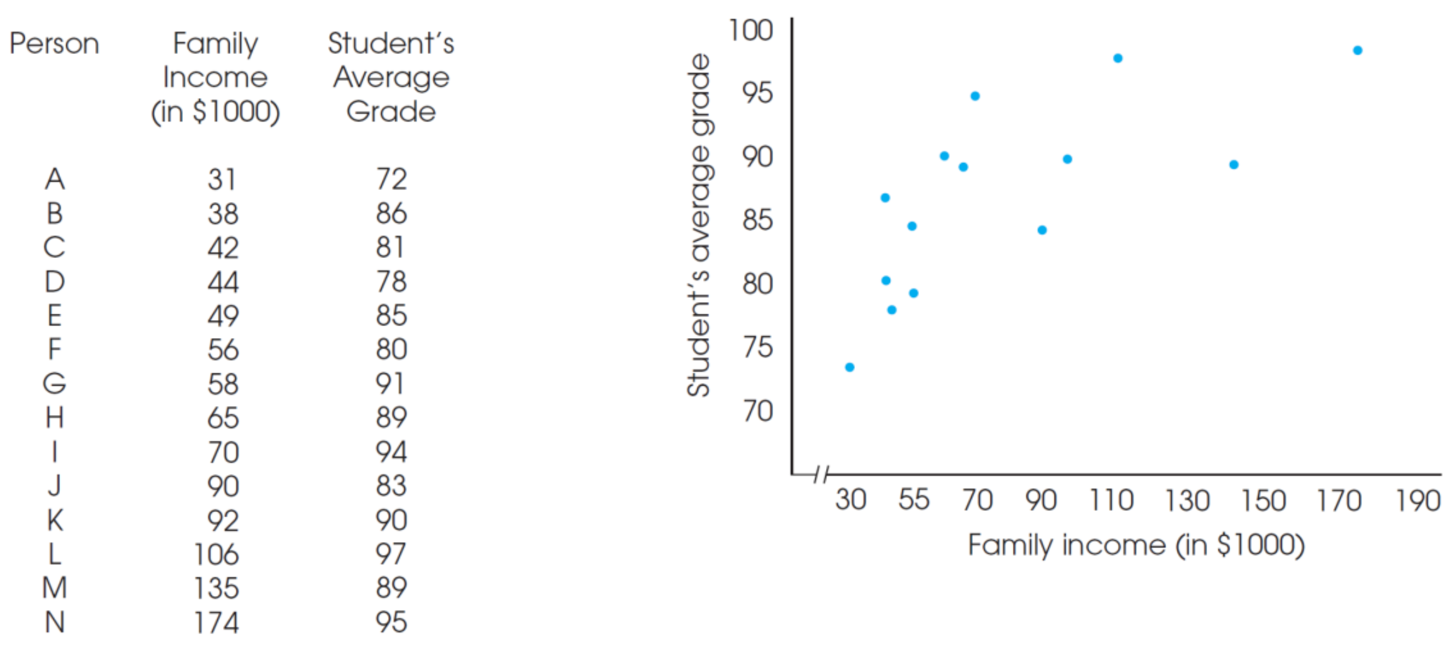
Lecture 13: Correlations

Outline

Correlation

- Correlation coefficient
  - Pearson correlation
  - Spearman correlation
  - Point-biserial and Phi coefficient
- One-sample testing for correlation coefficient

Example
family income & average grade



Person	Family Income (in $1000)	Student's Average Grade
A	31	72
B	38	86
C	42	81
D	44	78
E	49	85
F	56	80
G	58	91
H	65	89
I	70	94
J	90	83
K	92	90
L	106	97
M	135	89
N	174	95

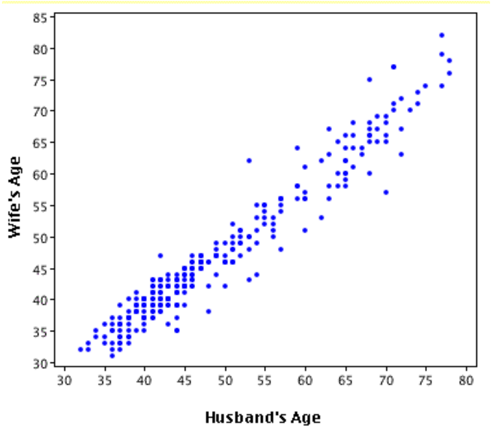
FIGURE 14.1

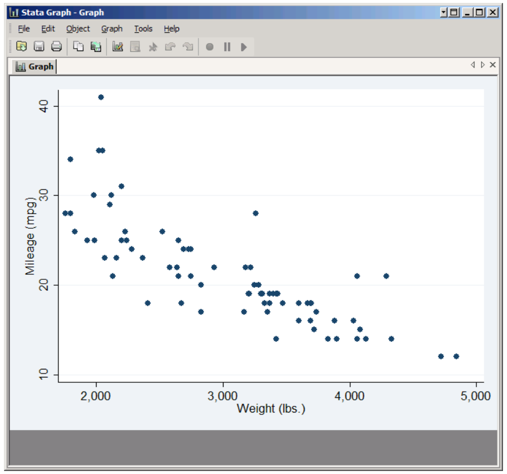
Correlational data showing the relationship between family income (X) and student grades (Y) for a sample of $n = 14$ high school students. The scores are listed in order from lowest to highest family income and are shown in a scatter plot.

