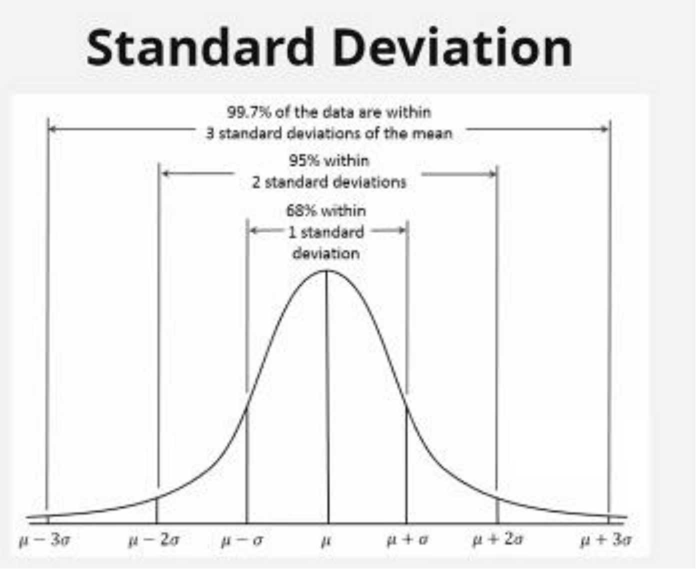
Statistics Fundamentals

Center – typical value of a data point.

- Median – middle value when the data are ordered from least to greatest.
- Mean – sum of values/number of values.

Spread – variation of the data.

- Range
- Standard deviation – measure of the spread of the distribution of data.
    - A low standard deviation indicates that much of the data lies close to the mean.
    - A high standard deviation means the data is spread out.


Standard Deviation
99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

-

Mean = The average

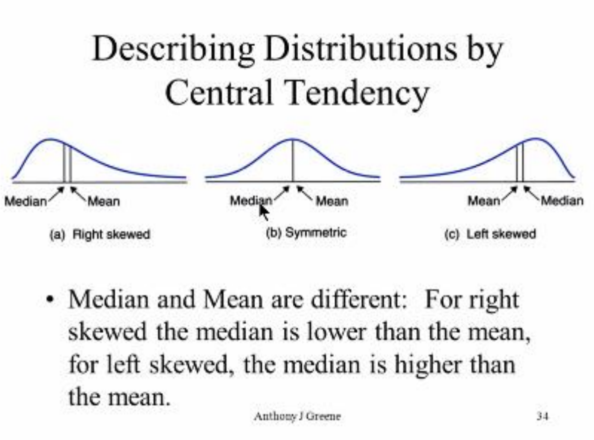
Median = the middle value

Skewed Distributions = Concentrated values on left/right side

Mean displays the average of the distribution.

Median displays the middle values. It splits the data in half and picks the centermost value.

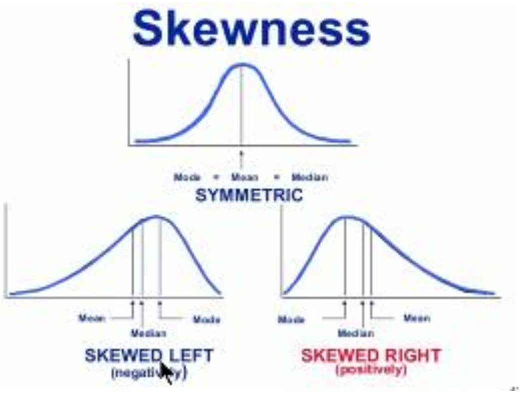
The mode is the value that appears most often in a set of data values.



Describing Distributions by Central Tendency

- Median and Mean are different: For right skewed the median is lower than the mean, for left skewed, the median is higher than the mean.

Anthony J Greene 34



Random Sampling

Representative – only member of the population you want are being studied.

Random – every member of the population has an equal chance to be sampled.

Estimate = population size * sample proportion

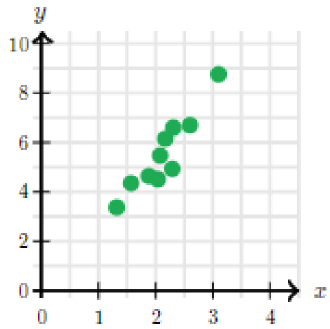
Correlation - the dependence of two variables on each other. It describes how two variables change in relation to each other.

X = independent variable – horizontal axis

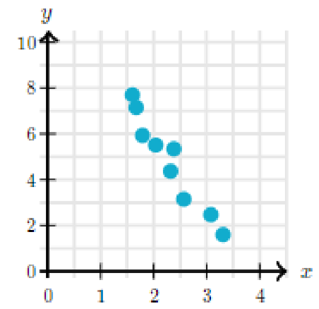
Y = dependent variable – vertical axis

In a negative correlation, as values of x increase, values of y decrease.

