1. **Analyzing categorical data**

* Mean
* Median
* Mode
* Range
* Mid-range
* Two way freaquency table
* Venn diagram
* Joint probability
* Marginal probability
* Conditional probability
* Independence

Link : <https://www.youtube.com/watch?v=SrEmzdOT65s&t=1s>

1. **Displaying and comparing quatitative data**

* Frequency table and dot plots
* Histograms
* Stem and leaf plots
* Shape of distributions
* Right skewed (Median to right side)
* Left skewed
* Symmetrical
* Cluster, gaps, peaks and outliers



* Line graphs

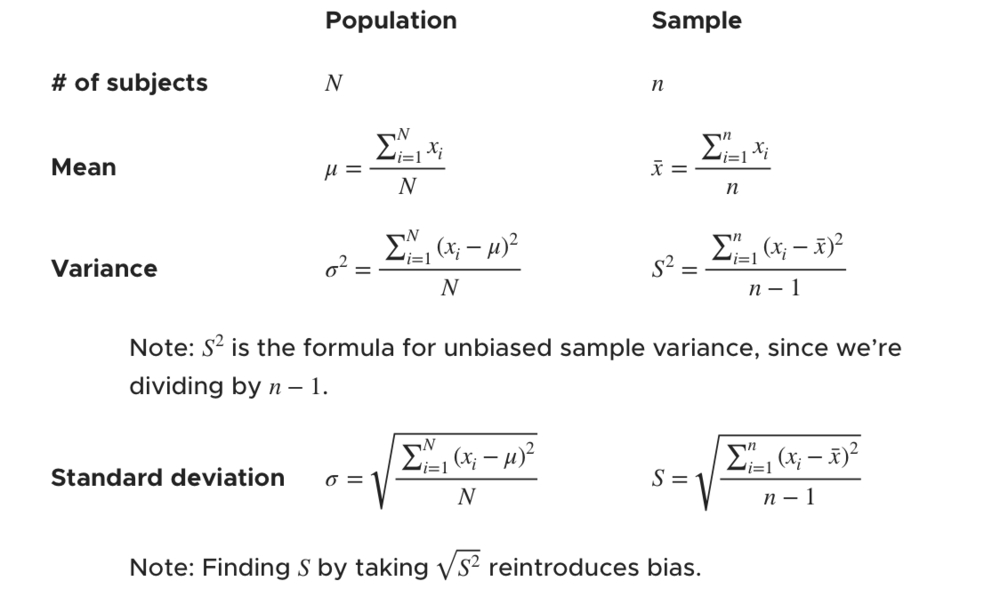
1. **Summarizing quantitative data**

* Mean , median, mode
* Impact of removing outliers on mean and median
* Finding missing value using mean
* Median & range assumptions

1. Iqr

Meassure of spred

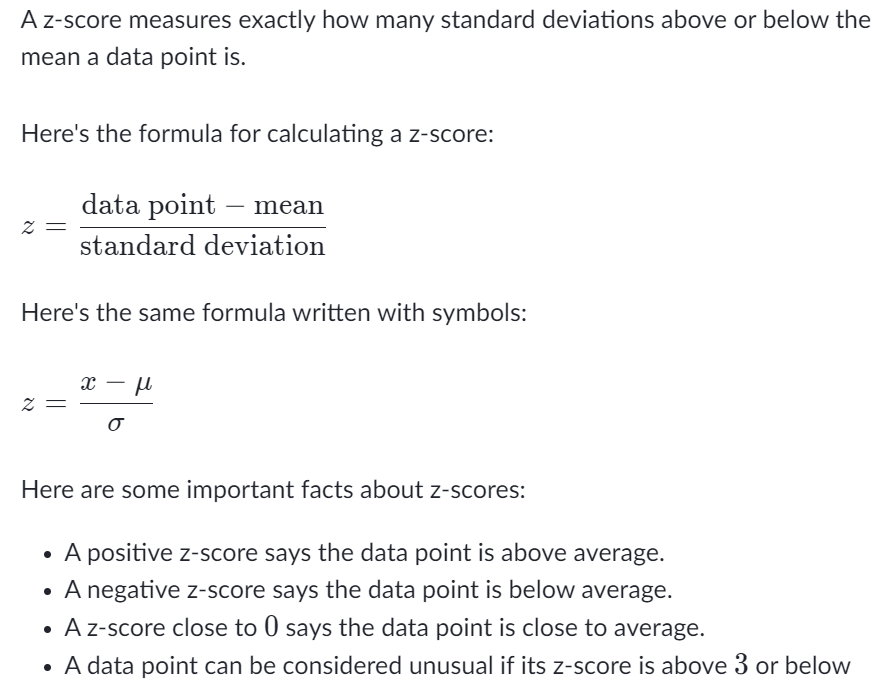
* Range
* Mid - range
* Variance
* Standard deviation



