# Elementary Statistics
**A Step by Step Approach**
**Sixth Edition**

by
Allan G. Bluman
http://www.mhhe.com/math/stat/blumanbrief

SLIDES PREPARED
BY
LLOYD R. JAISINGH
MOREHEAD STATE UNIVERSITY
MOREHEAD KY

Updated by
Dr. Saeed Alghamdi
King Abdulaziz University
www.kau.edu.sa/saalghamdy

# CHAPTER 10

## Correlation and Regression



## Objectives

10-1

- Draw a scatter plot for a set of ordered pairs.
- Compute the correlation coefficient.
- Compute the equation of the regression line.



## Notes

- ……………………………………………………...............
- ……………………………………………………...........
- ……………………………………………………...........
- ……………………………………………………...........
- ……………………………………………………...........

## Introduction

10-2

- Inferential statistics involves determining whether a relationship between two or more numerical or quantitative variables exists.
- *Correlation* is a statistical method used to determine whether a relationship between variables exists.
- *Regression* is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.



## Notes

- ……………………………………………………...............
- ……………………………………………………...........
- ……………………………………………………...........
- ……………………………………………………...........
- ……………………………………………………...........

## Introduction

10-3

### Statistical Questions

1. Are two or more variables related?
2. If so, what is the strength of the relationship?
3. What type or relationship exists?
4. What kind of predictions can be made from the relationship?



## Notes

- …………………………………………………...............
- …………………………………………………...........
- …………………………………………………...........
- …………………………………………………...........
- …………………………………………………...........

## Introduction

10-4

- A *correlation coefficient* is a measure of how variables are related.
- In a *simple relationship*, there are only two types of variables under study; an independent variable or explanatory variable or a predictor variable, and a *dependent variable* or an *outcome variable* or a *response variable*.



## Notes

- …………………………………………………...............
- …………………………………………………...........
- …………………………………………………...........
- …………………………………………………...........
- …………………………………………………...........

## Introduction

10-5

- Simple relationship can be positive or negative.
- A *positive relationship* exists when both variables increase or decrease at the same time.
- A *negative relationship* exists when one variable increases and the other variable decreases.



## Notes

- …………………………………………………...............
- …………………………………………………...........
- …………………………………………………...........
- …………………………………………………...........
- …………………………………………………...........

## Scatter Plots

10-6

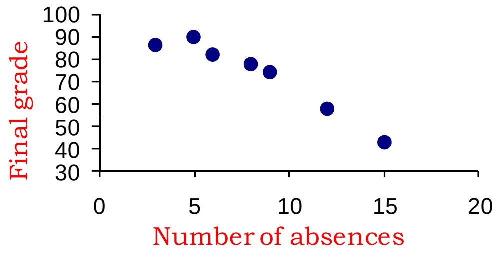
- A *scatter plot* is a graph of the ordered pairs ($x$,$y$) of numbers consisting of the independent variable, $x$, and the dependent variable, $y$.
- A *scatter plot* is a visual way to describe the nature of the relationship between the independent and dependent variables.



## Notes

## Scatter Plot Example

10-7




*See examples 10-1, 10-2 and 10-3

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

## Notes

## Correlation Coefficient

10-8

- The *correlation coefficient* computed from the sample data measures the strength and direction of a linear relationship between two variables.
- The symbol for the sample correlation coefficient is $r$.
- The symbol for the population correlation coefficient is $\rho$ (rho).

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

## Notes

## Correlation Coefficient

10-9

- The range of the correlation coefficient is from −1 to +1.
- If there is a *strong positive linear relationship* between the variables, the value of $r$ will be close to +1.
- If there is a *strong negative linear relationship* between the variables, the value of $r$ will be close to −1.

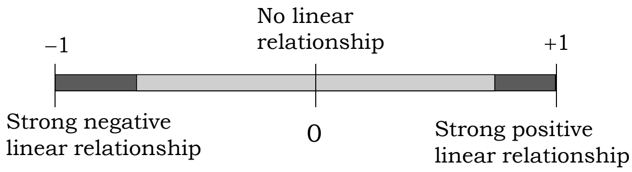


## Notes

## Correlation Coefficient

10-10

- When there is no linear relationship between the variables or only a weak relationship, the value of $r$ will be close to 0.




* See Figure 10-6 on page 534



## Notes

## Correlation Coefficient

10-11

| Correlation Coefficient Value | Meaning |
|---|---|
| +1 | Complete Positive Linear Relationship |
| 0.90 — 0.99 | Very Strong Positive Linear Relationship |
| 0.70 — 0.89 | Strong Positive Linear Relationship |
| 0.50 — 0.69 | Moderate Positive Linear Relationship |
| 0.30 — 0.49 | Weak Positive Linear Relationship |
| 0.01 — 0.29 | Very Weak Positive Linear Relationship |
| 0 | No Linear Relationship |
| -0.01 — -0.29 | Very Weak Negative Linear Relationship |
| -0.30 — -0.49 | Weak Negative Linear Relationship |
| -0.50 — -0.69 | Moderate Negative Linear Relationship |
| -0.70 — -0.89 | Strong Negative Linear Relationship |
| -0.90 — -0.99 | Very Strong Negative Linear Relationship |
| -1 | Complete Negative Linear Relationship |



## Notes

10-12

## Correlation Coefficient

Formula for the ***Pearson product moment correlation coefficient*** (*r*)

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}}$$

- where $n$ is the number of data pairs.

