AI Booster – Week 02
Session 03 – Bivariate stats

Franck JAOTOMBO & Antoine SCHERRER
ascherrer@em-lyon.com

Outline & Program

em lyon business school

- One week dedicated to improve your python programming skills and review basic statistical notions
- Day 1 => Introduction, data, data cleaning
- Day 2 => Univariate statistics
- Day 3 (today!) => Bivariate statistics
- Day 4 => Hypothesis testing and important distributions
- Day 5 => Review linear algebra

Teaching / learning materials

em lyon business school

- Every day will follow the same schedule
- 1h30 of lecture (or less)
- 1h30 of in-class pratice (live coding session)
- Afternoon dedicated to practice (Tues., Wed., Thur. With a teaching assistant)
- Evaluation => individual quizz at the end of the week + group project (at the end of week 3)

Representing data with tables

Tables and Frequency Distributions

Summarizing two categorical variables : Contingency Table

- A random sample of 400 invoices is drawn.
- Each invoice is categorized as a small, medium or large amount.
- Each invoice is also examined to identify if there are any errors.
- This data are then organized in the contingency table to the right.

Contingency Table Showing Frequency of Invoices Categorized By Size and The Presence Of Errors

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

Contingency Table Based On Percentage Of Overall Total

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

42.50% = 170 / 400
25.00% = 100 / 400
16.25% = 65 / 400

	No Errors	Errors	Total
Small Amount	42.50%	5.00%	47.50%
Medium Amount	25.00%	10.00%	35.00%
Large Amount	16.25%	1.25%	17.50%
Total	83.75%	16.25%	100.0%

83.75% of sampled invoices have no errors and 47.50% of sampled invoices are for small amounts.

Contingency Table Based On Percentage of Row Totals

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

89.47% = 170 / 190
71.43% = 100 / 140
92.86% = 65 / 70

	No Errors	Errors	Total
Small Amount	89.47%	10.53%	100.0%
Medium Amount	71.43%	28.57%	100.0%
Large Amount	92.86%	7.14%	100.0%
Total	83.75%	16.25%	100.0%

Medium invoices have a larger chance (28.57%) of having errors than small (10.53%) or large (7.14%) invoices.

Contingency Table Based On Percentage Of Column Total

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

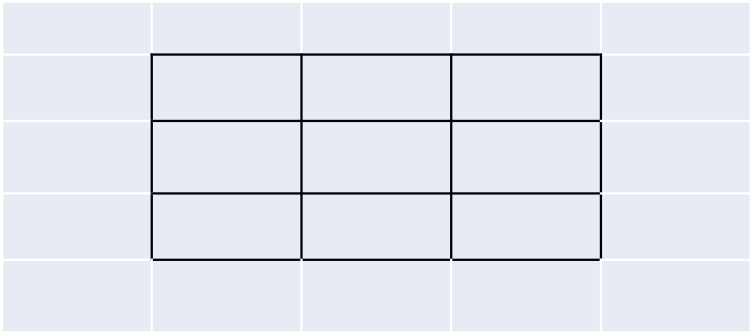
50.75% = 170 / 335
30.77% = 20 / 65

	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%

61.54% of invoices with errors are of medium size.

Investigate if variable are independant

- Consider a contingency table

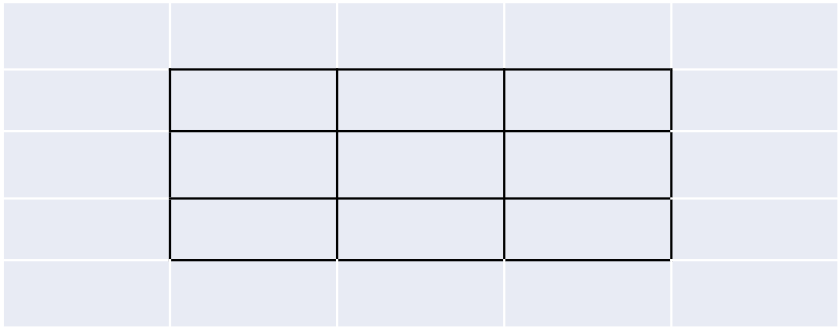


- indicates the actual (observed) frequency of the cell on row (line) and column
- indicates the expected value under assumption of independence.
- measures how much frequency are different from expected value under assumption of independence

The formula

- If the row and column variables are independent:

$$X^2 = \sum_{i,j} \frac{\left[n_{ij} - \frac{l_i c_j}{n} \right]^2}{\frac{l_i c_j}{n}}$$



  - and indicate respectively the total (marginal) frequency of row and column
  -

- Effect size :

