# ECON 57 - Lecture 4

## Multivariate Statistics

Augusto Gonzalez-Bonorino

Pomona College

Fall 2023

# Table of Contents

# Bidimensional Distributions

## Introduction

- A bidimensional distribution describes the joint behavior of two variables. It says nothing about causality!
- It helps us understand how the variables relate to each other. Say we are studying two variables Inflation (X) and Fed Funds Rate (Y)
  - What is the relationship, if any, between X and Y?
  - What is the economic interpretation of this relationship? Does it make sense?
- Denoted by $f_{X,Y}(x, y)$, where $X$ and $Y$ are the statistical variables.
- The goal is to study if two variables are related, not to deduce if such relationship goes in one direction or the other.

# Bidimensional Variables

## Definitions

Bidimensional variables are represented as $[x_i, y_j, n(x_i, y_j), p(x_i, y_j)]$, where

- Absolute Joint Frequency ($n(x_i, y_j)$): The count of occurrences of outcome pair $(X_i, Y_j)$. The actual counts in each cell of the joint frequency table.
- Relative Joint Frequency ($p(x_i, y_j)$): Proportion of total outcomes that are $(X_i, Y_j)$. That is, the proportions of the total observations in each cell.
- Conditional Frequency ($p_{X|Y}(x_i|y_j)$ or $p_{Y|X}(y_j|x_i)$): The proportions of observations within a specific category of one variable given a fixed value of the other variable.

$$p(x_i, y_j) = \frac{n(x_i, y_j)}{n}, \quad p_{Y_j|X_i} = \frac{n(x_i, y_j)}{n_i}$$

# Introduction

When working with data, a two-entry table of joint frequencies provides a way to organize and visualize the relationships between two variables (i.e., one bidimensional variable). Each cell in the table represents the frequency of occurrences for a specific combination of the variables.

| | Category A | Category B |
|---|---|---|
| Category X | $n_{11}$ | $n_{12}$ |
| Category Y | $n_{21}$ | $n_{22}$ |

Table: Joint Frequency Table

# Two-entry Table of Joint Frequencies

## Absolute Frequency Table

| **Industry (X) \ Region (Y)** | **North** | **South** | **East** | **West** | $n_X(x_i)$ |
|---|---|---|---|---|---|
| Agriculture | 120 | 80 | 60 | 90 | 350 |
| Manufacturing | 180 | 150 | 200 | 130 | 660 |
| Services | 250 | 220 | 180 | 270 | 920 |
| Finance | 80 | 60 | 110 | 70 | 320 |
| Technology | 130 | 110 | 90 | 120 | 450 |
| $n_Y(y_j)$ | 760 | 620 | 640 | 680 | **2700** |

**Total Region** $= n_Y(y_j) = \sum_j^h n(x_i, y_j) = n_Y(y_j)$

**Total Industry** $= n_X(x_i) = \sum_i^k n(x_i, y_j) = n_X(x_i)$.

**Sample Size** $= \sum_i^k \sum_j^h n(x_i, y_j) = 2700$

The collection of absolute frequencies is referred to as **marginal distribution of Y (or X)**.

# Two-entry Table of Joint Frequencies

## Relative Frequency Table

| **Industry \ Region** | **North** | **South** | **East** | **West** | $p_X(x_i)$ |
|---|---|---|---|---|---|
| Agriculture | 0.0444 | 0.0296 | 0.0222 | 0.0333 | 0.1296 |
| Manufacturing | 0.0667 | 0.0556 | 0.0741 | 0.0481 | 0.2444 |
| Services | 0.0926 | 0.0815 | 0.0667 | 0.1000 | 0.3407 |
| Finance | 0.0296 | 0.0222 | 0.0407 | 0.0259 | 0.1185 |
| Technology | 0.0481 | 0.0407 | 0.0333 | 0.0444 | 0.1667 |
| $p_Y(y_j)$ | 0.2815 | 0.2296 | 0.2363 | 0.2519 | **1.0000** |

