Data Science for Economists

Comparison and Correlation

Tamer Çetin
UC, Berkeley
2023

Motivation

- Many questions that data analysis can answer are based on comparing values of one variable, y, against values of another variable, x, and often other variables.
- Statistical terms for uncovering information related to one variable for different values of another variable:
- Conditioning, conditional comparison, and further conditioning
- Conditional probabilities, conditional distributions, and conditional means
- Dependence, mean-dependence, and correlation.
- We will use informative visualizations of the various kinds of comparisons.

Learning Outcomes:

After working through this chapter, you should be able to:

- Define the y and the x variable(s) and identify them in data,
- Understand the concepts of conditional probability, conditional distribution, and conditional mean,
- Create informative figures to visualize conditional means (bin scatter), other conditional statistics (box plots), and joint distributions of quantitative variables (scatterplots),
- Understand the concepts of dependence, mean-dependence, correlation,
- And produce and interpret correlation coefficients.

The y and the x

- Much of data analysis is built on comparing values of a y variable by values of an x variable, or more x variables.
- Such comparison can uncover the **patterns of association** between the two variables: whether and how observations with particular values of one variable (x) tend to have particular values of the other variable (y).
- The role of y is different from the role of x.
  - we are interested in the values of y.
  - we compare observations that are different in their x values.

The y and the x

- There are two ways/goals in the analysis of conditional y on conditioning x:
- Goal 1 (prediction): predicting the value of a y variable with the help of other variables – many x variables, such as x_1, x_2, ...
- The prediction itself takes place when we know the values of those other variables but not the y variable.
- To predict y based on the other variables we need a rule that tells us what the predicted y value is as a function of the values of the other variables.
- Such a rule can be devised by analyzing data where we know the y values.
- Goal 2 (causal inference): learn about the effect of a causal variable x on an outcome variable y.
- What the value of y would be if we could change x.

Comparison and conditioning

- Data analysis can help uncover such effects by examining data with both y and x and comparing values of y between observations with different values of x
- The word statisticians use for comparison is **conditioning**
- Conditioning is a statistical term for uncovering information related to one variable for different values of another variable.
- When we compare the values of y by the values of x, we condition y on conditioning x.
- y is also called the **outcome variable**; x is also called the **conditioning variable**
- Compare prices of hotels (y) with different cities (x) –>
  - price of hotel is the outcome
  - type of city is the conditioning variable

Conditional statistic

- The most important conditional statistic is the **conditional mean**, also known as the **conditional expectation**, which shows the mean (average, expected value) of y for each value of x.
- Conditional mean = mean of variable y for each value of the conditioning variable.
- The conditional expectation of variable y for different values of variable x is

$$E[y|x]$$

- This is a function: for a value of x, the conditional expectation gives number that is the expected value (mean, average) of variable y for observations that have that x value
- It gives different values based on the conditioning variable x because it is different from the overall mean of y, $E[y]$.

Conditional and joint distributions of two quantitative variables

- The probability of a value of a variable in a dataset is its relative frequency (percentage).
- The probability of an event is the likelihood that it occurs.
- The joint distribution of two variables shows the probabilities (frequencies) of each value combination of the two variables.
- **Conditional probability** is the probability of one event if another event happens.
- The event the conditional probability is about is called the **conditional event**; the other event is called the **conditioning event**.
- Conditional probabilities are denoted as $P(event_1|event_2)$: the probability of $event_1$ conditional on $event_2$.

Conditional and joint distributions of two quantitative variables

- Note that the conditional probability is not symmetrical: $P(event_1|event_2) \neq P(event_2|event_1)$ in general.
- Joint probabilities are related to conditional probabilities.
- The **joint probability** of two events is the probability that both occur: $P(event_1 \& event_2)$.
- When two events are mutually exclusive, their joint probability is zero (the two never happen together).

Dependence and independence

- Dependence of two variables are not the same in many cases:

Different Dependence Measures: For example, correlation measures linear dependence, but if the relationship is nonlinear, the correlation may be low even though the variables are dependent.

Conditional Dependence: Dependence between two variables X and Y can change when conditioning on a third variable Z.

Non-Stationarity: In time series data, the dependence between two variables may change over time.

Causality: Even if two variables are dependent, the nature of their dependence can be very different depending on causal structure underlying their relationship. For example, if X causes Y, intervening on X will change the distribution of Y, but not vice versa.

