Correlational Designs

Correlational Research

- Correlational research is used to describe the relationship between two or more naturally occurring variables.
  - Is age related to political conservativism?
  - Are highly extraverted people less afraid of rejection than less extraverted people?
  - Is depression correlated with hypochondriasis?
  - Is I.Q. related to reaction time?

Why Use a Correlational Design?

- Some factors are impossible to manipulate experimentally
  - Personality
  - Demographic categories
- It is unethical to manipulate some variables
  - Severe illness
  - Brain injury
- Most variables that cannot be studied experimentally can be studied correlationally
  - Variables are measured
  - Relationship among variables is assessed

A Note on Terminology

In correlational research

- The terms *predictor variable* and *criterion/outcome variable* are used to describe the variables
  - The terms IV and DV may be used but do not have the same meaning as when used in true experiments
    - In correlational research, independent variable is not manipulated
    - There is no presumption that dependent variable "depends on" the independent variable, only that a relationship exists
  - Therefore, one cannot draw causal conclusions from correlational research

CORRELATIONAL RESEARCH STUDIES

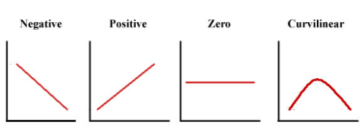
- Do imply that variables share something in common
- Do not imply a cause-and-effect relationship
  - But can show that one variable *does not* cause another if they do not co-vary
- 3rd Variable Problem
- Directionality Problem

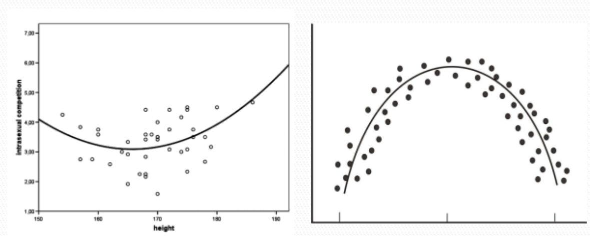
Correlational Hypotheses

- Nondirectional Hypothesis – predicts that two variables will be correlated but does not specify whether the correlation will be positive or negative
- Directional Hypothesis – predicts the direction of the correlation (i.e., positive or negative)

Assumptions of Correlational Statistics: Linearity

- Correlational analysis assumes that the relationship between the independent and dependent variables is linear
  - Can be graphed as a straight line
- Always plot before analyzing





If relationship between variables is nonlinear, correlation coefficient will be misleading
Curvilinear relationships have a correlational coefficient of zero

CORRELATION COEFFICIENT

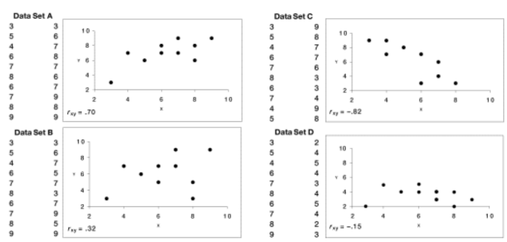
- Expresses degree of linear relatedness between two variables
- Varies between –1.00 and +1.00
- Strength of relationship is
  - Indicated by absolute value of coefficient
  - Stronger as shared variance increases
  - When r = .00, the variables are not related.
  - A correlation of .78 indicates that the variables are more strongly related than does a correlation of .30.
  - Magnitude is unrelated to the sign of r; two variables with a correlation of .78 are just as strongly related as two variables with a correlation of –.78.

TWO TYPES OF CORRELATION

If X…	And Y…	The correlation is	Example
Increases in value	Increases in value	Positive or direct	The taller one gets (X), the more one weighs (Y).
Decreases in value	Decreases in value	Positive or direct	The fewer mistakes one makes (X), the fewer hours of remedial work (Y) one participates in.
Increases in value	Decreases in value	Negative or inverse	The better one behaves (X), the fewer in-class suspensions (Y) one has.
Decreases in value	Increases in value	Negative or inverse	The less time one spends studying (X), the more errors one makes on the test (Y).

