Mathematical Statistics

Anna Janicka

Lecture I, 22.02.2021

DESCRIPTIVE STATISTICS, PART I

Technicalities

- Contact: ajanicka@wne.uw.edu.pl
- Office hours: Mondays after the lecture
- Course materials:

  wne.uw.edu.pl/**ajanicka**
- Mandatory readings: Lecture notes, Wackerly, Mendenhall, Scheaffer (library and online)
- Problem sets: **web page**
- **Homework sets: web page**

Rules

1. Presence during lectures expected. Those who skip the lecture must go through the material themselves.
2. The exam will cover material from the lecture and classes.
3. Presence during classes is mandatory (at most 2 absences)
4. Class grade: tests, homeworks & activity.
5. Exam: for all those who **passed classes.**
6. Exam: 8 problems, 2 points each.
   Exam grade = (number of exam points)/3
7. Final grade= 1/3* class grade +2/3* exam grade, rounded.

What to expect

- Course materials, problem sets, examples, old exams, etc. on the web page.
- Links to everything on moodle

What we will do during the semester

- *Index numbers*
- Descriptive statistics
- Statistical model, statistical inference, notion of a statistic
- Estimation. Estimator properties
- Verification of hypotheses, different kinds of tests
- Bayesian statistics

Plan for today

1. Introduction
2. Descriptive statistics:
   - basic terms
   - data presentation
   - sample characteristics
   - measures
     - central tendency

What is the difference between Statistics and Mathematical Statistics?

Statistics: gathering and analyzing data on *mass* phenomena

historically: ancient times, various censuses, a description of the state

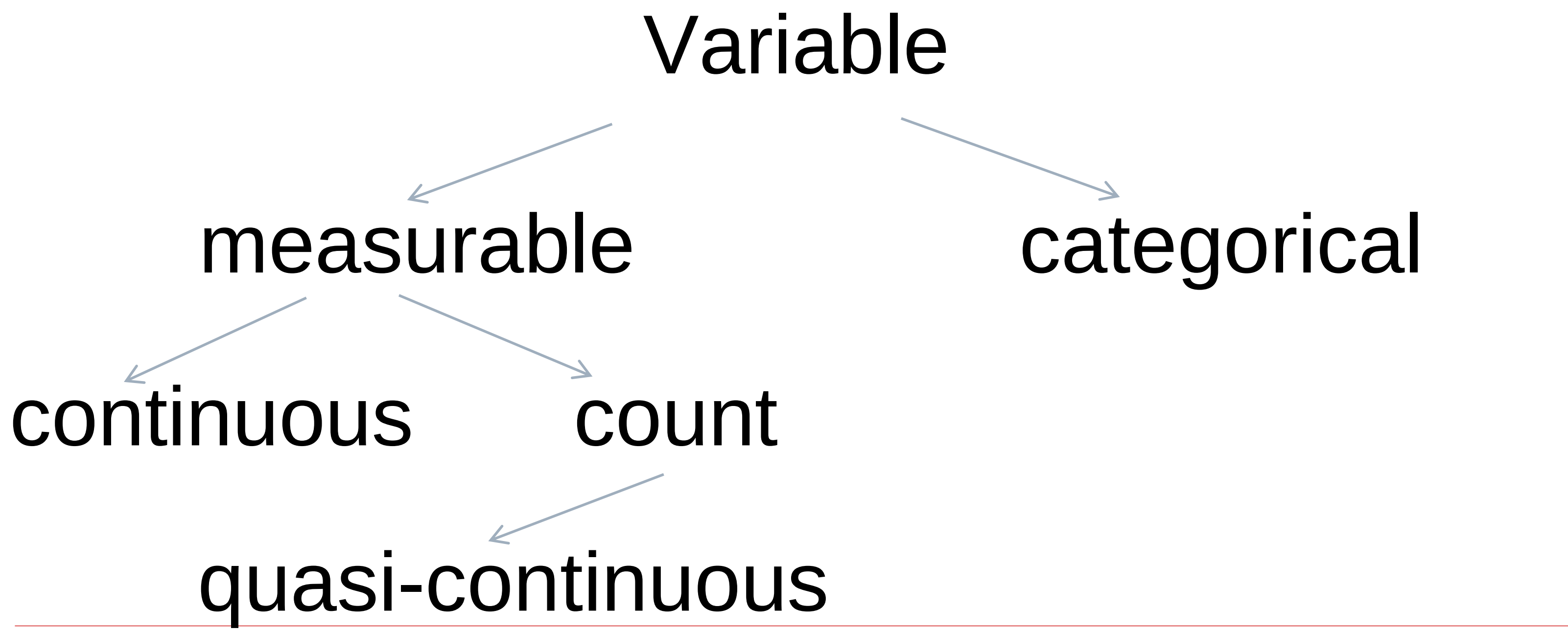
Mathematical Statistics: Statistics from a mathematical standpoint, i.e. a field of applied mathematics used to describe and analyze phenomena with mathematical tools, mainly probability theory

historically: with the beginning of probability calculus: Pascal, Fermat, Gauss

Descriptive Statistics

Quantitative description of data.

