Probability & Statistics Workbook Solutions

Data distributions

krista king MATH

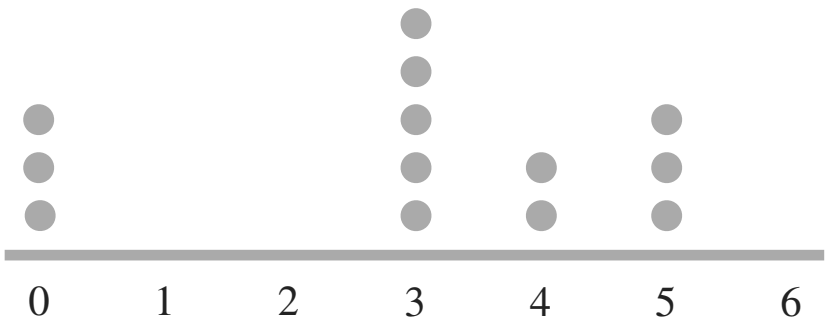
MEAN, VARIANCE, AND STANDARD DEVIATION

■ 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 to the nearest hundredth? Use the sample mean rounded to the nearest hundredth for your calculation.



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=414**

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

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

for the data set. 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 you're measuring. If, for example, you were measuring the price of a soft drink at a state fair, and you found a standard deviation of $5 that is a large standard deviation. It is large because soft drinks usually do not cost very much and this would tell you that you need to hunt around for the best price.

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

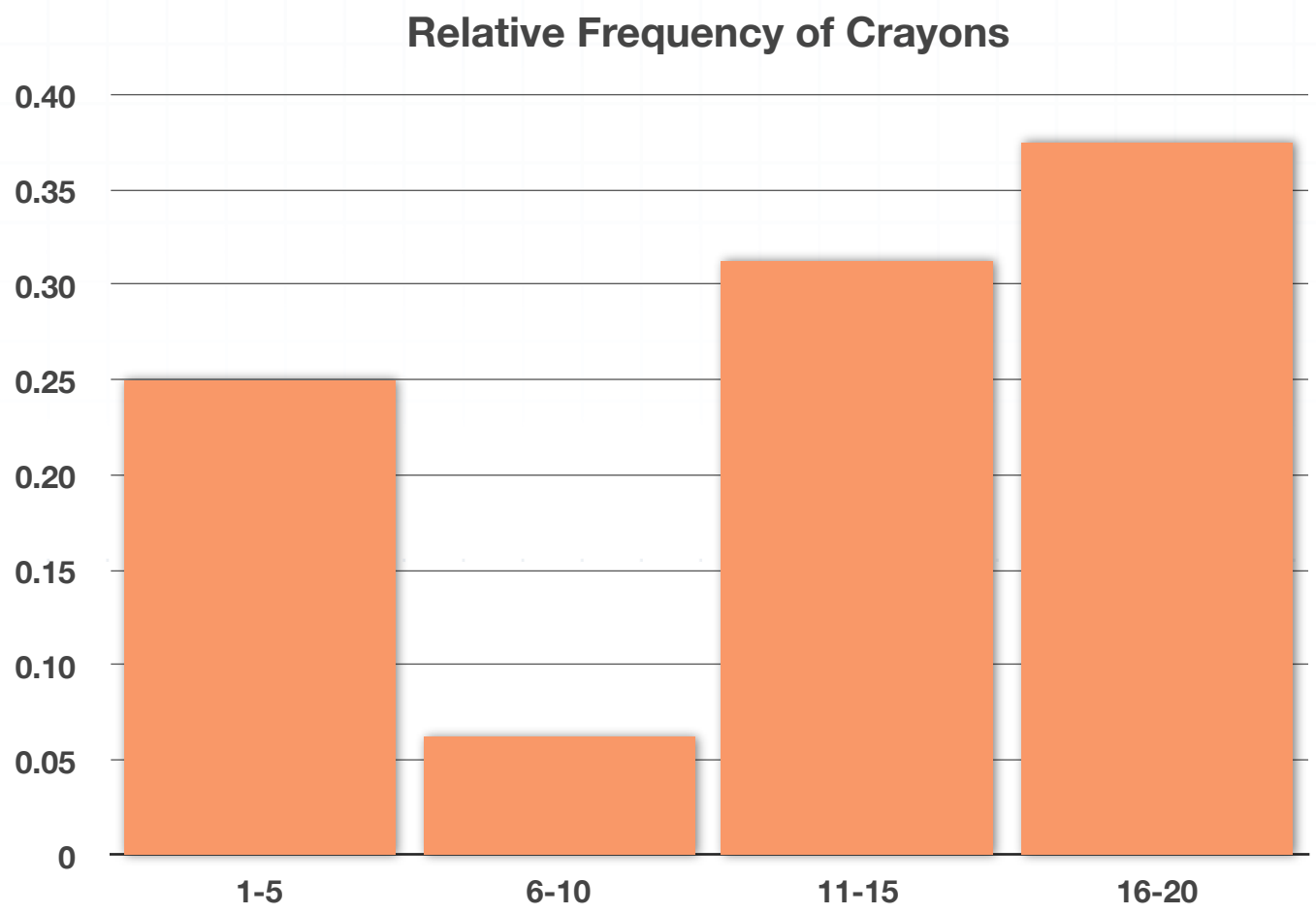
First count the number of items in each frequency interval and add that to the table, as well as calculate the total number of crayons.

