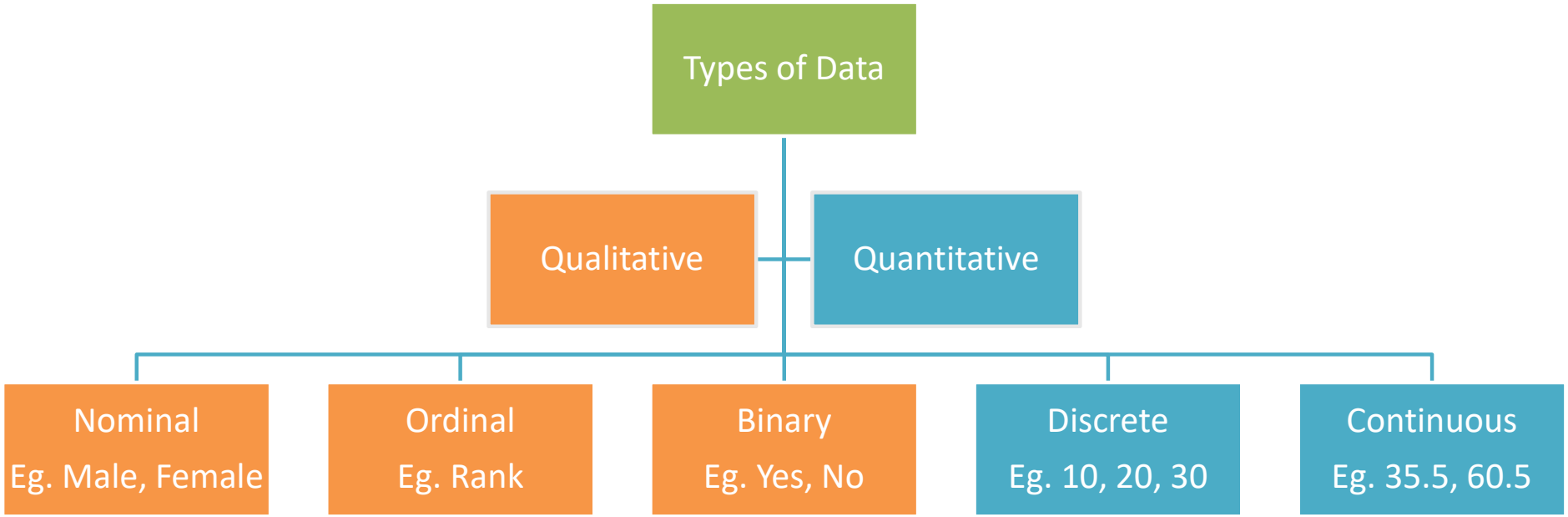
Amazing Journey of Statistics begins here

Types of data



ANALYZING CATEGORICAL DATA

How to identify individuals, variables and categorical variables in a data set?

Biscuits are snacks liked by people of every age . The nutritional data of some famous Indian biscuits is as follows:

Biscuits	Type	Calories	Carbs
Krack Jack	Sweet and Salty	16	2
Monaco	Salty	15	1.8
Marie	Sweet	25	6
Oreo	Sweet	53	8
Nice	Sweet	39	5

How to identify individuals, variables and categorical variables in a data set?

Q1. The individuals in this data set are:

i) Consumers

ii) Biscuits

iii) Calories

Biscuits	Type	Calories	Carbs
Krack Jack	Sweet and Salty	16	2
Monaco	Salty	15	1.8
Marie	Sweet	25	6
Oreo	Sweet	53	8
Nice	Sweet	39	5

Q2. The data set consists of:

i) 3 variables, 1 of which is categorical

ii) 3 variables, 2 of which are categorical

iii) 2 variables, 1 of which is categorical

Reading Pictographs

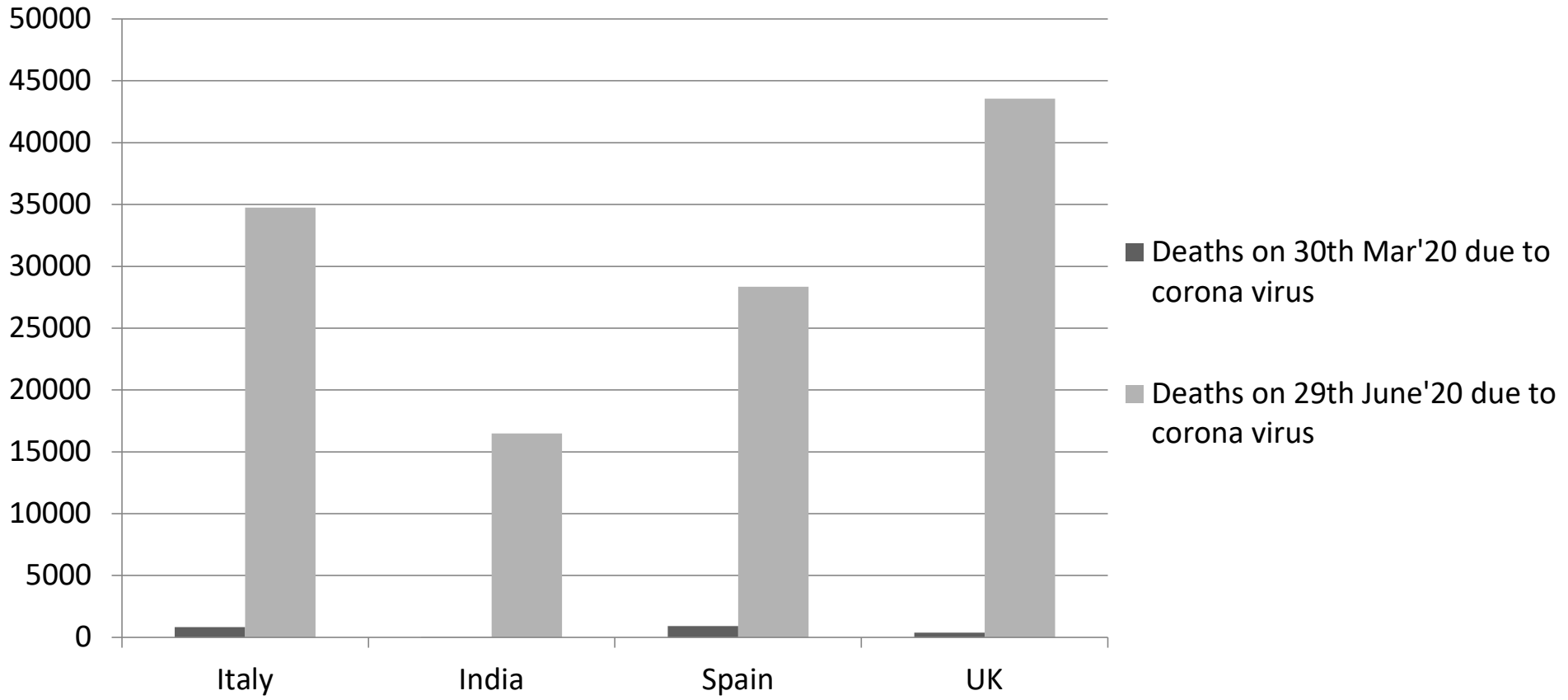
According to the pictograph below, how many survey respondents have white hair and many have brown hair?

Colour of hair	Number of People
Black	
Brown	
White	

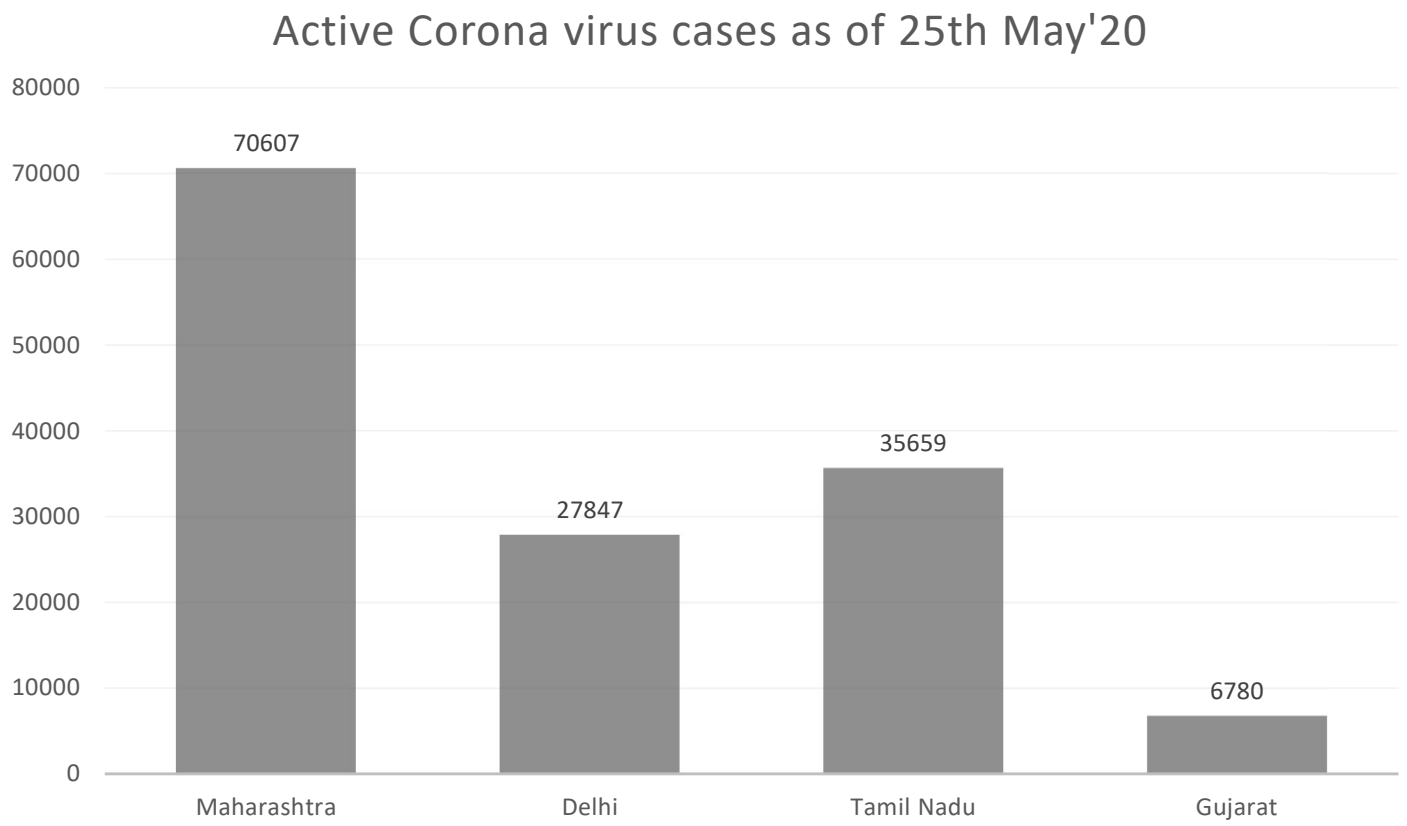
 = 2 People

Reading Bar Graphs

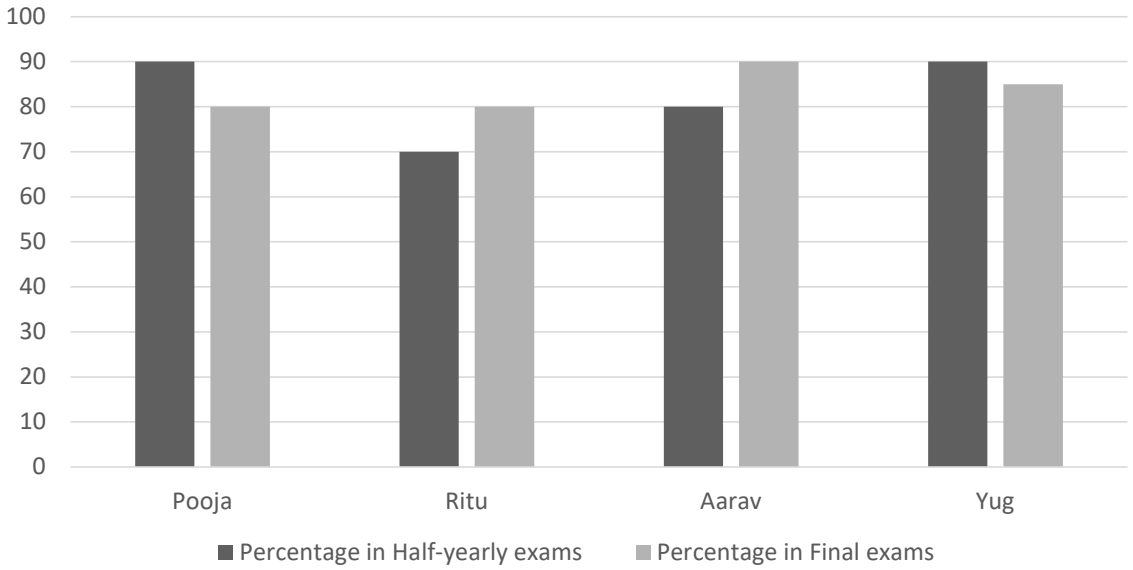
Based on the data below, which country's daily death count due to corona virus has increased the most between 30th Mar'20 and 29th Jun'20?



Reading Bar Graph: Example of few Indian states having corona virus cases



Reading bar chart: along with central tendency



Q1. What was the median percentage for the half-yearly exams?

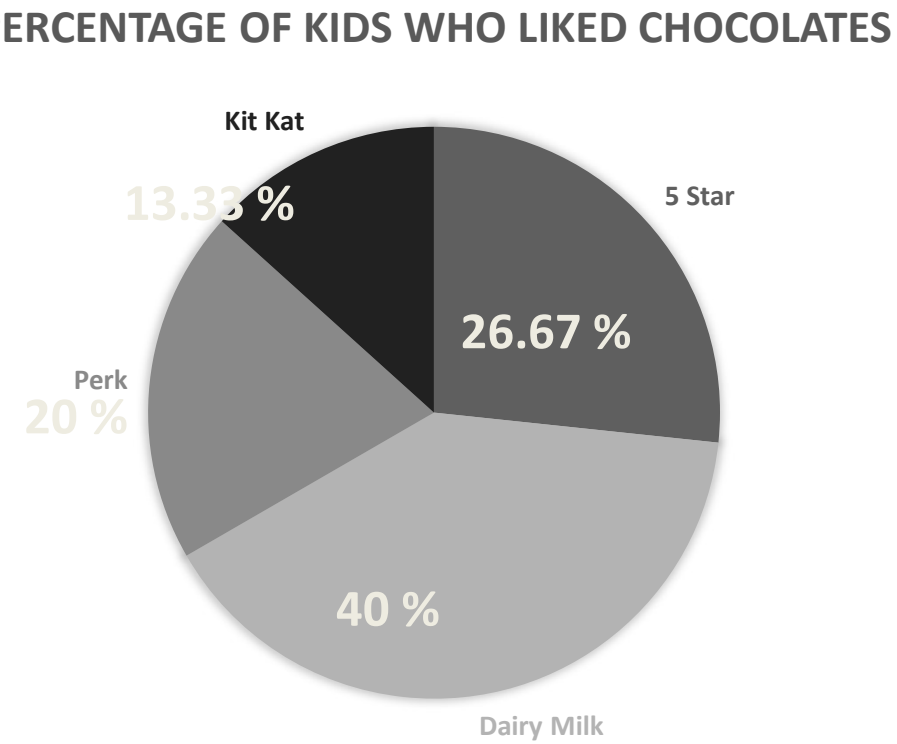
Q2. What is the midrange percentage of the final exams?

Q3. What was the average student percentage for the half- yearly exams?

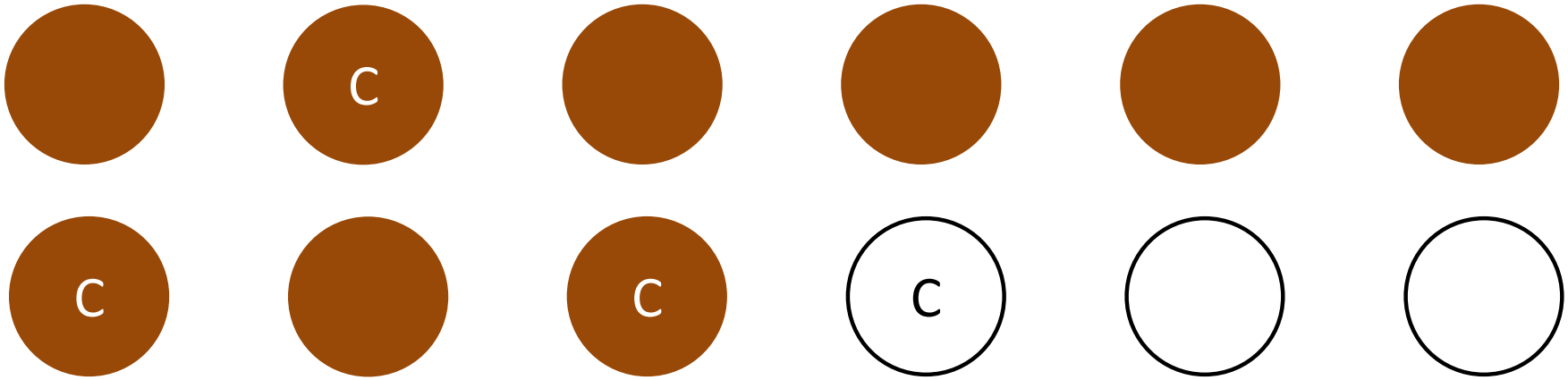
Q4. What is the range percentage of the final exams?

Reading Pie Chart

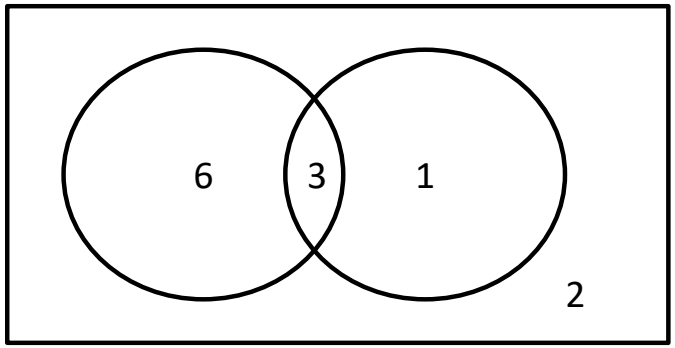
15 kids were asked about their favourite chocolate. 4 said ‘5 Star’, 6 said ‘Dairy Milk’, 3 said ‘Perk’ and 2 said ‘Kit-Kat’. Now let’s create a pie chart. Which are the most liked and the most disliked chocolates?



Two way frequency table and Venn Diagram



Venn Diagram



Two way Table

	Coconut	No Coconut	Total
Has Chocolate	3	6	9
No Chocolate	1	2	3
Total	4	8	

Marginal and Conditional Distribution

Score	Time Spend					
	0 - 20	**21 - 40**	**41 - 60**	**> 60**	**Total**	**MD**
80 – 100	0	5	20	20	45	21%
60 – 79	0	20	30	10	60	28%
40 – 59	2	4	32	32	70	33%
20 – 39	15	2	8	0	25	12%
0 – 19	2	0	0	12	14	6.5%
Total	19	31	90	74	214	
MD	8.9%	14.5%	42%	35%		

Marginal Distribution we can show by count or percentages

Distribution of score given that students study between 41-60 minutes?

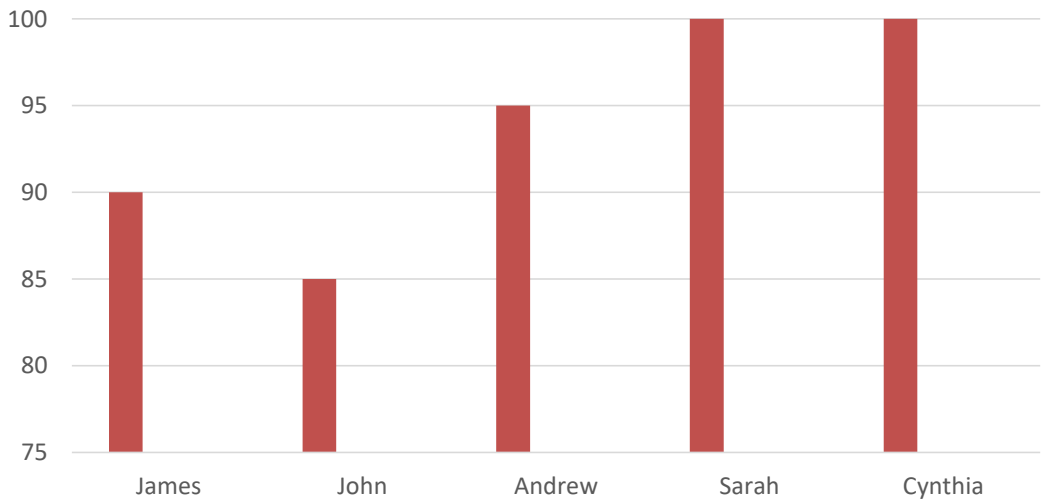
Time Spend

Score	41 - 60	CD
80 – 100	20	22%
60 – 79	30	33%
40 – 59	32	35.5%
20 – 39	8	8.9%
0 – 19	0	0%
Total	90	

Conditional Distribution we can show by percentages

Displaying Quantitative data with graphs

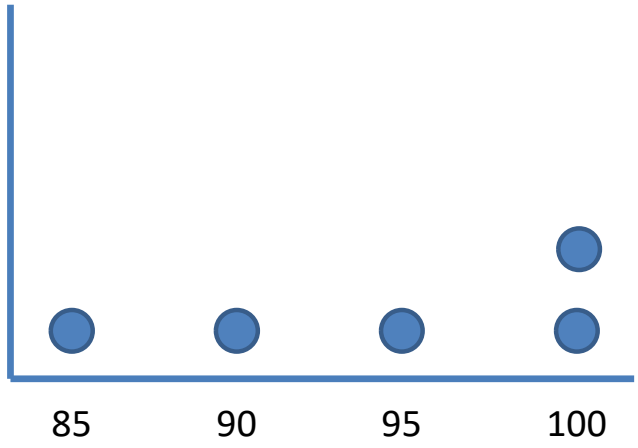
Name	Score
James	90
John	85
Andrew	95
Sarah	100
Cynthia	100



Range: Max – Min
100 – 85 = 15

Below 100: 3

Most Frequent: 100



Frequency Tables and Dot Plots

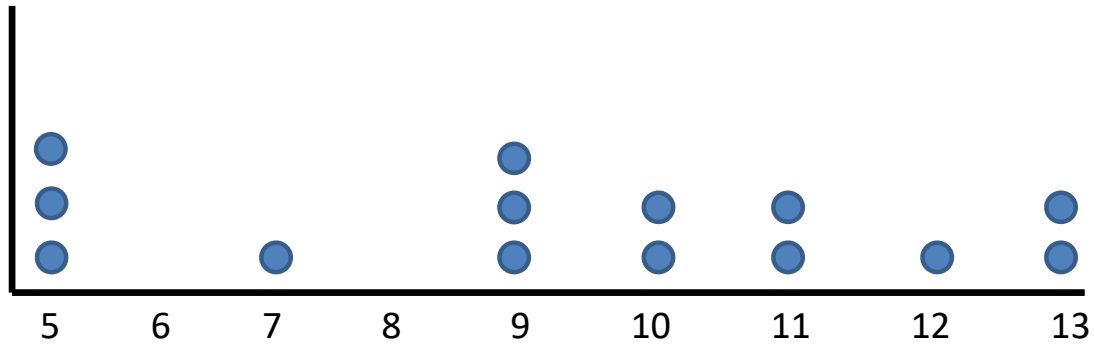
Age of students in class:

5, 7, 5, 9, 11, 12, 13, 9, 5, 10, 11, 13, 9, 10

Frequency Table

Age	Frequency
5	3
6	0
7	1
8	0
9	3
10	2
11	2
12	1
13	2

Dot Plot



Range: 13 – 5 = 8

How many older than 9? 7

Creating Histogram

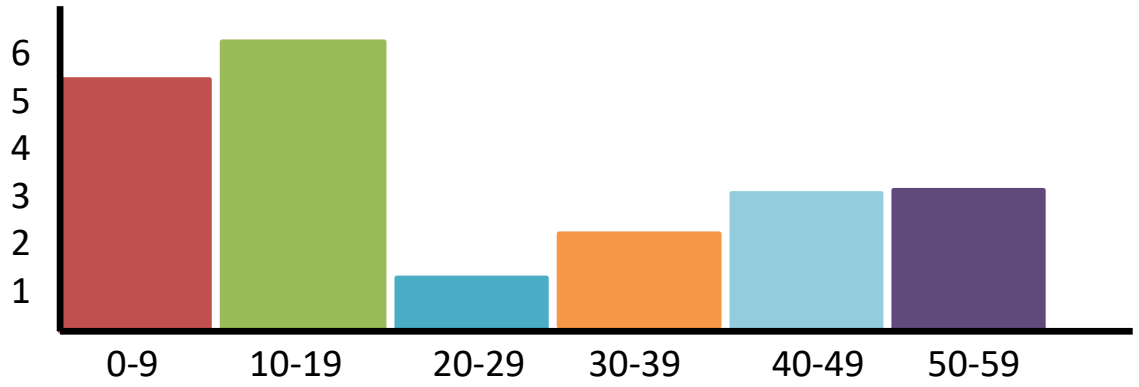
Age of students in class:

5, 7, 25, 19, 31, 42, 13, 9, 5, 10, 11, 13, 9, 10, 34, 45, 54, 55, 56, 43

Frequency Table

Bins	Frequency
0-9	5
10-19	6
20-29	1
30-39	2
40-49	3
50-59	3

Histogram



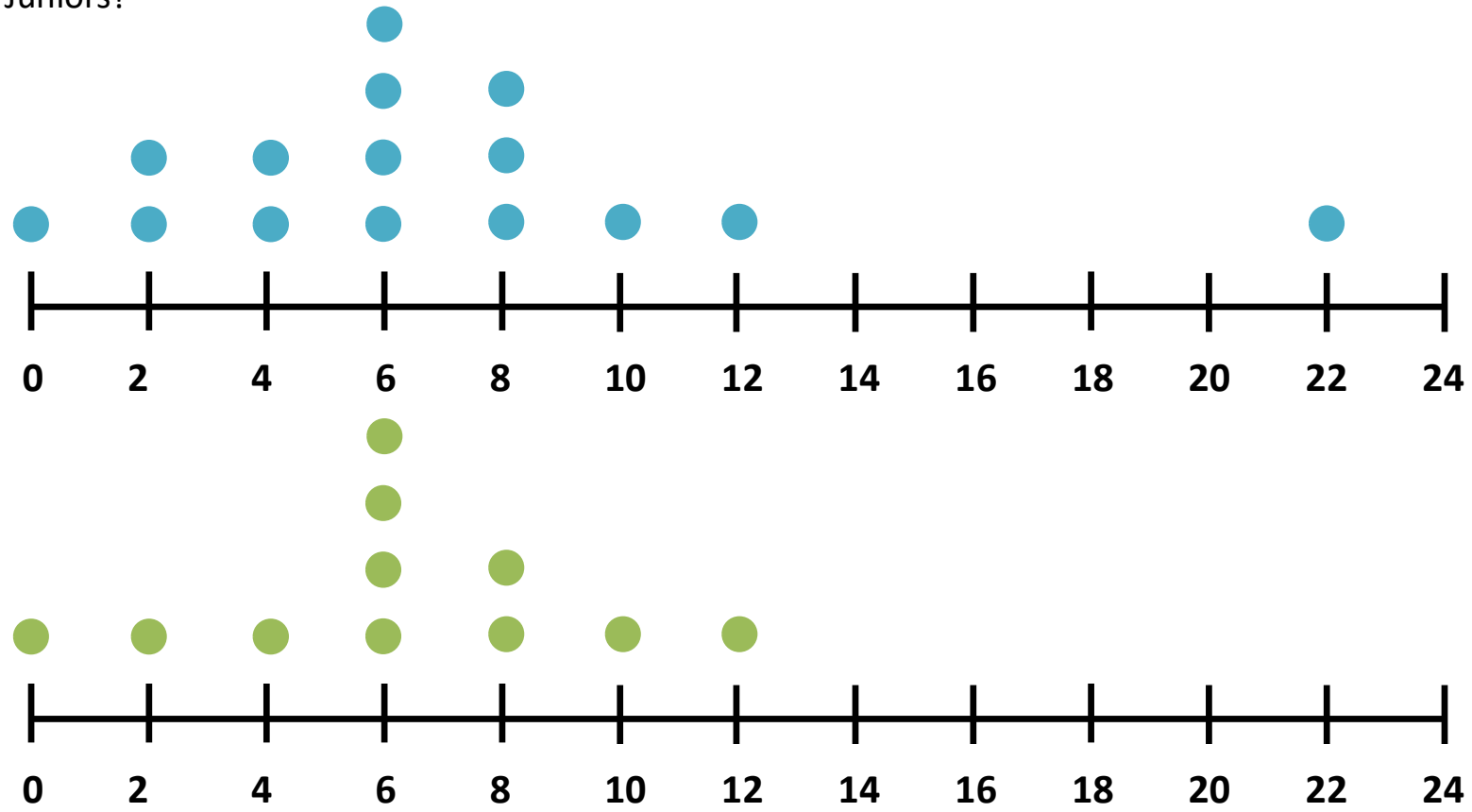
Measure of Central Tendency

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

23, 30, 21, 32, 24, 22, 33, 21, 25

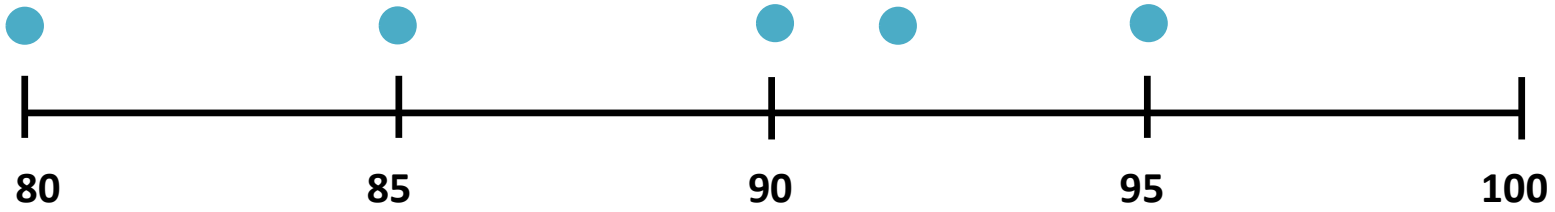
23, 30, 21, 32, 24, 22, 33, 21, 25, 32

James interviewed Juniors and Seniors at his university, asking them how many pieces of fruit they eat each day.
The results are shown below:
Which group has high mean?
Is mean a good measure for centre of the distribution for seniors? Can we use mean for centre of the distribution for Juniors?

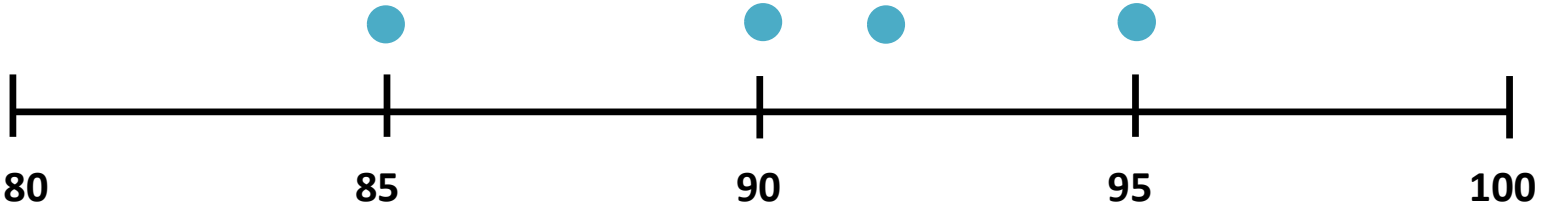


Atharv played 5 rounds of golf, and his lowest score was an 80.

The scores of the first 4 rounds and the lowest round are show in the following dot plot.



It was found that Atharv broke some rules when he scored 80, so that score will be removed from the data set.



How will the removal of the lowest score affect the mean and the median?

- Both the mean and the median will decrease , but the mean will decrease by more than the median.
- Both the mean and the median will decrease , but the median will decrease by more than the mean.
- Both the mean and the median will increase , but the mean will increase by more than the median.
- Both the mean and the median will increase , but the median will increase by more than the mean.

A group of 4 friends likes to bowl together, and each friend keeps track of his all time highest score in a single game. Their high scores are all between 180 and 220, except for Atharv, whose highest score is 250.

Atharv then bowls a great game and has a new high score of 290.

How will increasing Atharv's high score affect the mean and median?

- Both the mean and median will increase
- The median will increase, and the mean will stay the same
- The mean will increase, and the median will stay the same
- The mean will increase, and the median will decrease

At least one of the salesperson sold more than 10 cars?

- True
- False
- Don’t know

- There are 11 salespeople
- The median of sales is 6
- Range is 4

Measure of Spread

The following data points represent the number of chocolates in each kid's pocket.

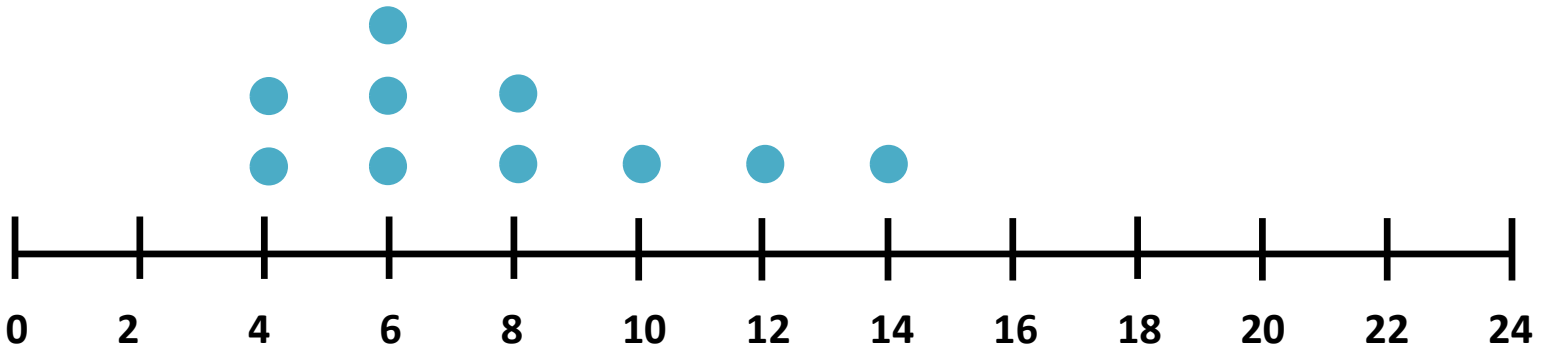
Sort the data from least to greatest

4, 5, 11, 10, 7, 12, 6, 11, 13

Find the interquartile range (IQR) of the data set.