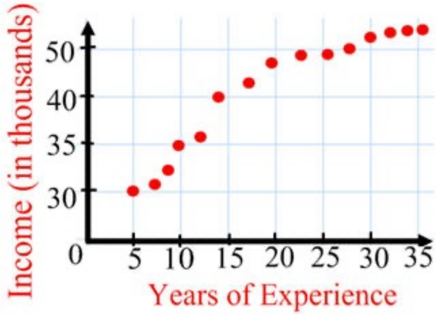
Association of two variables

■ How to compare variables ...

From a table of measurements (e.g. Family income & GPA) we construct a **scatterplot** to visualize the relation







Correlation

- Correlation is a statistical technique to measure the linear relationship between two variables in one sample.

E.g.: Height and Weight, Age and Systolic BP, the time needed to finish an exam and the grade, ...

- Properties:

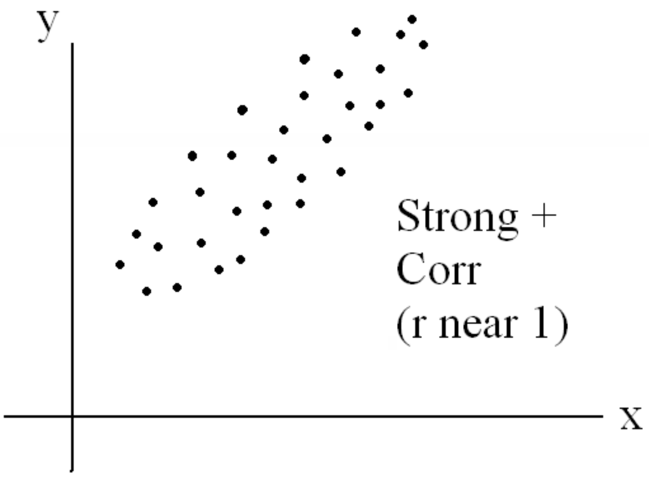
- Location and scale-free
- Quantitative measure, range: [-1,1]
- Negatively correlated/uncorrelated/positively correlated
- Only measures the degree of linear association.
- -1 or 1: deterministic linear relationship

Correlations: Measuring and Describing Relationships

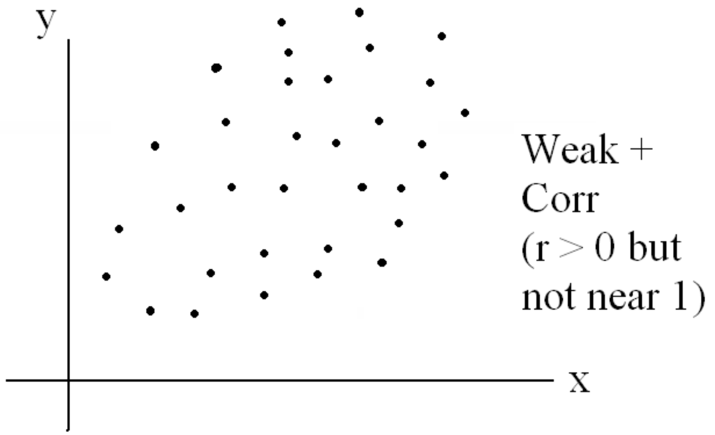
- **We consider three characteristics of the relationship**:

1. the direction of the relationship (positive, negative;正相关，负相关)
2. the form of the relationship (linear, quadratic, ..)
3. the degree of the relationship (magnitude，相关强度)

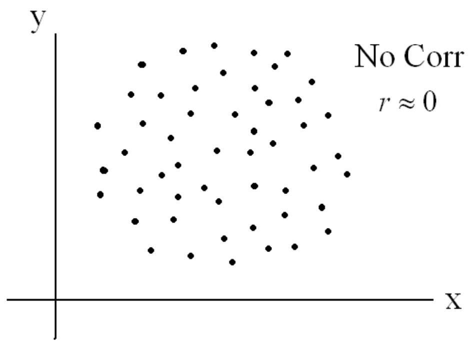
Direction of the relationship: negatively /uncorrelated/positively



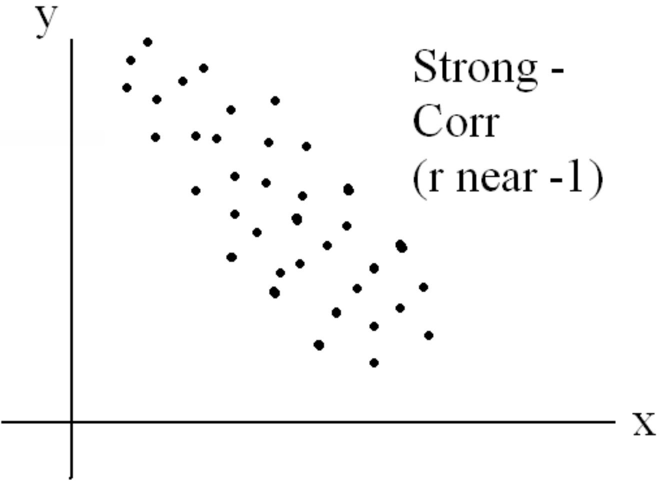
Strong + Corr (r near 1)



