Applied Statistics for the Behavioral Sciences

Chapter 6:
Correlation and Regression
Describing Relationships

Outline

- Bivariate distributions
- Types of correlations (+, -, 0)
- The Pearson Product-Moment Correlation Coefficient (r)
- Scatterplots/interpreting r
- Cautions (curvilinear & truncated range data)
- Correlation & regression
- Prediction of a variable
- Regression equations: lines represented as equations

History:Correlation/regression

- Sir Francis Galton
- interested in heredity
- thought psychological characteristics were inherited like physical
- set up an anthropometric lab in London
- invented the concepts of correlation and regression
- describe relationships between variables.
- Karl Pearson, put his ideas into formulae

Bivariate distributions

- using correlation or regression implies bivariate data
- one variable at a time-univariate analysis
- two scores paired somehow
- usual pairing is different scores for same individual
- how one variable varies as a function of the other

4

Types of correlations

- Correlation coefficients have a range of -1 to +1
- When variables are paired, three states of affairs can result
  - As one goes up, the other goes up (positive)
  - One goes up, other goes down (negative)
  - No particular pattern can be identified (0)

5

Positive Correlation

- regression line is the line of best fit
- With a 1.0 correlation, all points fall exactly on the line
- 1.0 correlation does not mean values identical
- the difference between them is identical

6

Negative Correlation

- If r=-1.0 all points fall directly on the regression line
- slopes downward from left to right
- sign of the correlation tells us the direction of relationship
- number tells us the size or magnitude

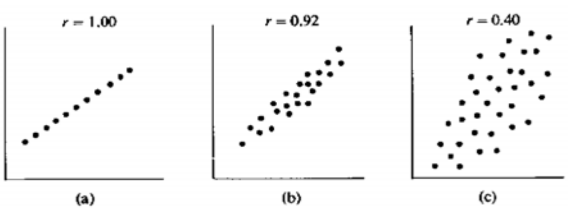
7

Zero correlation

- no relationship between the variables
- a positive or negative correlation gives us predictive power

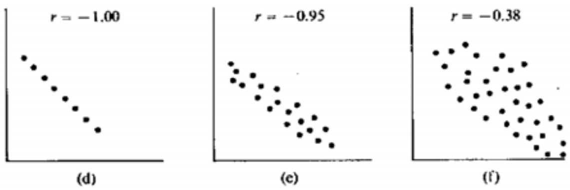
8

Direction and degree



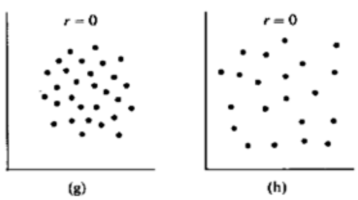
9

Direction and degree (cont.)



10

Direction and degree (cont.)



11

Correlation Coefficient

- r = Pearson Product-Moment Correlation Coefficient
- z_x = z score for variable x
- z_y = z score for variable y
- N = number of paired X-Y values
- Definitional formula (below)

$$r = \frac{\Sigma(z_x z_y)}{N}$$

12

Computational formulas

- "Blanched formula”
- means and standard deviations "cooked out"

$$r = \frac{\frac{\Sigma XY}{N} - (\bar{X})(\bar{Y})}{(S_x)(S_y)}$$

13

Raw score formula

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

14

Interpreting correlation coefficients

- comprehensive description of relationship
- direction and strength
- need adequate number of pairs
  - more than 30 or so
- same for sample or population
- population parameter is Rho (ρ)
- scatterplots and r
- more tightly clustered around line=higher correlation

15

Examples of correlations

- -1.0 negative limit
- -.80 relationship between juvenile street crime and socioeconomic level
- .43 manual dexterity and assembly line performance
- .60 height and weight
- 1.0 positive limit

16

Uses of r

- Reliability
  - test, retest - split half - parallel forms
  - Galton's height measurements reliability of .98
- Correlation as evidence of causation
  - necessary not sufficient condition
  - controlled experiments necessary for definitive evidence of causality

17

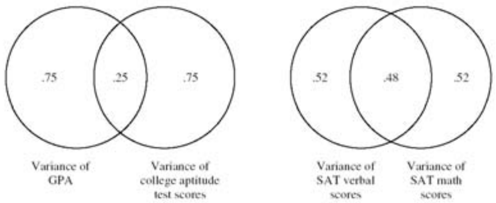
Effect size index

- Cohen's guidelines:
  - Small – r = .10
  - Medium – r = .30
  - Large – r = .50
- Very small correlations can be very important – e.g. physician's health study

19

Coefficient of determination

- r2
- % of shared variance
- between 0-1



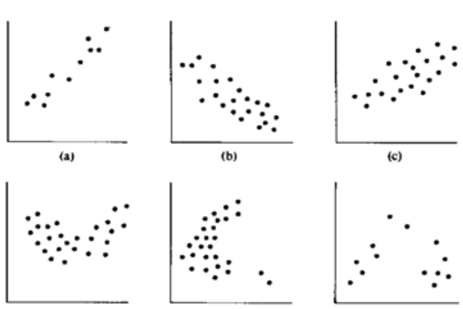
20

Nonlinearity and range restriction

- if relationship doesn't follow a linear pattern Pearson r useless
- r is based on a straight line function
- if variability of one or both variables is restricted the maximum value of r decreases

