



Probability & Statistics Quizzes

krista king
MATH

Topic: One-way tables**Question:** Which variable would be measured at the interval level?**Answer choices:**

- A Age
- B Race
- C Temperature in Fahrenheit or Celsius
- D Level of satisfaction



Solution: C

Temperature in Fahrenheit and Celsius is measured at the interval level because the interval change between each degree is equal, but the temperature scale doesn't have an absolute zero.

Age is measured at the ratio level because it also has a true zero.

Race is nominal data because it's a categorical variable and can take on only non-numerical values. Each value would represent a different category of the variable.

Level of satisfaction is also represented by categories, but those categories have a defined order, which means this variable is measured at the ordinal level.

Topic: One-way tables**Question:** Which of the following is a categorical variable?**Answer choices:**

- A The screen size of a cell phone, in inches
- B A customer's age, by age group (toddler, child, preteen, teen, adult)
- C The cost of an item, in dollars
- D The amount of electrical current in a wire



Solution: B

A categorical variable is a variable that uses words for variable names. They're non-numerical variables, which means they're not used for measurement or calculations.

A customer's age, by age range, would be considered categorical because we're calling them a toddler, child, preteen, and adult, even though age could be considered a number.

A quantitative variable uses numerical measurements. Inches, dollars and electrical current are all numerical amounts.



Topic: One-way tables

Question: A customer is comparing the fuel economy of different half-ton pickup trucks before making a purchase. Which variables are categorical?

Vehicles	Engine	HP / lb-ft	City MPG	Hwy. MPG	Comb. MPG
Chevy Silverado	4.3L V-6	285 / 305	18	24	20
Chevy Silverado	5.3L V-8	355 / 383	16	23	19
Ford F-150	3.7L V-6	302 / 278	17	23	19
Ford F-150 EB	3.5L V-6 TT	365 / 420	16	22	18
Ram 1500 HFE	3.6L V-6	305 / 269	18	25	21
Ram 1500 ED	3.0L V-6 TD	240 / 420	18	28	23
Toyota Tundra	4.0L V-6	270 / 278	16	20	17
Toyota Tundra	4.6L V-8	310 / 327	15	19	16

Answer choices:

- A The engine type
- B Horse power and pounds per foot
- C All of the MPG variables
- D All of the answer choices are quantitative

Solution: A

Categorical data divides things into categories. Even though engine type is reported as a number, these numbers are types of engines, so this variable is categorical.

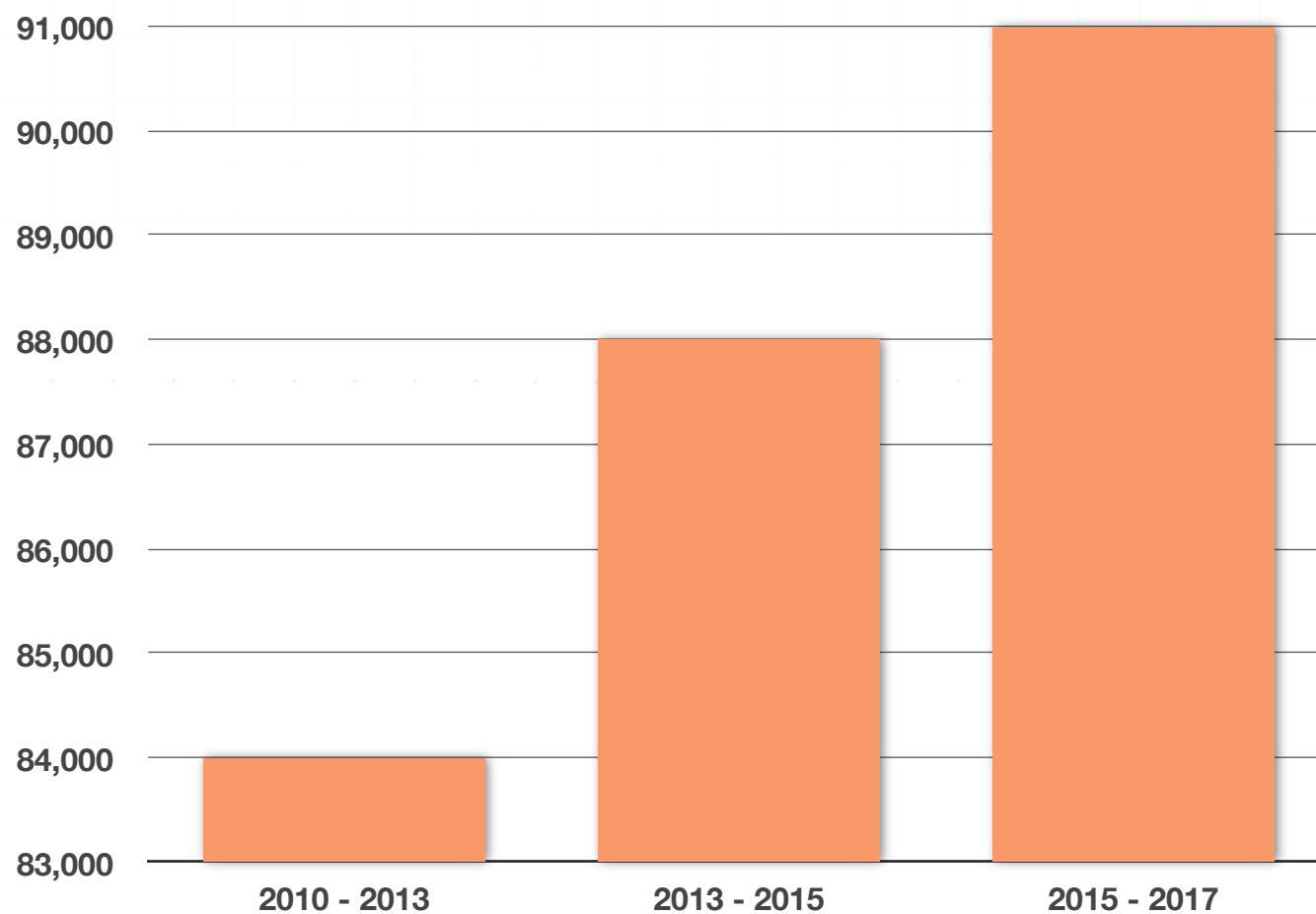
The other variables are measurements, so they're quantitative variables.

HP / lb-ft means Horsepower and pounds per foot, which is a measurement of energy that can be converted to watts. MPG means miles per gallon and is also a measurement.



Topic: Bar graphs and pie charts**Question:** Which statements can we know are true about the bar chart for total sales?

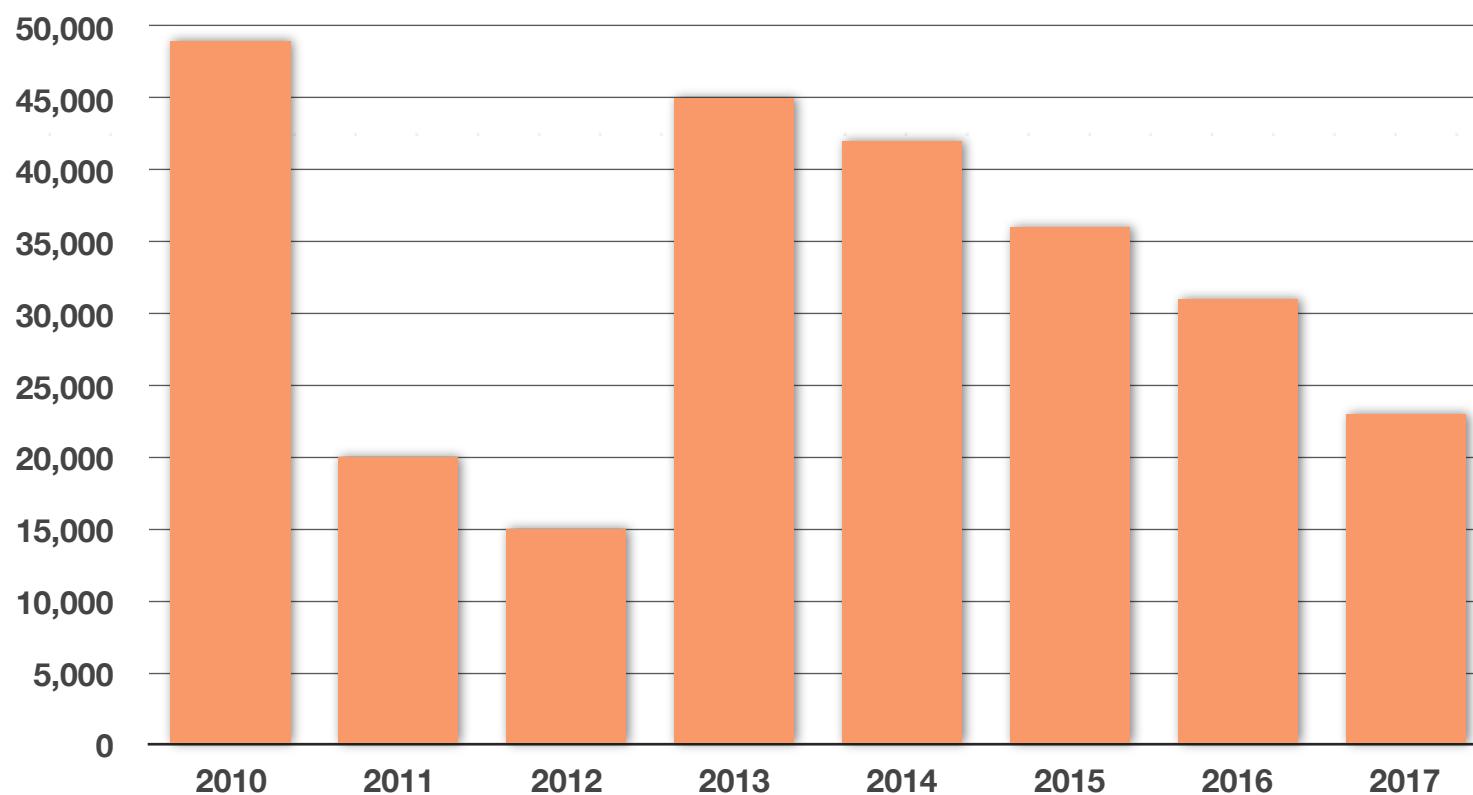
- I. Each year from 2010 to 2017, total sales have increased.
- II. The data on the horizontal axis should have been organized from largest to smallest.
- III. The labels on the horizontal axis have resulted in a misleading bar chart.

**Answer choices:**

- A I only
- B II only
- C III only
- D I and III only

Solution: C

The different intervals for the years on the horizontal axis make it impossible to get an inference from the data. Some of the years even overlap. Say the company broke down their total sales by year instead, and the bar chart actually looked like this:

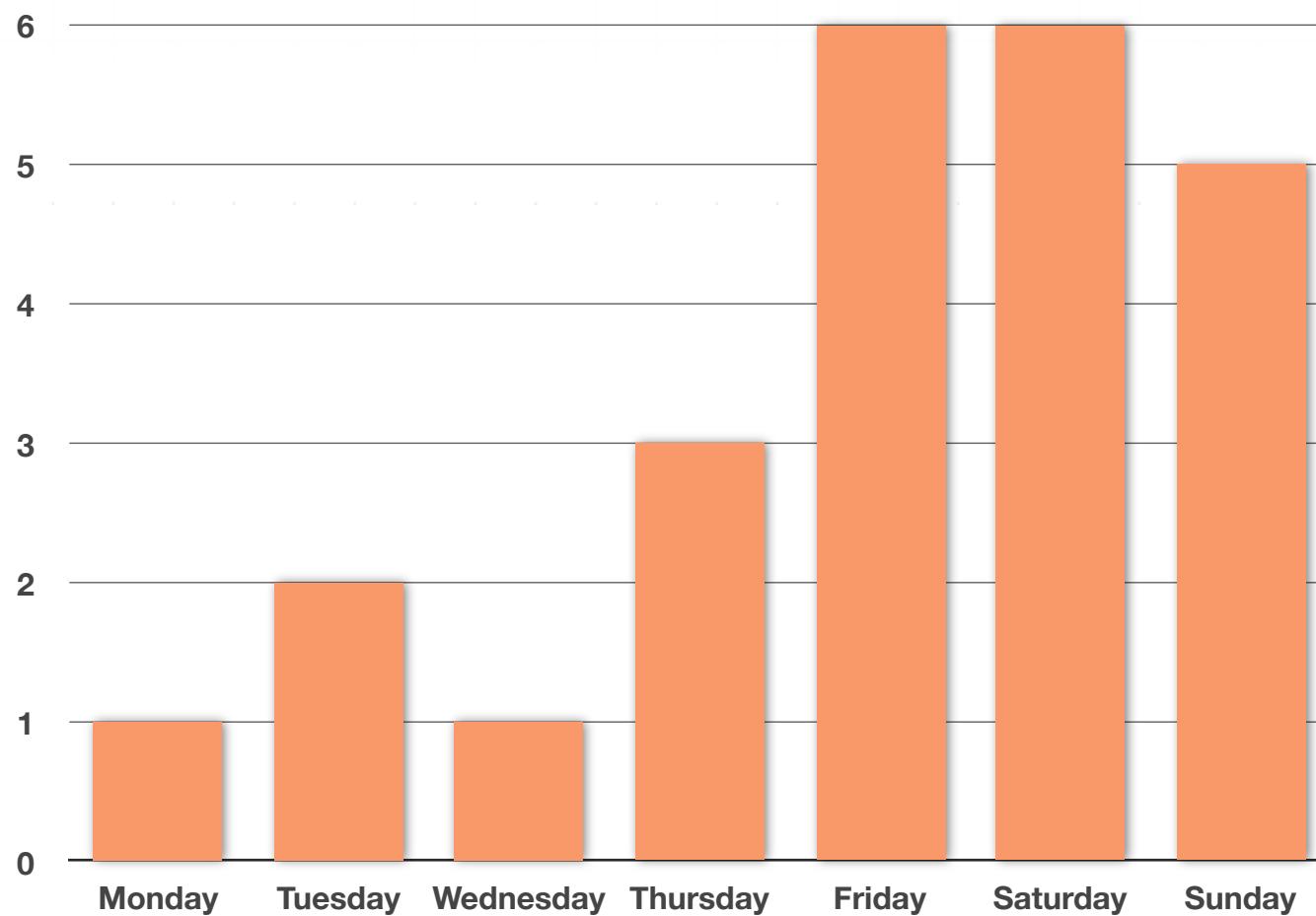


Grouping this chart together could result in the misleading bar chart you saw first, and lead consumers to think total sales were increasing, as opposed to decreasing year after year.



Topic: Bar graphs and pie charts**Question:** Which day of the week was graphed incorrectly?

Day	Rainfall (in cm)
Monday	1
Tuesday	3
Wednesday	1
Thursday	3
Friday	6
Saturday	6
Sunday	5



Answer choices:

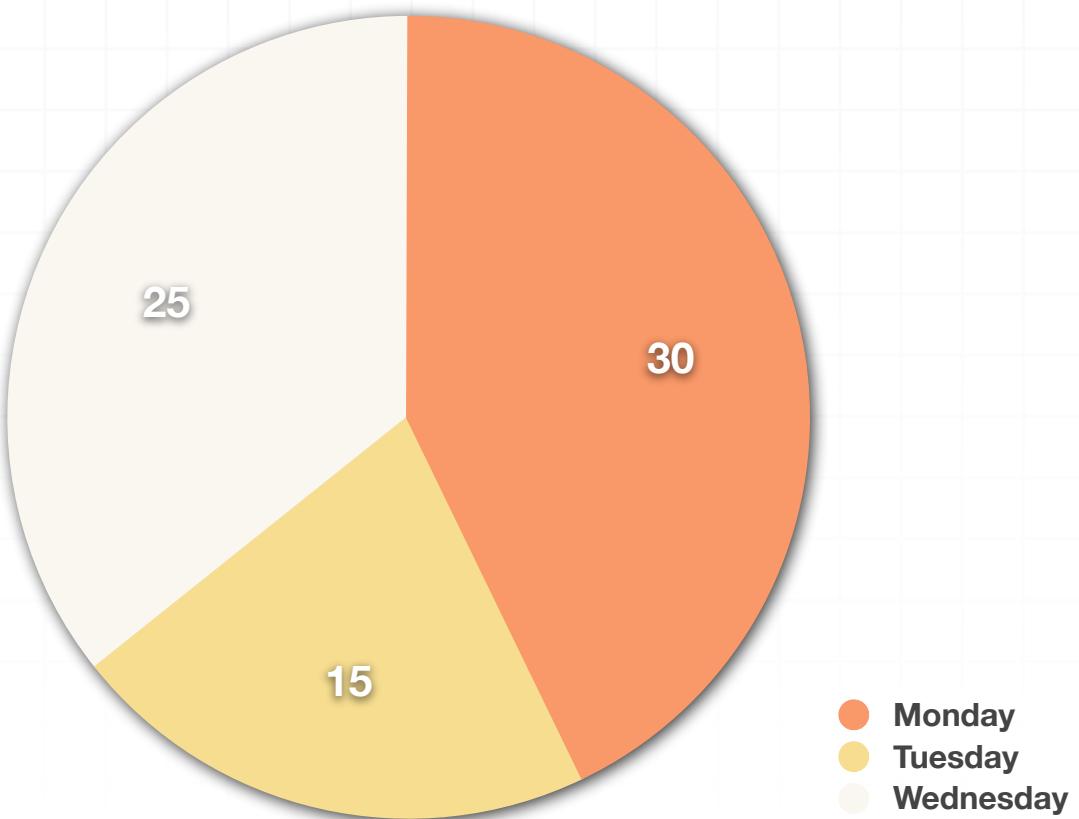
- A Monday
- B Tuesday
- C Wednesday
- D Thursday

Solution: B

The table says that it rained 3 cm on Tuesday. On the bar graph it says that it rained 2 cm on Tuesday. The data should be the same between the table and the bar graph, so Tuesday was graphed incorrectly.

Topic: Bar graphs and pie charts

Question: In the pie chart for ice cream cone sales by day, to the nearest tenth of a percent, what percentage of total sales were made on Monday?

**Answer choices:**

- A 21.4 %
- B 30.0 %
- C 35.7 %
- D 42.9 %

Solution: D

To create a pie chart, the data is divided into percentages to break up the circle. The total ice cream sales for Monday, Tuesday, and Wednesday are

$$30 + 15 + 25 = 70$$

and there were 30 ice cream cones sold on Monday. The percentage of ice cream cones sold on Monday is

$$\frac{30}{70} \cdot 100 = 42.9\%$$



Topic: Line graphs and ogives

Question: Which graph would be appropriate for changes in data over time?

Answer choices:

- A Ogive
- B Line graph
- C Circle graph
- D Bar graph



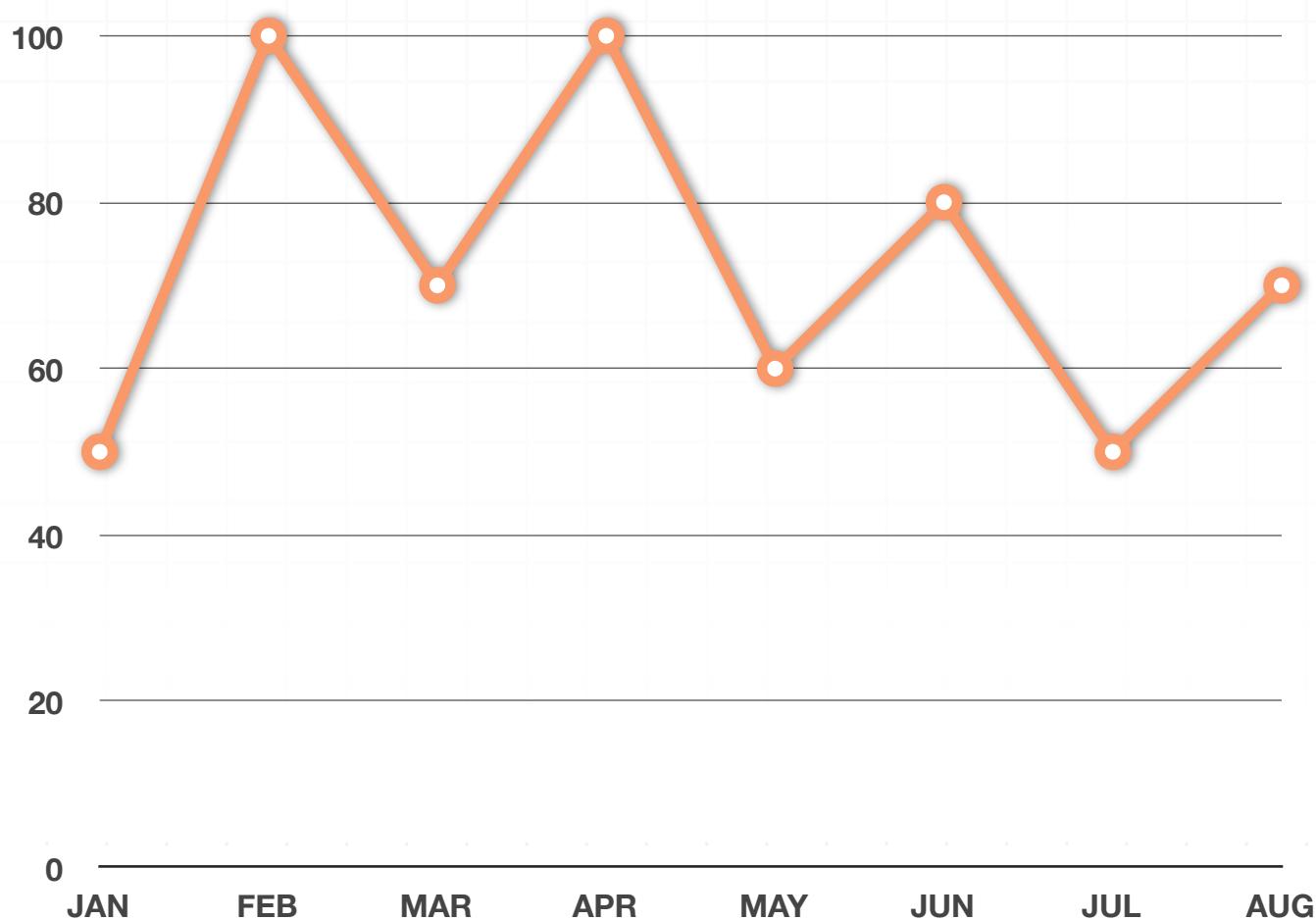
Solution: B

A line graph is the best type of graph to use to show changes over time because you can quickly read the graph to see the fluctuations.



Topic: Line graphs and ogives

Question: Stephanie made a line graph of the money she spent eating out each month from January to August. How would changing this graph to an ogive change the information?

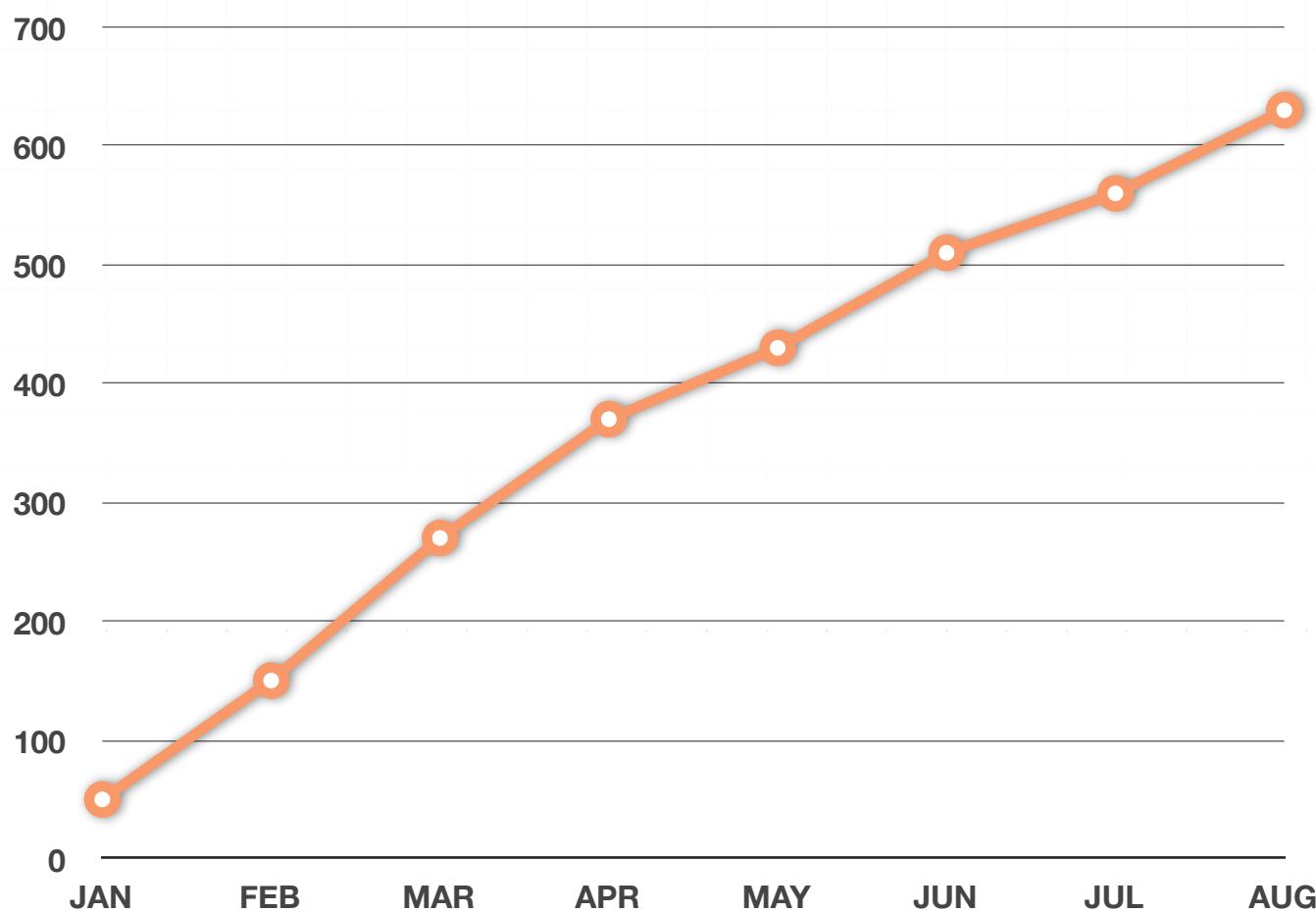
**Answer choices:**

- A The graph as an ogive would not give any relevant information.
- B The graph as an ogive would not change any of the information.
- C The graph as an ogive would allow you to read the amount Stephanie has spent each month.

- D The graph as an ogive would let you read the accumulated spending for Stephanie from month to month.

Solution: D

An ogive adds each month to the accumulated value and would let you read Stephanie's accumulated spending.



Topic: Line graphs and ogives**Question:** Is a line graph appropriate for analyzing rain fall per month?

Why or why not?

Answer choices:

- A Yes, because a line graph helps to quickly read the changes in monthly rainfall and make connections from month to month.
- B Yes, because you could quickly read the amount of rainfall that fell each month out of the total rainfall.
- C No, because you want to show data in equal intervals and you should use a histogram instead.
- D No, because you want to show data that's in multiples of months and you should use a pictograph.



Solution: A

A line graph is a good choice for analyzing rain fall per month because line graphs are used to show changes over time, or a connection between individuals.

Answer choice B is incorrect because comparing parts to a whole is accomplished in a circle graph rather than a line graph.



Topic: Two-way tables

Question: Which is an example of data that can be organized using a two-way table?

Answer choices:

- A The price of a Honda at a used car dealership is based on its mileage. A consumer report creates a table of the recent mileage paired with prices of each Honda sold so a customer can get an idea of what to pay for their new car.
- B Stacy's class is studying the temperature of their classroom in various parts of the room.
- C A flower company is creating a chart of the color of each flower after it blooms and using the information for its advertisements of fresh flower shipments to its customers.
- D Emma is taking a class survey. She asks each student who they would vote for in the upcoming student council election (Mark, Vanessa, or Anthony) and she plans on organizing the results by their grade (Junior or Senior).



Solution: D

Answer choice D is an example of data that can be organized using a two-way table because Emma is collecting and comparing information for two categorical variables.

She wants to compare who each student votes for as well as their grade. She could organize the information in a two-way table like this one:

Who will you vote for?	Mark	Vanessa	Anthony
Junior			
Senior			

Topic: Two-way tables

Question: A car magazine wants to create an article about how a driver's speed on the highway in miles per hour is correlated with their fuel consumption in gallons per mile. Which statement is true about the study?

Answer choices:

- A This is an example of two pairs of numerical data that could not be organized in a two-way table.
- B This is an example of categorical one-way data because they are studying individual drivers and multiple variables about each car.
- C This is an example data that would work in a two-way table because the drivers are independent of the speed and miles per gallon.
- D This is an example of data that would be best displayed in a pie chart.



Solution: A

There isn't a way to organize the driver's speed on the highway and their fuel consumption in a two-way table, because a two-way table organizes counts of data about two categorical variables.

Topic: Two-way tables

Question: Which question can be answered without using the two-way data table?

Favorite hobby for men:

Reading	Sports	Art	Total
14	28	15	57

Favorite hobby for men and women:

	Reading	Sports	Art	Total
Men	14	28	15	57
Women	8	13	7	28
Total	22	41	22	85

Answer choices:

- A What percentage of people prefer reading?
- B How many men preferred art?
- C How many women were in the study?
- D What percentage of women liked sports the best?

Solution: B

The questions that ask about two pieces of information, like gender and hobby, require the use of the two-way table. If the question only asks about people in general then you could use the one-way table.

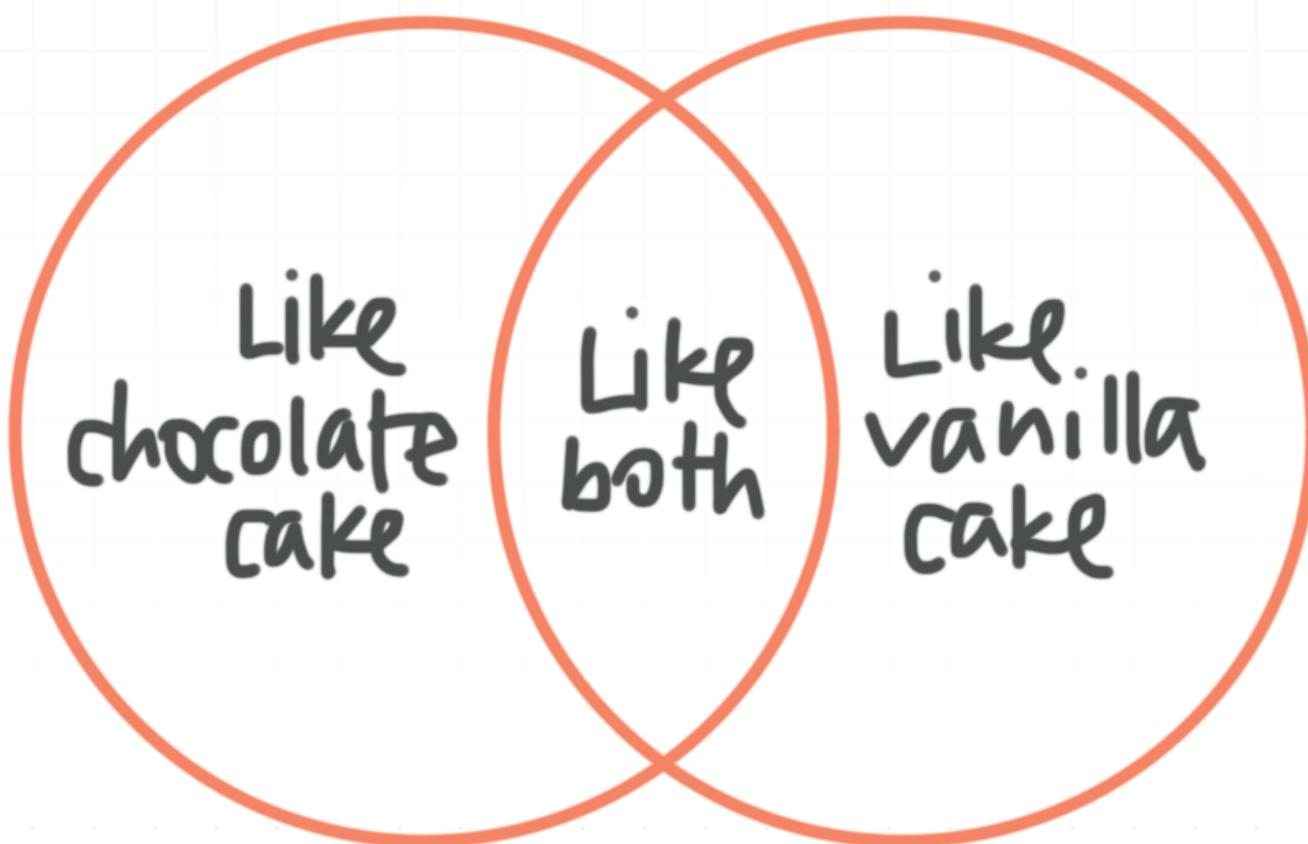


Topic: Venn diagrams**Question:** Which data set could be displayed in a Venn diagram?**Answer choices:**

- A The number of students who like chocolate cake, vanilla cake or both.
- B The list of student test scores on the last exam.
- C The data points from a scatterplot.
- D The average height of cat breeds verses the average heights of dog breeds.

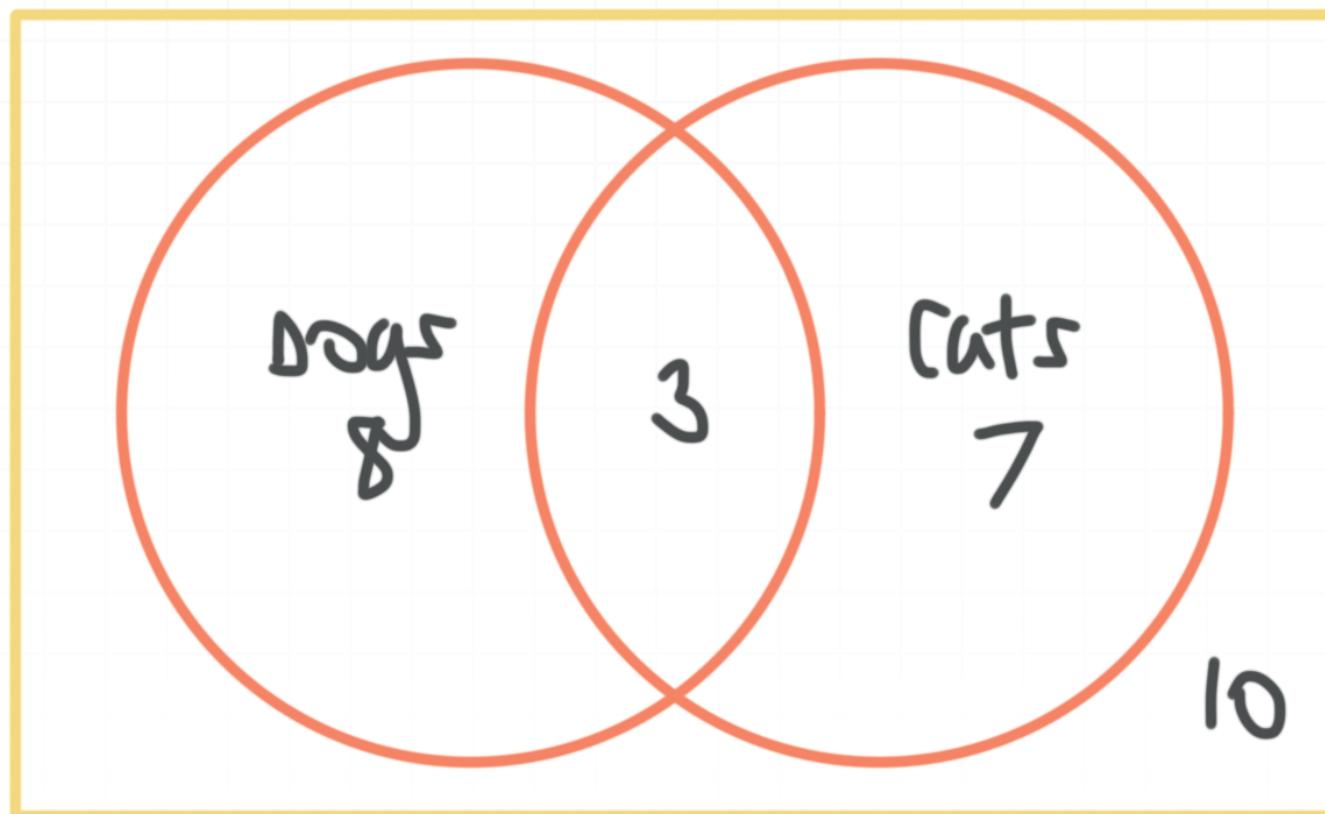
Solution: A

Venn diagrams are made with idea of displaying the difference or similarities between data sets, you are looking for the kind of information that could be easily displayed in a two-way table. You can't make a Venn diagram from a list of numbers or from points on a scatterplot because you need to have some categories that you can compare to one another.