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.

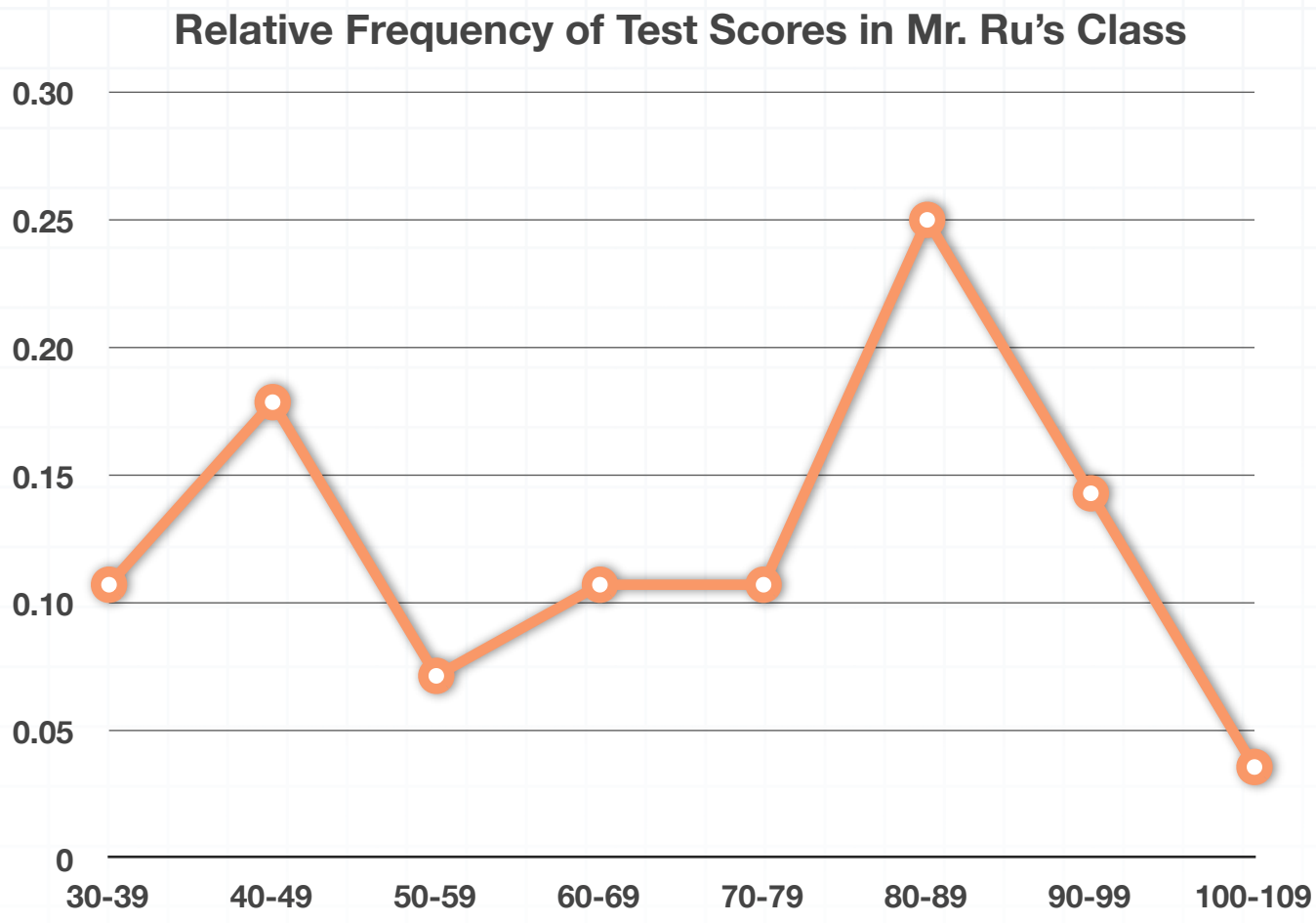
Score	Frequency	Relative Frequency
30-39		
40-49		
50-59		
60-69		
70-79		
80-89		
90-99		
100-109		
Totals:		

