

MAT 212  
Introduction to Business Statistics II  
Lecture Notes

Muhammad El-Taha  
Department of Mathematics and Statistics  
University of Southern Maine  
96 Falmouth Street  
Portland, ME 04104-9300

MAT 212, Spring 97, revised Fall 97, revised Spring 98

**MAT 212**

**Introduction to Business Statistics II**

**Course Content.**

Topic 1: Review and Background

Topic 2: Large Sample Estimation

Topic 3: Large-Sample Tests of Hypothesis

Topic 4: Inferences From Small Sample

Topic 5 The Analysis of Variance

Topic 6 Simple Linear Regression and Correlation

Topic 7: Multiple Linear Regression

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Review and Background</b>  | <b>4</b>  |
| <b>2</b> | <b>Large Sample Estimation</b>  | <b>6</b>  |
| 1        | Introduction . . . . .  | 6         |
| 2        | Point Estimators and Their Properties . . . . .   | 7         |
| 3        | Single Quantitative Population . . . . .  | 7         |
| 4        | Single Binomial Population . . . . .  | 9         |
| 5        | Two Quantitative Populations . . . . .  | 11        |
| 6        | Two Binomial Populations . . . . .  | 12        |
| <b>3</b> | <b>Large-Sample Tests of Hypothesis</b>   | <b>15</b> |
| 1        | Elements of a Statistical Test . . . . .  | 15        |
| 2        | A Large-Sample Statistical Test . . . . .   | 16        |
| 3        | Testing a Population Mean . . . . .   | 17        |
| 4        | Testing a Population Proportion . . . . .   | 18        |
| 5        | Comparing Two Population Means . . . . .  | 19        |
| 6        | Comparing Two Population Proportions . . . . .  | 20        |
| 7        | Reporting Results of Statistical Tests: P-Value . . . . .                                     | 22        |
| <b>4</b> | <b>Small-Sample Tests of Hypothesis</b>   | <b>24</b> |
| 1        | Introduction . . . . .  | 24        |
| 2        | Student's $t$ Distribution . . . . .  | 24        |
| 3        | Small-Sample Inferences About a Population Mean . . . . .                                     | 25        |
| 4        | Small-Sample Inferences About the Difference Between Two Means: Independent Samples . . . . . | 26        |
| 5        | Small-Sample Inferences About the Difference Between Two Means: Paired Samples . . . . .      | 29        |
| 6        | Inferences About a Population Variance . . . . .  | 31        |

|          |  |           |
|----------|--|-----------|
| 7        | Comparing Two Population Variances . . . . .                     | 32        |
| <b>5</b> | <b>Analysis of Variance</b>                                      | <b>34</b> |
| 1        | Introduction . . . . .   | 34        |
| 2        | One Way ANOVA: Completely Randomized Experimental Design . . . . | 35        |
| 3        | The Randomized Block Design . . . . .                            | 38        |
| <b>6</b> | <b>Simple Linear Regression and Correlation</b>                  | <b>43</b> |
| 1        | Introduction . . . . .   | 43        |
| 2        | A Simple Linear Probabilistic Model . . . . .                    | 44        |
| 3        | Least Squares Prediction Equation . . . . .                      | 45        |
| 4        | Inferences Concerning the Slope . . . . .                        | 48        |
| 5        | Estimating $E(y x)$ For a Given $x$ . . . . .                    | 50        |
| 6        | Predicting $y$ for a Given $x$ . . . . .                         | 50        |
| 7        | Coefficient of Correlation . . . . .                             | 50        |
| 8        | Analysis of Variance . . . . .                                   | 51        |
| 9        | Computer Printouts for Regression Analysis . . . . .             | 52        |
| <b>7</b> | <b>Multiple Linear Regression</b>                                | <b>56</b> |
| 1        | Introduction: Example . . . . .                                  | 56        |
| 2        | A Multiple Linear Model . . . . .                                | 56        |
| 3        | Least Squares Prediction Equation . . . . .                      | 57        |

# Chapter 1

## Review and Background

### I. Review and Background

**Probability:** A game of chance

**Statistics:** Branch of science that deals with data analysis

**Course objective:** To make decisions in the presence of uncertainty

#### Terminology

**Information:** A collection of numbers (data)

**Population:** set of all measurements of interest

(e.g. all registered voters, all freshman students at the university)

**Sample:** A subset of measurements selected from the population of interest

**Variable:** A property of an individual population unit (e.g. major, height, weight of freshman students)

**Descriptive Statistics:** deals with procedures used to *summarize* the information contained in a set of measurements.

**Inferential Statistics:** deals with procedures used to make inferences (predictions) about a population parameter from information contained in a sample.

#### Elements of a statistical problem:

- (i) A clear definition of the population and variable of interest.
- (ii) a design of the experiment or sampling procedure.
- (iii) Collection and analysis of data (gathering and summarizing data).
- (iv) Procedure for making predictions about the population based on sample information.
- (v) A measure of “goodness” or reliability for the procedure.

**Types of data:** quantitative vs qualitative

#### Descriptive statistics

## Graphical Methods

Frequency and relative frequency distributions (Histograms):

## Numerical methods

(i) Measures of central tendency

Sample mean:  $\bar{x} = \frac{\sum x_i}{n}$

Sample median: the middle number when the measurements are arranged in ascending order

Sample mode: most frequently occurring value

(ii) Measures of variability

Range:  $r = \max - \min$

Sample Variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Sample standard deviation:  $s = \sqrt{s^2}$

Population parameters vs sample statistics:

Z-score formula:

$$z = \frac{x - \mu_x}{\sigma_x}$$

## Standard normal distribution

Tabulated values

Examples

## The Central Limit Theorem

For large  $n$

(i) the sampling distribution of the sample mean is

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}};$$

(ii) the sampling distribution of the sample proportion is

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}.$$

# Chapter 2

## Large Sample Estimation

### Contents.

1. Introduction
2. Point Estimators and Their Properties
3. Single Quantitative Population
4. Single Binomial Population
5. Two Quantitative Populations
6. Two Binomial Populations
7. Choosing the Sample Size

### 1 Introduction

#### Types of estimators.

1. Point estimator
2. Interval estimator: (L, U)

#### Desired Properties of Point Estimators.

- (i) Unbiased: Mean of the sampling distribution is equal to the parameter.
- (ii) Minimum variance: Small standard error of point estimator.
- (iii) Error of estimation; distance between a parameter and its point estimate.

#### Desired Properties of Interval Estimators.

- (i) Confidence coefficient:  $P(\text{interval estimator will enclose the parameter})=1 - \alpha$ .
- (ii) Confidence level: Confidence coefficient expressed as a percentage.
- (iii) Margin of Error (Bound on the error of estimation).

#### Parameters of Interest.

Single Quantitative Population:  $\mu$

Single Binomial Population:  $p$

Two Quantitative Populations:  $\mu_1 - \mu_2$

Two Binomial Populations:  $p_1 - p_2$

## 2 Point Estimators and Their Properties

Parameter of interest:  $\theta$

Sample data:  $n, \hat{\theta}, \sigma_{\hat{\theta}}$

Point estimator:  $\hat{\theta}$

Estimator mean:  $\mu_{\hat{\theta}} = \theta$  (Unbiased)

Standard error:  $SE(\hat{\theta}) = \sigma_{\hat{\theta}}$

**Assumptions:** Large sample + others (to be specified in each case)

## 3 Single Quantitative Population

Parameter of interest:  $\mu$

Sample data:  $n, \bar{x}, s$

Other information:  $\alpha$

Point estimator:  $\bar{x}$

Estimator mean:  $\mu_{\bar{x}} = \mu$

Standard error:  $SE(\bar{x}) = \sigma/\sqrt{n}$  (also denoted as  $\sigma_{\bar{x}}$ )

**Confidence Interval (C.I.) for  $\mu$ :**

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

**Confidence level:**  $(1 - \alpha)100\%$  which is the probability that the interval estimator contains the parameter.

**Margin of Error.** ( or Bound on the Error of Estimation)

$$B = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

**Assumptions.**

1. Large sample ( $n \geq 30$ )
2. Sample is randomly selected

**Example 1.** We are interested in estimating the mean number of unoccupied seats per flight,  $\mu$ , for a major airline. A random sample of  $n = 225$  flights shows that the sample mean is 11.6 and the standard deviation is 4.1.

Data summary:  $n = 225$ ;  $\bar{x} = 11.6$ ;  $s = 4.1$ .

**Question 1.** What is the point estimate of  $\mu$  ( Do not give the margin of error)?

$$\bar{x} = 11.6$$

**Question 2.** Give a 95% bound on the error of estimation (also known as the margin of error).

$$B = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{4.1}{\sqrt{225}} = 0.5357$$

**Question 3.** Find a 90% confidence interval for  $\mu$ .

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$11.6 \pm 1.645 \frac{4.1}{\sqrt{225}}$$

$$11.6 \pm 0.45 = (11.15, 12.05)$$

**Question 4.** Interpret the CI found in Question 3.

The interval contains  $\mu$  with probability 0.90.

OR

If repeated sampling is used, then 90% of CI constructed would contain  $\mu$ .

**Question 5.** What is the width of the CI found in Question 3.?

The width of the CI is

$$W = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$W = 2(0.45) = 0.90$$

OR

$$W = 12.05 - 11.15 = 0.90$$

**Question 6.** If  $n$ , the sample size, is increased what happens to the width of the CI? what happens to the margin of error?

The width of the CI decreases.

The margin of error decreases.

**Sample size:**

$$n \simeq \frac{(z_{\alpha/2})^2 \sigma^2}{B^2}$$

where  $\sigma$  is estimated by  $s$ .

Note: In the absence of data,  $\sigma$  is sometimes approximated by  $\frac{R}{4}$  where  $R$  is the range.

**Example 2.** Suppose you want to construct a 99% CI for  $\mu$  so that  $W = 0.05$ . You are told that preliminary data shows a range from 13.3 to 13.7. What sample size should you choose?

