Probability & Statistics Workbook Solutions

Data distributions

krista king MATH

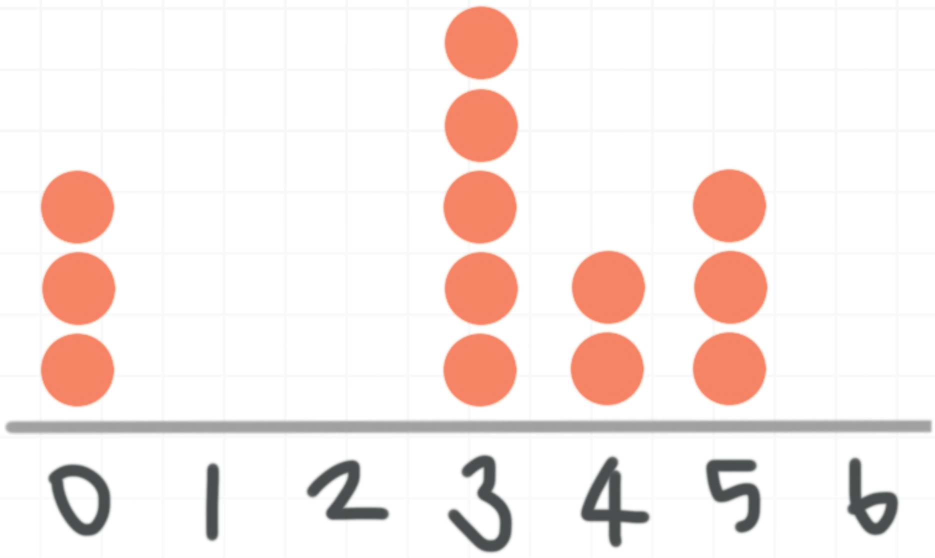
MEAN, VARIANCE, AND STANDARD DEVIATION

■ 1. Mrs. Bayer's students take a test on Friday. She grades their tests over the weekend and notes that the average test score is 68 points with a population standard deviation of 5 points. She decided to add 10 points to all of the tests. What are the new mean and population standard deviation?

Solution:

The population standard deviation will remain the same, because adding the 10 points won't change the spread of the data. The population standard deviation of the old and new data will both be 5. Adding 10 points to all of the tests will increase the mean by 10 points. The old mean is 68 points, so the new mean is 78 points.

■ 2. What is the sample variance of the data set, rounded to the nearest hundredth?



Solution:

The formula for the sample variance includes the sample mean, so we'll need to find that first. There are $n = 13$ data points in the dot plot, so the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{3(0) + 5(3) + 2(4) + 3(5)}{13}$$

$$\bar{x} = \frac{38}{13}$$

$$\bar{x} \approx 2.92$$

The sample variance is

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

$$S^2 = \frac{3(0-2.92)^2 + 5(3-2.92)^2 + 2(4-2.92)^2 + 3(5-2.92)^2}{13-1}$$

$$S^2 = \frac{40.9232}{12}$$

$$S^2 \approx 3.41$$

■ 3. Sometimes it can be helpful to calculate the standard deviation by using a table. Use the data to fill in the rest of the table and then use the table to calculate the sample standard deviation.

Data value	Data value - Mean	Squared difference
97		
110		
112		
121		
110		
98		
Total		

Solution:

We'll first calculate the mean of the data values given in the table.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{97 + 110 + 112 + 121 + 110 + 98}{6}$$

$$\bar{x} = \frac{648}{6}$$

$$\bar{x} = 108$$

Now we can fill out the table.

Data value	Data value - Mean	Squared difference
97	97-108=-11	$(-11)^2=121$
110	110-108=2	$2^2=4$
112	112-108=4	$4^2=16$
121	121-108=13	$13^2=169$
110	110-108=2	$2^2=4$
98	98-108=-10	$(-10)^2=100$
Total		**121+4+16+169+4+100=4**

The sum of the squared differences is 414. So sample variance is

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$S^2 = \frac{414}{5}$$

$$S^2 = 82.8$$

So the sample standard deviation is

$$\sqrt{S^2} = \sqrt{82.8}$$

$$S \approx 9.099$$

■ 4. The sum of the squared differences from the population mean for a data set is 212. If the data set has 25 items, what is the population standard deviation?

Solution:

The formula for population variance is

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

The numerator gives the sum of the squared differences, so we can plug in from the problem.

$$\sigma^2 = \frac{212}{25}$$

$$\sigma^2 = 8.48$$

The population standard deviation is therefore

$$\sqrt{\sigma^2} = \sqrt{8.48}$$

$$\sigma \approx 2.91$$

5. For the data set $40, 44, 47, 55, 60, 60, 65, 80$, find

$$\sum_{i=1}^{n} (x_i - \bar{x})$$

What does this say about why we square the $(x_i - \bar{x})$ in the variance and standard deviation formulas?

Solution:

The value of

$$\sum_{i=1}^{n} (x_i - \bar{x})$$

will be 0 for any data set. The sum of the deviations from the mean will always be 0, because the negative and positive values will cancel each another out. This is one of the reasons that $(x_i - \bar{x})$ is squared in the standard deviation formulas.

To prove that this value is 0 for this particular data set, we'll first find the mean.

$$\bar{x} = \frac{40 + 44 + 47 + 55 + 60 + 60 + 65 + 80}{8}$$

$$\bar{x} = 56.375$$

Then we can find the sum.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = (40 - 56.375) + (44 - 56.375) + (47 - 56.375) + (55 - 56.375)$$

$$+(60 - 56.375) + (60 - 56.375) + (65 - 56.375) + (80 - 56.375)$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = -16.375 - 12.375 - 9.375 - 1.375 + 3.625 + 3.625 + 8.625 + 23.625$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

6. Give an example of a situation where $\$5$ could represent a large standard deviation and another where $\$5$ could represent a small standard deviation.

Solution:

The idea of how large or small the standard deviation of a data set is really depends on what it is we're measuring. If, for example, we were measuring the price of a soft drink at a state fair, and we found a standard deviation of $\$5$, that's a large standard deviation. It's large because soft drinks usually do not cost very much and this would tell us that we need to hunt around for the best price.

On the other hand, if we were purchasing a specific type of car and we found that the standard deviation was $\$5$ among the dealerships we were considering, that standard deviation would be very small. Small enough, in fact, that it wouldn’t matter much where we bought the car because the prices would all be pretty much the same.

FREQUENCY HISTOGRAMS AND POLYGONS, AND DENSITY CURVES

1. A dog walking company keeps track of how many times each dog receives a walk. 40% of all the dogs walked by the company received between 25 and 40 walks, and no dogs received more than 40 walks. How many dogs received between 0 and 25 walks, if the company walks 400 dogs?

Solution:

Because no dogs received more than 40 walks, that means 100% of the dogs received between 0 and 40 walks. Since 40% of the dogs received between 25 and 40 walks, that must mean that $100\% - 40\% = 60\%$ of the 400 dogs received between 0 and 25 walks. This means $0.60(400) = 240$ dogs took between 0 and 25 walks.

2. The number of crayons in each student's pencil box is

$$4,\ 1,\ 5,\ 5,\ 9,\ 11,\ 15,\ 13,\ 15,\ 14,\ 16,\ 17,\ 20,\ 16,\ 16,\ 17$$

Complete the frequency and relative frequency tables for the data and use it to create a relative frequency histogram.

Crayons	Frequency	Relative Frequency
1-5		
6-10		
11-15		
16-20		
Totals:		**100%**

Solution:

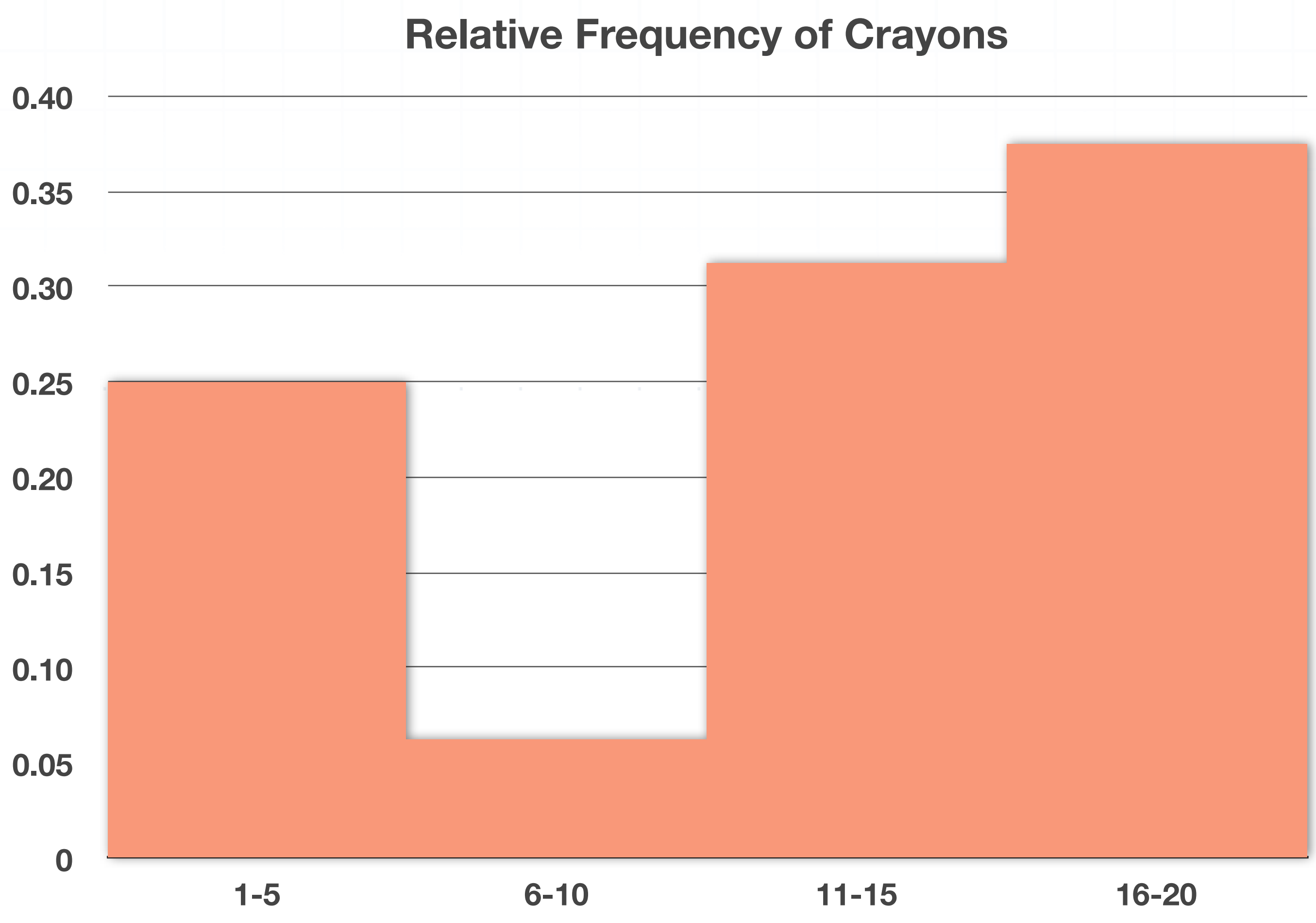
First count the number of items in each frequency interval and add that to the table, as well as calculate the total number of crayons.

Crayons	Frequency	Relative Frequency
1-5	4	
6-10	1	
11-15	5	
16-20	6	
Totals:	**16**	**100%**

Next calculate the relative frequencies in the table by dividing the frequency by the total number of crayons.

Crayons	Frequency	Relative Frequency
1-5	4	4/16=25%
6-10	1	1/16=6.25%
11-15	5	5/16=31.25%
16-20	6	6/16=37.5%
Totals:	**16**	**100%**

Use the intervals on the horizontal axis and the relative frequencies on the vertical axis to make the histogram.



3. The table shows the scores on the last history exam in Mr. Ru's class.

40	32	40	83
95	33	87	59
32	81	46	78
91	61	55	88
40	61	82	99
72	47	83	91
101	77	65	87

Complete the relative frequency table and create a frequency polygon for the data.

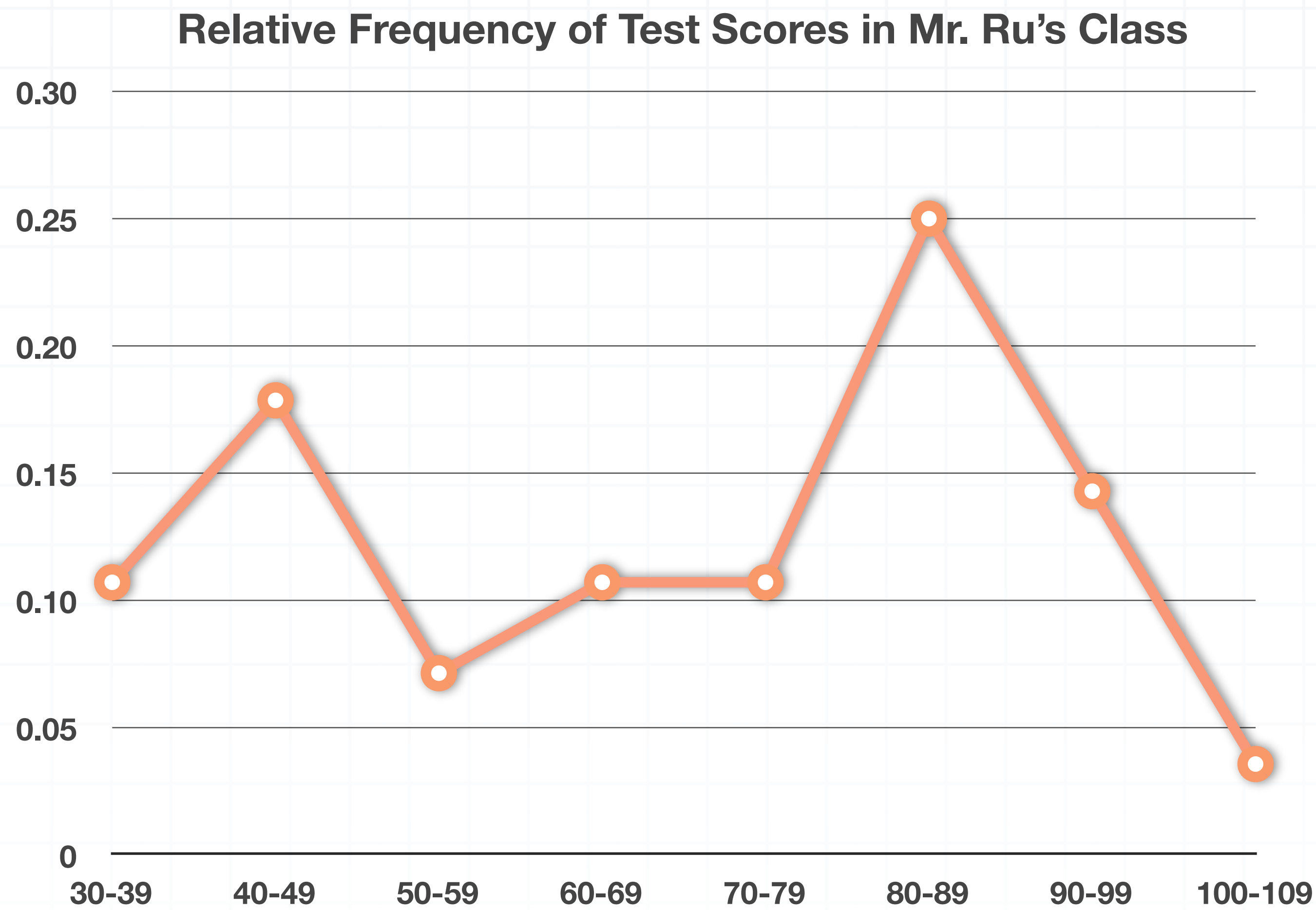
Score	Frequency	Relative Frequency
30-39		
40-49		
50-59		
60-69		
70-79		
80-89		
90-99		
100-109		
Totals:		

Solution:

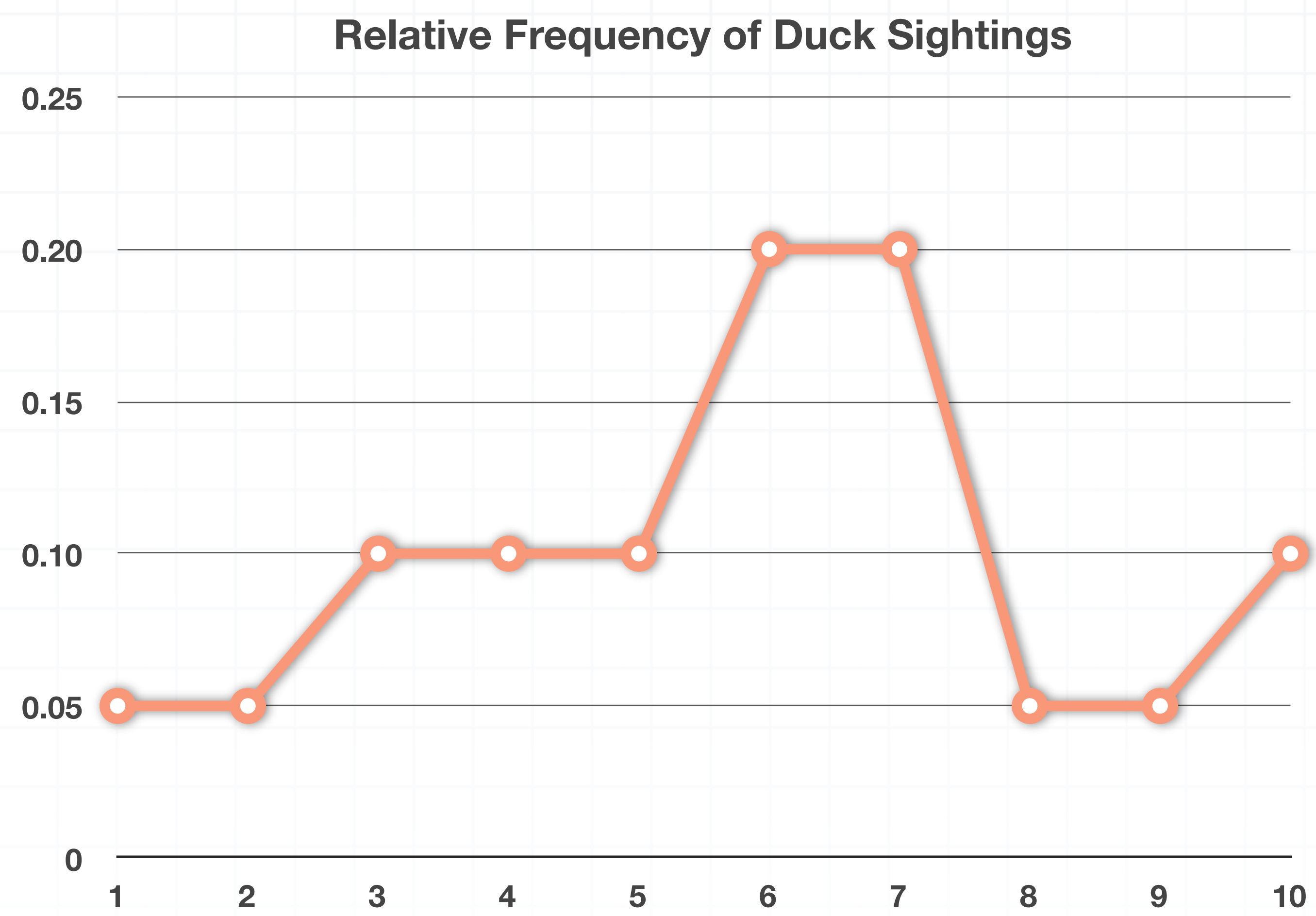
The first step to completing the frequency table is to count the scores in each interval, then use those frequencies and the total number of test scores to calculate the relative frequencies.

Score	Frequency	Relative Frequency
30-39	3	3/28=0.1071
40-49	5	5/28=0.1786
50-59	2	2/28=0.0714
60-69	3	3/28=0.1071
70-79	3	3/28=0.1071
80-89	7	7/28=0.2500
90-99	4	4/28=0.1429
100-109	1	1/28=0.0357
Totals:	**28**	**100%**

Use the intervals on the horizontal axis and the relative frequencies on the vertical axis to make the relative frequency polygon.



4. Becky kept track of the number of ducks she saw at her neighborhood pond at $6:30$ a.m. every morning for 365 days. On how many days did Becky see more than 5 ducks?



Solution:

We want to know on how many days Becky saw 6, 7, 8, 9, and 10 ducks. We can organize our data into a table to read the values we want. Read the relative frequencies from the frequency polygon.

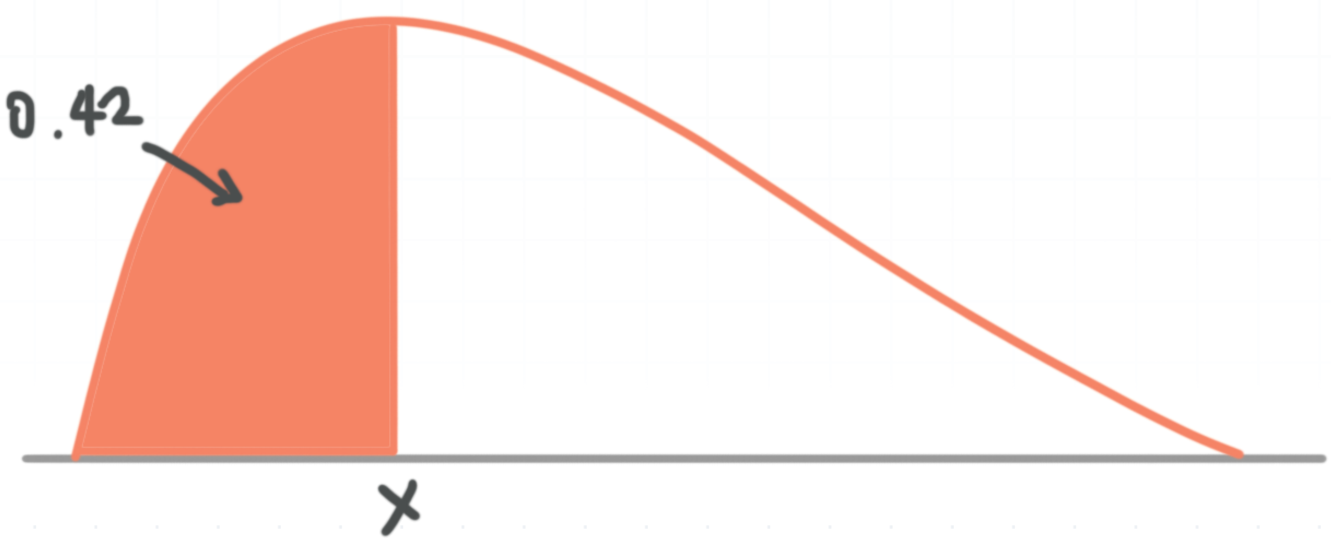
Ducks	Relative Frequency
6	0.20
7	0.20
8	0.05
9	0.05
10	0.10

Add the relative frequencies from 6 to 10. The cumulative relative frequency is

$$0.20 + 0.20 + 0.05 + 0.05 + 0.10 = 0.60 = 60\,\%$$

She took 365 days of data, which means she saw more than five ducks on $0.60(365) = 219$ days.

5. What percentage of the population is greater than x for the density curve?

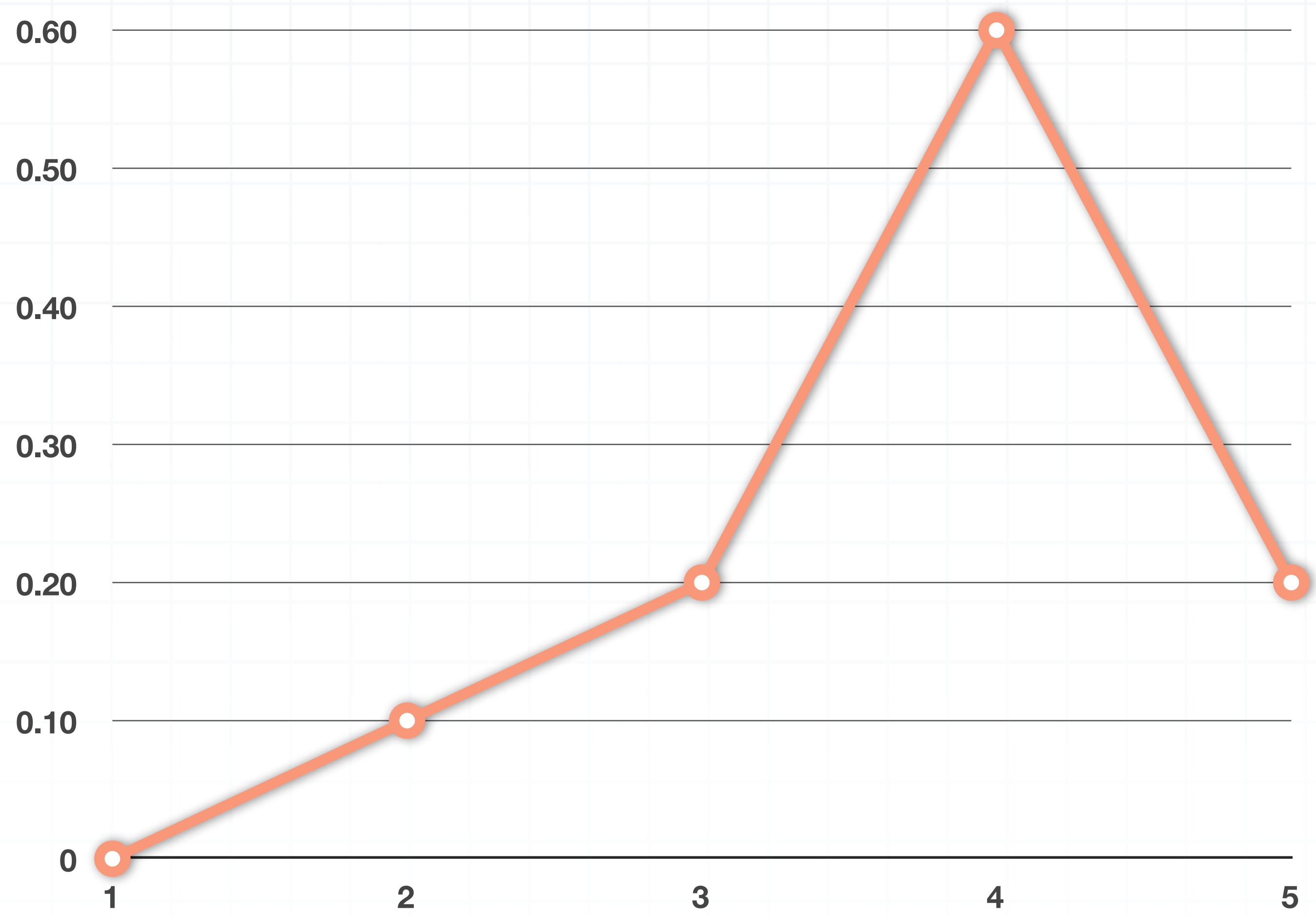


Solution:

Remember that the area under a density curve always adds to 1. Therefore everything greater than x must be

$$1 - 0.42 = 0.58 = 58\,\%$$

6. What percentage of the area in the density curve is between 3 and 5?



Solution:

We know that for a density curve, the area under the curve adds to 1. We can use area formulas to find the density under certain parts of the curve.

