

Measures

Central Tendency and Dispersion

- Index



- Mean

- Median

- Mode

- Range

- Statistical Dispersion

- Quartiles

- Interquartile Range

- Variance

- Standard Deviation

- Frequency Distribution

- Index

- Excel Formulas:

- Central tendency
- Dispersion
- Frequency Tables

Continuous Data (one variable)

- Normal Curve
- Inverse Normal
- Standard Normal Curve
 - Z-Score

- Index

- Continuous Data (two variables) ▫

- Scatter Plotting
- Line of best fit
- Residuals
- Correlation

- Mean

- The mean (average) of a data set is found by adding all numbers in the data set and then dividing by the number of values in the set.

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

f - frequency of item i

x - value of item i

n - number of terms = $\sum_{i=1}^k f_i$

- Median

- The median is the middle value when a data set is ordered from least to greatest.

Determine the median for the following sets of data.

$\{6, 7, 3, 2, 8, 1, 4\}$

$\{8, 12, 4, 10\}$

- Median

- The median is the middle value when a data set is ordered from least to greatest.

Determine the median for the following sets of data.

Put into ascending order first.

{1, 2, 3, 4, 6, 7, 8} 4

{4, 8, 10, 12} 9

- Mode

- The mode is the number that occurs most often in a data set.

Determine the mode for the following sets of data.

$\{4, 6, 2, 4, 1, 9, 4\}$

$\{1, 2, 3, 4\}$

$\{1, 2, 3, 2, 1, 5\}$

- Mode

- The mode is the number that occurs most often in a data set.

Determine the mode for the following sets of data.

{1, 2, 4, 4, 4, 6, 9}

Put into ascending order first.

4

{1, 2, 3, 4}

none

{1, 1, 2, 2, 3, 5}

1, 2 bimodal

Statistical Dispersion

Dispersion in statistics is a way of describing how spread out a set of data is. When a data set has a large value, the values in the set are widely scattered; when it is small the items in the set are tightly clustered.

- Range

- The range of a data set is the difference between the lowest and highest value.

- Interquartile Range (IQR)

- A measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, $IQR = Q_3 - Q_1$.

- Quartiles in Excel

- Quartile.exc - Excludes the median when finding the Q1 and Q3

Quartile.inc - Includes the median when finding the Q1 and Q3

- Quartiles in Excel

Ex:

1				
2				
2			EXC	INC
4		Median	4	
5		Q1	2	2
8		Q3	8	6.5
9				

Q3 Exc:

5, 8, 9

Q3 Inc:

4, 5, 8, 9

- Variance

Deviation refers to the distance between two values. In this case, between the values in the set and the mean.

Variance (σ^2) is the average squared deviation.

- Standard Deviation

- **Standard deviation** (σ) is the square root of variance. It brings the units of variance back to the units of the original data.

Practice:

$\{-10, 0, 10, 20, 30\}$

Mean:

Median:

Mode:

Range:

IQR:

Variance:

Standard Deviation:

$\{7, 8, 9, 9, 10, 10, 11, 12\}$

Mean:

Median:

Mode:

Range:

IQR:

Variance:

Standard Deviation:

Key:

$\{-10, 0, 10, 20, 30\}$

Mean: 10

Median: 10

Mode: none

Range: 40

IQR: 30

Variance: 200

Standard Deviation: $\sqrt{200} \approx 14.1$

$\{7, 8, 9, 9, 10, 10, 11, 12\}$

Mean: 9.5

Median: 9.5

Mode: 9, 10

Range: 5

IQR: 2

Variance: 2.25

Standard Deviation: 1.5

- Frequency Distribution Tables

Grades	Frequency
40-49	3
50-59	5
60-69	6
70-79	9
80-89	8
90-100	7

A way to group data to compare sizes of classes.
This is also how histograms are made.

There is a function in Excel for creating frequency tables quickly.

`=Frequency(data column, bins column)`

Working in Spreadsheets

- Measures of Central Tendency

- =Average(
=Median(
=Mode(
=Min(
=Max(

- Measures of Dispersion

- Range: =Max()-Min()

Quartile: =Quartile(Range,#)

IQR: =Quartile(Range,3)-Quartile(Range,1)

Variance: =Var(

Standard Deviation = SQRT(variance cell)

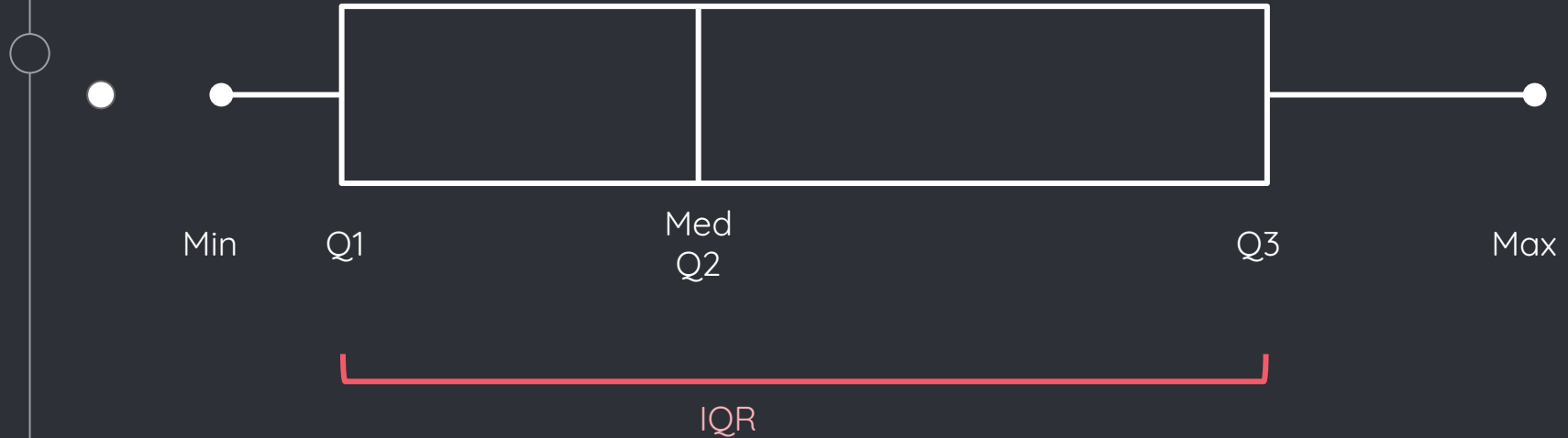
Or = STDEV(

- Outliers

- Outliers are extreme values of data that fall far away from the bulk of the data.