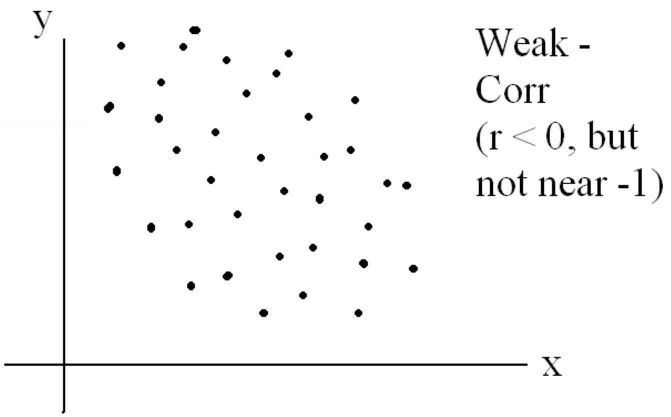
Weak + Corr ($r > 0$ but not near 1)



No Corr $r \approx 0$



Strong - Corr (r near -1)



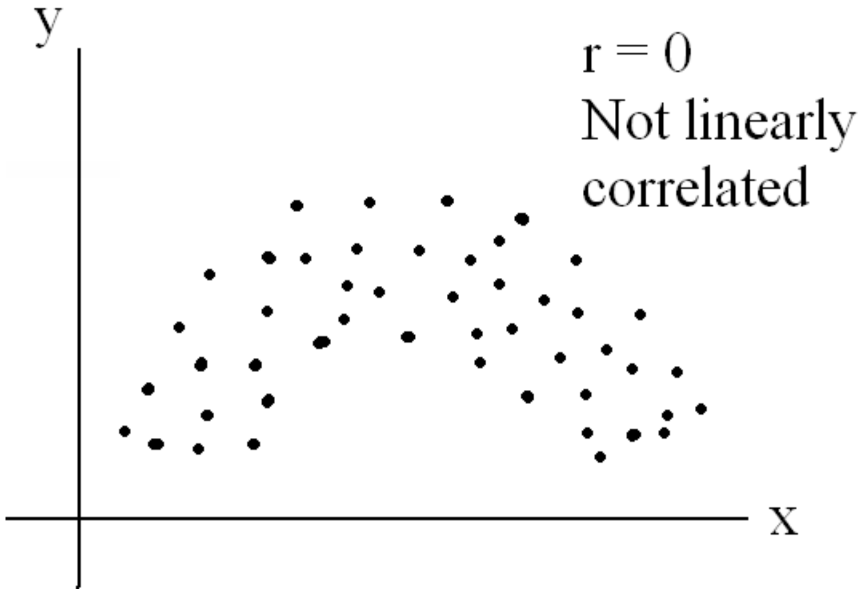
Weak - Corr ($r < 0$, but not near -1)

- The alignment of data cloud around the line is indicative of the correlation strength;
- Does not need to be 45 degree line.

Look at $|r|$ as measure of dependence betwen two variables.

(Linear Dependence)

e.g. pure parabolic relationship => r = 0



Correlations: Measuring and Describing Relationships (cont.)

- The most common **form** of relationship is a straight line or linear relationship which is measured by the Pearson correlation.
- Conceptually, it is computed by

$$
\begin{aligned}
r &= \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}} \\
&= \frac{\text{covariability of } X \text{ and } Y}{\text{variability of } X \text{ and } Y \text{ separately}}
\end{aligned}
$$

The Pearson Correlation

- The **Pearson correlation** measures the direction and degree of linear (straight line) relationship between two variables.
- To compute the Pearson correlation, you first measure the variability of X and Y scores separately by computing SS for the scores of each variable (SS_X and SS_Y).
- Then, the covariability (the tendency for X and Y to vary together) is measured by the sum of products (SP).
- The Pearson correlation is found by computing the ratio, $SP/\sqrt{(SS_X)(SS_Y)}$.

The Pearson Correlation (cont.)

- Thus the Pearson correlation is comparing the amount of covariability (variation from the relationship between X and Y) relative to the amount X and Y vary separately.
- The magnitude of the Pearson correlation ranges from 0 (indicating no linear relationship between X and Y) to 1.00 (indicating a perfect straight-line relationship between X and Y).
- The correlation can be either positive or negative depending on the direction of the relationship.

Inference for correlation coefficient

$$\rho_{xy} \quad \Leftarrow \quad r = \hat{\rho}_{xy}$$

population correlation coefficient

sample correlation coefficient

协方差

$$\rho_{xy} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x\sigma_y} = \frac{E(XY)-E(X)E(Y)}{\sigma_x\sigma_y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

协方差

$$r= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) / n}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