The area under the curve between 1 and 3 is a triangle, so the area can be found as

$$A = \frac{1}{2}bh = \frac{1}{2}(2)(0.2) = 0.2$$

Which means the area under the rest of the polygon is the area between 3 and 5 and must be

$$1 - 0.2 = 0.8$$

SYMMETRIC AND SKEWED DISTRIBUTIONS AND OUTLIERS

1. Which type of distribution is modeled in the box plot (symmetric, negatively skewed, or positively skewed)?

Solution:

This is an example of a symmetric distribution. The mean and the median are equal because the median of the data is in the middle of the box plot.

2. Which type of distribution is modeled in the box plot (symmetric, negatively skewed, or positively skewed)?

Solution:

This is an example of a positively skewed distribution. The median of the box plot is to the left of the middle of the box. This makes the mean greater than the median.

3. The ages (in months) that babies spoke for the first time are

$$6,\ 8,\ 9,\ 10,\ 10,\ 11,\ 11,\ 12,\ 12,\ 13,\ 15,\ 15,\ 18,\ 19,\ 20,\ 20,\ 21$$

Are there outliers in the data set? If so, state what they are. What is the best measure of central tendency for the data? What is the best measure of spread?

Solution:

This data has no outliers, so the best measure of central tendency is the mean, and the best measure of spread is the standard deviation. To find if there are outliers in the data, use the 1.5-IQR rule.

Low outliers are given by $Q_1 - 1.5(\text{IQR})$

High outliers are given by $Q_3 + 1.5(\text{IQR})$

In the data set, the median is 12. And the first and third quartiles are

$$Q_1 = \frac{10 + 10}{2} = 10$$

$$Q_3 = \frac{18 + 19}{2} = 18.5$$

The interquartile range is

$$Q_3 - Q_1 = 18.5 - 10 = 8.5$$

Now we can calculate the boundary for outliers.

Low outliers:

$$Q_1 - 1.5(\text{IQR})$$

$$10 - 1.5(8.5)$$

$$-2.75$$

High outliers:

$$Q_3 + 1.5(\text{IQR})$$

$$18.5 + 1.5(8.5)$$

$$31.25$$

Since the data set has no values below -2.75 or above 31.25, there are no outliers in the data set.

4. The number of text messages sent each day by Lucy's mom is

$$0,\ 18,\ 19,\ 20,\ 20,\ 20,\ 21,\ 23,\ 23,\ 23,\ 24,\ 24,$$

$$24,\ 24,\ 24,\ 25,\ 25,\ 25,\ 25,\ 25,\ 25,\ 30,\ 30,\ 31$$

Are there outliers in the data set? If so, state what they are. What is the best measure of central tendency for the data? What is the best measure of spread?

Solution:

This data has a low outlier of 0, so the best measure of central tendency is the median and the best measure of spread is the interquartile range. To see if there are outliers in the data use the 1.5-IQR rule.

Low outliers are given by $Q_1 - 1.5(\text{IQR})$

High outliers are given by $Q_3 + 1.5(\text{IQR})$

The median of the data set is 24. The first and third quartiles are

$$Q_1 = \frac{20+21}{2} = 20.5$$

$$Q_3 = \frac{25+25}{2} = 25$$

So the interquartile range is

$$Q_3 - Q_1 = 25 - 20.5 = 4.5$$

Now we can calculate where to look for outliers.

Low outliers:

$$Q_1 - 1.5(\text{IQR})$$

$$20.5 - 1.5(4.5)$$

$$13.75$$

High outliers:

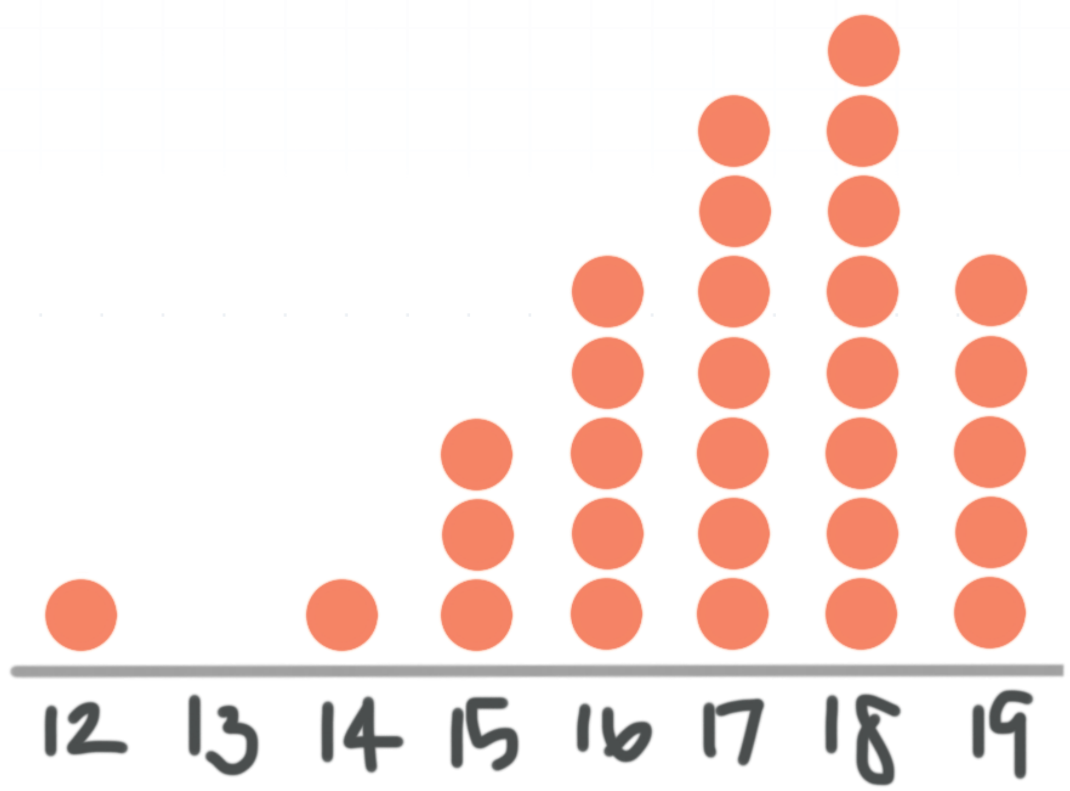
$$Q_3 + 1.5(\text{IQR})$$

$$25 + 1.5(4.5)$$

$$31.75$$

The data has a low outlier of 0 because it's less than 13.75. The data has no high outliers because no numbers in the set are greater than 31.75. Since the data has an outlier, the best measure of central tendency is the median and the best measure of spread is the interquartile range.

5. Describe the shape, center, and spread of the data. State if there are outliers and what they are if they exist.



Solution:

This data is negatively skewed, because it has a tail on the left-hand side with an outlier at 12. This means that the median will be the best measure of center and the interquartile range will be the best measure of spread. The median of the data is 17. The first and third quartile are $Q_1 = 16$ and $Q_3 = 18$, so the interquartile range is $Q_3 - Q_1 = 18 - 16 = 2$.

This means that low outliers are any values less than $Q_1 - 1.5(\text{IQR}) = 16 - 1.5(2) = 16 - 3 = 13$, and high outliers are any values greater than $Q_3 + 1.5(\text{IQR}) = 18 - 1.5(2) = 18 + 3 = 21$. Based on the dot plot, 12 is a low outlier, and there are no high outliers.

6. Describe the shape, center and spread of the data. State if there are outliers and what they are if they exist.

Solution:

This is a symmetric distribution that is approximately normal. There are no outliers in the data set. The best measure of center will be the mean (which is the same as the median) and the best measure of spread will be the standard deviation.