Interaction Effects: Time dependent X covariates

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.
- We use covariance and correlation to measure the changes/variation in $E[y|x]$ by the value of x.
- Correlation coefficient is the standardized version of covariance.

Covariance and correlation

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

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

- 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.

Covariance and correlation

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

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

- The correlation coefficient is the standardized version of covariance.
- Covariance measures how much two random variables vary together.
- It indicates the extent to which two variables increase or decrease in parallel.
- It does not include information about the individual variances of the variables and does not have an upper or lower bound.
- Correlation measures the strength and direction of the linear relationship between two variables.
- It provides a measure that is easy to understand and interpret.

Measuring a latent concept with many observed variables

- Latent variable: 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
- Proxy Variable: Used to represent latent variables
- The problem here is how to combine multiple observed variables representing one latent or proxy variable
- Data analysts use one of three main approaches:
  Use one of the observed variables
  Take the average (or sum) of the observed variables
  Use principal component analysis (PCA) to combine the 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

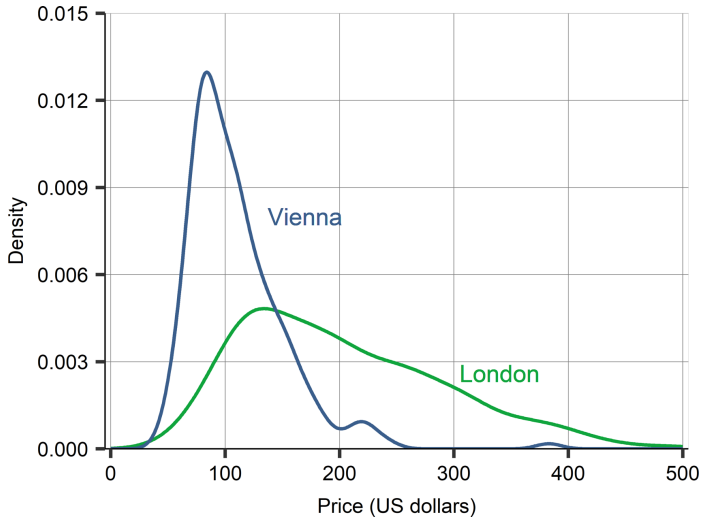
- Most data used in business, economics and policy analysis are observational.
- In observational data, no variable is fully controlled.
- Typical variables in such data are the results of the decisions
- The source of variation in these variables may have multiple sources
- People's choices, decisions,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

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 $y_i - \overline{y}$ and $x_i - \overline{x}$ make more sense when y and x are quantitative variables.

Comparisons and conditional distributions

- The conditional distribution of a variable is the distribution of the outcome variable given the conditioning variable.
- Straightforward concept if the conditioning variable is qualitative (simple if binary)
- Comparing histograms



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

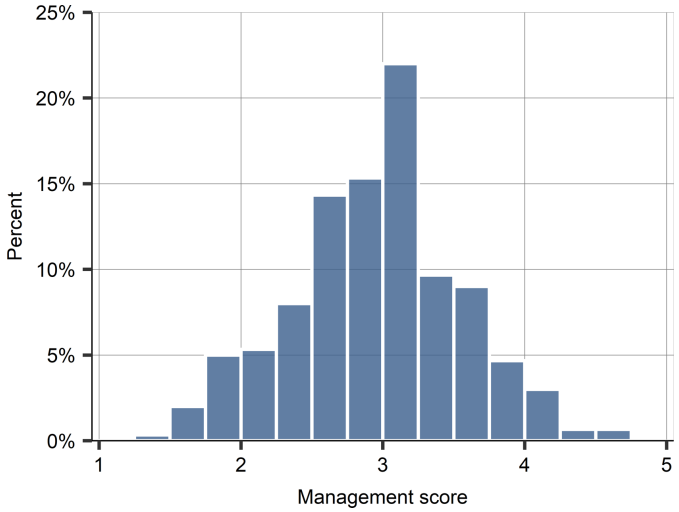
- Take 18 individual measures and average
- Measure of the quality of management is the simple average of these 18 scores = "the" management score.
- By construction, the range of the management score is between 1 and 5.