A. Data summary:  $\alpha = .01$ ;  $R = 13.7 - 13.3 = .4$ ;

so  $\sigma \simeq .4/4 = .1$ . Now

$B = W/2 = 0.05/2 = 0.025$ . Therefore

$$\begin{aligned}n &\simeq \frac{(z_{\alpha/2})^2 \sigma^2}{B^2} \\ &= \frac{2.58^2 (.1)^2}{0.025^2} = 106.50 .\end{aligned}$$

So  $n = 107$ . (round up)

**Exercise 1.** Find the sample size necessary to reduce  $W$  in the flight example to .6. Use  $\alpha = 0.05$ .

## 4 Single Binomial Population

Parameter of interest:  $p$

Sample data:  $n, x, \hat{p} = \frac{x}{n}$  ( $x$  here is the number of successes).

Other information:  $\alpha$

Point estimator:  $\hat{p}$

Estimator mean:  $\mu_{\hat{p}} = p$

Standard error:  $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$

**Confidence Interval (C.I.) for  $p$ :**

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

**Confidence level:**  $(1 - \alpha)100\%$  which is the probability that the interval estimator contains the parameter.

**Margin of Error.**

$$B = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

**Assumptions.**

1. Large sample ( $np \geq 5; nq \geq 5$ )
2. Sample is randomly selected

**Example 3.** A random sample of  $n = 484$  voters in a community produced  $x = 257$  voters in favor of candidate  $A$ .

Data summary:  $n = 484; x = 257; \hat{p} = \frac{x}{n} = \frac{257}{484} = 0.531$ .

**Question 1.** Do we have a large sample size?

$$n\hat{p} = 484(0.531) = 257 \text{ which is } \geq 5.$$

$$n\hat{q} = 484(0.469) = 227 \text{ which is } \geq 5.$$

Therefore we have a large sample size.

**Question 2.** What is the point estimate of  $p$  and its margin of error?

$$\hat{p} = \frac{x}{n} = \frac{257}{484} = 0.531$$

$$B = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.531)(0.469)}{484}} = 0.044$$

**Question 3.** Find a 90% confidence interval for  $p$ .

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.531 \pm 1.645 \sqrt{\frac{(0.531)(0.469)}{484}}$$

$$0.531 \pm 0.037 = (0.494, 0.568)$$

**Question 4.** What is the width of the CI found in Question 3.?

The width of the CI is

$$W = 2z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 2(0.037) = 0.074$$

**Question 5.** Interpret the CI found in Question 3.

The interval contains  $p$  with probability 0.90.

OR

If repeated sampling is used, then 90% of CI constructed would contain  $p$ .

**Question 6.** If  $n$ , the sample size, is increased what happens to the width of the CI? what happens to the margin of error?

The width of the CI decreases.

The margin of error decreases.

**Sample size.**

$$n \simeq \frac{(z_{\alpha/2})^2(\hat{p}\hat{q})}{B^2}.$$

Note: In the absence of data, choose  $\hat{p} = \hat{q} = 0.5$  or simply  $\hat{p}\hat{q} = 0.25$ .

**Example 4.** Suppose you want to provide an accurate estimate of customers preferring one brand of coffee over another. You need to construct a 95% CI for  $p$  so that  $B = 0.015$ . You are told that preliminary data shows a  $\hat{p} = 0.35$ . What sample size should you choose? Use  $\alpha = 0.05$ .

Data summary:  $\alpha = .05$ ;  $\hat{p} = 0.35$ ;  $B = 0.015$

$$\begin{aligned} n &\simeq \frac{(z_{\alpha/2})^2(\hat{p}\hat{q})}{B^2} \\ &= \frac{(1.96)^2(0.35)(0.65)}{0.015^2} = 3,884.28 \end{aligned}$$

So  $n = 3,885$ . (round up)

**Exercise 2.** Suppose that no preliminary estimate of  $\hat{p}$  is available. Find the new sample size. Use  $\alpha = 0.05$ .

**Exercise 3.** Suppose that no preliminary estimate of  $\hat{p}$  is available. Find the sample size necessary so that  $\alpha = 0.01$ .

## 5 Two Quantitative Populations

Parameter of interest:  $\mu_1 - \mu_2$

Sample data:

Sample 1:  $n_1, \bar{x}_1, s_1$

Sample 2:  $n_2, \bar{x}_2, s_2$

Point estimator:  $\bar{X}_1 - \bar{X}_2$

Estimator mean:  $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

Standard error:  $SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

**Confidence Interval.**

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

**Assumptions.**

1. Large samples (  $n_1 \geq 30; n_2 \geq 30$ )
2. Samples are randomly selected
3. Samples are independent

**Sample size.**

$$n \simeq \frac{(z_{\alpha/2})^2(\sigma_1^2 + \sigma_2^2)}{B^2}$$

## 6 Two Binomial Populations

Parameter of interest:  $p_1 - p_2$

Sample 1:  $n_1, x_1, \hat{p}_1 = \frac{x_1}{n_1}$

Sample 2:  $n_2, x_2, \hat{p}_2 = \frac{x_2}{n_2}$

$p_1 - p_2$  (unknown parameter)

$\alpha$  (significance level)

Point estimator:  $\hat{p}_1 - \hat{p}_2$

Estimator mean:  $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$

Estimated standard error:  $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

**Confidence Interval.**

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

**Assumptions.**

1. Large samples,

$$(n_1 p_1 \geq 5, n_1 q_1 \geq 5, n_2 p_2 \geq 5, n_2 q_2 \geq 5)$$

2. Samples are randomly and independently selected

**Sample size.**

$$n \simeq \frac{(z_{\alpha/2})^2(\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2)}{B^2}$$

For unknown parameters:

$$n \simeq \frac{(z_{\alpha/2})^2(0.5)}{B^2}$$

### Review Exercises: Large-Sample Estimation

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. A random sample of size  $n = 100$  is selected from a quantitative population. The data produced a mean and standard deviation of  $\bar{x} = 75$  and  $s = 6$  respectively.

(i) Estimate the population mean  $\mu$ , and give a 95% bound on the error of estimation (or margin of error). (Answer: B=1.18)

(ii) Find a 99% confidence interval for the population mean. (Answer: B=1.55)

(iii) Interpret the confidence interval found in (ii).

(iv) Find the sample size necessary to reduce the width of the confidence interval in (ii) by half. (Answer: n=400)

2. An examination of the yearly premiums for a random sample of 80 automobile insurance policies from a major company showed an average of \$329 and a standard deviation of \$49.

(i) Give the point estimate of the population parameter  $\mu$  and a 99% bound on the error of estimation. (Margin of error). (Answer: B=14.135)

(ii) Construct a 99% confidence interval for  $\mu$ .

(iii) Suppose we wish our estimate in (i) to be accurate to within \$5 with 95% confidence; how many insurance policies should be sampled to achieve the desired level of accuracy? (Answer: n=369)

3. Suppose we wish to estimate the average daily yield of a chemical manufactured in a chemical plant. The daily yield recorded for  $n = 100$  days, produces a mean and standard deviation of  $\bar{x} = 870$  and  $s = 20$  tons respectively.

(i) Estimate the average daily yield  $\mu$ , and give a 95% bound on the error of estimation (or margin of error).

(ii) Find a 99% confidence interval for the population mean.

(iii) Interpret the confidence interval found in (ii).

(iv) Find the sample size necessary to reduce the width of the confidence interval in (ii) by half.

4. Answer by True or False . (Circle your choice).

T F (i) If the population variance increases and other factors are the same, the width of the confidence interval for the population mean tends to increase.

T F (ii) As the sample size increases, the width of the confidence interval for the population mean tends to decrease.

T F (iii) Populations are characterized by numerical descriptive measures called *statistics*.

T F (iv) If, for a given C.I.,  $\alpha$  is increased, then the margin of error will increase.

T F (v) The sample standard deviation  $s$  can be used to approximate  $\sigma$  when  $n$  is larger than 30.

T F (vi) The sample mean always lies above the population mean.

# Chapter 3

## Large-Sample Tests of Hypothesis

### Contents.

1. Elements of a statistical test
2. A Large-sample statistical test
3. Testing a population mean
4. Testing a population proportion
5. Testing the difference between two population means
6. Testing the difference between two population proportions
7. Reporting results of statistical tests: p-Value

### 1 Elements of a Statistical Test

Null hypothesis:  $H_0$

Alternative (research) hypothesis:  $H_a$

Test statistic:

Rejection region : reject  $H_0$  if .....

Graph:

Decision: either “Reject  $H_0$ ” or “Do not reject  $H_0$ ”

Conclusion: At  $100\alpha\%$  significance level there is (in)sufficient statistical evidence to “favor  $H_a$ ” .

Comments:

\*  $H_0$  represents the status-quo

\*  $H_a$  is the hypothesis that we want to provide evidence to justify. We show that  $H_a$  is true by showing that  $H_0$  is false, that is proof by contradiction.

Type I error  $\equiv \{ \text{reject } H_0 | H_0 \text{ is true} \}$

Type II error  $\equiv \{ \text{do not reject } H_0 | H_0 \text{ is false} \}$

$\alpha = \text{Prob}\{\text{Type I error}\}$

$\beta = \text{Prob}\{\text{Type II error}\}$

Power of a statistical test:

$\text{Prob}\{\text{reject } H_0 \mid H_0 \text{ is false}\} = 1 - \beta$

**Example 1.**

$H_0$ : Innocent

$H_a$ : Guilty

$\alpha = \text{Prob}\{\text{sending an innocent person to jail}\}$

$\beta = \text{Prob}\{\text{letting a guilty person go free}\}$

**Example 2.**

$H_0$ : New drug is not acceptable

$H_a$ : New drug is acceptable

$\alpha = \text{Prob}\{\text{marketing a bad drug}\}$

$\beta = \text{Prob}\{\text{not marketing an acceptable drug}\}$

## 2 A Large-Sample Statistical Test

Parameter of interest:  $\theta$

Sample data:  $n, \hat{\theta}, \sigma_{\hat{\theta}}$

**Test:**

Null hypothesis ( $H_0$ ):  $\theta = \theta_0$

Alternative hypothesis ( $H_a$ ): 1)  $\theta > \theta_0$ ; 2)  $\theta < \theta_0$ ; 3)  $\theta \neq \theta_0$

Test statistic (TS):

$$z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

Critical value: either  $z_\alpha$  or  $z_{\alpha/2}$

Rejection region (RR) :

1) Reject  $H_0$  if  $z > z_\alpha$

2) Reject  $H_0$  if  $z < -z_\alpha$

3) Reject  $H_0$  if  $z > z_{\alpha/2}$  or  $z < -z_{\alpha/2}$

Graph:

Decision: 1) if observed value is in RR: "Reject  $H_0$ "

2) if observed value is not in RR: "Do not reject  $H_0$ "

Conclusion: At  $100\alpha\%$  significance level there is (in)sufficient statistical evidence to  
...

