Data Management

Comparison and correlation

Malka Guillot

HEC Liège | ECON2306

Motivation

- **Are larger companies better managed?**
- To answer such question, we need:
  - data (cf. **previously**)
  - **statistics**
    - summary measures? Interpretations?

In short: comparison & conditioning

- 2 variables:
  - x and y
- Objective:
  - Uncover the patterns of association between x and y
- We compare y, by x values
  - ie. we **condition** y on x (or y given x)
  - y= outcome variable

Table of contents

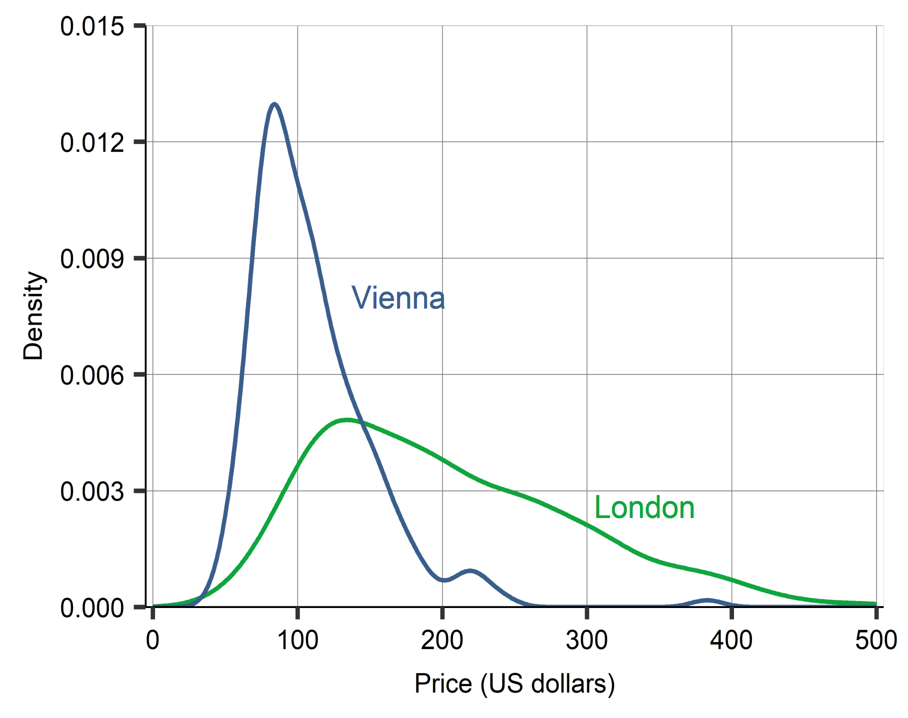
1. Conditional distributions
2. Exemple: Management quality and firm size
3. Correlation & Covariance

Comparisons and conditional distributions

- The conditional distribution of a variable is the distribution of the outcome variable given the conditioning variable.

If the conditioning variable is qualitative (or binary)

- Comparing histograms



Conditional statistic

- Conditional mean= mean of a variable for each value of the conditioning variable
- The conditional expectation of a variable y given x is:

$$E[y|x]$$

- This is a function
- In the case x is categorical:
  - for a value of x, the cond. exp. gives the expeted value (mean, average) of a y for observations that have that value of x

Conditional and joint distributions of 2 quantitative variables

- 2 variables ⇒ many values
- The joint distribution of 2 variables shows the probabilities (frequencies) of each value combination of the 2 variables.

⇒ Scatter plot

(binned) Scatter plot

- a 2D graph with the values of each of the 2 variables measured on its 2 axes
  - Scatter: Each dot correspond to 1 observation
  - Binscatter: averages of y by bins of x (based on quantiles)

Scatter	Binscatter
When dataset is *small*	For larger samples: we bin values

Management quality and firm size

- Management quality and firm size:
  - describing patterns of association
  - **Whether, and to what extent, larger firms are better managed?**
- Answering this question can help understand why some firms are better managed than others.
- Data from the World Management Survey

Measuring management quality

- Interviews by CEO/senior managers, based on that a score is given
- Each score is an assessment of management practices in a particular domain:
    - tracking and reviewing performance or
    - time horizon and breadth of targets, etc
- Measured on a scale of 1 (worst practice) to 5 (best practice).
- Management quality is = average of 18 scores.