Solution:

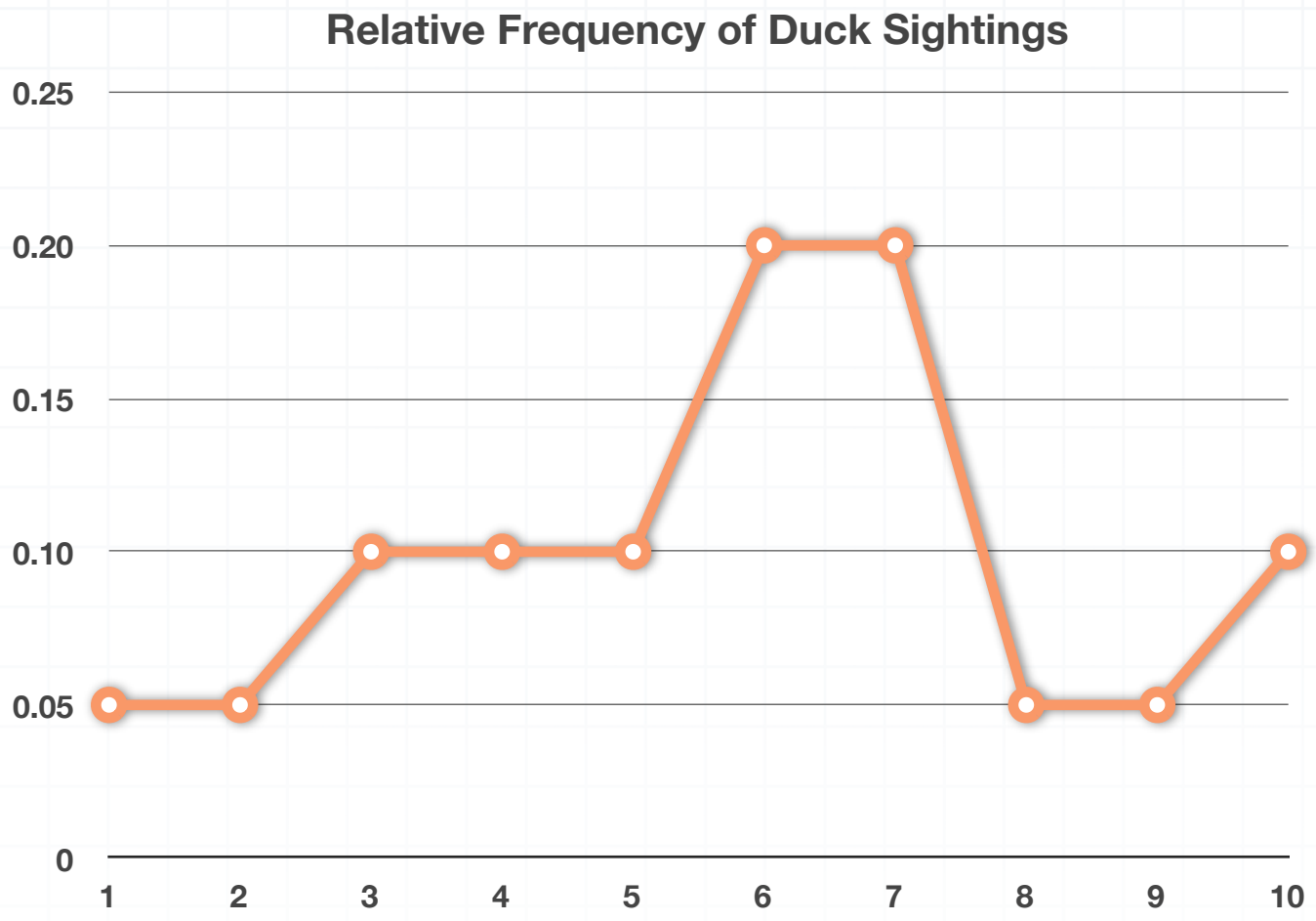
The first step to completing the frequency table is to count the scores in each interval, then use those frequencies and the total number of test scores to calculate the relative frequencies.

