# Statistics

## |Scatter Plots, Correlation (الارتباط)

Prof. Dr. Mostafa Elhosseini

https://youtube.com/drmelhosseini

# Agenda

- Understanding Quantitative Associations
- Response and Explanatory Variables
- Association (الارتباط )
- Looking for a Trend: The Scatterplot
- The Butterfly Ballot
- The Correlation ( معامل الارتباط )
- Example

# Introduction

- To examine the relationship between two **quantitative variables**, statisticians use various methods to assess how one variable may influence or relate to another. These methods help identify the form, direction, and strength of the association.

- **Scatter plots** for visual exploration,
- **Correlation** to measure strength, and
- **Regression analysis** to model and predict relationships.

# Response and Explanatory Variables

- Dependent variable (also called the response or outcome)
- Independent variables (also called Explanatory - predictors or features).

- Carbon dioxide (CO2) Level/Amount of gasoline use for cars
- College GPA/Number of hours a week spent studying

# Response and Explanatory Variables

- Explanatory variable → Response variable
  - Smoking status → Survival
  - Gasoline use → Carbon dioxide
  - Number of hours → GPA

| Explanatory Variable | Response Variable |
|---|---|
| Number of study hours | Exam score |
| Temperature (°C) | Ice cream sales |
| Advertising budget (USD) | Product sales |
| Age of a car (years) | Car resale value |
| Daily exercise duration (min) | Weight loss (kg) |
| Years of experience | Salary |
| Ambient air pollution level | Respiratory health incidents |
| Rainfall (mm) | Crop yield |
| Screen time (hours/day) | Sleep duration (hours/night) |
| Education level | Annual income |

# Response and Explanatory Variables

- Some studies regard either or **both variables as response** variables
  - Relationship between height and weight
- These are both quantitative biological measurements, and there is typically no clear causal direction — that is, we do not necessarily say that height causes weight or vice versa

- Blood Pressure and Age
- As people get older, their systolic blood pressure tends to increase. However, this is not a strict cause-effect relationship. Instead, it's a biological association observed across populations. Either variable can be modeled as a function of the other, depending on the study's goal.

# Association Between Two Variables

- **Association** refers to a relationship or link between two or more variables, where changes in one variable tend to be related to changes in another.
  - It is more likely that students with a high school GPA near 4 will have a college GPA above 3.5
  - So high school GPA and college GPA have an association
  - You are very happy, pretty happy, or not too happy — person's marital status (married, unmarried)

# Association Between Two Variables

- **Association** refers to a relationship or link between two or more variables, where changes in one variable tend to be related to changes in another.

| Type of Association | Description | Example | Graphical Pattern |
|---|---|---|---|
| **Positive Association** | Both variables increase or decrease together. | Study time ↑ → Exam score ↑ | ↗ Upward trend |
| **Negative Association** | One variable increases while the other decreases. | Exercise ↑ → Body fat ↓ | ↘ Downward trend |
| **No Association** | No consistent pattern or relationship between the variables. | Shoe size ↔ Intelligence | • • • Scattered points |

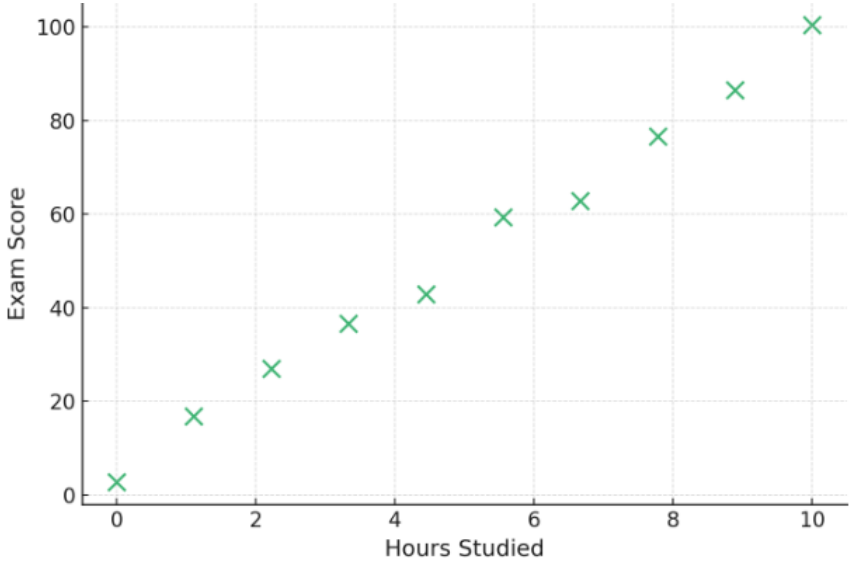
- Association does NOT imply causation.

