# CORRELATION

# RESEARCH PROBLEM

What about if we want to see if a relationship exists between two variables, but have too many categories/attributes within the variable?

What about if those categories were interval-ratio?

# RESEARCH PROBLEM

If interval-ratio, we can be more sophisticated… we ***can say much more than*** how the categories overlap.

- Equal intervals/steps between values means we can talk about degree of relationship between the variables
  - How the two variables move together (up, down, or opposite directions)
  - Can talk about the strength or direction of the association between two variables

# IT'S A CO-RELATION

Moving together

- "Co-relation":
  - The relationship between two interval-ratio variables
- Correlation:
  - Describes strength and direction of relationships in a linear fashion

# LOGIC OF CORRELATION

Examines *how much* two variables move together, *which direction* they're moving

(and the calculation *constrains* that value)

# LOGIC OF CORRELATION

Examines *how much* two variables move together, *which direction* they're moving (and the calculation *constrains* that value)

- Which **direction**?
  - As the X variable increases, does the Y increase or decrease?
- How much: how **strongly** are variables related/moving together?
  - Can we perfectly predict Y from X, or not?
    - For example, is Y = 2X?

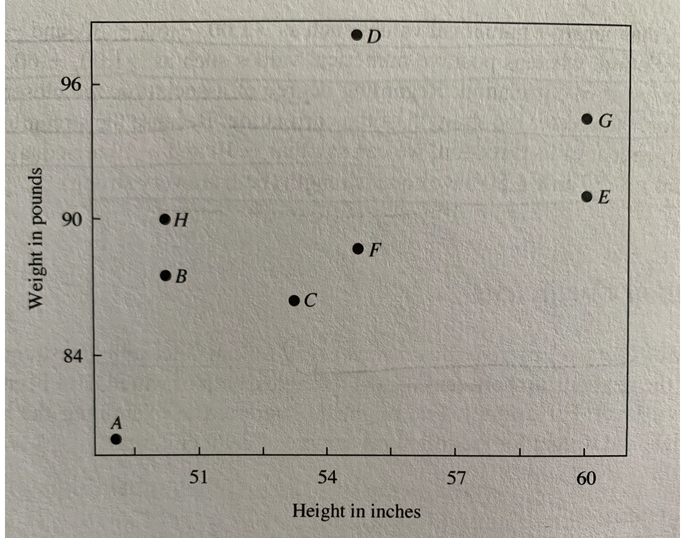



**FIGURE 10.4** ***Scatter Plot of Height and Weight***

*We know that both are associated because the taller a person is, the more they tend to weigh*

# TWO KEY COMPONENTS OF CORRELATION

Strength of relationship

Direction of relationship (linear)

# STRENGTH OF CORRELATION

Correlations vary in strength

Can visualize strength using scatterplot
- Independent variable (predictor) on X axis, dependent variable (outcome) on Y axis
- Easier to call one variable X (IV) and the other Y (DV)

Strength increases as the points on the scatterplot more closely form an imaginary line

