Statistics as a Tool in Scientific Research: Assessing the Relationship Between 2 Numerical Variables With Correlation

A. What a Correlation Is

Statistical test: Pearson correlation coefficient (r)

Pearson's r is Used For:
Analyzing the strength of the linear relationship between two numerical variables in order to answer research questions where scientists want to know whether or how strongly two variables are related to each other, but they do not have experimental control over those variables and rely on already existing conditions, e.g., Is there a relationship between TV viewing and obesity? Between age and severity of flu symptoms in adults? Between tire pressure and gas mileage? Between the amount of pressure and compression of insulation?

The nature of the research question is about the association between two variables, not about whether one causes the other. Is there an association between X and Y? As X increases, what does Y do? What is the type and strength of linear relationship between X and Y?

Correlation describes and quantifies the systematic linear relation between variables, but correlation ≠ causation.

Pearson's r is Used When: Both variables are numerical (interval or ratio) and the research question is about the type and strength of relation (not about causality)

Other Correlation Coefficients
There are many different correlation coefficients that are used for different types of variables, such as:

- Point biserial (r_{PB}): Used when one variable is nominal and has two levels (e.g., gender [male/female], type of car [gas-powered, hybrid]) and one variable is numerical (e.g., reaction time; miles per gallon)

- Spearman rank order (r_S): Used for ordinal data or numerical data that are not normally distributed or linear

This handout focuses on the most commonly used correlation coefficient (Pearson's r or more commonly known just as r)

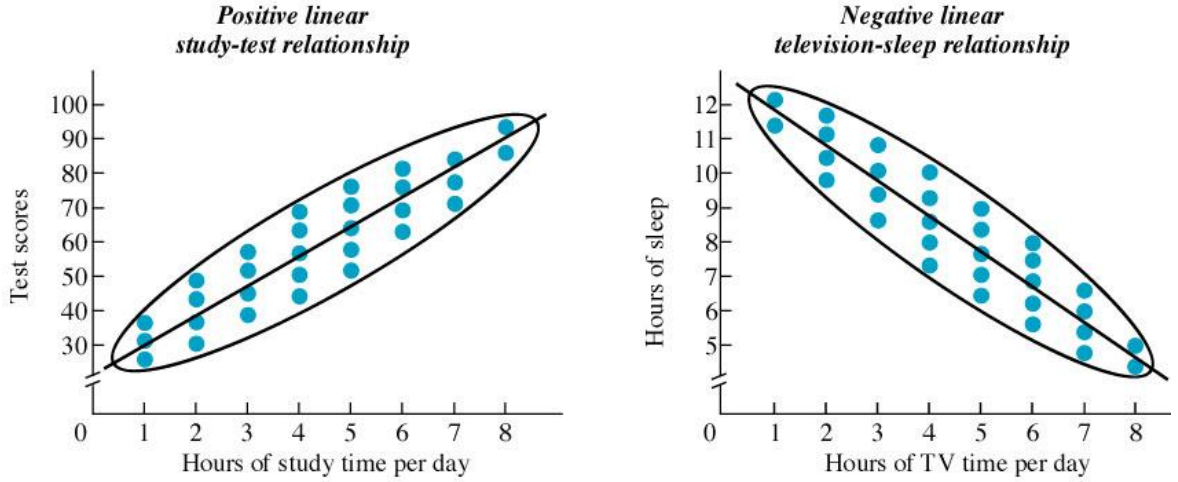
B. What a Correlation Tells You

The value of Pearson's r tells you about the type and strength of the relationship