# Looking for a Trend: The Scatterplot

- A scatter plot is a graphical representation used in statistics to display the relationship between two quantitative variables.
  - Each point on the graph represents a pair of values (x, y) — one for each variable.
  - The horizontal axis (X-axis) shows values of the explanatory (independent) variable.
  - The vertical axis (Y-axis) shows values of the response (dependent) variable.

# Looking for a Trend: The Scatterplot

- A scatter plot




| Pattern in Scatter Plot | Type of Association | Interpretation |
|---|---|---|
| Points trend **upward** | **Positive Association** | As X increases, Y tends to increase |
| Points trend **downward** | **Negative Association** | As X increases, Y tends to decrease |
| Points are **scattered randomly** | **No Association** | No consistent relationship between X and Y |
| Points form a **curved shape** | **Nonlinear Association** | Relationship exists but is not linear (e.g., quadratic) |
| Points form a **straight line** | **Strong Linear Association** | Perfect or near-perfect linear relationship |

# Looking for a Trend: The Scatterplot

- Example: Worldwide Internet and Facebook Use
  - # Worldwide Internet users and # users of social networking sites such as Facebook
  - **Investigate the relationship** between Internet penetration (use) and Facebook penetration
  - Source: Data from the World Bank (data.worldbank.org ) and www.internetworldstats.com for the year 2012

| Country | Internet Penetration | Facebook Penetration |
|---|---|---|
| Argentina | 55.8% | 48.8% |
| Australia | 82.4% | 51.5% |
| Belgium | 82.0% | 44.2% |

# Looking for a Trend: The Scatterplot

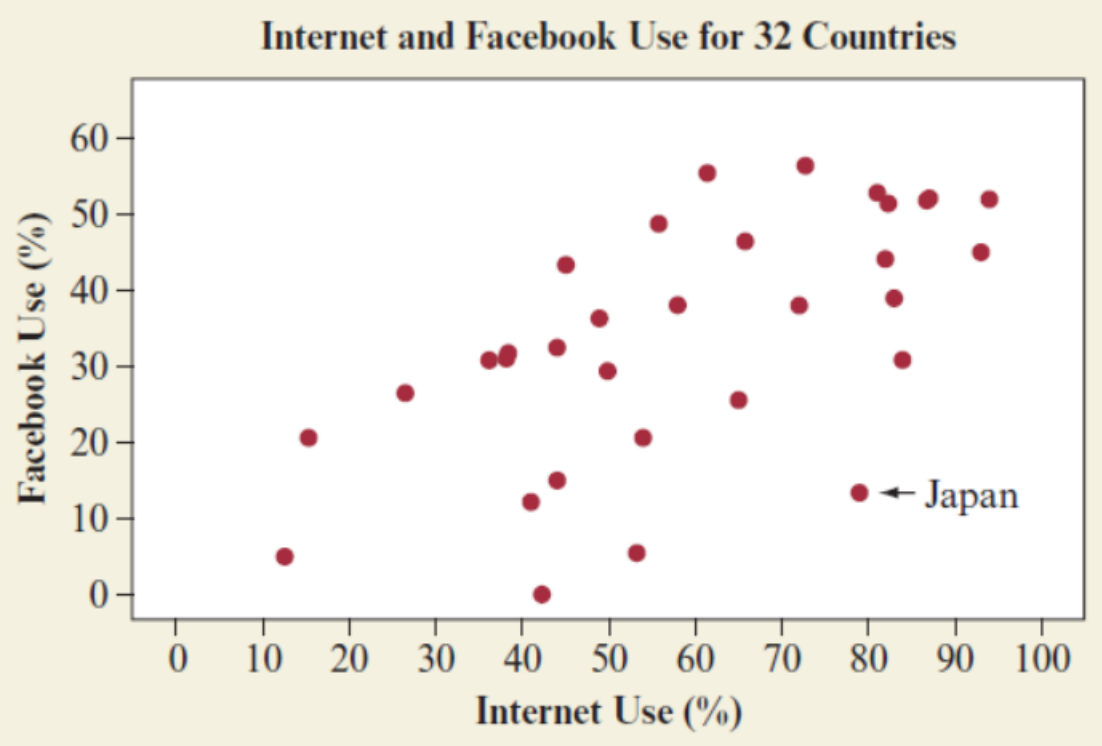
- Example: Worldwide Internet and Facebook Use

