Lecture Notes: Understanding Correlation Coefficients

Dr. Ratnesh Prasad Srivastava, CSIT, GGV, Bilaspur, C.G.

August 3, 2025

Abstract

These lecture notes provide a comprehensive overview of various correlation coefficients used in statistics. We will explore Pearson's r, Spearman's ρ, Kendall's τ, Point-Biserial r_{pb}, Phi ϕ, and Cramer's V. For each coefficient, we will define its purpose, provide the mathematical formula, and walk through a step-by-step calculation using a single illustrative dataset. The aim is to provide students with a clear understanding of when and how to apply each correlation measure, now enhanced with visual representations of the relationships.

1 Introduction to Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's important to note that correlation does not imply causation. There are various types of correlation coefficients, each suited for different types of data and relationships.

In these notes, we will use a hypothetical dataset of 10 students to demonstrate the calculation of different correlation coefficients.

1.1 Our Example Dataset

Consider the following data for 10 students:

Student ID	Hours_Studied (X)	Exam_Score (Y)	Tutoring_Attended (Z)	Pass_Fail (W)	Stress_Level (S)	Gender (G)
1	5	75	No	Pass	Medium	Male
2	8	90	Yes	Pass	Low	Female
3	3	60	No	Fail	High	Female
4	7	85	Yes	Pass	Medium	Male
5	4	65	No	Fail	High	Female
6	6	80	Yes	Pass	Low	Male
7	2	50	No	Fail	High	Female
8	9	95	Yes	Pass	Low	Male
9	5	70	No	Pass	Medium	Female
10	7	88	Yes	Pass	Low	Male

Where:

- **Hours_Studied (X)** and **Exam_Score (Y)** are continuous variables.
- **Tutoring_Attended (Z)** and **Pass_Fail (W)** are dichotomous variables (Yes/No, Pass/Fail).
- **Stress_Level (S)** is an ordinal variable (Low, Medium, High).
- **Gender (G)** is a nominal variable (Male/Female).

2 Types of Correlation Coefficients

2.1 Pearson Correlation Coefficient (r)

The Pearson product-moment correlation coefficient measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 (perfect negative linear correlation) to +1 (perfect positive linear correlation), with 0 indicating no linear correlation.

2.1.1 Formula

The formula for Pearson's r is:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Alternatively, using covariance and standard deviations:

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y}$$

Where:

- X_i, Y_i are individual data points.
- $\bar{X}, \bar{Y}$ are the means of variables X and Y.
- s_X, s_Y are the standard deviations of variables X and Y.

2.1.2 Example: Hours_Studied (X) vs. Exam_Score (Y)

Student ID	**X**	**Y**	$\mathbf{X - \bar{X}}$	$\mathbf{Y - \bar{Y}}$	$\mathbf{(X - \bar{X})(Y - \bar{Y})}$
1	5	75	-0.6	-6.8	4.08
2	8	90	2.4	8.2	19.68
3	3	60	-2.6	-21.8	56.68
4	7	85	1.4	3.2	4.48
5	4	65	-1.6	-16.8	26.88
6	6	80	0.4	-1.8	-0.72
7	2	50	-3.6	-31.8	114.48
8	9	95	3.4	13.2	44.88
9	5	70	-0.6	-11.8	7.08
10	7	88	1.4	6.2	8.68
Sum	**56**	**758**	**0**	**0**	**286.8**

2.1.3 Calculation Steps

1. **Calculate Means:** $\bar{X} = \frac{\sum X}{n} = \frac{56}{10} = 5.6$ $\bar{Y} = \frac{\sum Y}{n} = \frac{758}{10} = 75.8$

2. **Calculate Deviations and Product of Deviations:** We need $\sum(X_i - \bar{X})(Y_i - \bar{Y})$. From the table, this sum is 286.8.