# STRENGTH OF CORRELATION

Strong correlation: points are closer to imaginary line

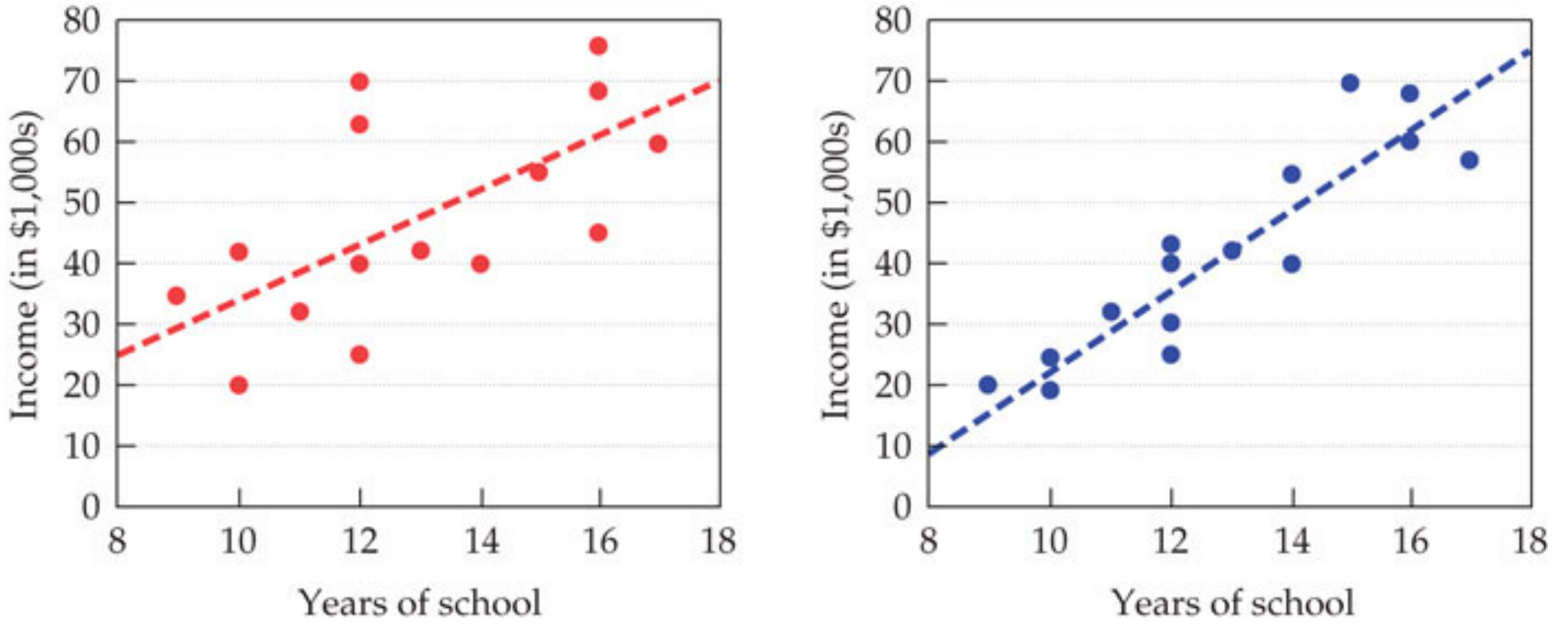
- Perfect correlation: each point falls directly on the imaginary line

Weak Correlation: points are further from imaginary line

- No correlation: no points touch the line

Perfect correlations and no correlations rarely seen in the real world

# STRENGTH OF CORRELATION




**(a)** Males

**(b)** Females

# DIRECTION OF CORRELATION

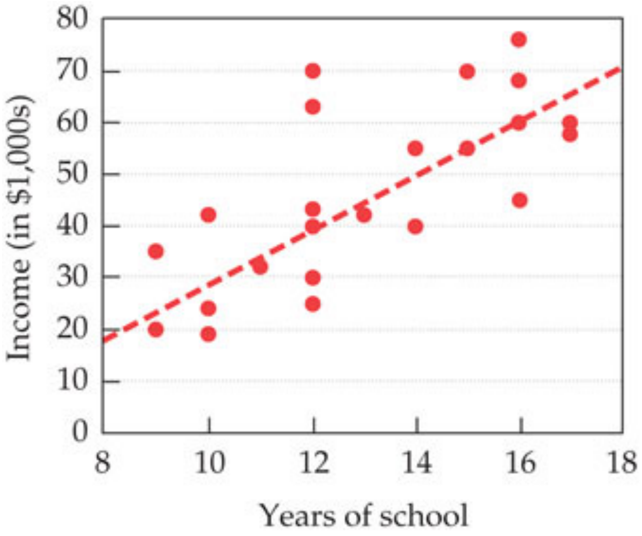
## Positive correlation

- Relationships in the SAME direction
  - As one variable increases, the other variable increases
  - As one variable decreases, the other variable decreases

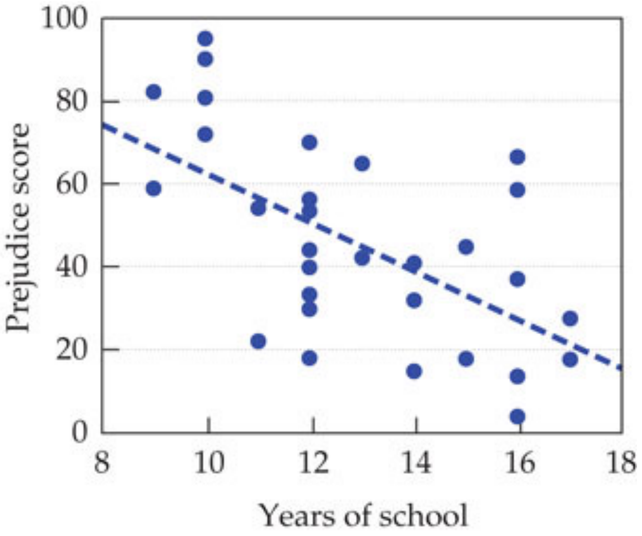
## Negative correlation

- relationships in the OPPOSITE direction; inverse relationship
  - as the score for one variable increases, the other decreases (vice versa)