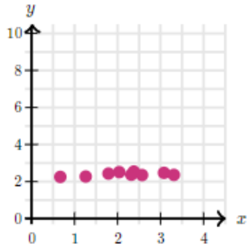
Positive correlation: As x increases, y tends to increase.

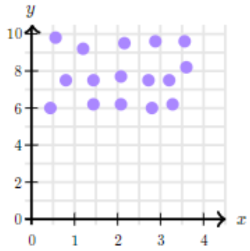


Negative correlation: As x increases, y tends to decrease.



No correlation: As x increases, y stays about the same or has no clear pattern.





Probability – likelihood that an event will occur.

Example: The probability of an event E is equal to the number of ways it can happen divided by the total number of outcomes. For example, when flipping a coin, the probability of heads is: 1/2 = 0.5

Mutually Exclusive/Disjoint Events – cannot occur at the same time.

Conditional Probability – event A occurs given that event B has occurred: $P(A|B)$.

Complement of an Event – event will not occur: $P(A^1)$

Intersection – events A and B both occur.

Union – events A or B occur.

Dependent – occurrence of event A changes the probability of event B.

Independent – occurrence of event A does not change the probability of event B.

Rule of Multiplication

The probability that Event A and Event B both occur is equal to the probability that Event A occurs multiplied by the probability that Event B occurs, given that A has occurred.

We have a box containing 10 red marbles and 5 blue marbles. We draw two marbles without replacement. What is probability both are red?
A being that the first marble is red, and B being that the second marble is red.

The probability of the first marble being red is P(A) = 10/15.

After the first marble is drawn, we have 14 marbles remaining, 9 of these marbles being red. So, P(B|A) = 9/14.

The probability that Event A or Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.

Rule of Multiplication

The probability that Event A and Event B both occur is equal to the probability that Event A occurs multiplied by the probability that Event B occurs, given that A has occurred.

We have a box containing 10 red marbles and 5 blue marbles. We draw two marbles without replacement. What is probability both are red?
A being that the first marble is red, and B being that the second marble is red.

The probability of the first marble being red is P(A) = 10/15.

After the first marble is drawn, we have 14 marbles remaining, 9 of these marbles being red. So, P(B|A) = 9/14.

The probability that Event A or Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.

Combination – collection of items where order is not considered.

n!/(r!(n-r)!)

Let us say we have 15 students and are choosing 10. The order does not matter. How many combinations are there?

15!/(10!(15-10)!)
3,003

In combinations where repetition does matter, it is a little different.

(r+n-1)!/r!(n-1)!

Let us say we have 10 options for pizza toppings, and we can select 3, with duplicates being allowed (e.g. we could have triple pepperoni, or double pepperoni with sausage).

(3+10-1)!/3!(10-1)! = 220

Permutation – order is important. Example: If your phone passcode is 1234, the same numbers in a different order (1324) will not work. This is heavily used in hacking and decoding. (n means number)

Permutations with repetition = n^2

For example, if we are making a 4-digit passcode and can choose from 10 numbers (0, 1, 2, 3, 4, 5, 6, 7, 8, 9), the number of possible permutations is:

10^4 = 10,000

That is not too many - maybe we want to make something a little bit more secure!

As r increases, we will have many more permutations. A 10-digit passcode has 10,000,000,000 possibilities!

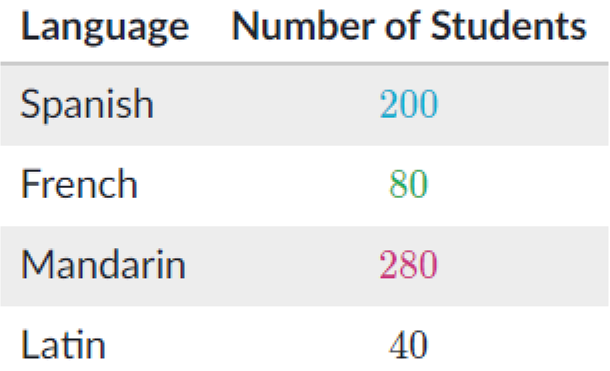
Permutations without repetition = n!/(n-r)!

For example, say we have 15 students and want to select 10 of them. Order matters. We want to know how many permutations are possible.

15!/(15-10)! = 10,897,286,400

(That is a lot of permutations!)

Table summarizes the data using rows and columns.