3. **Calculate Squared Deviations:**

X	$(\mathbf{X}-\bar{\mathbf{X}})^2$	Y	$(\mathbf{Y}-\bar{\mathbf{Y}})^2$
5	$(-0.6)^2 = 0.36$	75	$(-6.8)^2 = 46.24$
8	$(2.4)^2 = 5.76$	90	$(8.2)^2 = 67.24$
3	$(-2.6)^2 = 6.76$	60	$(-21.8)^2 = 475.24$
7	$(1.4)^2 = 1.96$	85	$(3.2)^2 = 10.24$
4	$(-1.6)^2 = 2.56$	65	$(-16.8)^2 = 282.24$
6	$(0.4)^2 = 0.16$	80	$(-1.8)^2 = 3.24$
2	$(-3.6)^2 = 12.96$	50	$(-31.8)^2 = 1011.24$
9	$(3.4)^2 = 11.56$	95	$(13.2)^2 = 174.24$
5	$(-0.6)^2 = 0.36$	70	$(-11.8)^2 = 139.24$
10	$(1.4)^2 = 1.96$	88	$(6.2)^2 = 38.44$
Sum	**44.36**	**Sum**	**2247.6**

So, $\sum(X_i - \bar{X})^2 = 44.36$ and $\sum(Y_i - \bar{Y})^2 = 2247.6$.

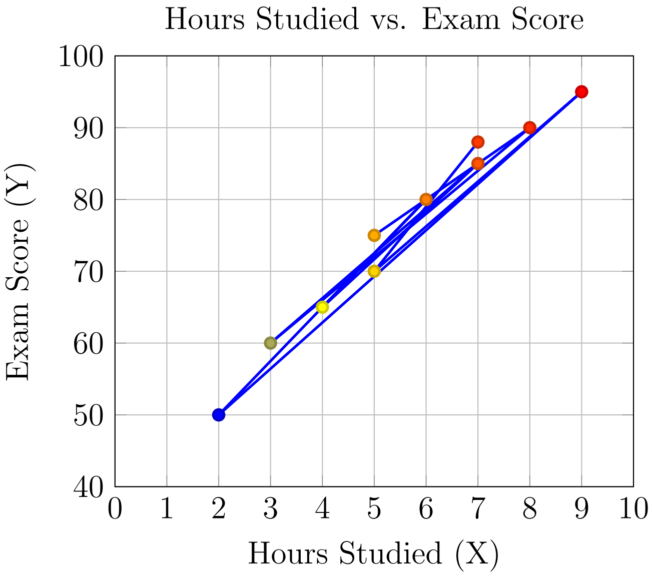
4. **Apply the Formula:**

$$r = \frac{286.8}{\sqrt{44.36 \times 2247.6}}$$
$$r = \frac{286.8}{\sqrt{99696.96}}$$
$$r = \frac{286.8}{315.748}$$
$$r \approx 0.908$$

There is a strong positive linear correlation between hours studied and exam score.

2.1.4 Visualization: Scatter Plot

A scatter plot is ideal for visualizing the relationship between two continuous variables like Hours Studied and Exam Score.



2.2 Spearman’s Rank Correlation Coefficient (ρ)

Spearman’s rank correlation coefficient assesses the strength and direction of a monotonic relationship between two ranked variables. It is suitable for ordinal data or when the assumptions for Pearson’s r (e.g., linearity, normality) are violated. It ranges from -1 to +1.

2.2.1 Formula

The formula for Spearman’s ρ is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where:

- d_i is the difference between the ranks of corresponding values for each observation.
- n is the number of observations.