|  |
| --- |
|  |
| * Boxplot * Removing ouliers * Low = q1 – iqr \* 1.5 * High = q3 + iqr \* 1.5   Outlier = data < low OR data > High |

* Mean Absaute deviation (MAD)

1. **Modeling Data Distribution**
2. Percentile
3. **Z scores**



1. Effect of skewed data on parameter

Mean and median changes with both shifting and scaling(shifting - adding same number into all )

Std, iqr and range changes with scaling(multiplying with same number)

1. Area of density curve

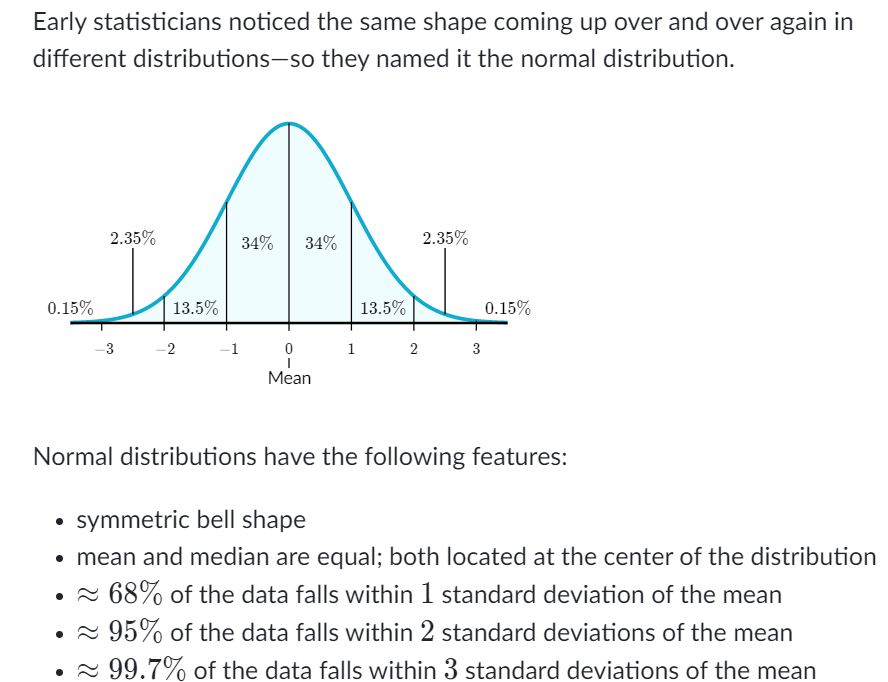
Area of trapizoid = ½ x (sum of the lengths of the parallel sides) x perpendicular distance between parallel sides

Area of squre

Area of triangle

1. Empirical rule and normal distribution

* 68–95–99.7 rule



1. Normal distribution calculations

* Using z score
* For percentages (find z score using percentage from table and put into z formula)

1. **Exploring Bi variate numerical data**
2. Scatter plot

* Lineariy, strength and dierection
* Positive and negative association
* Outliers and clusters

