What is “Correlation”?

The Pearson r “correlation coefficient” is a summary statistic that indicates both the strength and direction of the relationship between two variables

It has a value of between -1 and +1

- Values less than zero (e.g -0.8) indicate a negative correlation

- Values greater than zero (e.g. +0.8) indicate a positive correlation

Examples of Correlational Research Hypotheses

- The number of outpatient therapy sessions utilized is positively correlated with the extent of depression as measured by the total score of the Beck Depression Inventory
- The 242 final exam score is positively correlated with the number of hours students spend preparing for the exam
- The number of behavioral incidents by children in residential care is negatively correlated with the number of strength-based supportive comments from staff

Compare to other statistical hypotheses…

- Chi-Square: “ The variables are associated…”
- t-test: “ The group means differ…”
- One-way ANOVA: “ The group means differ…”
- Correlation: “ There is a positive [or negative] correlation between the two variables.”
- Multiple Regression: “ The independent variable is predictive of the dependent variable, controlling for additional factors”

Scatter plots (a.k.a. scattergrams)

- Scatter plot: the graphic representation of the relationship between two ratio or interval variables, plotting the value of one variable against another with one dot

- Useful as a preliminary step in visually inspecting data:
  - Seeing strength and direction of relationship
  - Seeing how linearly the variables are related
  - Seeing outliers

Scattergrams:
Examples of Various Correlations

What is Correlation?

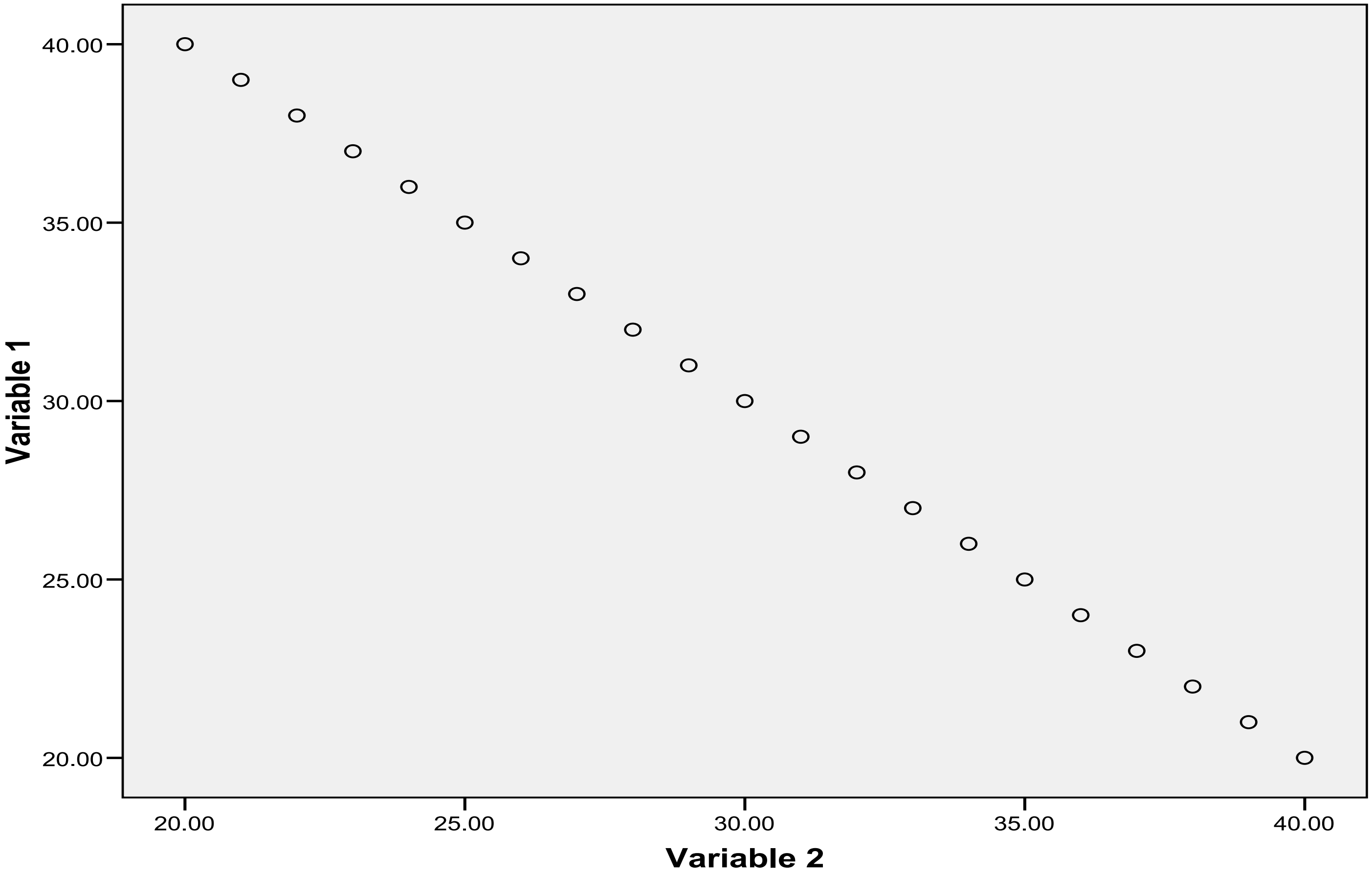
- Correlation tests the relationship between a continuous independent variable and a continuous dependent variable.

