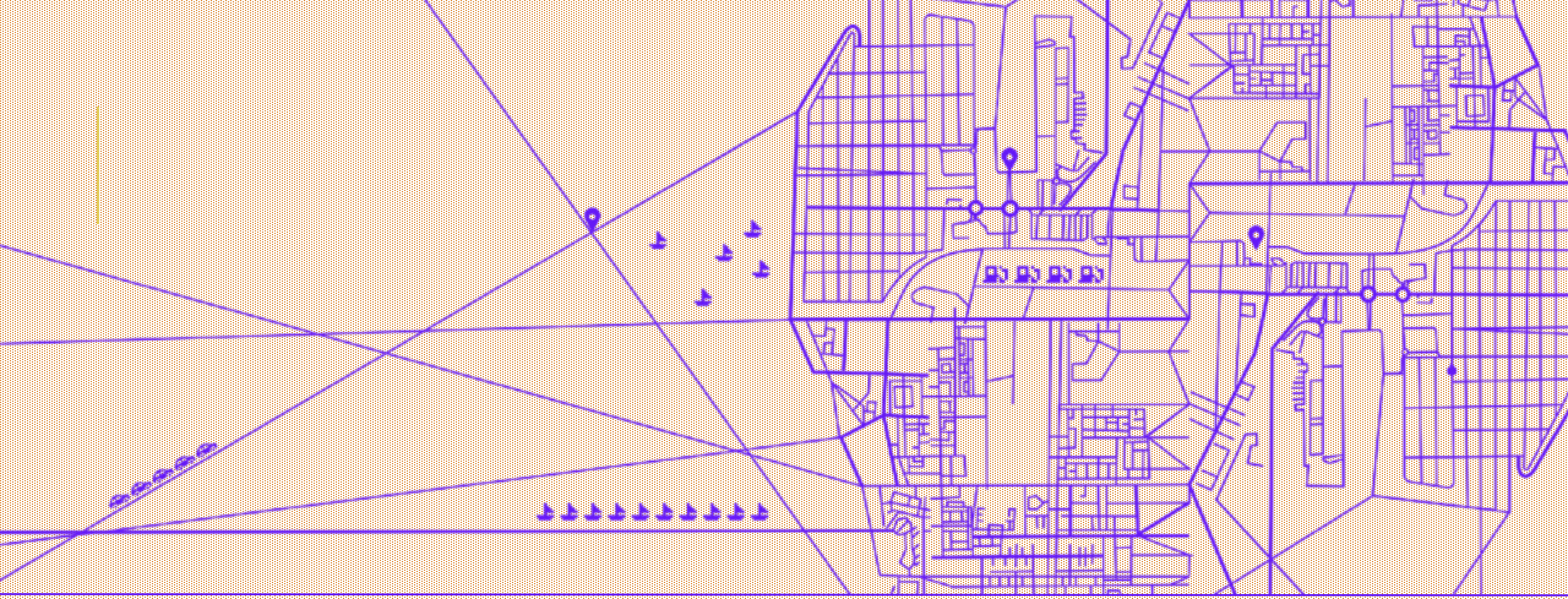


TOPIC 4: STATISTICS

42 أبوظبي
ABU DHABI

OUTLINE

- Mean
- Median
- Mode
- Range and Interquartile Range
- Variance
- Standard Deviation
- Data Representation Techniques

STATISTICS

Statistics is a branch of mathematics dealing with collection, analysis, interpretation, and representation of large numerical data.

The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.

MEAN

The mean is used to summarize a data set. It is the measure of the center of a data set. Mean is the "average" number; found by adding all data points and dividing by the number of data points.

Example 1: The mean of 4, 1, and 7 is $(4 + 1 + 7)/3 = 12/3 = 4$

There are many different types of mean, but usually when people say mean, they are talking about the arithmetic mean.

The arithmetic mean is the sum of all of the data points divided by the number of data points.

$$mean = \frac{sum\ of\ data}{\#\ of\ data\ points}$$

Formally:

$$mean = \frac{\sum x_i}{n}$$

Example 2: 'A' has 5 cookies, 'B' has 3 cookies, 'C' has 6 cookies, and 'D' has 2 cookies. Find the mean number of cookies.

Solution:

$$Mean = (5 + 3 + 6 + 2) / 4 = 4\, cookies$$

We can think of the mean as the number of cookies each girl would have if they were equally distributed among the four girls.