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.

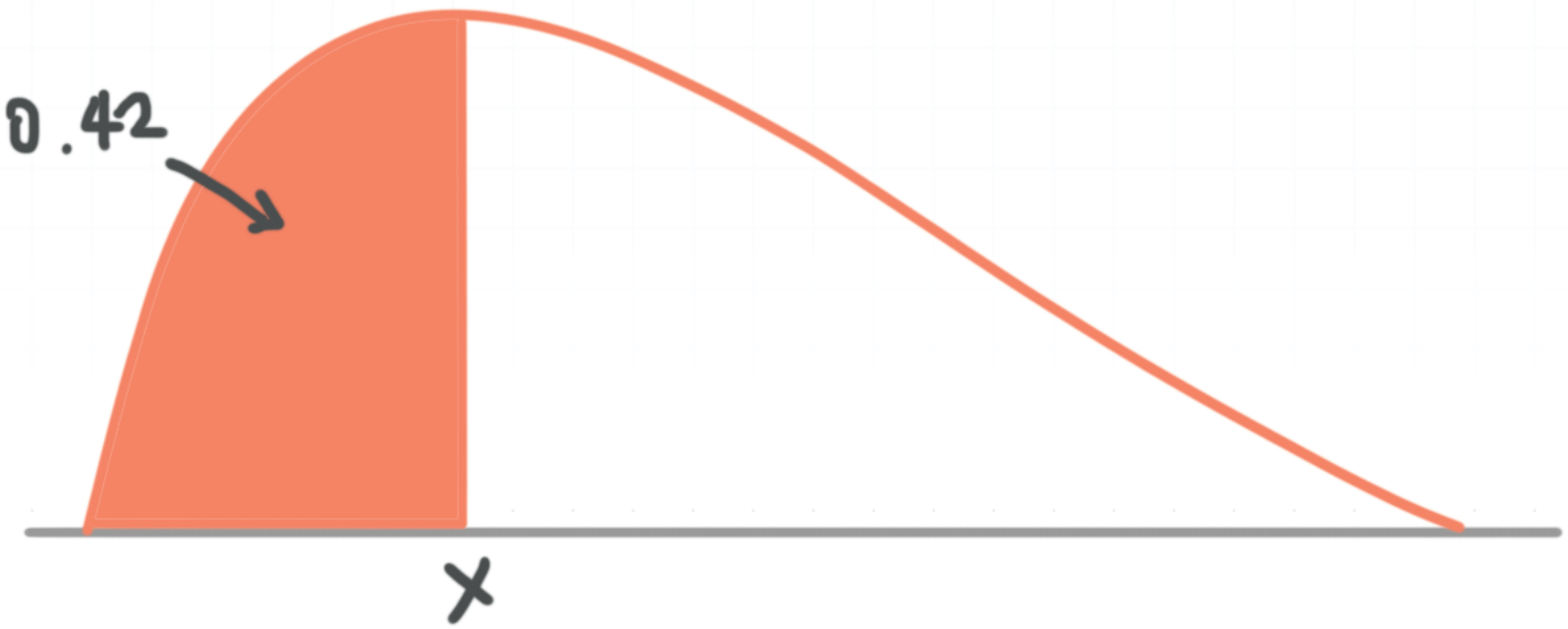
Ducks	Relative Frequency
6	0.20
7	0.20
8	0.05
9	0.05
10	0.10

