				
8				

Uni-variate analysis applied to qualitative variables

Frequency

x_i	n_i	f_i
Noun	137	0.457
Adjective	51	0.170
Adverb	20	0.067
Verb	92	0.307
	300	1.000

sort by f_i $\longrightarrow$

x_i	n_i	f_i
Noun	137	0.457
Verb	92	0.307
Adjective	51	0.170
Adverb	20	0.067
	300	1.000

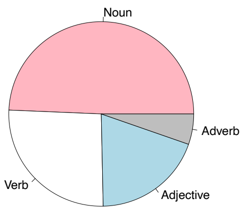


Figure 1: Pie chart

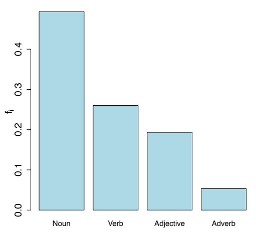


Figure 2: Bar plot

Uni-variate analysis applied to qualitative variables

Graphics

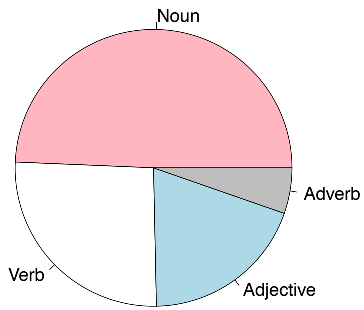


Figure 3: Pie chart

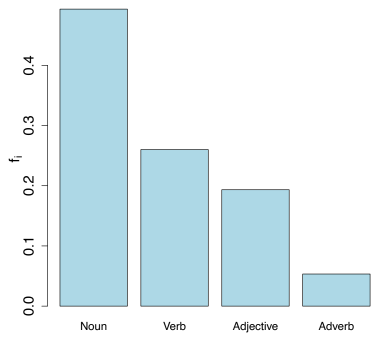


Figure 4: Bar plot

Exercise: given these figures, can you get a frequency table? (both n_i and f_i?)

Uni-variate analysis applied to qualitative variables

Variability

Example:
An analysis was carried out for a document in two languages.

N	N	C	N	N	N	N	C	N	N
C	N	N	N	N	N	N	N	N	V
N	N	N	N	N	V	N	N	N	N
N	C	N	N	N	N	N	N	V	V

Table 1: Language 1

C	C	C	N	C	C	V	C	V	C
C	C	C	N	N	V	N	V	C	N
N	V	V	C	N	C	V	V	N	C
C	V	N	N	V	N	C	C	C	N

Table 2: Language 2

Intuitively, which language shows the biggest variability?

Uni-variate analysis applied to qualitative variables

Variability

Computing variability:

Maximum $M = 1 - \max_{i=1}^{N} f_i$

Gini's index $G = \sum_{i=1}^{N} f_i(1 - f_i)$

Entropy $H = -\sum_{i=1}^{N} f_i \log_2(f_i)$

Uni-variate analysis applied to qualitative variables

Variability

Exercise: Assess, quantitatively, the variability of the following two sets.

N	N	C	N	N	N	N	C	N	N
C	N	N	N	N	N	N	N	N	V
N	N	N	N	N	V	N	N	N	N
N	C	N	N	N	N	N	N	V	V

Table 3: Set 1

C	C	C	N	C	C	V	C	V	C
C	C	C	N	N	V	N	V	C	N
N	V	V	C	N	C	V	V	N	C
C	V	N	N	V	N	C	C	C	N

Table 4: Set 2

fill the gaps	Set 1	Set 2
Maximum		
Gini's index		
Entropy		

Table 5: Variability of sets 1 and 2

Uni-variate analysis applied to quantitative variables

Exercise: The following table gathers the length of each sentence (variable X) in a text.

sentence i	length x_i
1	14
2	8
3	13
4	9
5	14
⋮	⋮
148	9
148	11
149	14
150	11

1. What is the type of variable X?
2. Do you find it appropriate a frequency table to explore this variable?

Uni-variate analysis applied to quantitative variables

sentence i	length x_i
1	14
2	8
3	13
4	9
5	14
⋮	⋮
148	9
148	11
149	14
150	11

Table 6: Data set

$\longrightarrow$

x_i	n_i	N_i	f_i	F_i
2	1	1	0.007	0.007
5	1	2	0.007	0.013
6	4	6	0.027	0.040
7	3	9	0.020	0.060
8	7	16	0.047	0.107
9	17	33	0.113	0.220
10	11	44	0.073	0.293
11	21	65	0.140	0.433
12	16	81	0.107	0.540
13	15	96	0.100	0.640
14	22	118	0.147	0.787
15	15	133	0.100	0.887
16	10	143	0.067	0.953
17	5	148	0.033	0.987
18	2	150	0.013	1.00

Table 7: Frequency table

Uni-variate analysis applied to quantitative variables

x_i	n_i	N_i	f_i	F_i
2	1	1	0.007	0.007
5	1	2	0.007	0.013
6	4	6	0.027	0.040
7	3	9	0.020	0.060
8	7	16	0.047	0.107
9	17	33	0.113	0.220
10	11	44	0.073	0.293
11	21	65	0.140	0.433
12	16	81	0.107	0.540
13	15	96	0.100	0.640
14	22	118	0.147	0.787
15	15	133	0.100	0.887
16	10	143	0.067	0.953
17	5	148	0.033	0.987
18	2	150	0.013	1.00

