Correlation

PATIRA DATA SCIENCE
DISCOVERING KNOWLEDGE FROM DATA

THE OHIO STATE UNIVERSITY

Correlational Analysis

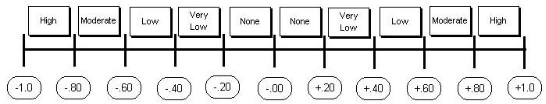
- The purpose is to measure the strength of a linear relationship between 2 variables.
- A correlation coefficient does not ensure "causation" (i.e. a change in X causes a change in Y)
- X is typically the input, measured, or Independent variable.
- Y is typically the output, predicted, or dependent variable.
- If X increases and there is a predictable shift in the values of Y, a correlation exists.

General Properties of Correlation Coefficients

- Correlation coefficients values ranges between +1 and -1.
- The value of the correlation coefficient represents the scatter of points on a scatterplot.
- You should be able to look at a scatterplot and estimate what the correlation would be.
- You should be able to look at a correlation coefficient and visualize the scatterplot.

Interpretation

- Depends on what the purpose of the study is… but here is a "*general guideline*"…



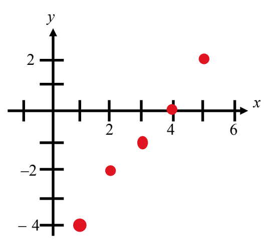
- Value = magnitude of the relationship
- Sign = direction of the relationship

Correlation graph

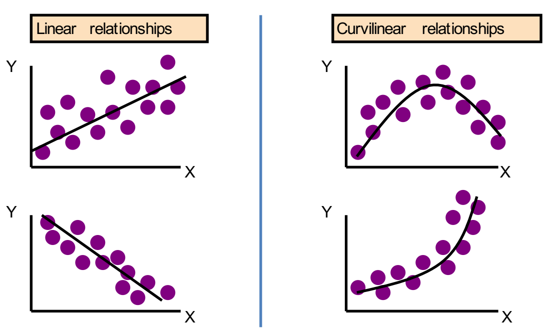
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

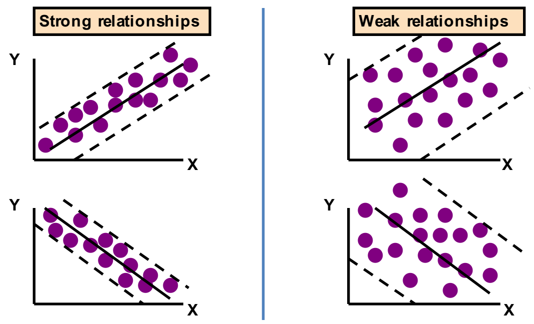
x	1	2	3	4	5
y	–4	–2	–1	0	2



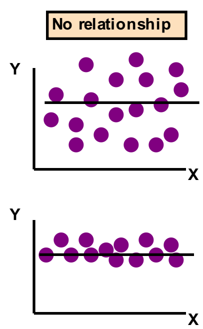
Correlation graph



Correlation graph

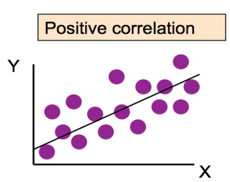


Correlation graph



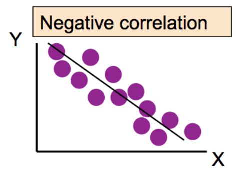
Positive Correlation

- Association between variables such that high scores on one variable tend to have high scores on the other variable
- A direct relation between the variables



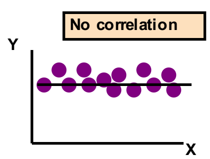
Negative Correlation

- Association between variables such that high scores on one variable tend to have low scores on the other variable
- An inverse relation between the variables



No correlation

- No association between variables such that high scores on one variable tend to have no relationship the other variable and vise versa



Pearson Correlation Coefficient (r)

- A statistic that quantifies a relation between two variables
- Can be either positive or negative
- Falls between -1.00 and 1.00
- The value of the number (not the sign) indicates the strength of the relation

The Pearson Correlation Coefficient

- Symbolized by the italic letter *r* when it is a statistic based on sample data.
- Symbolized by the italic letter *p* "rho" when it is a population parameter.

Correlation Coefficient

- The **correlation coefficient** is a measure of the **strength** and the **direction** of a linear relationship between two variables. The symbol *r* represents the sample correlation coefficient. The formula for *r* is

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

- The range of the correlation coefficient is –1 to 1. If *x* and *y* have a strong positive linear correlation, *r* is close to 1. If *x* and *y* have a strong negative linear correlation, *r* is close to –1. If there is no linear correlation or a weak linear correlation, *r* is close to 0.