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?

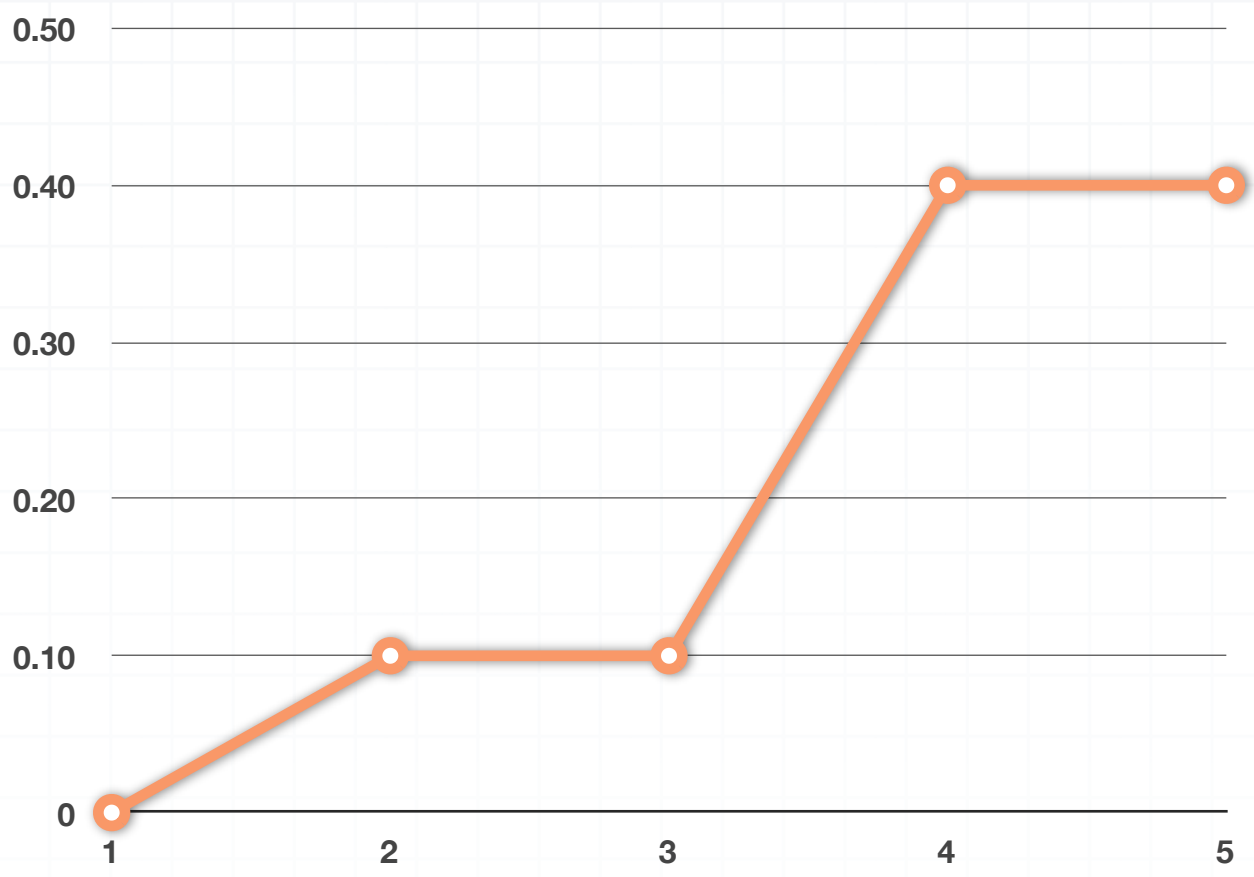


Solution:

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 2 is a triangle, so the area can be found as

$$A = \frac{1}{2}bh = \frac{1}{2}(1)(0.1) = 0.05$$

The area under the curve between 2 and 3 is a rectangle, so the area can be found as

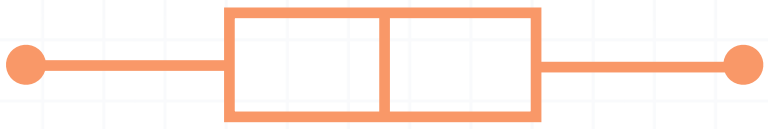
$$A = lw = (1)(0.1) = 0.1$$

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

$$1 - 0.1 - 0.05 = 0.85$$

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:

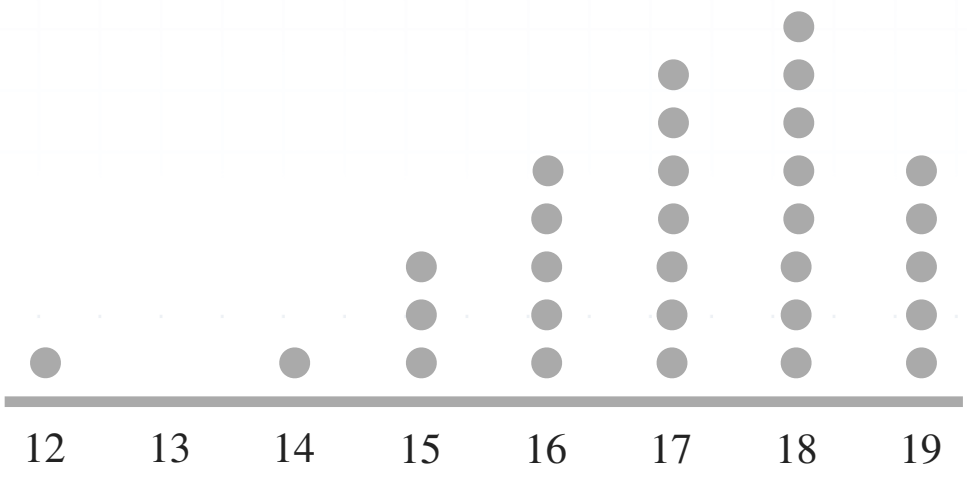
$$Q_3 + 1.5(\text{IQR})$$

$$25 + 1.5(4.5)$$

$$31.75$$

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.



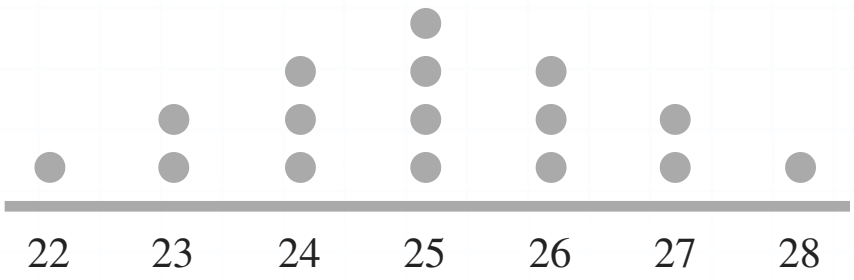
Solution:

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.