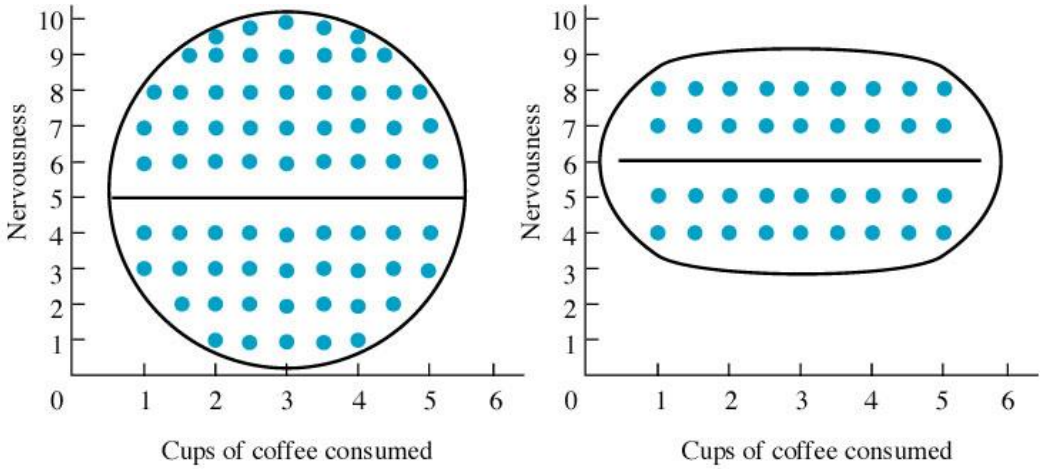
Direction/type:

- *Positive*: As X increased, Y increased; as X decreased, Y decreased

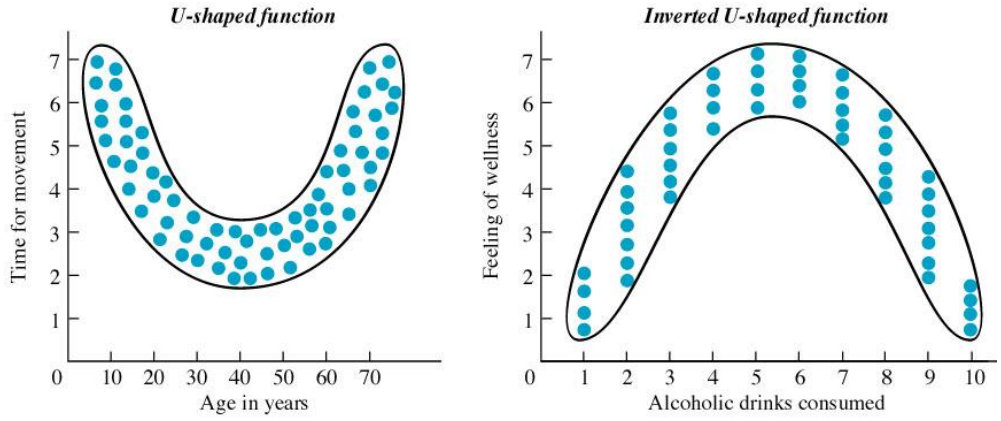
- *Negative*: As X increased, Y decreased; as X decreased, Y increased



- *No systematic relation between X and Y:* High values of X are associated with both high & low values of Y; Low values of X are associated with both high & low values of Y



- *Curvilinear*: Not straight
  U: As X increased, Y decreased up to a point then increased
  Inverted U: As X increased, Y increased up to a point, then decreased



Magnitude/strength: r values range from –1 to +1; the size of the r value tells you about the strength of the relationship. r values close to zero = weak or no correlation; r values closer to +1 or -1 = stronger correlation (the absolute value of r determines its strength; the sign indicates whether it is positive or negative -- that is as X increased, does Y increased (positive) or decreased (negative)?

|r|	**Conclusion About Relationship**
0 to .20	negligible to weak
.20 to .40	weak to moderate
.40 to .60	moderate to strong
.60 to 1.0	strong

C. Hypothesis Testing With Correlations

Correlations are descriptive: You can describe the type and strength of the linear relationship between two variables by examining the r value

But you can also test hypotheses because r values are associated with p values that reflect the probability that the obtained relationship is real or is more likely just due to chance

The key research question in a correlational design is: Is there a real linear relationship between the two variables or is the obtained pattern no different what we would expect just by chance?