Table 8: Frequency table

$\longrightarrow$

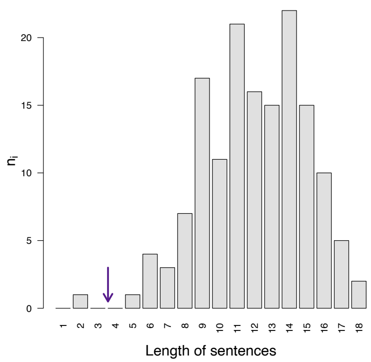


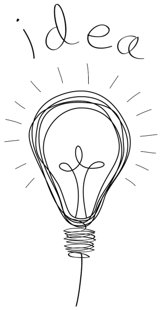
Figure 5: Bar plot of variable X

Uni-variate analysis applied to quantitative variables

Exercise: recap

1. How many instances did we have in our data set? (for how many sentences did we collect the length? N?)
2. How many different outcomes did we observe for variable X? (different lengths for sentences)
3. What is the type of variable X? (bear in mind that X is the length of a sentence)
4. How many different outcomes (x_i) could X have taken? Accordingly, what would be the length of the corresponding frequency table? Discuss about the relative frequency for each outcome (f_i). Do you find frequency table a sensitive way to summarize and analyze the data? Why?

Uni-variate analysis applied to quantitative variables



Both frequency tables and bar plots are good tools to summarize the data collected about X and get insights at a glance when the space of observations (the set of x_i) is discrete and small (much smaller than the size of the data-set).

Uni-variate analysis applied to quantitative variables

Histogram

x_i	n_i	N_i	f_i	F_i
2	1	1	0.007	0.007
5	1	2	0.007	0.013
6	4	6	0.027	0.040
7	3	9	0.020	0.060
8	7	16	0.047	0.107
9	17	33	0.113	0.220
10	11	44	0.073	0.293
11	21	65	0.140	0.433
12	16	81	0.107	0.540
13	15	96	0.100	0.640
14	22	118	0.147	0.787
15	15	133	0.100	0.887
16	10	143	0.067	0.953
17	5	148	0.033	0.987
18	2	150	0.013	1.00

Table 9: Frequency table

$\longrightarrow$

Group the data by intervals (turn to discretization or *binning*).
Exercise: fill in the table.

$[x_i, x_{i+1})$	n_i	N_i	f_i	F_i
$[1, 7)$	6	6	0.04	0.04
$[7, 11)$				
$[11, 16)$			0.59	
$[16, \infty)$	17	150		1.00

Table 10: Frequency table by intervals

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: given the relative frequency (f_i) for the variable Y, can you draw a bar plot?

y_i	n_i	N_i	f_i	F_i
$[1, 7)$	6	6	0.04	0.04
$[7, 11)$	38	44	0.25	0.29
$[11, 16)$	89	133	0.59	0.89
$[16, \infty)$	17	150	0.11	1.00

Table 11: Frequency table

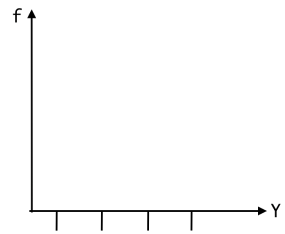


Figure 6: Bar plot of Y

Recap: a histogram of X is a bar plot of the intervals of X (in the example Y represents the intervals of X).

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: (yet another one about discretization)
A study collected the visual lexical decision latency for beginning readers (in seconds) = X

$\mathcal{D}$ comprises these data

	x_i		x_i		x_i
1	0.22	11	0.41	21	0.49
2	0.25	12	0.41	22	0.50
3	0.32	13	0.45	23	0.52
4	0.34	14	0.46	24	0.53
5	0.34	15	0.46	25	0.54
6	0.37	16	0.46	26	0.54
7	0.38	17	0.46	27	0.58
8	0.39	18	0.47	28	0.59
9	0.40	19	0.48	29	0.60
10	0.41	20	0.49	30	0.60

For $\mathcal{D}$ provide:

1. The size of the data-set.
2. The type of X.
3. The size (range) of the observation space of X (min, max).
4. The mode of X.
5. A bar plot.

Uni-variate analysis applied to quantitative variables

Histogram

Exercise:
I drew this bar plot for $\mathcal{D}$

1. is it useful?
2. what does it show?

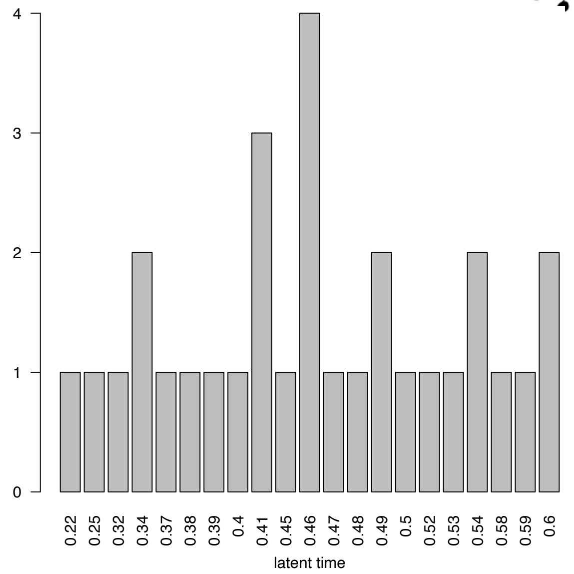


Figure 7: Bar plot of latency time (s) for beginner readers

Uni-variate analysis applied to quantitative variables

Histogram

Exercise:
I discretized X and drew the corresponding histogram

1. is it useful?
2. what does it show?