# DIRECTION OF CORRELATION




**(a)** Years of School and Income




**(b)** Years of School and Prejudice

# A NOTE ON NONLINEAR RELATIONSHIPS

Not all relationships between X and Y form a straight line/are linear

Curvilinear correlation

- one variable increases as the other increases until the relationship reverses itself

# CORRELATION

Pearson's Product-Moment Correlation Coefficient (*r*)

- Examines the *strength* and *direction* of two interval-ratio variables

- Constrained to range from -1.0 to +1.0

# CORRELATION (RESEARCH QUESTION)

Is variation in X related to variation in Y?

# CORRELATION (VARIABLE TYPES)

IV: interval-ratio (e.g. continuous)

DV: interval-ratio (e.g. continuous)

# CORRELATION (ASSUMPTIONS)

1. Linearity
   - Variables move together in a linear fashion.
     - Visual inspection of scatterplot

2. Normality
   - Distribution must be relatively normal
     - Visual inspection of…
       - Histogram
       - Box-and-Whiskers plots
       - Normality (Q-Q) plots

3. Absence of Range Restrictions
   - Values on variables cannot be restricted to small range

4. Absence of Heterogeneous Subsamples
   - Not having groups that have extremely different values (e.g. for which a t-test/ANOVA might appropriately identify)

# CORRELATION (HYPOTHESES)

Null hypothesis ($H_0$)

- *No relationship* between the variables (in the population)
  - $H_0$: $\rho = 0$

Research hypothesis ($H_1$)

- *There is a relationship* between the variables (in the population)
  - $H_1$: $\rho \neq 0$

Rejecting $H_0$ means:

- there is a significant relationship between the X and Y variables

# PEARSON'S CORRELATION COEFFICIENT (*r*)

Strength:

- The closer to ±1.0, the stronger the relationship

Direction:

- Ranges from -1.0 to +1.0
  - Negative: negative correlation
  - Positive: positive correlation

-0.7 and +0.7 have the same strength, but different directions

# CORRELATION STRENGTH CUTOFFS (COHEN 1988)

Weak/Small Correlation

- *r* less than/equal to |.29| ($r \leq |.29|$)

Moderate Correlation

- *r* between |.30| and |.49| ($|.30| \leq r \leq |.49|$)

Strong Correlation

- *r* greater than/equal to |.50| ($r \geq |.50|$)

# CALCULATING THE PEARSON'S CORRELATION COEFFICIENT (*r*)

Relies on concept of <u>covariance</u>: how much, on average, two variables *vary* together

$$cov_{XY} = \frac{\Sigma\,(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

- Uses the product of X and Y deviations from their means
  - Deviation of X ($X$-$\bar{X}$)
    - Example: tells us how much more or less education a person has from the mean education
  - Deviation of Y ($Y$-$\bar{Y}$)
    - Example: tells us how much more or less income a person makes than the mean income

# CALCULATING THE PEARSON'S CORRELATION COEFFICIENT (*r*)

But because $cov_{XY}$ is a function of the *SD* for each variable, we have to constrain it

- e.g. it is highly related to the variability within each variable – is extremely large when variables have large SDs

To constrain, we divide $cov_{XY}$ by the product of X and Y SDs, which is an estimate of how the variability of both variables moves together…

# CALCULATING THE PEARSON'S CORRELATION COEFFICIENT ($r$)

$$r = \frac{cov_{XY}}{SD_X SD_Y}$$

# ANOTHER CALCULATION FOR PEARSON'S *r*

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\ \Sigma(Y-\bar{Y})^2}} = \frac{SP}{\sqrt{SS_X SS_Y}}$$

| X | Y | (X-$\bar{X}$) | (Y-$\bar{Y}$) | (X-$\bar{X}$)(Y-$\bar{Y}$) | (X-$\bar{X}$)$^2$ | (Y-$\bar{Y}$)$^2$ |
|---|---|---|---|---|---|---|
| | | | | | | |