Effect size

- Measure the strength of the relationship between two variables
  - For categorical variables
  - For quantitative variables : **|r|**
- The effect size of these indices are given as following (rule of thumb) :
  - Under 0.1 : not significant
  - Between 0.1 and 0.3 : small (if significant)
  - Between 0.3 and 0.5 : moderate (if significant)
  - Between 0.5 and 0.7 : strong
  - Between 0.7 and 0.9 : very strong
  - Beyond 0.9 : colinearity or identity
    - One of the two variables should be removed from the analyses

Effect size

- Measure the strength of the relationship between two variables
  - For categorical variables
  - For quantitative variables : **|r|**
- The effect size of these indices are given as following (rule of thumb) :
  - Under 0.1 : not significant
  - Between 0.1 and 0.3 : small (if significant)
  - Between 0.3 and 0.5 : moderate (if significant)
  - Between 0.5 and 0.7 : strong
  - Between 0.7 and 0.9 : very strong
  - Beyond 0.9 : colinearity or identity
    - One of the two variables should be removed from the analyses

Covariance and correlation

How do we describe how much 2 variables are correlated?

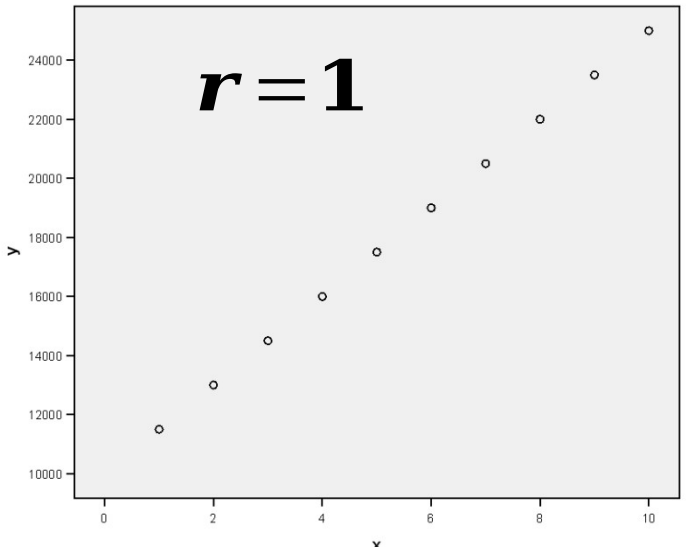
Covariance and correlation

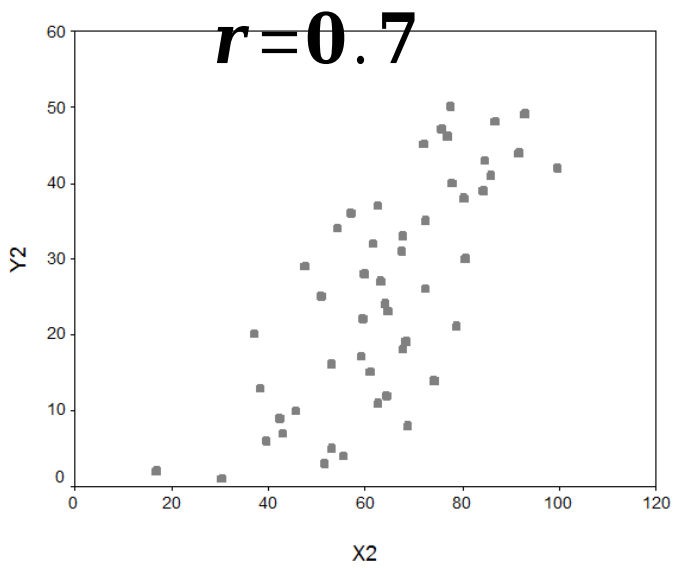
- The covariance between two variables X and Y indicates if there is an association between the variation of the two variables around their respective means

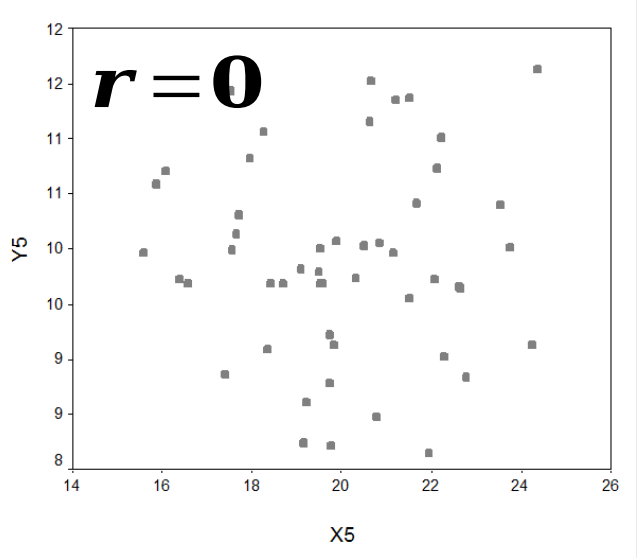
- Correlation is a standardized measure of the covariance