Outliers are located above and below the 3rd and 1st quartiles, respectively, a distance of $1.5 * IQR$

- Outliers - Box and Whisker Plot



Outliers are located above and below the 3rd and 1st quartiles, respectively, a distance of $1.5 * IQR$

- Outliers

- To find outliers, first identify the upper and lower bounds for reasonable data.

Upper: $=\text{Quartile}(\text{data}, 3) + 1.5 * \text{IQR cell}$

Lower: $=\text{Quartile}(\text{data}, 1) - 1.5 * \text{IQR cell}$

- Frequency Table

- First, we have to decide on a reasonable bin size for the table. I usually start by taking my range and dividing by a number of groups I'd like to count. Round to a nearby whole number and use that as your bin size. Bins can be edited to make them more informative.

- Frequency Table

- Create a column with your chosen bins' upper limits.

Ex:

Bins
0-2000
2001-4000
4001-6000
6001-8000

On Excel
2000
4000
6000
8000

- Frequency Table

- In the column to the side of your bins, in the first row:

=Frequency(data,bins)

Excel will automatically sort your data into frequencies.

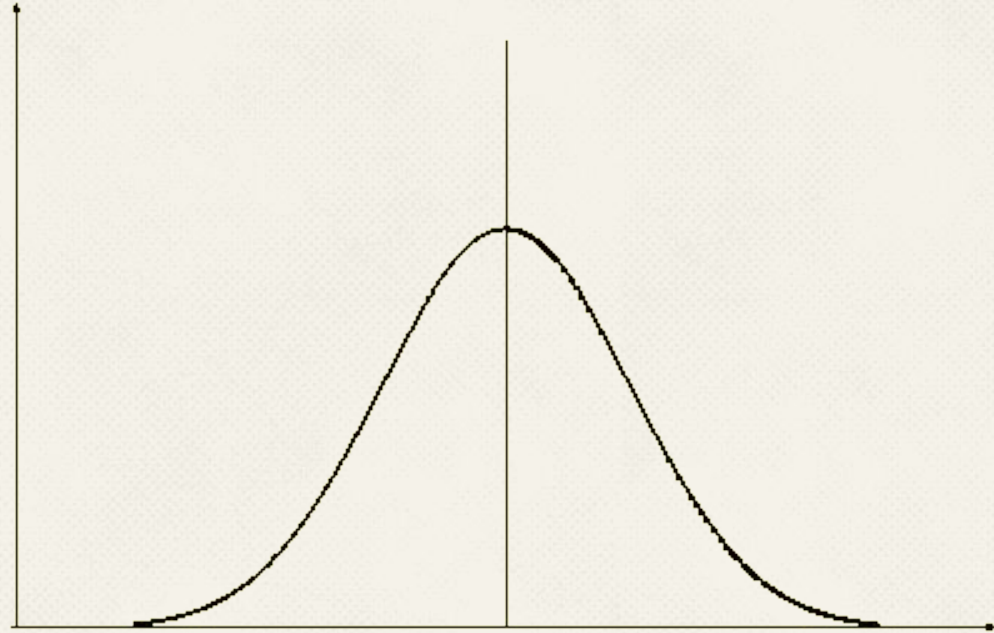
Normal Distributions

Representing Continuous Data

What is a normal distribution?

A normal distribution is one where the frequency of values in a data set is symmetric about the mean.

When graphed as the value along the x-axis and the frequency along the y-axis, the data forms a bell shaped curve.

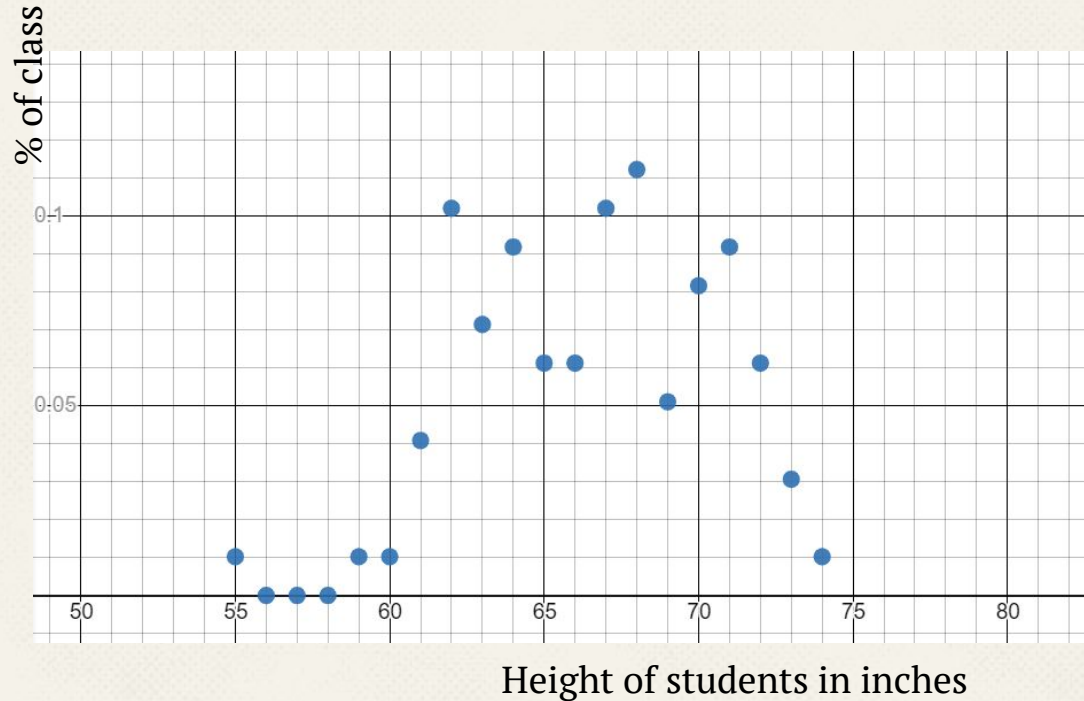


What is a normal distribution?

"A normal distribution arises in nature when many different factors affect the value of the variable."¹

To the right you can see the distribution of heights from students in an IB math class.