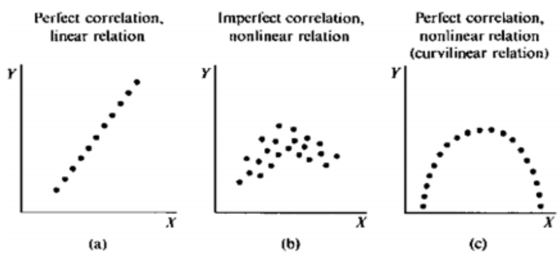
21

Linear vs. curvilinear relationships



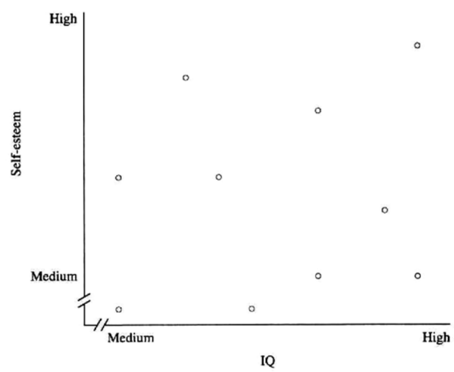
22

Linear vs. curvilinear (cont.)

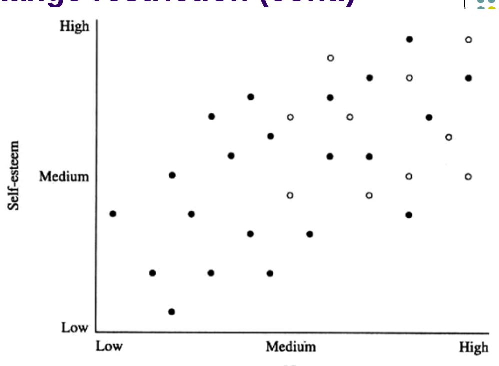


23

Range restriction



Range restriction (cont.)



Understanding r

Given all cases above the median on variable 1, percentage of cases above and below the median on variable 2

True Correlation	Percent Expected on Second Variable[a] Above Median	Below Median
.00	50.0	50.0
.10	53.1	46.9
.20	56.2	43.8
.30	59.5	40.5
.40	63.0	37.0
.50	66.5	33.5
.60	70.3	29.7
.70	74.5	25.5
.80	79.3	20.7
.90	85.3	14.7
1.00	100.0	0.0

26

Origin of regression concept

- Francis Galton studied inheritance of various physical traits (testing some of his cousin Darwin's hypotheses)
- Studying heights of parents and their children
- noted that children of both tall and short parents tended to regress toward the general population mean

27

Origin of regression (cont.)

- tall parents had children who were above average height, but not as tall as they were
- short parents had shorter than average children, but not as short as they were
- dropping back toward general mean was referred to as "the law of filial regression"
- regression came to mean any situation where two variables were studied

28

Moving to prediction

- statistically significant relationship between college entrance exam scores and GPA
- how can we use entrance scores to predict GPA?

- Regression equation:

  $\hat{Y} = bX + a$

  $\hat{Y}$ = predicted value of Y

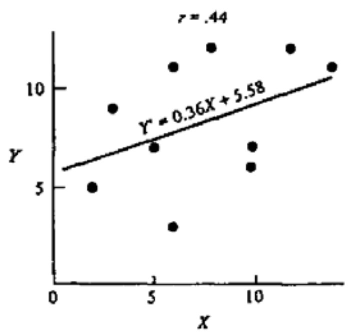
  b = slope of regression line

  a = y intercept

  X = value of X for which Y is being predicted

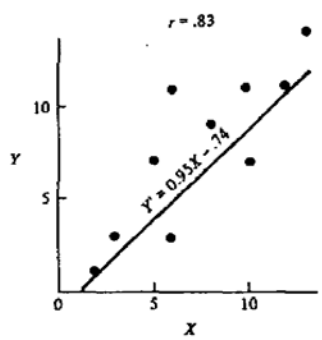
29

Best-fitting line (cont.)



30

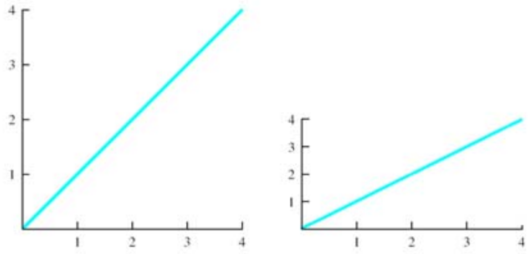
Best-fitting line (cont.)



31

Scaling of axes

- Can distort (misrepresent) relationship



Calculating the slope (b)

- N=number of pairs of scores, rest of the terms are the sums of the X, Y, X^2, Y^2, and XY columns we're already familiar with

$$b = \frac{N(\Sigma XY) - (\Sigma X)(\Sigma Y)}{N(\Sigma X^2) - (\Sigma X)^2}$$

33

Calculating Y-intercept (a)

- b = slope of the regression line
- $\overline{Y} =$ the mean of the Y values
- $\overline{X} =$ the mean of the X values

$$a = \overline{Y} - (b)\overline{X}$$

- Full example problem

34