The mean of the data set is $\mu = 25$ and the population standard deviation is 1.5811. To get the standard deviation, we would need to first calculate variance.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{1(22-25)^2 + 2(23-25)^2 + 3(24-25)^2 + 4(25-25)^2 + 3(26-25)^2 + 2(27-25)^2 + 1(28-25)^2}{16}$$

$$\sigma^2 = \frac{1(-3)^2 + 2(-2)^2 + 3(-1)^2 + 4(0)^2 + 3(1)^2 + 2(2)^2 + 1(3)^2}{16}$$

$$\sigma^2 = \frac{1(9) + 2(4) + 3(1) + 4(0) + 3(1) + 2(4) + 1(9)}{16}$$

$$\sigma^2 = \frac{9 + 8 + 3 + 0 + 3 + 8 + 9}{16}$$

$$\sigma^2 = \frac{40}{16}$$

$$\sigma^2 = 2.5$$

Now take the square root of the population variance to find the population standard deviation.

$$\sqrt{\sigma^2} = \sqrt{2.5}$$

$$\sigma \approx 1.5811$$

NORMAL DISTRIBUTIONS AND Z-SCORES

1. A population has a mean of 62 and a standard deviation of 5. What is the z-score for a value of 50?

Solution:

The formula for a z-score is:

$$z = \frac{x - \mu}{\sigma}$$

We know the mean is $\mu = 62$ and that the standard deviation is $\sigma = 5$. The value of interest is $x = 50$. So the z-score is

$$z = \frac{50 - 62}{5} = -\frac{12}{5} = -2.4$$

2. What percentile is a z-score of -1.68?

Solution:

To find the percentile, we'll look up the z-score in the z-table. The amount in the table is 0.0465, which rounds to about 5%, so the z-score is associated with approximately the 5th percentile.

3. A population has a mean of 170 centimeters and a standard deviation of 8 centimeters. What percentage of the population has a value less than 154 centimeters?

Solution:

The mean is $\mu = 170$ and that the standard deviation is $\sigma = 8$. The value we're interested in is $x = 154$. So the z-score is

$$z = \frac{154 - 170}{8} = -\frac{16}{8} = -2.00$$

If we look up this z-score in the z-table, we find 0.0228, so about 2.28% of the population has a value less than 154 centimeters.

4. The mean diameter of a North American Native Pine tree is $18''$ with a standard deviation of $4''$. What is the approximate diameter for a tree in the 21st percentile for this distribution? Assume an approximately normal distribution.

Solution:

We know that the mean is $\mu = 18$ and that the standard deviation is $\sigma = 4$. If we look up the 21st percentile, or 0.2100 in a z-table, we get a z-score of -0.81. Plugging all this into the z-score formula, we get

$$z = \frac{x-\mu}{\sigma}$$

$$-0.81 = \frac{x-18}{4}$$

$$-0.81(4) = x - 18$$

$$-3.24 = x - 18$$

$$14.76 = x$$

5. The mean diameter of a North American Native Pine tree is $18''$ with a standard deviation of $4''$. According to the Empirical Rule, $68\,\%$ of North American Native Pines have a diameter between which two values? Assume an approximately normal distribution.

Solution:

According to the Empirical Rule, $68\,\%$ of an approximately normal distribution is within one standard deviation of the mean. Since we know that $\mu = 18$ and $\sigma = 4$, $68\,\%$ of these pines have a diameter on the interval

$$(18 - 4, 18 + 4)$$

$$(14, 22)$$

6. IQ scores are normally distributed with a mean of 100 and a standard deviation of 16. What percentage of the population has an IQ score between and 120 and 140?

Solution:

First, we need to find the percentage of people who have an IQ of at most 120 and then the percentage of people with an IQ of at most 140, and then subtract those percentages. This means we find those z-scores and look up the percentages on the z-table.

Since we know that $\mu = 100$ and $\sigma = 16$, the z-score for 120 is

$$z = \frac{120 - 100}{16} = \frac{20}{16} = 1.25$$

which gives $.8944$ in the z-table. The z-score for 140 is

$$z = \frac{140 - 100}{16} = \frac{40}{16} = 2.5$$

which gives $.9938$ in the z-table. Therefore, we can say that

$$.9938 - .8944 = .0994 = 9.94$$

percent of people have an IQ between 120 and 140.

CHEBYSHEV'S THEOREM

1. If the Empirical Rule tells us that $95\,\%$ of the area under the normal distribution falls within two standard deviations of the mean, what will Chebyshev's Theorem say about the same number of standard deviations?

Solution:

Because Chebyshev's Theorem has to apply to distributions of all shapes, it's always more conservative than the Empirical Rule.

Therefore, we know that Chebyshev's Theorem will only be able to conclude that less than $95\,\%$ of the area under distribution will fall within two standard deviations of the mean.

In actuality, Chebyshev's Theorem says that at least $75\,\%$ of the data under the probability distribution will fall within two standard deviations of the mean.

2. A basket of strawberries has a mean weight of 2 ounces with a standard deviation of 0.35 ounces. What percentage of the strawberries in the basket have a weight between 1.5 and 2.5 ounces?

Solution:

Determine the distance from the mean of 1.5 and 2.5, in terms of standard deviations.

$$k = \frac{1.5 - 2}{0.35} = -\frac{0.5}{0.35} \approx -1.43$$

$$k = \frac{2.5 - 2}{0.35} = \frac{0.5}{0.35} \approx 1.43$$

With $k = 1.43$, Chebyshev's Theorem gives

$$1 - \frac{1}{k^2} = 1 - \frac{1}{1.43^2} \approx 1 - \frac{1}{2.04} \approx 0.51$$

Because we found 0.51, we know at least 51% of the strawberries have a weight between 1.5 and 2.5 ounces.

■ 3. A pod of 580 migrating whales travels a mean distance of $2{,}000$ miles each year, with a standard deviation of 175 miles. How many whales in the pod travel between $1{,}600$ and $2{,}400$ miles?

Solution:

Determine the distance from the mean of $1{,}750$ and $2{,}250$, in terms of standard deviations.

$$k = \frac{1{,}600 - 2{,}000}{175} = -\frac{400}{175} \approx -2.29$$

$$k = \frac{2{,}400 - 2{,}000}{175} = \frac{400}{175} \approx 2.29$$

With $k = 2{,}29$, Chebyshev's Theorem gives

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2.29^2} \approx 1 - \frac{1}{5.22} \approx 0.81$$

Because we found 0.81, we know at least $81\,\%$ of the whales migrate between $1{,}600$ and $2{,}400$ miles. Then we can say that $81\,\%$ of the 580-whale pod is approximately 469.8 whales.

Of course we can't take part of a whale, so in order not to overstate Chebyshev's Theorem, we round down to 469 whales.

■ 4. A hockey team of 20 boys have a mean height of 73 inches, with a standard deviation of 1.8 inches. Find the height range for the central $90\,\%$ of team members.

Solution:

Using Chebyshev's Theorem,

$$0.9 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.9$$

$$1 = 0.1k^2$$

$$k^2 = 10$$

$$k \approx 3.16$$

Approximately 3.16 standard deviations above the mean gives us a height of

$$73 + 3.16(1.8)$$

$$73 + 5.69$$

$$78.69$$

And 3.16 standard deviations below the mean gives us a height of

$$73 - 3.16(1.8)$$

$$73 - 5.69$$

$$67.31$$

So at least 90% of the players have heights between 67.31 and 78.69 inches.

■ 5. A university with $40{,}000$ students accepts an average of $10{,}000$ new students each year, with a standard deviation of 500 students. Find the values that make up the middle 75% of the yearly acceptance range.

Solution:

Using Chebyshev's Theorem,

$$0.75 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.75$$

$$1 = 0.25k^2$$

$$k^2 = 4$$

$$k = 2$$

Two standard deviations above the mean gives us the upper end of students accepted.

$$10{,}000 + 2(500)$$

$$10{,}000 + 1{,}000$$

$$11{,}000$$

And two standard deviations below the mean gives us the lower end of students accepted.

$$10{,}000 - 2(500)$$

$$10{,}000 - 1{,}000$$

$$9{,}000$$

So at least 75% of the time (75% of years), the school accepts between $9{,}000$ and $11{,}000$ students.

6. A pack of 26 wolves have a mean weight of 100 pounds, with a standard deviation of 24 pounds. Find the weight range for the central $82\,\%$ of the wolves.

Solution:

Using Chebyshev's Theorem,

$$0.82 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.82$$

$$1 = 0.18k^2$$

$$k^2 \approx 5.56$$

$$k \approx 2.36$$

Approximately 2.36 standard deviations above the mean gives us a weight of

$$100 + 2.36(24)$$

$$100 + 56.57$$

$$156.57$$

And 2.36 standard deviations below the mean gives us a weight of

$$100 - 2.36(24)$$

$$100 - 56.57$$

$$43.43$$

So at least 82% of the wolves have weights between 43.43 and 156.57 pounds.

COVARIANCE

1. A bakery records sales and number of customers for a sample of hours throughout the week. Calculate the covariance of customers and sales.

