Theory- 1

Wednesday, December 13, 2023 11:05 AM

- **Probabilities:** The distribution is known
- **Statistics:** The distribution is unknown and we try to find it via a sample
  - **Non-parametric:** We know nothing about the distribution
  - **Parametric:** We know that it belongs to a family of distributions and we **estimate** an unknown parameter

We care about this category

- **Definitions**

Probabilities	Statistics	Comments
Population(N)	Sample(ν)	A sample is drawn from the population via sampling
Parameter (e.g. expected value (μ) variance (σ^2))	Estimation (e.g. sample mean ($\bar{x}$) sample variance (s^2))	**Estimations** of parameters

- Let a sample of size ν, the we have ν independent r.v. all drawn from the same distribution and same parameters

Good / Bad sampling

↓

Good / Bad sample

↓

Good / Bad estimations

↓

Valid/Not-valid results

- Categories of statistics
  - **Quantitive** (#children, income, height)
  - **Qualitative** (color, gender, proffession)

• Representation methods

A) Plots

• Frequency tables

Let a random sample $X_1, \dots, X_v$
Let $y_1, \dots, y_k$ possible values $\Big\}\ k \leq v$

Frequency v_i της τιμής y_i: number of r.v. taking the value y_i

$$v_1 + v_2 + \dots + v_k = v$$

Relative frequency f_i: the corresponding percentage

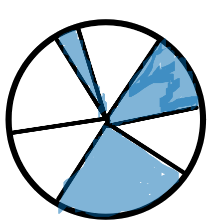
$$f_i = \frac{v_i}{v} \qquad 0 \leq f_i \leq 1 \qquad f_i = \frac{v_i}{v} \cdot 100\%$$

Cumulative frequency N_i: number of r.v. with values less or equal to y_i

Cumulative relative frequency F_i: corresponding percentage

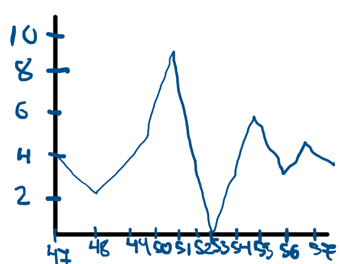
Frequency table: contains some or all of the above

Interval $[a_i, b_i)$	Frequency (v_i)	Relative frequency (f_i)	Cumulative frequency (N_i)	Cumulative Relative Frequency F_i
$-27 \leq HPR < 17$	1	0.025	1	0.025
$-17 \leq HPR < -7$	5	0.125	6	0.15
$-7 \leq HPR < 3$	8	0.2	14	0.35
$3 \leq HPR < 13$	15	0.375	29	0.725
$13 \leq HPR < 23$	9	0.225	38	0.95
$23 \leq HPR < 33$	2	0.05	40	1



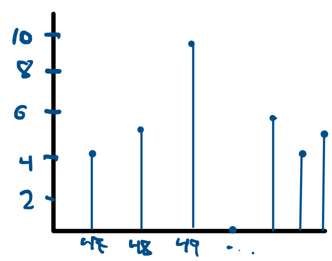
Pie chart

Angle $a_i = f_i \cdot 360°$

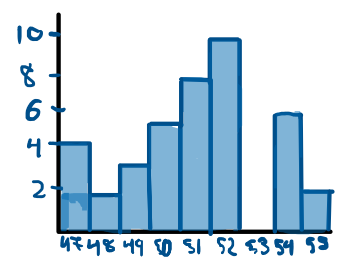


Frequency Polygon

Visualization of the change of frequencies



Bar plot



Histogram

- height of polygon such that area = f_i
- Total area = v (or 1 if relative frequencies)

- In cases that the possible values are too many, we group values to intervals
- Not that many, but not that few either

B) Arithmetic measures

Let $x_1, \dots, x_v$ the values of the sample and $y_1, \dots, y_k$ the possible values and v_i, f_i the corresponding frequencies