Question 1: Find the mean of the data set {7, 2, 6, 8, 7}

Solution: $Mean = (7 + 2 + 6 + 8 + 7) / 5 = 30/5 = 6$

Question 2: Given = {5, 2, *x*, 2, 4, 8}, mean = 4. Find the value of *x*?

Solution: $\frac{(x+2+2+4+5+8)}{6} = 4$

$x = 24 - 21 = 3$

Check! $(2 + 2 + 3 + 4 + 5 + 8) / 6 = \mathbf{4}$ (verified)

MEDIAN

The median is the middle point in a dataset - half of the data points are smaller than the median and half of the data points are larger.

To find the median:

- Arrange the data points from smallest to largest.
- If the number of data points is odd, the median is the middle data point in the list.
- If the number of data points is even, the median is the average of the two middle data points in the list.

Example 3: Find the median of this data set: 1, 4, 2, 0, 5

Solution: Put the data in order first: 0, 1, 2, 4, 5

There is an odd number of data points, so the median is the middle data point.

0, 1, 2, 4, 5

The median is **2.**

MODE

Mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.

Example 4: Find the mode of this data: {4, 2, 4, 3, 2, 2}

Solution: Look for the value that occurs the most:

{2, 2, 2, 3, 4, 4}

The mode is 2.

MEAN, MEDIAN, MODE

Question 3: Find the mean, median, and mode of the following set of numbers:

23, 29, 20, 32, 23, 21, 33, 25

Solution:

Rearrange: 20, 21, 23, 23, 25, 29, 32, 33

Mean: (20+21+23+23+25+29+32+33) / 8 = 25.75

Median: 20,21,23,**23**,**25**,29,32,33 = (23+25)/2 = 24

Mode: 23

RANGE AND INTERQUARTILE RANGE

Range and interquartile range (IQR) both measure the "spread" in a data set. Looking at spread lets us see how much data varies. Range is a quick way to get an idea of spread. It takes longer to find the IQR, but it sometimes gives us more useful information about spread.

Interquartile range is the amount of spread in the middle 50% of a dataset.

In other words, it is the distance between the first quartile (Q1) and the third quartile (Q3).

$IQR = Q_3 - Q_1$

How to find the IQR:

Step 1: Put the data in order from least to greatest.

Step 2: Find the median. If the number of data points is odd, the median is the middle data point. If the number of data points is even, the median is the average of the middle two data points.

Step 3: Find the first quartile Q_1. The first quartile is the median of the data points to the left of the median in the ordered list.

Step 4: Find the third quartile Q_3. The third quartile is the median of the data points to the right of the median in the ordered list.

Step 5: Calculate IQR by subtracting $Q_3 - Q_1$

Example 5: Essay's in French class are scored on a 6 point scale. Find the IQR of these scores: 1, 3, 3, 3, 4, 4, 4, 6, 6

Solution:

Step 1: The data is already in order.

Step 2: Find the median. There are 9 scores, so the median is the middle score.

1, 3, 3, 3, **4**, 4, 4, 6, 6

The median is 4.

Step 3: Find Q_1, which is the median of the data to the left of the median. There is an even number of data points to the left of the median, so we need the average of the middle two data points.

1, 3, 3, 3; $Q_1 = \frac{3+3}{2} = 3$

The first quartile is 3.

Step 4: Find Q_3, which is the median of the data to the right of the median. There is an even number of data points to the right of the median, so we need the average of the middle two data points.

4, 4, 6, 6

$$Q_3 = \frac{4+6}{2} = 5$$

The third quartile is 5.

Step 5: Calculate the IQR.

$IQR = Q_3 - Q_1 = 5 - 3 = 2$

The IQR is 2 points.

Question 4: Sara recorded the daily high temperatures for two different cities in a recent week in degree Celsius. The temperatures for each city are shown below.

Abu Dhabi: 23,**25**,28,**28**,32,**33**,35

Fujairah: 16,**24**,26,**26**,26,**27**,28

Calculate the range and IQR of the temperatures in Abu Dhabi and Fujairah?

Solution: Range: AD = 35-23 = 12 °C, Fujairah = 28-16 = 12 °C

IQR: AD = 33-25 = 8 °C, Fujairah = 27–24 = 3 °C

VARIANCE

Variance measures how far a data set is spread out. It is mathematically defined as the average of the squared differences from the mean.