Management quality and firm size

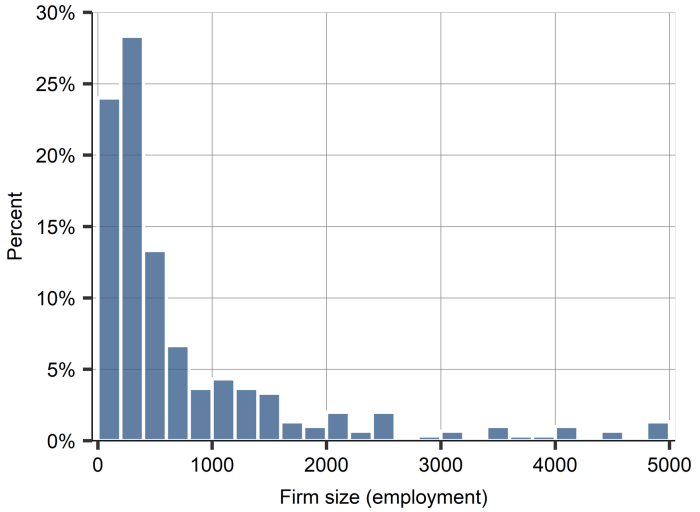
- Data from the World Management Survey to investigate our question.
- In this case study we analyze a cross-section of **Mexican** firms from the 2013 wave of the survey.
- Only firms with $100 - 5000$ employees, N=300
- The $y =$ measure of the quality of management. The $x =$ measure of firm size.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

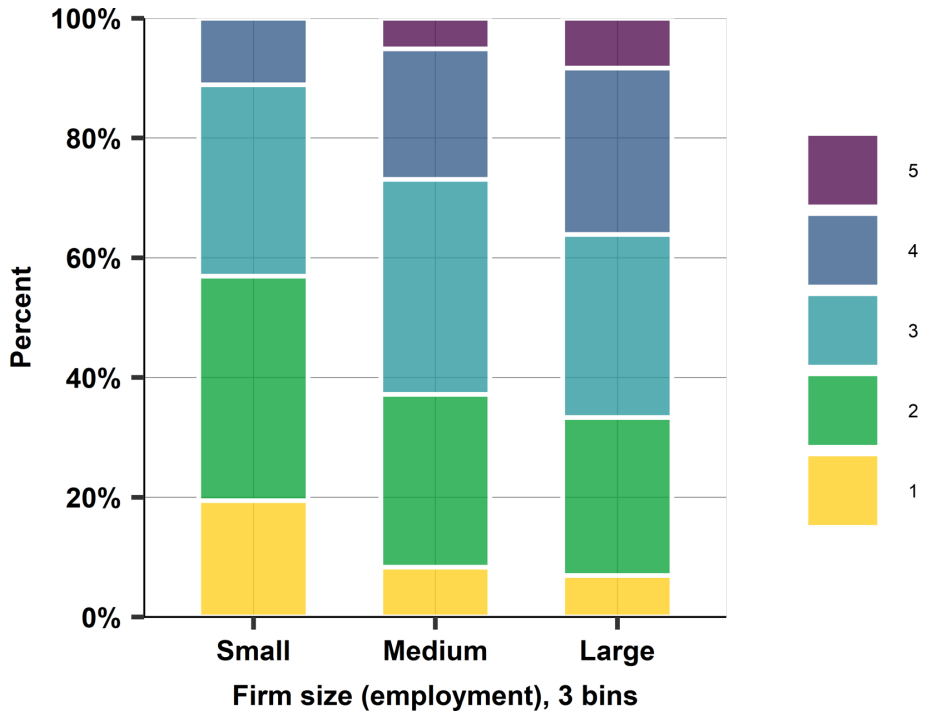
- Three bins of firm size. By number of employees: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- Take a single measure: Lean management score, with values 1,2,3,4,5.
- Thus, for each score variable we have 15 conditional probabilities: the probability of each of the 5 values of 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.