**Assumptions:** Large sample + others (to be specified in each case).

One tailed statistical test

Upper (right) tailed test

Lower (left) tailed test

Two tailed statistical test

### 3 Testing a Population Mean

Parameter of interest:  $\mu$

Sample data:  $n, \bar{x}, s$

Other information:  $\mu_0 =$  target value,  $\alpha$

**Test:**

$H_0 : \mu = \mu_0$

$H_a : 1) \mu > \mu_0; 2) \mu < \mu_0; 3) \mu \neq \mu_0$

T.S. :

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Rejection region (RR) :

1) Reject  $H_0$  if  $z > z_\alpha$

2) Reject  $H_0$  if  $z < -z_\alpha$

3) Reject  $H_0$  if  $z > z_{\alpha/2}$  or  $z < -z_{\alpha/2}$

Graph:

Decision: 1) if observed value is in RR: "Reject  $H_0$ "

2) if observed value is not in RR: "Do not reject  $H_0$ "

Conclusion: At  $100\alpha\%$  significance level there is (in)sufficient statistical evidence to  
"favor  $H_a$ ".

**Assumptions:**

Large sample ( $n \geq 30$ )

Sample is randomly selected

**Example:** Test the hypothesis that weight loss in a new diet program exceeds 20 pounds during the first month.

Sample data :  $n = 36, \bar{x} = 21, s^2 = 25, \mu_0 = 20, \alpha = 0.05$

$H_0 : \mu = 20$  ( $\mu$  is not larger than 20)

$H_a : \mu > 20$  ( $\mu$  is larger than 20)

T.S. :

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21 - 20}{5/\sqrt{36}} = 1.2$$

Critical value:  $z_\alpha = 1.645$

RR: Reject  $H_0$  if  $z > 1.645$

Graph:

Decision: Do not reject  $H_0$

Conclusion: At 5% significance level there is insufficient statistical evidence to conclude that weight loss in a new diet program exceeds 20 pounds per first month.

**Exercise:** Test the claim that weight loss is not equal to 19.5.

## 4 Testing a Population Proportion

Parameter of interest:  $p$  (unknown parameter)

Sample data:  $n$  and  $x$  (or  $\hat{p} = \frac{x}{n}$ )

$p_0$  = target value

$\alpha$  (significance level)

**Test:**

$H_0 : p = p_0$

$H_a$ : 1)  $p > p_0$ ; 2)  $p < p_0$ ; 3)  $p \neq p_0$

T.S. :

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

RR:

1) Reject  $H_0$  if  $z > z_\alpha$

2) Reject  $H_0$  if  $z < -z_\alpha$

3) Reject  $H_0$  if  $z > z_{\alpha/2}$  or  $z < -z_{\alpha/2}$

Graph:

Decision:

1) if observed value is in RR: "Reject  $H_0$ "

2) if observed value is not in RR: "Do not reject  $H_0$ "

Conclusion: At ( $\alpha$ )100% significance level there is (in)sufficient statistical evidence to "favor  $H_a$ ".

**Assumptions:**

1. Large sample ( $np \geq 5, nq \geq 5$ )
2. Sample is randomly selected

**Example.** Test the hypothesis that  $p > .10$  for sample data:  $n = 200, x = 26$ .

**Solution.**

$$\hat{p} = \frac{x}{n} = \frac{26}{200} = .13,$$

Now

$$H_0 : p = .10$$

$$H_a : p > .10$$

TS:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}} = \frac{.13 - .10}{\sqrt{(.10)(.90)/200}} = 1.41$$

RR: reject  $H_0$  if  $z > 1.645$

Graph:

Dec: Do not reject  $H_0$

Conclusion: At 5% significance level there is insufficient statistical evidence to conclude that  $p > .10$ .

**Exercise** Is the large sample assumption satisfied here ?

## 5 Comparing Two Population Means

Parameter of interest:  $\mu_1 - \mu_2$

Sample data:

Sample 1:  $n_1, \bar{x}_1, s_1$

Sample 2:  $n_2, \bar{x}_2, s_2$

**Test:**

$$H_0 : \mu_1 - \mu_2 = D_0$$

$$H_a : 1) \mu_1 - \mu_2 > D_0; 2) \mu_1 - \mu_2 < D_0;$$

$$3) \mu_1 - \mu_2 \neq D_0$$

T.S. :

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

RR:

1) Reject  $H_0$  if  $z > z_\alpha$

2) Reject  $H_0$  if  $z < -z_\alpha$

3) Reject  $H_0$  if  $z > z_{\alpha/2}$  or  $z < -z_{\alpha/2}$

Graph:

Decision:

Conclusion:

**Assumptions:**

1. Large samples (  $n_1 \geq 30; n_2 \geq 30$  )
2. Samples are randomly selected
3. Samples are independent

**Example:** (Comparing two weight loss programs)

Refer to the weight loss example. Test the hypothesis that weight loss in the two diet programs are different.

1. Sample 1 :  $n_1 = 36, \bar{x}_1 = 21, s_1^2 = 25$  (old)
2. Sample 2 :  $n_2 = 36, \bar{x}_2 = 18.5, s_2^2 = 24$  (new)

$D_0 = 0, \alpha = 0.05$

$H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0,$

T.S. :

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 2.14$$

Critical value:  $z_{\alpha/2} = 1.96$

RR: Reject  $H_0$  if  $z > 1.96$  or  $z < -1.96$

Graph:

Decision: Reject  $H_0$

Conclusion: At 5% significance level there is sufficient statistical evidence to conclude that weight loss in the two diet programs are different.

**Exercise:** Test the hypothesis that weight loss in the old diet program exceeds that of the new program.

**Exercise:** Test the claim that the difference in mean weight loss for the two programs is greater than 1.

## 6 Comparing Two Population Proportions

Parameter of interest:  $p_1 - p_2$

Sample 1:  $n_1, x_1, \hat{p}_1 = \frac{x_1}{n_1},$

Sample 2:  $n_2, x_2, \hat{p}_2 = \frac{x_2}{n_2}$ ,  
 $p_1 - p_2$  (unknown parameter)  
 Common estimate:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

**Test:**

$H_0 : p_1 - p_2 = 0$   
 $H_a : 1) p_1 - p_2 > 0$   
 2)  $p_1 - p_2 < 0$   
 3)  $p_1 - p_2 \neq 0$   
 T.S. :

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

RR:

- 1) Reject  $H_0$  if  $z > z_\alpha$
- 2) Reject  $H_0$  if  $z < -z_\alpha$
- 3) Reject  $H_0$  if  $z > z_{\alpha/2}$  or  $z < -z_{\alpha/2}$

Graph:

Decision:

Conclusion:

**Assumptions:**

Large sample ( $n_1p_1 \geq 5, n_1q_1 \geq 5, n_2p_2 \geq 5, n_2q_2 \geq 5$ )  
 Samples are randomly and independently selected

**Example:** Test the hypothesis that  $p_1 - p_2 < 0$  if it is known that the test statistic is  $z = -1.91$ .

**Solution:**

$H_0 : p_1 - p_2 = 0$   
 $H_a : p_1 - p_2 < 0$   
 TS:  $z = -1.91$   
 RR: reject  $H_0$  if  $z < -1.645$   
 Graph:  
 Dec: reject  $H_0$

Conclusion: At 5% significance level there is sufficient statistical evidence to conclude that  $p_1 - p_2 < 0$ .

**Exercise:** Repeat as a two tailed test

## 7 Reporting Results of Statistical Tests: P-Value

**Definition.** The p-value for a test of a hypothesis is the smallest value of  $\alpha$  for which the null hypothesis is rejected, i.e. the statistical results are significant.

The p-value is called the *observed significance level*

Note: The p-value is the probability ( when  $H_0$  is true) of obtaining a value of the test statistic as extreme or more extreme than the actual sample value in support of  $H_a$ .

**Examples.** Find the p-value in each case:

(i) Upper tailed test:

$$H_0 : \theta = \theta_0$$

$$H_a : \theta > \theta_0$$

$$\text{TS: } z = 1.76$$

$$\text{p-value} = .0392$$

(ii) Lower tailed test:

$$H_0 : \theta = \theta_0$$

$$H_a : \theta < \theta_0$$

$$\text{TS: } z = -1.86$$

$$\text{p-value} = .0314$$

(iii) Two tailed test:

$$H_0 : \theta = \theta_0$$

$$H_a : \theta \neq \theta_0$$

$$\text{TS: } z = 1.76$$

$$\text{p-value} = 2(.0392) = .0784$$

**Decision rule using p-value: (Important)**

Reject  $H_0$  for all  $\alpha > p - \text{value}$

### Review Exercises: Testing Hypothesis

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. A local pizza parlor advertises that their average time for delivery of a pizza is within 30 minutes of receipt of the order. The delivery time for a random sample of 64

orders were recorded, with a sample mean of 34 minutes and a standard deviation of 21 minutes.

(i) Is there sufficient evidence to conclude that the actual delivery time is larger than what is claimed by the pizza parlor? Use  $\alpha = .05$ .

$H_0$ :

$H_a$ :

T.S. (Answer: 1.52)

R.R.

Graph:

Dec:

Conclusion:

((ii) Test the hypothesis that  $H_a : \mu \neq 30$ .

2. Answer by True or False . (Circle your choice).

T F (v) If, for a given test,  $\alpha$  is fixed and the sample size is increased, then  $\beta$  will increase.