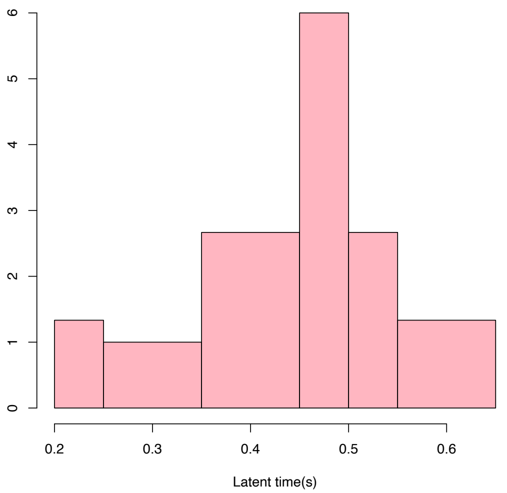


Figure 8: Histogram of latency time (s) for beginner readers

Uni-variate analysis applied to quantitative variables

Histogram

Histogram: properties

- For the quantitative variable X, the values (x_i) are bracketed together in k intervals such as $[x_i, x_{i+1})$ with $1 \leq i \leq k$ (X is discretized)
- The area corresponding to the interval $[x_i, x_{i+1})$ is proportional to the frequency of the observations within the interval.

$$[height_i \cdot (x_{i+1} - x_i)] \propto f_i$$

  - In particular, if all the intervals are equally spaced, then, the height of the interval is also proportional to the frequency (not only the area):

    $x_{i+m} - x_i = m \cdot (x_{i+1} - x_i) \quad 1 \leq i \leq k \quad 1 < i + m \leq k + 1 \quad \Rightarrow height_i \propto f_i$
- Appropriate for both
  - quantitative continuous variables (in general)
  - quantitative discrete variables when the size of the observation space is big

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: (yet another one about discretization)
A study collected the visual lexical decision latency for beginning readers (in seconds) = X

$\mathcal{D}$ comprises these data

	x_i		x_i		x_i
1	0.22	11	0.41	21	0.49
2	0.25	12	0.41	22	0.50
3	0.32	13	0.45	23	0.52
4	0.34	14	0.46	24	0.53
5	0.34	15	0.46	25	0.54
6	0.37	16	0.46	26	0.54
7	0.38	17	0.46	27	0.58
8	0.39	18	0.47	28	0.59
9	0.40	19	0.48	29	0.60
10	0.41	20	0.49	30	0.60

1. Draw the histogram after having discretized X in 5 bins by means of
   1.1 equal width binning
   1.2 equal frequency binning
2. Discuss about the area of the bars in each of the histograms.

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: graphic representation of data

2	1	1	1	0	3	1	3	1	1
2	2	0	4	0	1	2	2	0	1
2	3	1	0	2	3	2	6	3	1
4	2	6	2	3	3	4	2	4	2
2	1	2	2	0	2	2	1	0	4

Table 12: Data-set 1

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: graphic representation of data

16.3	24.5	31.7	42.6	48.8	51.2	54.9	55.3	55.3	61.7
62.7	64.4	66.2	66.5	67.7	69.9	70.2	72.6	72.8	73.3
73.7	74.8	75.4	75.8	77.1	78.9	79.0	79.4	80.4	80.6
81.2	82.5	83.1	83.2	83.5	85.1	87.0	87.4	88.1	88.1
88.9	89.3	90.0	90.3	90.7	91.2	91.5	91.5	91.6	91.6
92.0	95.2	97.5	97.7	98.3	98.9	99.1	99.6	101.6	104.7
104.9	105.4	105.5	106.2	106.8	108.9	110.0	110.5	111.0	113.9
114.3	115.7	116.1	116.3	117.7	117.9	120.1	120.3	121.2	122.2
124.2	124.9	125.1	126.3	126.9	128.8	130.0	131.2	133.9	136.8
136.9	140.2	140.4	140.4	140.6	141.6	147.8	149.0	159.9	165.8

Table 13: Data-set 2

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: graphic representation of data

54.90	70.20	74.80	85.10	87.00
89.30	90.30	92.00	92.20	105.50
116.10	120.30	121.20	124.90	165.80

Table 14: Data-set 3

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: analyze the graph and describe the data (variable type and observed outcomes, number of instances, . . .)

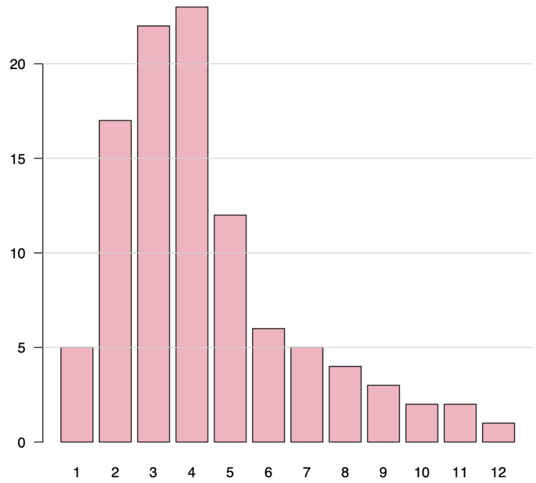


Figure 9: Graph 1

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: analyze the graph and describe the data (variable type and observed outcomes, number of instances, . . .)

