Practical 10

Dr. Sudeep Tanwar

What is Karl Pearson's Coefficient of Correlation?

Coefficient of Correlation

- A coefficient of correlation is generally applied in statistics to calculate a relationship between two variables.
- The correlation shows a specific value of the degree of a linear relationship between the X and Y variables.
- There are various types of correlation coefficients. However, Pearson's correlation (also known as Pearson's R) is the correlation coefficient that is frequently used in linear regression.

Types of Correlations

Depending on the direction of the relationship between variables, correlation can be of three types, namely –

1. **Positive Correlation (0 to +1) –** In this case, the direction of change between X and Y is the same. For instance, an increase in the duration of workout leads to an increase in the number of calories one burns.
2. **Negative Correlation (0 to -1) –** Here, the direction of change between X and Y variables is opposite. For example, when the price of a commodity increases its demand decreases.
3. **Zero Correlation (0) –** There is no relationship between the variables in this case. For instance, an increase in height has no impact on one's intelligence.

Pearson’s Coefficient Correlation

- Karl Pearson’s coefficient of correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linearly related variables.
- The coefficient of correlation is expressed by “r”.

Karl Pearson Correlation Coefficient

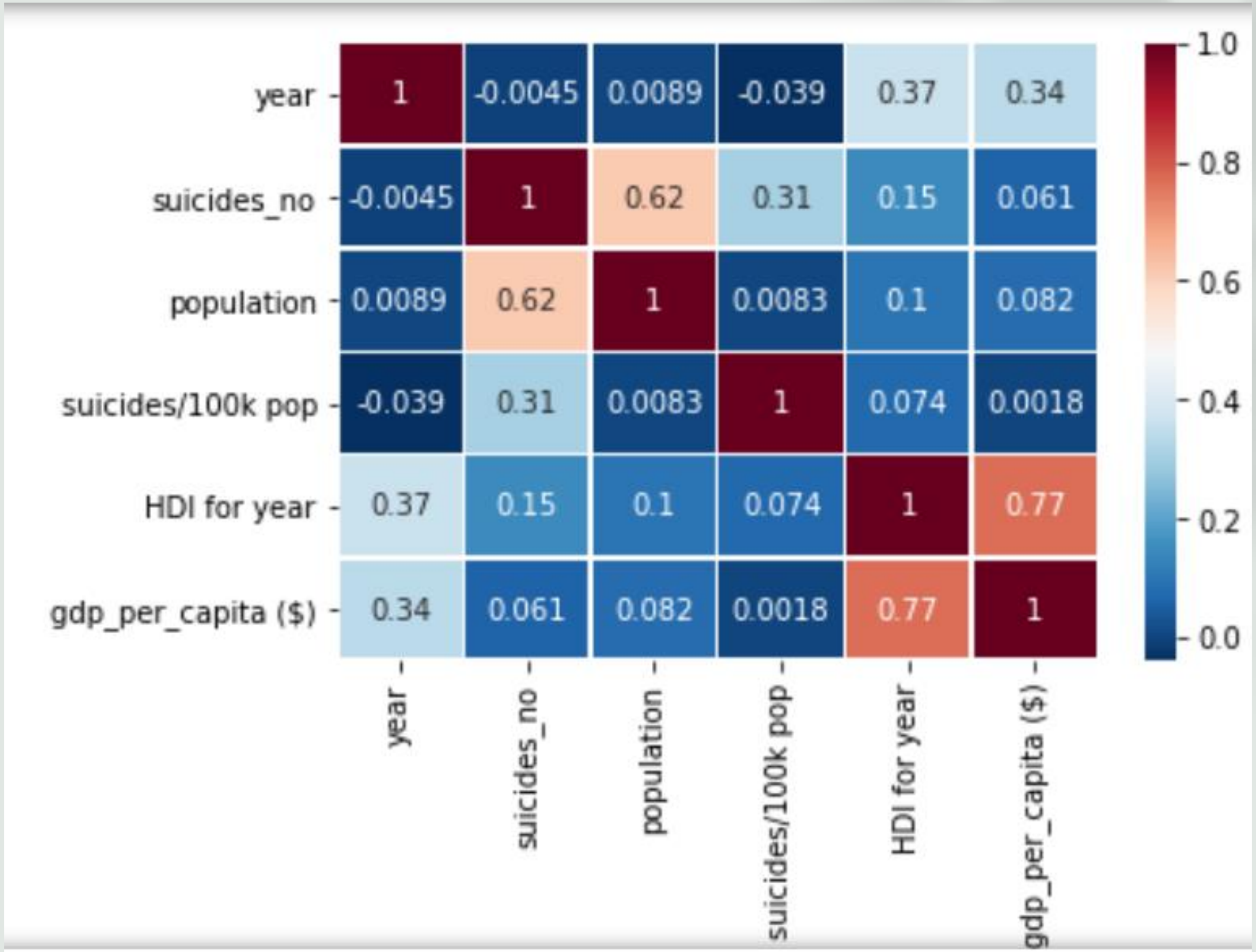
It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

