VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING

ITA5007-Data Mining and Business Intelligence

MCA

Topic :

1. Covariance calculation among two variables
2. Correlation coefficient calculation
3. Variance calculation

Description:

In mathematics and statistics, covariance is a measure of the relationship between two random variables.
The metric evaluates how much – to what extent – the variables change together.
In other words, it is essentially a measure of the variance between two variables.
However, the metric does not assess the dependency between variables.

Unlike the correlation coefficient, covariance is measured in units. The units are computed by multiplying the units of the two variables.

The variance can take any positive or negative values. The values are interpreted as follows:

Positive covariance: Indicates that two variables tend to move in the same direction.
Negative covariance: Reveals that two variables tend to move in inverse directions.

Formula for Covariance

The covariance formula is similar to the formula for correlation and deals with the calculation of data points from the average value in a dataset. For example, the covariance between two random variables X and Y can be calculated using the following formula (for population):

$$\mathrm{Cov}\,(X, Y) = \frac{\sum (X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

For a sample covariance, the formula is slightly adjusted:

$$\mathrm{Cov}\,(X, Y) = \frac{\sum (X_i - \bar{X})(Y_j - \bar{Y})}{n\ - 1}$$

Where:

Xi – the values of the X-variable
Yj – the values of the Y-variable
X̄ – the mean (average) of the X-variable
Ȳ – the mean (average) of the Y-variable
n – the number of data points
Covariance vs. Correlation
Covariance and correlation both primarily assess the relationship between variables. The closest analogy to the relationship between them is the relationship between the variance and standard deviation.

Covariance measures the total variation of two random variables from their expected values. Using covariance, we can only gauge (measure) the direction of the relationship (whether the variables tend to move in tandem (in the same direction) or show an inverse relationship). However, it does not indicate the strength of the relationship, nor the dependency between the variables.

On the other hand, correlation measures the strength of the relationship between variables. Correlation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

The relationship between the two concepts can be expressed using the formula below:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where:

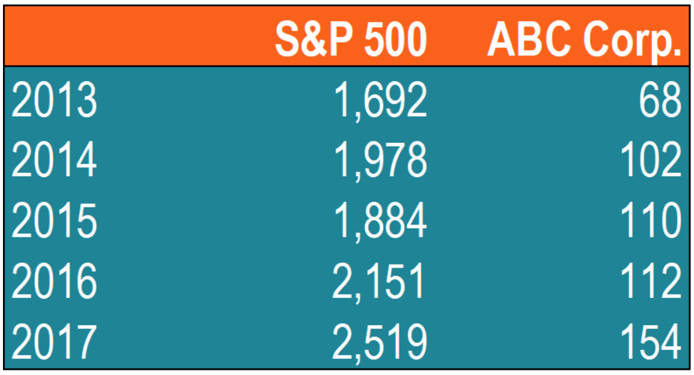
ρ(X,Y) – the correlation between the variables X and Y
Cov(X,Y) – the covariance between the variables X and Y
σX – the standard deviation of the X-variable
σY – the standard deviation of the Y-variable

Example of Covariance
John is an investor. His portfolio primarily tracks the performance of the S&P 500 and John wants to add the stock of ABC Corp. Before adding the stock to his portfolio, he wants to assess the directional relationship between the stock and the S&P 500.

John does not want to increase the unsystematic risk of his portfolio. Thus, he is not interested in owning securities in the portfolio that tend to move in the same direction.

John can calculate the covariance between the stock of ABC Corp. and S&P 500 by following the steps below:
1. Obtain the data.

