

WORKSHOP SERIES | DAY TWO

Correlation Analysis with SPSS

Bivariate and multiple Pearson correlation, from design and assumptions to APA reporting

Facilitator: Assoc. Prof. Enoch Oladunmoye

College of Humanities and Social Sciences

Kampala International University

Email: enoch.oladunmoye@kiu.ac.ug

Where we left off and where we are going

From day one

- Variables defined in Variable View
- Value labels and missing codes declared
- Data entered case by case
- Negative items reverse scored
- Scale scores computed with MEAN
- Descriptive statistics produced

1 Concept

What correlation is, what Pearson r measures and how to read it

2 Assumptions

The seven conditions that must hold, and how to test each one in SPSS

3 Design

Correlational design, sampling techniques and sample size

4 Procedure

Analyze, Correlate, Bivariate, for two variables and for a full matrix

5 Output

Reading the correlation table and the significance flags

6 Reporting

Writing the result in APA style, in text and in a formatted table

DEFINITION

What is correlation?

Correlation is a statistical technique that measures the strength and direction of the linear relationship between two continuous variables. It tells you whether the two variables move together, whether they move in opposite directions, or whether they are unrelated. It does not tell you that one causes the other.

Direction

Is the sign positive or negative? Do the variables rise together or move in opposite directions?

Strength

How closely do the points cluster around a straight line? This is the magnitude of r

Significance

Is the relationship large enough that it is unlikely to have arisen by chance in the population?

Pearson product moment correlation coefficient

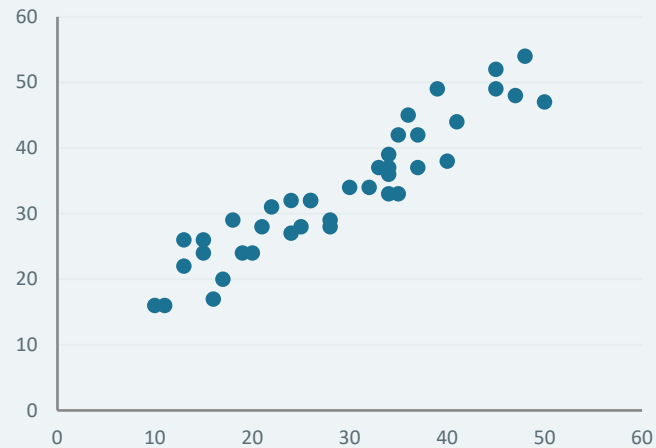
Denoted by r for a sample and by ρ for the population. It ranges from minus one, a perfect negative relationship, through zero, no linear relationship, to plus one, a perfect positive relationship.

FIGURE 1

The three possible directions

Positive

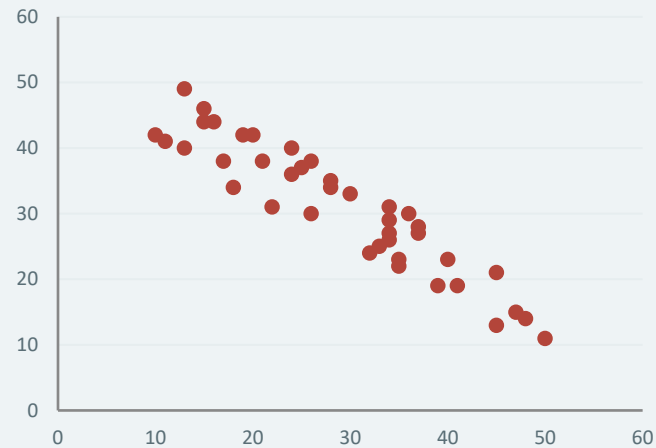
$r = +.82$



As one variable increases the other increases.
Example: hours of study and examination score.

