Learning Module 9: Parametric and Non Parametric Tests of Independence

LOS 9a: explain parametric and non-parametric tests of the hypothesis that the population correlation coefficient equals zero and determine whether the hypothesis is rejected at a given level of significance

Parametric versus Non-parametric Tests of Independence

A parametric test is a hypothesis test concerning a population parameter used when the data has **specific distribution assumptions**. If these assumptions are not met, **non-parametric tests** are used.

In summary, researchers use non-parametric testing when:

- Data do not meet distributional assumptions.
- There are outliers.
- Data is given in the form of ranks.
- The hypothesis test objective does not concern a parameter.

Hypotheses Concerning Population Correlation Coefficient

We frequently compare the population correlation coefficient to zero when testing for correlation. This helps us determine whether there's a relationship between the variables. The population correlation coefficient, represented by ρ, is used to test the relationship. There are three possible hypotheses:

- Two-sided; $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$.
- One-sided right side; $H_0 : \rho \leq 0$ versus $H_a : \rho > 0$.
- One-sided left side; $H_0 : \rho \geq 0$ versus $H_a : \rho < 0$.

Let's assume that we have variables X and Y. The sample correlation, r_{XY}, tests the above

hypotheses.

Parametric Test of a Correlation

The **parametric pairwise correlation coefficient**, also known as **Pearson correlation**, is used to test the correlation in a parametric test. The formula for the sample correlation involves the sample covariance between the X and Y variables and their respective standard deviations, which is expressed as:

$$r = \frac{S_{XY}}{S_X S_Y}$$

Where:

S_{XY} = Sample covariance between the X and Y variables.

S_X = Standard deviation of the X variable.

S_Y = Standard deviation of the Y variable.

A **t-test** can determine if the null hypothesis should be rejected using the sample correlation, r if the two variables are **normally distributed**. The formula for the t-test is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}}$$

Where:

r= Sample correlation.

n= Sample size.

$(n-2)$= Degrees of freedom.

The test statistic follows a t-distribution with $n-2$ degrees of freedom. From the equation above, it is easy to see that the sample size, n, increases, and the degrees of freedom increase. In other words, as the sample size n increases, the power of the test increases. This implies that a false

null hypothesis is more likely to be rejected as the sample size increases.

Example: Parametric Test of a Correlation

The table below shows the sample correlations between the monthly returns of five different sector-specific exchange-traded funds (ETFs) and the overall market index (Market 1). There are 48 monthly observations, and the following ETFs are included in the analysis:

	ETF 1	ETF 2	ETF 3	ETF 4	ETF 5	Market 1
ETF 1	1					
ETF 2	0.8214	1				
ETF 3	0.5672	0.6438	1			
ETF 4	0.4276	0.5789	0.4123	1		
ETF 5	0.7121	0.7942	0.6896	0.5614	1	
Market 1	0.8375	0.9096	0.7223	0.6954	0.7919	1

Using a 1% significance level and the following hypotheses: $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$, calculate the t-statistic for the correlation between ETF 2 and ETF 4. Based on the calculated t-statistic, draw a conclusion about the significance of the correlation using the following sample t-table:

df	p = 0.10	p = 0.05	p = 0.025	p = 0.01	p = 0.005
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724
36	1.306	1.688	2.028	2.434	2.719
37	1.305	1.687	2.026	2.431	2.715
38	1.304	1.686	2.024	2.429	2.712
39	1.304	1.685	2.023	2.426	2.708
40	1.303	1.684	2.021	2.423	2.704
41	1.303	1.683	2.020	2.421	2.701
42	1.302	1.682	2.018	2.418	2.698
43	1.302	1.681	2.017	2.416	2.695
44	1.301	1.680	2.015	2.414	2.692
45	1.301	1.679	2.014	2.412	2.690
46	1.300	1.679	2.013	2.410	2.687
47	1.300	1.678	2.012	2.408	2.685
48	1.299	1.677	2.011	2.407	2.682

Solution

To test the significance of the correlation between ETF 2 and ETF 4, we will use the t-test formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}}$$

Where:

r = Sample correlation coefficient (in this case, $r_{EFT2,ETF4} = 0.5789$).

n = Number of observations (48 in this case).

Now, let's calculate the t-statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}} = \frac{0.5789\sqrt{48-2}}{\sqrt{1-0.5789^2}} = 4.815$$

The calculated t-statistic for the correlation between ETF2 and ETF4 is 4.815.