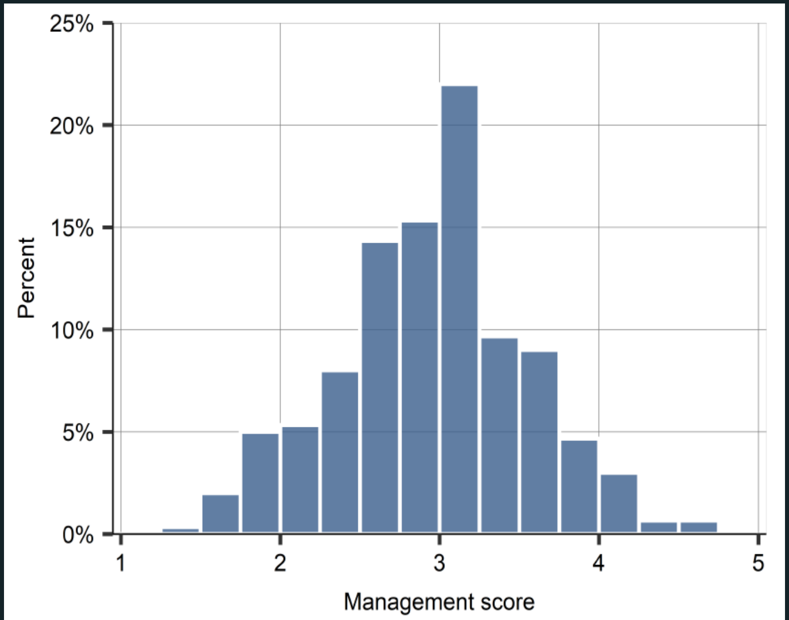
Management quality and firm size

Data

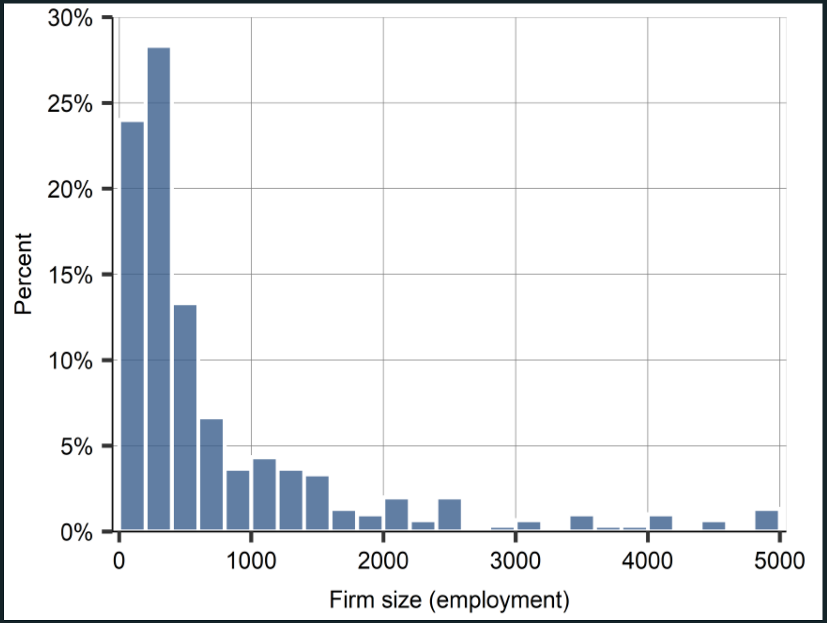
- Cross-sectionnal data of Mexican firms from the 2013 wage survey
- Sample: Only firms with 100-5000 employees ($N = 300$)
- $y =$ quality of management
- $x =$ firm size (number of employees)