Topic: Venn diagrams

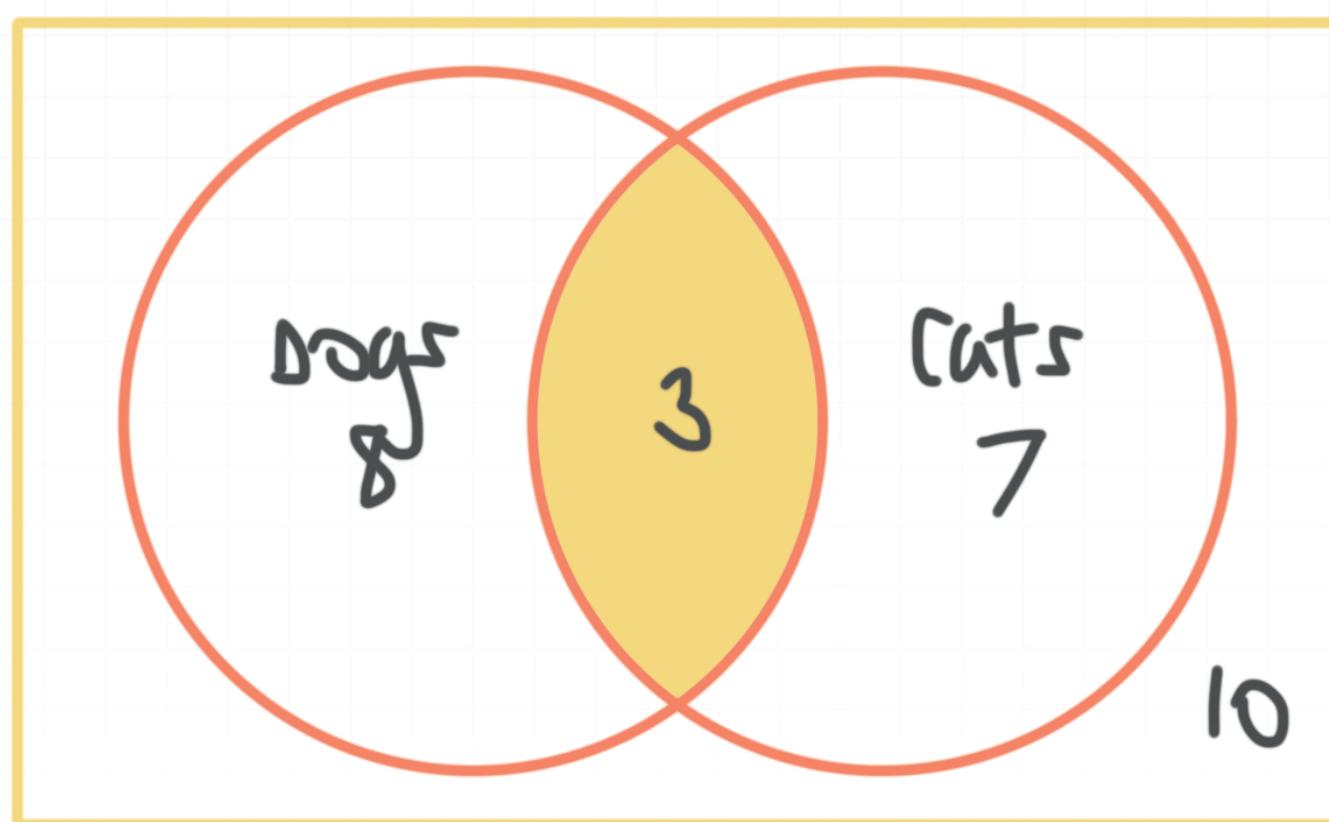
Question: The Venn diagram shows all the pets that the 3rd graders own. How many 3rd graders own both a cat and a dog?

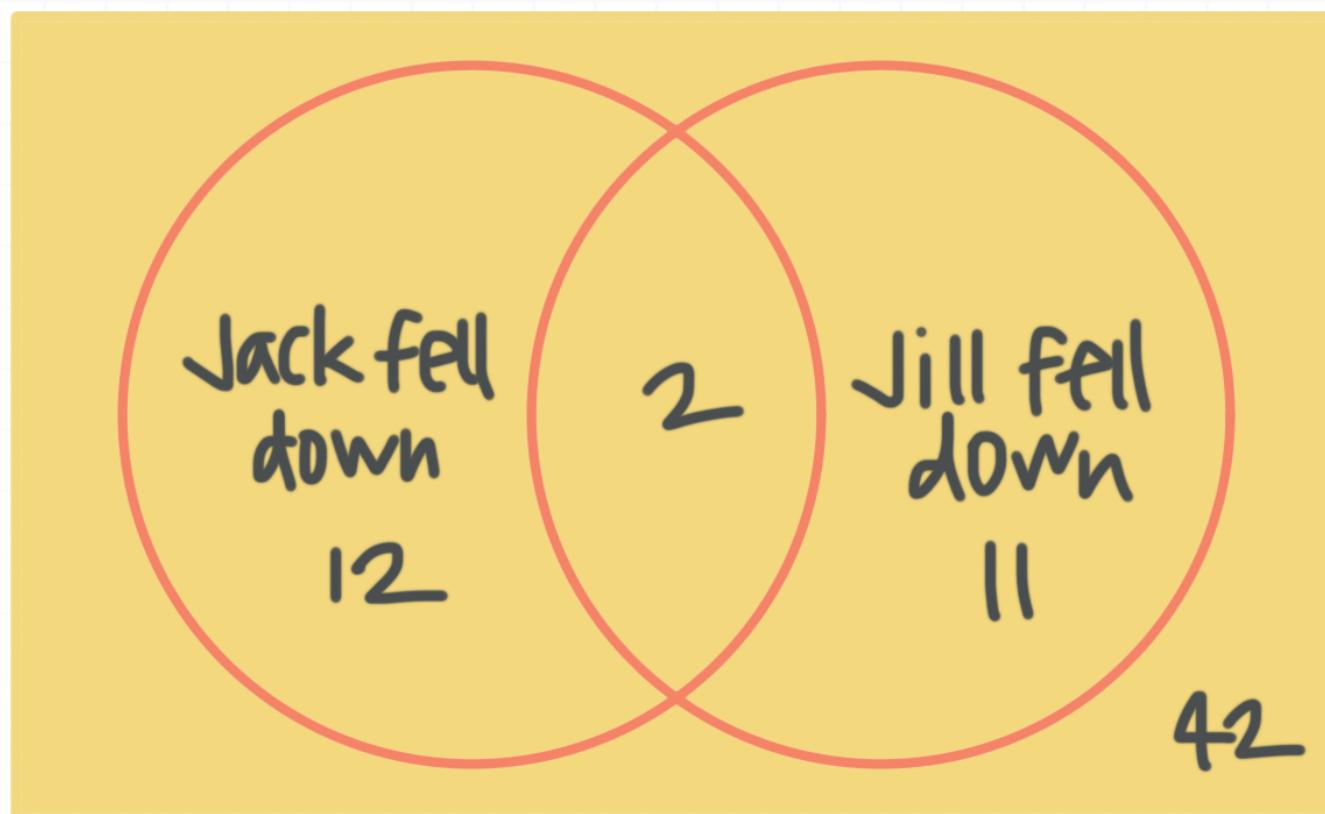
**Answer choices:**

- A 8
- B 3
- C 15
- D 10

Solution: B

The section in the middle of the Venn diagram is the one that shows the number of 3rd graders who own both a dog and a cat. So there are 3 students who own both a dog and a cat.

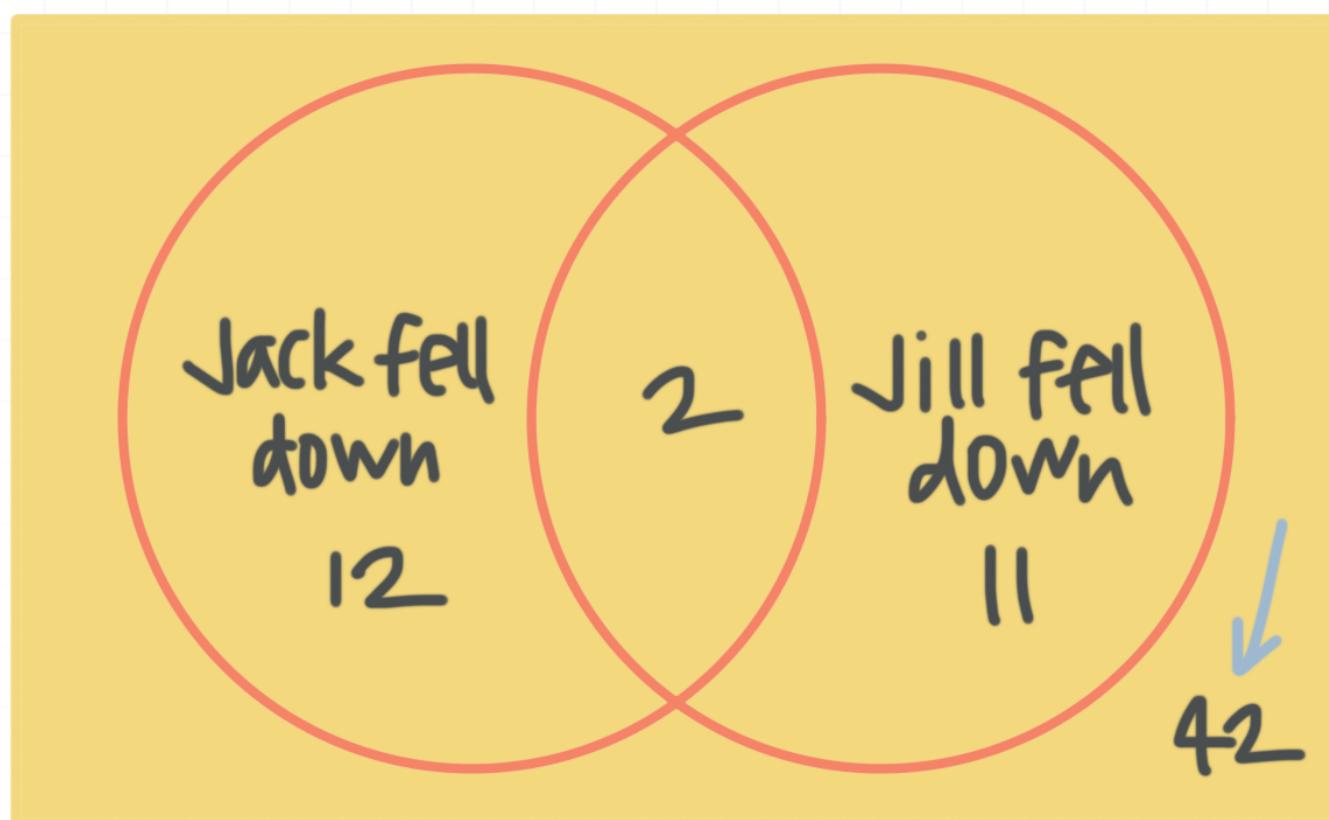


Topic: Venn diagrams**Question:** How many times did Jack and Jill not fall down?**Answer choices:**

- A 12
- B 2
- C 11
- D 42

Solution: D

The number in the Venn diagram that's outside both circles is the number that shows how many times neither Jack nor Jill fell down. Therefore, there were 42 times where neither of them fell.



Topic: Frequency tables and dot plots

Question: The following data shows the number of points each player earned in a game. What is the missing frequency in the table?

75, 75, 75, 81, 81, 81, 82, 84, 84, 84, 84, 84, 97, 97

Points	Frequency
75	3
81	
82	1
84	5
97	2

Answer choices:

- A 1
- B 2
- C 3
- D 4

Solution: C

A frequency table is a count of the number of times a data point occurs. Count the number of times that 81 occurs in the data set.

75, 75, 75, 81, 81, 81, 82, 84, 84, 84, 84, 84, 97, 97

The number 81 occurs 3 times in the data set, so that's the missing frequency in the table.

Points	Frequency
75	3
81	3
82	1
84	5
97	2

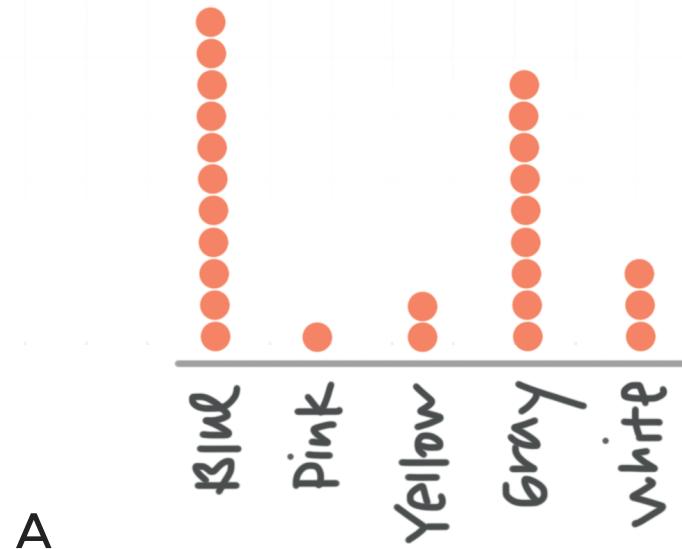


Topic: Frequency tables and dot plots

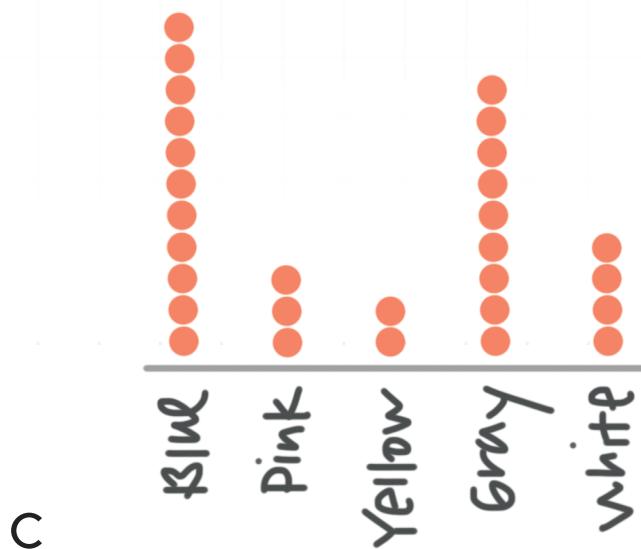
Question: The frequency table shows the number of times a bird chose to eat a different colored piece of bird seed. Choose the dot plot that matches the frequency table.

Color	Blue	Pink	Yellow	Gray	White
Frequency	11	1	2	9	3

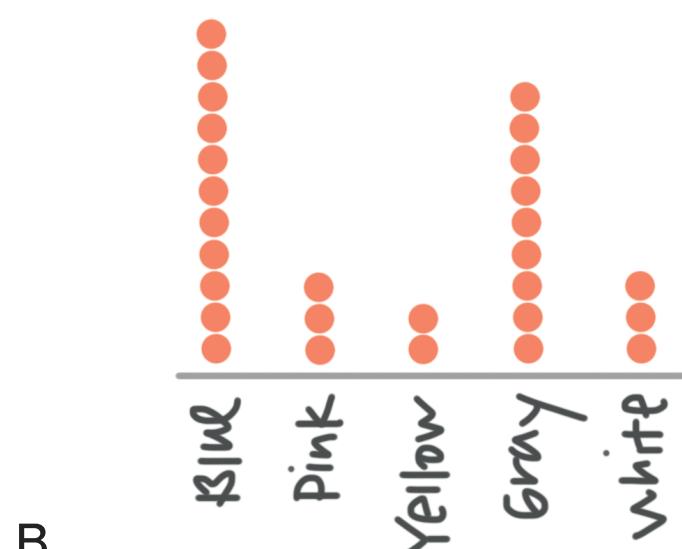
Answer choices:



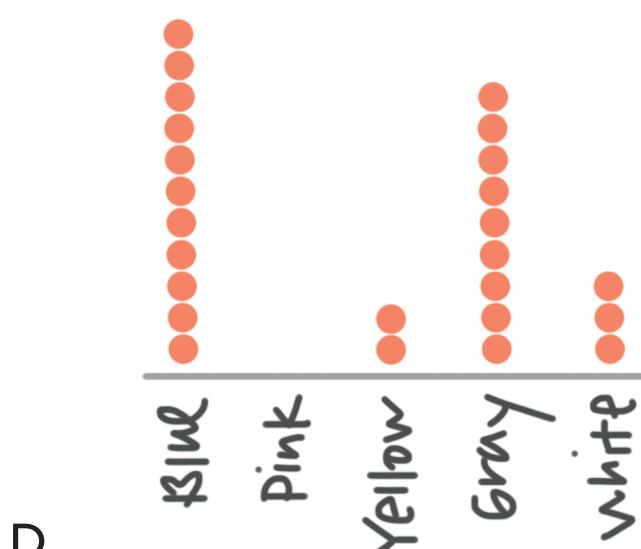
A



C



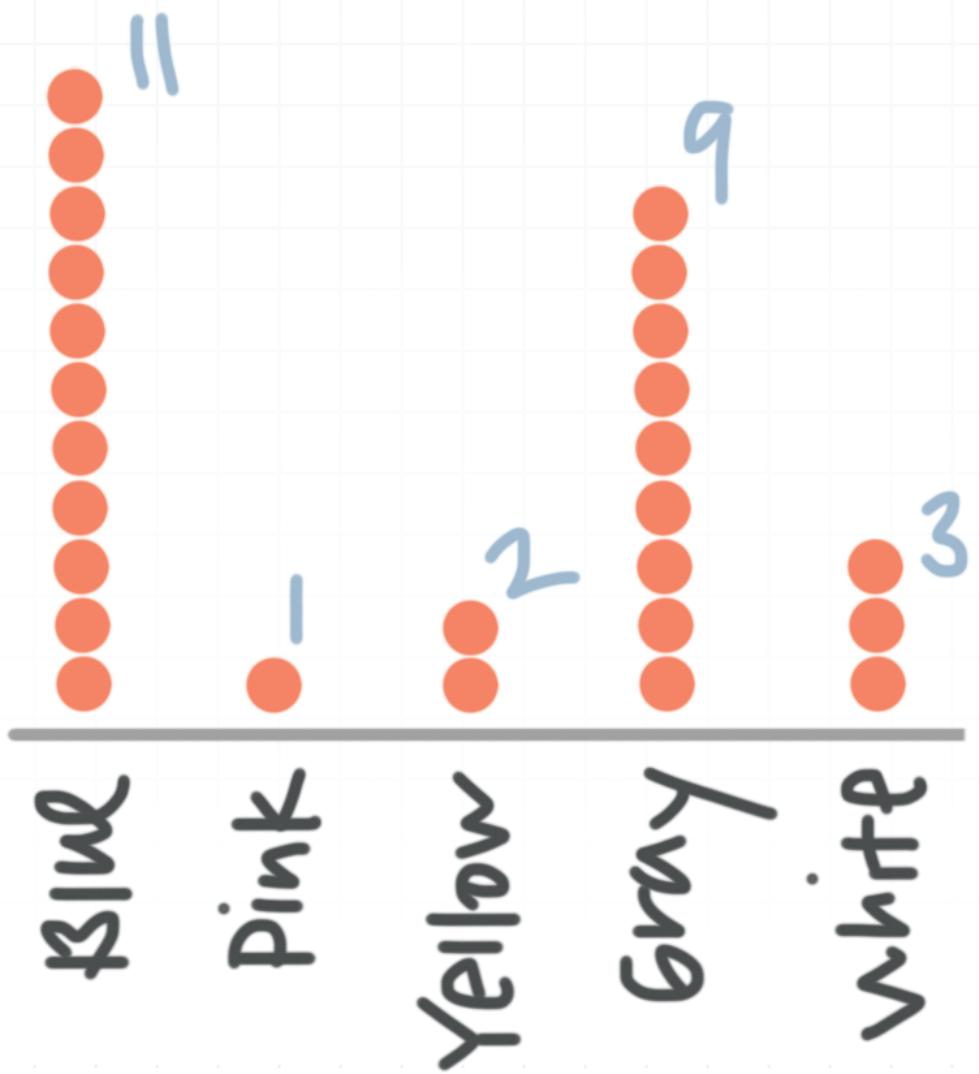
B



D

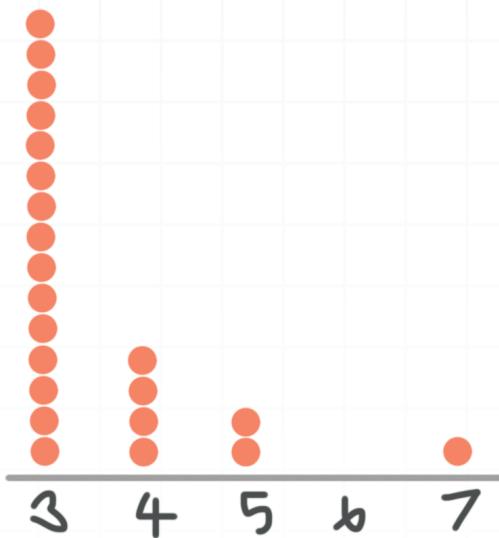
Solution: A

Choose the dot plot that has the same number of dots for each color as the frequencies in the table.



Topic: Frequency tables and dot plots

Question: The dot plot shows the number of leaves on each stem found in a small sample of clover. Which frequency table matches the dot plot?

**Answer choices:****A**

Number of leaves	Frequency
3	15
4	4
5	2
6	0
7	1

C

Number of leaves	Frequency
3	1
4	0
5	2
6	4
7	15

B

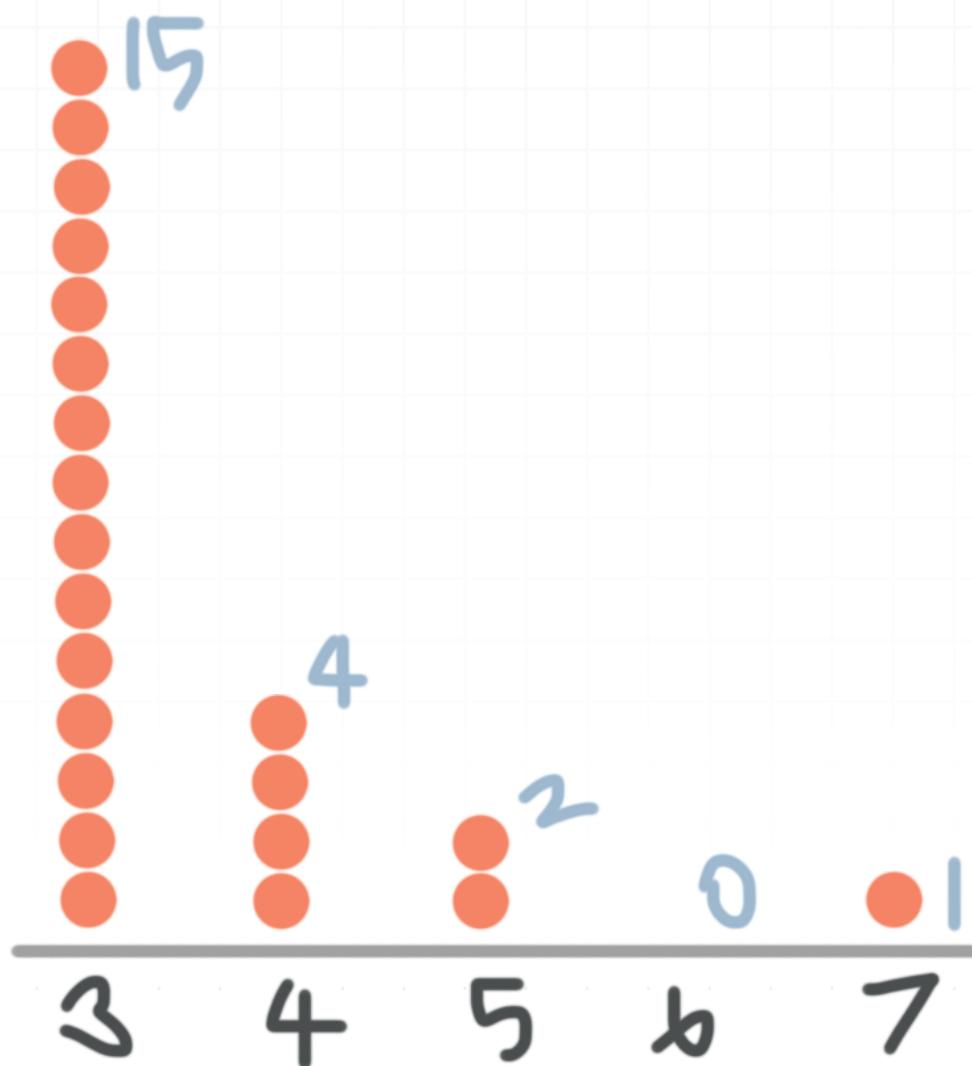
Number of leaves	Frequency
15	3
4	4
2	5
0	6
1	7

D

Number of leaves	Frequency
1	3
0	4
2	5
4	6
15	7

Solution: A

You want to count the number of points for each data point in the dot plot to find the frequencies in the table. Make sure the number of leaves goes on the left and the frequency on the right.



So the frequency table is

Number of leaves	Frequency
3	15
4	4
5	2
6	0
7	1

Topic: Relative frequency tables

Question: Every student at a certain high school needs to choose exactly one fine arts elective. The following table shows the enrollment of electives for all students. About what percentage of Sophomores were enrolled in art?

	Art	Architectu	Music	Total
Freshman	40	25	55	120
Sophomo	52	12	71	135
Junior	56	45	54	155
Senior	30	60	20	110
Total	178	142	200	520

Answer choices:

- A 29 %
- B 39 %
- C 10 %
- D 26 %

Solution: B

To find the percentage of sophomores who were enrolled in art, we need to find the row for the Sophomores,

	Art	Architectu	Music	Total
Sophomo	52	12	71	135

and turn it into a row-relative frequency.

	Art	Architectu	Music	Total
Sophomo	$52/135=38$	$12/135=8.9$	$71/135=52$	$135/135=1$

We can see from the row-relative frequency that about 39 % of the sophomores were enrolled in art.

	Art	Architectu	Music	Total
Sophomo	38.5%	8.9%	52.6%	100%

Topic: Relative frequency tables

Question: Every student at a certain high school needs to choose exactly one fine arts elective. The table shows the enrollment of electives for all students. Which type of relative frequency table would answer the question: “Of the students enrolled in architecture, what percentage of them are seniors?”

	Art	Architectu	Music	Total
Freshman	40	25	55	120
Sophomo	52	12	71	135
Junior	56	45	54	155
Senior	30	60	20	110
Total	178	142	200	520

Answer choices:

- A Column-relative frequency table
- B Row-relative frequency table
- C Total-relative frequency table
- D All of the above

Solution: A

Picking out the column for students enrolled in Architecture,

	Architecture
Freshman	25
Sophomore	12
Junior	45
Senior	60
Total	142

and then turning it into a column-relative frequency

	Architecture
Freshman	$25/142=18\%$
Sophomore	$12/142=8\%$
Junior	$45/142=32\%$
Senior	$60/142=42\%$
Total	$142/142=100\%$

would let us answer the question. We can see that, of the students enrolled in Architecture, 42 % are seniors.

Topic: Relative frequency tables

Question: Emily counted the shape and type of blocks that her little sister owns and organized the information into a table. Which question could she answer by looking at a row-relative frequency that she makes from this table?

		Block Shape		
Block Color	Red	Cube	Rectangular Prism	Total
		5	9	14
	Total	9	19	28

Answer choices:

- A What percentage of the blocks are red?
- B What percentage of the blocks are red cubes?
- C What percentage of the blocks are blue?
- D What percentage of the blue blocks are rectangular prisms?

Solution: D

To find the percentage of the blue blocks that are rectangular prisms, we want to find the number of rectangular prisms that are blue and the total number of blue blocks. This information is found in the row frequency,

	Cube	Rectangular Prism	Total
Blue	4	10	14

and the percentages would be found in the row-relative frequency.

	Cube	Rectangular Prism	Total
Blue	$4/14=29\%$	$10/14=71\%$	$14/14=100\%$

We can see the percentage of blue blocks that are rectangular prisms in this row-relative frequency table.

Topic: Joint distributions

Question: The table shows Addie's poll of the children in her neighborhood and their cartoon watching habits. Which description comes from the joint distribution section of the table?

	< 2 hrs cartoons	> 2 hrs cartoons	Total
Watched on the T.V.	37%	19%	56%
Watched on a different device	7%	37%	44%
Total	44%	56%	100%

Answer choices:

- A The percentage of children who watched less than 2 hours of cartoons
- B The percentage of children who watched on a different device
- C The percentage of children who watched less than 2 hours of cartoons on the T.V.
- D The percentage of children who watched 2 hours or more of cartoons

Solution: C

The joint distribution or the joint probability distribution is the probability that a pair of events can happen. All of the possible pairs of events happen in the body of the table. For example, the percentage of children who watched less than 2 hours of cartoons on the T.V. is 37%:

	< 2 hrs cartoons	> 2 hrs cartoons	Total
Watched on the T.V.	37%	19%	56%
Watched on a different device	7%	37%	44%
Total	44%	56%	100%

Answer choice C is the only choice that describes a pair of events.

Topic: Joint distributions

Question: The table shows favorite activities of college students. Which description comes from the marginal distribution section of the table?

	Movie	Bowling	Pizza Party	Total
Male	8%	15%	21%	44%
Female	13%	19%	24%	56%
Total	21%	34%	45%	100%

Answer choices:

- A The percentage of male student who prefer movies.
- B The percentage of students who preferred a pizza party.
- C The percentage of female students in the sample.
- D Both B and C



Solution: D

The marginal distribution comes from the total column or total row.

Answer choices B and C are both found in the totals sections of the table.

	Movie	Bowling	Pizza Party	Total
Male	8%	15%	21%	44%
Female	13%	19%	24%	56%
Total	21%	34%	45%	100%



Topic: Joint distributions

Question: Every student at a certain high school needs to choose exactly one fine arts elective. The following frequency table shows the enrollment of electives for all students. Jessica is answering a list of questions where she needs to find distributions based off of the table below. Which question would require a conditional distribution to answer?

		Extracurricular Activities			
		Art	Architect	Music	Total
Grade	Freshma	40	25	55	120
	Sophom	52	12	71	135
	Junior	56	45	54	155
	Senior	30	60	20	110
	Total	178	142	200	520

Answer choices:

- A The percentage of students who are freshman and take art.
- B The percentage of seniors enrolled in extracurricular activities.
- C The percentage of architecture students.
- D The percentage of juniors enrolled in music class of the total students enrolled in music.

Solution: D

Conditional distribution is calculated based on the total of the row or column. In this case we want to think of the problem as the probability of choosing a junior if we know the student is already enrolled in music.

		Extracurricular Activities			
		Art	Architect	Music	Total
Grade	Freshma	40	25	55	120
	Sophom	52	12	71	135
	Junior	56	45	54	155
	Senior	30	60	20	110
	Total	178	142	200	520

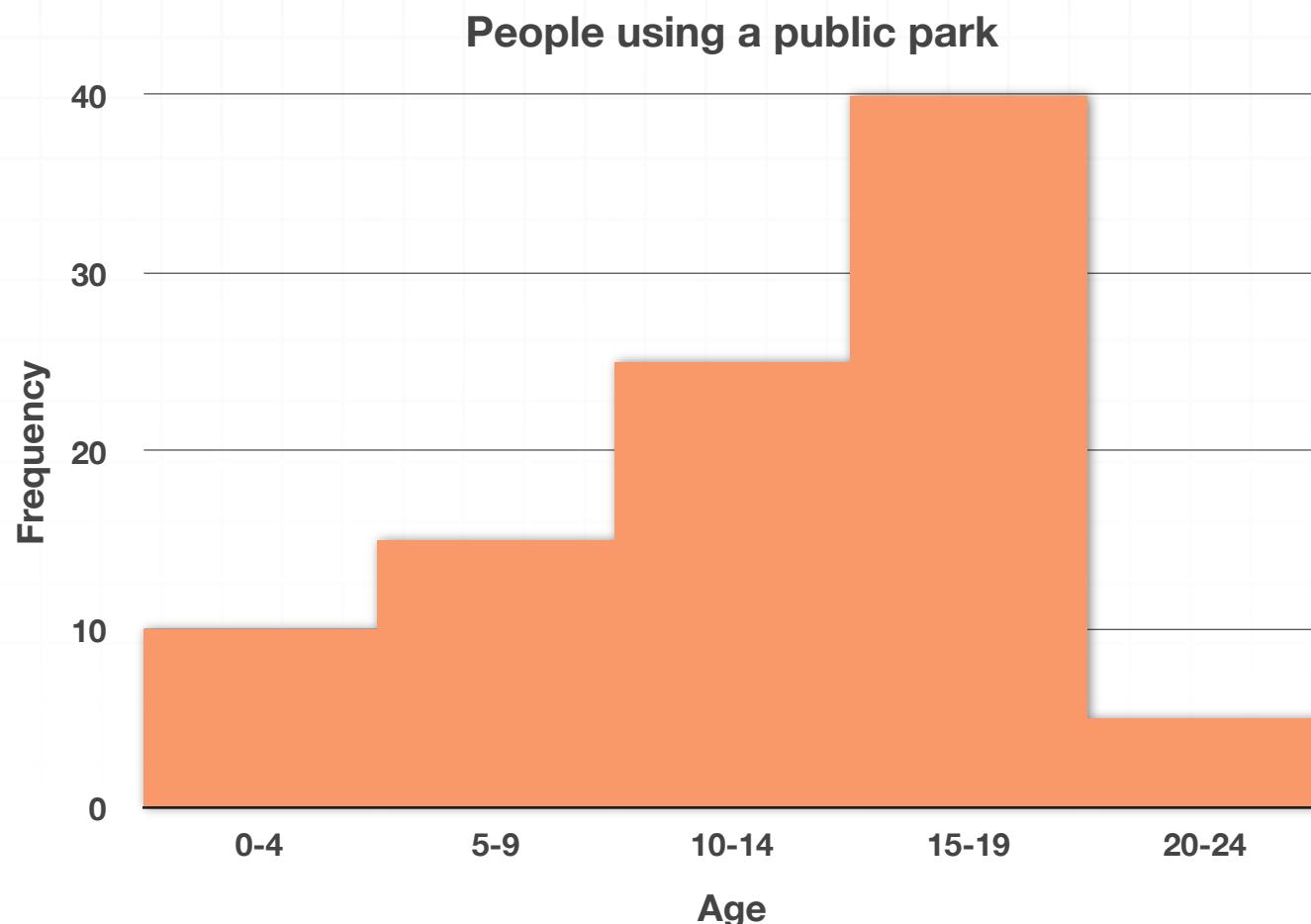
This would be calculated by taking

$$\frac{54}{200} = 27\%$$

So 27% of the students enrolled in music class are juniors. That's why it's a conditional probability.

Topic: Histograms and stem-and-leaf plots

Question: What is the length of the buckets that are used in the histogram?

**Answer choices:**

- A 4 years
- B 5 years
- C 10 years
- D 25 years

Solution: B

The age groups are put into buckets of 5-unit intervals.



Topic: Histograms and stem-and-leaf plots**Question:** Which data set would best be displayed in a histogram?**Answer choices:**

- A The percentage of people who like a certain brand of cola.
- B A kindergarten class's favorite colors.
- C The way rainfall changes each month.
- D Town population by age.



Solution: D

Town population by age would best be displayed in a histogram. It would be useful to group age ranges together to create a graph of the data, as opposed to graphing each age individually.



Topic: Histograms and stem-and-leaf plots

Question: A shopkeeper counted the number of candies in each basket and recorded the results in the stem-and-leaf plot. How many baskets have more than 35 candies?

1	3, 5
2	1, 4
3	5
6	2, 6

$$1 | 3 = 13$$

Answer choices:

- A 2
- B 3
- C 6
- D 62

Solution: A

We want to know how many baskets had more than 35 pieces of candy. If you look at the stem-and-leaf plot, you can see that two baskets had more than 35 pieces of candy.

1	3, 5
2	1, 4
3	5
6	2, 6

One basket had 62 pieces, and one basket had 66 pieces, so there were only two baskets that had more than 35 pieces.



Topic: Building histograms from data sets

Question: If we divide the data set into 6 classes, what will be the class width?

12, 7, 9, 5, 8, 17, 28, 31, 17, 25, 13, 14, 6, 2, 20, 39, 45, 16, 33, 28

Answer choices:

- A 6
- B 7
- C 8
- D 9



Solution: C

First, we need to put the data points in ascending order,

2, 5, 6, 7, 8, 9, 12, 13, 14, 16, 17, 17, 20, 25, 28, 28, 31, 33, 39, 45

so that we can calculate the range.

$$\text{Range} = 45 - 2 = 43$$

Now we divide the range by the number of classes in order to find the class width.

$$\frac{43}{6} \approx 7.2$$

Since we have to round up to the nearest integer (rounding down would mean that we wouldn't be able to span the entire data set), the class width will be 8.



Topic: Building histograms from data sets

Question: When constructing a histogram from the data set, which class interval will have the largest frequency if we use 6 classes?

12, 7, 9, 5, 8, 17, 28, 31, 17, 25, 13, 14, 6, 2, 20, 39, 45, 16, 33, 28

Answer choices:

- A 0 – 7
- B 8 – 15
- C 16 – 23
- D 32 – 39

Solution: B

Put the data in ascending order.

2, 5, 6, 7, 8, 9, 12, 13, 14, 16, 17, 17, 20, 25, 28, 28, 31, 33, 39, 45

Then the range is $45 - 2 = 43$. Divide the range by the number of classes to find the class width.

$$\frac{43}{6} \approx 7.2$$

We have to round up, so the class width will be 8. The smallest data value is 2, so we can start the first interval from 0, and the class intervals will be

Class interval	Frequency
0 - 7	4
8 - 15	5
16 - 23	4
24 - 31	4
32 - 39	2
40 - 47	1

Therefore, the class interval with the largest frequency will be 8 – 15.

