Correlation

Course: Statistics 1
Lecturer: Dr. Courtney Pindling

Measures of Association

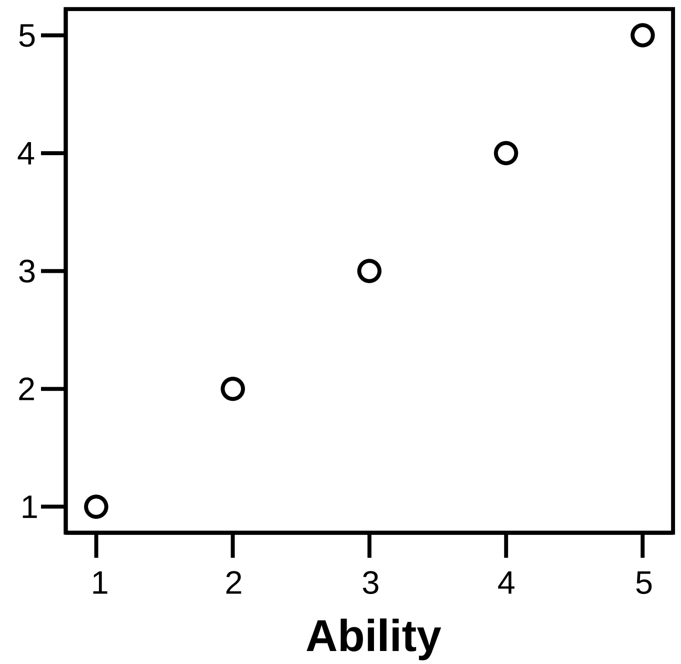
Between Two Variables

Measures of Linear Associations:

- Scatter Plots
- Covariance
- Correlation Coefficient
- Coefficient of Determination

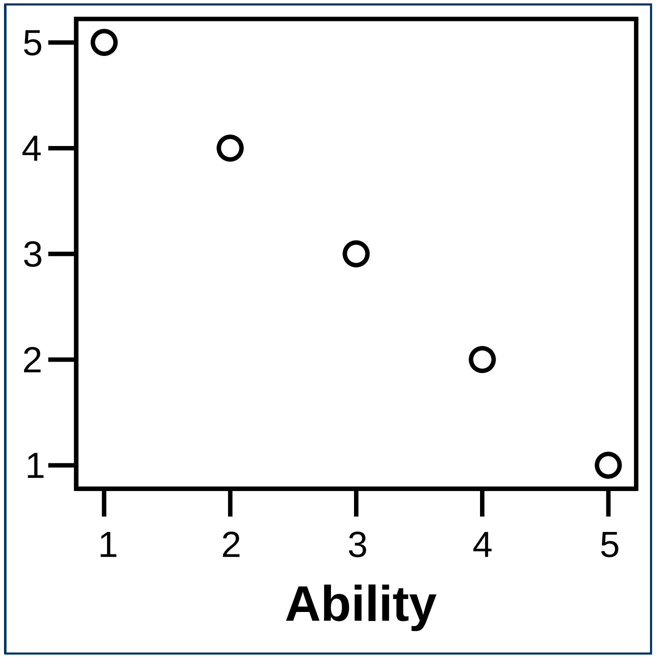
Scatter Plot 1

- Scatter Plot
- Positive linear association
- As the *Ability Index* increase so does value of the y-axis
- Positive correlation



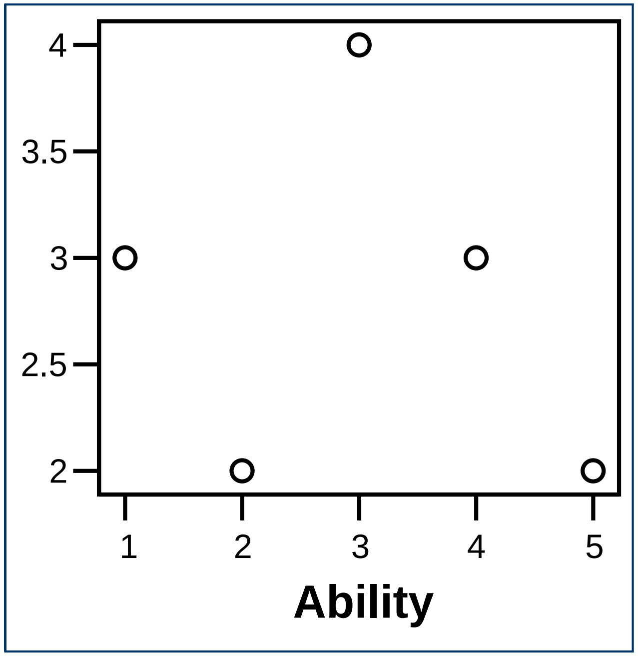
Scatter Plot 2

- Scatter Plot
- Negative linear association
- As the *Ability Index* increase so the value on the y-axis decreases
- Negative correlation



Scatter Plot 3

- Scatter Plot
- No linear association
- As the *Ability Index* increase there seem to be no trend in the value on the y-axis
- No correlation



Covariance

- A measure of the linear association between variables
  - ***Positive*** *indicates positive linear relationship*
  - ***Negative*** *indicates a negative linear relationship*
  - *Values close to* ***zero*** *indicates no linear relationship*
- It is dependent upon the units of measurement for x and y variables
  - Height in inches would give a larger covariance than height in feet; even with same degree of association
  - So the magnitute of the covariance is not significant