| Country | Internet Penetration | Facebook Penetration |
|---|---|---|
| Brazil | 49.9% | 29.5% |
| Canada | 86.8% | 51.9% |
| Chile | 61.4% | 55.5% |
| China | 42.3% | 0.1% |
| Colombia | 49.0% | 36.3% |
| Egypt | 44.1% | 15.1% |
| France | 83.0% | 39.0% |
| Germany | 84.0% | 30.9% |
| Hong Kong | 72.8% | 56.4% |
| India | 12.6% | 5.1% |
| Indonesia | 15.4% | 20.7% |
| Italy | 58.0% | 38.1% |
| Japan | 79.1% | 13.5% |
| Malaysia | 65.8% | 46.5% |
| Mexico | 38.4% | 31.8% |
| Netherlands | 93.0% | 45.1% |
| Peru | 38.2% | 31.2% |
| Philippines | 36.2% | 30.9% |
| Poland | 65.0% | 25.6% |
| Russia | 53.3% | 5.6% |
| Saudi Arabia | 54.0% | 20.7% |
| South Africa | 41.0% | 12.3% |
| Spain | 72.0% | 38.1% |
| Sweden | 94.0% | 52.0% |
| Thailand | 26.5% | 26.5% |
| Turkey | 45.1% | 43.4% |
| United Kingdom | 87.0% | 52.1% |
| United States | 81.0% | 52.9% |
| Venezuela | 44.1% | 32.6% |

# Looking for a Trend: The Scatterplot

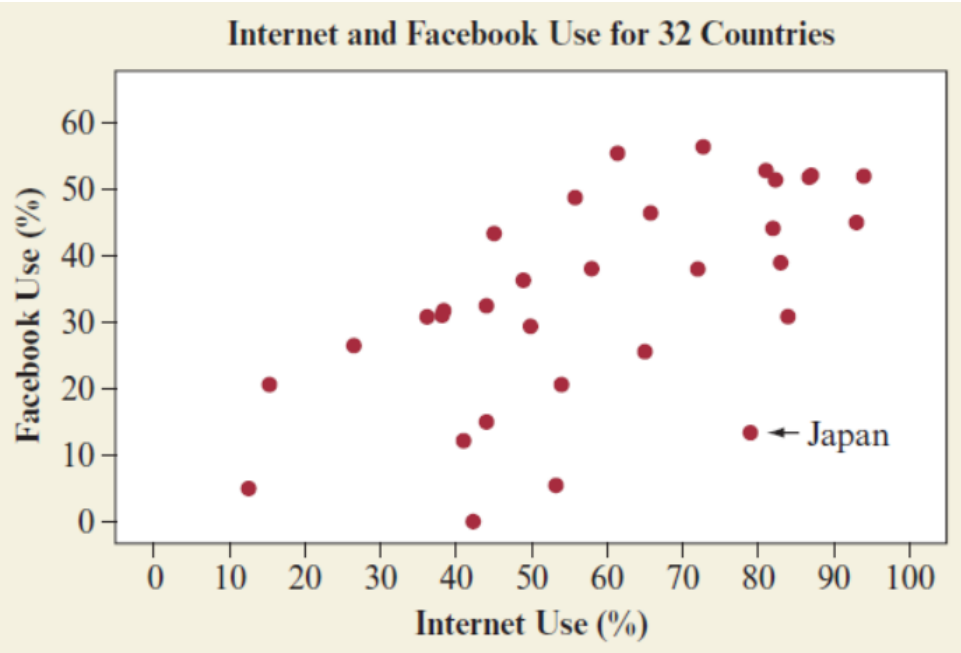
- Is it true that countries with higher Internet use tend to have higher Facebook use?

- How can we picture the association between the two variables on a single display



