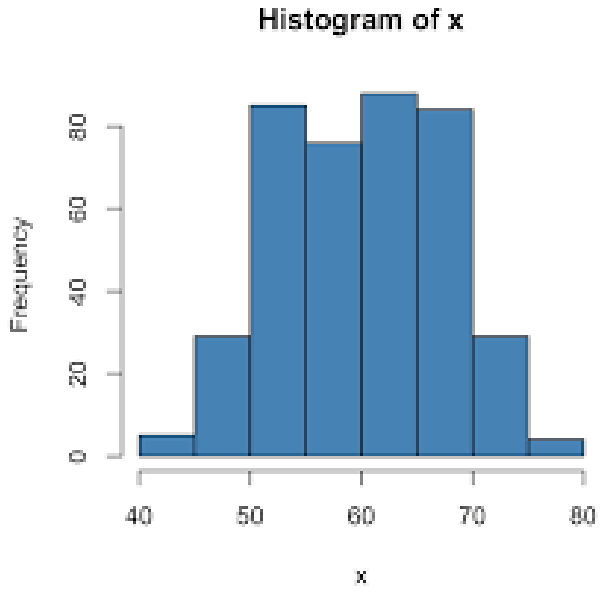
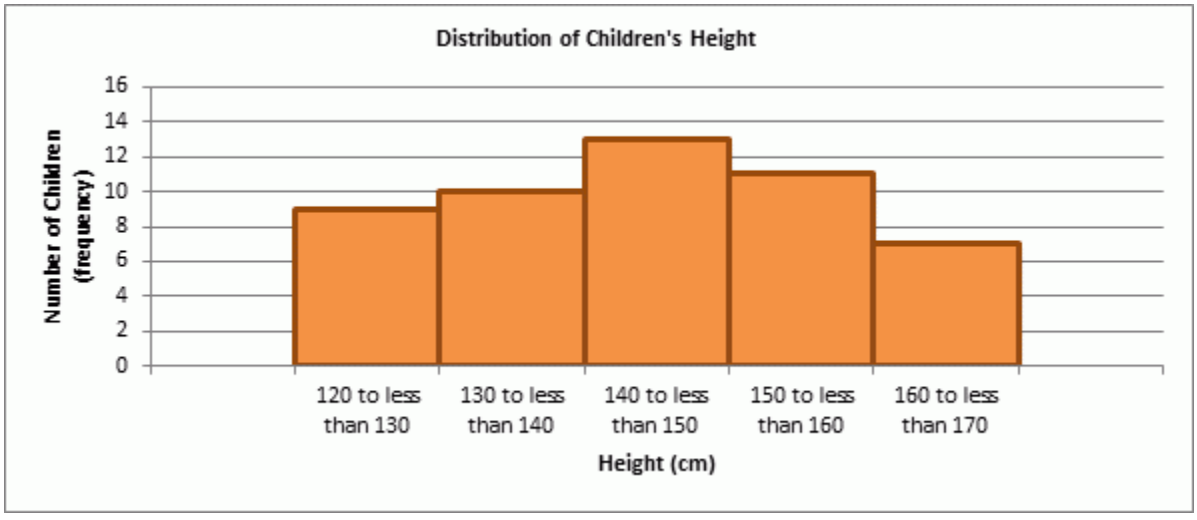
Study Guide: Histogram

What is a Histogram?



- It is a bar chart of a frequency distribution
- It displays data as a series of bars or bins, where each bar represents a range of values and the height of the bar corresponds to the frequency or count of data points within that range
- Histograms are used to visualize and understand the underlying patterns and characteristics of a dataset, including information about central tendencies, variability, and potential outliers
- They are commonly used in statistics and data analysis to gain insights into data distributions and make informed decisions
- Histograms are particularly useful for identifying data trends and making data-driven decisions in fields such as finance, healthcare, and research
- In a histogram, opposite to a bar chart, the boundaries touch. This is because the class boundaries are shared between the two adjacent classes

Let's talk more about frequency distribution:



A frequency distribution is a tabular or graphical representation of data that shows how often specific values or value ranges occur within a dataset. It summarizes the data by counting the frequency (or number of occurrences) of each unique value or value range

Here's how a frequency distribution is typically created:

1. **Data Collection:** Collect the dataset of interest
2. **Data Sorting:** Sort the data in ascending or descending order
3. **Data Binning or Grouping:** Group the data into intervals or bins, if necessary. This step is particularly useful when dealing with continuous data
4. **Counting Frequencies:** Count how many data points fall into each bin or have each unique value
5. **Creating the Distribution:** Present the frequency counts in a tabular format or visualize them using histograms, bar charts, or other graphical representations.

What are the benefits of Frequency Distribution?