Customers	4	7	12	2	3	9	15
Sales	45.75	36.00	58.5	20.00	15.80	39.95	123.45

Solution:

We'll find the mean number of customers,

$$\bar{x} = \frac{4+7+12+2+3+9+15}{7} \approx 7.43$$

and the mean revenue in dollars.

$$\bar{y} = \frac{45.75+36+58.50+20+15.80+39.95+123.45}{7} \approx 48.49$$

Now we'll use the means to find the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (4 - 7.43)(45.75 - 48.49) + (7 - 7.43)(36 - 48.49)$$

$$+(12 - 7.43)(58.50 - 48.49) + (2 - 7.43)(20 - 48.49)$$

$$+(3 - 7.43)(15.80 - 48.49) + (9 - 7.43)(39.90 - 48.49)$$

$$+(15 - 7.43)(123.45 - 48.49)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) \approx 913.9929$$

$$s_{xy} \approx \frac{913.9929}{7 - 1}$$

$$s_{xy} \approx 152.33$$

2. The cost of the stock of two unrelated companies over five days is recorded in the table. Calculate the covariance of the stocks.

Company X	13	13.75	12.70	13.15	14.80
Company Y	21.05	21.55	20.95	21.75	21.50

Solution:

We'll find the mean cost of the stock for company X,

$$\bar{X} = \frac{13 + 13.75 + 12.70 + 13.15 + 14.80}{5} = 13.48$$

and the mean cost of the stock for company Y.

$$\bar{Y} = \frac{21.05 + 21.55 + 20.95 + 21.75 + 21.50}{5} = 21.36$$

Now we'll use the means to find the sample covariance.

$$s_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (13 - 13.48)(21.05 - 21.36)$$

$$+(13.75 - 13.48)(21.55 - 21.36) + (12.70 - 13.48)(20.95 - 21.36)$$

$$+(13.15 - 13.48)(21.75 - 21.36) + (14.80 - 13.48)(21.50 - 21.36)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 0.576$$

$$s_{XY} = \frac{0.576}{5-1}$$

$$s_{XY} = 0.144$$

3. The following table represents temperatures, in Celsius, during a sample of 5 days in two cities with a distance of 50 miles between them. Calculate the covariance.

City X	25	23	24.5	20	18
City Y	23	24	21	18	22

Solution:

We'll find the mean temperature in city X,

$$\bar{X} = \frac{25 + 23 + 24.5 + 20 + 18}{5} = 22.1$$

and the mean temperature in city Y.

$$\bar{Y} = \frac{23 + 24 + 21 + 18 + 22}{5} = 21.6$$

Now we'll use the means to find the sample covariance.

$$s_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (25 - 22.1)(23 - 21.6) + (23 - 22.1)(24 - 21.6)$$

$$+(24.5 - 22.1)(21 - 21.6) + (20 - 22.1)(18 - 21.6)$$

$$+(18 - 22.1)(22 - 21.6)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 10.7$$

$$s_{XY} = \frac{10.7}{5 - 1}$$

$$s_{XY} = 2.675$$

4. David prepares for his annual math and physics exams and decides to take four practice tests for each subject. Calculate the covariance for his test scores for math and physics.

Math, X	85	89	89	93
Physics, Y	92	93	89	90

Solution:

We'll find the mean score of the math practice tests,

$$\bar{X} = \frac{85 + 89 + 89 + 93}{4} = 89$$

and the mean score of the physics practice tests.

$$\bar{Y} = \frac{92 + 93 + 89 + 90}{4} = 91$$

Now we'll use the means to find the sample covariance.

$$s_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (85 - 89)(92 - 91) + (89 - 89)(93 - 91)$$

$$+(89 - 89)(89 - 91) + (93 - 89)(90 - 91)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = -8$$

$$s_{XY} = \frac{-8}{4 - 1}$$

$$s_{XY} \approx -2.67$$

■ 5. Mark and John exercise daily and record their minutes of daily exercise over 10 days. Calculate the covariance.

Mark	53	57	63	55	45	50	65	60	59	70
John	65	55	60	53	30	45	25	65	57	50

Solution:

We'll find the mean amount of time, in minutes, that Mark spent exercising daily for the last 10 days,

$$\bar{M} = \frac{53 + 57 + 63 + 55 + 45 + 50 + 65 + 60 + 59 + 70}{10} = 57.7$$

and the mean amount of time, in minutes, that John spent exercising daily for the last 10 days.

$$\bar{J} = \frac{65 + 55 + 60 + 53 + 30 + 45 + 25 + 65 + 57 + 50}{10} = 50.5$$

Now we'll use the means to find the sample covariance.

$$s_{MJ} = \frac{\sum (M_i - \bar{M})(J_i - \bar{J})}{n - 1}$$

$$\sum (M_i - \bar{M})(J_i - \bar{J}) = (53 - 57.7)(65 - 50.5) + (57 - 57.7)(55 - 50.5)$$

$$+(63 - 57.7)(60 - 50.5) + (55 - 57.7)(53 - 50.5)$$

$$+(45 - 57.7)(30 - 50.5) + (50 - 57.7)(45 - 50.5)$$

$$+(65 - 57.7)(25 - 50.5) + (60 - 57.7)(65 - 50.5)$$

$$+(59 - 57.7)(57 - 50.5) + (70 - 57.7)(50 - 50.5)$$

$$\sum (M_i - \bar{M})(J_i - \bar{J}) = 124.5$$

$$s_{MJ} = \frac{124.5}{10 - 1}$$

$$s_{MJ} \approx 13.83$$

■ 6. An annual return on investment of two stocks over the last 7 years is recorded in the table. Calculate the covariance.

Stock X	3.5	2.4	1.4	-0.5	0.7	1.1	0.5
Stock Y	2.4	1.7	2.1	1.8	2.1	-0.4	0.8

Solution:

We'll find the mean annual return of stock X,

$$\bar{X} = \frac{3.5 + 2.4 + 1.4 + (-0.5) + 0.7 + 1.1 + 0.5}{7} = 1.3$$

and the mean annual return of stock Y.

$$\bar{Y} = \frac{2.4 + 1.7 + 2.1 + 1.8 + 2.1 + (-0.4) + 0.8}{7} = 1.5$$

Now we'll use the means to find the sample covariance.

$$s_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (3.5 - 1.3)(2.4 - 1.5) + (2.4 - 1.3)(1.7 - 1.5)$$

$$+(1.4 - 1.3)(2.1 - 1.5) + (-0.5 - 1.3)(1.8 - 1.5)$$

$$+(0.7 - 1.3)(2.1 - 1.5) + (1.1 - 1.3)(-0.4 - 1.5)$$

$$+(0.5 - 1.3)(0.8 - 1.5)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 2.3$$

$$s_{XY} = \frac{2.3}{7 - 1}$$

$$s_{XY} \approx 0.38$$

CORRELATION COEFFICIENT

■ 1. Calculate the correlation coefficient of the newborns‘ weight and body length, and then interpret the result.

Weight, kg	Body length, cm
3.55	51
4.01	54
3.05	50
5.35	60
4.22	52
6.12	61
7.45	63
5.95	59
6.35	68
6.98	74

Solution:

Start by calculating mean weight,

$$\bar{x} = \frac{3.55 + 4.01 + 3.05 + 5.35 + 4.22 + 6.12 + 7.45 + 5.95 + 6.35 + 6.98}{10}$$

$$\bar{x} = 5.303$$

and mean body length.

$$\bar{y} = \frac{51 + 54 + 50 + 60 + 52 + 61 + 63 + 59 + 68 + 74}{10}$$

$$\bar{y} = 59.2$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (3.55 - 5.303)(51 - 59.2)$$

$$+(4.01 - 5.303)(54 - 59.2) + (3.05 - 5.303)(50 - 59.2)$$

$$+(5.35 - 5.303)(60 - 59.2) + (4.22 - 5.303)(52 - 59.2)$$

$$+(6.12 - 5.303)(61 - 59.2) + (7.45 - 5.303)(63 - 59.2)$$

$$+(5.95 - 5.303)(59 - 59.2) + (6.35 - 5.303)(68 - 59.2)$$

$$+(6.98 - 5.303)(74 - 59.2)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 93.194$$

$$s_{xy} = \frac{93.194}{10 - 1}$$

$$s_{xy} \approx 10.35$$

Next we'll need the standard deviation for weight,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=(3.55-5.303)^2+(4.01-5.303)^2+(3.05-5.303)^2$$

$$+(5.35-5.303)^2+(4.22-5.303)^2+(6.12-5.303)^2+(7.45-5.303)^2$$

$$+(5.95-5.303)^2+(6.35-5.303)^2+(6.98-5.303)^2$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2\approx 20.60$$

$$s_x\approx\sqrt{\frac{20.60}{10-1}}$$

$$s_x\approx 1.513$$

and the standard deviation for body length.

$$s_y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=(51-59.2)^2+(54-59.2)^2+(50-59.2)^2+(60-59.2)^2$$

$$+(52-59.2)^2+(61-59.2)^2+(63-59.2)^2+(59-59.2)^2$$

$$+(68-59.2)^2+(74-59.2)^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=545.6$$

$$s_y=\sqrt{\frac{545.6}{10-1}}$$

$$s_y \approx 7.786$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{10.35}{1.513 \cdot 7.786}$$

$$r_{xy} \approx 0.8786$$

The value of the correlation coefficient indicates that there's a strong positive correlation between the weight and body length of newborns.

2. Oliver is wondering whether there's a correlation between the number of hours his classmates studied to prepare for the exam and their exam scores. He surveyed five classmates and recorded the data in a table. Calculate the correlation coefficient.

Study hours	6	2	11	7	5
Exam score	85	79	84	89	91

Solution:

Start by calculating mean number of study hours,

$$\bar{x} = \frac{6 + 2 + 11 + 7 + 5}{5}$$

$$\bar{x} = 6.2$$

and mean exam score.

$$\bar{y} = \frac{85 + 79 + 84 + 89 + 91}{5}$$

$$\bar{y} = 85.6$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (6 - 6.2)(85 - 85.6) + (2 - 6.2)(79 - 85.6)$$

$$+(11 - 6.2)(84 - 85.6) + (7 - 6.2)(89 - 85.6) + (5 - 6.2)(91 - 85.6)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 16.4$$

$$s_{xy} = \frac{16.4}{5 - 1}$$

$$s_{xy} = 4.1$$

Next we'll need the standard deviation for the number of study hours,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (6 - 6.2)^2 + (2 - 6.2)^2 + (11 - 6.2)^2 + (7 - 6.2)^2 + (5 - 6.2)^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 42.8$$

$$s_x = \sqrt{\frac{42.8}{5 - 1}}$$

$$s_x \approx 3.271$$

and the standard deviation for exam score.

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}}$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = (85 - 85.6)^2 + (79 - 85.6)^2 + (84 - 85.6)^2$$

$$+(89 - 85.6)^2 + (91 - 85.6)^2$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = 87.2$$

$$s_y = \sqrt{\frac{87.2}{5 - 1}}$$

$$s_y \approx 4.669$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{4.1}{3.271 \cdot 4.669}$$

$$r_{xy} \approx 0.2685$$

There's a weak positive correlation between the number of study hours and exam score.

3. Calculate the value of the Pearson correlation coefficient for the age, in years, and blood glucose levels, in mg/dL, then interpret the result.

Age	28	35	58	42	21	63	46
Blood glucose	101	93	95	105	93	89	100

Solution:

Start by calculating mean age,

$$\bar{x} = \frac{28 + 35 + 58 + 42 + 21 + 63 + 46}{7}$$

$$\bar{x} \approx 41.86$$

and mean blood glucose.

$$\bar{y} = \frac{101 + 93 + 95 + 105 + 93 + 89 + 100}{7}$$

$$\bar{y} \approx 96.57$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (28 - 41.86)(101 - 96.57)$$

$$+(35 - 41.86)(93 - 96.57) + (58 - 41.86)(95 - 96.57)$$

$$+(42 - 41.86)(105 - 96.57) + (21 - 41.86)(93 - 96.57)$$

$$+(63 - 41.86)(89 - 96.57) + (46 - 41.86)(100 - 96.57)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = -56.7286$$

$$s_{xy} = \frac{-132.4286}{7 - 1}$$

$$s_{xy} \approx -22.071$$

Next we'll need the standard deviation for age,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (28 - 41.86)^2 + (35 - 41.86)^2 + (58 - 41.86)^2$$

$$+(42-41.86)^2+(21-41.86)^2+(63-41.86)^2+(46-41.86)^2$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=1{,}398.8572$$

$$s_x=\sqrt{\frac{1{,}398.8572}{7-1}}$$

$$s_x\approx 15.269$$

and the standard deviation for blood glucose level.

$$s_y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=(101-96.57)^2+(93-96.57)^2+(95-96.57)^2$$

$$+(105-96.57)^2+(93-96.57)^2+(89-96.57)^2+(100-96.57)^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=187.7143$$

$$s_y=\sqrt{\frac{187.7143}{7-1}}$$

$$s_y\approx 5.593$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{-22.071}{15.269 \cdot 5.593}$$

$$r_{xy} \approx -0.2584$$

The value of the correlation coefficient indicates that there's a weak negative correlation between age and blood glucose levels.

■ 4. Maria likes discovering interesting correlations. She decides to choose random six days and record the data for shark attacks and ice cream sales in her coastal city. How should she interpret the correlation coefficient.

Shark attacks	4	2	8	11	5	9
Ice cream sales	38	30	55	61	38	42

Solution:

Start by calculating mean shark attacks,

$$\bar{x} = \frac{4+2+8+11+5+9}{6}$$

$$\bar{x} = 6.5$$

and mean ice cream sales.

$$\bar{y} = \frac{38 + 30 + 55 + 61 + 38 + 42}{6}$$

$$\bar{y} = 44$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (4 - 6.5)(38 - 44) + (2 - 6.5)(30 - 44) + (8 - 6.5)(55 - 44)$$

$$+(11 - 6.5)(61 - 44) + (5 - 6.5)(38 - 44) + (9 - 6.5)(42 - 44)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 175$$

$$s_{xy} = \frac{175}{6 - 1}$$

$$s_{xy} = 35$$

Next we'll need the standard deviation for shark attacks,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (4 - 6.5)^2 + (2 - 6.5)^2 + (8 - 6.5)^2 + (11 - 6.5)^2$$

$$+(5 - 6.5)^2 + (9 - 6.5)^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 57.5$$

$$s_x = \sqrt{\frac{57.5}{6-1}}$$

$$s_x \approx 3.391$$

and the standard deviation for ice cream sales.

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (38-44)^2 + (30-44)^2 + (55-44)^2 + (61-44)^2$$

$$+(38-44)^2 + (42-44)^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 682$$

$$s_y = \sqrt{\frac{682}{6-1}}$$

$$s_y \approx 11.679$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{35}{3.391 \cdot 11.679}$$

$$r_{xy} \approx 0.8838$$

There's strong positive correlation between the number of shark attacks and ice cream sales on a particular day.

However, it seems unlikely that an increase in ice cream sales is the cause of an increase in shark attacks. Instead, it seems more likely that a hot summer day causes more people to buy ice cream and more people to swim in the ocean, leading to more shark attacks.

■ 5. Calculate and interpret the correlation coefficient of the variables.

Hand length, cm	Height, cm
12	158
15	160
11	157
13	164
9	150
18	178
16	169
17	156

Solution:

Start by calculating mean hand length,

$$\bar{x} = \frac{12 + 15 + 11 + 13 + 9 + 18 + 16 + 17}{8}$$

$$\bar{x} = 13.875$$

and mean height.

$$\bar{y} = \frac{158 + 160 + 157 + 164 + 150 + 178 + 169 + 156}{8}$$

$$\bar{y} = 161.5$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (12 - 13.875)(158 - 161.5) + (15 - 13.875)(160 - 161.5)$$

$$+(11 - 13.875)(157 - 161.5) + (13 - 13.875)(164 - 161.5)$$

$$+(9 - 13.875)(150 - 161.5) + (18 - 13.875)(178 - 161.5)$$

$$+(16 - 13.875)(169 - 161.5) + (17 - 13.875)(156 - 161.5)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 138.5$$

$$s_{xy} = \frac{138.5}{8 - 1}$$

$$s_{xy} \approx 19.786$$

Next we'll need the standard deviation for hand length,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (12 - 13.875)^2 + (15 - 13.875)^2 + (11 - 13.875)^2$$

$$+(13 - 13.875)^2 + (9 - 13.875)^2 + (18 - 13.875)^2$$

$$+(16 - 13.875)^2 + (17 - 13.875)^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 68.875$$

$$s_x = \sqrt{\frac{68.875}{8-1}}$$

$$s_x \approx 3.137$$

and the standard deviation for height.

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = (158 - 161.5)^2 + (160 - 161.5)^2 + (157 - 161.5)^2$$

$$+(164 - 161.5)^2 + (150 - 161.5)^2 + (178 - 161.5)^2$$

$$+(169 - 161.5)^2 + (156 - 161.5)^2$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = 532$$

$$s_y = \sqrt{\frac{532}{8-1}}$$

$$s_y \approx 8.718$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r_{xy} \approx \frac{19.786}{3.137 \cdot 8.718}$$

$$r_{xy} \approx 0.7235$$

There's a strong positive correlation between hand length and height.

6. Calculate and interpret the value of the correlation coefficient for the correlation between systolic blood pressure, in mmHg, and weight, in lbs.

