Homework 2

Part 1:

1: Whiskers usually extend interquartile range 1.5 times / beyond 1st and 3rd quartiles / for normal distribution.

for non-normal distribution ~~the~~ range can result of extreme points being considered as outliers.

For instance: heavy tailed data;

~~1.5~~ 1.5 * IQR can exclude valid point as an outlier

2: Boxplot → symmetry → skewed data is not symmetric

skewness; multiple peaks can lead to a distorted conclusions; will mask true outliers.

~~Med-couple-based skewness correction;~~

quantile-based outlier detection / less sensitive to distribution

kernel density estimation

3. mean → sensitive to extreme values
average of all values

median → middle value
robust to outliers

Boxplots tend to represent typical value and the spread of the data
↳ minimal impact of extreme values

Multimodal distribution → median can be in a low-density region; it will not represent actual center of a data.

4. Strong right skewness →

→ overlying distribution with long tail extending to the right; most data concentrated to the left

- variance increase;
- skewness coefficient >0
  ↳ statistical models assuming normality become invalid.

log-normal; exponential distribution

5. Boxplots efficiently compare central tendency; spread across groups.

Overlapping distributions or small sample hide multimodality or exaggerate quartile variability

6. few bins -> merge modes.

-> oversmooth the data

many bins -> create noise

-> disturb pattern

-> overfit data.

In KDE bandwidth selection has similar role

too few -> noisy density estimate

too wide -> oversmooth data

7. Histograms → area of each bar represents the frequency of data falling in the bin.

Bar chart → hight of bar shows the frequency or value of specific category.

Bin choice in histogram is more crucial cause it determines how the continues data is grouped ⇒ affects the distribution.

Bar charts show discrete category.

8. Wide bins in case of bimodal distribution. If bin width is larger than separation between two modes., may show single peak.

KDE preserve true density structure.

9. density plots → probability density function; kernel function

↳ smooth representation of distribution

↳ kernel and band with is crucial → Poor choice can lead to over; under smoothing

Sparse data → challenge bandwidth

small → spikiness

large → obscure details

10. area under the curve represents a probability function / tot. probability of all values.

direct comparison of distribution regardless of sample size,

Part 2:

1. ECDF plot.

Data

-5; -2; 0; 3; 4; 5; 5; 6; 7; 7; 8; 9; 9; 10; 12; 15

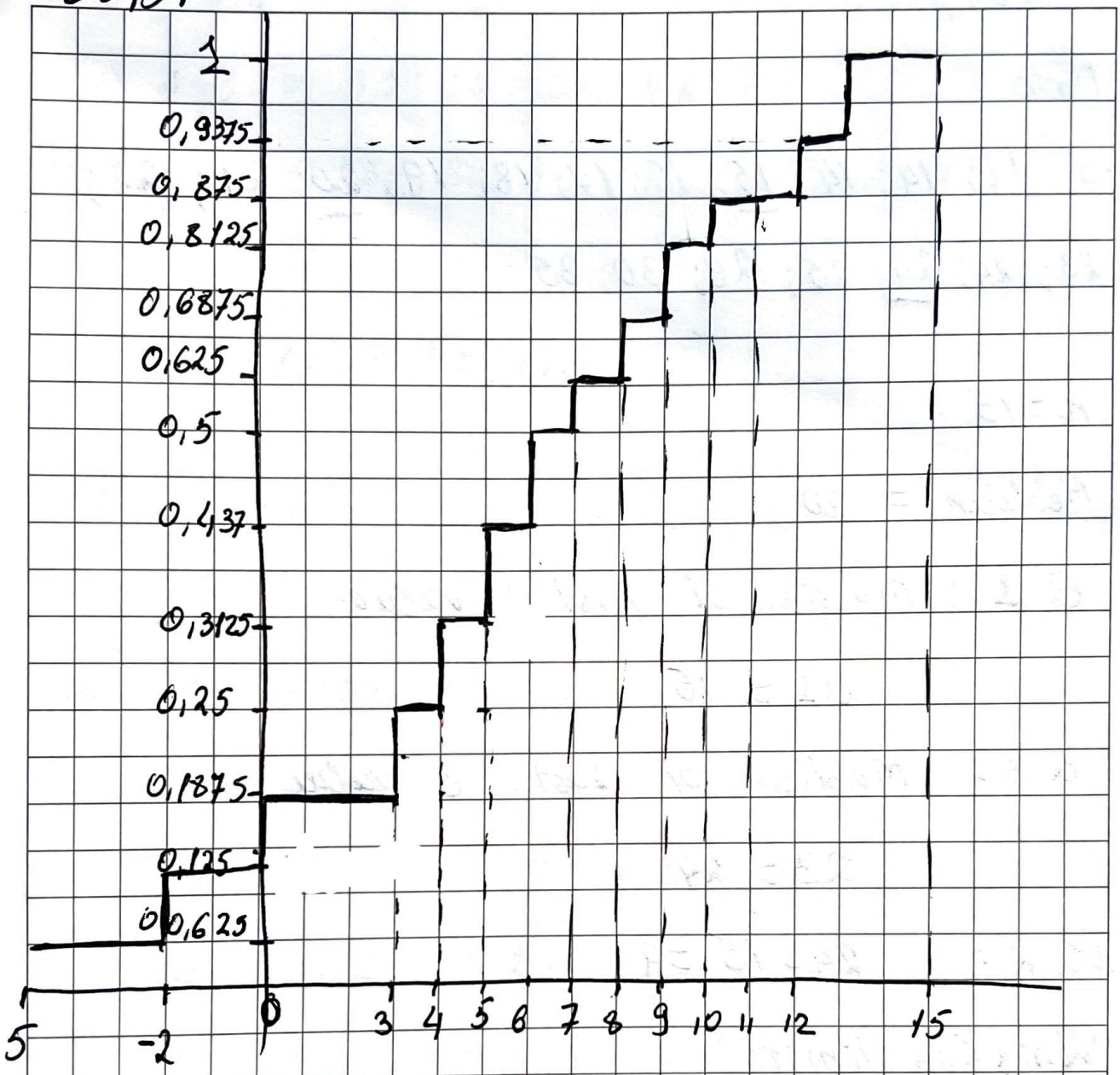
1) sort -> ✓

2) $n = 16$ data points

3)

-5	1/16
-2	2/16
0	3/16
3	4/16
4	5/16
5	7/16
6	8/16
7	10/16
8	11/16
9	13/16
10	14/16
12	15/16
15	16/16

ECDF


1
0,9375
0,875
0,8125
0,6875
0,625
0,5
0,437
0,3125
0,25
0,1875
0,125
0
0,625
5
-2
0
3
4
5
6
7
8
9
10
11
12
15

2. Boxplot:

Data:

-5; 12; 14; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 24; 25; 29; 30; 35

$n = 15$.

Median = 20.

Q1: Median of first 9 values.