Data = ***sample*** from a ***population***, for which a ***variable*** (or variables) are studied



Study

- **full** – concerns the full population
- **representative** – part of the population; the sample ≠ population

  in the latter case, inference about the whole population requires assumptions and the use of probability calculus tools

Presentation of data

- Aim: visibility
- depends on the characteristics of the variable

- tabular
- graphical

Example 1 – count variable

Some class grades for a FoES course

(248 individuals)

3 3 3.5 2 3.5 4 2 2 2 2 3 2 4 2 4 2 4 3.5 4 3 3 3.5 3.5 3.5 3
3.5 3 4 3 3 2 2 2 2 3.5 2 2 3 3 3.5 4.5 3.5 4 3 3 3 3.5 4 3.5
3.5 4 4 2 3.5 3.5 2 2 3.5 3.5 3 3 3 2 3 3.5 2 3 3 3.5 2 4 5 3 2
3 3 3 3 3 3 5 3 3 4.5 3 3.5 2 3 3.5 3.5 3.5 2 4.5 3 2 2 2 3 3
3.5 3.5 5 2 4 5 4 2 4 3 3 3 4 2 3 2 3.5 2 3.5 3 2 2 3.5 3 2 3.5
3.5 3 2 3.5 4 3.5 3 4.5 2 2 3.5 3 2 5 4 2 3 3 3 3.5 3.5 3 2 3.5
4 4 3 2 4.5 2 3 2 2 2 3 3 3 5 3 3 2 4 2 5 4 3 3.5 3.5 2 2 3 3 3
3.5 2 2 2 2 3.5 5 3.5 3 3 3 4 3 3 3 5 4 4.5 4 3.5 4.5 3 3.5 3 3
3 3.5 2 3.5 4 3 2 2 2 3 2 4.5 4.5 4 4 2 4 2 3 3 3.5 3 2 3 3 4
3.5 2 2 2 5 3.5 4 2 2 2 2 2 2 2 2 4 3 2 2

Frequency tables

Single value

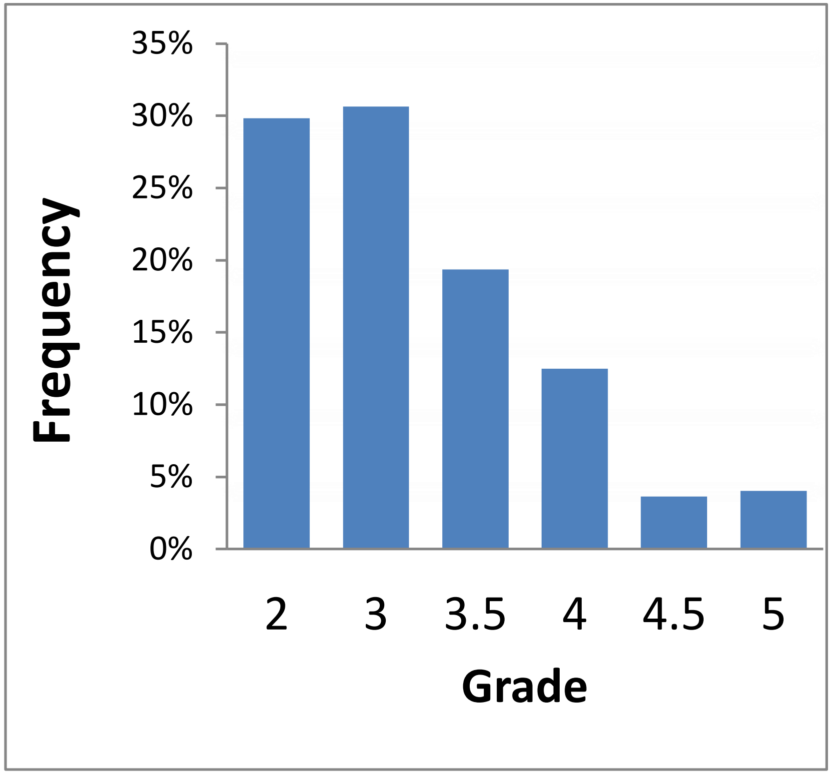
Value	*Number*	*Frequency*
x_1	n_1	$f_1=n_1/n$
x_2	n_2	$f_2=n_2/n$
x_3	n_3	$f_3=n_3/n$
...	...	...
x_k	n_k	$f_k=n_k/n$
Total	n	1

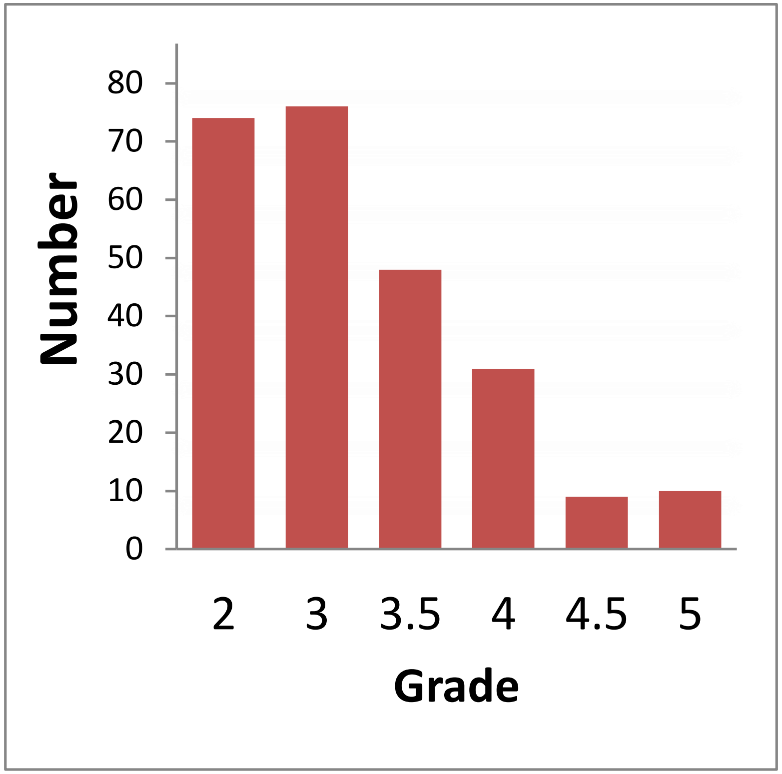
Example 1 – cont.