How to interpret *r*

- Range between [-1 1]: direction, magnitude

  0.1~0.29: small effect

  0.3~0.49: medium effect

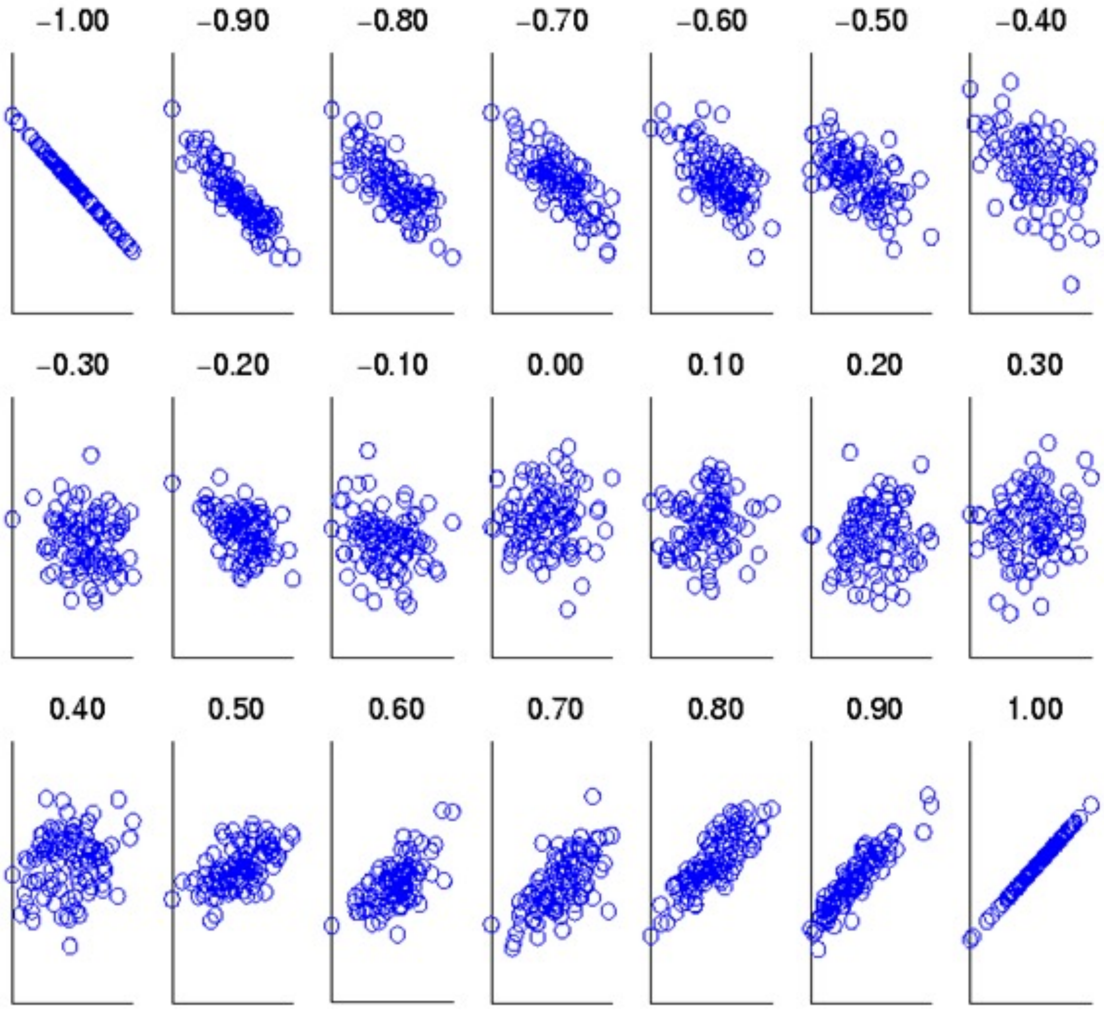
  >0.5: large effect

- Correlation does not imply causation

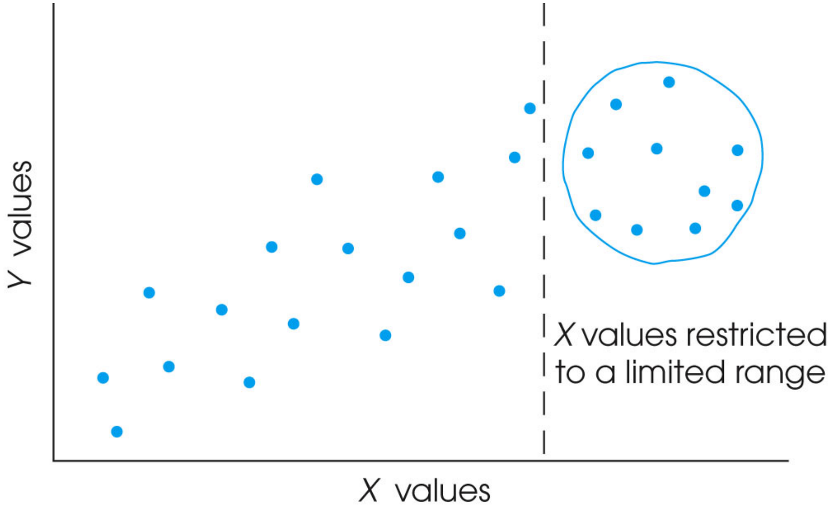
- Saturation effect, Range restriction (page 283, figure 14-6)
- **The bias from outliers (page 283, figure 14-7)**