Indices of Correlation

- **Pearson correlation coefficient** (r) is the most commonly used measure of correlation
  - r_{xy} = correlation between variables x and y
- **Spearman rank-order correlation** –used when variables are measured on an ordinal scale (the numbers reflect the rank ordering of participants on some attribute)
- **Phi coefficient** – used when both variables are dichotomous
- **Point-biserial correlation** – used when only one of the variables is dichotomous

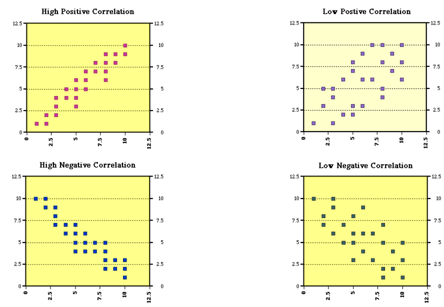


- When points are closer to a straight line, the correlation becomes stronger
- As slope of line approaches 45°, correlation becomes stronger

Factors Affecting the Correlation Coefficient

Outliers: Extreme scores
- Usually defined as scores more than three standard deviations above/below mean
- Can artificially lower a correlation
- If outliers are present, the researcher can:
  - mathematically transform the data
  - omit outliers
- Which option to choose depends on the probable meaning of the outliers

Scatterplots

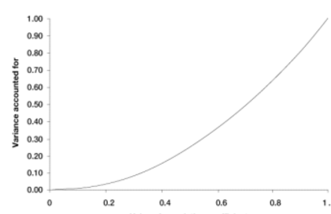


Coefficient of Determination

- Is the square of the correlation coefficient
- Indicates the proportion of variance in one variable that is accounted for by another variable.
  - The correlation between children's and parents' neuroticism scores is .25. If we square this correlation (.0625), the coefficient of determination tells us that 6.25% of the variance in children's neuroticism scores can be accounted for by their parent's scores.

RELATIONSHIP BETWEEN CORRELATION COEFFICIENT AND COEFFICIENT OF DETERMINATION

If r_{xy} Is	And r_{xy}^2 Is	Then the Change From	Is
0.1	0.01		
0.2	0.04	.1 to .2	3%
0.3	0.09	.2 to .3	5%
0.4	0.16	.3 to .4	7%
0.5	0.25	.4 to .5	9%
0.6	0.36	.5 to .6	11%
0.7	0.49	.6 to .7	13%
0.8	0.64	.7 to .8	15%
0.9	0.81	.8 to .9	17%



- The increase in the proportion of variance explained is not linear

INTERPRETING THE PEARSON CORRELATION COEFFICIENT

- Coefficient of alienation
  - 1 – coefficient of determination
  - Proportion of variance in one variable unexplained by variance in the other

INTERPRETING THE PEARSON CORRELATION COEFFICIENT

- "Eyeball" method

Correlations between	Are said to be
± .8 and 1.0	Very strong
± .6 and .8	Strong
± .4 and .6	Moderate
± .2 and .4	Weak
± 0 and .2	Very weak

Statistical Significance of r

- A correlation coefficient is statistically significant when the correlation calculated on a sample has a very low probability of being .00 in the population from which the sample came.

Statistical Significance of r is Affected by Three Things

1. Sample size
2. Magnitude of the correlation
3. How careful you want to be not to draw an inaccurate conclusion about whether the correlation is .00

Factors That Distort Correlation Coefficients

Reliability of a Measure -- the less reliable a measure is, the lower its correlations with other measures will be.

- If the true correlation between neuroticism in children and in their parents is .40, but you use a scale that is unreliable, the obtained correlation will not be .40 but rather near .00.
- Example:
  - True score correlation = 0.40
  - Reliability of two measures = 0.75
  - Maximum possible observed correlation = 0.30

Factors Affecting the Correlation Coefficient

Restriction in range: Occurs when the scores of one or both variables in a sample have a range of values that is less than the range of scores in the population

- Reduces the observed correlation
- As a variable's range narrows, the variable comes closer to being a constant
  - Correlation between a constant and a variable is zero
  - As a variable's sampled range becomes smaller, its maximum possible correlation with another variable becomes smaller

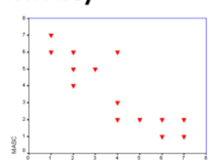
Hypothetical Self-Reported Masculinity and Femininity

Gender	Masc	Fem	Gender	Masc	Fem
Male	7	1	Female	1	6
Male	6	4	Female	2	5
Male	5	3	Female	2	6
Male	6	2	Female	2	7
Male	5	2	Female	1	7
Male	4	2	Female	2	5
Male	6	1	Female	3	4
Male	7	1	Female	2	4
Male	5	2	Female	2	5
Male	5	3	Female	3	4

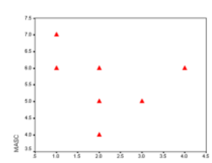
Ratings on 7-point scales where 1 = not at all and 7 = very

Hypothetical Self-Reported Masculinity and Femininity

- With all participants included, ratings of masculinity and femininity are negatively correlated ($r = -0.88$)
- With only male participants, $r = -0.42$
- Range is restricted on both measures when only male participants are assessed