First, John obtains the figures for both ABC Corp. stock and the S&P 500. The prices obtained are summarized in the table below:



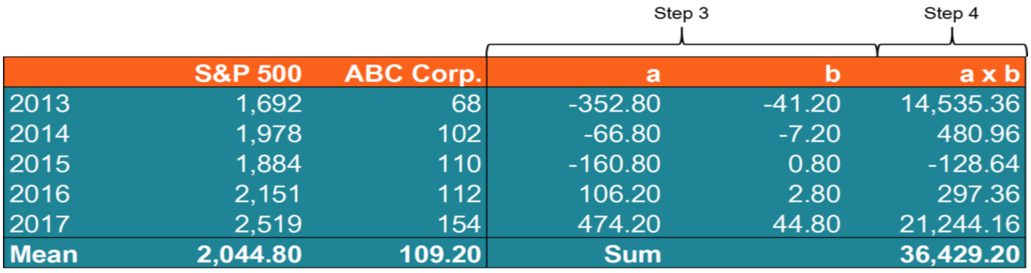
	S&P 500	ABC Corp.
2013	1,692	68
2014	1,978	102
2015	1,884	110
2016	2,151	112
2017	2,519	154

2. Calculate the mean (average) prices for each asset.

$$\text{Mean (S\&P 500)} = \frac{1{,}692 + 1{,}978 + 1{,}884 + 2{,}151 + 2{,}519}{5} = 2{,}044.80$$

$$\text{Mean (ABC Corp.)} = \frac{68 + 102 + 110 + 112 + 154}{5} = 109.20$$

3. For each security, find the difference between each value and mean price.



Step 3 Step 4

	S&P 500	ABC Corp.	a	b	a x b
2013	1,692	68	-352.80	-41.20	14,535.36
2014	1,978	102	-66.80	-7.20	480.96
2015	1,884	110	-160.80	0.80	-128.64
2016	2,151	112	106.20	2.80	297.36
2017	2,519	154	474.20	44.80	21,244.16
Mean	2,044.80	109.20	Sum		36,429.20

4. Multiply the results obtained in the previous step.

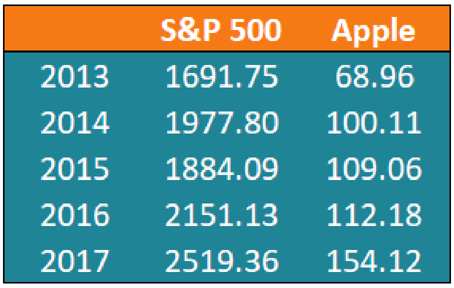
5. Using the number calculated in step 4, find the covariance.

$$\text{Cov(S\&P 500, ABC Corp.)} = \frac{36{,}429.20}{5 - 1} = 9{,}107.30$$

In such a case, the positive covariance indicates that the price of the stock and the S&P 500 tend to move in the same direction.

Correlation: Example of Correlation

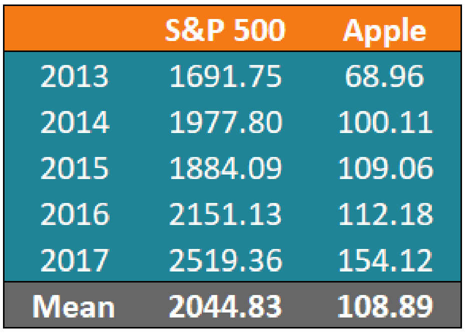
John is an investor. His portfolio primarily tracks the performance of the S&P 500 and John wants to add the stock of Apple Inc. Before adding Apple to his portfolio, he wants to assess the correlation between the stock and the S&P 500 to ensure that adding the stock won’t increase the systematic risk of his portfolio. To find the coefficient, John gathers the following prices for the last five years (Step 1):



	S&P 500	Apple
2013	1691.75	68.96
2014	1977.80	100.11
2015	1884.09	109.06
2016	2151.13	112.18
2017	2519.36	154.12