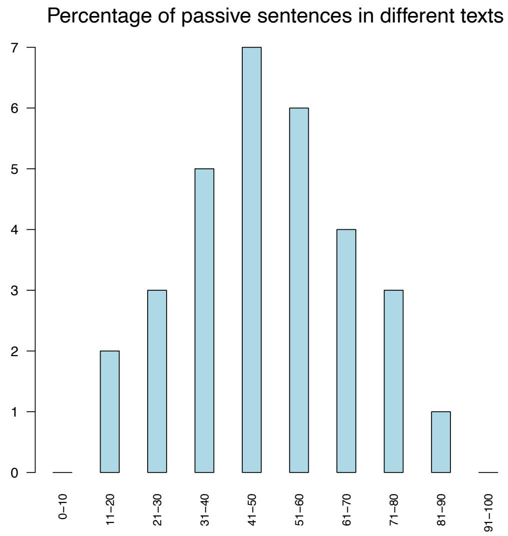


Figure 10: Graph 2

Uni-variate analysis applied to quantitative variables

Histogram

Exercise: analyze the graph and describe the data (variable type and observed outcomes, number of instances, . . .)

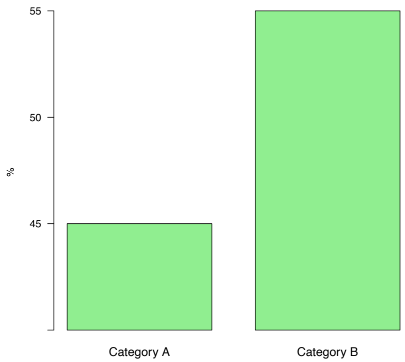


Figure 11: Graph 3

Uni-variate analysis applied to quantitative variables

Histogram

Counter-example

Source of the figure: `https://junkcharts.typepad.com`

Uni-variate analysis applied to quantitative variables
Statistics

Statistics

- Statistics provide quantitative summaries of the data set.
- Classification of statistics depending on the objectives they meet:
  1. measures of the location (position) of the data
  2. central tendency
  3. spread
  4. shape

Uni-variate analysis applied to quantitative variables
Statistics for the location of data

Measures of the location (position) of the data

- median
- quartiles
- percentiles
- interquartile-range

Graphic representations: boxplot

Uni-variate analysis applied to quantitative variables
Statistics for the location of data

Median (Me): after having **sorted** the observations in ascending order, the median is the midpoint value i.e. half of the observations (50%) are equal or smaller than the median and half are equal or larger. The median separate ordered observations into halves. The median may or may not be part of the data.

Uni-variate analysis applied to quantitative variables
Statistics for the location of data

Quartiles: divide ordered observations into quarters.

Q_1 (25%)

Q_2 (50%)

Q_3 (75%)

Percentiles: divide ordered observations into hundredths.

P_1 (1%)

P_2 (2%)

$\ldots$

P_{99} (99%)

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

- A percentile informs about the location of the data (relative position of sorted data)
- *n*% of the outcomes are less than or equal to the *n*th percentile
- Median and quartiles are particular cases of the percentiles

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Exercise: Compute the median of the latency to utter the first vocalization (s)

	Observations (s)
1	0.46
2	0.41
3	0.46
4	0.35
5	0.41
6	0.46
7	0.47
8	0.54
9	0.22
10	0.49
11	0.34
12	0.25
13	0.60
14	0.59
15	0.54

$\longrightarrow$

	Observations (s)
1	0.22
2	0.25
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	0.60

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Exercise: True/False
We can compute...

1. ...the Median as the second quantile
2. ...the Median as the 50th percentile
3. ...Q3 as the 75th percentile
4. ...Q3 as the median of the upper half of the sorted data

Exercise: X is a continuous numeric variable for which we gathered 1000 observations. We discretized X in four bins. We decided the bin range in such a way that all the bins contained the same number of observations (though the width of each bin was irregular). How would be the resulting bar-plot?

Exercise: How can I compute the median if the dataset contains an even number of observations?

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Box-plot: shows min, Q_1, median, Q_3, max statistics graphically. Aka box-and-whisker plot.

Example:

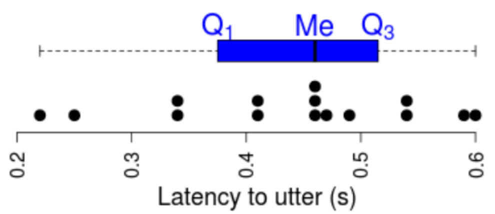


Figure 12: Box-plot of the latency to utter the first vocalization (s)

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Exercise: True/False

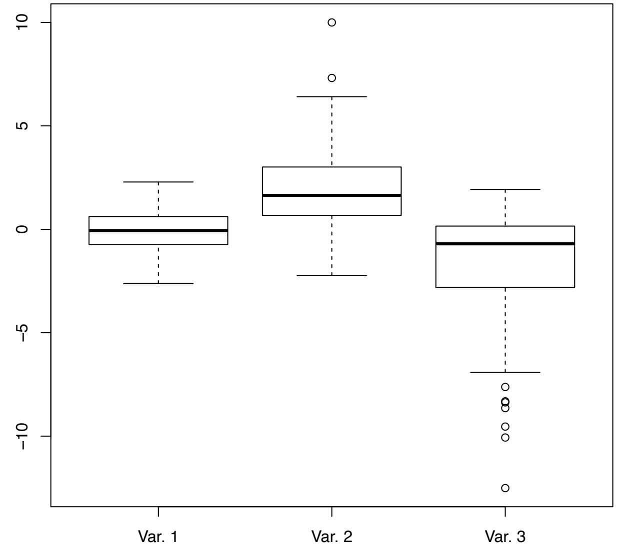
There are 100 students in this course, going through your transcript of records, you are the 90th percentile . . .