Find the interquartile range (IQR) of the data shown in the dot plot below:

Here the data is talking about Movies on each album in James's collection



Sample 1	Sample 2
-10, 0, 10, 20, 30	8, 9, 10, 11, 12

Mean:

Median:

Mode:

Range:

Variance:

Standard Deviation:

Here is the data which consists of the employees with the years of experience in data science, find the mean and the variance

1

3

6

7

15

Here is the data which consists of the employees with the years of experience in data science, find the standard deviation

1

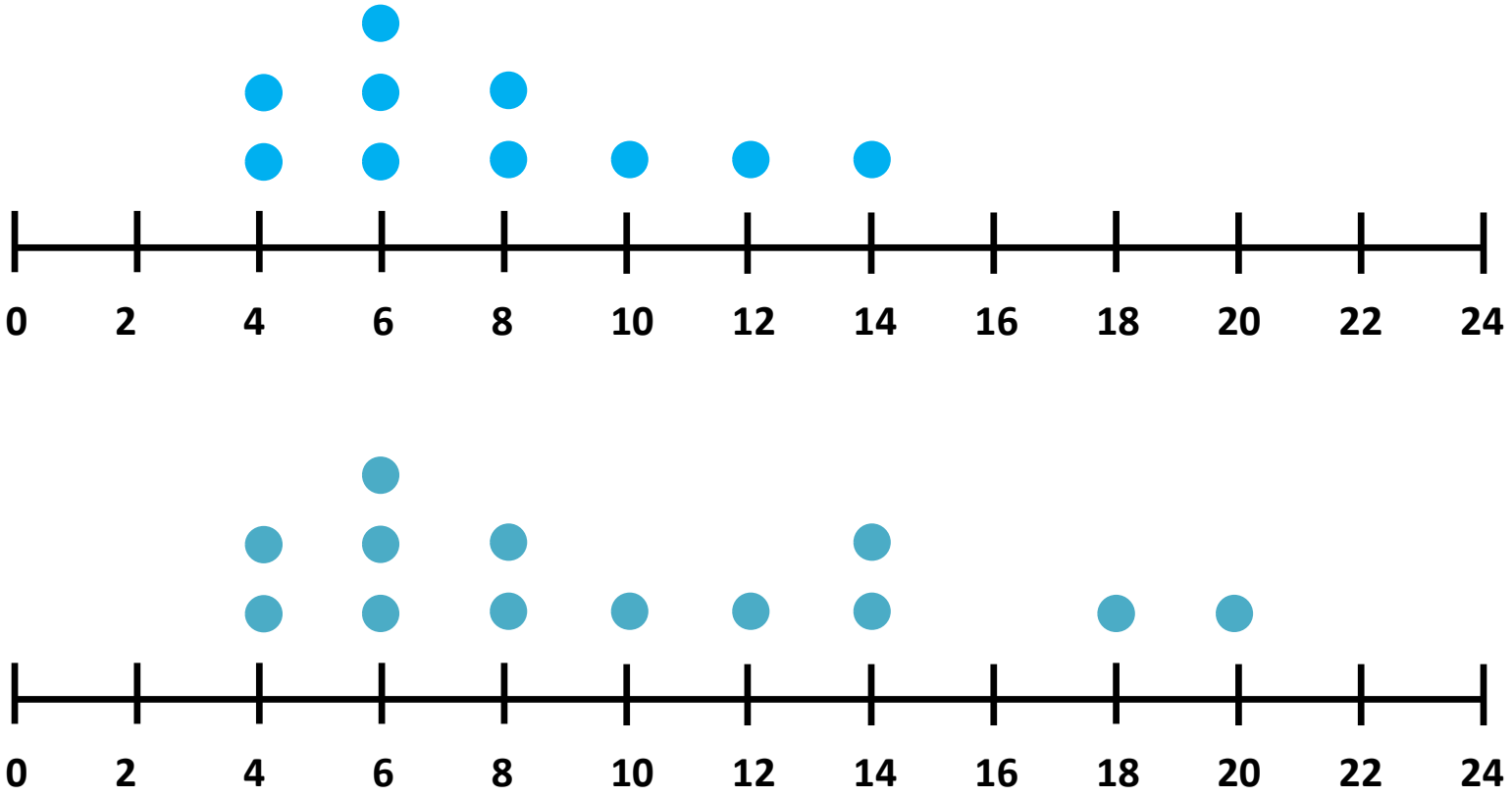
3

6

7

15

Which distribution has the highest standard deviation?



Mean and Standard Deviation vs. Median and Interquartile range (IQR)

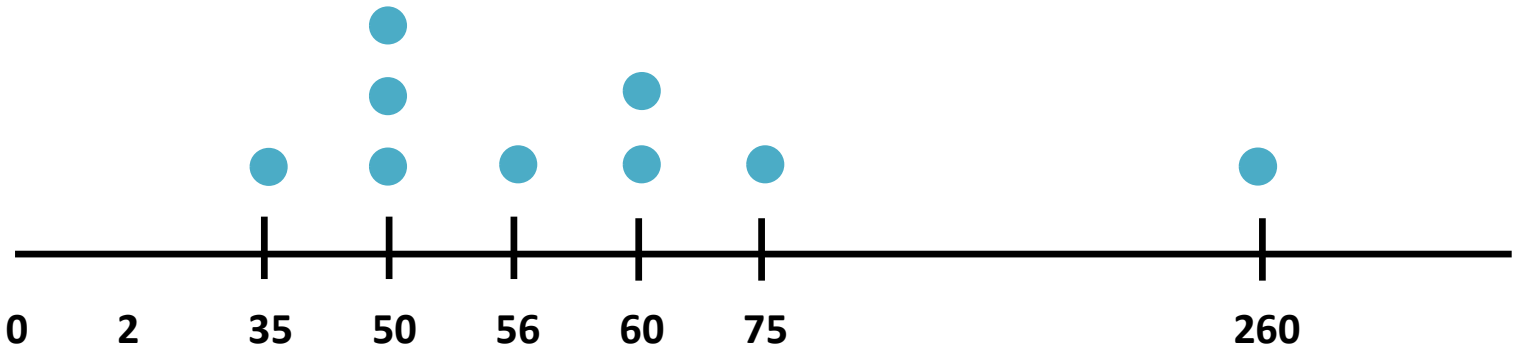
35, 50, 50, 50, 56, 60, 60, 75, 260

Mean:

Median:

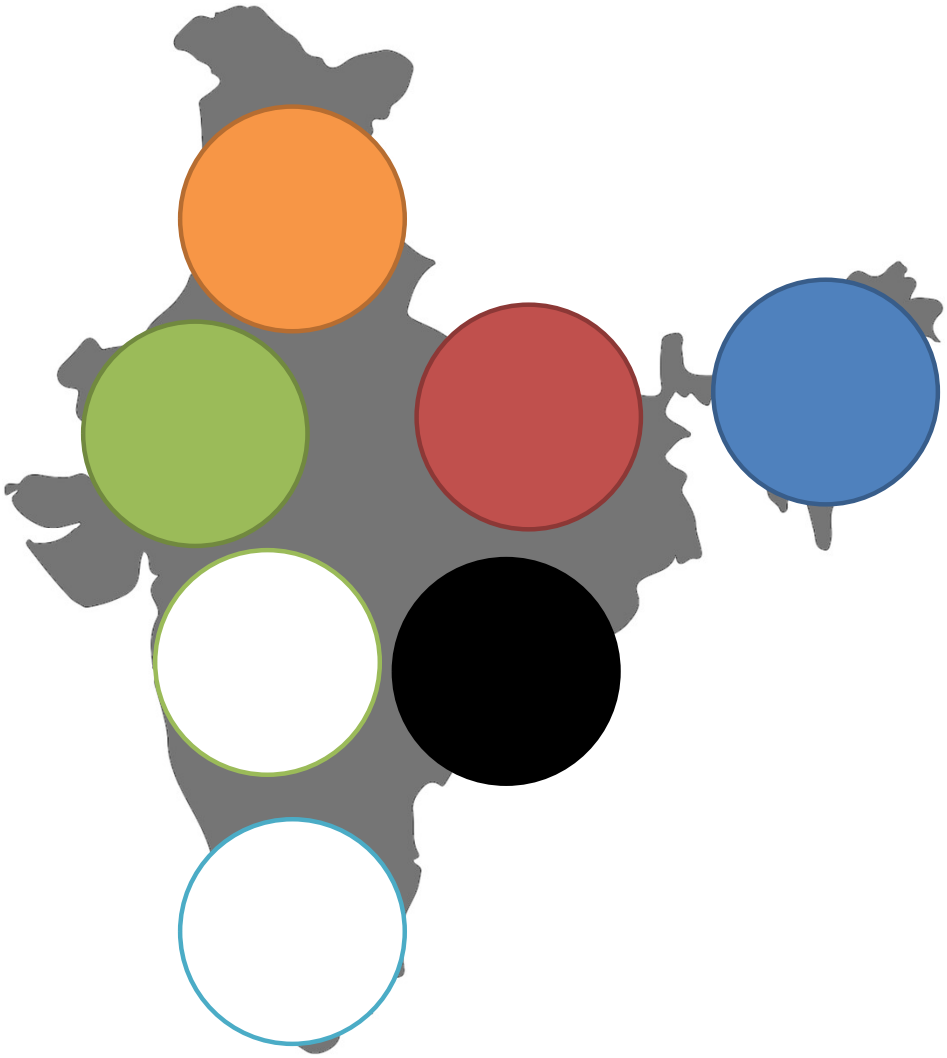
Standard Deviation:

Interquartile Range:



Alternate Variance Formula: Derivation

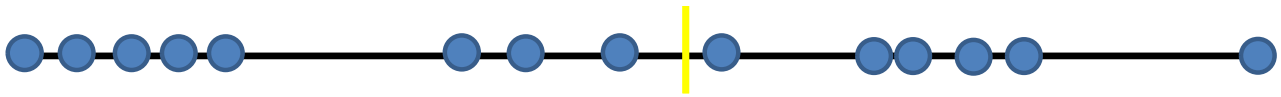
1,382,377,346

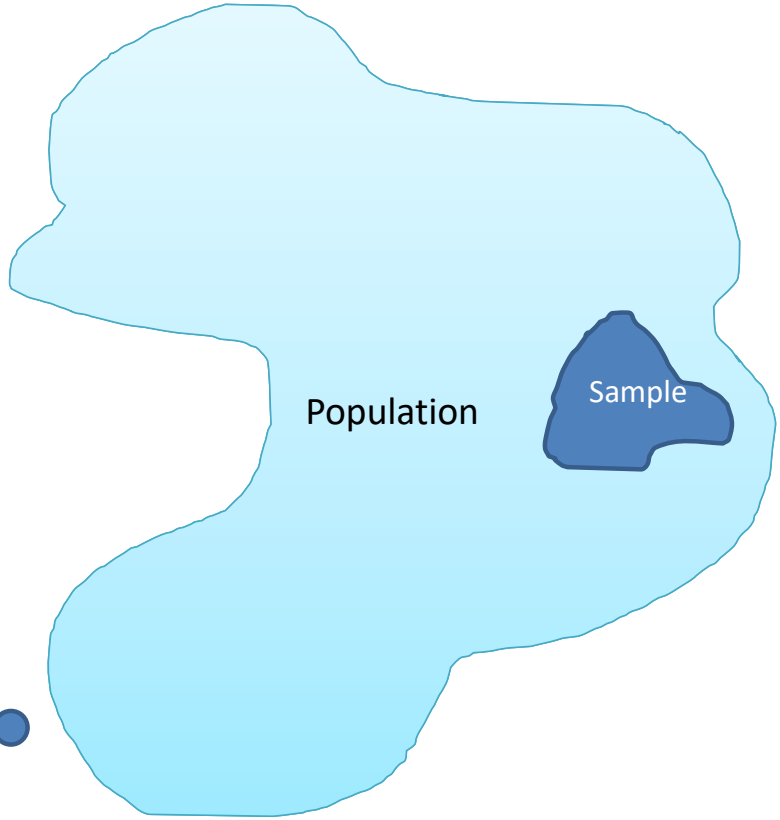


Population (Parameter) Sample (Statistic)

Mean:

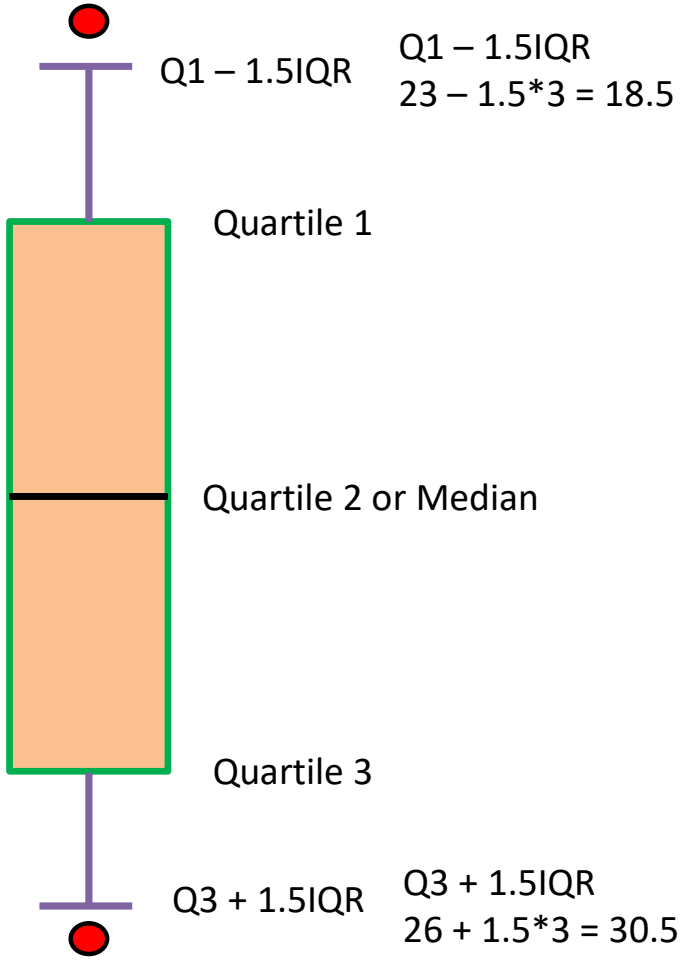
Variance:





N = 14
n = 3

Box Plot



Q1 – 1.5IQR
23 – 1.5*3 = 18.5

Q3 – Q1
26 – 23 = 3

Q3 + 1.5IQR
26 + 1.5*3 = 30.5

Age
4
23
24
24
26
26
60

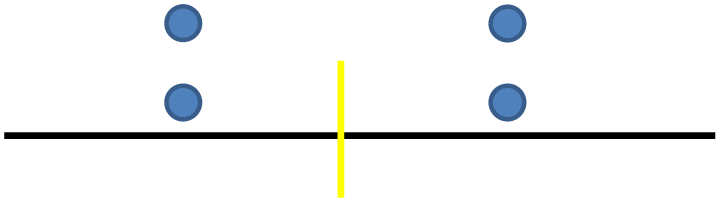
Find the range and the mid range of the following set of numbers:

66, 83, 74, 88, 95, 78, 66, 84, 81

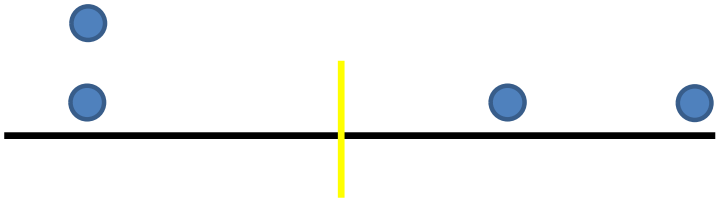
Mean Absolute Deviation (MAD)

2, 2, 4, 4

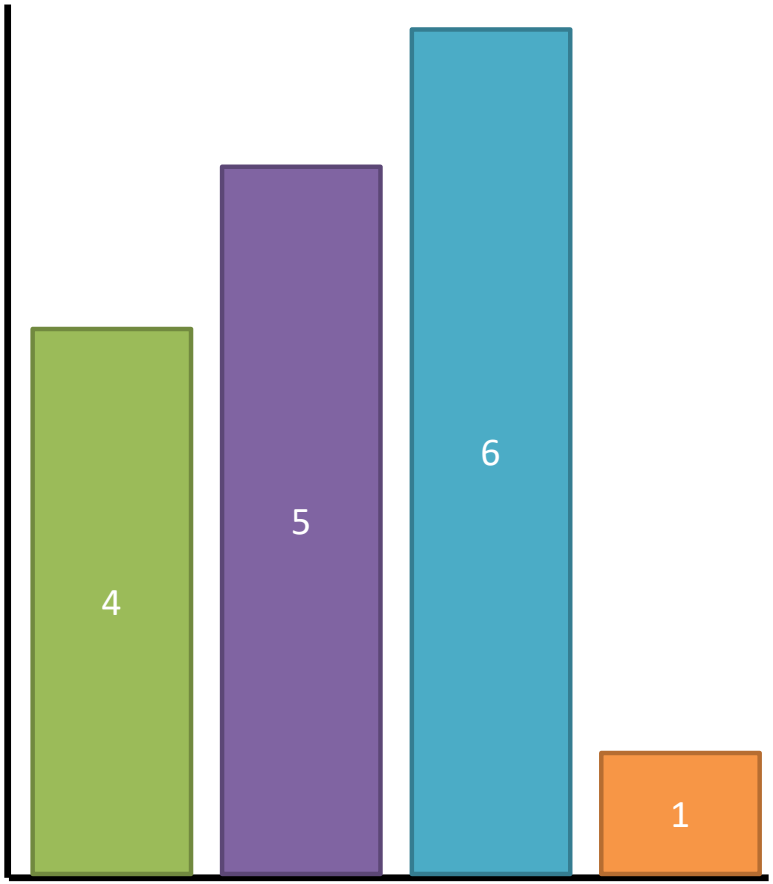
Mean:



1, 1, 6, 4

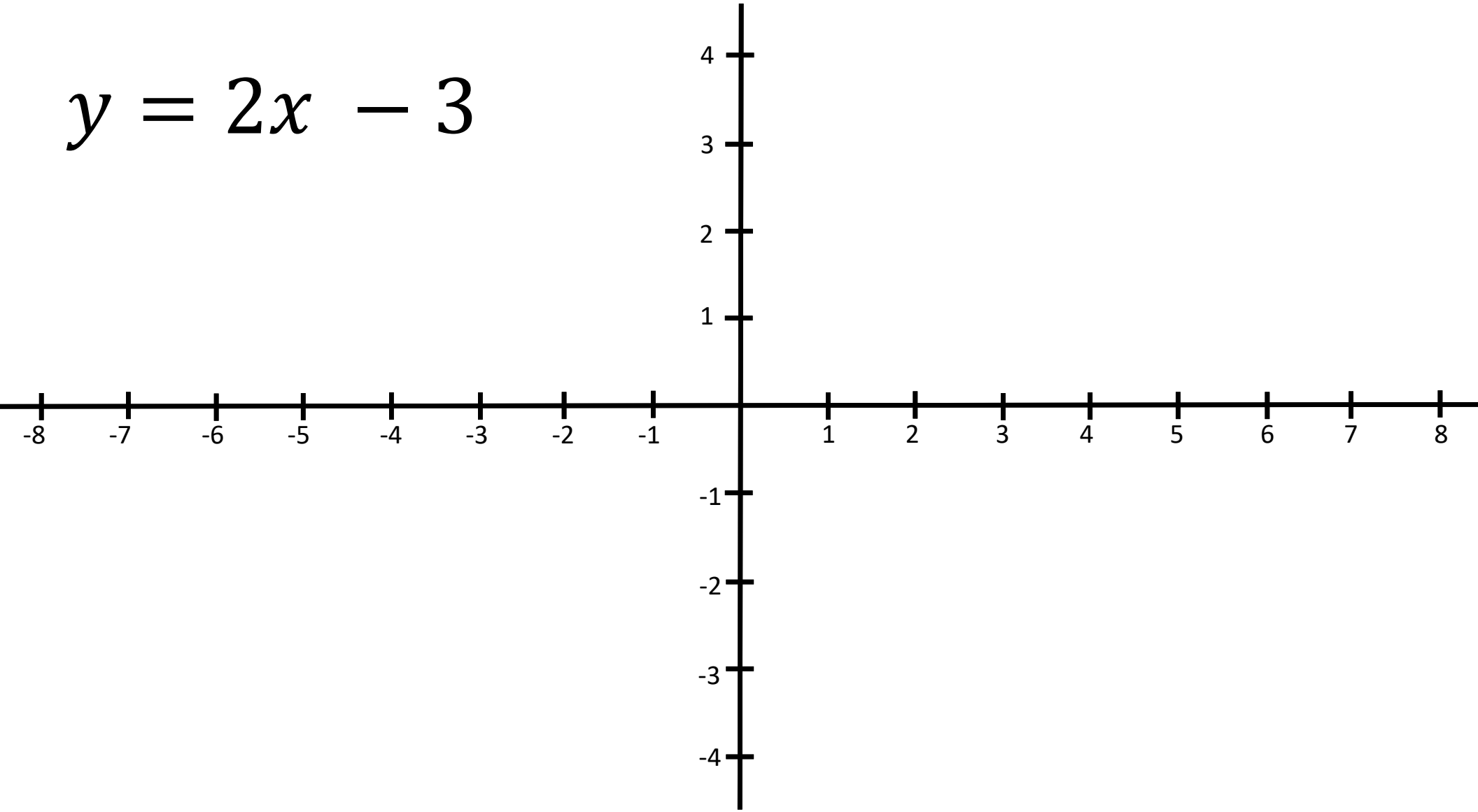


Calculate Mean Absolute Deviation (MAD)

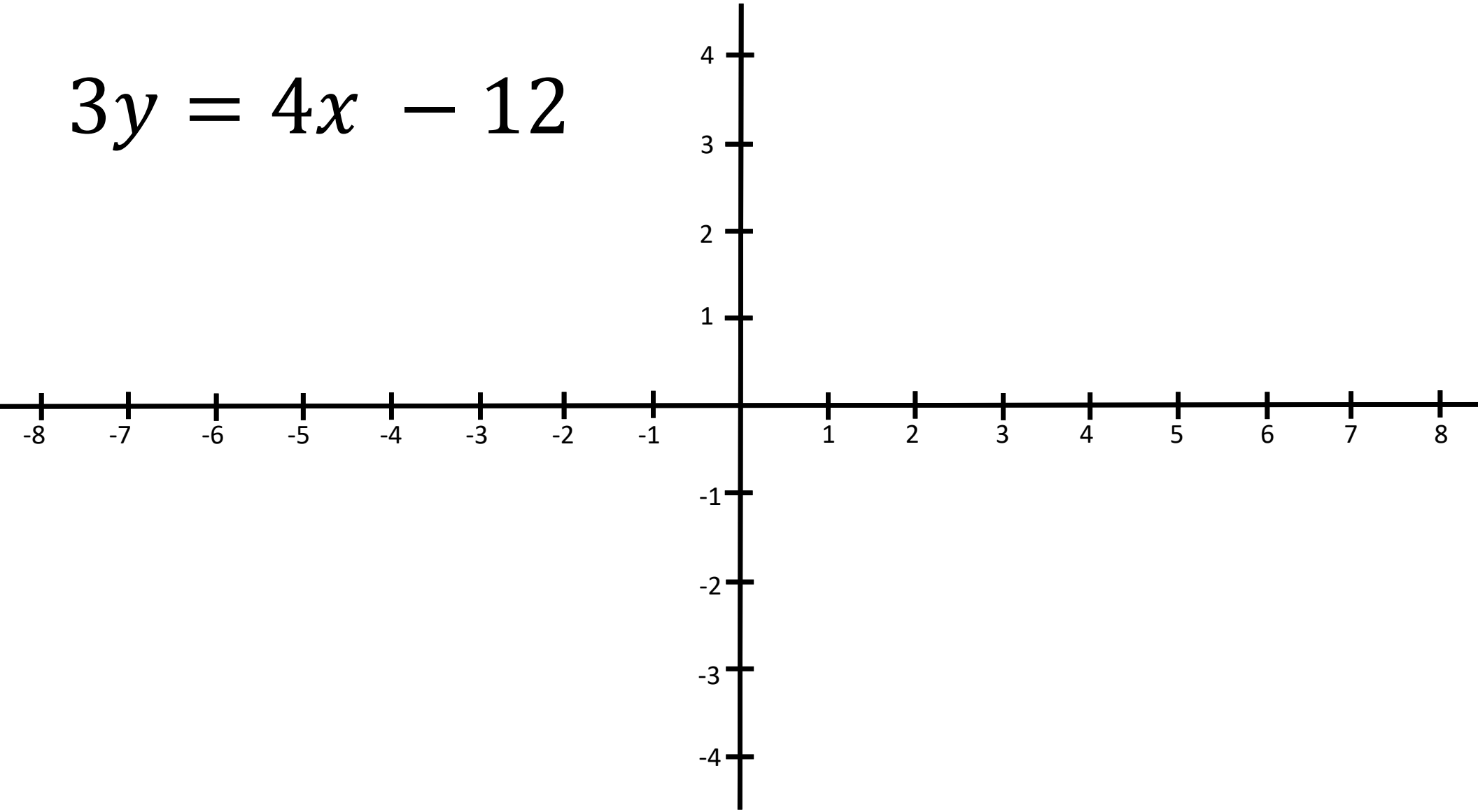


Exploring Bivariate Numerical Data

$y = 2x - 3$



$$3y = 4x - 12$$



Which of the ordered pairs is a solution of the equation mentioned below?

$$-3x - y = 6$$

- (-4, 4)
- (-3,3)
- Both
- Neither

Which of the ordered pairs is a solution of the equation mentioned below?

$$4x - 1 = 3y + 5$$

- (3, 2)
- (2,3)
- Both
- Neither

Fill in the blanks using below equation

$$-3x + 7y = 5x + 2y$$

- When x = -5, what is the value of y
- When y = 8, what is the value of x

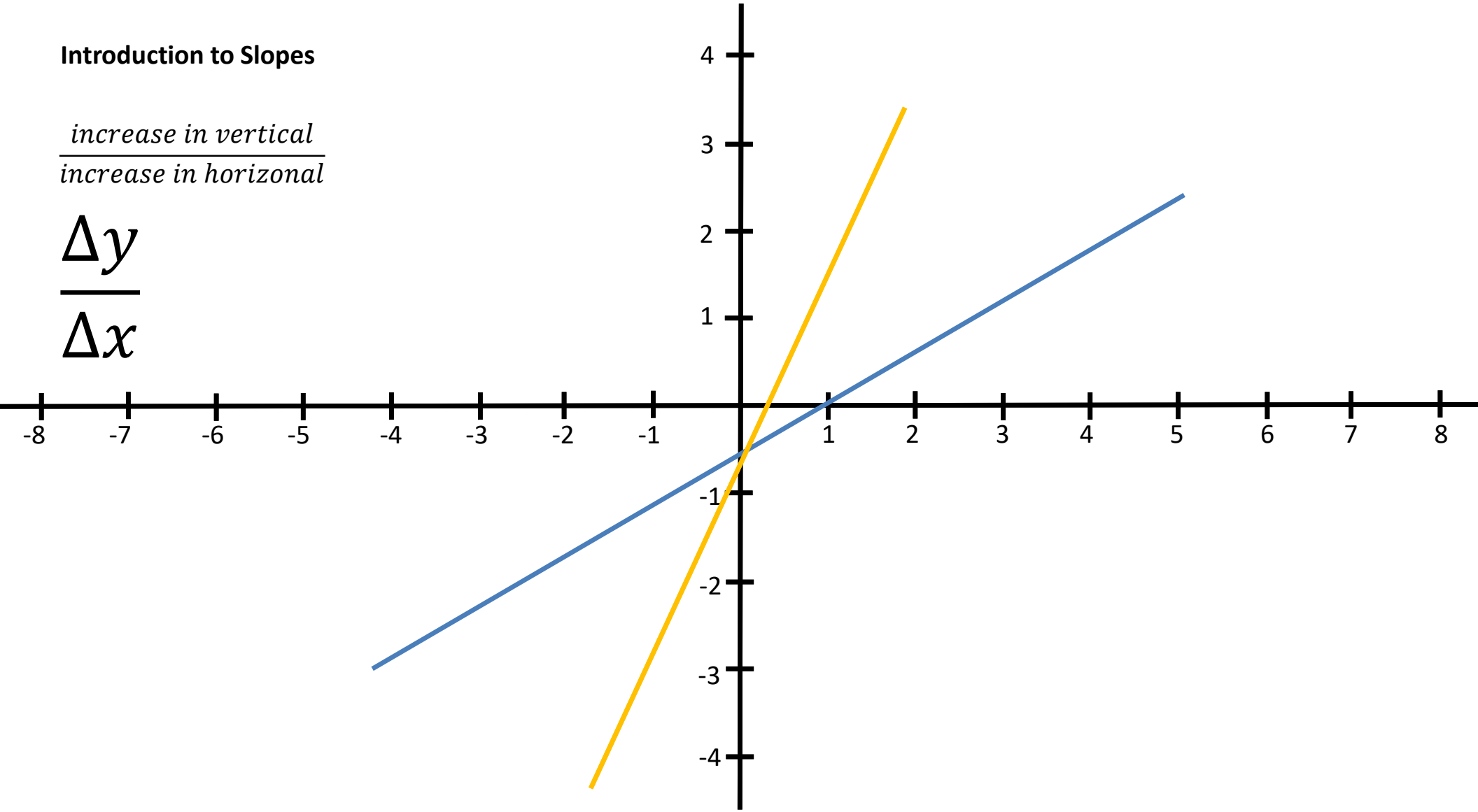
Two approaches:

- Put the value of x and y in the equation and solve
- Get x in one side and y in one side and find the relation between x and y and then pass the values

Introduction to Slopes

$$\frac{increase\ in\ vertical}{increase\ in\ horizonal}$$

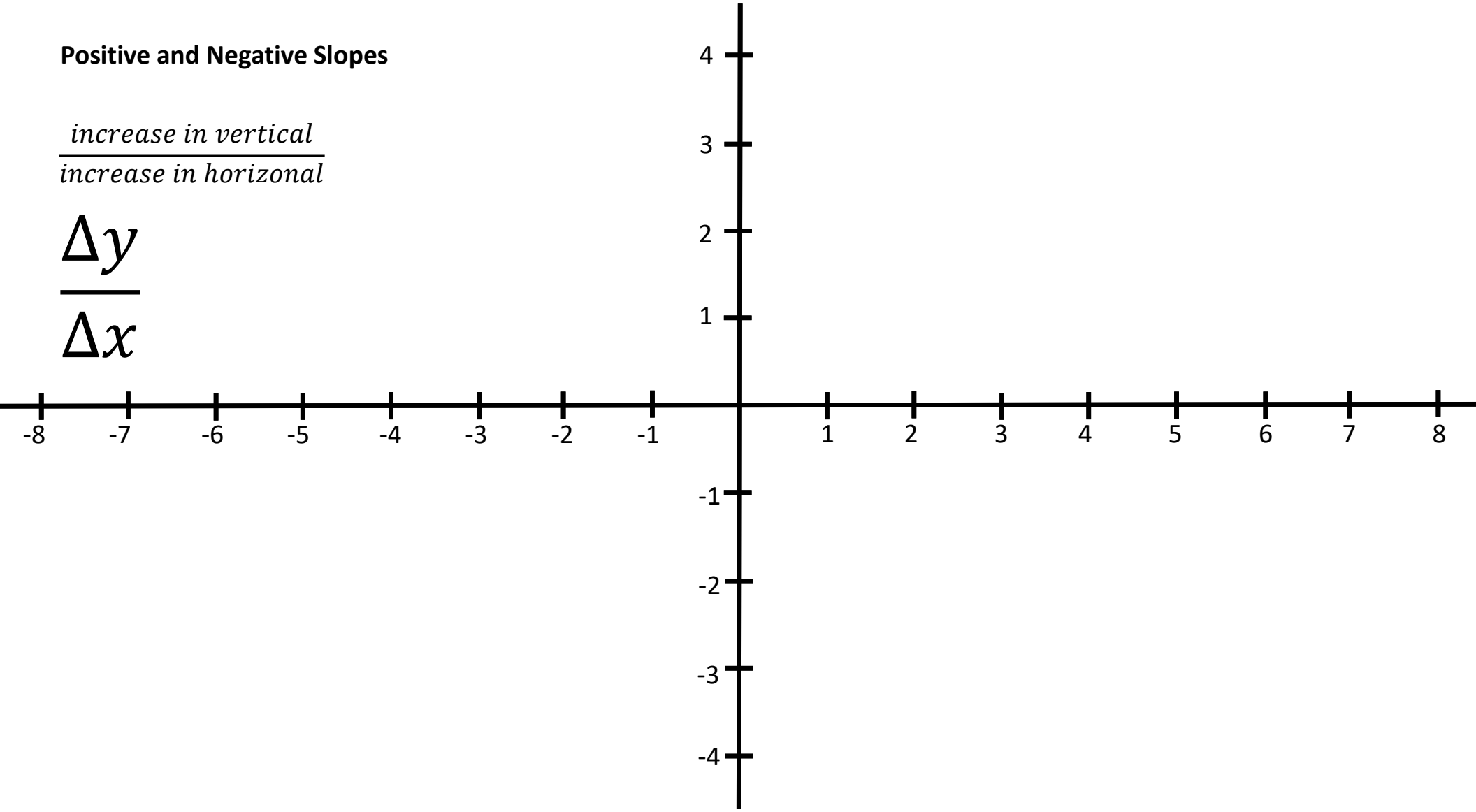
$$\frac{\Delta y}{\Delta x}$$



Positive and Negative Slopes

$$\frac{increase\ in\ vertical}{increase\ in\ horizonal}$$

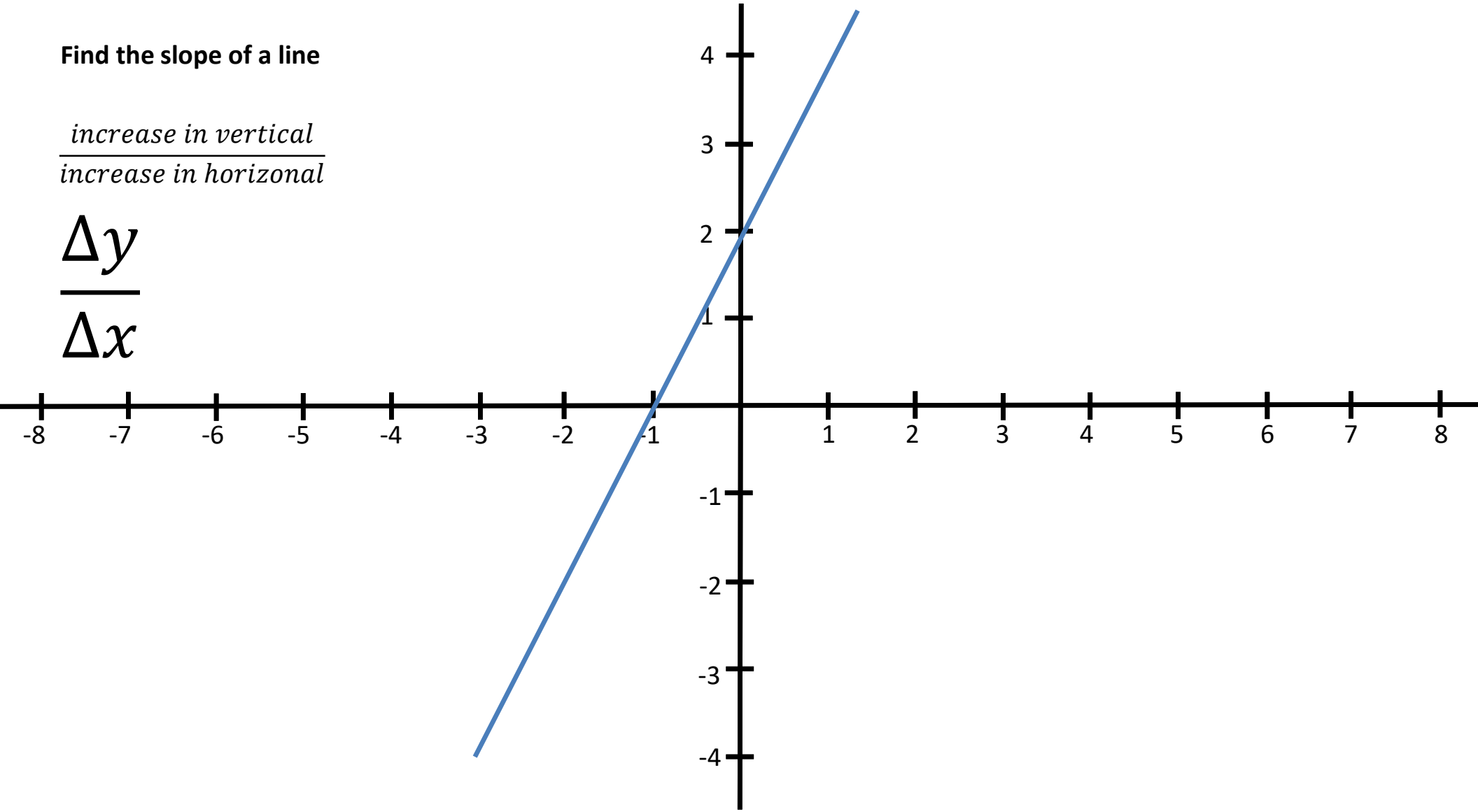
$$\frac{\Delta y}{\Delta x}$$



Find the slope of a line

$$\frac{increase\ in\ vertical}{increase\ in\ horizonal}$$

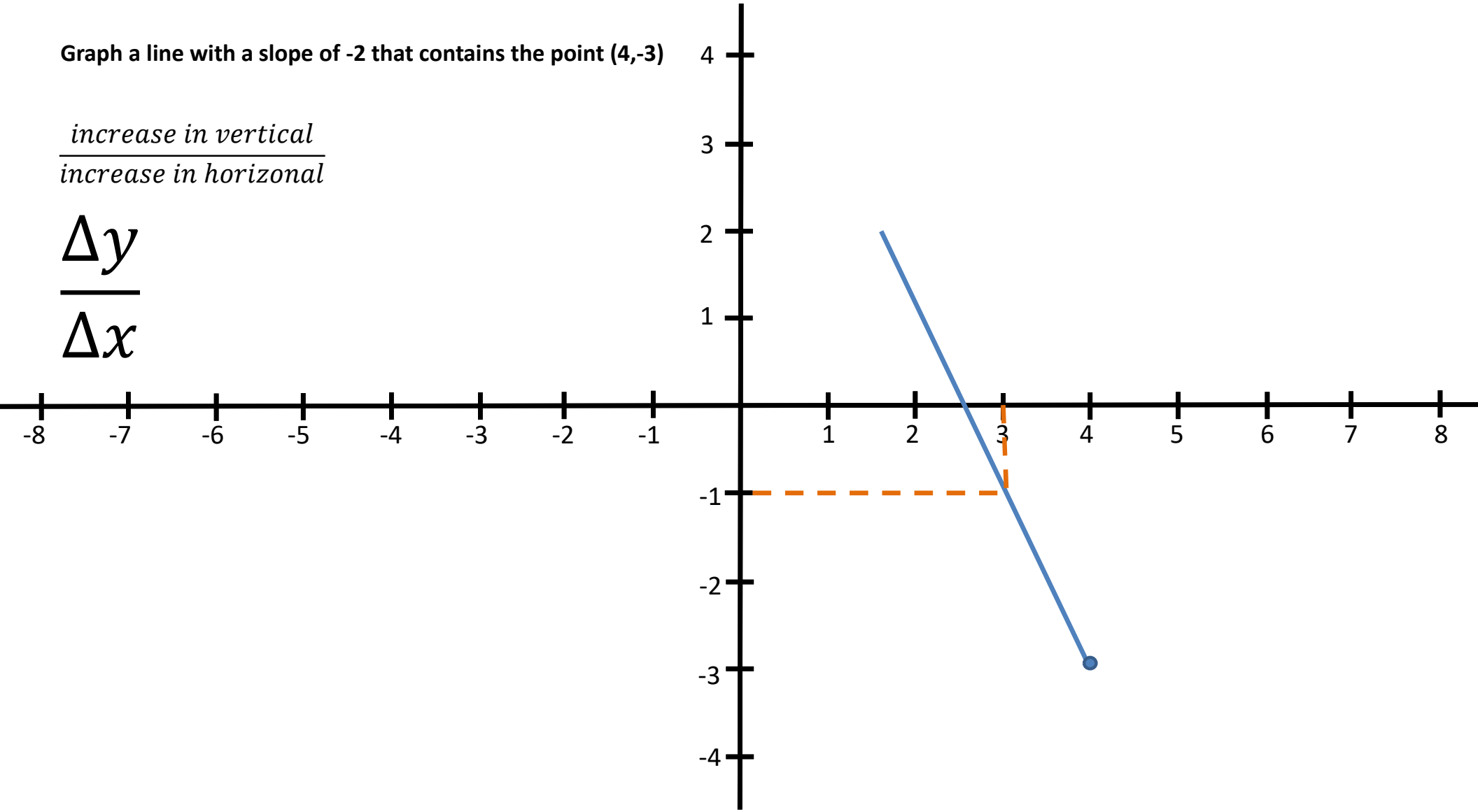
$$\frac{\Delta y}{\Delta x}$$



Graph a line with a slope of -2 that contains the point (4,-3)

$$\frac{increase\ in\ vertical}{increase\ in\ horizonal}$$

$$\frac{\Delta y}{\Delta x}$$



What is the slope of the line that contains following points

$$\frac{increase\ in\ vertical}{increase\ in\ horizonal}$$

$$\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

X	Y
2	4
3	1
4	-2
5	-5

What is the slope of the line that contains following points

$$\frac{increase\ in\ vertical}{increase\ in\ horizonal}$$

$$\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

X	Y
7	-1
-3	-1

What is the equation of the horizontal line through (-2,6)

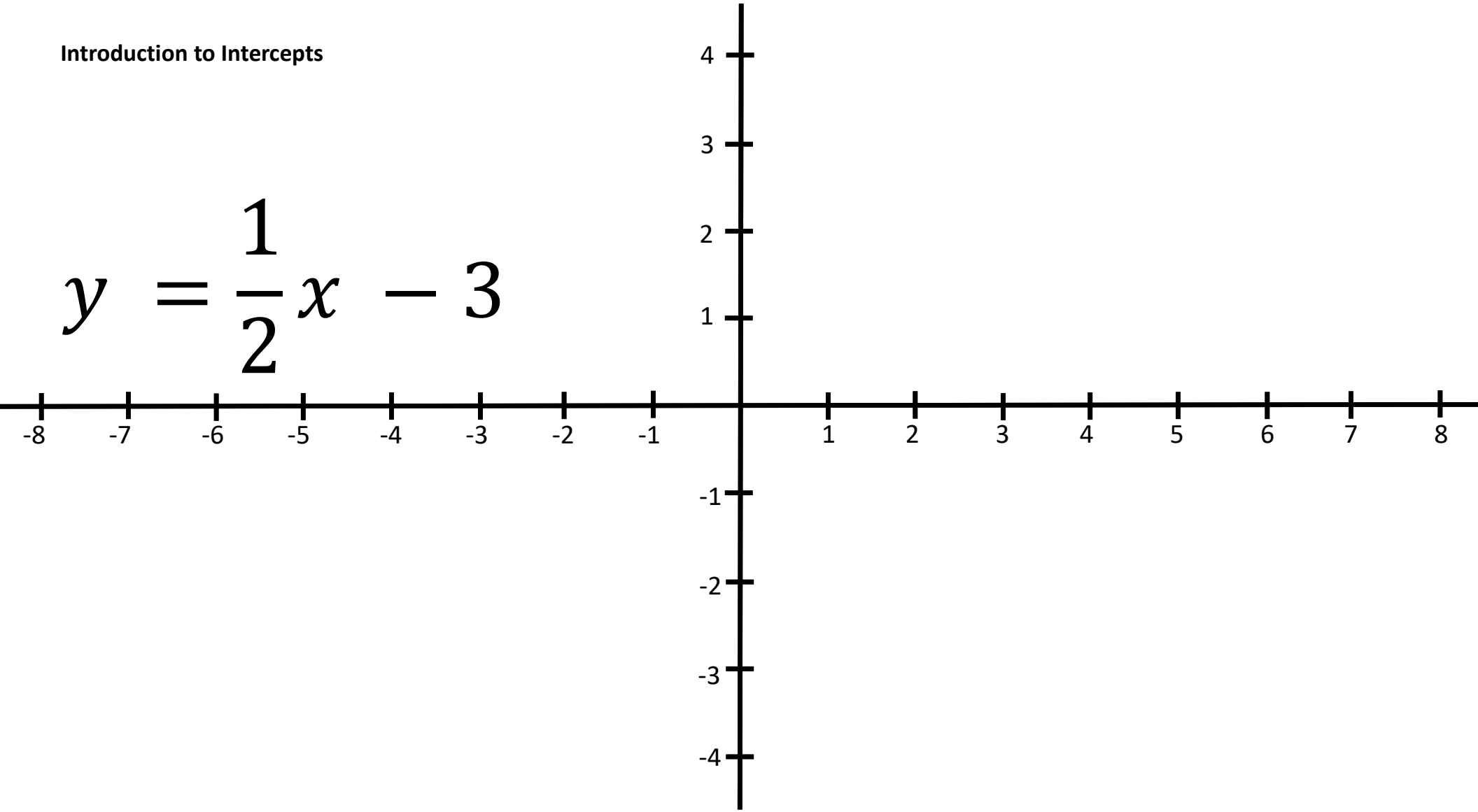
What is the slope of the line where y = -4?

What is the slope of the line where x = -3?

What is the equation of the vertical line through (-4, -2)?

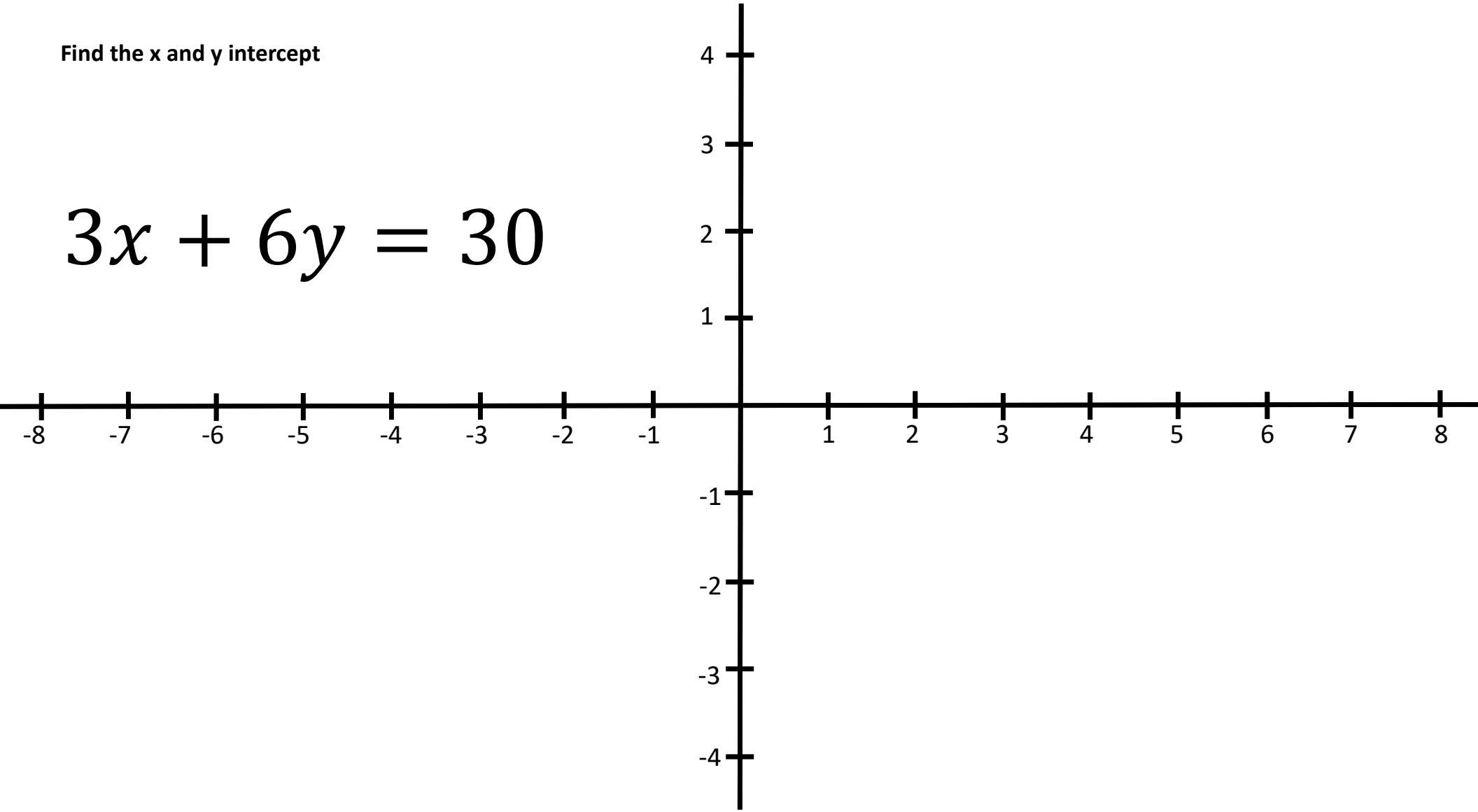
Introduction to Intercepts

$$y = \frac{1}{2}x - 3$$



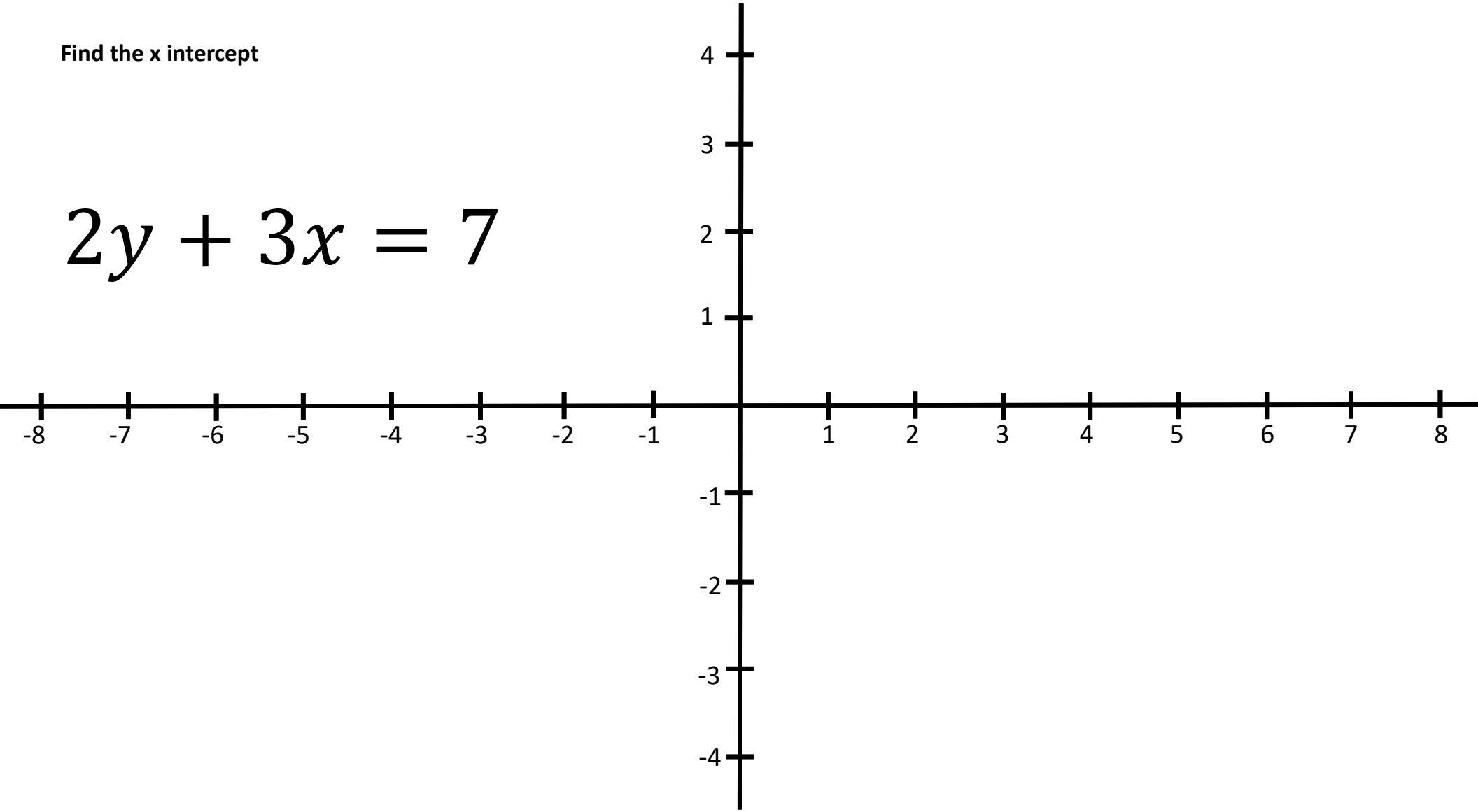
Find the x and y intercept

$$3x + 6y = 30$$



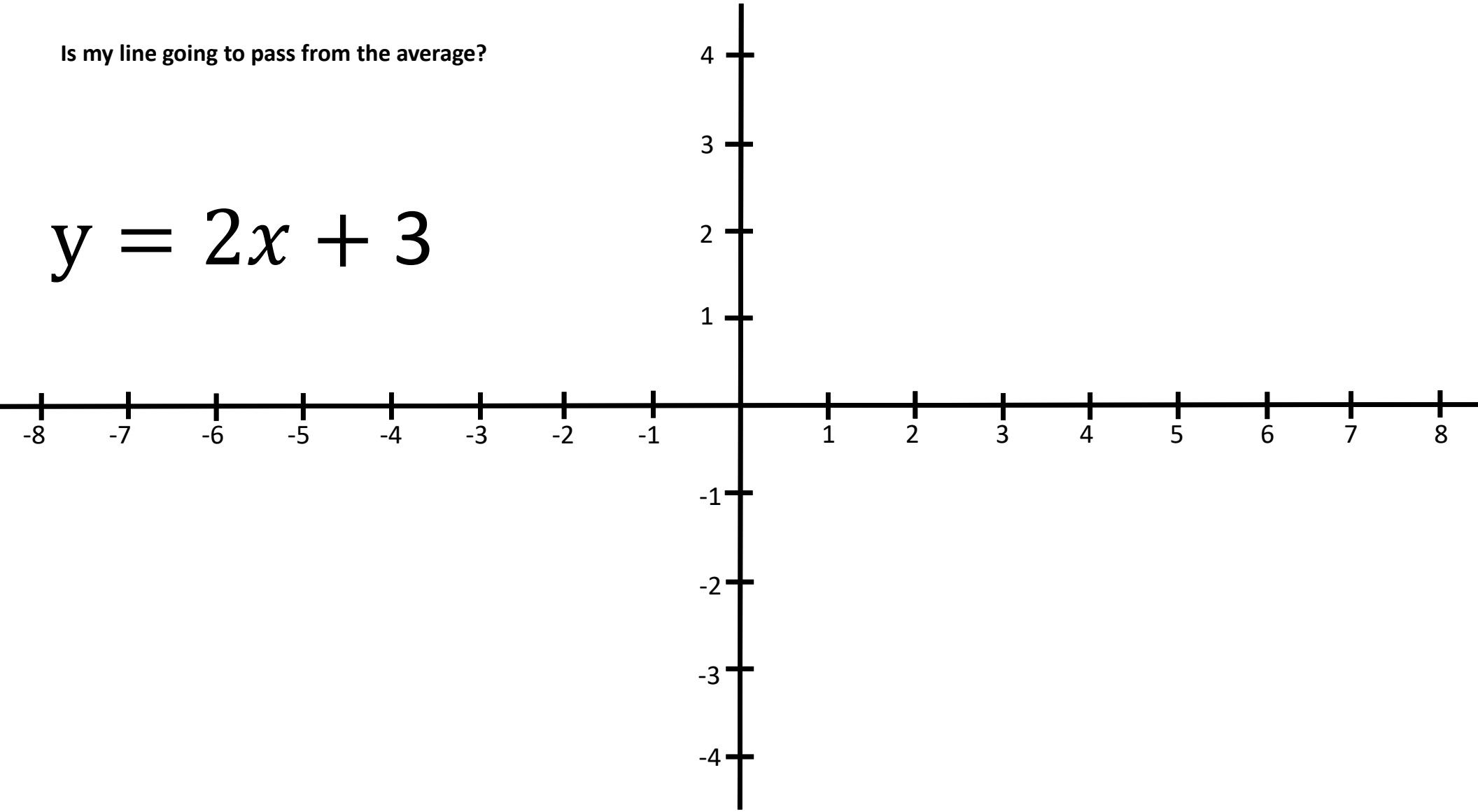
Find the x intercept

$$2y + 3x = 7$$



Is my line going to pass from the average?

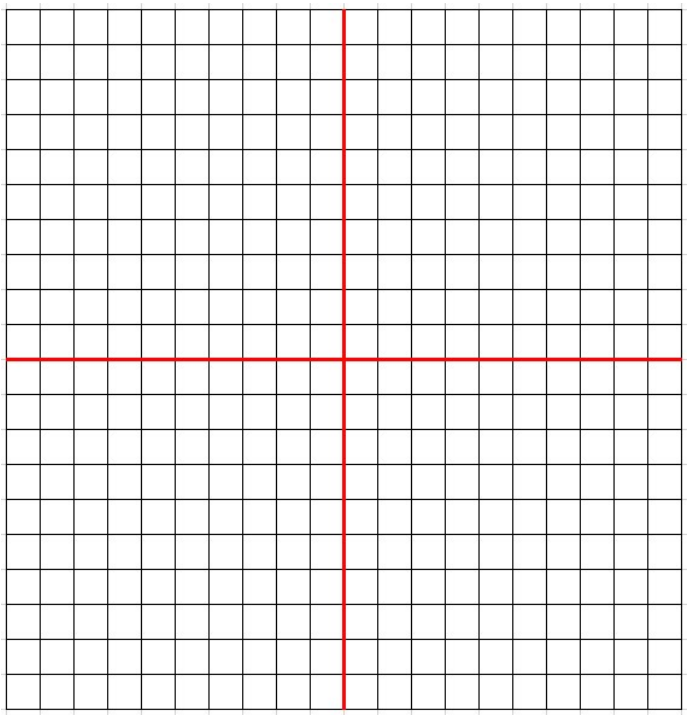
$$y = 2x + 3$$



Scatter Plots

- Nikki wanted to see if there is a connection between the time a given exam takes place and the score of this exam.
- She collected the data about exams from the previous year.

- Plot the data in the scatter plot

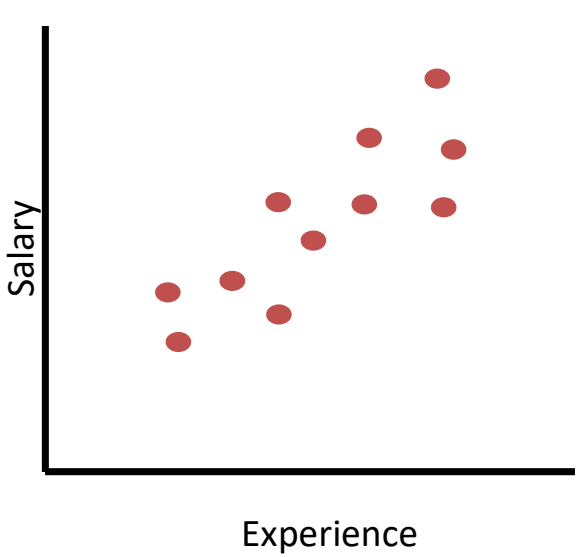


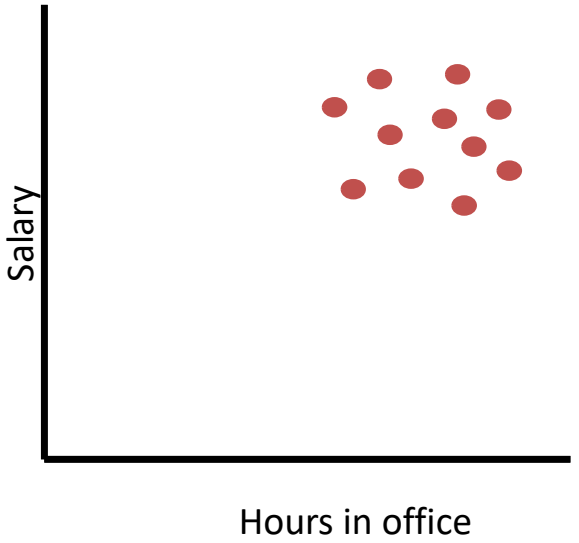
Class	English	History	Maths	Computers	Arts	Physics	Biology	Chemistry
Period	2	6	1	4	4	3	2	3
Score	9	9	7	6	8	6	7	9

Scatter Plots

The scatter plots show the salary of the employees based on the two different features that is Experience and how many hours you stay in office.

Choose the best description of the relationship between the graphs.



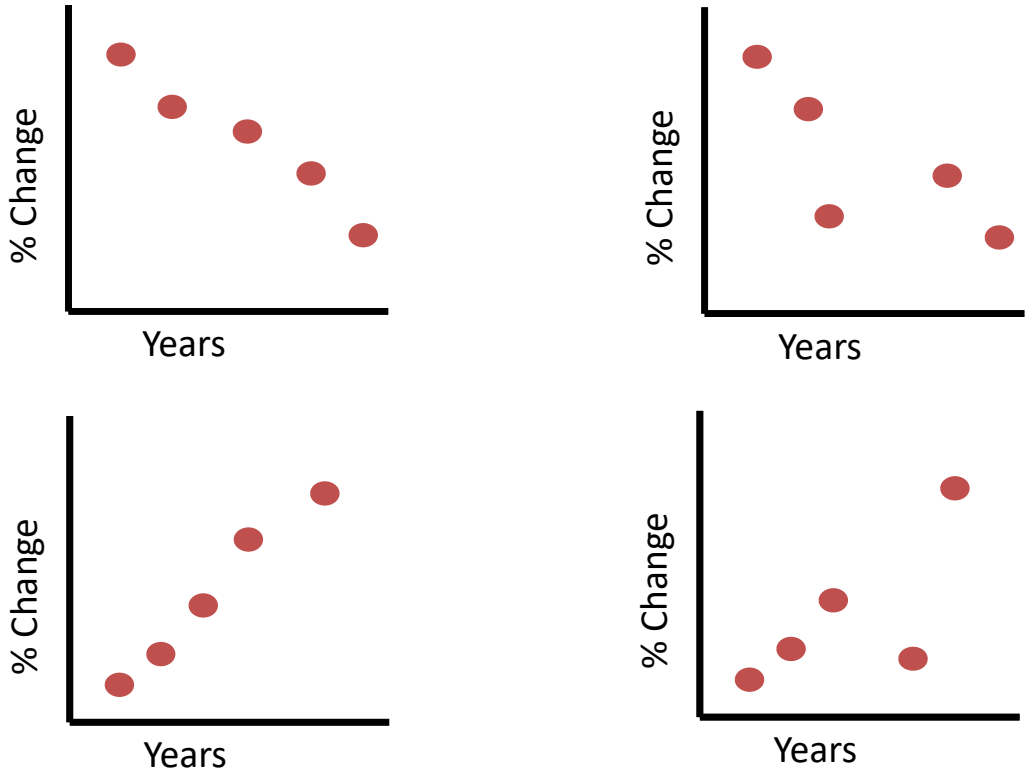


- There is a negative linear relationship between salary and experience and positive linear relationship between Salary and hours in office
- There is a non linear relationship between Salary and Experience, and a negative relationship between Salary and Hours in office
- Both the charts show positive linear trends of approximately equal strength
- There is a positive linear relationship between Salary and Experience and no relationship between Salary and Hours in office

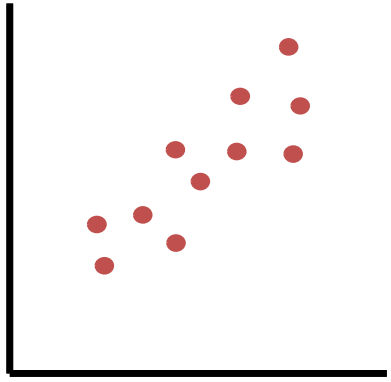
Scatter Plots

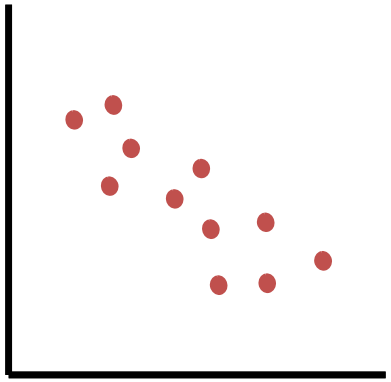
The percent of unemployment recorded every few years since 1980, suggests a negative linear association with no outliers. On average, the percent drops 1 point per year.

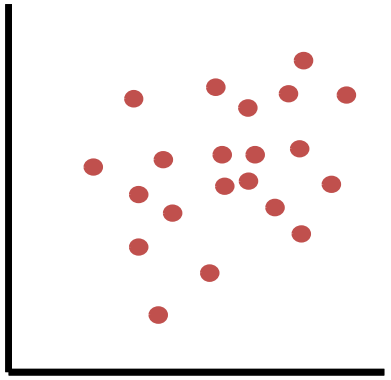
Which of the following suits the above description?