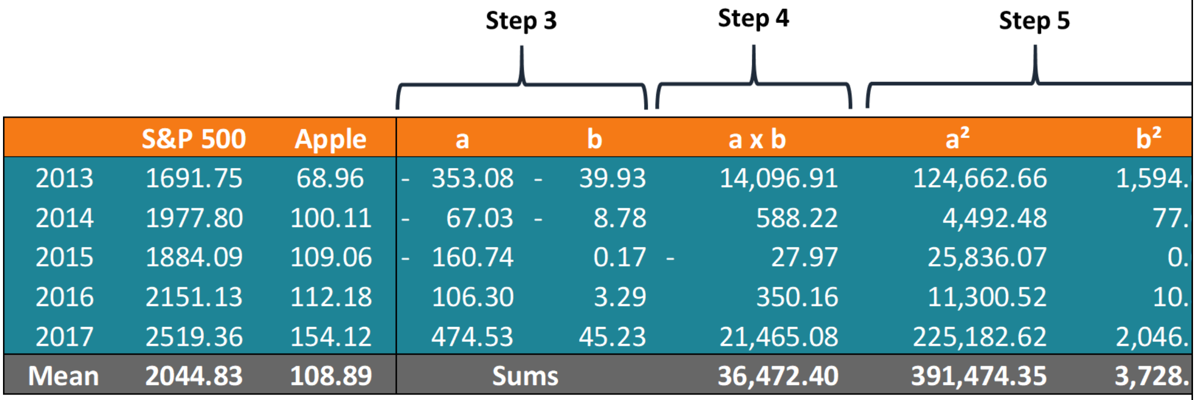
John can determine the correlation between the prices of the S&P 500 Index and Apple Inc.

First, John calculates the average prices of each security for the given periods (Step 2):



	S&P 500	Apple
2013	1691.75	68.96
2014	1977.80	100.11
2015	1884.09	109.06
2016	2151.13	112.18
2017	2519.36	154.12
Mean	2044.83	108.89

After the calculation of the average prices, we can find the other values. A summary of the calculations is given in the table below:



			Step 3		Step 4	Step 5	
	S&P 500	Apple	a	b	a x b	a²	b²
2013	1691.75	68.96	- 353.08	- 39.93	14,096.91	124,662.66	1,594.
2014	1977.80	100.11	- 67.03	- 8.78	588.22	4,492.48	77.
2015	1884.09	109.06	- 160.74	0.17	- 27.97	25,836.07	0.
2016	2151.13	112.18	106.30	3.29	350.16	11,300.52	10.
2017	2519.36	154.12	474.53	45.23	21,465.08	225,182.62	2,046.
Mean	2044.83	108.89	Sums		36,472.40	391,474.35	3,728.

Using the obtained numbers, John can calculate the coefficient:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

rxy – the correlation coefficient of the linear relationship between the variables x and y
xi – the values of the x-variable in a sample
x̅ – the mean of the values of the x-variable
yi – the values of the y-variable in a sample
ȳ – the mean of the values of the y-variable

$$r_{xy} = \frac{36{,}272.40}{\sqrt{391{,}474.35 \times 3{,}728.10}} = 0.95$$

The coefficient indicates that the prices of the S&P 500 and Apple Inc. have a high positive correlation. This means that their respective prices tend to move in the same direction. Therefore, adding Apple to his portfolio would, in fact, increase the level of systematic risk.

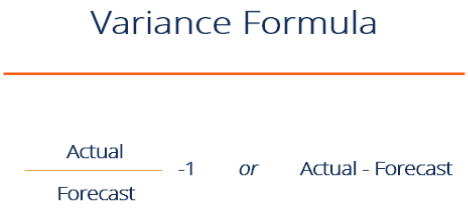
How to Calculate Variance?

The variance formula is used to calculate the difference between a forecast and the actual result. The variance can be expressed as a percentage or as an integer (dollar value or the number of units). Variance analysis and the variance formula play an important role in corporate financial planning and analysis (FP&A) to help evaluate results and make informed decisions for a business going forward.
What is the Variance Formula?
There are two formulas to calculate variance:

Variance % = Actual / Forecast – 1

or

Variance $ = Actual – Forecast



Variance Formula

$$\frac{\text{Actual}}{\text{Forecast}} -1 \quad \textit{or} \quad \text{Actual - Forecast}$$

