CORRELATION/REGRESSION I (DA_2022)

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

Lecture n. 22, Rome mon 23 th of May 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

Outline L 22

- General notion of correlated events
- General notion of correlation
- Pearson's Correlation Coefficient (sample)
- Population Correlation coefficient
- Rank correlation (Spearman)
- The mathematics of linearity: linear spaces of finite dimension, vectors,
- Linear transformations, matrices
- Discussion of geometric data analysis/dimensional reduction / classification/clustering)
- Intro to Principal Component analysis (PCA)

GENERAL EXPERIMENTAL SETTING
Population of objects
extract random samples from it
measure on each object of a sample two observables: X and Y
(at the same time or with a delay τ) dynamical approaches neurobiology
How much X and Y influence each other?
Is the order relevant? (this has to do with causal influence)

CORRELATED EVENTS

General Multiplication Rule (Bayes' Rule). If outcomes A and B occur with probabilities $p(A)$ and $p(B)$, the joint probability of events A AND B is

$$p(AB) = p(B \mid A)p(A) = p(A \mid B)p(B). \tag{1.11}$$

If events A and B happen to be independent, the pre-condition A has no influence on the probability of B. Then $p(B \mid A) = p(B)$, and Equation (1.11) reduces to $p(AB) = p(B)p(A)$, the multiplication rule for independent events. A probability $p(B)$ that is not conditional is called an *a priori* probability. The conditional quantity $p(B \mid A)$ is called an *a posteriori* probability. The general multiplication rule is general because independence is not required. It defines the probability of the *intersection* of events, $p(AB) = p(A \cap B)$.

General Addition Rule. A general rule can also be formulated for the union of events, $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, when we seek the probability of A OR B for events that are not mutually exclusive. When A and B *are* mutually exclusive, $p(A \cap B) = 0$, and the general addition rule reduces to the simpler addition rule on page 3. When A and B are independent, $p(A \cap B) = p(A)p(B)$, and the general addition rule gives the result in Example 1.6.

GENERAL NOTION: DEGREE OF CORRELATION BETWEEN EVENTS

DEGREE OF CORRELATION. The degree of correlation g between events A and B can be expressed as the ratio of the conditional probability of B, given A, to the unconditional probability of B alone. This indicates the degree to which A influences B:

$$g = \frac{p(B \mid A)}{p(B)} = \frac{p(AB)}{p(A)p(B)}. \tag{1.12}$$

The second equality in Equation (1.12) follows from the general multiplication rule, Equation (1.11). If $g = 1$, events A and B are independent and not correlated. If $g > 1$, events A and B are positively correlated. If $g < 1$, events A and B are negatively correlated. If $g = 0$ and A occurs, then B will not. If the *a priori* probability of rain is $p(B) = 0.1$, and if the conditional probability of rain, given that there are dark clouds, A, is $p(B \mid A) = 0.5$, then the degree of correlation of rain with dark clouds is $g = 5$. Correlations are important in statistical thermodynamics. For example, attractions and repulsions among molecules in liquids can cause correlations among their positions and orientations.

From: K.A: Dill & S. Bromberg, Molecular Driving Forces, chap. 1

THE CORRELATION COEFFICIENT

The discussion of linear-regression analysis in Sections 11.2–11.6 primarily focused on methods of predicting a dependent variable (y) as a function of an independent variable (x). Often we are interested not in predicting one variable from another but rather in investigating whether or not there is a relationship between two variables.

Cardiovascular Disease Serum cholesterol is an important risk factor in the etiology of cardiovascular disease. Much research has been devoted to understanding the environmental factors that cause elevated cholesterol levels. For this purpose, cholesterol levels were measured on 100 genetically unrelated spouse pairs. We are not interested in predicting the cholesterol level of a husband from that of his wife but instead would like some quantitative measure of the relationship between their levels. We will use the correlation coefficient for this purpose.

First, we discuss the related concept of covariance. The *covariance* is a measure used to quantify the relationship between two random variables.

The **covariance** between two random variables X and Y is denoted by $Cov(X, Y)$ and is defined by

$$Cov(X,Y) = E\left[\left(X-\mu_x\right)\left(Y-\mu_y\right)\right]$$

This is a POPULATION LEVEL probabilistic notion

which can also be written as $E(XY) - \mu_x\mu_y$, where μ_x is the average value of X, μ_y is the average value of Y, and $E(XY)$ = average value of the product of X and Y.