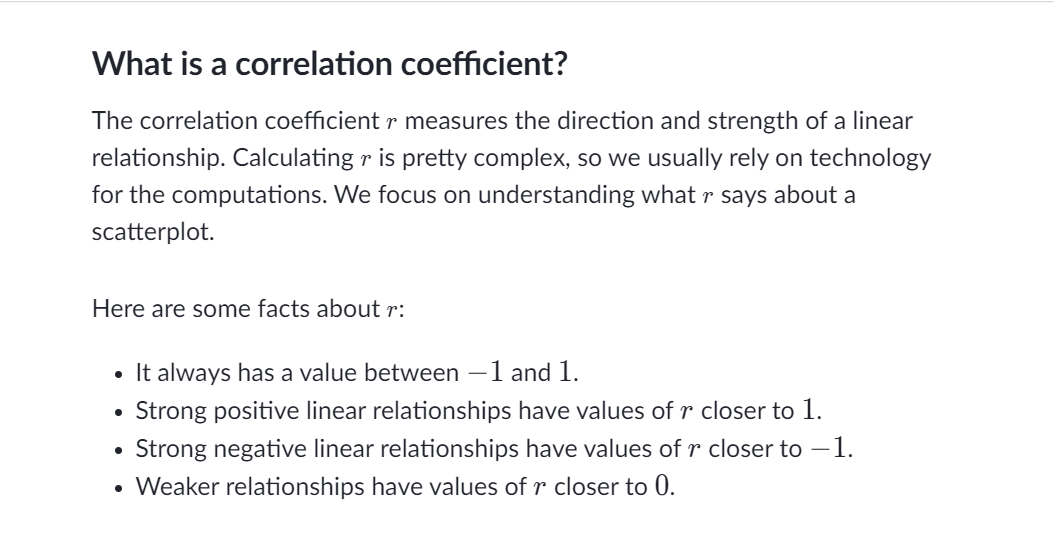
1. Trend lines

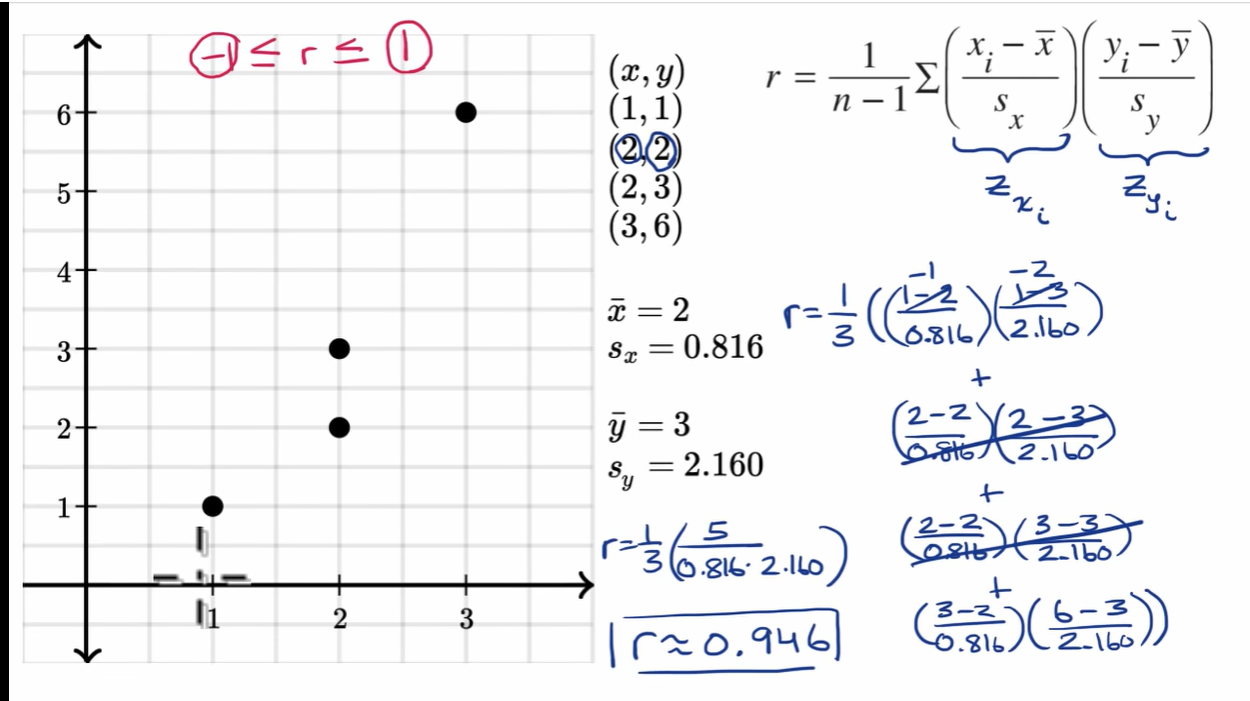
* Linear regression



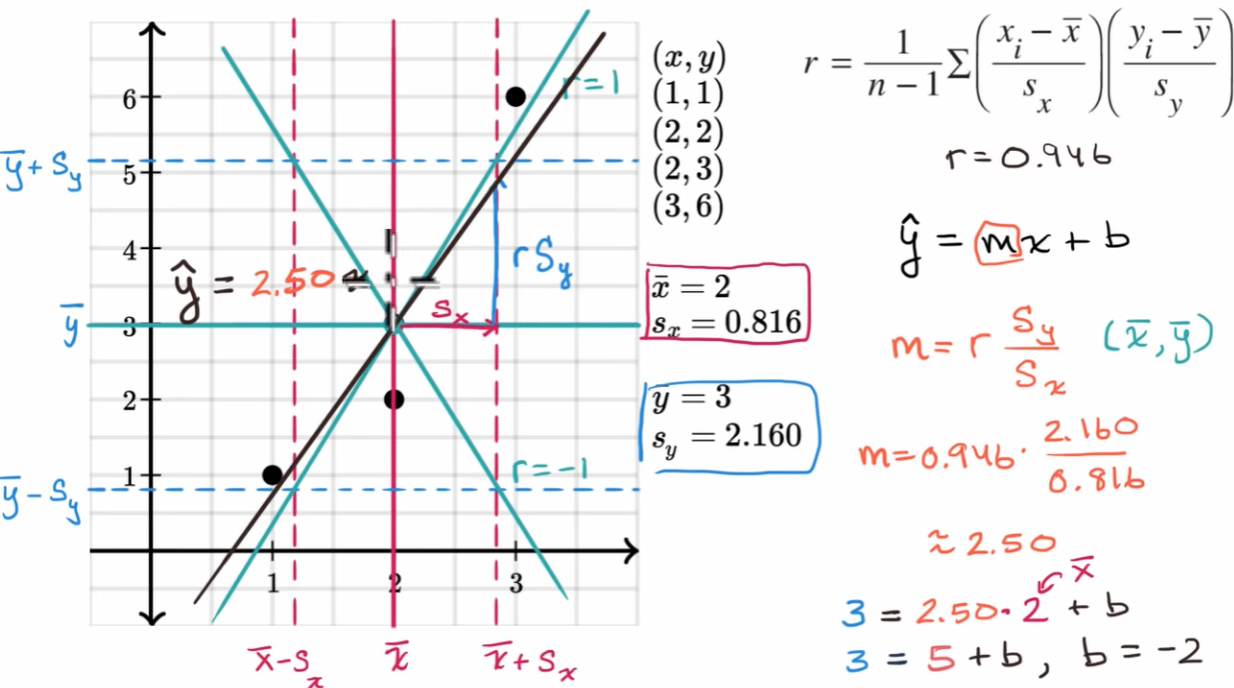
M = slop

b = y intercept

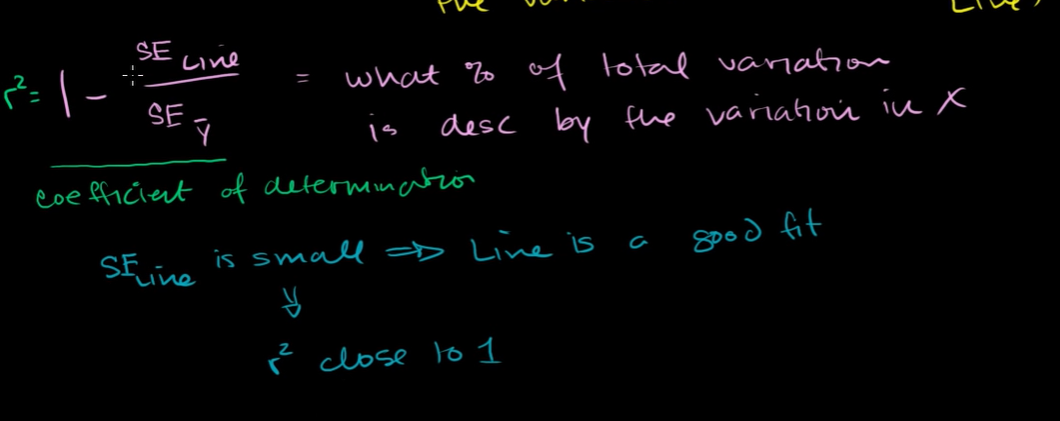
1. **Correlation coefficients ( r )** 



1. Residuals = actual – predicted y

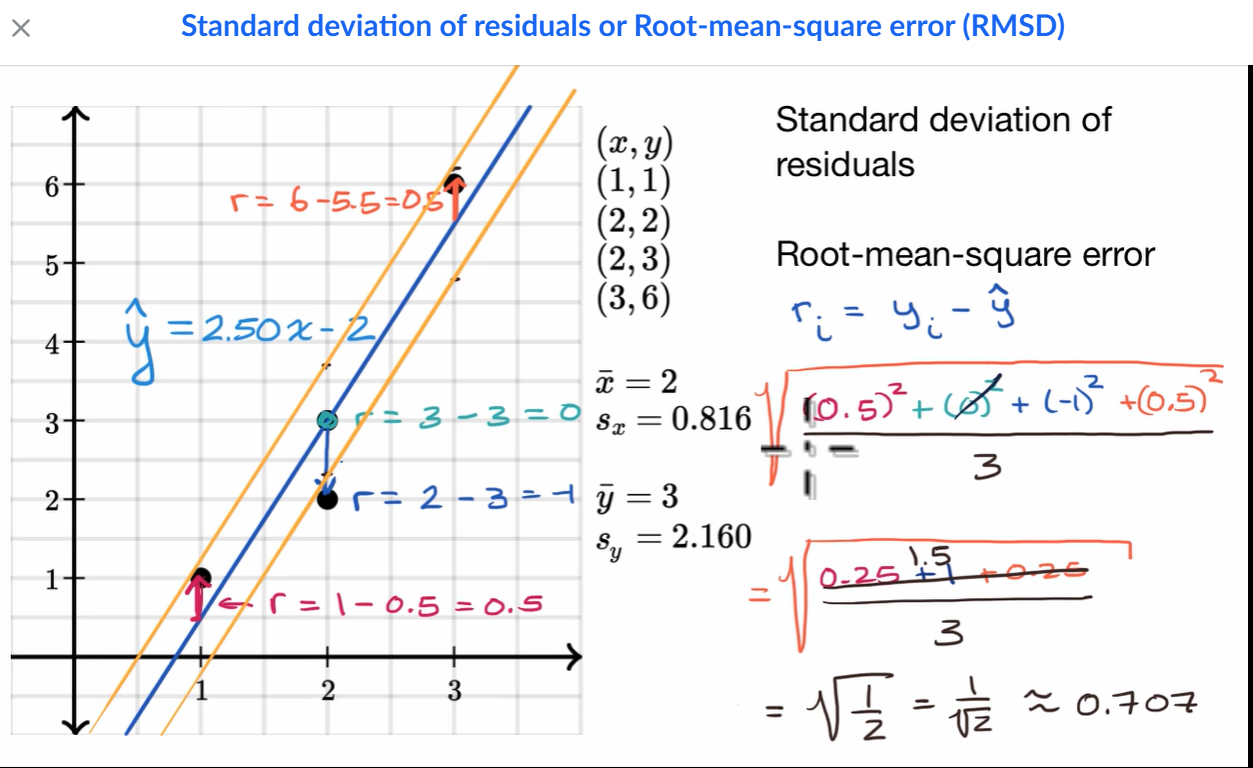
* Equation for regression line