The formula for population variance (σ) is

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

The formula for sample variance (σ) is

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

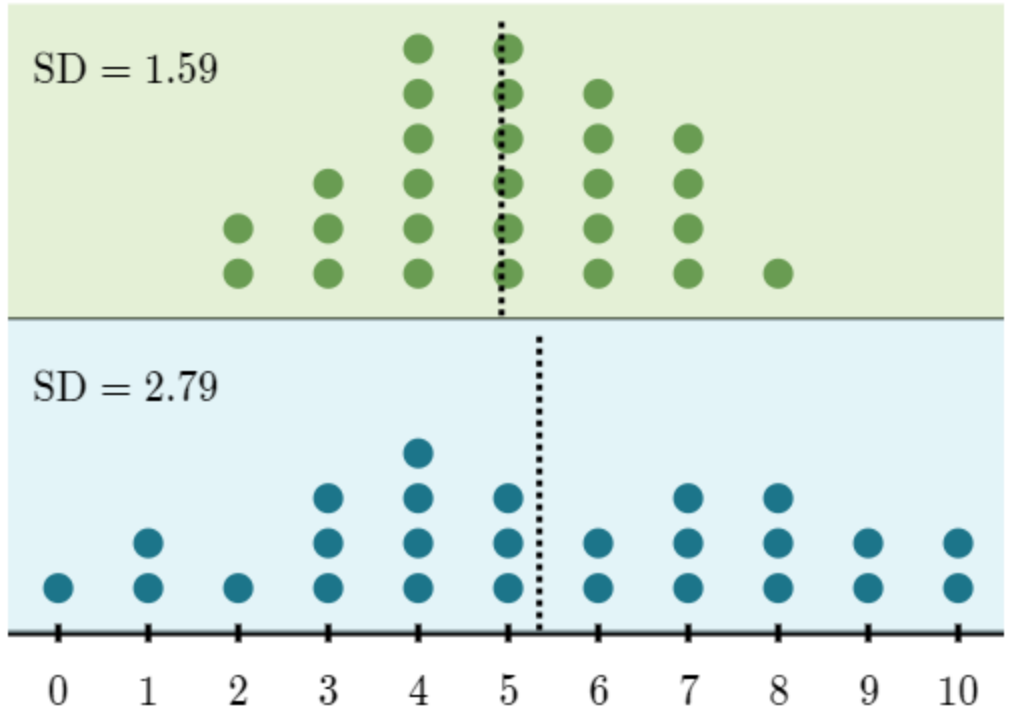
There is an $n-1$ in the denominator rather than the N. This is because we don't have the real mean, μ, available so we have to use the *estimate* of the mean based on the sample, i.e. $\bar{x}$. This is likely to be slightly better centered in the sample than the true mean (μ) of all the card values, hence the sum of the terms of the form $(x_i-\bar{x})^2$ will be slightly too small. The reduction in the denominator from n to $n-1$ compensates for this.

STANDARD DEVIATION

Standard deviation measures the spread of a data distribution. **The more spread out a data distribution is, the greater its standard deviation.**

For example, the blue distribution on bottom has a greater standard deviation (SD) than the green distribution on top:

Interestingly, standard deviation cannot be negative. A standard deviation close to 0 indicates that the data points tend to be close to the mean (shown by the dotted line). The further the data points are from the mean, the greater the standard deviation.



The formula for population standard deviation (SD) is

$$SD = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

The formula for sample standard deviation (SD) is

$$SD = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

where Σ means "sum of", x is a value in the data set, μ / $\bar{x}$ is the mean of the data set, and *N* / *n* is the number of data points in the population.

Steps to find the SD:

Step 1: Find the mean.

Step 2: For each data point, find the square of its distance to the mean.

Step 3: Sum the values from Step 2.

Step 4: Divide by the number of data points.

Step 5: Take the square root.

Example 6: Find the population standard deviation of the data set: 6, 2, 3, 1

Solution:

Step 1: Find the mean μ.

$$\mu = \frac{6+2+3+1}{4} = 3$$

Step 2: Find the square of the distance from each data point to the mean $(x-\mu)^2$.