- *Null hypothesis* (H_0): The population correlation coefficient is zero; the relationship we see is just due to chance

- *Research hypothesis* (H_A): The population correlation coefficient is not zero; the relationship is most likely real and is not due to chance

D. What Do We Mean by "Just Due to Chance"?

p value = probability of results being due to chance

When the p value is high ($p > .05$), the obtained relationship is probably due to chance: .99 .75 .55 .25 .15 .10 .07

When the p value is low ($p < .05$), the obtained relationship is probably NOT due to chance and more likely reflects a real relationship: .04 .03 .02 .01 .001

In science, a p value of .05 is a conventionally accepted cutoff point for saying when a result is more likely due to chance or more likely due to a real effect

- Not significant = the obtained relationship is probably due to chance; the relationship observed does not appear to really differ from what would be expected based on chance; $p > .05$

- Statistically significant = the obtained relationship is probably NOT due to chance and is likely a real linear relationship between the two variables; $p < .05$

E. Finding Pearson's r Using SPSS

Step 1: Make a scatterplot of the data. If there is a linear trend, go to Step 2
Graphs → Scatter/Dot → Simple Scatter and click "Define". Move the Y and X variables over and click Ok.

When to Calculate Pearson's r: Both variables are numerical (interval or ratio) and the research question is about the type and strength of relation (not about causality)

To run: Get the Pearson Correlation Coefficient (r) and test for significance
Analyze → Correlate → Bivariate; Move the X and Y variables over. Check "Pearson Correlation" and "Two-tailed" test of signficance". Then click Ok.

Output: Computer calculates r and the p value

Correlations

		SurfaceAreaToVolumeRatio	DrugReleaseRate
SurfaceAreaToVolumeRatio	Pearson Correlation	1	.989(**)
	Sig. (2-tailed)		.000
	N	6	6
DrugReleaseRate	Pearson Correlation	.989(**)	1
	Sig. (2-tailed)	.000	
	N	6	6

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

Report as: r = .989, p < .001 (Significant)

F. Reporting Results of Pearson's r Correlation

If the relationship was significant (p < .05)
(a) Say: A [say something about size: weak, moderate, strong] [say direction: positive/negative] correlation was found between [X] and [Y] that was statistically significant, r = .xx, p = .xx

(b) Describe the relation: Thus as X increased, Y [increased/decreased]
Note: If the r value is positive, then as X increased, Y increased; if r is negative, then as X increased, Y decreased

e.g., A strong positive correlation was found between surface area to volume ratio and the drug release

rate that was statistically significant, r = .99, p = .0001. As the surface area to volume ratio increased, the drug release rate increased.

If the relationship was not significant (p > .05)

Say: No statistically significant correlation between [X] and [Y] was found, r = .xx, p = .xx. Thus as X increased, Y neither increased nor decreased systematically.

e.g., No statistically significant correlation between oxide thickness and deposit time was found, r = .002, p = .99. Thus as oxide thickness increased, deposit time neither increased nor decreased systematically.

Other Things to Keep in Mind When Reporting Correlations

- An association does not imply causation! e.g., Average life expectancy and the average number of TVs per household are highly correlated, but you can't increase life expectancy by increasing the number of TVs. They are related, but it isn't a cause-and-effect relationship.

- r is unitless, it is NOT a proportion or percentage. Never read r = .15 as 15%! Say r = "point one five."

- You can only use the word "significant" only when you mean it (i.e., the probability the results are due to chance is less than 5%)

- Do not use the word "significant" with adjectives (i.e., it is a mistake to think one test can be "more significant" or "less significant" than another). "Significant" is a cutoff that is either met or not met -- Just like you are either found guilty or not guilty, pregnant or not pregnant. There are no gradients. Lower p values = less likelihood result is due to chance, not "more significant"

G. What About R^2?

Often in correlational research the value R^2 is reported.