2.2.2 Example: Hours_Studied (X) vs. Stress_Level (S)

First, we need to rank both variables. For ‘Stress$_{L}evel‘, weassignranks : Low = 1, Medium = 2, High = 3. Iftherearenties, weaveragetheranks.$

Student ID	**X**	**Rank(X) ($\mathbf{R_X}$)**	**S**	**Rank(S) ($\mathbf{R_S}$)**	$\mathbf{d_i = R_X - R_S}$	$\mathbf{d_i^2}$
1	5	4.5	Medium	6	-1.5	2.25
2	8	9	Low	2.5	6.5	42.25
3	3	2	High	9	-7	49
4	7	7.5	Medium	6	1.5	2.25
5	4	3	High	9	-6	36
6	6	6	Low	2.5	3.5	12.25
7	2	1	High	9	-8	64
8	9	10	Low	2.5	7.5	56.25
9	5	4.5	Medium	6	-1.5	2.25
10	7	7.5	Low	2.5	5	25
Sum					**0**	**297.75**

The sum $\sum d_i^2 = 297.75$.

2.2.3 Calculation Steps

(a) **Rank the Data:** As shown in the table above.

(b) **Calculate d_i and d_i^2:** As shown in the table above, $\sum d_i^2 = 297.75$.

(c) **Apply the Formula:** Given $n = 10$:

$$\rho = 1 - \frac{6 \times 297.75}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{1786.5}{10(100 - 1)}$$

$$\rho = 1 - \frac{1786.5}{990}$$

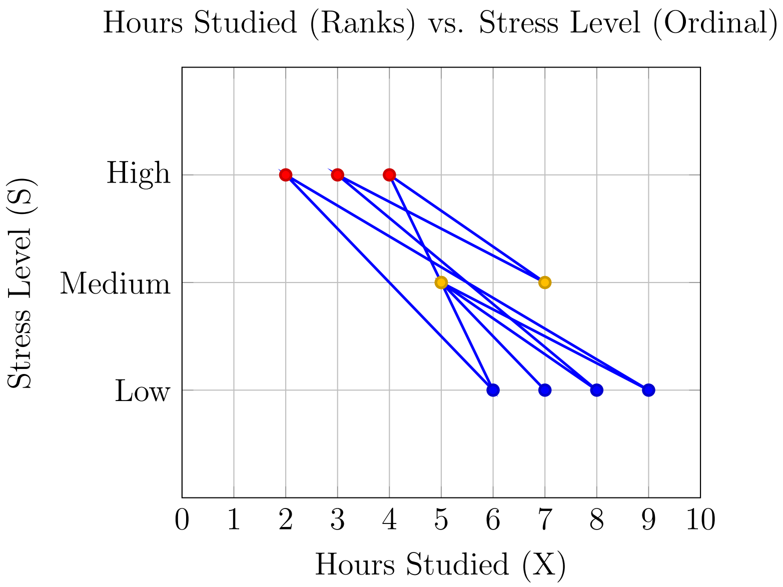
$$\rho = 1 - 1.8045$$

$$\rho \approx -0.8045$$

There is a strong negative monotonic relationship, meaning as hours studied increase, stress level tends to decrease.

2.2.4 Visualization: Scatter Plot of Ranks

To visualize the monotonic relationship for Spearman's correlation, we can plot the ranks of ‘Hours$_Studied‘againsttheranksof‘Stress_Level‘.Weusenumericalvaluesfor‘Str$
$1, Medium = 2, High = 3)torepresenttheordinalscale.$



2.3 Kendall's Tau (τ)

Kendall's Tau is another non-parametric measure of the strength and direction of association between two ranked variables. It is often preferred over Spearman's ρ when dealing with smaller sample sizes or data with many tied ranks. It measures the probability that two randomly selected observations will be in the same order (concordant) versus different orders (discordant).

2.3.1 Formula (Kendall’s Tau-b for ties)

$$\tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + N_{tX})(N_c + N_d + N_{tY})}}$$

Where:

- N_c = Number of concordant pairs.
- N_d = Number of discordant pairs.
- N_{tX} = Number of pairs tied on X only.
- N_{tY} = Number of pairs tied on Y only. /ul¿ $N_c + N_d + N_{tX} + N_{tY}$ is the total number of pairs, which is $\frac{n(n-1)}{2}$.