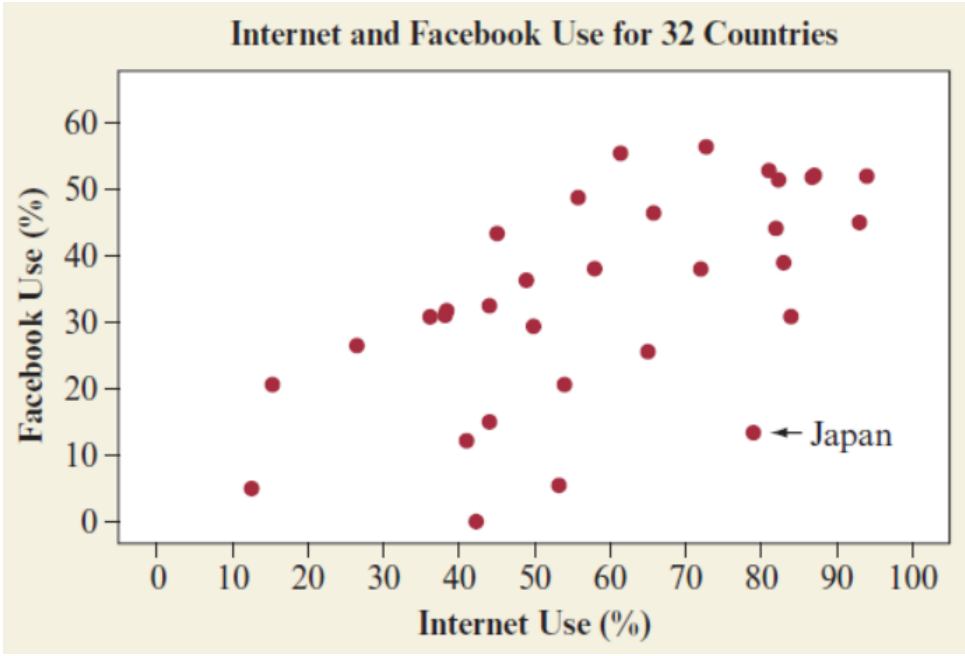
# Looking for a Trend: The Scatterplot

- Is it true that countries with higher Internet use tend to have higher Facebook use?



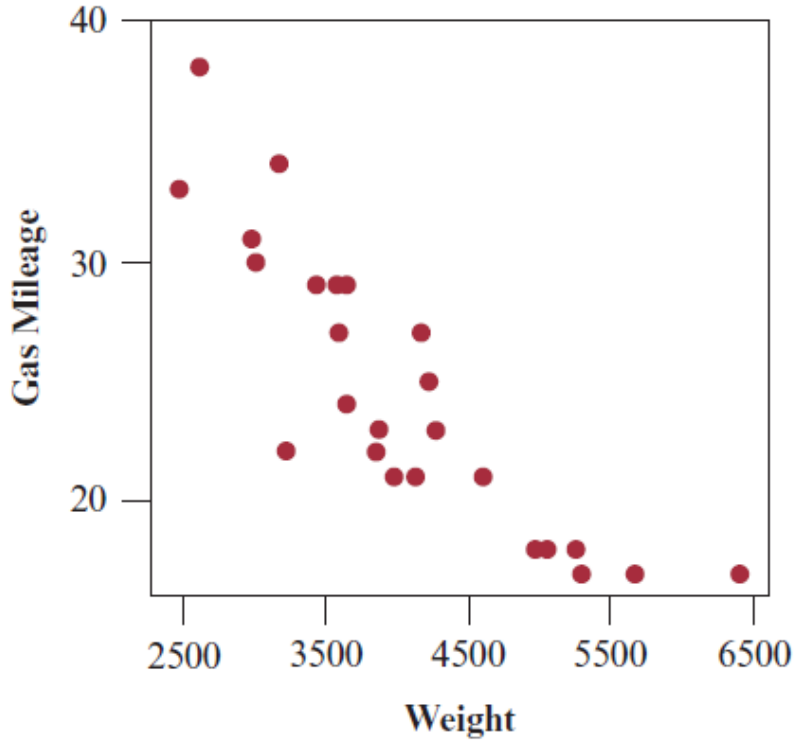
# Looking for a Trend: The Scatterplot

- You can describe the overall pattern of a scatterplot by the **trend**, **direction**, and **strength**.
  - **Trend**: linear, curved, clusters, no pattern
  - **Direction**: positive, negative, no direction
  - **Strength**: how closely the points fit the trend
  - Also look for **outliers** from the overall trend.



# Looking for a Trend: The Scatterplot

- Association between x = weight of car and y = gas mileage (in miles per gallon),
- Negative association: Heavier cars would tend to get poorer gas mileage, i.e., drive fewer miles per gallon of gas.
  - Negative association,
  - Trend: a straight line?



# The Butterfly Ballot (بطاقة الفراشة)

- **Al Gore** and **George W. Bush** were the Democratic and Republican candidates in the 2000 U.S. presidential election. In **Palm Beach** County, **Florida**, initial election returns reported 3407 votes for the Reform party candidate, **Pat Buchanan** (بات بيوكانن) . Political analysts thought this total seemed surprisingly large.
- They felt that **most of these votes** may have actually been intended for **Gore** (whose name was next to Buchanan's on the ballot) but were wrongly cast for Buchanan because of the **design of the butterfly ballot** used in that county, which some voters found confusing

