Part 1: Theoretical Questions

1. In a dataset with a non-normal distribution and potential extreme values, how are the whiskers in a boxplot determined, and what are the limitations of the standard IQR- based rule in such cases?

To find the lower whisker, we use this formula: $Q_1 - 1.5 \times IQR$
And for the upper whisker, we use this formula: $Q_3 + 1.5 \times IQR$
If we, let's say, have a skewed distribution, this rule might exclude the data points (for example extreme values) and consider them as outliers. These problems arise because the rule assumes that the distribution is normal.

2. Given a dataset with heavy skewness and multiple peaks, how can a boxplot misrepresent outliers, and what alternative methods exist for identifying them more accurately?

The $1.5 \times IQR$ rule assumes that the distribution is symmetric around the median, but if we, for example, have right-skewness, it may classify higher values as outliers, or if we have multiple peaks on the data, the boxplot might place whiskers in regions where there are gaps. An alternative to this problem might be to use density histograms (KDE) or normalize the data.

3. Explain the conceptual difference between median and mean in the context of non symmetric distributions. Why does a boxplot prioritize the median, and in what cases could this choice obscure important data characteristics?

Mean is the arithmetic average of the dataset (summing all the values and dividing by the number of observations). The median is the middle value of the data.
Boxplot prioritizes median because, in that case, it will be less affected by outliers and extreme skewness. If, for example, the data has a lot of peaks, the median might fall in these data (peak ones) and not represent the actual high-density regions.

4. If a boxplot exhibits strong right skewness, what can you infer about the underlying probability distribution? How would this skewness affect statistical measures such as variance, skewness coefficient, and potential model assumptions?

The distribution is positively skewed on the left side, having a long tail on the right side. The mean is greater than the median, and the median is close to Q1. There might be a higher variance, as more data is concentrated on one side. The Skewness Coefficient is positive.

5. Why are boxplots particularly useful for comparing multiple groups in high-dimensional data? What are the limitations of boxplots when dealing with overlapping distributions or categorical variables with small sample sizes?

We can easily compare multiple data distributions. Get insights about their median, mean, and outliers. It is good in high-dimensional datasets where we can visualize different groups.
There may be overlaps in quartiles, for example, which might be confusing in collecting insights, such as differentiating between the distributions (we had a similar question in quiz 1, given a graph where Q1 and median overlapped). Compared to histograms, they do not show the exact frequency of specific values, which makes it harder to identify the distribution model.

6. What are the theoretical consequences of selecting an inappropriate number of bins in a histogram, particularly in datasets with varying density regions or multimodal distributions? How does bin width selection affect kernel density estimation (KDE)?

If we choose a small number of bins, we can lose important insights into the distribution of the given dataset, such as skewness. Too many bins might cause overfitting, which may cause some unnecessary insights to be taken as a wrong pattern. In the case of small bin width, peaks, and valleys might be blurred if too small artificial peaks are created; also, theoretically speaking, a small width indicates low bias and high variance, and a large width indicates the opposite.

7. Histograms and bar charts both use rectangular bars to display data. How does the interpretation of frequency differ in these two visualizations, and why is bin choice irrelevant in bar charts but crucial in histograms?

A histogram visualizes numerical or continuous data. Each bar height indicates frequency or density within that bin. Bar charts are used for categorical or discrete data. The bar height shows the proportion of observations in that category. In histograms, the bin width is crucial because it shows

how it is grouped and what shape of distribution it has. Barcharts do not use bin width because each bin represents a certain category and not a continuous range of values.

8. Under what conditions might a histogram distort the perception of a dataset's distribution? Provide an example where binning choices lead to misleading conclusions, and explain how alternative visualizations (e.g., KDE or violin plots) could address these distortions.

As discussed already, too many or too few bins might cause a misunderstanding of the data distribution. As an example, let's take pizza consumption (per slice per week) and a group of people. Let's say we have a group of 110 people. We might have a bad visualization if we choose, for example, 0.5 with bins; in that case, this will cause inconsistency and visualize twice as many bins as there are unique consumption values. (Pizza consumption is measured in whole sizes, so any non-discrete value causes misunderstanding). If we set the bin width to 100, having 110 people, we will have a loss of details, as well as an oversimplified visualization of pizza consumption, which will not give us any valuable insights. In the case of Kernel, it reveals the distribution shape of the dataset more accurately without taking into account any arbitrary value set for bins.