Calculating a Correlation Coefficient

Calculating a Correlation Coefficient

	In Words	*In Symbols*
1.	Find the sum of the *x*-values.	$\sum x$
2.	Find the sum of the *y*-values.	$\sum y$
3.	Multiply each *x*-value by its corresponding *y*-value and find the sum.	xy
4.	Square each *x*-value and find the sum.	x^2
5.	Square each *y*-value and find the sum.	y^2
6.	Use these five sums to calculate the correlation coefficient.	$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$

Continued.

Correlation Coefficient

Example:
Calculate the correlation coefficient *r* for the following data.

x	*y*
1	– 3
2	– 1
3	0
4	1
5	2

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

Correlation Coefficient

Example:
Calculate the correlation coefficient *r* for the following data.

x	*y*	*xy*	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

Correlation Coefficient

Example:
Calculate the correlation coefficient *r* for the following data.

x	*y*	*xy*	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$x = 15$	$y = \ 1$	$xy = 9$	$x^2 = 55$	$y^2 = 15$

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}. = \frac{5(9) - (15)(-1)}{\sqrt{5(55) - 15^2}\sqrt{5(15) - (-1)^2}}$$

$$= \frac{60}{\sqrt{50}\sqrt{74}} \quad 0.986$$

There is a strong positive linear correlation between *x* and *y*.

Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.
a.) Display the scatter plot.
b.) Calculate the correlation coefficient *r*.

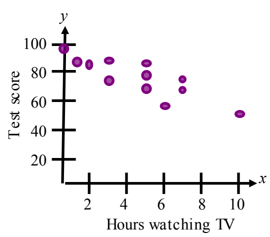
Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Continued.

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50



Continued.

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$x = 54$ $\quad y = 908$ $\quad xy = 3724$ $\quad x^2 = 332$ $\quad y^2 = 70836$

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} = \frac{12(3724) \quad (54)(908)}{\sqrt{12(332) \quad 54^2}\sqrt{12(70836) \quad (908)^2}} \qquad 0.831$$

There is a strong negative linear correlation (-0.831).

As the number of hours spent watching TV increases, the test scores tend to decrease.

Testing a Population Correlation Coefficient

- Once the sample correlation coefficient r has been calculated, we need to determine whether there is enough evidence to decide that the population correlation coefficient ρ is significant at a specified level of significance.
- One way to determine this is to use Critical Values of Pearson's Correlation Coefficient r Table
- If $|r|$ is greater than the critical value, there is enough evidence to decide that the correlation coefficient ρ is significant.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

For a sample of size $n = 6$, ρ is significant at the 5% significance level, if $|r| > 0.811$.

Testing a Population Correlation Coefficient

Finding the Correlation Coefficient ρ

In Words	*In Symbols*
1. Determine the number of pairs of data in the sample.	Determine n.
2. Specify the level of significance.	Identify α.
3. Find the critical value.	Use correlation Table .
4. Decide if the correlation is significant.	If $|r|$ > critical value, the correlation is significant. Otherwise, there is not enough evidence to support that the correlation is significant.
5. Interpret the decision in the context of the original claim.	

Testing a Population Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

The correlation coefficient $r \approx -0.831$.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Is the correlation coefficient significant at $\alpha = 0.01$?

Continued.

Testing a Population Correlation Coefficient

Example continued:

$r \approx -0.831$

$n = 12$

$\alpha = 0.01$

Correlation Table

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684

$|r| > 0.708$

Because, the population correlation is significant, there is enough evidenc at the 1% level of significance to conclude that there is a significant linear correlation between the number of hours of television watched during the weekend and the scores of each student who took a test the following Monday.

Significance Test for Correlation

- Hypotheses

H_0: $\rho = 0$ (no correlation)
H_A: $\rho \neq 0$ (correlation exists)

- Test statistic

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

(with n – 2 degrees of freedom)

Correlation and Causation

The fact that two variables are strongly correlated does not in itself imply a cause-and-effect relationship between the variables.

If there is a significant correlation between two variables, you should consider the following possibilities.

1. **Is there a direct cause-and-effect relationship between the variables?**
   Does x cause y?
2. **Is there a reverse cause-and-effect relationship between the variables?**
   Does y cause x?
3. **Is it possible that the relationship between the variables can be caused by a third variable or by a combination of several other variables?**
4. **Is it possible that the relationship between two variables may be a coincidence?**