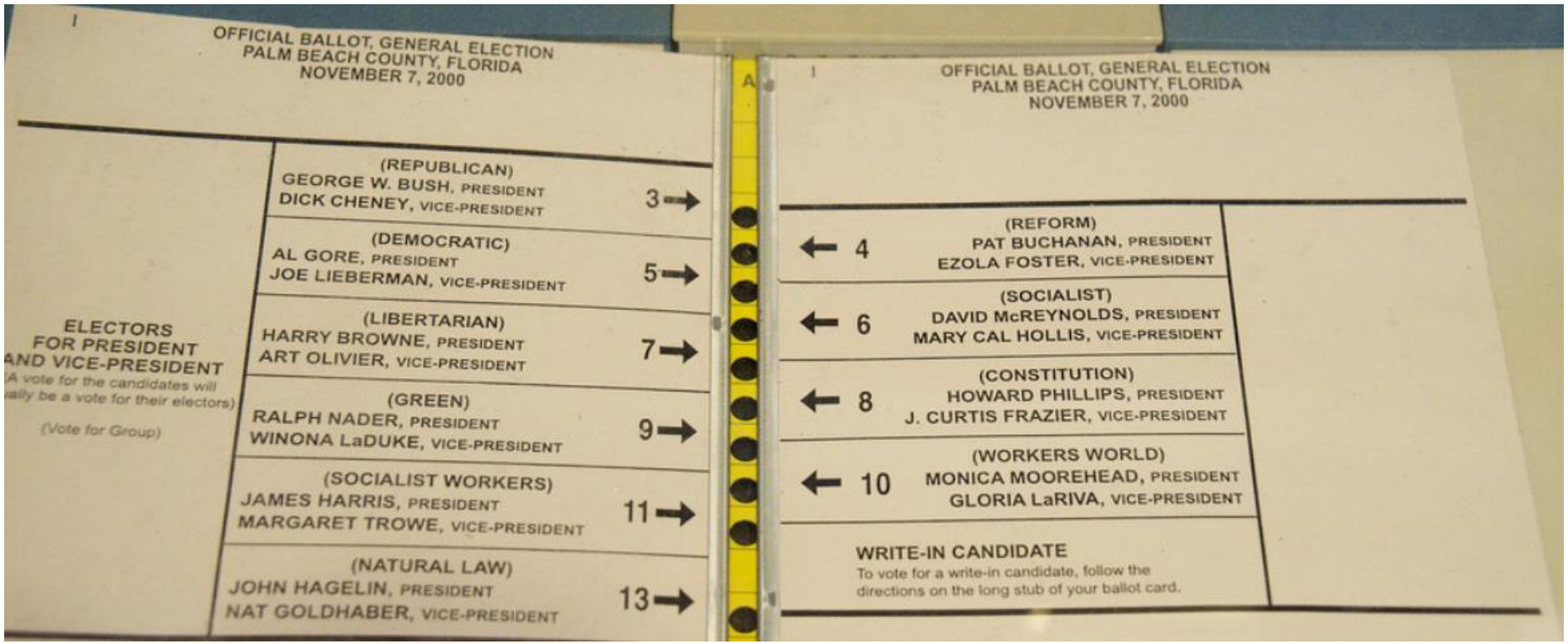
# The Butterfly Ballot (بطاقة الفراشة)

- The Butterfly Ballot refers to a controversial and confusing ballot design used during the 2000 U.S. Presidential Election, particularly in Palm Beach County, Florida, which is widely believed to have influenced the election’s outcome.

**📦ما هي بطاقة الفراشة؟**

بطاقة الفراشة كانت تصميماً غريباً ومعقداً للاقتراع استُخدم في مقاطعة "بالم بيتش" بولاية فلوريدا خلال الانتخابات الرئاسية الأمريكية عام 2000.كانت البطاقة مطوية من المنتصف بحيث تُدرج أسماء المرشحين على جانبي الورقة، بينما توضع ثقوب التصويت في العمود الأوسط.سُمّيت "بطاقة الفراشة" لأن تصميمها يشبه جناحي الفراشة عند الفتح.

# The Butterfly Ballot (بطاقة الفراشة)



# The Butterfly Ballot (بطاقة الفراشة)

❗ What Went Wrong?

- The confusing alignment of names and punch holes led many voters to **accidentally vote** for the wrong candidate.
- For example, some people intending **to vote for Al Gore** (Democratic candidate) mistakenly **punched the hole for Pat Buchanan** (Reform Party).
- Pat Buchanan received an unusually high number of votes in that county — far more than expected given his support nationally.

# The Butterfly Ballot (بطاقة الفراشة)

🔍 Impact on the 2000 Election:

- The 2000 election was extremely close, and the outcome hinged on Florida’s 25 electoral votes.

- فاز جورج دبليو بوش في نهاية المطاف بولاية فلوريدا – وبالتالي برئاسة الولايات المتحدة – بفارق ضئيل بلغ 537 صوتاً فقط، وذلك بعد قرار مثير للجدل من المحكمة العليا أوقف عملية إعادة فرز الأصوات.