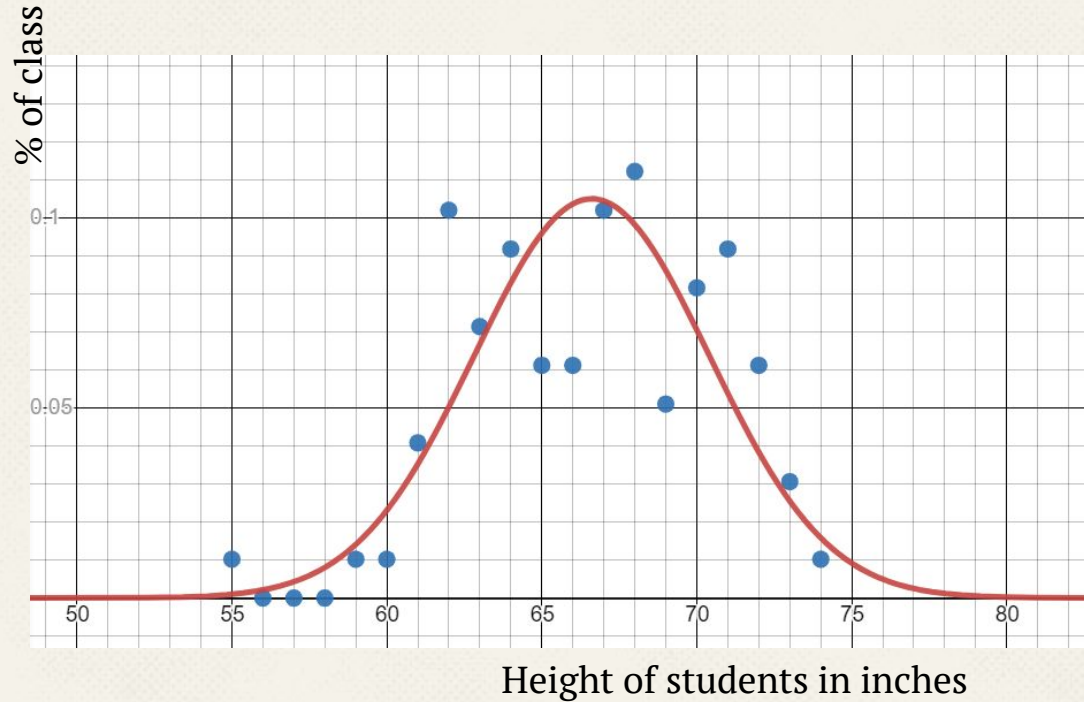
Do you think this data is normally distributed?



¹ Hease, Michael, et al. *Mathematics: Analysis and Approaches SL*. Hease Mathematics, 2019.

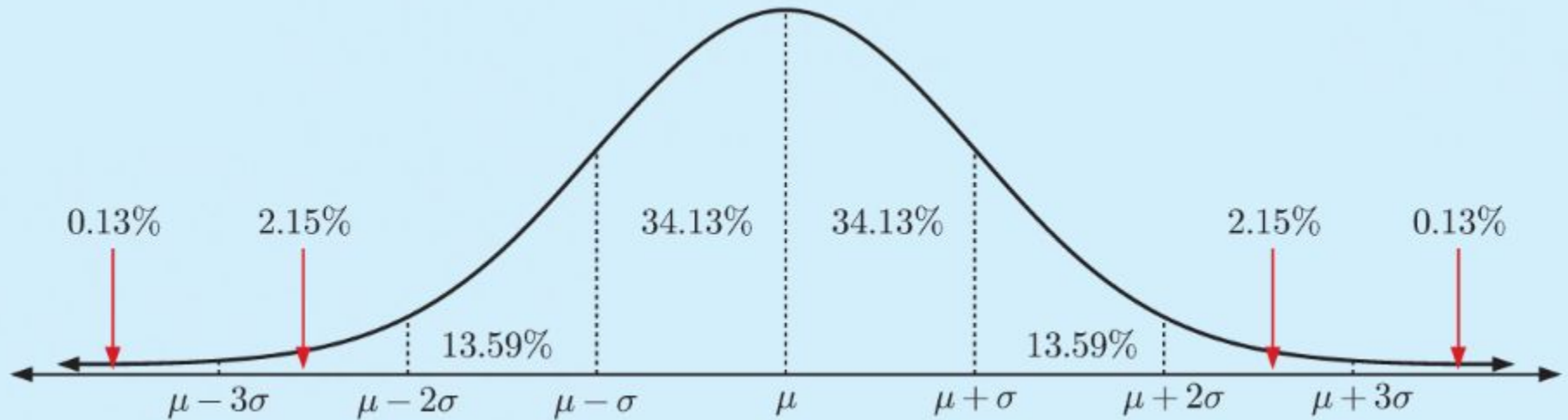
What is a normal distribution?

Now you can see the the normal curve should look like for the mean and standard deviation of our heights.



Empirical Rule

The empirical rule is a shortcut for approximating the proportion of the area under the normal curve based on how many standard deviations from the mean you are looking.



Example

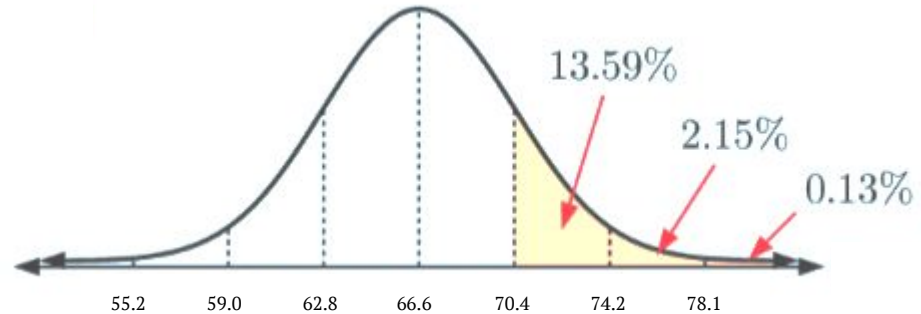
For the group of students represented on the previous slides, the mean was 66.6" and the standard deviation was 3.82".

State the approximate percentage of students whose height is between 62.78" and 70.42".

State the percentage students whose height is above 70.42".