- **Summarizing Data:** It provides a concise summary of the data, making it easier to understand and interpret.
- **Identifying Patterns:** Frequency distributions help identify patterns, trends, and central tendencies (like the mode, median, or mean) within the data.
- **Detecting Outliers:** Outliers or extreme values can be readily identified in a frequency distribution, helping to detect potential data errors or anomalies.
- **Comparing Datasets:** It allows for easy comparison of different datasets, helping analysts and researchers draw meaningful conclusions.
- **Statistical Analysis:** Frequency distributions are often a prerequisite for more advanced statistical analyses, such as hypothesis testing or regression analysis.
- **Data Visualization:** They serve as the basis for creating various types of charts and graphs, like histograms, bar charts, and pie charts, which can make complex data more accessible.

What is a Relative Frequency Histogram?

Distribution of Items Sold at Garage Sale, Including Relative Frequencies

Class	Interval	Frequency	Relative Frequency
1	$1-$5	8	.25
2	$6-$10	6	.1875
3	$11-$15	4	.125
4	$16-$20	2	.0625
5	$21-$25	4	.125
6	$26-$30	6	.1875
7	$31-$35	2	.0625

- A relative frequency histogram is a type of histogram that displays the relative frequencies or proportions of data points within specified intervals or bins
- Unlike a regular histogram, which shows the counts or frequencies of data points in each bin, a relative frequency histogram represents the data as percentages or proportions, allowing for a comparison of the distribution's relative contributions

How do you create a Relative Frequency Histogram?

1. **Collect and Organize Data:** Gather the dataset you want to analyze
2. **Determine the Intervals or Bins:** Decide on the intervals or bins into which you want to group your data. These intervals should cover the entire range of data values
3. **Count Data in Each Bin:** Calculate the number of data points that fall into each bin, just as you would for a regular histogram
4. **Calculate Relative Frequencies:** Instead of displaying the counts, calculate the relative frequency for each bin. To calculate the relative frequency, divide the frequency of each category by the total number of observations, and multiply by 100 to express it as a percentage. The formula for relative frequency is: Relative Frequency = (Frequency of Category / Total Observations) * 100

5. **Create the Histogram:** Construct the histogram by representing the relative frequencies as bars. The height of each bar represents the relative frequency of data points within the corresponding bin

Price Range	Frequency	Relative Frequency
$1 – $10	20	0.303
$11 – $20	21	0.318
$21 – $30	13	0.197
$31 – $40	8	0.121
$41 – $50	4	0.061

In this table, the third column represents the relative frequency for each category. The total number of items sold is 66, allowing us to calculate the relative frequency for each category. For example, the relative frequency for the price range $1 – $10 is calculated as (20 / 66) * 100 = 0.303.

Let's talk about Class Boundaries:

- Class boundaries are values that define the upper and lower limits of the intervals or bins used in a histogram or frequency distribution.
- They are essential for creating a well-structured histogram that accurately represents the distribution of data.
- Class boundaries are different from class limits, which represent the actual endpoints of the intervals.

Here's how to calculate class boundaries:

1. **Start with the Class Limits:** Determine the class limits, which are the endpoints of the intervals or bins. Class limits consist of a lower limit and an upper limit for each interval.
2. **Calculate Class Width:** Calculate the class width, which is the range of values covered by each interval. You can do this by subtracting the lower limit of one class from the lower limit of the next class.

Mathematically, the class width (C) is calculated as follows: Class Width (C) = Upper Limit of Class - Lower Limit of Class

3. **Calculate Class Boundaries:** To find the class boundaries, you add half of the class width to the upper limit of the previous class and subtract half of the class width from the lower limit of the current class. This ensures that the class boundaries are halfway between the adjacent class limits.
    - Upper Class Boundary = (Upper Limit of Current Class) + (0.5 * Class Width)
    - Lower Class Boundary = (Lower Limit of Current Class) - (0.5 * Class Width)

Here's a step-by-step example to illustrate how to calculate class boundaries:

Suppose you have data on the ages of individuals, and you want to create a frequency distribution with class limits and boundaries:

- Class Width (C) = 10 years (for example)
- Lower Limit of First Class = 20 years
- Upper Limit of Last Class = 80 years

Using the formulas above:

- Upper Class Boundary of First Class = 20 + (0.5 * 10) = 25 years
- Lower Class Boundary of First Class = 20 - (0.5 * 10) = 15 years
- Upper Class Boundary of Second Class = 30 + (0.5 * 10) = 35 years
- Lower Class Boundary of Second Class = 30 - (0.5 * 10) = 25 years

Repeat this process for each class to calculate the class boundaries.

Class boundaries

- The numbers that separate classes without forming gaps between them.
- The distance from the upper limit of the first class to the lower limit of the second class is 115 − 114 = 1.
- Half this distance is 0.5.
- First class lower boundary = **59 − 0.5 = 58.5**
- First class upper boundary = **114 + 0.5 = 114.5**

Class	Class Boundaries	Frequency, *f*
59 − **114**	**58.5 − 114.5**	5
115 − 170		8
171 − 226		6

Class boundaries are important for creating accurate histograms and frequency distributions because they ensure that data points are properly assigned to the correct intervals, and there is no ambiguity in the representation of the data.