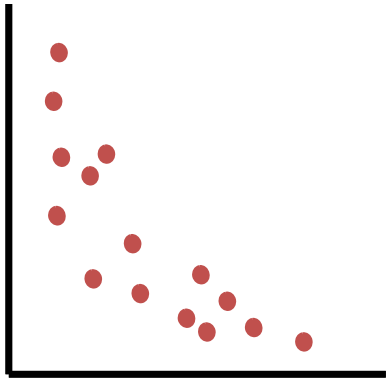


Scatter Plots







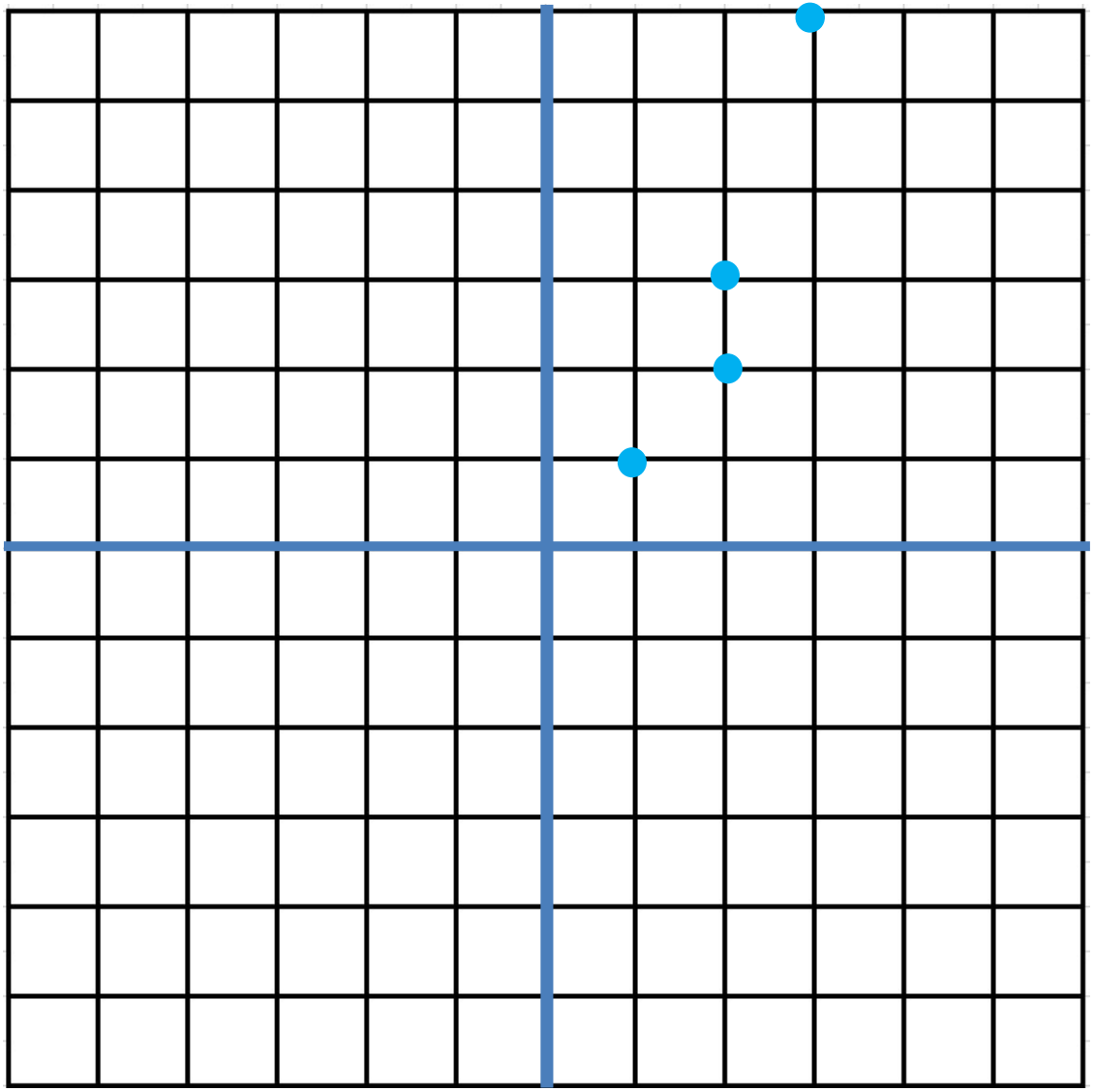


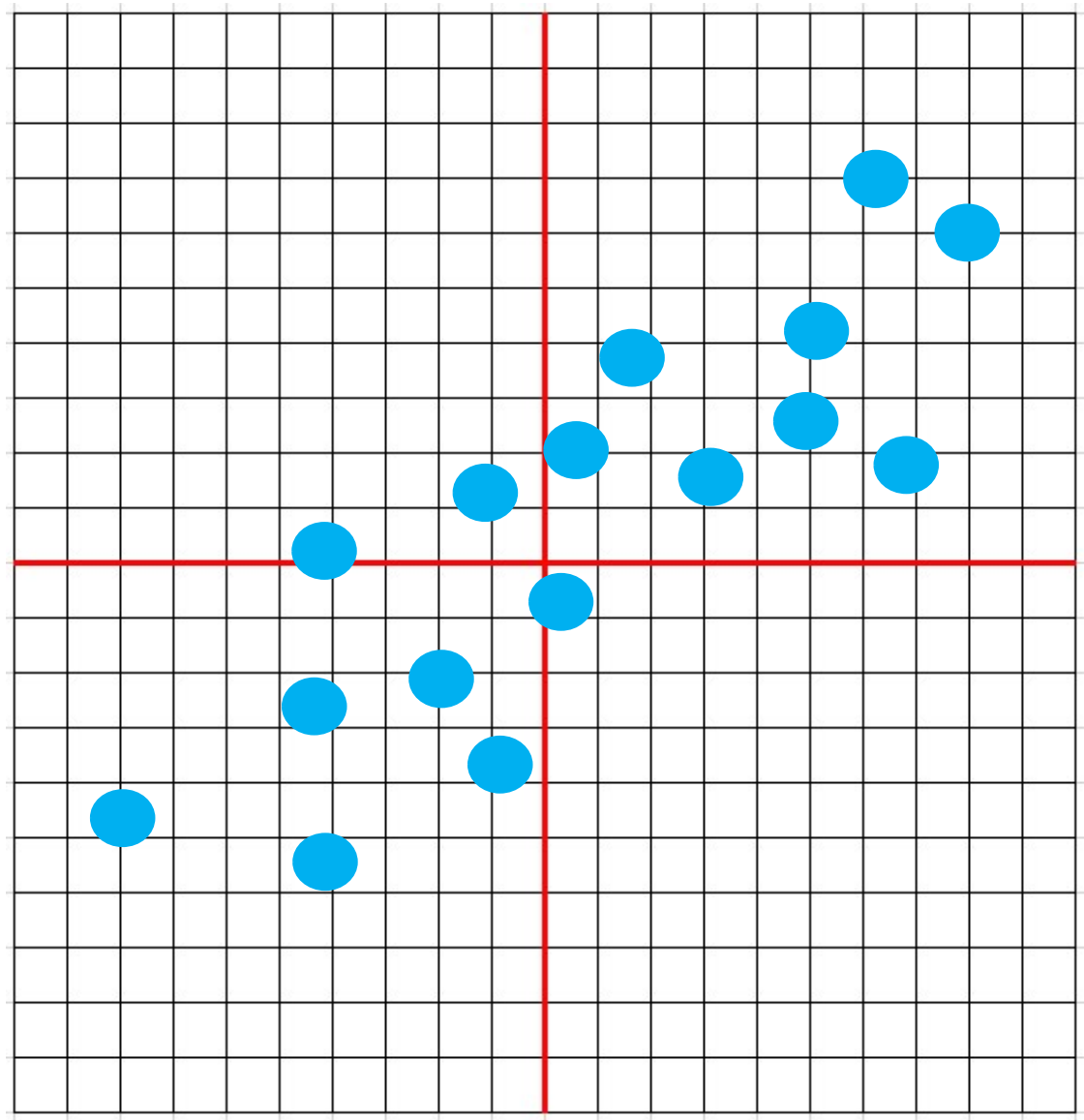
Covariance and Covariance Matrix

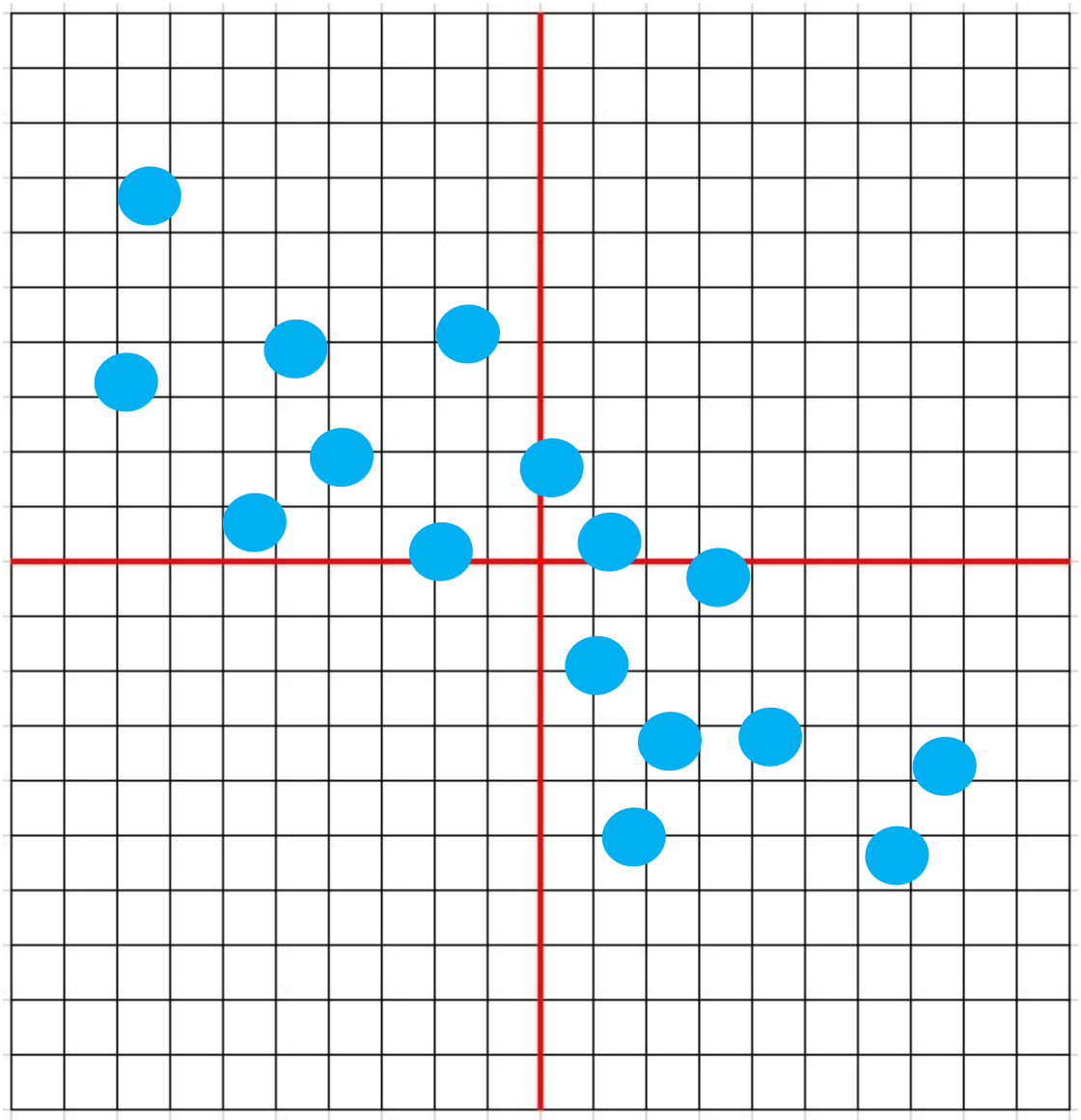
Correlation Coefficients

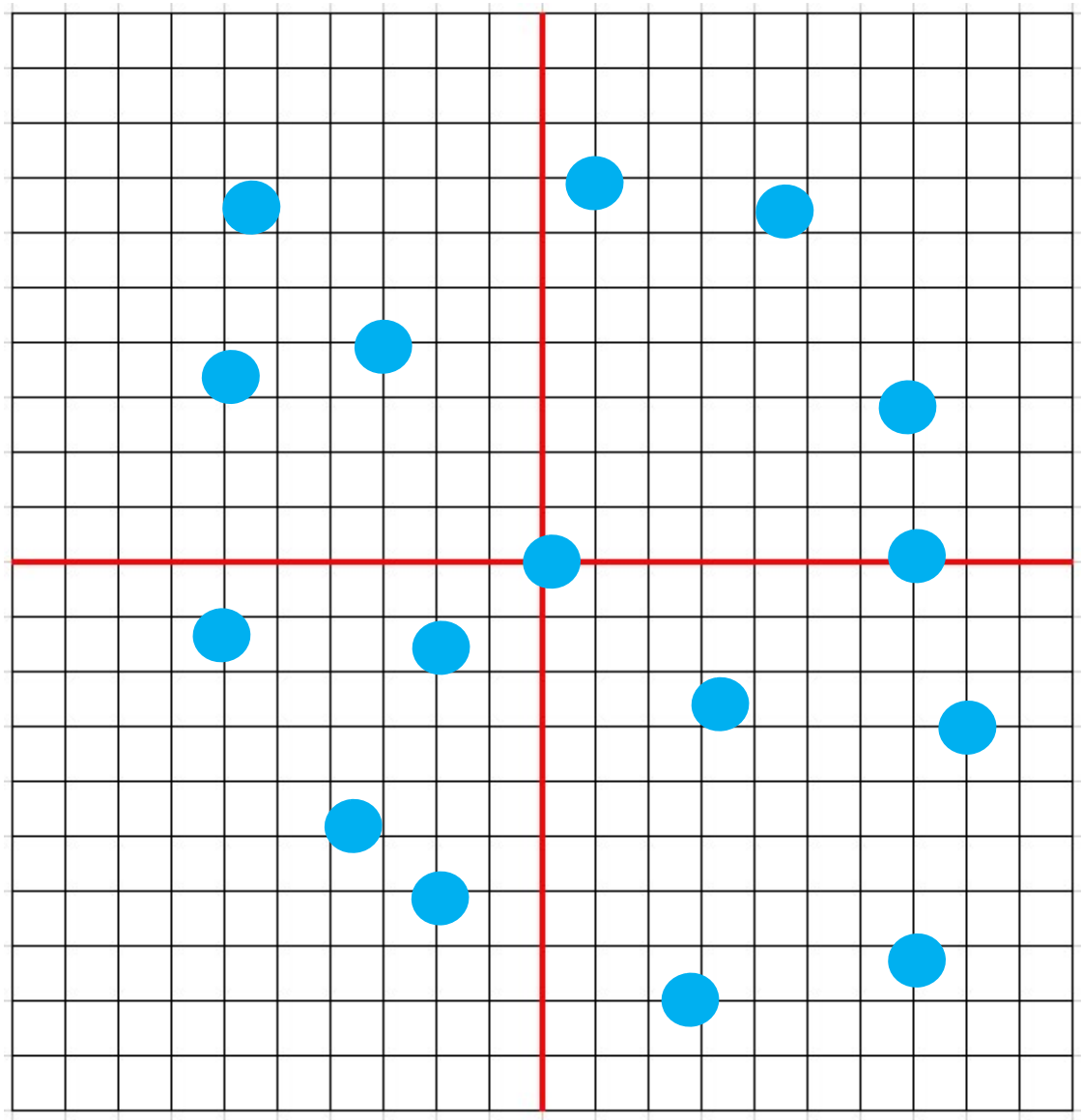
x	y
1	1
2	2
2	3
3	6

Coordinates	mean	Standard deviation
x	2	0.816
y	3	2.160



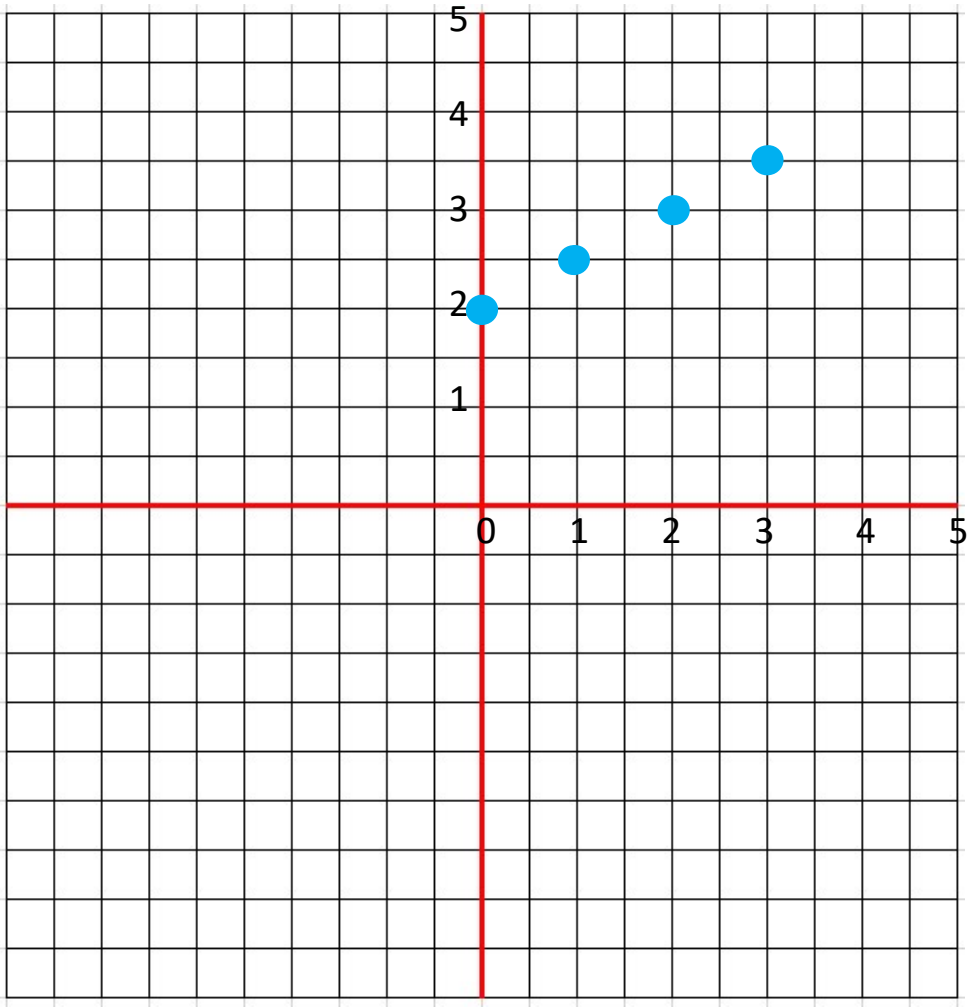






Spearman Rank Correlation Coefficient

Estimating with Linear Regression



Which of these linear equations best describes the given model?

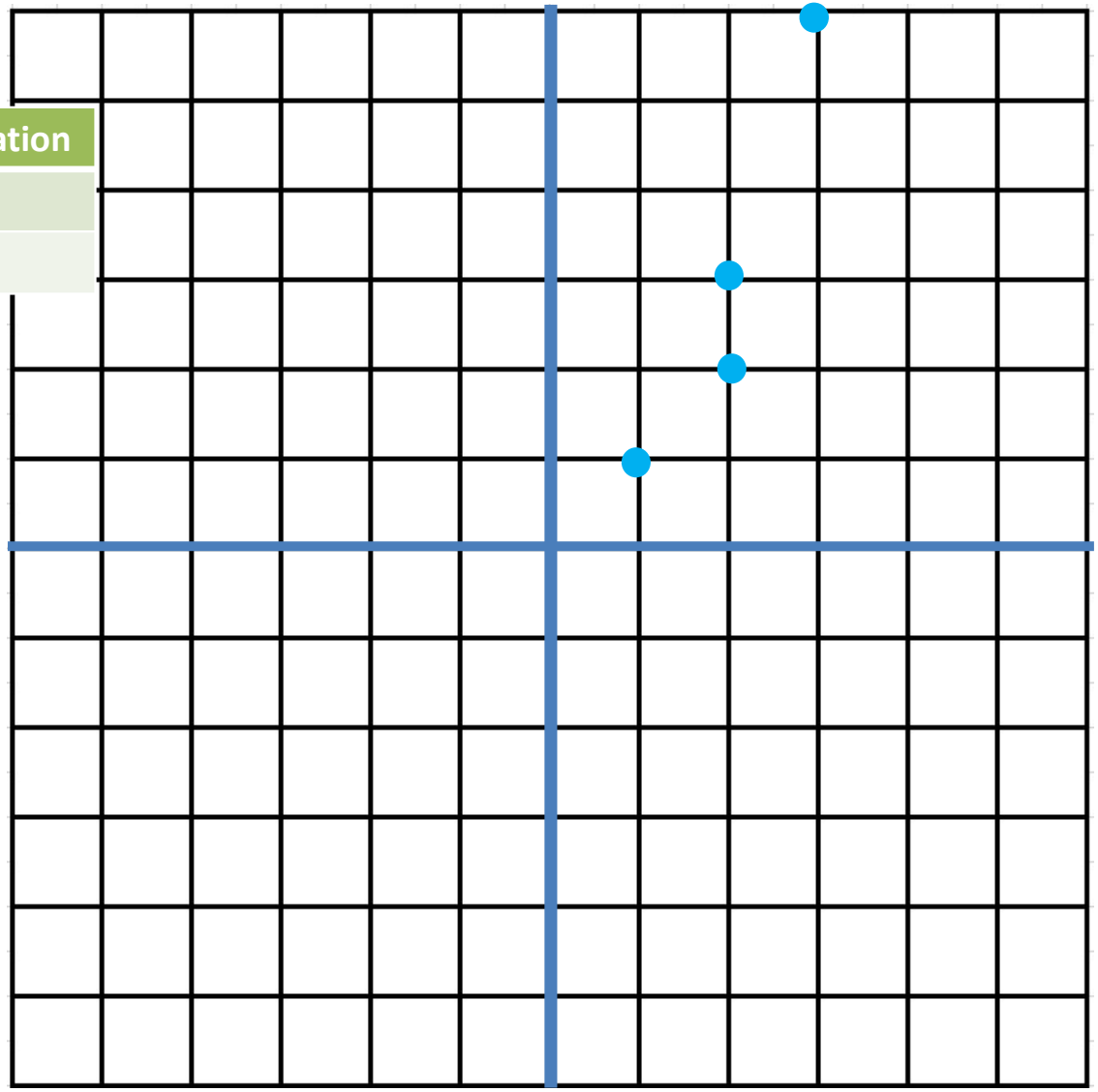
- Y = 0.5x + 2
- Y = x + 2
- Y = -0.5x + 2

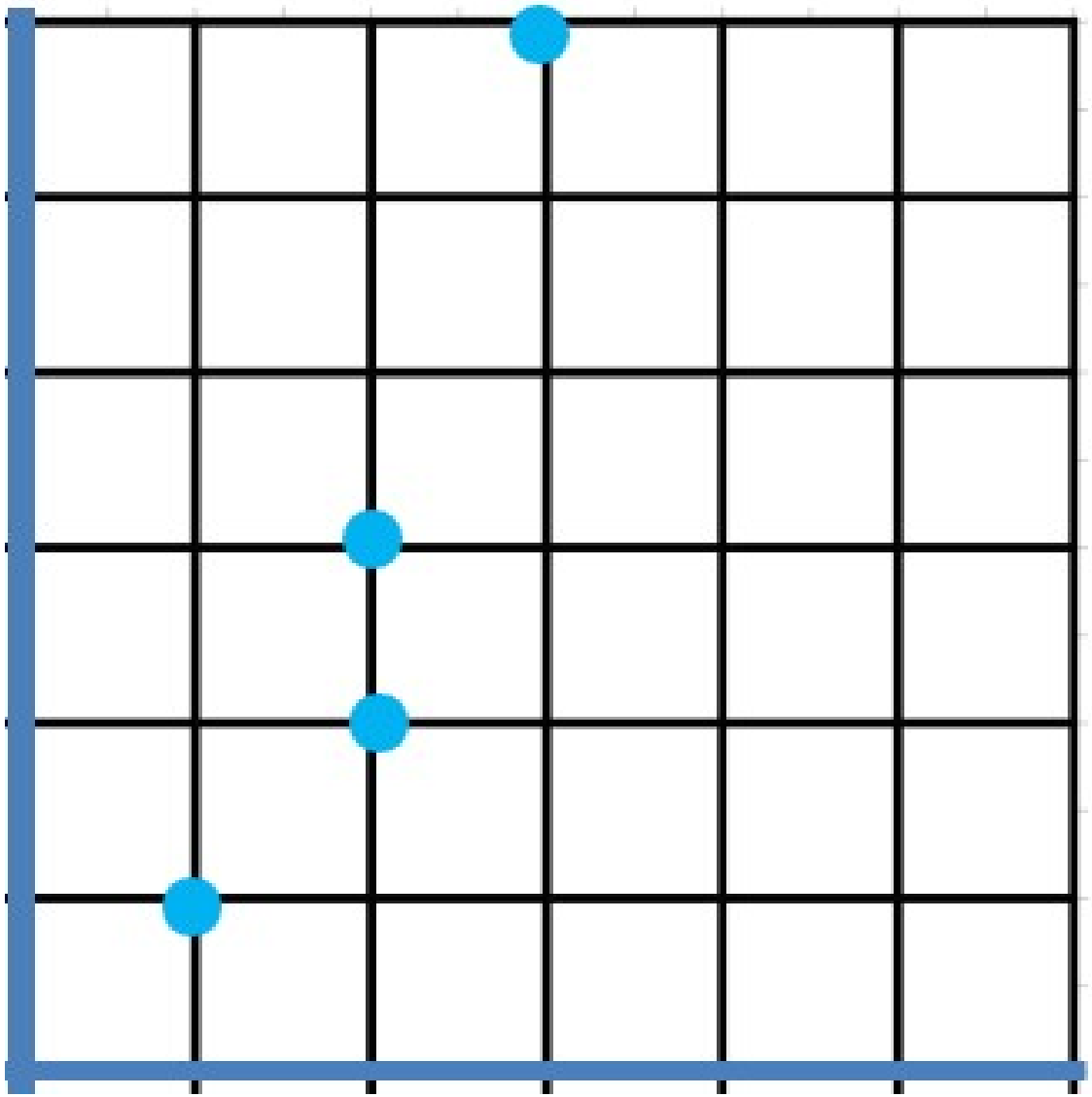
Based on the equation, estimate the score for x = 7.5

Calculating the equation of a regression line

x	y
1	1
2	2
2	3
3	6

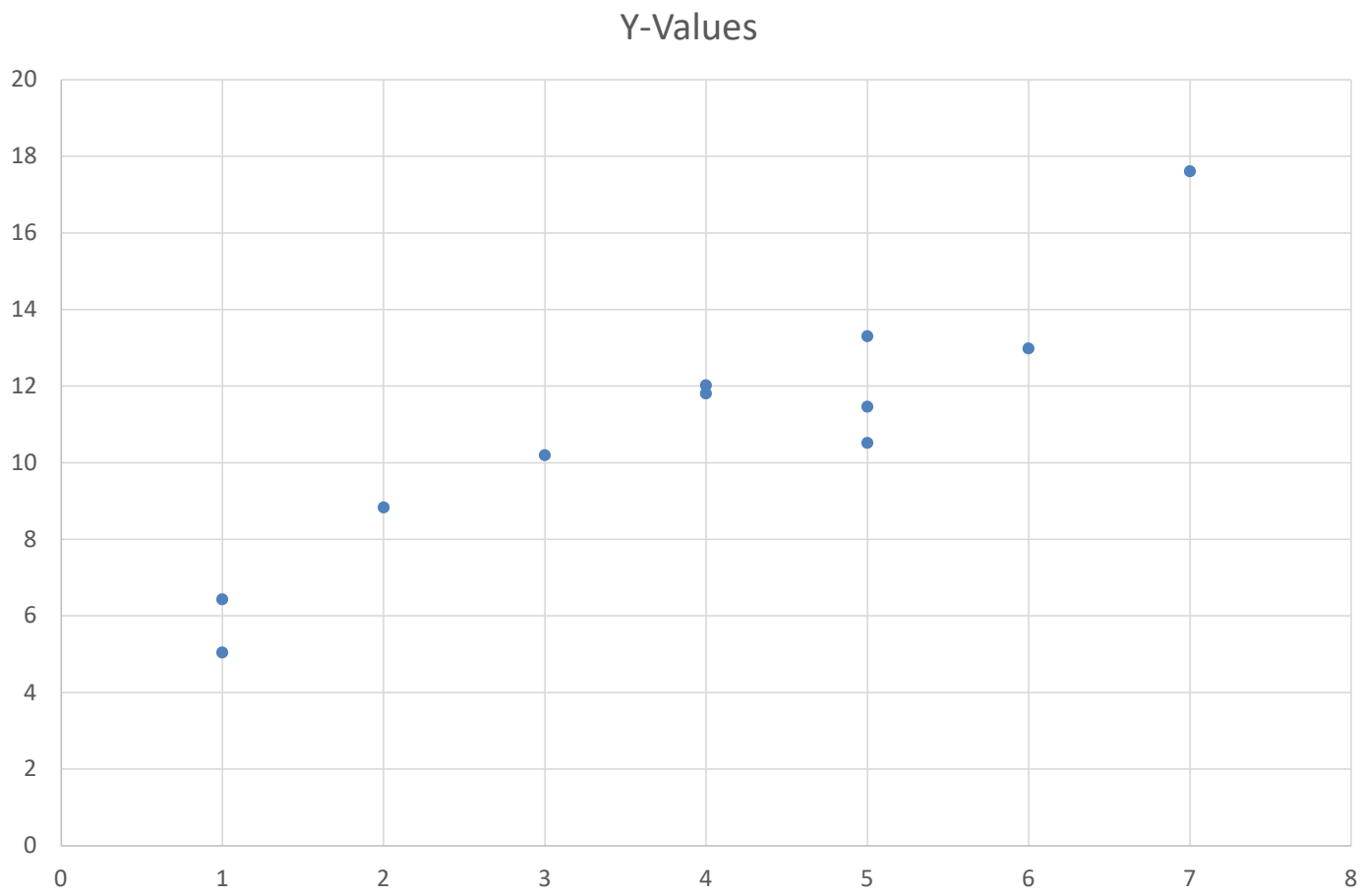
Coordinates	mean	Standard deviation
x	2	0.816
y	3	2.160



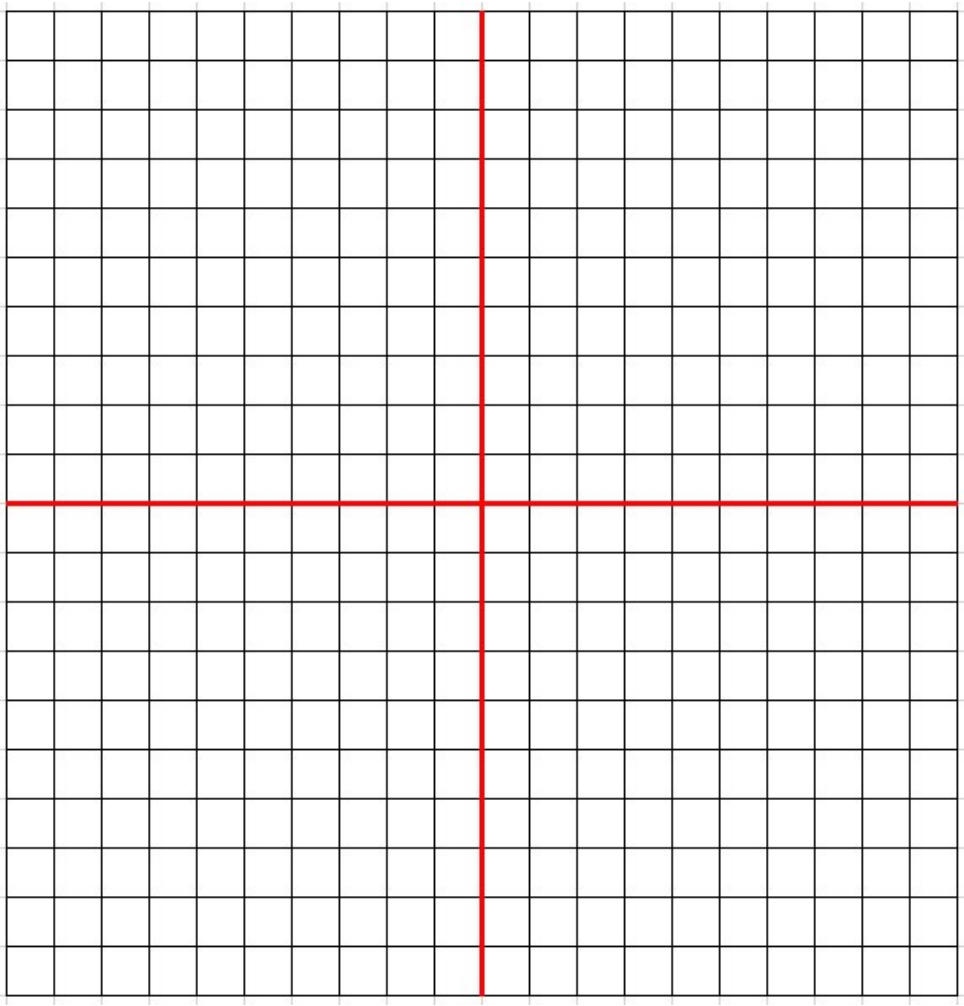


Fitting an line

X	Y
1	6.43
2	8.83
3	10.20
4	11.81
5	13.30
5	10.52
6	12.98
7	17.61
5	11.46
4	12.02
1	5.04



Introduction to residuals and least square regression



Calculating Residuals

Aarti rents bicycle to tourists. She recorded the height (in cm) of each customer and the frame size (in cm) of the bicycle that customer rented.

After plotting her results, Aarti noticed that the relationship between the two variables was fairly linear, so she used the data to calculate the following least squares regression equation for predicting bicycle frame size from the height of the customer.

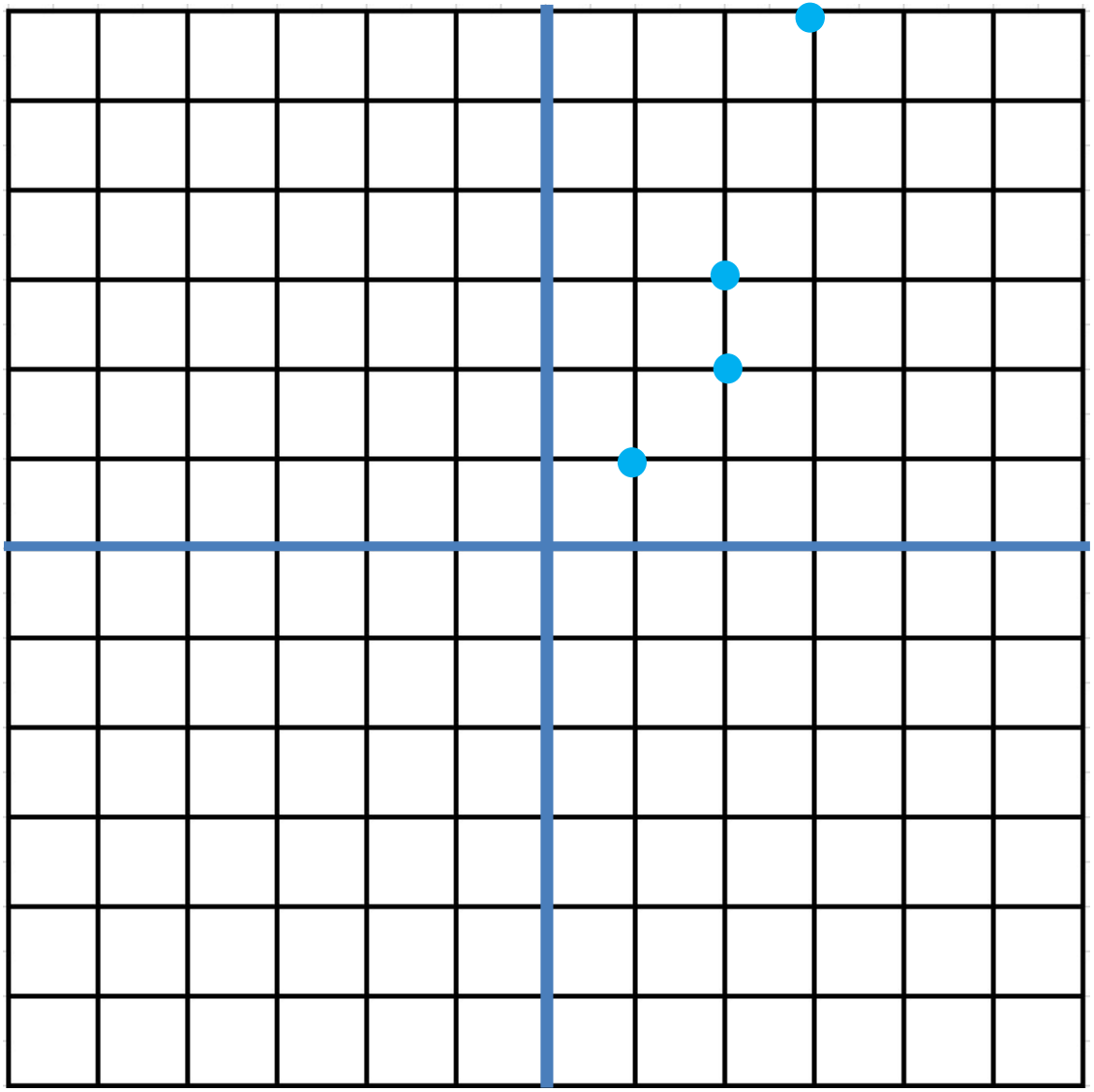
$$\hat{y} = \frac{1}{3} + \frac{1}{3}x$$

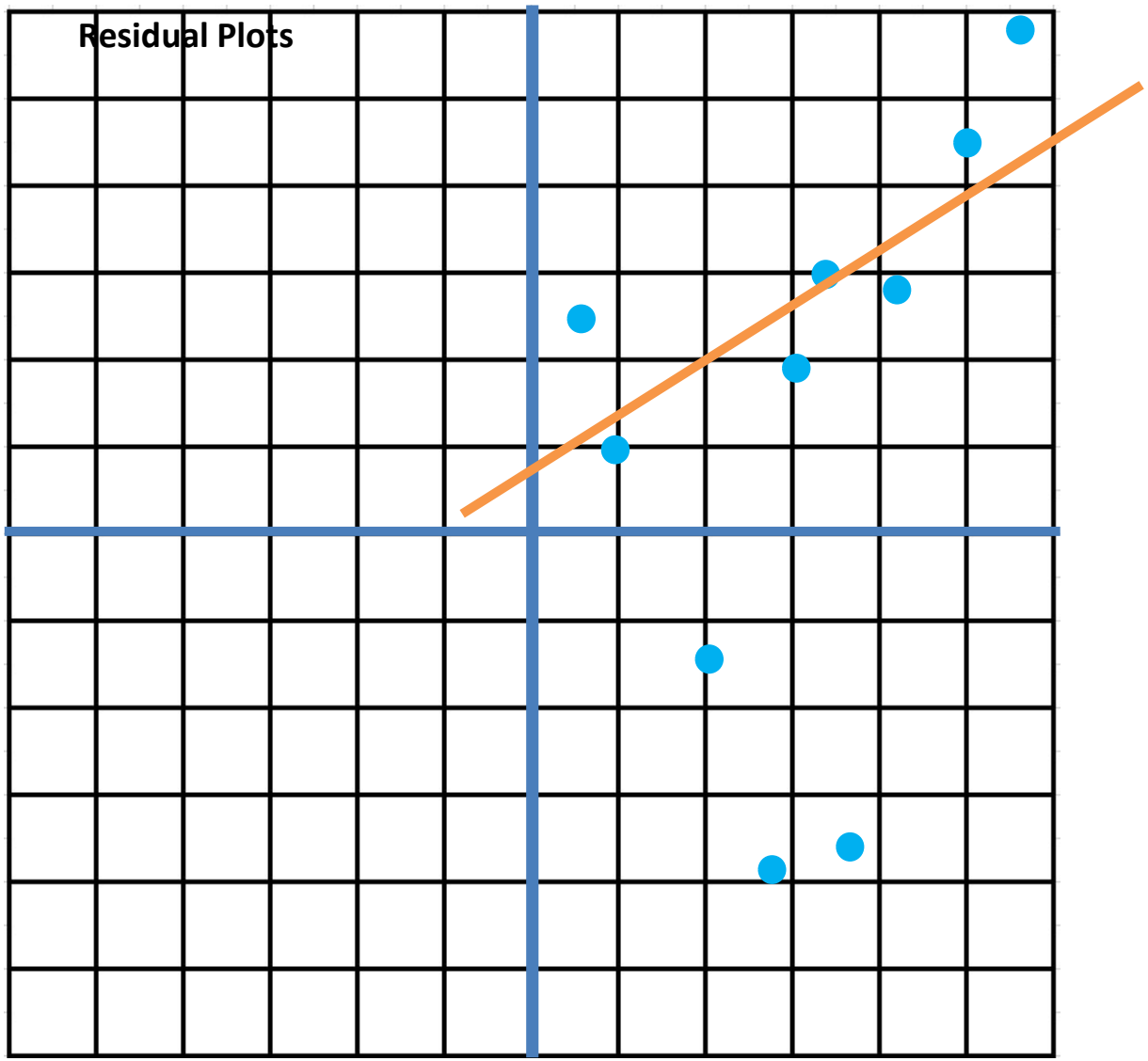
What is the residual of a customer with a height of 150 cm who rents a bike with a 50 cm frame?

