AdTrans project , 2023-1-PL01-KA220-HED-000158917

ADTRANS
ADVANCED DATA ANALYSIS FOR TRANSPORTATION AND MOBILITY ACADEMIC TEACHERS

Mathematical Statistics
Development of statistical material

Anna Gobis

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

Introduction

- **Mathematical statistics** deals with methods that allow us to learn about certain properties of a population based on the study of only that part of the population.
- Generalizing the results of partial research to the entire population is called **statistical inference** .
- Limiting oneself to partial studies results from the costs or time of the research, and sometimes it is not possible to study the entire population. Another situation is when a certain process can only be observed in a limited time period.

Population, sample, statistical unit

- **A population (collective)** is a set of any elements related to each other (having a common feature) but not identical from the point of view of the feature being studied. The population includes all elements that are subject to statistical observation (are subject to statistical study):
  - One-dimensional population – one feature of the population is examined,
  - Multidimensional community – many characteristics are considered in the study.
- **Sample** – a finite subset of the general population that is directly subject to study with respect to a specific characteristic.
  - Random sample – whose elements have been selected at random. Each unit has the same probability of being included in the sample.
- **Statistical unit** – the smallest unit of the statistical community covered by the study, a single observation

Example

- The student government decided to check what PG students think about the civic projects selected for implementation at the university. 100 people were randomly selected from the student database and interviews were conducted with them.
  - A) In this study, the population is all PG students whose names are in the university database; the sample consists of randomly selected 100 people from this population.
  - B) In this study, the population is randomly selected students (100 people). The sample size cannot be determined.
  - C) In this study, the population is the student government and the sample is 100 randomly selected students.

Statistical features

- **Statistical features** are divided into **constants** (common to all observations) and **variables** that are properties of the statistical units that make up the studied population.
- **Constant features** are statistical features that in a given statistical study constitute a common property of the population (they do not differentiate the units studied). They decide whether a unit is classified in a given community.
  - Material features – they define what is the subject of the study
  - Spatial features – determine where the studied units were located
  - Time features – define the moment when the study was conducted
- **Variable features are divided into measurable and immeasurable.**
  - Measurable (quantitative) – capable of being expressed numerically (e.g. height, age, income)
    - continuous – can take any value from a certain numerical range, e.g. height, weight
    - discrete – they take a finite (usually small) or countable number of values, e.g. number of children, number of employees
    - ordinal – e.g. education, size of city (large, medium, small)
  - Unmeasurable (qualitative) – not quantifiable, expressed descriptively (e.g. national origin, place of residence, gender)

Example

1) The study aimed to investigate the division of transport tasks among first-year PG students. The study covered 800 first-year PG students who took part in a survey of transport preferences and behaviours .
2) There are 2,500 first-graders in primary schools in Gdańsk. The growth of first-graders was analyzed. The study was conducted on a sample of 1,000 first-graders.
3) The statistical study concerns the monthly salaries of supermarket employees.
    - Population?
    - Attempt?
    - Statistical unit?
    - Variable and its type?

Preparation of statistical material

- **Detailed series** - an ordered sequence of observed values of the statistical feature being studied
- **Distribution series** – statistical material divided into groups (classes) according to a selected criterion, recorded in tabular form, specifying the number of people in each of the distinguished groups.
  - Division according to measurable characteristics:
    - Point
    - Compartment
  - Division according to non-measurable features

Preparation of statistical material

- **Detailed series** - an ordered sequence of observed values of the statistical feature being studied
- **Distribution series** – statistical material divided into groups (classes) according to a selected criterion, recorded in tabular form, specifying the number of people in each of the distinguished groups.
  - Division according to measurable characteristics:
    - Point
    - Compartment
  - Division according to non-measurable features

Resolution and detail series

Detailed series:

76; 81; 83; 85; 87; 91; 93; 94; 95; 97; 99; 104; 116; 118; 119; 122; 123; 125; 126; 127; 128; 128; 129; 133; 133; 135; 135; 136; 144; 146; 146; 155; 158; 159; 161; 167; 178; 179; 179; 182; 184; 184; 193; 198; 200.

Distribution series:

Compartment	
[68,84)	3
[84,100)	8
[100,116)	5
[116,132)	15
[132,148)	17
[148,164)	9
[164,180)	6
[180,196)	4
[196,212)	2

Preparation of statistical material

- *n* - sample size (**we assume the sample is small if n < 30**)
- $x_1, x_2, \cdots, x_n$ - successive observed values of the studied feature
- *x{(1)}, x{(2)}, ⋯, x{(n)}* - observed values of the studied feature arranged in ascending order
- *k* - number of classes (intervals) of the distribution series $k \approx \sqrt{n}$
- *i* - class (interval) width: $i = \frac{x_{max} - x_{min}}{k}$
- n_i - class size
- c_i - frequency: $c_i = \frac{n_i}{n}$

Example 1.1

- The number of employees in the surveyed companies is:

100; 125; 170; 144; 144; 235; 301; 100; 100; 170; 144; 235; 100;301; 170; 301; 125; 125; 235; 125; 125; 100; 144; 301; 144; 144; 170; 144; 144; 144.

Determine the type of feature. Arrange values into a detailed series. Create a point distribution series.

Example 1.2

- The data are the usable areas of the shops in m^2. Below is an ordered series of detailed feature values:

76; 81; 83; 85; 87; 91; 93; 94; 95; 97; 99; 104; 111; 112; 113; 114; 116; 118; 119; 120; 121; 122; 123; 125; 126; 127; 128; 128; 129; 130; 131; 132; 133; 133; 135; 135; 136; 137; 138; 138; 141; 141; 141; 143; 144; 146; 146; 148; 148; 152; 155; 158; 159; 161; 162; 163; 165; 166; 167; 178; 179; 179; 182; 184; 184; 193; 198; 200.

^{2} wide interval distribution series with the beginning of the first interval of 70 m^2.
Assume the interval is closed on the left side and open on the right side.

Graphical presentation of data

- Charts
  - Columnar
  - Wheeled
  - Linear
  - Cartograms

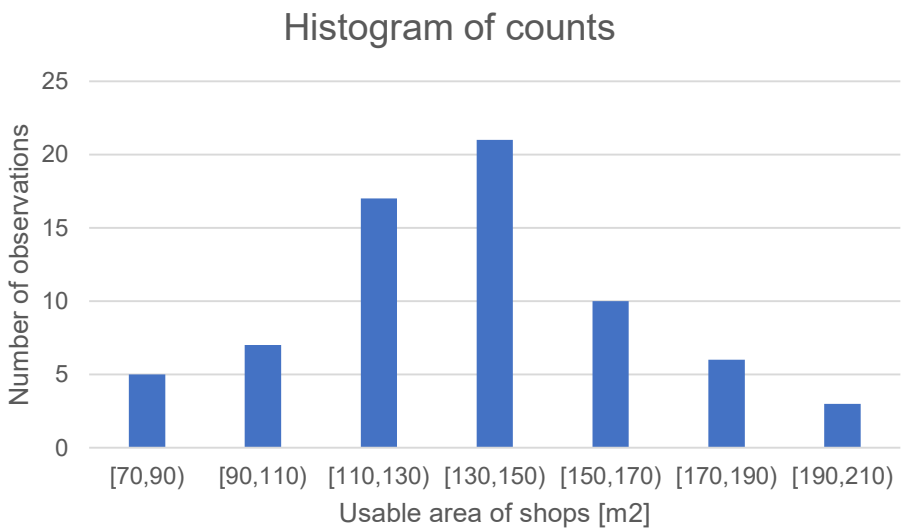
- Presentation of distributive series:
  - Histograms
  - Diagrams (polygons of counts, frequencies)
  - Crooked

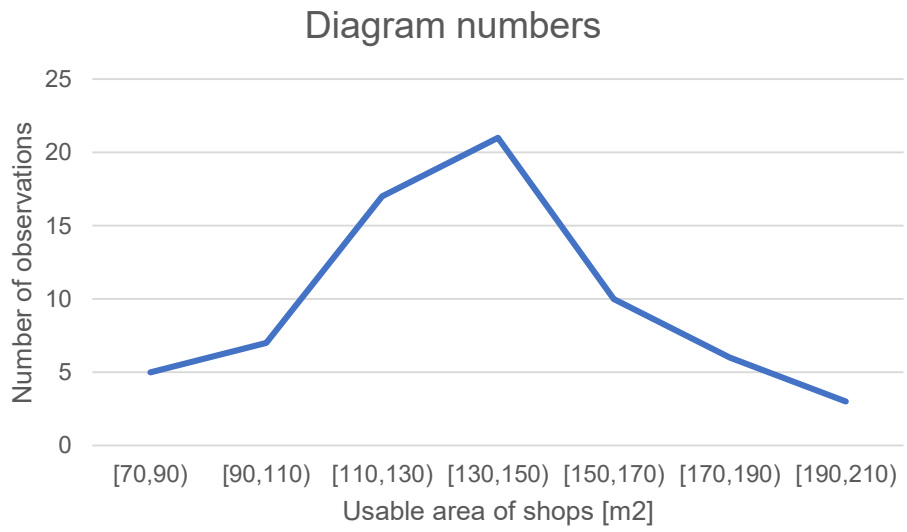
Example 1.3

- The distribution series with the usable area of the shops (previous example) should be illustrated with a histogram and a diagram.
  - Present a histogram of abundance and cumulative abundance.
  - Present a histogram of frequencies and cumulative frequencies.

Example 1.3

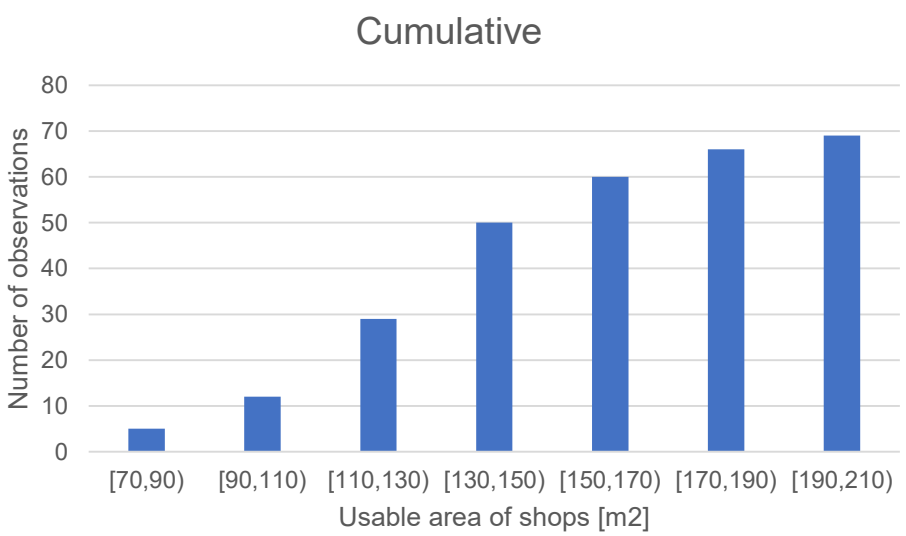
Compartment				
[70.90)	5	5	0.07	0.07
[90,110)	7	12	0.10	0.17
[110,130)	17	29	0.25	0.42
[130,150)	21	50	0.30	0.72
[150,170)	10	60	0.14	0.87
[170,190)	6	66	0.09	0.96
[190,210)	3	69	0.04	1.00
sum	69	-	1.00	-

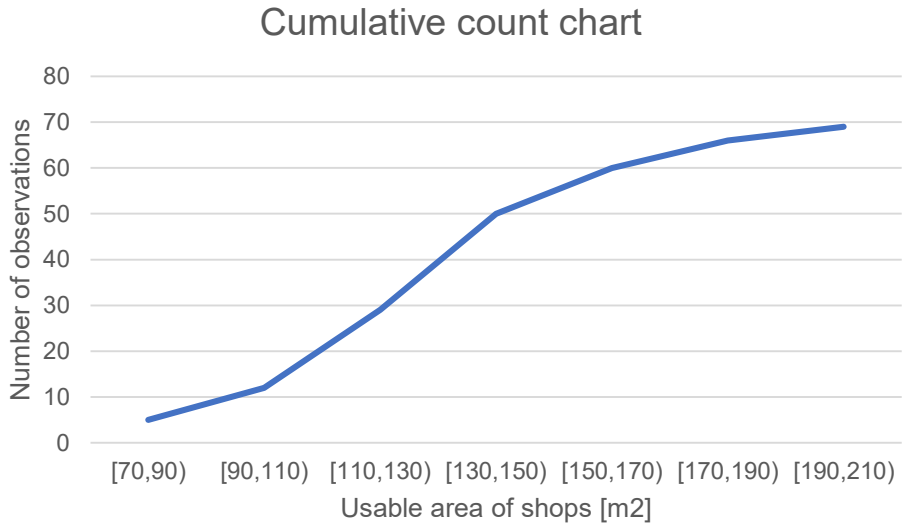




Example 1.3

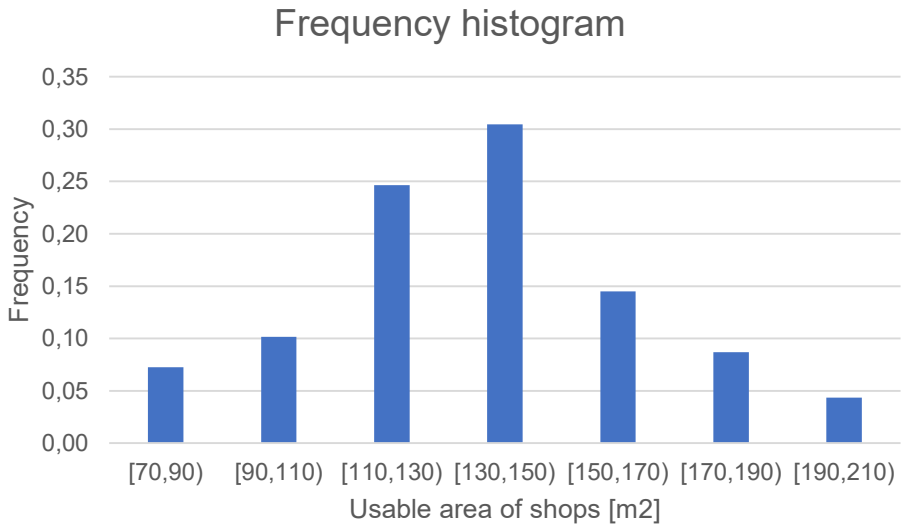
Compartment				
[70.90)	5	5	0.07	0.07
[90,110)	7	12	0.10	0.17
[110,130)	17	29	0.25	0.42
[130,150)	21	50	0.30	0.72
[150,170)	10	60	0.14	0.87
[170,190)	6	66	0.09	0.96
[190,210)	3	69	0.04	1.00
sum	69	-	1.00	-

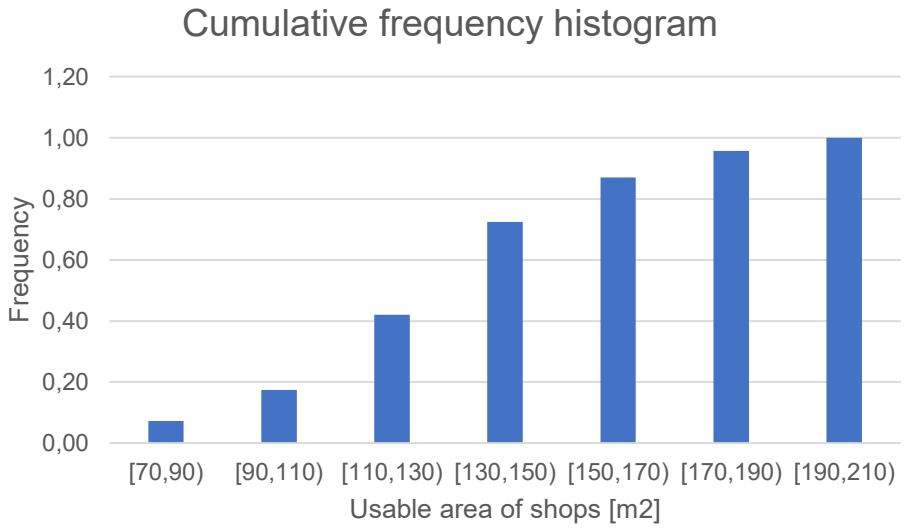




Example 1.3

Compartment				
[70.90)	5	5	0.07	0.07
[90,110)	7	12	0.10	0.17
[110,130)	17	29	0.25	0.42
[130,150)	21	50	0.30	0.72
[150,170)	10	60	0.14	0.87
[170,190)	6	66	0.09	0.96
[190,210)	3	69	0.04	1.00
sum	69	-	1.00	-





Example 1.4 – independent task

- Based on the ready-made point distribution series of the number of registered faults in make x cars in 2015, prepare a count/frequency histogram and a cumulative count/frequency histogram.