Covariance Formula

$$s_{xy} = \frac{\Sigma(X_i - M_x)(Y_i - M_y)}{n-1}$$

Where

X_i is data point i for X variable
Y_i is data point i for Y variable
M_x is mean for X variable
M_y is mean for Y variable
n is sample size

$$s_{xy} = \frac{\Sigma(X_i - M_x)(Y_i - M_y)}{n-1} = \frac{99}{9} = 11$$

Covariance Example

X_i	Y_i	X_i - M_x	Y_i - M_y	$(X_i - M_x)(Y_i - M_y)$
2	49	-1	-1	1
5	56	2	6	12
1	40	-2	-10	20
3	53	0	3	0
4	53	1	3	3
1	37	-2	-13	26
5	62	2	12	24
3	47	0	-3	0
4	58	1	8	8
2	45	-1	-5	5
2	49	-1	-1	1
M_x = 3	***M_y = 50***	***Sum = 0***	***Sum = 0***	***Sum = 99***

Correlation Coefficient

- A measure of the linear association between variables
  - ***Positive*** *indicates positive linear relationship*
  - ***Negative*** *indicates a negative linear relationship*
  - *Values close to* ***zero*** *indicates no linear relationship*
- It not affected by the units of measurement for x and y variables
  - Pearson product moment correlation coefficient or
  - *Sample correlation coefficient*, *r*

Correlation Coefficient Formula 1

$$r_{xy} = r = \frac{s_{xy}}{s_x s_y}$$

Where

r_{xy} = sample correlation coefficient,
s_{xy} = sample covariance,
s_x = sample standard deviation of x, and
s_y = sample standard deviation of y

r from Covariance

- Knowing the covariance and the standard deviations of each variable we can compute the sample correlation coefficient, r
- Covariance = 11, SD_x = 1.49, SD_y = 7.93
- So Pearson ***r*** = 11/(1.49 x 7.93) = **0.93**

Descriptive Statistics

	Mean	Std. Deviation	N
Y	50.0000	7.93025	10
X	3.0000	1.49071	10

Correlation Coefficient Formula 2

Computational Formula

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

Correlation Example

X_i	Y_i	X^2	Y^2	XY
2	49	4	2401	98
5	56	25	3136	280
1	40	1	1600	40
3	53	9	2809	159
4	53	16	2809	212
1	37	1	1369	37
5	62	25	3844	310
3	47	9	2209	141
4	58	16	3364	232
2	45	4	2025	90
$\sum X = 30$	$\sum Y = 500$	$\sum X^2 = 110$	$\sum Y^2 = 25566$	$\sum XY = 1599$

Correlation Example cont.

$$\Sigma X = 30 \qquad \Sigma Y = 500 \qquad \Sigma X^2 = 110 \qquad \Sigma Y^2 = 25566$$

$$\Sigma XY = 1599$$

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} = \frac{10(1599) - (30)(500)}{\sqrt{10(110) - 30^2} \cdot \sqrt{10(25566) - 500^2}}$$

$$r = \frac{990}{1063.96} = 0.93$$

Coefficient of Determination

- Tells us how much of the variation in the dependent variable, *Y* is due to change in the independent variable, *X*
- Coefficient of Determination is r^2
- For example, r^2 = 0.8649
  - *Therefore, 86.49% of the variation in Y is associated with the change in X or*
  - *13.51% of variation is Y is due to other factors*

Properties of r

- Required Interval or Ratio Scales
- Relationship between X and Y must be linear
- Requires pairs of values for X and Y
- The standard deviation about Y for a given value of X is about the same (homogeneity)
- The sample size, N, has little effect on r, but is used to make compute the significance of r

Limitations of *r*

- Correlation does not mean causality
  - *Patients' height may correlate with their blood pressure, but it does not mean that their height is the cause for their blood pressure*
- When *r* is based on sample data, you may get a strong positive or negative correlation purely by chance, even though there is no relationship between the two variables
  - *Patients' shoe size in the hospital may correlates with their blood pressure at time of admission, but there may be no relationship between the two*

Other Correlational Methods

- Pearson *r* is computed on interval and ratio scales
- Spearman *r*, is Pearson *r* computed for ordinal scale
- Other correlational methods based on modified Pearson *r* or probability functions for specific applications

Correlation Methods

Point-biserial r	One dichotomous variable (yes/no; male/female) and one interval or ratio variable
Biserial r	One variable forced into a dichotomy (grade distribution dichotomized to “pass” and “fail”) and one interval or ratio variable
Phi coefficient	Both variables are dichotomous on a nominal scale (male/female vs. high school graduate/dropout)
Tetrachoric r	Both variables are dichotomous with underlying normal distributions (pass/fail on a test vs. tall/short in height)
Correlation ratio	There is a curvilinear rather than linear relationship between the variables (also called the eta coefficient)
Partial correlation	The relationship between two variables is influenced by a third variable (e.g., mental age and height, which is influenced by chronological age)
Multiple R	The maximum correlation between a dependent variable and a combination of independent variables (a college freshman’s GPA as predicted by his high school grades in Math, chemistry, history, and English)