Grade	*Number*	*Frequency*
2	74	29.84%
3	76	30.65%
3.5	48	19.35%
4	31	12.50%
4.5	9	3.63%
5	10	4.03%
Total	248	100%

[Mean – examples]
[Median – examples]
[Mode – examples]
[Quartile – examples]

Example 1 – cont. (2).
Bar charts of numbers and frequencies

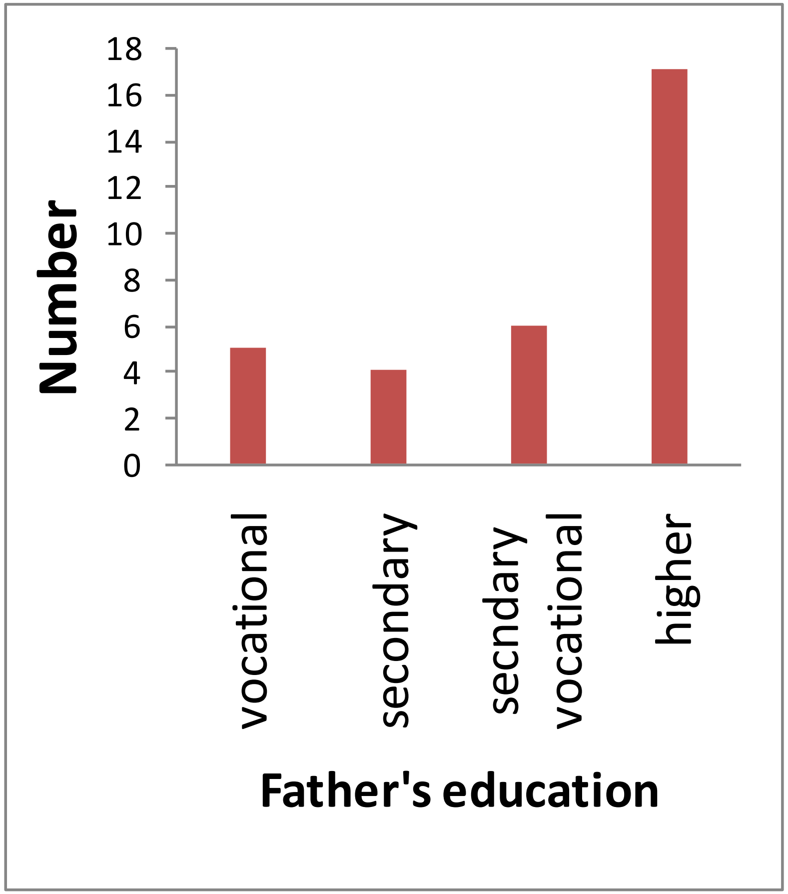




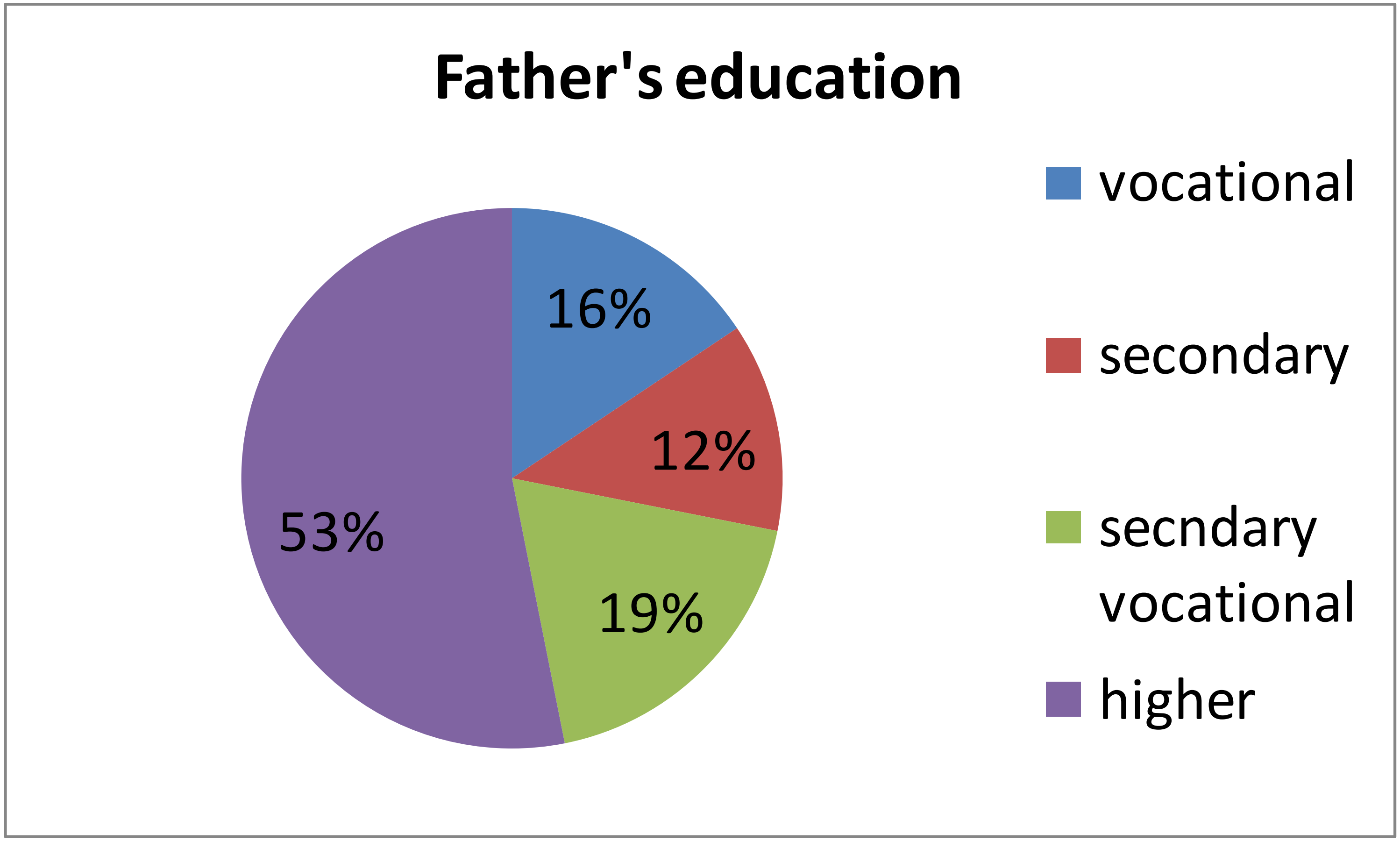
Example 2 – categorical variable

Father's educational attainment for a sample of 32 students

Father's education	*Number*	*Frequency*
vocational	5	0.16
secondary	4	0.13
secondary vocational	6	0.19
higher	17	0.53
Total	32	1.00



Example 2 – cont.
Pie chart



Example 3 – continuous or quasi-continuous variable

Apartment surface area, *n*=100

32.45	33.21	34.36	35.78	37.79	38.54	38.91	38.96	39.50	39.67
39.80	41.45	41.55	42.27	42.40	42.45	44.25	44.50	44.70	44.83
44.90	45.10	45.90	46.52	47.65	48.10	48.55	48.90	49.00	49.24