R^2 = Proportion of variability of Y accounted for by X and the linear model

For example, if R^2 = .64 then we can say:

- 64% of the variance in the Y scores can be predicted from Y's (linear) relation with X

- Predictions are 64% more accurate using the linear regression equation to make predictions (Y') than when we use MY to make predictions

R^2 is literally the correlation coefficient, r, times itself.

$R^2 = (r)^2$; $0 \leq R^2 \leq 1$

$r = Sqrt(R^2)$ or $r = -Sqrt(R^2)$; The correlation (r) has the same sign as the slope of the least squares line

Low r $\leftrightarrow$ low R^2 $\leftrightarrow$ little improvement over M_Y

High r $\leftrightarrow$ high R^2 $\leftrightarrow$ more accurate predictions using Y'

H. What is Linear Regression?

Linear regression is used for:
Making predictions of Variable Y based on knowing the value of Variable X

You can predict one variable from the other if there is a strong correlation. Stronger correlation = better prediction

Low or no correlation: Best prediction of Y is the mean of Y (knowing X doesn't add anything)

High correlation: Best prediction of Y is based on knowing X

It is important to make a scatterplot of the data (see above under correlation) to see if a linear model is appropriate and/or the best model!

Terminology
X variable: predictor (independent variable, explanatory variable, what we know)

Y variable: criterion (dependent variable, response variable, what we want to predict)

Regression line: Best fitting straight line that summarizes a linear relation; Comprised of the predicted values of *Y* (denoted by *Y'*)

What Do We Mean by the "Best Fit" Line?

Error or residual = observed Y – predicted Y = Y – Y'

The *least-squares* line is the "best fit" line that minimizes the sum of the square errors/residuals

Used to predict Y using a single predictor X and a linear model

Y' = bX + a
- Y' = predicted y-value
- X = known x-value
- b = slope (ratio of how much Y changes relative to a change in one unit of X; same sign as the correlation)
- a = y-intercept or the point where line crosses y-axis (where X=0)

I. Running Linear Regression Using SPSS

Linear Regression: When you want to predict a numerical variable Y from another numerical variable X

Need: Enter X and Y data in SPSS

To get linear model: Analyze → Regression → Linear; Move the Y variable to Dependent and the X

variable to Independent; Click Ok.

Output: Least squares line

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	7.209	2.584		2.790	.049
	SurfaceAreaToVolumeRatio	35.916	2.697	.989	13.317	.000

a Dependent Variable: DrugReleaseRate

So, the linear model is: Y = 7.209 + 35.916X; Y = Drug Release Rate, X = Surface Area to Volume Ratio

J. Things to Keep in Mind When Using Regression to Make Predictions

- You can't extrapolate outside the given domain of the X's. You can only make predictions in the range of the given data.
- Be careful of the effect of potential outliers
- Association does not mean causation!

K. Hypothesis Test for Simple Linear Regression

Assumed: $Y = \alpha + \beta X + \varepsilon$ where ε is normally distributed with mean 0 and variance σ^2

The F test allows a scientist to determine whether their research hypothesis is supported

Null hypothesis H_0:
- There is not a linear relationship between X and Y
- There is no correlation between X and Y
- $\beta = 0$

Research hypothesis H_A:
- X is linearly related to Y
- The correlation between X and Y is different from 0
- $\beta \neq 0$

Examples:
Null hypothesis: Pain level is not linearly related to weight
Research hypothesis: Pain level is linearly related to weight

Null hypothesis: There is no linear relationship between age and heart rate
Research hypothesis: There is a linear relationship between age and heart rate

Sources of Variation: $SS_{Total} = SS_{Regression} + SS_{Residual}$

$$\Sigma(Y_i - M_Y)^2 = \Sigma(Y_i' - M_Y)^2 + \Sigma(Y_i - Y_i')^2$$

SS_{Total} = Total Sums of Squares = How much do all the individual scores differ from the grand mean

$SS_{Regression}$ = Regression Sums of Squares = How much do all the predicted values differ from the grand mean

$SS_{Residual}$ = Residual Sums of Squares = How much do the individual scores differ from the predicted values

Each F test has certain values for degrees of freedom (df), which is based on the sample size (N) and number of conditions, and the F value will be associated with a particular p value
SPSS calculates these numbers.