# Chapter 4

## Small-Sample Tests of Hypothesis

### Contents:

1. Introduction
2. Student's  $t$  distribution
3. Small-sample inferences about a population mean
4. Small-sample inferences about the difference between two means: Independent Samples
5. Small-sample inferences about the difference between two means: Paired Samples
6. Inferences about a population variance
7. Comparing two population variances

### 1 Introduction

When the sample size is small we only deal with normal populations.  
For non-normal (e.g. binomial) populations different techniques are necessary

### 2 Student's $t$ Distribution

RECALL

For small samples ( $n < 30$ ) from normal populations, we have

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

If  $\sigma$  is unknown, we use  $s$  instead; but we no more have a  $Z$  distribution  
**Assumptions.**

1. Sampled population is normal
2. Small random sample ( $n < 30$ )
3.  $\sigma$  is unknown

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

**Properties of the  $t$  Distribution:**

- (i) It has  $n - 1$  degrees of freedom (df)
- (ii) Like the normal distribution it has a symmetric mound-shaped probability distribution
- (iii) More variable (flat) than the normal distribution
- (iv) The distribution depends on the degrees of freedom. Moreover, as  $n$  becomes larger,  $t$  converges to  $Z$ .
- (v) Critical values (tail probabilities) are obtained from the  $t$  table

**Examples.**

- (i) Find  $t_{0.05,5} = 2.015$
- (ii) Find  $t_{0.005,8} = 3.355$
- (iii) Find  $t_{0.025,26} = 2.056$

### 3 Small-Sample Inferences About a Population Mean

Parameter of interest:  $\mu$

Sample data:  $n, \bar{x}, s$

Other information:  $\mu_0 =$  target value,  $\alpha$

Point estimator:  $\bar{x}$

Estimator mean:  $\mu_{\bar{x}} = \mu$

Estimated standard error:  $\sigma_{\bar{x}} = s/\sqrt{n}$

**Confidence Interval for  $\mu$ :**

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \left( \frac{s}{\sqrt{n}} \right)$$

Test:

$H_0 : \mu = \mu_0$

$H_a : 1) \mu > \mu_0; 2) \mu < \mu_0; 3) \mu \neq \mu_0.$

Critical value: either  $t_{\alpha, n-1}$  or  $t_{\frac{\alpha}{2}, n-1}$

$$\text{T.S. : } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

RR:

- 1) Reject  $H_0$  if  $t > t_{\alpha, n-1}$
- 2) Reject  $H_0$  if  $t < -t_{\alpha, n-1}$
- 3) Reject  $H_0$  if  $t > t_{\frac{\alpha}{2}, n-1}$  or  $t < -t_{\frac{\alpha}{2}, n-1}$

Graph:

Decision: 1) if observed value is in RR: “Reject  $H_0$ ”

2) if observed value is not in RR: “Do not reject  $H_0$ ”

Conclusion: At  $100\alpha\%$  significance level there is (in)sufficient statistical evidence to “favor  $H_a$ ” .

### Assumptions.

1. Small sample ( $n < 30$ )
2. Sample is randomly selected
3. Normal population
4. Unknown variance

**Example** For the sample data given below, test the hypothesis that weight loss in a new diet program exceeds 20 pounds per first month.

1. Sample data:  $n = 25, \bar{x} = 21.3, s^2 = 25, \mu_0 = 20, \alpha = 0.05$

Critical value:  $t_{0.05, 24} = 1.711$

$H_0 : \mu = 20$

$H_a : \mu > 20,$

T.S.:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21.3 - 20}{5/\sqrt{25}} = 1.3$$

RR: Reject  $H_0$  if  $t > 1.711$

Graph:

Decision: Do not reject  $H_0$

Conclusion: At 5% significance level there is insufficient statistical evidence to conclude that weight loss in a new diet program exceeds 20 pounds per first month.

**Exercise.** Test the claim that weight loss is not equal to 19.5, (i.e.  $H_a : \mu \neq 19.5$ ).

## 4 Small-Sample Inferences About the Difference Between Two Means: Independent Samples

Parameter of interest:  $\mu_1 - \mu_2$

Sample data:

Sample 1:  $n_1, \bar{x}_1, s_1$

Sample 2:  $n_2, \bar{x}_2, s_2$

Other information:  $D_0$  = target value,  $\alpha$

Point estimator:  $\bar{X}_1 - \bar{X}_2$

Estimator mean:  $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

**Assumptions.**

1. Normal populations
2. Small samples ( $n_1 < 30; n_2 < 30$ )
3. Samples are randomly selected
4. Samples are independent
5. Variances are equal with common variance

$$\sigma^2 = \sigma_1^2 = \sigma_2^2$$

**Pooled estimator for  $\sigma$ .**

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Estimator standard error:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Reason:

$$\begin{aligned}\sigma_{\bar{X}_1 - \bar{X}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \\ &= \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\end{aligned}$$

**Confidence Interval:**

$$(\bar{x}_1 - \bar{x}_2) \pm (t_{\alpha/2, n_1 + n_2 - 2}) \left( s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

**Test:**

$$H_0 : \mu_1 - \mu_2 = D_0$$

$H_a$  : 1)  $\mu_1 - \mu_2 > D_0$ ; 2)  $\mu_1 - \mu_2 < D_0$ ;

3)  $\mu_1 - \mu_2 \neq D_0$

T.S. :

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

RR: 1) Reject  $H_0$  if  $t > t_{\alpha, n_1+n_2-2}$

2) Reject  $H_0$  if  $t < -t_{\alpha, n_1+n_2-2}$

3) Reject  $H_0$  if  $t > t_{\alpha/2, n_1+n_2-2}$  OR  $t < -t_{\alpha/2, n_1+n_2-2}$

Graph:

Decision:

Conclusion:

**Example.**(Comparison of two weight loss programs)

Refer to the weight loss example. Test the hypothesis that weight loss in a new diet program is different from that of an old program. We are told that that the observed value is 2.2 and the we know that

1. Sample 1 :  $n_1 = 7$

2. Sample 2 :  $n_2 = 8$

$\alpha = 0.05$

**Solution.**

$H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0$

T.S. :

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2.2$$

Critical value:  $t_{.025, 13} = 2.160$

RR: Reject  $H_0$  if  $t > 2.160$  or  $t < -2.160$

Graph:

Decision: Reject  $H_0$

Conclusion: At 5% significance level there is sufficient statistical evidence to conclude that weight loss in the two diet programs are different.

**Exercise:** Test the claim that the difference in mean weight loss for the two programs is greater than 0.

Minitab Commands: A twosample  $t$  procedure with a pooled estimate of variance

MTB> twosample C1 C2;

SUBC>pooled;

SUBC> alternative 1.

Note: alternative : 1=right-tailed; -1=left tailed; 0=two tailed.

## 5 Small-Sample Inferences About the Difference Between Two Means: Paired Samples

Parameter of interest:  $\mu_1 - \mu_2 = \mu_d$

Sample of paired differences data:

Sample :  $n$  = number of pairs,  $\bar{d}$  = sample mean,  $s_d$

Other information:  $D_0$ = target value,  $\alpha$

Point estimator:  $\bar{d}$

Estimator mean:  $\mu_{\bar{d}} = \mu_d$

### Assumptions.

1. Normal populations
2. Small samples (  $n_1 < 30; n_2 < 30$ )
3. Samples are randomly selected
4. Samples are paired (not independent)

Sample standard deviation of the sample of  $n$  paired differences

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$

Estimator standard error:  $\sigma_{\bar{d}} = s_d/\sqrt{n}$

### Confidence Interval.

$$\bar{d} \pm t_{\alpha/2, n-1} s_d/\sqrt{n}$$

### Test.

$H_0 : \mu_1 - \mu_2 = D_0$  (equivalently,  $\mu_d = D_0$ )

$H_a : 1) \mu_1 - \mu_2 = \mu_d > D_0; 2) \mu_1 - \mu_2 = \mu_d < D_0;$

3)  $\mu_1 - \mu_2 = \mu_d \neq D_0,$

T.S. :

$$t = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

RR:

1) Reject  $H_0$  if  $t > t_{\alpha, n-1}$

2) Reject  $H_0$  if  $t < -t_{\alpha, n-1}$

3) Reject  $H_0$  if  $t > t_{\alpha/2, n-1}$  or  $t < -t_{\alpha/2, n-1}$

Graph:

Decision:

Conclusion:

**Example.** A manufacturer wishes to compare wearing qualities of two different types of tires,  $A$  and  $B$ . For the comparison a tire of type  $A$  and one of type  $B$  are randomly assigned and mounted on the rear wheels of each of five automobiles. The automobiles are then operated for a specified number of miles, and the amount of wear is recorded for each tire. These measurements are tabulated below.

| Automobile | Tire A | Tire B |
|------------|--------|--------|
| 1          | 10.6   | 10.2   |
| 2          | 9.8    | 9.4    |
| 3          | 12.3   | 11.8   |
| 4          | 9.7    | 9.1    |
| 5          | 8.8    | 8.3    |

$$\bar{x}_1 = 10.24 \quad \bar{x}_2 = 9.76$$

Using the previous section test we would have  $t = 0.57$  resulting in an insignificant test which is inconsistent with the data.

| Automobile | Tire A | Tire B | d=A-B |
|------------|--------|--------|-------|
| 1          | 10.6   | 10.2   | .4    |
| 2          | 9.8    | 9.4    | .4    |
| 3          | 12.3   | 11.8   | .5    |
| 4          | 9.7    | 9.1    | .6    |
| 5          | 8.8    | 8.3    | .5    |

$$\bar{x}_1 = 10.24 \quad \bar{x}_2 = 9.76 \quad \bar{d} = .48$$

Q1: Provide a summary of the data in the above table.

Sample summary:  $n = 5, \bar{d} = .48, s_d = .0837$

Q2: Do the data provide sufficient evidence to indicate a difference in average wear for the two tire types.

**Test.** (parameter  $\mu_d = \mu_1 - \mu_2$ )

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0$$

T.S. :

$$t = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} = \frac{.48 - 0}{.0837/\sqrt{5}} = 12.8$$

RR: Reject  $H_0$  if  $t > 2.776$  or  $t < -2.776$  ( $t_{.025,4} = 2.776$ )

Graph:

Decision: Reject  $H_0$

Conclusion: At 5% significance level there is sufficient statistical evidence to conclude that the average amount of wear for type A tire is different from that for type B tire.