- Correlation tests produce an r value and a p value.

- The r value is always between -1 and +1

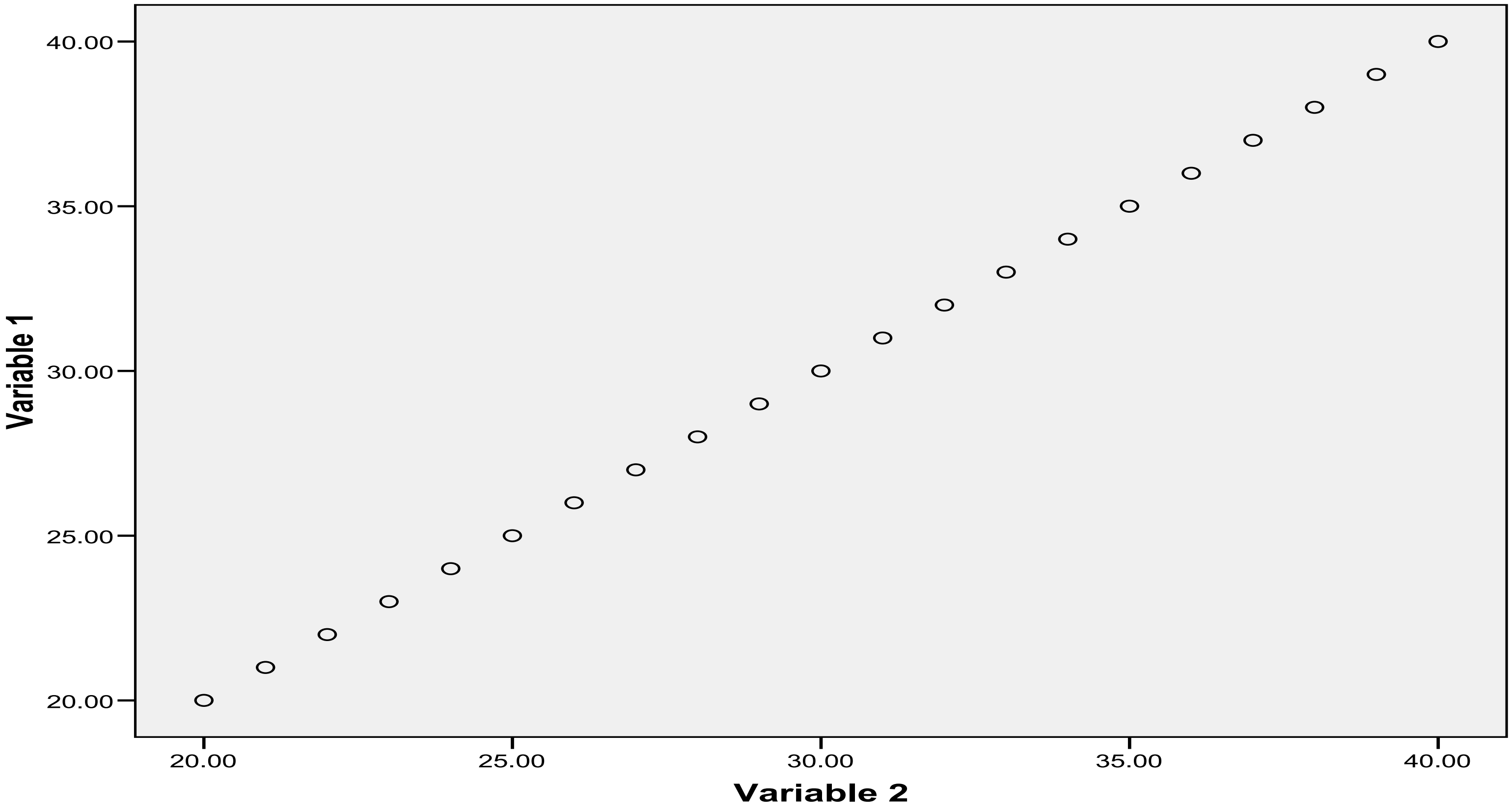
A **negative r value** indicates that as the value of one variable increases, the value of the other variable decreases (referred to as a negative correlation)

Example of a Graph with a Negative Correlation:



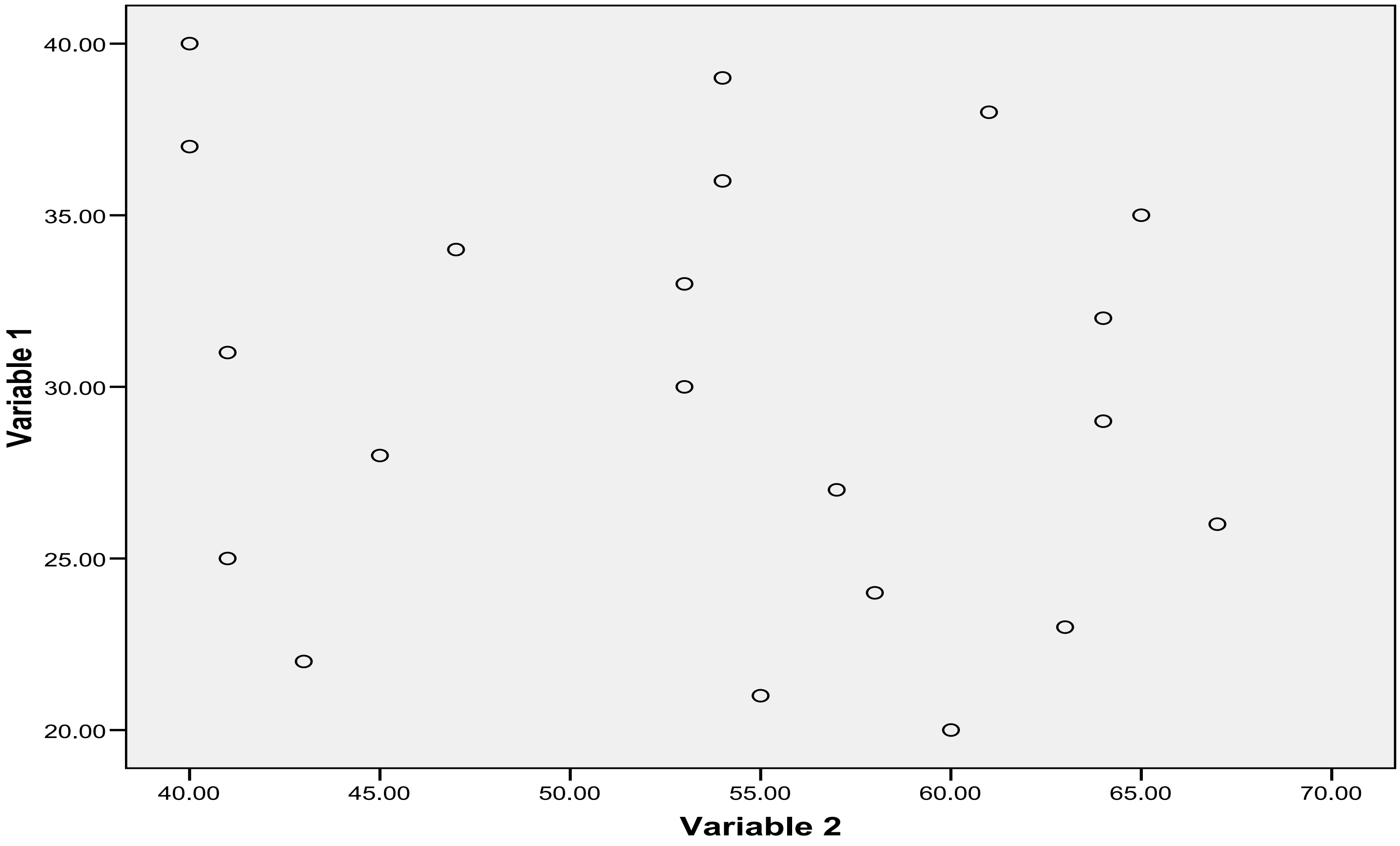
A **positive r value** indicates that as one variable increases, the other variable also increases (referred to as a positive correlation)

Example of a Graph with a Positive Correlation:



An **r value of zero** indicates no relationship between variables

Example of a graph indicating no correlation between variables



What are the Requirements to use r?