Male and Female Participants



Male Participants Only

Factors Affecting the Correlation Coefficient

Subgroup differences: The participant sample on which a correlation is based contains two or more subgroups

- The combined group correlation will *not* accurately reflect the subgroup correlations if
  - the correlation differs within the subgroups

  or
  - the mean scores differ within the subgroups

Hypothetical Ratings of Masculinity and Femininity

	Males			Females		
	Mean	SD	r_{MF}	Mean	SD	r_{MF}
Masc	5.6	0.97	−0.42	2.1	0.99	−0.72
Fem	2.0	0.66		5.3	1.16	

- The means and standard deviations differ for male and female participants
- The correlation between masculinity and femininity is stronger for female participants

Multifaceted Constructs

Constructs that are composed of two or more subordinate components

- Each component can be distinguished from the others and measured separately
- Components are distinguishable even though they are related to each other both logically and empirically
- E.g., 3 Factor Model of Narcissism

Multidimensional Constructs

- It is important to determine whether a construct is multifaceted or multidimensional
  - Components of multidimensional constructs are not related to one another
    - E.g., depression – activity, sleep, mood
- Facets should *not* be combined if:
  - they are theoretically or empirically related to different DVs or different facets of a DV
  - the theory of the construct predicts an interaction among the facets

Combining Facets

Facets can be combined in some circumstances, such as

- when the researcher is interested in the latent variable represented by the facets
  - e.g., Narcissism, Dark Triad
- And when the latent variable is:
  - more important
  - more interesting
  - represents a more appropriate level of abstraction
  - a better predictor of the DV

Guidelines for Correlational Research

- Use only the most reliable measures available
- Look for restricted range
  - Check the ranges of the scores for your sample against published norms
- Plot the scores for the subgroups and the combined group before computing r
- Compute subgroup correlations and means
- When using multifaceted constructs, avoid combining facets unless there is a good reason to do so

Bivariate Regression

- r is an index of the relationship between two variables
  - indicates the accuracy with which scores on one variable can predict the other
- Prediction is assessed by bivariate regression
  - An equation is developed to predict one variable (Y) from the other (X)

    $Y = a + bX$, where
- a is the intercept
  - the value of Y when X is zero
- b is the slope
  - the amount of change in Y for each unit change in X

Partial Correlation Analysis

- Examines the extent to which the correlation between two variables (X and Y) can be accounted for by their mutual correlation with an extraneous variable (Z)
  - That is, Z is correlated with both X and Y
- Tests what the correlation of X and Y would be if Z were not also correlated with them (e.g., all participants had the same score on Z)
  - pr represents the strength of the relationship when the effect of Z is removed or held constant

Zero-Order Correlations		
	Masculinity	Self-Esteem
Depression	−0.26*	−0.52*
Self-esteem	0.67*	
Partial Correlations		
Correlation of	Controlling for	
Masculinity and Depression	Self-esteem	0.14
Masculinity and Self-esteem	Depression	0.64*
Depression and Self-esteem	Masculinity	−0.48*
Source: Feather, 1985		*$p<0.001$

Feather's (1985) study of the correlation between masculinity and depression if self-esteem is controlled

Partial Correlation Analysis

- Results show the correlation between
  - masculinity and depression becomes nonsignificant when self-esteem is controlled
  - self-esteem and depression is virtually unaffected when masculinity is controlled
- Suggests the relationship between masculinity and depression can be accounted for by masculinity's correlation with self-esteem
  - The masculinity-depression relationship is spurious

Multiple Regression Analysis (MRA)

- Extends simple and partial correlations to situations in which there are more than two IVs
- Used for two purposes
  1. To derive an equation that predicts scores on some criterion variable from a set of predictor variables
  2. To explain variation in a DV in terms of its degree of association with members of a set of IVs

Simultaneous MRA

Purpose is to derive the equation that most accurately predicts a criterion variable from a set of predictor variables

- Uses all predictors in a set
- Not designed to determine which predictor does the best job
- Instead, used to determine the *best* predictive equation using an entire set of predictors

Hierarchical MRA

- Similar to partial correlation analysis
- Allows as many variables to be partialed as the investigator needs
- Researcher creates a regression equation by entering variables into the equation
- Allows investigator to test hypotheses about relationships between predictor variables and a criterion variable with other variables controlled

Information Provided by MRA

- Multiple correlation coefficient (*R*): An index of the degree of association between the predictor variables as a set and the criterion variable
  - Provides no information about the relationship of any one predictor variable to the criterion variable
  - R^2 represents the proportion of variance in the criterion variable accounted for by its relationship with the *total set* of predictors