Summary Table for Simple Linear Regression

Source	*Sum of Squares (SS)*	*df*	*Mean Square (MS)*	*F*
Regression	$\Sigma(Y_i' - M_Y)^2$	1	$\frac{SS_{Regression}}{df_{Regression}}$	$\frac{MS_{Regression}}{MS_{Residual}}$
Residual (Error)	$\Sigma(Y_i - Y_i')^2$	N - 2	$\frac{SS_{Residual}}{df_{Residual}}$	
Total	$\Sigma(Y_i - M_Y)^2$	N - 1 N= # samples		

To report, use the format: $F(df_{Regression}, df_{Residual}) = x.xx$, p _____.

A test for simple linear regression gives you an F ratio

- The bigger the F value, the <u>less</u> likely the relationship between X and Y is just due to chance
- The bigger the F value, the <u>more</u> likely the relationship between X and Y is not just due to chance and is due to a real relationship
- So <u>big</u> values of F will be associated with small p values that indicate the linear relationship is significant ($p < .05$)
- <u>Little</u> values of F (i.e., close to 1) will be associated with larger p values that indicate the linear relationship is not significant ($p > .05$)

Based on p value, determine whether you have evidence to conclude the relationship was probably real or was probably due to chance: Is the research hypothesis supported?

<u>$p < .05$: Significant</u>

- Reject null hypothesis and support research hypothesis (the relationship was probably real; X is linearly related to Y)

<u>$p > .05$: Not significant</u>

- Retain null hypothesis and reject research hypothesis (any relationship was probably due to chance; X is not linearly related to Y)

L. Running the F Test for Simple Linear Regression in SPSS

Linear Regression: When you want to test whether there is a linear relationship between two numerical variables (X and Y)

Need: Enter X and Y data in SPSS

To get linear model: Analyze → Regression → Linear;
Move the Y variable to Dependent and the X variable to Independent; Click Ok.

Output: F and p-value. (Sig = p-value)

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	718.624	1	718.624	177.335	.000(a)
	Residual	16.209	4	4.052		
	Total	734.833	5			

a Predictors: (Constant), SurfaceAreaToVolume
b Dependent Variable: DrugReleaseRate

Report as: $F(1,4) = 177.34$, $p < .001$ (Significant)

M. Reporting the Results of the F Test for Simple Linear Regression in SPSS

Step 1: Write a sentence that clearly indicates what statistical analysis you used
An F test was used to determine if there was a linear relationship between [X] and [Y].
Or, an F test was used to determine if [X] is linearly related to [Y].

Examples:
An F test was used to determine if pain level is linearly related to weight.
An F test was used to determine if there is a linear relationship between age and heart rate.

Step 2: Report whether the linear relationship was significant or not
The linear relationship between [X] and [Y] was significant [or not significant], $F(df_{Regression}, df_{Residual})$ = X.XX [fill in F], p = xxxx.

Examples:
The linear relationship between pain level and weight was significant, $F(2, 40) = 7.31$, $p = .01$.
The linear relationship between age and heart rate was not significant, $F(2, 120) = 2.35$, $p = .10$.

N. Assumptions of the F Test for Simple Linear Regression

- Linearity
  Is a linear model appropriate? Is a linear model the best model?

Look at a scatterplot to look for an overall trend.

- Randomness and Normality

  Are the residuals independent and normally distributed?
  The points should be randomly scattered about the regression line. There should be no patterns in the residual plot.

- Homoscedasticity

  Is the variability about the regression line constant?
  Does the model fit consistently well for all X's?