Empirical Rule

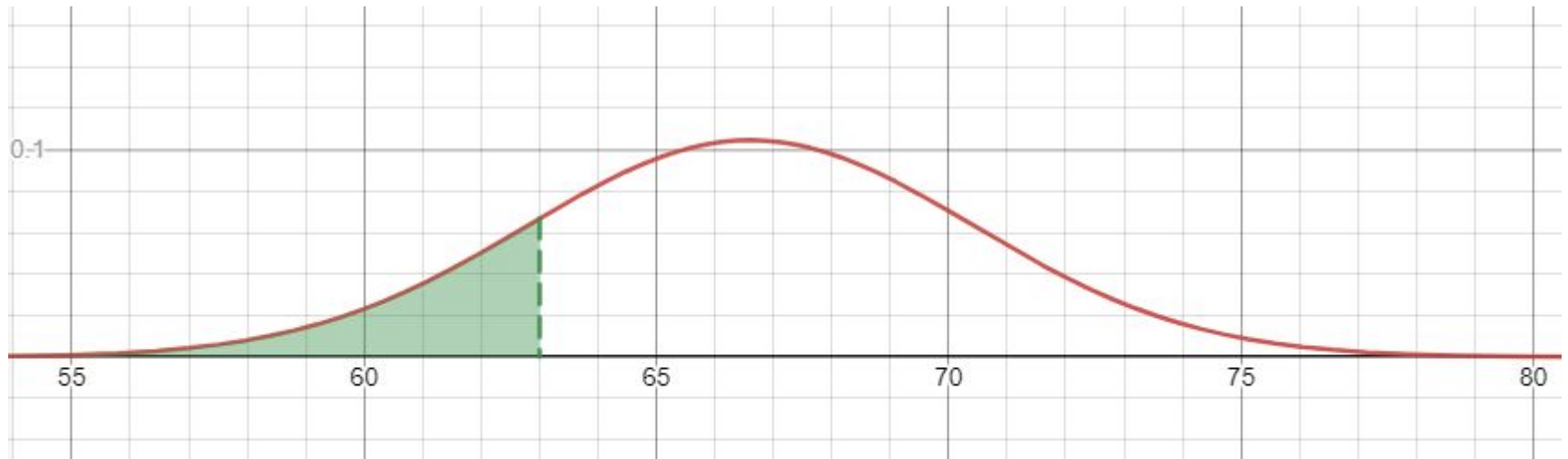
You should have found that because 70.42 is one standard deviation above the mean, the proportion of the population whose heights are more than 70.42" is approximately 15.87%. This also means that there is a 15.87% chance (or probability of 0.1587) that a particular person will have a height above 70.42 in.



Using Formulas to Find Proportions

Not all questions will refer to values that fall on a standard deviation, however. Let's proceed with a formula in Excel to find the area under the Normal Curve at other values.

This time, I'm interested to see what percentage of the population has a height below 5'3".



Using Formulas to Find Proportions

```
=Norm.dist(63,mean,st.dev,TRUE)
```

Use True if you want less than or equal to 63.

Use False if you want exactly equal to 63.

[illegible]

Inverse Normal Distribution

The inverse of a function switches the functions domain and range or input and output.

The normal distribution can input a value that the variable can take on and output a probability.

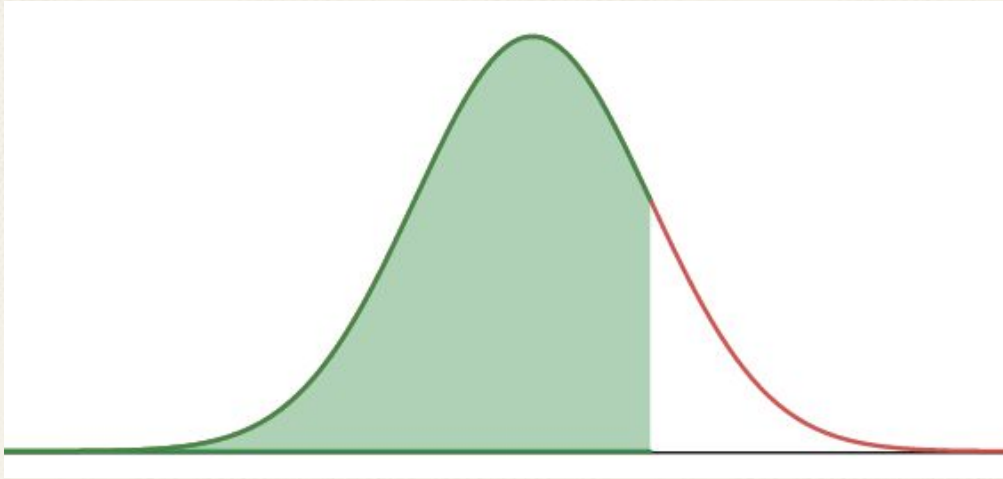
The inverse normal distribution can input a probability and tell you the boundary value that produced it.

To learn more about the two distributions, check out this video.



Inverse Normal Distribution

Find k for which $P(X < k) = .84$



=Norm.INV(.84,mean,
st.dev)

The number it provides is
the height of the person
who is at the 84th
percentile.

Standard Normal Distribution

The standard normal curve is one with a mean of 0 and a standard deviation of 1.

That's all.

The purpose of using the standard normal curve is to compare distributions that normally have different means and standard deviations. This process is called standardizing the curves.

Standardizing

Standardizing is just a series of linear transformations to transform any normal distribution into a proportional standard normal distribution.

Standardizing

Standardizing is just a series of linear transformations to transform any normal distribution into a proportional standard normal distribution.

Consider a normal distribution with $\mu = 215$ and $\sigma = 3$.

Suppose $x = 211$.

What is the proportional x value on a standard normal distribution?

Standardizing

Consider a normal distribution with $\mu = 215$ and $\sigma = 3$.

Suppose $x = 211$.

What is the proportional x value on a standard normal distribution?



Standardizing

The whole graph needs to be shifted left 215 units so that the curve is centered at 0.

Then the graph is scaled down by $1/3$ so that the standard deviation is 1.

Standardizing

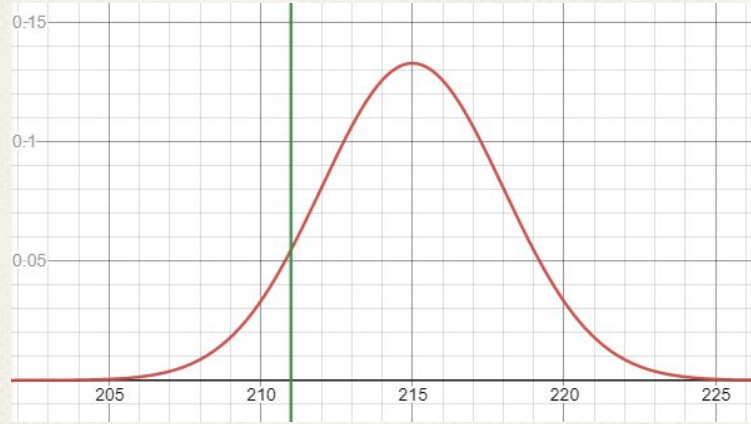
The whole graph needs to be shifted left 215 units so that the curve is centered at 0.