Steps 3, 4 and 5:

x	$\lvert x-\mu\rvert^2$
6	$\lvert 6-3\rvert^2 = 3^2 = 9$
2	$\lvert 2-3\rvert^2 = 1^2 = 1$
3	$\lvert 3-3\rvert^2 = 0^2 = 0$
1	$\lvert 1-3\rvert^2 = 2^2 = 4$

$$SD = \sqrt{\frac{\sum(x-\mu)^2}{N}} = \sqrt{\frac{9+1+0+4}{4}} = \sqrt{\frac{14}{4}} = \sqrt{3.5} = 1.87$$

Question 5:

Ten cards are selected individually, noted and replaced in the pack (that has no picture cards). This gives a sample size $n = 10$. The values of the cards are:

10 3 4 3 5 4 1 5 8 5

Estimate the median, mode, mean, Range, IQR, and standard deviation of the values in the complete pack based on this sample.

Solution: 1 3 3 4 4 5 5 5 8 10

Median = (4+5)/2 = 4.5; Mode = 5, Range = 10-1 = 9; IQR = $Q_3 - Q_1 = 5 - 3 = 2$

Mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{10+3+4+3+5+4+1+5+8+5}{10} = \frac{48}{10} = 4.8$

Standard Deviation: $SD = \sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}} =$

$$\sqrt{\frac{(10-4.8)^2+(3-4.8)^2+(4-4.8)^2+(3-4.8)^2+(5-4.8)^2+(4-4.8)^2+(1-4.8)^2+(5-4.8)^2+(8-4.8)^2+(5-4.8)^2}{9}} = 2.5734$$

DATA REPRESENTATION TECHNIQUES

In statistics, data is represented in tables, charts and graphs.

A **stem-and-leaf plot** is a method of organizing the data that includes sorting the data and graphing it at the same time. This type of graph uses a stem as the leading part of a data value and a leaf as the remaining part of the value. The result is a graph that displays the sorted data in groups, or classes. A stem-and-leaf plot is used most when the number of data values is large.

Example 7:

At a local veterinarian school, the number of animals treated each day over a period of 20 days was recorded. Construct a stem-and-leaf plot for the data set, which is as follows, and find the mean, mode, median, and standard deviation:

28 34 23 35 16 17 47 05 60 26

39 35 47 35 38 35 55 47 54 48

Step 1: Create the stem-and-leaf plot.

Some people prefer to arrange the data in order before the stems and leaves are created. This will ensure that the values of the leaves are in order. However, this is not necessary and can take a great deal of time if the data set is large. We will first create the stem-and-leaf plot, and then we will organize the values of the leaves.

Stem	Leaf
0	5
1	6, 7
2	8, 3, 6
3	4, 5, 9, 5, 5, 8, 5
4	7, 7, 7, 8
5	5, 4
6	0

The leading digit of a data value is used as the stem, and the trailing digit is used as the leaf. The numbers in the stem column should be consecutive numbers that begin with the smallest class and continue to the largest class. If there are no values in a class, do not enter a value in the leaf (leave it blank).

Step 2: Organize the values in each leaf row.

Stem	Leaf
0	5
1	6, 7
2	3, 6, 8
3	4, 5, 5, 5, 5, 8, 9
4	7, 7, 7, 8
5	4, 5
6	0

Mean: $\mu = \frac{5+16+17+23+26+28+34+35+35+35+35+38+39+47+47+47+48+54+55+60}{20} = \frac{724}{20} = 36.2$

Mode: 35

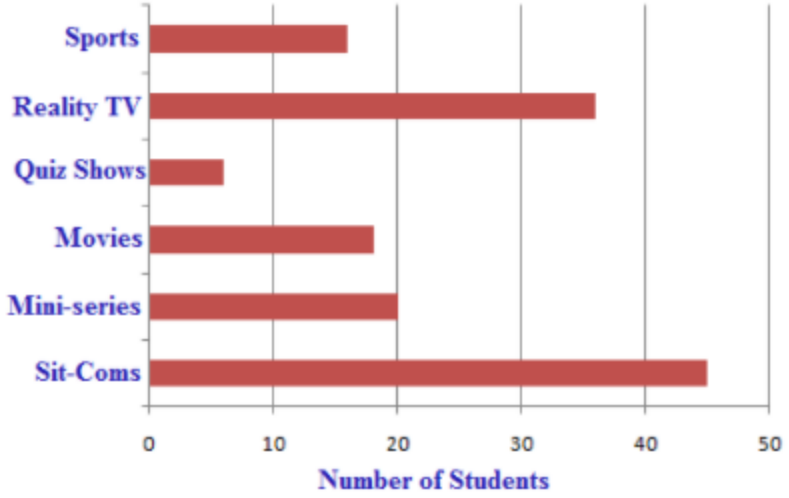
Median: $\frac{35+35}{2} = \frac{70}{2} = 35$