Information Provided by MRA

- Regression coefficient: The value by which the score on a predictor variable is multiplied to predict the score on the criterion variable
- Represents the amount of change in *Y* brought about by a change in *X*

Information Provided by MRA

Regression coefficients can be standardized (β) or unstandardized (*B*)

- If standardized:
  - βs for all IVs in an analysis are on the same scale
    - Have a mean of 0 and *SD* of 1
  - can be used to compare the degree to which different IVs in an analysis are predictive of the DV

Information Provided by MRA: Regression Coefficients

- If unstandardized:
  - *B*s have same units regardless of the sample
  - coefficients can be used to compare the predictive utility of an IV across samples
  - *t*-tests can be used to determine whether these comparisons are statistically significant

Information Provided by MRA

Change in R^2

- used in hierarchical MRA
- represents the increase in the proportion of variance in the DV that is accounted for by adding another IV to the regression equation
- addresses whether adding that IV helps predict Y better than the equation without that IV
- can fluctuate as a function of the order in which the variable is entered into the equation
- entering a variable earlier will generally result in larger change in R^2 than entering it later
  - Especially likely if variable has high correlation with other predictor variables

The Problem of Multicollinearity

Multicollinearity

- is a condition that arises when two or more predictor variables are highly correlated with each other
- can adversely affect results of MRA
- researchers must check to see if it is affecting their data

Effects of Multicollinearity

- Can inflate the standard errors of regression coefficients
  - Can lead to nonsignificant statistical outcomes
  - Researcher may erroneously conclude that the criterion and predictor variables are unrelated
- Can lead to misleading conclusions about changes in R^2

Causes of Multicollinearity

- Including multiple measures of one construct in set of predictor variables
  - If using multiple measures, better to use a latent variables analysis
- Using variables that are naturally correlated
- Using measures of conceptually different constructs that are highly correlated
- Sampling error

Detecting Multicollinearity

- Create a correlation matrix of predictor variables *before* conducting MRA
  - Look for correlations ≥ 0.80
- Examine the pattern of correlations among several predictors
- Compute the *variance inflation factor (VIF)*
  - Look for *VIF*s ≥ 10
  - $VIF = 1 / (1 - R^2)$

Dealing with Multicollinearity

- When planning a study, avoid including redundant variables
- Combine multiple measures of a construct into indexes
- If measures of conceptually-different constructs are highly correlated, use measures that show the lowest *r*

Dealing with Multicollinearity

- When source of multicollinearity is sampling error, collect more data to reduce error
- Delete IVs that might be source of the problem
  - Consider whether this results in valuable information loss
- Conduct a factor analysis to empirically determine which variables to combine in an index

MRA as Alternative to ANOVA

To use continuous IVs in ANOVA, they must be transformed into categorical variables

- Usually done with median split
  - People scoring above median classified as "high"
  - People scoring below median classified as "low"
- However, doing so creates conceptual, empirical, and statistical problems

Problems are avoided by using MRA

- IV can be treated as continuous rather than categorical
- For categorical variables use dummy coding
  - Assign values to experimental and control conditions

MRA as Alternative to ANOVA

- In ANOVA, assumption is that IVs are uncorrelated
  - Not an assumption of MRA
- If IVs are correlated, use MRA

Logistic Regression Analysis

- Used when DV is categorical
- Has same purpose as MRA, but does *not* assume that
  - variables are normally distributed
  - the relationship between the IVs and DVs are linear

Logistic Regression Analysis

- Produces an *odds ratio* (*OR*)
  - Describes the likelihood that a research participant is a member of one category rather than the others
  - *OR* of 1 indicates that scores are unrelated to membership in the DV categories
  - *OR* > 1 indicates that high scoring participants are more likely to be in a target group
  - *OR* < 1 indicates that high scoring participants are more likely to be in the other group

Multiway Frequency Analysis

- Allows a researcher to examine the pattern of relationships among a set of nominal level variables
  - Most familiar example: chi-square test of association

Multiway Frequency Analysis

- Loglinear analysis extends the principles of chi-square to situations in which there are more than two variables
- Logit analysis is used when one of the variables in loglinear analysis is considered to be the IV
  - Analogous to ANOVA for nominal level DVs
  - Allows tests of main effects and interactions for IVs

Data Types and Data Analysis

	Independent Variable	
Dependent Variable	*Categorical*	*Continuous*
Categorical	Chi-square Analysis	Logistic Regression Analysis
	Loglinear Analysis	
	Logit Analysis	
Continuous	Analysis of Variance	Multiple Regression Analysis