Then the graph is scaled down by $1/3$ so that the standard deviation is 1.

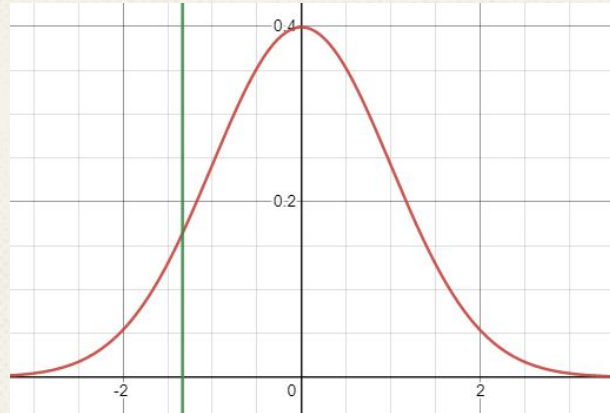
All x-values on the graph, including 211, will be scaled accordingly.

To find the new proportional x-value: $\frac{211 - 215}{3} = -\frac{4}{3}$

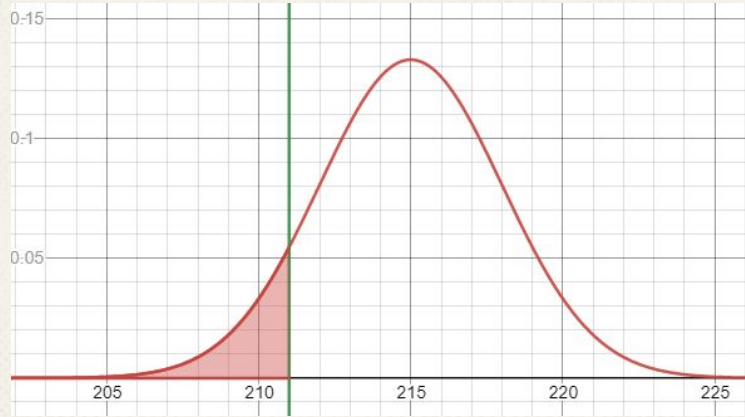
Standardizing



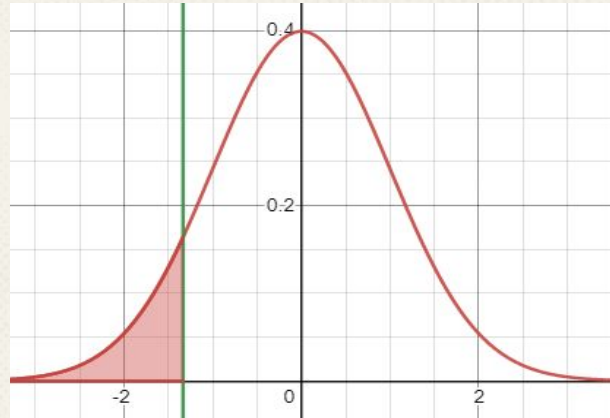
The area under the curve around proportional x-values is equal.



Standardizing



The area under the curve around proportional x-values is equal.



Z-Score

The proportional x-value on a standard normal distribution is called the **Z-score** of the x-value.

We would say that the **z-score** of $x=211$ on the normal distribution with $\mu = 215$ and $\sigma = 3$ is **$-4/3$** .

Z-Score

The proportional x-value on a standard normal distribution is called the Z-score of the x-value.

We would say that the z-score of $x=211$ on the normal distribution with $\mu = 215$ and $\sigma = 3$ is $-4/3$.

The z-score formula is just the transformations we performed:

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

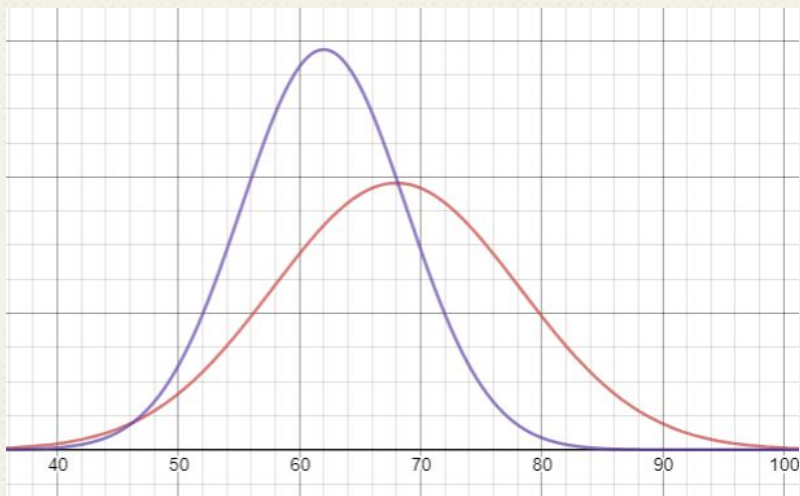
Example 1

Kelly scored 73% on a History exam, where the class mean was 68% and the standard deviation was 10.2%. In Mathematics she score 66% on an exam, where the class mean was 62% and the standard deviation was 6.8%. In which subject did Kelly perform better compared with the rest of her class?

In this case, we have two class sets of data which have different means and standard deviations. We could say Kelly performed better in History, but how did she do compared to the rest of the class?

Example 1

Below are the graphs of the two distributions. History class in red and Mathematics in purple. The area under the curve represents all of the student's and their scores because the y-axis represents frequency of each score.



The further to the right of the mean, the better Kelly has performed compared to the class because more students scored below her.

Example 1

A z-score tells you how many standard deviations away from the mean and in what direction a certain value is.

To find Kelly's z-score in History class, fill in all of the values related to History class.

$$X = 73, \quad \mu = 68, \quad \sigma = 10.2$$

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{73 - 68}{10.2}$$

$$Z = 0.490$$

Example 1

A z-score tells you how many standard deviations away from the mean and in what direction a certain value is.

To find Kelly's z-score in History class, fill in all of the values related to History class.

$$X = 73, \quad \mu = 68, \quad \sigma = 10.2$$

$$\begin{aligned} Z &= \frac{X - \mu}{\sigma} \\ Z &= \frac{73 - 68}{10.2} \\ Z &= 0.490 \end{aligned}$$

This result means that Kelly was 0.49 or approximately half a standard deviation above the average class score.