1. Coefficient of determination (r^2)

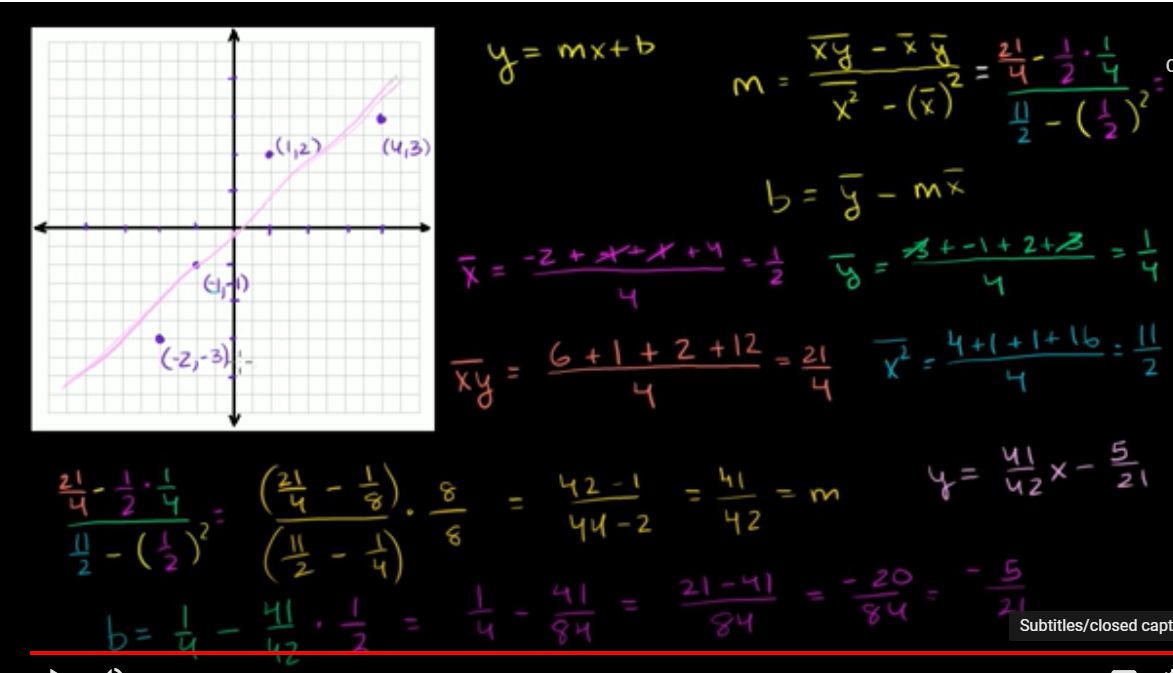


SEline = Squared error of regression line

1. Rmse (less rmse will give a better model)



* Equation for regression line
* Calculate r squared for regression line

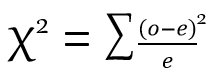


* Covariance : covariance is a measure of the joint variability of two random variables

**Chi-squared test**

* Condition for chi squre test

1. Sample must be random
2. Expected values must be >= 5
3. Variables under study must be categorical
4. Sample not more than 10 % of the population
5. Contigency table chi-square test



O = observed value

e = expected

degree of freedom = (c-1)(r-1)

* n = number of coloms
* r = number of rows
* Assume a significance level alpha , and using the alpha and degrees of freedom find critical chi square value from chi squre distribution table
* If chi- square found is less than critical chi – square value there is a significance for null hypothesis else reject the null hypothesis

Example video link: <https://www.youtube.com/watch?v=hpWdDmgsIRE&t=44s>

1. Goodness of fit