



RASHTREEYA SIKSHANA SAMITHI TRUST
RV INSTITUTE OF MANAGEMENT
CA 17, 26 Main, 36th Cross, 4th T Block, Jayanagar
Bengaluru, Karnataka 560 041



Course Docket

21MBA212- APPLICATION OF STATISTICS IN BUSINESS

Term – I Semester

Batch – 2021 - 2023

January 2022- May 2022

Course Facilitators

Dr. Suresh N

Dr. Santhosh M

Dr. Jahnvi M



Director

R V INSTITUTE OF MANAGEMENT BANGALORE

COURSE OUTLINE

Programme	Master of Business Administration
Batch	2021- 2023
Semester	I
Course Title	APPLICATION OF STATISTICS IN BUSINESS
Course Code	21MBA212
Credits	3
Sessions	36
Course Facilitator	Dr. Santhosh M Dr. Suresh N Dr. Jahnvi M

PART A

INTRODUCTION

The Business Statistics course is designed to elevate students' awareness of data in everyday life and prepare them for a career in today's age of information. The course facilitates to develop statistical literacy skills in students in order to comprehend and practice statistical ideas to solve the business problems. The course aims to promote the practice of the scientific methods amongst the students' community and the ability to identify questions, collect evidence (data), discover and apply tools to interpret the data, and communicate and exchange results. At the end of this course, students will be able to find ways to move beyond the-what of statistics to the how and why of statistics.

COURSE OUTCOMES

Post completion of the course student should able to:

- CO1** Understand and apply the central tendency, Dispersion and Skewness for data Interpretation
- CO2** Apply correlation and regression tools for data analysis.
- CO3** Develop appropriate probability models for decision making.
- CO4** Test the hypothesis using appropriate statistical methods
- CO5** Construct decision tree on the basis of probability distribution

PROGRAM OUTCOMES

- PO1:** Apply knowledge of management theories and practices to solve business problems.
- PO2:** Foster Analytical and critical thinking abilities for data-based decision making
- PO3:** Ability to develop Value based Leadership
- PO4:** Ability to understand, analyze and communicate global, economic, societal, cultural, legal and ethical aspects of business

PO5: Ability to lead themselves and others in the achievement of organizational goals, contributing effectively to a team environment

PO6: Ability to identify business opportunities, frame innovative solutions and launch new business ventures or be an entrepreneur

PO7: Ability to deal with contemporary issues using multi-disciplinary approach with the help of advanced Management and IT tools and techniques

PO8: Ability to apply domain specific knowledge and skills to build competencies in their respective functional area

PO9: Ability to engage in research and development work with cognitive flexibility to create new knowledge and be a lifelong learner

PO10: Ability to understand social responsibility and contribute to the community for inclusive growth and sustainable development of society through ethical behavior

PO11: Ability to function effectively as individuals and in teams through effective communication and Negotiation skills

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11
CO1	1	3	1	-	--	2	3	-	3	-	--
CO2	-	3	1	-	1		3	3	3	1	
CO3	2	3	1	-	-		3	1	3	1	-
CO4		3	-	-	-	3	3	1	3	1	1
CO5	1	3	2	-	-	1	3	1	3	-	-
LEVEL	3-Substantial			2-Moderate			1-Slight		- No Co-relation		

KEY CONCEPTS

Module 1: Measures of Central tendency, Dispersion, Skewness and Kurtosis

- Introduction to basic measures of central tendency;
- Missing value cases in basic measures;
- Problems on missing frequency;
- Empirical relationships between basic measures;
- Application of central tendency in functional areas of business-

- Comparison between various measures of dispersion;
- Standard deviation; Coefficient of variance- Partition values
- Karl Pearson co efficient of Skewness
- Bowley's co efficient of Skewness
- Measures of kurtosis

Module 2: Correlation and Regression Analysis

- Introduction and significance of correlation and Regression
- Methods of correlation Analysis
- Scatter diagram
- Karl Pearson's coefficient of Correlation for Uni-variate and Bi-variate series
- Spearman's Rank Correlation
- Concurrent deviation method
- Simple regression analysis; Least square method

Module 3: Probability and Theoretical Distribution

- Concept and Definition of probability and theories of probability
- Relevance of Permutations and Combinations to Probability
- Rules of Probability; Bayes' theorem & its applications
- Theoretical Probability Distributions; Binomial, Poisson and Normal distribution

Module 4: Testing of Hypotheses

- Hypothesis Testing; Concept; Formulation of Hypotheses; Errors: Type I and II error;
- Parametric tests; z-test; t-test; f-test; Chi-Square test; Analysis of Variance (ANOVA) - one and two way
- Non-parametric tests - Sign test; Wilcoxon test; Mann-Whitney U test; Median test Run test; Kolmogorov –Smirnov one sample test

Module 5: Business Statistics in Decision making

- Time series analysis: Concepts and components
- Identification of Trend; Methods of measuring: Semi averages; Moving averages; Method of Least squares; Non-linear trend
- Decision Theory concepts-; Decision under certainty; Decision making under risk (EMV criteria); Decision making under uncertainty; Decision tree

TEACHING PEDAGOGY

- Class room discussions
- SPSS lab sessions
- Project based learning
- Workshop from practitioners
- Case based teaching

TEXT BOOKS AND REFERENCE MATERIALS

ESSENTIAL READINGS

- James R. Evans, “Business Analytics – Methods, Models and Decisions”, Prentice Hall
- N Srivastava, Shailaja Rego, “Statistics for Management”, Tata McGraw Hill
- SP Gupta, “Statistical Methods”, Sultan Chand & Sons
- Glynn Davis and Branko Pecar, “Business Statistics using excel”, Oxford University Press
- J K Sharma, “Fundamentals of Business Statistics”, Vikas Publication

REFERENCES

- Keller/Arora, “BSTAT: A South-Asian Perspective”, Cengage Learning
- S C Gupta, “Fundamentals of Statistics”, Himalaya Publications
- N D Vohra, “Business Statistics”, Tata McGraw Hill
- Levin & Rubin, “Statistics for Management”, Prentice-Hall

- Richard I. Levin, David S. Rubin, Masood H. Siddiqui, Sanjay Rastogi, “Statistics for Management”, Pearson India
- Amir D Aczel, Jayavel Sounderpandian, Palaniswamy Saravanan, Rohit Joshi, “Complete Business Statistics”, McGraw Hill Education
- Statistics for Managers Using Microsoft Excel, 9th Edition, David M. Levine, Baruch College, Zicklin School of Business, City University of New York, Pearson

CASES AND ARTICLES

- ANOVA
- Decision tree
- Descriptive statistics
- Hypothesis testing
- Paired t test
- Moving average
- Measures of central tendency
- Correlation and regression analysis
- Statistics for managers using MS Excel
- Test of Independent sample
- Time series forecasting

SUPPORTING READINGS

- <https://www.coursera.org/learn/basic-statistics?specialization=social-science>
- <https://www.edx.org/learn/statistics>
- Pearson e – library; <https://elibrary.in.pearson.com/bookshelfDashboard>
- EBSCO: <https://www.ebsco.com/search?search=supplychainmodel>
- Jgate: <https://jgateplus.com/home/resources/>
- www.capitaline.com

EVALUATION PLAN

SL NO	PARTICULARS	MARKS
1	SEMESTER END EXAMINATION	50
2	CIA	50
	1. Attendance and class participation	05
	2. Mid –term Test	20
	3. Quiz (5 quiz, 1 from each module)	05
	4. Assignment (Total 5 assignments from 5 modules – Numerical problems – at least 5 questions for each assignment)	10
	5. SPSS Lab component	10
	Total	100

SPSS lab component: (10 marks)

1. Lab journal, participation and execution of all exercises – 7.5 marks
2. Final assignment on SPSS data set – 2.5 Marks

SPSS Exercise:

1. Introduction to SPSS
2. Descriptive statistics
3. Correlation and regression
4. Testing of Hypothesis
5. Z test
6. t test
7. chi-square test
8. ANOVA

COURSE FACILITATORS

Dr. Santhosh M

Department of Marketing

Email: santhoshm.rvim@rvei.edu.in Mobile : 9739945333

Dr. Suresh N

Department of General Management

Email: suresh.rvim@rvei.edu.in Mobile: 9449637363

Dr. Jahnvi M

Department of Finance

E-mail: jahnvim.rvim@rvei.edu.in Mobile : 9353939778

PART- B

SESSION PLAN

I SEMESTER A, B, C SECTION

SESSION NO.	COVERAGE OF THE KEY CONCEPTS	TEACHING PEDAGOGY	READING MATERIAL TO BE REFERRED
1	Module -1: Introduction to basic measures of central tendency	<ul style="list-style-type: none">• Lecture• Class room discussion	Study Material Recommended book no. 2 and 4
2	Mean – Missing frequency	<ul style="list-style-type: none">• Lecture• Class room discussion	Study Material Recommended book no. 2 and 4
3	Median and problems on missing frequency	<ul style="list-style-type: none">• Lecture• Class room discussion	Study Material Recommended book no. 2 and 4
4	Mode and partition values	<ul style="list-style-type: none">• Lecture• Class room discussion• Articles on descriptive Statistics	Study Material Recommended book no. 2 and 4
5	Measures of Dispersion	<ul style="list-style-type: none">• Lecture• Class room discussion	Study Material Recommended book no. 2 and 4
6	Standard deviation and CV	<ul style="list-style-type: none">• Lecture• Class room discussion	Study Material Recommended book no. 2 and 4

7	Co efficient of Skewness	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
8	Measures of Kurtosis	<ul style="list-style-type: none"> • Lecture • Class room discussion • SPSS Lab session on Module -1 • Quiz 	Study Material Recommended book no. 2 and 4
9	Module 2: Introduction and significance of correlation	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
10	Methods of correlation Analysis	<ul style="list-style-type: none"> • Lecture • Class room discussion • Articles on Correlation coefficient 	Study Material Recommended book no. 2 and 4
11	Scatter diagram; Karl Pearson's coefficient of Correlation for Uni-variate and Bi-variate series	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
12	Spearman's Rank Correlation, Concurrent deviation method	<ul style="list-style-type: none"> • Lecture • Class room discussion • SPSS lab session on Correlation coefficient 	Study Material Recommended book no. 2 and 4
13	Simple regression analysis	<ul style="list-style-type: none"> • Lecture • Class room discussion • Articles on regression models 	Study Material Recommended book no. 2 and 4

14	Least square method	<ul style="list-style-type: none"> • Lecture • Class room discussion • Quiz 	Study Material Recommended book no. 2 and 4
15	Module 3 : Concept and Definition of probability and theories of probability	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
16	Relevance of Permutations and Combinations to Probability -Rules of Probability; Bayes' theorem & its applications	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
17	Theoretical Probability Distributions	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
18	Binomial Distribution	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
19	Poisson Distribution	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
20	Normal Distribution	<ul style="list-style-type: none"> • Lecture • Class room discussion • Articles on theory of probability • Quiz 	Study Material Recommended book no. 2 and 4
21	Module 4: - Hypothesis Testing; Concept; Formulation of Hypotheses, Errors: Type I and II error	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
22	Z test	<ul style="list-style-type: none"> • Lecture • Class room 	Study Material Recommended book

		<ul style="list-style-type: none"> • Discussion • SPSS lab session 	no. 2 and 4
23	t test	<ul style="list-style-type: none"> • Lecture • Class room discussion • SPSS lab session 	Study Material Recommended book no. 2 and 4
24	Chi-square test	<ul style="list-style-type: none"> • Lecture • Class room discussion • SPSS lab session 	Study Material Recommended book no. 2 and 4
25	ANOVA One Way	<ul style="list-style-type: none"> • Lecture • Class room discussion • SPSS lab session 	Study Material Recommended book no. 2 and 4
26	ANOVA Two Way	<ul style="list-style-type: none"> • Lecture • Class room discussion • SPSS lab session 	Study Material Recommended book no. 2 and 4
27	Sign test	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
28	Wilcoxon test	<ul style="list-style-type: none"> • Lecture • Class room discussion • Articles on Non parametric tests 	Study Material Recommended book no. 2 and 4
29	Mann-Whitney U test; Median test Run test	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
30	Kolmogorov –Smirnov one sample test	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4

		<ul style="list-style-type: none"> • Quiz 	
31	Module 5: Time series analysis: Concepts and components	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
32	Identification of Trend; Methods of measuring: Semi averages; Moving averages;	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
33	Method of Least squares; Non-linear trend	<ul style="list-style-type: none"> • Lecture • Class room discussion • Articles on time series data 	Study Material Recommended book no. 2 and 4
34	Decision Theory concepts-; Decision under certainty; Decision making under risk	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
35	Decision making under uncertainty	<ul style="list-style-type: none"> • Lecture • Class room discussion 	Study Material Recommended book no. 2 and 4
36	Decision tree	<ul style="list-style-type: none"> • Lecture • Class room discussion • Quiz 	Study Material Recommended book no. 2 and 4

Course Title	APPLICATION OF STATISTICS IN BUSINESS
Term/Semester	1
Course ID	21MBA212
Credits	3

Introduction:

The Business Statistics course is designed to elevate students' awareness of data in everyday life and prepare them for a career in today's age of information. The course facilitates to develop statistical literacy skills in students in order to comprehend and practice statistical ideas to solve the business problems. The course aims to promote the practice of the scientific methods amongst the students' community and the ability to identify questions, collect evidence (data), discover and apply tools to interpret the data, and communicate and exchange results. At the end of this course, students will be able to find ways to move beyond the-what of statistics to the how and why of statistics.