* See examples 10-4 and 10-5



## Notes

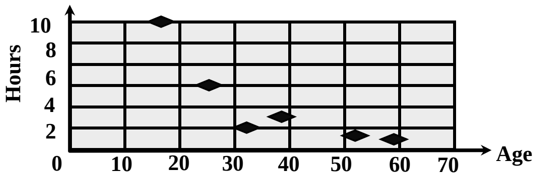
- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................

10-13

**A researcher wishes to determine if a person's age is related to the number of hours he or she exercises per week. The data for the sample are shown below.**

| Age $x$ | 18 | 26 | 32 | 38 | 52 | 59 |
|---|---|---|---|---|---|---|
| Hours $y$ | 10 | 5 | 2 | 3 | 1.5 | 1 |

**a. Draw the scatter plot for the variables.**






## Notes

- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................

10-14

**b. Compute the value of the correlation coefficient.**

| | | | | | | | $\sum$ |
|---|---|---|---|---|---|---|---|
| Age $x$ | 18 | 26 | 32 | 38 | 52 | 59 | 225 |
| Hours $y$ | 10 | 5 | 2 | 3 | 1.5 | 1 | 22.5 |
| $x^2$ | 324 | 676 | 1024 | 1444 | 2704 | 3481 | 9653 |
| $y^2$ | 100 | 25 | 4 | 9 | 2.25 | 1 | 141.25 |
| $x \times y$ | 180 | 130 | 64 | 114 | 78 | 59 | 625 |

$$r = \frac{n\left(\Sigma xy\right) - \left(\Sigma x\right)\left(\Sigma y\right)}{\sqrt{\left[n\left(\Sigma x^2\right) - \left(\Sigma x\right)^2\right]\left[n\left(\Sigma y^2\right)\left(\Sigma y\right)^2\right]}}$$

$$r = \frac{6\left(625\right) - \left(225\right)\left(22.5\right)}{\sqrt{\left[6\left(9653\right) - \left(225\right)^2\right]\left[6\left(141.25\right) - \left(22.5\right)^2\right]}} = -0.832$$

**Thus, there is a strong negative linear relationship which means that older people tend to exercise less on average.**



## Notes

- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................

## Regression Line

10-15

- If the value of the correlation coefficient is significant (will not be discussed here), the next step is to determine the equation of the *regression line* which is the data's line of best fit.
- *Best fit* means that the sum of the squares of the vertical distance from each point to the line is at a minimum. See Figure 10-12 page 545.

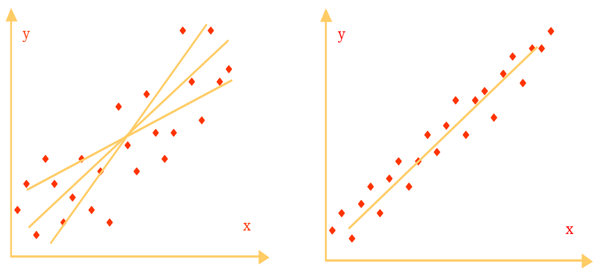


## Notes

- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………

## Scatter Plot and Regression Line

10-16






## Notes

- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………

## Equation of a Line

10-17

- The equation of the regression line is written as $y' = a + bx$, where $b$ is the slope of the line and $a$ is the $y'$ intercept.

* See Figure 10-13 page 546.

- The regression line can be used to predict a value for the <u>dependent</u> variable ($y$) for a given value of the <u>independent</u> variable ($x$).
- <u>Caution:</u> Use $x$ values within the experimental region when predicting $y$ values.



## Notes

- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………

## Regression Line

10-18

- Formulas for the regression line $y' = a + bx$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

where $a$ is the $y'$ intercept and $b$ is the slope of the line.

* See examples 10-9, 10-10 and 10-11



## Notes

- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............

## Assumptions for Valid Predictions in Regression

10-19

1. For any specific value of the independent variable $x$, the value of the dependent variable $y$ must be normally distributed about the regression line.
2. The standard deviation of each of the dependent variables must be the same for each value of the independent variable.

* See Figure 10-16 page 549.

Note: When $r$ is not significantly different from 0, the best predictor of $y$ is the mean of the data values of $y$.

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

## Notes

- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............

10-20

Find the equation of the regression line and find the y′ value for the specified x value. Remember that no regression should be done when $r$ is not significant.

Ages and Exercise

| Age x | 18 | 26 | 32 | 38 | 52 | 59 |
|---|---|---|---|---|---|---|
| Hours y | 10 | 5 | 2 | 3 | 1.5 | 1 |

Find y ′ when x = 35 years.

$$y' = a + bx$$

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

## Notes

- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............
- ……………………………………………………….............