Negative

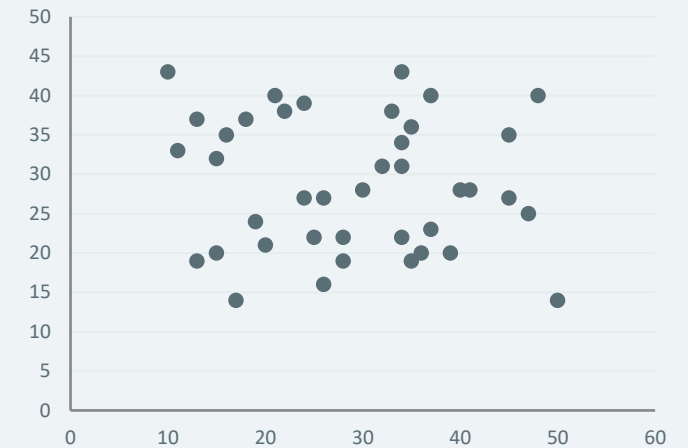
$r = -.79$



As one variable increases the other decreases.
Example: academic stress and academic performance.

No relationship

$r = .04$



Knowing one variable tells you nothing about the other.
Example: shoe size and intelligence.

Figure 1. Scatterplots illustrating a positive, a negative and a negligible linear relationship.

TABLE 1

Interpreting the size of r

Absolute value of r	Conventional description	r squared	Shared variance explained
.10 to .29	Small or weak	.01 to .08	About 1 to 8 per cent
.30 to .49	Medium or moderate	.09 to .24	About 9 to 24 per cent
.50 to 1.00	Large or strong	.25 to 1.00	About 25 per cent and above

Table 1. Conventional guidelines for describing the magnitude of a correlation coefficient, following Cohen.

The coefficient of determination

Square r and multiply by one hundred to obtain the percentage of variance the two variables share. If r equals .45 then r squared equals .2025, so the two variables share about twenty per cent of their variance. This is often more informative than r itself.

Two cautions on these labels

First, these are conventions, not laws. In some fields an r of .20 is substantial and in others it is trivial, so interpret against the literature in your own area. Second, statistical significance and practical magnitude are different questions. A very large sample can make a trivial r significant.

Bivariate correlation and the correlation matrix

Bivariate correlation

One pair of variables at a time

- You enter exactly two variables
- SPSS produces a two by two table
- Only one coefficient is of interest, and it appears twice
- Used when a single hypothesis concerns one specific pair

Example: Is there a relationship between academic stress and academic performance?

Multiple correlation matrix

Several variables simultaneously

- You enter three or more variables at once
- SPSS produces a square symmetrical matrix
- Every variable is correlated with every other variable
- Standard practice before regression, and standard in an APA results chapter

Example: What are the interrelationships among stress, self esteem, social support and performance?

The seven assumptions of Pearson correlation

1

Level of measurement

Both variables must be continuous, that is measured at interval or ratio level

2

Related pairs

Each case must supply a score on both variables. A case missing either score contributes nothing

3

Independence of observations

One respondent must not influence another. Each row is one independent case

4

Normality

The two variables should be approximately normally distributed

5

Linearity

The relationship between the two variables must be a straight line, not a curve

6

Homoscedasticity

The spread of scores on one variable should be roughly equal across all values of the other

7

Absence of outliers

Extreme scores must be identified, because a single outlier can inflate or suppress r

Assumptions one to three are settled by your design and data collection. Assumptions four to seven are checked in the data, before the correlation is run, and the evidence is reported.