- Analysts and journalists widely believe that the butterfly ballot cost Gore thousands of votes in Palm Beach County alone.

# The Butterfly Ballot (بطاقة الفراشة)

📚 Why It Matters:

- The butterfly ballot became a symbol of poor ballot design and its potential to undermine democracy.
- It sparked nationwide debates on voting technology, ballot design, election fairness, and electoral reform.
- It also led to the passage of the Help America Vote Act (HAVA) in 2002 to modernize voting systems.

---

- أدت هذه الحادثة إلى دعوات كبيرة لإصلاح نظم التصويت، وأسهمت في إقرار قانون "مساعدة أمريكا على التصويت" في عام 2002 لتحديث الأنظمة الانتخابية.

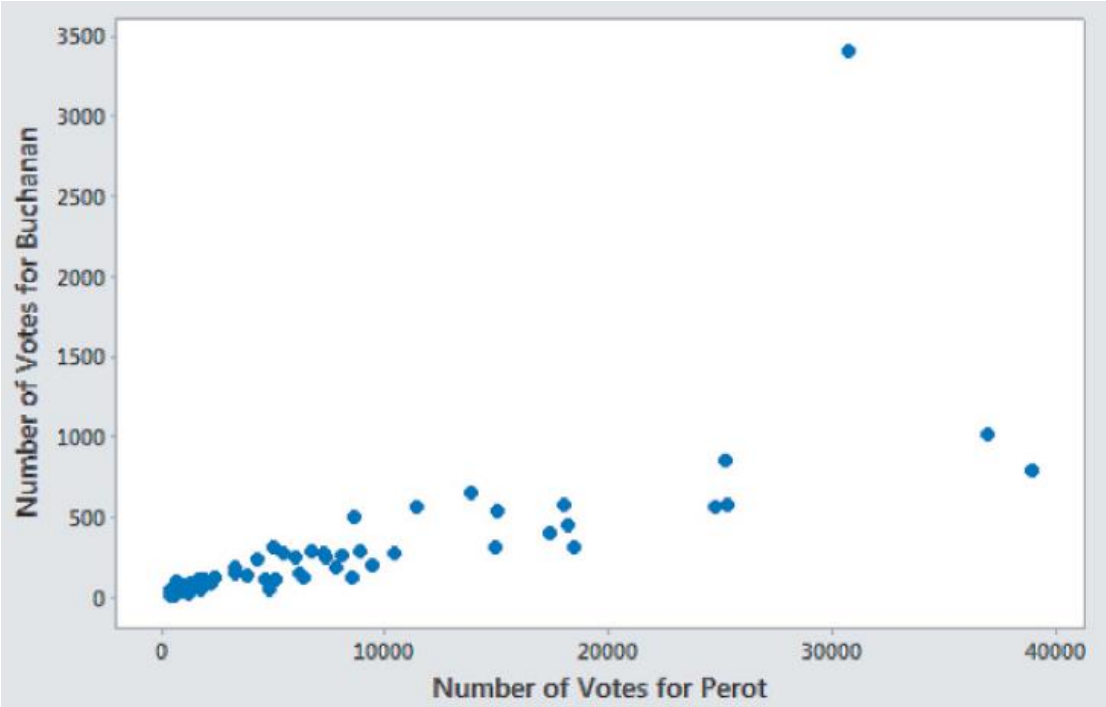
# The Butterfly Ballot (بطاقة الفراشة)

- Determine whether the Buchanan vote in Palm Beach County was an outlier compared to the typical relationship between 1996 Reform Party votes and 2000 Buchanan votes.
- هل كانت أصوات بيوكانن في مقاطعة بالم بيتش عام 2000 مرتفعة بشكل غير طبيعي مقارنة بأصوات حزب الإصلاح في عام 1996 (روس بيرو)؟
- Explore whether Pat Buchanan's vote in Palm Beach County in 2000 was unexpectedly high
  - Compare Buchanan's votes in 2000 to the Reform Party's votes in 1996 (Ross Perot) across all 67 counties in Florida

# The Butterfly Ballot (بطاقة الفراشة)

**Scatter Plot**:

- X-axis: Reform Party votes in 1996 (Perot)
- Y-axis: Buchanan votes in 2000
- Each point: Represents one of the 67 counties




- One point falls well above the others. This severe outlier is the observation for Palm Beach County

# The Butterfly Ballot (بطاقة الفراشة)

- Add a **Regression Line** (Line of Best Fit)
  - Fit a linear regression line to the scatter plot.
  - Check how far Palm Beach County is from this line (residual).

✅ A large positive residual for Palm Beach means Buchanan received more votes than expected, based on 1996 voting patterns.

# The Butterfly Ballot (بطاقة الفراشة)

**Residual Plot**:

- Plot residuals from the regression model (i.e., actual - predicted votes).
- Palm Beach County's residual will show how much it deviated from the expected trend.

# The Butterfly Ballot (بطاقة الفراشة)

**Box Plot of Buchanan**:

- Votes You can also make a box plot of Buchanan votes across all counties.
- This can highlight whether Palm Beach is an extreme outlier.

# The Correlation ( معامل الارتباط )

📌 Summarizing the Strength of Association:

- When the data points show a pattern that roughly follows a straight line, the two variables are said to have an **approximately linear relationship**.
- Sometimes, the points **lie very close** to a straight line, while other times there is more variability (**scatter**) around that trend.
- To quantify the **strength and direction** of this linear relationship, we use a summary statistic called the **correlation**.

# The Correlation ( معامل الارتباط )

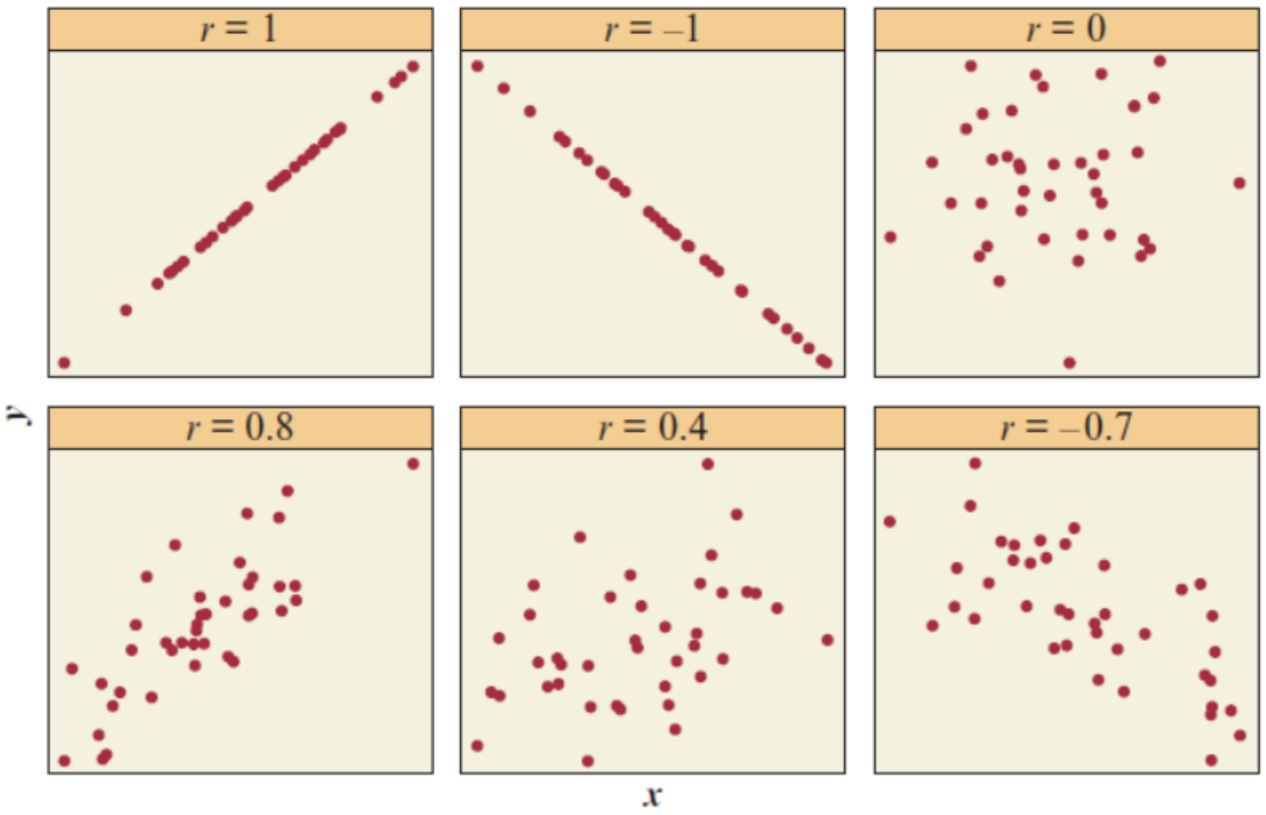
✅ What is Correlation:

- Correlation (often denoted as $r$) measures the strength and direction of a linear association between two quantitative variables.
- The value of $r$ ranges from: $-1 \leq r \leq 1$
- **Interpretation**:
  - $r = 1$: Perfect positive linear relationship.
  - $r = -1$: Perfect negative linear relationship.
  - $r = 0$: No linear relationship.