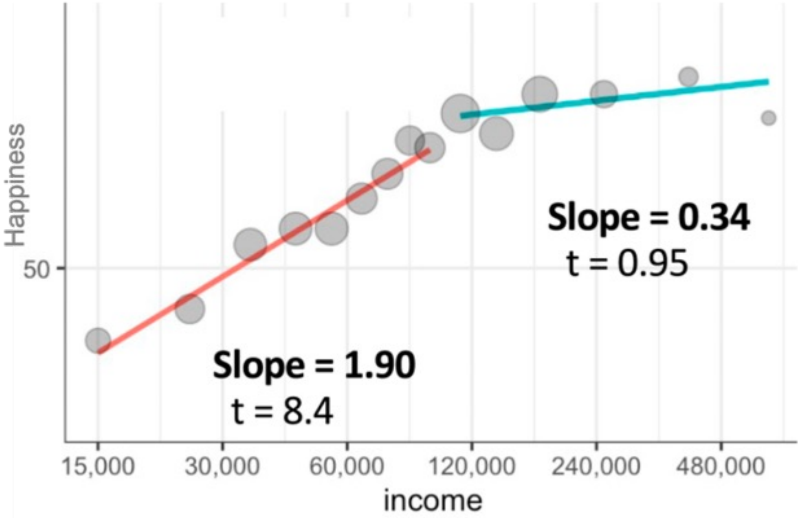
- Cautious about interpreting your correlation results

Get some sense of varying r



Range restriction

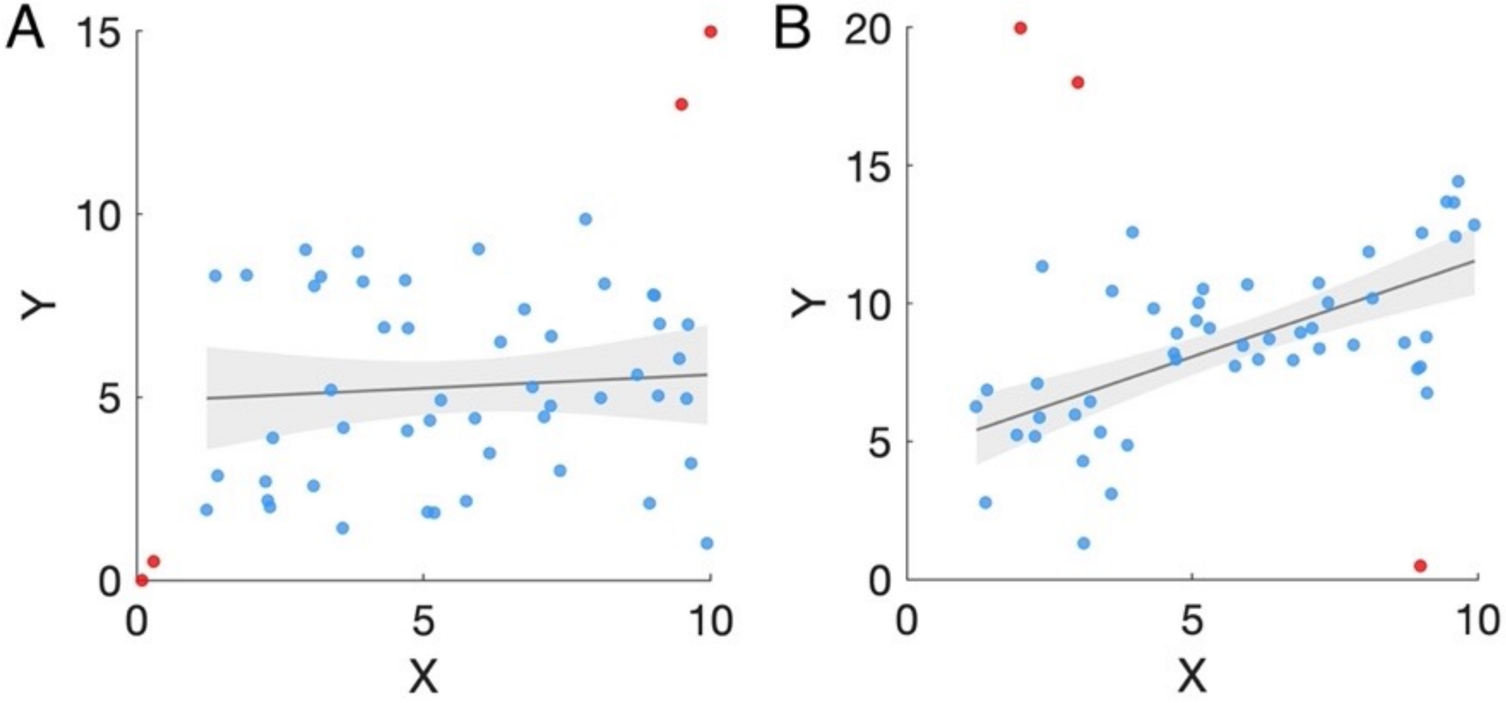




- Overall there is a correlation
- But in a restricted range, no correlation

- More money, more happiness?
- Only in a certain range
- Here shows a saturation effect

Outliers can impact the value of r



Without outliers: r = 0.08
With: r = 0.32

With outliers: r = 0.08
Without: r = 0.32

How to test if correlation is significant

我们也可以用查表得方法, 见附表B.6

Example

Suppose correlation between cholesterol levels of spouses is estimated to be $r = 0.25$ based on 100 spouses.