**Exercise.** Construct a 99% confidence interval for the difference in average wear for the two tire types.

## 6 Inferences About a Population Variance

**Chi-square distribution.** When a random sample of size  $n$  is drawn from a normal population with mean  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of  $S^2$  depends on  $n$ . The standardized distribution of  $S^2$  is called the chi-square distribution and is given by

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Degrees of freedom (df):  $\nu = n - 1$

Graph: Non-symmetrical and depends on df

Critical values: using  $\chi^2$  tables

**Test.**

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \sigma^2 \neq \sigma_0^2 \text{ (two-tailed test).}$$

T.S. :

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

RR: Reject  $H_0$  if  $\chi^2 > \chi_{\alpha/2}^2$  or  $\chi^2 < \chi_{1-\alpha/2}^2$  where  $\chi^2$  is based on  $(n-1)$  degrees of freedom.

Graph:

Decision:

Conclusion:

**Assumptions.**

1. Normal population
2. Random sample

**Example:**

Use text

## 7 Comparing Two Population Variances

**F-distribution.** When independent samples are drawn from two normal populations with equal variances then  $S_1^2/S_2^2$  possesses a sampling distribution that is known as an **F distribution**. That is

$$F = \frac{s_1^2}{s_2^2}$$

Degrees of freedom (df):  $\nu_1 = n_1 - 1; \nu_2 = n_2 - 1$

Graph: Non-symmetrical and depends on df

Critical values: using F tables

**Test.**

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2 \text{ (two-tailed test).}$$

T.S. :  $F = \frac{s_1^2}{s_2^2}$  where  $s_1^2$  is the larger sample variance.

Note:  $F = \frac{\text{larger sample variance}}{\text{smaller sample variance}}$

RR: Reject  $H_0$  if  $F > F_{\alpha/2}$  where  $F_{\alpha/2}$  is based on  $(n_1 - 1)$  and  $(n_2 - 1)$  degrees of freedom.

Graph:

Decision:

Conclusion:

**Assumptions.**

1. Normal populations
2. Independent random samples

**Example.** (Investment Risk) Investment risk is generally measured by the volatility of possible outcomes of the investment. The most common method for measuring investment volatility is by computing the variance ( or standard deviation) of possible outcomes. Returns over the past 10 years for first alternative and 8 years for the second alternative produced the following data:

Data Summary:

Investment 1:  $n_1 = 10, \bar{x}_1 = 17.8\%; s_1^2 = 3.21$

Investment 2:  $n_2 = 8, \bar{x}_2 = 17.8\%; s_2^2 = 7.14$

Both populations are assumed to be normally distributed.

Q1: Do the data present sufficient evidence to indicate that the risks for investments 1 and 2 are unequal ?

**Solution.**

Test:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2 \text{ (two-tailed test).}$$

T.S. :

$$F = \frac{s_2^2}{s_1^2} = \frac{7.14}{3.21} = 2.22$$

RR: Reject  $H_0$  if  $F > F_{\alpha/2}$  where

$$F_{\alpha/2, n_2-1, n_1-1} = F_{.025, 7, 9} = 4.20$$

Graph:

Decision: Do not reject  $H_0$

Conclusion: At 5% significance level there is insufficient statistical evidence to indicate that the risks for investments 1 and 2 are unequal.

**Exercise.** Do the upper tail test. That is  $H_a : \sigma_1^2 > \sigma_2^2$ .

# Chapter 5

## Analysis of Variance

### Contents.

1. Introduction
2. One Way ANOVA: Completely Randomized Experimental Design
3. The Randomized Block Design

## 1 Introduction

Analysis of variance is a statistical technique used to compare more than two population means by isolating the sources of variability.

**Example.** Four groups of sales people for a magazine sales agency were subjected to different sales training programs. Because there were some dropouts during the training program, the number of trainees varied from program to program. At the end of the training programs each salesperson was assigned a sales area from a group of sales areas that were judged to have equivalent sales potentials. The table below lists the number of sales made by each person in each of the four groups of sales people during the first week after completing the training program. Do the data present sufficient evidence to indicate a difference in the mean achievement for the four training programs?

**Goal.** Test whether the means are equal or not. That is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a : \text{Not all means are equal}$$

Definitions:

- (i) Response: variable of interest or dependent variable (sales)
- (ii) Factor: categorical variable or independent variable (training technique)
- (iii) Treatment levels (factor levels): method of training;  $t = 4$

| Training Group |           |           |           |           |          |
|----------------|-----------|-----------|-----------|-----------|----------|
|                | 1         | 2         | 3         | 4         |          |
|                | 65        | 75        | 59        | 94        |          |
|                | 87        | 69        | 78        | 89        |          |
|                | 73        | 83        | 67        | 80        |          |
|                | 79        | 81        | 62        | 88        |          |
|                | 81        | 72        | 83        |           |          |
|                | 69        | 79        | 76        |           |          |
|                |           | 90        |           |           |          |
|                | $n_1 = 6$ | $n_2 = 7$ | $n_3 = 6$ | $n_4 = 4$ | $n = 23$ |
| $T_i$          | 454       | 549       | 425       | 351       | GT= 1779 |
| $\bar{T}_i$    | 75.67     | 78.43     | 70.83     | 87.75     |          |
| parameter      | $\mu_1$   | $\mu_2$   | $\mu_3$   | $\mu_4$   |          |

- (iv) ANOVA: ANalysis OF VAriance
- (v) N-Way ANOVA: studies N factors.
- (vi) experimental unit: (trainee)

## 2 One Way ANOVA: Completely Randomized Experimental Design

| ANOVA Table     |    |         |       |      |         |
|-----------------|----|---------|-------|------|---------|
| Source of error | df | SS      | MS    | F    | p-value |
| Treatments      | 3  | 712.6   | 237.5 | 3.77 |         |
| Error           | 19 | 1,196.6 | 63.0  |      |         |
| Totals          | 22 | 1909.2  |       |      |         |

### Inferences about population means

#### Test.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$H_a$  : Not all means are equal

$$\text{T.S. : } F = \frac{MST}{MSE} = 3.77$$

where F is based on (t-1) and (n-t) df.

RR: Reject  $H_0$  if  $F > F_{\alpha, t-1, n-t}$

i.e. Reject  $H_0$  if  $F > F_{0.05, 3, 19} = 3.13$

Graph:

Decision: Reject  $H_0$

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate a difference in the mean achievement for the four training programs.

**Assumptions.**

1. Sampled populations are normal
2. Independent random samples
3. All  $t$  populations have equal variances

**Computations.**

| ANOVA Table |     |     |               |         |         |
|-------------|-----|-----|---------------|---------|---------|
| S of error  | df  | SS  | MS            | F       | p-value |
| Treatments  | t-1 | SST | MST=SST/(t-1) | MST/MSE |         |
| Error       | n-t | SSE | MSE=SSE/(n-t) |         |         |
| Totals      | n-1 | TSS |               |         |         |

| Training Group |             |             |             |             |      |
|----------------|-------------|-------------|-------------|-------------|------|
|                | 1           | 2           | 3           | 4           |      |
|                | $x_{11}$    | $x_{21}$    | $x_{31}$    | $x_{41}$    |      |
|                | $x_{12}$    | $x_{22}$    | $x_{32}$    | $x_{42}$    |      |
|                | $x_{13}$    | $x_{23}$    | $x_{33}$    | $x_{43}$    |      |
|                | $x_{14}$    | $x_{24}$    | $x_{34}$    | $x_{44}$    |      |
|                | $x_{15}$    | $x_{25}$    | $x_{35}$    |             |      |
|                | $x_{16}$    | $x_{26}$    | $x_{36}$    |             |      |
|                |             | $x_{27}$    |             |             |      |
|                | $n_1$       | $n_2$       | $n_3$       | $n_4$       | $n$  |
| $T_i$          | $T_1$       | $T_2$       | $T_3$       | $T_4$       | $GT$ |
| $\bar{T}_i$    | $\bar{T}_1$ | $\bar{T}_2$ | $\bar{T}_3$ | $\bar{T}_4$ |      |
| parameter      | $\mu_1$     | $\mu_2$     | $\mu_3$     | $\mu_4$     |      |

Notation:

TSS: sum of squares of total deviation.

SST: sum of squares of total deviation *between* treatments.

SSE: sum of squares of total deviation *within* treatments (error).

CM: correction for the mean

GT: Grand Total.

Computational Formulas for TSS, SST and SSE:

$$TSS = \sum_{i=1}^t \sum_{j=1}^{n_i} x_{ij}^2 - CM$$

$$SST = \sum_{i=1}^t \frac{T_i^2}{n_i} - CM$$

$$SSE = TSS - SST$$

Calculations for the training example produce

$$CM = (\sum \sum x_{ij})^2 / n = 1,779^2 / 23 = 137,601.8$$

$$TSS = \sum \sum x_{ij}^2 - CM = 1,909.2$$

$$SST = \sum \frac{T_i^2}{n_i} - CM = 712.6$$

$$SSE = TSS - SST = 1,196.6$$

Thus

ANOVA Table

| Source of error | df | SS      | MS    | F    | p-value |
|-----------------|----|---------|-------|------|---------|
| Treatments      | 3  | 712.6   | 237.5 | 3.77 |         |
| Error           | 19 | 1,196.6 | 63.0  |      |         |
| Totals          | 22 | 1909.2  |       |      |         |

### Confidence Intervals.

Estimate of the common variance:

$$s = \sqrt{s^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-t}}$$

CI for  $\mu_i$ :

$$\bar{T}_i \pm t_{\alpha/2, n-t} \frac{s}{\sqrt{n_i}}$$

CI for  $\mu_i - \mu_j$ :

$$(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2, n-t} \sqrt{s^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

### MINITAB

MTB> aovoneway C1-C4.

**Exercise.** Produce a Minitab output for the above example.

### 3 The Randomized Block Design

Extends paired-difference design to more than two treatments.

A *randomized block design* consists of  $b$  blocks, each containing  $t$  experimental units. The  $t$  treatments are randomly assigned to the units in each block, and each treatment appears once in every block.

**Example.** A consumer preference study involving three different package designs (treatments) was laid out in a randomized block design among four supermarkets (blocks). The data shown in Table 1. below represent the number of units sold for each package design within each supermarket during each of three given weeks.