HeatMap



sb.heatmap(pearsoncorr, xticklabels=pearsoncorr.columns, yticklabels=pearsoncorr.columns, cmap='RdBu_r', annot=True, linewidth=0.5)

HeatMap

import seaborn as sb

- Pandas **dataframe.corr()** is used to find the pairwise correlation of all columns in a dataframe.
- Any **na** values are automatically excluded.
- Any **non-numeric** data type column in the dataframe will be ignored.
- dataframe.corr parameters:
  dataframe.corr(method='',min_periods=1)
- method: {‘pearson’, ‘kendall’, ‘spearman’} or callable

HeatMap

SuicideRate.head()

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

HeatMap

```
pearsoncorr = SuicideRate.corr(method='pearson’)
pearsoncorr
```

	year	suicides_no	population	suicides/100k pop	HDI for year	gdp_per_capita ($)
year	1.000000	-0.004546	0.008850	-0.039037	0.366786	0.339134
suicides_no	-0.004546	1.000000	0.616162	0.306604	0.151399	0.061330
population	0.008850	0.616162	1.000000	0.008285	0.102943	0.081510
suicides/100k pop	-0.039037	0.306604	0.008285	1.000000	0.074279	0.001785
HDI for year	0.366786	0.151399	0.102943	0.074279	1.000000	0.771228
gdp_per_capita ($)	0.339134	0.061330	0.081510	0.001785	0.771228	1.000000

Chi-Square Test

Chi-Square Test

- Groups and Numbers
  - You research two groups and put them in categories of single, married or divorced:
- The numbers are definitely different, but ...
  - Is that just random chance?
  - Or have you found something interesting?
- The **Chi-Square Test** gives a "p" value to help you decide!

Finding P-Value

- **P stands for probability here.**
- **To calculate the p-value, the chi-square test is used in statistics. The different values of p indicates the different hypothesis interpretation, are given below:**
  - **P≤ 0.05; Hypothesis rejected**
  - **P>.05; Hypothesis Accepted**
- **Hypothesis**: A statement that might be true, which can then be tested.

Example

- The two **hypotheses** are.
  - Gender and preference for cats or dogs are **independent**.
  - Gender and preference for cats or dogs are **not independent**.
- Lay the data out in a table:

	Cat	**Dog**
Men	207	282
Women	231	242

- Add up rows and columns:

	Cat	**Dog**	
Men	207	282	**489**
Women	231	242	**473**
	438	**524**	**962**

- Calculate "**Expected Value**" for each entry:

	Cat	Dog	
Men	$\frac{489 \times 438}{962}$	$\frac{489 \times 524}{962}$	**489**
Women	$\frac{473 \times 438}{962}$	$\frac{473 \times 524}{962}$	**473**
	438	**524**	**962**

Which gives us:

	Cat	Dog	
Men	222.64	266.36	**489**
Women	215.36	257.64	**473**
	438	**524**	**962**

Example

- Subtract expected from observed, square it, then divide by expected:
- In other words, use formula (O−E)/E where,

	Cat	**Dog**	
Men	$(207-222.64)^2/222.64$	$(282-266.36)^2/266.36$	**489**
Women	$(231-215.36)^2/215.36$	$(242-257.64)^2/257.64$	**473**
	438	**524**	**962**

- Which gets us:

	Cat	Dog	
Men	1.099	0.918	**489**
Women	1.136	0.949	**473**
	438	**524**	**962**

- Now add up those calculated values:

$$1.099 + 0.918 + 1.136 + 0.949 = 4.102$$

Chi-Square is 4.102

Degree of Freedom = (rows − 1) × (columns − 1) =1

p = 0.04283

Code Snippet

- # defining the table
- data = [[207, 282, 241], [234, 242, 232]]
- stat, p, dof, expected = chi2_contingency(data)

- # interpret p-value
- alpha = 0.05
- print("p value is " + str(p))
- if p <= alpha:
-     print('Dependent (reject H0)')
- else:
-     print('Independent (H0 holds true)')