$SD = \sqrt{\frac{\sum_{i=1}^{N}(x-\mu)^2}{N}} = \sqrt{\frac{(5-36.2)^2+\cdots\ldots\ldots(60-36.2)^2}{20}} = \sqrt{\frac{3843.2}{20}} = \sqrt{192.16} = 13.86$

A **bar graph** is a plot made of bars whose heights (vertical bars) or lengths (horizontal bars) represent the frequencies of each category. There is 1 bar for each category, with space between each bar, and the data that is plotted is discrete data. Each category is represented by intervals of the same width. When constructing a bar graph, the category is usually placed on the horizontal axis, and the frequency is usually placed on the vertical axis. These values can be reversed if the bar graph has horizontal bars.

Example 8:

The following bar graph represents the results of a survey to determine the type of TV shows watched by high school students:



Use the bar graph to answer the following questions:

1. What type of show is watched the most?
2. What type of show is watched the least?
3. Approximately how many students participated in the survey?
4. Does the graph show the differences between the preferences of males and females?

Solution:

1. Sit-coms are watched the most.

2. Quiz shows are watched the least.

3. Approximately 45+20+18+6+35+16=140 students participated in the survey.

4. No, the graph does not show the differences between the preferences of males and females.

A **histogram** is a type of vertical bar graph in which the bars represent grouped continuous data. While there are similarities between a bar graph and a histogram, such as each bar being the same width, a histogram has no spaces between the bars. The quantitative data is grouped according to a determined bin size, or interval. The bin size refers to the width of each bar, and the data is placed in the appropriate bin.

The bins, or groups of data, are plotted on the x-axis, and the frequencies of the bins are plotted on the y-axis. A grouped frequency distribution is constructed for the numerical data, and this table is used to create the histogram.

Example 9:

Construct a frequency distribution table with a bin size of 10 for the following data, which represents the ages of 30 lottery winners:

38 41 29 33 40 74 66 45 60 55

25 52 54 61 46 51 59 57 66 62

32 47 65 50 39 22 35 72 77 49

Solution:

Step 1: Determine the range of the data by subtracting the smallest value from the largest value.

Range: $77-22=55$

Step 2: Divide the range by the bin size to ensure that you have at least 5 groups of data. A histogram should have from 5 to 10 bins to make it meaningful: $[55/10]=5.5\approx6$. Since you cannot have 0.5 of a bin, the result indicates that you will have at least 6 bins.

Step 3: Construct the table.

Step 4: Determine the sum of the frequency column to ensure that all the data has been grouped.

$3+5+6+8+5+3=30$

Note: When data is grouped in a frequency distribution table, the actual data values are lost. The table indicates how many values are in each group, but it doesn't show the actual values.

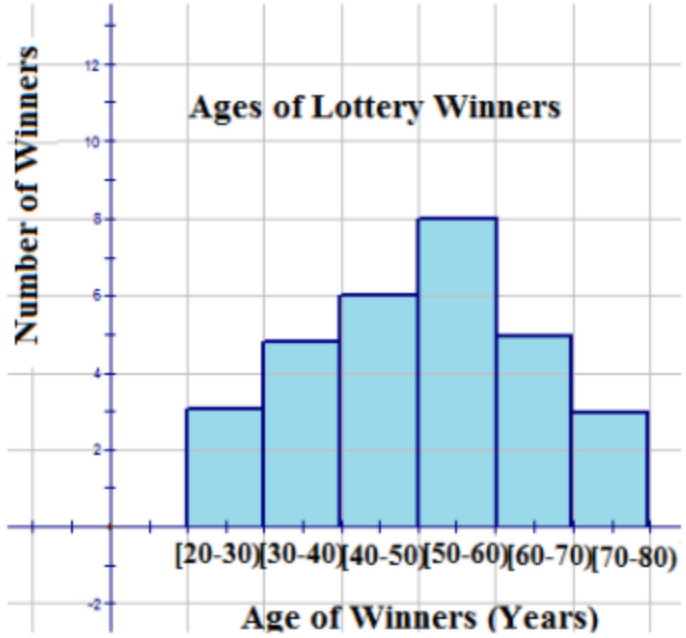
Bin	Frequency
$[20-30)$	3
$[30-40)$	5
$[40-50)$	6
$[50-60)$	8
$[60-70)$	5
$[70-80)$	3

Example 10:

Use the data from Example 9 that displays the ages of the lottery winners to construct a histogram.

