**Statistics**

**Measures of Central Tendency:**

* 1. **Mean (Average):** Also called the arithmetic mean or average

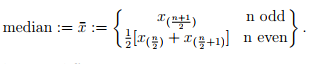
The sum of all the values in the sample divided by the number of values in the sample/population

μ is the mean of the whole population;

͞x is the mean of the sample taken from the population



* 1. **Median**: The middle value when the data are ordered, so that 50% of the data are above and 50% are below



* 1. **Mode**: the most frequently occurring value

**Measures of Dispersion:**

1. **Range**: The minimum and maximum values

xmax − xmin measures dispersion

1. **Variance**: Measures dispersion around the mean

Determined by averaging the squared differences of all the values from the mean

Variance of a population is, σ² =

Variance of a sample is s2(note the n-1), s2 =

1. **Standard Deviation**: square root of the variance

Also measures dispersion around the mean but in the same units as the values (instead of square units with variance)

‘σ’ is the standard deviation of the population;

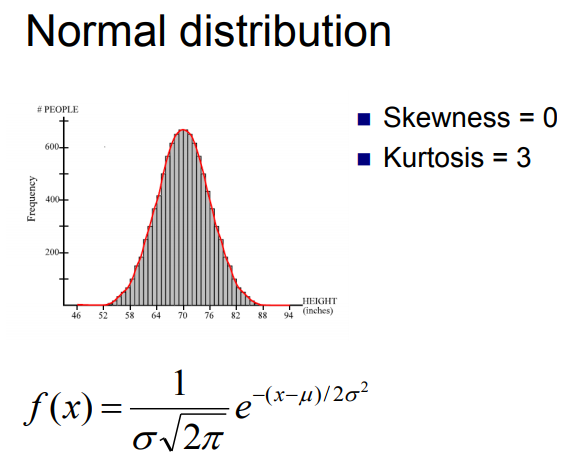
‘s’ is the standard deviation of the sample

1. **Skewness:** Skewness is defined such that if more datapoints lie below the mean of the dataset it is known as negatively skewed. If the distribution is depleted of values below the mean it is positively skewed.

**Nature of Skewness**  
Skewness can be positive or negative or zero.

1. When the values of mean, median and mode are equal, there is no skewness.
2. When mean > median > mode, skewness will be positive.
3. When mean < median < mode, skewness will be negative.
4. **Kurtosis**: The second deviation from symmetry is known as kurtosis and compares the population of the tails of the dataset to that of the central region. If a distribution is more peaked than a Gaussian and/or the tails are more populated than a Gaussian then it has a positive kurtosis. If a distribution has tails less populated and/or is less peaked than a Gaussian then it has a negative kurtosis.

****



**Measures of Association:**

1. **Covariance:** Degree with which y depends on x. Covariance is a measure of how much two random variables vary together.

Suppose X and Y are random variables with means µX and µY. The covariance of X and Y is defined as

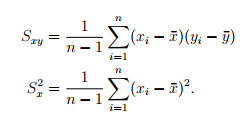
Cov(X, Y ) = E((X − µX)(Y − µY ))

1. **Correlation Coefficient:** This is also called Pearson Correlation Coefficient

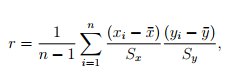
Measures the strength of a linear relationship between two variables

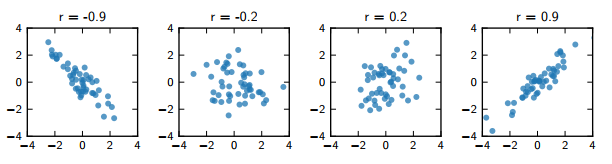
Sample correlation coefficient, r =

where



We can rewrite r as





* r ∈ [−1, 1] and is ±1 only when data fall along a straight line
* sign(r) indicates the slope of the line
* If r is -1, means oppositely related
* If r is 0, means there is no dependency

1. **Linear relationship:** **Regression** is a set of technique for estimating relationships. One of the simplest type of relationship is **linear regression**.

Let’s take equation of the line, **y = mx + c**

x – independent variable or predictor variable

y – dependent variable or response variable

m – slope of the line

If ‘m’ is close to 0 indicates little to no relationship

Value of m with large positive or negative values indicate large positive or negative relationships respectively

c – constant, intercept of the line

**Problem:** **Find whether the Stock x and y are linearly correlated?**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Stock (x)** | **Stock (y)** | **x -͞x** | **(x -͞x)2** | **y-͞y** | **(y-͞y)2** | **(x -͞x)( y-͞y)** |
| 2001 | 5 | -2 | 3 | 9 | -5 | 25 | -15 |
| 2002 | 3 | 1 | 1 | 1 | -2 | 4 | -2 |
| 2003 | -2 | 6 | -4 | 16 | 3 | 9 | -12 |
| 2004 | 2 | 7 | 0 | 0 | 4 | 16 | 0 |
| **Sum** | **8** | **12** |  | **26** |  | **54** | **-29** |
|  | ͞x =2 | ͞y=3 |  | = = 8.67 |  | = = 18 | = = -9.66 |
|  |  |  |  | =2.94 |  | =4.24 |  |