Residual Plots

x	y
1	1
2	2
2	3
3	6

Coordinates	mean	Standard deviation
x	2	0.816
y	3	2.160




Residual Plots

$$R^2 = 1 - \frac{SSE}{SST}$$

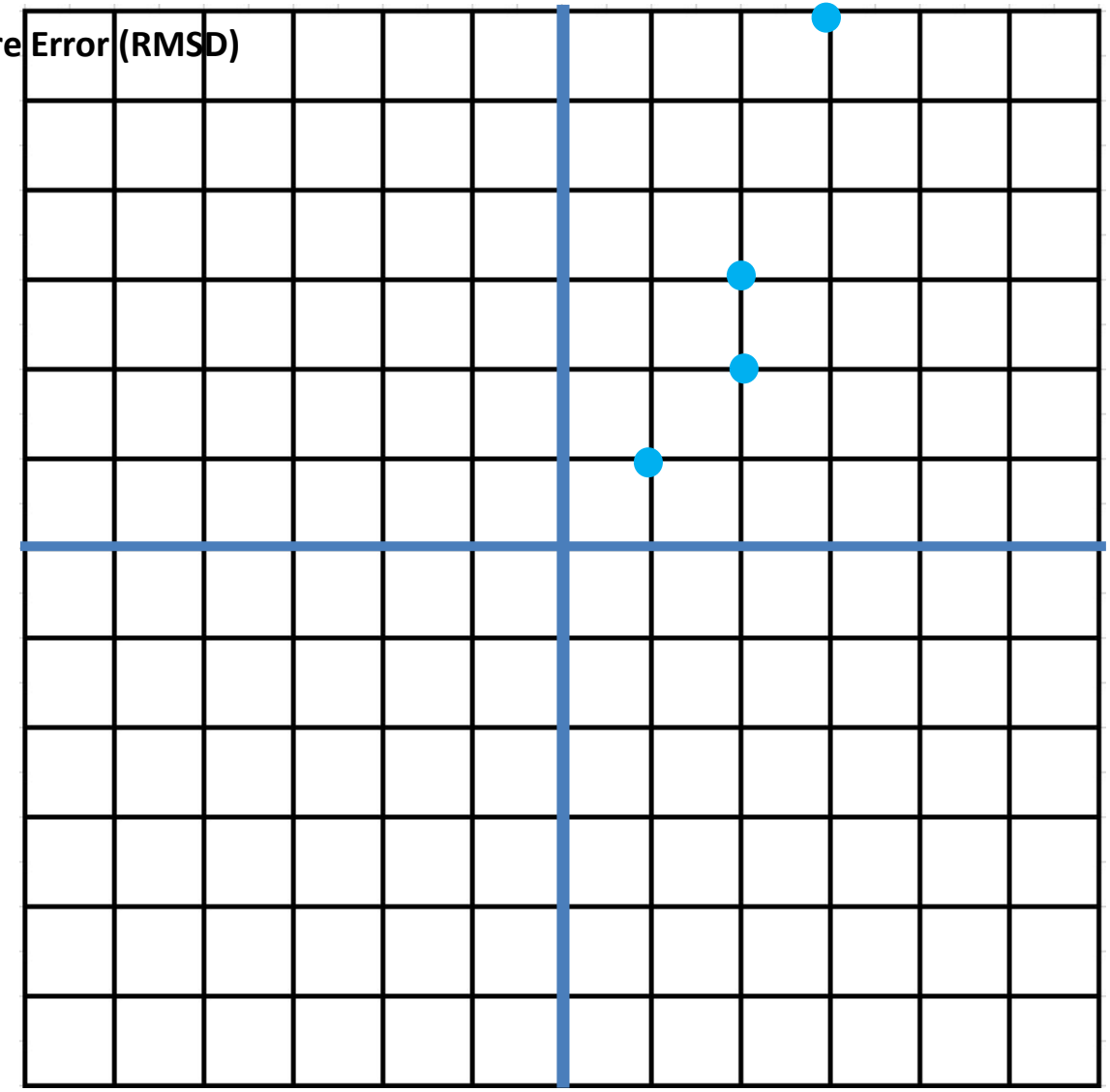
$Experience$	$Salary\ (y)$	$Avg\ Salary$	$S - SAvg$	$(S - SAvg)^2$ SST – Sum Squared Total	$\hat{y} = 2250 * Experience + 5750$	$y - \hat{y}$	$(y - \hat{y})^2$ SSE - Sum Squared errors
2	10000	12500	-2500	6250000	10250	-250	62500
3	13000	12500	500	250000	12500	500	250000
4	14500	12500	2000	4000000	14750	250	62500
				1,05,00,000			**3,75,000**

$$R^2 = 1 - \frac{375000}{10500000} = 0.9642$$

Standard Deviation of residuals or Root Mean Square Error (RMSD)

x	y
1	1
2	2
2	3
3	6

Coordinates	mean	Standard deviation
x	2	0.816
y	3	2.160



Sum of Square Errors

$$\sum (Y_A - Y_P)^2$$

$$Y_P = \beta_1 X_i + \beta_0$$

$$\sum (Y_i - Y_p)^2 = \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Derive Equation for the Estimators

1) Take the partial derivatives with respect to $\beta_0 \; and \; \beta_1$

2) Set the partial derivatives equal to 0

3) Solve for $\beta_0 \; and \; \beta_1$

Derive Equation for the Estimators

Taking the partial derivative with respect to β_0

$$\frac{\partial}{\partial \beta_0} \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$$

$$\sum \frac{\partial}{\partial \beta_0} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

$$\sum 2 \left(Y_i - (\beta_0 + \beta_1 X_i)\right)(-1)$$

$$-2 \sum \left(Y_i - (\beta_0 + \beta_1 X_i)\right)$$

Derive Equation for the Estimators

Taking the partial derivative with respect to β_1

$$\frac{\partial}{\partial \beta_1} \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$$

$$\sum \frac{\partial}{\partial \beta_1} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

$$\sum 2 \left(Y_i - (\beta_0 + \beta_1 X_i)\right)(-X_i)$$

$$-2 \sum X_i \left(Y_i - (\beta_0 + \beta_1 X_i)\right)$$

Derive Equation for the Estimators

Setting the partial derivatives equal to 0

$$-2\sum\left(Y_i - (\beta_0 + \beta_1 X_i)\right) = 0$$

$$-2\sum X_1\left(Y_i - (\beta_0 + \beta_1 X_i)\right) = 0$$

Derive Equation for the Estimators

Solving for β_0

$$-2\sum\left(Y_i - (\beta_0 + \beta_1 X_i)\right) = 0$$

$$\beta_0 = Y_a - \beta_1 X_a$$

Derive Equation for the Estimators

Solving for β_1

$$-2\sum X_i\left(Y_i-(\beta_0+\beta_1 X_i)\right)=0$$

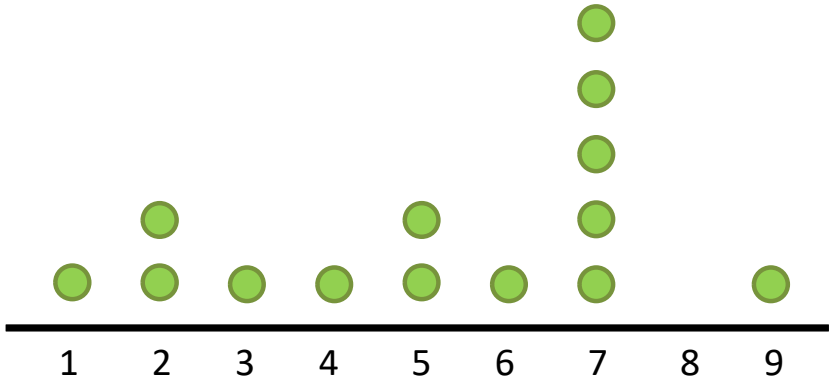
Substitute in the equation $\beta_0 = Y_a - \beta_1 X_a$

$$\beta_1 = \frac{\sum X_i(Y_i - Y_a)}{\sum X_i\,(X_i - X_a)}$$

$$\beta_1 = \frac{\sum (X_i - X_a)\,(Y_i - Y_a)}{\sum\,(X_i - X_a)^{\,2}}$$

Modeling data distributions

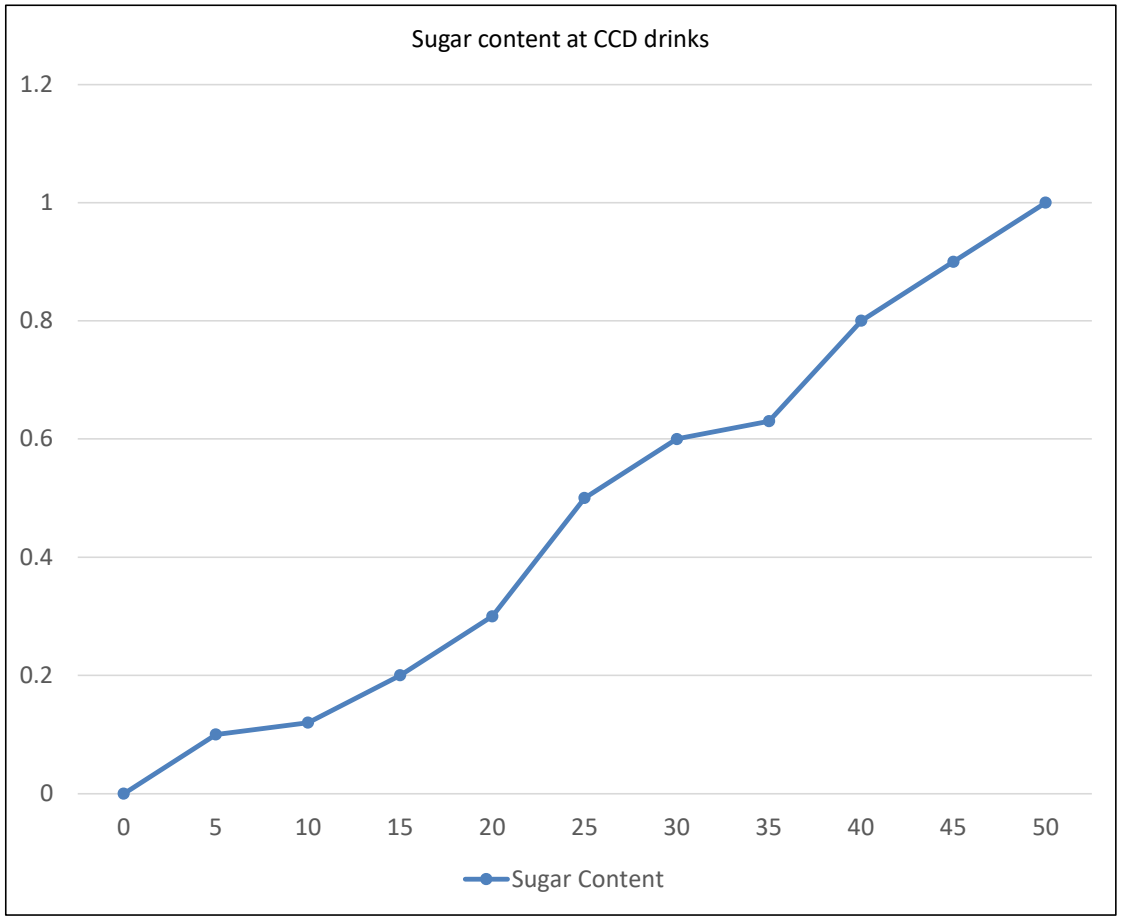
A dot plot shows the number of hours of daily study time for 14 students. Each dot represents a student.



Which of the following is the closest estimate to the percentile rank for the student with a daily study time of 6 hours?

1. 40 Percentile
2. 55 Percentile
3. 70 Percentile
4. 85 Percentile

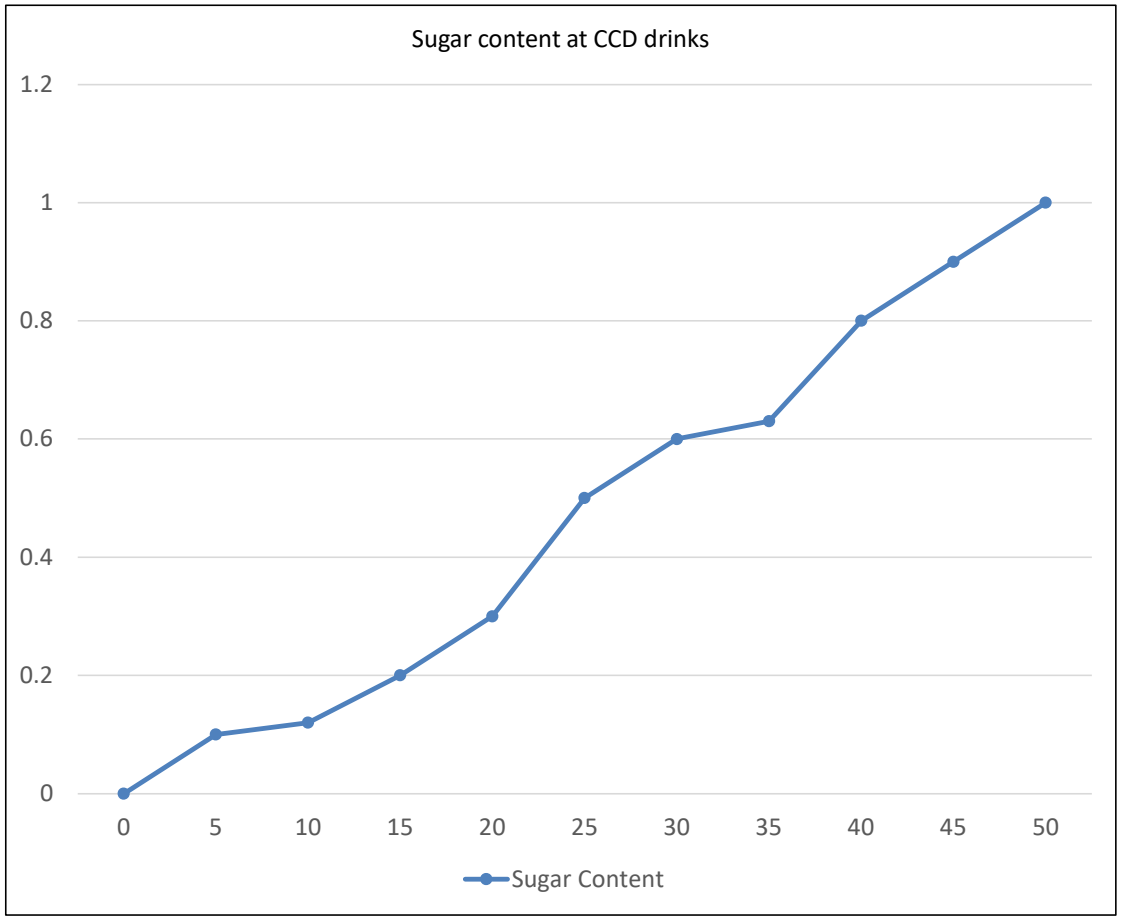
Nutritionists measured the sugar content (in grams) for 32 drinks at CCD. A cumulative relative frequency graph for the data is shown below



An cold coffee has 15 grams of sugar.

Estimate the percentile of this drink to nearest whole percent.

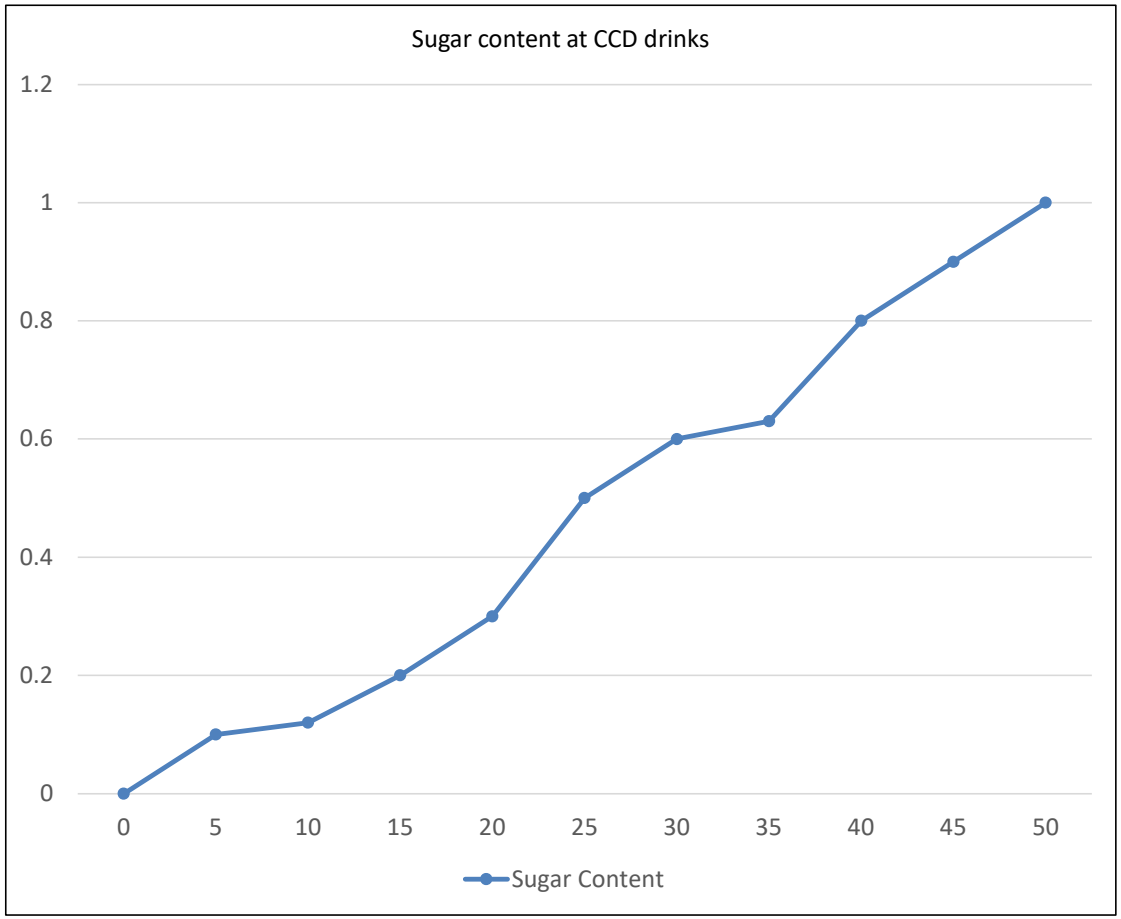
Nutritionists measured the sugar content (in grams) for 32 drinks at CCD. A cumulative relative frequency graph for the data is shown below



Estimate the median of the distribution of drinks.

Median in grams?

Nutritionists measured the sugar content (in grams) for 32 drinks at CCD. A cumulative relative frequency graph for the data is shown below



What is the best estimate for the interquartile range (IQR) of the distribution?

1. 10 grams
2. 20 grams
3. 30 grams
4. 40 grams

Z - Scores

The marks on a statistics exam for a high school are normally distributed with mean of 81 and standard deviation of 6.3
Calculate the z-score for each of the following exam marks.

1. 63
2. 83
3. 93
4. 100

Z - Scores

John appeared for two exams. Here are some summary statistics for each exam.

Section	Mean	Standard Deviation
Exam 1	$\mu = 151$	$\sigma = 10$
Exam 2	$\mu = 25.1$	$\sigma = 6.4$

John scored 172 in Exam 1 and 37 in exam 2. Which exam did he do relatively better on?

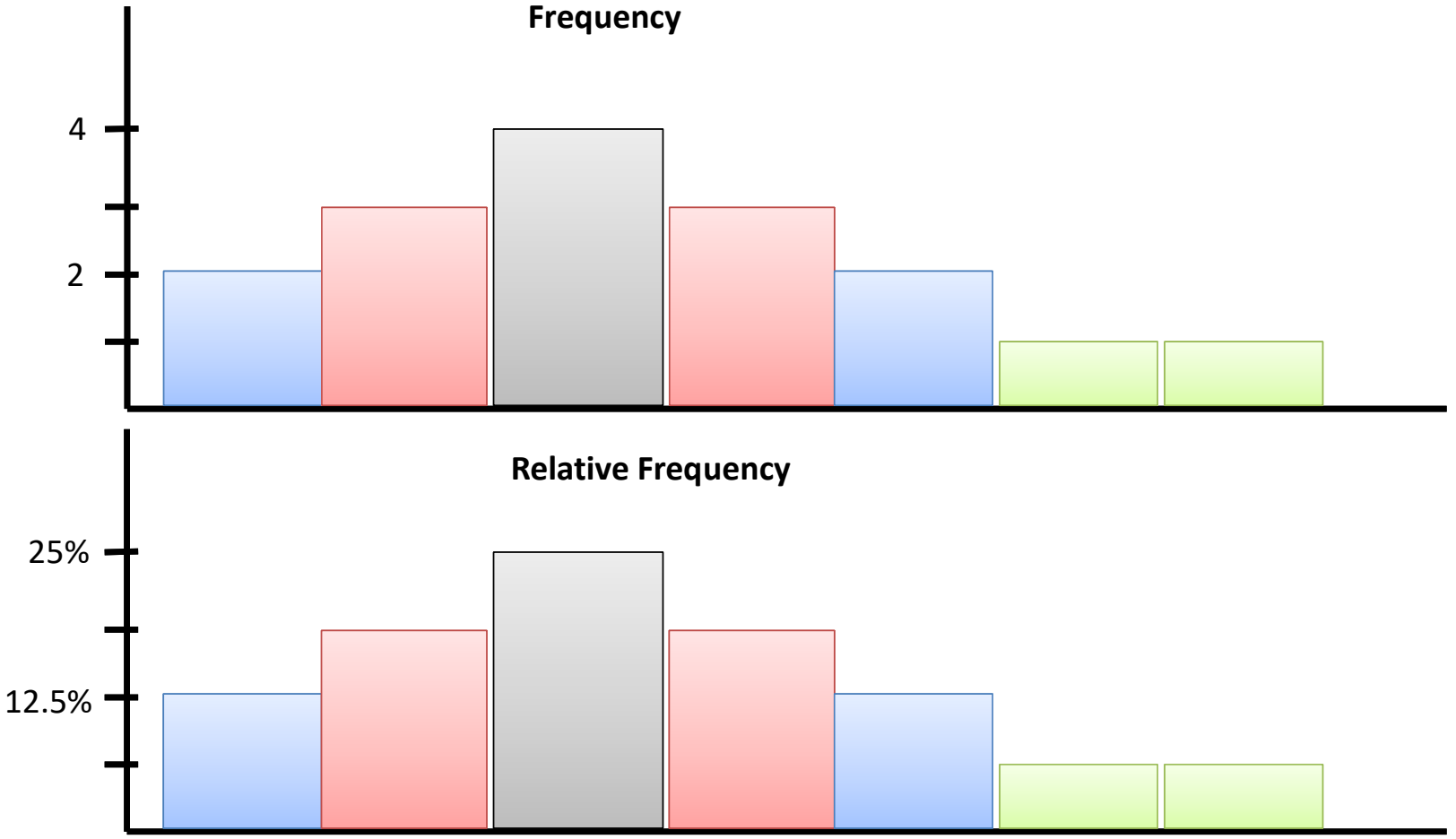
How parameters change as data is shifted and scaled

Data	Data + 5	Data * 5
7	12	35
7	12	35
5	10	25
8	13	40
10	15	50
13	18	65
5	10	25
3	8	15
2	7	10
3	8	15
5	10	25
6	11	30

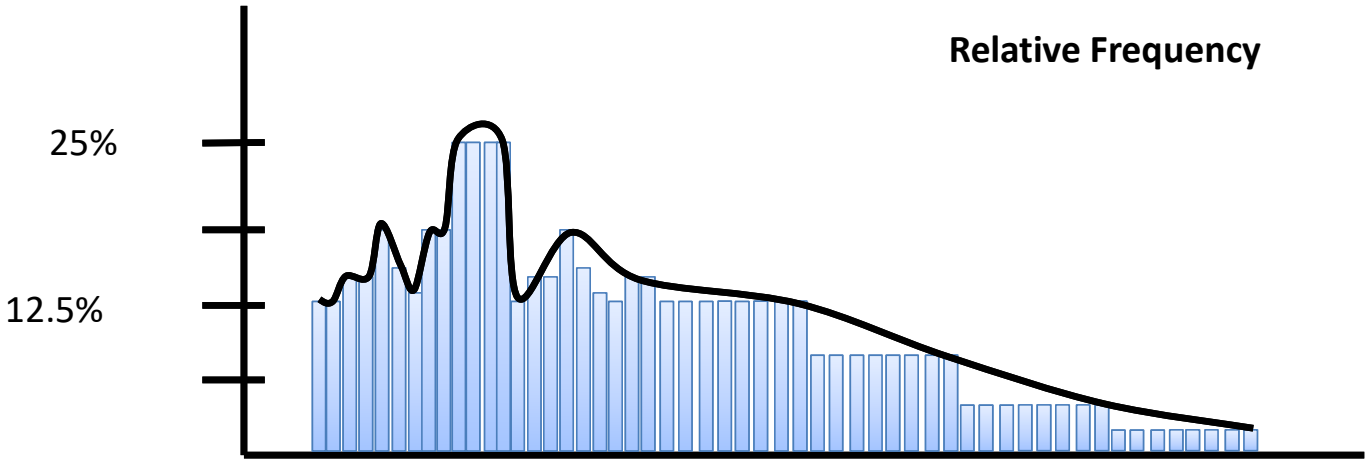
Find the mean, Standard deviation, Median and IQR?

Density curve

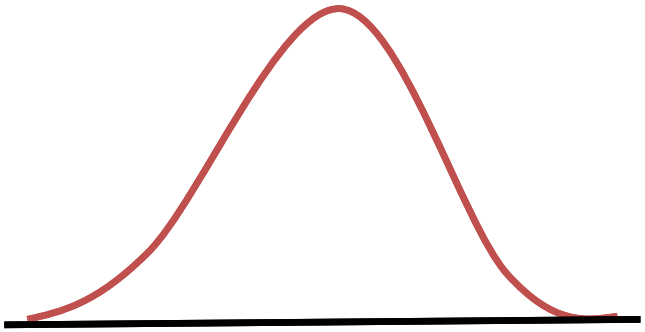
0.5, 0.7, 2.1, 2.2, 2.9, 3.1, 3.1, 3.3, 3.6, 4.5, 4.6, 4.8, 5.2, 5.3, 6.7, 8.1



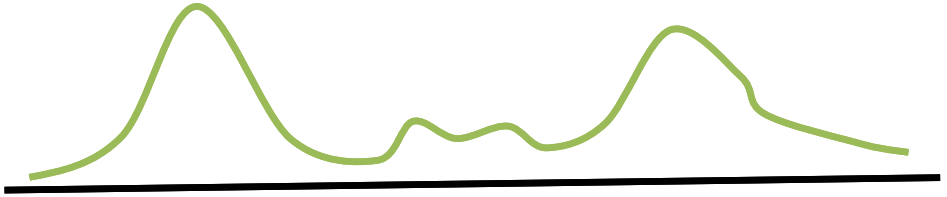
Density curve



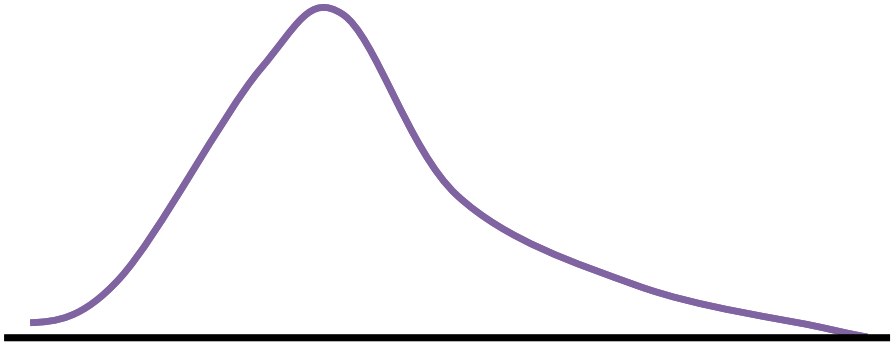
Density curve



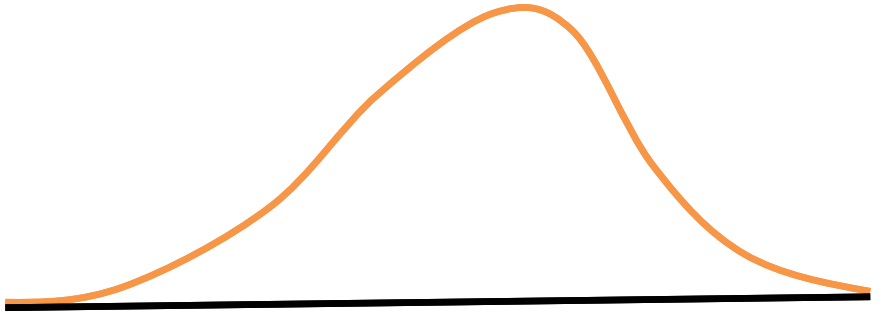
Normal Distribution



Bimodal Distribution



Right Skewed



Left Skewed

Normal Distribution – Empirical Rule

Assume that the mean weight of 1 year old girls in the USA is normally distributed, with a mean of about 9.5 kilograms and a standard deviation of approximately 1.1 kilograms.

a) Less than 8.4 kg
b) Between 7.3 kg and 11.7 kg
c) More than 12.8 kg

Normal Distribution – Empirical Rule

For a standard normal distribution, place the following in order from smallest to largest.

a) The percentage of data below 1
b) The percentage of data below −1
c) The mean
d) The standard deviation
e) The percentage of data above 2

Normal Distribution – Empirical Rule

The 2007 AP Statistics examination scores were not normally distributed, with $\mu=2.8$ and $\sigma=1.34$. What is the approximate z-score that corresponds to an exam score of 5? (The scores range from 1 to 5.)

a) 0.786
b) 1.46
c) 1.64
d) 2.20
e) A z-score cannot be calculated because the distribution is not normal.

Normal Distribution – Empirical Rule

The heights of 5th grade boys in the USA is approximately normally distributed, with a mean height of 143.5 cm and a standard deviation of about 7.1 cm. What is the probability that a randomly chosen 5th grade boy would be taller than 157.7 cm?

Normal Distribution – Empirical Rule (Ztable)

A set of middle school students heights are normally distributed with a mean of 150 centimetres and a standard deviation of 20 centimetres. John is a middle school student with a height of 161.4 centimetres.

What proportion of student heights are lower than John's height.

Normal Distribution – Empirical Rule

A set of middle school students heights are normally distributed with a mean of 150 centimetres and a standard deviation of 20 centimetres. John is a middle school student with a height of 161.4 centimetres.

What proportion of student heights are lower than John's height.

Give answer with 4 decimal places.

Normal Distribution – Empirical Rule

A set of Science exam scores are normally distributed with a mean of 40 points and a standard deviation of 3 points. James got a score of 47.5 points on the exam.

What proportion of exam scores are higher than James's score?

Give answer with 4 decimal places.

Normal Distribution – Empirical Rule

A set of laptop prices are normally distributed with a mean of 750 dollars and a standard deviation of 60 dollars.

What proportion of laptop prices are between 624 dollars and 768 dollars.

Give answer with 4 decimal places.

Normal Distribution – Empirical Rule

The distribution of resting pulse rates of all the students at Saint Angel's high school was approximately normal with a mean of 80 beats per minute and standard deviation 9 beats per minute.

The school nurse plans to provide additional screening to the students whose resting pulse rates are in the top 30% of the students who were tested.

What is the minimum resting pulse rate at that school for students who will receive additional screening?
Round to the nearest whole number

Normal Distribution – Empirical Rule

The distribution of average wait times in drive through restaurant lines in one town was approximately normal with the mean of 185 seconds and standard deviation of 11 seconds.