49.55	49.65	49.70	49.90	50.90	51.40	51.50	51.65	51.70	51.80
51.98	52.00	52.10	52.30	53.65	53.89	53.90	54.00	54.10	55.20
55.30	55.56	55.62	56.00	56.70	56.80	56.90	56.95	57.13	57.45
57.70	57.90	58.00	58.50	58.67	58.80	59.23	63.40	63.70	64.20
64.30	64.60	65.00	66.29	66.78	67.80	68.90	69.00	69.50	73.20
76.80	77.10	77.80	78.90	79.50	82.70	83.40	84.50	84.90	85.00
86.00	89.10	89.60	93.00	96.70	98.78	103.00	107.90	112.70	118.90

Source: A. Boratyńska, *Wykłady ze statystyki matematycznej*

Grouped frequency table

Interval	*Class mark*	*Number of. obs.*	*Frequency*	*Cumulative number* cn_i	*Cumulative frequency* cf_i
$(c_0,c_1]$	$\bar{c}_1$	n_1	$f_1=n_1/n$	n_1	f_1
$(c_1, c_2]$	$\bar{c}_2$	n_2	$f_2=n_2/n$	n_1+n_2	f_1+f_2
$(c_2, c_3]$	$\bar{c}_3$	n_3	$f_3=n_3/n$	$n_1+n_2+n_3$	$f_1+f_2+f_3$
...		...	...		
$(c_{k-1}, c_k]$	$\bar{c}_k$	n_k	$f_k=n_k/n$	$\Sigma n_i = n$	$\Sigma f_i=1$
Total		n	1		