Percent Variance Formula

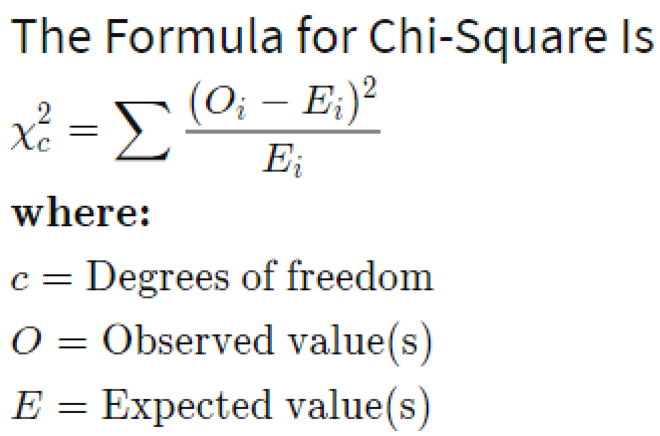
As the name implies, the percent variance formula calculates the percentage difference between a forecast and an actual result.
In the example analysis above we see that the revenue forecast was $150,000 and the actual result was $165,721. Therefore, we take $165,721 divided by $150,000, less one, and express that number as a percentage, which is 10.5%.

This is an example of outperformance, a positive variance, or a favourable variance.

Chi square test:

- A chi-square (χ^2) statistic is a measure of the difference between the observed and expected frequencies of the outcomes of a set of events or variables.
- χ^2 depends on the size of the difference between actual and observed values, the degrees of freedom, and the samples size.
- χ^2 can be used to test whether two variables are related or independent from one another or to test the goodness-of-fit between an observed distribution and a theoretical distribution of frequencies.

The Formula for Chi-Square Is



The Formula for Chi-Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c = \text{Degrees of freedom}$

$O = \text{Observed value(s)}$

$E = \text{Expected value(s)}$

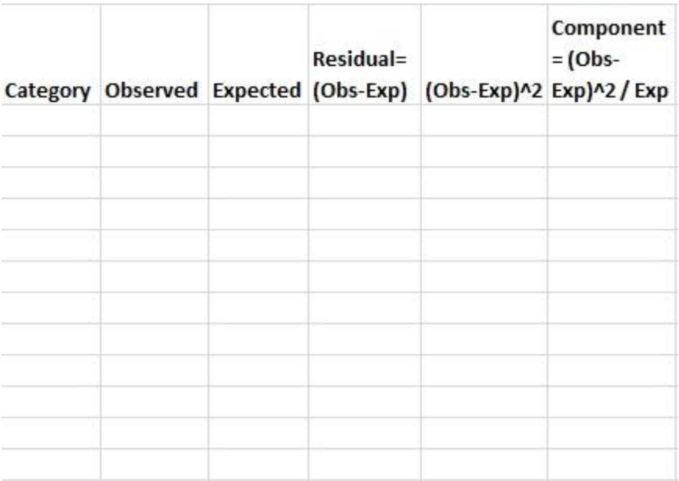
What Does a Chi-Square Statistic Tell You?
There are two main kinds of chi-square tests: the test of independence, which asks a question of relationship, such as, "Is there a relationship between student sex and course choice?"; and the goodness-of-fit test, which asks something like "How well does the coin in my hand match a theoretically fair coin?"

- A **very small chi square test statistic** means that your observed data fits your expected data extremely well. In other words, there is a relationship.
- A **very large chi square test statistic** means that the data does not fit very well. In other words, there isn’t a relationship.

Example question: 256 visual artists were surveyed to find out their zodiac sign. The results were: Aries (29), Taurus (24), Gemini (22), Cancer (19), Leo (21), Virgo (18), Libra (19), Scorpio (20), Sagittarius (23), Capricorn (18), Aquarius (20), Pisces (23). Test the hypothesis that zodiac signs are evenly distributed across visual artists.

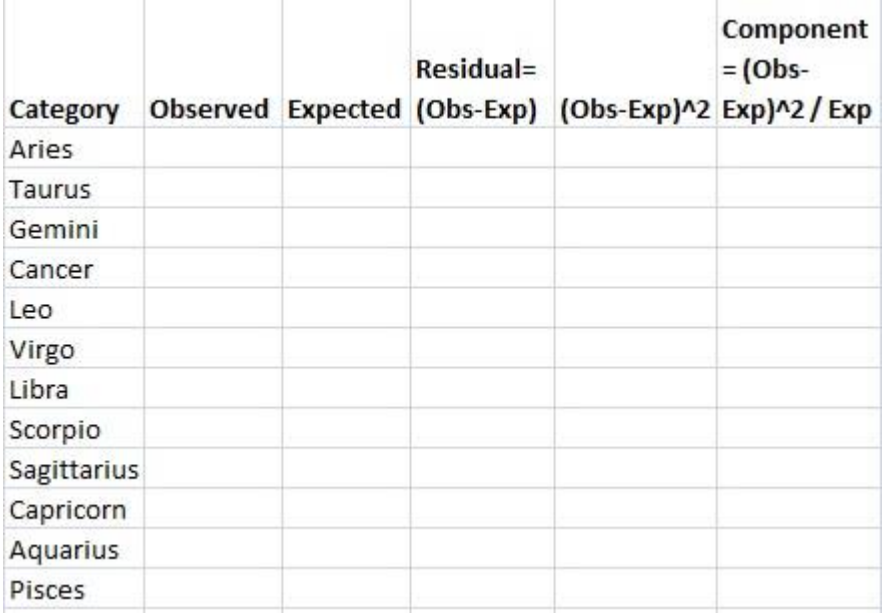
Step 1: Make a table with columns for “Categories,” “Observed,” “Expected,” “Residual (Obs-Exp)”, “(Obs-Exp)2” and “Component

(Obs-Exp)2 / Exp.”



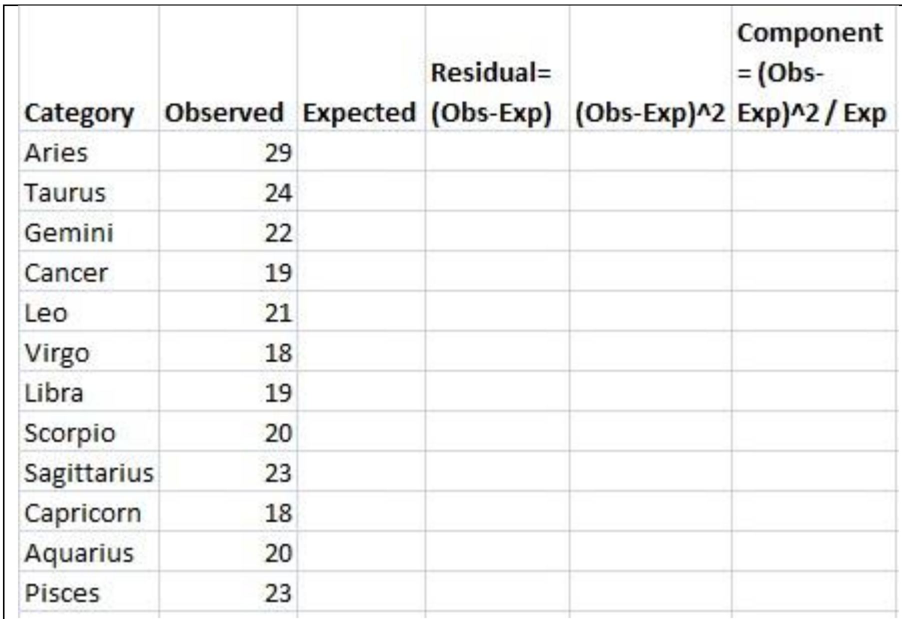
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp

Step 2: Fill in your categories. Categories should be given to you in the question. There are 12 zodiac signs, so:



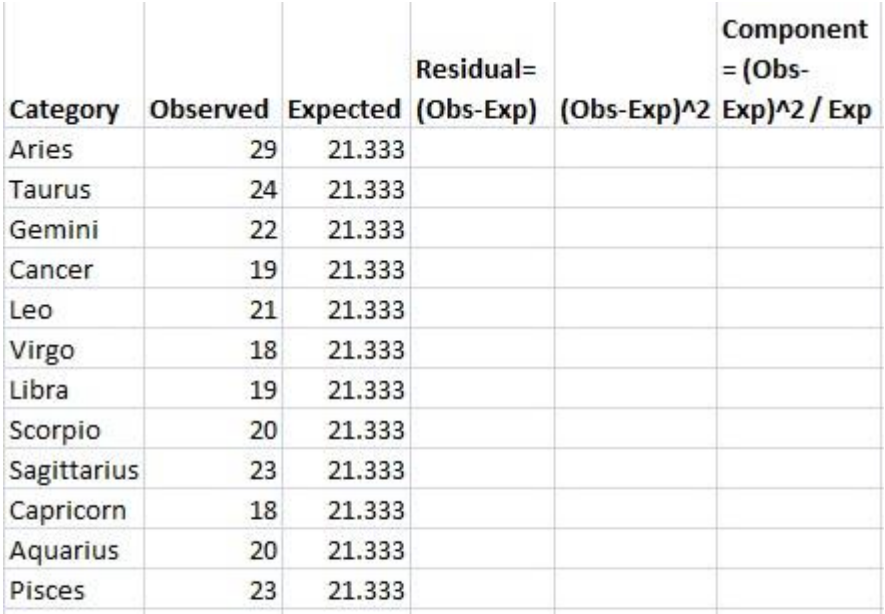
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries					
Taurus					
Gemini					
Cancer					
Leo					
Virgo					
Libra					
Scorpio					
Sagittarius					
Capricorn					
Aquarius					
Pisces					

Step 3: Write your counts. Counts are the number of each items in each category in column 2. You’re given the counts in the question:



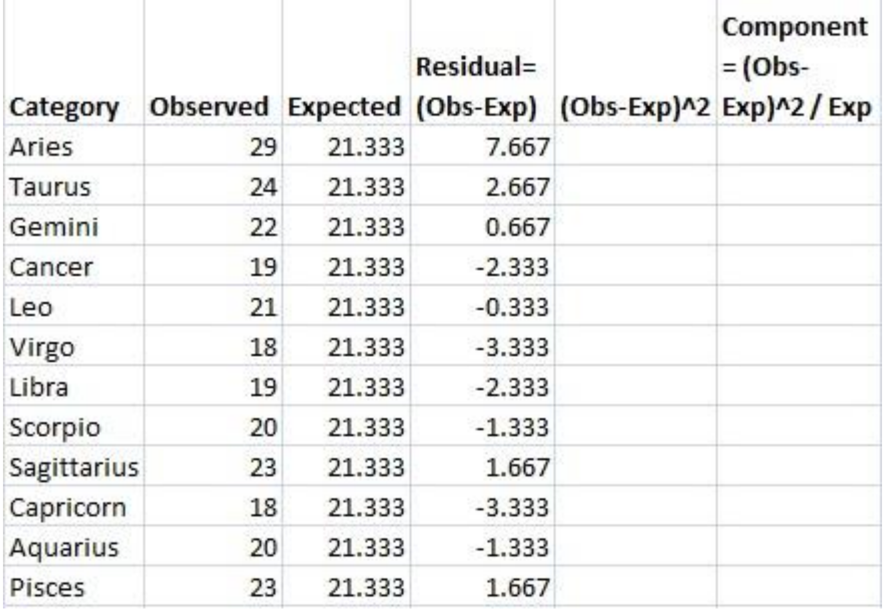
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29				
Taurus	24				
Gemini	22				
Cancer	19				
Leo	21				
Virgo	18				
Libra	19				
Scorpio	20				
Sagittarius	23				
Capricorn	18				
Aquarius	20				
Pisces	23				

Step 4: Calculate your expected value for column 3. In this question, we would expect the 12 zodiac signs to be evenly distributed for all 256 people, so 256/12=21.333. Write this in column 3.