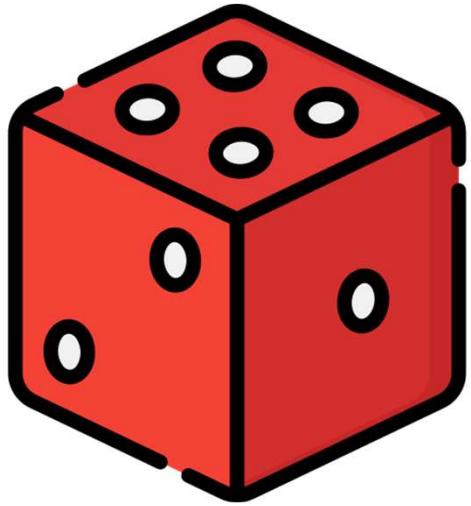
Aman only likes to use the drive through for restaurants where the average wait time is in the bottom 10% for that town.

What is the maximum average wait time for restaurants where Aman likes to use the drive through?
Round to the nearest whole number

Symmetric Distribution, Skewness, Kurtosis and KDE

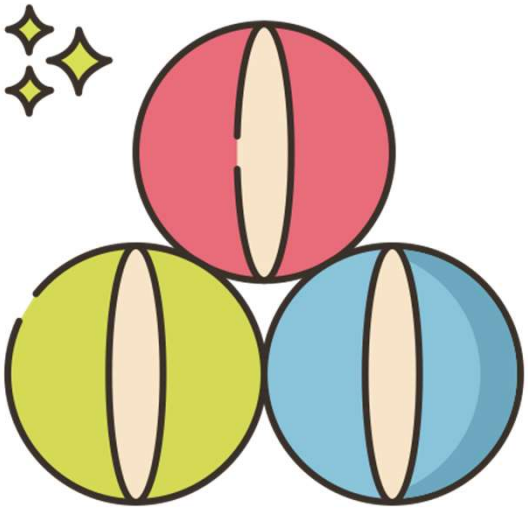
Probability

Probability – An Introduction



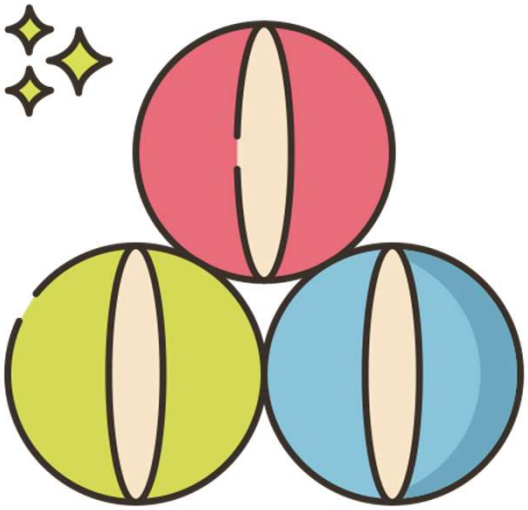
Simple Probability

Find the probability of pulling a yellow marble from a bag with 3 yellow, 2 pink and 2 blue marble



Simple Probability

We have a bag of 9 pink marbles, 2 blue marbles and 3 yellow marbles in it. What is the probability of randomly selecting a non blue marble from the bag?



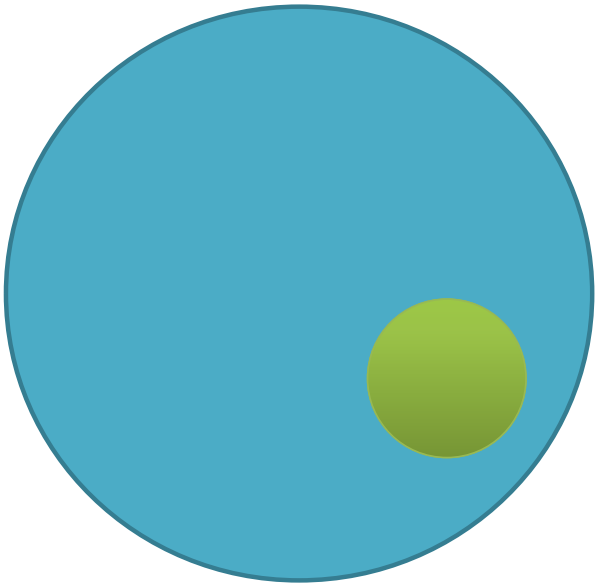
Simple Probability

If a number is randomly chosen from the following list, what is the probability that the number is a multiple of 5?

34, 51, 55, 32, 35, 34, 56, 55, 50, 30, 45, 3, 23, 25

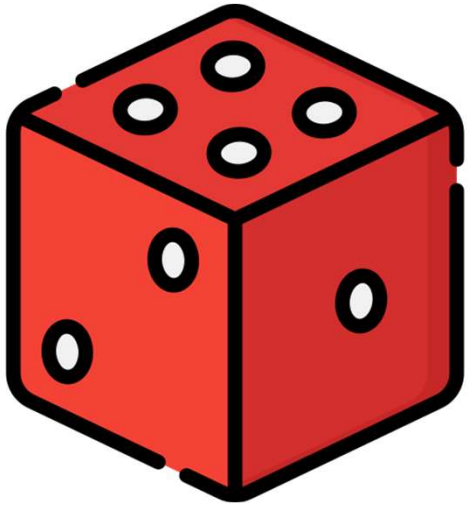
Simple Probability

The circumference if a circle is 42Pi. Contained in that circle is a smaller circle with the area of 24Pi. A point is selected at random from inside the larger circle. What is the probability that the point also lies in the smaller circle?



Probability

What is the maximum and minimum probability?

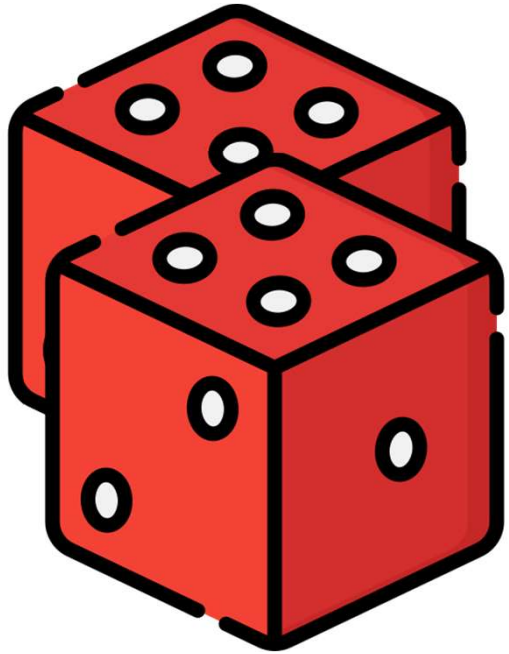


Probability – With Counting Outcomes

Find the probability of flipping exactly two heads on three coins?

Probability – Die rolling

Find the probability of rolling doubles on two six sided dice numbered from 1 to 6?



Subsets of sample spaces

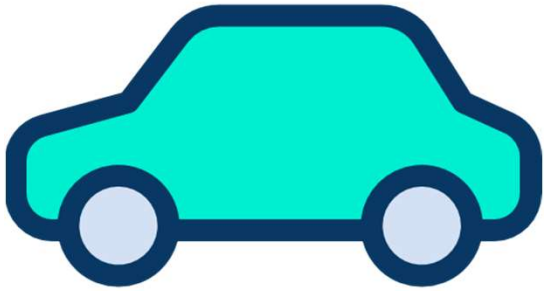
Amit is at Honda showroom and wanted to purchase a new car.

The types of cars available are City, Civic, WRV and BRV. The colors are Silver, Golden, White, Black and Red.

Based on the sample space of possible outcomes listed below, what is more likely?

	Silver	Golden	White	Black	Red
City	City Silver	City Golden	City White	City Black	City Red
Civic	Civic Silver	Civic Golden	Civic White	Civic Black	Civic Red
WRV	WRV Silver	WRV Golden	WRV White	WRV Black	WRV Red
BRV	BRV Silver	BRV Golden	BRV White	BRV Black	BRV Red

- The Car that Amit will select is City or Golden color?
- The Car that Amit will select is City and Golden color?



Intersection and Union of sets

$X = \{3,12,5,13\}$

$Y = \{14,15,6,3\}$

$X \cap Y$

$X \cup Y$

Relative Complement or Difference between sets

X = {5, 3, 17, 12, 19}

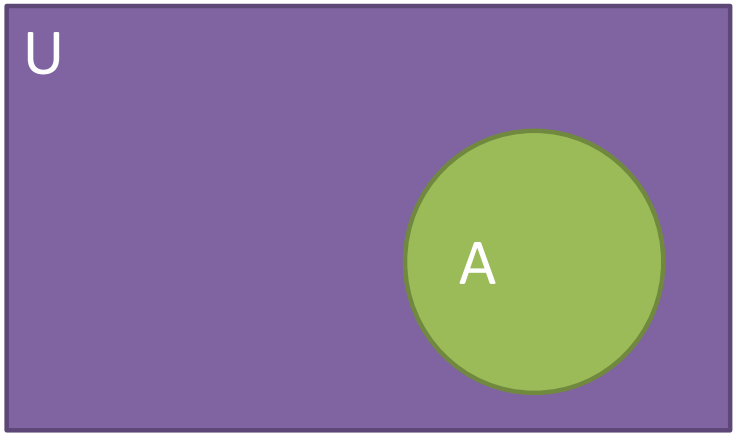
Y = {17, 19, 6}

$X - Y$

$Y - X$

$X - X$

Universal Set and Absolute complement



$$A' = U - A$$

Subset, Strict, Subset and Superset

$A = \{1,3,5,7,18\}$

$B = \{1,7,18\}$

$C = \{18,7,1,19\}$

Set Operations together

$A = \{3,7,-5,0,13\}$

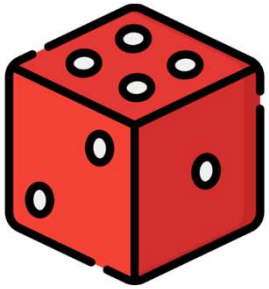
$B = \{0,17,3, Blue,*\}$

$C = \{Pink,*, 3, 17\}$

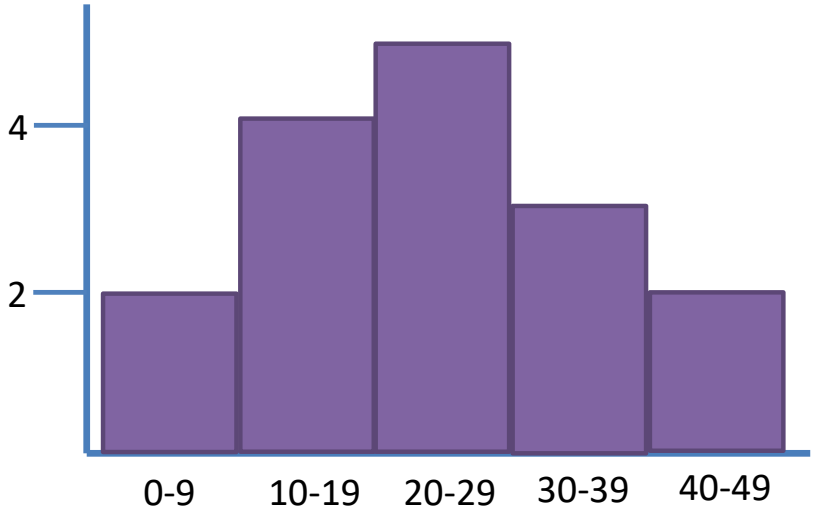
$$A - (A \cap (B - C)') \cup (B \cap C)$$

Experimental Probability

Theoretical Probability



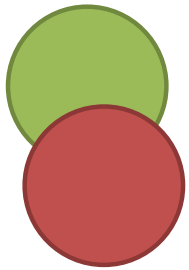
Experimental Probability



P(Points >= 30) = ?

Experimental Probability

What is the probability of getting red marble if there are 50 green marbles and 50 red marbles in the bucket?



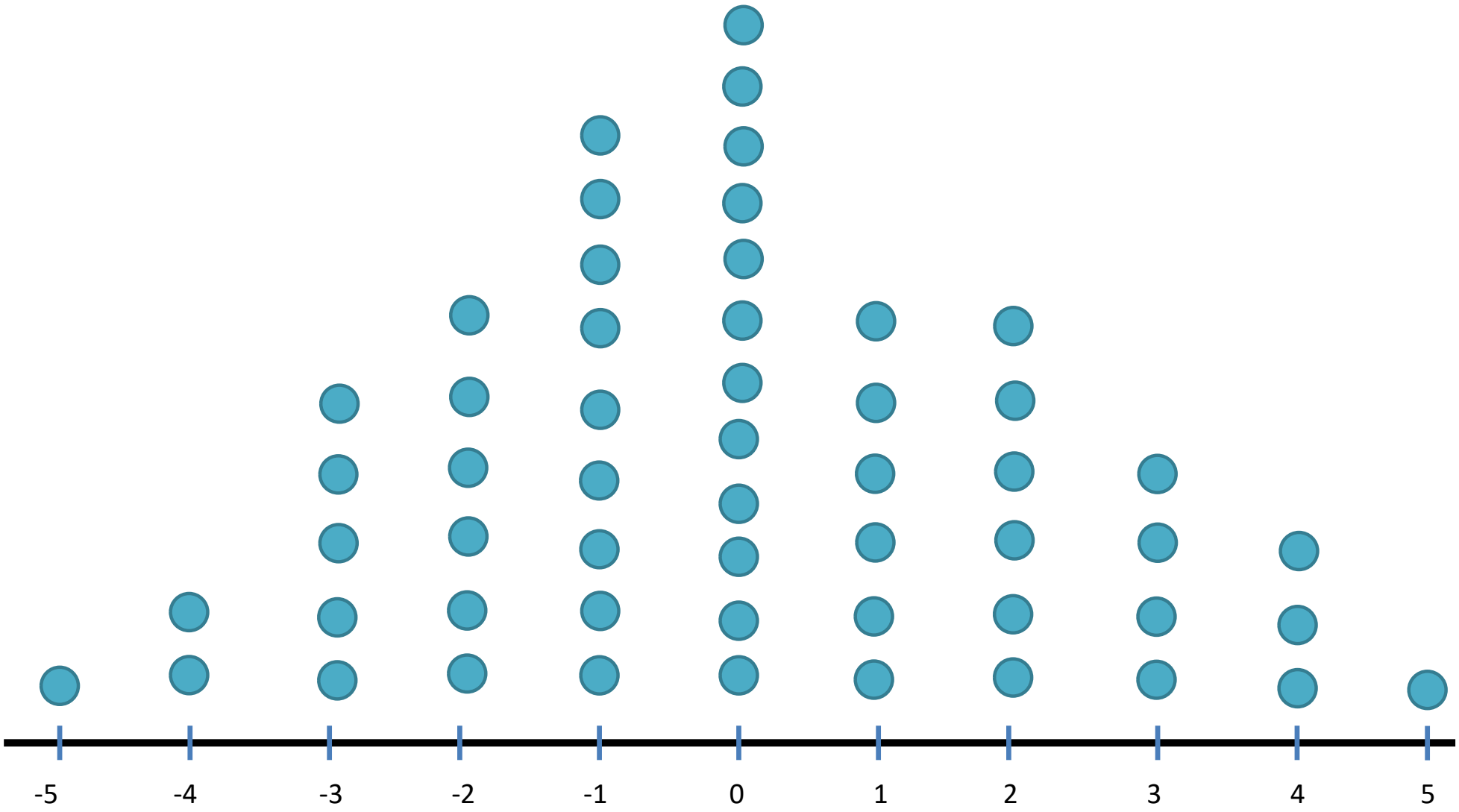
Statistical Significance of experiment

In an experiment aimed at studying the effect of advertising on eating behavior in children, a group of 500 children 6 to 12 years old, were randomly assigned to two different groups. After randomization, each child was asked to watch a cartoon in a private room, containing a large bowl of popcorns. The cartoon included 2 commercial breaks.

The first group watched food commercials, while the second group watched non-food commercials. Once the child finished watching the cartoon, the conductors of the experiment weighed the crackers bowl to measure how many grams of popcorns the child ate. They found that the mean amount of popcorns eaten by the children who watched food commercials is 5 grams greater than the mean amount of crackers eaten by the children who watched non-food commercials.

They re-randomized the results into two new groups and measured the difference between the means of the new groups. They repeated this for 55 times, and plotted the resulting differences as show in the image.

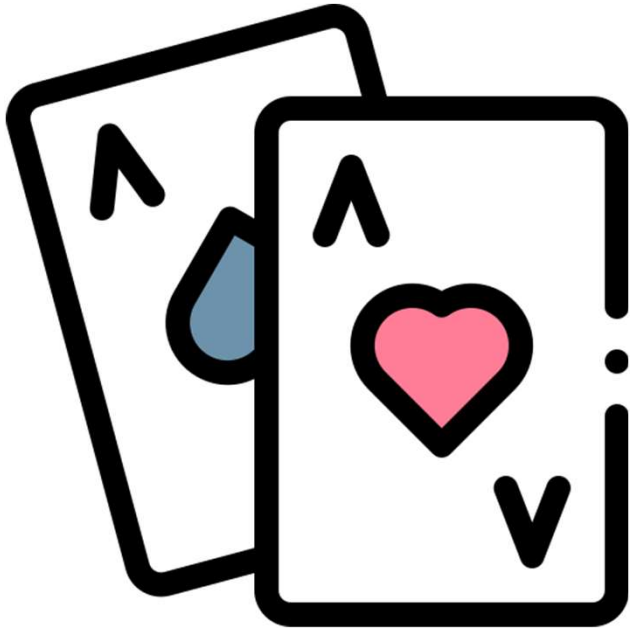
According to the multiple experiments, is the result of the experiment significant?


-5
-4
-3
-2
-1
0
1
2
3
4
5

Probability with Venn Diagrams

Total Cards: 52

A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, K, Q



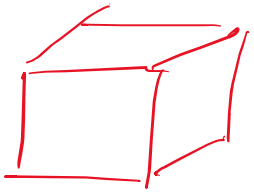
P(Jack)

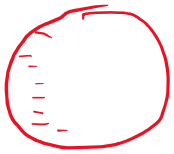
P(Hearts)

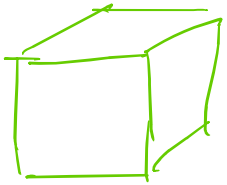
P(Jack and Hearts)

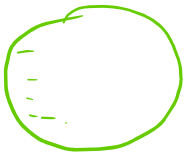
P(Jack or Heart)

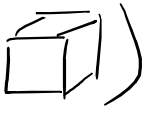
Addition Rule for Probability

 × 8 (Red cubes)

 × 9 (Red Spheres)

 × 5 (Green cubes)

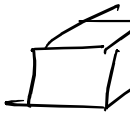
 × 7 (Green Spheres)

$P($  $) = ?$

$P(G) = ?$

$P($  $) = ?$

Green

$P(G$ or  $) = ?$

Sample Space for compound space

Compound probability of independent events

Probability of a compound event

Coin flipping probability

Probability without equally likely events

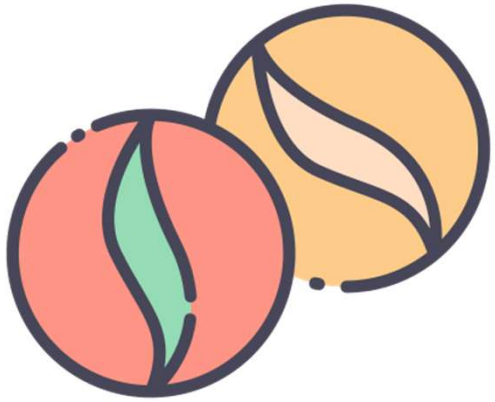
Independent events - Exam

On a multiple choice test, question 1 has 4 choices and question 2 has 3 choices. Each problem has only one correct answer. What is the probability of randomly guessing the correct answer to both questions.

Die rolling probability with independent events

Find the probability of rolling even numbers three times, using a six sided die numbered from 1 to 6

Dependent Probability Introduction



Dependent Probability: Coins

You have 8 coins in a bag. 3 of them are unfair in that and they have 60% chance of coming up heads when flipped. You randomly choose one coin from the bag and flip it 2 times. What is the percentage probability of getting 2 heads?

Dependent Probability: Coins

You have 4 coins in a bag. 3 of them are unfair in that and they have 45% chance of coming up tails when flipped. You randomly choose one coin from the bag and flip it 4 times. What is the percentage probability of getting 4 heads?

Dependent Probability

Suppose Amit simultaneously rolls a six sided die and a four sided die. Let A be the even that he rolls doubles and B is the event that the four sided dis is 4. Use the sample space of possible outcomes shown to answer the following.

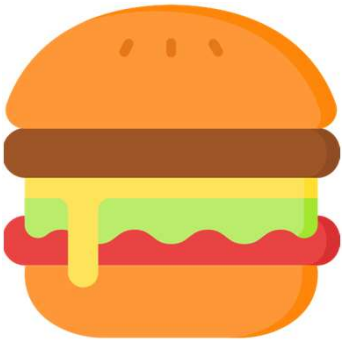
- What is P(A), the probability that Amit rolls double?
- What is P(B), the probability that the four sided die is 4?
- What is P(A|B), the probability that Amit rolls doubles, given that the four-sided die is 4?
- What is P(B|A), the probability that the four sided die is 4, given that Amit rolls doubles?
- What is P(A and B), the probability that Amit rolls doubles and the second die is 4?
- What is P(A).P(B|A)?
- What is P(B).P(A|B)?

Conditional Probability

Rahul's two favorite foods are burger and pizza. Let **A** represent the event that he eats a burger for breakfast and **B** represents the event that he eats pizza for lunch.

On a randomly selected day, the probability that Rahul will eat a burger for breakfast, P(A) is 0.6, the probability that he will eat pizza for lunch, P(B), is equal to 0.5, and the conditional probability that he eats a burger for breakfast, given that he eats pizza for lunch, P(A|B) is equal to 0.7.

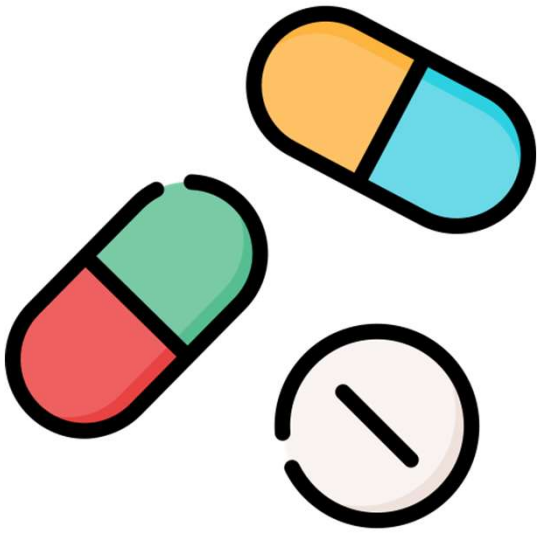
Based on this information, what is the P(B|A), the conditional probability that Rahul eats pizza for lunch, given that he eats a burger for breakfast, rounded to nearest hundredth.



Conditional Probability tree diagram

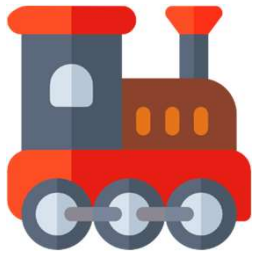
A company screens job applicants for illegal drug use at a certain stage in their hiring process. The specific test they use has a false positive rate of 2% and a false negate rate of 1%. Suppose that 5% of all their applicants are actually using illegal drugs, and we randomly select an applicant.

Given the applicant tests positive, what is the probability that they are actually on drugs?



Conditional Probability and Independence

John is interested in the weather conditions and whether the downtown train he sometimes takes runs on time. For a year, John records whether each day is sunny, cloudy, rainy or snowy, as well as whether this train arrives on time or is delayed. His results are displayed in the table below:



For these days, are the events “Delayed” and “Snowy” independent?

	On Time	Delayed	Total
Sunny	167	3	170
Cloudy	115	5	120
Rainy	40	15	55
Snowy	8	12	20
Total	330	35	365

Counting, Permutations and combinations

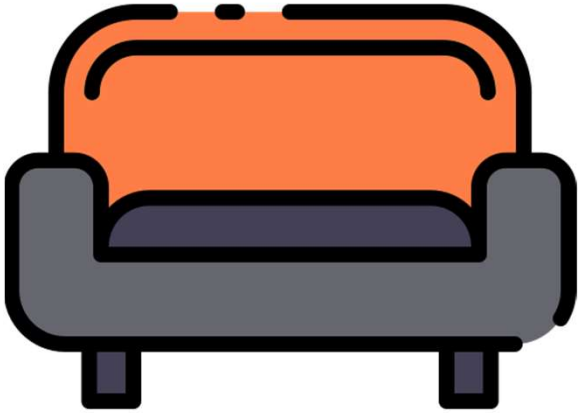
Count outcomes using tree diagram

Count outcomes

Permutation Formula

Zero Factorial or 0!

Factorial and counting



Possible three letter words

A B C D ..... Z

Arrange Colors

In a game, a code made using different colors is created by one of the players and the other player tries to guess the code. The code builder gives hints about whether the colors are correct and in the right position or not.

The possible colors are Blue, Yellow, White, Red, Orange and Green. How many 4 color codes can be made if the colors cannot be repeated?

Ways to pick officers

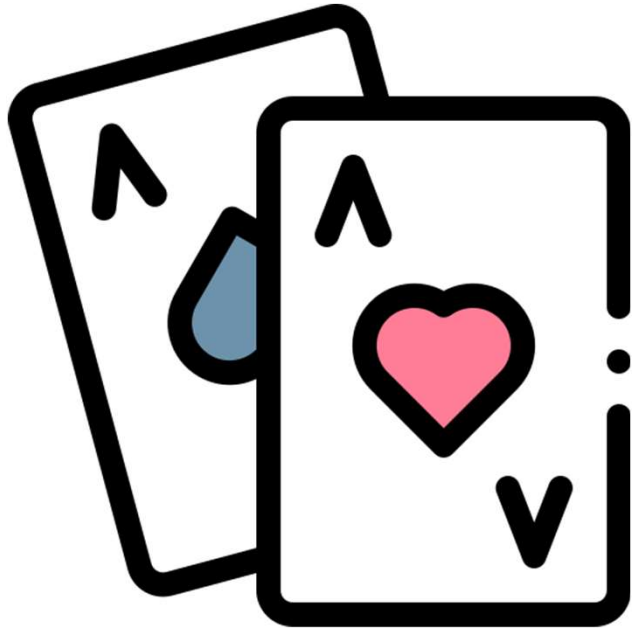
A club of 9 people to choose a board of three officers: President, Vice President and Secretary. How many ways are there to choose the board from the 9 people?

Introduction to Combinations

Handshaking combinations

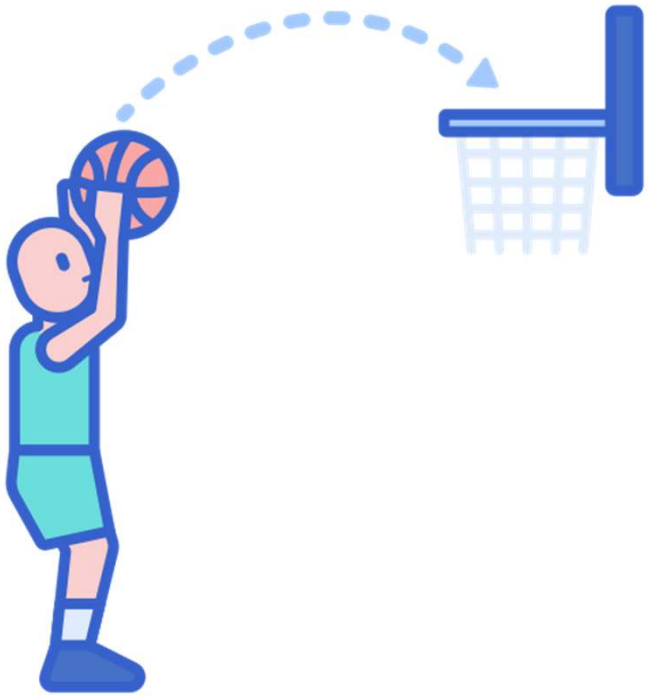
Cards Example

A card game using 36 unique cards: four suits (Diamond, Hearts, Spades and Clubs) with cards numbered from 1 to 9 in each suit. A hand is collection of 9 cards, which can be sorted however the players chooses. How many card hands are possible?



Probability using combinations

Probability using combinations

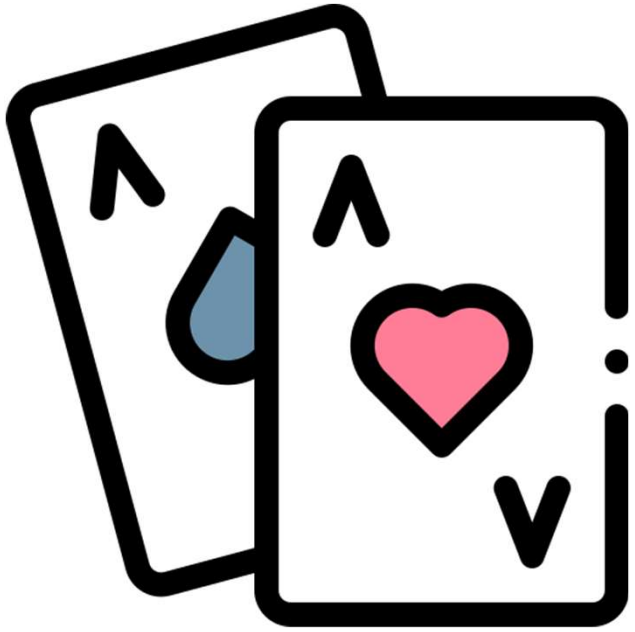


Probability using combinations

A club of 9 people to choose a board of three officers: President, Vice President and Secretary. Assuming the officers are chosen at random, what is the probability that the officers are Atharv for President, Aarav for vice president and Richa for secretary?

Cards Example

A card game using 36 unique cards: four suits (Diamond, Hearts, Spades and Clubs) with cards numbered from 1 to 9 in each suit. A hand is collection of 9 cards, which can be sorted however the players chooses. What is the probability of getting all four of the 1s?



Random Variables

Continuous and Discrete and its distribution

Frozen Yogurt

Valid discrete probability distribution

Aman is analyzing his basketball statistics. The following table shows a probability model for the results from his next two free throws.

Outcome	Probability
Miss both free throws	0.2
Make exactly one free throw	0.5
Make both free throws	0.1

Is this a valid probability model?

Probability with discrete random variable

Harshal plans to buy packs of baseball cards until he gets the card of his favorite player, but he only has enough money to buy at most 4 packs. Suppose that each pack has probability 0.2 for containing the card Harshal is hoping for.

Let the random variable X be the number of packs of cards Harshal buys. Here is the probability distribution.

X	1	2	3	4
P(X)	0.2	0.16	0.128	?

Find the probability where
P(X>=2) = ?

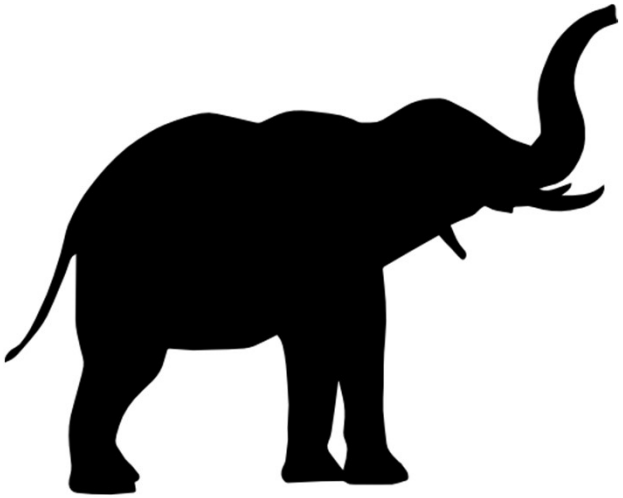
Mean (Expected value) of a discrete random variable

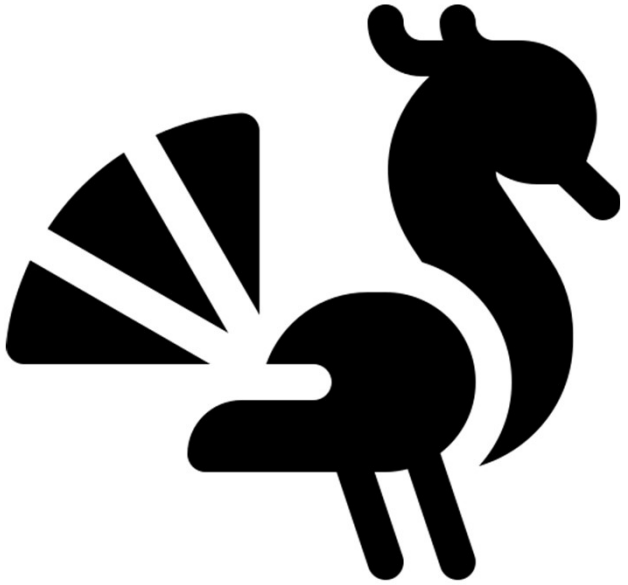
X	P(X)
0	0.1
1	0.15
2	0.4
3	0.25
4	0.1

Variance and Standard Deviation of a discrete random variable

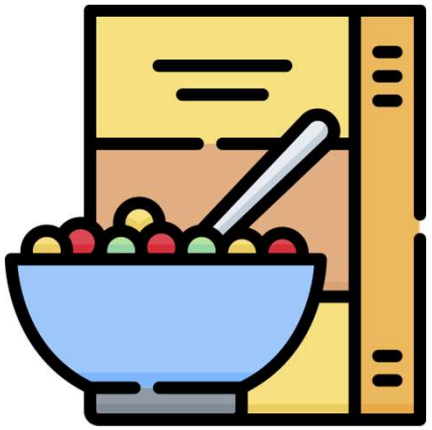
X	P(X)
0	0.1
1	0.15
2	0.4
3	0.25
4	0.1

Mean of sum and difference of random variables





Variance of sum and difference of random variables



Intuition for why independence matters for variance of sum

Deriving the variance if the difference of random variables

Analyzing distribution of sum of two normally distributed random variables

Shikhar commutes to work and he worries about running out of fuel. The amount of fuel he uses follows a normal distribution for each part of his commute, but the amount of fuel he uses on the way home varies more. The amounts of fuel he uses for each part of the commute are also independent of each other. Here are summary statistics for the amount of fuel Shikhar uses for each part of his commute.

	Mean	Standard Deviation
To work	$\mu_w = 10L$	$\sigma_w = 1.5L$
To Home	$\mu_H = 10L$	$\sigma_H = 2L$

Suppose that Shikhar has 25L of fuel in his tank, and he intends to drive to work and back home.