- The Pearson r is a *parametric* statistic. Why?
  - The variables should be normally distributed in the population
    - For larger samples there can be some relaxation of this requirement
    - There are non-parametric tests for non-normally distributed variables, and those other than continuous
  - Also, the variables should be related linearly—either positively or negatively
- Why is this important? (Hint: see previous slides)

Inference Testing in Correlation

- The r statistic can be located in a table of critical values
- The logic of inference testing is the same as other statistics:
  - If the p value given by SPSS is equal to or less than the alpha, then we reject the Null Hypothesis
  - We also need to interpret the correlation coefficient (r) and inspect the scatter plot:
    - Is it in the same direction as hypothesized?
    - Does the strength of the correlation support the alternate hypothesis?
    - Are the variables linearly related?

Can We Imply Causality from Correlation?

- Remember the requirements for causality:

1. Time ordering (The IV should precede the DV chronologically)

2. Correlation between variables

3. No other rival hypothesis (effect of 3rd variable)

- What might be missing in the correlation?

--Other confounding variables!

Multivariate Regression Statistics

What are multivariate statistics?

- **Multivariate statistics allow you to determine the impact of an independent variable on a dependent variable while factoring out the influence of potentially confounding (i.e. extraneous) variables.**

Types of multivariate statistics:

- **Bivariate: It predicts the value of a dependent (or outcome) variable from an observed independent (or predictor) variable**

- **Multivariate: It predicts the value of a dependent (or outcome) variable from an observed independent (or predictor) variable,controlling for other variables**

Coding in Multiple Linear Regression and Binomial Logistic Regression:

- **If an independent/control variable is categorical, then dummy coding (AKA creating indicator variables) is necessary. This involves creating a separate variable for each category within the categorical variable and using a "baseline" category to compare categories.**

- **For instance, race/ethnicity is a very common demographic variable that is included in many multivariate statistical tests. We normally think of race/ethnicity as one categorical variable with multiple categories within it (i.e. White, African American, Latino, Asian/Pacific Islander etc…). However, to include race/ethnicity in a multivariate model, we need to use a procedure called dummy coding (AKA creating indicator variables). To do this, the one variable of race/ethnicity is re-coded (in SPSS) into 4 indicator variables:**

  1. **White: Value labels: 0 = Not White 1 = White**
  2. **African American: Value Labels: 0 = Not African American 1 = African American**
  3. **Latino: Value labels: 0 = Not Latino, 1 = Latino**
  4. **Asian/Pacific Islander: Value labels: 0 = Not API, 1 = API**

- One indicator variable is chosen as the “baseline” to which all other racial/ethnic categories are then compared. For instance, if White is chosen as the baseline, then the statistical output provided by SPSS will indicate a comparison between African Americans and Whites, Latinos and Whites, and APIs and Whites with respect to the dependent variable.

Sample size requirements for multivariate statistics

- General rule of thumb is there needs to be at least 10 people in the sample for every independent or control variable included in the model.

What Does "Controlling For" Mean?

- Controlling for a variable (e.g. gender or race) means:

  1. We collect data on that variable

  2. We include that variable in the list of independent variables in our model

  3. The regression analysis separates out the effects of each attribute (male, female)

  4. You can interpret the resulting statistics for all other variables as if by saying "regardless of gender"

- Even though only some variables might be labeled control variables in the hypothesis, multiple regression analysis uses the same process on all independent variables in the model:
  - You can say about any variable, "controlling for the effects of all other variables…"

Typical Uses

- While regression can be used to predict outcomes, the procedure is most often used to:
  - Determine whether the relationship between the IV and DV is likely due to sampling error
  - Determine the strength and direction of the relationship between the primary IV and DV (as in correlation)
  - Determine the effects of other independent variables (such as control variables) in the relationship between the primary IV and DV