TABLE 2

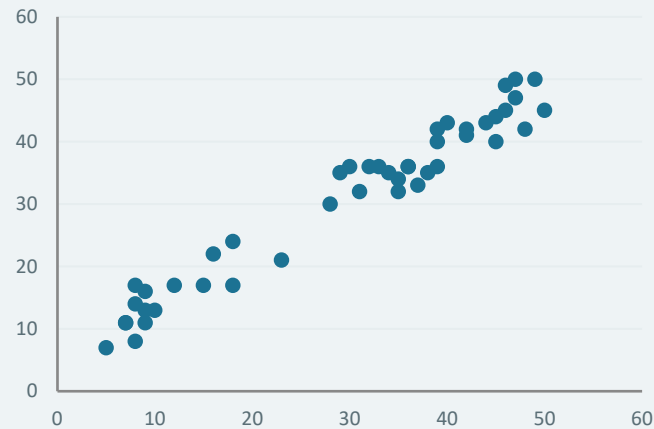
How to check each assumption in SPSS

Assumption	How you check it	What you are looking for
Level of measurement	Inspect the Measure column in Variable View	Both variables set to Scale
Related pairs and independence	Inspect your data collection procedure and the Data View	One row per independent respondent, with both scores present
Normality	Analyze, Descriptive Statistics, Explore. Also request a histogram with the normal curve	Skewness and kurtosis within about plus or minus two, and a non significant Shapiro Wilk in smaller samples
Linearity	Graphs, Legacy Dialogs, Scatter or Dot, then Simple Scatter. Add a fit line	Points falling roughly along a straight line rather than a curve
Homoscedasticity	The same scatterplot	A roughly even band of points, not a fan or funnel shape
Outliers	The same scatterplot, plus boxplots from Explore	Isolated points far from the cloud. Investigate before deciding to remove

FIGURE 2

Reading the scatterplot before you read the coefficient

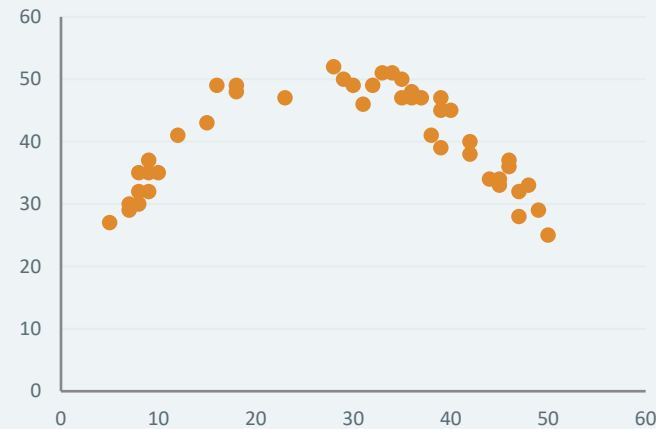
Linear and even



PROCEED

Assumptions met. Pearson r is appropriate.

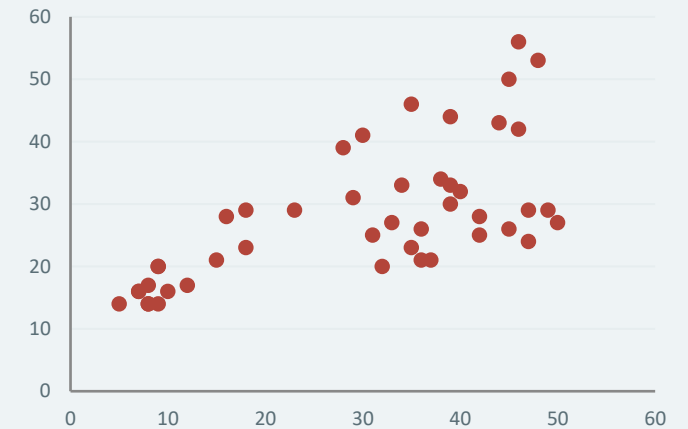
Curvilinear



DO NOT USE r

A real relationship exists but it is not linear.
Pearson r will underestimate it badly.

Heteroscedastic



INTERPRET WITH CAUTION

The spread widens across the range. The fan shape violates homoscedasticity.

Figure 2. Scatterplot patterns indicating whether the assumptions of Pearson correlation are tenable.

What to do when an assumption is violated

Problem	Appropriate response
One or both variables are ordinal	Use Spearman rank order correlation, or Kendall tau b if the sample is small with many tied ranks
Both variables are dichotomous	Use the phi coefficient, obtained through Crosstabs
One variable is dichotomous and the other continuous	Use the point biserial correlation, which SPSS computes as an ordinary Pearson r
Marked non normality	Consider a transformation, or move to Spearman, which does not require normality
The relationship is curvilinear	Do not force a Pearson r. Consider eta, polynomial regression, or report the pattern descriptively
Genuine outliers present	Verify against the raw questionnaire first. Report the analysis with and without the outlier if you remove it

In SPSS all of these except phi live in the same dialogue box. Analyze, Correlate, Bivariate offers Pearson, Kendall tau b and Spearman as three tick boxes side by side.