1. it means that you received 90% as the average score
2. it means that 90% of the scores in this course are the same or less than yours
3. it means that 10% of the test scores are the same or greater than yours

Uni-variate analysis applied to quantitative variables

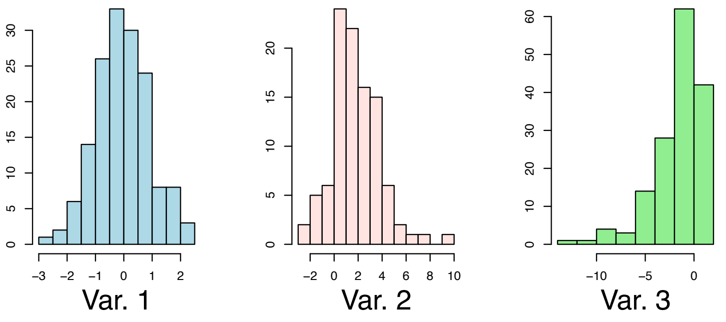
Statistics for the location of data

Exercise: relation between box-plot and bar-plot. Given a box-plot can you draw a compatible bar-plot? And the other way around?



Uni-variate analysis applied to quantitative variables

Statistics for the location of data



Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Exercise:

1. Enumerate 5 journals that get papers within the area of Computational Linguistics, Text Mining, Natural Language Processing, Natural Language Understanding, etc.
2. Regarding the latest Journal Impact Factor, which of them are Q_1 according to the JCR?

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Exercise: draw a box-plot for each data-set

1. $\mathcal{D}_1$={5, 5, 5, 5, 5, 5, 5, 5, 5, 5}
2. $\mathcal{D}_2$={5, 5, 5, 5, 5, 5, 5, 6, 6, 7}
3. $\mathcal{D}_3$={59, 60, 61, 62, 62, 63, 63, 64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 67, 67, 68, 68, 69, 70, 70, 70, 70, 70, 71, 71, 72, 72, 73, 74, 74, 75, 77}

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Exercise: True/False: can you find a data-set for which. . .

1. the minimum value is equal to the first quartile and display the corresponding box-plot? if so, display the corresponding box-plot
2. the median is equal to the third quartile? if so, display the corresponding box-plot
3. the first and third quartiles are equal

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Exercise: We collected max mel cepstrum coeficient for two sets of speakers (A) with and (B) without Alzheimer:

- A ={ 69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94}
- B ={90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100}

1. Draw the box-plot for each set in the same scale (keep the ordinate)
2. How distinct are both sets of speakers regarding these observations?

Uni-variate analysis applied to quantitative variables

Statistics for the location of data

Exercise: fill in the table and represent the data by means of box-plot and find the 65th percentile

x	n	f	N	F
2	12			
3	14			
4		0.25		
5	4		40	

Uni-variate analysis applied to quantitative variables
Statistics for the location of data

Exercise:

- Form groups with 3 students per group
- Each student has to write a frequency table and draw in separate pieces of paper a histogram representing the data and the box-plot
- Collect the histograms and the box-plots (6 pieces of papers) and exchange with another group. Next, match each histogram with a box-plot.

Uni-variate analysis applied to quantitative variables
Central tendency

Statistics

- Statistics provide quantitative summaries of the data set.
- Classification of statistics depending on the objectives they meet:
  1. measures of the location (position) of the data
  2. central tendency
  3. spread
  4. shape

Uni-variate analysis applied to quantitative variables
Central tendency

Statistics of Central Tendency of the Data:

- Goal: what is a representative observation like?
- These statistics try to provide a prototype that would represent the sample.

Uni-variate analysis applied to quantitative variables
Central tendency

Statistics of Central Tendency of the Data: what is a representative observation like?

- Median: after having sorted the observations in ascending order, the median is the mid point value
- Mean ($\bar{x}$): given a sample $x_1, x_2, \ldots, x_n$, the arithmetic mean ($\bar{x}$) is

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

- Mode: the most frequent observation. We can have more than one mode in a sample.

Uni-variate analysis applied to quantitative variables
Central tendency

What is a representative observation like? Median or Mean?

- Both, median and mean are in the interval $[x_{min}, x_{max}]$
- Mean is regarded as the *prototype* of the sample in the sense of a *balance point* or *mass center*
- Generally, the median is a better measure of the center...
  - ...in severely asymmetric distributions:
    - ☹ the mean tends to be located towards the tail
    - ☺ the median would be a better representative observation than the mean
  - ...in samples with potential outliers:
    - ☹ The mean is affected by the value of the outliers
    - ☺ The median is insensitive to the value that the outliers take

Uni-variate analysis applied to quantitative variables
Central tendency

	Observations (s)
1	0.46
2	0.41
3	0.46
4	0.35
5	0.41
6	0.46
7	0.47
8	0.54
9	0.22
10	0.49
11	0.34
12	0.25
13	0.60
14	0.59
15	0.54

Table 15: Latency to utter the first vocalization (s)

Exercise: compute the mean latency to utter the first vocalization on this sample

$$n = 15$$
$$\sum_{i=1}^{n} x_i = 6.18$$
$$\bar{x} = 0.439$$

Uni-variate analysis applied to quantitative variables
Central tendency

x_i	n_i	N_i	f_i	F_i
2	1	1	0.007	0.007
5	1	2	0.007	0.013
6	4	6	0.027	0.040
7	3	9	0.020	0.060
8	7	16	0.047	0.107
9	17	33	0.113	0.220
10	11	44	0.073	0.293
11	21	65	0.140	0.433
12	16	81	0.107	0.540
13	15	96	0.100	0.640
14	22	118	0.147	0.787
15	15	133	0.100	0.887
16	10	143	0.067	0.953
17	5	148	0.033	0.987
18	2	150	0.013	1.00

Table 16: Frequency table with N=150 total observations of the variable X and m=18 different outcomes registered

Exercise: Describe the central tendency of this sample. Mean, Median, Mode.

Uni-variate analysis applied to quantitative variables
Central tendency

x_i	n_i	N_i	f_i	F_i
2	1	1	0.007	0.007
5	1	2	0.007	0.013
6	4	6	0.027	0.040
7	3	9	0.020	0.060
8	7	16	0.047	0.107
9	17	33	0.113	0.220
10	11	44	0.073	0.293
11	21	65	0.140	0.433
12	16	81	0.107	0.540
13	15	96	0.100	0.640
14	22	118	0.147	0.787
15	15	133	0.100	0.887
16	10	143	0.067	0.953
17	5	148	0.033	0.987
18	2	150	0.013	1.00

Table 17: Frequency table with N=150 total observations of the variable X and m=18 different outcomes registered

Exercise: Given a frequency table, can we compute the mean as follows? justify your answer

$$\bar{x} = \sum_{i=1}^{m} (f_i x_i)$$

Mean is defined as:

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

then...

Uni-variate analysis applied to quantitative variables

Central tendency

- The sample mean ($\bar{x}$ is an statistic to estimate the population mean (μ).
- According to The Law of Large Numbers, the mean of a random sample ($\bar{x}$) is likely to get closer to the mean of the population (μ) as the size of the sample increases.

Uni-variate analysis applied to quantitative variables

Spread of data

Statistics

- Statistics provide quantitative summaries of the data set.
- Classification of statistics depending on the objectives they meet:
  1. measures of the location (position) of the data
  2. central tendency
  3. spread
  4. shape

Uni-variate analysis applied to quantitative variables

Spread of data

Spread of data:

- Goal: are the observations stretched or squeezed?
- Spread aka dispersion, variability, scatter

Uni-variate analysis applied to quantitative variables

Spread of data

Exercise: These two samples show the same mean value but a different spread. Depict the histogram.

A	B
6	4
6.5	5
7	6
7	8
7.5	9
8	10

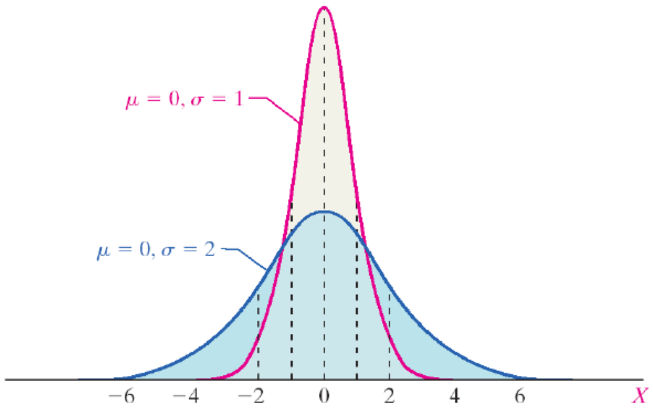
Compute:

- $\bar{x}_A =$
- $\bar{x}_B =$

Uni-variate analysis applied to quantitative variables

Spread of data

Example:



Uni-variate analysis applied to quantitative variables

Spread of data

Statistics to compute the spread of the data:

- Range (R): the difference between minimum and maximum value

$$R = x_{max} - x_{min}$$

- Inter-quartile range (IQR): the spread of the 50% of the data that are in the middle

$$IQR = Q_3 - Q_1$$

- Variance (s^2): given the observations $x_1, x_2, \ldots, x_n$

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

- Standard deviation (s):

$$s = \sqrt{s^2}$$

Uni-variate analysis applied to quantitative variables

Spread of data

Exercise: These two samples show the same mean value but a different spread. Compute the standard deviation to assess the spread of the data.

A	B
6	4
6.5	5
7	6
7	8
7.5	9
8	10

$\bar{x}_A = 7.00 \qquad \bar{s}_A =$

$\bar{x}_B = 7.00 \qquad \bar{s}_B =$

Uni-variate analysis applied to quantitative variables

Spread of data

Inter-quartile range is often used to detect outliers (x):

$$x \notin [Q_1 - 1.5IQR, Q_3 + 1.5IQR]$$

Uni-variate analysis applied to quantitative variables
Spread of data

Exercise: We measured the length of the dialogue-turns (s) but we feel that we might have introduced data from another speaker by mistake. Can you find potential outliers?

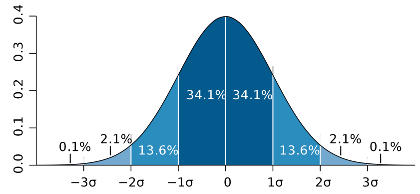
1	2	3	4	5	6	7	8	9	10	11
68.5	33.0	69.0	54.0	54.0	28.0	120.0	42.0	72.0	40.5	64.5

Uni-variate analysis applied to quantitative variables
Spread of data

Properties: ($\bar{x}$: mean and s standard deviation)
The percentage of observations that are within the interval

$$(\bar{x} - 2s, \bar{x} + 2s)$$

- are at least 75% (in general)
- approximately 95% if the distribution is bell-shaped

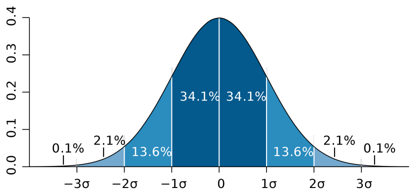


Source: Wikimedia Commons

Uni-variate analysis applied to quantitative variables
Spread of data

Empirical rule: with a sample of size n in which the distribution of the variable is bell shaped, then, approximately

- %68 of the observations are within $(\bar{x} - s, \bar{x} + s)$
- %95 of the observations are within $(\bar{x} - 2s, \bar{x} + 2s)$
- nearly all the observations are within $(\bar{x} - 3s, \bar{x} + 3s)$



Source: Wikimedia Commons

Uni-variate analysis applied to quantitative variables
Spread of data

x_i	n_i	N_i	f_i	F_i
2	1	1	0.007	0.007
5	1	2	0.007	0.013
6	4	6	0.027	0.040
7	3	9	0.020	0.060
8	7	16	0.047	0.107
9	17	33	0.113	0.220
10	11	44	0.073	0.293
11	21	65	0.140	0.433
12	16	81	0.107	0.540
13	15	96	0.100	0.640
14	22	118	0.147	0.787
15	15	133	0.100	0.887
16	10	143	0.067	0.953
17	5	148	0.033	0.987
18	2	150	0.013	1.00

Frequency table with N=150 total observations of the variable X and m=18 different outcomes registered

Example:

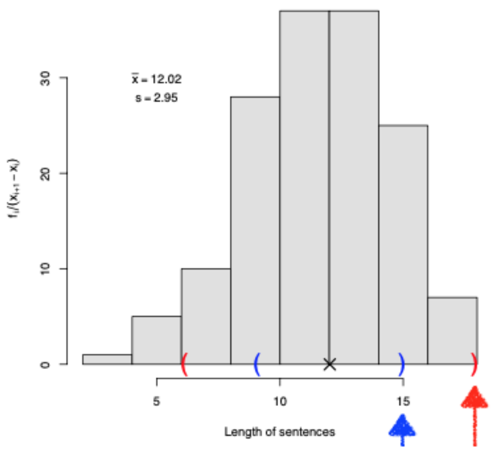
- X: length of the sentences
- n=150
- $\bar{x} = 12.02$
- $s_x = 2.95$

Uni-variate analysis applied to quantitative variables

Spread of data

Example:

- X: length of the sentences
- n=150
- $\bar{x} = 12.02$
- $s_x = 2.95$



Uni-variate analysis applied to quantitative variables

Shape of the data

Statistics

- Statistics provide quantitative summaries of the data set.
- Classification of statistics depending on the objectives they meet:
  1. measures of the location (position) of the data
  2. central tendency
  3. spread
  4. shape

Uni-variate analysis applied to quantitative variables

Shape of the data

Shape of data:

- Goal: are the observations skewed or symmetric?

Uni-variate analysis applied to quantitative variables

Shape of the data

Exercise: How is the mean compared to the median in these samples?

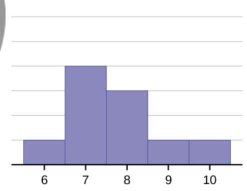


Figure 13: Skewed to the left

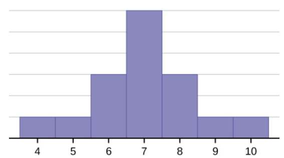


Figure 14: Symmetric

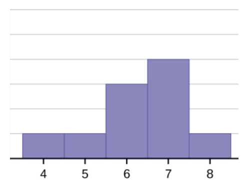
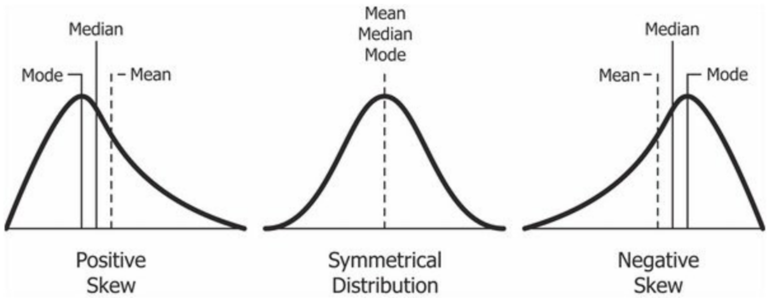


Figure 15: Skewed to the right

Uni-variate analysis applied to quantitative variables

Shape of the data



Source: Wikimedia Commons

Uni-variate analysis applied to quantitative variables

Linear transformations applied to data

Linear transformation

Given a set of observations $x_1, x_2, \ldots, x_n$ with $x_i \in \mathbb{R}$, applying a linear transformation, transformed values $y_1, y_2, \ldots, y_n$ are obtained with:

$$y_i = ax_i + b$$

with $a, b \in \mathbb{R}$ being a pair of constant values.

Uni-variate analysis applied to quantitative variables

Linear transformations applied to data

Exercise: $x_1, x_2, \ldots, x_n$ is a set of observations of a variable X and Y is a linear transformation of X. Compute $\bar{y}$.