10-21

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

| Age $x$ | 18 | 26 | 32 | 38 | 52 | 59 | $\Sigma$ 225 |
|---|---|---|---|---|---|---|---|
| Hours $y$ | 10 | 5 | 2 | 3 | 1.5 | 1 | 22.5 |
| $x^2$ | 324 | 676 | 1024 | 1444 | 2704 | 3481 | 9653 |
| $y^2$ | 100 | 25 | 4 | 9 | 2.25 | 1 | 141.25 |
| $x \times y$ | 180 | 130 | 64 | 114 | 78 | 59 | 625 |

$$a = \frac{(22.5)(9653) - (225)(625)}{6(9653) - (225)^2} = \boxed{10.499}$$



Notes

- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................

10-22

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

| Age $x$ | 18 | 26 | 32 | 38 | 52 | 59 | $\Sigma$ 225 |
|---|---|---|---|---|---|---|---|
| Hours $y$ | 10 | 5 | 2 | 3 | 1.5 | 1 | 22.5 |
| $x^2$ | 324 | 676 | 1024 | 1444 | 2704 | 3481 | 9653 |
| $y^2$ | 100 | 25 | 4 | 9 | 2.25 | 1 | 141.25 |
| $x \times y$ | 180 | 130 | 64 | 114 | 78 | 59 | 625 |

$$b = \frac{6(625) - (225)(22.5)}{6(9653) - (225)^2} = \boxed{-0.18}$$

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

Notes

- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................

10-23

**Find y ′ when x = 35 years.**

**Ages and Exercise**

| **Age x** | **18** | **26** | **32** | **38** | **52** | **59** |
|---|---|---|---|---|---|---|
| **Hours y** | **10** | **5** | **2** | **3** | **1.5** | **1** |

$a = 10.499 \quad b = -0.18$

$y' = a + bx$

$\boxed{y' = 10.499 - 0.18x}$

$y' = 10.499 - 0.18(35)$

$y' = \boxed{4.199 \text{ hours}}$

**Thus, a person who is 35 years old tends to exercise 4.199 hours per weak on average.**

Dr. Saeed Alghamdi, Statistics Department, Faculty of Sciences, King Abdulaziz University

Notes

- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................
- ........................................................................

# Elementary Statistics

**A Step by Step Approach**
**Sixth Edition**

by
Allan G. Bluman
http://www.mhhe.com/math/stat/blumanbrief

SLIDES PREPARED
BY
LLOYD R. JAISINGH
MOREHEAD STATE UNIVERSITY
MOREHEAD KY

Updated by
Dr. Saeed Alghamdi
King Abdulaziz University
www.kau.edu.sa/saalghamdy

# CHAPTER 13

# Nonparametric Statistics



## The Spearman Rank Correlation Coefficient

13-1

- When the assumption that the populations from which the samples are obtained are normally distributed cannot be met, the nonparametric equivalent of *Pearson product moment correlation coefficient* is *Spearman rank correlation coefficient*.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where $d$ = difference in the ranks and

$n$ = number of data pairs

* See example 13-7



## Notes

- ……………………………………………………………...............
- ……………………………………………………………...........
- ……………………………………………………………...........
- ……………………………………………………………...........
- ……………………………………………………………...........

13-2

**The table shows the total number of tornadoes that occurred in states from 1962 to 1991 and the record high temperatures for the same states.**
**Is there a relationship between the number of tornadoes and the record high temperatures?**

| State | Tornadoes | Record High Temp |
|---|---|---|
| AL | 668 | 112 |
| CO | 781 | 118 |
| FL | 1590 | 109 |
| IL | 798 | 117 |
| KS | 1198 | 121 |
| NY | 169 | 108 |
| PA | 310 | 111 |
| TN | 360 | 113 |
| VT | 21 | 105 |
| WI | 625 | 114 |



## Notes

- ……………………………………………………………...............
- ……………………………………………………………...........
- ……………………………………………………………...........
- ……………………………………………………………...........
- ……………………………………………………………...........

| Tornado | $R_1$ | Temp | $R_2$ | $R_1 - R_2$ | $d^2$ |
|---|---|---|---|---|---|
| 668 | 6 | 112 | 5 | 1 | 1 |
| 781 | 7 | 118 | 9 | –2 | 4 |
| 1590 | 10 | 109 | 3 | 7 | 49 |
| 798 | 8 | 117 | 8 | 0 | 0 |
| 1198 | 9 | 121 | 10 | –1 | 1 |
| 169 | 2 | 108 | 2 | 0 | 0 |
| 310 | 3 | 111 | 4 | –1 | 1 |
| 360 | 4 | 113 | 6 | –2 | 4 |
| 21 | 1 | 105 | 1 | 0 | 0 |
| 625 | 5 | 114 | 7 | –2 | 4 |

$n = 10$

$\sum d^2 = 64$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(64)}{10(10^2-1)}$$

$$= 0.612$$

**There is a moderate positive linear relationship between the number of tornados and the record high temperatures.**



Notes

- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………
- ……………………………………………………………