The required research method

Correlation requires a quantitative, non experimental approach. The researcher measures variables as they naturally occur and does not manipulate any of them.

1 Quantitative

Variables are expressed numerically so that a coefficient can be computed

2 Non experimental

No treatment is administered and no condition is assigned to participants

3 Cross sectional

Data are usually collected at a single point in time from the whole sample

4 Survey based

A questionnaire or standardised scale is the typical instrument

5 Field setting

Data are gathered in the natural setting rather than a laboratory

6 Deductive

Hypotheses are stated in advance and then tested against the data

Correlational design

Definition

A correlational design examines the degree to which two or more variables co vary in a defined population, without manipulating any variable and without assigning participants to conditions. Its purpose is to establish whether a relationship exists, how strong it is and in which direction it runs.

1

Relationship studies Establish whether and how strongly variables are associated

2

Prediction studies Use a measured predictor to forecast an outcome measured later

3

Explanatory studies Explore patterns of association among a set of constructs to refine theory

Strengths

- Permits study of variables that cannot ethically be manipulated
- Allows several variables to be examined at once
- Efficient in time and cost, and high in external validity

Limitations

- Cannot establish causation, and is open to the third variable problem
- The direction of the relationship is often ambiguous
- Sensitive to restriction of range and to common method bias

THE CARDINAL RULE

Correlation does not establish causation

Directionality

A correlation between X and Y is equally consistent with X causing Y and with Y causing X

Third variable

An unmeasured variable may be producing both, making the observed link spurious

Coincidence

With enough variables tested, some coefficients will reach significance by chance alone

Write your findings as an association, never as an effect. Say related to, associated with, or predicted by, not caused by or led to.

Sampling techniques for a correlational study

Because you intend to generalise from your sample to a population, probability sampling is the preferred approach. Every member of the population must have a known, non zero chance of selection.

1

Simple random sampling

Every member has an equal chance. Draw names or use a table of random numbers

Best when: A homogeneous and accessible population with a complete sampling frame

2

Systematic sampling

Select every kth element from a list after a random start

Best when: A long, ordered list such as a register of enrolled students

3

Stratified random sampling

Divide the population into strata, then sample randomly within each stratum

Best when: When faculty, year of study or sex must be proportionally represented

4

Cluster sampling

Sample intact groups such as classes or halls, then study all members or a sample within them

Best when: A geographically dispersed population with no individual level frame

5

Multi stage sampling

Combine techniques in successive stages, for example cluster then stratified then random

Best when: Large scale institutional or regional studies

Non probability sampling and sample size

Non probability techniques

Purposive

Cases selected deliberately because they possess the characteristic of interest

Convenience

Whoever is available and willing. Common but weakest for generalisation

Snowball

Existing participants recruit others. Useful for hidden or hard to reach populations

Quota

Fixed numbers set for each subgroup, then filled non randomly

If you use a non probability technique, say so plainly and treat it as a limitation on generalisability rather than hiding it.

Sample size	Smallest r detectable	Comment
30	About .45	Adequate only for strong relationships
50	About .35	A common practical minimum
85	About .30	Detects a medium relationship
150	About .23	Comfortable for most social science work
200 and above	About .20 or less	Preferred when a correlation matrix is reported

Table 3. Approximate sample sizes needed to detect a correlation at the five per cent level with about eighty per cent power, two tailed. Treat these as planning guides, not rules.

A general working rule for correlational work is a minimum of about thirty cases, with one hundred or more strongly preferred.

Before you run the analysis

1

Confirm your scale scores exist

Each construct must be a single computed variable, not a set of items. Use Transform, Compute Variable if not

2

Confirm Measure is set to Scale

Correlation dialogues will not treat a variable properly if it is flagged as Nominal

3

Check reliability

Analyze, Scale, Reliability Analysis. Report Cronbach alpha for each scale before correlating

4

Run descriptives

Obtain N, mean, standard deviation, skewness and kurtosis for every variable to be correlated

5

Inspect the scatterplots

One simple scatterplot for each pair of central interest, with a fit line added

6

Decide how missing data will be handled

Choose listwise or pairwise deliberately, and be prepared to state which you used

A correlation table is only as trustworthy as the scale scores that went into it.

