STA 1013 : Statistics through Examples

Lecture 30: Correlation Coefficient

Hwiyoung Lee

November 22, 2019

Department of Statistics, Florida State University

Correlation

A **correlation** exists between two variables

- when the values of one variable are somehow associated with the values of the other variable
- when higher values of one variable consistently go with higher values of another variable
- when higher values of one variable consistently go with lower values of another variable

Examples

1. amount of smoking and likelihood of lung cancer
   : heavier smokers were more likely to get lung cancer

2. height and weight for people
   : taller people tend to weigh more than shorter people

3. practice time and skill among piano player
   : those who practice more tend to be more skilled

4. demand for apples and price of apples
   :demand tends to decrease as price increases

Types of Correlation

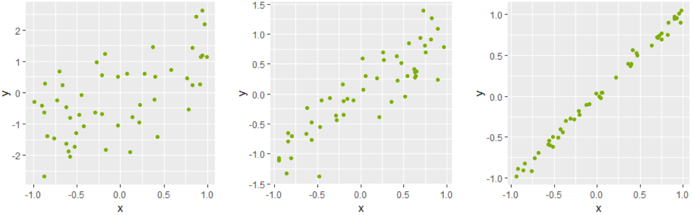
1. **Positive correlation:** Both variables tend to increase (or decrease) together.
2. **Negative correlation:** The two variables tend to change in opposite directions, with one increasing while the other decreases.
3. **No correlation:** There is no apparent (linear) relationship between the two variables.
4. **Nonlinear relationship:** The two variables are related, but the relationship results in a scatterplot that does not follow a straight-line pattern.

Scatter plot

- A scatterplot is a graph in which each point represents the values of two variables
- We can identify relation between two variables

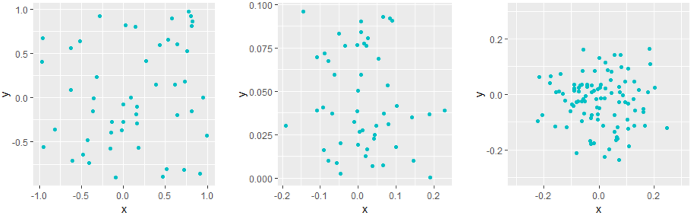
Types of Correlation

- Positive correlation

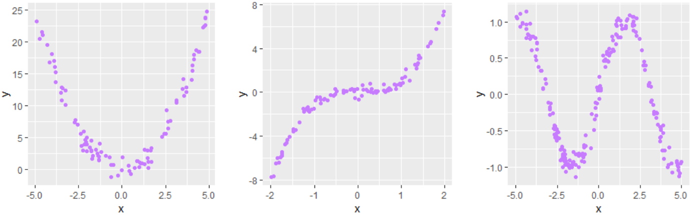


- Negative correlation

- No relation



- Nonlinear relation



Measuring the Strength of a Correlation

Statisticians measure the strength of a **linear correlation** with a number called the correlation coefficient.

Correlation coefficient

$$
\begin{aligned}
r &= \frac{\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}\right)}{n-1} \\
&= \frac{n(\sum_{i=1}^{n} x_i y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}
\end{aligned}
$$

Use Calculator

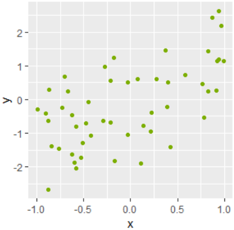
Properties of a Correlation

- The correlation coefficient, r, is a measure of the strength of a correlation. Its value can range only from -1 to 1.

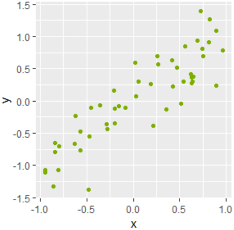
$$-1 \leqslant r \leqslant 1$$

- If there is no correlation, the value of r is close to 0.
- If there is a positive correlation, the correlation coefficient is positive $(0 < r \leqslant 1)$. Values of r close to 1 indicate a strong positive correlation and positive values closer to 0 indicate a weak positive correlation.
- If there is a negative correlation, the correlation coefficient is negative$(-1 \leqslant r < 0)$: Values of r close to -1 indicate a strong negative correlation and negative values closer to 0 indicate a weak negative correlation.