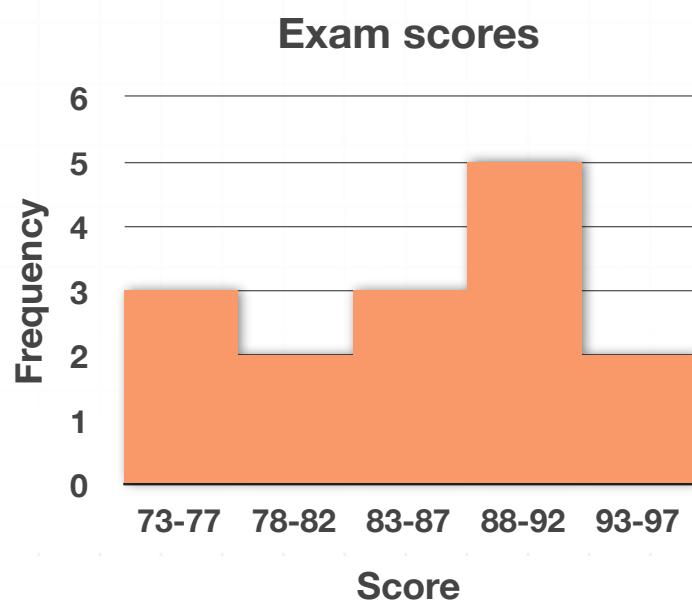


Topic: Building histograms from data sets

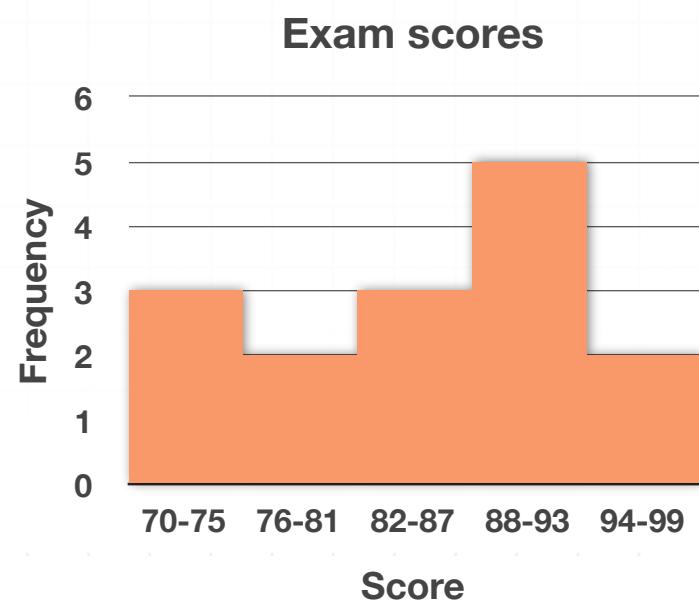
Question: The exam scores for 15 students are listed below. If we want to use 5 bins to organize the data, which chart is the correct histogram?

85, 91, 94, 74, 88, 98, 83, 73, 86, 89, 93, 80, 77, 79, 95

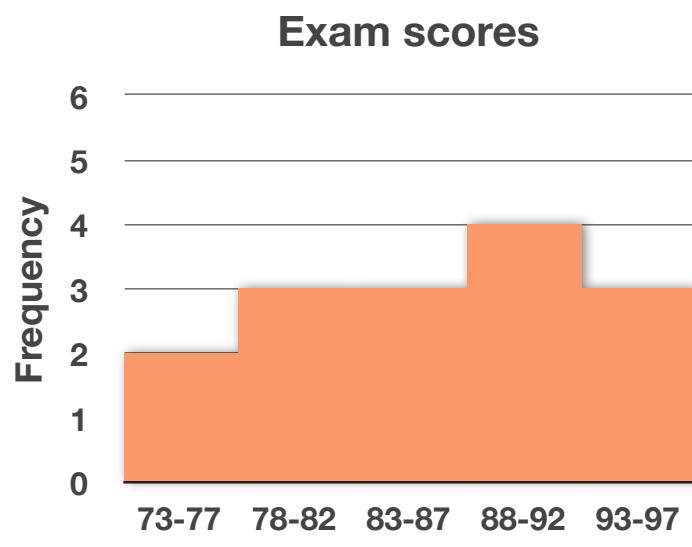
Answer choices:



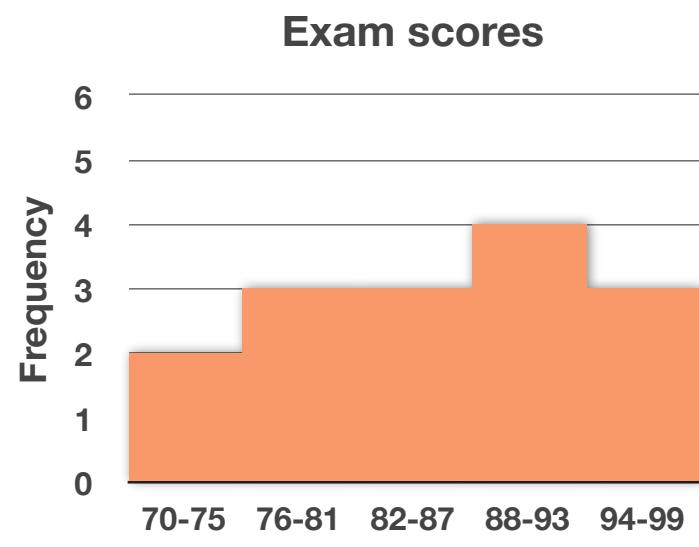
A



C



B



D

Solution: D

Put the data points in ascending order,

73, 74, 77, 79, 80, 83, 85, 86, 88, 89, 91, 93, 94, 95, 98

and calculate the range.

$$98 - 73 = 25$$

To find class width, divide the range by the number of classes.

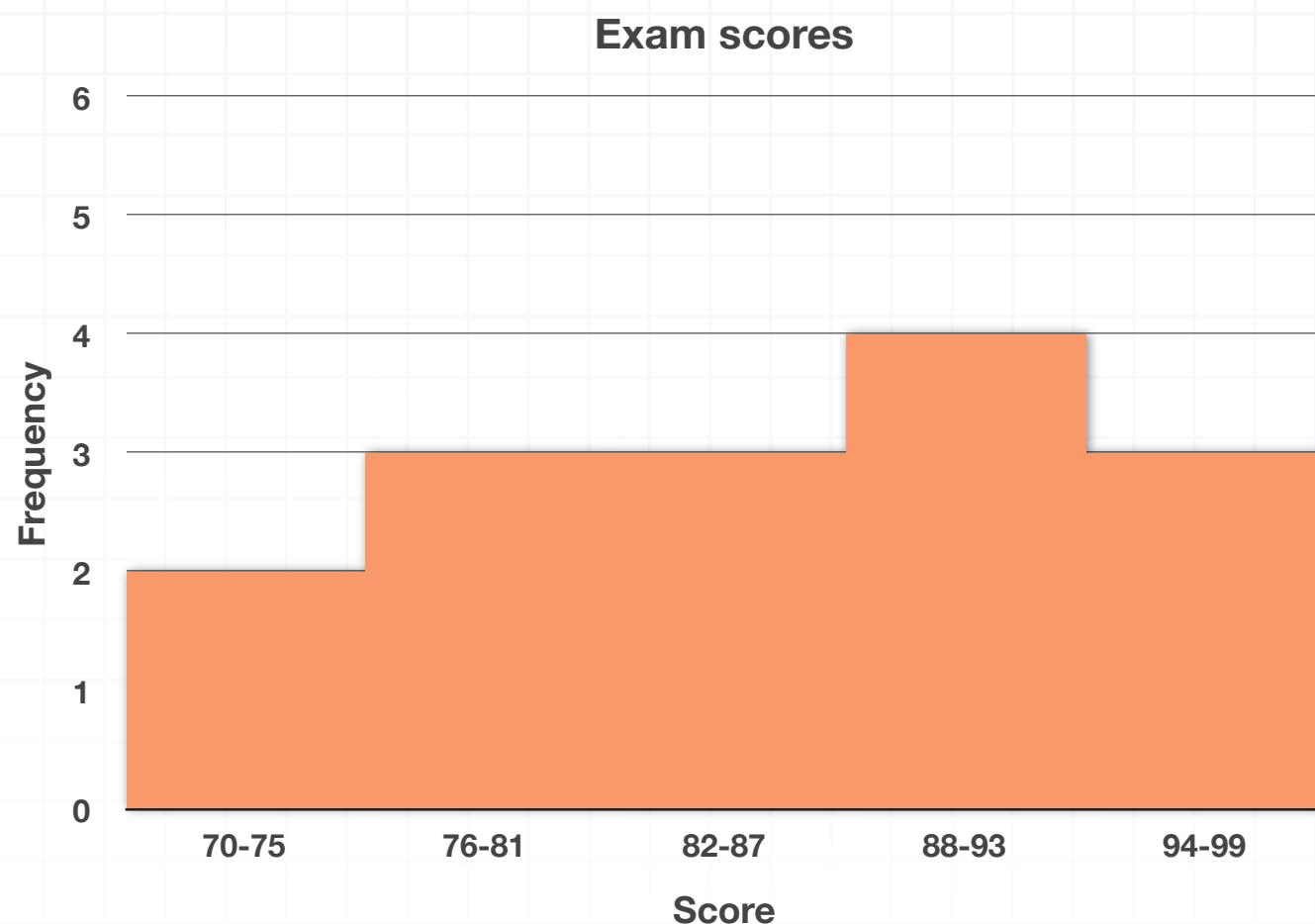
$$\frac{25}{5} = 5$$

But if we use this class width, we'll fall just short of including the complete range of the data. So to make sure we include the smallest and largest values, we'll use a class width of 6.

Now we can set up the class intervals and count their respective frequencies.

Class interval	Frequency
70 - 75	2
76 - 81	3
82 - 87	3
88 - 93	4
94 - 99	3

Now we can sketch the histogram.



Topic: Measures of central tendency

Question: Heather built 7 buildings with her toy blocks and measured their heights in millimeters. The heights of her block buildings were 750 mm, 850 mm, 700 mm, 650 mm, 750 mm, 900 mm, and 950 mm. What is the mean height of the buildings?

Answer choices:

- A 750 mm
- B 793 mm
- C 250 mm
- D 4,735 mm

Solution: B

To find the mean, add the heights of each building, and then divide by the number of buildings.

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

$$\mu = \frac{750 + 850 + 700 + 650 + 750 + 900 + 950}{7}$$

$$\mu \approx 793 \text{ mm}$$

Topic: Measures of central tendency**Question:** Which statement is a true statement about the data set?

2, 2, 2, 4, 4, 6, 6, 7, 8, 9, 10, 10, 10

Answer choices:

- A The data is bi-modal.
- B The mean is larger than the median.
- C Both A and B are true.
- D Both A and B are false.



Solution: C

The mode is the number that appears most often in a data set. This data has two modes, 2 and 10, because they both appear three times in the data set. This means the data is bi-modal.

To find the median, find the middle number. One way to do this is to cross off numbers on the right and left until you find the middle number. Here you can see the median is 6.

$$2, 2, 2, 4, 4, 6, 7, 8, 9, 10, 10, 10$$

To find the mean, add all of the numbers in the data set together and divide by how many there are.

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

$$\mu = \frac{2 + 2 + 2 + 4 + 4 + 6 + 6 + 7 + 8 + 9 + 10 + 10 + 10}{13}$$

$$\mu \approx 6.15$$

Since $6.15 > 6$, the mean is larger than the median. This means answer choices A and B are both true statements.



Topic: Measures of central tendency**Question:** Given the data set and its mean, what is the value of x ?

$$61, 80, x, 91$$

$$\mu = 78$$

Answer choices:

- A 78
- B 77.25
- C 80
- D 85.6



Solution: C

The mean of the data set is $\mu = 78$, and the data set contains the four numbers 61, 80, x , 91.

To find the mean, add all of the numbers in the data set together and divide by how many data points are in the set. This time we know the mean is 78, so we can set up the equation:

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

$$78 = \frac{61 + 80 + x + 91}{4}$$

When we solve for x we get:

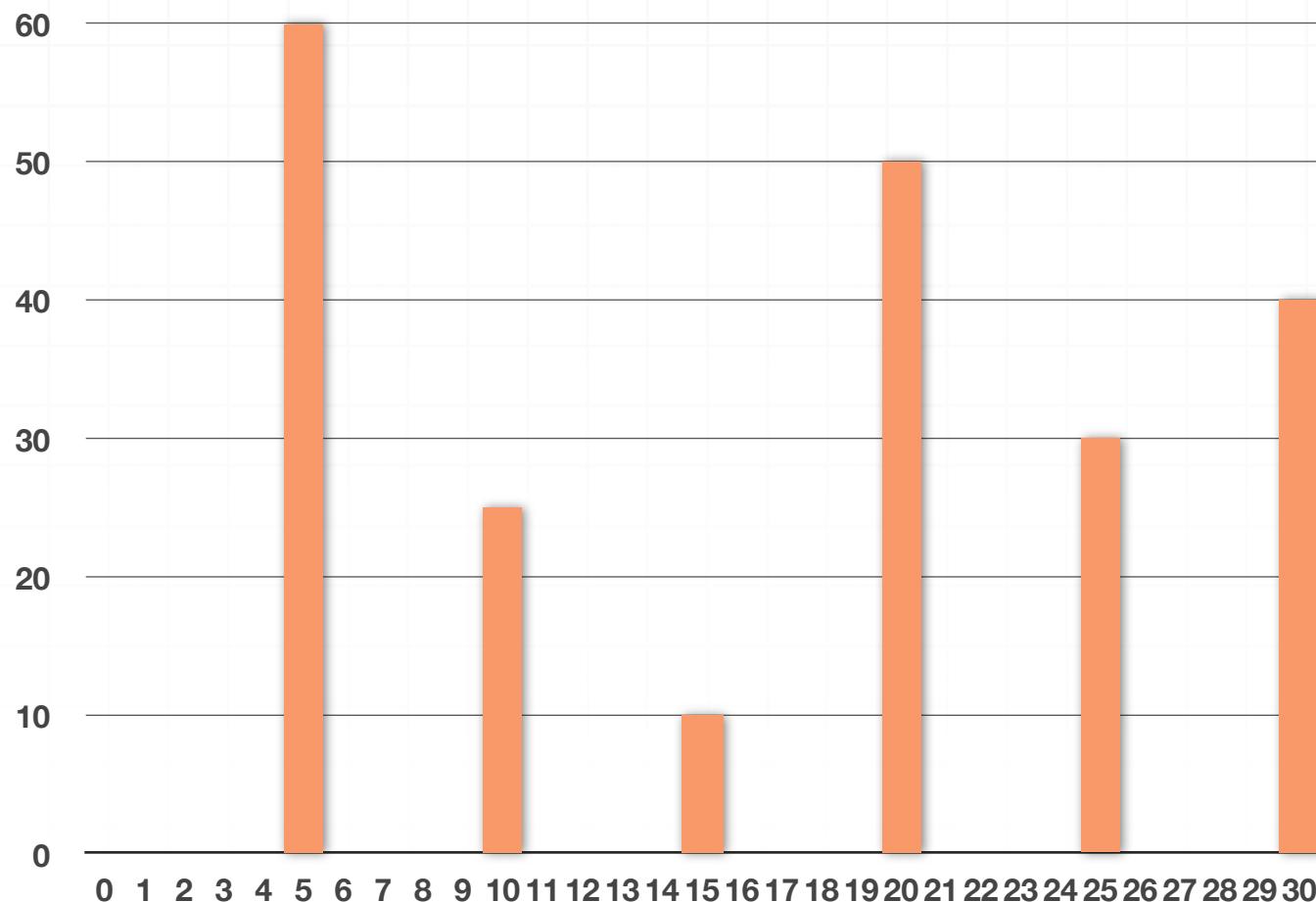
$$78(4) = 61 + 80 + x + 91$$

$$312 = 61 + 80 + x + 91$$

$$312 - 61 - 80 - 91 = x$$

$$x = 80$$



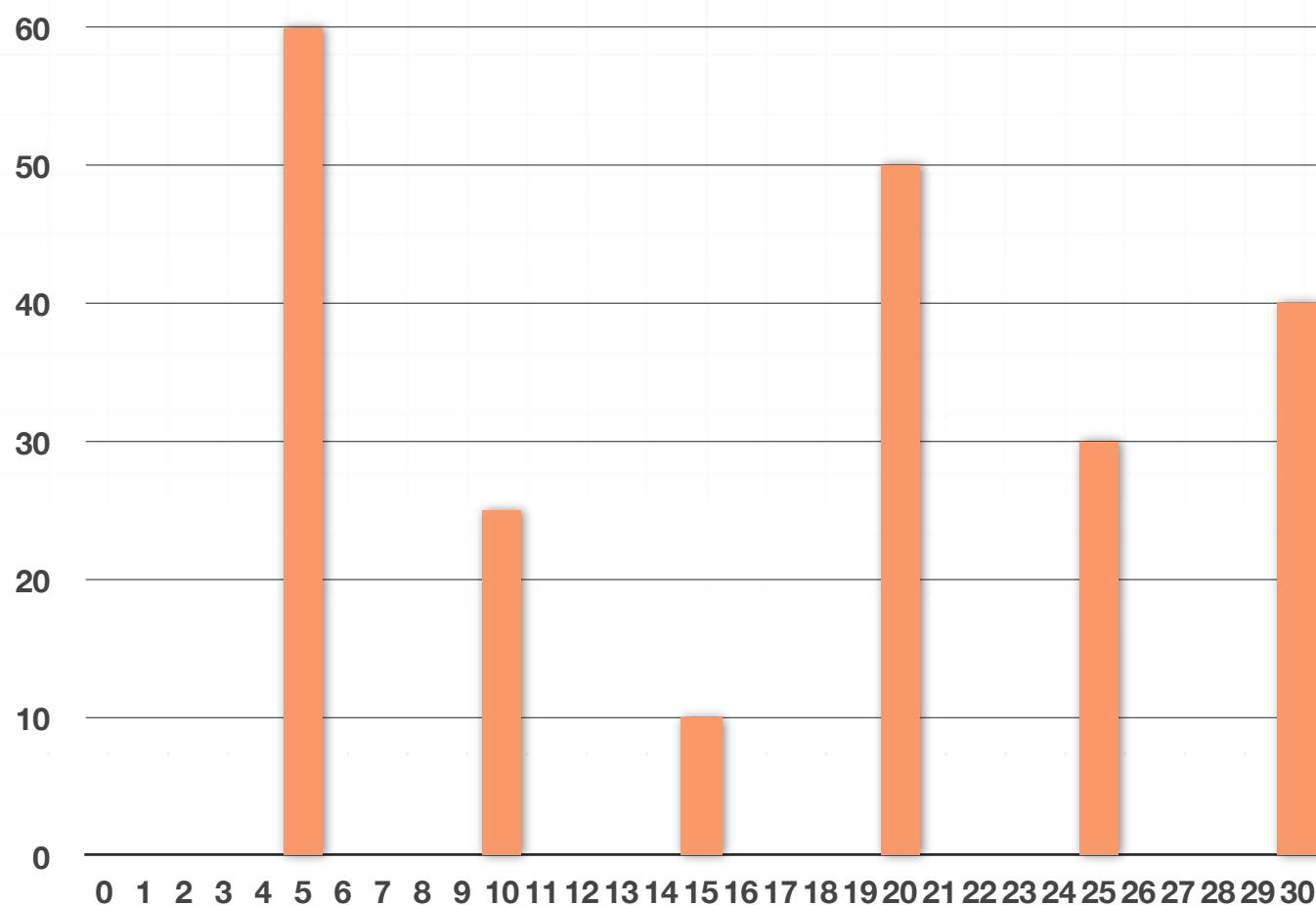
Topic: Measures of spread**Question:** What is the range of the data set shown in the graph?**Answer choices:**

- A 5
- B 10
- C 25
- D 50

Solution: C

The range of a data set is the difference between the largest value and the smallest value in the data set.

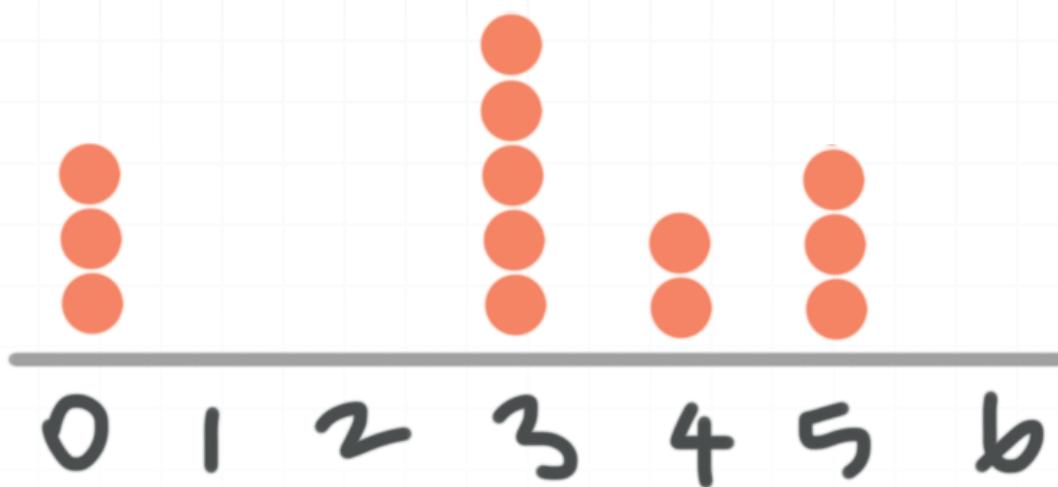
Here, the largest number is 30 and the smallest number is 5. The frequency at which 5 or 30 occurred (60 and 40, respectively) isn't relevant to determining the range.



This makes the range $30 - 5 = 25$. Again, be careful not to use the frequencies to calculate the range, they just tell us how many of a specific number we have. For example, in this data set we have 60 fives, 25 tens, and so on.

Topic: Measures of spread

Question: The dot plot shows the number of emails sent on Monday by each employee. What is the IQR of the data?

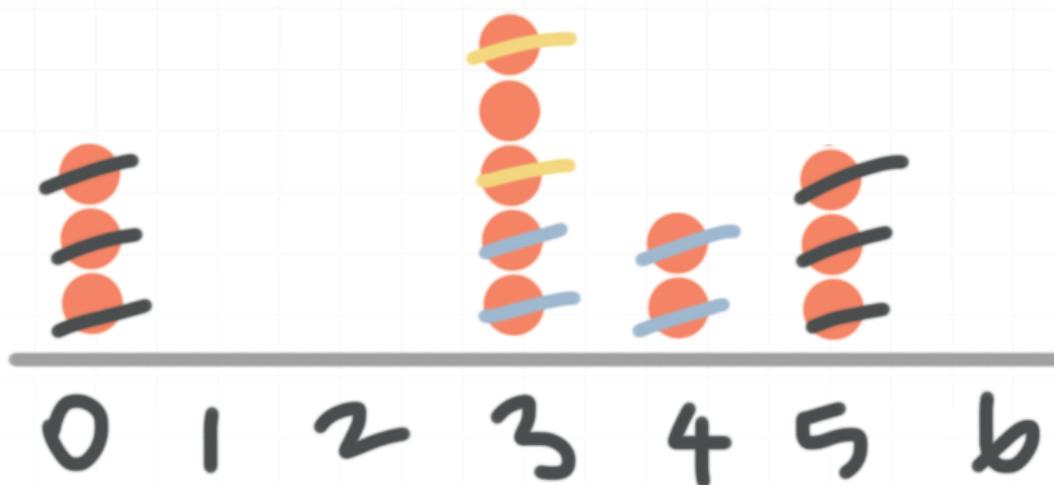


Answer choices:

- A 2
- B 3
- C 5
- D 6

Solution: B

To find the IQR of a data set, we need to find the median of the upper half and the median of the lower half. First let's find the median of the full data set. There are 13 items in the data set, so if we cross off six data items from each side,



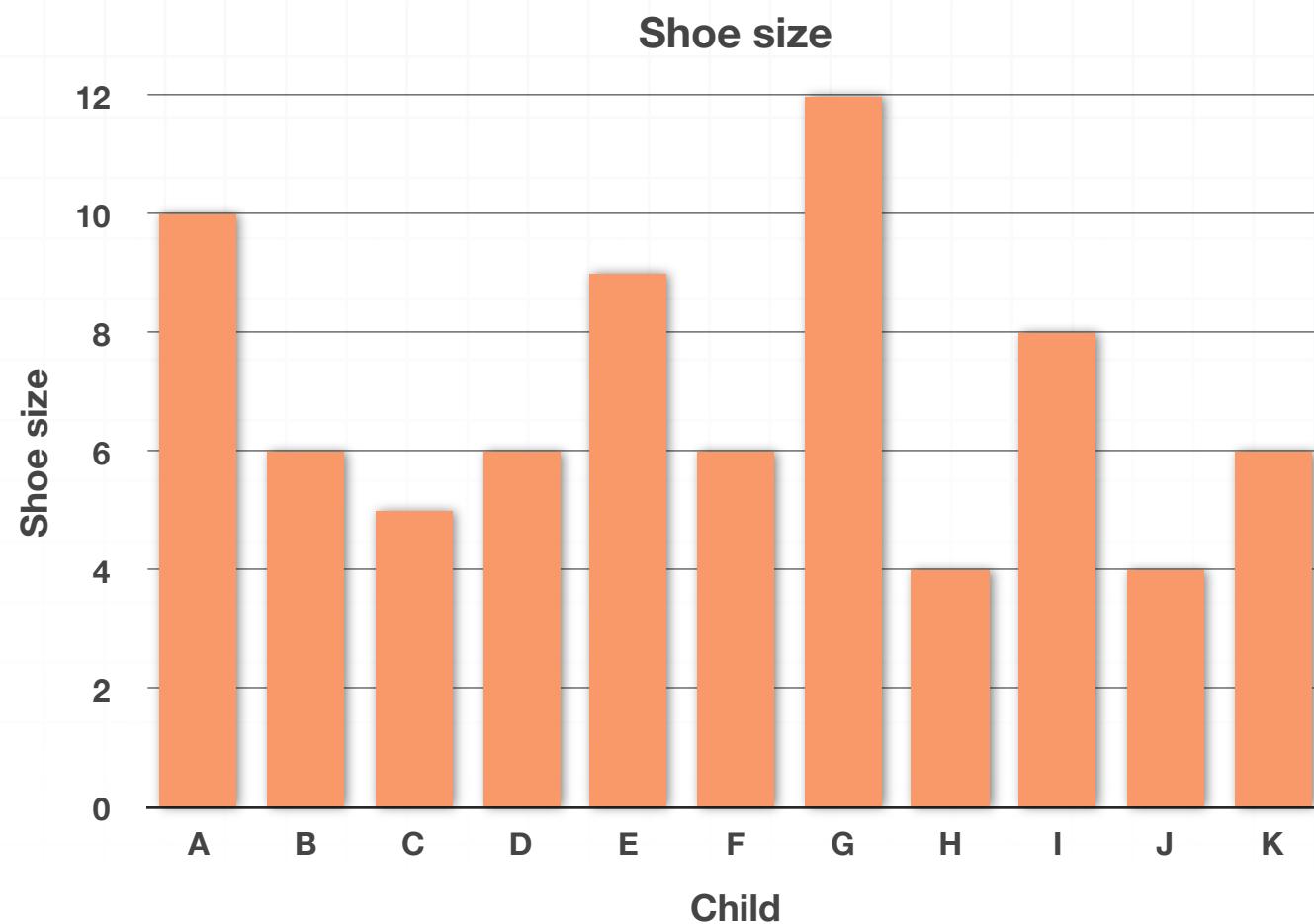
we see that the median is 3. The lower half of the data (everything below the median) is 0, 0, 0, 3, 3, 3, and the median of that lower half is

$$\frac{0+3}{2} = 1.5$$

The upper half of the data (everything above the median) is 3, 4, 4, 5, 5, 5, and the median of that upper half is

$$\frac{4+5}{2} = 4.5$$

Now we can say that the IQR is $4.5 - 1.5 = 3$.

Topic: Measures of spread**Question:** What is the range of the data set shown in the graph?**Answer choices:**

- A 4
- B 8
- C 14
- D 12

Solution: B

The range of a data set is the largest number minus the smallest number. Child G has the largest shoe size (size 12), and children H and J share the smallest shoe size (size 4). This means the range is $12 - 4 = 8$.



Topic: Changing the data and outliers

Question: A company measures the width of its copper tubing in inches. The mean diameter of a tube is $1/8$ of an inch. They want to change the measurement to centimeters for shipment overseas. Given $1 \text{ in} = 2.54 \text{ cm}$, how does the conversion affect the mean diameter?

Answer choices:

- A The new mean is $(1/8) + 2.54$ centimeters.
- B The new mean is $(1/8)(2.54)$ centimeters.
- C The new mean is $(1/8) \div (2.54)$ centimeters.
- D The mean remains the same.

Solution: B

To convert the inches to centimeters, you need to multiply. Scaling a data set by multiplying changes the mean by the same factor.



Topic: Changing the data and outliers

Question: The IQR of a data set is 57. How will subtracting 5 from each data point in the data set affect the IQR?

Answer choices:

- A The IQR will increase by 5.
- B The IQR will decrease by 5.
- C The IQR will be divided by 5.
- D The IQR will stay the same.



Solution: D

Subtracting a constant (like 5) from each data point in a data set will have no effect on the IQR.



Topic: Changing the data and outliers

Question: The students in an English class ended up with a mean score on their recent exam of 65 points. The range of the scores was 25 points. If each score is increased by 5 points, what are the new mean and range?

Answer choices:

- A The new mean is 65; the new range is 25
- B The new mean is 65; the new range is 30
- C The new mean is 70; the new range is 25
- D The new mean is 70; the new range is 30

Solution: C

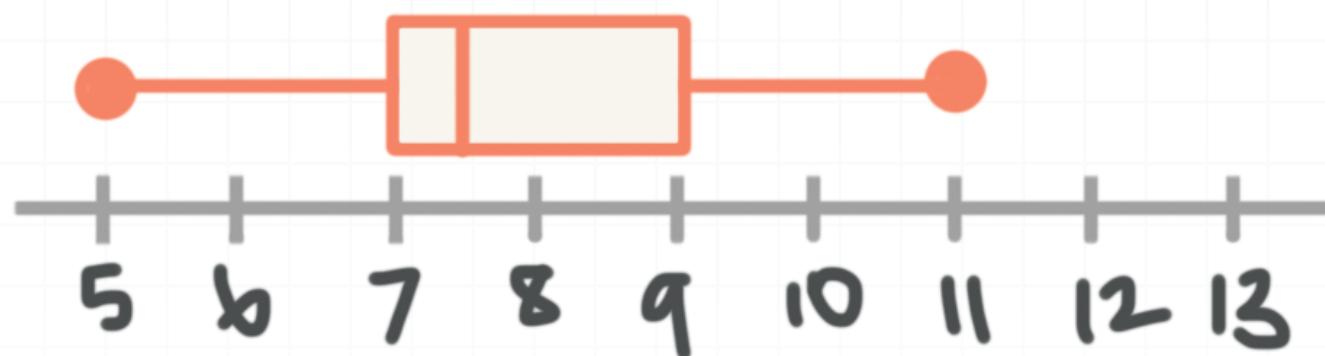
Adding 5 points to all of the exam scores increases all of the scores by 5 points, but the distances between the exam scores remain the same. That is why the mean increases by 5 but the range stays the same.

The original mean is 65 and the new mean is $65 + 5 = 70$. The original range is 25 and the new range is 25.



Topic: Box-and-whisker plots

Question: The box plot shows the number of hours slept on vacation. What is the median of the data?

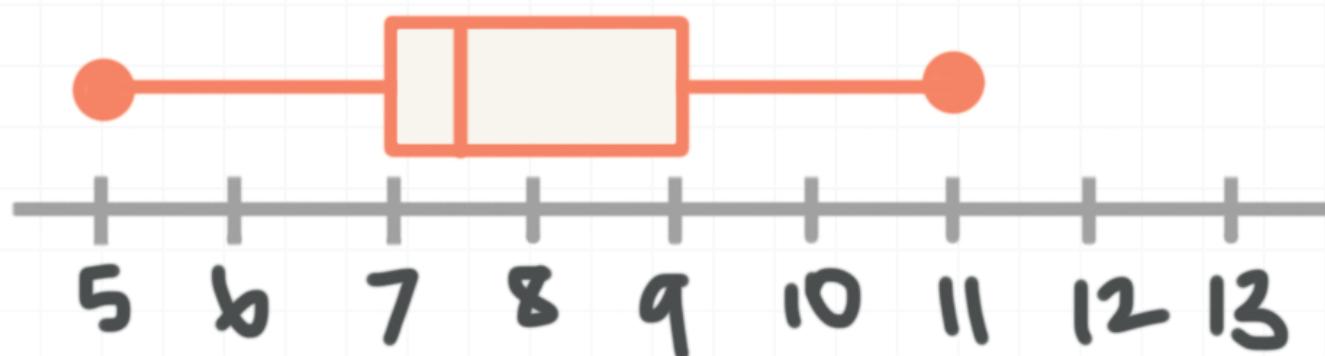


Answer choices:

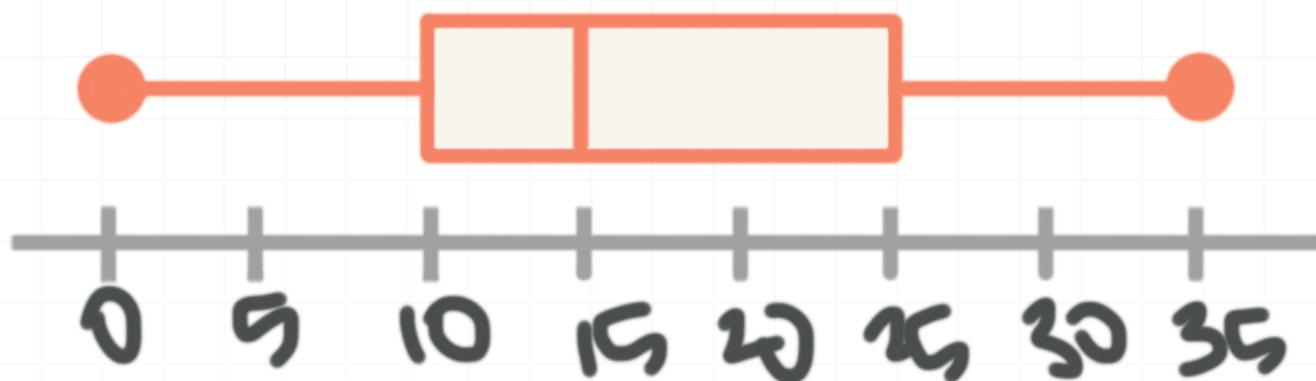
- A 2
- B 6
- C 7.5
- D 8.5

Solution: C

The median of a box-and-whisker plot is the line inside of the box.



Since that line is sitting between 7 and 8, answer choice C of 7.5 can be the only correct choice.

Topic: Box-and-whisker plots**Question:** Build the 5-number summary for the box plot.**Answer choices:****A**

Min	Q ₁	Median	Q ₃	Max
5	10	15	20	25

B

Min	Q ₁	Median	Q ₃	Max
35	25	15	10	0

C

Min	Q ₁	Median	Q ₃	Max
0	10	15	25	35

D

Min	Q_1	Median	Q_3	Max
0	10	25	30	35

Solution: C

The 5-number summary of a data set includes the minimum value, the first quartile, the median, the third quartile, and the maximum value of a data set.

In a box plot, those values are given by the lower edge of the left whisker, the lower edge of the box, the vertical line inside the box, the upper edge of the box, and the upper edge of the right whisker, respectively.

Therefore, the 5-number summary that matches the box plot is

Min	Q_1	Median	Q_3	Max
0	10	15	25	35

Topic: Box-and-whisker plots

Question: If the first quartile is 15 and the third quartile is 25, which statement is true?

Answer choices:

- A The median is 20.
- B The mean is between 15 and 25.
- C Both statements are true.
- D With the information provided, it's impossible to say whether A or B is true or false.



Solution: D

If $Q_1 = 15$ and $Q_3 = 25$, we know the median is between 15 and 25.

But we can't determine its exact value. The mean could be larger than 25 or smaller than 15 if there was an extreme outlier in the data set.



Topic: Mean, variance, and standard deviation**Question:** Find the mean and population standard deviation for the data set.

6, 3, 3, 2, 2

Answer choices:

- A $\mu = 3.2, \sigma \approx 1.4697$
- B $\mu = 3.2, \sigma \approx 1.6431$
- C $\mu = 3.0, \sigma \approx 1.4697$
- D $\mu = 3.0, \sigma \approx 1.6431$

Solution: A

The formula for the population mean is

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

This means we add up all of our numbers and divide by how many there are. Our numbers are 6, 3, 3, 2, 2, and we have 5 of them, so $N = 5$.

$$\mu = \frac{6 + 3 + 3 + 2 + 2}{5}$$

$$\mu = \frac{16}{5}$$

$$\mu = 3.2$$

To find the population standard deviation, we first need to find the population variance and then take the square root. The variance is

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

This means we take each value and subtract the mean, then square it and add the sum. Then we divide the sum by the number of items in the data set. We know that $\mu = 3.2$ and $N = 5$, so we get

$$\sigma^2 = \frac{(6 - 3.2)^2 + (3 - 3.2)^2 + (3 - 3.2)^2 + (2 - 3.2)^2 + (2 - 3.2)^2}{5}$$

$$\sigma^2 = \frac{(2.8)^2 + (-0.2)^2 + (-0.2)^2 + (-1.2)^2 + (-1.2)^2}{5}$$



$$\sigma^2 = \frac{7.84 + 0.04 + 0.04 + 1.44 + 1.44}{5}$$

$$\sigma^2 = \frac{10.8}{5}$$

$$\sigma^2 = 2.16$$

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

$$\sqrt{\sigma^2} = \sqrt{2.16}$$

$$\sigma = \sqrt{2.16}$$

$$\sigma \approx 1.4697$$



Topic: Mean, variance, and standard deviation**Question:** Find the mean and sample standard deviation for the data set.

2, 4, 7, 9, 10

Answer choices:

- A $\bar{x} = 6.4, S \approx 3.0067$
- B $\bar{x} = 6.4, S \approx 3.3615$
- C $\bar{x} = 8.0, S \approx 3.0067$
- D $\bar{x} = 8.0, S \approx 3.3615$

Solution: B

The formula for the sample mean is

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

This means we add up all of our numbers and divide by how many there are. Our numbers are 2, 4, 7, 9, 10, and we have 5 of them, so $n = 5$.

$$\bar{x} = \frac{2 + 4 + 7 + 9 + 10}{5}$$

$$\bar{x} = \frac{32}{5}$$

$$\bar{x} = 6.4$$

To find the sample standard deviation, we first need to find the sample variance and then take the square root. The formula for the sample variance is

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

This means we take each value and subtract the mean, then square it and add the sum. Then we divide the sum by the number of items in the data set, minus 1. We know that $\bar{x} = 6.4$ and $n - 1 = 5 - 1 = 4$, so we get

$$S^2 = \frac{(2 - 6.4)^2 + (4 - 6.4)^2 + (7 - 6.4)^2 + (9 - 6.4)^2 + (10 - 6.4)^2}{4}$$

$$S^2 = \frac{(-4.4)^2 + (-2.4)^2 + (0.6)^2 + (2.6)^2 + (3.6)^2}{4}$$



$$S^2 = \frac{19.36 + 5.76 + 0.36 + 6.76 + 12.96}{4}$$

$$S^2 = \frac{45.2}{4}$$

$$S^2 = 11.3$$

Now take the square root of the sample variance to get sample standard deviation.

$$\sqrt{S^2} = \sqrt{11.3}$$

$$S \approx 3.3615$$

Topic: Mean, variance, and standard deviation

Question: Consider the small population: 1, 2, 1. If each number is increased by 4, how will the population standard deviation change?