FIGURE 3

Running the analysis: Analyze, Correlate, Bivariate

- 1 From the menu bar click Analyze, then Correlate, then Bivariate
- 2 Move your variables from the left hand list into the Variables box
- 3 For a bivariate correlation move exactly two variables
- 4 For a correlation matrix move three or more
- 5 Under Correlation Coefficients tick Pearson
- 6 Under Test of Significance choose Two tailed for a non directional hypothesis
- 7 Tick Flag significant correlations
- 8 Click Options and tick Means and standard deviations
- 9 Under Missing Values choose Exclude cases pairwise or listwise, then Continue
- 10 Click OK and read the Correlations table in the Output Viewer

Bivariate Correlations

id
sex
age
faculty
level

Variables:

stress
selfesteem
support
performance

Correlation Coefficients

☒ Pearson
☐ Kendall tau-b
☐ Spearman
☒ Flag significant correlations

Test of Significance

☒ Two-tailed
☐ One-tailed

Options **OK**

Figure 3. Layout of the Bivariate Correlations dialogue box.

Producing the scatterplot that must accompany it

- 1 Click Graphs, then Legacy Dialogs, then Scatter or Dot
- 2 Select Simple Scatter and click Define
- 3 Move the dependent or outcome variable into the Y Axis box
- 4 Move the independent or predictor variable into the X Axis box
- 5 Click OK to produce the plot in the Output Viewer
- 6 Double click the chart to open the Chart Editor
- 7 Click Elements, then Fit Line at Total, and close the editor
- 8 Inspect the plot for linearity, even spread and isolated outliers

What to record from the plot

- The general shape: straight, curved or shapeless
- The direction of the slope, upward or downward
- How tightly the points cluster around the line
- Whether the band of points is even or fans out
- Any point sitting well away from the main cloud
- Whether the range of scores is restricted at either end

In an APA results chapter the scatterplot is usually described in the text rather than reproduced, unless the pattern is unusual.

TABLE 4

Reading the bivariate output

		Academic stress	Academic performance
Academic stress	Pearson Correlation	1	-.462**
	Sig. (2-tailed)		.000
	N	196	193
Academic performance	Pearson Correlation	-.462**	1
	Sig. (2-tailed)	.000	
	N	193	195

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4. SPSS Correlations output for a single pair of variables.

Never report $p = .000$

SPSS rounds the probability to three decimal places, so .000 means the value is smaller than .0005, not that it is zero. In APA style this is written as $p < .001$. A probability can never be exactly zero.

What to look at

The diagonal of ones

Every variable correlates perfectly with itself. These cells carry no information

The mirror image

The value above the diagonal repeats below it. Report the coefficient once, not twice

The asterisks

One asterisk means significant at .05 and two asterisks mean significant at .01

The N row

Watch for a smaller N on some pairs. That is pairwise deletion at work

Read this table as: $r = -.46$, $p < .001$, $N = 193$.

TABLE 5

Reading the correlation matrix output

Variable	Statistic	1 Stress	2 Self esteem	3 Support	4 Performance
1. Academic stress	r	1	-.383**	-.271**	-.462**
	Sig.		.000	.000	.000
2. Self esteem	r	-.383**	1	.415**	.508**
	Sig.	.000		.000	.000
3. Social support	r	-.271**	.415**	1	.194*
	Sig.	.000	.000		.013
4. Academic performance	r	-.462**	.508**	.194*	1
	Sig.	.000	.000	.013	

*. Correlation is significant at the 0.05 level (2-tailed). **. Correlation is significant at the 0.01 level (2-tailed). N = 190 (listwise).

Table 5. SPSS Correlations output for four variables entered simultaneously.

The matrix is symmetrical about the diagonal, so only one triangle carries unique information. Four variables yield six unique coefficients, not sixteen. In your report you present one triangle only.