- (i) Provide a data summary.
- (ii) Do the data present sufficient evidence to indicate a difference in the mean sales for each package design (treatment)?
- (iii) Do the data present sufficient evidence to indicate a difference in the mean sales for the supermarkets?

|    | weeks  |        |        |
|----|--------|--------|--------|
|    | w1     | w2     | w3     |
| s1 | (1) 17 | (3) 23 | (2) 34 |
| s2 | (3) 21 | (1) 15 | (2) 26 |
| s3 | (1) 1  | (2) 23 | (3) 8  |
| s4 | (2) 22 | (1) 6  | (3) 16 |

#### Remarks.

(i) In each supermarket (block) the first entry represents the design (treatment) and the second entry represents the sales per week.

(ii) The three designs are assigned to each supermarket completely at random.

(iii) An alternate design would be to use 12 supermarkets. Each design (treatment) would be randomly assigned to 4 supermarkets. In this case the difference in sales could be due to more than just differences in package design. That is larger supermarkets would be expected to have larger overall sales of the product than smaller supermarkets. The *randomized block design* eliminates the store-to-store variability.

For computational purposes we rearrange the data so that

**Data Summary.** The treatment and block totals are

$t = 3$  treatments;  $b = 4$  blocks

| Treatments |       |       |       |       |
|------------|-------|-------|-------|-------|
|            | t1    | t2    | t3    | $B_i$ |
| s1         | 17    | 34    | 23    | $B_1$ |
| s2         | 15    | 26    | 21    | $B_2$ |
| s3         | 1     | 23    | 8     | $B_3$ |
| s4         | 6     | 22    | 16    | $B_4$ |
| $T_i$      | $T_1$ | $T_2$ | $T_3$ |       |

$$T_1 = 39, T_2 = 105, T_3 = 68$$

$$B_1 = 74, B_2 = 62, B_3 = 32, B_4 = 44$$

Calculations for the training example produce

$$CM = (\sum \sum x_{ij})^2 / n = 3,745.33$$

$$TSS = \sum \sum x_{ij}^2 - CM = 940.67$$

$$SST = \sum \frac{T_i^2}{b} - CM = 547.17$$

$$SSB = \sum \frac{B_i^2}{t} - CM = 348.00$$

$$SSE = TSS - SST - SSB = 45.50$$

MINITAB.(Commands and Printouts)

MTB> Print C1-C3

| ROW | UNITS | TRTS | BLOCKS |
|-----|-------|------|--------|
| 1   | 17    | 1    | 1      |
| 2   | 34    | 2    | 1      |
| 3   | 23    | 3    | 1      |
| 4   | 15    | 1    | 2      |
| 5   | 26    | 2    | 2      |
| 6   | 21    | 3    | 2      |
| 7   | 1     | 1    | 3      |
| 8   | 23    | 2    | 3      |
| 9   | 8     | 3    | 3      |
| 10  | 6     | 1    | 4      |
| 11  | 22    | 2    | 4      |
| 12  | 16    | 3    | 4      |

MTB> ANOVA C1=C2 C3

| ANOVA Table     |    |        |        |       |         |
|-----------------|----|--------|--------|-------|---------|
| Source of error | df | SS     | MS     | F     | p-value |
| Treatments      | 2  | 547.17 | 273.58 | 36.08 | 0.000   |
| Blocks          | 3  | 348.00 | 116.00 | 15.30 | 0.003   |
| Error           | 6  | 45.50  | 7.58   |       |         |
| Totals          | 11 | 940.67 |        |       |         |

**Solution** to (ii)

**Test.**

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$H_a$  : Not all means are equal

$$\text{T.S. : } F = \frac{MST}{MSE} = 36.09$$

where F is based on (t-1) and (n-t-b+1) df.

RR: Reject  $H_0$  if  $F > F_{\alpha, t-1, n-t-b+1}$

i.e. Reject  $H_0$  if  $F > F_{0.05, 2, 6} = 5.14$

Graph:

Decision: Reject  $H_0$

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate a real difference in the mean sales for the three package designs.

Note that  $n - t - b + 1 = (t - 1)(b - 1)$ .

**Solution** to (iii)

**Test.**

$H_0$  : Block means are equal

$H_a$  : Not all block means are equal (i.e. blocking is desirable)

$$\text{T.S.: } F = \frac{MSB}{MSE} = 15.30$$

where F is based on (b-1) and (n-t-b+1) df.

RR: Reject  $H_0$  if  $F > F_{\alpha, b-1, n-t-b+1}$

i.e. Reject  $H_0$  if  $F > F_{0.005, 3, 6} = 12.92$

Graph:

Decision: Reject  $H_0$

Conclusion: At .5% significance level there is sufficient statistical evidence to indicate a real difference in the mean sales for the four supermarkets, that is the data supports our decision to use supermarkets as blocks.

**Assumptions.**

1. Sampled populations are normal
2. Dependent random samples due to blocking
3. All  $t$  populations have equal variances

**Confidence Intervals.**

Estimate of the common variance:

$$s = \sqrt{s^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-t-b+1}}$$

**CI for  $\mu_i - \mu_j$ :**

$$(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2, n-t-b+1} s \sqrt{\frac{2}{b}}$$

**Exercise.** Construct a 90% C.I. for the difference between mean sales from package designs 1 and 2.

# Chapter 6

## Simple Linear Regression and Correlation

### Contents.

1. Introduction: Example
2. A Simple Linear probabilistic model
3. Least squares prediction equation
4. Inferences concerning the slope
5. Estimating  $E(y|x)$  for a given  $x$
6. Predicting  $y$  for a given  $x$
7. Coefficient of correlation
8. Analysis of Variance
9. Computer Printouts

### 1 Introduction

Linear regression is a statistical technique used to predict (forecast) the value of a variable from known related variables.

**Example.** (Ad Sales) Consider the problem of predicting the gross monthly sales volume  $y$  for a corporation that is not subject to substantial seasonal variation in its sales volume. For the predictor variable  $x$  we use the the amount spent by the company on advertising during the month of interest. We wish to determine whether advertising is worthwhile, that is whether advertising is actually related to the firm's sales volume. In addition we wish to use the amount spent on advertising to predict the sales volume. The data in the table below represent a sample of advertising expenditures,  $x$ , and the associated sales

volume,  $y$ , for 10 randomly selected months.

| Month | y(y\$10,000) | x(x\$10,000) |
|-------|--------------|--------------|
| 1     | 101          | 1.2          |
| 2     | 92           | 0.8          |
| 3     | 110          | 1.0          |
| 4     | 120          | 1.3          |
| 5     | 90           | 0.7          |
| 6     | 82           | 0.8          |
| 7     | 93           | 1.0          |
| 8     | 75           | 0.6          |
| 9     | 91           | 0.9          |
| 10    | 105          | 1.1          |

### Definitions.

- (i) Response: dependent variable of interest (sales volume)
- (ii) Independent (predictor) variable ( Ad expenditure)
- (iii) Linear equations (straight line):  $y = a + bx$

Scatter diagram:

Best fit straight line:

Equation of a straight line:

(y-intercept and slope)

## 2 A Simple Linear Probabilistic Model

### Model.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where

x: independent variable (predictor)

y: dependent variable (response)

$\beta_0$  and  $\beta_1$  are unknown parameters.

$\epsilon$  : random error due to other factors not included in the model.

### Assumptions.

1.  $E(\epsilon) := \mu_\epsilon = 0$ .
2.  $Var(\epsilon) := \sigma_\epsilon^2 = \sigma^2$ .

3. The r.v.  $\epsilon$  has a normal distribution with mean 0 and variance  $\sigma^2$ .
4. The random components of any two observed  $y$  values are independent.

### 3 Least Squares Prediction Equation

The least squares prediction equation is sometimes called the *estimated regression equation* or the *prediction equation*.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

This equation is obtained by using the method of *least squares*; that is

$$\min \sum (y - \hat{y})^2$$

#### Computational Formulas.

Objective: Estimate  $\beta_0, \beta_1$  and  $\sigma^2$ .

$$\bar{x} = \sum x/n; \bar{y} = \sum y/n$$

$$SS_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2/n$$

$$SS_{yy} = \sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2/n$$

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - (\sum x)(\sum y)/n$$

$$\hat{\beta}_1 = SS_{xy}/SS_{xx}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

To estimate  $\sigma^2$

$$\begin{aligned} SSE &= SS_{yy} - \hat{\beta}_1 SS_{xy} \\ &= SS_{yy} - (SS_{xy})^2/SS_{xx}. \end{aligned}$$

$$s^2 = \frac{SSE}{n-2}$$

#### Remarks.

(i)  $\hat{\beta}_1$  : is the slope of the estimated regression equation.

(ii)  $s^2$  provides a measure of spread of points  $(x, y)$  around the regression line.

#### Ad Sales example

**Question 1.** Do a scatter diagram. Can you say that  $x$  and  $y$  are linearly related?

**Answer.**

**Question 2.** Use the computational formulas to provide a data summary.

**Answer.**

**Data Summary.**

$$\bar{x} = 0.94; \bar{y} = 95.9$$

$$SS_{xx} = .444$$

$$SS_{xy} = 23.34$$

$$SS_{yy} = 1600.9$$

Optional material

Ad Sales Calculations

| Month            | x        | y                | $x^2$      | xy        | $y^2$      |
|------------------|----------|------------------|------------|-----------|------------|
| 1                | 1.2      | 101              | 1.44       | 121.2     | 10,201     |
| 2                | 0.8      | 92               | 0.64       | 73.6      | 8,464      |
| 3                | 1.0      | 110              | 1.00       | 110.0     | 12,100     |
| 4                | 1.3      | 120              | 1.69       | 156.0     | 14,400     |
| 5                | 0.7      | 90               | 0.49       | 63.0      | 8,100      |
| 6                | 0.8      | 82               | 0.64       | 65.6      | 6,724      |
| 7                | 1.0      | 93               | 1.00       | 93.0      | 8,649      |
| 8                | 0.6      | 75               | 0.36       | 45.0      | 5,625      |
| 9                | 0.9      | 91               | 0.81       | 81.9      | 8,281      |
| 10               | 1.1      | 105              | 1.21       | 115.5     | 11,025     |
| Sum              | $\sum x$ | $\sum y$         | $\sum x^2$ | $\sum xy$ | $\sum y^2$ |
|                  | 9.4      | 959              | 9.28       | 924.8     | 93,569     |
| $\bar{x} = 0.94$ |          | $\bar{y} = 95.9$ |            |           |            |

$$\bar{x} = \sum x/n = 0.94; \bar{y} = \sum y/n = 95.9$$

$$SS_{xx} = \sum x^2 - (\sum x)^2/n = 9.28 - \frac{(9.4)^2}{10} = .444$$

$$SS_{xy} = \sum xy - (\sum x)(\sum y)/n = 924.8 - \frac{(9.4)(959)}{10} = 23.34$$

$$SS_{yy} = \sum y^2 - (\sum y)^2/n = 93,569 - \frac{(959)^2}{10} = 1600.9$$

**Question 3.** Estimate the parameters  $\beta_0$ , and  $\beta_1$ .

**Answer.**

$$\hat{\beta}_1 = SS_{xy}/SS_{xx} = \frac{23.34}{.444} = 52.5676 \simeq 52.57$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 95.9 - (52.5676)(.94) \simeq 46.49.$$