Answer choices:

- A The population standard deviation will increase by 4 also.
- B The population standard deviation will increase by 16.
- C The population standard deviation will be multiplied by 4.
- D The population standard deviation will be the same for both data sets.



Solution: D

The population standard deviation is meant to measure how far apart numbers are from one another. It's a measure of how much the data is spread out.

Adding 4 to each number doesn't change how far apart they are. You can calculate both population standard deviations to be sure.

For the original data set:

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

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

$$\sigma^2 = \frac{\left(-\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2}{3}$$

$$\sigma^2 = \frac{\frac{1}{9} + \frac{4}{9} + \frac{1}{9}}{3}$$

$$\sigma^2 = \frac{\frac{6}{9}}{3}$$

$$\sigma^2 = \frac{6}{27}$$

$$\sigma^2 = \frac{2}{9}$$

$$\sigma = \frac{\sqrt{2}}{3}$$

For the new data set where we add 4 to each data point:

$$\mu = \frac{(1+4) + (2+4) + (1+4)}{3} = \frac{5+6+5}{3} = \frac{16}{3}$$

$$\sigma^2 = \frac{\left(5 - \frac{16}{3}\right)^2 + \left(6 - \frac{16}{3}\right)^2 + \left(5 - \frac{16}{3}\right)^2}{3}$$

$$\sigma^2 = \frac{\left(\frac{15}{3} - \frac{16}{3}\right)^2 + \left(\frac{18}{3} - \frac{16}{3}\right)^2 + \left(\frac{15}{3} - \frac{16}{3}\right)^2}{3}$$

$$\sigma^2 = \frac{\left(-\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2}{3}$$

$$\sigma^2 = \frac{\frac{1}{9} + \frac{4}{9} + \frac{1}{9}}{3}$$

$$\sigma^2 = \frac{\frac{6}{9}}{3}$$

$$\sigma^2 = \frac{6}{27}$$

$$\sigma^2 = \frac{2}{9}$$

$$\sigma = \frac{\sqrt{2}}{3}$$

Both population standard deviations are the same.



Topic: Frequency histograms and polygons, and density curves

Question: Which interval would you use on the horizontal axis to create a frequency polygon for the data?

8, 8, 9, 10, 11, 12, 20, 25, 28, 29, 30, 31, 45

51, 55, 65, 67, 68, 68, 70, 72, 78, 86, 90, 91, 100

Answer choices:

- A 2
- B 5
- C 10
- D 20



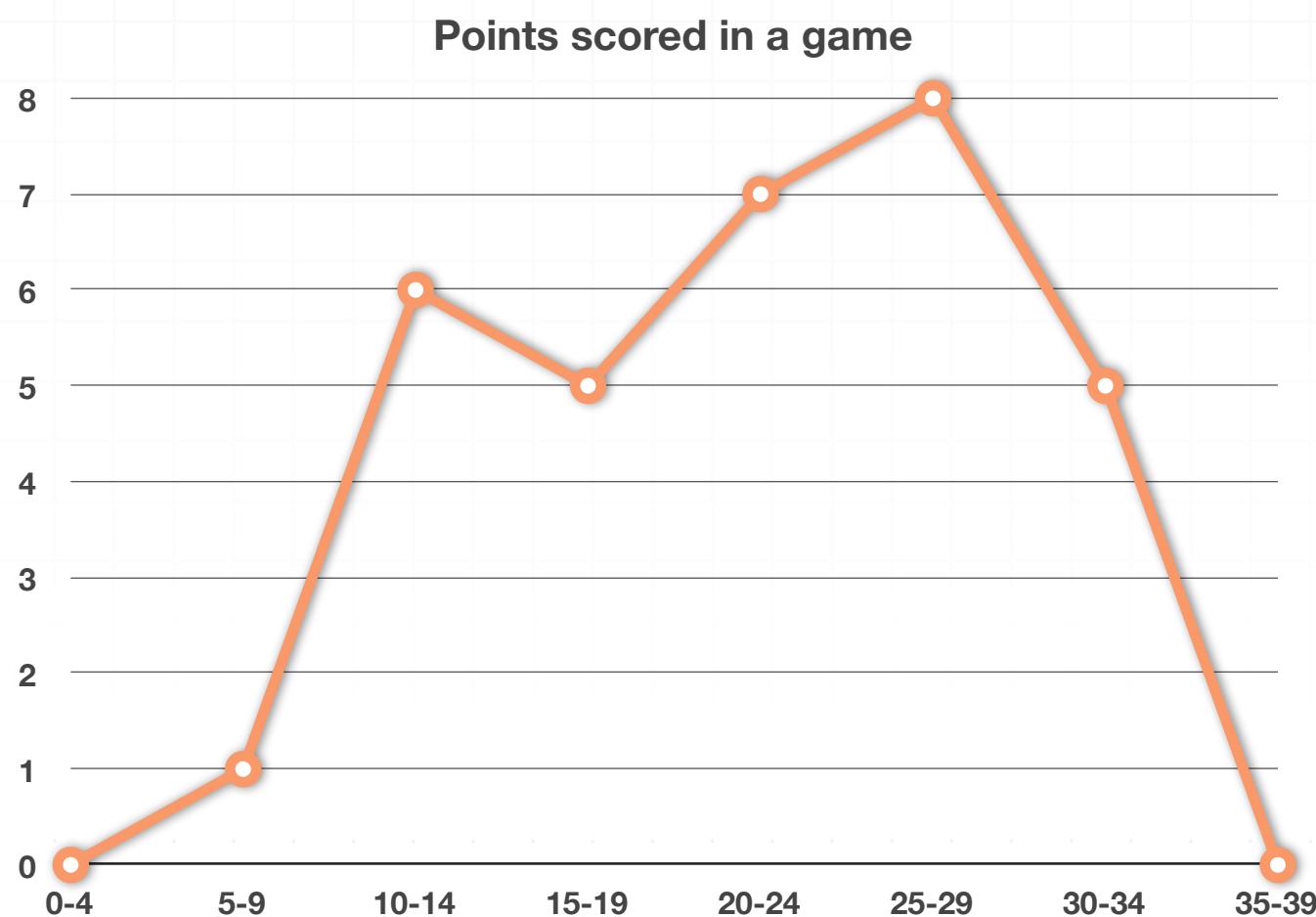
Solution: C

Using tens as the interval for the horizontal axis would give you a good idea for the shape of the data. This set of data has numbers in each of those intervals. Choosing a smaller interval could result in data that was harder to read. Choosing a larger interval could result in a graph that didn't have enough information.



Topic: Frequency histograms and polygons, and density curves

Question: Keith is in charge of a game at the school fair. He keeps track of the points scored by each individual player and creates a frequency polygon. How many times did someone score between 0 – 9 points?

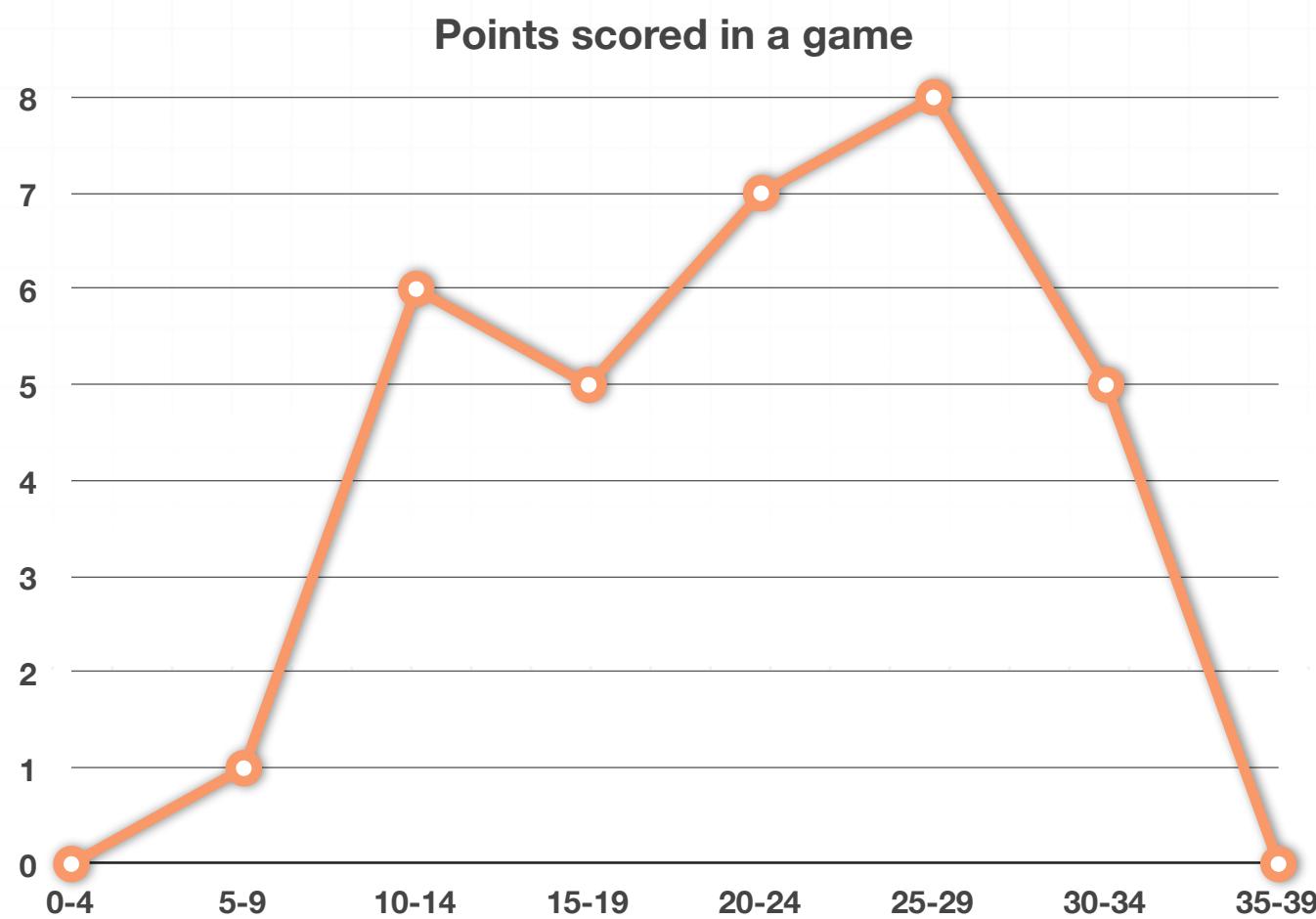
**Answer choices:**

- A 1 time
- B 5 times
- C 7 times

D 32 times

Solution: A

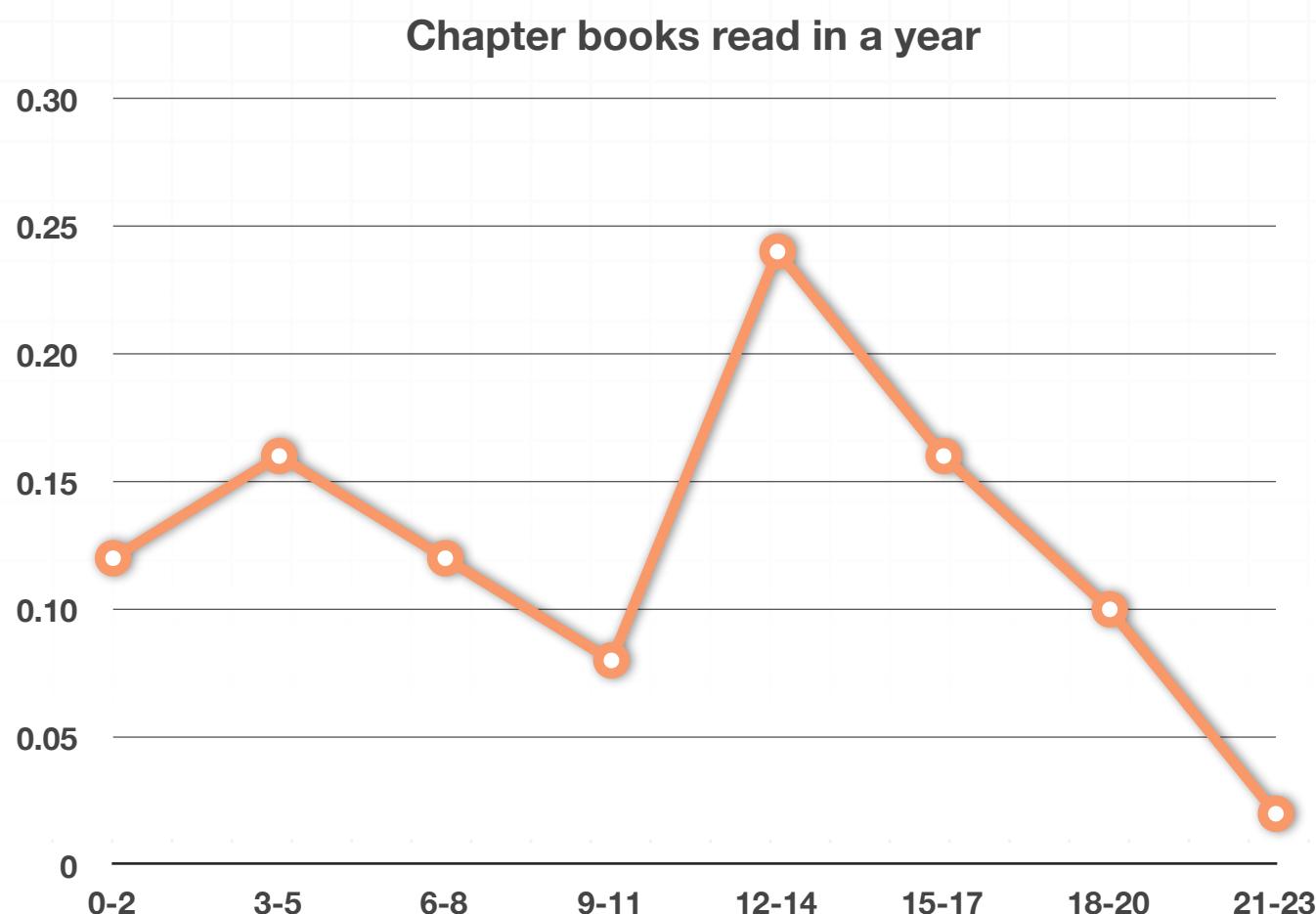
To look at how many times someone scored 0 – 9 points, look at the intervals for 0 – 4 and 5 – 9.



No players scored 0 – 4 points, and one player scored 5 – 9 points.

Topic: Frequency histograms and polygons, and density curves

Question: Mr. Moore created a relative frequency polygon for the number of chapter books read in a year by 50 third graders at his school. How many students read 18 – 20 books during the year?

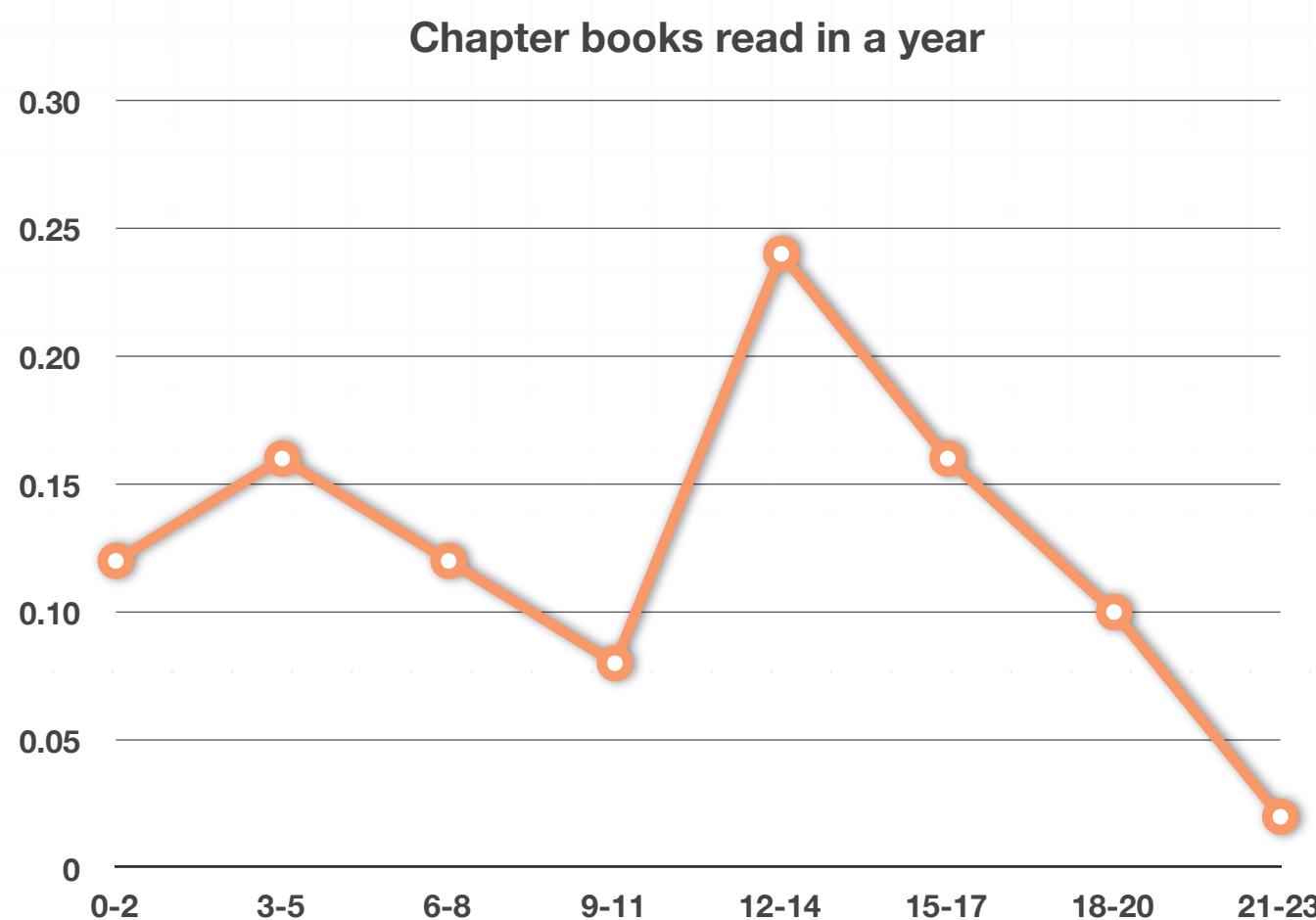
**Answer choices:**

- A 0.10
- B 5
- C 10

- D There is not enough information to answer the question.

Solution: B

We're told the relative frequency polygon contains information on the class of 50 third grade students.



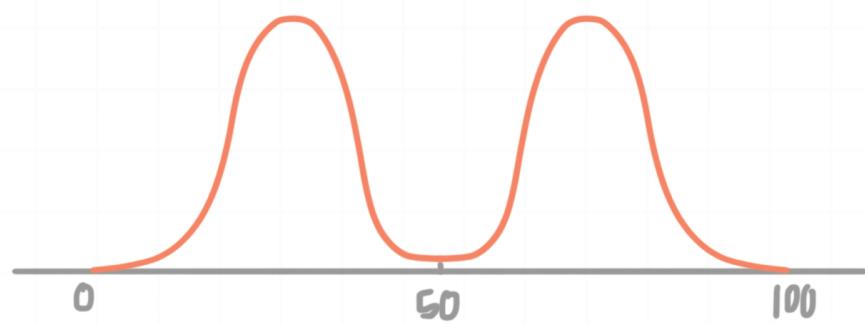
If we look at the 18 – 20 interval, 0.10 or 10 % of the 50 third grade students read 18 – 20 books during the year. We can calculate the number of students by multiplying.

$$0.10(50) = 5$$

This means 5 students read 18 – 20 books during the year.

Topic: Symmetric and skewed distributions and outliers**Question:** Which of the statements are true about the given distributions?

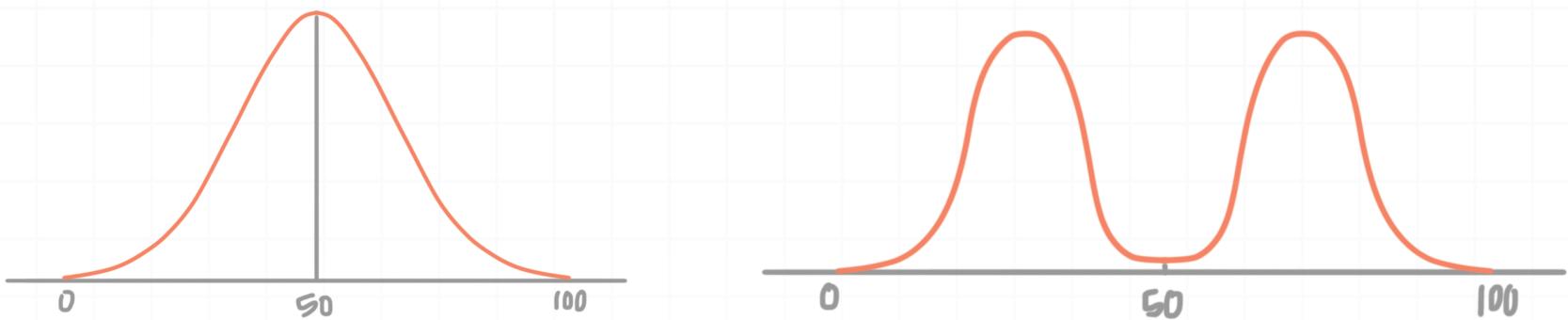
- I. Both distributions have the same mean.
- II. Both distributions have the same range.
- III. Both distributions have the same standard deviation.

**Answer choices:**

- A I only
- B I and II
- C I and III
- D I, II, and III

Solution: B

Both distributions are symmetric, so the mean and median are both in the middle. The mean of both distributions is 50. The range is the largest number in the data set, minus the smallest number. Both of these distributions have a range of $100 - 0 = 100$.



The standard deviation measures the spread of the data set. These data sets are spread out in different ways so their standard deviations will be different.

Topic: Symmetric and skewed distributions and outliers**Question:** Which of the following statements are true?

- I. A right skewed distribution has a mean that's greater than the median.
- II. A symmetric distribution is always normally distributed.
- III. The only reason for a distribution to have a tail is if it has an outlier.

Answer choices:

- A I only
- B I and II
- C I and III
- D I, II, and III

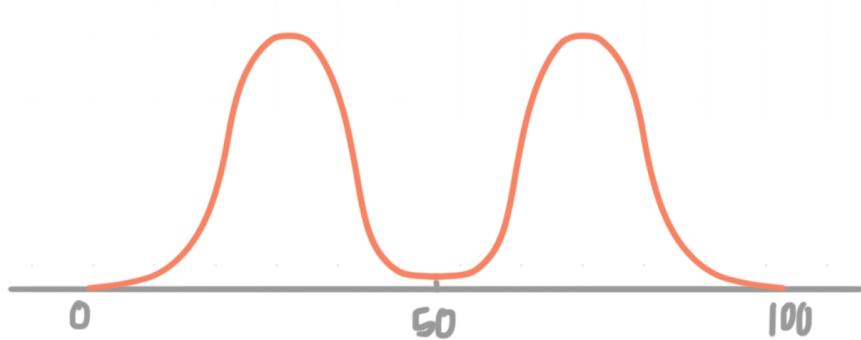
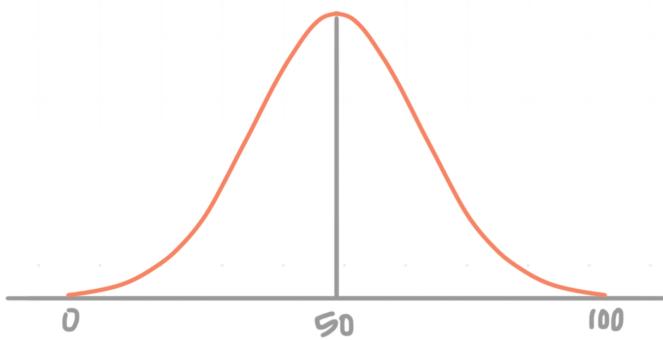


Solution: A

The first statement is true. A right-skewed distribution has a tail on the right. The mean is further to the right than the median.



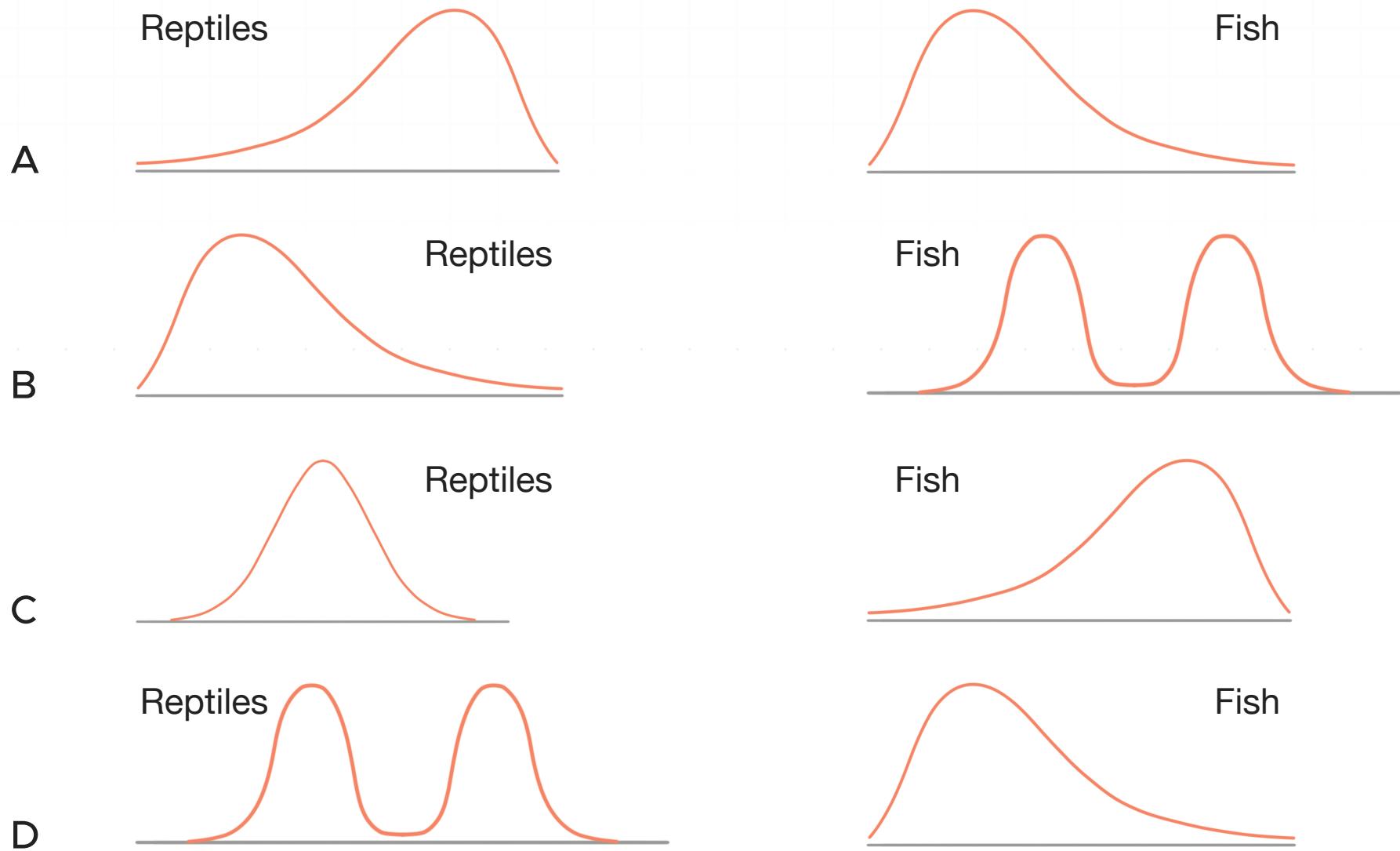
The second statement is false. Symmetric distributions are not always normal distributions. Both of these are examples of symmetric distributions but the one on the left is a normal distribution and the one on the right is a bimodal distribution.



The third statement is false. An outlier could cause a tail in a data set, but some distributions naturally taper off to one side or another, not necessarily because of an outlier.

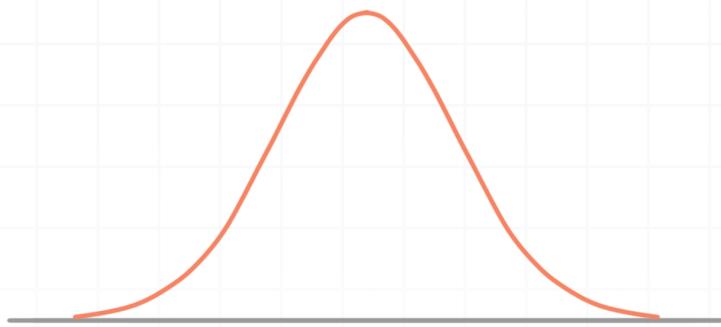
Topic: Symmetric and skewed distributions and outliers

Question: A pet store conducted a survey to determine how much money its customers spent on taking care of reptiles and fish. The amount of money the customers spent on reptiles is symmetrically distributed, but the amount spent on fish has a mean that's less than the median. Which set of distributions could match the descriptions?

Answer choices:

Solution: C

In answer choice C, the distribution for reptiles is a normal distribution, which is one type of symmetric distribution.



The distribution for fish is a left-skewed distribution, also called a negatively skewed distribution. The tail is on the left and the mean is less than the median.



Topic: Normal distributions and z-scores

Question: Given an approximately normal distribution with a mean of 150 and a standard deviation of 28, approximately what percentage of the values fall between 94 and 206? (Note: We can write this interval of values in interval notation as (94,206).)

Answer choices:

- A 68 %
- B 95 %
- C 99.7 %
- D There is not enough information



Solution: B

The empirical rule tells us approximately how much data is within one, two or three standard deviations from the mean.

We know the mean of the data is 150. Since the standard deviation is 28, the interval around one standard deviation is

$$(150 - 28, 150 + 28)$$

$$(122, 178)$$

The interval around two standard deviations is

$$(150 - 28 - 28, 150 + 28 + 28)$$

$$(94, 206)$$

The interval around three standard deviations is

$$(150 - 28 - 28 - 28, 150 + 28 + 28 + 28)$$

$$(66, 234)$$

We are asked about the interval (94,206), which is the interval around two standard deviations from the mean. According to the empirical rule, that interval contains 95 % of the data.

Topic: Normal distributions and z-scores

Question: A third grade class has a mean height of 50" with a standard deviation of 3". What is the approximate percentile of a third grader who is 53" tall?

Answer choices:

- A 16 %
- B 32 %
- C 68 %
- D 84 %



Solution: D

We're given a mean of 50 and a standard deviation of 3, and we want to know the percentile for someone who is 53" tall. We could do this calculation either with the empirical rule or with a z -score.

If you use the empirical rule, you'll need to know that 53 is one standard deviation above the mean. We know there is 68% of the data within one standard deviation, which means there is exactly half of that within one standard deviation, but above the mean only.

$$\frac{68\%}{2} = 34\%$$

Adding this to the 50% of the data below the mean, we can say that someone who is 53" tall is in the $50\% + 34\% = 84\%$ percentile.

If we do this with a z -score, then we use the formula to find the z -score and look up the percentage in the table.

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

We know the mean is $\mu = 50$ and that the standard deviation is $\sigma = 3$. If we plug all of this, plus 53", into the z -score formula, we get

$$z = \frac{53 - 50}{3} = \frac{3}{3} = 1$$

We can look up 1.00 in a z -table, we find the value .8413, which tells us that we're in the 84.13 percentile.

Topic: Normal distributions and z-scores

Question: One of the values in a standard normal distribution is 12, and its z -score is -0.80 . If the mean of the distribution is 14, what is the standard deviation of the distribution?

Answer choices:

- A -0.2119
- B 0.2119
- C 0.25
- D 2.5

Solution: D

The formula for a z -score is

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

We know the mean is $\mu = 14$, and we're looking for the standard deviation. The value of interest is 12, and we know the z -score is -0.80 . Set up the formula and solve for the standard deviation.

$$-0.80 = \frac{12 - 14}{\sigma}$$

$$-0.80\sigma = 12 - 14$$

$$-0.80\sigma = -2$$

$$\sigma = \frac{-2}{-0.80}$$

$$\sigma = 2.5$$



Topic: Chebyshev's Theorem

Question: The shape of a probability distribution is unknown. At least what percentage of the data falls within 2.5 standard deviations of the mean?

Answer choices:

- A 78 %
- B 80 %
- C 84 %
- D 86 %

Solution: C

Using Chebyshev's Theorem with $k = 2.5$, the percentage of data that falls within 2.5 standard deviations of the mean is

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

$$1 - \frac{1}{2.5^2}$$

$$1 - \frac{1}{6.25}$$

$$1 - 0.16$$

$$0.84$$

$$84\%$$

Topic: Chebyshev's Theorem

Question: Find the interval, in terms of standard deviations, that contains at least 90 % of the data in a probability distribution, regardless of the shape of the distribution.

Answer choices:

- A ± 3.02 standard deviations
- B ± 3.04 standard deviations
- C ± 3.10 standard deviations
- D ± 3.17 standard deviations

Solution: D

To find the number of standard deviations that contain a specific percentage of the data in a probability distribution, we can set Chebyshev's expression equal to the percentage we're interested in. Since we're looking for the interval for 90 % of the data, we'll set Chebyshev's expression equal to 0.9.

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

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

$$1 = (1 - 0.9)k^2$$

$$k = \pm \sqrt{\frac{1}{1 - 0.9}}$$

Now that we've solved for k , we can simplify.

$$k = \pm \sqrt{\frac{1}{0.1}}$$

$$k = \pm \sqrt{10}$$

$$k \approx \pm 3.17$$



Topic: Chebyshev's Theorem

Question: A particular school of fish has a mean body length of 8 inches, with a standard deviation of 0.75 inches. What's the minimum body length of a fish in the middle 85% of the school?

Answer choices:

- A 6.07 inches
- B 6.10 inches
- C 6.12 inches
- D 6.15 inches



Solution: A

Using Chebyshev's Theorem,

$$0.85 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.85$$

$$1 = 0.15k^2$$

$$k^2 = \frac{1}{0.15}$$

$$k \approx 2.58$$

Approximately 2.58 standard deviations above the mean gives us a body length of

$$8 + 2.58(0.75)$$

$$9.94$$

And 2.58 standard deviations below the mean gives us a body length of

$$8 - 2.58(0.75)$$

$$6.07$$

So at least 85 % of the fish fell between 6.07 and 9.94 inches.



Topic: Covariance**Question:** Calculate the covariance of the sample.

X	2	3	4	5	6
Y	3	5	7	9	11

Answer choices:

- A $s_{XY} = 8$
- B $s_{XY} = 7$
- C $s_{XY} = 5$
- D $s_{XY} = 4$

Solution: C

Find the mean of X ,

$$\bar{X} = \frac{2 + 3 + 4 + 5 + 6}{5}$$

$$\bar{X} = \frac{20}{5}$$

$$\bar{X} = 4$$

and then the mean of Y .

$$\bar{Y} = \frac{3 + 5 + 7 + 9 + 11}{5}$$

$$\bar{Y} = \frac{35}{5}$$

$$\bar{Y} = 7$$

Now use the means to find the sample covariance.

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

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (2 - 4)(3 - 7) + (3 - 4)(5 - 7)$$

$$+ (4 - 4)(7 - 7) + (5 - 4)(9 - 7) + (6 - 4)(11 - 7)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = -2(-4) - 1(-2) + 0(0) + 1(2) + 2(4)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 8 + 2 + 2 + 8$$



$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 20$$

$$s_{XY} = \frac{20}{5 - 1}$$

$$s_{XY} = 5$$



Topic: Covariance

Question: If X takes on the sample values $\{2, 4, 6, 8, 10, 15\}$, and Y takes on the sample values $\{12, 17, 23, 25, 33, 40\}$, find the covariance of X and Y .

Answer choices:

- A 39.2
- B 46
- C 47
- D 56



Solution: C

Find the mean of X ,

$$\bar{X} = \frac{2 + 4 + 6 + 8 + 10 + 15}{6}$$

$$\bar{X} = \frac{45}{6}$$

$$\bar{X} = 7.5$$

and then the mean of Y .

$$\bar{Y} = \frac{12 + 17 + 23 + 25 + 33 + 40}{6}$$

$$\bar{Y} = \frac{150}{6}$$

$$\bar{Y} = 25$$

Now use the means to find the sample covariance.

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

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (2 - 7.5)(12 - 25) + (4 - 7.5)(17 - 25)$$

$$+ (6 - 7.5)(23 - 25) + (8 - 7.5)(25 - 25)$$

$$+ (10 - 7.5)(33 - 25) + (15 - 7.5)(40 - 25)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 235$$



$$s_{XY} = \frac{235}{6 - 1}$$

$$s_{XY} = 47$$



Topic: Covariance

Question: Two corporations record their stock returns between 2010 and 2014. From the sample, calculate the covariance of their stock returns.

	2010	2011	2012	2013	2014
X	2%	1%	-2%	4%	-1%
Y	3%	0%	1%	2%	1%

Answer choices:

- A 0.95
- B 1.08
- C 1.35
- D 1.49

Solution: C

Find the mean of X ,

$$\bar{X} = \frac{2 + 1 + (-2) + 4 + (-1)}{5}$$

$$\bar{X} = \frac{4}{5}$$

$$\bar{X} = 0.8$$

and then the mean of Y .

$$\bar{Y} = \frac{3 + 0 + 1 + 2 + 1}{5}$$

$$\bar{Y} = \frac{7}{5}$$

$$\bar{Y} = 1.4$$

Now use the means to find the sample covariance.