2.3.2 Example: Hours_Studied (X) vs. Stress_Level (S)

We will use the same data as for Spearman’s. First, sort the data by the first variable (X).

Student ID	Hours_Studied (X)	Stress_Level (S)
7	2	High
3	3	High
5	4	High
1	5	Medium
9	5	Medium
6	6	Low
4	7	Medium
10	7	Low
2	8	Low
8	9	Low

2.3.3 Calculation Steps

(a) **Sort Data:** Sort the data by ‘Hours$_{S}tudied‘(X)inascendingorder.$**Identify Pairs and C**

(b) (c) **Concordant (C):** If S of subsequent observation is higher.

(d) **Discordant (D):** If S of subsequent observation is lower.

(e) **Tied on S (T_S):** If S is the same.

(f) **Tied on X (T_X):** If X is the same (handled by sorting, but affects denominator).

Let’s use numerical representation for S: Low=1, Medium=2, High=3.

Student 7 (X=2, S=3):

– (3,3): Tied on S
– (5,3): Tied on S
– (1,2): D
– (9,2): D
– (6,1): D
– (4,2): D
– (10,1): D
– (2,1): D
– (8,1): D

C=0, D=7, T_S=2 (pairs with S=3)

Student 3 (X=3, S=3): (compared to 8 subsequent)

- (5,3): Tied on S
- (1,2): D
- (9,2): D
- (6,1): D
- (4,2): D
- (10,1): D
- (2,1): D
- (8,1): D

C=0, D=6, T_S=1

Student 5 (X=4, S=3): (compared to 7 subsequent)

- (1,2): D
- (9,2): D
- (6,1): D
- (4,2): D
- (10,1): D
- (2,1): D
- (8,1): D

C=0, D=7, T_S=0

Student 1 (X=5, S=2): (compared to 6 subsequent)

- (9,2): Tied on S
- (6,1): D
- (4,2): Tied on S
- (10,1): D
- (2,1): D
- (8,1): D

C=0, D=4, T_S=2

Student 9 (X=5, S=2): (compared to 5 subsequent)

- (6,1): D
- (4,2): Tied on S
- (10,1): D
- (2,1): D
- (8,1): D

C=0, D=4, T_S=1

Student 6 (X=6, S=1): (compared to 4 subsequent)

- (4,2): C
- (10,1): Tied on S
- (2,1): Tied on S
- (8,1): Tied on S

C=1, D=0, T_S=3

Student 4 (X=7, S=2): (compared to 3 subsequent)

- (10,1): D
- (2,1): D
- (8,1): D

C=0, D=3, T_S=0

Student 10 (X=7, S=1): (compared to 2 subsequent)

- (2,1): Tied on S
- (8,1): Tied on S

C=0, D=0, T_S=2

Student 2 (X=8, S=1): (compared to 1 subsequent)

- (8,1): Tied on S

C=0, D=0, T_S=1

Student 8 (X=9, S=1): (compared to 0 subsequent) C=0, D=0, T_S=0

Total Counts: $N_c = 0+0+0+0+0+1+0+0+0+0 = 1$ $N_d = 7+6+7+4+4+0+3+0+0+0 = 31$

Count Tied Pairs for X (N_{tX}): X=5 (2 students: 1, 9). Number of pairs tied on X = $\frac{2(2-1)}{2} = 1$. X=7 (2 students: 4, 10). Number of pairs tied on X = $\frac{2(2-1)}{2} = 1$. So, $N_{tX} = 1+1 = 2$.

Count Tied Pairs for S (N_{tY}): S=High (3 students: 7, 3, 5). Number of pairs tied on S = $\frac{3(3-1)}{2} = 3$. S=Medium (3 students: 1, 9, 4). Number of pairs tied on S = $\frac{3(3-1)}{2} = 3$. S=Low (4 students: 6, 10, 2, 8). Number of pairs tied on S = $\frac{4(4-1)}{2} = 6$. So, $N_{tY} = 3+3+6 = 12$.