**Marginal relative frequency of Industry** $= \sum_i^k p(x_i, y_j) = p_X(x_i)$

**Marginal relative frequency of Region** $= \sum_j^h p(x_i, y_j) = p_Y(y_j)$

**Total probability** $= \sum_i^k \sum_j^h p(x_i, y_j) = 1$. Recall that $p(x_i, y_j) = \frac{n(x_i, y_j)}{n}$

# Two-entry Table of Joint Frequencies

## Conditional Frequency Table

| **X = Finance \ Y = East** | **Frequency** |
|---|---|
| Yes | 110 |
| No | 530 |
| **Total** | **640** |

Probability of Finance Activity given East Region:

$$P(\text{Finance}|\text{East}) = \frac{\text{Freq of Finance in East}}{\text{Total Freq in East}} = \frac{110}{640} = \frac{0.0407}{0.2363} \approx 0.1719$$

**Exercise:** Write the conditional frequency table for $X = \textit{Technology} \mid Y = \textit{South}$ (i.e., the probability of Technology activity given South region) and $Y = \textit{North} \mid X = \textit{Agriculture}$ (i.e., the probability of North region given Agriculture activity).

# Statistical Independence

- Two random variables $X$ and $Y$ are statistically independent if knowing the outcome of one provides no information about the outcome of the other.
- In other words, X is statistically independent of Y if the conditional distributions of X does not change when Y changes.
- Mathematically: $p(x_i|y_J) = p_X(x_i) \cdot p_Y(y_j), \forall i, j$
- If independent, $p_{Y|X}x_i|y_j = p_Y(y)$ and $p_{X|Y}(x_i|y_j)$
- *Whiteboard example*

# Statistical Independence

Two categorical variables are considered statistically independent if the occurrence of one variable does not affect the occurrence of the other.

*Consider two dice rolls: $X$ represents the outcome of the first roll, and $Y$ represents the outcome of the second roll. Are the outcomes of these rolls independent?*

Intuitively, the outcome of the first roll does not affect the outcome of the second roll, and vice versa. Therefore, the variables $X$ and $Y$ are statistically independent.

Another way to think about it is that after conditioning one of the variables you can still observe any of the elements of the other variable with positive frequency. In contrast, two variables X and Y are said to be statistically dependent if conditioning Y limits the values of X (i.e., the frequency of some value of X is 0).

# Mathematical Explanation

Mathematically, for independent variables, the joint probability is the product of the marginal probabilities:

$$P(X = x_i, Y = y_j) = P(X = x_i) \times P(Y = y_j)$$

If this equation holds for all possible values of $x_i$ and $y_j$, then $X$ and $Y$ are independent.

# Covariance & Correlation

## Conceptual Explanation

The covariance between two variables X and Y is an absolute index of association between X and Y. It measures the directional relationship between two variables (i.e., how the means values of two variables move together). For instance, returns of AAPL and MSFT, Fed Funds Rate and Inflation, or Investment and GDP.

Correlation normalizes covariance, providing a more interpretable measure. Correlation helps us study and measure the direction and intensity of relationship among variables. It measures co-variation not causation. It does not imply cause and effect relation.

Covariance measures the degree of joint variability between two random variables. For two variables $X$ and $Y$ with $n$ data points, the covariance is given by:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

where $\bar{X}$ and $\bar{Y}$ are the means of $X$ and $Y$, respectively.
The sign of the covariance indicates the direction of the relationship:

- Positive Covariance: Both variables tend to increase or decrease together.
- Negative Covariance: One variable tends to increase when the other decreases and vice versa.
- Zero Covariance: There is no linear relationship between the variables.