What is the probability that Shikhar runs out of the fuel?

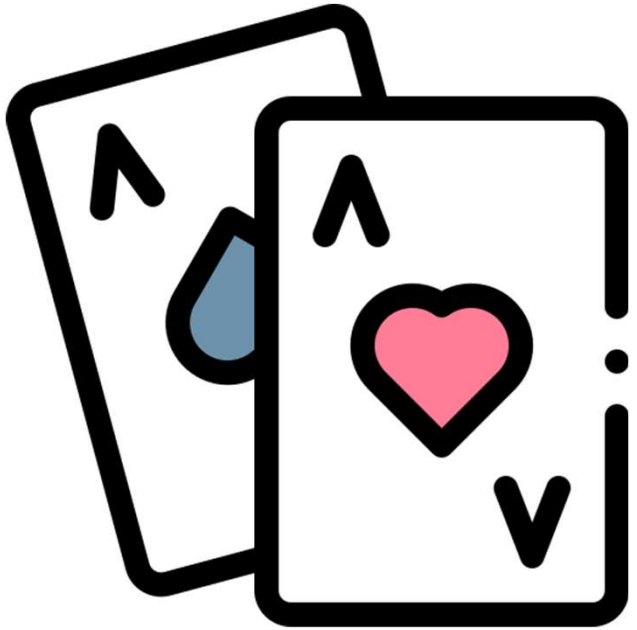
Analyzing the difference in distributions

Suppose that

- Men have a mean height of 178 cm with a standard deviation of 8 cm.
- Woman have a mean height of 170 cm with a standard deviation of 6 cm.
- The male and female heights are each normally distributed.
- We independently, randomly selected a man and a woman.

What is the probability that the woman is taller than the man?

Binomial Variables



Recognizing Binomial variables

A manager oversees 11 female employees and 9 male employees. They need to pick 3 of these employees to go on a business trip, so the manager places all 20 names in a bowl and chooses at random. Let X = the number of female employees chosen.

Is X a binomial variable? Why or Why not?

- Each trail isn't being classified as a success or failure, so X is not a binomial variable.
- There is no fixed number of trails, so X is not a binomial variable.
- The trails are not independent, so X is not a binomial variable.
- This situation satisfies each of the conditions for a binomial variables, so X has a binomial distribution.

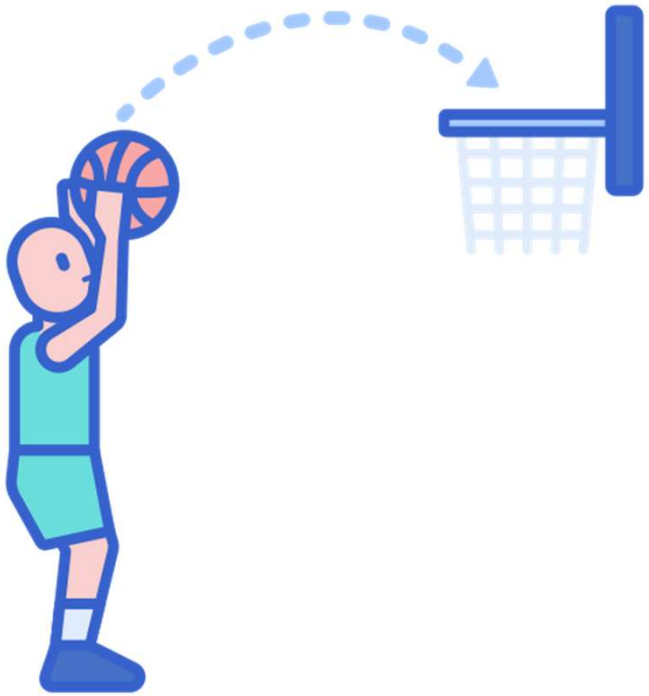
10% Rule of assuming independence between trails

Let X = #of boys from 3 trials of selecting from a classroom of n students, where 50% of the class is a boy and 50% is a girl.

n	P(X=3) with replacement	P(X=3) without replacement	3 as % of n
20			
30			
100			
10000			

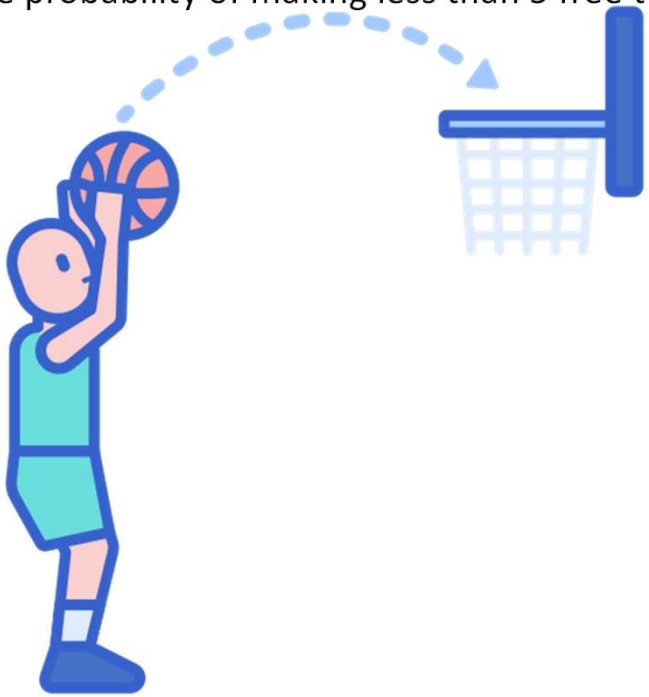
Binomial Distribution

Binomial Probability Example



Binompdf and binomcdf functions

I have a 0.35 probability of making a free throw. What is the probability of making 4 out of 7 free throws? What is the probability of making less than 5 free throws?



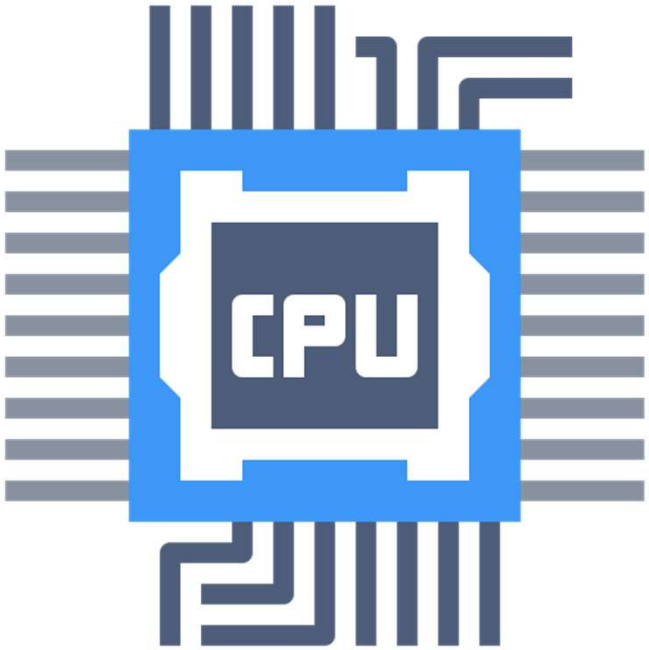
Mean and variance of Bernoulli distribution

Expected value and variance of Binomial variable

Finding mean and standard deviation of a binomial random variable

A company produces processing chips for cell phones. At one of its large factories, 2% of the chips produced are defective. A quality check involves randomly selecting and testing 500 chips.

What are the mean and standard deviation of the number of defective processing chips in these samples?



Geometric random variables - Introduction

Probability of a geometric random variable

Johnson makes 25% of the three point shots he attempts. For practice, Johnson likes to shoot three point shots until he successfully makes one. Let M be the number of shots it takes Johnson to successfully make his first three point. Assume that the results of each shot are independent.

Find the probability that Johnson's first successful shot occurs on his 3rd attempt.

Cumulative geometric probability

Richa registers vehicles for the department of transportation. Sports utility vehicles (SUVs) make up 12% of the vehicles she registers. Let V be the number of vehicles Richa registers in a day until she first registers an SUV. Assume the type of each vehicle is independent.

Find the probability that Richa registers more than 4 vehicles before she registers for SUV.

Cumulative geometric probability

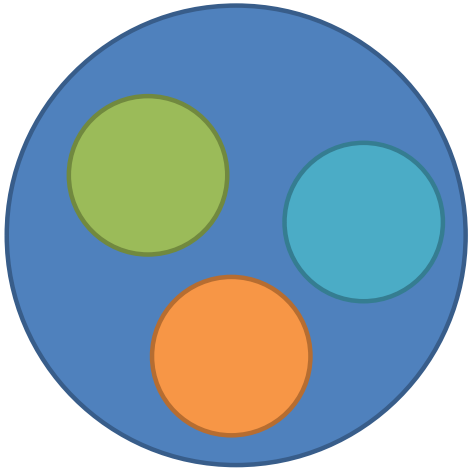
Leena runs a cake decorating business, for which 10% of her orders come over the telephone. Let C be the number of cake orders Leena receives in a month until she first gets an order over the telephone. Assume the method of placing each cake order is independent.

Find the probability that it takes fewer than 5 orders for Leena to get her first telephone order of the month.

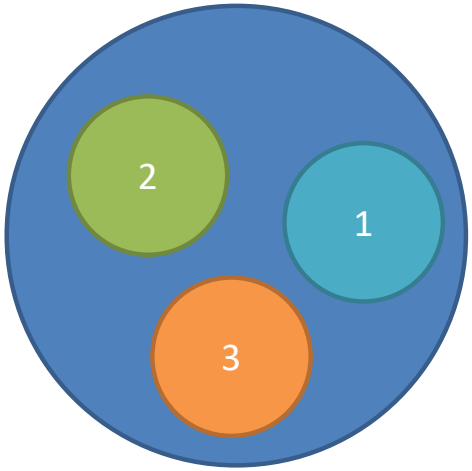
Proof of expected value of geometric random variables

Sampling distributions

Sampling Distributions



What is the distribution of the sample mean?



#'s Pick	Mean
1,1	1
1,2	1.5
1,3	2
2,1	1.5
2,2	2
2,3	2.5
3,1	2
3,2	2.5
3,3	3

Sampling distribution for the sample mean for the sample size = 2

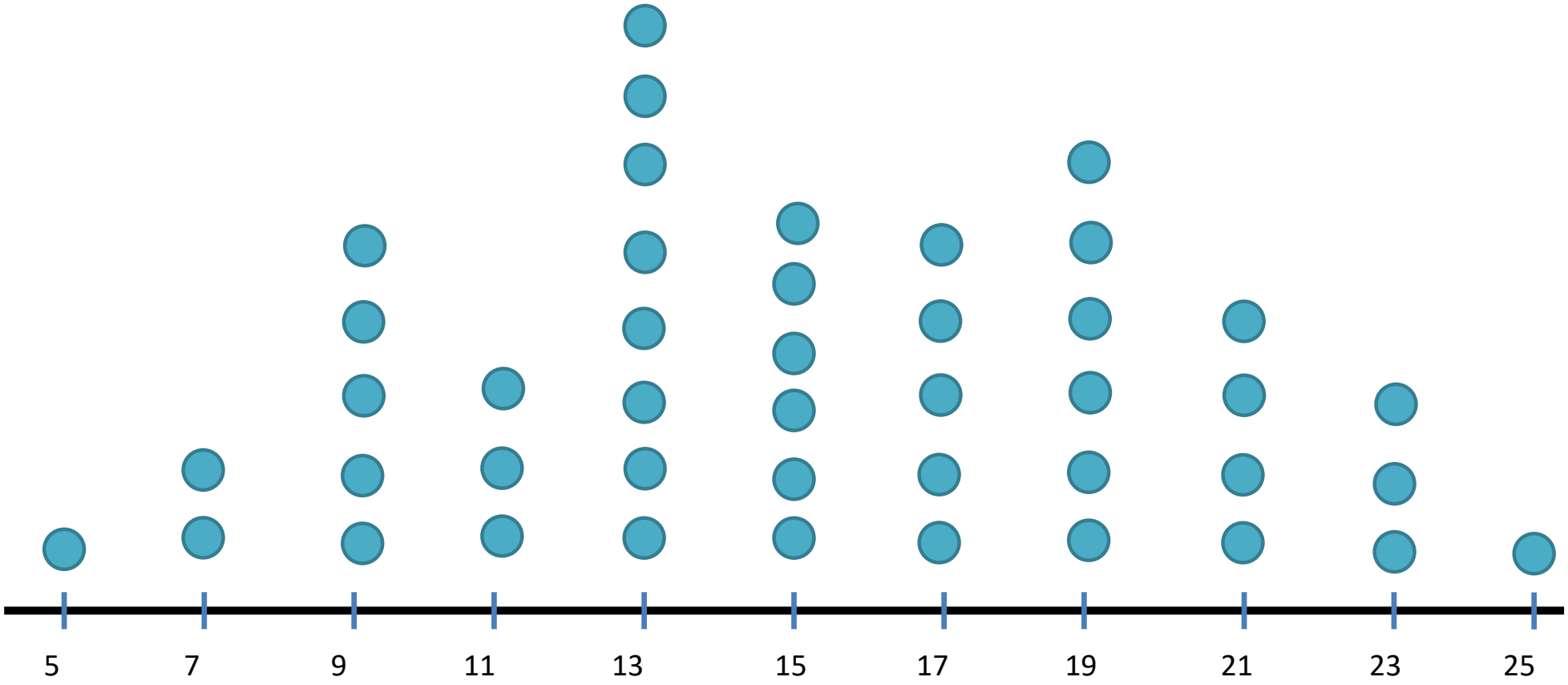
Sample Statistics Bias

Amit was curious if sample mean was an unbiased estimator of population median. He placed tennis balls numbered from 0 to 32 in a drum and mixed them well. Note that the median of the population is 16.

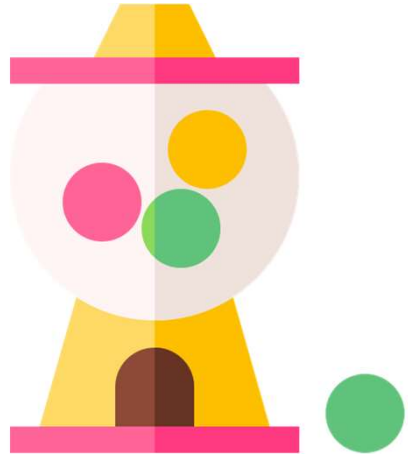
He then took a random sample of 5 balls and calculated the median of the sample. He replaced the balls and repeated this process for total of 50 times. His results are summarized in the table below, where each value represents the sample mean from a sample of 5 balls.

Based on these results, does the sample median appear to be a biased or unbiased estimator of population median?

Sample Statistics Bias



Sampling distribution of sample proportion



Normal conditions for sampling distributions of sample proportions

Emi runs a restaurant that receives a shipment of 50 Oranges every day. According to supplier, approximately 12% of the population of these oranges is overripe. Suppose that Emi calculates the daily proportion of overripe oranges in her sample of 50. We can assume the supplier's claim is true, and that the oranges each day represent a random sample.

What will be the shape of the sampling distribution of the daily proportion of overripe oranges?

- Skewed to the left
- Skewed to the right
- Approximately Normal
- Uniform

Normal conditions for sampling distributions of sample proportions

According to the survey, radio reaches 88% of children each week. Suppose we tool weekly random samples of n =125 children from this population and computed the proportion of children in each sample whom radio reaches?

What will be the shape of the sampling distribution of the proportions of children the radio reaches?

- Skewed to the left
- Skewed to the right
- Approximately normal
- Uniform

Probability of sample proportions

Suppose that 15% of the 1750 students at a school have experienced extreme levels of stress during the past month. A high school newspaper doesn't know this figure, but they are curious what it is, so they decide to ask a simple random sample of 160 students if thy have experienced extreme levels of stress during the past month. Subsequently, they find that 10% of the sample replied "Yes" to the questions.

Assuming the true proportion is 15%, what is the approximate probability that more than 10% of the sample would report that they experienced extreme levels of stress during the past month?

- $P(\hat{p} > 0.10) \approx 0.92$
- $P(\hat{p} > 0.10) \approx 0.94$
- $P(\hat{p} > 0.10) \approx 0.96$
- $P(\hat{p} > 0.10) \approx 0.98$

Inferring population mean from sample mean

Central Limit Theorem

Standard Error of the mean

Example

The average male drinks 2 Lts of water when active outdoors (with a standard deviation of 0.7 Lts). You are planning a full day nature trip for 50 men and will bring 110 Lts of water. What is the probability that you will run out?

Confidence Interval

Confidence intervals and margin of error

Interpretation of Confidence interval

A zookeeper took a random sample of 30 days and observed how much food an elephant ate on each of those days. The sample mean was 350 Kg and the sample standard deviation was 25 Kg. The resulting 90% confidence interval for the mean amount of food was (341, 359) kg.

Which of the following statements is a correct interpretation of the 90% confidence interval?

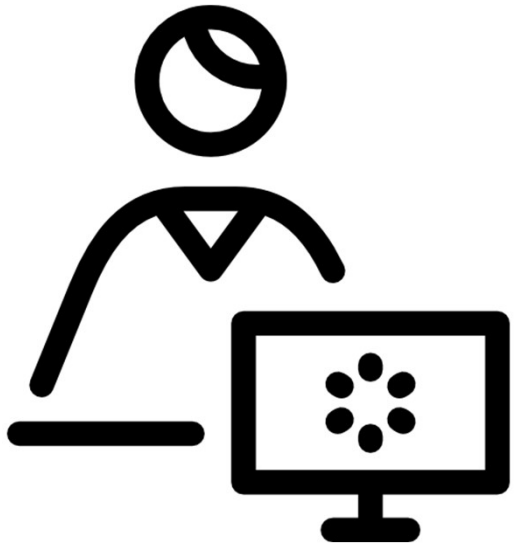
Choose 1 answer:

- The elephant ate between ate between 341 kg and 359 kg on 90% of all the days.
- There is a 0.90 probability that the true mean amount of food is between 341 kg and 359 kg.
- In repeated sampling, this method produces intervals that capture the population mean in about 90% of samples.
- In repeated sampling, this method produces a sample mean between 341 kg and 359 kg in about 90% of samples.

Interpretation of Confidence interval

In a local teaching district a technology grant is available to teachers in order to install a cluster of four computers in their classrooms. From the 6250 teachers in the district, 250 were randomly selected and asked if they felt that computers were an essential teaching tool for their classroom. Of those selected, 142 teachers felt that computers were an essential teaching tool.

- Calculate a 99% confidence interval for the proportion of teachers who felt that computers are an essential teaching tool.
- How could the survey be changed to narrow the confidence interval but to maintain the 99% confidence interval?



Margin of Error

Condition for valid confidence interval for a proportion

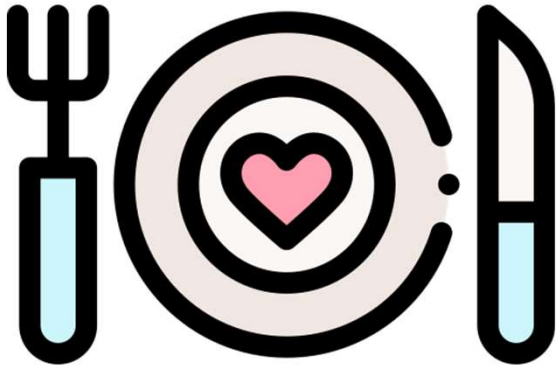
Example

Amit is in charge of the dinner menu for his senior prom, and he wants to use a one sample z interval to estimate what proportion of seniors would order a vegetarian option. He randomly selects 30 of the 150 total seniors and finds that 7 of those sampled would order the vegetarian option.

Which conditions for constructing this confidence interval did Amit's sample meet?

Choose all apply:

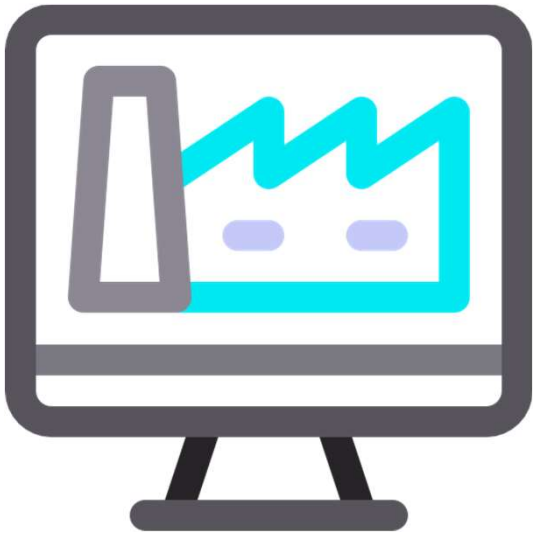
- The data is a random sample from the population of interest
- It is normal
- Individual observations can be considered independent



Example

Eva wants to build a one sample z interval to estimate what proportion of computers produced at a factory have a certain defect. She chooses a confidence level of 94%. A random sample of 200 computers shows that 12 computers have the defect.

What critical value z^* should Eva use to construct this confidence interval?



Example

Divya has over 500 songs on her mobile phone, and she wants to estimate what proportion of the songs are by a female artist. She takes an Simple random sample of 50 songs on her phone and finds that 20 of the songs sampled are by a female artist.

Based on this sample, which of the following is a 99% confidence interval for the proportion of songs on her phone that are by a female artist?

$20 \pm 1.96\sqrt{\frac{20(30)}{50}}$

$0.40 \pm 1.96\sqrt{\frac{0.40(0.60)}{50}}$

$20 \pm 2.576\sqrt{\frac{20(30)}{50}}$

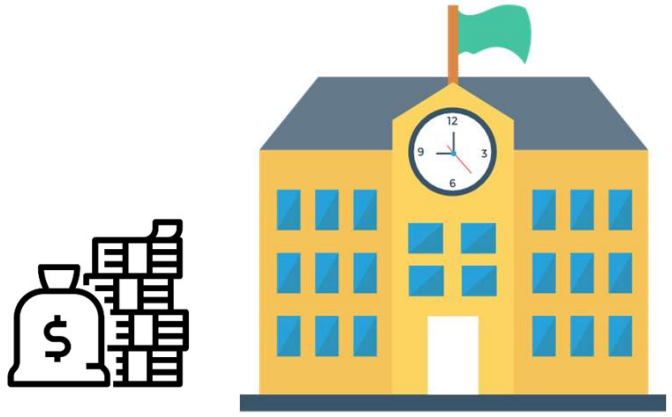
$0.40 \pm 2.576\sqrt{\frac{0.40(0.60)}{50}}$

Example

Divya wants to make a one sample z interval to estimate what proportion of her community members favor a tax increase for more local school funding. She wants her margin of error to be not more than $\pm 2\%$ at the 95% confidence interval.

What is the smallest sample size required to obtain the desired margin of error?

- 267
- 601
- 1068
- 2401



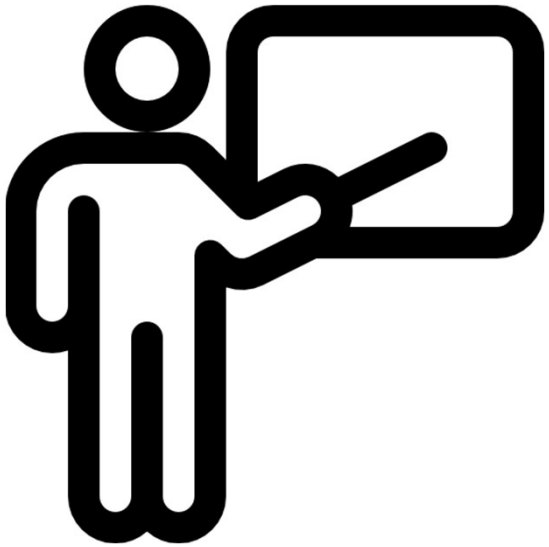
Introduction to t statistics

Conditions for valid t intervals

Faiz wanted to estimate the mean age of the faculty members at his large university. He took a simple random sample of 20 of the approximately 700 faculty members, and each faculty member in the sample provided Faiz with their age. The data were skewed to the right with a sample mean of $\bar{x} = 38.75$. He is considering using his data to make a confidence interval to estimate the mean age of faculty members at his university.

Which conditions for constructing a t interval have been met?

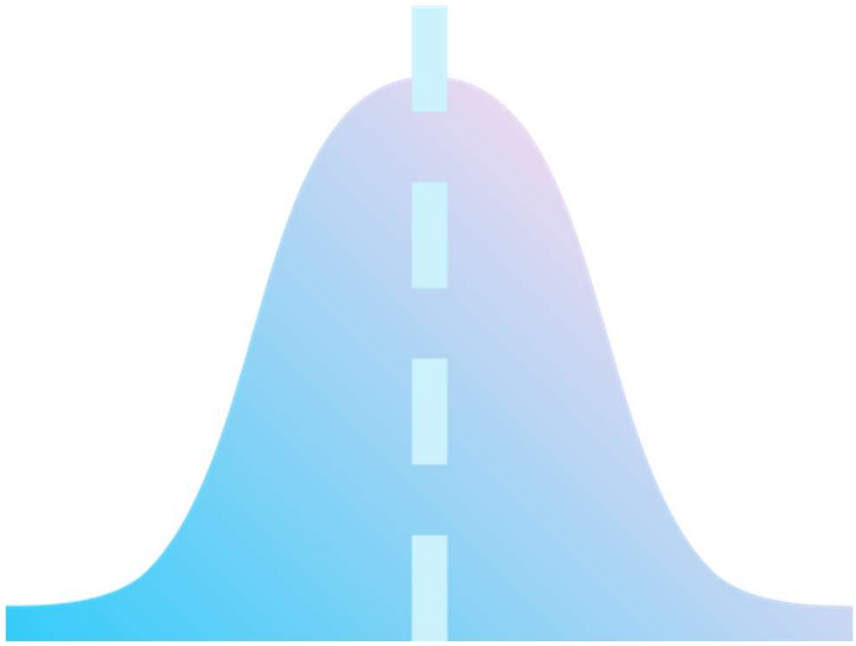
- The data is a random sample from the population of interest
- The sampling distribution of $\bar{x}$ is approximately normal
- Individual observations can be considered independent



Find critical t value

What is the critical value t^* for constructing a 98% confidence interval for a mean from a sample size of n = 15 observations?

- $t^* = 2.249$
- $t^* = 2.264$
- $t^* = 2.602$
- $t^* = 2.624$
- $t^* = 2.977$



Example

A nutritionist wants to estimate the average caloric content of the wrap at a popular restaurant. They obtain a random sample of 14 wraps and measure their caloric content. Their sample data are roughly symmetric with a mean of 700 calories and a standard deviation of 50 calories.

Based on the sample, which of the following is a 95% confidence interval for the mean caloric content of these wraps?

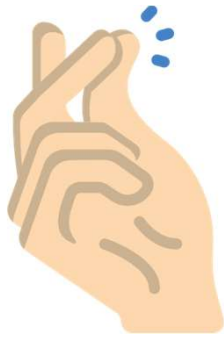
- 700 ± 23.7
- 700 ± 26.2
- 700 ± 28.7
- 700 ± 28.9
- 700 ± 100

Confidence interval for a mean with paired data

A group of friends wondered how much faster they could snap their fingers on one hard versus the other hand. Each person snapped their fingers with their dominant hand for 10 seconds and their non dominant hand for 10 seconds. Each participants flipped a coin to determine which hand they would use first. Here are the data for how many snaps they performed with each hand, the difference for each participant, and summary statistics.

Participant	Muzzammil	Ankush	Rishabh	Priyanka	Roshan
Dominant	44	42	40	37	42
Non Dominant	35	37	32	31	36
Difference (Dom-Non)	9	5	8	6	6

Summary	Mean	Standard Deviation	Sample Size
Dominant	$\overline{x_{dom}} = 41$	$S_{dom} \approx 2.65$	5
Non Dominant	$\overline{x_{non}} = 34.2$	$S_{dom} \approx 2.59$	5
Difference (dom-non)	$\overline{x_{diff}} = 6.8$	$S_{dom} \approx 1.64$	5



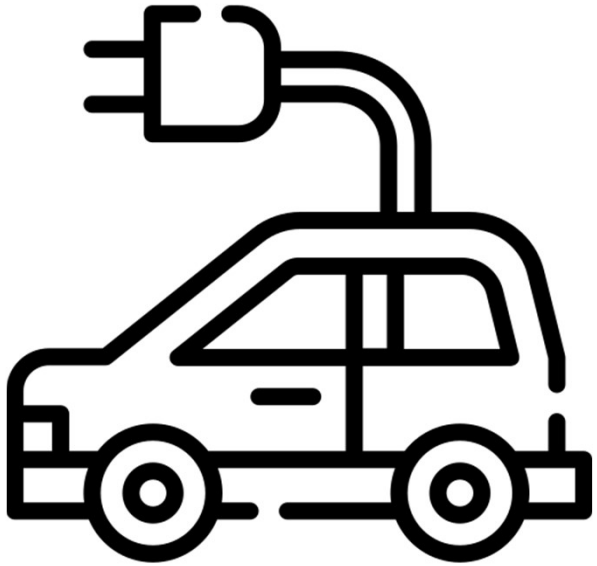
Create and interpret a 95% confidence interval for mean difference in number of snaps for these participants.

Sample size for a given margin of error for a mean

Naina wants to create a confidence interval to estimate the mean driving range for her company's new electric vehicle. She wants the margin of error to be no more than 10 kilometers at a 90% level of confidence. A pilot study suggests that the driving ranges for this type of vehicle have a standard deviation of 15 kilometers.

Which of these is the smallest approximate sample size required to obtain the desired margin of error?

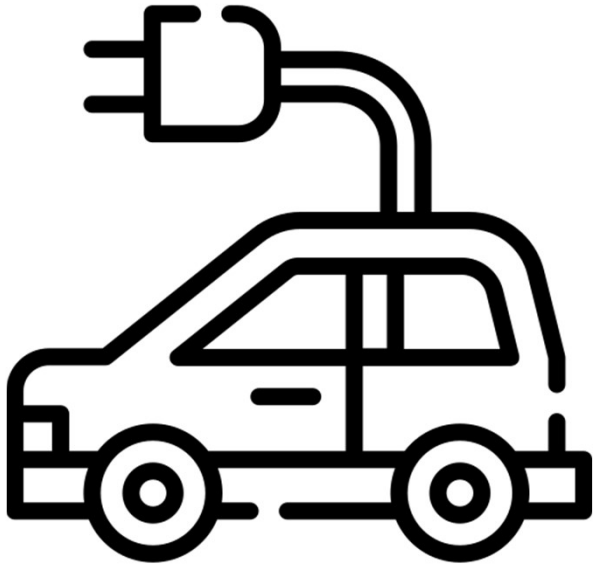
- 5
- 7
- 10
- 15



T-statistic confidence interval

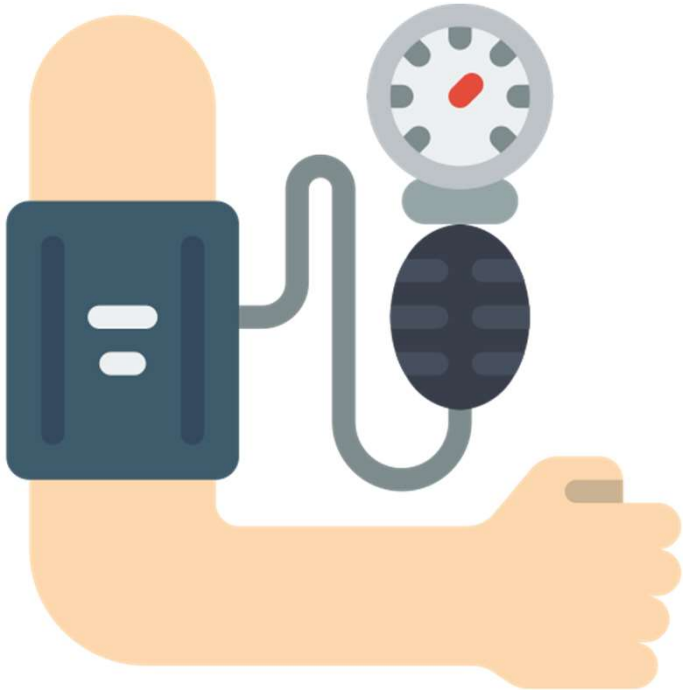
The mean emission of all engines of a new design needs to be below 20 ppm if the design is to meet new emission requirements. Ten engines are manufactured for testing purposes and the emission level of each is determined. The emission data is:

15.6, 16.2, 22.5, 20.5, 16.4, 19.4, 16.6, 17.9, 12.7, 13.9



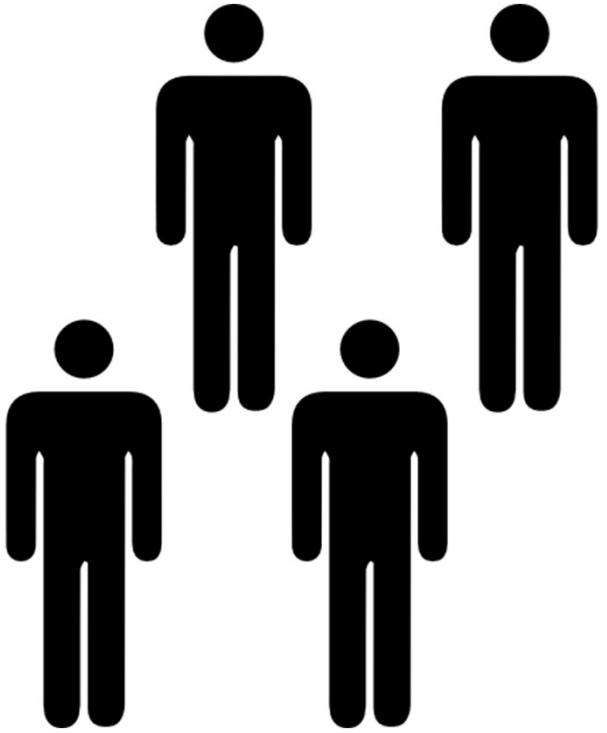
Small sample size confidence intervals

7 patients blood pressures have been measured after having been given a new drug for 3 months. They had blood pressure increases of 1.5, 2.9, 0.9, 3.0, 3.2, 2.1 and 1.9. Construct a 95% confidence interval for the true expected blood pressure increase for all patients in a population.