One issue is that the covariance between two random variables X and Y is in the units of X multiplied by the units of Y. Thus, it is difficult to interpret the strength of association between two variables from the magnitude of the covariance. To obtain a measure of relatedness independent of the units of X and Y, we consider the *correlation coefficient*.

1.16 The **correlation coefficient** between two random variables X and Y is denoted by $Corr(X, Y)$ or ρ and is defined by

$$\rho = Corr(X,Y) = Cov(X,Y)/(\sigma_x \sigma_y)$$

where σ_x and σ_y are the standard deviations of X and Y, respectively.

Unlike the covariance, the correlation coefficient is a dimensionless quantity that is independent of the units of X and Y and ranges between -1 and 1. For random variables that are approximately linearly related, a correlation coefficient of 0 implies independence. A correlation coefficient close to 1 implies nearly perfect positive dependence with large values of X corresponding to large values of Y and small values of X corresponding to small values of Y. An example of a strong posi-

NOTE: CORRELATION IS LINEAR REGRESSION WITHOUT THE MODEL

The **raw sum of squares for x** is defined by

$$\sum_{i=1}^{n} x_i^2$$

The **corrected sum of squares for x** is denoted by L_{xx} and defined by

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n$$

It represents the sum of squares of the deviations of the x_i from the mean. Similarly, the **raw sum of squares for y** is defined by

$$\sum_{i=1}^{n} y_i^2$$

The **corrected sum of squares for y** is denoted by L_{yy} and defined by

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n$$

Notice that L_{xx} and L_{yy} are simply the numerators of the expressions for the sample variances of x (i.e., s_x^2) and y (i.e., s_y^2), respectively, because

$$s_x^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 \Big/ (n-1) \text{ and } s_y^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 \Big/ (n-1)$$

The **raw sum of cross products** is defined by

$$\sum_{i=1}^{n} x_i y_i$$

The **corrected sum of cross products** is defined by

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

which is denoted by L_{xy}.

It can be shown that a short form for the corrected sum of cross products is given by

$$\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n$$

NOTE: THIS L_{xy} TERM HAS TO DO WITH CORRELATION
(degree of equivariance of y and x)

The sample (Pearson) correlation coefficient (r) is defined by

$$L_{xy} / \sqrt{L_{xx} L_{yy}}$$

The correlation is not affected by changes in location or scale in either variable and must lie between −1 and +1. The sample correlation coefficient can be interpreted in a similar manner to the population correlation coefficient ρ.

Interpretation of the Sample Correlation Coefficient

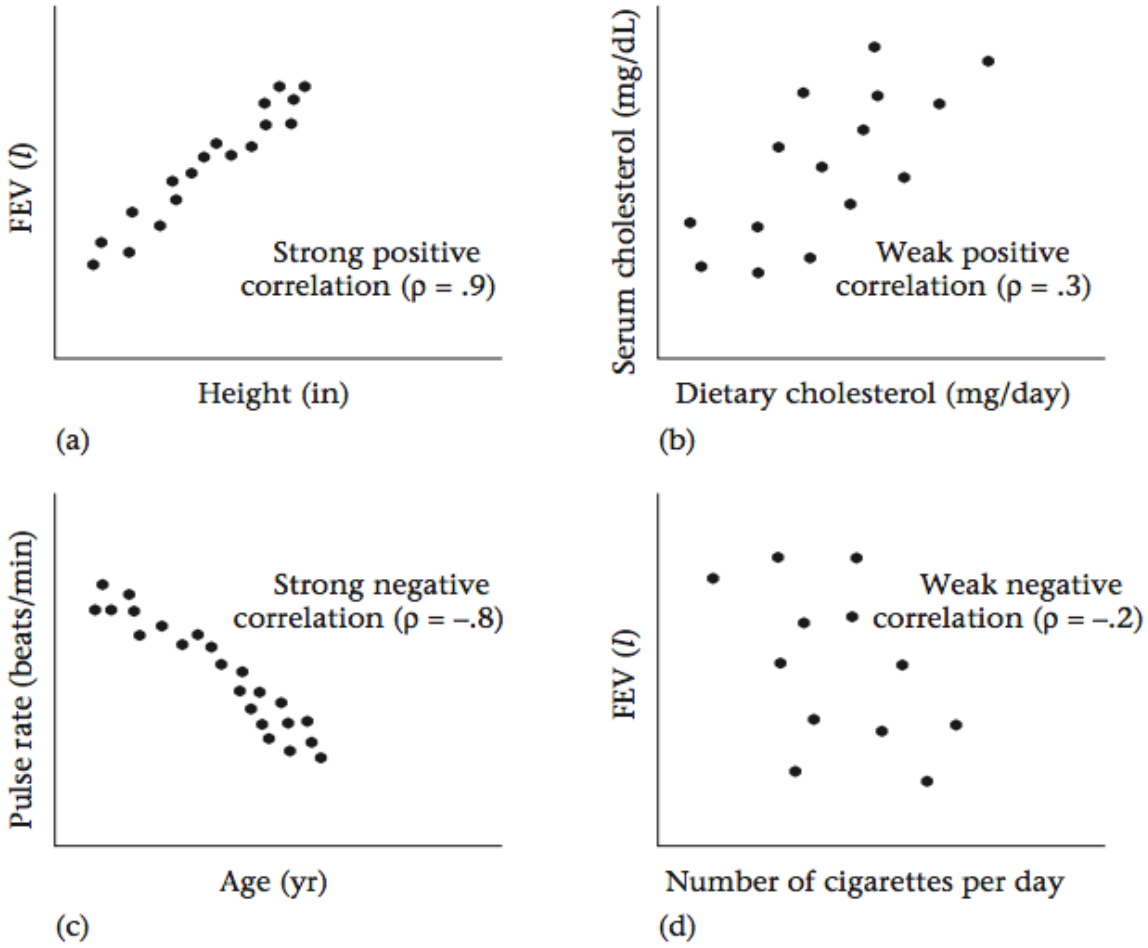
(1) If the correlation is greater than 0, such as for birthweight and estriol, then the variables are said to be **positively correlated.** Two variables (x, y) are positively correlated if as x increases, y tends to increase, whereas as x decreases, y tends to decrease.

(2) If the correlation is less than 0, such as for pulse rate and age, then the variables are said to be **negatively correlated.** Two variables (x, y) are negatively correlated if as x increases, y tends to decrease, whereas as x decreases, y tends to increase.

(3) If the correlation is exactly 0, such as for birthweight and birthday, then the variables are said to be **uncorrelated.** Two variables (x, y) are uncorrelated if there is no linear relationship between x and y.

Thus the sample correlation coefficient provides a *quantitative* estimate of the dependence between two variables: the closer $|r|$ is to 1, the more closely related the variables are; if $|r| = 1$, then one variable can be predicted exactly from the other.

Interpretation of various degrees of correlation



As was the case for the population correlation coefficient (ρ), interpreting the sample correlation coefficient (r) in terms of degree of dependence is only correct if the variables x and y are normally distributed and in certain other special cases. If the variables are not normally distributed, then the interpretation may not be correct.

Relationship Between the Sample Correlation Coefficient (r) and the Population Correlation Coefficient (ρ)

We can relate the sample correlation coefficient r and the population correlation coefficient ρ more clearly by dividing the numerator and denominator of r by $(n-1)$ in Definition 11.17, whereby

$$r = \frac{L_{xy}/(n-1)}{\sqrt{\left(\frac{L_{xx}}{n-1}\right)\left(\frac{L_{yy}}{n-1}\right)}}$$

Useful exercise !

We note that $s_x^2 = L_{xx}/(n-1)$ and $s_y^2 = L_{yy}/(n-1)$. Furthermore, if we define the *sample covariance* by $s_{xy} = L_{xy}/(n-1)$, then we can re-express Equation 11.16 in the following form.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\text{sample covariance between } x \text{ and } y}{(\text{sample standard deviation of } x)(\text{sample standard deviation of } y)}$$

This is completely analogous to the definition of the population correlation coefficient ρ given in Definition 11.16 with the population quantities, $Cov(X, Y)$, σ_x, and σ_y replaced by their sample estimates s_{xy}, s_x, and s_y.

Relationship Between the Sample Regression Coefficient (b) and the Sample Correlation Coefficient (r)

The relationship between the sample regression coefficient (b) and the sample correlation coefficient (r) is given in Equation 11.18.

$$b = \frac{rs_y}{s_x}$$

HERE DISCUSS

- Intro to Principal Component Analysis