9. How does a density plot differ from a histogram in terms of its mathematical foundation and interpretability? What challenges arise when choosing a kernel function and bandwidth for density estimation, particularly in sparse datasets?

A histogram represents itself as a binning of a dataset into discrete bins and counting the frequency of the data within each bin. Here, the number of bins and bin width are crucial parameters. (As discussed above). A density plot (KDE) estimates the probability density function of the dataset. This relies on bandwidth and provides a more clear view of the dataset's distribution and skewness. For example, if we choose a Gaussian kernel for data, it will give us better insights about the smoothness of data rather than choosing, for example, a uniform kernel. The same as in question 6, there might be a problem of overfitting or over-smoothing the data.

10. Explain why the area under a density plot is always equal to 1. How does this property relate to probability theory, and what implications does it have for comparing distributions with different sample sizes?

The area under the density plots is always equal to 1 because it represents the total probability of the dataset distribution. This comes from probability theory, stating that no probability can be greater than 1 (100%). Since it is 1, KDE is normalized by default, which means two or more datasets' distributions can be easily and correctly compared.

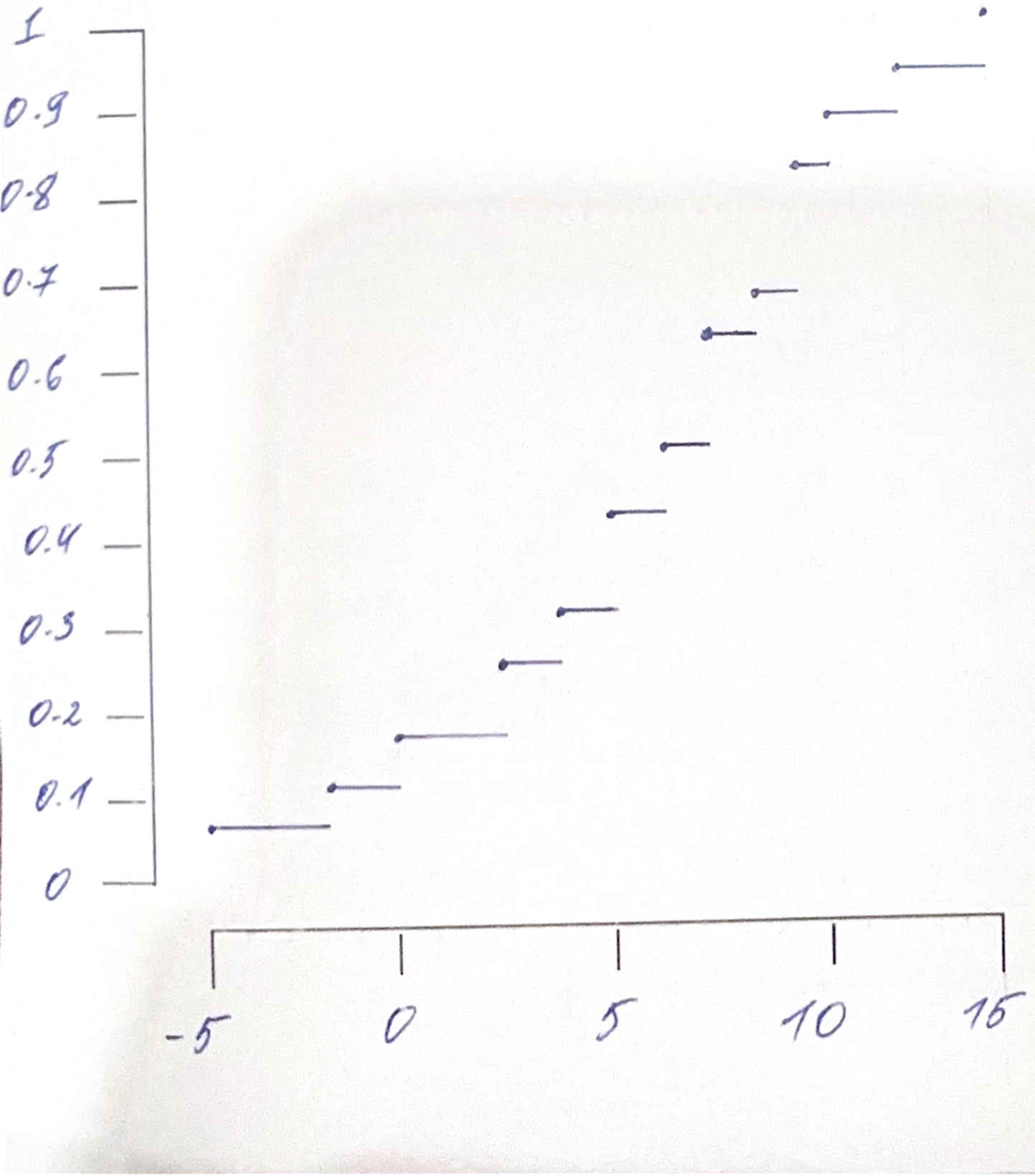
Part 2: Hand-Drawn Graphs

Create graphs by hand using the provided datasets.

1. Given the numbers: -5, -2, 0, 3, 4, 5, 5, 6, 7, 7, 8, 9, 9, 10, 12, 15, draw an ECDF plot.

Value	Frequency	Relative Freq.	Cumulative	
-5	1	1/16	1/16	= 0.0625
-2	1	1/16	1/16 + 1/16 = 2/16	= 0.125
0	1	1/16	2/16 + 1/16 = 3/16	= 0.1875
3	1	1/16	3/16 + 1/16 = 4/16	= 0.25
4	1	1/16	4/16 + 1/16 = 5/16	= 0.3125
5	2	2/16	5/16 + 2/16 = 7/16	= 0.4375
6	1	1/16	7/16 + 1/16 = 8/16	= 0.5
7	2	2/16	8/16 + 2/16 = 10/16	= 0.625
8	1	1/16	10/16 + 1/16 = 11/16	= 0.6875
9	2	2/16	11/16 + 2/16 = 13/16	= 0.8125
10	1	1/16	13/16 + 1/16 = 14/16	= 0.875
12	1	1/16	14/16 + 1/16 = 15/16	= 0.9375
15	1	1/16	15/16 + 1/16 = 16/16	= 1

The ECDF Graph:



2. Given the dataset: -5, 12, 14, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 24, 25, 29, 30, 35, create a boxplot. Indicate the median, quartiles, and any potential outliers.

median (Q_2)

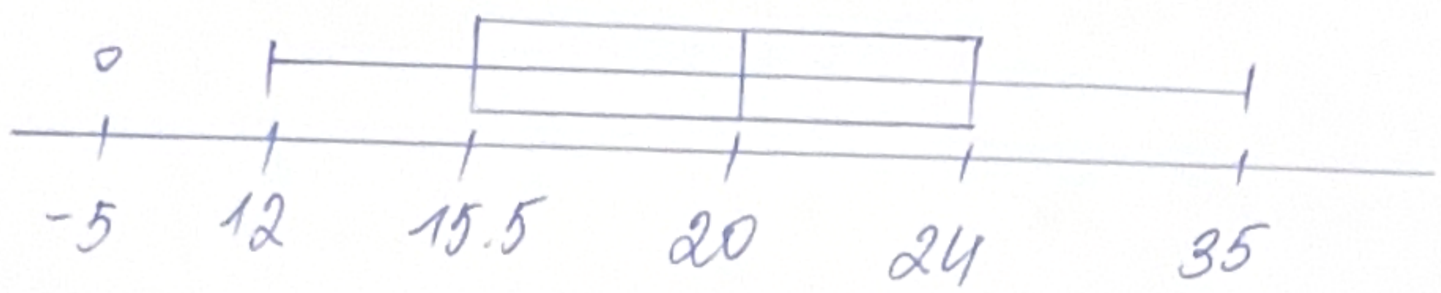
$Q_1 = \frac{15+16}{2} = 15.5$ $\quad Q_3 = 24$

$IQR = Q_3 - Q_1 = 24 - 15.5 = 8.5$

$I = [15.5 - 8.5 \cdot 1.5 ; 24 + 8.5 \cdot 1.5] = [2.75 ; 36.75]$

$W_1 = 12$, $W_2 = 35$, -5 is an outlier

The Boxplot:



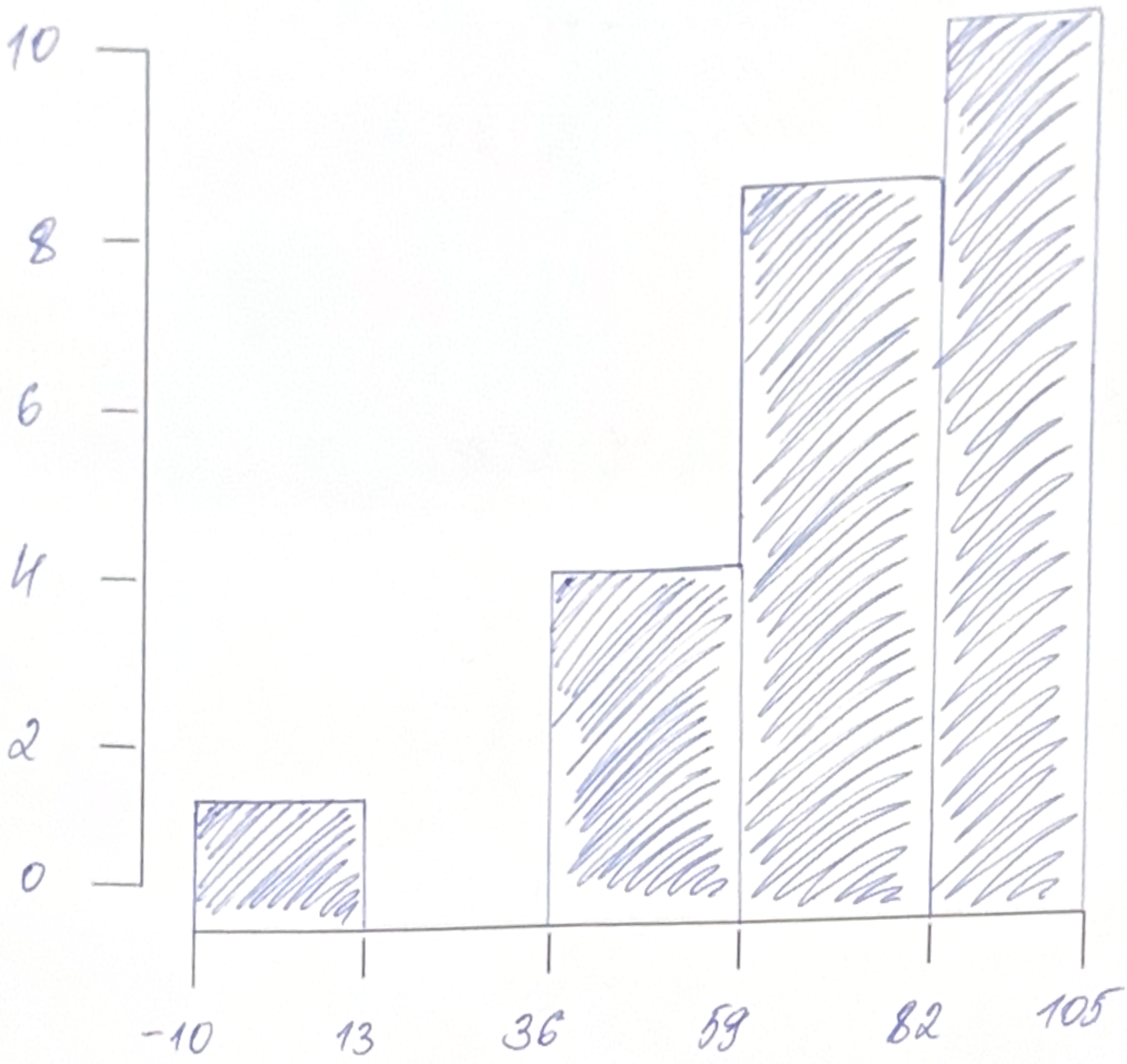
3. Given the test scores: -10, 45, 50, 55, 55, 60, 62, 65, 68, 70, 73, 74, 80, 80, 82, 85, 88, 90, 91, 92, 94, 97, 100, 105, create a histogram using 5 bins and label the axes.

5 bins => 5 intervals:

$[-10;13)$ $[13;36)$ $[36;59)$ $[59;82)$ $[82;105]$

Frequency	1	0	4	9	10
Relative Frequency	$\frac{1}{24} = 0.042$	0	$\frac{4}{24} = 0.166$	$\frac{9}{24} = 0.375$	$\frac{10}{24} = 0.417$
Density	$\frac{0.042}{23} = 0.008$	0	$\frac{0.166}{23} = 0.0072$	$\frac{0.375}{23} = 0.0163$	$\frac{0.417}{23} = 0.01813$

the Frequency Histogram



the Density Histogeann