# Probabilities and Statistics - Lecture 13

Olariu E. Florentin

June 3, 2024

# Table of contents

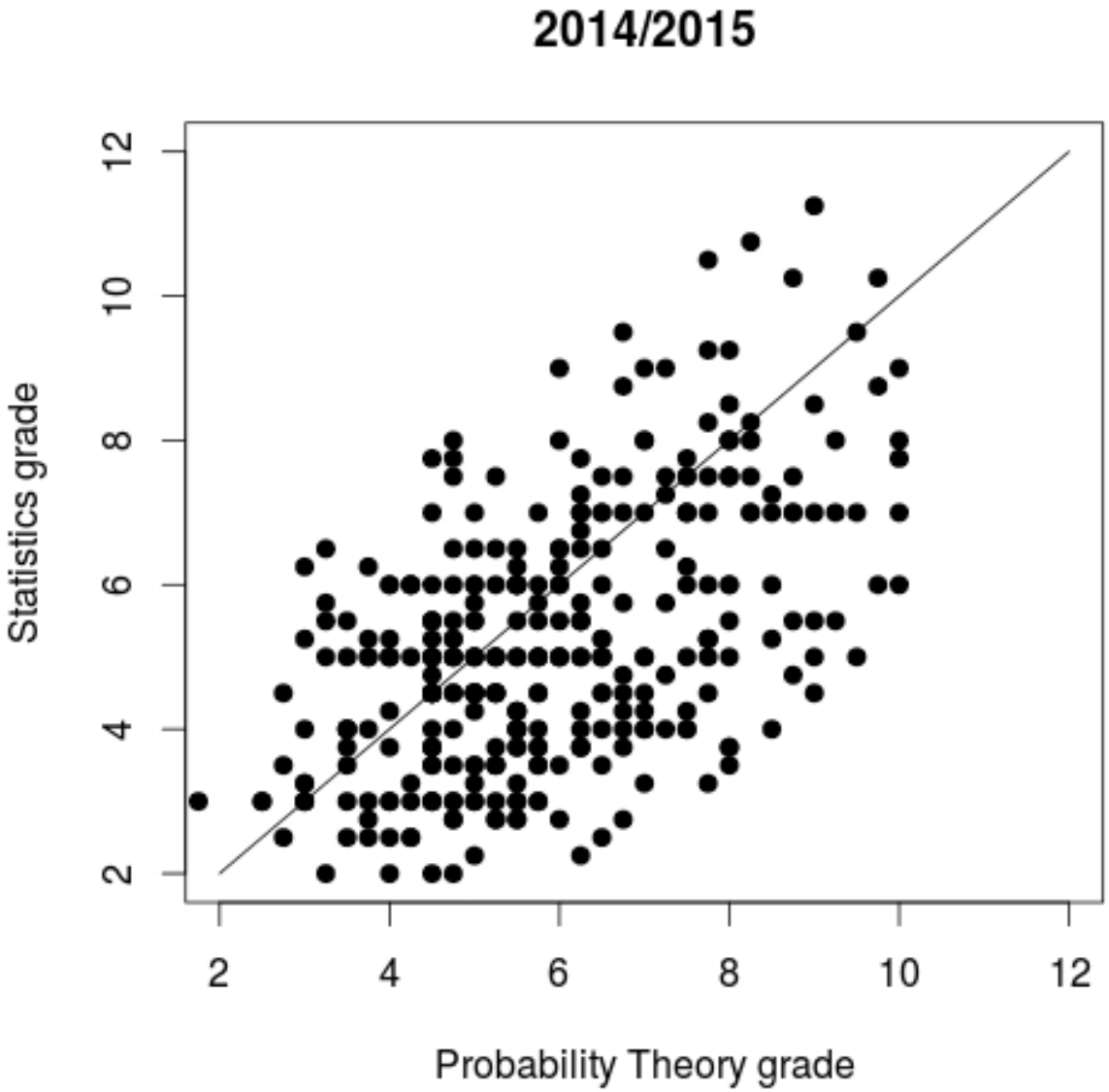
# Linear Correlation

- The *correlation* is a method for studying the relationship between two variables, it is part of *bivariate statistics*.
- The first statistician who made real progress on this matter was Francis Galton by studying the degree to which children resemble their parents.
- It was a trend of the time to study the hereditary influences by means of statistical and mathematical tools.
- As part of a study carried by Karl Pearson (a Galton's disciple), the heights of 1078 fathers and their sons at maturity were measured.
- The relationship betwen the two variables (father's and son's height) can be brought out in a scatter diagram.

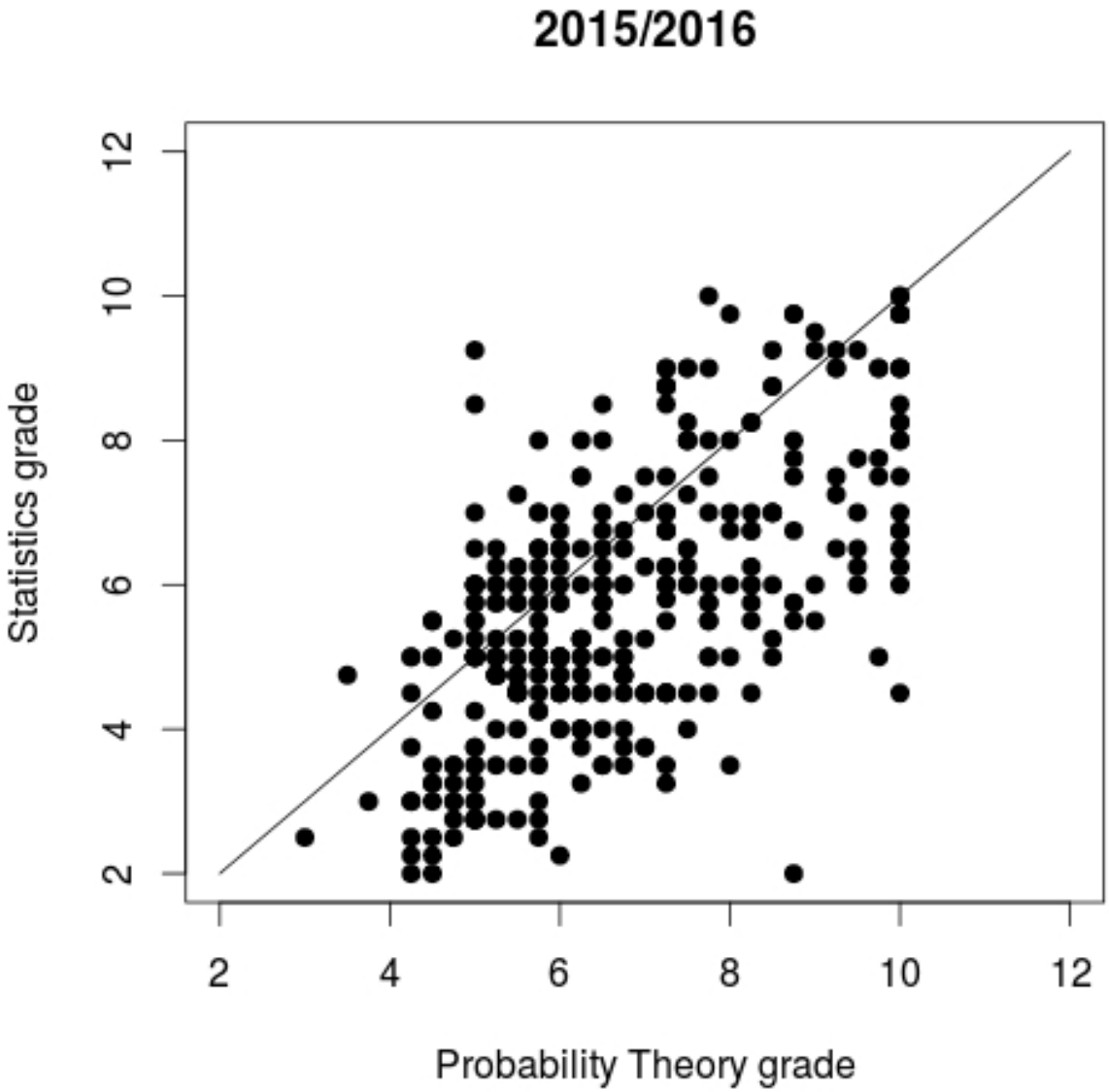
# Correlation - an example

- This time we will carry another study concerning the relationship between the grades obtained by freshmen at FII on the half-semester examination (at Probability Theory) and at the final session examination (at Statistics).
- The grades of 335 students from 2014/2015 year and 355 students from 2015/2016 year who take both exams are scattered in following figures.
- Each dot represents one pair of grades: the $y$-coordinate gives the Statistics grade and the $x$-coordinate gives the Probability Theory grade.
- We plot also the 45-degree line; this line corresponds to that students which have the same grade on both exams.
- If for a student its Probability's grade on exam is close to his Statistic's grade, then the corresponding point on the scatter diagram will be close to this line.

# Correlation - an example



# Correlation - an example



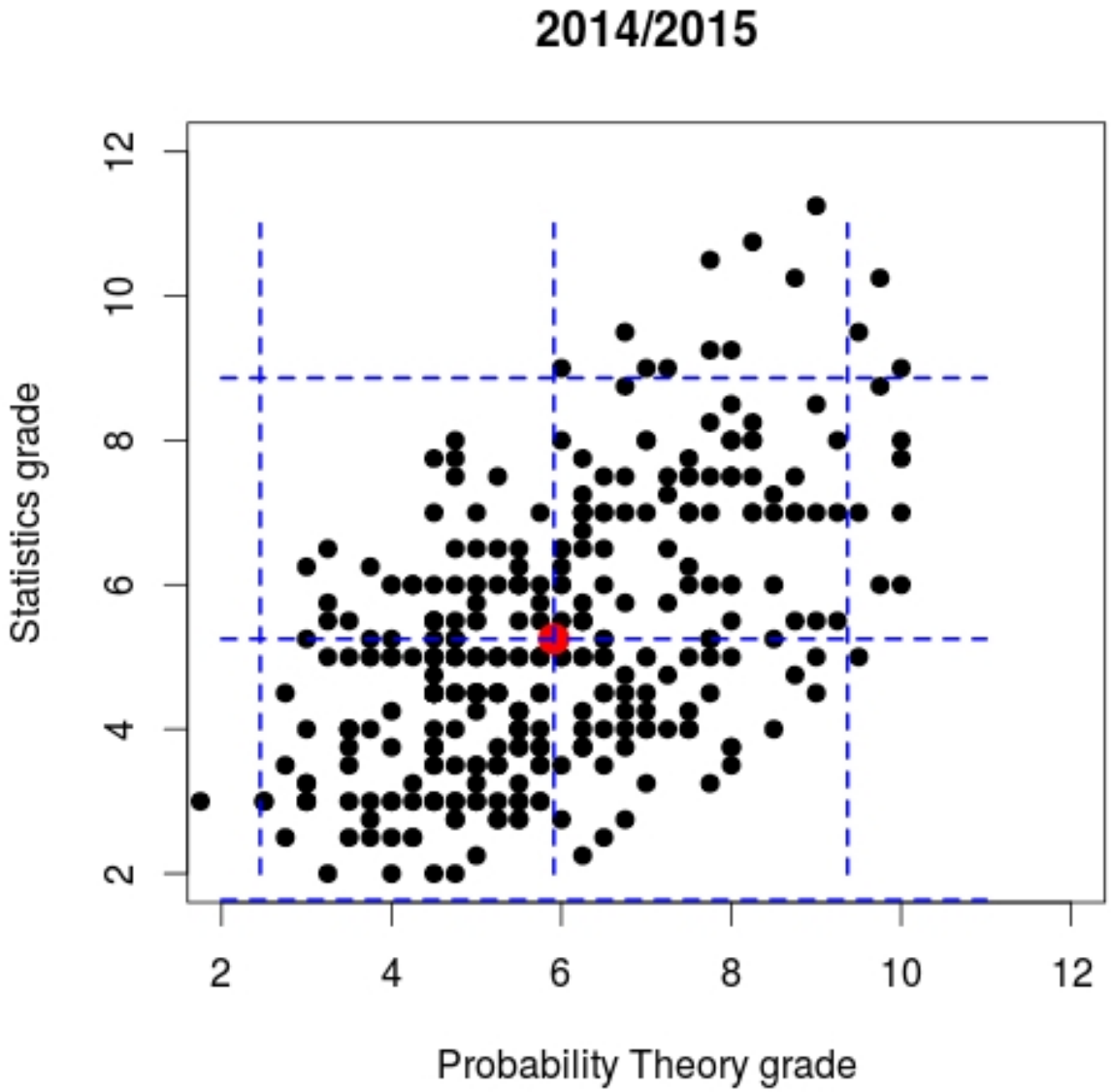
# Correlation - an example

- The points below the 45-degree line are the points corresponding to students that have a better Probability grade: this is the area where most of the points are.
- There is alot more spread around 45-degree line which shows a relationship not very strong between the two grades.
- If there is a very strong association between the two variables, then knowing one helps a lot in predicting the other.
- But when there is a weak association, information about one variable doesn't help much in predicting the other.
- Our problem is to try to guess (to predict) the Statistics grade from the Probability grade.

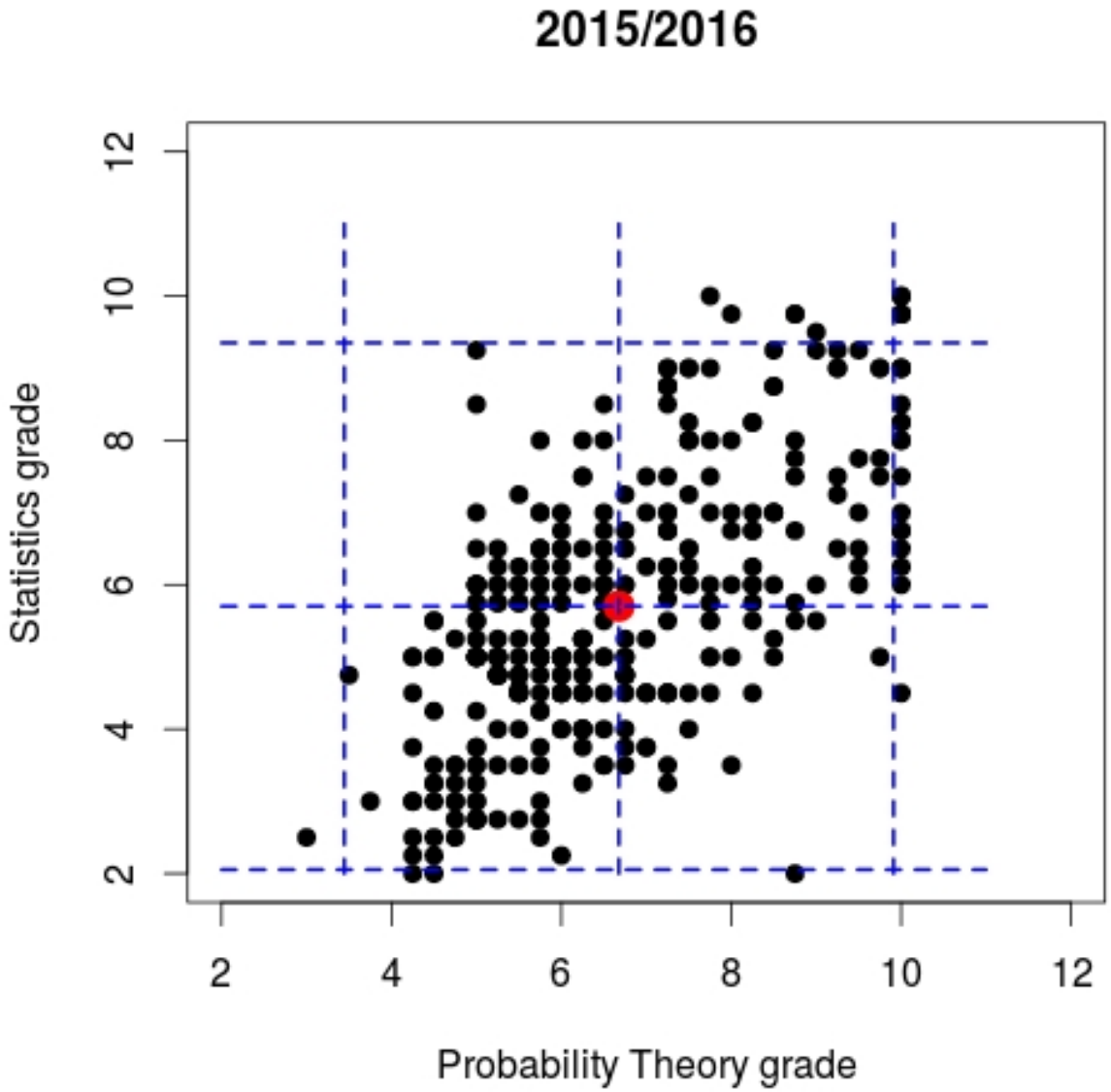
# Correlation - an example

- The scatter diagram is more like an ellipse-shaped cloud of points. How we can summarize the relationship between the two variables?
- The first step is to mark a point showing the average of $x$-values and the average of $y$-values: this is the *point of averages* located in the center of the cloud.
- The second step would be to measure the spread of the cloud from side to side. This would be done using the standard deviations of the two samples.
- Most of the points will be within 2 standard deviations (vertically or horizontally).
- These statistics don't show the strength of the *association* between the two variables.

# Correlation - an example



# Correlation - an example



# The correlation coefficient

- The most common statistic that measures the dependence between two variables is the correlation coefficient or the Pearson product-moment correlation coefficient.
- For two random variables $X$ and $Y$ the correlation coeficient is

$$\rho[X, Y] = \frac{cov[X, Y]}{\sigma_X \sigma_Y},$$

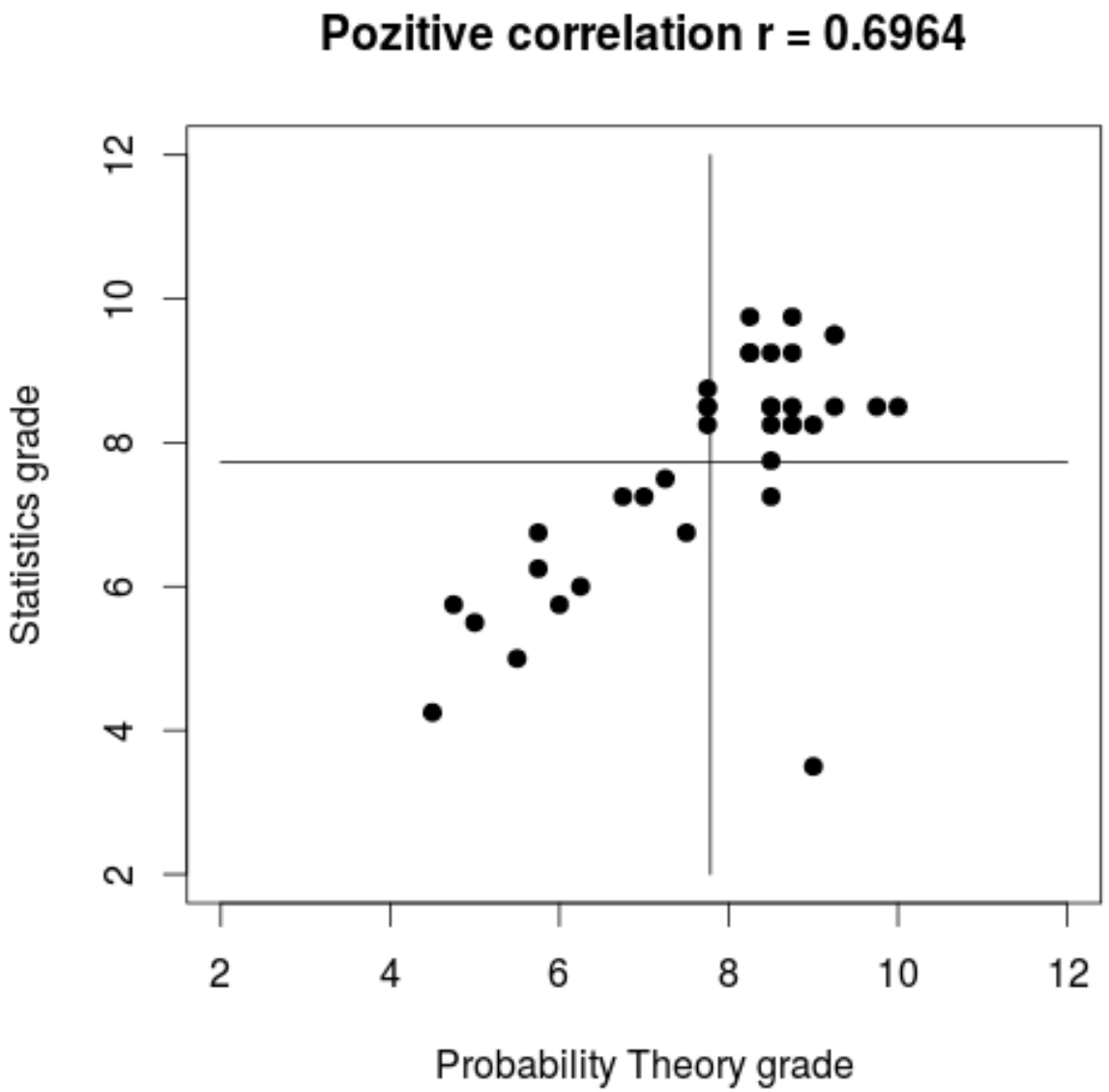
where $\sigma_X^2$ and $\sigma_Y^2$ are the variances of the two variables.

- For two random samples $x = \{x_1, x_2, \ldots x_n\}$ and $y = \{y_1, y_2, \ldots y_n\}$ is

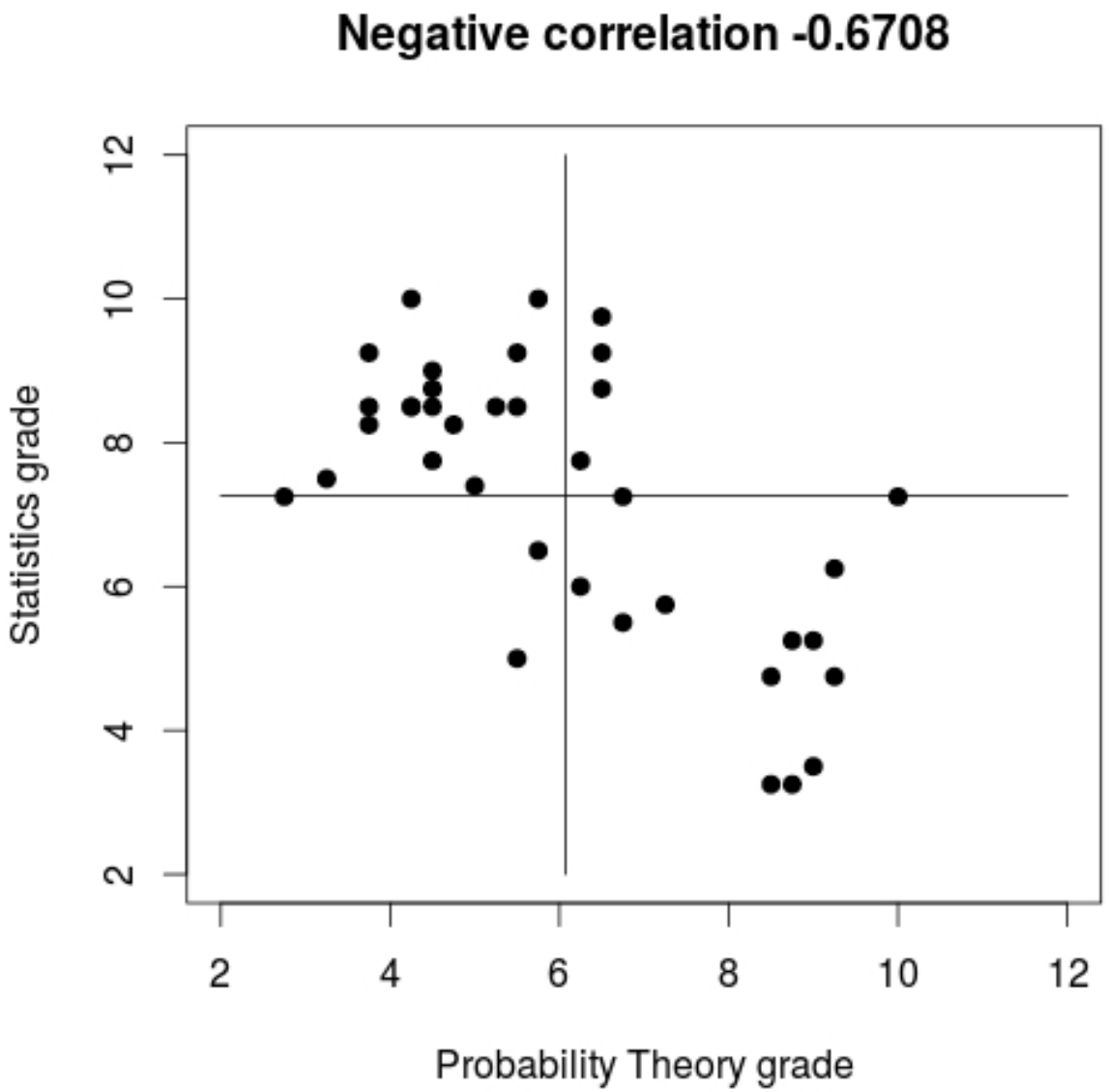
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right]\left[\sum_{i=1}^{n}(y_i - \overline{y}_n)^2\right]}},$$

where $\overline{x}_n$ and $\overline{y}_n$ are the sample means.

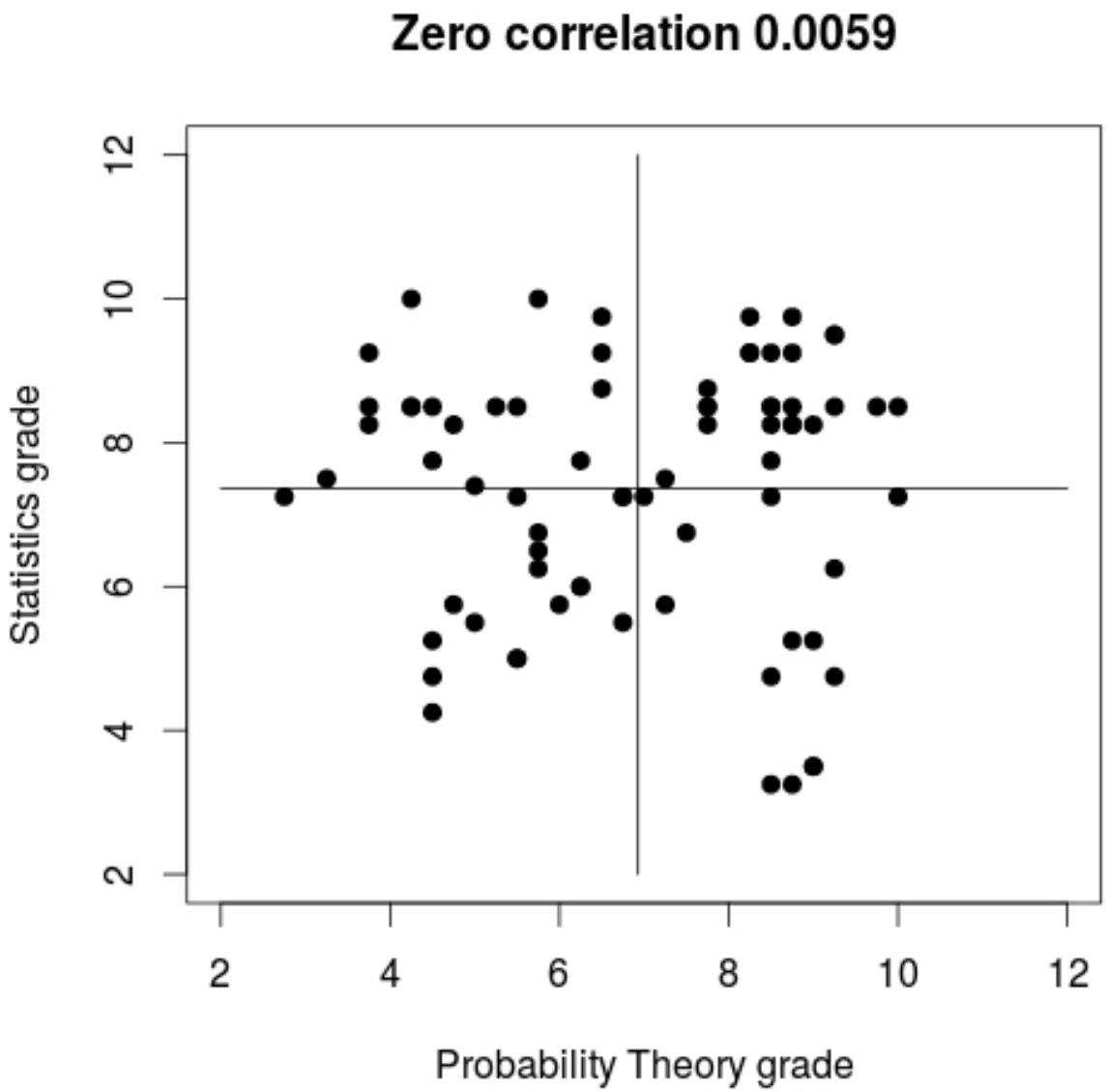
# The correlation coefficient



# The correlation coefficient



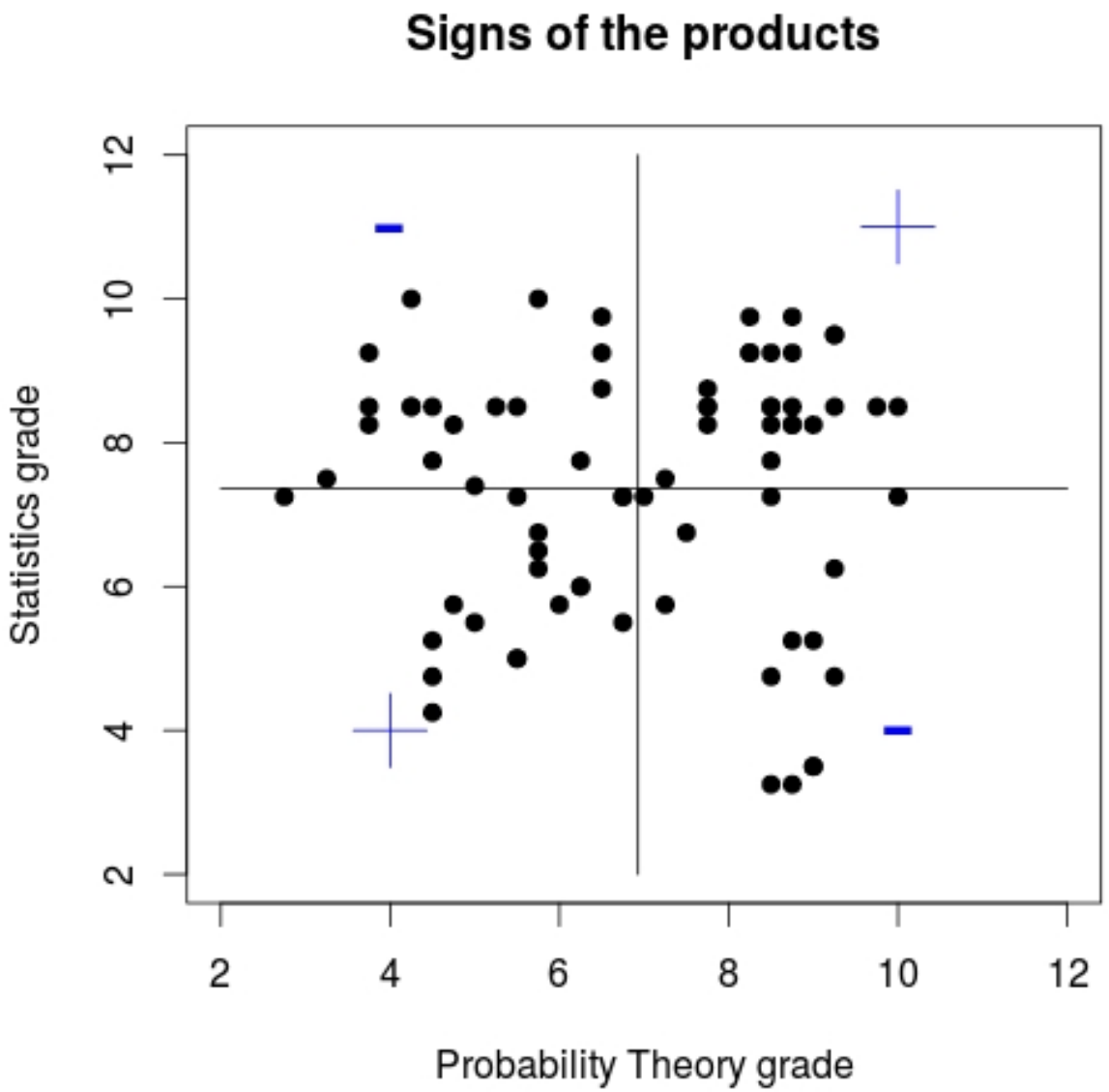
# The correlation coefficient



# The correlation coefficient

- How the correlation coefficient work as a measure of association?
- The two lines drawn through the point of averages divide the diagram in four quadrants:
  - in the lower left quadrant both variables are below average: $(x_i - \overline{x}_n)(y_i - \overline{y}_n) > 0$;
  - in the upper right quadrant both variables are above average: the product will be positive too;
  - in the lower right quadrant the $x$-variable is above and the $y$-variable is below average: $(x_i - \overline{x}_n)(y_i - \overline{y}_n) < 0$;
  - in the remaining quadrant $y$-variable is above and the $x$-variable is below average: the product will be negative also.

# The correlation coefficient



# The correlation coefficient

- The average of all these products is the correlation coefficient; if $r$ is negative, points in the two negative quadrants will predominate; if f $r$ is positive, will predominate points in the two positive quadrants.
- The correlation coefficient is not affected by
  - interchanging the two variables;
  - adding the same number to all values of one variable;
  - multiplying all the values of one variable by the same positive number.
- The correlation coefficient has values between $-1$ and 1; values close to zero usually means a very weak linear association.

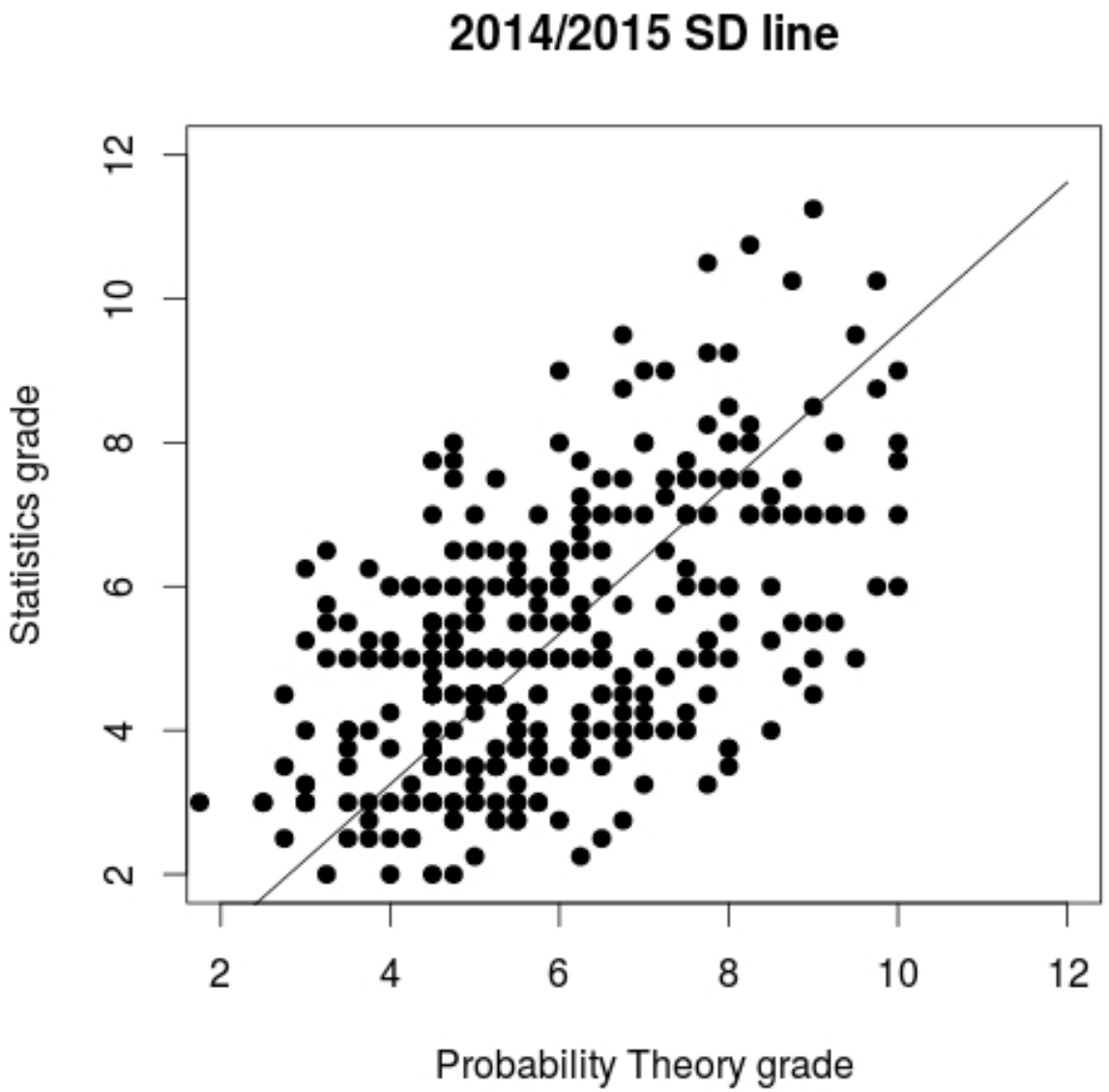
## The SD line

- The points in a scatter diagram generally seem to cluster around the *standard deviation line* (*SD line*).
- The SD line goes through the point of averages, and it goes through all points which are an equal number of SD's away from average.
- In other words it has a slope equal (in absolute value) with the fraction between the standard deviation of the $y$-values and the standard deviation of the $x$-values: $m = s_Y/s_X$ for positive correlation, and $m = -s_Y/s_X$, for negative correlation.
- The equation of this line is

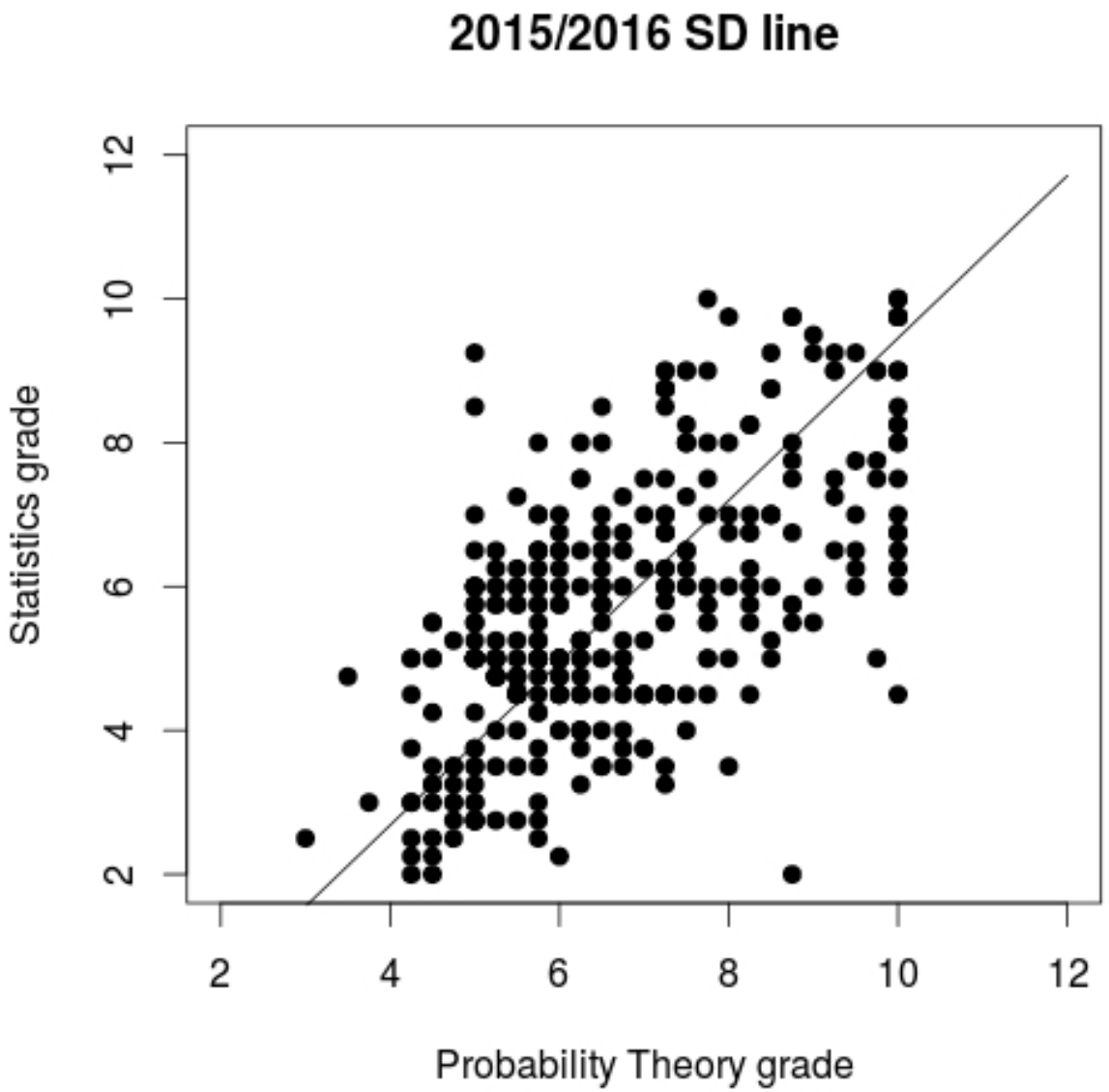
$$y - \overline{y}_n = m(x - \overline{x}_n),$$

where $\overline{x}_n$ and $\overline{y}_n$ are the sample means.

# The SD line


2014/2015 SD line
Statistics grade
Probability Theory grade
2
4
6
8
10
12

# The SD line


2015/2016 SD line
Statistics grade
Probability Theory grade
2
4
6
8
10
12

# Summary

- When the scatter diagram is tight clustered to the SD line there is a strong *linear association* between the variables.
- The scatter diagram can be summarized using five statistics
  - the sample mean and the sample standard deviation of the $x$-values;
  - the sample mean and the sample standard deviation of the $y$-values;
  - the correlation coefficient.
- *Positive association* is indicate by a plus-sign of the correlation coefficient or by the slope of the cloud going up.
- *Negative association* is indicate by a minus-sign of the correlation coefficient or by the slope of the cloud going down.
- The correlation coefficient ranges from $-1$ (when all points lie on a line which slopes down) to $+1$ (when all points lie on a line which slopes up).

# Summary

- The perfect positive (negative) association, $r = +1$ ($r = -1$), means that between the two variables there is a linear dependence with a positive (negative) slope:

$$Y = mX + n,$$

$m > 0$ for positive association and $m < 0$ for negative association.

- If $|r|$ is close to 1, then a typical point is only a small fraction of an $y$-standard deviation (respectively $x$) above or below (respectively to the left or to the right) of the SD line.
- The relation between the correlation coefficient and the typical distance above or below the SD line can be expressed mathematically.
- The spread around the SD line is about $\sqrt{2(1-|r|)} \cdot s_Y$ of a vertical SD. Horizontally the spread around the SD line is about $\sqrt{2(1-|r|)} \cdot s_X$.

# Summary

- The correlation coefficient is an useful statistics for ellipse-shaped scatter diagrams, because for other types of diagrams, the correlation can be misleading.
- This behaviour can be caused by outliers or by other, non-linear, association.
- The correlation coefficient measures the linear association, not association in general.
- We return to our example: the correlation coefficient of the 2014/2015 year is 0.5418, and that of the 2015/2016 year is 0.6313.
- There are very few outlier-like values, hence the correlation coefficient is a good measure for a prezumtive linear association.

# Summary

- The spread around the SD line for the year 2014/2015 is 165% vertically, and 173% horizontally. The spread around the SD line for the year 2015/2016 is 138% vertically, and 156% horizontally.
- Definitely the two years have a positive correlation, but with a large spread spread around the SD line.
- We can, however, detect a greater positive correlation for the second year with much smaller spread around the SD line.
- The SD slopes may indicate that the trend is for similar grades on the two exams.
- We can consider that there is a linear association between the two grades, although is not very strong.

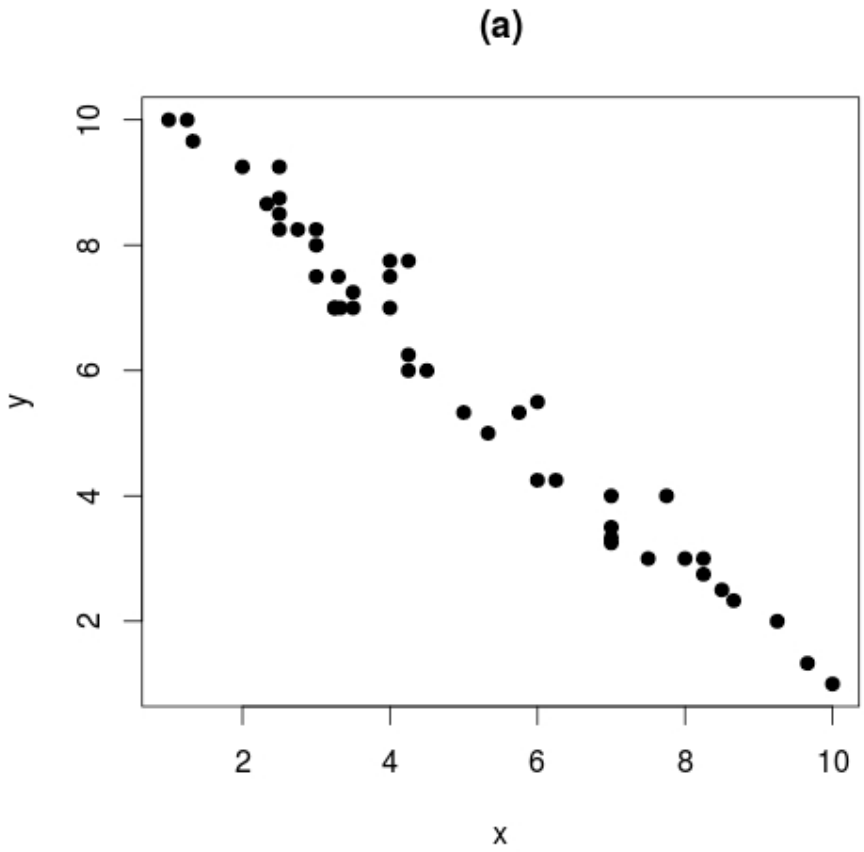
# Correlation - Exercises

1. Suppose men always marry women who were exctly 8% shorter. What would be the correlation between their heights be?
2. For a representative sample of cars, would the correlation between the age of a car and its fuel consumption be positive or negative?
3. The figures below contain four scatter diagrams for hypothetical data. The correlation coefficients, in a scrambled order, are

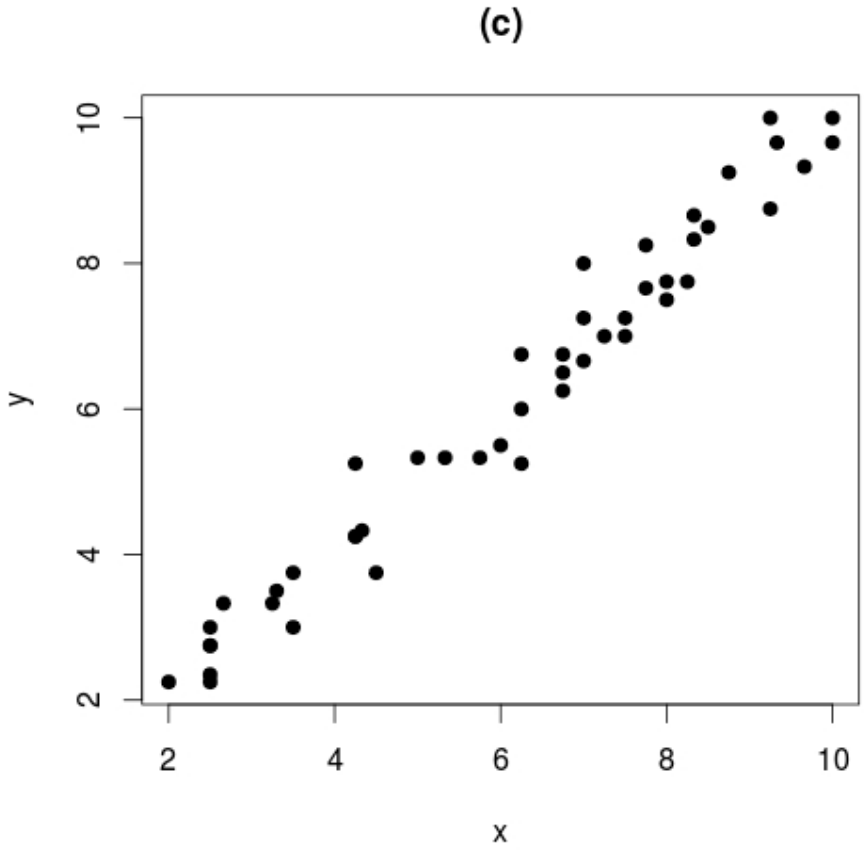
$$1 \quad -0.9833 \quad 0.9829 \quad -0.0760$$

Match the scatter diagrams with the correlation coefficients.

# Correlation - Exercises




# Correlation - Exercises




# Regression

- If the correlation tries to detect linear association between two variables, regression describes how one variable depends on another.
- There are two *regression lines*: the regression line for $y$ on $x$ that estimates the average value for $y$ corresponding to each value of $x$, and the regression line for $x$ on $y$. We will mainly discuss the first type of regression.
- The regression method can be stated as follows: *associated with each increase of one standard deviation (SD) in $x$ there is an increase of only $r$ standard deviations (SD) in $y$, on the average*.
- The linear regression method finds a line which "best fits" all the point in the scatter diagram, by going through the point of averages.

# Regression line

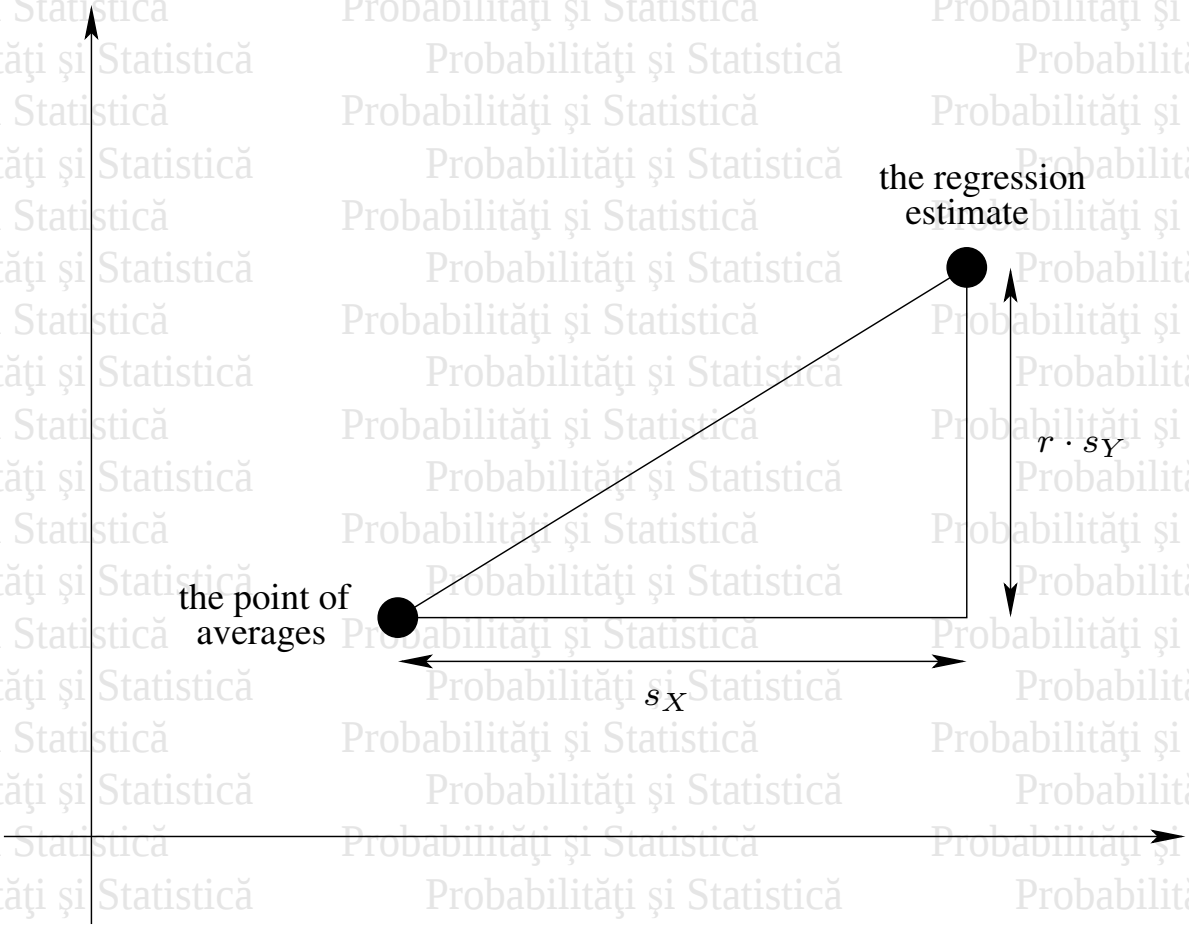



Figure: Regression estimate

# Regression line

- More formally, if the line is $y = px + q$, then $p$ and $q$ must minimize the following sum of square distances from all points (the best fit is found by the least-square method):

$$\sum_{i=1}^{n} (y_i - px_i - q)^2$$

- The solution to this minimization problem is

$$p = rs_y/s_x, q = \overline{y}_n - \overline{x}_n rs_y/s_x.$$

- Knowing this line we can predict the value of one variable from another.
- Regression lines should not be used where there is a non-linear association between variables: if there is a non-linear association between the two variables the regression line will passed it by.

# Regression line - examples

- We go back to our Probability/Statistics grades
- For the 2014/2015 year:

$$\overline{x}_n = 5.9116, \overline{y}_n = 5.2492, s_X = 1.7281, s_Y = 1.8082, r = 0.5418$$

$$\text{The regression line is } Y = 0.5669X + 1.8978.$$

- If we take at random a student from that year, and he had a grade of 5.25 at the Probability theory exam, than we can predict that his Statistics grade was 4.8215.
- For the 2015/2016 year:

$$\overline{x}_n = 6.6772, \overline{y}_n = 5.7029, s_X = 1.6163, s_Y = 1.8243, r = 0.6313$$

$$\text{The regression line is } Y = 0.7125X + 0.9451.$$

- If we take at random a student from that year, and he had a grade of 5.00 at the Statistics exam, than we can predict that his Probability grade was 5.6910.

# Regression - Exercises

1. An university has made a statistical analysis of the relationship between Math SAT scores (ranging from 200 to 800) and the first-year GPAs (Grade Point Average, ranging from 0 to 40, for students which complete the first year). The resuts are:

$$\text{average Math SAT score } = 550, s = 80$$

$$\text{average first-year GPA score } = 2.6, s = 0.6, r = 0.4$$

The scatter diagram is ellipse-shaped. A student is chosen at random and has an SAt of 650. Predict his first-year GPA.

2. An instructor standardizes her midterm and final exams so the class average 50 and the standard deviation is 10 on both tests. The correlation between the tests is 0.60. Estimate the score for the second test of a student which scored below 30 on the first test. (The diagram is ellipse-shaped.)

# Regression - Exercises

3. In a study of the stability of IQ scores, a large group of individuals is tested once at age 18 and again at age 35. The following results are obtained

$$\text{age 18: average score} = 100,\ s = 15$$

$$\text{age 35: average score} = 100,\ s = 15,3 = 0.80$$

Estimate the score at age 35 for an individual who scored 115 at age 18. (The diagram is ellipse-shaped.)

4. In a study of 1000 families:

$$\text{average height of husband} = 68in, s = 2.7in$$

$$\text{average height of husband} = 63in, s = 2.5in, r = 0.25$$

Predict the height of a wife when the height of her husband is (a) 72 inches; (b) 64 inches; (c) 68 inches. (The diagram is ellipse-shaped.)

# Bibliography


- Freedman, D., R. Pisani, R. Purves, *Statistics*, W. W. Norton & Company, 4th edition, 2007.
- Johnson, R., P. Kuby, *Elementary Statistics*, Brooks/Cole, Cengage Learning, 11th edition, 2012.