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

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (2 - 0.8)(3 - 1.4) + (1 - 0.8)(0 - 1.4)$$

$$+ (-2 - 0.8)(1 - 1.4) + (4 - 0.8)(2 - 1.4) + (-1 - 0.8)(1 - 1.4)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 5.4$$

$$s_{XY} = \frac{5.4}{5 - 1}$$



$$s_{XY} = 1.35$$



Topic: Correlation coefficient

Question: Given the sample covariance $s_{xy} = 0.456$ and the sample standard deviations $s_x = 2.53$ and $s_y = 0.25$, calculate the value of the correlation coefficient.

Answer choices:

- A 0.0480
- B 0.2132
- C 0.2884
- D 0.7209

Solution: D

We can substitute the sample covariance $s_{xy} = 0.456$ and the sample standard deviations $s_x = 2.53$ and $s_y = 0.25$ into the formula for the correlation coefficient.

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

$$r_{xy} = \frac{0.456}{2.53 \cdot 0.25}$$

$$r_{xy} = 0.7209$$

Topic: Correlation coefficient

Question: Two corporations record their stock returns between 2010 and 2014. Calculate the value of the correlation coefficient and interpret the result.

	2010	2011	2012	2013	2014
X	2%	1%	-2%	4%	-1%
Y	3%	0%	1%	2%	1%

Answer choices:

- A $r_{XY} \approx 1.35$
- B $r_{XY} \approx 0.4961$
- C $r_{XY} \approx 0.8644$
- D $r_{XY} \approx -0.8644$

Solution: B

Find the mean of X ,

$$\bar{X} = \frac{2 + 1 + (-2) + 4 + (-1)}{5}$$

$$\bar{X} = \frac{4}{5}$$

$$\bar{X} = 0.8$$

and then the mean of Y .

$$\bar{Y} = \frac{3 + 0 + 1 + 2 + 1}{5}$$

$$\bar{Y} = \frac{7}{5}$$

$$\bar{Y} = 1.4$$

Now use the means to find the sample covariance.

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

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = (2 - 0.8)(3 - 1.4) + (1 - 0.8)(0 - 1.4)$$

$$+ (-2 - 0.8)(1 - 1.4) + (4 - 0.8)(2 - 1.4) + (-1 - 0.8)(1 - 1.4)$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 5.4$$

$$s_{XY} = \frac{5.4}{5 - 1}$$

$$s_{XY} = 1.35$$

Next we'll need the standard deviation for corporation X ,

$$s_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= (2 - 0.8)^2 + (1 - 0.8)^2 + (-2 - 0.8)^2 \\ &\quad + (4 - 0.8)^2 + (-1 - 0.8)^2 \end{aligned}$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = 22.8$$

$$s_X = \sqrt{\frac{22.8}{5 - 1}}$$

$$s_X \approx 2.387$$

and the standard deviation for corporation Y .

$$s_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= (3 - 1.4)^2 + (0 - 1.4)^2 \\ &\quad + (1 - 1.4)^2 + (2 - 1.4)^2 + (1 - 1.4)^2 \end{aligned}$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = 5.2$$



$$s_Y = \sqrt{\frac{5.2}{5 - 1}}$$

$$s_Y \approx 1.140$$

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

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

$$r_{XY} \approx \frac{1.35}{2.387 \cdot 1.140}$$

$$r_{XY} \approx 0.4961$$

The correlation coefficient tells us that there's a moderate positive correlation between the annual stock return of the two corporations.

From what we know generally about the stock market, we might suspect that it's not necessarily the return of one stock that's causing the return of the other, but instead that broader market forces might be causing the returns of both stocks.



Topic: Correlation coefficient

Question: Given the sample covariance $s_{XY} = -0.12$ of the data set, calculate the value of the correlation coefficient.

X	12	5	8	18	6	5	11
Y	0.5	0.8	0.3	0.4	0.55	0.25	0.67

Answer choices:

- A -0.3455
- B -0.1282
- C 0.1282
- D 0.3455

Solution: B

Find the mean of X ,

$$\bar{X} = \frac{12 + 5 + 8 + 18 + 6 + 5 + 11}{7}$$

$$\bar{X} \approx 9.286$$

and then the mean of Y .

$$\bar{Y} = \frac{0.5 + 0.8 + 0.3 + 0.4 + 0.55 + 0.25 + 0.67}{7}$$

$$\bar{Y} \approx 0.496$$

Next we'll need the standard deviation of X ,

$$s_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = (12 - 9.286)^2 + (5 - 9.286)^2 + (8 - 9.286)^2$$

$$+ (18 - 9.286)^2 + (6 - 9.286)^2 + (5 - 9.286)^2 + (11 - 9.286)^2$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 \approx 135.429$$

$$s_X = \sqrt{\frac{135.429}{7 - 1}}$$

$$s_X \approx 4.751$$



and the standard deviation of Y .

$$s_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = (0.5 - 0.496)^2 + (0.8 - 0.496)^2 + (0.3 - 0.496)^2$$

$$+(0.4 - 0.496)^2 + (0.55 - 0.496)^2$$

$$+(0.25 - 0.496)^2 + (0.67 - 0.496)^2$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 \approx 0.234$$

$$s_Y = \sqrt{\frac{0.234}{7 - 1}}$$

$$s_Y \approx 0.197$$

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

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

$$r_{XY} \approx \frac{-0.12}{4.751 \cdot 0.197}$$

$$r_{XY} \approx -0.1282$$



Topic: Weighted means and grouped data

Question: A website asks visitors to rate their user experience on a scale from 1 to 10, with 1 being the worst experience, and 10 being the best experience. They record 50 responses. Calculate the mean satisfaction score.

Rating	Number of users
1	2
2	1
3	3
4	5
5	7
6	4
7	8
8	7
9	12
10	1

Answer choices:

- A 5.5
- B 5.85
- C 6.44
- D 7.5

Solution: C

We can calculate the weighted sample mean for the user satisfaction scores.

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

$$\bar{x} = \frac{2(1) + 1(2) + 3(3) + 5(4) + 7(5) + 4(6) + 8(7) + 7(8) + 12(9) + 1(10)}{2 + 1 + 3 + 5 + 7 + 4 + 8 + 7 + 12 + 1}$$

$$\bar{x} = 6.44$$

The mean rating is 6.44.



Topic: Weighted means and grouped data

Question: Mark's grade points are 3.0 for English, which corresponds to 4 credits, 4.0 for Physics, which corresponds to 6 credits, 3.5 for Chemistry, which corresponds to 5 credits, and 3.8 for History, which corresponds to 3 credits. Calculate his grade point average.

Answer choices:

- A 3.58
- B 3.61
- C 3.95
- D 4.53

Solution: B

Each subject has its corresponding weight, represented by credits, so we can calculate the weighted grade point average.

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

$$\bar{x} = \frac{(3)(4) + (4)(6) + (3.5)(5) + (3.8)(3)}{4 + 6 + 5 + 3}$$

$$\bar{x} \approx 3.61$$

Mark's grade point average is approximately 3.61.

Topic: Weighted means and grouped data

Question: Given the frequency table of grouped data, calculate the sample mean, variance, and standard deviation.

X	Frequency
1 - 3	2
4 - 6	5
7 - 9	12
10 - 12	4
13 - 15	6

Answer choices:

- A $\bar{x} \approx 8.724$, $s^2 \approx 12.635$, and $s \approx 3.555$
- B $\bar{x} \approx 8.724$, $s^2 \approx 353.793$, and $s \approx 18.809$
- C $\bar{x} \approx 9.036$, $s^2 \approx 12.635$, and $s \approx 3.555$
- D $\bar{x} \approx 9.036$, $s^2 \approx 353.793$, and $s \approx 18.809$

Solution: A

First we need to find the midpoint for each class.

X	Midpoint	Frequency
1 - 3	2	2
4 - 6	5	5
7 - 9	8	12
10 - 12	11	4
13 - 15	14	6

Then the estimate of the sample mean is

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

$$\bar{x} = \frac{2(2) + 5(5) + 12(8) + 4(11) + 6(14)}{2 + 5 + 12 + 4 + 6}$$

$$\bar{x} \approx 8.724$$

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

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

$$\sum_{i=1}^n f_i (M_i - \bar{x})^2 = 2(2 - 8.724)^2 + 5(5 - 8.724)^2$$

$$+ 12(8 - 8.724)^2 + 4(11 - 8.724)^2 + 6(14 - 8.724)^2$$

$$\sum_{i=1}^n f_i (M_i - \bar{x})^2 = 353.793104$$



$$s^2 = \frac{353.793104}{29 - 1}$$

$$s^2 \approx 12.635$$

The standard deviation of the sample will be the square root of the variance.

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

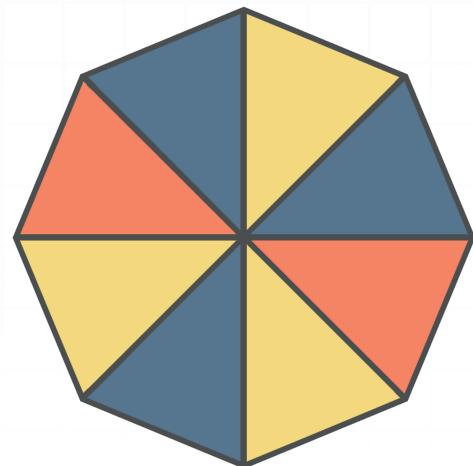
$$s \approx \sqrt{12.635}$$

$$s \approx 3.555$$

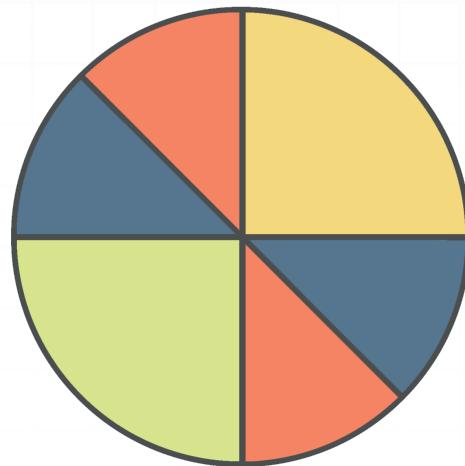


Topic: Simple probability

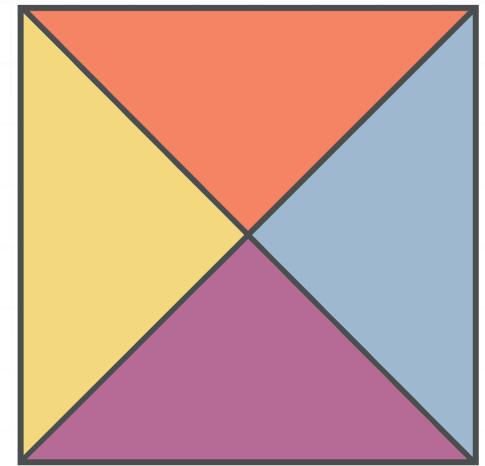
Question: Which spinner has a $1/4$ probability of landing on red? Hint: Consider the size of each section, not just the number of sections in each spinner.

Answer choices:

A



B



C

- D All of the spinners have a $1/4$ probability of landing on red.

Solution: D

To find the probability of the octagon spinner landing on red, consider that the triangles are the same size and shape. This means there are 2 out of 8 total ways to land on red.

$$P(\text{red}) = \frac{2}{8} = \frac{1}{4}$$

To find the probability of the circle spinner landing on red, consider that the red parts each make up 1/8 of the circle, since they are half of a quarter of the circle. We have 2 of them, so

$$P(\text{red}) = \frac{2}{8} = \frac{1}{4}$$

To find the probability of the square spinner landing on red, consider that it's divided into 4 equal triangles, one of which is red.

$$P(\text{red}) = \frac{1}{4}$$



Topic: Simple probability

Question: Brett bought a box of chocolates. 2 of them are coconut, 3 are orange, 4 are caramels, and 1 is raspberry. All of the chocolates look the same. What's the probability that Brett chooses a coconut chocolate?

Answer choices:

- A 20 %
- B 30 %
- C 50 %
- D 60 %



Solution: A

To answer the question, use the definition of simple probability.

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

In this case, the outcomes that meet our criteria are the 2 coconut chocolates. All possible outcomes can be found by adding all of the types together.

$$2 + 3 + 4 + 1 = 10$$

Therefore, the probability of choosing a coconut chocolate is

$$P(\text{coconut}) = \frac{2}{10} = \frac{1}{5} = 0.2 = 20\%$$



Topic: Simple probability

Question: Linel is playing the game Go Fish with her daughter Leah. She needs a picture of a blue whale to win the game. In the pile there are 4 whales, 6 tuna fish, 1 dolphin, and 3 mackerel. To the nearest percent, what is the probability that she draws a whale and wins the game?

Answer choices:

- A 7 %
- B 18 %
- C 35 %
- D 29 %

Solution: D

Remember the definition of simple probability.

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

In this case, the outcomes that meet our criteria are the 4 whales. All possible outcomes can be found by adding all of the types of cards in the deck together:

$$4 + 6 + 1 + 3 = 14$$

Therefore, the probability of pulling a whale from the deck is

$$P(\text{whale}) = \frac{4}{14} \approx 0.29 \approx 29\%$$



Topic: The addition rule, and union vs. intersection**Question:** Which events are mutually exclusive?**Answer choices:**

- A The probability of rolling a sum that's divisible by 5 and 2 when two dice are thrown.
- B The probability of drawing a blue gum ball or a pink gum ball from a jar.
- C The probability of drawing a black card or an ace from a deck of cards.
- D The probability of rolling a 10 or a double when a pair of dice is rolled.



Solution: B

The events in answer choice A are not mutually exclusive, because if you roll a 10, it's divisible by both 5 and 2, so both events could occur at the same time.

The events in answer choice B are mutually exclusive, because you can only draw one gum ball at a time, so you can't draw a blue gum ball and a pink gum ball at the same time.

The events in answer choice C are not mutually exclusive, because if you draw a black ace, you're drawing a black card and an ace, so both events could occur at the same time.

The events in answer choice D are not mutually exclusive, because if you could roll two 5s, your roll is a double and sums to 10, so both events could occur at the same time.



Topic: The addition rule, and union vs. intersection

Question: Which events are not mutually exclusive?

Answer choices:

- A The probability of drawing an ace and the probability of drawing a king from a deck of cards.
- B The probability of rolling a sum that is either 8 or 10 when two dice are thrown.
- C The probability of rolling three 5s in a row when you roll a six-sided die and the probability you roll three 1s in a row when you roll a six-sided die.
- D The probability of selecting a small dog from an animal shelter and the probability of selecting a brown dog from an animal shelter.

Solution: D

The events in answer choices A, B, and C are mutually exclusive because they can't happen at the same time.

But the events in answer choice D are not mutually exclusive, because you could choose a small, brown dog, which means you can choose a small dog and a brown dog at the same time. The events can occur at the same time, so they are not mutually exclusive.



Topic: The addition rule, and union vs. intersection

Question: If we roll one standard 6-sided die, what's the probability that the outcome is both odd and divisible by 3?

Answer choices:

- A $\frac{1}{3}$
- B $\frac{1}{2}$
- C $\frac{1}{6}$
- D Can't be determined

Solution: C

Let A be the event of rolling an odd number. Then

$$A = \{1, 3, 5\}$$

Then let B be the event of rolling a number that's divisible by 3. Then

$$B = \{3, 6\}$$

The probability that the outcome is both odd and divisible by 3 is the intersection of A and B , which means we need to find any common outcomes in both event spaces. The only overlap is 3, so the intersection is

$$A \cap B = \{3\}$$

Then the probability of rolling a 3 is

$$P(A \cap B) = \frac{1}{6}$$



Topic: Independent and dependent events and conditional probability

Question: Events A and B are independent events. Find $P(B)$ if $P(A \text{ and } B) = 0.25$ and $P(A) = 0.5$.

Answer choices:

- A $P(B) = 0.125$
- B $P(B) = 0.45$
- C $P(B) = 0.5$
- D Not enough information



Solution: C

Since the events are independent, events we know that

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

We can plug in $P(A \text{ and } B) = 0.25$ and $P(A) = 0.5$ and solve for $P(B)$.

$$0.25 = 0.5 \cdot P(B)$$

$$P(B) = \frac{0.25}{0.5}$$

$$P(B) = 0.5$$

Topic: Independent and dependent events and conditional probability

Question: Events A and B are dependent events. If $P(A \text{ and } B) = 0.7$ and $P(B) = 0.875$, what is $P(A)$?

Answer choices:

- A $P(A) = 0.875$
- B $P(A) = 0.8$
- C $P(A) = 0.6125$
- D Not enough information



Solution: D

These events are dependent events, so we can say

$$P(A \text{ and } B) = P(A) \cdot P(B|A)$$

We know that $P(A \text{ and } B) = 0.7$, but we would also need to know $P(B|A)$ in order to be able to solve for $P(A)$. Therefore, we don't have enough information to solve the problem.



Topic: Independent and dependent events and conditional probability

Question: Suppose that Katie rolls a six-sided die twice. Event A is that the first roll is a 6, so $P(A)$ is the probability that the first roll is a 6. Event B is that the second roll is a 6, so $P(B)$ is the probability that the second roll is a 6. Which statement is false?

Answer choices:

- A The events are independent.
- B The events are dependent.
- C $P(A \text{ and } B) = P(A) \cdot P(B | A)$
- D $P(A \text{ and } B) = P(A) \cdot P(B)$

Solution: B

Events can't be independent and dependent at the same time, so either answer choice A is false or answer choice B is false.

The rolls are independent if we can show that $P(A \text{ and } B) = P(A) \cdot P(B)$. If events are independent, it doesn't necessarily mean that

$P(A \text{ and } B) = P(A) \cdot P(B|A)$ is a false statement. It just means that

$P(B) = P(B|A)$.

$P(A)$ is the probability that the first die lands on 6, so $P(A) = 1/6$. $P(B)$ is the probability that the second die lands on 6, so $P(B) = 1/6$. $P(A \text{ and } B)$ is the probability of rolling a 6 on both dice, so $P(A \text{ and } B) = 1/36$. Now we can check for independence.

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

$$\frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6}$$

$$\frac{1}{36} = \frac{1}{36}$$

Because this equation is true, the events are independent, not dependent.



Topic: Bayes' Theorem**Question:** When should you use Bayes' Theorem?**Answer choices:**

- A When you have $P(A \cap B)$ but want to find $P(A)$.
- B When you have $P(A | B)$ but want to find $P(B | A)$.
- C When you have $P(A)$ but want to find $P(B)$.
- D When you have $P(A | B)$ but want to find $P(A)$.

Solution: B

Bayes' Theorem is used when you have a conditional probability of two events, and you're interested in the reversed conditional probability. For example, when you have $P(A | B)$ but want to find $P(B | A)$.



Topic: Bayes' Theorem

Question: Three factories A , B , and C produce car seats. What is the probability that a defective car seat comes from factory C , given that factory C produces 40 % of all the car seats, that there's a 1 % chance that any given car seat is defective, and that the defective rate at factory C is 0.8 %?

Answer choices:

- A 28 %
- B 32 %
- C 36 %
- D 40 %

Solution: B

We could name these events.

A represents a car seat from factory A

B represents a car seat from factory B

C represents a car seat from factory C

D represents a defective car seat

We're looking for $P(C|D)$, the probability that a car seat came from factory C , given that it was defective. We know

$$P(C) = 0.4$$

$$P(D) = 0.01$$

$$P(D|C) = 0.008$$

Bayes' Theorem therefore tells us that the probability of $P(C|D)$ is given by

$$P(C|D) = \frac{P(D|C) \cdot P(C)}{P(D)}$$

$$P(C|D) = \frac{(0.008)(0.4)}{0.01}$$

$$P(C|D) = \frac{0.0032}{0.01}$$

$$P(C|D) = 0.32$$



Topic: Bayes' Theorem**Question:** Which choice is equivalent to $P(C|D)$?**Answer choices:**

A
$$\frac{P(D|C) \cdot P(C)}{P(D)}$$

B
$$\frac{P(C \cap D)}{P(D)}$$

C
$$\frac{P(C \cup D)}{P(D)}$$

D Both A and B



Solution: D

Bayes' Theorem is

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

This problem uses different variables. If you replace A with C and B with D , then Bayes' Theorem is

$$P(C|D) = \frac{P(D|C) \cdot P(C)}{P(D)}$$

For dependent events, the multiplication rule says that

$P(C \cap D) = P(C) \cdot P(D|C)$, which means we could also write Bayes' Theorem as

$$P(C|D) = \frac{P(D|C) \cdot P(C)}{P(D)} = \frac{P(C \cap D)}{P(D)}$$



Topic: Discrete probability

Question: A red and blue die are rolled. Both are six-sided fair dice. Let X represent the sum of the dice. Which of the following is the correct probability distribution for X ?

Answer choices:**A**

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

B

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12	1/12

C

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

D

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/36	1/36	1/36	1/36	1/36	1/36	1/36	1/36	1/36	1/36	1/36

Solution: C

Because the smallest value we can roll on each die is 1, the smallest sum we can get is $X = 1 + 1 = 2$. The largest value we can roll on each die is 6, so the largest sum we can get is $X = 6 + 6 = 12$.

Therefore, the sample space for X is $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. This table shows all possible ways to roll two dice and the sum of each roll.

		Red die					
		1	2	3	4	5	6
Blue die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

There are 11 possible sums (from 2 to 12) and 36 different pairs (from (1,1) all the way to (6,6)).

By looking at the table, we can start calculating probabilities for each sum.

$$P(\text{sum of } 2) = \frac{1}{36}$$

$$P(\text{sum of } 3) = \frac{2}{36} = \frac{1}{18}$$

...

Already, the only probability distribution that matches these calculations is the table from answer choice C.

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/3 C	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/3 C

Topic: Discrete probability

Question: You purchase a raffle ticket for \$125. In exchange, you'll be allowed to participate in two drawings. In each drawing, you blindly pick one of three tokens. One token is worth \$0, one is worth \$50, and one is worth \$100. Let Y be the profit you make from a raffle ticket. Find the expected value for Y after the two drawings.

Answer choices:

- A -\$50
- B -\$25
- C \$50
- D \$75

Solution: B

In the first drawing, you can pick \$0, \$50, or \$100. The same is true with the second drawing. So after two drawings, your possible earnings are given in the table.

		Second drawing		
		0	50	100
First drawing	0	0	50	100
	50	50	100	150
	100	100	150	200

The sample space is therefore \$0, \$50, \$100, \$150, or \$200. Because there are 9 possible combinations, from (0,0) to (100,100), the probability of winning each amount of money is

Y	0	50	100	150	200
P(Y)	1/9	2/9	3/9	2/9	1/9

But your ticket cost you \$125, which means we need to adjust the probability distribution by subtracting the cost from each potential profit.

Y	-125	-75	-25	25	75
P(Y)	1/9	2/9	3/9	2/9	1/9

Therefore, the expected value for Y is

$$E(Y) = -125 \left(\frac{1}{9} \right) - 75 \left(\frac{2}{9} \right) - 25 \left(\frac{3}{9} \right) + 25 \left(\frac{2}{9} \right) + 75 \left(\frac{1}{9} \right)$$

$$E(Y) = -\frac{125}{9} - \frac{150}{9} - \frac{75}{9} + \frac{50}{9} + \frac{75}{9}$$

$$E(Y) = -\frac{225}{9}$$

$$E(Y) = -25$$

Topic: Discrete probability

Question: The following table shows the 2017 AP Statistics Exam score distribution for all students taking the test in the United States. Let Z represent the exam score. Find μ_Z and σ_Z .

Score	1	2	3	4	5
Probability	0.136	0.159	0.248	0.202	0.255

Answer choices:

- A $\mu_Z = 3.281$ and $\sigma_Z = 1.846$
- B $\mu_Z = 2.719$ and $\sigma_Z = 1.846$
- C $\mu_Z = 3.281$ and $\sigma_Z = 1.359$
- D $\mu_Z = 2.719$ and $\sigma_Z = 1.359$

Solution: C

Z is a discrete random variable with sample space $\{1, 2, 3, 4, 5\}$. The percentage of students taking the exam who received each of those scores is given in the table as

Score	1	2	3	4	5
Probability	0.136	0.159	0.248	0.202	0.255

We'll find the mean of this discrete random variable as

$$\mu_Z = 1(0.136) + 2(0.159) + 3(0.248) + 4(0.202) + 5(0.255)$$

$$\mu_Z = 0.136 + 0.318 + 0.744 + 0.808 + 1.275$$

$$\mu_Z = 3.281$$

We'll find the variance in order to get to standard deviation. The variance of Z is

$$\sigma_Z^2 = \sum_{i=1}^5 (Z_i - \mu_Z)^2 P(Z_i)$$

$$\begin{aligned} \sigma_Z^2 &= (1 - 3.281)^2(0.136) + (2 - 3.281)^2(0.159) + (3 - 3.281)^2(0.248) \\ &\quad + (4 - 3.281)^2(0.202) + (5 - 3.281)^2(0.255) \end{aligned}$$

$$\sigma_Z^2 = 1.846$$

So the standard deviation of Z is

$$\sqrt{\sigma_Z^2} = \sqrt{1.846}$$

$$\sigma_Z \approx 1.359$$



Topic: Transforming random variables**Question:** Let X be a random variable with $\mu_X = 75$ and $\sigma_X = 8$. Let $Y = 6 + 2X$.Find μ_Y and σ_Y .**Answer choices:**

- A $\mu_Y = 75$ and $\sigma_Y = 8$
- B $\mu_Y = 156$ and $\sigma_Y = 14$
- C $\mu_Y = 156$ and $\sigma_Y = 16$
- D $\mu_Y = 156$ and $\sigma_Y = 22$



Solution: C

Since $Y = 6 + 2X$, we'll scale each value in our data set by the constant 2 and then add a constant of 6. The scaling by 2 will effect the mean and standard deviation, but the shifting by 6 will only effect the mean. The mean of Y will therefore be

$$\mu_Y = 6 + 2(\mu_X)$$

$$\mu_Y = 6 + 2(75)$$

$$\mu_Y = 156$$

And the standard deviation of Y will be

$$\sigma_Y = 2(\sigma_X)$$

$$\sigma_Y = 2(8)$$

$$\sigma_Y = 16$$



Topic: Transforming random variables

Question: An employee at a candy store has the job of cutting fudge into 1-inch cubes and weighing them. After a day at work she finds the mean and standard deviation of her fudge cubes to be 1.5 ounces and 0.3 ounces, respectively. At the end of the day her boss realizes the scale was not calibrated correctly. The weights of all the pieces were actually 0.1 less than recorded. Find the actual mean and standard deviation for the fudge that was cut today.

Answer choices:

- A $\mu = 1.5$ and $\sigma = 0.3$
- B $\mu = 1.4$ and $\sigma = 0.2$
- C $\mu = 1.5$ and $\sigma = 0.2$
- D $\mu = 1.4$ and $\sigma = 0.3$

Solution: D

If all the pieces of fudge actually weighed 0.1 ounces less than what was recorded, the new mean will be lower than originally calculated. In fact, it will be 0.1 ounces less than the original.

$$\mu_{\text{new}} = \mu_{\text{old}} - 0.1$$

$$\mu_{\text{new}} = 1.5 - 0.1$$

$$\mu_{\text{new}} = 1.4 \text{ ounces}$$

The standard deviation won't change when we shift the data by a constant.

$$\sigma_{\text{new}} = \sigma_{\text{old}}$$

$$\sigma_{\text{new}} = 0.3 \text{ ounces}$$



Topic: Transforming random variables

Question: A chemistry teacher gave a test to his students and calculated the following statistics for the class: median of 68, IQR of 12, and range of 58. He decides to “curve” the scores by using this formula:

$$\text{New score} = 1.05(\text{Old score}) + 4$$

Find the median, IQR, and range for the new set of test scores.

Answer choices:

- A Median of 75.4, IQR of 12.6, and Range of 60.9
- B Median of 71.4, IQR of 12.6, and Range of 60.9
- C Median of 75.4, IQR of 16.6, and Range of 64.9
- D Median of 75.4, IQR of 16, and Range of 62



Solution: A

Each student's old score was substituted into the formula to give a new score for that student. For example, if a student had an old score of 75, we would compute his new score as

$$\text{New score} = 1.05(\text{Old score}) + 4$$

$$\text{New score} = 1.05(75) + 4$$

$$\text{New score} = 78.75 + 4$$

$$\text{New score} = 82.75$$

All old scores would be transformed into new, curved scores in this same way. We're only given the median, IQR, and range for the old test scores. The median measures the center for the test scores and the IQR and range both measure the spread in the scores. The median will be transformed in the same way as each individual score.

$$\text{New median} = 1.05(\text{Old median}) + 4$$

$$\text{New median} = 1.05(68) + 4$$

$$\text{New median} = 71.4 + 4$$

$$\text{New median} = 75.4$$

The measures of spread are transformed using only the scale factor of 1.05, but are not affected by adding 4 to each value. Remember that adding a constant k will move all of values up by k units, but won't make



the data any more or less spread out. Therefore, we'll convert the IQR and range this way:

$$\text{New IQR} = 1.05(\text{Old IQR})$$

$$\text{New IQR} = 1.05(12)$$

$$\text{New IQR} = 12.6$$

and

$$\text{New range} = 1.05(\text{Old range})$$

$$\text{New range} = 1.05(58)$$

$$\text{New range} = 60.9$$



Topic: Combinations of random variables

Question: Let A and B be independent, continuous random variables with $\mu_A = 85$, $\sigma_A = 6$, $\mu_B = 92$, and $\sigma_B = 8$. Find μ_{A+B} and σ_{A+B} .

Answer choices:

- A $\mu_{A+B} = 177$ and $\sigma_{A+B} = 14$
- B $\mu_{A+B} = 177$ and $\sigma_{A+B} = 100$
- C $\mu_{A+B} = 177$ and $\sigma_{A+B} = 10$
- D $\mu_{A+B} \approx 125.26$ and $\sigma_{A+B} = 10$

Solution: C

The mean for the sum of two or more random variables is the sum of the respective means.

$$\mu_{A+B} = \mu_A + \mu_B$$

$$\mu_{A+B} = 85 + 92$$

$$\mu_{A+B} = 177$$

The variance for the sum of independent random variables is the sum of the respective variances. Finding the square root of that value gives us the standard deviation of the sum.

$$\sigma_{A+B} = \sqrt{\sigma_A^2 + \sigma_B^2}$$

$$\sigma_{A+B} = \sqrt{6^2 + 8^2}$$

$$\sigma_{A+B} = \sqrt{100}$$

$$\sigma_{A+B} = 10$$



Topic: Combinations of random variables

Question: Let X be the mass of a farm fresh egg. Assume X follows a normal distribution with $\mu_X = 64$ grams and $\sigma_X = 6$ grams. Suppose we package a dozen eggs into an egg carton, and the egg carton itself has a fixed mass of 40 grams. Let C be the mass of a carton full of a dozen farm fresh eggs. Find the mean and standard deviation for C .

Answer choices:

- A $\mu_C = 808$ grams and $\sigma_C \approx 20.78$ grams
- B $\mu_C = 768$ grams and $\sigma_C \approx 20.78$ grams
- C $\mu_C = 808$ grams and $\sigma_C \approx 72$ grams
- D $\mu_C = 808$ grams and $\sigma_C \approx 60.78$ grams



Solution: A

Let T be the total mass of the dozen eggs:

$$T = X_1 + X_2 + X_3 + \dots + X_{12}$$

Since the mass of the eggs follow a normal distribution with $\mu_X = 64$ grams and $\sigma_X = 6$ grams, the mean for the total mass of the dozen eggs can be found using the following formula:

$$\mu_T = \mu_{X_1} + \mu_{X_2} + \dots + \mu_{X_{12}}$$

$$\mu_T = 64 + 64 + \dots + 64$$

$$\mu_T = 12(64)$$

$$\mu_T = 768 \text{ grams}$$

The mean mass for the entire carton filled with eggs is

$$C = \mu_T + 40$$

$$C = 768 + 40$$

$$C = 808 \text{ grams}$$

The variance for the total mass of the dozen eggs is

$$\sigma^2_T = \sigma^2_{X_1} + \sigma^2_{X_2} + \dots + \sigma^2_{X_{12}}$$

$$\sigma^2_T = 6^2 + 6^2 + \dots + 6^2$$

$$\sigma^2_T = 12(6^2)$$

$$\sigma^2_T = 432$$

Which means the standard deviation is

$$\sqrt{\sigma^2}_T = \sqrt{432}$$

$$\sigma_T \approx 20.78 \text{ grams}$$

The standard deviation for the entire carton filled with eggs will not change when we add the fixed mass of the 40 grams since this is just a shift and does not change the variability in the overall mass of the carton of eggs.

$$\sigma_C \approx 20.78 \text{ grams}$$

Topic: Combinations of random variables

Question: Sandwiches can be purchased in the school cafeteria. Bread is baked each day and the sandwich is topped with meat and cheese, then sold for \$3.50.

The weight of the bread used for each sandwich is normally distributed with mean of 2.3 ounces and standard deviation of 0.4 ounces. The weight of the meat and cheese used for each sandwich is normally distributed with mean of 2.5 ounces and standard deviation of 0.6 ounces. Suppose you purchase a sandwich at random from the school cafeteria. What is the probability that the overall weight of the sandwich exceeds 6 ounces? Assume the two variables are independent.

Answer choices:

- A 0.0961
- B 0.0485
- C 0.1151
- D 0.0000

Solution: B

Let B be the weight of the bread and let M be the weight of the meat and cheese. Because we'll be creating a sandwich by combining the weights together, we'll let T be the total weight of the sandwich.

$$T = B + M$$

For bread, $\mu_B = 2.3$ and $\sigma_B = 0.4$, and for meat and cheese, $\mu_M = 2.5$ and $\sigma_M = 0.6$. The mean for the sum of two or more random variables is the sum of the respective means.

$$\mu_T = \mu_{B+M}$$

$$\mu_T = \mu_B + \mu_M$$

$$\mu_T = 2.3 + 2.5$$

$$\mu_T = 4.8 \text{ ounces}$$

The variance for the sum of independent random variables is the sum of the respective variances. Finding the square root of that value gives us the standard deviation of the sum.

$$\sigma_T = \sigma_{B+M}$$

$$\sigma_T = \sqrt{\sigma_B^2 + \sigma_M^2}$$

$$\sigma_T = \sqrt{(0.4)^2 + (0.6)^2}$$

$$\sigma_T = \sqrt{0.52}$$



$$\sigma_T \approx 0.7211 \text{ ounces}$$

Because B and M were both normally distributed, the new random variable T will also be normally distributed.

To find the probability that the overall weight of the sandwich exceeds 6 ounces, the value of 6 is converted to a z -score and then the normal model is used to find the corresponding area under the curve.

$$P(T > 6) = P\left(Z > \frac{6 - 4.8}{0.7211}\right)$$

$$P(T > 6) = P(Z > 1.66)$$

$$P(T > 6) \approx 0.0485$$



Topic: Permutations and combinations

Question: Out of 30 students in a math class, how many study groups of 5 students can be formed from the class members?

Answer choices:

- A 6 groups
- B 150 groups
- C 142,506 groups
- D 17,100,720 groups



Solution: C

This is a combination question where $n = 30$ and $k = 5$. The order in which we choose the 5 study group members doesn't matter in this situation.

If Person A, B, C, D , and E end up in a group together, this is equivalent to Person E, D, C, B , and A ending up in a group together.

We'll use the combination formula.

$${}_nC_k = \binom{n}{k} = \binom{30}{5} = \frac{n!}{k!(n-k)!} = \frac{30!}{5!25!} = 142,506 \text{ groups}$$

Topic: Permutations and combinations

Question: Four children are sledding in a toboggan. How many ways can the children arrange themselves on the toboggan?

Answer choices:

- A 4 ways
- B 16 ways
- C 24 ways
- D 256 ways

Solution: C

This is a permutation question. We have 4 people we're arranging and we'll arrange those 4 people as many different ways as we can. Set $n = 4$ and $k = 4$ and use the permutations formula.

$${}_nP_k = \frac{n!}{(n-k)!} = \frac{4!}{0!} = \frac{4!}{1} = 4! = (4)(3)(2)(1) = 24 \text{ ways}$$



Topic: Permutations and combinations

Question: Sawyer is taking a 5-question biology test, and the test only requires him to answer 3 out of the 5 questions. He gets to choose which 3 he answers. How many different ways could he choose exactly 3 of the 5 questions?

Answer choices:

- A 10 ways
- B 15 ways
- C 60 ways
- D 125 ways



Solution: A

To figure out how many different ways could Sawyer could answer exactly 3 of the 5 questions, we need the formula for combinations.

We have 5 questions and want to know how many ways we can pick 3 of the 5 questions. The order won't matter, which is why we need the combination, and not the permutation. For example, answering questions #1, #2, and #3 is the same as answering questions #2, #1, and #3.

Therefore, we find the combination ${}_5C_3$.

$${}_nC_k = \binom{n}{k} = \binom{5}{3} = \frac{n!}{k!(n-k)!} = \frac{5!}{3!2!} = 10 \text{ ways}$$

There are 10 different ways that Sawyer could answer exactly 3 of the 5 questions.



Topic: Binomial random variables**Question:** Which random variable X follows a binomial distribution?**Answer choices:**

- A X is the number of attempts it takes a basketball player to make a free throw
- B X is the amount of time it takes a runner to complete a marathon
- C X is the number of red cards you're dealt in a 5-card hand of poker
- D X is the number of times out of 10 tries that you roll a 4 on a six-sided die



Solution: D

In order for X to be a binomial random variable,

1. each trial must be independent,
2. each trial can be called a “success” or a “failure,”
3. there are a fixed number of trials, and
4. the probability of success on each trial is constant.

Answer choice A isn’t a binomial random variable because we don’t have a fixed number of trials.

Answer choice B isn’t a binomial random variable because there’s no success or failure, but rather a continuous numeric random variable.

Answer choice C isn’t a binomial random variable because the trials aren’t independent and the probability on each trial isn’t constant. As you draw cards out of a deck, the probability of drawing a red changes because the number of cards you’re drawing from decreases and your probability of getting a red card on any draw depends on what you were already dealt.

Answer choice D is a binomial random variable. A trial consists of rolling a die. Trials are independent when we roll a die because what we roll on each trial has no influence on what we’ll roll next. Rolling a 4 will be considered a success. There are a fixed number of trials, $n = 10$. And the probability of success on each trial remains constant at $p = 1/6$.