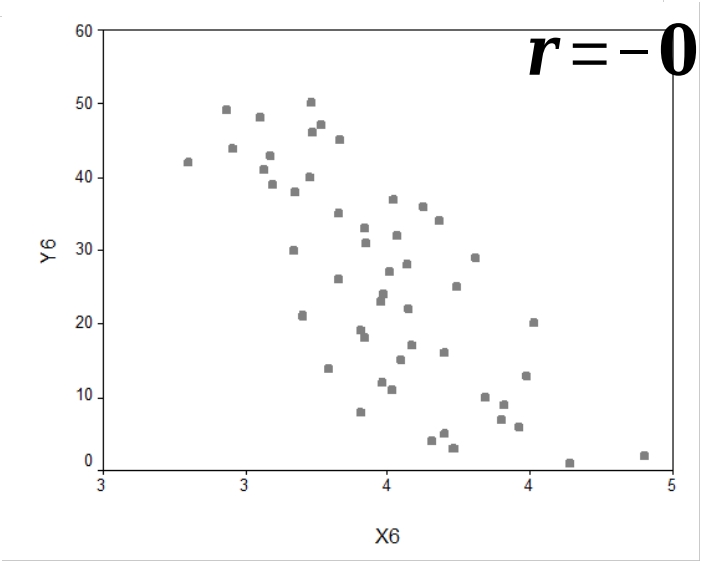
and are the respective standard deviations of X and Y

Examples (scatter plots)



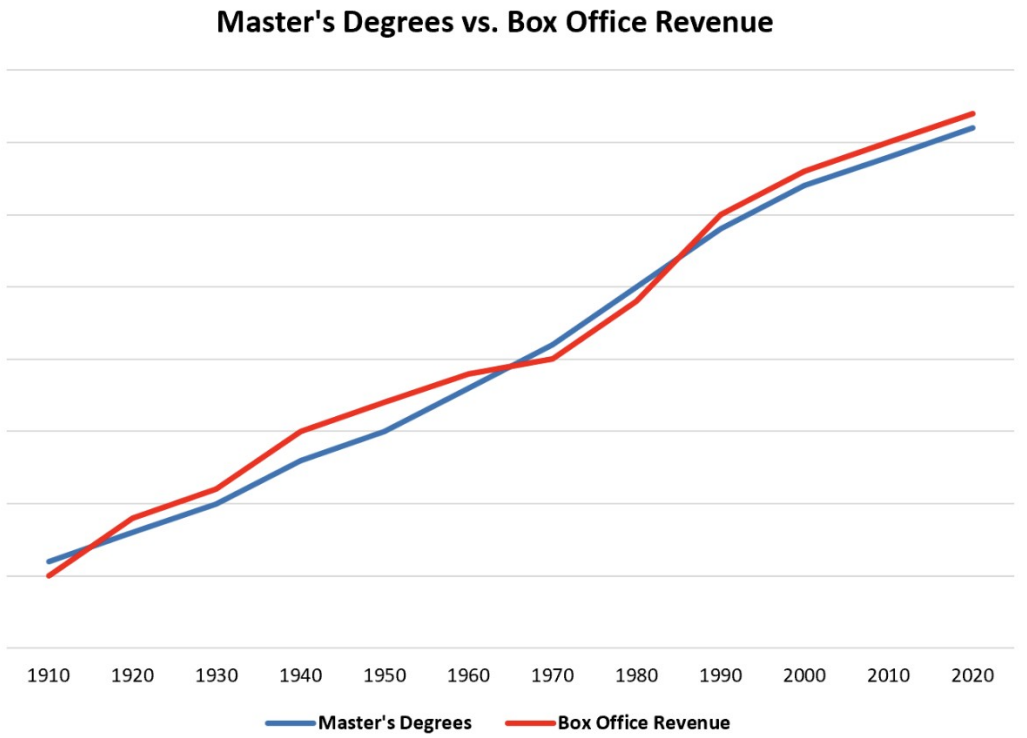






Correlation vs causation

- Correlation does not imply causation !!!
- Major pitfall in interpretations of statistics...



Any questions ? +
Let’s start coding !