Lecture 4 : Comparison and Correlation

Sakib Anwar

AN7914 Data Analytics and Modelling

University of Winchester

2024

Motivation

Are larger companies better managed? Answering this question may help in benchmarking management practices in a specific company, assessing the value of a company, or estimating the potential benefits of a merger between two companies.

To answer this question you downloaded data from the World Management Survey. How should you use the data to measure firm size and the quality of management? How should you assess whether larger companies are better managed?

The Relationship Between y and x

- Data analysis often involves comparing the values of a dependent variable (y) across different values of one or more independent variables (x).
- The goal is to identify patterns of association—to understand if and how the values of x are associated with the values of y.
- The roles of y (dependent variable) and x (independent variable) are distinct:
  - We analyze the variation in y to understand outcomes or effects.
  - We observe these variations across different groups defined by x.
- Choosing y, the variable of interest, is a crucial step in our analysis.

Analysis Goals: Understanding y through x

- The focus on different variables stems from the specific objectives of our analysis.
- *Goal 1*: Use multiple independent variables (x_1, x_2, ...) to predict the value of a dependent variable (y). This prediction is particularly valuable when we have the values of the x variables but not the value of y.
- *Goal 2*: Investigate the influence of a specific independent variable (x) on a dependent variable (y) to understand the causal relationship. This involves exploring hypothetical scenarios of how y would change if we could manipulate x.

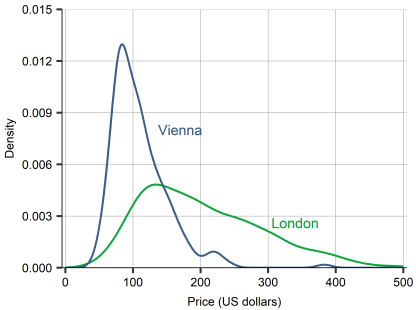
Comparison and Conditioning

- When we analyze the relationship between two variables, we condition the outcome variable (y) on the values of another variable (x), effectively examining y *given* x.
  - The variable x, which dictates the basis of our comparisons, is known as the conditioning variable.
  - The variable y, whose values we are interested in, is the outcome variable.

- For example, comparing hotel prices (y) across different cities (x) illustrates this concept:
  - Here, the price of the hotel represents the outcome variable.
  - The city in which the hotel is located acts as the conditioning variable.

Comparisons and Conditional Distributions

- The *conditional distribution* of a variable shows how the outcome variable's distribution varies, given specific values of the conditioning variable.
- This concept is particularly straightforward when the conditioning variable is qualitative, especially if it is binary.
- One common method to visualize these comparisons is through comparing histograms.



Understanding Conditional Statistics

- The *Conditional Mean* represents the average value of a variable within groups defined by each value of the conditioning variable.
- The *Conditional Expectation* of a variable y, given different values of a variable x, is expressed mathematically as:

$$E[y|x]$$

- This represents a function where, for any given value of x, it returns the expected value (mean or average) of y for observations corresponding to that x value.
- Hence, the conditional expectation varies with different values of the conditioning variable x, providing a detailed insight into how y changes with x.

Conditional Statistics Examples

- Students' Test Scores
  - $E[\text{test score}|\text{grade level} = 10]$: Average test score for 10th grade students.
- Employee Salaries
  - $E[\text{salary}|\text{department} = \text{Marketing}]$: Average salary in the Marketing department.
- Daily Temperatures
  - $E[\text{daily high}|\text{month} = \text{July}]$: Average daily high temperature in July.

Understanding Distributions of Two Quantitative Variables

- Exploring relationships between two variables involves analyzing numerous value combinations.
- The *joint distribution* maps out how often each pair of variable values occurs together, illuminating the pattern of their relationship.
- A *scatter plot* visualizes this relationship by plotting dots for each data point based on the values of these two variables, with one variable on each axis.
- This method is effective for datasets of manageable size, allowing for a clear visual representation of the data points.
- For larger datasets, an alternative approach involves grouping the data into bins and plotting *bin scatter*. This technique displays the conditional means within each bin, offering a simplified view of the data's distribution and trends.

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 to investigate our question.

Management quality and firm size

- Interviews by CEO/senior managers, based on that a score is given
- Management quality is measured as management score.
- Each score is an assessment by the survey interviewers 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 and firm size

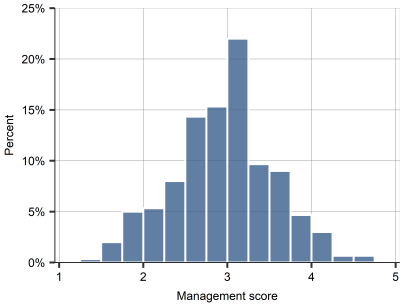
- Take 18 individual measures and average
- Measure of the quality of management is the simple average of these 18 scores = "the" management score.
- By construction, the range of the management score is between 1 and 5.