**Question 4.** Estimate  $\sigma^2$ .

**Answer.**

$$\begin{aligned} SSE &= SS_{yy} - \hat{\beta}_1 SS_{xy} \\ &= 1,600.9 - (52.5676)(23.34) = 373.97 . \end{aligned}$$

Therefore

$$s^2 = \frac{SSE}{n-2} = \frac{373.97}{8} = 46.75$$

**Question 5.** Find the least squares line for the data.

**Answer.**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 46.49 + 52.57x$$

**Remark.** This equation is also called the *estimated regression equation* or *prediction line*.

**Question 6.** Predict sales volume,  $y$ , for a given expenditure level of \$10,000 (i.e.  $x = 1.0$ ).

**Answer.**

$$\hat{y} = 46.49 + 52.57x = 46.49 + (52.57)(1.0) = 99.06.$$

So sales volume is \$990,600.

**Question 7.** Predict the mean sales volume  $E(y|x)$  for a given expenditure level of \$10,000,  $x = 1.0$ .

**Answer.**

$$E(y|x) = 46.49 + 52.57x = 46.49 + (52.57)(1.0) = 99.06$$

so the mean sales volume is \$990,600.

**Remark.** In **Question 6** and **Question 7** we obtained the same estimate, the bound on the error of estimation will, however, be different.

## 4 Inferences Concerning the Slope

Parameter of interest:  $\beta_1$

Point estimator:  $\hat{\beta}_1$

Estimator mean:  $\mu_{\hat{\beta}_1} = \beta_1$

Estimator standard error:  $\sigma_{\hat{\beta}_1} = \sigma/\sqrt{SS_{xx}}$

**Test.**

$H_0 : \beta_1 = \beta_{10}$  (no linear relationship)

$H_a : \beta_1 \neq \beta_{10}$  (there is linear relationship)

T.S. :

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s/\sqrt{SS_{xx}}}$$

RR:

Reject  $H_0$  if  $t > t_{\alpha/2, n-2}$  or  $t < -t_{\alpha/2, n-2}$

Graph:

Decision:

Conclusion:

**Question 8.** Determine whether there is evidence to indicate a linear relationship between advertising expenditure,  $x$ , and sales volume,  $y$ .

**Answer.**

**Test.**

$H_0 : \beta_1 = 0$  (no linear relationship)

$H_a : \beta_1 \neq 0$  (there is linear relationship)

T.S. :

$$t = \frac{\hat{\beta}_1 - 0}{s/\sqrt{SS_{xx}}} = \frac{52.57 - 0}{6.84/\sqrt{.444}} = 5.12$$

RR: ( critical value:  $t_{.025, 8} = 2.306$ )

Reject  $H_0$  if  $t > 2.306$  or  $t < -2.306$

Graph:

Decision: Reject  $H_0$

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate a linear relation ship between advertising expenditure,  $x$ , and sales volume,  $y$ .

**Confidence interval for  $\beta_1$ :**

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{s}{\sqrt{SS_{xx}}}$$

**Question 9.** Find a 95% confidence interval for  $\beta_1$ .

**Answer.**

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{s}{\sqrt{SS_{xx}}}$$

$$52.57 \pm 2.306 \frac{6.84}{\sqrt{.444}}$$

$$52.57 \pm 23.57 = (28.90, 76.24)$$

## 5 Estimating $E(y|x)$ For a Given $x$

The *confidence interval (CI)* for the expected (mean) value of  $y$  given  $x = x_p$  is given by

$$\hat{y} \pm t_{\alpha/2, n-2} \sqrt{s^2 \left[ \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right]}$$

## 6 Predicting $y$ for a Given $x$

The *prediction interval (PI)* for a particular value of  $y$  given  $x = x_p$  is given by

$$\hat{y} \pm t_{\alpha/2, n-2} \sqrt{s^2 \left[ 1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right]}$$

## 7 Coefficient of Correlation

In a previous section we tested for a linear relationship between  $x$  and  $y$ .

Now we examine how strong a linear relationship between  $x$  and  $y$  is.

We call this measure *coefficient of correlation* between  $y$  and  $x$ .

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

**Remarks.**

- (i)  $-1 \leq r \leq 1$ .

- (ii) The population coefficient of correlation is  $\rho$ .
- (iii)  $r > 0$  indicates a positive correlation ( $\hat{\beta}_1 > 0$ )
- (iv)  $r < 0$  indicates a negative correlation ( $\hat{\beta}_1 < 0$ )
- (v)  $r = 0$  indicates no correlation ( $\hat{\beta}_1 = 0$ )

**Question 10.** Find the coefficient of correlation,  $r$ .

**Answer.**

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{23.34}{\sqrt{0.444(1,600.9)}} = 0.88$$

### Coefficient of determination

Algebraic manipulations show that

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

**Question 11.** By what percentage is the sum of squares of deviations of  $y$  about the mean ( $SS_{yy}$ ) is reduced by using  $\hat{y}$  rather than  $\bar{y}$  as a predictor of  $y$ ?

**Answer.**

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 0.88^2 = 0.77$$

$r^2$  = is called the *coefficient of determination*

## 8 Analysis of Variance

Notation:

$TSS := SS_{yy} = \sum(y - \bar{y})^2$  (Total SS of deviations).

$SSR = \sum(\hat{y} - \bar{y})^2$  (SS of deviations due to regression or explained deviations)

$SSE = \sum(y - \hat{y})^2$  (SS of deviations for the error or unexplained deviations)

$$TSS = SSR + SSE$$

**Question 12.** Give the ANOVA table for the AD sales example.

**Answer.**

**Question 13.** Use ANOVA table to test for a significant linear relationship between sales and advertising expenditure.

| Source | df | SS        | MS        | F     | p-value |
|--------|----|-----------|-----------|-------|---------|
| Reg.   | 1  | 1,226.927 | 1,226.927 | 26.25 | 0.0001  |
| Error  | 8  | 373.973   | 46.747    |       |         |
| Totals | 9  | 1,600.900 |           |       |         |

| Source | df  | SS  | MS            | F       | p-value |
|--------|-----|-----|---------------|---------|---------|
| Reg.   | 1   | SSR | MSR=SSR/(1)   | MSR/MSE |         |
| Error  | n-2 | SSE | MSE=SSE/(n-2) |         |         |
| Totals | n-1 | TSS |               |         |         |

**Answer.**

**Test.**

$H_0 : \beta_1 = 0$  (no linear relationship)

$H_a : \beta_1 \neq 0$  (there is linear relationship)

T.S.:  $F = \frac{MSR}{MSE} = 26.25$

RR: (critical value:  $F_{.005,1,8} = 14.69$ )

Reject  $H_0$  if  $F > 14.69$

(OR: Reject  $H_0$  if  $\alpha > \text{p-value}$ )

Graph:

Decision: Reject  $H_0$

Conclusion: At 0.5% significance level there is sufficient statistical evidence to indicate a linear relationship between advertising expenditure,  $x$ , and sales volume,  $y$ .

## 9 Computer Printouts for Regression Analysis

Store  $y$  in C1 and  $x$  in C2.

MTB> Plot C1 C2. : Gives a scatter diagram.

MTB> Regress C1 1 C2.

Computer output for Ad sales example:

More generally we obtain:

The regression equation is

$$y = 46.5 + 52.6x$$

| Predictor | Coef   | Stdev | t-ratio | P     |
|-----------|--------|-------|---------|-------|
| Constant  | 46.486 | 9.885 | 4.70    | 0.000 |
| x         | 52.57  | 10.26 | 5.12    | 0.000 |

$$s = 6.837$$

$$R\text{-sq} = 76.6\%$$

$$R\text{-sq(adj)} = 73.7\%$$

Analysis of Variance

| Source | df | SS        | MS        | F     | p-value |
|--------|----|-----------|-----------|-------|---------|
| Reg.   | 1  | 1,226.927 | 1,226.927 | 26.25 | 0.000   |
| Error  | 8  | 373.973   | 46.747    |       |         |
| Totals | 9  | 1,600.900 |           |       |         |

### Review Exercises: Linear Regression

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. Given the following data set

|   |    |    |   |   |   |
|---|----|----|---|---|---|
| x | -3 | -1 | 1 | 1 | 2 |
| y | 6  | 4  | 3 | 1 | 1 |

- Plot the scatter diagram, and indicate whether  $x$  and  $y$  appear linearly related.
- Show that  $\sum x = 0$ ;  $\sum y = 15$ ;  $\sum x^2 = 16$ ;  $\sum y^2 = 63$ ;  $SS_{xx} = 16$ ;  $SS_{yy} = 18$ ; and  $SS_{xy} = -16$ .
- Find the regression equation for the data. (Answer:  $\hat{y} = 3 - x$ )
- Plot the regression equation on the same graph as (i); Does the line appear to provide a good fit for the data points?
- Compute  $SSE$  and  $s^2$ . (Answer:  $s^2 = 2/3$ )
- Estimate the expected value of  $y$  when  $x = -1$
- Find the correlation coefficient  $r$  and find  $r^2$ . (Answer:  $r = -.943$ ,  $r^2 = .889$ )

The regression equation is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

| Predictor        | Coef            | Stdev                    | t-ratio   | P       |
|------------------|-----------------|--------------------------|-----------|---------|
| Constant         | $\hat{\beta}_0$ | $\sigma_{\hat{\beta}_0}$ | TS: t     | p-value |
| x                | $\hat{\beta}_1$ | $\sigma_{\hat{\beta}_1}$ | TS: t     | p-value |
| $s = \sqrt{MSE}$ |                 | $R - sq = r^2$           | R-sq(adj) |         |

Analysis of Variance

| Source | df  | SS  | MS            | F       | p-value |
|--------|-----|-----|---------------|---------|---------|
| Reg.   | 1   | SSR | MSR=SSR/(1)   | MSR/MSE |         |
| Error  | n-2 | SSE | MSE=SSE/(n-2) |         |         |
| Totals | n-1 | TSS |               |         |         |

2. A study of middle to upper-level managers is undertaken to investigate the relationship between salary level,  $Y$ , and years of work experience,  $X$ . A random sample of 20 managers is chosen with the following results (in thousands of dollars):  $\sum x_i = 235$ ;  $\sum y_i = 763.8$ ;  $SS_{xx} = 485.75$ ;  $SS_{yy} = 2,236.1$ ; and  $SS_{xy} = 886.85$ . It is further assumed that the relationship is linear.