Example 1

Find Kelly's z-score for Mathematics. In which class did she outperform more of her classmates?


$$X = 66, \quad \mu = 62, \quad \sigma = 6.8$$

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{66 - 62}{6.8}$$

$$Z = 0.588$$

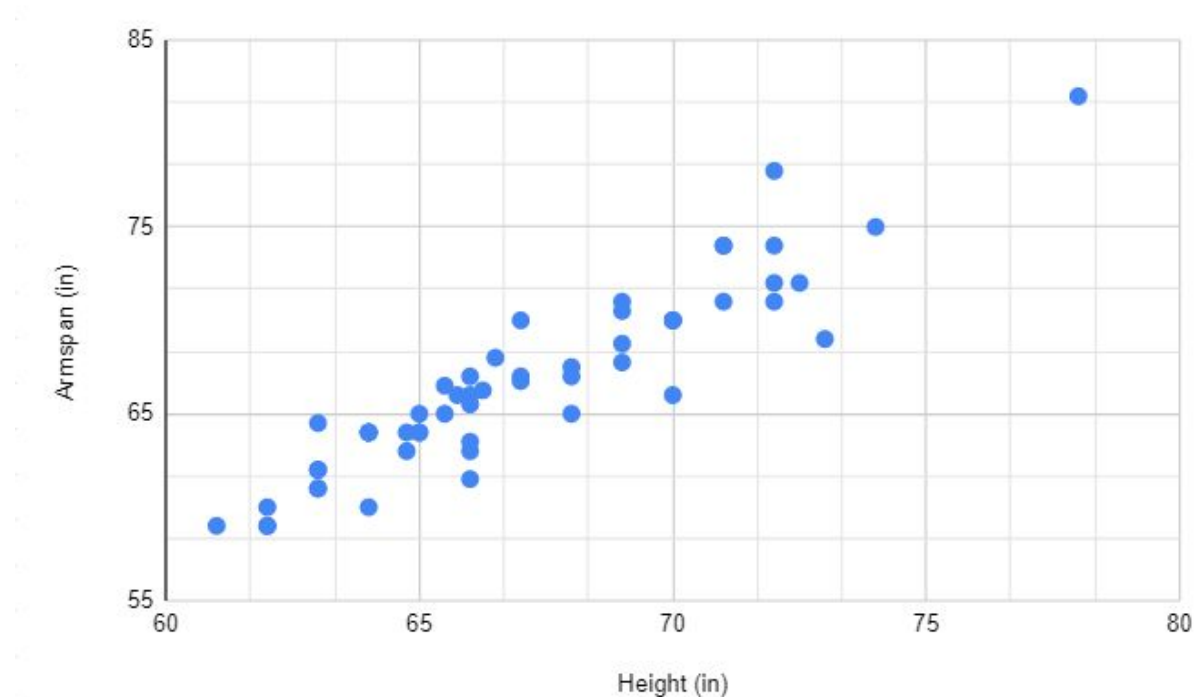
Based on these results, we can conclude that Kelly performed better, compared to her classmates, in Mathematics.



Scatter Plotting & Linear Correlation

Line of Best Fit

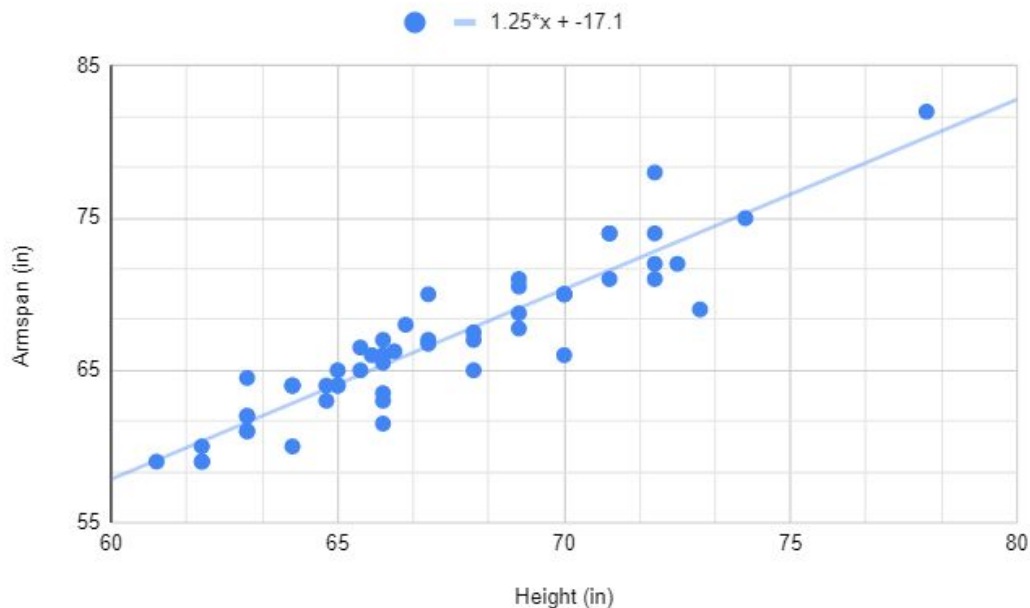
To the right are the class measurements of height and armspan from my IB class in 2019.



Line of Best Fit

The proper line of best fit should have the following features:

1. Goes through the mean point (\bar{x} , \bar{y})
2. Has the minimum absolute distance to the points in the scatter plot.



Residuals



Residuals

A residual is the difference between the observed value and the expected.

We will treat the y-coordinate of the point on the scatter diagram as the observed value, and the y-coordinate on the line of best fit as the expected value.