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

$$\sigma = 1.5811$$

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:

For this problem we need to find the z-score and then find the percentage from the z-table. We know that the mean is $\mu = 170$ and that the standard deviation is $\sigma = 8$. And the value we're interested in is $x = 154$. So the z-score is

$$z = \frac{154 - 170}{8} = \frac{16}{8} = 2$$

To find the percentage, look up the z-score in the z-table. The amount in the table is 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

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

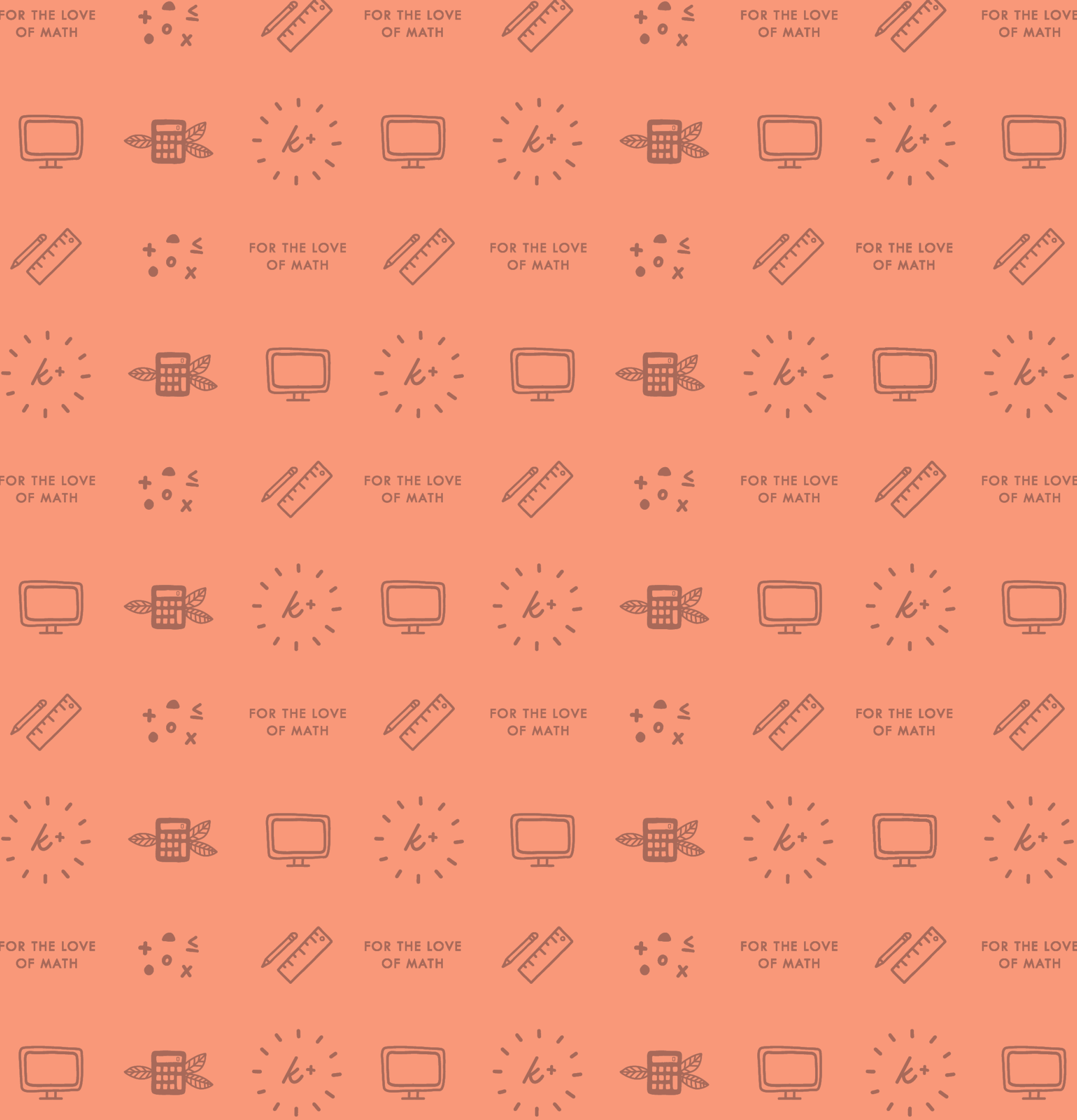
which gives $.8944$ in the z-table. The z-score for 140 is

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

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.



WWW.KRISTAKINGMATH.COM