Topic: Binomial random variables

Question: Let X be a binomial random variable with $n = 15$ and $p = 0.45$.
Find $P(X = 9)$.

Answer choices:

- A 0.000006
- B 0.6
- C 0.1048
- D 6.75

Solution: C

The goal is to find the probability of exactly $k = 9$ successes.

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Start by finding the number of ways to have exactly 9 successes in 15 trials using the combination formula.

$$\binom{15}{9} = {}_{15}C_9 = \frac{15!}{9!(15-9)!} = 5,005$$

Now plug this value into the probability formula.

$$P(X = 9) = \binom{15}{9} (0.45)^9 (1 - 0.45)^6$$

$$P(X = 9) = (5,005)(0.45)^9 (1 - 0.45)^6$$

$$P(X = 9) \approx 0.1048$$



Topic: Binomial random variables

Question: Suppose 35 % of our nation's high school seniors will be taking at least one AP Exam this year. We select 80 students at random from our nation. What is the probability that exactly 30 will be taking at least one exam?

Answer choices:

- A 0.0824
- B 0.375
- C 0.1406
- D 0.35



Solution: A

Let X be the number of seniors that take at least one AP exam out of 80 trials.

X follows a binomial distribution with a trial representing choosing a random student from our nation and recording whether or not they're taking an AP Exam. These trials will be independent and the probability of success remains constant at $p = 0.35$. And there is a fixed number of trials, $n = 80$.

$$X \sim B(80, 0.35)$$

The goal is to find the probability of exactly $k = 30$ successes.

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} (p)^k (1 - p)^{n-k}$$

Start by finding the number of ways to have exactly 30 successes in 80 trials using the combination formula.

$$\binom{80}{30} = {}_{80}C_{30} = \frac{80!}{30!(80-30)!}$$

Now we can find the probability.

$$P(x = 30) = \binom{80}{30} (0.35)^{30} (1 - 0.35)^{50} \approx 0.0824$$

Topic: Poisson distributions

Question: A carpenter is able to build 3 chairs per day, on average. Find the probability that he can build 5 chairs tomorrow.

Answer choices:

- A $P(5) \approx 0.1008$
- B $P(5) \approx 0.1404$
- C $P(5) \approx 0.0136$
- D $P(5) \approx 0.2729$

Solution: A

We know this is a Poisson experiment with the following given values:

$\lambda = 3$, the average number of chairs built in a day

$x = 5$, the number of chairs required to be built tomorrow

The Poisson probability is

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$P(5) = \frac{3^5 e^{-3}}{5!}$$

$$P(5) \approx 0.1008$$

So the probability the carpenter will build five chairs tomorrow is approximately 0.1008 or 10.08 % .

Topic: Poisson distributions

Question: Let X be the number of typos on a page in a printed book, with a mean of 3 typos per page. What is the probability that a randomly selected page has at most one typo on it?

Answer choices:

- A $P(X \leq 1) \approx 0.0498$
- B $P(X \leq 1) \approx 0.1404$
- C $P(X \leq 1) \approx 0.1494$
- D $P(X \leq 1) \approx 0.1991$

Solution: D

The probability that there is at most one typo on the page is the probability that there are no typos on the page ($X = 0$), plus the probability that there is one typo on the page ($X = 1$).

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

$$P(X \leq 1) = \frac{3^0 e^{-3}}{0!} + \frac{3^1 e^{-3}}{1!}$$

$$P(X \leq 1) = e^{-3} + 3e^{-3}$$

$$P(X \leq 1) = 4e^{-3}$$

$$P(X \leq 1) \approx 0.1991$$

There is an approximately 0.1991 or 19.91 % chance of finding at most one typo on a randomly selected page, when the average number of typos per page is 3.

Topic: Poisson distributions

Question: There are 40 students in a college math course, and each one of them has a 4.5 % chance of forgetting their calculator on any given day. What is the probability that exactly 5 of them will forget their calculator today?

Answer choices:

- A $P(5) \approx 95.26\%$
- B $P(5) \approx 97.4\%$
- C $P(5) \approx 2.6\%$
- D $P(5) \approx 36.97\%$

Solution: C

This is a binomial experiment with $n = 40$, $p = 0.045$, and $x = 5$. Because we have at least 20 “attempts,” and because the probability of a “success” is less than 5 % , we can use the Poisson formula to estimate this binomial probability.

$$P(x) = \frac{(np)^x e^{-np}}{x!}$$

$$P(5) = \frac{(40 \cdot 0.045)^5 e^{-40 \cdot 0.045}}{5!}$$

$$P(5) = \frac{1.8^5 e^{-1.8}}{120}$$

$$P(5) \approx 0.0260$$

So the chance that exactly 5 of the college math students forget their calculator today is approximately 2.6 % .



Topic: "At least" and "at most," and mean, variance, and standard deviation

Question: Let X be a binomial random variable with $n = 15$ and $p = 0.45$. Find $P(X \leq 10)$.

Answer choices:

- A 0.0515
- B 0.9745
- C 0.9231
- D 0.0255

Solution: B

X follows a binomial distribution, but instead of finding the probability of exactly k successes in n trials, we're asked to find the probability of k or fewer successes in n trials. Specifically, find the chance of 10 or fewer successes in 15 trials, where the probability of success on any one trial is $p = 0.45$.

Find the probability of 0 success, 1 success, 2 successes, etc., up to 10 successes and then find the sum of those probabilities.

$$P(X \leq 10) = P(X = 0) + P(X = 1) + P(X = 2) + \dots + P(X = 10)$$

To find the probability for each value of k , we use the binomial probability formula.

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{n-k}$$

The probability is

$$\begin{aligned} P(X \leq 10) &= \binom{15}{0} (0.45)^0 (1 - 0.45)^{15} + \binom{15}{1} (0.45)^1 (1 - 0.45)^{14} \\ &\quad + \dots + \binom{15}{10} (0.45)^{10} (1 - 0.45)^5 \end{aligned}$$

$$P(X \leq 10) \approx 0.9745$$



Topic: "At least" and "at most," and mean, variance, and standard deviation

Question: 32 % of all internet users have Instagram accounts. Suppose 20 random internet users are selected. What's the probability that at least half of them have Instagram accounts?

Answer choices:

- A 0.0440
- B 0.9721
- C 0.0719
- D 0.0279

Solution: C

Let X be the number of Instagram users out of 20 internet users.

X follows a binomial distribution, but instead of finding the probability of exactly k successes in n trials, we're finding the probability of k or more successes in n trials. Specifically, we're finding the chance of 10 or more successes in 20 trials, where the probability of success on any one trial is $p = 0.32$.

We could find the probability of 10 success, 11 success, 12 successes, etc., up to 20 successes, and then find the sum of those probabilities. But it would be easier to find the complement, $P(X \leq 9)$. We'll use the formula

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{n-k}$$

and then calculate the probability of 0 successes, 1 success, 2 successes, etc., all the way up to 9 successes.

$$P(X \leq 9) = \binom{20}{0} (0.32)^0 (1 - 0.32)^{20} + \binom{20}{1} (0.32)^1 (1 - 0.32)^{19}$$

$$+ \dots + \binom{20}{9} (0.32)^9 (1 - 0.32)^{11} = 0.9281$$

$$P(X \leq 9) \approx 0.9281$$

And now find the probability of at least 10 successes out of 20 by subtracting $P(X \leq 9)$ from 1.

$$P(X \geq 10) = 1 - P(X \leq 9)$$

$$P(X \geq 10) = 1 - 0.9281$$

$$P(X \geq 10) = 0.0719$$



Topic: "At least" and "at most," and mean, variance, and standard deviation

Question: In January, 2018, 17% of teens said Instagram is the most important social media site. If we select an SRS of 150 teens, how many are expected to say Instagram is the most important social media site? With what standard deviation?

Answer choices:

- A $\mu = 17$ and $\sigma = 2.55$
- B $\mu = 25.5$ and $\sigma = 21.165$
- C $\mu = 25.5$ and $\sigma = 8.824$
- D $\mu = 25.5$ and $\sigma = 4.601$

Solution: D

Let X be the number of teens who say Instagram is the most important social media site.

X follows a binomial distribution, with a fixed number of trials $n = 150$ and a probability of success $p = 0.17$. We can find the expected value for the distribution.

$$\mu_X = E(X) = np$$

$$\mu_X = (150)(0.17)$$

$$\mu_X = 25.5 \approx 26 \text{ teens}$$

Find the variance for the distribution.

$$\sigma_X^2 = Var(X) = np(1 - p)$$

$$\sigma_X^2 = (150)(0.17)(1 - 0.17)$$

$$\sigma_X^2 = 21.165$$

Then we can find standard deviation.

$$\sigma_X = \sqrt{21.165}$$

$$\sigma_X \approx 4.601 \text{ teens}$$



Topic: Bernoulli random variables**Question:** Which of the following is not an example of Bernoulli trials?**Answer choices:**

- A Flip a coin and observe whether it landed on heads or tails
- B Randomly choose restaurants and observe whether or not they have a children's menu
- C Select cards from a standard 52-card deck without replacement and observe whether or not each card is a face card
- D Buy lottery tickets and observe whether or not they paid out any money

Solution: C

In order for trials to be Bernoulli trials,

- each trial must be independent,
- each trial can be called a “success” or a “failure,” and
- the probability of success on each trial is constant.

Answer choices A, B, and D are all examples of Bernoulli trials because the trials are independent, we can observe the outcome of each trial as a “success” or “failure,” and the probability of each trial remains constant.

Answer choice C is not an example of Bernoulli trials because, while we can observe the outcome of each trial as a “success” or “failure,” the probability of success does not remain constant and the trials are not independent. When drawing cards out of a deck without replacement, our trials are dependent.

Topic: Bernoulli random variables

Question: Suppose we're conducting Bernoulli trials with a probability of success $p = 0.345$. If we conduct three trials in a row, find the probability that all three are failures.

Answer choices:

- A 0.655
- B 0.0411
- C 1.965
- D 0.2810

Solution: D

If each trial has a probability of success of $p = 0.345$, then the probability of failure on any one trial is

$$1 - p$$

$$1 - 0.345$$

$$0.655$$

If we want three trials in a row to be failures and our trials are independent, we can multiply the probabilities to find the chance of three failures in a row.

$$P(3 \text{ failures}) = (0.655)^3$$

$$P(3 \text{ failures}) \approx 0.2810$$

Topic: Bernoulli random variables

Question: 68 % of U.S. households own a pet. Suppose we start randomly surveying households and asking whether or not they are pet owners. When a household is selected, we can consider it a Bernoulli trial because each trial is independent, a success can be noted as finding a household with a pet, and the probability of success remains constant for each trial. What is the mean and standard deviation for each trial?

Answer choices:

- A $\mu = 0.68$ and $\sigma = 0.2176$
- B $\mu = 0.68$ and $\sigma = 0.4665$
- C $\mu = 0.68$ and $\sigma = 0.32$
- D $\mu = 0.68$ and $\sigma = 6.8$

Solution: B

Each trial will result in either a success or a failure (either the household owns a pet, or it doesn't). But the average can be found for each trial as $\mu = p$. So for these Bernoulli trials, $\mu = 0.68$.

The standard deviation is

$$\sigma = \sqrt{p(1 - p)}$$

So for these Bernoulli trials,

$$\sigma = \sqrt{0.68(1 - 0.68)}$$

$$\sigma = \sqrt{0.68(0.32)}$$

$$\sigma = \sqrt{0.2176}$$

$$\sigma \approx 0.4665$$



Topic: Geometric random variables

Question: Which of the following random variables follow a geometric distribution?

Answer choices:

- A X is the number of attempts it takes a baseball player to get a hit
- B X is the amount of time it takes a runner to complete a marathon
- C X is the number of red cards you're dealt in a 5-card hand of poker
- D X is the number of times out of 10 tries that you roll a 4 on a six-sided die



Solution: A

In order for X to be a geometric random variable,

- each trial must be independent,
- each trial can be called a “success” or a “failure,” and
- the probability of success on each trial is constant.

Answer choice A is a geometric random variable if we assume each attempt at a hit is a trial, and that these trials are independent. Consider a hit to be a success and anything else to be a failure. And assume the probability of a hit is constant.

Answer choice B is not a geometric random variable because there's no success or failure, but rather a continuous numeric random variable.

Answer choice C is not a geometric random variable because the trials are not independent and the probability of success on each trial is not constant. As you draw cards out of a deck, the probability of drawing a red card changes as the number of cards you're drawing from decreases and your probability of getting a red card on any draw depends on what you were already dealt.

Answer choice D is a binomial random variable. A trial consists of rolling a die. Trials are independent when we roll a die because what we roll on each trial has no influence on what we'll roll next. Rolling a 4 will be considered a success. There are a fixed number of trials, $n = 10$. And the probability of success on each trial remains constant at $p = 1/6$.



Topic: Geometric random variables**Question:** Let X be a geometric random variable with $p = 0.30$. Find $P(X = 4)$.**Answer choices:**

- A 0.0081
- B 0.1029
- C 1.2
- D 0.0720

Solution: B

Let X be the trial where we get the first success. X follows a geometric distribution, and we want to find the probability of getting our first success on exactly the 4th trial, knowing the chance of success on any trial is $p = 0.30$.

To find the probability that a success S occurs on the n th trial, when a success has a probability of p , and therefore failure has a probability of $1 - p$, we'll use

$$P(S = n) = p(1 - p)^{n-1}$$

In this case, we'll set $n = 4$ and get

$$P(S = 4) = 0.30(1 - 0.30)^{4-1}$$

$$P(S = 4) = 0.1029$$

Topic: Geometric random variables

Question: Suppose 35 % of our nation's high school seniors will be taking at least one AP Exam this year. Suppose we select students at random and ask them if they'll be taking an AP Exam. What's the probability that we'll need to ask exactly 3 people to find someone who is taking the exam?

Answer choices:

- A 0.0429
- B 0.2389
- C 0.1479
- D 1.05

Solution: C

Let X be the trial when we find our first person taking at least one AP Exam. X follows a geometric distribution with a trial representing choosing a random student and recording whether he's taking an AP Exam or not. These trials will be independent and the probability of success remains constant at $p = 0.35$. There is not a fixed number of trials.

To find the probability that a success S occurs on the n th trial when a success has a probability of p , and therefore failure has a probability of $1 - p$, we'll use

$$P(S = n) = p(1 - p)^{n-1}$$

In this case, we'll set $n = 3$ and get

$$P(S = 3) = 0.35(1 - 0.35)^{3-1}$$

$$P(S = 3) \approx 0.1479$$

Topic: Types of studies

Question: In statistics, there are observational and experimental studies. Which of the following studies represents an observational study?

Answer choices:

- A A hospital conducting a study measuring the effectiveness of a new drug places people into two groups. One group is the control group and is given a placebo. The other group is given the new drug and results are compared.
- B A scientist conducting a study on a new mattress type and quality of sleep places people into two groups where one group sleeps on a standard spring mattress and the other group sleeps on the new mattress. The sleep cycles of each group are recorded and compared.
- C A class conducts a survey that asks students to record their height and shoe size. The class creates a chart of the results and analyzes it for correlation.
- D An educational researcher studying how praise affects student success groups students into three categories for the school year. One group is given no praise on their academics, the next group is only told that they are smart if they do well in the class, and the



final group is praised on how hard they have worked to achieve academic success.

Solution: C

An observational study analyzes information that's already there, while an experimental study manipulates what's happening to try to establish causality (people are put into at least two different groups and the results are compared).

A class comparing height and shoe size is not manipulating any information and just showing the correlation of height and shoe size, so it's an observational study.

The other three choices are types of experimental studies because people are placed into two or more groups so that one or more groups can be manipulated and the results can be analyzed.



Topic: Types of studies

Question: A study conducted in the Indian Ocean records the total number of plastic objects found in the ocean each year for 10 years. Classify the study as observational or experimental and the data collected in the study as data that forms a one-way or two-way table.

Answer choices:

- A Observational and one-way
- B Observational and two-way
- C Experimental and one-way
- D Experimental and two-way

Solution: A

The study on the total number of plastic objects found in the Indian Ocean is an observational study since we're observing the number of objects and not manipulating the data to find causality. The data is suitable for building a one-way table because we're only analyzing one variable, the number of plastic objects.

An observational and two-way study would be analyzing two variables, an experimental and one-way study can exist if you analyze one variable within the data, and an experimental and two-way study would be a study between two groups with one acting as a control group.

Topic: Types of studies

Question: Which of the following studies represents a matched pairs experiment?

Answer choices:

- A A study recording the gender of Galapagos tortoises worldwide.
- B A study analyzing the effects of one cup of coffee per day on blood pressure where people are placed into three groups. In the first group no one drinks coffee, in the second group people drink one cup of coffee per day, and in the third group they drink two or more cups of coffee per day. After two months blood pressure is recorded and compared to initial blood pressure.
- C A study comparing the pollution emissions on five models of SUV vehicles.
- D A study of the blood pressure medicine where the control group and experimental group only includes women of ages 40 – 65 years old.



Solution: D

The study of the blood pressure medicine on women ages 40 – 65 years old is an example of matched pairs experiment since the control group and treatment group consist of the same gender and age range. This controls the variables of gender and age.



Topic: Sampling and bias

Question: A restaurant wants to know how the new chef's special is received. The owner of the restaurant polls 20 patrons on whether they enjoyed the special by asking the question, "Did you enjoy our chef's new delicious special today?" What type of bias does the sample have?

Answer choices:

- A Measurement bias
- B Response bias
- C Leading questions
- D Selection bias

Solution: C

The sample has leading question biased. Since the words “enjoy” and “delicious” have been included in the question, people will feel more pressure to answer in favor of the new chef’s special.

Instead, we could have asked, “What do you think about the new chef’s special?” This wording does not lead people to answer in a certain way.



Topic: Sampling and bias

Question: A political poll conducts a survey on current presidential approval in the United States. Which choice would be the best representative sample of the US population?

Answer choices:

- A The poll asks 100 people in Alabama if they approve of the president.
- B The poll asks 100,000 people across the United States if they approve of the president.
- C The poll asks 50,000 people in the eastern United States if they approve of the president.
- D The poll asks 100 people across the United States if they approve of the president.

Solution: B

A poll that asks 100,000 people across the United States will be the best representative sample, because it's large and it surveys people across the country.

Answer choices A and C are only surveying people in narrow regions of the country. If we want a representative sample of the United States, then all 50 states should be included.

Answer choice D doesn't survey enough people to give a clear representation of the total population, even though it's surveying people in all 50 states.



Topic: Sampling and bias

Question: The state of Colorado conducts a study on the GPA of college students throughout the state. They randomly sample GPA for an equal number of students within the majors of English, Science, Mathematics, Business, and Arts. Which sampling technique was used in the experiment?

Answer choices:

- A Simple random sample
- B Stratified random sample
- C Clustered random sample
- D Voluntary random sample

Solution: B

The study used a stratified random sample since it divided the population into groups according to type of major, and then sampled within those groups.

A simple random sample would have been a random collection of GPAs with no regard to type of major. A clustered random sample would have divided the population into clusters with a mix of majors in each cluster. A voluntary random sample would have been a sample of individuals who voluntarily respond to a general survey asking about their GPA.



Topic: Sampling distribution of the sample mean

Question: If the original population is normally distributed, then the sampling distribution of the sample mean...

Answer choices:

- A will also be normally distributed, if the sample size is large enough.
- B will also be normally distributed, regardless of the sample size.
- C will not be normally distributed.
- D will have an unknown shape.



Solution: B

The Central Limit Theorem tells us that, if the original population is normally distributed, then the SDSM will also be normally distributed, regardless of the sample size n that we use.

If the original population is not normally distributed, or if we don't know the shape of the population distribution, then the SDSM is only guaranteed to be normally distributed when we use a sample size of at least $n = 30$.

Topic: Sampling distribution of the sample mean

Question: A hospital finds that the average birth weight of a newborn is 7.5 lbs with a standard deviation of 0.4 lbs. What is the standard deviation of the sampling distribution, if the hospital randomly selects 45 newborns to test this claim?

Answer choices:

- A $\sigma_{\bar{x}} = 0.0596$
- B $\sigma_{\bar{x}} = 1.118$
- C $\sigma_{\bar{x}} = 0.0533$
- D $\sigma_{\bar{x}} = 0.0089$



Solution: A

To find the standard deviation of the sampling distribution, we'll plug population standard deviation $\sigma = 0.4$ and the sample size $n = 45$ into the formula for the standard error.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{0.4}{\sqrt{45}}$$

$$\sigma_{\bar{x}} \approx 0.0596$$

Topic: Sampling distribution of the sample mean

Question: A group of marathon runners have the following finishing times in hours: 3.2, 3.5, 3.8, 4.2, 4.5. Given a sample of size 2 if we're sampling with replacement, find the standard error $\sigma_{\bar{x}}$.

Answer choices:

- A $\sigma_{\bar{x}} \approx 0.2104$
- B $\sigma_{\bar{x}} \approx 0.2504$
- C $\sigma_{\bar{x}} \approx 0.2904$
- D $\sigma_{\bar{x}} \approx 0.3304$

Solution: D

Let's first determine the total number of possible samples, using N^n , given $N = 5$ and $n = 2$.

$$N^n = 5^2 = 25$$

The complete sample space, and the mean for each sample, is

Sample (Sample mean)				
3.2, 3.2 (3.20)	3.5, 3.2 (3.35)	3.8, 3.2 (3.50)	4.2, 3.2 (3.70)	4.5, 3.2 (3.85)
3.2, 3.5 (3.35)	3.5, 3.5 (3.50)	3.8, 3.5 (3.65)	4.2, 3.5 (3.85)	4.5, 3.5 (4.00)
3.2, 3.8 (3.50)	3.5, 3.8 (3.65)	3.8, 3.8 (3.80)	4.2, 3.8 (4.00)	4.5, 3.8 (4.15)
3.2, 4.2 (3.70)	3.5, 4.2 (3.85)	3.8, 4.2 (4.00)	4.2, 4.2 (4.20)	4.5, 4.2 (4.35)
3.2, 4.5 (3.85)	3.5, 4.5 (4.00)	3.8, 4.5 (4.15)	4.2, 4.5 (4.35)	4.5, 4.5 (4.50)

Build a table for the probability distribution of the sample mean. Because there are 25 total samples, the probability of each sample mean will be given by the number of times that sample mean occurs, divided by the total number of possible samples, so “count/25.”

Sample mean	$P(x_i)$
3.20	1/25
3.35	2/25
3.50	3/25
3.65	2/25
3.70	2/25
3.80	1/25
3.85	4/25
4.00	4/25
4.15	2/25
4.20	1/25
4.35	2/25
4.50	1/25

Now we can calculate the mean of the sampling distribution of the sample mean, $\mu_{\bar{x}}$, where \bar{x}_i is a given sample mean, $P(\bar{x}_i)$ is the probability of that particular sample mean occurring, and N is the number of samples.

$$\mu_{\bar{x}} = \sum_{i=1}^N \bar{x}_i P(\bar{x}_i)$$

$$\mu_{\bar{x}} = 3.20 \left(\frac{1}{25} \right) + 3.35 \left(\frac{2}{25} \right) + 3.50 \left(\frac{3}{25} \right) + 3.65 \left(\frac{2}{25} \right) + 3.70 \left(\frac{2}{25} \right)$$

$$+ 3.80 \left(\frac{1}{25} \right) + 3.85 \left(\frac{4}{25} \right) + 4.00 \left(\frac{4}{25} \right) + 4.15 \left(\frac{2}{25} \right)$$

$$+4.20\left(\frac{1}{25}\right) + 4.35\left(\frac{2}{25}\right) + 4.50\left(\frac{1}{25}\right)$$

$$\begin{aligned}\mu_{\bar{x}} = \frac{3.20}{25} + \frac{6.70}{25} + \frac{10.50}{25} + \frac{7.30}{25} + \frac{7.40}{25} + \frac{3.80}{25} + \frac{15.40}{25} \\ + \frac{16.00}{25} + \frac{8.30}{25} + \frac{4.20}{25} + \frac{8.70}{25} + \frac{4.50}{25}\end{aligned}$$

$$\mu_{\bar{x}} = \frac{96}{25}$$

$$\mu_{\bar{x}} = 3.84$$

Because we're sampling with replacement, we would expect this mean of the SDSM to be equivalent to the mean of the population, $\mu_{\bar{x}} = \mu$, and we can see that it is if we calculate the mean of the population.

$$\mu = \frac{3.2 + 3.5 + 3.8 + 4.2 + 4.5}{5} = \frac{19.2}{5} = 3.84$$

Both means are $\mu_{\bar{x}} = \mu = 3.84$. The population variance is

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

$$\sigma^2 = \frac{(3.2 - 3.84)^2 + (3.5 - 3.84)^2 + (3.8 - 3.84)^2 + (4.2 - 3.84)^2 + (4.5 - 3.84)^2}{5}$$

$$\sigma^2 = \frac{(-0.64)^2 + (-0.34)^2 + (-0.04)^2 + 0.36^2 + 0.66^2}{5}$$

$$\sigma^2 = \frac{0.4096 + 0.1156 + 0.0016 + 0.1296 + 0.4356}{5}$$

$$\sigma^2 = \frac{1.092}{5}$$

$$\sigma^2 = 0.2184$$

which means that the population standard deviation is

$$\sigma = \sqrt{0.2184}$$

$$\sigma \approx 0.4673$$

Because we're sampling with replacement, we can use the simplified formula for standard error (the one *without* the FPC).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{0.4673}{\sqrt{2}}$$

$$\sigma_{\bar{x}} \approx 0.3304$$



Topic: Conditions for inference with the SDSM

Question: A school finds that time spent studying for a test isn't normally distributed for their 2,800 students. They would like to use the Central Limit Theorem to normalize the data. Which sample allows them to use the Central Limit Theorem?

Answer choices:

- A The school samples 300 different students randomly
- B The school samples 200 different students randomly
- C The school samples 200 different students from an Honors classes
- D The school samples 25 different students randomly

Solution: B

The random sample of 200 different students (“different” tells us that we’re sampling without replacement) will allow the school to use the Central Limit Theorem. The sample is random, less than 10% of the population, and greater than 30, so it’s large enough to normalize the data.

The other answer choices wouldn’t allow the school to use the Central Limit Theorem. Answer choice A samples more than 10% of the population which doesn’t maintain independence, answer choice C isn’t random because selecting students in Honors classes will skew the data, and answer choice D doesn’t have a big enough sample because a sample size smaller than 30 won’t normalize the data.



Topic: Conditions for inference with the SDSM

Question: A company produces tires in a factory. Individual tires are filled to an approximate pressure of 36 PSI (pounds per square inch), with a standard deviation of 0.8 PSI. The pressure in the tires is normally distributed. The company randomly selects 125 tires to check their pressure. What is the probability that the mean pressure in the tires is within 0.1 PSI of the population mean?

Answer choices:

- A 8.38 %
- B 91.62 %
- C 71.55 %
- D 83.84 %

Solution: D

To verify normality, our sample must be random, no more than 10 % of the population, and (if the population is not normal) the sample size must be greater than 30.

The sample was collected randomly. It's safe to assume that 125 tires is less than 10 % of the total tires produced in the factory. The population is normal, so the sample size doesn't have to be greater than 30, but 125 is greater than 30 anyway. The sample space meets the conditions of normality.

Find the standard deviation of the sampling distribution.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{0.8}{\sqrt{125}}$$

$$\sigma_{\bar{x}} \approx 0.07155$$

We want to know the probability that the sample mean \bar{x} is within 0.1 PSI of the population mean. We need to express 0.1 in terms of standard deviations.

$$\frac{0.1}{0.07155} \approx 1.39762 \approx 1.40$$

This means we want to know the probability $P(-1.40 < z < 1.40)$. Using a z -table, a z -value of -1.40 gives 0.0808 and a z -value of 1.40 gives 0.9192. The probability under the normal curve between these z -scores is



$$P(-1.40 < z < 1.40) = 0.9192 - 0.0808$$

$$P(-1.40 < z < 1.40) = 0.8384$$

$$P(-1.40 < z < 1.40) = 83.84 \%$$

There's an 83.84 % chance that our sample mean will fall within 0.1 PSI of the population mean of 36 PSI.



Topic: Conditions for inference with the SDSM

Question: A large cookie company knows that the weight of their tins of Christmas cookies is normally distributed with a mean weight of 1 pound and a standard deviation of 0.2 pounds. If they take a random 50-tin sample, what is the probability that the sample mean \bar{x} is within 0.05 pounds of the population mean?

Answer choices:

- A 94.16 %
- B 92.32 %
- C 89.14 %
- D 84.98 %

Solution: B

We were told that the sample was taken randomly. Our sample size is $n \geq 30$. We're sampling without replacement, but we can safely assume that this large cookie company makes more than 500 Christmas cookie tins. Therefore, we've met the random, normal, and independence conditions, respectively.

To answer the probability question, we'll start by finding the mean of the SDSM, but we know it'll be equal to the population mean, so $\mu = \mu_{\bar{x}} = 1$. The standard error will be

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{0.2}{\sqrt{50}}$$

$$\sigma_{\bar{x}} \approx 0.028$$

We want to know the probability that the sample mean \bar{x} is within 0.05 pounds of the population mean, 1. A 0.05 interval around 1 gives us the interval 0.95 to 1.05, so

$$P(0.95 < \bar{x} < 1.05) \approx P\left(\frac{0.95 - 1}{0.028} < z < \frac{1.05 - 1}{0.028}\right)$$

$$P(0.95 < \bar{x} < 1.05) \approx P\left(-\frac{0.05}{0.028} < z < \frac{0.05}{0.028}\right)$$

$$P(0.95 < \bar{x} < 1.05) \approx P(-1.77 < z < 1.77)$$

Which means we want to know the probability of $P(-1.77 < z < 1.77)$. In a z -table, a z -value of 1.77 gives 0.9616,

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706

and a z -value of -1.77 gives 0.0384.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455

Which means the probability under the normal curve between these z -scores is

$$P(0.95 < \bar{x} < 1.05) \approx 0.9616 - 0.0384$$

$$P(0.95 < \bar{x} < 1.05) \approx 0.9232$$

$$P(0.95 < \bar{x} < 1.05) \approx 92.32\%$$

So there's an approximately 92.32% chance that the mean \bar{x} of the 50-tin sample the company takes will fall within 0.05 pounds of the population mean of $\mu = 1$ pound.

Topic: Sampling distribution of the sample proportion

Question: In the case of a population proportion, the original population will be modeled by...

Answer choices:

- A a binomial distribution.
- B a normal distribution.
- C a skewed distribution.
- D a uniform distribution.

Solution: A

For any proportion, a response is always classified as either a “success” or a “failure.” Because exactly two outcomes are possible, the probability distribution for a proportion is always binomial.



Topic: Sampling distribution of the sample proportion

Question: A population proportion is $p = 0.7$. Find the standard error of the proportion for samples of size $n = 100$.

Answer choices:

A $\sigma_{\hat{p}} \approx 0.0337$

B $\sigma_{\hat{p}} \approx 0.0458$

C $\sigma_{\hat{p}} \approx 0.0548$

D $\sigma_{\hat{p}} \approx 0.0837$

Solution: B

The standard deviation of the sampling distribution of the sample proportion $\sigma_{\hat{p}}$, also called the standard error of the proportion, is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.7(1-0.7)}{100}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.7(0.3)}{100}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.21}{100}}$$

$$\sigma_{\hat{p}} \approx \frac{0.4583}{10}$$

$$\sigma_{\hat{p}} \approx 0.0458$$

Topic: Sampling distribution of the sample proportion

Question: A group of 3 siblings have the following eye color: blue, blue, green. Find the mean $\mu_{\hat{p}}$ and standard error $\sigma_{\hat{p}}$ of the sampling distribution of the sample proportion for the proportion of siblings with blue eyes, if we take 2-sibling samples, with replacement.

Answer choices:

- A $\mu_{\hat{p}} = 1/3$ and $\sigma_{\hat{p}} = 1/3$
- B $\mu_{\hat{p}} = 1/3$ and $\sigma_{\hat{p}} = 2/3$
- C $\mu_{\hat{p}} = 2/3$ and $\sigma_{\hat{p}} = 1/3$
- D $\mu_{\hat{p}} = 2/3$ and $\sigma_{\hat{p}} = 2/3$

Solution: C

Determine the total number of possible samples, using N^n , given $N = 3$ and $n = 2$.

$$N^n = 3^2 = 9$$

The complete sample space, and the proportion for each sample, is

Sample	Sample proportion
blue, blue	1
blue, blue	1
blue, green	1/2
blue, blue	1
blue, blue	1
blue, green	1/2
green, blue	1/2
green, blue	1/2
green, green	0

Build a table for the probability distribution of the sample proportion. Because there are 9 total samples, the probability of each sample proportion will be given by the number of times that sample proportion occurs, divided by the total number of possible samples, so “count/9.”

Sample proportion	P(p_i)
0	1/9
1/2	4/9
1	4/9

Now we can calculate the mean of the sampling distribution of the sample proportion, $\mu_{\hat{p}}$, where \hat{p}_i is a given sample proportion, $P(\hat{p}_i)$ is the probability of that particular sample proportion occurring, and N is the number of samples.

$$\mu_{\hat{p}} = \sum_{i=1}^N \hat{p}_i P(\hat{p}_i)$$

$$\mu_{\hat{p}} = 0 \left(\frac{1}{9} \right) + \frac{1}{2} \left(\frac{4}{9} \right) + 1 \left(\frac{4}{9} \right)$$

$$\mu_{\hat{p}} = \frac{2}{9} + \frac{4}{9}$$

$$\mu_{\hat{p}} = \frac{6}{9}$$

$$\mu_{\hat{p}} = \frac{2}{3}$$

Because we're sampling with replacement, we would expect this mean of the SDSP to be equivalent to the population proportion, $\mu_{\hat{p}} = p$, and we can see that it is if we calculate the population proportion.

$$p = \frac{2 \text{ people with blue eyes}}{3 \text{ people in the population}} = \frac{2}{3}$$



Both proportions are $\mu_{\hat{p}} = p = 2/3$. The variance of the SDSP would be

$$\sigma_{\hat{p}}^2 = \sum_{i=1}^N (\hat{p}_i - p)^2 P(\hat{p}_i)$$

$$\sigma_{\hat{p}}^2 = \left(0 - \frac{2}{3}\right)^2 \left(\frac{1}{9}\right) + \left(\frac{1}{2} - \frac{2}{3}\right)^2 \left(\frac{4}{9}\right) + \left(1 - \frac{2}{3}\right)^2 \left(\frac{4}{9}\right)$$

$$\sigma_{\hat{p}}^2 = \left(-\frac{2}{3}\right)^2 \left(\frac{1}{9}\right) + \left(-\frac{1}{6}\right)^2 \left(\frac{4}{9}\right) + \left(\frac{1}{3}\right)^2 \left(\frac{4}{9}\right)$$

$$\sigma_{\hat{p}}^2 = \left(\frac{4}{9}\right) \left(\frac{1}{9}\right) + \left(\frac{1}{36}\right) \left(\frac{4}{9}\right) + \left(\frac{1}{9}\right) \left(\frac{4}{9}\right)$$

$$\sigma_{\hat{p}}^2 = \frac{4}{81} + \frac{1}{81} + \frac{4}{81}$$

$$\sigma_{\hat{p}}^2 = \frac{9}{81}$$

$$\sigma_{\hat{p}}^2 = \frac{1}{9}$$

and then the standard error would be

$$\sigma_{\hat{p}} = \sqrt{\frac{1}{9}}$$

$$\sigma_{\hat{p}} = \frac{1}{3}$$

Topic: Conditions for inference with the SDSP

Question: A math class wants to know how many of the 2,000 students in their school carry a blue backpack. Which sample meets all conditions of a normal sampling distribution?

Answer choices:

- A The class randomly surveys 250 students during lunch
- B The class surveys 100 students from freshmen classes
- C The class surveys 500 students from senior classes
- D The class randomly surveys 100 students during passing period



Solution: D

Surveying 100 students randomly during a passing period is a valid sampling distribution because it's random and keeps the number of subjects in the sample below 10 % , which maintains independence.

The other choices either don't keep the sample below 10 % , and/or aren't random because they're surveying only freshmen or seniors.



Topic: Conditions for inference with the SDSP

Question: A restaurant wants to know the percentage of their customers who order desert. The restaurant has 1,500 customers in one week and finds by randomly surveying 100 of them that 35 order desert. What is the standard error of the SDSP?

Answer choices:

- A $\sigma_{\hat{p}} \approx 0.047697$
- B $\sigma_{\hat{p}} \approx 0.015096$
- C $\sigma_{\hat{p}} \approx 0.052303$
- D $\sigma_{\hat{p}} \approx 0.084900$

Solution: A

To verify normality, our sample space should be random, no more than 10% of the population, and the expected number of successes and failures should each be at least 5.

$$\text{Independence: } \frac{100}{1,500} = 0.067 = 6.7\% \leq 10\%$$

$$\text{Successes: } 100(0.35) = 35 \geq 5$$

$$\text{Failures: } 100(0.65) = 65 \geq 5$$

The sample space was random, so we've met the conditions of normality. The standard error, or sampling distribution of the sample proportion, is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.35(1 - 0.35)}{100}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.35(0.65)}{100}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.2275}{100}}$$

$$\sigma_{\hat{p}} \approx 0.047697$$



Topic: Conditions for inference with the SDSP

Question: A group of scientists is studying 10,000 manatees and finds that 20 % are calves. We want to verify their claim, but can't conduct a study of all 10,000, so we randomly sample just 500. What's the probability that our results are within 5 % of the scientists' study?