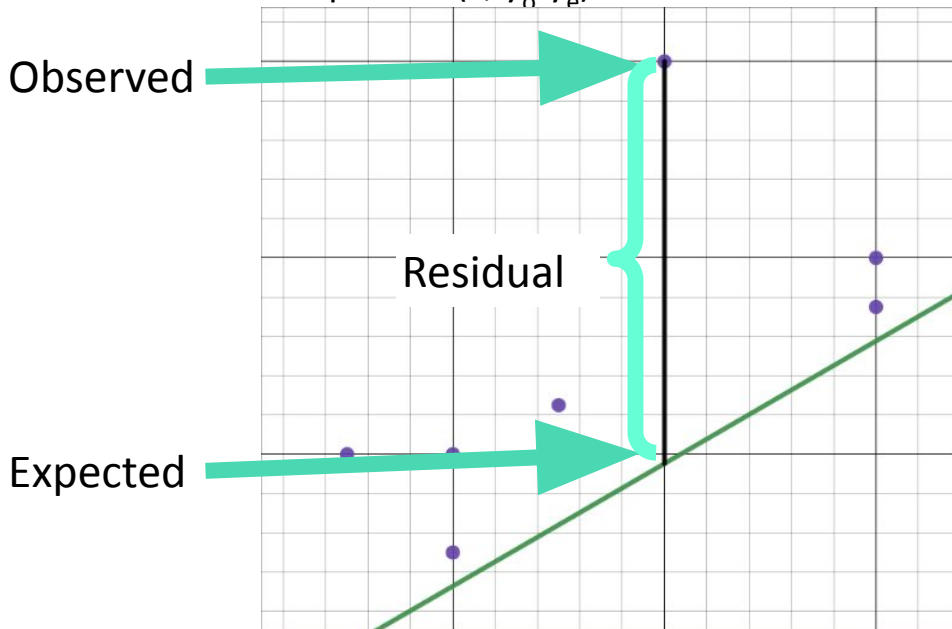
Residuals are plotted $(x, y_o - y_e)$

Residuals

A residual is the difference between the observed value and the expected.

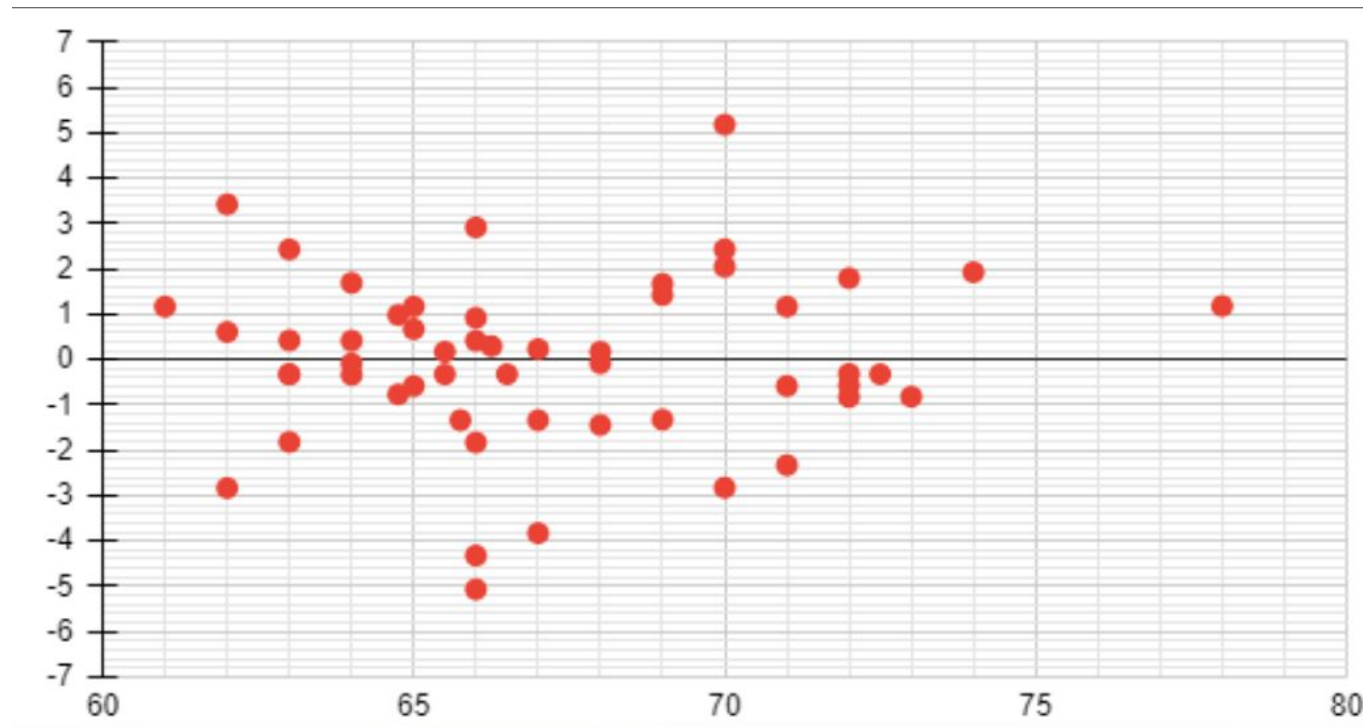
We will treat the y-coordinate of the point on the scatter diagram as the observed value, and the y-coordinate on the line of best fit as the expected value.

Residuals are plotted $(x, y_o - y_e)$



Residuals

The goal with a good line of best fit is for the residuals to be evenly spread above and below the x-axis and to have no distinguishable pattern.

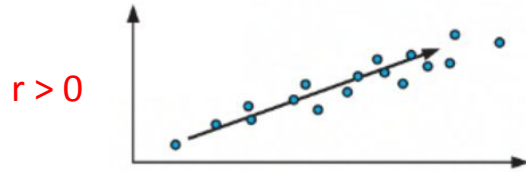


Correlation

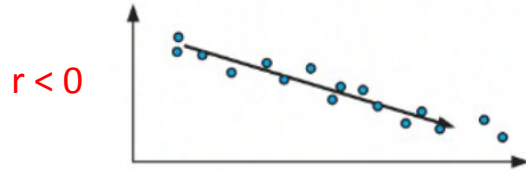


Correlation refers to the relationship between two variables.

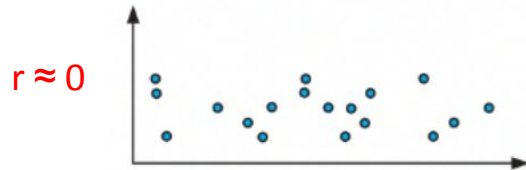
We can describe the direction of correlation:



For a generally *upward* trend, we say that the correlation is **positive**. An increase in the independent variable generally results in an increase in the dependent variable.



For a generally *downward* trend, we say that the correlation is **negative**. An increase in the independent variable generally results in a decrease in the dependent variable.