Management quality and firm size

Histograms



(a) Management score



(b) Firm size (number of employees

Management quality and firm size

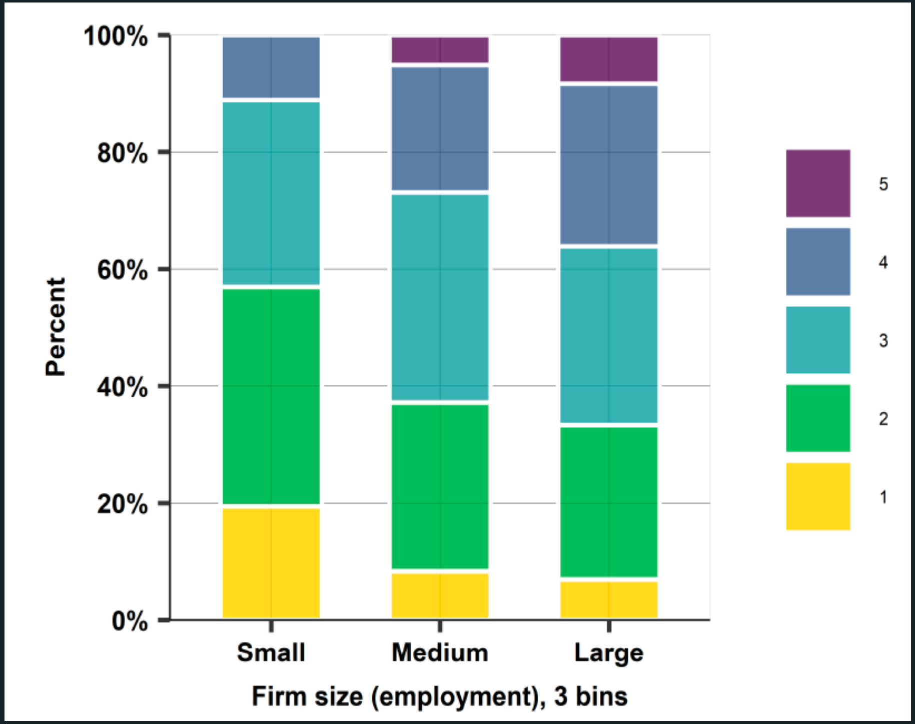
Conditional probabilities in data

- 3 bins of firm size
    - small: 100–199, N=72
    - medium: 200–999, N=156
    - large: 1000, N=72
- For each score variable we have 15 conditional probabilities
    - the probability of each of the 5 values of y by each of the three values of x
    - e.g. $P(y = 1|x = small)$

Management quality and firm size

Conditional probabilities

- Lean management score 1–5
- Firm size: small, medium, large
- Conditional probability:
  - $p(y = 1|x = small) = 20\%$.
  - $p(y = 5|x = large) = 10\%$.
- Shows a pattern of association



Management quality and firm size

Conditional statistic: conditional mean.

Mean given firm size:

- Mean management score is
  - For small firms: 2.68
  - For medium firms: 2.94
  - For large firms: 3.18
- First simple evidence:
  - Larger firms have better management

Management quality and firm size

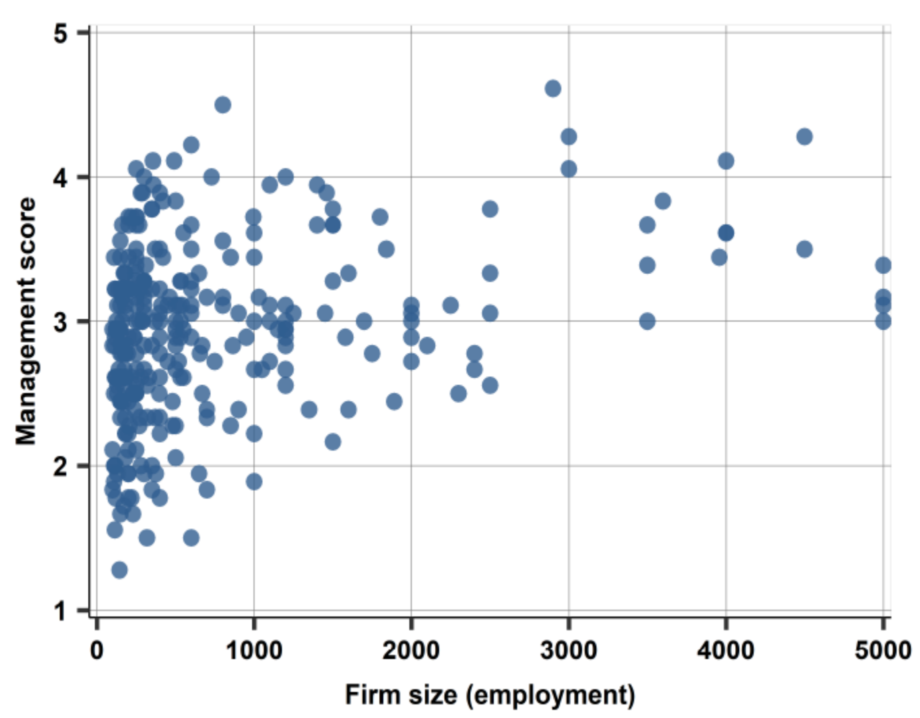
Joint distribution

- How is management quality related to the firm size?
  - $y =$ management score
  - $x =$ employment
- Graphical analysis:
  1. scatterplot
  2. bin scatter

Management quality and firm size

Scatterplot

- Both x- and y- axis quantitative
- Firm size: small, medium, large
- Each dot is an observation:
- Full information on association