$H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$

Assumption:

- Cholesterol levels of husbands are Normally Distributed.
- Same for Wives

How to test if correlation is significant

- t statistics is used for hypothesis testing.

$$t = \frac{\text{sample statistic} \;-\; \text{population parameter}}{\text{standard error}}$$

$$\text{standard error for } r = s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Thus, the complete t statistic is

$$t = \frac{r - \rho}{\sqrt{\frac{(1 - r^2)}{(n - 2)}}}$$

The df of t is n-2

Assumptions for Pearson Correlation

- **Assumption**: the pairs (x_i,y_i) are selected randomly from a population of (X,Y)
  - Both X and Y should be continuous variables (interval or ratio data), and follow a normal distribution
  - Each observation (a pair of x and y) should be independent. In other words, no related measures are allowed.

  Requirements:

  - *No outliers*
  - *The effect is indeed linear*

What if X (or Y or both) is ordinal or even binary data?

Non-parametric tests for Correlations

- The **Spearman correlation** is used in two general situations:

  (1) It measures the relationship between two ordinal variables; that is, X and Y both consist of ranks.

  (2) It measures the consistency of the direction of the relationship between two variables. In this case, the two variables must be converted to ranks before the Spearman correlation is computed.

The Spearman Correlation (cont.)

The calculation of the Spearman correlation requires:

1. Two variables are observed for each individual.
2. The observations for each variable are rank ordered. Note that the X values and the Y values are ranked separately.
3. After the variables have been ranked, the Spearman correlation is computed by either:

    a. Using the Pearson formula with the ranked data.

    b. Using the special Spearman formula (assuming there are few, if any, tied ranks).

Correlation between IQ and performance

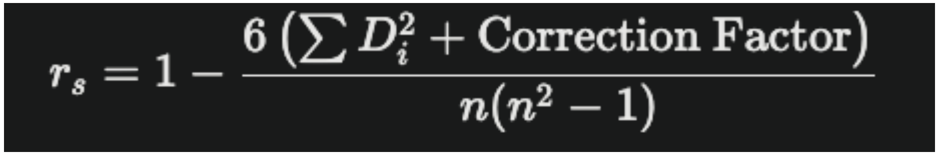
IQ	Performance	IQ rank	Performance rank	D	D^2
98	6	1	4	−3	9
99	3	2	2	0	0
101	9. 5	3	9	−6	36
103	1	4	1	3	9
104	8. 5	5	7	−2	4
107	10	6	10	−4	16
108	9	7	8	−1	1
111	12	8	13	−5	25
112	7	9	5	4	16
113	5	10	3	7	49
120	10. 5	11	11	0	0
122	8	12	6	6	36
130	11	13	12	1	1

Pearson correlation = .486
Spearman correlation = .445

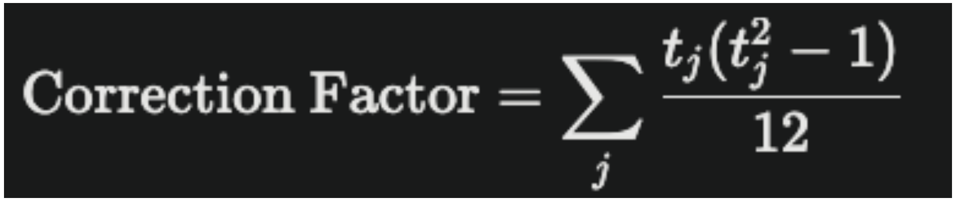
Spearman correlation formula:

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

Spearman correlation with correction for tied ranks

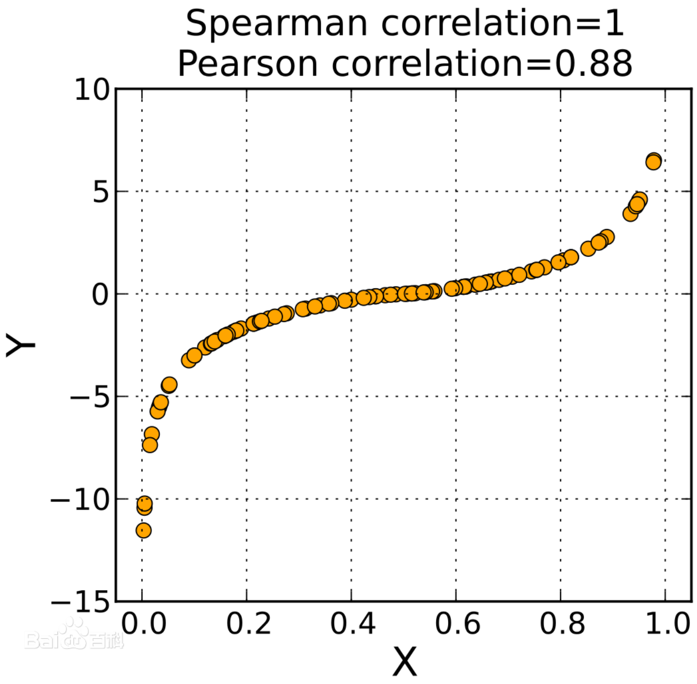


$$r_s = 1 - \frac{6\left(\sum D_i^2 + \text{Correction Factor}\right)}{n(n^2-1)}$$



$$\text{Correction Factor} = \sum_j \frac{t_j(t_j^2-1)}{12}$$

- ***t_j*** is the number of tied observations in each group j.
- For example, three observations are tied, and they would originally occupy ranks 4, 5, and 6. Then, this group *j* has ***t_j*** = 3



Spearman correlation measures the consistency of the direction of the relationship

The Point-Biserial Correlation and the Phi Coefficient

- The Pearson correlation formula can also be used to measure the relationship between two variables when one or both of the variables are dichotomous.

- A dichotomous variable is one for which there are exactly two categories: for example, men/women or succeed/fail.

The Point-Biserial Correlation and the Phi Coefficient (cont.)

With either one or two dichotomous variables the calculation of the correlation precedes as follows:

1. Assign numerical values to the two categories of the dichotomous variable(s). Traditionally, one category is assigned a value of 0 and the other is assigned a value of 1.

2. Use the regular Pearson correlation formula to calculate the correlation.

The Point-Biserial Correlation and the Phi Coefficient (cont.)

- In situations where one variable is dichotomous and the other consists of regular numerical scores (interval or ratio scale), the resulting correlation is called a **point-biserial correlation**.
- When both variables are dichotomous, the resulting correlation is called a **phi-coefficient**.

The Point-Biserial Correlation and the Phi Coefficient (cont.)

- The point-biserial correlation is closely related to the independent-measures t test introduced before.
- When the data consists of one dichotomous variable and one numerical variable, the dichotomous variable can also be used to separate the individuals into two groups.
- Then, it is possible to compute a sample mean for the numerical scores in each group.
- $r_{pb} = \frac{M_1 - M_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}}, \qquad s_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}.$

The Point-Biserial Correlation and the Phi Coefficient (cont.)

- In this case, the independent-measures t test can be used to evaluate the mean difference between groups.

- If the effect size for the mean difference is measured by computing r^2 (the percentage of variance explained), the value of r^2 will be equal to the value obtained by squaring the point-biserial correlation.