Answer choices:

- A 13.5 %
- B 73.72 %
- C 99.48 %
- D 99.8 %

Solution: C

To verify normality, our sample space should be random, no more than 10% of the population, and the expected number of successes and failures should each be at least 5.

$$\text{Independence: } \frac{500}{10,000} = 0.05 = 5\% \leq 10\%$$

$$\text{Successes: } 500(0.2) = 100 \geq 5$$

$$\text{Failures: } 500(0.8) = 400 \geq 5$$

The sample space was random, so we've met the conditions of normality. Now we'll find the mean and standard deviation of the sampling distribution of the sample proportion.

$$\mu_{\hat{p}} = p = 0.2$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.2(0.8)}{500}} \approx 0.0179$$

We need to find the probability that our results are within 5% of the population proportion $p = 20\%$. In other words, how likely is it that the sample proportion falls between 15% and 25%?

$$\frac{0.05}{0.0179} \approx 2.79$$

We want to know the probability of $P(-2.79 < z < 2.79)$. Using a z -table, -2.79 gives us 0.0026



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036

and 2.79 gives us 0.9974.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981

So the probability we want to find is

$$P(-2.79 < z < 2.79) = 0.9974 - 0.0026$$

$$P(-2.79 < z < 2.79) = 0.9948$$

There's a 99.48 % chance that our sample proportion will fall within 5 % of the first study's claim.

Topic: The student's t-distribution

Question: Compared to the standard normal distribution, the *t*-distribution for $n < 30$ is...

Answer choices:

- A flatter and wider
- B flatter and narrower
- C taller and wider
- D taller and narrower



Solution: A

For smaller samples ($n < 30$), the student's t -distribution is flatter and wider than the standard normal distribution (z -distribution).

As a result, the standard deviation is larger, because there's more area in the tails of the t -distribution.



Topic: The student's t-distribution

Question: What value in the t -table is associated with a sample size $n = 15$ and an upper-tail probability of 0.025?

Answer choices:

- A 2.131
- B 2.145
- C 2.602
- D 2.624

Solution: B

A sample size $n = 15$ is associated with 14 degrees of freedom.

If we locate 14 degrees of freedom down the left side of the t -table, and that row's intersection with a 0.025 upper-tail probability, the t -table returns 2.145.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Topic: The student's t-distribution

Question: Find the upper-tail probability associated with a confidence level of 99 % .

Answer choices:

- A 0.05
- B 0.025
- C 0.1
- D 0.005

Solution: D

At a 99 % confidence level,

$$\alpha = 1 - 0.99$$

$$\alpha = 0.01$$

The entire region of rejection will constitute 1 % of the area under the distribution. But this area includes both the upper and lower tails, and we want to find just upper-tail probability.

So we split the alpha value in half, and the upper-tail probability associated with a 99 % confidence level will be

$$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$$



Topic: Confidence interval for the mean

Question: The height of students in our school is normally distributed with a standard deviation of $\sigma = 4$ inches. We sample 50 of our classmates (with replacement) and get a sample mean of $\bar{x} = 66$ inches. What is the confidence interval for a confidence level of 95 % ?

Answer choices:

- A $(a, b) \approx (64.54, 67.46)$
- B $(a, b) \approx (64.89, 67.11)$
- C $(a, b) \approx (65.07, 66.93)$
- D $(a, b) \approx (65.74, 66.26)$

Solution: B

A 95% confidence level is associated with z -scores of $z = \pm 1.96$.

If we plug everything we know into the confidence interval formula for a known population standard deviation, we get

$$(a, b) = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

$$(a, b) = 66 \pm 1.96 \cdot \frac{4}{\sqrt{50}}$$

$$(a, b) \approx 66 \pm 1.1087$$

Therefore, we can say that the confidence interval is

$$(a, b) \approx (66 - 1.1087, 66 + 1.1087)$$

$$(a, b) \approx (64.8913, 67.1087)$$

$$(a, b) \approx (64.89, 67.11)$$

We could also express this as the sample mean plus or minus the margin of error, 66 ± 1.1087 inches. We're 95% certain that the actual mean height of students in our school is between 64.89 inches and 67.11 inches.



Topic: Confidence interval for the mean

Question: The weight of chickens on a farm is normally distributed with a standard deviation of $\sigma = 3.5$ ounces. What is the smallest sample we can take if we want a margin of error of ± 2.5 ounces, and we want to be 99 % confident?

Answer choices:

- A $n = 10$ chickens
- B $n = 13$ chickens
- C $n = 14$ chickens
- D $n = 30$ chickens

Solution: C

Solve the margin of error formula for n .

$$ME = z^* \frac{\sigma}{\sqrt{n}}$$

$$ME\sqrt{n} = z^*\sigma$$

$$\sqrt{n} = \frac{z^*\sigma}{ME}$$

$$n = \left(\frac{z^*\sigma}{ME} \right)^2$$

Now we can plug the values we were given into this equation, remembering that a confidence level of 99 % is associated with critical values of $z = \pm 2.58$.

$$n = \left(\frac{2.58 \cdot 3.5}{2.5} \right)^2$$

$$n \approx 13.05$$

Because we can't sample 0.05 of a chicken, we round up to $n = 14$ chickens. Then we can say that, to meet that threshold, and keep a margin of error of ± 2.5 at 99 % confidence, we'd need to take a sample size of at least $n = 14$ chickens.



Topic: Confidence interval for the mean

Question: We want to know the mean number of daylight hours (the time between sunrise and sunset) in a day in our city over the course of a year. We take a random sample of 30 days throughout the year and get a sample mean of $\bar{x} = 13.15$ hours and a sample standard deviation of $s = 0.85$ hours. What is the confidence interval for a confidence level of 90 % ?

Answer choices:

- A $(a, b) \approx (12.89, 13.41)$
- B $(a, b) \approx (12.85, 13.45)$
- C $(a, b) \approx (12.75, 13.55)$
- D $(a, b) \approx (12.28, 14.02)$

Solution: A

Because population standard deviation is unknown, we have to use the t -distribution instead of the z -distribution.

A 90% confidence level with $n - 1 = 30 - 1 = 29$ degrees of freedom is associated with t -scores of $t = \pm 1.699$.

$$(a, b) = \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$$(a, b) = 13.15 \pm 1.699 \cdot \frac{0.85}{\sqrt{30}}$$

$$(a, b) \approx 13.15 \pm 0.2637$$

Therefore, we can say that the confidence interval is

$$(a, b) \approx (13.15 - 0.2637, 13.15 + 0.2637)$$

$$(a, b) \approx (12.8863, 13.4137)$$

$$(a, b) \approx (12.89, 13.41)$$

We could also express this as the sample mean plus or minus the margin of error, 13.15 ± 0.2637 hours. We're 90% certain that the actual population mean of hours of daylight in a day is between 12.89 hours and 13.41 hours.

Topic: Confidence interval for the proportion

Question: A study shows that 78 % of patients who try a new medication for migraines feel better within 30 minutes of taking the medicine. If the study involved 120 patients, construct a 95 % confidence interval for the proportion of patients who feel better within 30 minutes of taking the medicine.

Answer choices:

- A $(a, b) \approx (0.73, 0.83)$
- B $(a, b) \approx (0.72, 0.84)$
- C $(a, b) \approx (0.71, 0.85)$
- D $(a, b) \approx (0.68, 0.88)$

Solution: C

We know that the sample proportion is $\hat{p} = 0.78$, and that the confidence level is 95 %. The critical value for this confidence level is $z^* = 1.96$. The sample size is $n = 120$. So we can plug these values into the confidence interval formula.

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$(a, b) = 0.78 \pm 1.96 \sqrt{\frac{0.78(1 - 0.78)}{120}}$$

$$(a, b) \approx 0.78 \pm 0.0741$$

$$(a, b) \approx (0.78 - 0.0741, 0.78 + 0.0741)$$

$$(a, b) \approx (0.71, 0.85)$$

This means that we're 95 % confident that the proportion of patients who feel better within 30 minutes of taking the medicine is between 71 % and 85 % .

Topic: Confidence interval for the proportion

Question: A study shows that 243 of 500 randomly selected households were using a family member to care for their young children. Build a 90 % confidence interval for the proportion of households using a family member to care for young children.

Answer choices:

- A $(a, b) \approx (0.45, 0.52)$
- B $(a, b) \approx (0.44, 0.53)$
- C $(a, b) \approx (0.43, 0.54)$
- D $(a, b) \approx (0.42, 0.55)$

Solution: A

The population proportion is

$$\hat{p} = \frac{243}{500} = 0.486$$

With $\hat{p} = 0.486$, a confidence level of 90 % and therefore a critical value of $z^* = 1.65$, and a sample size of $n = 500$, the confidence interval will be

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$(a, b) = 0.486 \pm 1.65 \sqrt{\frac{0.486(1 - 0.486)}{500}}$$

$$(a, b) \approx 0.486 \pm 0.0369$$

$$(a, b) \approx (0.486 - 0.0369, 0.486 + 0.0369)$$

$$(a, b) \approx (0.45, 0.52)$$

This means that we're 90 % confident that the proportion of households using a family member to care for young children is between 45 % and 52 %.

Topic: Confidence interval for the proportion

Question: Bentley is a dog who helps police find drugs, but he has a low success rate. In Bentley's time on the job, it's estimated that he correctly identified drugs 59 % of the time. How many different trials would the police need to put him through to verify that this is his actual success rate, at a 95 % confidence level and with a margin of error of 0.05?

Answer choices:

A $n = 264$

B $n = 372$

C $n = 413$

D $n = 645$

Solution: B

The minimum sample size n that gives a fixed margin of error is given by

$$n = \left(\frac{z^* \sqrt{p(1-p)}}{ME} \right)^2$$

Bentley's success rate is $p = 59\% = 0.59$, the confidence level is 95%, and the critical value for this confidence level is $z^* = 1.96$. The margin of error is $ME = 0.05$. If we plug these values into the formula, we can find the minimum number of trials we need to verify Bentley's success rate.

$$n = \left(\frac{1.96 \sqrt{0.59(1-0.59)}}{0.05} \right)^2$$

$$n = \left(\frac{1.96 \sqrt{0.59(0.41)}}{0.05} \right)^2$$

$$n = \left(\frac{1.96 \sqrt{0.2419}}{0.05} \right)^2$$

$$n \approx 371.71$$

Since we need more than 371 trials to test Bentley, we have to round up to 372 trials, so we can say $n = 372$.



Topic: Inferential statistics and hypotheses**Question:** What is the null hypothesis in a test procedure?**Answer choices:**

- A The claim for which we're trying to provide support
- B The claim we're trying to reject in order to provide support for the opposite claim
- C A step that we can skip because it's empty
- D The claim that the test characteristic is greater than the hypothesized value



Solution: B

The null hypothesis in a statistical hypothesis test or test procedure is the claim that's assumed to be true. In order to provide statistical support for some claim, we want to show that our data is outside of the null hypothesis. This means the point of the statistical test is to reject the null hypothesis in order to provide support for the opposite claim.

In answer choice A, the claim for which we're trying to provide support is our alternative hypothesis.

In answer choice B, the claim we're trying to reject in order to provide support for the opposite claim (the alternative hypothesis) is the null hypothesis.

Answer choice C is a false statement. We can't skip the step where we write our null and alternative hypotheses, because we need these hypothesis statements in order to know clearly what we're trying to test.

Answer choice D is an example of a possible alternative hypothesis.

Topic: Inferential statistics and hypotheses**Question:** Which statement is not a possible alternative hypothesis?**Answer choices:**

- A population parameter > hypothesized value
- B population parameter < hypothesized value
- C population parameter \neq hypothesized value
- D population parameter = hypothesized value



Solution: D

The null hypothesis always takes one of the following forms,

H_0 : population parameter = hypothesized value

H_0 : population parameter \geq hypothesized value

H_0 : population parameter \leq hypothesized value

and the associated alternative hypotheses are

H_a : population parameter \neq hypothesized value

H_a : population parameter $<$ hypothesized value

H_a : population parameter $>$ hypothesized value

respectively. Answer choices A, B, and C are all possible alternative hypotheses, which means D is the only answer choice that can't represent an alternative hypothesis.

Topic: Inferential statistics and hypotheses

Question: A researcher thinks that a new drug will decrease the number of people who experience severe symptoms of asthma. Currently, 286 out of every 2,000 asthma patients experience severe symptoms. Which pair of hypothesis statements shows a correct way to set up the hypothesis test?

Answer choices:

- A $H_0 : \mu = 286$ and $H_a : \mu \neq 286$
- B $H_0 : p = 286$ and $H_a : p \neq 286$
- C $H_0 : p = 0.143$ and $H_a : p > 0.143$
- D $H_0 : p \geq 0.143$ and $H_a : p < 0.143$

Solution: D

The proportion of people who experience severe asthma symptoms is

$$p = \frac{x}{N}$$

$$p = \frac{286}{2,000}$$

$$p = 0.143$$

Since the researcher believes that the drug will decrease the proportion of people who experience severe symptoms, his alternative hypothesis will be

$$H_a : p < 0.143$$

which means his null hypothesis is

$$H_0 : p \geq 0.143$$

Topic: Significance level and type I and II errors**Question:** As the alpha level gets lower, which error rate also gets lower?**Answer choices:**

- A The Type I error rate
- B The Type II error rate
- C Both the Type I and Type II error rates
- D Neither the Type I nor Type II error rates



Solution: A

The Type I error rate is the α level. The lower the alpha level, the lower the Type I error rate.



Topic: Significance level and type I and II errors

Question: It's been shown many times that on a certain memory test, recognition (recognizing something familiar) is substantially better than recall (pulling something from memory). WE run our own test but fail to reject the null hypothesis that recall and recognition produce the same results. What type of error did we make?

Answer choices:

- A Type I error
- B Type II error
- C Both a Type I and Type II error
- D Neither a Type I nor Type II error

Solution: B

There's a difference in the population between recognition and recall, but we didn't find a significant difference in our sample. Which means the null hypothesis was false, but we failed to reject it. Failing to reject a false null hypothesis is a Type II error.



Topic: Significance level and type I and II errors

Question: In a population, there's no difference between men and women on a certain test. However, we found a significant difference in our sample and therefore rejected the null hypothesis. What type of error did we make?

Answer choices:

- A Type I error
- B Type II error
- C Both a Type I and Type II error
- D Neither a Type I nor Type II error



Solution: A

There's no difference in the population, but we found a difference in our sample. Which means the null hypothesis was true, but we rejected it, thinking it was false. Rejecting a true null hypothesis is a Type I error.



Topic: Test statistics for one- and two-tailed tests

Question: If we're conducting a two-tailed test, what are the signs in the null and alternative hypotheses?

Answer choices:

- A = and \neq
- B \leq and $>$
- C \geq and $<$
- D \leq and \geq

Solution: A

When the null and alternative hypotheses use the = and \neq signs, we'll use a two-tailed test.

When using a two-tailed test, we need to use the null hypothesis of no difference, =, and an alternative hypothesis that states there is a difference between the population parameter and the hypothesized value, \neq .



Topic: Test statistics for one- and two-tailed tests

Question: Consider the hypothesis, “Reading more about statistics changes the reader’s desire to learn about statistics.” What type of test would be used to investigate this hypothesis?

Answer choices:

- A One-tailed test
- B Two-tailed test
- C Both a one- and two-tailed test
- D Neither a one- nor two-tailed test

Solution: B

The hypothesis is non-directional because the hypothesis stated that the desire to learn statistics will change after reading more about statistics, but it doesn't specifically say whether that desire will specifically increase or decrease.

If we tested this hypothesis statistically, the test would be a two-tailed test, in which the null hypothesis would include an $=$ sign, and the alternative hypothesis would include a \neq sign.



Topic: Test statistics for one- and two-tailed tests

Question: We've decided to give all of our friends a small box of homemade cookies. We've already baked a variety of cookies and randomly placed them into boxes. We want to make sure that each box is close to 0.5 pounds, so we sample 10 boxes and find a mean of 0.54 pounds and a standard deviation of $s = 0.3$ pounds. Assuming that the weights of all the boxes are normally distributed, calculate the test statistic.

Answer choices:

- A $t \approx -0.42$
- B $t \approx 0.42$
- C $t \approx -0.58$
- D $t \approx 0.58$

Solution: B

Because population standard deviation is unknown (we were only given sample standard deviation), and because we were told that we can assume the population is normally distributed, the test statistic will be

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{0.54 - 0.5}{\frac{0.3}{\sqrt{10}}}$$

$$t = \frac{0.04\sqrt{10}}{0.3}$$

$$t \approx 0.42$$

Topic: The p-value and rejecting the null

Question: Which pair of p -value and significance level would lead us to reject the null hypothesis of the test?

Answer choices:

- A A lower-tailed test with $p = 0.002$ and $\alpha = 0.001$
- B An upper-tailed test with $p = 0.925$ and $\alpha = 0.95$
- C A two-tailed test with $p = 0.07$ and $\alpha = 0.05$
- D A lower-tailed test with $p = 0.085$ and $\alpha = 0.05$

Solution: B

We reject, or fail to reject, the null hypothesis based on the relationship between the p -value and the α level, regardless of the type of test.

If $p \leq \alpha$, we reject the null hypothesis

If $p > \alpha$, we don't reject the null hypothesis

In answer choice A with $p = 0.002$ and $\alpha = 0.001$, $0.002 > 0.001$ so $p > \alpha$, which means we fail to reject the null hypothesis.

In answer choice B with $p = 0.925$ and $\alpha = 0.95$, $0.925 < 0.95$ so $p \leq \alpha$, which means we reject the null hypothesis.

In answer choice C with $p = 0.07$ and $\alpha = 0.05$, $0.07 > 0.05$ so $p > \alpha$, which means we fail to reject the null hypothesis.

In answer choice D with $p = 0.085$ and $\alpha = 0.05$, $0.085 > 0.05$ so $p > \alpha$, which means we fail to reject the null hypothesis.

Topic: The p-value and rejecting the null

Question: The smaller the *p*-value...

Answer choices:

- A the more significant the result.
- B the less likely it is that we found this result purely by chance.
- C the more likely we are to reject the null hypothesis.
- D All of these



Solution: D

The smaller the p -value is in a statistical significance test, the more likely we are to reject the null hypothesis and make a claim that the alternative hypothesis is true.

If we find a very smaller p -value, it means it was unlikely that we obtained our result by chance, which means the conclusion is significant at a higher level.



Topic: The p-value and rejecting the null

Question: If we're running an upper-tailed test and find $p = 0.0643$, what is the z -value that gives the boundary between the region of acceptance and the region of rejection?

Answer choices:

- A $z = -1.85$
- B $z = -1.52$
- C $z = 1.52$
- D $z = 1.85$

Solution: C

In an upper-tailed test, the entire region of rejection will lie in the upper tail, with the region of acceptance (non-rejection region) to the left of the region of rejection.

Which means the full $p = 0.0643$ will lie in the upper tail. If we subtract this value from 1, we'll get the value that we'll be looking for in the body of the z -table.

$$1 - 0.0643$$

$$0.9357$$

If we look for 0.9357 in the body of the z -table, we find $z = 1.52$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545

Topic: Hypothesis testing for the population proportion

Question: We've heard that 10% of people are left-handed. We want to verify this claim, so we collect a random sample of 500 people and find that 43 of them are left-handed. What can we conclude at a significance level of $\alpha = 0.10$?

Answer choices:

- A We'll reject the null hypothesis; our result is significant at the $\alpha = 0.10$ level
- B We'll reject the null hypothesis; our result is significant at the $\alpha = 0.05$ level
- C We'll reject the null hypothesis; our result is significant at the $\alpha = 0.01$ level
- D We'll fail to reject the null hypothesis

Solution: D

First, build the hypothesis statements.

$$H_0: 10\% \text{ of people are left-handed, } p = 0.1$$

$$H_a: \text{The proportion of left-handed people is different than } 10\%, \\ p \neq 0.1$$

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{43}{500} = 0.086$$

Then the standard error of the proportion is

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.1(1 - 0.1)}{500}} \approx 0.0134$$

Now we have enough to find the z -test statistic.

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.086 - 0.10}{0.0134} \approx -1.04$$

The critical z -values for 90% confidence with a two-tailed test are $z = \pm 1.65$. Since the test statistic we found is negative ($z = -1.04$), we'll compare it to $z = -1.65$.

Our z -value of $z = -1.04$ is not less than $z = -1.65$, and therefore falls in the region of acceptance, which means we'll fail to reject the null hypothesis and fail to conclude that the proportion of left-handed people is different than 10%.



Topic: Hypothesis testing for the population proportion

Question: A breakfast company claims at least 80% of Americans eat breakfast. We want to verify this claim, so we collect a random sample of 650 Americans and find that 496 of them eat breakfast. What can we conclude at a significance level of $\alpha = 0.05$?

Answer choices:

- A We'll reject the null hypothesis; our result is significant at the $p = 0.0094$ level
- B We'll reject the null hypothesis; our result is significant at the $p = 0.9500$ level
- C We'll reject the null hypothesis; our result is significant at the $p = 0.9864$ level
- D We'll fail to reject the null hypothesis



Solution: A

First, build the hypothesis statements.

H_0 : At least 80% of Americans eat breakfast, $p \geq 0.8$

H_a : Fewer than 80% of Americans eat breakfast, $p < 0.8$

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{496}{650} \approx 0.7631$$

and the standard error of the proportion is

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.8(1 - 0.8)}{650}} \approx 0.0157$$

Now we have enough to find the z -test statistic.

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.7631 - 0.8}{0.0157} \approx -2.35$$

The critical z -value for 95% confidence with a lower-tailed test is $z = -1.65$.

Our z -value of $z \approx -2.35$ falls to the left of $z = -1.65$, and therefore falls in the region of rejection, which means we'll reject the null hypothesis and conclude that the proportion of Americans who eat breakfast is less than 80%.

We know our findings are significant at $\alpha = 0.05$, but we can find the p -value to state a higher level of significance that corresponds to $z \approx -2.35$

and not just $z = -1.65$. The test statistic $z \approx -2.35$ gives a value of 0.0094 in the z -table.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110

Which means the conclusion isn't only significant at $\alpha = 0.05$, but it's actually significant at 0.0094. As long as $\alpha \geq 0.0094$, we'll be able to reject H_0 .

Topic: Hypothesis testing for the population proportion

Question: The NBA (National Basketball Association) claims that no more than 25% of NBA players started playing basketball before age 5. We want to verify this claim, so we collect a random sample of 117 NBA players and find that 34 of them started playing before age 5. What can we conclude at a significance level of $\alpha = 0.01$?

Answer choices:

- A We'll reject the null hypothesis; our result is significant at the $\alpha = 0.01$ level
- B We'll reject the null hypothesis; our result is significant at the $\alpha = 0.05$ level
- C We'll reject the null hypothesis; our result is significant at the $\alpha = 0.10$ level
- D We'll fail to reject the null hypothesis

Solution: D

First, build the hypothesis statements.

H_0 : At most 25 % of NBA players started playing before 5, $p \leq 0.25$

H_a : More than 25 % of NBA players started playing before 5, $p > 0.25$

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{34}{117} \approx 0.2906$$

and the standard error of the proportion is

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.25(1 - 0.25)}{117}} \approx 0.0400$$

Now we have enough to find the z -test statistic.

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.2906 - 0.25}{0.0400} \approx 1.0142$$

The critical z -value for 99 % confidence with an upper-tailed test is $z = 2.33$.

Our z -value of $z = 1.015$ falls to the left of $z = 2.33$, and therefore falls in the region of acceptance, which means we'll fail to reject the null hypothesis and fail to conclude that the proportion of NBA players who started playing basketball before age 5 is more than 25 %.



Topic: Confidence interval for the difference of means

Question: A professor is interested in whether exam scores differ between two nearby colleges. He selects a simple random sample of 20 students each from both colleges and finds a mean test score of 350 with a standard deviation of 15 at the first college, and a mean test score of 390 with a standard deviation of 30 at the second college. Assuming exam scores are normally distributed at both colleges, find a 95 % confidence interval around the difference in exam scores.

Answer choices:

- A $(-55.52, -24.48)$
- B $(-55.39, -24.61)$
- C $(24.64, 55.36)$
- D $(24.61, 55.39)$

Solution: D

We don't know the population standard deviations, and the sample sizes are smaller than 30. The sample variances are $s_1^2 = 30^2 = 900$ and $s_2^2 = 15^2 = 225$. We assign the populations this way so that we end up with a positive result for the difference of means.

Not only were the samples taken from different populations, but 900 is more than double 225, so we can conclude that we're working with unequal population variances.

Therefore, our confidence interval formula will be

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{with } df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)^2}$$

Let's start with degrees of freedom.

$$df = \frac{\left(\frac{900}{20} + \frac{225}{20} \right)^2}{\frac{1}{20 - 1} \left(\frac{900}{20} \right)^2 + \frac{1}{20 - 1} \left(\frac{225}{20} \right)^2}$$



$$df = \frac{\left(\frac{1,125}{20}\right)^2}{\frac{1}{19} \left(\frac{900}{20}\right)^2 + \frac{1}{19} \left(\frac{225}{20}\right)^2}$$

$$df = \frac{\frac{1,265,625}{400}}{\frac{810,000}{7,600} + \frac{50,625}{7,600}}$$

$$df = \frac{1,265,625}{400} \left(\frac{7,600}{860,625} \right)$$

$$df \approx 27.94$$

Rounding down to the nearest whole number in order to keep the estimate conservative, we find 27 degrees of freedom. Together with a 95% confidence level, the t -table gives $t_{\alpha/2} = 2.052$.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Then the confidence interval will be

$$(a, b) = (390 - 350) \pm 2.052 \sqrt{\frac{30^2}{20} + \frac{15^2}{20}}$$

$$(a, b) = 40 \pm 2.052 \sqrt{\frac{900}{20} + \frac{225}{20}}$$



$$(a, b) = 40 \pm 2.052 \sqrt{\frac{1,125}{20}}$$

$$(a, b) = 40 \pm 2.052(7.5)$$

$$(a, b) = 40 \pm 15.39$$

Therefore, we can say that the confidence interval is

$$(a, b) = (40 - 15.39, 40 + 15.39)$$

$$(a, b) = (24.61, 55.39)$$

So we can say that we're 95 % confident that the true difference between mean exam scores is between 24.61 and 55.39.



Topic: Confidence interval for the difference of means

Question: Two college directors want to determine whether there's a difference in the amount that their students spend annually on textbooks. They sample 200 students from college A and 230 from college B and find mean spends of $\bar{x}_A = \$1,258$ and $\bar{x}_B = \$1,150$. Assuming both populations are normally distributed with $\sigma_A = \$52$ and $\sigma_B = \$64$, find a 99 % confidence interval around the difference in annual textbook spend.

Answer choices:

- A $(-122.44, -93.56)$
- B $(97.03, 118.97)$
- C $(93.56, 122.44)$
- D $(93.35, 122.65)$

Solution: C

Because population standard deviations are known, our confidence interval formula will be

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

At 99 % confidence, we use a critical z -value of $z_{\alpha/2} = 2.58$. Then the confidence interval will be

$$(a, b) = (1,258 - 1,150) \pm 2.58 \sqrt{\frac{52^2}{200} + \frac{64^2}{230}}$$

$$(a, b) = 108 \pm 2.58 \sqrt{\frac{2,704}{200} + \frac{4,096}{230}}$$

$$(a, b) = 108 \pm 2.58 \sqrt{\frac{338}{25} + \frac{2,048}{115}}$$

$$(a, b) = 108 \pm 2.58 \sqrt{\frac{18,014}{575}}$$

$$(a, b) \approx 108 \pm 2.58(5.597)$$

$$(a, b) \approx 108 \pm 14.441$$

Therefore, we can say that the confidence interval is

$$(a, b) \approx (108 - 14.441, 108 + 14.441)$$

$$(a, b) \approx (93.559, 122.441)$$

So we can say that we're 99 % confident that the true difference between mean textbook spend is between 93.559 and 122.441.



Topic: Confidence interval for the difference of means

Question: The owners of two restaurants on the same street are interested in whether or not their daily earnings differ. They take simple random samples of earnings over 15 days, and find mean daily earnings of \$1,365 with a standard deviation of \$48 for the first restaurant, and mean daily earnings of \$1,230 with a standard deviation of \$28 for the second restaurant. Assuming daily earnings at both restaurants follow a normal distribution, find a 90 % confidence interval around the difference in daily earnings.

Answer choices:

- A (111.33,158.67)
- B (134.38,135.62)
- C (116.16,153.84)
- D (110.36,159.64)

Solution: D

Because population standard deviations are unknown, and because we have small samples $n_1, n_2 < 30$, we'll need to use a critical t -value instead of a critical z -value.

Our samples were taken from different populations, and one sample variance is more than twice the other, $48^2 = 2,304 > 2(28^2) = 2(784) = 1,568$, so we'll say that the population variances are unequal.

Find the number of degrees of freedom.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

$$df = \frac{\left(\frac{48^2}{15} + \frac{28^2}{15}\right)^2}{\frac{1}{15 - 1} \left(\frac{48^2}{15}\right)^2 + \frac{1}{15 - 1} \left(\frac{28^2}{15}\right)^2}$$

$$df = \frac{\left(\frac{48 \cdot 48 + 28 \cdot 28}{15}\right)^2}{\frac{1}{14} \left(\frac{48 \cdot 48 \cdot 48 \cdot 48}{15 \cdot 15}\right) + \frac{1}{14} \left(\frac{28 \cdot 28 \cdot 28 \cdot 28}{15 \cdot 15}\right)}$$

$$df = \frac{\frac{(48 \cdot 48 + 28 \cdot 28)(48 \cdot 48 + 28 \cdot 28)}{15 \cdot 15}}{\frac{24 \cdot 48 \cdot 48 \cdot 48}{7 \cdot 15 \cdot 15} + \frac{2 \cdot 28 \cdot 28 \cdot 28}{15 \cdot 15}}$$

$$df = \frac{\frac{(48 \cdot 48 + 28 \cdot 28)(48 \cdot 48 + 28 \cdot 28)}{15 \cdot 15}}{\frac{24 \cdot 48 \cdot 48 \cdot 48 + 2 \cdot 7 \cdot 28 \cdot 28 \cdot 28}{7 \cdot 15 \cdot 15}}$$

$$df = \frac{(48 \cdot 48 + 28 \cdot 28)(48 \cdot 48 + 28 \cdot 28)}{15 \cdot 15} \left(\frac{7 \cdot 15 \cdot 15}{24 \cdot 48 \cdot 48 \cdot 48 + 2 \cdot 7 \cdot 28 \cdot 28 \cdot 28} \right)$$

$$df = \frac{7(48 \cdot 48 + 28 \cdot 28)(48 \cdot 48 + 28 \cdot 28)}{24 \cdot 48 \cdot 48 \cdot 48 + 2 \cdot 7 \cdot 28 \cdot 28 \cdot 28}$$

$$df = \frac{7(48^4 + 2(28^2 \cdot 48^2) + 28^4)}{24(48^3) + 2(7)(28^3)}$$

$$df \approx 22.54$$

Rounding down to the nearest whole number in order to keep the estimate conservative, we find 22 degrees of freedom. Together with a 90% confidence level, the t -table gives $t_{\alpha/2} = 1.717$.

	Upper-tail probability p									
df	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Then the confidence interval will be

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(a, b) = (1,365 - 1,230) \pm 1.717 \sqrt{\frac{48^2}{15} + \frac{28^2}{15}}$$

$$(a, b) = 135 \pm 1.717 \sqrt{\frac{2,304}{15} + \frac{784}{15}}$$

$$(a, b) = 135 \pm 1.717 \sqrt{\frac{3,088}{15}}$$

$$(a, b) \approx 135 \pm 24.64$$

Therefore, we can say that the confidence interval is

$$(a, b) \approx (135 - 24.64, 135 + 24.64)$$

$$(a, b) \approx (110.36, 159.64)$$

Based on the confidence interval, we're 90 % confident that the true difference between the mean daily earnings of the two restaurants is between \$110.36 and \$159.64.

Topic: Hypothesis testing for the difference of means

Question: Two math teachers, Mr. Johnson and Mr. Adams, want to determine whose students performed better on a recent exam. Mr. Johnson sampled 40 of his students and found a mean score of 84 with the standard deviation of 3.5, while Mr. Adams sampled 38 of his students and found a mean score of 82 with the standard deviation of 3.9. Assuming the exam scores are normally distributed, what can they conclude at 0.01 level of significance?

Answer choices:

- A They reject the null; the result is significant at $\alpha = 0.01$.
- B They reject the null; the result isn't significant at $\alpha = 0.01$.
- C They fail to reject the null; the result is significant at $\alpha = 0.01$.
- D They fail to reject the null; the result isn't significant at $\alpha = 0.01$.

Solution: D

The teachers are looking for a difference in exam scores, without any suspicion about the direction of the difference, so they'll use a two-tailed test, and their hypothesis statements will be

$$H_0 : \mu_J - \mu_A = 0; \text{ the exam scores don't differ significantly}$$

$$H_a : \mu_J - \mu_A \neq 0; \text{ the exam scores differ significantly}$$

The samples are large, $n_1 \geq 30$ and $n_2 \geq 30$, and neither population variance is more than twice the other, so we'll assume equal population variances. Therefore, we'll start by calculating pooled variance.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(40 - 1)3.5^2 + (38 - 1)3.9^2}{40 + 38 - 2}}$$

$$s_p = \sqrt{\frac{39(12.25) + 37(15.21)}{76}}$$

$$s_p \approx 3.70$$

Then the test statistic will be

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



Because $\mu_1 - \mu_2 = 0$, the test statistic formula simplifies.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$z = \frac{84 - 82}{3.70 \sqrt{\frac{1}{40} + \frac{1}{38}}}$$

$$z = \frac{2}{3.70 \sqrt{\frac{1}{40} + \frac{1}{38}}}$$

$$z \approx 2.39$$

We find $z \approx 2.39$ in the z -table, and we get 0.9916.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

The p -value is therefore $p = 2(1 - 0.9916) = 0.0168$, and we have $p = 0.0168 > \alpha = 0.01$, so the teachers will fail to reject the null hypothesis. They can't conclude that the mean exam scores differ.

Topic: Hypothesis testing for the difference of means

Question: An engine manufacturer claims their new engine consumes at least 0.5 gallons less gasoline per 100 miles, compared to older engines. An independent auditing company wants to test this claim and randomly selects 25 cars with new engines and 25 cars with old engines and finds that the new engine consumes 2.95 gallons of gasoline per 100 miles with a standard deviation of 0.14, while the old engine consumes 3.24 gallons of gasoline per 100 miles with a standard deviation of 0.19. What can the auditing company conclude at a 0.05 level of significance?

Answer choices:

- A They fail to reject the null; the result is significant at $\alpha = 0.05$.
- B They fail to reject the null; the result is not significant at $\alpha = 0.05$.
- C They reject the null; the result is significant at $\alpha = 0.05$.
- D They reject the null; the result is not significant at $\alpha = 0.05$.

Solution: B

Because the manufacturer suspects that the new engine consumes “at least 0.5 gallons less gasoline,” they’ll use a one-tailed test (specifically a lower-tailed test), and their hypothesis statements will be

$H_0 : \mu_1 - \mu_2 \geq -0.5$; the new engine doesn’t consume at least 0.5 gallons less gasoline

$H_a : \mu_1 - \mu_2 < -0.5$; the new engine consumes at least 0.5 gallons less gasoline

Neither sample variance is more than twice the other, so we can assume that the population variances are equal. Therefore, with small samples $n_1, n_2 < 30$, pooled variance is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(25 - 1)0.14^2 + (25 - 1)0.19^2}{25 + 25 - 2}}$$

$$s_p = \sqrt{\frac{24(0.0196) + 24(0.0361)}{48}}$$

$$s_p = \sqrt{\frac{0.0196 + 0.0361}{2}}$$

$$s_p = \sqrt{\frac{0.0557}{2}}$$



$$s_p \approx 0.167$$

Then the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{(2.95 - 3.24) - (-0.5)}{0.167 \sqrt{\frac{1}{25} + \frac{1}{25}}}$$

$$t \approx \frac{0.21}{0.167(0.2828)}$$

$$t \approx 4.45$$

The degrees of freedom is

$$df = n_1 + n_2 - 2$$

$$df = 25 + 25 - 2$$

$$df = 48$$

From the t -table, $df = 48$ at a 0.05 level of significance for a lower-tailed test, we get -1.677 . The t -value we found would need to be less than (more negative than) -1.677 , so the auditing company fails to verify the claim of the manufacturer.



Topic: Hypothesis testing for the difference of means

Question: A pollster wants to determine whether male financial directors earn more than female financial directors. It's known that the standard deviations of mean monthly salaries for men and women are \$459 and \$430, respectively. The pollster samples 35 male directors and 35 female directors and found that mean monthly salaries were \$8,504 and \$7,845, respectively. What can the pollster conclude at a 0.10 level of significance?

Answer choices:

- A They fail to reject the null; the result is significant at $\alpha = 0.10$.
- B They fail to reject the null; the result isn't significant at $\alpha = 0.10$.
- C They reject the null; the result is significant at $\alpha = 0.10$.
- D They reject the null; the result isn't significant at $\alpha = 0.10$.

Solution: C

The pollster believes that male financial directors earn more than female directors, so they'll use an upper-tailed test and their hypothesis statements will be

$$H_0 : \mu_1 - \mu_2 \leq 0; \text{ men don't earn more than women}$$

$$H_a : \mu_1 - \mu_2 > 0; \text{ men earn more than women}$$

The samples are large $n_1, n_2 \geq 30$, and neither population variance is more than twice the other, so we'll calculate pooled variance as

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(35 - 1)459^2 + (35 - 1)430^2}{35 + 35 - 2}}$$

$$s_p = \sqrt{\frac{34(459^2) + 34(430^2)}{68}}$$

$$s_p \approx 444.74$$

and then the test statistic will be

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



$$z = \frac{8,504 - 7,845}{444.74\sqrt{\frac{1}{35} + \frac{1}{35}}}$$

$$z \approx \frac{659}{444.74(0.239)}$$

$$z \approx 6.20$$

This z -value is well above the largest value in the z -table, so the pollster will reject the null hypothesis and conclude that the mean monthly salary for men is greater than the mean monthly salary for women.

Topic: Matched-pair hypothesis testing**Question:** Which samples are used in a matched-pair test?**Answer choices:**

- A Independent samples
- B Dependent samples
- C Both independent and dependent samples
- D Neither independent nor dependent samples

Solution: B

Matched-pair tests are conducted with dependent samples.

Dependent samples are related to one another in the sense that they contain the same subjects, and each subject produces a value for each sample.

One common type of matched-pair test is “change over time” or “before and after,” where each subject in the test produces a “before” value for the first sample and an “after” value for the second sample.