= = = -0.77

< = = 0.98 i.e. <0.98

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SUMMARY OUTPUT** | | |  |  |  |  |  |  |
| ***Regression Statistics*** |  |  |  |  |  |  |  |  |
| Multiple R | **0.773952743** | **Pearson Coefficient, r = 0.773952743** |  |  |  |  |  |  |
| R Square | 0.599002849 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.398504274 |  |  |  |  |  |  |  |
| Standard Error | 3.290429011 |  |  |  |  |  |  |  |
| Observations | 4 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 32.34615385 | 32.34615385 | 2.987566607 | 0.226047257 |  |  |  |
| Residual | 2 | 21.65384615 | 10.82692308 |  |  |  |  |  |
| Total | 3 | 54 |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 5.230769231 | 2.091031143 | 2.501526219 | 0.129482995 | -3.766211626 | 14.22775009 | -3.766211626 | 14.22775009 |
| Stock (x) | -1.115384615 | 0.645306221 | -1.72845787 | 0.226047257 | -3.891913187 | 1.661143956 | -3.891913187 | 1.661143956 |

**Problem:** Annual revenues and profits of an IT company for the years 2004 to 2014 are given. Find the correlation coefficient.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Revenue (x)** | **Profit (y)** | **x -͞x** | **(x -͞x)2** | **y-͞y** | **(y-͞y)2** | **(x -͞x)( y-͞y)** | **x²** | **xy** |
| 2004 | 225 | 42 | -39.0909091 | 1528.09917 | -13 | 169 | 508.1818182 | 50625 | 9450 |
| 2005 | 237 | 43 | -27.0909091 | 733.917355 | -12 | 144 | 325.0909091 | 56169 | 10191 |
| 2006 | 245 | 48 | -19.0909091 | 364.46281 | -7 | 49 | 133.6363636 | 60025 | 11760 |
| 2007 | 222 | 40 | -42.0909091 | 1771.64463 | -15 | 225 | 631.3636364 | 49284 | 8880 |
| 2008 | 265 | 60 | 0.909090909 | 0.82644628 | 5 | 25 | 4.545454545 | 70225 | 15900 |
| 2009 | 270 | 56 | 5.909090909 | 34.9173554 | 1 | 1 | 5.909090909 | 72900 | 15120 |
| 2010 | 254 | 53 | -10.0909091 | 101.826446 | -2 | 4 | 20.18181818 | 64516 | 13462 |
| 2011 | 280 | 60 | 15.90909091 | 253.099174 | 5 | 25 | 79.54545455 | 78400 | 16800 |
| 2012 | 290 | 62 | 25.90909091 | 671.280992 | 7 | 49 | 181.3636364 | 84100 | 17980 |
| 2013 | 305 | 65 | 40.90909091 | 1673.55372 | 10 | 100 | 409.0909091 | 93025 | 19825 |
| 2014 | 312 | 76 | 47.90909091 | 2295.28099 | 21 | 441 | 1006.090909 | 97344 | 23712 |
| **Sum** | 2905 | 605 |  | 9428.90909 |  | 1232 | 3305 | 776613 | 163080 |
|  | 264.0909091 | 55 |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |  |  |  |  |
| Multiple R | 0.969695512 | **Pearson Coefficient, r = 0.969695512159858** |  |  |  |  |  |  |
| R Square | 0.940309386 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.933677096 |  |  |  |  |  |  |  |
| Standard Error | 2.858492922 |  |  |  |  |  |  |  |
| Observations | 11 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 1158.461 | 1158.461164 | 141.777475 | 8.23E-07 |  |  |  |
| Residual | 9 | 73.53884 | 8.170981786 |  |  |  |  |  |
| Total | 10 | 1232 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -37.56855126 | 7.821902 | -4.80299424 | 0.00096967 | -55.2629 | -19.8742 | -55.26292339 | -19.8742 |
| Revenue (x) | 0.35051775 | 0.029438 | 11.90703468 | 8.227E-07 | 0.283925 | 0.417111 | 0.283924659 | 0.417111 |

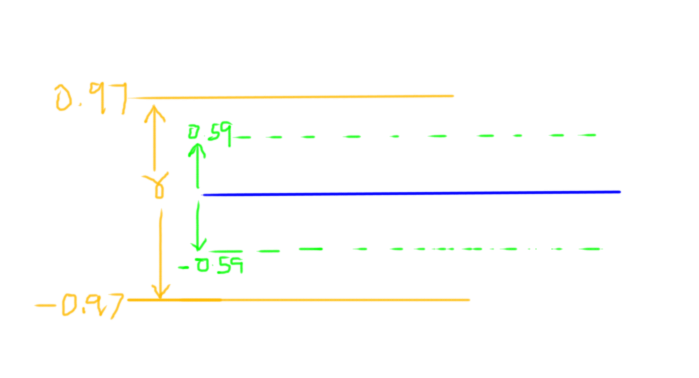
**Sx** = = = = 30.69

S**y** = = = 11.09

S**xy** = = = 330.5

r**xy** = = = 0.971

Threshold = = = = 0.5921



**Problem:** In the above problem, if the revenue for the year 2015 is $350 million. Forecast the profit for 2015 in terms of millions of dollars.

776613 m + 2905 c = 163080

2905 m + 11 c = 605

m = 0.35

c = -37.57

If x = 350, then y = $ 84.9 million

**Problem:** In the above problem, find the profit for the year 2004

Y = mx + c

= 0.35 \* 225 – 37.57

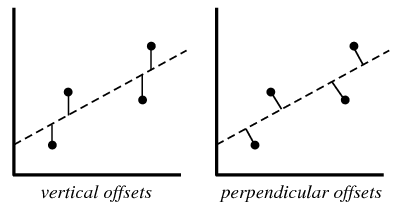
= 78.75 – 37.57

= 41.18

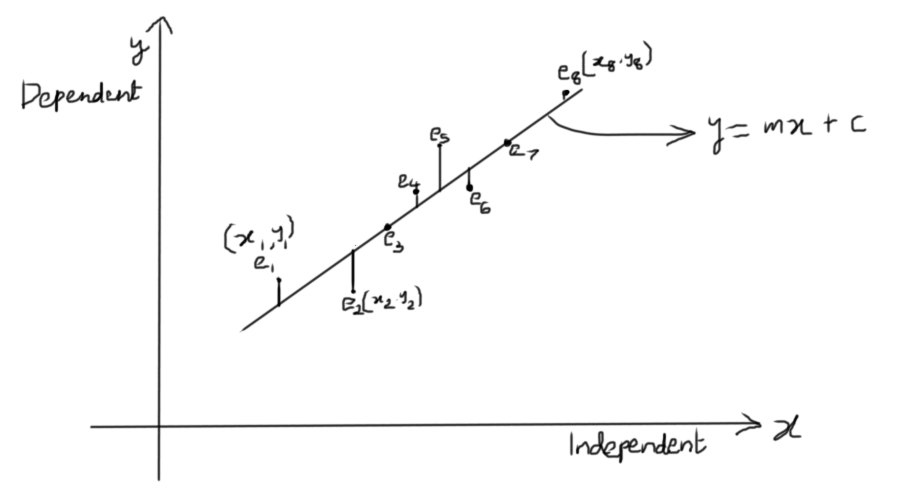
**Steps to follow in solving a problem:**

1. Scatter Plot
2. Identify Trend and Outliers
3. Transform data (if required)
4. Validate the Trend
5. Fit a curve
6. Forecast
7. Verify the forecasted value

**Least square technique: -**

The least squares method is a form of mathematical regression analysis that finds the *line of best fit* for a dataset, providing a visual demonstration of the relationship between the data points. Each point of data is representative of the relationship between a known independent variable and an unknown dependent variable.  
  
In regression analysis, dependent variables are designated on the vertical Y axis and independent variables are designated on the horizontal X axis. These designations will form the equation for the line of best fit, which is determined from the least squares method.  


In practice, the ***vertical offsets*** from a line (polynomial, surface, hyperplane, etc.) are almost always minimized instead of the ***perpendicular offsets***. In addition, the fitting technique can be easily generalized from a best-fit *line* to a best-fit *polynomial* when sums of vertical distances are used. In any case, for a reasonable number of noisy data points, the difference between vertical and perpendicular fits is quite small.



The method of least squares gives a way to find the best estimate, assuming that the errors (i.e. the differences from the true value) are random and unbiased.

Let’s us assume the equation of the line as

y = mx + c

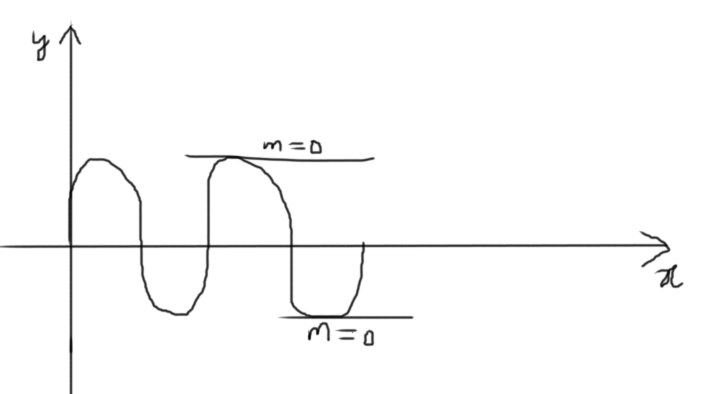
Error,

We need to find the sum of squares of the deviation is minimal.

i.e., should be minimum

which means should be minimum

is minimum

At maximum point or minimum point of wave curve, the slope should be zero

= 0

= 2 = 0

= 2 = 0

= - = 0

- = 0

m ∑ + c =

m ∑ + c. n = ∑

We have 2 equations and 2 unknowns, so we would be able to solve.