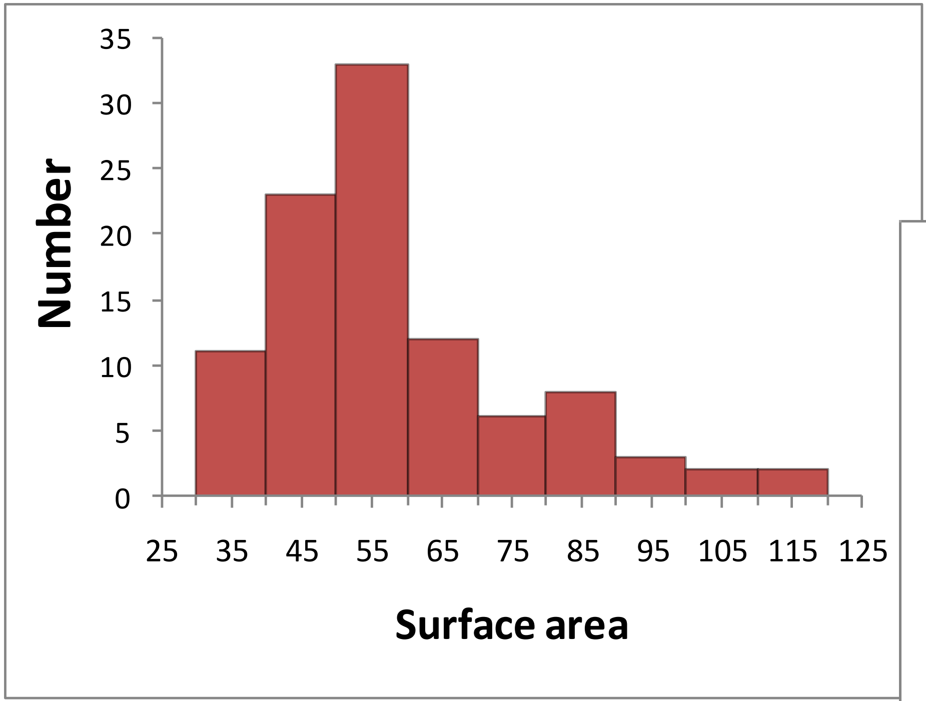
Choice of classes (interval ranges, bins): usually equal length or similar frequency

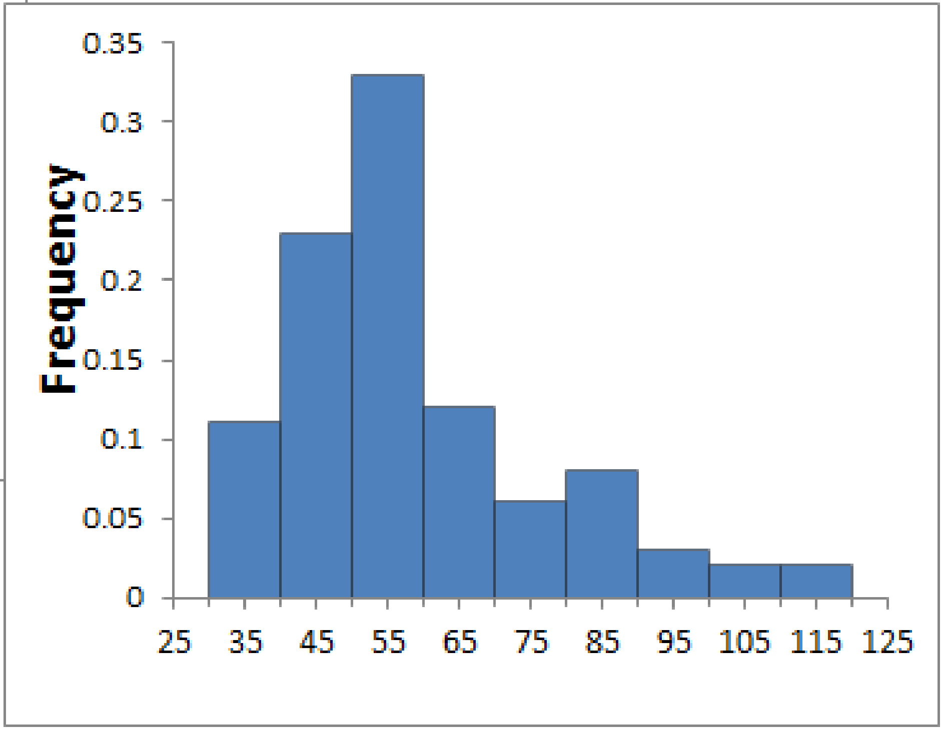
Example 3 – cont.

Interval	*Class mark*	*Number*	*Frequency*	*Cumulative number* cn_i	*Cumulative frequency* cf_i
(30,40]	35	11	0.11	11	0.11
(40,50]	45	23	0.23	34	0.34
(50,60]	55	33	0.33	67	0.67
(60,70]	65	12	0.12	79	0.79
(70,80]	75	6	0.06	85	0.85
(80,90]	85	8	0.08	93	0.93
(90,100]	95	3	0.03	96	0.96
(100,110]	105	2	0.02	98	0.98
(110,120]	115	2	0.02	100	1.00
Total		100	1		

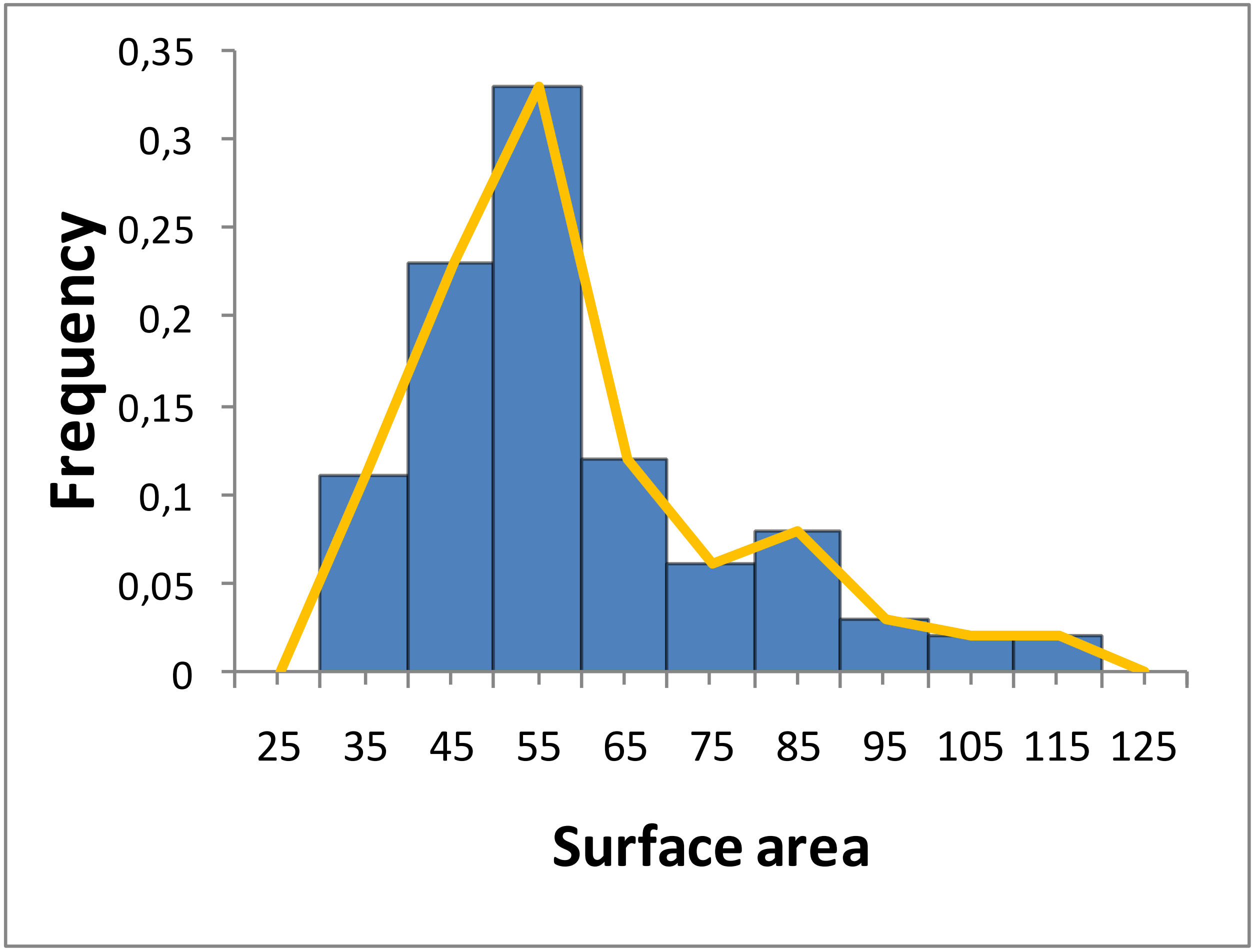
Mean – example
Median – example
Mode – example
Quartile – example
Variance – example