(i) Find  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and the estimated regression equation.

(Answer:  $\hat{y} = 16.73 + 1.826x$ )

(ii) Find the correlation coefficient,  $r$ . (Answer:  $r = .85$ )

(iii) Find  $r^2$  and interpret it value.

3. The *Regress* Minitab's command has been applied to data on family income,  $X$ , and last year's energy consumption,  $Y$ , from a random sample of 25 families. The income data are in thousands of dollars and the energy consumption are in millions of BTU. A portion of a linear regression computer printout is shown below.

|           |            |                 |         |       |
|-----------|------------|-----------------|---------|-------|
| Predictor | Coef       | stdev           | t-ratio | P     |
| Constant  | 82.036     | 2.054           | 39.94   | 0.000 |
| X         | 0.93051    | 0.05727         | 16.25   | 0.000 |
| s=        | R-sq=92.0% | R-sq(adj)=91.6% |         |       |

Analysis of Variance

| Source     | DF | SS   | MS     | F      | P     |
|------------|----|------|--------|--------|-------|
| Regression |    |      | 7626.6 | 264.02 | 0.000 |
| Error      | 23 |      |        |        |       |
| Total      |    | 8291 |        |        |       |

(i) Complete all missing entries in the table.

(ii) Find  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and the estimated regression equation.

(iii) Do the data present sufficient evidence to indicate that  $Y$  and  $X$  are linearly related? Test by using  $\alpha = 0.01$ .

(iv) Determine a point estimate for last year's mean energy consumption of all families with an annual income of \$40,000.

4. Answer by True or False . (Circle your choice).

T F (i) The correlation coefficient  $r$  shows the degree of association between  $x$  and  $y$ .

T F (ii) The coefficient of determination  $r^2$  shows the percentage change in  $y$  resulting from one-unit change in  $x$ .

T F (iii) The last step in a simple regression analysis is drawing a scatter diagram.

T F (iv)  $r = 1$  implies no linear correlation between  $x$  and  $y$ .

T F (v) We always estimate the value of a parameter and predict the value of a random variable.

T F (vi) If  $\beta_1 = 1$ , we always predict the same value of  $y$  regardless of the value of  $x$ .

T F (vii) It is necessary to assume that the response  $y$  of a probability model has a normal distribution if we are to estimate the parameters  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$ .

# Chapter 7

## Multiple Linear Regression

### Contents.

1. Introduction: Example
2. Multiple Linear Model
3. Analysis of Variance
4. Computer Printouts

### 1 Introduction: Example

Multiple linear regression is a statistical technique used predict (forecast) the value of a variable from multiple known related variables.

### 2 A Multiple Linear Model

#### Model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where

$x_i$  : independent variables (predictors)

$y$ : dependent variable (response)

$\beta_i$  : unknown parameters.

$\epsilon$  : random error due to other factors not included in the model.

#### Assumptions.

1.  $E(\epsilon) := \mu_\epsilon = 0$ .
2.  $Var(\epsilon) := \sigma_\epsilon^2 = \sigma^2$ .
3.  $\epsilon$  has a normal distribution with mean 0 and variance  $\sigma^2$ .

4. The random components of any two observed  $y$  values are independent.

### 3 Least Squares Prediction Equation

Estimated Regression Equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

This equation is obtained by using the method of *least squares*

| Multiple Regression Data |       |          |          |          |
|--------------------------|-------|----------|----------|----------|
| Obser.                   | $y$   | $x_1$    | $x_2$    | $x_3$    |
| 1                        | $y_1$ | $x_{11}$ | $x_{21}$ | $x_{31}$ |
| 2                        | $y_2$ | $x_{12}$ | $x_{22}$ | $x_{32}$ |
| ...                      | ...   | ...      | ...      | ...      |
| $n$                      | $y_n$ | $x_{1n}$ | $x_{2n}$ | $x_{3n}$ |

#### Minitab Printout

The regression equation is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

| Predictor | Coef            | Stdev                    | t-ratio | P       |
|-----------|-----------------|--------------------------|---------|---------|
| Constant  | $\hat{\beta}_0$ | $\sigma_{\hat{\beta}_0}$ | TS: t   | p-value |
| $x_1$     | $\hat{\beta}_1$ | $\sigma_{\hat{\beta}_1}$ | TS: t   | p-value |
| $x_2$     | $\hat{\beta}_2$ | $\sigma_{\hat{\beta}_2}$ | TS: t   | p-value |
| $x_3$     | $\hat{\beta}_3$ | $\sigma_{\hat{\beta}_3}$ | TS: t   | p-value |

$$s = \sqrt{MSE} \qquad R^2 = r^2 \qquad R^2(\text{adj})$$

#### Analysis of Variance

| Source | df      | SS  | MS            | F       | p-value |
|--------|---------|-----|---------------|---------|---------|
| Reg.   | 3       | SSR | MSR=SSR/(3)   | MSR/MSE |         |
| Error  | $n - 4$ | SSE | MSE=SSE/(n-4) |         |         |
| Totals | $n - 1$ | TSS |               |         |         |

| Source | df | SS         |
|--------|----|------------|
| $x_1$  | 1  | $SSx_1x_1$ |
| $x_2$  | 1  | $SSx_2x_2$ |
| $x_3$  | 1  | $SSx_3x_3$ |

Unusual observations (ignore)

## MINITAB.

Use REGRESS command to regress  $y$  stored in C1 on the 3 predictor variables stored in C2 – C4.

```
MTB> Regress C1 3 C2-C4;  
SUBC> Predict x1 x2 x3.
```

The subcommand PREDICT in Minitab, followed by fixed values of  $x_1, x_2$ , and  $x_3$  calculates the estimated value of  $\hat{y}$  (Fit), its estimated standard error (Stdev.Fit), a 95% CI for  $E(y)$ , and a 95% PI for  $y$ .

**Example.** A county assessor wishes to develop a model to relate the market value,  $y$ , of single-family residences in a community to the variables:

$x_1$  : living area in thousands of square feet;  
 $x_2$  : number of floors;  
 $x_3$  : number of bedrooms;  
 $x_4$  : number of baths.

Observations were recorded for 29 randomly selected single-family homes from residences recently sold at fair market value. The resulting prediction equation will then be used for assessing the values of single family residences in the county to establish the amount each homeowner owes in property taxes.

A Minitab printout is given below:

```
MTB> Regress C1 4 C2-C5;  
SUBC> Predict 1.0 1 3 2;  
SUBC> Predict 1.4 2 3 2.5.
```

The regression equation is

$$y = -16.6 + 7.84x_1 - 34.4x_2 - 7.99x_3 + 54.9x_4$$

| Predictor | Coef.  | Stdev | t-ratio | P     |
|-----------|--------|-------|---------|-------|
| Constant  | -16.58 | 18.88 | -0.88   | 0.389 |
| $x_1$     | 7.839  | 1.234 | 6.35    | 0.000 |
| $x_2$     | -34.39 | 11.15 | -3.09   | 0.005 |
| $x_3$     | -7.990 | 8.249 | -0.97   | 0.342 |
| $x_4$     | 54.93  | 13.52 | 4.06    | 0.000 |

$s = 16.58$

$R^2 = 88.2\%$

$R^2(adj) = 86.2\%$

Analysis of Variance

| Source | df | SS    | MS    | F     | p-value |
|--------|----|-------|-------|-------|---------|
| Reg.   | 4  | 49359 | 12340 | 44.88 | 0.000   |
| Error  | 24 | 6599  | 275   |       |         |
| Totals | 28 | 55958 |       |       |         |

| Source | df | SS    |
|--------|----|-------|
| $x_1$  | 1  | 44444 |
| $x_2$  | 1  | 59    |
| $x_3$  | 1  | 321   |
| $x_4$  | 1  | 4536  |

| Fit    | Stdev.Fit | 95% <i>C.I.</i>  | 95% <i>P.I.</i>  |
|--------|-----------|------------------|------------------|
| 113.32 | 5.80      | (101.34, 125.30) | (77.05, 149.59)  |
| 137.75 | 5.48      | (126.44, 149.07) | (101.70, 173.81) |

Q1. What is the prediction equation ?

The regression equation is

$$y = -16.6 + 7.84x_1 - 34.4x_2 - 7.99x_3 + 54.9x_4$$

Q2. What type of model has been chosen to fit the data?

Multiple linear regression model.

Q3. Do the data provide sufficient evidence to indicate that the model contributes information for the prediction of  $y$ ? Test using  $\alpha = 0.05$ .

**Test:**

$H_0$  : model not useful

$H_a$  : model is useful

T.S. : p-value=0.000

DR. Reject  $H_0$  if  $\alpha > p - value$

Graph:

Decision: Reject  $H_0$

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate that the model contributes information for the prediction of  $y$ .

Q4. Give a 95% CI for  $E(y)$  and PI for  $y$  when  $x_1 = 10$ ,  $x_2 = 1$ ,  $x_3 = 3$ , and  $x_4 = 2$ .

CI: (101.34, 125.30)

PI: (77.05, 149.59)

**Non-Linear Models**

**Example.**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \hat{\beta}_3x_1^2x_2$$