- Central tendency measures

- **Mean value:** $\bar{x} = \frac{1}{v} \sum_{i=1}^{v} x_i = \sum_{i=1}^{v} \frac{v_i}{v} \cdot y_i = \sum_{i=1}^{v} f_i \cdot v_i$

  equal

- **Mode:** Value with the largest frequency

  $M_o = \max(x_i \mid \forall j: v_j \leq v_i)$

- **Median:** The value splitting the sample into two equal parts

  Sorted sample $\longrightarrow x_{(1)}, x_{(2)}, \dots, x_{(v)}$

  $$\delta = \begin{cases} x_{(v)} & , v \text{ odd} \\ \frac{x_{(v)} + x_{(v+1)}}{2} & , v \text{ even} \end{cases}$$

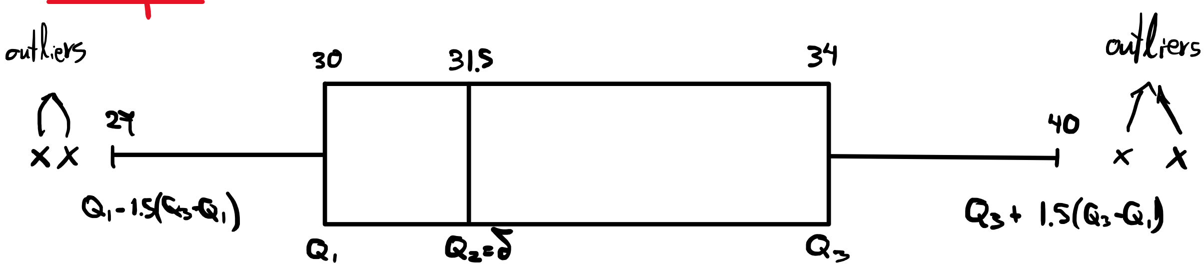
- **Percentile:** the a-percentile is the value for which the $a \cdot 100\%$ of the values are less or equal to this, while the $(1-a)100\%$ is larger
  - $P_{0.25} = Q_1$: first quantile
  - $P_{0.5} = Q_2 = \delta$: second quantile
  - $P_{0.75} = Q_3$: third quantile

- Variability measures

- **Range R:** Difference of largest value to the smaller
- **Interquantile range:** $Q_3 - Q_1$
  - Contains 50% of the values
  - $Q_3 - Q_1 \downarrow \Rightarrow$ large concentration of the values $\Leftrightarrow$ small variance
- **Mean deviation:** $MD = \frac{1}{v} \sum_{i=1}^{v} |x_i - \bar{x}|$

- Mean deviation: $MD = \frac{1}{v}\sum_{i=1}^{v}|x_i - \bar{x}|$
    - $MD \downarrow \Rightarrow$ high concentration around the mean
- Sample variance: $s^2 = \frac{1}{v-1}\sum_{i=1}^{v}(x_i - \bar{x})^2$
- Sample standard deviation: $s = \sqrt{s^2}$
- Gini index: $d = \frac{1}{v^2}\sum_{i=1}^{v}\sum_{j=1}^{v}|x_i - x_j|$
    - Mean difference of pairs (uniformity measures)
- Coefficient of variation: $CV = \frac{s}{\bar{x}}$
    - If $CV \leq 0.1 \Rightarrow$ uniform

- Box plot



Box plot length: R

Box length: $Q_3 - Q_1$: interquantile range

C) Pairs of statistic data

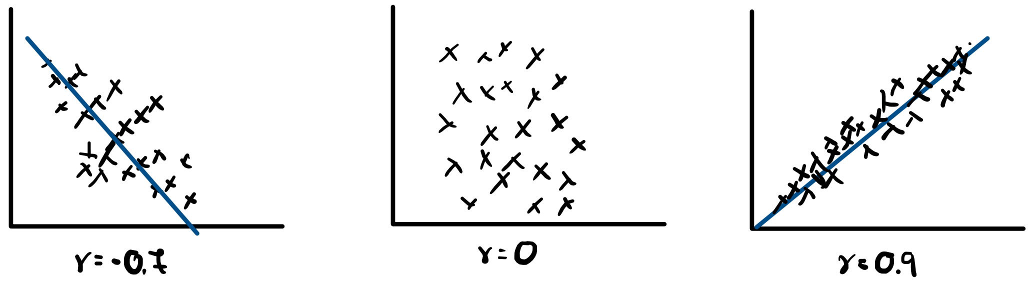
Let a sample of pairs (x_i, y_i) and we explore if the values x_i are related to the values y_i