SBP, mmHg	Weight, lbs
138	167
125	153
145	149
156	165
132	170
148	175
160	180

Solution:

Start by calculating mean systolic blood pressure,

$$\bar{x} = \frac{138 + 125 + 145 + 156 + 132 + 148 + 160 + 135 + 150 + 155}{10}$$

$$\bar{x} = 144.4$$

and mean weight.

$$\bar{y} = \frac{167 + 153 + 149 + 165 + 170 + 175 + 180 + 140 + 190 + 155}{10}$$

$$\bar{y} = 164.4$$

Calculate the sample covariance.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (138 - 144.4)(167 - 164.4) + (125 - 144.4)(153 - 164.4)$$

$$+(145 - 144.4)(149 - 164.4) + (156 - 144.4)(165 - 164.4)$$

$$+(132 - 144.4)(170 - 164.4) + (148 - 144.4)(175 - 164.4)$$

$$+(160 - 144.4)(180 - 164.4) + (135 - 144.4)(140 - 164.4)$$

$$+(150 - 144.4)(190 - 164.4) + (155 - 144.4)(155 - 164.4)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 687.4$$

$$s_{xy} = \frac{687.4}{10 - 1}$$

$$s_{xy} \approx 76.378$$

Next we'll need the standard deviation for systolic blood pressure,

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (138 - 144.4)^2 + (125 - 144.4)^2 + (145 - 144.4)^2$$

$$+(156 - 144.4)^2 + (132 - 144.4)^2 + (148 - 144.4)^2 + (160 - 144.4)^2$$

$$+(135 - 144.4)^2 + (150 - 144.4)^2 + (155 - 144.4)^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1194.4$$

$$s_x = \sqrt{\frac{1194.4}{10 - 1}}$$

$$s_x \approx 11.52$$

and the standard deviation for weight.

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}}$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = (167 - 164.4)^2 + (153 - 164.4)^2 + (149 - 164.4)^2$$

$$+(165-164.4)^2+(170-164.4)^2+(175-164.4)^2+(180-164.4)^2$$

$$+(140-164.4)^2+(190-164.4)^2+(155-164.4)^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=2{,}100.4$$

$$s_y=\sqrt{\frac{2{,}100.4}{10-1}}$$

$$s_y\approx 15.277$$

Now we can plug the covariance and standard deviations into the formula for correlation.

$$r_{xy}=\frac{s_{xy}}{s_xs_y}$$

$$r_{xy}\approx\frac{76.378}{11.52\cdot 15.277}$$

$$r_{xy}\approx 0.4340$$

There's a moderate positive correlation between systolic blood pressure and weight.

WEIGHTED MEANS AND GROUPED DATA

1. An investor purchases shares of a particular stock on the same date every month for 12 months. He records the price and number of shares each month. Calculate the mean share price.

Stock price	Shares
8	30
10	12
14	10
9	25
6	35
13	15
18	10
21	5
25	7
27	10
28	8
31	4

Solution:

We can calculate the weighted sample mean share price.

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$\sum_{i=1}^{n} w_i x_i = 30(8) + 12(10) + 10(14) + 25(9) + 35(6) + 15(13) + 10(18)$$

$$+5(21) + 7(25) + 10(27) + 8(28) + 4(31)$$

$$\sum_{i=1}^{n} w_i x_i = 2{,}208$$

$$\bar{x} = \frac{2{,}208}{30 + 12 + 10 + 25 + 35 + 15 + 10 + 5 + 7 + 10 + 8 + 4}$$

$$\bar{x} \approx 12.91$$

The mean share price is $\$12.91$.

■ 2. A chemistry course teacher weights class discussions at 0.05, quizzes at 0.10, and group projects at 0.40. Given the grades for one student in the table below, calculate her final grade.

Assignment	Grade	Weight
Quiz 1	88	0.10
Discussion 1	92	0.05
Quiz 2	93	0.10
Discussion 2	90	0.05
Quiz 3	85	0.10
Discussion 3	94	0.05
Quiz 4	97	0.10
Discussion 4	80	0.05
Group Project	85	0.40

Solution:

This is the entire population of scores for the single student, so we'll calculate the weighted population mean.

$$\mu = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

$$\sum_{i=1}^{N} w_i x_i = 0.10(88) + 0.05(92) + 0.10(93) + 0.05(90) + 0.10(85)$$

$$+0.05(94) + 0.10(97) + 0.05(80) + 0.40(85)$$

$$\sum_{i=1}^{N} w_i x_i = 88.1$$

$$\mu = \frac{88.1}{0.1 + 0.05 + 0.1 + 0.05 + 0.1 + 0.05 + 0.1 + 0.05 + 0.4}$$

$$\mu = 88.1$$

The student's final grade is 88.1.

■ 3. Given the dataset $\{12, 15, 8, 21, 25, 14, 16, 18, 10\}$, divide it into four groups and find the sample variance and standard deviation.

Solution:

First, we need to put the data set into ascending order.

$$\{8, 10, 12, 14, 15, 16, 18, 21, 25\}$$

Now let's set up a table with four groups, find the midpoint of each group, and then the frequency of each group.

Group	Midpoint	Frequency
8 - 12	10	3
13 - 17	15	3
18 - 22	20	2
23 - 27	25	1

Then the estimate of the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i M_i}{n}$$

$$\bar{x} = \frac{(3)(10) + (3)(15) + (2)(20) + (1)(25)}{9}$$

$$\bar{x} = 15.55$$

We can use this mean to estimate the variance of the sample,

$$s^2 = \frac{\sum_{i=1}^{n} f_i (M_i - \bar{x})^2}{n - 1}$$

$$s^2 = \frac{3(10 - 15.55)^2 + 3(15 - 15.55)^2 + 2(20 - 15.55)^2 + 1(25 - 15.55)^2}{9 - 1}$$

$$s^2 \approx 27.778$$

and then the standard deviation of the sample will be the square root of the variance.

$$s = \sqrt{s^2}$$

$$s \approx \sqrt{27.778}$$

$$s \approx 5.270$$

■ 4. A sample of book club members record the number of books they read last year. Calculate the mean number of books per member.

Number of books	0	1	2	3	4	5
Number of people	25	15	18	5	12	3

Solution:

We can calculate the weighted sample mean.

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$\bar{x} = \frac{25(0) + 15(1) + 18(2) + 5(3) + 12(4) + 3(5)}{25 + 15 + 18 + 5 + 12 + 3}$$

$$\bar{x} \approx 1.654$$

Each book club member read 1.654 books, on average, during the last year.

5. The frequency distribution represents the number of pizza orders a local pizza restaurant received each day over the last 20 days. Calculate the weighted sample mean, variance, and standard deviation.

Orders	Number of days
5 - 7	5
8 - 10	4
11 - 13	4
14 - 16	3
17 - 19	3
20 - 22	1

Solution:

First, we need to find the midpoint of each class.

Number of orders	Midpoint	Number of days
5 - 7	6	5
8 - 10	9	4
11 - 13	12	4
14 - 16	15	3
17 - 19	18	3
20 - 22	21	1

Then the estimate of the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i M_i}{n}$$

$$\bar{x} = \frac{5(6) + 4(9) + 4(12) + 3(15) + 3(18) + 1(21)}{5 + 4 + 4 + 3 + 3 + 1}$$

$$\bar{x} = 11.7$$

We can use this mean to estimate the variance of the sample,

$$s^2 = \frac{\sum_{i=1}^{n} f_i (M_i - \bar{x})^2}{n - 1}$$

$$\sum_{i=1}^{n} f_i (M_i - \bar{x})^2 = 5(6 - 11.7)^2 + 4(9 - 11.7)^2 + 4(12 - 11.7)^2$$

$$+3(15 - 11.7)^2 + 3(18 - 11.7)^2 + 1(21 - 11.7)^2$$

$$\sum_{i=1}^{n} f_i (M_i - \bar{x})^2 \approx 430.198$$

$$s^2 \approx \frac{430.198}{20 - 1}$$

$$s^2 \approx 22.642$$

and then the standard deviation of the sample will be the square root of the variance.

$$s = \sqrt{s^2}$$

$$s \approx \sqrt{22.642}$$

$$s \approx 4.758$$

■ 6. Use the sample data to find the mean, variance, and standard deviation of commute time.

Commute time	Number of people
1 - 5	1
6 - 10	4
11 - 15	6
16 - 20	3
21 - 25	10
26 - 30	13

Solution:

First, we need to find the midpoint of each group.

Commute time	Midpoint	Number of people
1 - 5	3	1
6 - 10	8	4
11 - 15	13	6
16 - 20	18	3
21 - 25	23	10
26 - 30	28	13

Then the estimate of the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i M_i}{n}$$

$$\bar{x} = \frac{1(3) + 4(8) + 6(13) + 3(18) + 10(23) + 13(28)}{1 + 4 + 6 + 3 + 10 + 13}$$

$$\bar{x} \approx 20.568$$

We can use this mean to estimate the variance of the sample,

$$s^2 = \frac{\sum_{i=1}^{n} f_i (M_i - \bar{x})^2}{n - 1}$$

$$\sum_{i=1}^{n} f_i (M_i - \bar{x})^2 = 1(3 - 20.578)^2 + 4(8 - 20.578)^2 + 6(13 - 20.578)^2$$

$$+3(18 - 20.578)^2 + 10(23 - 20.578)^2 + 13(28 - 20.578)^2$$

$$\sum_{i=1}^{n} f_i (M_i - \bar{x})^2 = 2{,}081.088$$

$$s^2 = \frac{2{,}081.088}{37 - 1}$$

$$s^2 \approx 57.808$$

and then the standard deviation of the sample will be the square root of the variance.

$$s = \sqrt{s^2}$$

$$s \approx \sqrt{57.808}$$

$$s \approx 7.603$$