Solution:

From looking at the tops of the bars, you can see how many winners were in each category, and by adding these numbers, you can determine the total number of winners. You can also determine how many winners were within a specific category. For example, you can see that 8 winners were 60 years of age or older. The graph can also be used to determine percentages. For example, it can answer the question, "What percentage of the winners were 50 years of age or older?" as follows:



16/30=0.533 ; (0.533)(100%)≈53.3%.

A **frequency polygon** is a graph constructed by using lines to join the midpoints of each interval, or bin. The heights of the points represent the frequencies. A frequency polygon can be created from the histogram or by calculating the midpoints of the bins from the frequency distribution table. The **midpoint** of a bin is calculated by adding the upper and lower boundary values of the bin and dividing the sum by 2.

Example 11:

The following distribution table represents the number of miles run by 20 randomly selected runners during a recent road race:

Using this table, construct a frequency polygon?

Bin	Frequency
$[5.5 - 10.5)$	1
$[10.5 - 15.5)$	3
$[15.5 - 20.5)$	2
$[20.5 - 25.5)$	4
$[25.5 - 30.5)$	5
$[30.5 - 35.5)$	3
$[35.5 - 40.5)$	2

Solution:

Step 1: Calculate the midpoint of each bin by adding the 2 numbers of the interval and dividing the sum by 2.

Midpoints: (5.5+10.5)/2=16/2=8

(10.5+15.5)/2=26/2=13

(15.5+20.5)/2=36/2=18

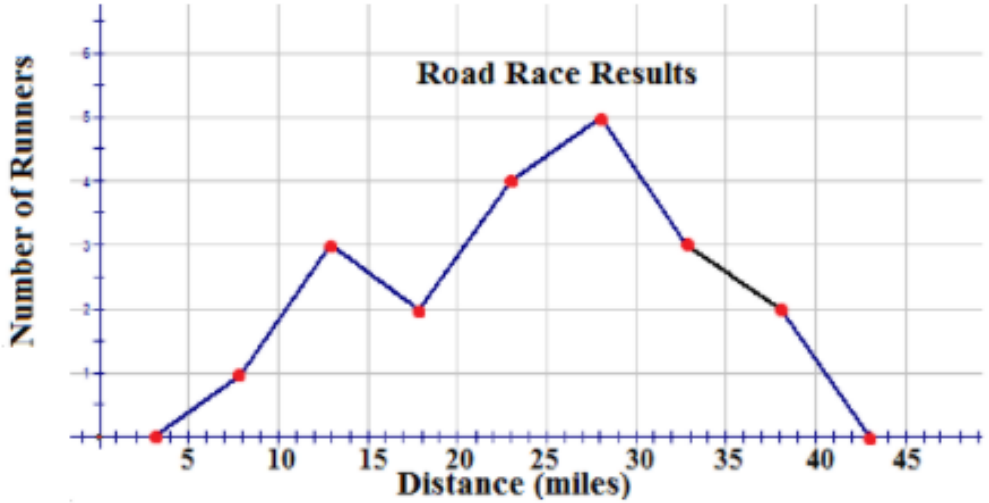
(20.5+25.5)/2=46/2=23

(25.5+30.5)/2=56/2=28

(30.5+35.5)/2=66/2=33

(35.5+40.5)/2=76/2=38

Step 2: Plot the midpoints on a grid, making sure to number the x-axis with a scale that will include the bin sizes. Join the plotted midpoints with lines.



A frequency polygon usually extends 1 unit below the smallest bin value and 1 unit beyond the greatest bin value. This extension gives the frequency polygon an appearance of having a starting point and an ending point, which provides a view of the distribution of data. If the data set were very large so that the number of bins had to be increased and the bin size decreased, the frequency polygon would appear as a smooth curve.

Resources:
https://statanalytica.com/blog/importance-of-statistics/

42 أبوظبي
ABU DHABI