At the 1% significance level, with a two-tailed test and degrees of freedom,

df = $n-2 = 46$, the critical t-value is approximately ± 2.687.

Conclusion: We reject the null hypothesis since our calculated t-statistic (4.815) is greater than the critical value (+2.687). This indicates sufficient evidence to suggest that the correlation between ETF 2 and ETF 4 significantly differs from zero.

Non-Parametric Test of Correlation: The Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient, r_S, is a non-parametric test used to examine the relationship between two data sets when the population deviates from normality.

The Spearman rank correlation coefficient is like the Pearson correlation coefficient. The difference is that the Spearman coefficient is calculated based on the ranks of variables in the samples.

Consider two variables, X and Y. We need to calculate Spearman's Rank Correlation r_S.

Steps of Calculating Spearman's Rank Correlation Coefficient, r_S

- Rank the observations of each variable X and Y in descending order. Note that when there are tied values in the data, their ranks are calculated by taking the average of the ranks that would have been assigned to those values if they were not tied.
- Find the difference between the ranks for each pair of observations.
- Square the difference and calculate the sum of the difference, that is $\sum d_i$.
- Use the following formula to find r_S:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Where;d_i=The difference between the ranks for each pair of observationsn= Sample size.

Example: Calculating Spearman's Rank Correlation Coefficient

An analyst is studying the relationship between returns for two sectors, steel and cement, over the past 5 years by using Spearman's rank correlation coefficient. The hypotheses are $H_0 : r_S = 0$ and $H_a : r_S \neq 0$. The returns of both sectors are provided below.

Year	Steel sector returns	Cement sector returns
1	10%	8%
2	6%	7%
3	9%	5%
4	12%	6%
5	8%	9%

The Spearman's rank correlation coefficient is *closest to*:

Solution

Year	Steel sector returns (X)	Cement sector returns (Y)	Rank order for X	Rank order for Y	D	d^2
1	10%	8%	2	2	0	0
2	6%	7%	5	3	2	4
3	9%	5%	3	5	−2	4
4	12%	6%	1	4	−3	9
5	8%	9%	4	1	3	9
					Sum=	26

We can now use the formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - [\frac{(6 \times 26)}{5 \times (5^2-1)}] = 1 - 1.3$$
$$r_s = -0.3$$

This indicates a very weak negative correlation between the returns of the steel and cement sectors.

Hypothesis Test for the Spearman Rank Correlation

The hypothesis test on the Spearman Rank depends on the sample size. If the sample size is small ($n \leq 30$), we would need a specialized table of critical value. On the other hand, if the sample size is large ($n > 30$), we can perform a t-test using the test statistic similar to that of Pearson correlation:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{(1-r_s^2)}}$$

Consider the above example. Assume we want to conduct a hypothesis test at a 5% significance level. The hypotheses statement is $H_0 : r_S = 0$ and $H_a : r_S \neq 0$

Question

Assume an investment analyst, John Smith, is studying the relationship between two stocks, X and Y. Based on 100 observations, he has found that $S_{XY} = 10, S_X = 2$, and $S_Y = 8$. Smith needs to find the sample correlation r_{XY} and use it to perform a t-test to determine if there is a significant correlation between the returns of stocks X and Y. The critical value for the test statistic at the 0.05 level of significance is approximately 1.96. He should conclude that the statistical relationship between X and Y is:

A. Significant because the test statistic falls outside the range of the critical values.
B. Significant, because the absolute value of the test statistic is less than the critical value.
C. Insignificant because the test statistic falls outside the range of the critical values.

Solution

The correct answer is A.

Note that the sample correlation coefficient, r_{XY} is calculated using the following formula:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

Substituting the given values in this formula, we get:

$$r_{XY} = \frac{10}{2 \times 8} = 0.625$$

To test the significance of the sample correlation, we can use a t-test with the following null and alternative hypotheses: $H_0 = \rho = 0$ and $H_\propto = \rho \neq 0$

The test statistic for this test is calculated using the following formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where:

r = Sample correlation coefficient.

n = The Number of observations.

Substituting the given values into this formula, we get:

$$t = \frac{0.625\sqrt{100-2}}{\sqrt{1-0.625^2}} \approx \frac{6.1872}{0.7806} = 7.9262$$

The critical value for the test statistic at the 0.05 level of significance is approximately 1.96.