Management quality and firm size

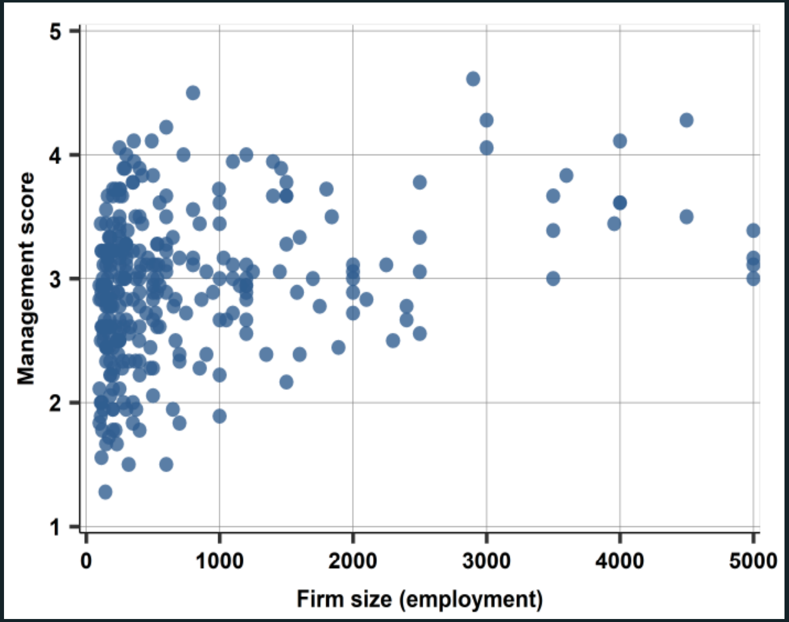
Bin scatter plot

1. Divide x into 10 bins with similar nb. of observations (deciles)
2. Calculate the mean of y conditional on the 10 bins of x.
3. Plot the previous average on the y-axis with bin values on the x-axis
    - i.e. Average management score as a point corresponding to the mean in the employment bin (e.g., 110 for the 100–120 bin).

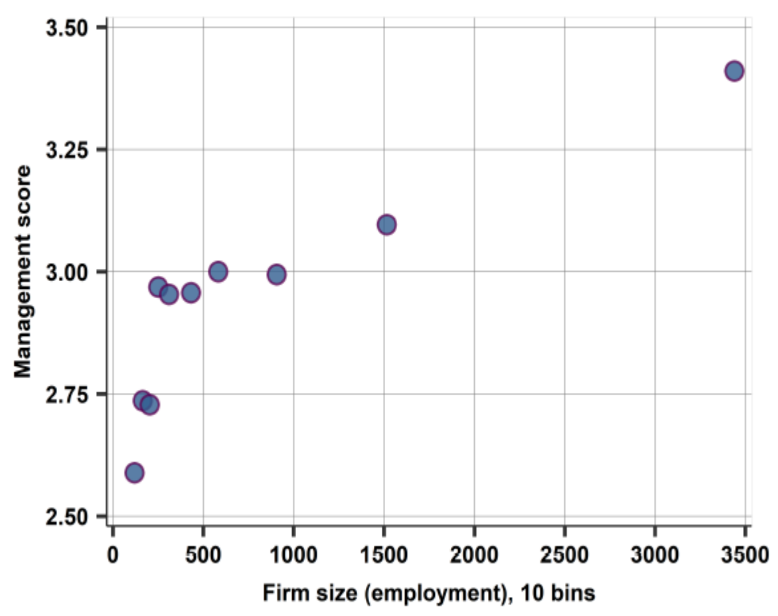
Dots NOT equally spread out - more frequent where more observations!

Management quality and firm size

Joint distributions



(a) Scatterplot



(b) 10 bin-scatter

Dependence and independence

- y is independent of x when the distribution of y does not depend on the conditionning on x
- y is dependent of x when the distribution of y depends on the conditionning on x
  - may take many forms

Mean Dependence

- mean-dependence:
  - conditional expectation $E[y|x]$ varies with the value of x.
  - the extent to which conditional expectations (means) differ.
- measured by covariance and correlation coefficient

Covariance

$$Cov(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Correlation coefficient

$$Corr(x,y) = \frac{Cov(x,y)}{Std(x)Std(y)}$$

- The correlation coefficient is the standardized version of the Covariance
- sum over the observations: $i = 1, \cdots, n$

Management quality and firm size

Correlation of management quality and firm size by industry

Industry	Correlation	# Observations
Auto	0.50	26
Chemicals	0.05	69
Electronics	0.33	24
Food, drinks, tobacco	0.05	34
Materials, metals	0.32	50
Textile, apparel	0.29	43
Wood, furniture, paper	0.28	29
Other	0.44	25
All	**0.30**	**300**

Summary: correlation?

- The correlation coefficient captures a simple measure of mean dependence.
- Qualitative variables:
  - Summarize conditional probabilities (frequencies).
- Quantitative variables:
  - Scatterplots offer a visual insight to the pattern of the relationship.