Course Outcomes (COs):

Having successfully completed this course student will be able to:

- CO1** Understand and apply the central tendency, Dispersion and Skewness for data Interpretation
- CO2** Apply correlation and regression tools for data analysis.
- CO3** Develop appropriate probability models for decision making.
- CO4** Test the hypothesis using appropriate statistical methods
- CO5** Construct decision tree on the basis of probability distribution

Course content and Structure: (36 hours)

Module 1: Measures of Central tendency, Dispersion, Skewness and Kurtosis

8 Hours

- Introduction to basic measures of central tendency; Missing value cases in basic measures; Problems on missing frequency; Empirical relationships between basic measures; Application of central tendency in functional areas of business-
- Comparison between various measures of dispersion; Standard deviation; Coefficient of variance- Partition values
- Karl Pearson co efficient of Skewness; Bowley's co efficient of Skewness
- Measures of kurtosis

Module 2: Correlation and Regression Analysis (6 Hours)

- Introduction and significance of correlation and Regression
- Methods of correlation Analysis; Scatter diagram; Karl Pearson's coefficient of Correlation for Uni-variate and Bi-variate series; Spearman's Rank Correlation, Concurrent deviation method
- Simple regression analysis; Least square method

Module 3: Probability and Theoretical Distribution (6 Hours)

- Concept and Definition of probability and theories of probability
- Relevance of Permutations and Combinations to Probability
- Rules of Probability; Bayes' theorem & its applications
- Theoretical Probability Distributions; Binomial, Poisson and Normal distribution

Module 4: Testing of Hypotheses

(10 Hours)

- Hypothesis Testing; Concept; Formulation of Hypotheses; Errors: Type I and II error;
- Parametric tests; z-test; t-test; f-test; Chi-Square test; Analysis of Variance (ANOVA) -one and two way
- Non-parametric tests - Sign test; Wilcoxon test; Mann-Whitney U test; Median test Run test; Kolmogorov –Smirnov one sample test

Module 5: Business Statistics in Decision making

(6 Hours)

- Time series analysis: Concepts and components
- Identification of Trend; Methods of measuring: Semi averages; Moving averages; Method of Least squares; Non-linear trend
- Decision Theory concepts-; Decision under certainty; Decision making under risk (EMV criteria); Decision making under uncertainty; Decision tree

Pedagogy:

- Class room discussions
- SPSS lab sessions
- Project based learning
- Workshop from practitioners
- Case based teaching

Teaching Learning Resources:

Recommended Books

- James R. Evans, “Business Analytics – Methods, Models and Decisions”, Prentice Hall T N Srivastava, Shailaja Rego, “Statistics for Management”, Tata McGraw Hill
- SP Gupta, “Statistical Methods”, Sultan Chand & Sons
- Glynn Davis and Branko Pecar, “Business Statistics using excel”, Oxford University Press
- J K Sharma, “Fundamentals of Business Statistics”, Vikas Publication

Reference Books

- Keller/Arora, “BSTAT: A South-Asian Perspective”, Cengage Learning
- S C Gupta, “Fundamentals of Statistics”, Himalaya Publications
- N D Vohra, “Business Statistics”, Tata McGraw Hill
- Levin & Rubin, “Statistics for Management”, Prentice-Hall
- Richard I. Levin, David S. Rubin, Masood H. Siddiqui, Sanjay Rastogi, “Statistics for Management”, Pearson India

- Amir D Aczel, Jayavel Sounderpandian, Palaniswamy Saravanan, Rohit Joshi, “Complete Business Statistics”, McGraw Hill Education
- Statistics for Managers Using Microsoft Excel, 9th Edition, David M. Levine, Baruch College, Zicklin School of Business, City University of New York, Pearson

Supplementary Reading:

<https://www.coursera.org/learn/basic-statistics?specialization=social-science>

<https://www.edx.org/learn/statistics>

Pearson e – library; <https://elibrary.in.pearson.com/bookshelfDashboard>

EBSCO: <https://www.ebsco.com/search?search=supplychainmodel>

Jgate: <https://jgateplus.com/home/resources/>

www.capitaline.com

CO-PO Mapping:

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11
CO1	1	3	1	-	--	2	3	-	3	-	--
CO2	-	3	1	-	1		3	3	3	1	
CO3	2	3	1	-	-		3	1	3	1	-
CO4		3	-	-	-	3	3	1	3	1	1
CO5	1	3	2	-	-	1	3	1	3	-	-

LEVEL 3-Substantial 2-Moderate 1-Slight - No Co-relation

Course Evaluation Plan:

Sl. No.	Evaluation Item	Unit of Evaluation	Marks Allotted	Timeline
1	End Term Exam	Individual	50	At the end of the semester
2	Mid - Term Test	Individual	20	After completion of 2-3 modules
3	Attendance and Class participation	Individual	5	At the end of the semester
4	Remaining assignments (Quiz, Individual assignment, Cap-Stone project, Major or minor project, Group assignments etc.)	Individual	25	Full Semester



I Semester M.B.A. Degree Examination, Jan./Feb. 2019
(CBCS Scheme)
2014-15 and Onwards
Paper – 1.4 : STATISTICS FOR MANAGEMENT

Time : 3 Hours

Max. Marks : 70

Instruction : Calculators and Tables are **allowed**.

SECTION – A

Answer **any five** questions. **Each** question carries **five** marks. (5×5=25)

1. What is meant by one tailed and two tailed tests ? Illustrate and explain.
2. What is meant by asymmetric distribution ? Explain the types with suitable illustrations.
3. A certain medicine was given to a village population to prevent mosquito related diseases. Use the Chi-square test and a 5 percent level of significance to determine whether the medicine was effective or not.

	Took ill	Did not take ill	Total
Took the Medicine	380	360	740
Did not take the Medicine	420	540	960
	800	900	1700

4. Find the linear trend through the method of least squares and forecast the sales for the next two years. A graph is not necessary.

Year	2013	2014	2015	2016	2017	2018
Sales in 00'000 Rs.	45	56	60	64	75	80



5. An investor wishes to buy shares from between the two companies given below. The market value of the shares of these two companies over ten days is given below. If the investor is looking for consistency in the shares he buys, use the coefficient of variation and advise him on which company shares he should buy.

	Day One	Day Two	Day Three	Day Four	Day Five	Day Six	Day Seven	Day Eight	Day Nine	Day Ten
Company Uno	110	190	175	185	105	115	190	170	180	120
Company Dno	145	155	150	160	155	165	160	150	140	150

6. What is meant by normal distribution ? Using illustrations, explain the above concept and how you will determine the entire area within the normal distribution curve. Also demonstrate how you will find the area to the right of $Z = 0.16$ with an illustration.
7. What are non parametric tests ? Briefly explain some non parametric tests and their advantages.

SECTION – B

Answer **any three** questions. **Each** question carries **ten** marks. **(3×10=30)**

8. From the data given below, you are required to :
- Calculate the correlation coefficient
 - Find the probable error and discuss the significance of correlation
 - Find the two regression equations.

Sales	42	44	50	54	60	70
Purchases	26	29	35	36	44	50



9. Three varieties of ore was examined by four Geologists and their sulphur content was found as below :

Ore Variety	Geologist One	Geologist Two	Geologist Three	Geologist Four
A	9	6	6	7
B	8	7	8	8
C	3	7	6	5

Use ANOVA with a 5 percent level of significance to determine whether there is any significant variation in the opinion of the Geologists.

- 10. What is the importance of a Hypothesis ? How is it set up ? What are the other factors involved in testing a hypothesis statistically ? Use illustrations in your answer.
- 11. What is meant by sampling ? Discuss the various methods of sampling.

**SECTION – C
(Compulsory)**

12. Case Study.

(1×15=15)

Calculate the index numbers of price by all five methods and prove that the Fischer's index number satisfies the factor reversal test and the time reversal test.

Commodity	P ₀	P ₁	Q ₀	Q ₁
A	13	14	10	12
B	16	17	13	15
C	12	14	14	16
D	17	19	18	19
E	18	20	19	20

I Semester M.B.A. Examination, February 2019
(CBCS Scheme)

Management

Paper – 1.4 : STATISTICS FOR MANAGEMENT

Time : 3 Hours

Max. Marks : 70

- Instructions :** 1) Calculators and appropriate statistical tables are **allowed**.
2) **Provide** the graph sheet.

SECTION – A

Answer **any five** questions. **Each** question carries **five** marks. **(5×5=25)**

1. Explain the importance of statistics in management.
2. Write short notes on :
 - a) Null hypothesis
 - b) Alternative hypothesis
 - c) Type I and Type II error.
3. A bag contains 5 white and 6 red marbles. Another bag contains 4 white and 7 red marbles. Two marbles are drawn from the selected bag. What is the probability that selected bag contains (a) white marbles (b) one white and one red marble.
4. Fit a linear trend for the following data and forecast for the next two years (A graph is necessary).

Year	2012	2013	2014	2015	2016	2017
Sale of sugar '000 kgs	80	90	92	94	96	98
5. Derive Chi-square statistic by stating suitable null and alternative hypothesis. Use 1% level of significance.

The data given below is regarding infavour of against and indifferent to a National Policy on FDI.

Occupation	Favour of	Against	Indifferent
Social workers	80	30	10
Lawyers	70	60	20
University students	60	60	40



6. Calculate Karl Pearson's and Bowley's coefficient of skewness for the marks obtained by students of 2 institutions.

Measure	Institution A	Institution B
Mean	65	70
Standard deviation.	10	14
Middle quartile	65	64
Third quartile	87	102
First quartile	28	35

7. The average height of 1000 students are normally distributed. Its mean is 72 inches and standard deviation is 2 feet. Find
- The number of students whose height is more than 68 inches.
 - The number of students whose height will be between 5.5 feet and 6.25 feet.

SECTION – B

Answer **any three** questions. **Each** question carries **10** marks. **(3×10=30)**

- What is non parametric test ? Explain the different types of test used in the statistical analysis.
- Calculate the ideal index and test for the time reversal and factor reversal test for the following data.

Commodity	2017		2018	
	Price	Expenditure	Price	Expenditure
A	30	1350	22	990
B	32	1344	24	840
C	30	1200	25	1200
D	35	2100	27	1161
E	36	900	28	1036



10. An investment company speculates about the relationship between family incomes and their allocation for investments. A survey of 8 randomly selected families gives the following data.

Annual income									
in '000 Rs.	:	18	21	19	34	23	30	36	39
Percent allocation									
for investment	:	28	36	32	40	35	55	60	70

- a) Develop the regression equations to describe the data.
- b) What could be the percentage of income allocated to investment by a family earnings Rs. 27,500 per annum ?

11. A businessman from Delhi wishes to sell his products in Bangalore. He can set up a showroom in the city or can sell through a wholesaler. Setting up a showroom will entail cost of Rs. 7,25,000 with a 65% probability of success. If the showroom succeeds, he can get a net profit of Rs. 12,25,000 per year. If it fails, he can either shutdown the showroom or rent it out for an annual rent of Rs. 4,25,000 (for rest of the year). The probability that he gets rent for the showroom is 45%.

If he sells through a wholesaler, he incurs Rs. 3,25,000 initial costs. The chances of selling successfully are 48% with a net profit of Rs. 6,20,000 per year.

- a) Advise the businessman on the best decision.
- b) How is the decision tree analysis useful in business decision ?

SECTION – C

12. **Compulsory :** **(1×15=15)**

A manufacturer of perfumes wishes to launch a new perfume in 4 different fragrances. Test marketing in 5 different cities has given the following data. Is there a significant difference in the sales figures of the various fragrances ?

	Lavender	Rose	Lily	Daisy
City A	80	100	95	70
City B	82	110	90	75
City C	88	105	100	82
City D	85	115	105	88
City E	75	90	80	65



I Semester M.B.A. Degree Examination, February 2016
(CBCS) (2014-15 & Onwards)

Management

Paper – 1.4 : STATISTICS FOR MANAGEMENT

Time : 3 Hours

Max. Marks : 70

Instruction : Calculators and tables are allowed.

SECTION – A

Answer any five questions. Each question carries five marks. (5×5=25)

In the frequency distribution of 100 families given below, the median is known to be 50. Find the missing frequencies.

Expenditure	No. of families
0 – 20	14
20 – 40	–
40 – 60	27
60 – 80	–
80 – 100	15
Total	100

An analysis of the monthly wages paid to workers in two firms A and B belonging to the same industry that gave the following results.

	A	B
Number of wage earners	566	648
Average monthly wage	52.50	47.50
Variance of the distribution	100	121

- Which firm pays the larger amount as monthly wages ?
- In which firm you find greater variability in individual wages ?



3. What is Correlation Analysis ? List and explain its types and uses.
4. Following data are available in respect of sales and advertisement expenditure.

	Sales	Advertisement Expenditure
Mean	70,000	15,000
Standard Deviation	15,000	3,000

Coefficient of correlation is + 0.8

Find the regression equations.

5. Explain Decision Theory along with its advantages and limitations.
6. Two sample polls of votes for two candidates A and B for a public office are taken, one from among residents of rural area and one from urban areas. The results are given below. Examine, whether the nature of the area is related to the voting preference in this election.

Votes for Area	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

7. Explain Bayes theorem and its applications.



SECTION – B

Answer any three of the following questions. Each question carries ten marks. (3×10=30)

- 8. Explain different methods of sampling with examples.
- 9. Compute Laspeyres, Paasche's and Fisher's price index number for 2015, using the following data concerning three commodities :

Commodity	2014		2015	
	Price (Rs.)	Quantity (Kg)	Price (Rs.)	Quantity (Kg)
A	15	15	22	12
B	20	5	27	4
C	4	10	7	5

Also show that it satisfies both Time Reversal Test and Factor Reversal Test.

- 10. A company appoints four salesmen, A, B, C, D and observes their sales in three seasons – summer, winter and monsoon. The figures (in lakhs) are given in the following table :

Season	Salesman				Total
	A	B	C	D	
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Total	90	93	81	96	360

Carry out an analysis of variance.



11. In Bangalore, 400 persons were considered regular consumers of pizzas out of a sample of 1000 persons. In Mangalore, 350 were regular consumers of pizzas out of sample of 800 persons. Test at 1% level of significance, whether there is a significant difference between the two towns as far as the proportion of pizza-eating habits are concerned.

SECTION – C

Compulsory.

(1×15=15)

12. A dietician wants to test 3 different types of diet plans to see if all these plans have similar weight reducing effects or not. He selected a homogenous group of 23 persons and placed them into 3 sub-groups, each sub-group trying a different diet plan. Each plan was tried for a period of 30 days.

The following observations of weight losses in kgs were recorded for members of each group after this period of 30 days.

Diet Plan 1	Diet Plan 2	Diet Plan 3
4.0	3.6	6.5
3.8	5.2	7.2
3.7	2.8	5.9
6.2	3.0	5.5
5.6	3.8	6.8
4.2	5.0	7.7
	3.9	8.0
	5.5	8.2
		7.0



I Semester M.B.A. Degree Examination, February 2017
(CBCS)

Management

Paper – 1.4 : STATISTICS FOR MANAGEMENT

Time : 3 Hours

Max. Marks : 70

Instruction : Statistical tables and calculators are **allowed**.

SECTION – A

Answer **any five** questions. Each question carries **five** marks. (5×5=25)

1. Explain the role of statistics in managerial decision-making. Illustrate with examples.
2. A bowler's scores for six games were 182, 168, 184, 190, 170 and 174. Using these data as a sample, compute the following descriptive statistics.
 - a) Standard Deviation
 - b) Variance
 - c) Coefficient of variation.
3. What is Sampling ? Explain the different methods of sampling.
4. Five students P, Q, R, S and T are given a problem to solve. The probabilities are $\frac{1}{3}$, $\frac{1}{5}$, $\frac{1}{6}$, $\frac{1}{8}$ and $\frac{1}{9}$ of solving the problem. What is the probability that the problem will be solved ?
5. The mean circumference of 1500 shafts manufactured in a company is 15 cm and the deviation from the mean is 3 cm. Assuming normal distribution find out how many shafts have a circumference
 - a) greater than 13 cm
 - b) lesser than 19 cm.

P.T.O.



6. From the following data, find the straight line trend and forecast the production figures for the next two years of a certain company. A graph is not necessary.

Year	2007	2008	2009	2010	2011	2012	2013	2014
Production ('000 kgs)	64	70	82	69	75	88	90	94

7. Using the chi square test, determine whether a new drug discovered for preventing poultry disease is successful or not, based on the data given below : You may use a 5% degree of significance.

	Got disease	Did not get disease
Administered the drug	175	810
Did not administer the drug	215	620

SECTION - B

Answer any three questions. Each question carries ten marks. (3×10=30)

8. Construct Laspeyre's, Paache's and Fischer's ideal index for the following data and prove that ideal index satisfies the time reversal and factor reversal tests for the data below :

Commodity	2015		2016	
	Price	Quantity	Price	Quantity
A	3	9	5	8
B	6	12	7	9
C	4	14	5	10
D	2	18	3	15



9. A study was carried out on the advertising methods of a brand of product. The unit sales achieved by five stores were recorded as under.

	Store - A	Store - B	Store - C	Store - D	Store - E
Method I	78	85	82	88	79
Method II	81	92	77	83	81
Method III	79	83	71	78	80

Calculate the F-ratio, using ANOVA and 15% level of significance. Establish there is a significant difference between the sales in the different stores.

10. Explain the following concepts briefly with suitable diagrams : (5x5=25)
- a) One tailed and two tailed tests
 - b) Type I and Type II errors
 - c) Skewness
 - d) Kurtosis.

11. Find the coefficient of correlation and the probable error for the following data.

X	12	24	30	45	56	70	83
Y	29	31	44	56	72	88	90

Comment on the significance of the correlation.

SECTION - C

12. Case study (compulsory) : (1x15=15)

Anil has 2 investment options, but he can take up only one option at a time.

Option one : He can start a restaurant for an investment of Rs. 8,00,000. The outcome will be success (probability of 90%) with a cash inflow of Rs. 10,00,000. If he fails he incurs a loss of Rs. 2,00,000. If he succeeds he can decide to open a fast food joint for Rs. 6,00,000. The outcome would be success (probability 70%) with a cash inflow of Rs. 8,00,000. Failure means he can still salvage Rs. 3,00,000.

Option two : He can start a readymade dress showroom for Rs. 8,00,000. The outcome will be success (probability 80%) with a cash inflow of Rs. 11,00,000. Failure means he can still salvage Rs. 5,00,000. Draw a decision tree and a pay off table. Advise Anil on the most profitable option to undertake.

Q.P. Code : 62205



Q.P. Code : 62205

**First Semester M.B.A. (Day/Evening) Degree Examination,
February/March 2020**

(CBCS Scheme)

Management

Paper 1.5 — BUSINESS STATISTICS

Time : 3 Hours]

[Max. Marks : 70

Note : Calculators and Statistical Tables are allowed.

SECTION - A

Answer any **FIVE** questions. Each question carries **5** marks : **(5 × 5 = 25)**

1. Explain the scope of Statistics in Managerial Decision Making.
2. Define the following concepts:
 - (a) Null hypothesis and alternative hypothesis
 - (b) Type I and Type II errors
 - (c) Confidence limits
 - (d) One tailed and two tailed tests
3. Five students A, B, C, D and E are given a problem to solve. The probabilities are 0.25, 0.20, 0.35, 0.52 and 0.65 of solving problem. What is the probability that atleast one of the student solves the problem?
4. For a group of 20 items $\sum x = 1452$, $\sum x^2 = 144280$ and median = 69.3. Find the Pearsonian Co-efficient of skewness.
5. Fit a linear trend by the method of least squares and estimate the number of patients for the years 2016 and 2017 from the following data.

Years :	2009	2010	2011	2012	2013	2014	2015
Patients in lakhs :	19	21	25	29	26	27	32



6. From the following information, find whether mentoring has an impact on the performance index.

Performance Index	Mentoring Done	No Mentoring
High	200	50
Average	150	70
Very Low	35	30

7. Test the hypothesis of no difference between the ages of male and female employees of a certain company using U test for the sample data. Use 0.05 level of significance.

Males :	31	25	38	33	42	40	44	26	43	35
Females :	44	30	34	47	35	32	48	34		

SECTION - B

Answer any **THREE** questions. Each question carries **10** marks : **(3 × 10 = 30)**

8. (a) What is business forecasting? Mention its methods.
 (b) Distinguish between variance and co-efficient of variance.
 (c) Define and explain the main characteristics of a Binomial Distribution.
 (d) Graphs and diagrams have an advantage over written reports. Comment briefly.
9. The following data gives the work experience of machine operators in a factory and the number of units of production turned out per day.

Machine Operator :	1	2	3	4	5	6	7	8	9
Work Experience in years :	6	8	7	5	2	1	3	9	10
Units of production :	50	60	54	47	25	20	41	62	70

- (a) Calculate the regression lines and estimate the probable units of production of a machine operator with an experience of 12 years.
 (b) Estimate the probable years of experience of a machine operator whose daily production is 85 units.

Q.P. Code : 62205

10. Calculate the consumer price index by the method of (a) Aggregative expenditure and (b) Family budget for the given data.

Commodity	Quantity in 2016	Price in 2016	Price in 2018
A	200	12	17.00
B	40	10	12.50
C	15	8	9.25
D	30	50	60.00
E	35	25	27.50
F	50	15	30.00

11. A company 'X' has 2 options to sell its products. It can set up a show room in the city or can sell from his factory outlet. Setting up a showroom will cost Rs. 7,00,000 with a 60% probability of success. If the showroom succeeds, it can gross a net profit of Rs. 15,00,000 per year. If it fails, it can close the showroom or rent it out for an annual rent of Rs. 3,50,000. The probability of getting rent is 80%.

If it sells from the factory outlet, it has to incur Rs. 75,000 as renovation charges. The chances of successful selling here is 45% with a net profit of Rs. 4,75,000 per year.

What will be your advise to the company? How a decision tree will help the company?

SECTION - C

12. Case Study **Compulsory** : **(1 × 15 = 15)**

Four judges of soft skills assessment test gave the following marks to six candidates. Test whether there is a significant difference in

- (a) the performance of the six candidates
(b) the judgement of the four judges.

Judges	Candidates					
	A	B	C	D	E	F
1	11	12	13	15	10	12
2	14	13	16	15	17	15
3	10	12	14	16	15	18
4	16	14	15	19	16	14



I Semester M.B.A. Degree Examination, January/February 2018
(CBCS) (2014-15 and Onwards)
MANAGEMENT

Paper – 1.4 : Statistics for Management

Time : 3 Hours

Max. Marks : 70

Instruction : Calculators and statistical tables are **allowed**.

SECTION – A

Answer **any five** questions from the following. Each question carries **five** marks. (5x5=25)

1. Briefly explain with illustrations how tables and graphs may be used to present data.
2. Explain the concepts of skewness and kurtosis with suitable illustrations.
3. Calculate the straight line trend through the method of least squares for the data given below :

Year	2013	2014	2015	2016	2017
Production in M.T	186	194	210	225	235

Also find the possible production figures of 2018 and 2019.

4. Using the Chi Square Test, determine whether the medicine given to cattle was effective or not.

Details	Took Medicine	Did not take Medicine	Total
Fell ill	150	230	380
Did not fall ill	375	420	795
Total	525	650	1175

You may use a 5 percent level of significance.

X	26	40	55	60	80
Y	42	58	57	73	30



5. Use the coefficient of variation to determine which of the 2 students are consistent in performance

Details	Maths	Science	History	Geography
Student A	55	65	80	70
Student B	93	87	30	40

6. What is meant by sampling ? Explain the different methods of sampling.
7. A company manufactures metal boxes. The monthly production is 4500 boxes. If the average diameter of the boxes is 6 cm and the standard deviation is 3 cm, find
- How many boxes have a diameter between 9 cm and 12 cm.
 - How many boxes have a diameter between 5 cm and 2 cm.
- Illustrate every answer with a suitable diagram.

SECTION – B

Answer **any 3** questions. **Each** carries **10** marks. **(3×10=30)**

8. A businessman has 2 options for investment

Option A : He can open a restaurant for Rs. 10,00,000. He can expect success with a cash inflow of Rs. 14,00,000 at a probability of 75 per cent. If he fails, he can still salvage Rs. 6,00,000.

When he succeeds he can open a fast food kiosk for Rs. 7,00,000. The chances of success are 80 per cent with a cash inflow of Rs. 6,00,000. If he fails, he loses Rs. 1,00,000.

Option B : He can open a Gym for Rs. 12,00,000. The chances of success are 60 per cent with a cash inflow of Rs. 8,00,000. If he fails, he can still salvage Rs. 6,00,000.

You are expected to

- Draw a decision tree.
- Construct a pay off table and state. Your decision as to which option is profitable for the businessman.



9. Find Fischer's ideal index for the following data and prove that it satisfies the factor reversal and time reversal tests.

Components	P_0	P_1	Q_0	Q_1
Rice	40	50	10	12
Wheat	45	55	9	10
Oil	70	75	10	11
Fuel	80	90	12	15
Clothing	30	40	15	20

10. Explain in detail the process of setting up and testing a hypothesis. You are expected to explain with suitable illustrations all the involved concepts.

11. A common exam was taken by 3 students in four different cities.

Using the ANOVA test, decide whether there is a significant difference in the academic performance of the students in different cities

Cities/Students	Marks of Student A	Marks of Student B	Marks of Student C
City One	60	70	45
City Two	70	65	55
City Three	75	55	85
City Four	85	90	75

You may use a 5 per cent level of significance.

SECTION – C

Compulsory case study. (1×15=15)

12. For the data given herein, you are required to :

- a) Find the coefficient of correlation
- b) Find the probable error and comment on the significance of correlation.
- c) Find the regression equations.
- d) Find Y when X = 50 and find X when Y = 45.

X	25	40	55	60	80
Y	42	58	67	73	90

PAPER • OPEN ACCESS

A Study on the Application of Decision Tree Algorithm in Mobile Marketing

To cite this article: Danning He 2021 *J. Phys.: Conf. Ser.* **2037** 012033

View the [article online](#) for updates and enhancements.

You may also like

- [Detection Method of Fast Flux Service Network Based on Decision Tree Algorithm](#)
Jiajia Wang and Yu Chen
- [Application of English Score Management System Based on Spark-Decision Tree Algorithm](#)
Yisha Zhang
- [Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques](#)
I.A. Pasadana, D. Hartama, M. Zarlis et al.



The Electrochemical Society
Advancing solid state & electrochemical science & technology

242nd ECS Meeting

Oct 9 – 13, 2022 • Atlanta, GA, US

Abstract submission deadline: **April 8, 2022**

Connect. Engage. Champion. Empower. Accelerate.

MOVE SCIENCE FORWARD



Submit your abstract



A Study on the Application of Decision Tree Algorithm in Mobile Marketing

Danning He^{1,*}

¹Sichuan Vocational College of Culture and Communication, 611230, Chongzhou, Sichuan, People R China.

*Corresponding author e-mail: hedanningl@scvc.edu.cn

Abstract. The decision tree algorithm is an inductive learning algorithm based on examples. It infers a classification rule represented by a tree structure through the learning of the training set. Mobile telecom operators have accumulated a large amount of user information in their long-term operations. Using decision tree algorithms to perform data mining on user information can accurately analyze user needs and user satisfaction with services. Based on the research of decision tree algorithm, this paper conducts analysis and mining based on mobile user data in order to improve the level of mobile marketing services and implement customer refined management strategies.

Keywords: Decision Tree Algorithm, Data Mining, Marketing

After entering the information age, the mobile telecommunications market has ushered in an open and highly competitive new business collaboration environment. Telecom operators insist on customer demand-oriented when conducting marketing. While maintaining the development of traditional services, they also formulate targeted marketing methods through user data mining. A large amount of customer demand information is carried in customer data. How to mine valuable information from the massive information depends on data mining technology. From the perspective of application, data mining technology mainly realizes the classification, prediction, estimation, association grouping, and homogeneous grouping of target data. Common data mining methods include decision tree algorithm, KNN algorithm, SVM method, Bayesian method, etc. Different algorithms follow different principles. The specific collection is as follows: Table 1.



Table 1. Data mining methods and ideas

Number	Data Mining Method	Design Idea
1	Decision tree algorithm	Generates an easy-to-understand tree structure and classification rules by inductively learning training data.
2	KNN algorithm	By calculating the distance between the current data and the other data in the data set, find the closest k data, and judge the category of the current data according to the majority of the k data.
3	SVM method	Find a hyperplane in the data set that can separate the data set. At the same time, it is required that the vertical distance between this plane and the boundary of the data set should be the largest. This maximizes the distance between the classes in the data set.
4	Bayesian method	First find the probability of the data under each possible classification condition, and then select the category with the highest probability is the category of the current data.

1. The Application Status of Data Mining in Mobile Marketing

1.1. The Development History of Data Mining Technology

Data mining technology was first applied to the financial industry and telecommunications industry, and played an important role in this field. Financial operation requires accurate and effective data support, which requires data mining technology to have a complete theoretical foundation. With the rapid development of the mobile telecommunications industry, a large amount of user data has been accumulated in a short period of time. The use of data mining technology to find valuable information can provide guarantee for the sustainable development of the mobile telecommunications industry. For example, building customers through data mining technology and the churn prediction model develops an effective retention strategy for users who may churn.

In recent years, under the guidance of user-oriented strategies, on the one hand, it has promoted the innovative development of data mining technology. On the other hand, it has accelerated the integration of data mining technology and the telecommunications industry. Traditional single data mining models (for example, customer relationship management models, customer segmentation models) have become common in the mobile telecommunications industry, but after all, a single data mining model has a limit to the degree of user segmentation. So multiple models must be developed. Only the combined data mining technology can meet the future application needs of the mobile telecommunications industry.

1.2. Application Status of Data Mining Technology

At present, the application of data mining in the telecommunications industry is mainly concentrated in user segmentation, churn user prediction, customer relationship management, etc., and less involved in the field of mobile marketing. From a practical point of view, using data mining technology to understand the needs of customers in the mobile market, customer segmentation based on this, and formulating corresponding marketing strategies for customers with different needs can effectively

increase customer retention. In this context, this article applies data mining technology to analyze the basic attributes of mobile market customers to predict potential customers, fake customers and sticky customers, and provide corresponding products and services to different customers, so as to achieve better marketing effects .

2. Introduction and Implementation of Decision Tree Algorithm

2.1. Introduction to Decision Tree Algorithm

The decision tree algorithm analyzes the internal information of the known classification data set, thereby extracting a set of formal and simple classification rules, which are expressed as a tree structure. And each node corresponds to a decision point of a certain attribute, which leads to There are multiple classification branches, and each branch accurately classifies the data. When it reaches the leaf node of the decision tree, the data can be classified completely and accurately.

The core advantage of the decision tree algorithm is that it can learn from known historical data, thereby establishing a tree model that can reveal the internal information and rules of the data and has a high information density. According to this model, the target data can be classified. After a quick analysis of the classification data set, a simple, intuitive, clear and understandable tree structure model is established based on the decision tree algorithm. Any branch between the root node and the leaf node marks an accurate classification rule. In addition, the decision tree algorithm has strong scalability, not only can quickly process small data sets, but also can deal with large data sets.

The current common decision tree algorithms include ID3 algorithm, C4.5 algorithm, SLIQ algorithm, SPRINT algorithm and so on. Among them, the ID3 algorithm follows the idea that small decision trees are better than large decision trees, and uses a top-down recursive method to construct trees. At the same time, the ID3 algorithm also borrows the idea of information entropy in information theory, by calculating the information entropy of each attribute in the current internal node classification in the candidate attribute set. So it can obtain the attribute with the smallest information entropy and use this attribute as Classification basis. In addition, it can obtain classification results containing valuable information.

2.2. Implementation of Decision Tree Algorithm

First, we create a node to distinguish the attribute value of the data partition label. If the attribute value of the partition data label belongs to the same type, return to the node and select this type of label. If the candidate set is an empty set, return to the node and select. Most classes mark this node, otherwise the information entropy of each candidate attribute under this node needs to be solved. The minimum information entropy is selected by comparing the information entropy, and a branch is created for each value according to the attribute of the minimum information entropy, so as to achieve the purpose of expanding the decision tree. Then call its own method on each branch to create its own child nodes, and finally generate a decision tree.

In the implementation path of the ID3 algorithm, the key link is to first calculate the information entropy of each attribute in the candidate attribute set, and finally select the attribute with the smallest information entropy value as the classification attribute through information entropy comparison. In this regard, suppose there is a sample data set X , the size of the data set is represented by $|X|$, the X set contains a target attribute R , which contains m values, and the sample data set X is divided into m data according to the target attribute R Subset, $P(X_i)$ represent the probability that each data in the X set belongs to the i -th subset. At this time, the expected information required to classify the data of the data sample set X is the formula (1):

$$\text{Info}(\mathbf{X}) = -\sum_{i=1}^v P(X_i) \log_2 P(X_i) \quad (1)$$

However, the actual situation is that these branches contain a lot of data of other classes, so in order to get accurate classification, you need to get the entropy of the subset:

$$\text{Info}_T(\mathbf{X}) = \sum_{j=1}^m \frac{|X_j|}{|\mathbf{X}|} \times \text{Info}(X_j) \quad (2)$$

Information gain is the difference between the two demands:

$$\text{Gain}(T) = \text{Info}(X) - \text{Info}_T(X) \quad (3)$$

$\text{Gain}(T)$ represents the information increment after dividing by the attribute T, and the attribute with the highest increment is selected as the best attribute for splitting. According to formula (3), it can be seen that the entropy value of each node classified by the target attribute $\text{Info}(X)$ is certain, and the smaller the value $\text{Info}_T(X)$, the maximum value $\text{Gain}(T)$. Therefore, in order to reduce the amount of calculation $\text{Info}_T(X)$, only the calculated value is needed in the actual algorithm, and the minimum value is selected as the best attribute for classification.

3. Defects and Optimization of ID3 Algorithm Based on Mobile Marketing Applications

3.1. Application Defects of ID3 Algorithm in Mobile Marketing

First, since the ID3 algorithm selects the split attribute based on the size of the classification information entropy of the attribute, the algorithm will calculate the classification information entropy of the attribute at each internal node $\text{Info}_A(\mathbf{X}) = -\sum_{i=1}^n P(X_i) \cdot \log P(X_i)$, which requires frequent calls to the system function `Math.log` for calculation. However, frequent calling of system functions greatly reduces the efficiency of the algorithm and causes a lot of waste of time.

Secondly, the ID3 algorithm is based on the highest information gain $\text{Gain}(T)$ as the criterion for the selection of classification attributes. It can be seen from the formula $\text{Gain}(T) = \text{Info}(X) - \text{Info}_T(X)$ that the higher the value $\text{Gain}(T)$, the smaller the value $\text{Info}_T(X)$ is required, and the attribute with more attribute values $\text{Info}_T(X)$ will be smaller by calculation. The attribute T with the most attribute value will be considered as the best classification attribute and will be selected finally. However, in practical applications, the attributes with more attribute values are not the sticky attributes that are of interest to actual problems. For example, the customer attributes that operators are more interested in include "whether the user is the lowest consumer", "the user uses the service brand", etc. The information gain of the latter is significantly higher than that of the former, so the ID3 algorithm will choose the attribute of "users use service brand", but this will deviate from the original intention.

3.2. Application-oriented ID3 Algorithm Optimization

For the application scenario of mobile marketing, this section achieves the goal of simplifying the calculation by decomposing the information entropy formula of the classification attribute A. Assuming that the label set E of the data target attribute value in the data set X has two values (e_1, e_2) ,

the size of each data set is m and n in turn, and the sum of the two is the size of the total data set $|X|$. The attribute A has v values (a_1, a_2, \dots, a_v) , and when the attribute $A = a_i$ is the attribute, there are m_i bars of data $E = e_1$ and n_i bars of data $E = e_2$. Thus, the information entropy of A attribute:

$$\text{Info}_A(X) = \sum_{i=1}^v \frac{m_i + n_i}{m + n} \cdot \log \text{Info}(X) P(X_i) \quad (4)$$

$$\text{Info}(X) = -\frac{m_i}{m_i + n_i} \log_2 \frac{m_i}{m_i + n_i} - \frac{n_i}{m_i + n_i} \log_2 \frac{n_i}{m_i + n_i} \quad (5)$$

After integration and simplification, we get:

$$\text{Info}_A(X) = \sum_{i=1}^v \frac{1}{(m + n) \ln 2} \left(-m_i \ln \frac{m_i}{m_i + n_i} - n_i \ln \frac{n_i}{m_i + n_i} \right) \quad (6)$$

Knowing that $\ln 2$ is a constant, and $m+n=|X|$, the above formula (6) is simplified:

$$\text{Info}_A(X) = -\sum_{i=1}^v \left(m_i \ln \frac{m_i}{m_i + n_i} + n_i \ln \frac{n_i}{m_i + n_i} \right) \quad (7)$$

Use the definitions of Taylor formula and McLaughlin formula, a large number of log functions are used in the formula for calculating information entropy. Let $f(x) = \ln(1+x)$, then we can get

$$f(x) = \ln(1+x) \approx x \quad (8)$$

To further simplify the above formula (7), we get:

$$\text{Info}_A(X) = \sum_{i=1}^v \frac{2m_i n_i}{m_i + n_i} \quad (9)$$

After the above formula optimization processing, the simplified information entropy formula (9) of the classification attribute A is finally obtained. The calculation amount of this formula is greatly reduced, especially the log calculation is omitted. There is no need to frequently call the system function `Math` during the execution of the ID3 algorithm `log`, which significantly improves computational efficiency.

This chapter reveals the shortcomings of the ID3 algorithm in mobile marketing applications, that is, the amount of calculation is large and the selection of attributes is biased. By decomposing the information entropy formula of the classification attribute A , an improved information entropy formula is finally obtained for user classification in marketing can be more effective in practice.

4. Conclusion

At this stage, data mining technology has been widely used in the mobile telecommunications industry, and has achieved fruitful applications in many aspects. This paper takes data mining

technology as a starting point, collects and analyzes common algorithms in data mining, and focuses on the idea and process of ID3 algorithm. At the same time, the ID3 algorithm is improved according to the characteristics of the user's own attributes in mobile marketing. The ID3 algorithm not only improves computing efficiency, but also closes the relationship between operators and customer groups, thereby guiding the formulation of more targeted marketing strategies.

References

- [1] Zhang Meitu(2011). The application of data mining in the mobile communication market. *Modern Telecommunications Technology*, no.1, pp: 110-114.
- [2] Xiao Mingwei(2012). Research on the application of data mining technology based on customer share of telecommunication groups. *Computer and Telecommunications*, no.6, pp: 38-39,42.
- [3] Safaa O A, Jassim F S(2013). Evaluation of Different Data Mining Algorithms with KDD CUP 99 Data Set[J]. *Journal of Babylon University*, vol.21, no.8, pp: 2663-2681.
- [4] Chen Ping(2009). Parallel algorithm design and performance analysis of decision tree in data mining grid. *Journal of Beijing University of Posts and Telecommunications*, no.1.
- [5] Wang Yongmei(2011). Research on ID3 algorithm in decision tree. *Journal of Anhui University (Natural Science Edition)*, no.3.
- [6] Xue Yongning(2013). Research on Data Mining Technology. *Chinese and Foreign Entrepreneurs*, vol.32, pp: 236.
- [7] Zhao Wei(2013). Research on Data Mining Technology of Computer Semi-structured Data Source. *Computer CD Software and Application*, no.8.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/233988919>

Determination of Trading Points using the Moving Average Methods

Conference Paper · June 2010

CITATIONS

10

READS

26,183

1 author:



[Puchong Praekhaow](#)

King Mongkut's University of Technology Thonburi

6 PUBLICATIONS 10 CITATIONS

SEE PROFILE

Determination of Trading Points using the Moving Average Methods

Puchong Praekhaow^{1,*}

¹ Department of Mathematics, Faculty of Science, King Mongkut's University of
Technology Thonburi, Bangkok 10140, Thailand

*Corresponding author: puchong.pra@kmutt.ac.th

Abstract

This paper is focused on the net profit of trading points using three moving average techniques based on pattern determinations. Stocks' samples were selected by simple random sampling method from the stocks in the set 50 stocks indexes of the Thailand Stock Market. The net profits of the investment in each trading methods on the experiments were compared. The results indicated that all of the moving average technique can make profit for trading. Moreover, the simple moving average technique can give profit-making up to 9% ,which is better than other methods.

Keywords: Moving Average, Trading, Profit.

1. INTRODUCTION

Nowadays, the investments in stock markets around the world have become more complicated. This is because of the dynamic fast pace nature of the business world. Hence, the simple system to provide current information and predict future stock prices is desirable to inventors. Moreover, these systems need to be back-ended with suitable mathematical and statistical tools.

Since, the investor's problem was trading points for profitable investment in the stock markets, the moving average technique was one of the tools generally used to facilitate decision making in buying and selling stocks. Therefore, the researchers are interested in the moving average trading rule for very sophisticate and simple systems. For example, the moving average rules have predictive power and can illustrate the patterns of the stock prices for profitable trading. (Metghalchi, M., X. Garza-Gomez, et al. 2008), the variable moving averages (VMAs) and the fixed moving averages (FMAs) in the China, Thailand, Taiwan, Malaysia, Singapore, Hong Kong, Korea, and Indonesia stock markets. The length of 20 days and 60 days appeared to be the most profitable for variable and fixed moving averages, respectively. (Ming-Ming, L. and L. Siok-Hwa. 2006), the Adaptive Moving Average over fixed length Simple Moving Average (SMA) trading systems is an ability to automatically respond to changing market. (Ellis, C. A. and S. A. Parbery 2005), The total of successive trading is a non-stationary on the average in about three out of four cases trading rule signals are false, a fact that leaves a lot of space for improved trading rule performance if trading rule signals are combined with other information.(e.g. filters, or volume of trade) (Milionis, A. E. and E. Papanagiotou 2008). The performance of the moving average trading rule is improved, if it is combined with other indicators. (Gen,cay andStengos 1998), the past information on the volume of trade improves the performance of the moving average trading rule, (Fang and Xu 2003). All stated above were performed by using the moving average of the stock prices as determinant for selling periods or buying periods.

At present, the subjects in other research papers are about the

trading rules and the crucial rules to the investment strategy for investors. In this research the main purpose is on the development of trading rule with the model that is invented by the three methods of the moving average as follows, simple moving average (SMA), weighted moving average (WMA), and exponential moving average (EMA). Besides, this research provides comparative regard to the profit for trading in the stock market of all three moving average methods.

2. METHODOLOGY

The daily closing prices of the stock were forecasted by the moving average three methods in this research are calculated with formulas as follows,

1. Simple Moving Average (SMA)

$$SMA_{t+1} = SMA_t - \frac{P_{t-n+1}}{n} + \frac{P_t}{n}$$

Where, SMA_{t+1} is defined as the simple moving average with the n time length is calculated at time $(t+1)$,

$(P_i, i = t, t-1, \dots, t-n+1)$ are observed values of the stock prices at time (t) .

2. Weighted Moving Average (WMA)

$$WMA_{t+1} = \frac{nP_t + (n-1)P_{t-1} + \dots + 2P_{t-n+2} + P_{t-n+1}}{n + (n-1) + \dots + 2 + 1}$$

Where, WMA_{t+1} is defined as the weight moving average with n time length is calculated at time $(t+1)$,

$(P_i, i = t, t-1, \dots, t-n+1)$ are observed values of the stock prices at time (t) .

3. Exponential Moving Average (EMA)

$$EMA_{t+1} = \alpha P_t + \alpha(1-\alpha)P_{t-1} + \alpha(1-\alpha)^2 P_{t-2} + \dots \\ \dots + \alpha(1-\alpha)^{t-n+2} P_{t-n+2} + (1-\alpha)^{t-n+1} EMA_{t-n+1}$$

Where, EMA_{t+1} is defined as the exponential moving average

with n time length is calculated at time $(t + 1)$,

$(P_i, i = t, t - 1, \dots, t - n + 1)$ is observed value of the stock prices at time (t) . and α is a constant smoothing factor calculated from the Pearson correlation coefficient of closing stock prices today and yesterday.

The signals of trading rule model were determined as follows (see in Fig. 1).

1. The buying signals are generated at the crossover which the daily closing stock price is moving up higher than the moving average method of the daily closing stock price. Therefore, the investor will buy the stock on the opening price in the next day.

2. The selling signals are generated at the crossover which the daily closing stock price is moving down lower than the moving average method of the daily closing stock price. The investor will sell the stock on the opening price in the next day, which there is the profit. If the stock was not the profit, the investor will sell the stock in the next cycle of trading.

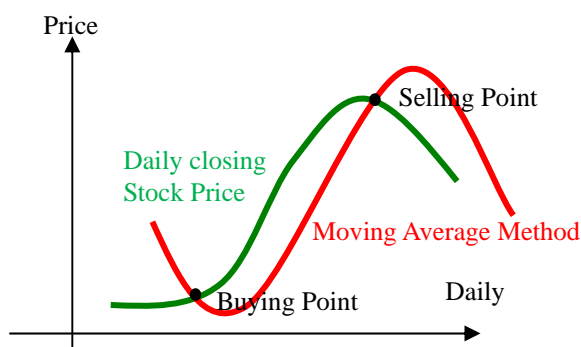


Fig. 1 The signal points of trading rule

The signal points of trading rule model were examined with the experiment by sampling from 4 stocks in the set 50 stocks indexes of Thailand Stock Exchange between January 1st, 2009 and July 31st, 2009. These stocks in the set 50 stocks indexes were the common shares with high market capitalization and trading liquidity, which were consistently high to be chosen for calculation. The four stocks were the sample size of the experiment calculated by Yamane's formula (Yamane, T., 1967) with the error of 5% of the profit. These samples were collected by simple random sampling, with energy utilities, bank, media publishing, and building stock, respectively. The sample stocks in this research are PTT public company limited (PTT), Siam commercial bank public company limited (SCB), BEC world public company limited (BEC), and Italian Thai development public company limited (ITD). The data are displayed on the graphs of the daily closing price and the price forecasting with moving average to find the trading signal spots of shares with time length 5 days, 10 days and 20 days during January 1st, 2009 to July 31st, 2009. The percentages of the net profit of determination of Trading Points was calculated the three moving average methods. (Following the methodology proposed by Brock, W., Lakonishok, J., LeBaron, B., 1992)

The net profits of selling in this research were calculated with the formulas as follows,

$$\text{Net Profit} = \text{Sale Price} - \text{Cost Price}$$

$$\text{Cost Price} = \text{Market Price} + \text{Commission} + \text{Taxes}$$

$$\text{Sale Price} = \text{Market Price} - \text{Commission} - \text{Taxes}$$

$$\text{Percentage of the net profit} = \frac{\text{Net Profit}}{\text{Cost Price}} \times 100$$

The net profits of the investment in each trading methods on the experiments were compared by the analysis of variance test (ANOVA). (R.A. Fisher, Referred by Montgomery, D. C., 1991) If the statistics test was conducted at 0.05 significant level, the net profits of three moving average methods will be multiple compared with Duncan Multiple Rang test. (Duncan, Referred by Montgomery, D. C., 1991)

3. RESULTS

Between January 1st, 2009 and July 15th, 2009, the price of the sample stocks (BEC, ITD, PTT, and SCB) had the closing price average, maximum price and minimum price value as shown in Table. 1.

Table 1 Price of the sample stocks

Stock	Min Price(Bath)	Max Price(Bath)	Average Price(Bath)
BEC	18.1	21.4	19.39
ITD	1.96	3.49	2.55
PTT	143.15	257.32	184.97
SCB	49.47	80.27	80.48

In the table 1, it shows each that the stock price of each increased during the experiment, PTT's share increased in the highest rate, 79.75% and BEC's share increased in the lowest rate, 18.23%, while the ITD and SCB shares increased in the rates 78.06% and 62.26%, respectively.

3.1 Buying and Selling Points

The signals of buying and selling were determined by trading rule model with three moving average methods for the experiment as follows (see Fig. 2 – 4).

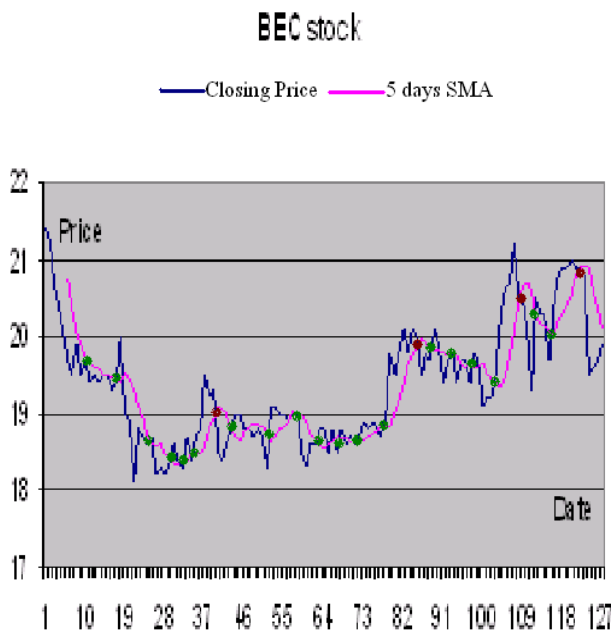


Fig. 2 Trading signals for 5 days Simple Moving Average of BEC.

From fig. 2, it can be observed that there are 19 buying points shown with the green points (●) and 4 selling points shown with the red points (●). The investors can take these trading points to make decisions for the net profit.

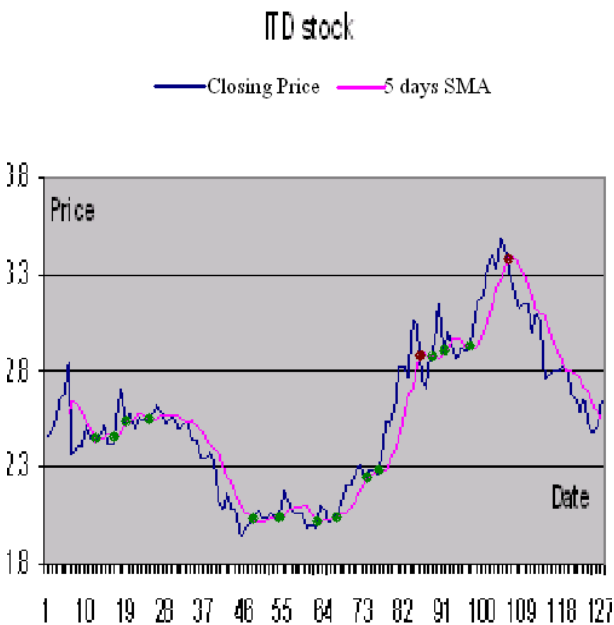


Fig. 3 Trading signals for 5 days Simple Moving Average of ITD.

From fig. 3, it can be observed that there are 13 buying points shown with the green points (●) and 2 selling points shown with the red points (●). The investors can take these trading points to make decisions for the net profit.

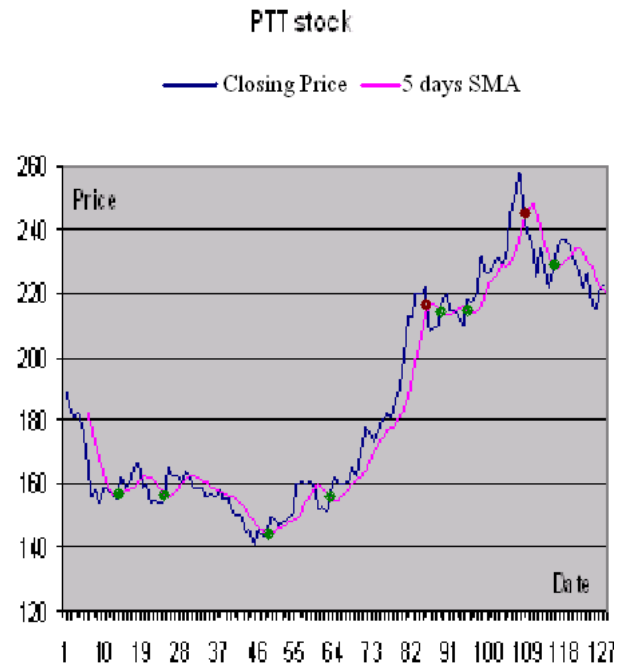


Fig. 4 Trading signals for 5 days Simple Moving Average of PTT.

From fig.4, it can be observed that there are 7 buying points shown with the green points (●) and 2 selling points shown with the red points (●). The investors can take these trading points to make decisions for the net profit.

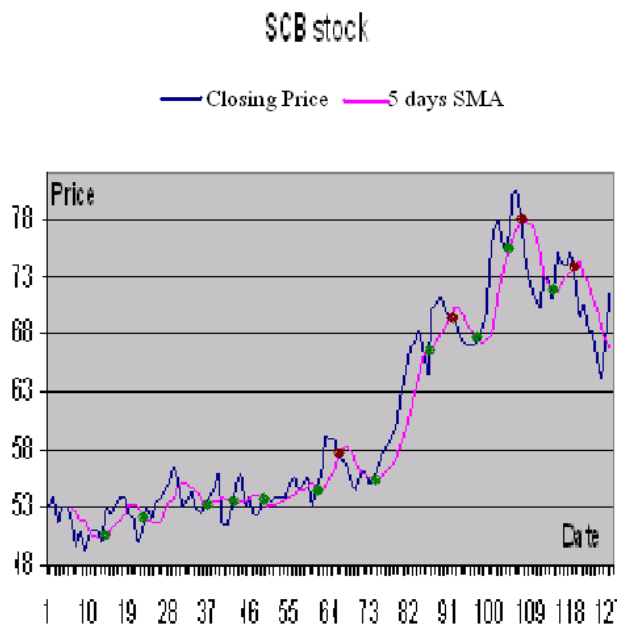


Fig. 5 Trading signals for 5 days Simple Moving Average of SCB.

From fig.5, it can be observed that there are 12 buying points shown with the green points (●) and 4 selling points shown with the red points (●). The investors can take these trading points to make decisions for the net profit.

The signals of buying and selling determined by trading rule model of the SMA at interval 10 days, 20 days, the WMA at interval 5 days, 10 days, 20 days, and the EMA at interval 5 days, 10 days, 20 days, they were calculated in the same methods.

3.2 Returns of Trading

The net profits were trading with the MA types of the experiment as shown in Table 2.

Table 2 Average of the net profit with three MA.

MA Types	Mean	N	Std. Deviation
SMA	11.8025	12	9.31474
WMA	10.3608	12	8.57753
EMA	4.0283	12	2.06732
Total	8.7306	36	7.96636

From Table 2, it was found that the average rates of return were 8.7306% and the net profit of SMA was better than other methods.

The net profits were trading with the interval 5 days, 10 days, and 20 days MA of the experiment as shown in Table 3.

Table 3 Average of the net profit with the interval-days MA

Interval	Mean	N	Std. Deviation
5-day	7.8092	12	6.23628
10-day	7.5058	12	6.37809
20-day	10.8767	12	10.71304
Total	8.7306	36	7.96636

Table 3 shows that the average rates of return were 8.7306% and the net profit of 20 days SMA was better than other methods.

Table 4 The percentage of the net profit of all the methods.

Stock	Type of MA	the percentage of the net profit		
		5 days	10 days	20 days
BEC	SMV	1.72	2.05	2.12
	WMV	2.00	1.91	1.77
	EMV	0.99	1.28	1.62
ITD	SMV	12.98	10.54	23.26
	WMV	12.31	12.17	26.07
	EMV	7.61	4.61	5.17
PTT	SMV	18.55	21.45	29.32
	WMV	18.15	16.55	21.31
	EMV	3.18	3.34	3.53
SCB	SMV	6.65	5.66	7.34
	WMV	4.44	4.52	3.14
	EMV	5.14	6.00	5.87

Table 4 shows that the PTT stock that determined 20 days SMA give to a higher net profit better than other methods, while the 5 days EMS of BEC stock give to a lowest net profit

3.3 Comparing the Analysis

Using the one – way ANOVA method to analyze the relation of the net profit depended on the all three moving average methods. The statistics test was conducted at 0.05 significant levels (p-value < 0.05) as shown in Table 5.

Table 5 Result of ANOVA

Sour of Variance	Sum of Square	df	Mean Square	F
Between Groups	410.467	2	205.233	0.034*
Within Groups	1810.73	33	54.871	
Total	2221.20	35		

* Significant at 0.05

From Table 5, there was playing significant of the F statistics test show that the net profit depends on the type of the moving average methods. Then, the net profits of all three moving average methods were multiple compared by using Duncan Multiple Rang test. (Refer to Montgomery, D. C., 1991) The results are presented in the Table 6.

Table 6 Multiple Comparison of the returns

Type of MA	SMA (11.08)	WMA (10.36)	EMA (4.03)
SMA(11.08)	-	-0.72	-7.05*
WMA(10.36)		-	-6.33*
EMA(4.03)			-

* Significant at 0.05,
(Number) is a percentage of the net profit

Table 6 shows that the percentage' net profit in the EMA is different from the SMA and WMA at 0.05 significant levels.

4. DISCUSSION

The results of the net profits were trading by the three moving average rules in Thailand stock markets; the causes of the net profits might be from the increasing of stocks prices in the interval of the experiment. The SMA can give to produce the highest returns among the various rules during the period of the experiment in the time test and there is the ability to generate above average returns of the net profit. This is because the SMA is not sensitive to vary the stocks prices while to other is better. The interval length of about 20 days in the moving average methods of the investment was emerged as the most profitable. It might cause the market condition at that time belong to the investment on long time.

5. Conclusion

This study presents development of the moving average trading rules in Thailand Stock Exchange. The experiment of the research using a sample was chosen by simple random sampling method from the stocks in the set 50 stocks indexes on the Thailand Stock Market. The results found that the time length of 20 days emerged as the most profitable for SMA and EMA. Also, it indicated that profits generated from SMA were

higher than WMA and EMA. The stocks in set 50 stocks indexes of Thailand Stock Exchange showed the most promising net profits from the applying of both the three moving averages and the trading points of the time length 20 days. There are the consistence with market efficiency, and the moving average trading rules make a better opportunity for buying and selling of the stocks profitable at any time. Future researches should be on other popular technical indicators such as the variables that relative strength indicator on the individual stock in these markets would be of interest as well as to measure additional insights. Also, the study should include the process monitoring the basic factor of the stock. It could make determination of trading points to be used much more widely.

pp.519-544.

5. ACKNOWLEDGMENTS

I thank the editor, the associate editor and two anonymous referees for helpful comments which improved the results in the paper. I also thank the Faculty of Science, King Mongkut's University of Technology Thonburi for the support.

6. REFERENCES

- [1] Andrada-Félix, J. and F. Fernández-Rodríguez (2008). "Improving moving average trading rules with boosting and statistical learning methods." *Journal of Forecasting* 27(5): pp.433-449.
- [2] Metghalchi, M., X. Garza-Gomez, et al. (2008). "Are moving average trading rules profitable? Evidence from the Mexican stock market." *Journal of Applied Business Research* 24(1): pp.115-128.
- [3] Milionis, A. E. and E. Papanagiotou (2008). "On the use of the moving average trading rule to test for weak form efficiency in capital markets." *Economic Notes* 37(2): pp.181-201.
- [4] Ellis, C. A. and S. A. Parbery (2005). "Is smarter better? A comparison of adaptive, and simple moving average trading strategies." *Research in International Business and Finance* 19(3): pp.399-411.
- [5] Gunasekarage, A. and D. M. Power (2001). "The profitability of moving average trading rules in South Asian stock markets." *Emerging Markets Review* 2(1): pp.17-33.
- [7] Kuo, G. W., E. Acar, et al. (2002). Some exact results for moving-average trading rules with applications to UK indices. *Advanced Trading Rules (Second Edition)*. Oxford, Butterworth-Heinemann: pp.152-173.
- [8] Kwon, K.-Y. and R. J. Kish (2002). "A comparative study of technical trading strategies and return predictability: an extension of using NYSE and NASDAQ indices." *The Quarterly Review of Economics and Finance* 42(3): pp.611-631.
- [9] Ming-Ming, L. and L. Siok-Hwa (2006). "The profitability of the simple moving averages and trading range breakout in the Asian stock markets." *Journal of Asian Economics* 17(1): pp.144-170.
- [10] Parisi, F. and A. Vasquez (2000). "Simple technical trading rules of stock returns: evidence from 1987 to 1998 in Chile." *Emerging Markets Review* 1(2): pp.152-164.
- [11] Zhu, Y. and G. Zhou (2009). "Technical analysis: An asset allocation perspective on the use of moving averages." *Journal of Financial Economics* 92(3):

Dependent T-Test Statistics

Introduction

The dependent t-test (called the paired-samples t-test in SPSS Statistics) compares the means between two related groups on the same continuous, dependent variable. For example, you could use a dependent t-test to understand whether there was a difference in smokers' daily cigarette consumption before and after a 6 week hypnotherapy programme (i.e., your dependent variable would be "daily cigarette consumption", and your two related groups would be the cigarette consumption values "before" and "after" the hypnotherapy programme). If your dependent variable is dichotomous, you should instead use [McNemar's test](#).

This "quick start" guide shows you how to carry out a dependent t-test using SPSS Statistics, as well as interpret and report the results from this test. However, before we introduce you to this procedure, you need to understand the different assumptions that your data must meet in order for a dependent t-test to give you a valid result. We discuss these assumptions next.

SPSS Statistics^{top ^}

Assumptions

When you choose to analyse your data using a dependent t-test, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using a dependent t-test. You need to do this because it is only appropriate to use a dependent t-test if your data "passes" four assumptions that are required for a dependent t-test to give you a valid result. In practice, checking for these four assumptions just adds a little bit more time to your analysis, requiring you to click a few more buttons in SPSS Statistics when performing your analysis, as well as think a little bit more about your data, but it is not a difficult task.

Before we introduce you to these four assumptions, do not be surprised if, when analysing your own data using SPSS Statistics, one or more of these assumptions is violated (i.e., is not met). This is not uncommon when working with real-world data rather than textbook examples, which often only show you how to carry out a dependent t-test when everything goes well! However, don't worry. Even when your data fails certain assumptions, there is often a solution to overcome this. First, let's take a look at these four assumptions:

- **Assumption #1:** Your **dependent variable** should be measured on a **continuous** scale (i.e., it is measured at the **interval** or **ratio** level). Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. You can learn more about continuous variables in our article: [Types of Variable](#).
- **Assumption #2:** Your **independent variable** should consist of two **categorical**, "**related groups**" or "**matched pairs**". "Related groups" indicates that the same subjects are present in both groups. The reason that it is possible to have the same subjects in each group is because each subject has been measured on two occasions on the same dependent variable. For example, you might have measured 10 individuals' performance in a spelling test (the dependent variable) before and after they underwent a new form of computerised teaching method to improve spelling. You would like to know if the computer training improved their spelling performance. The first related group consists of the subjects at the beginning of (prior to) the computerised spelling training and the second related group consists of the same subjects, but now at the end of the computerised training. The dependent t-test can also be used to compare different subjects, but this does not happen very often. Nonetheless, to learn more about the different study designs that can be analysed using a dependent t-test, see our enhanced dependent t-test guide.
- **Assumption #3:** There should be **no significant outliers** in the **differences** between the two related groups. Outliers are simply single data points within your data that do not follow the usual pattern (e.g., in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very

unusual, and may even put her in the top 1% of IQ scores globally). The problem with outliers is that they can have a negative effect on the dependent t-test, reducing the validity of your results. In addition, they can affect the statistical significance of the test. Fortunately, when using SPSS Statistics to run a dependent t-test on your data, you can easily detect possible outliers. In our enhanced dependent t-test guide, we (a) show you how to use SPSS Statistics to compute the difference scores, (b) show you how to detect outliers using SPSS Statistics, and (c) discuss some of the options you have in order to deal with outliers.


- **Assumption #4:** The distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed. We talk about the dependent t-test only requiring approximately normal data because it is quite "robust" to violations of normality, meaning that the assumption can be a little violated and still provide valid results. You can test for normality using the Shapiro-Wilk test of normality, which is easily tested for using SPSS Statistics. In addition to showing you how to do this in our enhanced dependent t-test guide, we also explain what you can do if your data fails this assumption (i.e., if it fails it more than a little bit).

You can check assumptions #3 and #4 using SPSS Statistics. Before doing this, you should make sure that your data meets assumptions #1 and #2, although you don't need SPSS Statistics to do this. When moving on to assumptions #3 and #4, we suggest testing them in this order because it represents an order where, if a violation to the assumption is not correctable, you will no longer be able to use a dependent t-test (although you may be able to run another statistical test on your data instead). Just remember that if you do not run the statistical tests on these assumptions correctly, the results you get when running a dependent t-test might not be valid. This is why we dedicate a number of sections of our enhanced dependent t-test guide to help you get this right. You can find out about our enhanced content as a whole [here](#), or more specifically, learn how we help with testing assumptions [here](#).

In the section, [Test Procedure in SPSS Statistics](#), we illustrate the SPSS Statistics procedure required to perform a dependent t-test assuming that no assumptions have been violated.

First, we set out the example we use to explain the dependent t-test procedure in SPSS Statistics.

Get access to all the enhanced SPSS guides in Laerd Statistics



“ Thanks again for making such a helpful website and making a subscription completely affordable ”
Darlynda, USA

“ I just signed up for Laerd Statistics and must say that your website is just what I needed! Your guides are very detailed and easy to follow, which is just perfect! ”
Anne, Denmark

TAKE THE TOUR
PLANS & PRICING

SPSS Statistics^{top ^}

Example

A group of Sports Science students ($n = 20$) are selected from the population to investigate whether a 12-week plyometric-training programme improves their standing long jump performance. In order to test whether this training improves performance, the students are tested for their long jump performance before they undertake a plyometric-training programme and then again at the end of the programme (i.e., the dependent variable is "standing long jump performance", and the two related groups are the standing long jump values "before" and "after" the 12-week plyometric-training programme).

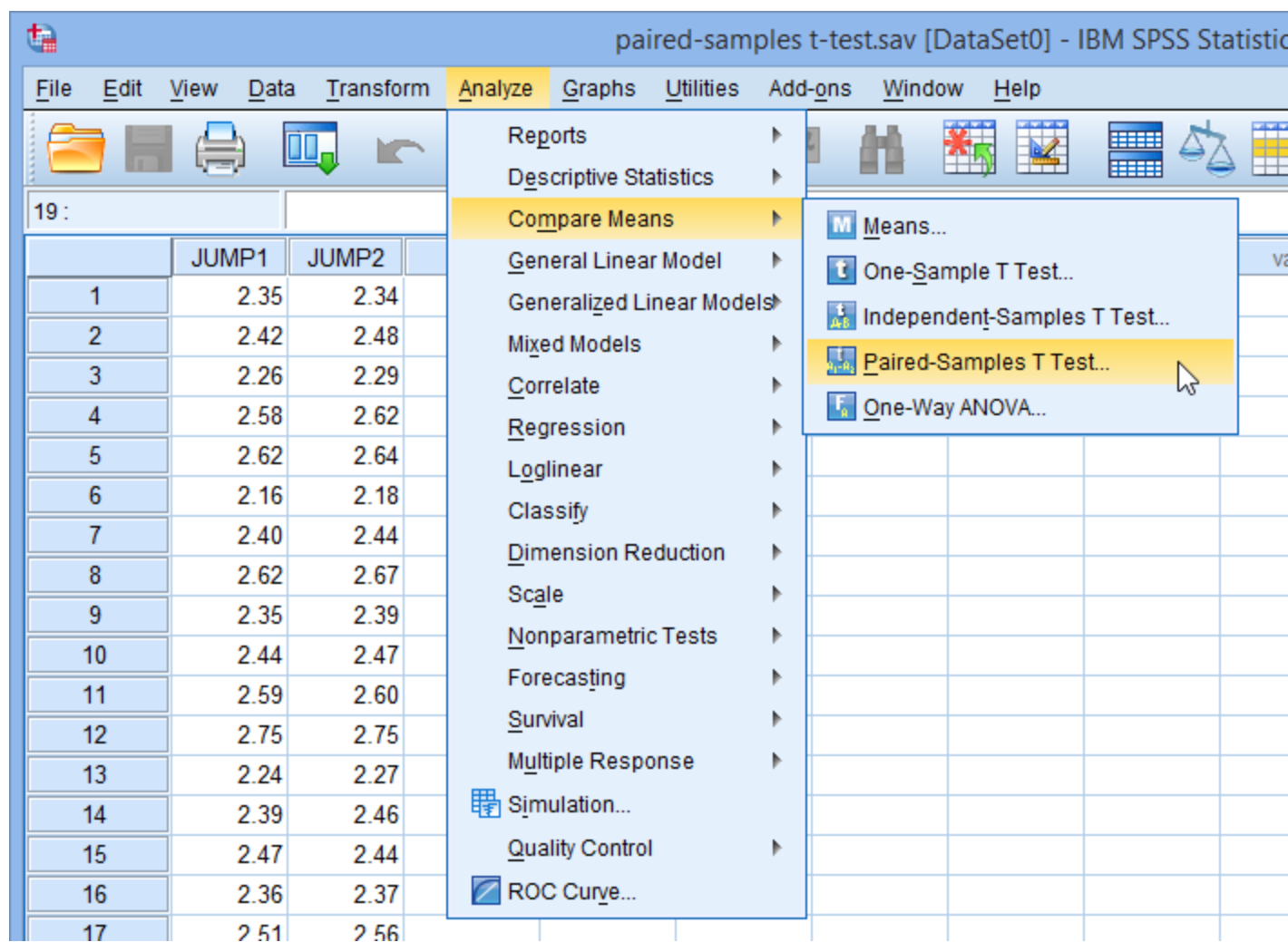
SPSS Statistics^{top ^}

Test Procedure in SPSS Statistics

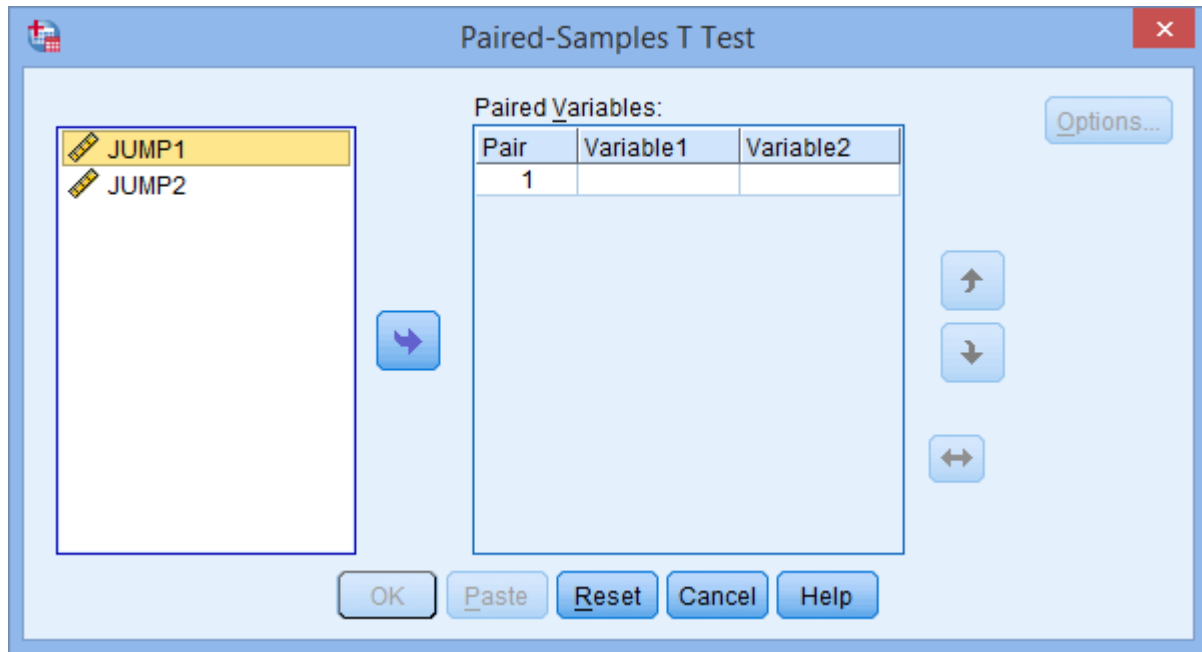
The six steps below show you how to analyse your data using a dependent t-test in SPSS Statistics when the four assumptions in the previous section, [Assumptions](#), have not been violated. At the end of these six steps, we show you how to interpret the results from this

test. If you are looking for help to make sure your data meets assumptions #3 and #4, which are required when using a dependent t-test, and can be tested using SPSS Statistics, you can learn more in our enhanced guides [here](#). We also show you how to correctly enter your data into SPSS Statistics in order to run a dependent t-test, as well as explaining how to deal with missing values (e.g., if a participant completed a pre-test but failed to turn up to the post-test). However, in this "quick start" guide, we focus on the six steps required to run the dependent t-test procedure using SPSS Statistics.


- Click **Analyze > Compare Means > Paired-Samples T Test...** on the top menu, as shown below:

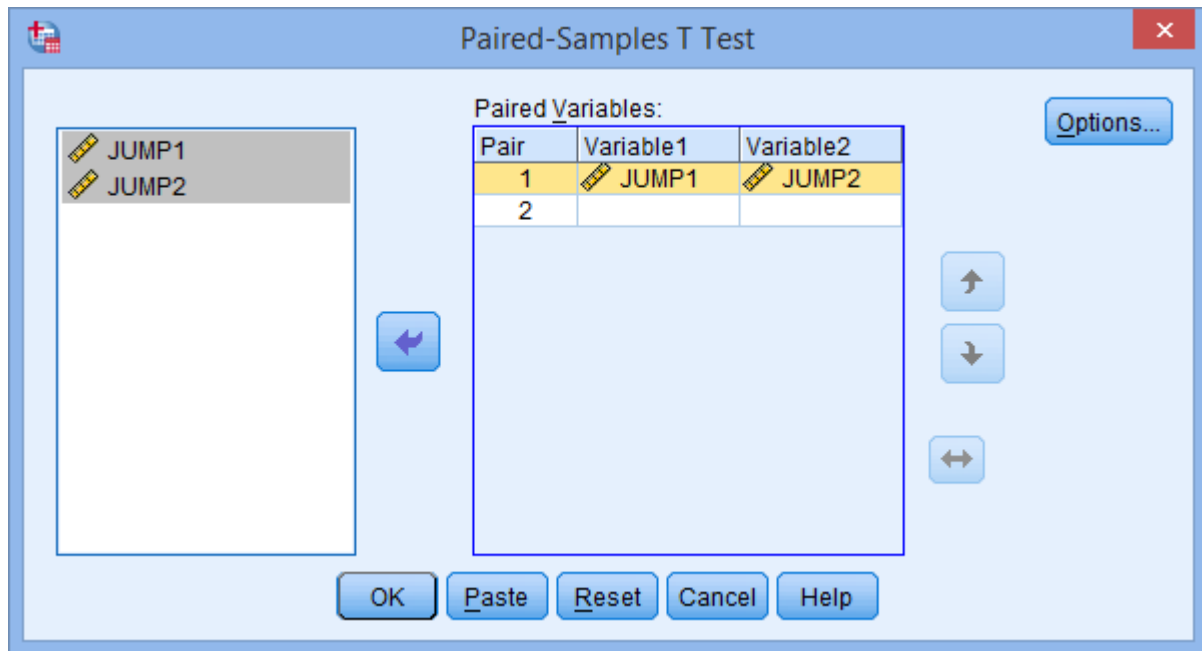


- You will be presented with the **Paired-Samples T Test** dialogue box, as shown below:






Published with written permission from SPSS Statistics, IBM Corporation.

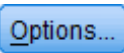
- Transfer the variables **JUMP1** and **JUMP2** into the Paired Variables: box. There are two ways to do this: (a) click on both variables whilst holding down the shift key (which highlights them) and then pressing the  button; or (b) drag-and-drop each variable separately into the boxes. If you are using older versions of SPSS Statistics, you will need to transfer the variables using the former method. You will end up with a screen similar to the one shown below:

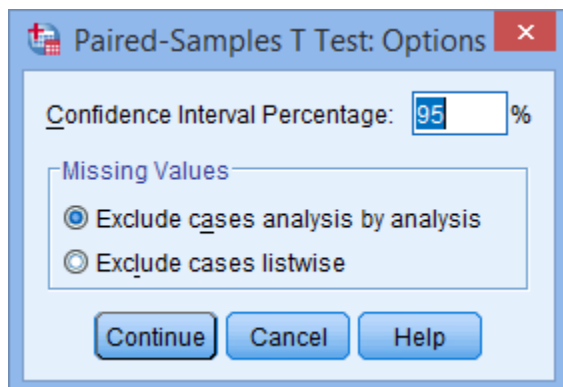



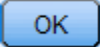
Published with written permission from SPSS Statistics, IBM Corporation.

Note:

-  button shifts the pair of variables you have highlighted down one level.
-  button shifts the pair of variables you have highlighted up one level.
-  button shifts the order of the variables within a variable pair.

- If you need to change the confidence level limits or exclude cases, click on the  button. You will be presented with the **Paired-Samples T Test: Options** dialogue box, as shown below:



- Click the  button. You will be returned to the **Paired-Samples T Test** dialogue box.
- Click the  button.

Join the 1,000s of students, academics and professionals who rely on Laerd Statistics. [TAKE THE TOUR](#) [PLANS & PRICING](#)

SPSS Statistics^{top} ^

Output of the Dependent T-Test in SPSS Statistics

SPSS Statistics generates three tables in the **Output Viewer** under the title "T-Test", but you only need to look at two tables: the **Paired Samples Statistics** table and the **Paired Samples Test** table. In addition, you will need to interpret the boxplots that you created to check for outliers and the output from the Shapiro-Wilk test for normality, which you used to determine whether the distribution of the differences in the dependent variable between the two related groups were approximately normally distributed. This is explained in our enhanced guide. However, in this "quick start" guide, we focus on the two main tables you need to understand if your data has met all the necessary assumptions:

Paired Sample Statistics Table

The first table, titled **Paired Samples Statistics**, is where SPSS Statistics has generated descriptive statistics for your variables. You could use the results here to describe the characteristics of the first and second jumps in your write-up.

	Mean	N	Std. Deviation	Std. Error Mean
Pair 1 JUMP1	2.4815	20	.16135	.03608
JUMP2	2.5155	20	.15982	.03574

Paired Samples Test Table

The **Paired Samples Test** table is where the results of the dependent t-test are presented. A lot of information is presented here and it is important to remember that this information refers to the **differences** between the two jumps (the subtitle reads "Paired Differences"). As such, the columns of the table labelled "**Mean**", "**Std. Deviation**", "**Std. Error Mean**" and "**95% Confidence Interval of the Difference**" refer to the mean difference between the two jumps and the standard deviation, standard error and 95% confidence interval of this mean difference, respectively. The last three columns express the results of the dependent t-test, namely the *t*-value ("**t**"), the degrees of freedom ("**df**") and the significance level ("**Sig. (2-tailed)**").

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	JUMP1 - JUMP2	-.03400	.03185	.00712	-.04891	-.01909	-4.773	19	.000

Published with written permission from SPSS Statistics, IBM Corporation.

You are essentially conducting a [one-sample t-test](#) on the differences between the groups.

SPSS Statistics^{top ^}

Reporting the Output of the Dependent T-Test

You might report the statistics in the following format: $t(\text{degrees of freedom}) = t\text{-value}, p = \text{significance level}$. In our case this would be: $t(19) = -4.773, p < 0.0005$. Due to the means of the two jumps and the direction of the *t*-value, we can conclude that there was a statistically significant improvement in jump distance following the plyometric-training programme from 2.48 ± 0.16 m to 2.52 ± 0.16 m ($p < 0.0005$); an improvement of 0.03 ± 0.03 m.

Note: SPSS Statistics can output the results to many decimal places, but you should understand your measurement scale in order to know whether it is appropriate to report your results with such precision.

In our enhanced dependent t-test guide, we show you how to write up the results from your assumptions tests and dependent t-test procedure if you need to report this in a dissertation/thesis, assignment or research report. We do this using the Harvard and APA styles. It is also worth noting that in addition to reporting the results from your assumptions

and dependent t-test, you are increasingly expected to report effect sizes. Whilst there are many different ways you can do this, we show you how to calculate effect sizes from your SPSS Statistics results in our enhanced dependent t-test guide. Effect sizes are important because whilst the dependent t-test tells you whether differences between group means are "real" (i.e., different in the population), it does not tell you the "size" of the difference. Providing the effect size in your results helps to overcome this limitation. You can learn more about our enhanced content [here](#).

Join the 1,000s of students, academics and professionals who rely on Laerd Statistics. **TAKE THE TOUR PLANS & PRICING**

Measures of Dispersion

Definitions:

• Dispersion or Variation:

The degree to which numerical data tend to spread about an average value. Various measures of dispersion or variation are available:

Range, Mean Deviation, Semi-interquartile range, 10-90 percentile range, and the Standard Deviation.

• Mean Deviation or (Average Deviation):

$$\text{Mean Deviation} = M.D = \frac{\sum_{j=1}^N |X_j - \bar{X}|}{N}$$

where $N = X_1, X_2, \dots, X_N$

\bar{X} = Arithmetic Mean

$$\bar{X} = \frac{2 + 3 + 6 + 8 + 11}{5} = \frac{30}{5} = 6$$

Ex: Find M.D of the set 2, 3, 6, 8, 11

Sol.:

$$M.D = \frac{|2 - 6| + |3 - 6| + |6 - 6| + |8 - 6| + |11 - 6|}{5} = 2.8$$

•**Note:** If X_1, X_2, \dots, X_k occur with frequencies f_1, f_2, \dots, f_k respectively, then

$$M.D = \frac{\sum_{j=1}^k f_j |X_j - \bar{X}|}{N} = \frac{\sum f |X - \bar{X}|}{N}$$

N= Total frequency

Since the mean deviation shown above is the “Mean Deviation about the Median”, so it is more appropriate to use the terminology, “Mean Absolute Deviation” rather than “Mean Deviation”.

Ex: Find the M.D or M.A.D of the average ages for a number of persons, which defined in the following table”

Average Age (yrs) Number of Persons (Freq)

45	3
20	10
30	5
60	<u>7</u>
	N= 25

Sol.:

$$\bar{X} = \frac{3(45) + 10(20) + 30(5) + 7(60)}{25} = 36.2 \cong 36$$

$$M.D = \frac{(3)|45 - 36| + (10)|20 - 36| + (5)|30 - 36| + (7)|60 - 36| + |11 - 6|}{25} = 15.4$$

Note: There is also a “Mean Deviation about the Median”, so it is more appropriate to use the terminology “Mean Absolute Deviation” rather than “Mean Deviation”.

•**The Standard Deviation:**

For a set of N numbers X_1, X_2, \dots, X_N the Standard Deviation is denoted by s and is defined by:

$$s = \sqrt{\frac{\sum_{j=1}^N (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

So s is called “Root Mean Square Deviation”.

If X_1, X_2, \dots, X_k occur with frequencies f_1, f_2, \dots, f_k respectively, the standard deviation is equal to:

$$s = \sqrt{\frac{\sum_{j=1}^N f_j (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum f (X - \bar{X})^2}{N}}$$

Where $N = \sum f$ total frequency

Ex: Find the standard deviation of the weights of 100 product boxes of the following data:

<u>Mass (kg)</u>	<u>Class Mark</u>	<u>Frequency (f)</u>
60-62	61	5
63-65	64	18
66-68	67	42
69-71	70	27
72-74	73	<u>8</u>
		N= 100

Sol.:

$$\bar{X} = \frac{(5)(61) + (18)(64) + (42)(67) + (27)(70) + (8)(73)}{100} = 67.45 \text{ kg} = \text{mean}$$

$$s = \sqrt{\frac{(5)(61 - 67.5)^2 + (18)(64 - 67)^2 + (42)(67 - 67.5)^2 + (27)(70 - 67.5)^2 + (8)(73 - 67.5)^2}{100}}$$

$$= 2.92 \text{ kg}$$

•**Variance:**

It is the Square of standard deviation,

$$\text{Variance} = s^2$$

S^2 = variance of sample

σ^2 = variance of population

•**Note:** Semi-interquartile Range = 2/3 (Standard Deviation)

•**Note:** M.D = 4/5 (Standard Deviation)

• **Coefficient of Variation (V):**

It is a relative dispersion which is also called “Coefficient of Dispersion” given by:

$$\frac{s}{\bar{X}} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

•**Standardized Variable (Standard Score):**

It is denoted by the variable

$z = \frac{X - \bar{X}}{s}$ which measures the deviation from the mean in units of the standard deviation.

Ex: A student received 84 marks in Math for which the mean mark = 76 and the standard deviation = 10.

In Physics he received 90 for which the mean mark = 82 and the standard deviation = 16. In which subject was his relative standing higher?

Sol:

The standardized variable z measures the deviation of X from the mean \bar{X} in terms of standard deviation s .

$$\text{For Math } z = (84 - 76) / 10 = 0.8$$

$$\text{For Physics } z = (90 - 82) / 16 = 0.5$$

Thus his relative standing is higher in Math, which it is mean that he has a grade of 0.8 of standard deviation above the mean in Math and he has a grade of 0.5 of standard deviation above the mean in Physics.

Moments, Skewness, and Kurtosis:

•Moments:

For X_1, X_2, \dots, X_N we define the r^{th} moment by:

$$\overline{X^r} = \frac{X_1^r + X_2^r + \dots + X_N^r}{N} = \frac{\sum_{j=1}^N X_j^r}{N} = \frac{\sum X^r}{N}$$

So, the 1st moment with $r = 1$ is the arithmetic mean \bar{X} .

•The r th moment about the mean \bar{X} is defined as:

$$m_r = \frac{\sum_{j=1}^N (X_j - \bar{X})^r}{N} = \frac{\sum (X - \bar{X})^r}{N} = \overline{(X - \bar{X})^r}$$

* The r th moment about any origin A is defined as

$$m_r = \frac{\sum (X - A)^r}{N} = \overline{(X - A)^r}$$

Ex: Find the 1st, 2nd, 3rd and 4th moments of the set of numbers 2, 3, 7, 8, 9, 10.

Sol:

$$\bar{X} = \frac{\sum X}{N} = \frac{2+3+7+8+10}{5} = 6 = \text{1st moment mean}$$

$$\overline{X^2} = \frac{\sum X^2}{N} = \frac{2^2+3^2+7^2+8^2+10^2}{5} = 45.2 = \text{2nd moment}$$

Ex: Find 1st, 2nd, 3rd, and 4th moments about mean for the set of above numbers.

Sol:

$$m_1 = (X - \bar{X}) = [(2-6)+(3-6)+(7-6)+(8-6)+(10-6)]/5 = 0$$

So the 1st moment (m_1) about mean = 0 always

$$m_2 = \frac{\sum(X - \bar{X})^2}{N} = 9.2$$

Since

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{N}}$$

So that the $m_2 = s^2 = \text{variance}$

Ex: Find the 1st four moments about the mean of the mass distribution of the following

<u>Mass (kg)</u>	<u>f</u>
60-62	5
63-65	18
66-68	42
69-71	27
72-74	<u>8</u>

N= total f = 100

Sol:

1. Find the X = class mark 61 , 64 , 67 , 70 , 73
2. Determine the Mean (Arithmetic mean) by applying

$$\bar{X} = \frac{\sum fX}{N}$$

3. Find m_1 , m_2 , m_3 , and m_4 by applying the following equation

$$m_r = \frac{\sum f(X - \bar{X})^r}{N} \quad \text{then } m_1 = 0, m_2 = 8.5275, m_3 = 2.6932, \text{ and } m_4 = 199.3759$$

Skewness:

It is the degree of asymmetry, or departure from symmetry.

- If the frequency curve of a distribution has a longer “tail” to the right of the central maximum than to the left, the distribution is said to be “skewed to the right” or “positive skewed”.

figure

- If the reverse is true, is said to be “skewed to the left” or to have “negative skewed”.

figure

$$\text{Skewness} = \frac{\text{mean} - \text{mode}}{\text{standard deviation}} = \frac{\bar{X} - \text{mode}}{s}$$

Also, by using the empirical formula:

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{s} = \frac{3(\bar{X} - \text{median})}{s}$$

Note the above two measures are called the
“**First and Second coefficients of skewness**”

Ex: Find the a) first and b) second coefficient of skewness for the distribution of

Mean = 79.76 , Median = 79.06

Mode = 77.5 , $s = 15.6$

Sol:

1st coefficient of skewness = $(79.76 - 77.5)/15.6 = 0.1448$

2nd coefficient of skewness = $3(79.76 - 79.06)/15.6 = 0.1346$

Since the coefficients are positive value, the distribution is skewed positively to the right.

•**Note** An important measures of skewness uses the 3rd moment about the mean, which is given by:

Moment coefficient of skewness = $a_3 = m_3 / s^3$ and since $s^2 = m_2$

so $s = \sqrt{m_2}$

So $a_3 = \frac{m_3}{(\sqrt{m_2})^3} = \frac{m_3}{\sqrt{m_2^3}}$

***Note**

For the normal distribution $a_3 = 0$.

•Kurtosis

It is the degree of **Peakdness** of a normal distribution.

With High Peak called Leptokurtic
figure

Flat-topped called Platykurtic
figure

Not very peaked or very flat topped called Mesokurtic
figure

Note one measures of kurtosis uses the “4th moment about the mean” given by:

Moment Coefficient of Kurtosis = a_4

$$a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} \quad \text{since} \quad s^2 = m_2$$

Ex: If the 2nd and 4th moments about the mean are equal to 199.3759, 8.5275 respectively, find the moment coefficient of kurtosis.

Sol:

$$a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} = 199.375/(8.5275)^2 = 2.74$$

* For Normal Distribution $a_4 = 3$.

TWO SAMPLE t-Test

A Review of the t-test

The t-test is used for testing differences between two means. In order to use a t-test, the **same variable** must be measured in different groups, at different times, or in comparison to a known population mean. Comparing a sample mean to a known population is an unusual test that appears in statistics books as a transitional step in learning about the t-test. The more common applications of the t-test are testing the difference between independent groups or testing the difference between dependent groups.

A t-test for independent groups is useful when the same variable has been measured in two independent groups and the researcher wants to know whether the difference between group means is statistically significant. "Independent groups" means that the groups have different people in them and that the people in the different groups have not been matched or paired in any way. A t-test for related samples or a t-test for dependent means is the appropriate test when the same people have been measured or tested under two different conditions or when people are put into pairs by matching them on some other variable and then placing each member of the pair into one of two groups.

The t-test For Independent Groups on SPSS

A t-test for independent groups is useful when the researcher's goal is to compare the difference between means of two groups on the same variable. Groups may be formed in two different ways. First, a preexisting characteristic of the participants may be used to divide them into groups. For example, the researcher may wish to compare college GPAs of men and women. In this case, the **grouping variable** is biological sex and the two groups would consist of men versus women. Other preexisting characteristics that could be used as grouping variables include age (under 21 years vs. 21 years and older or some other reasonable division into two groups), athlete (plays collegiate varsity sport vs. does not play), type of student (undergraduate vs. graduate student), type of faculty member (tenured vs. nontenured), or any other variable for which it makes sense to have two categories. Another way to form groups is to randomly assign participants to one of two experimental conditions such as a group that listens to music versus a group that experiences a control condition.

Regardless of how the groups are determined, one of the variables in the SPSS data file must contain the information needed to divide participants into the appropriate groups. SPSS has very flexible features for accomplishing this task.

Like all other statistical tests using SPSS, the process begins with data. Consider the fictional data on college GPA and weekly hours of studying used in the correlation example. First, let's add information about the biological sex of each participant to the data base. This requires a numerical code. For this example, let a "1" designate a female and a "2" designate a male. With the new variable added, the data would look like this:

Participant	Current GPA	Weekly Study Time	Sex
Participant #01	1.8	15 hrs	2
Participant #02	3.9	38 hrs	1
Participant #03	2.1	10 hrs	2
Participant #04	2.8	24 hrs	1
Participant #05	3.3	36 hrs	.
Participant #06	3.1	15 hrs	2
Participant #07	4.0	45 hrs	1
Participant #08	3.4	28 hrs	1
Participant #09	3.3	35 hrs	1
Participant #10	2.2	10 hrs	2
Participant #11	2.5	6 hrs	2

With this information added to the file, two methods of dividing participants into groups can be illustrated. Note that Participant #05 has just a single dot in the column for sex. This is the standard way that SPSS indicates missing data. This is a common

occurrence, especially in survey data, and SPSS has flexible options for handling this situation. Begin the analysis by entering the new data for sex. Use the arrow keys or mouse to move to the empty third column on the spreadsheet. Use the same technique as previously to enter the new data. When data is missing (such as Participant #5 in this example), hit the <ENTER> key when there is no data in the top line (you will need to <DELETE> the previous entry) and a single dot will appear in the variable column. Once the data is entered, click **Data > Define Variable** and type in the name of the variable, "Sex." Then go to "value" And type a "1" in the box. For "Value Label," type "Female." Then click on **ADD**. Repeat the sequence, typing "2" and "male" in the appropriate boxes. Then click **ADD** again. Finally, click **CONTINUE >OK** and you will be back to the main SPSS menu.

