Advanced Statistics

Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- **Elementary Statistics: Picturing the World,** 6^{th} Edition, Ron Larson and Betsy Farber
- **Elementary Statistics**, 13^{th} Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

- Elementary Statistics, 14th Edition, Mario F. Triola

These notes contain material from the above resources.

Linear Correlation Coefficient

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

(Good format for calculations)

$$r = \frac{\sum Z_x Z_y}{n-1}$$ **(Good format for understanding)**

where Z_x denotes the ***z* score** for an individual sample value x and Z_y is the ***z* score** for the corresponding sample value y.

Example: Calculating *r* Using the simple random sample of data given in the table, find the value of the **linear correlation coefficient *r***.

Chocolate Consumption and Nobel Laureates

Chocolate	5	6	4	4	5
Nobel	6	9	3	2	11

x	y	xy	x^2	y^2
5	6	30	25	36
6	9	54	36	81
4	3	12	16	9
4	2	8	16	4
5	11	5	25	121
$\sum x = 24$	$\sum y = 31$	$\sum xy = 159$	$\sum x^2 = 118$	$\sum y^2 = 251$

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

$$r = \frac{5(159) - (24)(31)}{\sqrt{5(118) - (24)^2}\sqrt{5(251) - (31)^2}}$$

$$r = \frac{51}{\sqrt{14}\sqrt{294}} = 0.795$$

Method 2

$$r = \frac{\sum Z_x Z_y}{n-1}$$

$$Z_x = \frac{x - \bar{x}}{s_x}$$

$$Z_y = \frac{y - \bar{y}}{s_y}$$

Method 2

$$s_x = \sqrt{\frac{\sum(x - \overline{x})^2}{n - 1}}$$

or

$$s_x = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2\}}$$

$$s_y = \sqrt{\frac{\sum(y - \overline{y})^2}{n - 1}}$$

or

$$s_y = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} \mathrm{y}^2{}_i - (\sum_{i=1}^{n} y_i)^2\}}$$

$$s_x = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i\ )^2\}}$$

$$s_x = \sqrt{\frac{1}{5(5-1)}\{5(118) - (24)^2\}}$$

$$= 0.8367$$

$$s_y = \sqrt{\frac{1}{n(n-1)}\{n\sum_{i=1}^{n} y^2{}_i - (\sum_{i=1}^{n} y_i\ )^2\}}$$

$$s_y = \sqrt{\frac{1}{5(5-1)}\{5(251) - (31)^2\}}$$

$$= 3.8341$$

$\bar{x} = 4.8000$

$\bar{y} = 6.2000$

$$Z_x = \frac{\mathrm{x} - \bar{x}}{s_x}$$

$$= \frac{5 - 4.8}{0.8367} = 0.2390$$

$$Z_x = \frac{\mathrm{x} - \bar{x}}{s_x}$$

$$= \frac{5 - 4.8}{0.8367} = 0.239046$$

$$Z_y = \frac{\mathrm{y} - \bar{y}}{s_y}$$

$$Z_y = \frac{6.0000-6.2000}{3.8341}$$

$$Z_y = -0.052164$$

x	y	$Z_x = \frac{x - \bar{x}}{s_x}$	$Z_y = \frac{y - \bar{y}}{s_y}$	$Z_x Z_y$
5	6	0.239046	-0.052164	-.012470
6	9	1.434274	0.730297	1.047446
4	3	-0.956183	-0.834625	0.798054
4	2	-0.956183	-1.095445	1.047446
5	11	0.239046	1.251937	0.299270
				$\sum(Z_x Z_y) =$ 3.179746