Since our calculated test statistic (7.9262) is greater than the upper bound of the critical values for the test statistic (1.96), we reject the null hypothesis. This indicates sufficient evidence to suggest that the correlation between X and Y is significantly different from zero.

Therefore, John Smith should conclude that the statistical relationship between X and Y is significant because the test statistic falls outside the range of the critical values (**Option A**).

LOS 9b: Explain tests of independence based on contingency table data

With categorical or discrete data, correlation is not suitable for assessing relationships between variables. Instead, we use a non-parametric test called the chi-square test of independence, which employs a chi-square distributed test statistic.

We employ a contingency table to structure the data when examining the connection between two categorical variables. Subsequently, we apply a test of independence utilizing a chi-square distribution to assess whether a noteworthy relationship exists between these variables. The test statistic is calculated as follows:

$$\chi = \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{(E_{ij})}$$

Where:

$$E_{ij} = \frac{(\text{Total row i}) \times (\text{Total column j})}{\text{Overall Total}}$$

m= Number of cells in the table, the Number of groups in the first class, multiplied by the number of groups in the second class.

O_{ij}= Number of observations in each cell of row i and column j (i.e., observed frequency).

E_{ij}= Expected number of observations in each cell of row i and column j, assuming independence (i.e., expected frequency).

The degrees of freedom are given by:

$$\text{Degrees of freedom} = (r - 1)(c - 1)$$

Where:

r= Number of rows.

c= Number of columns.

Example: Testing Independence Based on Contingency Table Data

The following contingency table shows the responses of two categories of investors (employed vs. retired) with regard to their primary investment objectives (growth, income, or both). The total sample size is 173.

	Growth	Income	Both	Total
Employed	52	25	10	87
Retired	32	47	7	86
Total	84	72	17	173

Use a 95% significance level to test whether there is any significant difference between employed and retired investors concerning primary investment objectives.

Solution

H_0: There is no significant difference between employed and retired investors with regard to primary investment objectives.

H_α: There is a significant difference between employed and retired investors with regard to primary investment objectives.

Step 1: We calculate the expected frequency of investors by their category (employed vs. retired) and investment objective using the following formula:

$$E_{ij} = \frac{\text{(Total row i)} \times \text{(Total column j)}}{\text{Overall Total}}$$

	Growth	Income	Both	Total
Employed	$\frac{(87\times84)}{173} = 42.24$	$\frac{(87\times72)}{173} = 36.20$	$\frac{(87\times17)}{173} = 8.55$	87
Retired	$\frac{(86\times84)}{173} = 41.75$	$\frac{(86\times72)}{173} = 35.79$	$\frac{(86\times17)}{173} = 8.45$	86
Total	84	72	17	173

Step 2: We calculate the scaled squared deviation for each combination of investor category and investment objective as follows:

	Growth	Income	Both
Employed	$\frac{(52-42)^2}{42} = 2.254$	$\frac{(25-36)^2}{36} = 0.469$	$\frac{(10-9)^2}{9} = 0.246$
Retired	$\frac{(32-42)^2}{42} = 2.280$	$\frac{(47\times36)^2}{36} = 3.510$	$\frac{(7-8)^2}{8} = 0.349$
Total	4.534	6.979	0.495

Step 3: We calculate the value of χ^2:

$$\chi^2 = 4.534 + 6.979 + 0.495 = 12.008$$

Step 4: The critical value of X^2 is 5.99. It is determined as follows:

- There are $(r-1)(c-1) = (2-1) \times (3-1) = 2$ degrees of freedom.
- It is a one-sided test with a 5% level of significance.

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Decision rule: The calculated value of $\chi^2 = 12.008$ is greater than the critical value of 5.99. As such, there is sufficient evidence to support the conclusion that retired investors and employed investors have different primary investment objectives.

Question

Regarding the chi-square test of independence, which statement is accurate? The chi-square test of independence is:

A. A parametric hypothesis test.
B. Used to test whether two categorical variables are related to each other.
C. Used to test whether two continuous variables are related to each other.

Solution

The correct answer is B. The chi-square test of independence is a non-parametric hypothesis test that can be used to test whether two categorical variables are related.

A is incorrect because the chi-square test of independence is non-parametric, not parametric.

C is incorrect because the chi-square test of independence is used for categorical variables, not continuous variables.