Example 3 – cont. (2)
Number histogram, frequency histogram

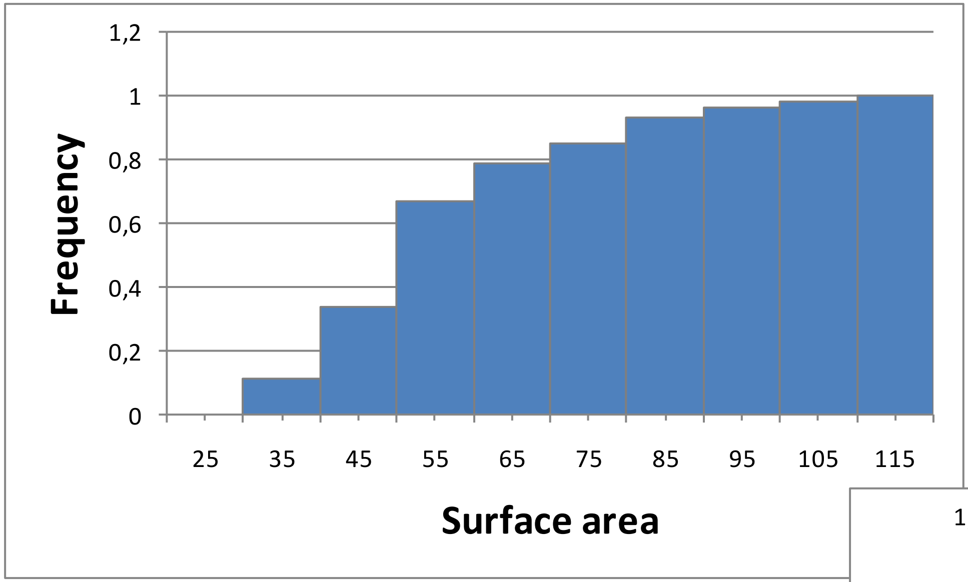


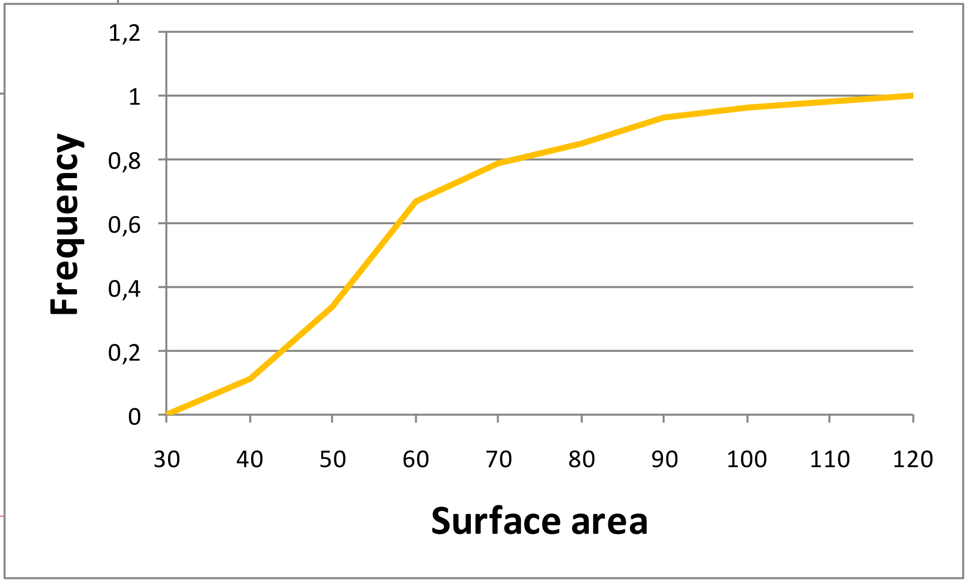


Example 3 – cont. (3)
Frequency histogram and frequency polygon

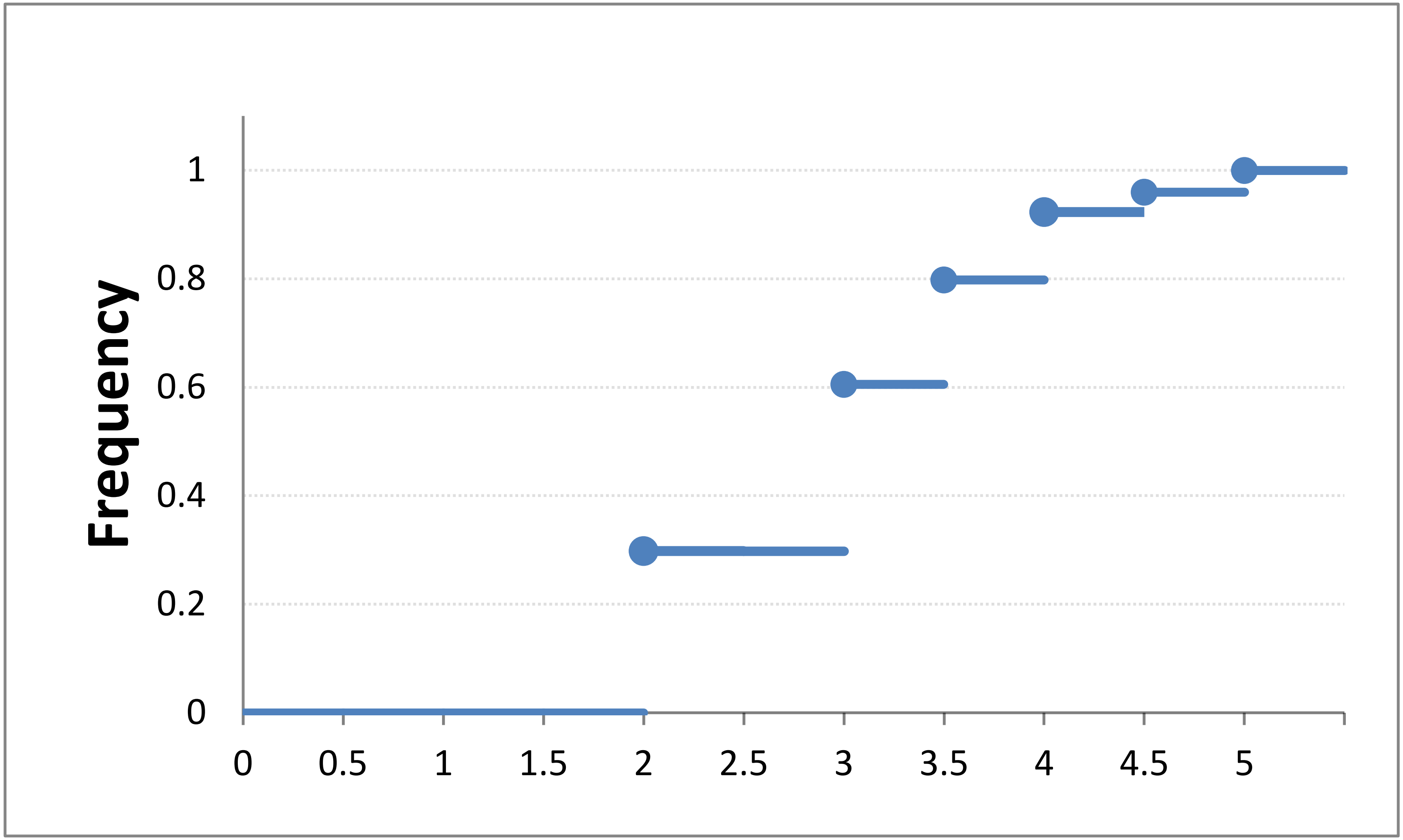


Example 3 – cont. (4) Cumulative frequency histogram and cumulative frequency polygon





Example 1 – cont. (3)
Empirical CDF



Sample characteristics

Describe different properties of measurable variables

Measures of

- central tendency
- variability (dispersion, spread)
- asymmetry
- concentration

Types:

- based on moments – classic
- based on measures of position

Central tendency

- Classic:
  - arithmetic mean
- Position (order, rank):
  - median
  - mode
  - quartile

Arithmetic mean

- raw data:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- grouped data:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$$

- grouped class interval data:

$$\overline{X} \cong \frac{1}{n}\sum_{i=1}^{k} \overline{c}_i \cdot n_i$$

Arithmetic mean – examples

Example 1 – cont.