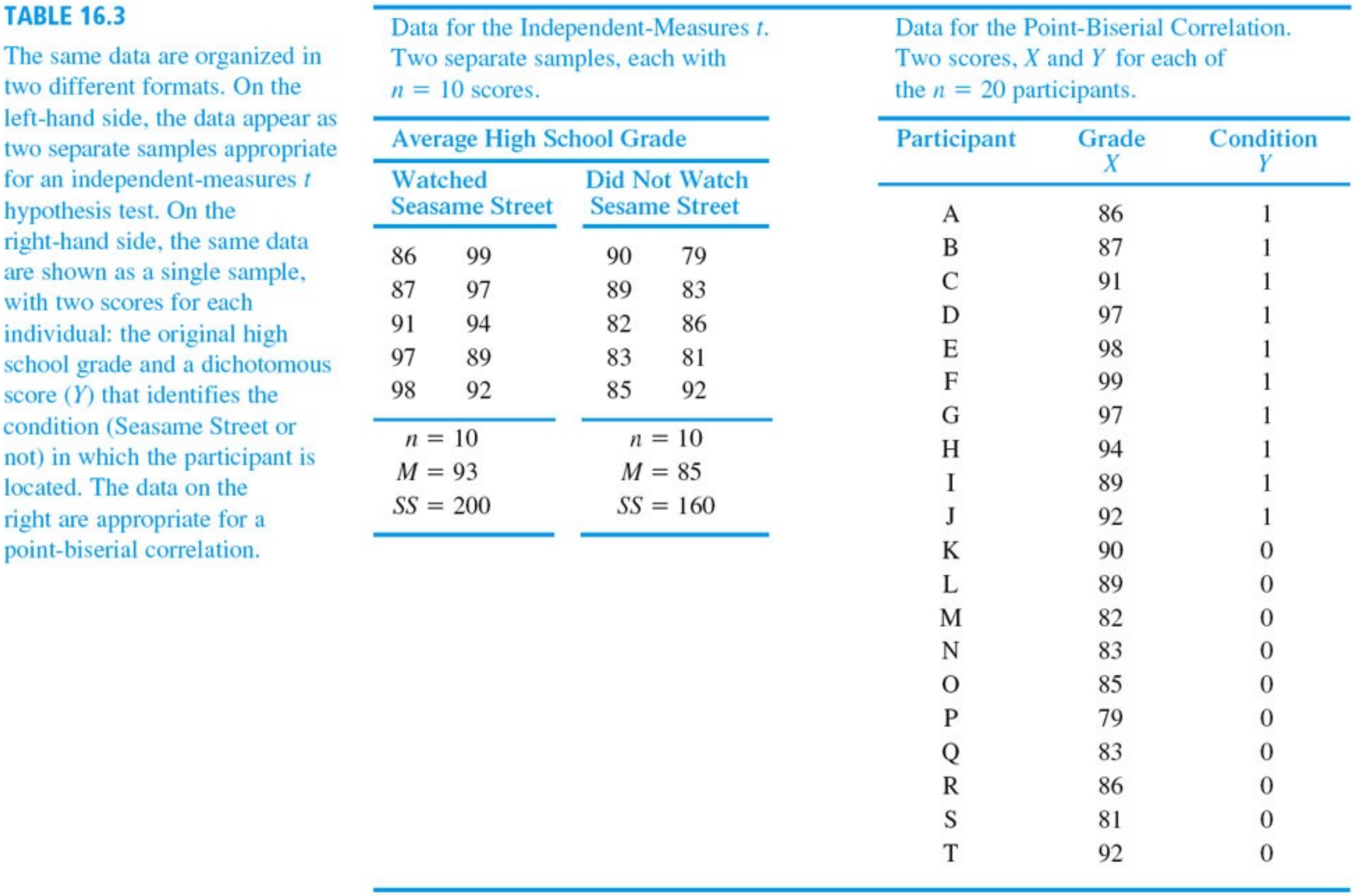


TABLE 16.3

The same data are organized in two different formats. On the left-hand side, the data appear as two separate samples appropriate for an independent-measures *t* hypothesis test. On the right-hand side, the same data are shown as a single sample, with two scores for each individual: the original high school grade and a dichotomous score (*Y*) that identifies the condition (Seasame Street or not) in which the participant is located. The data on the right are appropriate for a point-biserial correlation.

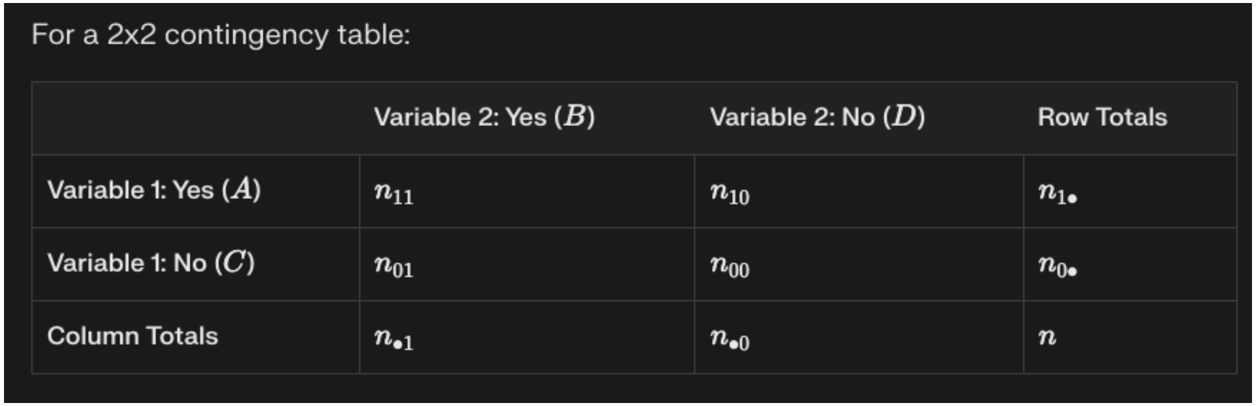
Data for the Independent-Measures *t*. Two separate samples, each with $n = 10$ scores.

Average High School Grade			
Watched Seasame Street		Did Not Watch Sesame Street	
86	99	90	79
87	97	89	83
91	94	82	86
97	89	83	81
98	92	85	92
$n = 10$ $M = 93$ $SS = 200$		$n = 10$ $M = 85$ $SS = 160$	

Data for the Point-Biserial Correlation. Two scores, *X* and *Y* for each of the $n = 20$ participants.

Participant	Grade *X*	Condition *Y*
A	86	1
B	87	1
C	91	1
D	97	1
E	98	1
F	99	1
G	97	1
H	94	1
I	89	1
J	92	1
K	90	0
L	89	0
M	82	0
N	83	0
O	85	0
P	79	0
Q	83	0
R	86	0
S	81	0
T	92	0

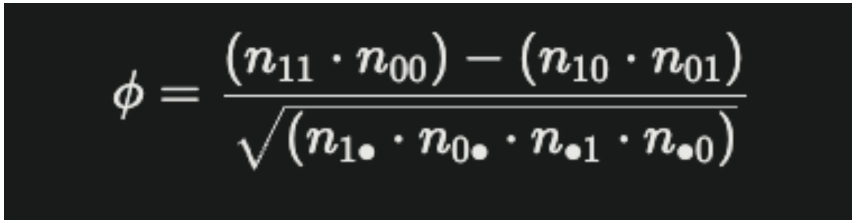
Phi coefficient



For a 2x2 contingency table:

	Variable 2: Yes (B)	Variable 2: No (D)	Row Totals
Variable 1: Yes (A)	n_{11}	n_{10}	$n_{1\bullet}$
Variable 1: No (C)	n_{01}	n_{00}	$n_{0\bullet}$
Column Totals	$n_{\bullet 1}$	$n_{\bullet 0}$	n

The data is a frequency table called contingency table



$$\phi = \frac{(n_{11} \cdot n_{00}) - (n_{10} \cdot n_{01})}{\sqrt{(n_{1\bullet} \cdot n_{0\bullet} \cdot n_{\bullet 1} \cdot n_{\bullet 0})}}$$

- Phi coefficient is computed by using frequencies.

- Phi (**Φ**) is interpreted in the same way as *r*

R codes

- Simple correlation
  - Cor(x, y, method = “pearson”/ “spearman”)
  - Or, use Corr() from the *bruceR* library
  - phi() from *psych* library for phi coefficient
  - cor.test() from *stats* library can do all the tests
- multivariate correlation
  - Cor(matrix)
- Visualize the correlation
  - Package (PerformanceAnalytics) – there will be more packages available
  - chart.Correlation(matrix)