Topic: Matched-pair hypothesis testing

Question: A test prep company believes their new SAT prep program will improve learners SAT scores by at least 150 points. How should they define their populations and write their hypothesis statements?

Answer choices:

- A Population 1 will be the SAT scores of students after they complete the prep program, and Population 2 will be the SAT scores of students before they complete the prep program. Then the hypothesis statements are $H_0 : \mu_2 - \mu_1 > 150$ and $H_a : \mu_2 - \mu_1 \leq 150$.
- B Population 1 will be the SAT scores of students before they complete the prep program, and Population 2 will be the SAT scores of students after they complete the prep program. Then the hypothesis statements are $H_0 : \mu_2 - \mu_1 > 150$ and $H_a : \mu_2 - \mu_1 \leq 150$.
- C Population 1 will be the SAT scores of students after they complete the prep program, and Population 2 will be the SAT scores of students before they complete the prep program. Then the hypothesis statements are $H_0 : \mu_2 - \mu_1 \leq 150$ and $H_a : \mu_2 - \mu_1 > 150$.
- D Population 1 will be the SAT scores of students before they complete the prep program, and Population 2 will be the SAT scores of students after they complete the prep program. Then the hypothesis statements are $H_0 : \mu_2 - \mu_1 \leq 150$ and $H_a : \mu_2 - \mu_1 > 150$.



Solution: D

The test prep company will define the SAT scores of students before they've completed the prep program as Population 1, and define the SAT scores of students after they've completed the prep program as Population 2.

Then their null and alternative hypothesis will be

$$H_0 : \mu_2 - \mu_1 \leq 150$$

$$H_a : \mu_2 - \mu_1 > 150$$

where μ_1 is the mean SAT score before the students complete the prep program, and μ_2 is the mean SAT score after the students complete the prep program.

Topic: Matched-pair hypothesis testing

Question: A pharmaceutical company believes their new weight-loss drug produces more weight loss than the current market-leader, which produces a mean monthly weight loss of 10 pounds. They record the before and after weights of 10 people who volunteer to try the new drug.

Participant	1	2	3	4	5	6	7	8	9	10
Before x_1	150	157	206	344	193	188	168	272	245	222
After x_2	142	145	198	330	191	185	150	252	233	206
Difference, d	8	12	8	14	2	3	18	20	12	16
d^2	64	144	64	196	4	9	324	400	144	256

Can the company conclude at 5 % significance that their drug works better than the market-leading weight loss drug?

Answer choices:

- A Yes, they can conclude that the new drug works better
- B No, they can't conclude that the new drug works better

Solution: B

The pharmaceutical company will define the “before” responses as Population 1, and the “after” responses as Population 2. The samples are dependent because it’s reasonable to see how a participant’s “after” response could be affected by their “before” response.

Then their null and alternative hypotheses will be

$$H_0 : \mu_1 - \mu_2 \leq 10$$

$$H_a : \mu_1 - \mu_2 > 10$$

where μ_1 is the mean starting weight before participants begin taking the new weight-loss drug, and μ_2 is the mean ending weight. And because $\mu_1 - \mu_2$ is the difference in weight, the hypothesis statements could also be written as

$$H_0 : \mu_d \leq 10$$

$$H_a : \mu_d > 10$$

where μ_d is the mean difference between the two populations.

To find the mean difference, we’ll sum the differences and divide by the number of matched-pairs in our sample, $n = 10$.

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{8 + 12 + 8 + 14 + 2 + 3 + 18 + 20 + 12 + 16}{10} = \frac{113}{10} = 11.3$$

So the sample mean tells us that mean weight loss is 11.3. Then the sample standard deviation is



$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$

To calculate this, we'll first find

$$\sum_{i=1}^n (d_i - \bar{d})^2$$

$$(8 - 11.3)^2 + (12 - 11.3)^2 + (8 - 11.3)^2 + (14 - 11.3)^2 + (2 - 11.3)^2 \\ + (3 - 11.3)^2 + (18 - 11.3)^2 + (20 - 11.3)^2 + (12 - 11.3)^2 + (16 - 11.3)^2 \\ (-3.3)^2 + 0.7^2 + (-3.3)^2 + 2.7^2 + (-9.3)^2 + (-8.3)^2 + 6.7^2 + 8.7^2 + 0.7^2 + 4.7^2 \\ 10.89 + 0.49 + 10.89 + 7.29 + 86.49 + 68.89 + 44.89 + 75.69 + 0.49 + 22.09$$

$$328.1$$

Then the sample standard deviation is

$$s_d = \sqrt{\frac{328.1}{9}}$$

$$s_d \approx \sqrt{36.456}$$

$$s_d \approx 6.038$$

Because the population standard deviations are unknown, and/or because both sample sizes are small, $n_1, n_2 < 30$, the test statistic will be

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$



$$t \approx \frac{11.3 - 10}{\sqrt{\frac{6.038}{10}}}$$

$$t \approx 1.3 \cdot \frac{\sqrt{10}}{6.038}$$

$$t \approx 0.681$$

and the degrees of freedom are

$$df = n - 1 = 10 - 1 = 9$$

At a significance level of 5% for an upper-tail test, and $df = 9$, the t -table gives 1.833.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

The company's t -test statistic $t \approx 0.681$ doesn't meet the threshold $t = 1.833$, so the critical value approach tells them that they can't reject the null hypothesis, and therefore can't conclude that their new weight-loss drug produces more weight loss than the current market-leading drug.



Topic: Confidence interval for the difference of proportions

Question: A video game developer wants to know how the number of male video game players compares to the number of female players. They randomly select 500 males and 500 females and find that 368 of the males play video games, while 230 of the females play video games. Find a 95 % confidence interval around the difference between the number of male and female players.

Answer choices:

- A (0.227,0.325)
- B (0.218,0.334)
- C (0.160,0.392)
- D (0.460,0.736)



Solution: B

The sample proportions for the males and females, respectively, are

$$\hat{p}_1 = \frac{368}{500} = 0.736$$

$$\hat{p}_2 = \frac{230}{500} = 0.460$$

The value of $z_{\alpha/2}$ for 95 % confidence in a two-tailed test is 1.96, so the confidence interval is

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.736 - 0.460) \pm 1.96 \sqrt{\frac{0.736(1 - 0.736)}{500} + \frac{0.460(1 - 0.460)}{500}}$$

$$(a, b) = 0.276 \pm 1.96 \sqrt{\frac{0.736(0.264)}{500} + \frac{0.460(0.540)}{500}}$$

$$(a, b) = 0.276 \pm 1.96 \sqrt{\frac{0.194304}{500} + \frac{0.2484}{500}}$$

$$(a, b) = 0.276 \pm 1.96 \sqrt{\frac{0.442704}{500}}$$

$$(a, b) \approx 0.276 \pm 0.058$$

Therefore, the 95 % confidence interval is

$$(a, b) \approx (0.276 - 0.058, 0.276 + 0.058)$$



$$(a, b) \approx (0.218, 0.334)$$

We can be 95% confident that the true difference of population proportions of males who play video games and females who play video games is between 0.218 and 0.334. Notice that both ends of the confidence interval are positive, so we can conclude that more males than females play video games.



Topic: Confidence interval for the difference of proportions

Question: A college director wonders about the difference between the number of male and female students who scored higher than 90 on a recent final exam. He randomly selects 25 males and 25 females and finds that 15 males and 12 females scored more than 90. Find a 99 % confidence interval around the true difference of male students and female students who scored higher than 90 on the recent exam.

Answer choices:

- A $(-0.481, 0.241)$
- B $(-0.152, 0.391)$
- C $(-0.241, 0.481)$
- D $(0.480, 0.600)$

Solution: C

The sample proportions for males and females, respectively, are

$$\hat{p}_1 = \frac{15}{25} = 0.60$$

$$\hat{p}_2 = \frac{12}{25} = 0.48$$

The value of $z_{\alpha/2}$ for 99 % confidence in a two-tailed test is 2.58, so the confidence interval is

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.60 - 0.48) \pm 2.58 \sqrt{\frac{0.6(1 - 0.6)}{25} + \frac{0.48(1 - 0.48)}{25}}$$

$$(a, b) = 0.12 \pm 2.58 \sqrt{\frac{0.6(0.4)}{25} + \frac{0.48(0.52)}{25}}$$

$$(a, b) = 0.12 \pm 2.58 \sqrt{\frac{0.24}{25} + \frac{0.2496}{25}}$$

$$(a, b) = 0.12 \pm 2.58 \sqrt{\frac{0.4896}{25}}$$

$$(a, b) \approx 0.12 \pm 0.361$$

Therefore, the 99 % confidence interval is

$$(a, b) \approx (0.12 - 0.361, 0.12 + 0.361)$$

$$(a, b) \approx (-0.241, 0.481)$$

We can be 99 % confident that the true difference of population proportions of males and females who scored higher than 90 on the recent exam is between –0.241 and 0.481. But because the confidence interval includes 0, we can't conclude that there's a significant difference between the number of male and female students who scored higher than 90.



Topic: Confidence interval for the difference of proportions

Question: Directors at colleges A and B are interested whether there's a difference in the number of students who work while attending their colleges. They take a random sample of 100 students from each college and find that 37 students at college A and 40 students at college B are currently working. Find a 90 % confidence interval around the difference of population proportions.

Answer choices:

- A $(-0.14, 0.08)$
- B $(-0.165, 0.105)$
- C $(-0.1, 0.04)$
- D $(0.37, 0.40)$

Solution: A

The sample proportions for colleges A and B , respectively, are

$$\hat{p}_1 = \frac{37}{100} = 0.37$$

$$\hat{p}_2 = \frac{40}{100} = 0.40$$

The value of $z_{\alpha/2}$ for 90% confidence in a two-tailed test is 1.65, so the confidence interval is

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.37 - 0.40) \pm 1.65 \sqrt{\frac{0.37(1 - 0.37)}{100} + \frac{0.40(1 - 0.40)}{100}}$$

$$(a, b) = -0.03 \pm 1.65 \sqrt{\frac{0.37(0.63)}{100} + \frac{0.40(0.60)}{100}}$$

$$(a, b) = -0.03 \pm 1.65 \sqrt{\frac{0.2331}{100} + \frac{0.24}{100}}$$

$$(a, b) = -0.03 \pm 1.65 \sqrt{\frac{0.4731}{100}}$$

$$(a, b) \approx -0.03 \pm 0.113$$

Therefore, the 90% confidence interval is

$$(a, b) \approx (-0.03 - 0.113, -0.03 + 0.113)$$

$$(a, b) \approx (-0.143, 0.083)$$

We can be 90% confident that the true difference of population proportions of the number of working students between the two colleges is between -0.143 and 0.083 . But because the confidence interval includes 0 , we can't conclude that there's a significant difference between the number of working students at each college.



Topic: Hypothesis testing for the difference of proportions

Question: Assuming $p_1 - p_2 = 0$, find the value of the z -test statistic, given $\hat{p}_1 = 0.295$ for $n_1 = 130$, and $\hat{p}_2 = 0.226$ for $n_2 = 110$.

Answer choices:

- A $z \approx -1.21$
- B $z \approx 1.18$
- C $z \approx 1.21$
- D $z \approx 1.27$

Solution: C

First, we'll calculate the proportion of the combined sample.

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

$$\hat{p} = \frac{0.295(130) + 0.226(110)}{130 + 110}$$

$$\hat{p} = \frac{38.35 + 24.86}{240}$$

$$\hat{p} = \frac{63.21}{240}$$

$$\hat{p} \approx 0.263$$

Now we can find the value of the z -test statistic.

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{0.295 - 0.226}{\sqrt{0.263(1 - 0.263)\left(\frac{1}{130} + \frac{1}{110}\right)}}$$

$$z = \frac{0.069}{\sqrt{0.263(0.737)\left(\frac{11}{1,430} + \frac{13}{1,430}\right)}}$$

$$z = \frac{0.069}{\sqrt{0.193831 \left(\frac{24}{1,430} \right)}}$$

$$z \approx \frac{0.069}{\sqrt{0.003}}$$

$$z \approx 1.21$$

Topic: Hypothesis testing for the difference of proportions

Question: A scientist wants to test how fast two flu drugs help patients recover from flu. He randomly assigns 100 patients each to two groups, and gives group 1 the first drug and group 2 the second drug. In the first group, 57 patients recovered from flu 5 days, while 49 patients in the second group recovered from flu in 5 days. Using a critical value approach at a 99 % confidence level, can the scientist conclude that either drug is more effective than the other?

Answer choices:

- A Yes, the first drug is more effective than the second drug
- B Yes, the second drug is more effective than the first drug
- C No, there's no significant difference in their effectiveness
- D None of these

Solution: C

We want to determine if either of the flu drugs is significantly more effective than the other, which means we need to perform a two-tailed test, and our hypothesis statements are

$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 \neq 0$$

The proportion of the combined sample is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\hat{p} = \frac{57 + 49}{100 + 100}$$

$$\hat{p} = \frac{106}{200}$$

$$\hat{p} = 0.53$$

The sample proportions are

$$\hat{p}_1 = \frac{57}{100} = 0.57$$

$$\hat{p}_2 = \frac{49}{100} = 0.49$$

Then the z -test statistic will be



$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{0.57 - 0.49}{\sqrt{0.53(1 - 0.53)\left(\frac{1}{100} + \frac{1}{100}\right)}}$$

$$z = \frac{0.08}{\sqrt{0.53(0.47)\left(\frac{1}{50}\right)}}$$

$$z = \frac{0.08}{\sqrt{0.004982}}$$

$$z \approx 1.13$$

For two-tailed test with $\alpha = 0.01$, the critical value of z is 2.576. It means that we can reject the null hypothesis if z -statistic is larger than 2.58 or smaller than -2.58 . Since $1.13 < 2.58$, the test statistic falls into the region of acceptance, so we fail to reject the null hypothesis, and we can't conclude that either of the drugs is more effective than the other.

Topic: Hypothesis testing for the difference of proportions

Question: 50 randomly chosen well-prepared students, and 50 randomly chosen poorly-prepared students, all took a math test. The response to a specific question was examined by the professor, who was interested whether the proportion of well-prepared students who answered the question correctly was more than 18 % higher than the proportion of poorly-prepared students who answered the question correctly. The professor found that 39 of the well-prepared and 35 of the poorly-prepared students answered the question correctly. What can he conclude at a 95 % confidence level?

Answer choices:

- A He fails to reject the null hypothesis, so he concludes that the proportion of well-prepared students who answers the question correctly is more than 18 % higher than the proportion of poorly-prepared students.
- B He fails to reject the null hypothesis, so he can't conclude that the proportion of well-prepared students who answers the question correctly is more than 18 % higher than the proportion of poorly-prepared students.
- C He rejects the null hypothesis, so he concludes that the proportion of well-prepared students who answers the question correctly isn't 18 % higher than the proportion of poorly-prepared students.
- D He rejects the null hypothesis, so he can't conclude that the proportion of well-prepared students who answers the question correctly is 18 % higher than the proportion of poorly-prepared students.



Solution: B

The professor is using an upper-tailed test because he's investigating whether the proportion of well-prepared students who answer the question correctly is more than 18 % larger than the proportion of poorly-prepared students who answer correctly.

$$H_0 : p_1 - p_2 \leq 0.18$$

$$H_a : p_1 - p_2 > 0.18$$

The sample proportions are

$$\hat{p}_1 = \frac{39}{50} = 0.78$$

$$\hat{p}_2 = \frac{35}{50} = 0.70$$

so, with $p_1 - p_2 = 0.18$, the z -test statistic will be

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}}$$

$$z = \frac{0.78 - 0.70 - 0.18}{\sqrt{\frac{0.78(1 - 0.78)}{50} + \frac{0.70(1 - 0.70)}{50}}}$$

$$z = \frac{-0.10}{\sqrt{\frac{0.78(0.22)}{50} + \frac{0.70(0.30)}{50}}}$$

$$z = \frac{-0.10}{\sqrt{\frac{0.3816}{50}}}$$

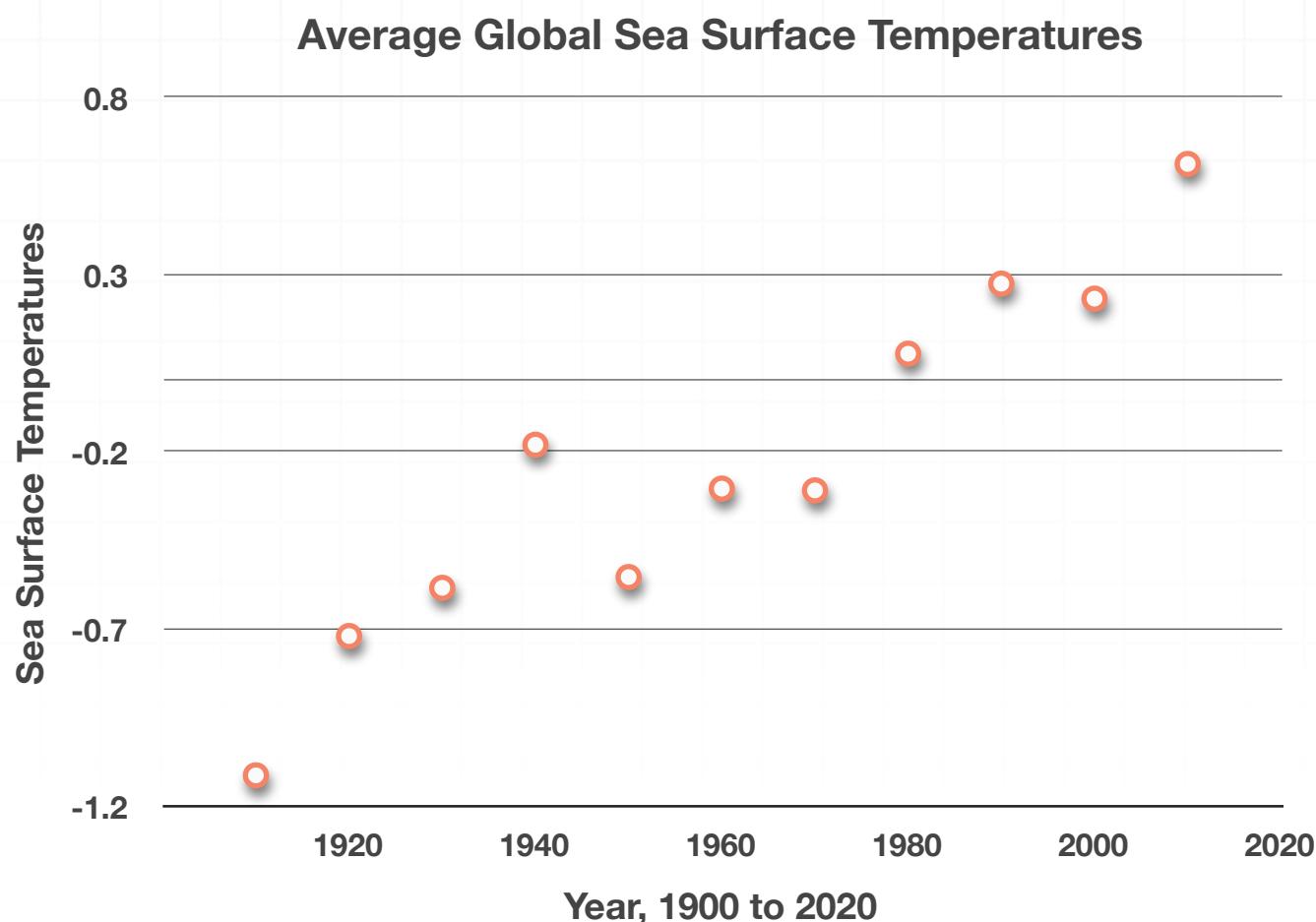
$$z = -0.10 \sqrt{\frac{50}{0.3816}}$$

$$z \approx -1.14$$

For an upper-tailed test at a 95 % confidence level, the critical z -value is 1.65. Since $-1.14 < 1.65$, the professor fails to reject the null hypothesis, and therefore can't conclude that there's more than an 18 % difference between the well-prepared and poorly-prepared students.

Topic: Scatterplots and regression

Question: Based on the scatterplot, which choice best explains the trend in the data?



Answer choices:

- A There appears to be a strong, positive, linear relationship between sea surface temperatures and time.
- B There appears to be a strong, negative, linear relationship between sea surface temperatures and time.

- C There appears to be a weak, positive, linear relationship between sea surface temperatures and time.
- D There appears to be a weak, negative, linear relationship between sea surface temperatures and time.

Solution: A

There appears to be a strong, positive, linear relationship between sea surface temperatures and time.

As time increases, the sea surface temperatures are also increasing. This means there's a positive relationship between time and temperature because as one goes up the other also goes up.

The points are relatively well clustered together so this implies a strong relationship. The section of the graph that we're looking at is linear in nature since the points trend upward from left to right.



Topic: Scatterplots and regression**Question:** Use the table to find the sum.

$$\sum xy$$

x	54	57	62	77	81	93	98
y	0.162	0.127	0.864	0.895	0.943	1.206	1.372

Answer choices:

- A $\sum xy = 522$
- B $\sum xy = 5.569$
- C $\sum xy = 461.467$
- D $\sum xy = 2,907.018$

Solution: C

The sum tells us to multiply each x -value by its corresponding y -value and then add those together.

x	y	xy
54	0.162	54(0.162)=8.748
57	0.127	57(0.127)=7.239
62	0.864	62(0.864)=53.568
77	0.895	77(0.895)=68.915
81	0.943	81(0.943)=76.386
93	1.206	93(1.206)=112.158
98	1.372	98(1.372)=134.456

Now add to find the sum.

$$\sum xy = 8.748 + 7.239 + 53.568 + 68.915 + 76.386 + 112.158 + 134.456$$

$$\sum xy = 461.467$$

Topic: Scatterplots and regression

Question: Corrina is conducting a study of how sleep and GPA are related. She surveys the students in her statistics class and creates a table comparing hours of sleep to GPA. Then she calculates the trend line of the data to be $y = 0.2069x + 1.817$. What is the correct interpretation of the slope of the trend line of the data?

Sleep	4	4	5	7	7	8	8	8	8	9	9	10
GPA	3	3	3	3	3	3	4	3	4	4	4	4

Answer choices:

- A For each additional hour of sleep, GPA decreased by 1.817 points.
- B For each additional hour of sleep, GPA decreased by 0.2069 points.
- C For each additional hour of sleep, GPA increased by 1.817 points.
- D For each additional hour of sleep, GPA increased by 0.2069 points.

Solution: D

The slope is the change in the y -values divided by the change in the x -values. In the case of this data, the y -value is the GPA and the x -values are the hours of sleep.

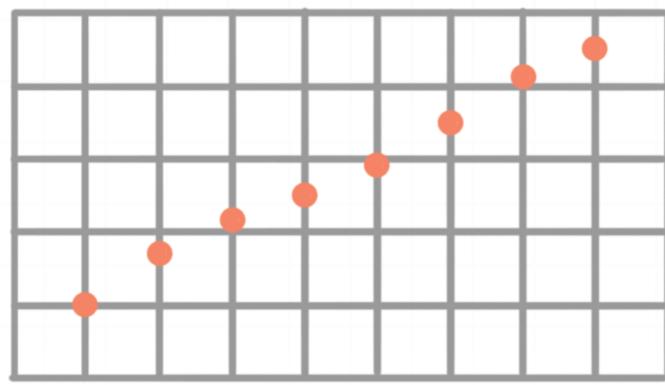
This means the slope has units “GPA points per hour.” This says that the interpretation of the slope of the best fit line is

“For each additional hour of sleep, GPA increased by 0.2069 points.”

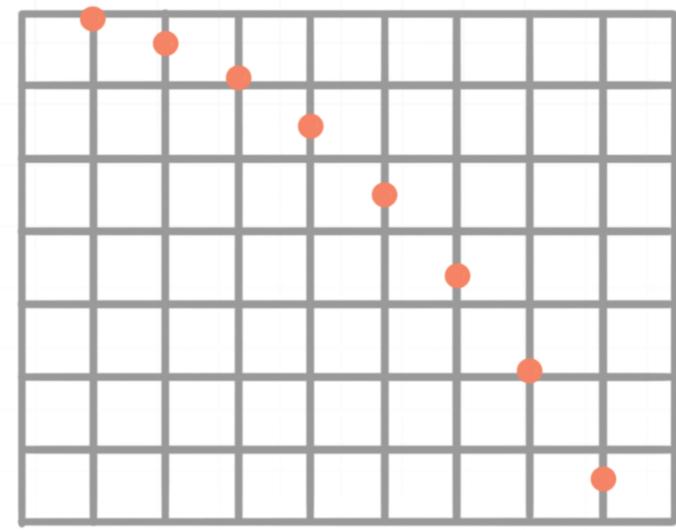


Topic: Correlation coefficient and the residual**Question:** Which scatterplot has a correlation coefficient closest to -1 ?**Answer choices:**

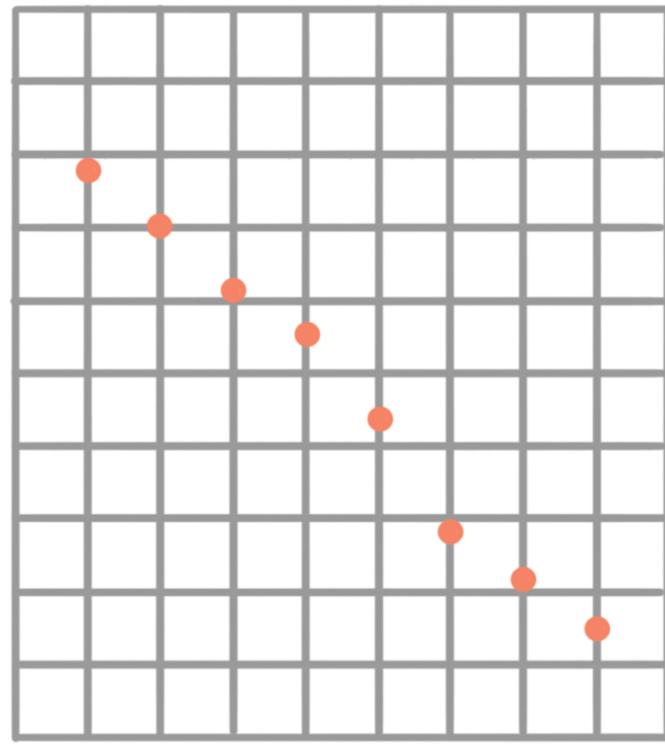
A



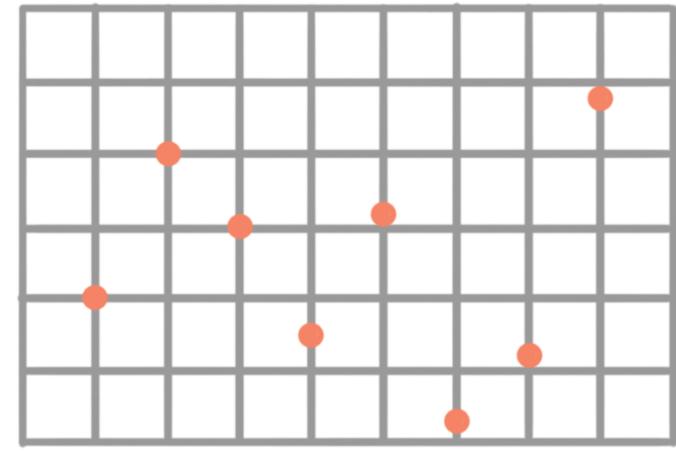
B



C



D



Solution: C

A correlation coefficient of -1 means the trend line has a negative slope and the points are very close to forming a line.

Answer choice A is an example of an R -value close to $+1$ because the points rise from left to right and are close to forming a line.

Answer choice B is an example of points that have a quadratic relationship. This might be estimated with a trend line and give a negative R -value, but there's a better choice.

Answer choice C has points that are more linear than choice B and they decrease from left to right, so this is the best answer.

Answer choice D has points that are spread out, so the correlation coefficient is likely very close to 0.



Topic: Correlation coefficient and the residual

Question: The line of best fit for a data set is given by $y = 2.0929x - 0.4429$. What's the value for the residual of the point (6,12)?

Answer choices:

- A 12.1145
- B -12.1145
- C 0.1145
- D -0.1145

Solution: D

The residual of a point is its actual value minus its predicted value.

We can calculate the predicted value by plugging in the x -coordinate of the point (6,12), into the line of best fit. The value $x = 6$ gives

$$\hat{y} = 2.0929(6) - 0.4429$$

$$\hat{y} = 12.1145$$

This is the predicted value. The actual value is the y -value of the point (6,12). The residual is

$$\text{actual} - \text{predicted}$$

$$12 - 12.1145$$

$$-0.1145$$



Topic: Correlation coefficient and the residual**Question:** Which statement is false?**Answer choices:**

- A An R -value close to 1 or -1 means the line of best fit is perfect at making predictions from your data, you do not need any other checks.
- B Always draw a scatterplot of your data points first before you perform a linear regression to get an idea of the shape of the data.
- C When you check the residual plot, it needs to be in a random pattern with points above and below 0 in order for the line of best fit to be a good predictor of the data.
- D The value of a data point is always equal to the value from the linear model, plus the residual value.

Solution: A

An R -value close to 1 or -1 could nicely describe the data that you got, but you need to be careful here.

Sometimes the R -value can be close to 1, but the data is just a small snapshot of another type of function. That's why it's important to both draw the scatter plot, as well as to calculate the residuals and look at the residual plot. These graphs, especially the graph of the residuals, can show you what the R -value can't.

For instance, the graph could tell you if your data is doing something unexpected, or if it's a small section of an exponential or parabolic function.



Topic: Coefficient of determination and RMSE

Question: If the correlation coefficient of a data set is $r = 0.92$, what percent of the variation in y can be explained by the variation in x -values?

Answer choices:

- A 92 %
- B 84.64 %
- C 8 %
- D 0.64 %

Solution: B

The percent of the variation in y that can be explained by the variation in x -values is the coefficient of determination, or the r^2 value. To find this, we square the r -value.

$$r^2 = 0.92^2$$

$$r^2 = 0.8464$$

As a percentage, this is 84.64 % .



Topic: Coefficient of determination and RMSE**Question:** When you calculate the RMSE, a smaller RMSE means that ...**Answer choices:**

- A ... the model is probably a good fit.
- B ... the model is probably a bad fit.
- C ... the r^2 value will decrease.
- D None of these

Solution: A

The root mean square error is the standard deviation of the residuals. If the residuals are closer to the line of best fit, then the standard deviation will be smaller.

This also means that the line of best fit is better correlated, so the r^2 value will increase as the RMSE decreases.



Topic: Coefficient of determination and RMSE

Question: Which RMSE will have the weakest correlation in the data for the line of best fit?

Answer choices:

- A RMSE = 0.6873
- B RMSE = 0.0871
- C RMSE = 0.9423
- D RMSE = 0.0001



Solution: C

Remember that the larger the RMSE (standard deviation of the residuals from the least squares line), the more scattered the data points are around the line of best fit, and the weaker the linear correlation in the data.



Topic: Chi-square tests

Question: A university is graduating 5,000 seniors and wants to know if their graduation rate is affected by student involvement in extracurriculars. They randomly sampled seniors as they graduated (or failed to graduate), and asked them about their extracurricular activities. What can the university conclude using a chi-square test at 95 % confidence?

	Number of extracurriculars			
	0-2	3-5	6+	Totals
Graduating	221	118	41	380
Not graduating	11	75	34	120
Totals	232	193	75	500

Answer choices:

- A Number of extracurriculars doesn't affect graduation rate
- B Number of extracurriculars affects graduation rate
- C The university can't use a chi-square test because their sample doesn't meet the large counts condition
- D The university can't use a chi-square test because their sample doesn't meet the independence condition

Solution: B

Start by computing expected values.

$$\text{Expected Graduating/0 - 2: } (380 \cdot 232)/500 = 176.32$$

$$\text{Expected Graduating/3 - 5: } (380 \cdot 193)/500 = 146.68$$

$$\text{Expected Graduating/6+: } (380 \cdot 75)/500 = 57.00$$

$$\text{Expected Not graduating/0 - 2: } (120 \cdot 232)/500 = 55.68$$

$$\text{Expected Not graduating/3 - 5: } (120 \cdot 193)/500 = 46.32$$

$$\text{Expected Not graduating/6+: } (120 \cdot 75)/500 = 18$$

Then fill in the table.

		Number of extracurriculars			
		0-2	3-5	6+	Totals
Graduating	221 (176.32)	118 (146.68)	41 (57.00)	380	
	11 (55.68)	75 (46.32)	34 (18.00)	120	
Totals		232	193	75	500

Now we'll check our sampling conditions. The problem told us that we took a random sample, and all of our expected values are at least 5, so we've met the random sampling and large counts conditions. And even though we're sampling without replacement, there are 5,000 students in



the graduating class, and we're sampling 10% of them, so we've met the independence condition as well.

We'll state the null hypothesis.

H_0 : Graduation rate is not affected by number of extracurriculars.

H_a : Graduation rate is affected by number of extracurriculars.

Calculate χ^2 .

$$\begin{aligned}\chi^2 = & \frac{(221 - 176.32)^2}{176.32} + \frac{(118 - 146.68)^2}{146.68} + \frac{(41 - 57)^2}{57} \\ & + \frac{(11 - 55.68)^2}{55.68} + \frac{(75 - 46.32)^2}{46.32} + \frac{(34 - 18)^2}{18}\end{aligned}$$

$$\chi^2 \approx 11.32 + 5.61 + 4.49 + 35.85 + 17.76 + 14.22$$

$$\chi^2 \approx 89.25$$

The degrees of freedom are

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$df = (2 - 1)(3 - 1)$$

$$df = (1)(2)$$

$$df = 2$$

With $df = 2$ and $\chi^2 \approx 89.25$, the χ^2 -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.003	0.001	5E-04
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the right, which means we will definitely exceed the alpha level $\alpha = 0.05$. Therefore, the university will reject the null hypothesis, and conclude that number of extracurricular activities affects graduation rate.

Topic: Chi-square tests

Question: A restaurant wants to know whether a diner's choice to order dessert is affected by whether or not they ordered an appetizer. On an evening when they served 2,000 diners, they randomly sampled diners as they finished their meals, and recorded whether they had ordered an appetizer and/or dessert. What can the restaurant conclude using a chi-square test at 95 % confidence?

	Dessert	No dessert	Totals
Appetizer	70	42	112
No appetizer	50	38	88
Totals	120	80	200

Answer choices:

- A Ordering an appetizer doesn't affect the dessert order
- B Ordering an appetizer affects the dessert order
- C The restaurant can't use a chi-square test because their sample doesn't meet the large counts condition
- D The restaurant can't use a chi-square test because their sample doesn't meet the independence condition

Solution: A

Start by computing expected values.

$$\text{Expected Appetizer/Dessert: } (112 \cdot 120)/200 = 67.2$$

$$\text{Expected Appetizer/No dessert: } (112 \cdot 80)/200 = 44.8$$

$$\text{Expected No appetizer/Dessert: } (88 \cdot 120)/200 = 52.8$$

$$\text{Expected No appetizer/No dessert: } (88 \cdot 80)/200 = 35.2$$

Then fill in the table.

	Dessert	No dessert	Totals
Appetizer	70 (67.2)	42 (44.8)	112
No appetizer	50 (52.8)	38 (35.2)	88
Totals	120	80	200

Now we'll check our sampling conditions. The problem told us that we took a random sample, and all of our expected values are at least 5, so we've met the random sampling and large counts conditions. And even though we're sampling without replacement, there are 2,000 diners, and we're sampling 10% of them, so we've met the independence condition as well.

We'll state the null hypothesis.

H_0 : Whether or not a diner orders dessert is not affected by whether or not they ordered an appetizer.



H_a : Whether or not a diner orders dessert is affected by whether or not they ordered an appetizer.

Calculate χ^2 .

$$\chi^2 = \frac{(70 - 67.2)^2}{67.2} + \frac{(42 - 44.8)^2}{44.8} + \frac{(50 - 52.8)^2}{52.8} + \frac{(38 - 35.2)^2}{35.2}$$

$$\chi^2 \approx 0.12 + 0.18 + 0.15 + 0.22$$

$$\chi^2 \approx 0.67$$

The degrees of freedom are

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$df = (2 - 1)(2 - 1)$$

$$df = (1)(1)$$

$$df = 1$$

With $df = 1$ and $\chi^2 \approx 0.67$, the χ^2 -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.003	0.001	5E-04
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the left, which means we will definitely not exceed the alpha level $\alpha = 0.05$. Therefore, the restaurant will fail to reject the null hypothesis, and conclude whether or not a diner orders an appetizer does not affect whether or not they order dessert.

Topic: Chi-square tests

Question: A company wants to know if the number of sick days taken by its employees is affected by quarter. They recorded sick days taken each quarter. What can the company conclude using a chi-square test at 95 % confidence?

Quarter	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Total
Sick days	44	49	45	42	180

Answer choices:

- A Sick days taken is not affected by quarter
- B Sick days taken is affected by quarter
- C The company can't use a chi-square test because their sample doesn't meet the large counts condition
- D The company can't use a chi-square test because their sample doesn't meet the independence condition

Solution: A

With 180 total sick days, the expected number of sick days in each quarter would be $180/4 = 45$.

Quarter	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	Total
Sick days	44	49	45	42	180
Expected	45	45	45	45	180

We'll state the null hypothesis.

H_0 : Number of sick days taken is not affected by quarter.

H_a : Number of sick days taken is affected by quarter.

Calculate χ^2 .

$$\chi^2 = \frac{(44 - 45)^2}{45} + \frac{(49 - 45)^2}{45} + \frac{(45 - 45)^2}{45} + \frac{(42 - 45)^2}{45}$$

$$\chi^2 = \frac{1}{45} + \frac{16}{45} + \frac{0}{45} + \frac{9}{45}$$

$$\chi^2 \approx 0.02 + 0.36 + 0.00 + 0.20$$

$$\chi^2 \approx 0.58$$

The degrees of freedom are $n - 1 = 4 - 1 = 3$. With $df = 3$ and $\chi^2 = 0.58$, the χ^2 -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.003	0.001	5E-04
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

We're off the chart on the left, which means we will definitely not exceed the alpha level $\alpha = 0.05$. Therefore, the company will fail to reject the null hypothesis, and conclude that number of sick days taken is not affected by quarter.