$$\mathbf{r} = \frac{\sum Z_x Z_y}{n-1}$$

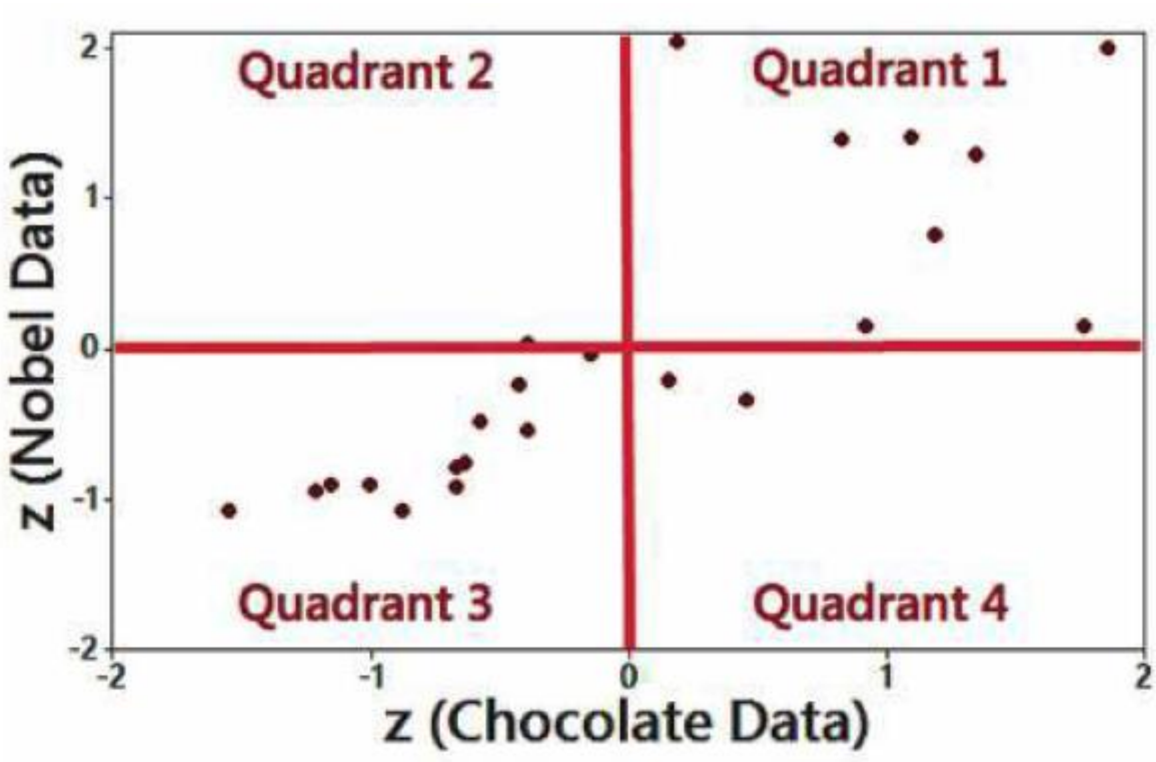
$$\mathbf{r} = \frac{3.179746}{4}$$

$$r = 0.795$$

- If the points of the **scatterplot approximate** an **uphill line** (as in Figure next slides), individual values of the product $Z_x Z_y$ tend to be **positive** (because most of the points are found in the **first and third quadrants**, where the values of Z_x and Z_y are either both positive or both negative), so $\sum(Z_x Z_y)$tends to be **positive**.

- If the points of the **scatterplot approximate** a **downhill line**, most of the points are in the **second and fourth quadrants**, where $Z_x$and Z_y are **opposite in sign**, so $\sum(Z_x Z_y)$tends to be **negative**.

- Points that follow **no linear pattern** tend to be scattered among the **four quadrants**, so the value of $\sum(Z_x Z_y)$tends to **be close to 0**.


Quadrant 2
Quadrant 1
Quadrant 3
Quadrant 4
z (Nobel Data)
z (Chocolate Data)
2
1
0
-1
-2
-2
-1
0
1
2

Using $\sum(Z_x Z_y)$ as a measure of how the points are configured among the four quadrants, we get the following:

- **Positive Correlation:** A large **positive value of** $\sum(Z_x Z_y)$ suggests that the points are predominantly in the **first and third quadrants** (corresponding to a positive linear correlation).
- **Negative Correlation:** A large **negative value of** $\sum(Z_x Z_y)$ suggests that the points are predominantly in the **second and fourth quadrants** (corresponding to a negative linear correlation).
- **No Correlation:** A value of $\sum(Z_x Z_y)$ **near 0** suggests that the points are **scattered among the four quadrants** (with no linear correlation).

Spurious Correlation (In class quiz)

Table1 lists paired data consisting of per capita consumption of **margarine (pounds)** in the United States and the divorce rate in Maine (divorces per 1000 people in Maine). Each pair of data is from a different year. The data are from the U.S. Census Bureau and the U.S. Department of Agriculture. Is there a linear correlation? What do you conclude?

Table 1 **U.S. Margarine Consumption and Divorces in Maine**

Margarine	8.2	7.0	6.5	5.3	5.2	4.0	4.6	4.5	4.2	3.7
Divorces	5.0	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1

Spurious Correlation

Here are the key points about the data in Table 1:

- The requirements appear to be **satisfied**.
- A **scatterplot** shows a very clear pattern of points that is close to a **straight-line pattern**, and there are no outliers.
- The linear correlation coefficient r is equal to 0.993.
- The P-value is 0.000.
- The critical values are $r = \pm 0.632$ (assuming a 0.05 significance level).

Spurious correlation

- Based on these results, we should support a **claim that there is a linear correlation between margarine consumption and the divorce rate in Maine**.

- But, come on! Common sense strongly suggests **that there is no real association between those two variables**.

- It would be totally **ridiculous to argue that one of the variables is the cause of the other**.

- Statistics is so much more than blindly running data through formulas and procedures—**it requires *critical thinking*!**

- A **spurious correlation** is a correlation that doesn’t have an actual association, as In the previous Example.

- Note: **Spurious correlations** will become more common with the **increased use of big data**, and they are more likely to occur with **time-series data that have similar trends**.