- Sample correlation coefficient: $r = \frac{\sum_{i=1}^{v}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$

absolute negative correlation $\leftarrow -1 \leq r \leq 1 \rightarrow$ absolute positive correlation

- $b > 0$, a constant and $y_i = bx_i + a \Rightarrow r = 1$
- $b < 0$, a constant and $y_i = bx_i + a \Rightarrow r = -1$
- If r the correlation coefficient of x_i and y_i, then the c.c. of $a + bx_i$ and $c + dy_i$ is also r and

the c.c. of $a+bx$, and $c+dy_i$ is also r and b and d have the same sign.



- **Property:** $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)$$

Exercise 1

Wednesday, December 11, 2024 2:42 PM

Fill the tables:

α)

x	vi	fi	Ni	Fi
1	10	0.2	10	0.2
2	20	0.4	30	0.6
4	17	0.34	47	0.94
6	3	0.06	50	1
Σύνολο	50	1	–	–

$$v = \sum_{i=1} v_i = 10+20+17+3 = 50$$

$$f_i = \frac{v_i}{v} \Rightarrow$$

β)

xi	vi	fi	Ni	Fi
1	12	0.2	12	0.2
2	21	0.35	33	0.55
3	24	0.4	57	0.95
5	3	0.05	60	1
Σύνολο	60	1	–	–

$$v = \frac{v_i}{f_i} = \frac{24}{0.4} = 60$$

$$f_i = \frac{v_i}{v} \Rightarrow \quad v_i = f_i \cdot v \Rightarrow$$

$$v_5 = 0.05 \cdot 60 = 3$$

$$v_2 = 60 - 12 - 24 - 3 = 21$$

Exercise 2

Wednesday, December 13, 2023 11:06 AM

Τιμή (y_i)	Συχνότητα (v_i)
1	9
2	8
3	5
4	5
5	6
6	7

Ο παρακάτω πίνακας συχνοτήτων δίνει τις τιμές που ήρθαν μετά από 40 ρίψεις ενός ζαριού. Να βρεθούν:

The table on the left contains the values of 40 rolls of a dice. Find:

α) the mean			στ) the interquantile range,
β) the median			ζ) the range,
γ) the mode			η) the mean deviation,
δ) the $P_{0.20}$ percentile,			θ) the sample variance, ·
ε) the quantiles,			ι) the sample standard deviation

α) $\bar{x} = \sum_{i=1}^{v} f_i \cdot y_i = \frac{\sum_{i=1}^{v} v_i \cdot y_i}{v} = \frac{1\cdot 9 + 2\cdot 8 + 3\cdot 5 + 4\cdot 5 + 5\cdot 6 + 6\cdot 7}{40} = 3.05$

β) $\delta = \begin{cases} x_{(r)} , & v \text{ odd} \\ \frac{x_{(r)} + x_{(r+1)}}{2} , & v \text{ even} \end{cases}$

$\delta = \frac{3+3}{2} = 3$

γ) $M_0 = 1$

δ) $P_{0,20}$: value for which the 20% of the values lies below this value and the 80% over this value

1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, ... , 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6

Thus $P_{0.20} = 1$

ε) $Q_1 = P_{0.25} = 2$

$Q_2 = P_{0.5} = \delta = 3$

$Q_3 = P_{0.75} = 5$

στ) $Q = Q_3 - Q_1 = 5 - 2 = 3$

ζ) $R = 6 - 1 = 5$

η) $MD = \frac{1}{v} \cdot \sum_{i=1}^{6} v_i |y_i - \bar{y}| = \frac{1}{40} \cdot \sum_{i=1}^{6} v_i |y_i - 3.05| \approx 0.4$

θ) $s^2 = \frac{1}{v-1} \sum_{i=1}^{v} (x_i - \bar{x})^2 = \frac{1}{v-1} \left(\left(\sum_{i=1}^{v} x_i \right)^2 - v \cdot \bar{x}^2 \right) \approx 5.023$

ι) $s = \sqrt{s^2} = \sqrt{5.023} \simeq 2.24$

Exercise 3

Wednesday, December 13, 2023 11:08 AM

Let the temperatures at 14:00 for the first 10 days of June:

32.5, 32.1, 32.3, 31.8, 31.7, 32.4, 32, 31.7, 32.2, 31.9

Calculate the sample variance.

$s^2 = ?$

1st way

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

$$\bar{x} = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i = \frac{1}{10}\sum_{i=1}^{\nu} x_i = 32.06$$

2nd way

x_i	32.5	32.1	32.3	31.8	31.7	32.4	32	31.7	32.2	31.9
-32	.5	.1	.3	-.2	-.3	.4	0	-.3	.2	-.1
y_i: $\times 10$	5	1	3	-2	-3	4	0	-3	2	-1

$$y_i = 10(x_i - 32)$$

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{\sum_{i=1}^{10}(x_i - 32)10}{10} = \sum_{i=1}^{10}\left(\frac{x_i}{10} - 3.2\right)\cdot 10 = (\bar{x} - 32)10$$

$$\bar{y} = \frac{5+1+3-2-3+4+0-3+2-1}{10} = 0.6$$

$$(\bar{x} - 32)10 = \bar{y} \Rightarrow \bar{x} - 32 = \frac{\bar{y}}{10} \Rightarrow \bar{x} = \frac{\bar{y}}{10} + 32 = \frac{0.6}{10} + 32 \Rightarrow$$

$$\boxed{\bar{x} = 32.06}$$

$$\boxed{\bar{x} = 32.06}$$

$$s_y^2 = \frac{1}{v-1}\sum_{i=1}^{v}(y_i-\bar{y})^2 = \frac{1}{v-1}\left(\sum_{i=1}^{v} y_i^2 - v\bar{y}^2\right) \Rightarrow$$

$$\Rightarrow s_y^2 = \frac{1}{10-1}\left[5^2+1^2+3^2+2^2+3^2+4^2+0^2+3^2+2^2+1^2 - 10\cdot 0.6^2\right] =$$

$$= \frac{1}{9}(78-3.6) \simeq 8.26$$

$$\boxed{y_i = ax_i + b \Rightarrow s_y^2 = a^2 s_x^2}$$

Thus: $y_i = 10x_i - 320 \Rightarrow s_y^2 = 10^2 s_x^2 \Rightarrow s_x^2 = \frac{s_y^2}{100} = \frac{8.26}{100} \Rightarrow \boxed{s_x^2 = 0.0826}$

Exercise 4

Wednesday, December 13, 2023 11:08 AM

Let that at the exams of 2018 for the lesson of Probabilities the mean grade was 7 and the mean standard deviation was 1.8, while in 2019 the mean grade was 6 and the sample standard deviation was 0.5. In which year these was more uniformity?

Coefficient of Variation: $CV = \frac{s}{\bar{x}}$

$\bar{x}_1 = 7 \qquad \bar{x}_2 = 6$

$s_1 = 1.8 \qquad s_2 = 0.5$

$CV_1 = \frac{s_1}{\bar{x}_1} = \frac{1.8}{7} \simeq 0.26$ or $CV_1 = 26\%$

$CV_2 = \frac{s_2}{\bar{x}_2} = \frac{0.5}{6} \simeq 0.083$ or $CV_2 = 8.3\%$

The sample with less CV is more uniform

Thus, $CV_2 < CV_1 \Rightarrow$ In 2019 more uniformity