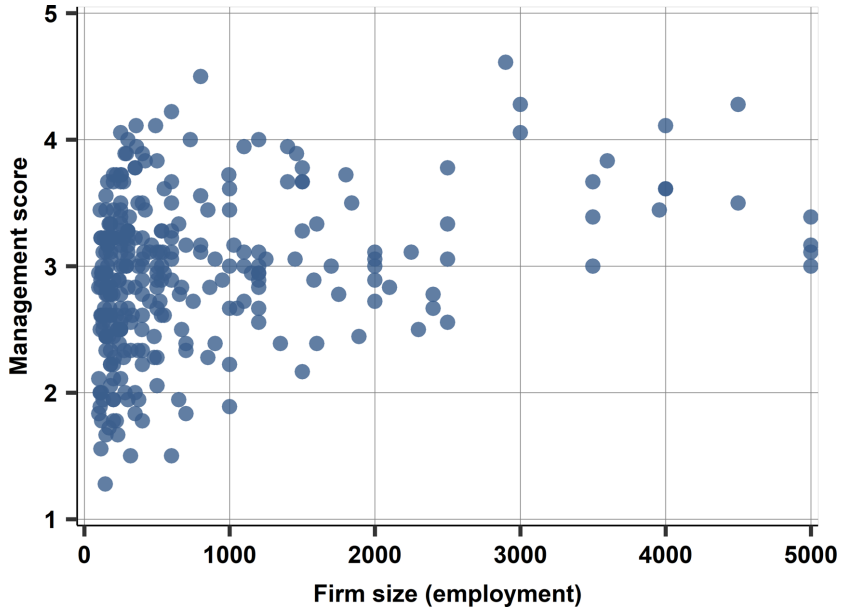
- Can calculate the mean given firm size.
- Three bins of employment: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- Mean management score is 2.68 for small firms, 2.94 for medium sized ones, and it is 3.18 for large.
- First simple evidence: larger firms have better management.

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
- What is the relationship?
- Is it clear?

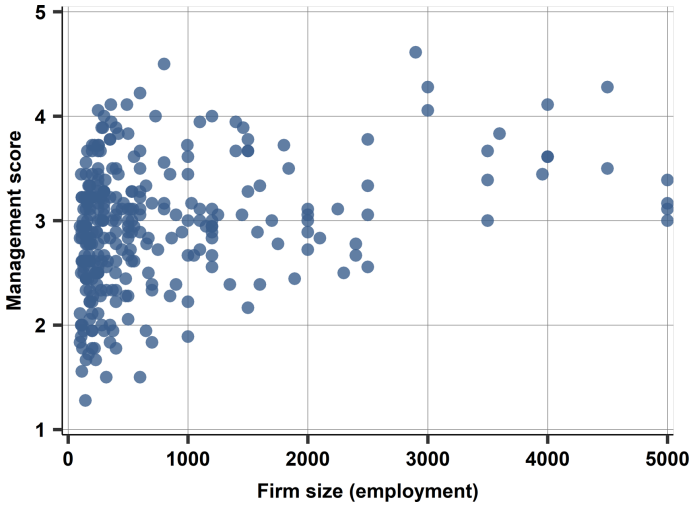
Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees. wms-management-survey data. Mexican sample, n=300.*

Management quality and firm size

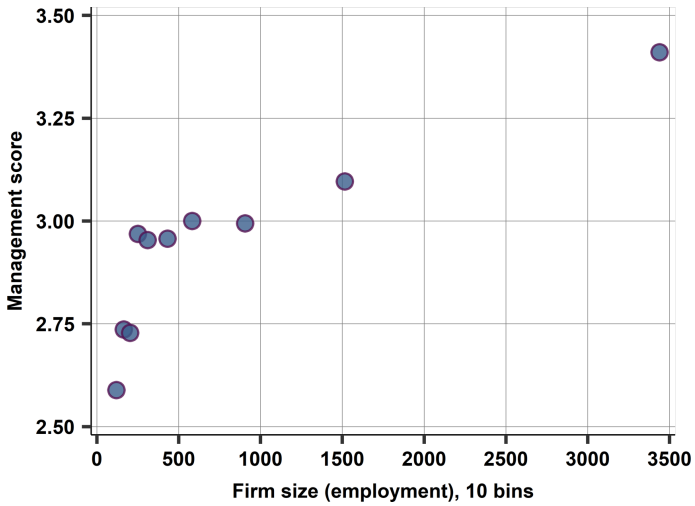
- Bin-scatter: calculate the mean of y conditional on ten bins of x.
- Bin-scatter: cut x's distribution into 10 parts, with equal number of firms. (remember - percentiles)
- Show average management score as a point corresponding to the midpoint in the employment bin (e.g., 110 for the 100–120 bin).
- 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-surveydata. 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

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-homedataset*

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.