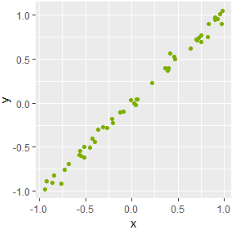
- Positive correlation



$r = 0.2$

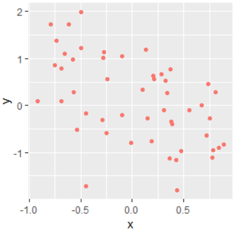


$r = 0.6$

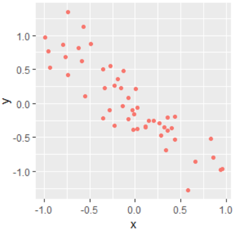


$r = 0.9$

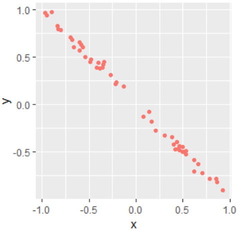
- Negative correlation



$r = -0.2$



$r = -0.6$



$r = -0.9$

Testing a Linear relation between two variables

Is there a Linear Correlation ?

- To claim that there is a linear correlation is to claim that the population linear correlation coefficient ρ is different from 0.
- Hypothesis

$$H_0 : \rho = 0 \text{ (There is no linear correlation)}$$
$$H_a : \rho \neq 0 \text{ (There is a linear correlation)}$$

- Test statistic

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

- **Use a p-value approach**

Calculating the Correlation Coefficient r

LinRegTTest

1. Press the $\boxed{\text{stat}}$ and highlight **TESTS**
2. Scroll down to F: **LinRegTTest**
3. Enter values
   - Xlist : L_1
   - Ylist : L_2
   - Freq : 1
   - β & ρ : $\neq 0, < 0, > 0$

Note : To view the Correlation Coefficient, **turn on**

- **DiaGnosticOn** : [2nd] "Catalog" (above the '0'). Scroll to DiaGnosticOn. [Enter] [Enter]
- or **STATDIAGNOSITICS** : [mode] Scroll to STATDIAGNOSITICS : ON

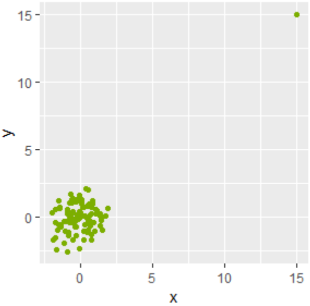
Example

Brain Size	IQ	Brain Size	IQ
965	90	1,077	97
1,029	85	1,037	124
1,030	86	1,068	125
1,285	102	1,176	102
1,049	103	1,105	114

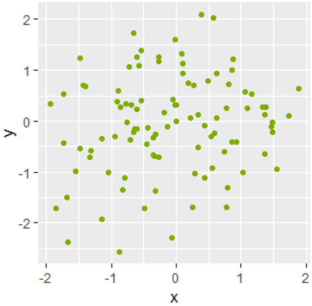
- Calculate the sample correlation coefficient r
- Does the value of r indicate that brain size is related to IQ (Test linear correlation with $\alpha = 0.05$)

Beware of outliers

Correlation is very sensitive to outliers



$r = 0.76$



$r = 0.1$

- The left panel contains an outlier : $r = 0.76$
- Outlier is removed in the right panel : $r = 0.1$

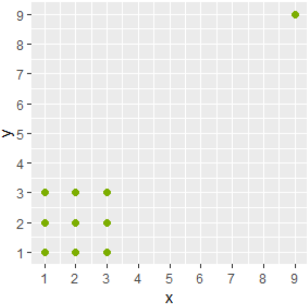
Example : Beware of outliers

X	Y	X	Y
1	1	2	3
1	2	3	1
1	3	3	2
2	1	3	3
2	2	9	9

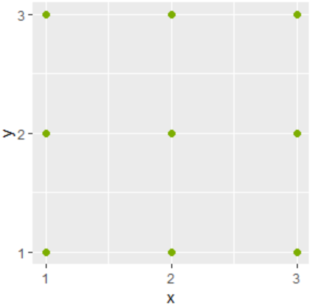
1. Draw a scatter plot
2. Find the correlation coefficient
3. Remove the last observation (9,9), then find the correlation coefficient

Solution : Beware of outliers

1. Draw a scatter plot



$r = 0.88$



$r = 0$

2. Find the correlation coefficient

$$r = 0.88 \text{ ; the left panel}$$

3. Remove the last observation (9,9), then find the correlation coefficient

$$r = 0 \text{ ; the right panel}$$

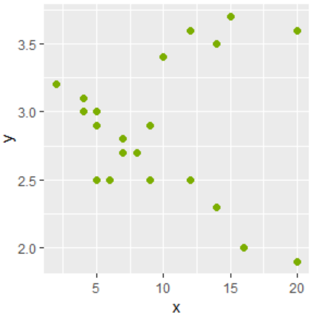
Beware of inapproate grouping

Hours of TV and high School GPA data

hours per week of TV	GPA	hours per week of TV	GPA
2	3.2	9	2.5
4	3.0	9	2.9
4	3.1	10	3.4
5	2.5	12	3.6
5	2.9	12	2.5
5	3.0	14	3.5
6	2.5	14	2.3
7	2.7	15	3.7
7	2.8	16	2.0
8	2.7	20	3.6
		20	1.9

