

Chapters 9 and 10

Review for Exam

Chapter 9

Correlation and Regression

Overview

Paired Data

- ❖ is there a relationship
- ❖ if so, what is the equation
- ❖ use the equation for prediction

Definition

❖ Correlation

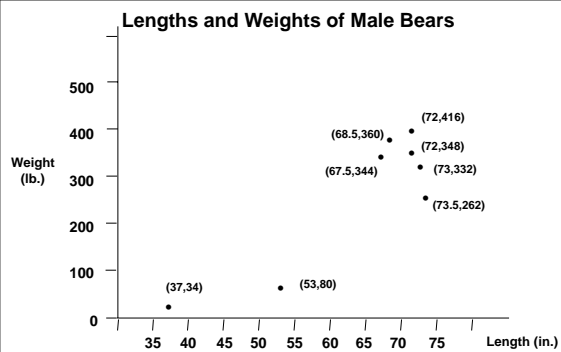
exists between two variables when one of them is related to the other in some way

Definition

❖ Scatterplot (or scatter diagram)

is a graph in which the paired (x,y) sample data are plotted with a horizontal x axis and a vertical y axis. Each individual (x,y) pair is plotted as a single point.

Scatter Diagram of Paired Data



Positive Linear Correlation

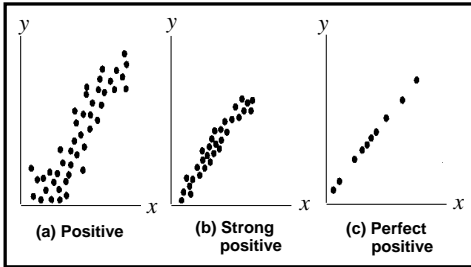


Figure 9-2 Scatter Plots

Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley

7

Negative Linear Correlation

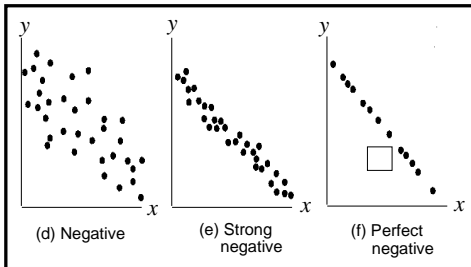


Figure 9-2 Scatter Plots

Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley

8

No Linear Correlation

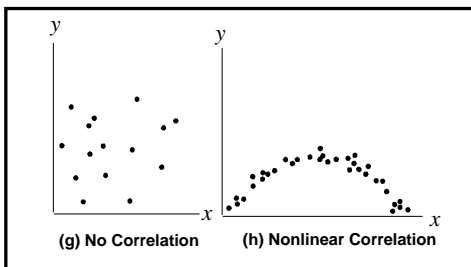


Figure 9-2 Scatter Plots

Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley

9

Definition

❖ Linear Correlation Coefficient r

measures strength of the linear relationship between paired x - and y -quantitative values in a sample

Definition

Linear Correlation Coefficient r

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

Formula 9-1

Calculators can compute r

ρ (rho) is the linear correlation coefficient for all paired data in the population.

Rounding the Linear Correlation Coefficient r

- ❖ Round to three decimal places so that it can be compared to critical values in Table A-5
- ❖ Use calculator or computer if possible

Interpreting the Linear Correlation Coefficient

- ❖ If the absolute value of r exceeds the value in Table A - 5, conclude that there is a significant linear correlation.
- ❖ Otherwise, there is not sufficient evidence to support the conclusion of significant linear correlation.

TABLE A-5 Critical Values of the Pearson Correlation Coefficient r

n	$\alpha = .05$	$\alpha = .01$
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

Properties of the Linear Correlation Coefficient r

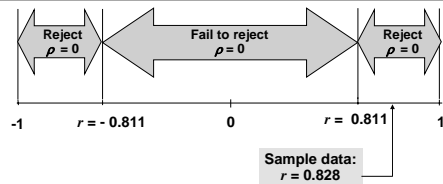
1. $-1 \leq r \leq 1$
2. Value of r does not change if all values of either variable are converted to a different scale.
3. The value of r is not affected by the choice of x and y . Interchange x and y and the value of r will not change.
4. r measures strength of a linear relationship.

Formal Hypothesis Test

- ❖ To determine whether there is a significant linear correlation between two variables
- ❖ Two methods
- ❖ Both methods let $H_0: \rho = 0$
(no significant linear correlation)
 $H_1: \rho \neq 0$
(significant linear correlation)

Method 2: Test Statistic is r (uses fewer calculations)

- ❖ Test statistic: r
- ❖ Critical values: Refer to Table A-5
(no degrees of freedom)



Is there a significant linear correlation?

Data from the Garbage Project								
x Plastic (lb)	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
y Household	2	3	3	6	4	2	1	5

$$n = 8 \quad \alpha = 0.05 \quad H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Test statistic is $r = 0.842$

Regression

Definition

❖ Regression Equation

Given a collection of paired data, the regression equation

$$\hat{y} = b_0 + b_1x$$

algebraically describes the relationship between the two variables

❖ Regression Line

(line of best fit or least-squares line)

the graph of the regression equation

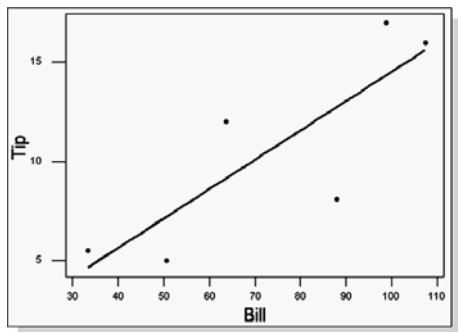
The Regression Equation

x is the independent variable
(predictor variable)

\hat{y} is the dependent variable
(response variable)

$$\hat{y} = b_0 + b_1x \quad b_0 = y - \text{intercept}$$
$$y = mx + b \quad b_1 = \text{slope}$$

Regression Line Plotted on Scatter Plot



Formula for b_1 and b_0

Formula 9-2 $b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$ (slope)

Formula 9-3 $b_0 = \bar{y} - b_1 \bar{x}$ (y-intercept)

calculators or computers can compute these values

Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley 25

Rounding the y-intercept b_0 and the slope b_1

- ❖ Round to three significant digits
- ❖ If you use the formulas 9-2 and 9-3, try not to round intermediate values or carry to at least six significant digits.

Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley 26

Example: Lengths and Weights of Male Bears

x Length (in.) 53.0 67.5 72.0 72.0 73.5 68.5 73.0 37.0

y Weight (lb) 80 344 416 348 262 360 332 34

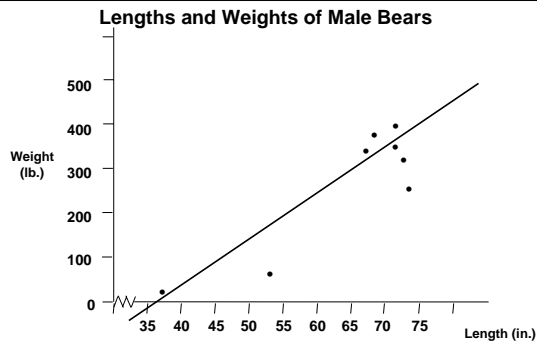
$b_0 = -352$ (rounded)

$b_1 = 9.66$ (rounded)

$\hat{y} = -352 + 9.66x$

Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley 27

Scatter Diagram of Paired Data



Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley

28

Predictions

In predicting a value of y based on some given value of x ...

1. If there is not a significant linear correlation, the best predicted y -value is \bar{y} .
2. If there is a significant linear correlation, the best predicted y -value is found by substituting the x -value into the regression equation.

Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley

29

Guidelines for Using The Regression Equation

1. If there is no significant linear correlation, don't use the regression equation to make predictions.
2. When using the regression equation for predictions, stay within the scope of the available sample data.
3. A regression equation based on old data is not necessarily valid now.
4. Don't make predictions about a population that is different from the population from which the sample data was drawn.

Chapter 9. Section 9-1 and 9-2. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley

30

Example: Lengths and Weights of Male Bears

x Length (in.) 53.0 67.5 72.0 72.0 73.5 68.5 73.0 37.0

y Weight (lb.) 80 344 416 348 262 360 332 34

$$\hat{y} = - 352 + 9.66x$$

What is the weight of a bear that is 60 inches long?

Since the data does have a significant positive linear correlation, we can use the regression equation for prediction.

Example: Lengths and Weights of Male Bears

x Length (in.) 53.0 67.5 72.0 72.0 73.5 68.5 73.0 37.0

y Weight (lb.) 80 344 416 348 262 360 332 34

$$\hat{y} = - 352 + 9.66 (60)$$

$$\hat{y} = 227.6 \text{ pounds}$$

Example: Lengths and Weights of Male Bears

x Length (in.) 53.0 67.5 72.0 72.0 73.5 68.5 73.0 37.0

y Weight (lb.) 80 344 416 348 262 360 332 34

A bear that is 60 inches long will weigh approximately 227.6 pounds.

Example: Lengths and Weights of Male Bears

x Length (in.) 53.0 67.5 72.0 72.0 73.5 68.5 73.0 37.0

y Weight (lb.) 80 344 416 348 262 360 332 34

If there were no significant linear correlation, to predict a weight for any length:

use the average of the weights (y-values)

$$\bar{y} = 272 \text{ lbs}$$

10-2

Multinomial Experiment

Definition

Goodness-of-fit test

used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution

Goodness-of-Fit Test Notation

O represents the observed frequency of an outcome

E represents the expected frequency of an outcome

k represents the number of different categories or outcomes

n represents the total number of trials

Expected Frequencies

If all expected frequencies are equal:

$$E = \frac{n}{k}$$

the sum of all observed frequencies divided by the number of categories

Expected Frequencies

If all expected frequencies are not all equal:

$$E = n p$$

each expected frequency is found by multiplying the sum of all observed frequencies by the probability for the category

Key Question

Are the differences between the observed values (O) and the theoretically expected values (E) statistically significant?

Key Question

We need to measure the discrepancy between O and E; the test statistic will involve their difference:

$$O - E$$

Test Statistic

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Critical Values

1. Found in Table A-4 using k-1 degrees of freedom
where k = number of categories
2. Goodness-of-fit hypothesis tests are always right-tailed.

Multinomial Experiment: Goodness-of-Fit Test

H_0 : No difference between
observed and expected
probabilities

H_1 : at least one of the
probabilities is different
from the others

Categories with Equal Frequencies (Probabilities)

H_0 : $p_1 = p_2 = p_3 = \dots = p_k$

H_1 : at least one of the probabilities is
different from the others

Example: A study was made of 147 industrial accidents that required medical attention. Test the claim that the accidents occur with equal proportions on the 5 workdays.

Frequency of Accidents					
Day	Mon	Tues	Wed	Thurs	Fri
Observed accidents	31	42	18	25	31

Claim: Accidents occur with the same proportion (frequency); that is, $p_1 = p_2 = p_3 = p_4 = p_5$

H_0 : $p_1 = p_2 = p_3 = p_4 = p_5$

H_1 : At least 1 of the 5 proportions is different from others

Categories with Unequal Frequencies

(Probabilities)

H_0 : $p_1, p_2, p_3, \dots, p_k$ are as claimed

H_1 : at least one of the above proportions is different from the claimed value

Example: Mars, Inc. claims its M&M candies are distributed with the color percentages of 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% blue. At the 0.05 significance level, test the claim that the color distribution is as claimed by Mars, Inc.

Claim: $p_1 = 0.30, p_2 = 0.20, p_3 = 0.20, p_4 = 0.10, p_5 = 0.10, p_6 = 0.10$

H_0 : $p_1 = 0.30, p_2 = 0.20, p_3 = 0.20, p_4 = 0.10, p_5 = 0.10, p_6 = 0.10$

H_1 : At least one of the proportions is different from the claimed value.

Example: Mars, Inc. claims its M&M candies are distributed with the color percentages of 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% blue. At the 0.05 significance level, test the claim that the color distribution is as claimed by Mars, Inc.

Frequencies of M&Ms						
	Brown	Yellow	Red	Orange	Green	Blue
Observed frequency	33	26	21	8	7	5

$n = 100$ **Brown** $E = np = (100)(0.30) = 30$

Yellow $E = np = (100)(0.20) = 20$

Red $E = np = (100)(0.20) = 20$

Orange $E = np = (100)(0.10) = 10$

Green $E = np = (100)(0.10) = 10$

Blue $E = np = (100)(0.10) = 10$

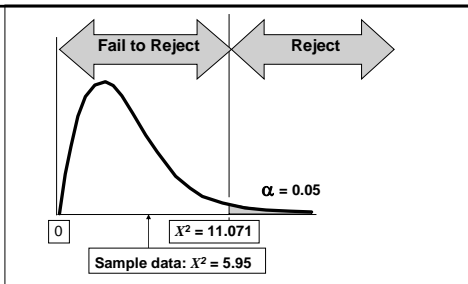
Frequencies of M&Ms

	Brown	Yellow	Red	Orange	Green	Blue
Observed frequency	33	26	21	8	7	5
Expected frequency	30	20	20	10	10	10
(O - E) ² /E	0.3	1.8	0.05	0.4	0.9	2.5

Test Statistic

$$X^2 = \sum \frac{(O - E)^2}{E} = 5.95$$

Critical Value $X^2 = 11.071$
 (with $k-1 = 5$ and $\alpha = 0.05$)



**Test Statistic does not fall within critical region;
 Fail to reject H_0 : percentages are as claimed**
**There is not sufficient evidence to warrant rejection of the
 claim that the colors are distributed with the given
 percentages.**

10-3 Contingency Tables

Definition

- ❖ **Contingency Table** (or two-way frequency table)
a table in which frequencies correspond to two variables.

(One variable is used to categorize rows, and a second variable is used to categorize columns.)

Contingency tables have at least two rows and at least two columns.

Definition

- ❖ **Test of Independence**
tests the null hypothesis that there is no association between the row variable and the column variable.

(The null hypothesis is the statement that the row and column variables are independent.)

Tests of Independence

H_0 : The row variable is independent of the column variable

H_1 : The row variable is dependent (related to) the column variable

This procedure cannot be used to establish a direct cause-and-effect link between variables in question.

Dependence means only there is a relationship between the two variables.

**Test of Independence
Test Statistic**

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Critical Values

1. Found in Table A-4 using degrees of freedom = $(r - 1)(c - 1)$
r is the number of rows and c is the number of columns
2. Tests of Independence are always right-tailed.

$$E = \frac{(\text{row total})(\text{column total})}{(\text{grand total})}$$



**Total number of all observed frequencies
in the table**

Is the type of crime independent of whether the criminal is a stranger?

	Homicide	Robbery	Assault	Row Total
Stranger	12	379	727	1118
Acquaintance or Relative	39	106	642	787
Column Total	51	485	1369	1905

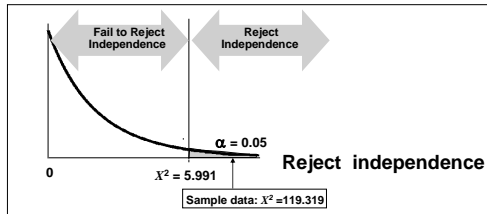
H₀: Type of crime is independent of knowing the criminal

H₁: Type of crime is dependent with knowing the criminal

Test Statistic: $\chi^2 = 119.319$

with $\alpha = 0.05$ and $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ degrees of freedom

Critical Value: $\chi^2 = 5.991$ (from Table A-4)

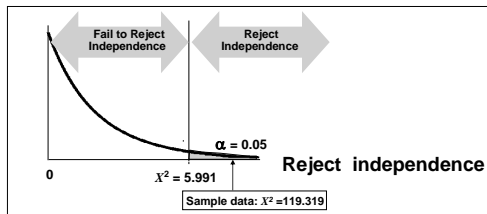


H_0 : The type of crime and knowing the criminal are independent
 H_1 : The type of crime and knowing the criminal are dependent

Test Statistic: $\chi^2 = 119.319$

with $\alpha = 0.05$ and $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ degrees of freedom

Critical Value: $\chi^2 = 5.991$ (from Table A-4)



It appears that the type of crime and knowing the criminal are related.

Definition

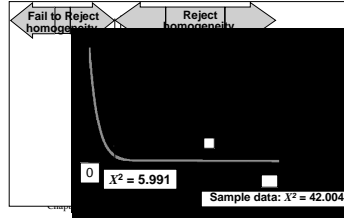
Test of Homogeneity

tests the claim that *different populations* have the same proportions of some characteristics

Example - Test of Homogeneity

		Seat Belt Use in Taxi Cabs		
		New York	Chicago	Pittsburgh
Taxi has	Yes	3	42	2
usable	No	74	87	70
seat belt?				

Claim: The 3 cities have the same proportion of taxis with usable seat belts
 H_0 : The 3 cities have the same proportion of taxis with usable seat belts
 H_a : The proportion of taxis with usable seat belts is not the same in all 3 cities



There is sufficient evidence to warrant rejection of the claim that the 3 cities have the same proportion of usable seat belts in taxis; appears from Table Chicago has a much higher proportion.