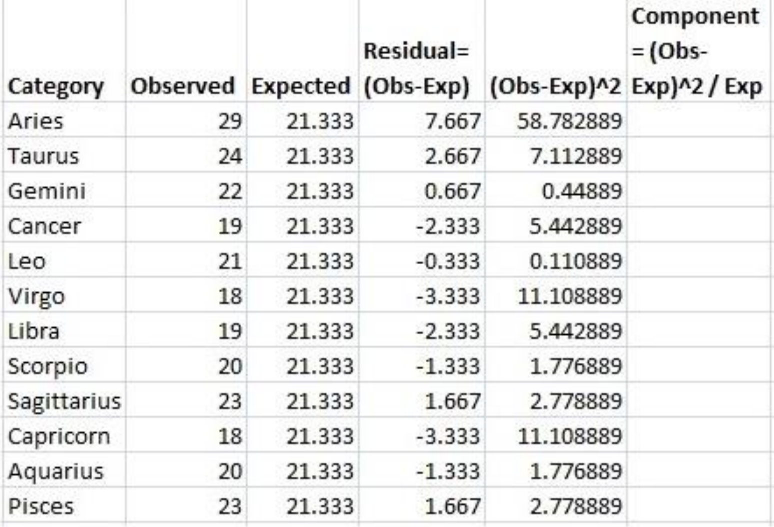
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333			
Taurus	24	21.333			
Gemini	22	21.333			
Cancer	19	21.333			
Leo	21	21.333			
Virgo	18	21.333			
Libra	19	21.333			
Scorpio	20	21.333			
Sagittarius	23	21.333			
Capricorn	18	21.333			
Aquarius	20	21.333			
Pisces	23	21.333			

Step 5: Subtract the expected value (Step 4) from the Observed value (Step 3) and place the result in the “Residual” column. For example, the first row is Aries: 29-21.333=7.667.



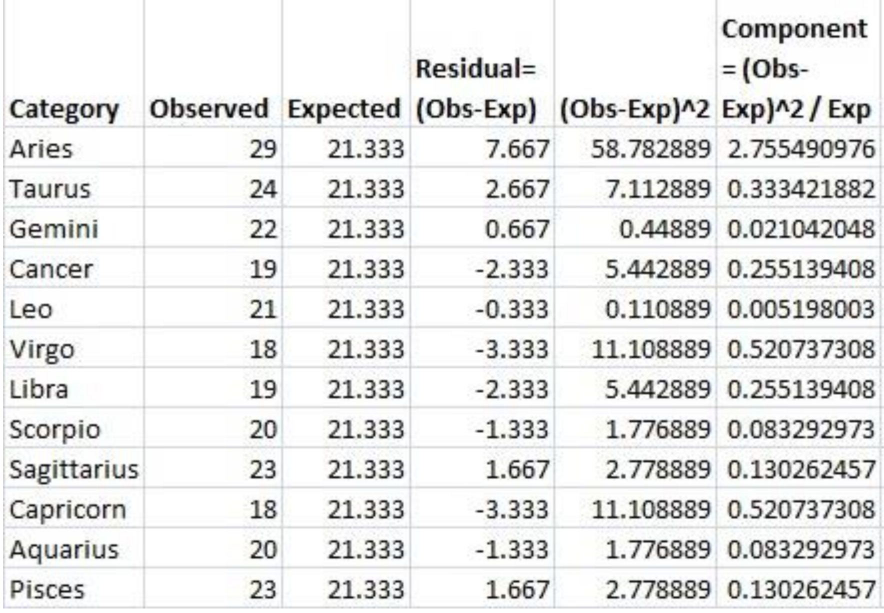
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667		
Taurus	24	21.333	2.667		
Gemini	22	21.333	0.667		
Cancer	19	21.333	-2.333		
Leo	21	21.333	-0.333		
Virgo	18	21.333	-3.333		
Libra	19	21.333	-2.333		
Scorpio	20	21.333	-1.333		
Sagittarius	23	21.333	1.667		
Capricorn	18	21.333	-3.333		
Aquarius	20	21.333	-1.333		
Pisces	23	21.333	1.667		

Step 6: **Square your results from Step 5** and place the amounts in the (Obs-Exp)^2 column.