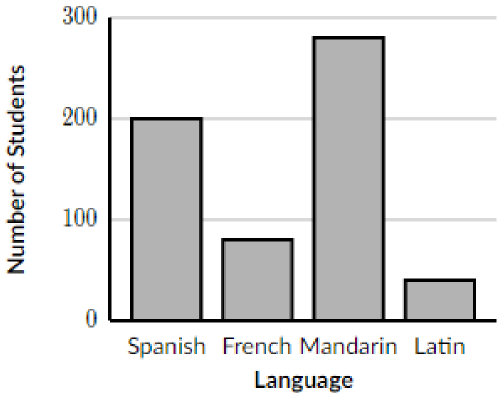


Language	Number of Students
Spanish	200
French	80
Mandarin	280
Latin	40

Explanation: The left column contains the languages, and the right column contains the number of students who want to study each language.

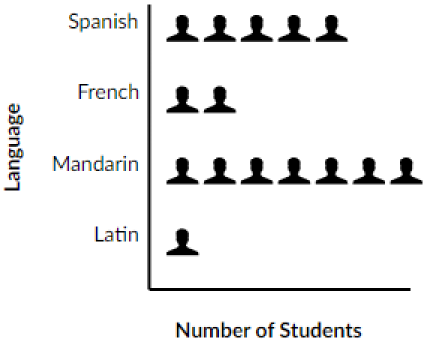
Using the table, we can conclude that the number of students want to study Mandarin 280start and the number of students who want to study Spanish or French (200+80) are the same.

Vertical bar chart lists the categories of the qualitative variable along a horizontal axis and uses the heights of the bars on the vertical axis to show the values of the quantitative variable.



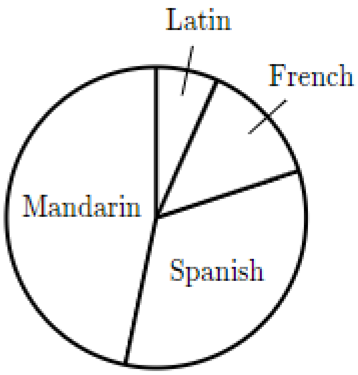
Explanation: The heights of the bars show the number of students who want to study each language. Using the bar chart, we can conclude that the greatest number of students want to study Mandarin and the least number of students want to study Latin.

Pictograph is like a horizontal bar chart but uses pictures instead of the lengths of bars to represent the values of the quantitative variable.



Explanation: Each 👤 represents 404040 students. The number of pictures shows the number of students who want to study each language. Using the pictograph, we can conclude that twice as many students want to study French as want to study Latin.

Circle graph (or pie chart) is a circle that is divided into as many sections as there are categories of the qualitative variable.



Explanation: The area of each section represents the fraction of students who want to study that language. Using the circle graph, we can conclude that just under1/2 the students want to study Mandarin and about 1/3 want to study Spanish.

<u>Random Samples</u>

A sample provides information about a population without having to survey the entire group. Representative means that the sample includes only members of the population being studied.

Random means that every member of the population being studied has an equal chance to be selected for the sample.

A good sample is representative and random.

- Representative means that the sample includes only members of the population being studied.
- Random means that every member of the population being studied has an equal chance to be selected for the sample.

Overfitting – high variance - Data that overrepresents a population.

Underfitting – high bias.

The sampling methods below lead to bad samples, ones that are not representative of the feelings of all students at the high school:
Method 1: Surveying all athletes at the high school
This sample overrepresents the feelings of student athletes. Student athletes may feel differently about the proposed cuts than other students.

Method 2: Emailing a survey to all students and using the responses
This sample includes only students who respond to the email. Respondents are likely to have stronger feelings about the topic than students who do not respond. This sample overrepresents students with stronger feelings.

Method 3: Calling a random sample of families with students enrolled in the school district
This sample includes data from outside the population being studied, all students at a certain high school.

<u>Inference –</u> an educated conclusion.

<u>How can we use sample data to make an estimate?</u>
We can use data from a random sample to estimate the number in the population having a particular attribute. To make an estimate, we need to know the sample proportion and the population size.

estimate = population size x sample proportion

Example: 6060 seniors at a particular high school are randomly selected for a survey. 666 report driving themselves to school at least once a week. If there are 400400400 seniors at this high school, what is a reasonable estimate of the total number of seniors that drive themselves to school at least once a week?

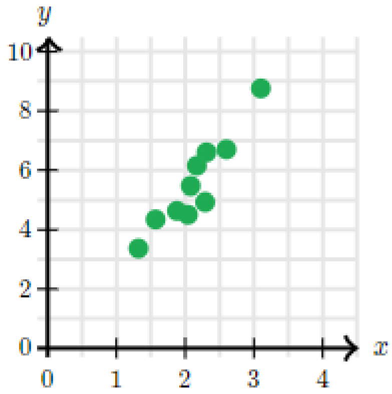
The population is all seniors at the high school. The sample proportion is the fraction of seniors who reported driving themselves to school at least once a week.