# ANOTHER CALCULATION FOR PEARSON'S *r*

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

# CORRELATION AND THE *r*-DISTRIBUTION

The r-distribution (sort of like normal distribution) has multiple curves

- Each curve based on degrees of freedom (e.g. sample size)
  - Looks more like normal distribution if sample size is large enough ($N \geq 30$)

$$r = \frac{cov_{XY}}{SD_X SD_Y}$$

$$df = N - 2$$

# HYPOTHESIS TESTING (IS THE *r* EXTREME?)

The logic is the same as usual, compare our calculated r (obtained) value to the critical r value (*r* Table)

- If $r_{obtained} \geq r_{critical}$, reject the null hypothesis
- If $r_{obtained} < r_{critical}$, fail to reject the null hypothesis

# HYPOTHESIS TESTING (IS THE *r* EXTREME?)

To find critical *r*, we need alpha ($\alpha$) and degrees of freedom *df*.

- Select the column based on $\boldsymbol{\alpha}$ (usually $\boldsymbol{\alpha} = .05$)
- Select the row based on *df* ($df = N\text{-}2$)
  - Where they intersect is the critical *r* value, $r_{critical}$

If $r_{obtained} \geq r_{critical}$, reject $H_0$

# HYPOTHESIS TESTING (IS THE *r* EXTREME?)

If $|r_{obtained}| \geq |r_{critical}|$, then…

- Relationship between X and Y is so extremely different from 0 (no relationship) that we can't blame it on sampling error, therefore… $H_0$ is probably not true, so…

- $r_{obtained}$ is in rejection region
- Reject $H_0$
- $p \leq \alpha$
- Statistically significant relationship

# HYPOTHESIS TESTING (IS THE *r* EXTREME AS A *t*-TEST)

We can also convert our $r_{obtained}$ test into a $t$-test, and use the $t$-test instead ($t$ Table)

- If $t_{obtained} \geq t_{critical}$, reject the null hypothesis
- If $t_{obtained} < t_{critical}$, fail to reject the null hypothesis

# HYPOTHESIS TESTING (IS THE *r* EXTREME AS A *t*-TEST)

To convert *r* test into a *t*-test, we do the following:

$$t = \frac{r\,\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{r\,\sqrt{df}}{\sqrt{1-r^2}}$$

# HYPOTHESIS TESTING (IS THE *r* EXTREME AS A *t*-TEST)

Then we need to find critical *t*, using alpha ($\alpha$) and degrees of freedom *df*.

- Select the column based on $\alpha$ (usually $\alpha$ = .05)
- Select the row based on *df* ($df = n1 + n2 - 2$; $df = N - 2$)
  - Where they intersect is the critical *t* value, $t_{critical}$

If $t_{obtained} \geq t_{critical}$, reject $H_0$

# HYPOTHESIS TESTING (IS THE *r* EXTREME AS A *t*-TEST)

So, applied to *t*-Test:

- If $|t_{obtained}| \geq |t_{critical}|$, then…
  - Relationship between X and Y is so extremely different from 0 (no relationship) that we can't blame it on sampling error, therefore… $H_0$ is probably not true, so…

  - $t_{obtained}$ is in rejection region
  - Reject $H_0$
  - $p \leq \alpha$
  - Statistically significant difference

# REPORTING R

Report

- The test used
- If you reject or fail to reject the null hypothesis
- The variables used in the analysis
- The degrees of freedom, calculated value of the test, and p-value
  - $r$(df) = $r_{obtained}$, p-value

- "Using the Pearson correlation, I reject/fail to reject the null hypothesis that there is no relationship between the independent variable and the dependent variable, in the population, $r$(?) = ?, $p$ ? .05"

- (if **significant**, follow with…)
  - "We have a [strength] [direction] relationship between [X] and [Y]"

# PEARSON'S *r* AND *r*$^2$ (EFFECT SIZE)

Pearson's product moment correlation coefficient is the basis for regression analysis.

- In correlation we find r, in regression we use r, but we square it to tell us "how much variation in Y is explained by variation in X"
- This is known as effect size ($r^2$), tells us how much X affects Y

$r^2$ tells us how much percent of variation in Y is explained by variation in X

- $r^2$ is a proportion, so convert it to percentage
  - If $r^2 = .159$, that means that 15.9% of the variation in Y is explained by variation in X