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\ &= \text{fill in} \\ &= \end{aligned}$$

Uni-variate analysis applied to quantitative variables

Linear transformations applied to data

Properties of linear transformations:

- Applying a linear transformation to a random variable (X), a new random variable (Y) is obtained.
- How does a linear transformation affect the mean, variance and standard deviation?

$$\begin{aligned} \bar{y} &= a\bar{x} + b \\ S_y^2 &= a^2 S_x^2 \\ S_y &= |a| S_x \end{aligned}$$

Uni-variate analysis applied to quantitative variables

Linear transformations applied to data

Applications of linear transformations:

To center your data apply this linear transformation

$$y_i = x_i - \bar{x}$$

with this linear transformation we get:

$$\bar{y} = 0$$
$$S_y = S_x$$

To standardize your data apply this linear transformation

$$z_i = \frac{x_i - \bar{x}}{S_x}$$

with this linear transformation we get:

$$\bar{z} = 0$$
$$S_z = 1$$

Uni-variate analysis applied to quantitative variables

Linear transformations applied to data

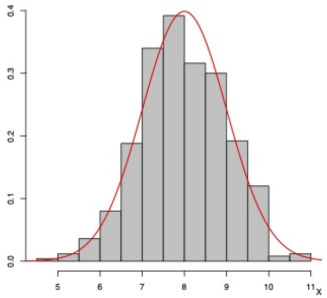
Exercise: A linear transformation is a transformation of the type $y = ax + b$ with a and b constant values.

1. How are those a and b for the transformation used to center the data?
2. How are those a and b for the transformation used to standardize the data?

Uni-variate analysis applied to quantitative variables

Models

- Some events (described through random variables) follow (or nearly follow) particular distributions (models). Example:



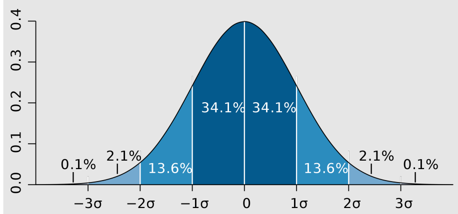
- Models are often used to summarize the general behavior of a variable.
- Examples of models: normal distribution, Poisson, binomial, etc.

Uni-variate analysis applied to quantitative variables

Models

Gaussian (or Normal) distribution: $\mathcal{N}(\mu, \sigma)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Probability density function

Source: Wikimedia Commons

$\mathcal{N}(\mu, \sigma)$

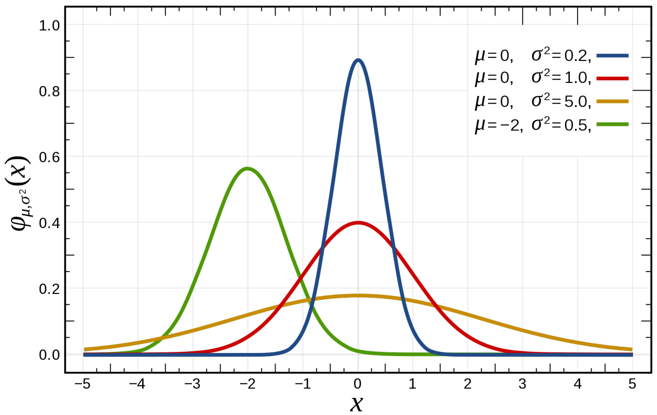
- Mean: μ
- Median: μ
- Mode: μ
- Standard deviation: σ
- Variance: σ^2
- Skewness: 0

Characteristic parameters: μ, σ

Uni-variate analysis applied to quantitative variables

Models

Exercise: applying linear transformations to a normal distribution



Source: Wikimedia Commons

Which of them is standardized?

Uni-variate analysis applied to quantitative variables

Models

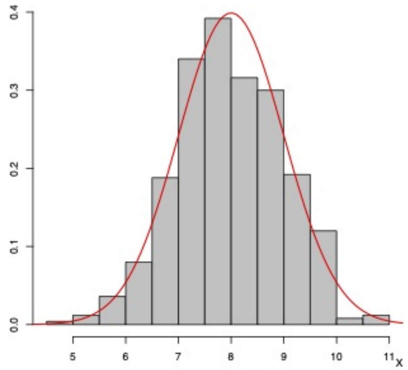
Statistical Inference:

- $\mathcal{N}(\mu, \sigma)$ is characterized by two parameters: μ, σ
- Estimate the model's parameters that best fit the data, i.e. given a sample $x_1, x_2, \ldots, x_n$, the aim is to compute the estimated values (maximum likelihood estimation): $\widehat{\mu}, \widehat{\sigma}$

Estimation

$$\widehat{\mu} = \bar{X}$$
$$\widehat{\sigma} = S_x$$

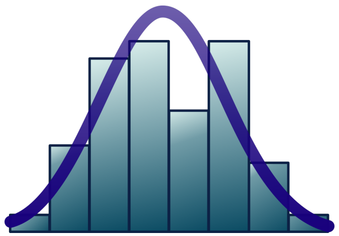
$\longrightarrow$



Uni-variate analysis applied to quantitative variables

Models

Exercise: with X=sepal-width for the iris versicolor instances(iris dataset)



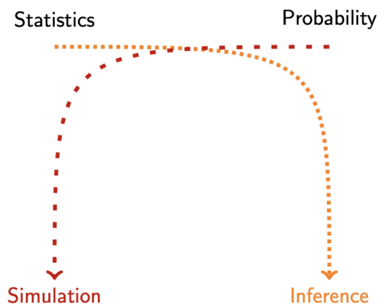
Source: Wikimedia Commons

Compute $\widehat{\mu}, \widehat{\sigma}$

Concluding remarks

Concluding remarks

- Descriptive statistics: describe or summarize characteristics of the sample.
- Inferential statistics: infer characteristics of a population given a sample.
  - Bayesian estimation
  - Maximum likelihood estimation
  - Hypothesis testing



Concluding remarks

Descriptive statistics

- Qualitative variable
  - Frequency table
  - Pie-chart, Bar-plot
- Quantitative variable
  - Discretization (binning)
  - Histogram, box-plot
  - Statistics:
    - measures of the location (position) of the data
    - central tendency
    - spread
    - shape
  - Linear transformations
  - Models Inferential statistics

Bibliography I