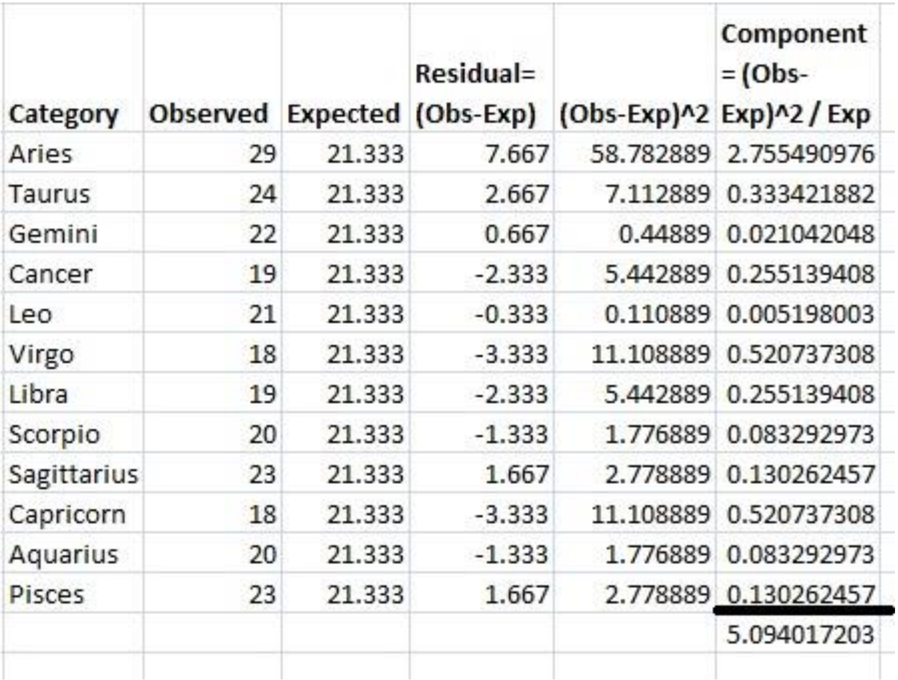
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	
Taurus	24	21.333	2.667	7.112889	
Gemini	22	21.333	0.667	0.44889	
Cancer	19	21.333	-2.333	5.442889	
Leo	21	21.333	-0.333	0.110889	
Virgo	18	21.333	-3.333	11.108889	
Libra	19	21.333	-2.333	5.442889	
Scorpio	20	21.333	-1.333	1.776889	
Sagittarius	23	21.333	1.667	2.778889	
Capricorn	18	21.333	-3.333	11.108889	
Aquarius	20	21.333	-1.333	1.776889	
Pisces	23	21.333	1.667	2.778889	

Step 7: Divide the amounts in Step 6 by the expected value (Step 4) and place those results in the final column.



Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457

Step 8: Add up (sum) all the values in the last column.

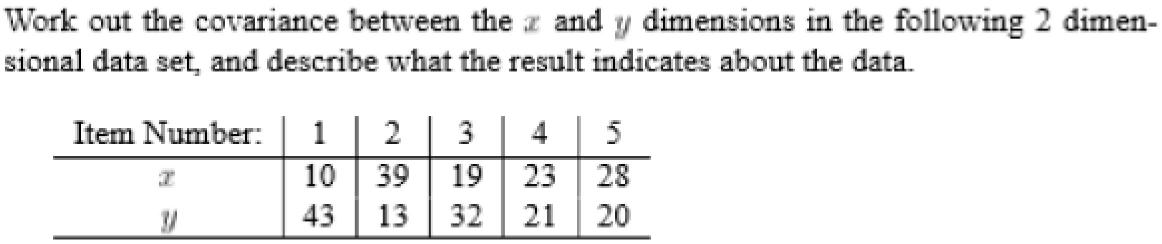


Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457
					5.094017203

This is the chi-square statistic: 5.094.

Questions:

1.



Work out the covariance between the x and y dimensions in the following 2 dimensional data set, and describe what the result indicates about the data.

Item Number:	1	2	3	4	5
x	10	39	19	23	28
y	43	13	32	21	20

Courtesy:
https://corporatefinanceinstitute.com/resources/knowledge/finance/correlation/