- **Apply the Formula:**

$$\tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + N_{tX})(N_c + N_d + N_{tY})}}$$

$$\tau_b = \frac{1-31}{\sqrt{(1+31+2)(1+31+12)}}$$

$$\tau_b = \frac{-30}{\sqrt{(34)(44)}}$$

$$\tau_b = \frac{-30}{\sqrt{1496}}$$

$$\tau_b = \frac{-30}{38.678}$$

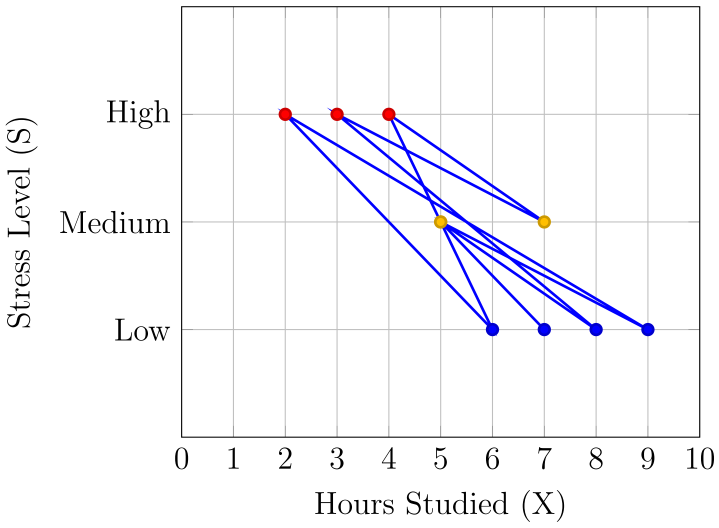
$$\tau_b \approx -0.775$$

This indicates a strong negative association, consistent with Spearman's ρ.

2.3.4 Visualization: Scatter Plot (Same as Spearman for visual representation)

The visualization for Kendall's Tau is similar to Spearman's as both deal with monotonic relationships between ranked or ordinal variables. We use the same scatter plot as for Spearman's to illustrate the trend.

Hours Studied (Ranks) vs. Stress Level (Ordinal)



2.4 Point-Biserial Correlation (r_{pb})

The Point-Biserial correlation coefficient is used to measure the strength and direction of the association between a continuous variable and a dichotomous variable (a variable with two categories, often coded as 0 and 1).

2.4.1 Formula

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_Y} \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

Alternatively, a simpler form when Y is the continuous variable and X is the dichotomous variable (coded 0/1):

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_Y} \sqrt{p_1 p_0}$$

Where:
- $\bar{Y}_1$ = Mean of the continuous variable (Y) for group 1.
- $\bar{Y}_0$ = Mean of the continuous variable (Y) for group 0.
- s_Y = Standard deviation of the continuous variable (Y) for the entire dataset.
- n_1 = Number of observations in group 1.
- n_0 = Number of observations in group 0.
- n = Total number of observations ($n_1 + n_0$).
- $p_1 = n_1/n$ = Proportion of observations in group 1.
- $p_0 = n_0/n$ = Proportion of observations in group 0.

2.4.2 Example: Exam_Score (Y) vs. Tutoring_Attended (Z)

Let's code 'Tutoring$_Attended`:Yes=1,No=0$.

Student ID	Exam_Score (Y)	Tutoring_Attended (Z)
1	75	0
2	90	1
3	60	0
4	85	1
5	65	0
6	80	1
7	50	0
8	95	1
9	70	0
10	88	1

2.4.3 Calculation Steps

(a) **Identify Groups and Counts:** Group 1 (Tutoring_Attended = Yes): Students 2, 4, 6, 8, 10. $n_1 = 5$. Group 0 (Tutoring_Attended = No): Students 1, 3, 5, 7, 9. $n_0 = 5$. Total $n = 10$.