Reporting in APA style: the in text sentence

The standard APA format is: $r(df) = \text{value}$, $p = \text{value}$. Degrees of freedom for a correlation are N minus two. Report r to two decimal places with no leading zero, because r cannot exceed one.

A significant negative relationship

There was a strong negative correlation between academic stress and academic performance, $r(191) = -.46$, $p < .001$, with high levels of stress associated with lower academic performance. Stress accounted for approximately 21 per cent of the variance in performance.

A significant positive relationship

Self esteem was positively and significantly related to academic performance, $r(191) = .51$, $p < .001$, indicating that students who reported higher self esteem also achieved higher academic performance.

A non significant relationship

The relationship between age and academic performance was weak and not statistically significant, $r(191) = .08$, $p = .271$. The null hypothesis was therefore not rejected.

Note the conventions: no leading zero on r or p , italic r and p in your final document, and p reported as less than .001 rather than as .000.

TABLE 6

Reporting in APA style: the correlation table

Table 6

Means, Standard Deviations and Intercorrelations Among the Study Variables (N = 190)

Variable	M	SD	1	2	3	4
1. Academic stress	3.42	0.71	—			
2. Self esteem	3.51	0.68	-.38**	—		
3. Social support	3.87	0.62	-.27**	.42**	—	
4. Academic performance	3.19	0.74	-.46**	.51**	.19*	—

Note. M = mean; SD = standard deviation. Coefficients are Pearson product moment correlations. Listwise deletion was applied.

* p < .05. ** p < .01.

- Horizontal rules only, never vertical lines
- Table number and italic title above the table
 - One triangle only, with dashes on the diagonal
 - Means and standard deviations included as columns
 - Probability notes placed beneath the general note

Writing up a hypothesis test

1

Restate the hypothesis

Hypothesis one stated that there would be no significant relationship between academic stress and academic performance among undergraduates.

2

State the test and its justification

Pearson product moment correlation was used because both variables were measured at interval level and preliminary analyses confirmed linearity, normality and homoscedasticity.

3

Present the result

The analysis revealed a significant negative relationship, $r(191) = -.46, p < .001$.

4

State the decision

The null hypothesis was therefore rejected.

5

Interpret in context

This indicates that undergraduates reporting higher levels of academic stress tended to record lower academic performance, with stress accounting for about 21 per cent of the variance in performance.

Keep interpretation in the results chapter brief and factual. Save the discussion of why the relationship exists for the discussion chapter.

CAUTIONS

Ten errors to avoid in correlation analysis

1 Running the correlation without ever plotting the data

2 Correlating individual items instead of computed scale scores

3 Writing $p = .000$ instead of $p < .001$

4 Putting a leading zero before the decimal point in r

5 Using causal verbs such as caused, led to or influenced

6 Reporting every coefficient in a large matrix as though all were hypothesised

7 Ignoring a curvilinear pattern and reporting r as near zero

8 Deleting outliers without checking the original questionnaire

9 Confusing statistical significance with practical importance

10 Pasting raw SPSS output into the dissertation instead of an APA table

Practice task and what comes next

Practice task

- Open your own file with its computed scale scores
- Run Explore on each variable and record skewness and kurtosis
- Produce a scatterplot with a fit line for your main pair
- Run Analyze, Correlate, Bivariate for that pair
- Then run it again with all your study variables to obtain the matrix
- Write the in text APA sentence for your main hypothesis
- Build the APA correlation table in Word and bring it to the next session

Coming in the next sessions

- Partial correlation, controlling for a third variable
- Simple linear regression and the prediction equation
- Multiple regression: joint and relative contribution
- Independent samples t test and one way ANOVA
- Chi square for categorical relationships
- Mediation and moderation analysis
- Building the complete results chapter in APA format

Questions and discussion

Plot before you compute, check before you conclude, and describe an association rather than a cause.

Assoc. Prof. Enoch Oladunmoye

Department of Applied Psychology

College of Humanities and Social Sciences, Kampala International University

Email: enoch.Oladunmoye@kiu.ac.ug