Management quality and firm size

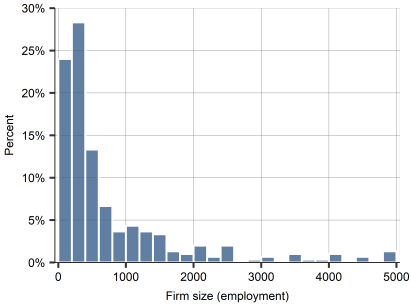
- Data from the World Management Survey to investigate our question.
- In this case study we analyze a cross-section of Mexican firms from the 2013 wave of the survey.
- Only firms with 100 – 5000 employees, N=300
- The y = measure of the quality of management. The x = measure of firm size.
- Firm size = number of employees

Management quality and firm size

(a) Management score



(b) Firm size (number of employees)



Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Management quality and firm size

- Management score: The mean is 2.9, the median is 3, and the standard deviation is 0.7.
- Firm size: The range of employment is 100 to 5000. The mean is 760 and the median is 350, skewness with a long right tail. Some large firms, but not extreme, kept as is.

Management quality and firm size

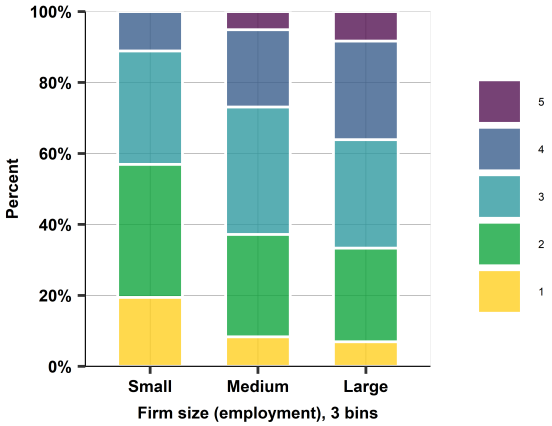
Conditional probabilities in data.

- Three bins of firm size. By number of employees: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- Take a single measure: Lean management score, with values 1,2,3,4,5.
- Thus, 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

- Lean management score 1–5
- Firm size: small, medium, large
- Conditional probability:
  - share of score=1 conditional on being a small firm is about 20%.
  - share of score=5 conditional on being a large firm is about 10%.
- Shows a pattern of association

Note: *Source: Management quality is an average score of 18 variables. Firm size is number of emoployees.* `wms-management-survey` *data. Mexican sample, n=300.*



Management quality and firm size

Conditional statistic - conditional mean.

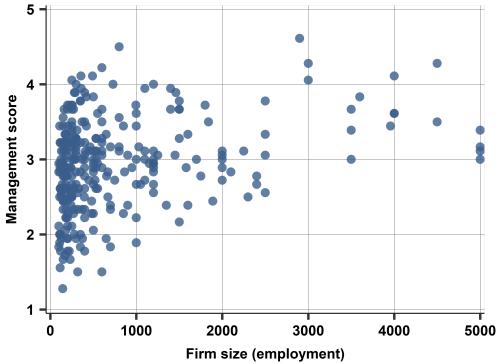
- Can calculate the mean given firm size.
- Three bins of employment: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- Mean management score is 2.68 for small firms, 2.94 for medium sized ones, and it is 3.18 for large.
- First simple evidence: larger firms have better management.

Management quality and firm size

- Conditional mean and joint distribution
- How our management quality variable
  - y: the management score

  is related to our firm size variable
  - x: employment
- Scatterplot
- Bin-scatter

Management quality and firm size



- Scatterplot
- Both x and y axis qualitative
- Each dot is an observation
- Full information on association

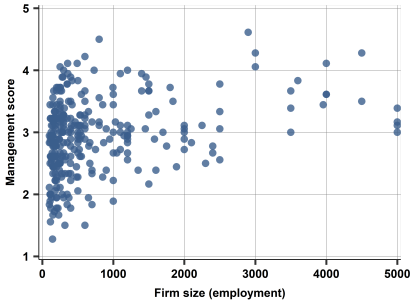
Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Management quality and firm size

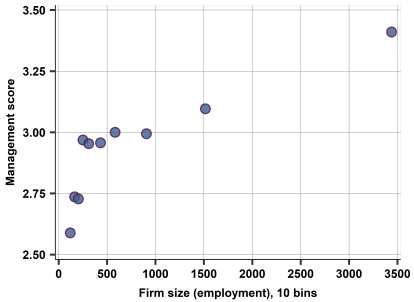
- Bin-scatter: calculate the mean of y conditional on ten bins of x.
- Bin-scatter: cut x's distribution into 10 parts, with equal number of firms. (remember - percentiles)
- Show average management score as a point corresponding to the midpoint in the employment bin (e.g., 110 for the 100–120 bin).
- Dots NOT equally spread out - more frequent where more observations!