(b) **Calculate Means for Each Group:** $\bar{Y}_1 = \frac{90+85+80+95+88}{5} = \frac{438}{5} = 87.6$ $\bar{Y}_0 = \frac{75+60+65+50+70}{5} = \frac{320}{5} = 64.0$

(c) **Calculate Overall Standard Deviation of Y (s_Y):** From Pearson's calculation, we found $\sum(Y_i - \bar{Y})^2 = 2247.6$. The sample standard deviation formula is $s_Y = \sqrt{\frac{\sum(Y_i-\bar{Y})^2}{n-1}}$. $s_Y = \sqrt{\frac{2247.6}{10-1}} = \sqrt{\frac{2247.6}{9}} = \sqrt{249.733}$ $s_Y \approx 15.799$

(d) **Calculate Proportions:** $p_1 = n_1/n = 5/10 = 0.5$ $p_0 = n_0/n = 5/10 = 0.5$

(e) **Apply the Formula:**

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_Y}\sqrt{p_1 p_0}$$

$$r_{pb} = \frac{23.6}{15.799}\sqrt{0.25}$$

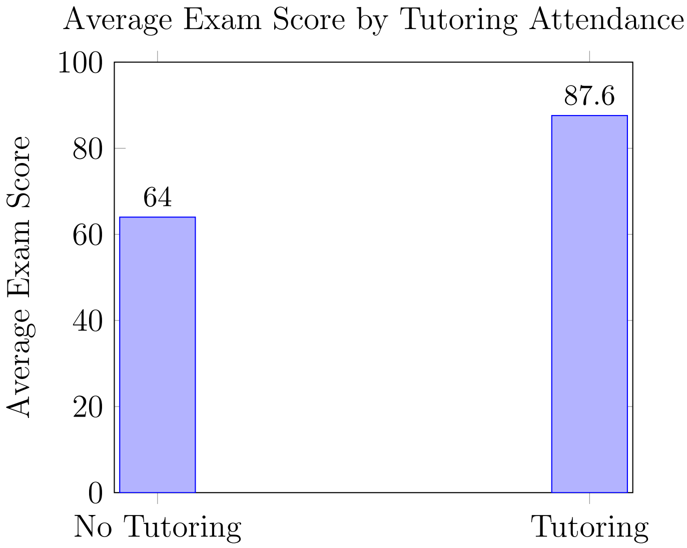
$$r_{pb} = 1.4937 \times 0.5$$

$$r_{pb} \approx 0.747$$

This indicates a strong positive correlation: students who attended tutoring tend to have higher exam scores.

2.4.4 Visualization: Bar Chart of Means

A bar chart effectively shows the difference in the mean ‘Exam$_{score}‘betweenstudentswhoatt$



2.5 Phi Coefficient (ϕ)

The Phi coefficient is used to measure the association between two dichotomous variables. It is essentially a Pearson correlation coefficient calculated on two binary variables. It ranges from -1 to +1, but values are often between 0 and 1 for positive associations.

2.5.1 Formula

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Where a, b, c, d are the cell frequencies in a 2×2 contingency table:

	Variable 2 (Category 1)	**Variable 2 (Category 2)**
Variable 1 (Category 1)	a	b
Variable 1 (Category 2)	c	d

2.5.2 Example: Tutoring_Attended (Z) vs. Pass_Fail (W)

Let's create a contingency table. 'Tutoring$_Attended`:Yes,No`Pass_Fail`:Pass,Fail$
Raw Data Review:

- Student 1: No, Pass
- Student 2: Yes, Pass
- Student 3: No, Fail
- Student 4: Yes, Pass
- Student 5: No, Fail
- Student 6: Yes, Pass
- Student 7: No, Fail
- Student 8: Yes, Pass
- Student 9: No, Pass
- Student 10: Yes, Pass

Contingency Table:

	Pass	**Fail**	**Total**
Tutoring_Yes	$a = 5$	$b = 0$	$a + b = 5$
Tutoring_No	$c = 2$	$d = 3$	$c + d = 5$
Total	$a + c = 7$	$b + d = 3$	$N = 10$

2.5.3 Calculation Steps

(a) **Construct the** 2×2 **Table:** As shown above.