Number of faults	Number of cars
0	32
1	119
2	184
3	148
4	124
5 and more	88

Inferential Statistics

- **Hypothesis testing**
  - **Key Words**
    - Compare Processes
    - Equal variances
    - Variance
    - Paired population
    - Mean
  - **Kind of Hypothesis**
    - H0 : Null hypothesis
    - Ha : Alternative hypothesis
  - **Statistical test**
    - Rejection region
    - Acceptance region
  - **Rejection criterion**
    - Unilateral (One Tail Test)
    - Bilateral (Two Tail Test)
  - **Risk of a wrong decision**
    - Error Type I
    - Error Type II
- **Estimate**
  - **By point and by interval**
    - Mean
    - Variance
    - Proportion
- **Probability distributions**
  - **Distribution**
    - **used in quality control**
      - binomial distribution
      - geometric distribution
      - hypergeometric distribution
      - Poisson
    - **derived from sampling**
      - Chi-square distribution
      - Student's t-distribution
      - Distribution F
- **Parameters and Statistics**
  - **Parameters**
    - Representative and descriptive value of a population as ...
      - μ: Mean
      - σ : Standard deviation
  - **Statistics**
    - Actions or functions of the sample data that help characterize the distribution of such data
- **Population and sample**
  - **Population**
    - Finite
      - Population is small and It is possible to measure all individuals
    - Infinite
      - Population is too big and It is impossible or unaffordable to measure all individuals
  - **Sample**
    - Representative Sample
      - 
      - Part of a population that retains the characteristics of this

https://www.mindomo.com/mindmap/irnfeential-statistics-mind-map-cb11cc20a85345489531b8d006bd3666

Additional tasks

Task 1.1

In a certain company the production time of 30 prefabricated elements was measured and the following results were obtained

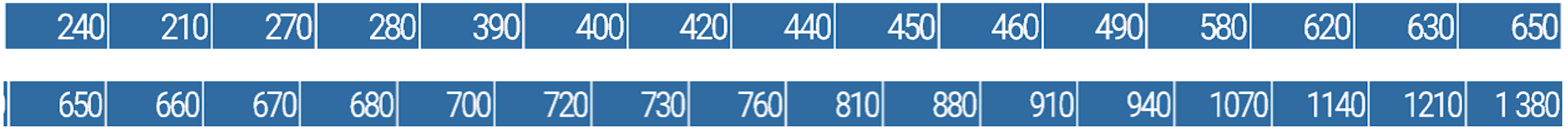
420; 240; 490; 670; 450; 270; 390; 210; 580; 400; 280; 730; 440; 660; 1210; 620; 700; 810; 1070; 680; 940; 760; 630; 880; 650; 1140; 460 1380; 720; 910

a) build a detailed distribution series

b) build a compartment distribution series

Answer 1.1

a)



b)

Range	Frequency
210 – 405	6
406 – 601	6
602 – 797	10
798 – 993	4
994 – 1189	2
1190 – 1385	2

Task 1.2

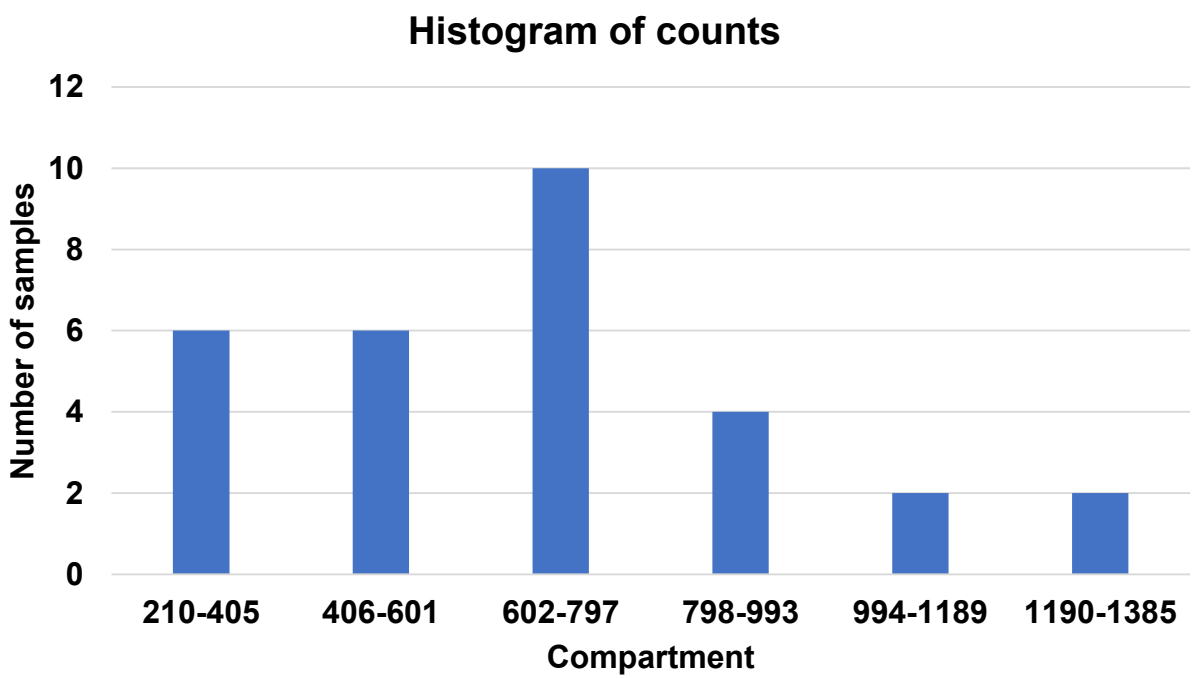
For the data from task 1.1, draw:

a) Histogram of counts

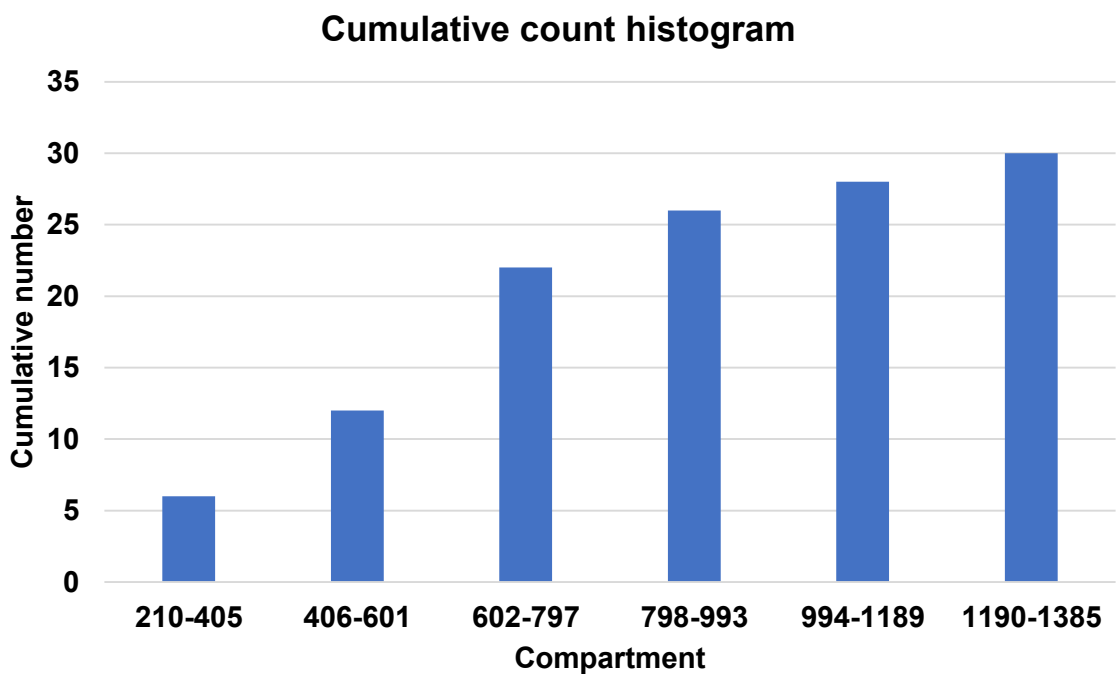
b) Cumulative count histogram

Answer 1.2

and)



b)



Task 1.3

For the data from task 1.1, calculate

a) Frequency

b) Cumulative frequency

Reply 1.3

Production Time Interval	Frequency	Relative Frequency	Cumulative Frequency
210 – 405	6	6/30=0.2	0.2
406 – 601	6	6/30=0.2	0.4
602 – 797	10	10/30=0.33	0.73
798 – 993	4	4/30=0.13	0.86
994 – 1189	2	2/30=0.07	0.93
1190 – 1385	2	2/30=0.07	1.0

Sources:

- Kot, SM, Jakubowski, J., Sokołowski, A.: Statistics. Second revised edition. Difin Publishing House . Warsaw, 2011.
- Stanisz, A.: Accessible statistics course using STATISTICA PL on examples from medicine. Volume 1. Basic statistics. StatSoft Polska Publishing House. Cracow, 2006.
- Flisikowski , K.: Statistical methods in business. Materials of the AdvancedPhD project . Gdańsk University of Technology, 2015.
- Zimny, A.: Descriptive statistics. Supporting materials for exercises. Second revised edition. State Higher Vocational School in Konin, 2010.