$Q1 = 15$

Q3: Median of last 9 value

$Q3 = 24$

IQR: $24 - 15 = 9$

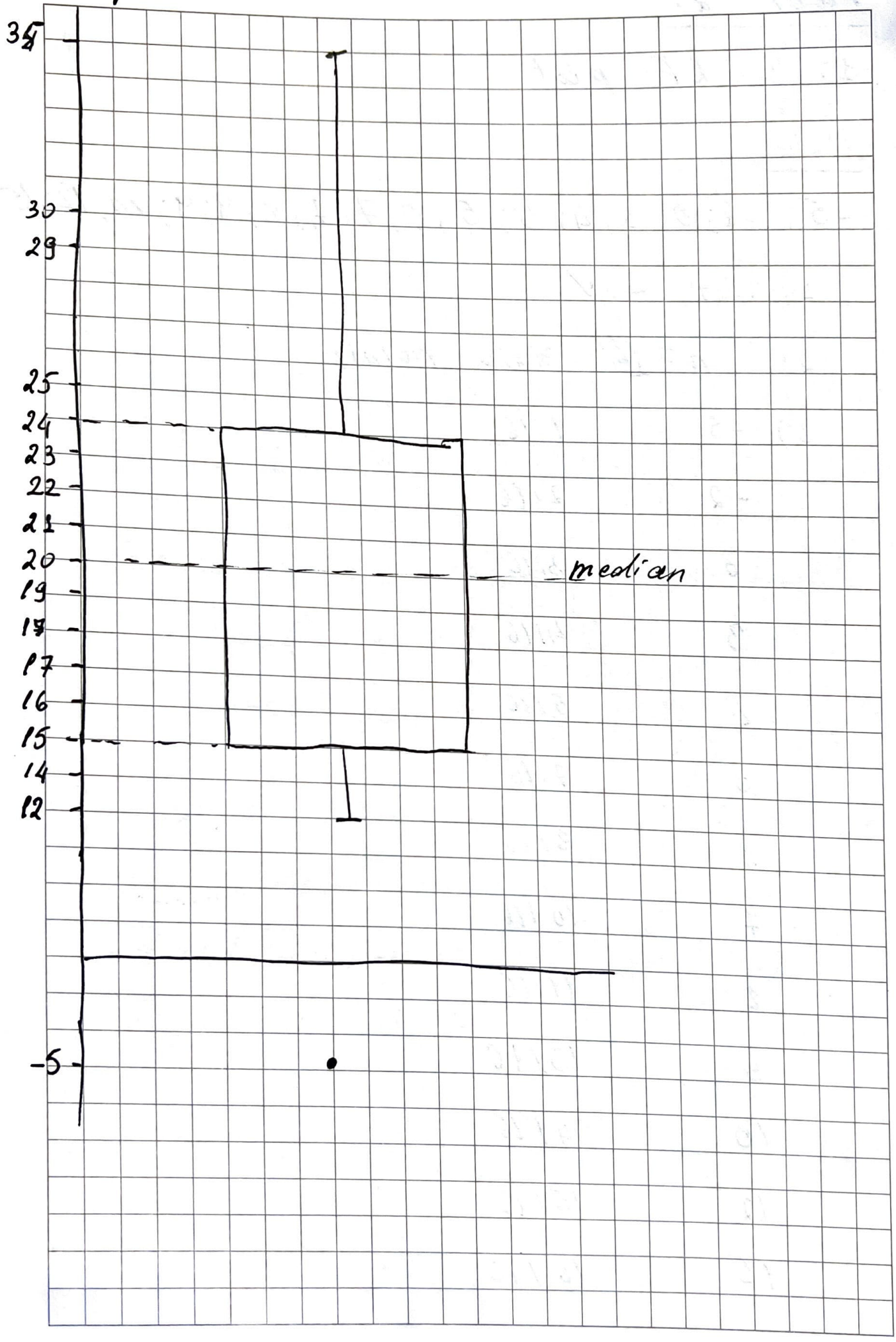
Whisker limits:

L: $Q1 - 1,5 \cdot IQR = 1,5$ -5 outlyer

U: $Q3 + 1,5 \cdot IQR = 37,5$ 35 in range

Outliers: ... -5 ...

Box plot



Histogram

-10; 45; 50; 55; 55; 60; 62; 65; 68; 70; 73; 74; 80; 80; 82; 85; 88; 90; 91; 92; 94; 97; 100; 105

Range : 105 - (-10) = 115

Bin Width: $\frac{\text{Range}}{\text{number of bins}} = \frac{115}{5} = 23$

		Frequency
Bin 1:	-10; -> 13	1
2:	13; -> 36	0
3:	36 -> 59	4
4:	59 -> 82	10
5;	82 -> 105	9

Histogram