Management quality and firm size

(a) Scatterplot



(b) 10 Bin-scatter



Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Management quality and firm size

- Some positive association is shown, but not easy to read
- Bin-scatter - positive overall, but most for small vs medium.
- Difference in mean management quality tends to be smaller when comparing bins of larger size

Dependence and independence

- *Dependence* of two variables - y and x means that the conditional distributions of y - conditional on x - are not the same (x is the conditioning variable).
- *Independence* of y and x means the opposite: the distribution of y on x is the same, regardless of the value of x.
- Dependence of y and x, may take many forms.

Mean dependence

- Mean-dependence: conditional expectation $E[y|x]$ varies with the value of x.
- *Mean-dependence* is the extent to which conditional expectations (means) differ.
- Two variables are positively mean-dependent if the average of one variable tends to be larger when the value of the other variable is larger, too.
- *Covariance* and *Correlation Coefficient* are measures of mean dependence.

Covariance

The formula for the covariance between two variables x and y both observed in a data table with n observations is:

$$Cov[x, y] = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n} \tag{1}$$

- for each observation $i = 1 \ldots n$
- The product within the sum in the numerator multiplies the deviation of x from its mean $(x_i - \bar{x})$ with the deviation of y from its mean $(y_i - \bar{y})$
- The entire formula is the average of these products across all observations.

The correlation coefficient

$$Corr[x, y] = \frac{Cov[x, y]}{Std[x]Std[y]} \tag{2}$$

$$-1 \leq Corr[x, y] \leq 1 \tag{3}$$

- The correlation coefficient is the standardized version of the covariance.
- The covariance may be any positive or negative number, while the correlation coefficient is bound to be between negative one and positive one.

Dependence, correlation

- Covariance or the correlation coefficient allow for all kinds of variables, including binary variables and ordered qualitative variables as well as quantitative variables.
- However, they are more appropriate measures for quantitative variables. That's because the differences $y_i - \bar{y}$ and $x_i - \bar{x}$ make more sense when y and x are quantitative variables.

Management quality and firm size

- The covariance between firm size and the management score is 177.
- The standard deviation of firm size is 977, the standard deviation of management score is 0.6.
- Positive mean-dependence: firm size tends to be higher at firms with better management.
- the correlation coefficient is 0.30 $(177/(977 * 0.6))$.
- This suggests a positive and moderately strong association.
- Management quality–firm size correlation varies considerably across industries?

Management quality and firm size

Table: Measures of management quality and their correlation with size by industry

Industry	Management–employment 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

Note: *Employee retention rates: The probability of staying with the firm, in the two experimental groups.*
Source: `working-from-home` *dataset*

Measuring a latent concept with many observed variables

- Often a concept is hard, even impossible, to measure.
- Latent variables - while we can think of them as a variable there is no single observed variable to measure them.
- Quality of management at a firm
- IQ

- The problem here is how to combine multiple observed variables

Condensing information: Using a sum

- Taking the average of all measured variables makes use of all information.
- If all measured using the *same scale* this approach, simple and a natural interpretation

- When variables measured in different scales, simple average is difficult to interpret and meaningless

Condensing information: Using a sum

- Taking the average of all measured variables makes use of all information.
- If all measured using the *same scale* this approach, simple and a natural interpretation

- When variables measured in different scales, simple average is difficult to interpret and meaningless
- Need bring it to common scale - standardization: subtracting the mean and dividing with the standard deviation
- The result is a series of variables with zero mean and standard deviation of one
- This standardized measure is called a *"z-score"* or *"score"*

Comparison and variation in x

- What is the "source of variation" in the conditioning variable
- Or put it differently, why values of the conditioning variable may differ across observations.
- Option 1: experimental data - perfect control
- Option 2: observational data - no perfect control

Comparison in Experimental data

- We have an intervention or treatment.
- Value of the conditioning variable differs across observations because the person running the experiment made them different. Hence the name: 'treatment variable'.
- There is controlled variation - a rule deciding treatment
- Experiment - comparing one or more outcome variables across the various values of a treatment variable

- Example: drug trial

Comparison with observational data

- Most data used in business, economics and policy analysis are observational.
- In observational data, no variable is fully controlled.
- Typical variables in such data are the results of the decisions
- The source of variation in these variables may have multiple sources
- People's choices, decisions, interactions, expectations, etc.
- Compare the value of the outcome variable for different values of the conditioning variable.
- Much harder interpretation

Summary

- For qualitative variables, correlation can be shown by summarizing conditional probabilities (frequencies).
- For quantitative variables, scatterplots offer a visual insight to the pattern of the relationship.
- The correlation coefficient captures a simple measure of mean dependence.
- In some cases, we measure a phenomenon with many variables. In such cases a standardized summary variable (the score) could be used to capture the essence.

Essential Reading for this week: Read Chapter 4 of Gabor Bekes Data Analysis.