Correlation measures the strength and direction of the linear relationship between two variables. It allows us to make statements such as "X is strongly correlated to Y", while the covariance allows us to only make statements about the direction of the relationship.
For two variables $X$ and $Y$ with $n$ data points, the Pearson correlation coefficient is given by:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$, respectively.
The value of $r$ ranges from -1 to 1:

- $\rho = 1$: Perfect positive correlation
- $\rho = -1$: Perfect negative correlation
- $\rho = 0$: No linear correlation

**Covariance:**

- Linearity: $\text{cov}(aX, Y) = a \cdot \text{cov}(X, Y)$
- Bilinearity:
  $\text{cov}(X+W, Y+Z) = \text{cov}(X, Y)+\text{cov}(X, Z)+\text{cov}(W, Y)+\text{cov}(W, Z)$
- Symmetry: $\text{cov}(X, Y) = \text{cov}(Y, X)$
- If $Cov(X, Y) = 0$, then X and Y are statistically independent
- $Cov(X, a) = 0, \forall a \in \mathbb{R}$
- $Var(X \mp Y) = Var(X) + Var(Y) \mp 2 \cdot Cov(X, Y)$

**Correlation:**

- Range: $-1 \leq \rho(X, Y) \leq 1$
- Normalized Covariance: $\rho(X, Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$
- Independence: $\rho(X, Y) = 0$ if and only if $X$ and $Y$ are independent.

# Covariance & Correlation

## Example

Consider the following two-way table representing the data points of two variables $X$ and $Y$:

| $X$ | $Y$ |
|---|---|
| 2 | 5 |
| 4 | 7 |
| 6 | 8 |
| 8 | 12 |
| 10 | 10 |

**Step 1: Calculate the means**

$$\bar{X} = 6, \quad \bar{Y} = 8.4$$

**Step 2: Calculate the covariance**

$$\text{Cov}(X, Y) = \frac{1}{5} \sum_{i=1}^{5} (X_i - 6)(Y_i - 8.4)$$

**Step 3: Calculate the correlation**

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$.

# Conceptual Exercises

1. Explain the difference between covariance and correlation. When is one preferred over the other in an economic analysis?
2. Consider two variables: hours of study ($X$) and exam score ($Y$). How might their relationship influence their covariance and correlation?
3. Can covariance be negative if the correlation is positive? Why or why not?
4. Explain the concept of statistical independence using a real-world example.

# Mathematical Exercises

**Exercise 1:** Given the following data points for variables $X$ and $Y$:

$$X = [4, 7, 2, 5, 8]$$

$$Y = [10, 12, 6, 9, 15]$$

Calculate the covariance between $X$ and $Y$.

**Exercise 2:** For the same data points, compute the Pearson correlation coefficient between $X$ and $Y$.

**Solution 1:** First, calculate the means of $X$ and $Y$:

$$\bar{X} = \frac{4+7+2+5+8}{5} = 5.2$$

$$\bar{Y} = \frac{10+12+6+9+15}{5} = 10.4$$

Then compute the covariance using the formula:

$$\text{Cov}(X,Y) = \frac{1}{5}\sum_{i=1}^{5}(X_i - 5.2)(Y_i - 10.4)$$

This will give you the covariance between $X$ and $Y$.

**Solution 2:** Compute the standard deviations of $X$ and $Y$:

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{5}(X_i - \bar{X})^2}{4}}$$

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{5}(Y_i - \bar{Y})^2}{4}}$$

Then use the formula for the Pearson correlation coefficient:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

This will give you the correlation coefficient between $X$ and $Y$.

# Economic Applications

Bivariate descriptive statistics and frequency tables are widely used in economics to explore relationships between variables and analyze data. Here are some economic applications:

- **Covariance and Investment Diversification:** Covariance helps investors assess the relationship between the returns of different assets. Diversifying the portfolio with negatively correlated assets can reduce overall risk (because risk is measured as the variability of the assets, high risk = high variance).
- **Correlation and Economic Indicators:** Correlation analysis can reveal the relationships between economic indicators, such as GDP growth and unemployment rate or Price and Demand for a commodity, helping policymakers make informed decisions.

# Economic Applications

## Example

Consider a study on the relationship between advertising spending and sales for a company. You collect data over several months and find a positive correlation between the two variables. How can this information be useful for the company's marketing strategy?

- Correlation suggests a potential linear relationship.
- Company can adjust advertising spending based on desired sales targets.
- However, correlation does not imply causation. Other factors may influence sales.