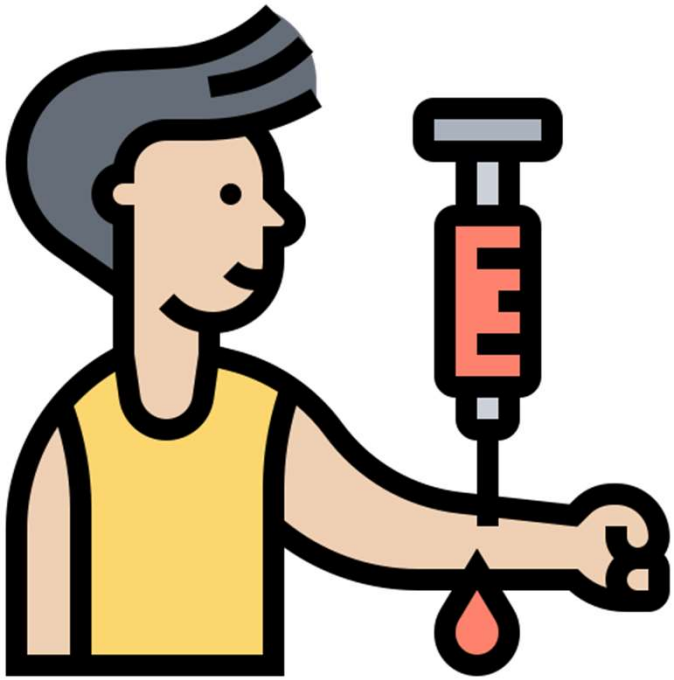


Significance Tests (Hypothesis Testing)

Simple Hypothesis Testing



Idea Behind Hypothesis Testing



Example of null and alternative hypothesis

A restaurant owner installed a new automated drink machine. The machine is designed to dispense 530 mL of liquid on the medium size setting. The owner suspects that the machine may be dispensing too much in medium drinks. They decide to take a sample of 30 medium drinks to see if the average amount is significantly greater than 530 mL. What are appropriate hypotheses for their significance test?

$H_0: p = 530\ mL$
$H_a: p > 530\ mL$
(where p is the proportion of liquid dispensed on this setting)

$H_0: p = 530\ mL$
$H_a: p < 530\ mL$
(where p is the proportion of liquid dispensed on this setting)

$H_0: \mu = 530\ mL$
$H_a: \mu > 530\ mL$
(where μ is the mean of liquid dispensed on this setting)

$H_0: \mu = 530\ mL$
$H_a: \mu < 530\ mL$
(where μ is the mean of liquid dispensed on this setting)

Example of null and alternative hypothesis

The national sleep foundation recommends that teenagers aged 14 to 17 years old get at least 8 hours of sleep per night for proper health and wellness.

A statistics class at a large high school suspects that students at their school are getting less than 8 hours of sleep on average. To test their theory, they randomly sample 42 of these students and ask them how many hours of sleep they get per night. The mean from the sample is $\bar{x} = 7.5\ hours$.

Here is their alternative hypothesis:

H_a: $The\ average\ amount\ of\ sleep\ at\ their\ school\ get\ per\ night\ is\ \ldots.$

P-values and significance tests

Comparing P-values to different significance levels

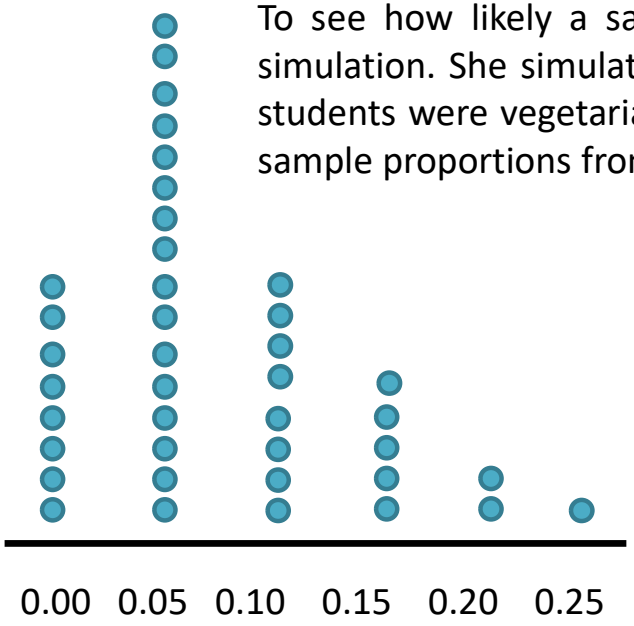
Raja heard that spinning – rather than flipping a coin raises the probability above 50% that the coin lands showing heads. He tested this by spinning 10 different coins 10 times each. His hypothesis were $H_0 : p = 0.50$ versus $H_a: p > 0.50$, where p is the true proportion of spins that a penny would land showing "heads".

In his 100 spins, the coin landed showing "heads" in 59 spins. Raja calculated that the statistic $\hat{p} = \frac{59}{100} = 0.59$ had an associated p-value of approximately 0.036

Estimating a P-value from a simulation

Eva read an article that said 6% of teenagers were vegetarians, but she thinks its higher for students at her large school. To test her theory, Eva took a random sample of 25 students at her school, and 20% of them were vegetarians.

To see how likely a sample like this was to happen by random chance alone, Eva performed a simulation. She simulated 40 samples of n = 25 students from a large population where 6% of the students were vegetarian. She recorded the proportion of vegetarians in each sample. Here are the sample proportions from her 40 samples.



Estimating a P-value from a simulation

Eva wants to test $H_0: p = 6\%\ vs. H_a: p > 6\%$ where p is the true proportion of students who are vegetarian at her school.

Based on these simulated results, what is the approximate p-value of the test?
Note: The sample result was $\hat{p} = 20\%$.

- P-value ≈ 0.01
- P-value ≈ 0.025
- P-value ≈ 0.03
- P-value ≈ 0.075

Type 1 and Type 2 Errors

Type 1 and Type 2 Errors

A large nationwide poll recently showed an unemployment rate of 9% in India. The MLA of a local town wonders if this national results holds true for his town, so he plans on taking a sample of his residents to see if the unemployment rate is significantly different than 9% in his town.

Let p represent the unemployment rate in his town. Here are the hypothesis he will use:

$H_0: p = 0.09$
$H_a: p \neq 0.09$

Under which of the following conditions would the MLA commit a type 1 error?

- He concludes the town's unemployment rate is not 9% when it actually is
- He concludes the town's unemployment rate is not 9% when it actually is not
- He concludes the town's unemployment rate is 9% when it actually is
- He concludes the town's unemployment rate is 9% when it actually is not

Type 1 and Type 2 Errors

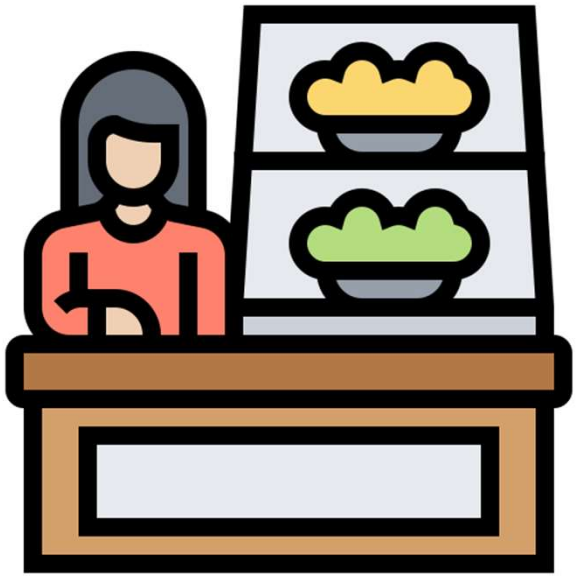
A large IT office is curious if they should build another cafeteria. They plan to survey a sample of their employees to see if there is strong evidence that the proportion interested in a meal plan is higher than 40%, in which case they will consider building a new cafeteria.

Let p represent the proportion of students interested in a meal plan. Here are the hypothesis they will use:

$H_0: p \leq 0.40$
$H_a: p > 0.40$

What would be the consequence of a Type II error in this context?

- They don't consider building a new cafeteria when they should
- They don't consider building a new cafeteria when they shouldn't
- The consider building a new cafeteria when they shouldn't
- The consider building a new cafeteria when they should



Introduction to Power in significance tests

Power in significance test

A significance test is going to be performed using a significance level of $\alpha = 0.05$. Suppose that the null hypothesis is actually false.

If the significance level was lowered to $\alpha = 0.01$, which of the following would be true?

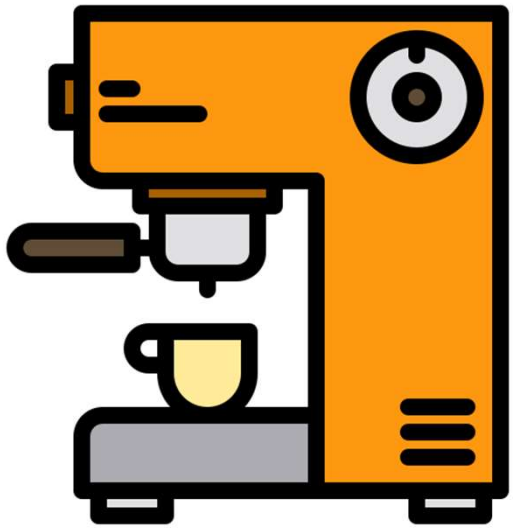
- Both the probability of a Type II error and the power of the test would decrease
- Both the probability of a Type II error and the power of the test would increase
- The probability of a Type II error would increase and the power of the test would decrease
- The probability of a Type II error would decrease and the power of the test would increase

Power in significance test

Anu owns a car wash and is trying to decide whether or not to purchase a vending machine so customers can buy coffee while they wait. She will get the machine if she is convinced that more than 30% of the customers would buy coffee. She plans on taking a random sample of n customers and asking them whether or not they would buy coffee from the machine, and she will then do a significance test using $\alpha = 0.05$ to see if the sample proportion who say "Yes" is significantly greater than 30%.

Which situation below would result in the highest power for her test?

- She uses a sample size of n = 100 and 32% of all customers would actually buy coffee.
- She uses a sample size of n = 200 and 32% of all customers would actually buy coffee.
- She uses a sample size of n = 50 and 50% of all customers would actually buy coffee.
- She uses a sample size of n = 200 and 50% of all customers would actually buy coffee.

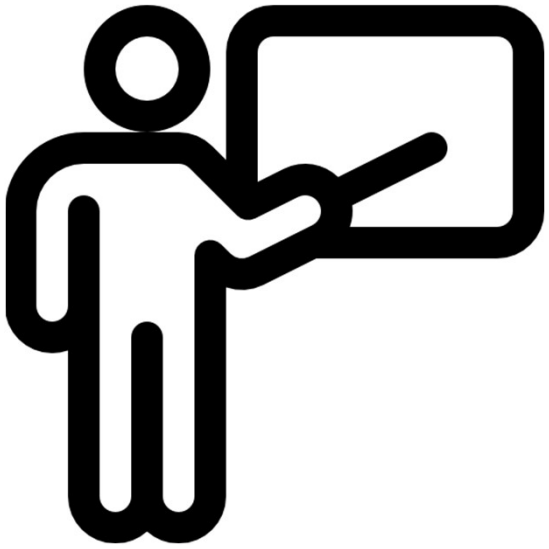


Constructing hypothesis for a significance test about a proportion

Aman read a report saying that 49% of teachers in the India were members of a labor union. He wants to test whether this holds true for teachers in his state, so he is going to take a random sample of these teachers and see what percent of them are members of a union.

Let p represent the proportion of teachers in his state that are members of a union.

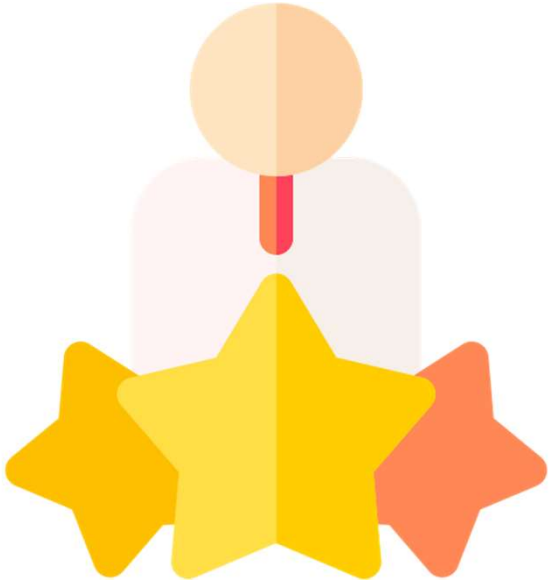
Write an appropriate set of hypothesis for his significance test.



Conditions for a z test about a proportion

Jana works on a small team of 40 employees. Each employee receives an annual rating, the best of which is "exceeds expectations." Management claimed that 10% of employees earn this rating, but Jana suspected it was actually less common. She obtained an anonymous random sample of 10 ratings for employees on her team. She wants to use the sample data to test $H_o: p = 0.1 \; versus \; H_a: p < 0.1$, where p is the proportion of all employees on her team who earned "exceeds expectations."

Which conditions for performing this type of test did Jana's sample meet?



Calculating a z statistic in a test about a proportion

The MLA of a town saw an article that claimed the national unemployment rate is 8%. They wondered if this held true in their town, so they took a sample of 200 residents to test $H_0: p = 0.08 \; versus \; H_a: p \neq 0.08$, where p is the proportion of residents in the town that are unemployed. The sample included 22 residents who were unemployed.

Assuming that the conditions for inference have been met, identify the correct test statistic for this significance test.

Calculating a P-value given a z statistic

Faiz read an article that said 26% of Indians can speak more that one language. He was curious if this figure was higher in his city, so he tested $H_0: p = 0.26 \; vs. \; H_a: p > 0.26$, where p represents the proportion of people in his city that can speak more than one language.

He found that 40 of 120 people sampled could speak more than one language. The test statistic for these results was z ≈ 1.83.

Assuming that the necessary conditions are met, what is the approximate P-value for Faiz's test?

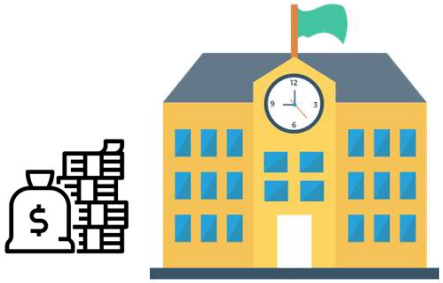
Making conclusions in a test about a proportion

A public opinion survey investigated whether a majority (more than 50%) of adults supported a tax increase to help fund the local school system. A random sample of 200 adults showed that 113 of those sampled supported the tax increase.

Researchers used these results to test $H_0: p = 0.5 \; versus \; H_a: p > 0.5, where \; p$ is the true proportion of adults that support the tax increase. They calculated a test statistic of $z \approx 1.84$ and a corresponding P-value of approximately 0.033.

Assuming the conditions for inference were met, which of these is an appropriate conclusion?

- At the $\alpha = 0.01$ significance level, they should conclude that more than 50% of the adults support the tax increase
- At the $\alpha = 0.01$ significance level, they should conclude that less than 50% of the adults support the tax increase
- At the $\alpha = 0.05$ significance level, they should conclude that more than 50% of the adults support the tax increase
- At the $\alpha = 0.05$ significance level, they should conclude that less than 50% of the adults support the tax increase



Writing Hypothesis for a significance test about a mean

A quality control expert at a drink factory took a random sample of bottles from a batch and measured the amount of liquid in each bottle in the sample. The amounts in the sample had a mean of 503 mL and a standard deviation of 5 mL. They want to test if this is convincing evidence that the mean amount for bottles in this batch is different than the target value of 500 mL.

Let μ be the mean amount of liquid in each bottle in the batch.

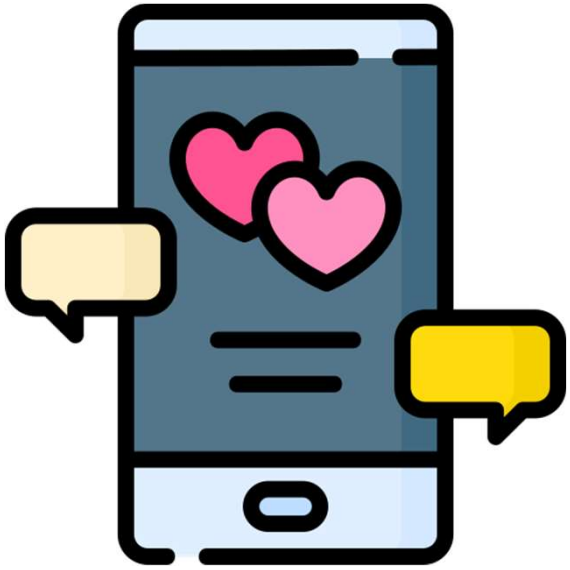
Write an appropriate set of hypothesis for their significance test.

Condition for a t test about a mean

Amit and his friends have been using a group messaging app for over a year to chat with each other. He suspects that, on average, they send each other more than 100 messages per day.

Amit takes a random sample of 7 days from their chat history and records how many messages were sent on those days. The sample data are strongly skewed to the right with a mean of 125 messages and a standard deviation of 44 messages. He wants to use these sample data to conduct a t test about the mean.

Which conditions for performing this type of significance test have been met?



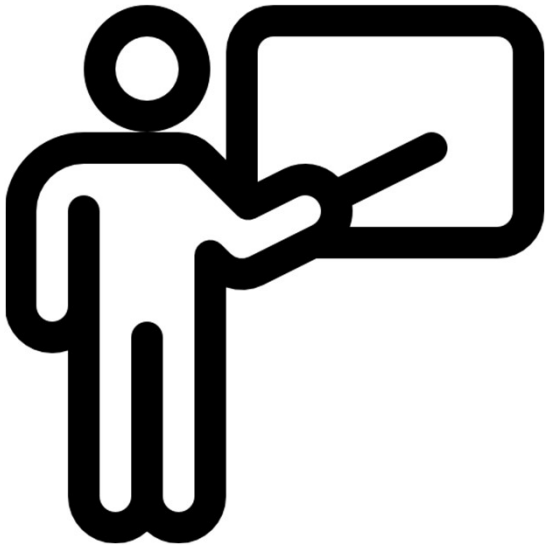
When to use z or t statistics in significance tests

Calculating t statistic for a test about a mean

Roy suspects that teachers in his school district have less than 5 years of experience on average. He decided to test H_o: $\mu = 5$ $versus$ H_a: $\mu < 5$ using a sample of 25 teachers. His sample mean was 4 years and his sample standard deviation was 2 years.

Roy wants to use these sample data to conduct a t test on the mean. Assume that all conditions for inference have been met.

Calculate the test statistic for Roy's test.



Calculate p-value from t statistic

Mary was testing H_0: $\mu = 18$ $versus$ H_a: $\mu < 18$ with a sample of 7 observations. Her test statistic was t = -1.9. Assume that the conditions for inference were met.

What is the approximate p-value for Mary's test?

Calculate P-value from t statistic

Cynthia was testing H_o: $\mu = 0$ $versus$ H_a: $\mu \neq 0$ with a sample of 6 observations. Her test statistic was t = 2.75. Assume that the conditions for inference were met.

What is the approximate p-value for Cynthia's test?

Comparing P-value from t statistic to significance level

James was curious if the automated machine at his restaurant was filling drinks with the proper amount. He filled a sample of 20 drinks to test H_o: $\mu = 530\ mL\ versus\ H_a$: $\mu \neq 530\ mL\ where\ \mu$ is the mean filling amount.

The drinks in the sample contained a mean amount of 528 mL with a standard deviation of 4 mL. There results produced a test statistic of t = -2.236 and a p-value of approximately 0.038.

Assuming the conditions for inference were met, what is an appropriate conclusion at the $\alpha = 0.05$ significance level?

- Reject H_0. This is a strong evidence that the mean filling amount is different than 530 mL.
- Reject H_0. This is not enough evidence to conclude that the mean filling amount is different than 530 mL.
- Fail to reject H_0. This is strong evidence that the mean filling amount is different than 530 mL.
- Fail to reject H_o. This is not enough evidence to conclude that the mean filling amount is different than 530 mL.

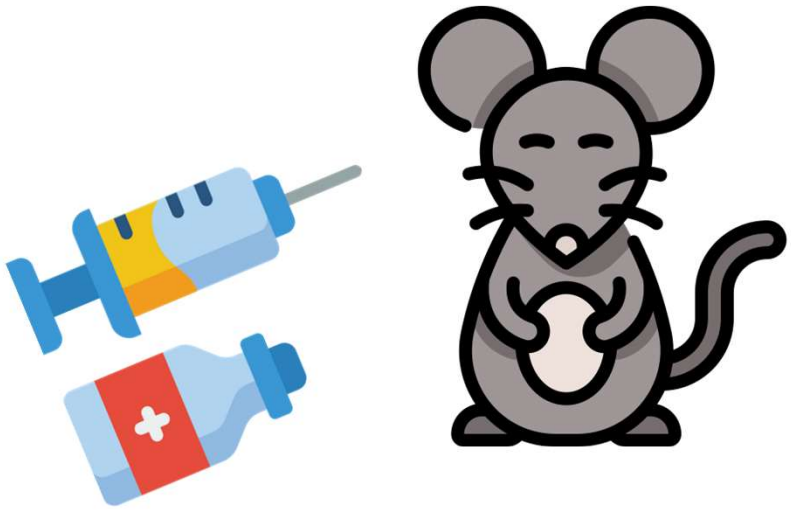
Significance Test for Mean

Regulations require that product labels on containers of food that are available for sale to the public accurately state the amount of food in those containers. Specifically, if milk containers are labelled to have 128 fluid ounces and the mean number of fluid ounces of milk in the containers is at least 128, the milk processor is considered to be in compliance with the regulations. The filling machines can be set to the labelled amount. Variability in the filling process causes the actual contents of milk containers to be normally distributed. A random sample of 12 containers of milk was drawn from the milk processing line in a plant and the amount of milk in each container was recorded.

The sample mean and standard deviation of this sample of 12 containers of milk were 127.2 ounces and 2.1 ounces respectively. Is there sufficient evidence to conclude that the packaging plant is not in compliance with the regulations? Provide statistical justification for your answer.

Hypothesis testing and p-values

A neurologist is testing the effect of a drug on a response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats response times is 1.05 seconds with a sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time?

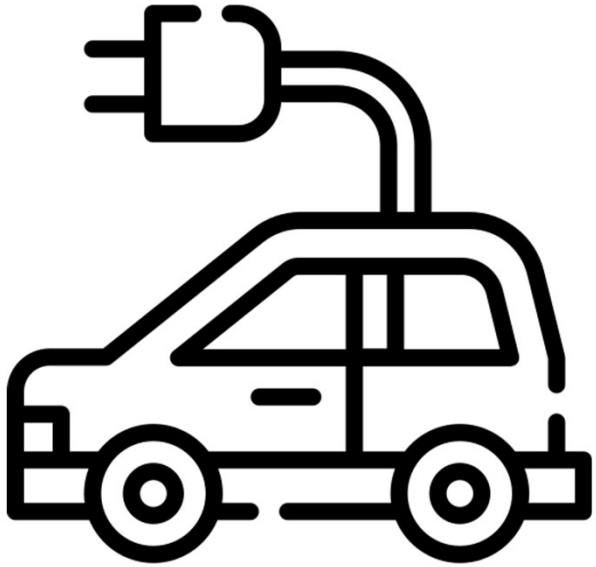


Small sample hypothesis test

The mean emission of all the engines of a new design needs to be below 20 ppm if the design is to meet new emission requirements. Ten engines are manufactured for testing purposes, and the emission level of each is determined. The emission data is

15.6, 16.2, 22.5, 20.5, 16.4, 19.4, 16.6, 17.9, 12.7, 13.9

Does the data supply sufficient evidence to conclude that this type of engine meets the new standard? Assume we are willing to risk a Type 1 error with probability of 0.01



Large Sample proportion hypothesis testing

We want to test the hypothesis that more than 30% of U.S. households have internet access (with a significance level of 5%). We collect a sample of 150 households and find that 57 have access.

Comparing Two Proportions

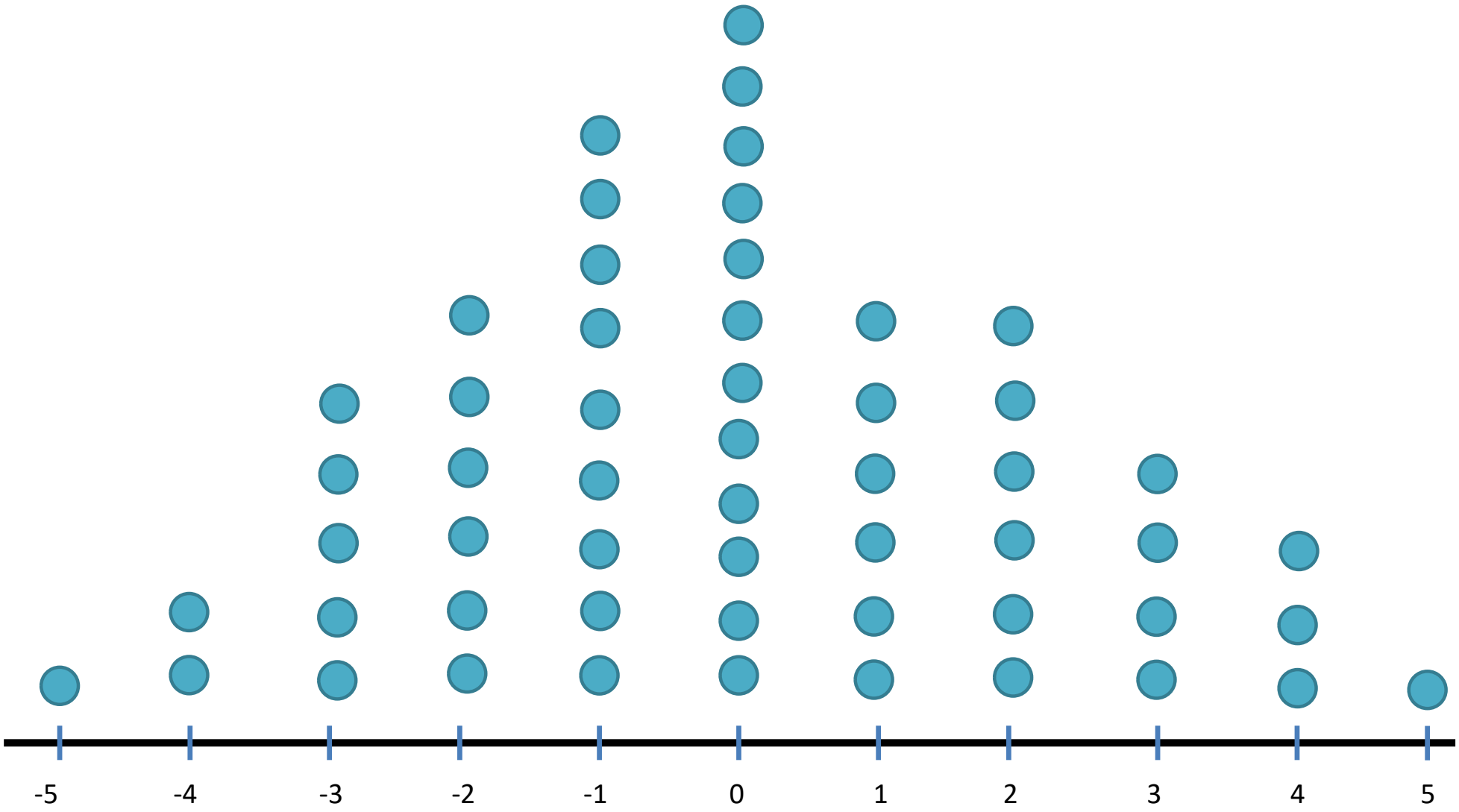
Statistical Significance of experiment

In an experiment aimed at studying the effect of advertising on eating behavior in children, a group of 500 children 6 to 12 years old, were randomly assigned to two different groups. After randomization, each child was asked to watch a cartoon in a private room, containing a large bowl of popcorns. The cartoon included 2 commercial breaks.

The first group watched food commercials, while the second group watched non-food commercials. Once the child finished watching the cartoon, the conductors of the experiment weighed the popcorn bowl to measure how many grams of popcorns the child ate. They found that the mean amount of popcorns eaten by the children who watched food commercials is 5 grams greater than the mean amount of popcorns eaten by the children who watched non-food commercials.

They re-randomized the results into two new groups and measured the difference between the means of the new groups. They repeated this for 55 times, and plotted the resulting differences as show in the image.

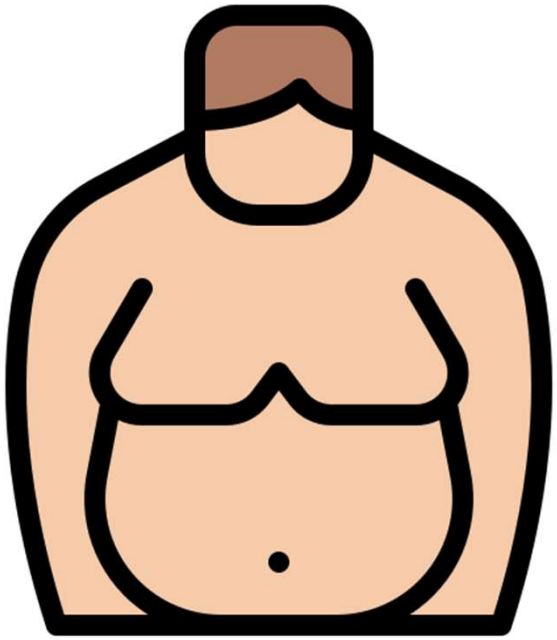
According to the multiple experiments, is the result of the experiment significant?


-5
-4
-3
-2
-1
0
1
2
3
4
5

Difference of Sample mean distribution

Confidence interval of difference of means

We are trying to test whether a new, low fat diet actually helps obese people lose weight. 100 randomly assigned obese people are assigned to group 1 and put once the low fat diet. Another 100 randomly assigned obese people are assigned to group 2 and put on a diet of approximately the same amount of food, but not as low in fat. After 4 months, the mean weight loss was 9.31 lbs for group 1 (s=4.67) and 7.40 lbs (s=4.04) for group 2.



Hypothesis test for difference of means

Chi-Square distribution introduction

Pearson's Chi-Square test (Goodness of fit)

Day	Mon	Tue	Wed	Thu	Fri	Sat
Expected	10	10	15	20	30	15
Observed	30	14	34	45	57	20

Chi-Square statistic for Hypothesis testing

Chi-square goodness of fit example

In the game rock-paper scissors, Kim expects to win, tie, and lose with equal frequency. Kim plays rock paper scissors often, but he suspected his own games were not following that pattern, so he took a random sample of 24 games and recorded their outcomes. Here are his results:

Outcome	Win	Loss	Tie
Games	4	13	7

He wants to use these results to carry out a χ^2 goodness of fit test to determine if the distribution of his outcomes disagrees with an even distribution.

What are the values of the test statistic and P-value for Kim's test?

(A) $\chi^2 = 5.25$

$0.05 < \text{P-value} < 0.10$

(B) $\chi^2 = 5.25$

$0.15 < \text{P-value} < 0.2$

(C) $\chi^2 = 21.875$

$\text{P-value} < 0.0005$

(D) $\chi^2 = 21.875$

$0.0005 < \text{P-value} < 0.001$

Filling out frequency table for independent events

One rainy Sunday morning, Amar woke up to hear his mom complaining about the house being dirty. “Mom is always grouchy when it rains.” Amar’s brother said to him.

So Amar decided to figure out if this statement was actually true. For the next year, he charted every time it rained and every time his mom was grouchy. What he found was very interesting rainy days and his mom being grouchy were entirely independent events. Some of his data are shown in the table below

Fill in the missing values from the frequency table

	Raining	Not Raining	Row Total
Grouchy			73
Not Grouchy			292
Column Total	35	330	365

Contingency table chi-square test

	Herb 1	Herb 2	Placebo	
#Sick	20	30	30	80
#Not Sick	100	110	90	300
Total	120	140	120	380

Chi-Squared test for Homogeneity

	Right	Left	Total
STEM	30	10	40
Humanities	15	25	40
Equal	15	5	20
Total	60	40	100

Chi-Squared test for association (Independence)

	Right Foot Longer	Left Foot Longer	Both Feet Same	Total
Right Hand Longer	11	3	8	
Expected				
Left Hand Longer	2	9	14	
Expected				
Both Hands Same	12	13	28	
Expected				
Total				

Advanced Regression (Inference and Transforming)

Introduction to inference about slope in linear regression

Conditions for Inference on slope

Confidence interval for the slope of a regression line

Atharv is interested in the relationship between hours spent studying and caffeine consumption among students at his school. He randomly selects 20 students at his school and records their caffeine intake (mg) and the amount of time spent studying in a given week. Here is computer output from a least squares regression analysis on his sample:

Predictor	Coef	SE Coef	T	P
Constant	2.544	0.134	18.955	0.000
Caffeine	0.164	0.057	2.862	0.010

S = 1.532
R-sq = 60%

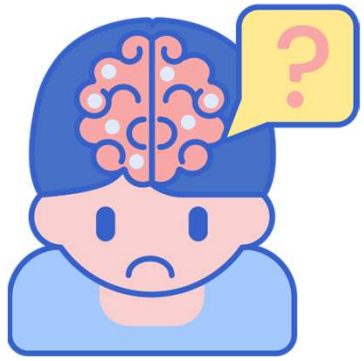
Assume that all conditions for inference have been met.
What is the 95% confidence interval for the slope of the least squares regression line?

Calculating t statistic for slope of regression line

James obtained a random sample of data on how long it took each of 24 students to complete a timed reaction game and a time memory game. He noticed a positive linear relationship between the times on each task. Here is computer output on the sample data:

Variable	n	Mean	StdDev	SE Mean
X = Reaction time	24	0.398	0.133	0.027
Y = Memory time	24	43.042	8.554	1.746

Predictor	Coef	SE Coef
Constant	37.200	5.579
Reaction	14.686	13.329



S = 8.515
R-sq = 5.2%

Assume that all conditions for inference have been met.

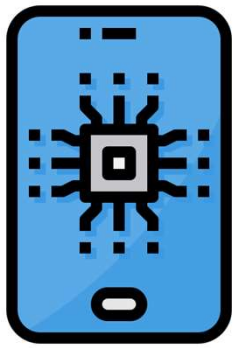
Calculate the test statistic that should be used for testing a null hypothesis that the population slope is actually 0?

Using a P-value to make conclusions in a test about slope

Andrea took a random sample of mobile phones and found a positive linear relationship between their processor speeds and their prices. Here is computer output from a least squares regression analysis on her sample:

Predictor	Coef	SE Coef	T	P
Constant	127.092	57.507	2.210	0.032
Speed	6.084	2.029	2.999	0.004

Andrea wants to test H_0: $\beta = 0$ $vs.$ H_a: $\beta > 0$. Assume that all conditions for inference have been met.
At the $\alpha = 0.01$ level of significance, is there sufficient evidence to conclude a positive linear relationship between these variables for all mobile phones? Why?



Using a confidence interval to test slope

Hashem obtained a random sample of students and noticed a positive linear relationship between their ages and there backpack weights. A 95% confidence interval for the slope of the regression line was 0.39 ± 0.23.

Hashem wants to use this interval to test H_0: $\beta = 0$ vs H_a: $\beta \neq 0$ at the $\alpha = 0.05$ level of significance. Assume that all conditions for inference have been met.

Anova – Calculating SST, SSW and SSB

Hypothesis testing with F-statistic