Example 1:

$$\overline{X} = \frac{2\cdot 74 + 3\cdot 76 + 3.5\cdot 48 + 4\cdot 31 + 4.5\cdot 9 + 5\cdot 10}{248} \approx 3.06$$

Example 3:

Example3 – cont.

$$\overline{X} \cong$$

$$\cong \frac{35\cdot 11 + 45\cdot 23 + 55\cdot 33 + 65\cdot 12 + 75\cdot 6 + 85\cdot 8 + 95\cdot 3 + 105\cdot 2 + 115\cdot 2}{100}$$

$$= 58.7$$

while in reality: $\overline{X} = 59.58$

only if raw data not available

Median

(any) number such that at least half of the observations are less than or equal to it and at least half of the observations are greater than or equal to it

- raw data:

$$Med = \begin{cases} X_{\frac{n+1}{2}:n} & n \text{ odd} \\ \frac{1}{2}\left(X_{\frac{n}{2}:n} + X_{\frac{n}{2}+1:n}\right) & n \text{ even} \end{cases}$$

where $X_{i:n}$ is the ***i*-th order statistic**, i.e. the i-th smallest value of the sample

Median – cont.

- for grouped class interval data:

$$Med \cong c_L + \frac{b}{n_M}\left(\frac{n}{2} - \sum_{i=1}^{M-1} n_i\right)$$

where:

M – number of the median's class

c_L – lower end of the median's class interval

b – length of the median's class interval

Median – examples

[Example 1 – cont.](#)

Example 1:

$$Med = \frac{X_{124:248} + X_{125:248}}{2} = 3$$

[Example 3 – cont.](#)

Example 3:

M=3, n_3=33, c_L=50, b=10

$$Med \cong 50 + \tfrac{10}{33}(50 - 34) \approx 54.85$$

in reality: Med = 55.25

Mode

the value that appears most often

- for grouped data:

  Mo = most frequent value

- for grouped class interval data:

$$Mo \cong c_L + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot b$$

where

n_{Mo} – number of elements in mode's class,

c_L, b – analogous to the median

Mode – examples

Example 1 – cont.

Example 1:

$Mo = 3$

Example 3 – cont.

Example 3:

the mode's interval is (50,60], with 33 elements

$n_{Mo} = 33$, $c_L = 50$, $b = 10$, $n_{Mo\text{-}1} = 23$, $n_{Mo+1} = 12$

$$Mo \cong 50 + \frac{33 - 23}{(33 - 23) + (33 - 12)} \cdot 10 \approx 53.23$$

Which measure should we choose?

- Arithmetic mean: for typical data series (single max, monotonous frequencies)
- Mode: for typical data series, grouped data (the lengths of the mode's class and neighboring classes should be equal)
- Median: no restrictions. The most robust (in case of outlier observations, fluctuations etc.)

Quantiles, quartiles

- *p*-th quantile (quantile of rank *p*): number such that the fraction of observations less than or equal to it is at least *p*, and values greater than or equal to it at least 1-*p*

- Q_1 : first quartile = quantile of rank ¼
- Second quartile = median

  = quantile of rank ½
- Q_3: Third quartile = quantile of rank ¾

Quantiles – cont.

Empirical quantile of rank *p*:

$$Q_p = \begin{cases} \dfrac{X_{np:n} + X_{np+1:n}}{2} & np \in Z \\ X_{[np]+1:n} & np \notin Z \end{cases}$$

Quartiles – cont.

- Quantiles for p = ¼ and p = ¾.
- For grouped class interval data – analogous to the median

$$Q_k \cong c_L + \frac{b}{n_{M_k}}\left(\frac{k\cdot n}{4} - \sum_{i=1}^{M_k - 1} n_i\right)$$

for k=1 or 3

where M_1, M_3 – number of the quartile's class

b – length of quartile class interval

c_L – lower end of the quartile class interval

Quartiles – examples

Example1 – cont.

Example 1:

$$248 \cdot 1/4 = 62 \qquad 248 \cdot 3/4 = 186$$

Example 3 – cont.

so

$$Q_1 = \frac{X_{62:248} + X_{63:248}}{2} = 2, \qquad Q_3 = \frac{X_{186:248} + X_{187:248}}{2} = 3.5$$

Example 3:

$$100 \cdot 1/4 = 25 \qquad 100 \cdot 3/4 = 75$$

$$M_1 = 2, \quad M_3 = 4 \quad \text{so}$$

$$Q_1 \cong 40 + \frac{10}{23}(25 - 11) \approx 46.09 \qquad Q_3 \cong 60 + \frac{10}{12}(75 - 67) \approx 66.67$$