Population size = 400

Sample proportion = 6/60

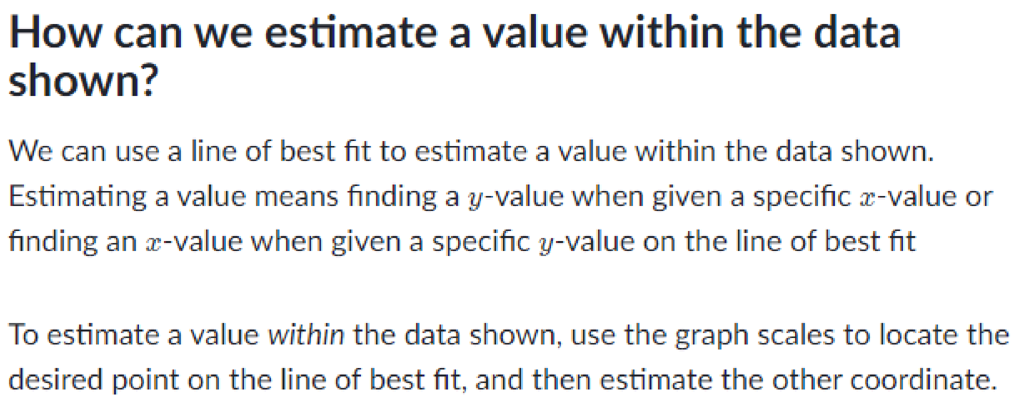
Estimate = 400 x 6/60 = 40

A **scatterplot** displays data about two variables as a set of points in the xy- plane.



A scatterplot is a key tool to determine if there is a relationship between the values of two variables.

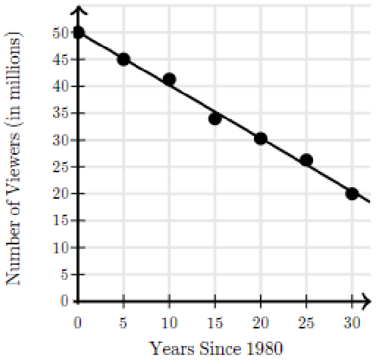
The purpose of **Supervised Learning** is Predictive Analytics using past structured data



How can we estimate a value within the data shown?

We can use a line of best fit to estimate a value within the data shown. Estimating a value means finding a y-value when given a specific x-value or finding an x-value when given a specific y-value on the line of best fit

To estimate a value *within* the data shown, use the graph scales to locate the desired point on the line of best fit, and then estimate the other coordinate.



The scatterplot shows data on the average number of Americans who viewed the nightly news from 1980 to 2010. The line of best fit is shown on the scatterplot. Based on the line of best fit, which of the following is the best estimate of the number of viewers in 1987?

We must use the line of best fit to find the point corresponding to an x-value of 7 years since 1980.

We can use a line of best fit to estimate a value beyond the data shown. Estimating a value means finding a y-value when given a specific x*x*x-value or finding an x-value when given a specific y-value on the line of best fit.

To estimate a value *beyond* the data shown, extend the graph scale and line of best fit to include the desired point and then estimate the value of the other coordinate.

A line of best fit usually shows two key features.

- The **y-intercept**, b, is the y-value when x = 0.
- The **slope**, *m*, is the change in *y* when *x* increases by 1.

Slope-intercept form, y=mx+b

Things to remember:

A line of best fit can be estimated by drawing a line so that the number of points above and below the line is about equal.

We can use a line of best fit to estimate values within or beyond the data shown.

- To estimate a value *within* the data shown, use the graph scales to locate the desired point on the line of best fit, and then estimate the other coordinate.
- To estimate a value *beyond* the data shown, extend the graph scale and line of best fit to include the desired point, and then estimate the value of the other coordinate.

The equation for a line of best fit is: y=m(x) + b where (x,y) represents any point which satisfies this equation.

- The ***y*-intercept**, *b*, is the *y*-value when x = 0.
- The **slope**, *m*, is the change in *y* when *x* increases by 1.

In context, the meaning of the slope of a line of best fit must be explained with the appropriate units.

- The slope specifies the change in *y* when *x* increases by 1.

How is the probability of an event calculated?

An event could be the outcome of any random process such as the toss of a fair coin, the roll of a fair number cube, or the random selection of an item from a group.

$$P(A) = \frac{\text{ways } A \text{ can happen}}{\text{possible outcomes}}$$

Artificial Intelligence -> Machine Learning -> Supervised Learning/Unsupervised Learning

Learning/Unsupervised Learning = Unstructured Data

Unsupervised Learning is the preprocessing step for Supervised Learning

Supervised Learning – data that is Structured and/or Labeled

Deep Learning = The advanced Form of Machine Learning

General purpose of classification is to predict categories