Beware of inapproate grouping

Hours of TV and high School GPA data

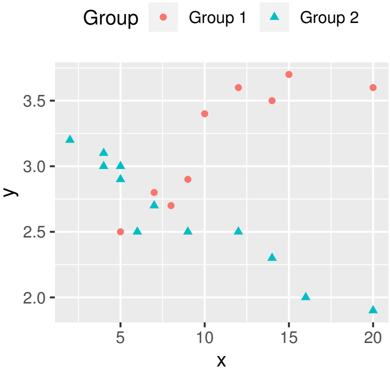


- The scatterplot shows virtually no correlation
- the correlation coefficient for the data is about $r = -0.063$
- The lack of correlation seems to suggest that TV viewing habits are unrelated to academic achievement

Beware of inapproate grouping

However, one astute researcher realizes that

- some of the students watched mostly **educational programs**
- while others tended to watch **comedies, dramas, and movies**

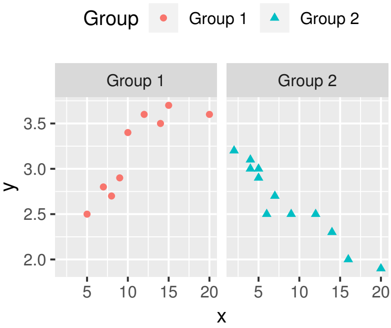


Beware of inapproate grouping

Hours of TV and high School GPA data by Group

Group 1 : Educational programs		**Group 2 : watched regular TV**	
hours per week of TV	GPA	hours per week of TV	GPA
5	2.5	2	3.2
7	2.8	4	3.0
8	2.7	4	3.1
9	2.9	5	2.9
10	3.4	5	3.0
12	3.6	6	2.5
14	3.5	7	2.7
15	3.7	9	2.5
20	3.6	12	2.5
		14	2.3
		16	2.0
		20	1.9

Beware of inapproate grouping



- A strong positive correlation for the students who watched educational programs ($r = 0.855$)
- A strong negative correlation for the other students ($r = -0.951$)
- **Correlations can also be misinterpreted when data are grouped inappropriately**

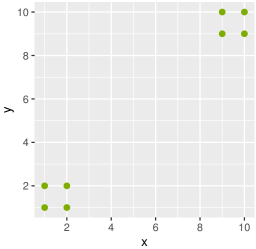
Example : Effects of Clusters

Group 1		Group 2	
X	Y	X	Y
1	1	9	9
1	2	9	10
2	1	10	9
2	2	10	10

1. Draw a scatter plot
2. Find the correlation coefficient of the whole data (using all eight points)
3. Find the correlation coefficient of the group 1 (using only the four points in the lower left corner)
4. Find the correlation coefficient of the group 2 (using the four points in the upper right corner)

Solution : Effects of Clusters

1. Draw a scatter plot



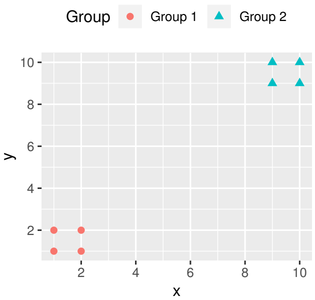
2. The correlation coefficient of the whole data : $r = 0.98$

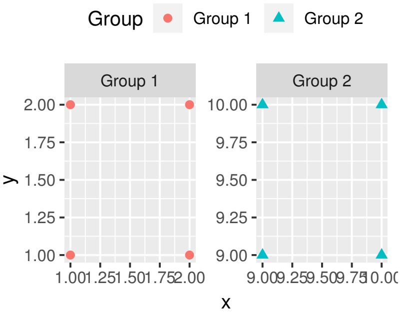
Note :

- The apparent correlation of the full data set occurs because of the separation between the two clusters of points
- The data set as a whole shows a strong correlation

Solution : Effects of Clusters

3. Find the correlation coefficient of the each group





- Group 1 : $r = 0$, Group 2 : $r = 0$

Note : If we analyze these subgroups separately, neither shows any correlation: