Department of Statistics
Faculty of Science
Yarmouk University

SATS 101

Introduction to Probability and Statistics

Yarmouk University

Second Semester

2009/2010

Done by: Osama Alkhoun
Mobile: 079648461З

Chapter 3
Describing Bivariate Data

Bivariate Data

- When two variables are measured on a single experimental unit, the resulting data are called **Bivariate data.**
- You can describe each variable individually, and you can also explore the **relationship** between the two variables.
- Bivariate data can be described with
  – **Graphs**
  – **Numerical Measures**

Graphs for Qualitative Variables

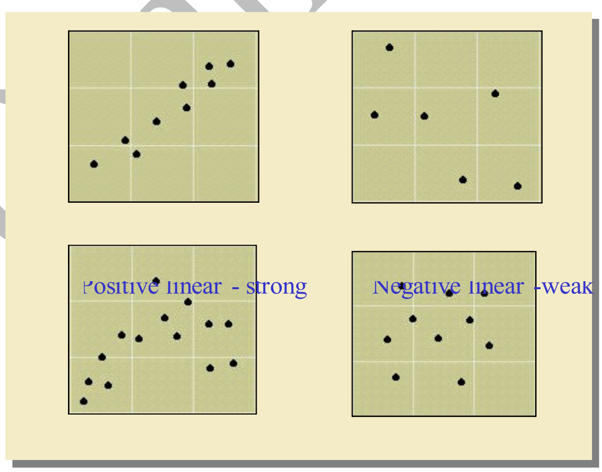
- When at least one of the variables is qualitative, you can use comparative pie charts or bar charts.

Two Quantitative Variables

When both of the variables are quantitative, call one variable *x* and the other *y*. A single measurement is a pair of numbers (*x*, *y*) that can be plotted using a two-dimensional graph called a **scatterplot.**

Describing the Scatterplot

- What **pattern** or **form** do you see?
  - Straight line upward or downward
  - Curve or no pattern at all
- How **strong** is the pattern?
  - Strong or weak
- Are there any **unusual observations**?
  - Clusters or outliers



Curvilinear No relationship

Numerical Measures for Two Quantitative Variables

- Assume that the two variables x and y exhibit a **linear pattern** or **form**.
- There are two numerical measures to describe
  - The **strength** and **direction** of the relationship between x and y.
  - The **form** of the relationship.

The Correlation Coefficient

- The strength and direction of the relationship between x and y are measured using the **correlation coefficient, *r*.**

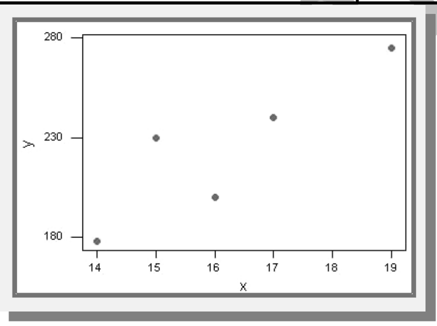
$$r = \frac{s_{xy}}{s_x s_y} \quad \text{where} \quad s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1}$$

s_x = standard deviation of the x's
s_y = standard deviation of the y's

Example

- Living area x and selling price y of 5 homes.

Residence	1	2	3	4	5
x (thousand sq ft)	14	15	17	19	16
y ($000)	178	230	240	275	200



•The scatterplot indicates a positive linear relationship.

x	y	xy
14	178	2492
15	230	3450
17	240	4080
19	275	5225
16	200	3200
81	1123	18447

Calculate

$\bar{x} = 16.2 \quad s_x = 1.924$

$\bar{y} = 224.6 \quad s_y = 37.360$

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1}$$

$$r = \frac{s_{xy}}{s_x s_y}$$

$$= \frac{18447 - \frac{(81)(1123)}{5}}{4} = 63.6$$

$$= \frac{63.6}{1.924(37.36)} = .885$$

Interpreting *r*

- $-1 \le r \le 1$ Sign of r indicates direction of the linear relationship.
- $r \approx 0$ Weak relationship; random scatter of points
- $r \approx 1$ or -1 Strong relationship; either positive or negative
- $r = 1$ or -1 All points fall exactly on a straight line.

- يمكن تصنيف نوع العلاقة (حسب أشارة "r") كالآتي:

إشارة " r "	نوع العلاقة
سالبة $(r < 0)$	عكسية
موجبة $(r > 0)$	طردية
$(r = 0)$	لا توجد علاقة

The Regression Line

- Sometimes *x* and *y* are related in a particular way—the value of *y* depends on the value of *x*.
  - *y* = dependent variable
  - *x* = independent variable
- The form of the linear relationship between *x* and *y* can be described by fitting a line as best we can through the points. This is the **regression line,**

  y = a + bx.
  - *a* = *y*-intercept of the line
  - *b* = slope of the line

- To find the slope and *y*-intercept of the best fitting line, use:

$$b = r\frac{s_y}{s_x} \quad \text{OR} \quad b = \frac{S_{xy}}{S_x^2}$$

$$a = \bar{y} - b\bar{x}$$

x	y	xy
14	178	2492
15	230	3450
17	240	4080
19	275	5225
16	200	3200
81	1123	18447

Recall

$\bar{x} = 16.2 \quad s_x = 1.9235$

$\bar{y} = 224.6 \quad s_y = 37.3604$

$r = .885$

$$b = r\frac{s_y}{s_x} = (.885)\frac{37.3604}{1.9235} = 17.189$$

$$a = \bar{y} - b\bar{x} = 224.6 - 17.189(16.2) = -53.86$$

Regression Line : $y = -53.86 + 17.189x$

معامل الارتباط الخطى البسيط " لبيرسون" Pearson

- يمكن قياس الارتباط بين متغيرين كميين (x,y) بطريقة "بيرسون" Pearson
- ولحساب معامل الارتباط في العينة ، نستخدم القانون:

$$r = \frac{S_{xy}}{S_x\, S_y} = \frac{\dfrac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)}}{\sqrt{\dfrac{\sum (x-\bar{x})^2}{(n-1)}}\sqrt{\dfrac{\sum (y-\bar{y})^2}{(n-1)}}} \qquad (١-٦)$$

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} \qquad (٢-٦)$$

حيث :

$S_{xy} = \sum (x-\bar{x})(y-\bar{y})/(n-1)$ هو التباين " Covariance" بين (x,y).

$S_x = \sqrt{\sum (x-\bar{x})^2/(n-1)}$ هو الانحراف المعياري لقيم (x)

$S_y = \sqrt{\sum (y-\bar{y})^2 / (n-1)}$ هو الانحراف المعياري لقيم (y)

مثال:

فيما يلي مساحة الأعلاف الخضراء بالألف هكتار، وإجمالي إنتاج اللحوم بالألف طن، خلال الفترة من 1995 حتى عام 2002 . والمطلوب: حساب معامل الارتباط بين المساحة والكمية، والتعليق.

[illegible]	1995	1996	1997	1998	1999	2000	2001	2002
(x) [illegible]	305	313	297	289	233	214	240	217
(y) [illegible]	592	603	662	607	635	699	719	747

- حساب الوسط الحسابي لكل من المساحة، والكمية:

$$\bar{x} = \frac{\sum x}{n} = \frac{2108}{8} = 2635 \; , \; \bar{y} = \frac{\sum y}{n} = \frac{5264}{8} = 658$$

- نحسب المجاميع كما في الجدول:

		x		y		x $\bar{x}$
305	592	41.5	1722.25	-66	4356	-2739
313	603	49.5	2450.25	-55	3025	-2722.5
297	662	33.5	1122.25	4	16	134
289	607	25.5	650.25	-51	2601	-1300.5
233	635	-30.5	930.25	-23	529	701.5
214	699	-49.5	2450.25	41	1681	-2029.5
240	719	-23.5	552.25	61	3721	-1433.5
217	747	-46.5	2162.25	89	7921	-4138.5
2108	5264	0	12040	0	23850	-13528

$$(x-\bar{x})^2$$

- نطبق المعادلة (6-2) ونحسب " r " كما يلي:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \frac{-13528}{\sqrt{12040}\;\sqrt{23850}}$$

$$= \frac{-13528}{(109.727)(154.434)} = \frac{-13528}{16945.619} = -0.798$$

Example:

- Predict the selling price for another residence with 1600 square feet of living area.

Predict: $y = -53.86 + 17.189x$

$= -53.86 + 17.189(16) = 221.16$ or \$221,160