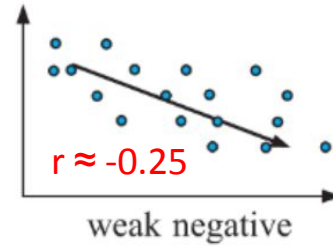
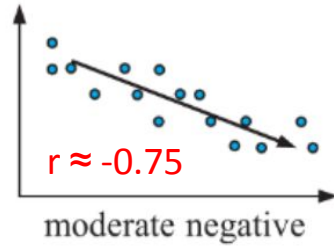
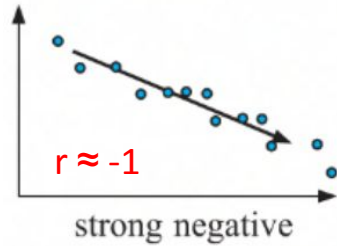
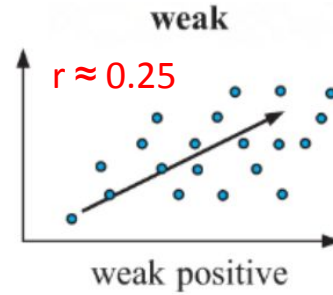
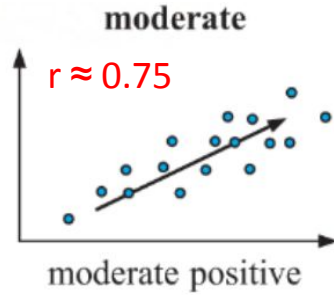
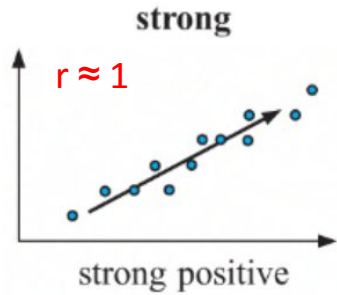


For *randomly scattered* points, with no upward or downward trend, we say there is **no correlation**.

Correlation

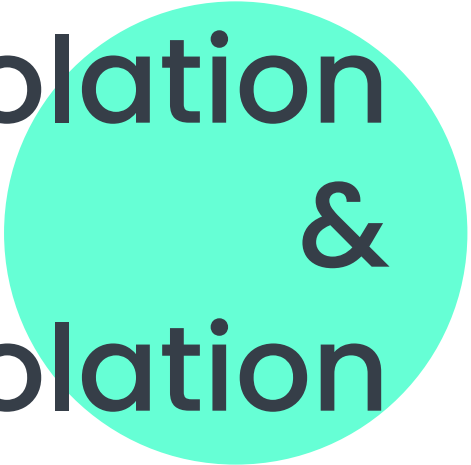
Correlation refers to the relationship between two variables.

We can describe the strength of correlation:



Correlation

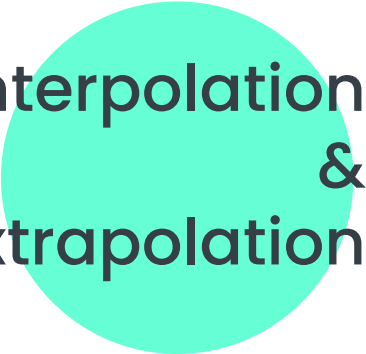
Interpolation & Extrapolation

A large teal circle is positioned on the right side of the image, partially overlapping the text. The circle is a solid teal color and has a diameter that is approximately one-third of the image width. It overlaps the right side of the word 'Interpolation' and the word '&', and it also overlaps the right side of the word 'Extrapolation'.

The data points on our scatter plot with the lowest and highest values are called **poles**.

Interpolation is the act of predicting a y-value for an x-value **between** the **poles**.

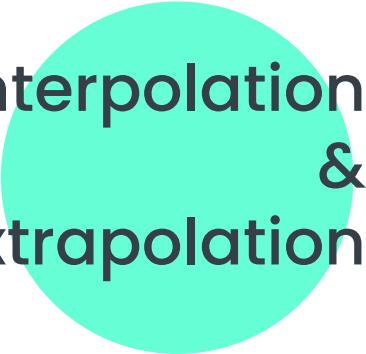
Extrapolation is the act of predicting a y-value for an x-value **outside** the **poles**.



Interpolation
&
Extrapolation

The accuracy of **interpolation** depends on the accuracy of the line of best fit, which can be determined using the correlation coefficient.

The accuracy of **extrapolation** depends not only on the accuracy of the line of best fit, but also on the assumption that the linear trend continues past the **poles**.



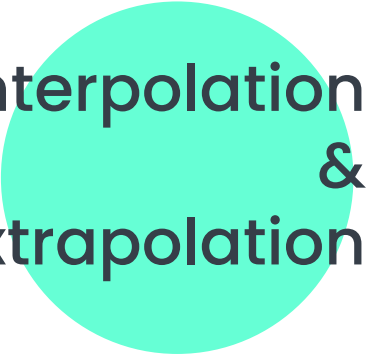
Interpolation
&
Extrapolation

Consider the Example 1.

The line of best fit modelled the age of a player and the distance they can throw a discus.

Use your line of best fit to predict how far a player aged 14 and aged 50 can throw the discus.

$$y_1 = 3.29x - 20.3$$



Interpolation
&
Extrapolation

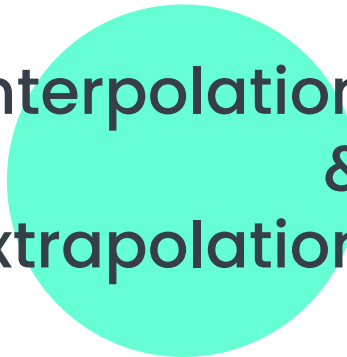
Consider the Example 1.

The line of best fit modelled the age of a player and the distance they can throw a discus.

Use your line of best fit to predict how far a player aged 14 and aged 50 can throw the discus.

$$y_1 = 3.29x - 20.3$$

(14, 25.71)
(50, 144.13)



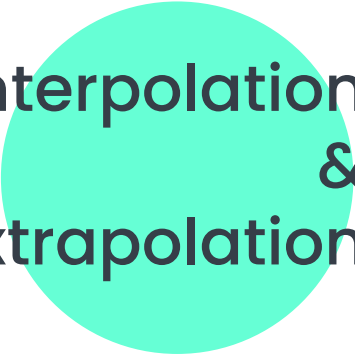
Interpolation
&
Extrapolation

Consider the Example 1.

(14, 25.71) Since 14 is within the **poles**, it is reasonable to assume a 14 year old could throw a discus 25.71 m.

(50, 144.13) Since 50 is without the **poles**, it may not be possible for a 50 year old to throw a discus 144.13 m.

Note: The current world record for discus throwing is 76.8 m.



Interpolation
&
Extrapolation