# The Correlation ( معامل الارتباط )

✅ What is Correlation:

- The value of the correlation does not depend on the variables' units
- The correlation does not depend on which variable is treated as the response and which as the explanatory variable



# Pearson Correlation Coefficient Formula

✅ Method 1: Covariance-Based Formula (Raw Score Formula):

- Given two variables $X$ and $Y$, with $n$ observations:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- $x_i, y_i$ are individual data points
- $\bar{x}, \bar{y}$ are the means of $X$ and $Y$
- $r \in [-1,1]$

# Pearson Correlation Coefficient Formula

✅ Method 2: Z-Score Based Formula:

- Let $z_x$ denote the z-score for an observation x on the explanatory variable

$$z_x = \frac{x - \bar{x}}{\sigma}$$

- To obtain $r$, you calculate the product $z_x z_y$ for each observation and then find a typical value (a type of average) of those products

$$r = \frac{1}{n-1}\sum z_x z_y$$

# Graph Data to Check


- The correlation is designed for **straight-line relationships**
- The correlation is not valid for describing association when the points cluster around a curve rather than around a straight line

- For this U-shaped relationship, **the correlation is 0** (or close to 0), even though the variables are **strongly associated**
- If we merely used software without looking at the scatterplot — mistakenly conclude that the variables have no association.
- They do have one (and a fairly strong one), but it is not a straight-line association

# Example 01

✅ Method 1: Covariance-Based Formula (Raw Score Formula):

- Given two variables $X$ and $Y$, with $n$ observations:

| $X$ | $Y$ |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 1.3 |
| 4 | 3.75 |
| 5 | 2.25 |

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

# Example 01

| $X$ | $Y$ |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 1.3 |
| 4 | 3.75 |
| 5 | 2.25 |

### ✅ Step 1: Calculate Means

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3.0, \quad \bar{y} = \frac{1+2+1.3+3.75+2.25}{5} = 2.06$$

### ✅ Step 2: Create a Table

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 1 | 1 | -2.0 | -1.06 | 2.12 | 4.00 | 1.1236 |
| 2 | 2 | -1.0 | -0.06 | 0.06 | 1.00 | 0.0036 |
| 3 | 1.3 | 0.0 | -0.76 | 0.00 | 0.00 | 0.5776 |
| 4 | 3.75 | 1.0 | 1.69 | 1.69 | 1.00 | 2.8561 |
| 5 | 2.25 | 2.0 | 0.19 | 0.38 | 4.00 | 0.0361 |
| **Σ** | | | | **4.25** | **10.00** | **4.597** |

# Example 01

| $X$ | $Y$ |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 1.3 |
| 4 | 3.75 |
| 5 | 2.25 |

✅ **Step 2: Create a Table**

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 1 | 1 | -2.0 | -1.06 | 2.12 | 4.00 | 1.1236 |
| 2 | 2 | -1.0 | -0.06 | 0.06 | 1.00 | 0.0036 |
| 3 | 1.3 | 0.0 | -0.76 | 0.00 | 0.00 | 0.5776 |
| 4 | 3.75 | 1.0 | 1.69 | 1.69 | 1.00 | 2.8561 |
| 5 | 2.25 | 2.0 | 0.19 | 0.38 | 4.00 | 0.0361 |
| **Σ** | | | | **4.25** | **10.00** | **4.597** |

✅ **Step 3: Plug into Formula**

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}} = \frac{4.25}{\sqrt{10.0 \cdot 4.597}} = \frac{4.25}{\sqrt{45.97}} \approx \frac{4.25}{6.78} \approx 0.626$$

# Example 01

🎯 Final Answer:

- $r \approx 0.63$ (Moderate positive correlation):
- This result shows that:
  - There is a positive trend between $X$ and $Y$,
  - But it's not perfectly linear — the points deviate from a straight line.

# References


- Montgomery, D.C. and Runger, G.C., 2010. Applied statistics and probability for engineers. John wiley & sons.
- Devore, J.L., 2011. Probability and Statistics for Engineering and the Sciences. Cengage learning.
- Gubner, J.A., 2006. Probability and random processes for electrical and computer engineers. Cambridge University Press.
- Agresti, A. and Franklin, C., 2007. The art and science of learning from data. Upper Saddle River, New Jersey, 88.
- Navidi, W.C., 2008. Statistics for engineers and scientists (Vol. 2, No. 4). New York, NY, USA: McGraw-Hill Higher Education.