(b) **Identify Cell Frequencies:** $a = 5$ (Tutoring Yes, Pass) $b = 0$ (Tutoring Yes, Fail) $c = 2$ (Tutoring No, Pass) $d = 3$ (Tutoring No, Fail)

(c) **Apply the Formula:**

$$\phi = \frac{(5)(3) - (0)(2)}{\sqrt{(5+0)(2+3)(5+2)(0+3)}}$$

$$\phi = \frac{15 - 0}{\sqrt{(5)(5)(7)(3)}}$$

$$\phi = \frac{15}{\sqrt{525}}$$

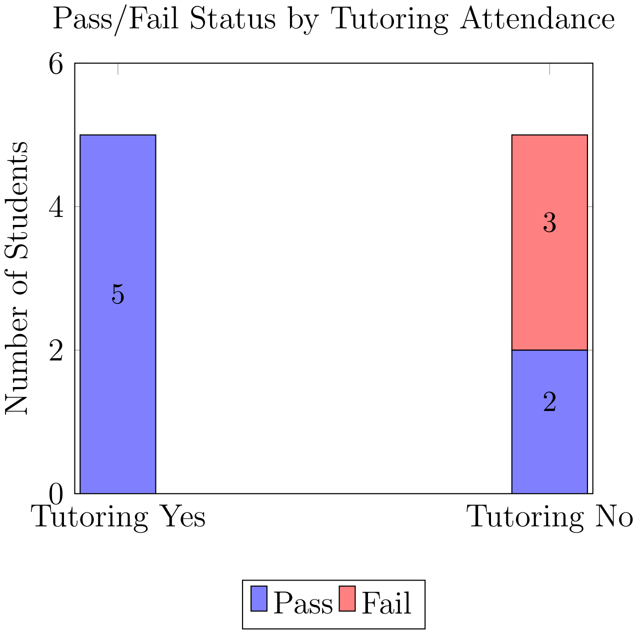
$$\phi = \frac{15}{22.913}$$

$$\phi \approx 0.655$$

This indicates a moderately strong positive association: attending tutoring is associated with passing the exam.

2.5.4 Visualization: Stacked Bar Chart

A stacked bar chart is useful for visualizing the frequencies within a 2×2 contingency table, showing how the ‘Pass$_{F}ail‘outcomeisdistributedacross‘Tutoring_{A}ttended‘g$



2.6 Cramer's V (V)

Cramer's V is a measure of association between two nominal variables, or between a nominal and an ordinal/interval variable treated as nominal. It is a more general measure than the Phi coefficient, applicable to contingency tables larger than 2×2. It ranges from 0 to 1, where 0 indicates no association and 1 indicates a perfect association.

2.6.1 Formula

$$V = \sqrt{\frac{\chi^2}{N \times \min(k-1, r-1)}}$$

Where:

- χ^2 = Chi-square statistic.
- N = Total number of observations.
- k = Number of columns in the contingency table.
- r = Number of rows in the contingency table.
- $\min(k-1, r-1)$ = The minimum of (number of columns - 1) and (number of rows - 1). This represents the degrees of freedom for a 2×2 table, or the smaller of the two for larger tables.

2.6.2 Example: Gender (G) vs. Pass_Fail (W)

Raw Data Review:

- Student 1: Male, Pass
- Student 2: Female, Pass
- Student 3: Female, Fail
- Student 4: Male, Pass
- Student 5: Female, Fail
- Student 6: Male, Pass
- Student 7: Female, Fail
- Student 8: Male, Pass
- Student 9: Female, Pass
- Student 10: Male, Pass

Contingency Table (Observed Frequencies, O_{ij}):

	Pass	**Fail**	**Row Total**
Male	$O_{11} = 5$	$O_{12} = 0$	$R_1 = 5$
Female	$O_{21} = 2$	$O_{22} = 3$	$R_2 = 5$
Column Total	$C_1 = 7$	$C_2 = 3$	$N = 10$

2.6.3 Calculation Steps

(a) **Construct the Contingency Table:** As shown above.

(b) **Calculate Expected Frequencies (E_{ij}):** $E_{ij} = \frac{\text{(Row Total)} \times \text{(Column Total)}}{\text{Grand Total}}$

- E_{11} (Male, Pass) $= \frac{5 \times 7}{10} = 3.5$
- E_{12} (Male, Fail) $= \frac{5 \times 3}{10} = 1.5$
- E_{21} (Female, Pass) $= \frac{5 \times 7}{10} = 3.5$
- E_{22} (Female, Fail) $= \frac{5 \times 3}{10} = 1.5$

(c) **Calculate Chi-square (χ^2):**

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 = \frac{(1.5)^2}{3.5} + \frac{(-1.5)^2}{1.5} + \frac{(-1.5)^2}{3.5} + \frac{(1.5)^2}{1.5}$$

$$\chi^2 = \frac{2.25}{3.5} + \frac{2.25}{1.5} + \frac{2.25}{3.5} + \frac{2.25}{1.5}$$

$$\chi^2 = 0.6428 + 1.5 + 0.6428 + 1.5$$

$$\chi^2 = 4.2856$$

(d) **Apply the Cramer's V Formula:** $N = 10$, $k = 2$ (columns: Pass, Fail), $r = 2$ (rows: Male, Female). $\min(k-1, r-1) = \min(2-1, 2-1) = \min(1, 1) = 1$.

$$V = \sqrt{\frac{\chi^2}{N \times \min(k-1, r-1)}}$$

$$V = \sqrt{\frac{4.2856}{10 \times 1}}$$

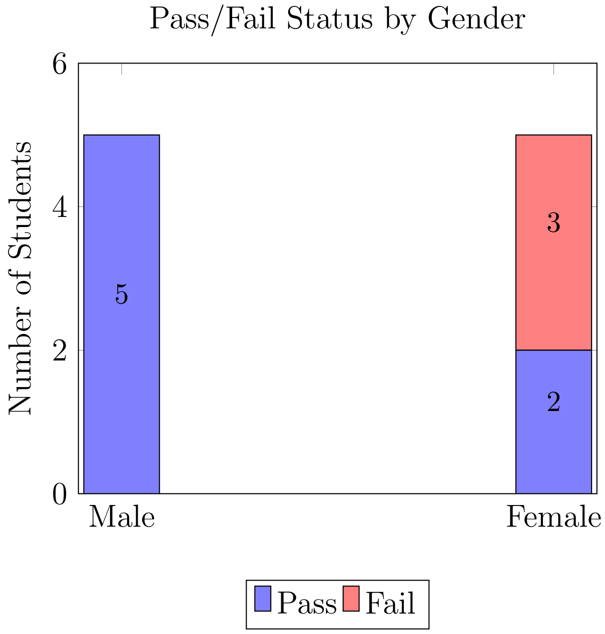
$$V = \sqrt{0.42856}$$

$$V \approx 0.6546$$

This indicates a moderately strong association between gender and passing/failing the exam. Note that for a 2×2 table, Cramer's V is equal to the absolute value of the Phi coefficient.

2.6.4 Visualization: Stacked Bar Chart

Similar to the Phi coefficient, a stacked bar chart helps visualize the association between two nominal variables like ‘Gender‘ and ‘Pass$_Fail$‘.



3 Conclusion

This lecture note has provided a practical, step-by-step guide to calculating various correlation coefficients using a single dataset, now complemented with relevant visualizations. Understanding when to use each coefficient based on the type of data (continuous, ordinal, dichotomous, nominal) and the nature of the relationship (linear, monotonic, association) is crucial for accurate statistical analysis. Remember that correlation measures association, not causation.