[Back to the Top of the Page](#)

To request the t-test, click **Statistics > Compare Means > Independent Samples T Test**. Use the right-pointing arrow to transfer COLGPA to the "Test Variable(s)" box. Then highlight Sex in the left box and click the bottom arrow (pointing right) to transfer sex to the "Grouping Variable" box. Then click **Define Groups**. Type "1" in the Group 1 box and type "2" in the Group 2 box. Then click **Continue**. Click **Options** and you will see the confidence interval or the method of handling missing data can be changed. Since the default options are just fine, click **Continue > OK** and the results will quickly appear in the output window. Results for the example are shown below:

T-Test

Group Statistics

	Variable	N	Mean	Std. Deviation	Std. Error Mean
SEX	1.00 Female	5	3.4800	.487	.218
	2.00 Male	5	2.3400	.493	.220

Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
SEX	Equal variances assumed	.002	.962

	Equal Variances not assumed		
--	-----------------------------	--	--

		t-test for Equality of Means			
		t	df	Sig. (2-tailed)	Mean Difference
SEX	Equal variances assumed	3.68	8	.021	.1750
	Equal variances not assumed	3.68	8.00	.025	.1750

The output begins with the means and standard deviations for the two variables which is key information that will need to be included in any related research report. The "Mean Difference" statistic indicates the magnitude of the difference between means. When combined with the confidence interval for the difference, this information can make a valuable contribution to explaining the importance of the results. "Levene's Test for Equality of Variances" is a test of the homogeneity of variance assumption. When the value for F is large and the P -value is less than .05, this indicates that the variances are heterogeneous which violates a key assumption of the t-test. The next section of the output provides the actual t-test results in two formats. The first format for "Equal" variances is the standard t-test taught in introductory statistics. This is the test result that should be reported in a research report under most circumstances. The second format reports a t-test for "Unequal" variances. This is an alternative way of computing the t-test that accounts for heterogeneous variances and provides an accurate result even when the homogeneity assumption has been violated (as indicated by the Levene test). It is rare that one needs to consider using the "Unequal" variances format because, under most circumstances, even when the homogeneity assumption is violated, the results are practically indistinguishable. When the "Equal" variances and "Unequal" variances formats lead to different conclusions, seek consultation. The output for both formats shows the degrees of freedom (df) and probability (2-tailed significance). As in all statistical tests, the basic criterion for statistical significance is a "2-tailed significance" less than .05. The .021 probability in this example is clearly less than .05 so the difference is statistically significant.

A second method of performing an independent groups t-test with SPSS is to use a noncategorical variable to divide the test variable (college GPA in this example) into

groups. For example, the group of participants could be divided into two groups by placing those with a high number of study hours per week in one group and a low number of study hours in the second group. Note that this approach would begin with exactly the same information that was used in the correlation example. However, converting the Studyhrs data to a categorical variable would cause some detailed information to be lost. For this reason, caution (and consultation) is needed before using this method. To request the analysis, click **Statistics > Compare Means > Independent Samples T Test...** Colgpa will remain the "Test Variable(s)" so it can be left where it is. Alternately, other variables can be moved into this box. Click "Sex(1,2)" to highlight it and remove it from the "Grouping Variable" box by clicking the bottom arrow which now faces left because a variable in the box has been highlighted. Next, highlight "Studyhrs" and move it into the "Grouping Variable" box. Now click **Define Groups...** and click the **Cut point** button. Enter a value (20 in this case) into the box. All participants with values less than the cutpoint will be in one group and participants with values greater than or equal to the cutpoint will form the other group. Click **Continue > OK** and the output will quickly appear. The results from the example are shown below:

Group Statistics

	Studyhours	N	Mean	Std.Deviation	Std. Error Mean
COLGPA College GPA for Fall 1997	Studyhours >= 20.00	6	3.4500	.4416	.1803
	Studyhours < 20.00	5	2.3400	.4930	.2205

The "Group Statistics" table provides the means and standard deviations along with precise information regarding the formation of the groups. This can be very useful as a check to ensure that the cutpoint was selected properly and resulted in reasonably similar sample sizes for both groups. The remainder of the output is virtually the same as the previous example.

[Back to the Top of the Page](#)

The t-test For Dependent Groups on SPSS

The t-test for **dependent** groups requires a different way of approaching the data. For this type of test, each case is assumed to have two measures of the same variable taken at different times. Each "Case" would therefore consist of a **single person**. This would be what is called a repeated measures design. Alternately, each case could contain the same information about two **different** individuals who have been paired or matched on a variable. In the repeated measures situation, one might collect GPA information at two different points in the careers of a group of students. The table below shows how this situation might appear in the fictional example. In this case, GPA data have been collected at the end of each participant's first year (Colgpa1) and senior year (Colgpa2).

Participant	Colgpa1	Weekly Study Time	Sex	Colgpa2
Participant #01	1.8	15 hrs	2	.
Participant #02	3.9	38 hrs	1	3.88
Participant #03	2.1	10 hrs	2	2.80
Participant #04	2.8	24 hrs	1	3.20
Participant #05	3.3	36 hrs	.	3.60
Participant #06	3.1	15 hrs	2	3.57
Participant #07	4.0	45 hrs	1	4.00
Participant #08	3.4	28 hrs	1	3.35
Participant #09	3.3	35 hrs	1	3.66
Participant #10	2.2	10 hrs	2	2.55
Participant #11	2.5	6 hrs	2	2.67

One thing to note about the new data is that the GPA of the first participant is missing. Given the 1.8 GPA at the first assessment, it seemed reasonable that this person might not remain in college for the entire four years. This is a common hazard of repeated measures designs and the implication of such missing data needs to be considered before interpreting the results.

To request the analysis, click **Statistics > Compare Means > Paired-Samples T Test ...** A window will appear with a list of variables on the left and a box labeled "Paired Variables" on the right. Highlight **two** variables (Colgpa and Colgpa2, in this example) and transfer them to the "Paired Variables" box by clicking the right-

pointing arrow between the boxes. Several pairs of variables can be entered at this time. The **Options...** button opens a window that allows control of the confidence interval and missing data options. Click **Continue** (if you opened the **Options...** window) > **OK** to complete the analysis. The output will appear in an Output window. Results for the example problem are shown below:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Colgpa1	3.0600	10	.6552	.2072
	Colgpa2	3.3280	10	.5091	.1610

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Colgpa1 - Colgpa2	10	.944	.000

Paired Samples Test

		Paired Differences					t
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		
					Lower	Upper	
Pair 1	Colgpa1 - Colgpa2	-.2680	.2419	7.649E-02	.4410	-9.50E-02	-3.504

Paired Samples Test

	df	Sig. (2-tailed)
Pair 1 Colgpa1 - Colgpa2	9	.007

The output is similar to the independent groups t-test. The first table of the output shows the means and standard deviations for the two groups and the second table shows the correlation between the paired variables. The next table shows the mean of the differences, standard deviation of the differences, standard error of the mean, the confidence interval for the difference, and the obtained value for t. The 2-tailed Sig[nificance] which is stated as a probability is shown in the last table. As usual, probabilities **less than** .05 indicate that the null hypothesis should be rejected. In this case, the interpretation would be that GPA increased significantly from firstyear to senior year, $t(9) = 3.50$, $p = .007$.

[Back to the Top of the Page](#)



Copyright 2000 The McGraw-Hill Companies. All rights reserved. Any use is subject to the [Terms of Use](#) and [Privacy Policy](#).
McGraw-Hill Higher Education is one of the many fine businesses of [The McGraw-Hill Companies](#).

If you have a question or a problem about a specific book or product, please fill out our [Product Feedback Form](#).

For further information about this site contact mhhe_webmaster@mcgraw-hill.com or let us know what you think by filling out our [Site Survey](#).



Preliminary communication
(accepted May 5, 2018)

TIME SERIES FORECASTING USING A MOVING AVERAGE MODEL FOR EXTRAPOLATION OF NUMBER OF TOURIST

Zoran Ivanovski¹
Ace Milenkovski
Zoran Narasanov

Abstract

Time series is a collection of observations made at regular time intervals and its analysis refers to problems in correlations among successive observations. Time series analysis is applied in all areas of statistics but some of the most important include macroeconomic and financial time series. In this paper we are testing forecasting capacity of the time series analysis to predict tourists' trends and indicators. We found evidence that the time series models provide accurate extrapolation of the number of guests, quarterly for one year in advance. This is important for appropriate planning for all stakeholders in the tourist sector. Research results confirm that moving average model for time series data provide accurate forecasting the number of tourist guests for the next year.

Keywords: seasonality, trend, regression, forecasting, centered moving average.

Jel Classification: C3; C32

INTRODUCTION

Contemporary business has to deal with uncertainty and operations in terms of complex and dynamic changes of business environment. Companies and public stakeholders use planning function in order to organize their work and provide necessary assets for the future periods. One of the most important parts of planning function is financial planning, as a continuous process of directing and allocating financial resources to meet companies' strategic goals and objectives. Financial planning usually starts with planning company's sales (revenues) that will be realized in next period.

¹ **Zoran Ivanovski**, PhD, Full Professor; **Ace Milenkovski**, PhD, Full Professor, University of Tourism and Management in Skopje; **Zoran Narasanov**, PhD, Assistant Professor, Winner Insurance, Vienna Insurance Group, Macedonia.

In tourist industry forecasting future revenues depends on capability to make prediction of number of tourists, having in mind that there are available statistics' data for average tourist spending per day. When all stakeholders in tourist industry have relevant and accurate forecasts of trends and number of tourists that will visit the country, they can provide reliable input for their financial planning. In order to do that it is important to use quantitative methods and statistics tools that can help to deal with past data where seasonality, irregularity and sometimes random variables appears.

There are many comprehensive theoretical studies in literature that suggest the necessity to use quantitative models and techniques for various types of data. In this paper we focus on a time series analysis. A time series is a collection of observations made sequentially in time. Examples are daily mortality counts, particulate air pollution measurements, and temperature data. Time series analysis refers to problems in which observations are collected at regular time intervals and there are correlations among successive observations as explained by Bartholomew (1971), Janacek (2009) and many other authors. Applications cover virtually all areas of statistics but some of the most important include macroeconomic time series as analyzed by Nelson and Plosser (1982), like monthly data for unemployment, hospital admissions, etc., in finance (eg. daily exchange rate, a share price, etc.), environmental (e.g., daily rainfall, air quality readings); medicine (e.g., ECG brain wave activity every 2^{-8} seconds).

Time series analysis include autoregressive and moving average processes and some discussion of the effect of time series correlations on the other kinds of statistical inference, such as the estimation of means and regression coefficients. The methods of time series analysis pre-date those for general stochastic processes and Markov Chains. The aims of time series analysis are to describe and summarize time series data, fit low-dimensional models, and make forecasts. One simple method of describing a series is a classical decomposition into four elements: Trend (Tt) — long term movements in the mean; Seasonal effects (It) — cyclical fluctuations related to the calendar; Cycles (Ct) — other cyclical fluctuations (such as a business cycles); Residuals (Et) — other random or systematic fluctuations. The idea is to create separate models for these four elements and then combine them, either additively:

$$X_t = T_t + I_t + C_t + E_t \quad (1)$$

or multiplicatively

$$X_t = T_t \cdot I_t \cdot C_t \cdot E_t \quad (2)$$

One of the most popular and frequently used stochastic series models is the Autoregressive Integrated Moving Average model (ARIMA) elaborated by many authors like Chatfield (1996), Box and Jenkins (1971), Cochrane (1997), Cottrell et al. (1995). One of the most used ARIMA model is Moving Average Model introduced by Armstrong (2006).

In this paper we are testing the ability and accuracy of application of the time series analysis model for prediction of some important tourist trends and indicators. Many authors have analyzed accuracy of time series analysis like Fildes and Makridakis (1995), Chernick (1994), Montgomery, Jennings and Kulahci (2008). Gardner and Mckenzie (1985) in their paper argue that forecast accuracy can be improved by either damping or ignoring altogether trends which have a low probability of persistence. We

are testing research hypothesis that moving average model for the time series data provide accurate forecasting of the number of tourist guests in next year.

The remainder of this paper is structured into three sections. In Section 1, we present a theoretical review and methodology. Section 2 presents research findings from a time series analysis using moving average model and last section summarizes the main conclusions.

1. LITERATURE REVIEW AND METHODOLOGY

There are many methods available for forecasting, so it is important to know their applicability and reliability to make appropriate selection before using in specific situation. Time series modeling is a very popular tool among researchers and practitioners for providing accurate forecast.

In fact, main task of time series modeling is to analyze past observations in order to develop an appropriate model for the specific data series and to use this as a model for forecasting future values for the series. A time series is a set of observations on the values that a variable takes at different times, collected at regular time intervals, such as monthly (eg. CPI), weekly (eg. money supply), quarterly (eg. GDP) or annually (eg. State budget). Time series models are used for the modelling and prediction of data collected sequentially in time in order to provide specific techniques for handling data (Brockwell and Davis, 2013) and Diggle (1990).

Depending on the frequency of the data (hourly, daily, weekly, monthly, quarterly, annually, etc) different patterns emerge in the data set which forms the component to be modeled. Sometimes the time series may just be increasing or decreasing over time with a constant slope or they may be patterns around the increasing slope (Bollerslev 1986). The pattern in a time series is sometimes classified into trend, seasonal, cyclical and random components. We can define trend as a long term relatively smooth pattern that usually persist for more than one year. Seasonal is a pattern that appears in a regular interval wherein the frequency of occurrence is within a year or even shorter, as it can be monthly arrival of tourist to a country. Cyclical means the repeated pattern that appears in a time-series but beyond a frequency of one year. It is a wavelike pattern about a long-term trend that is apparent over a number of years. Cycles are rarely regular and appear in combination with other components (eg. business cycles that record periods of economic recession and inflation, cycles in the monetary and financial sectors, etc.). Random is the component of a time-series that is obtained after these three patterns have been “extracted” out of the series. When we plot the residual series we can indicate a random pattern around a mean value.

Time series modeling is a dynamic research area which has become popular not only in academic circles but also widely used by business sector over last few decades. The main aim of time series modeling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model then can be used to generate future values for the series, i.e. to make forecasts as Adhikari and Agrawal argue in their paper (2013). Idea for univariate time series modeling based on situations where an appropriate economic theory to the relationship between series may not be available and hence we consider only the statistical relationship of the given series with its past values.

Sometimes even when the set of explanatory variables may be known it may not be possible to obtain the entire set of such variables required to estimate a regression model and we can use only a single series of the dependent variable to forecast the future values.

There are different time series processes:

- White Noise: A series is called white noise if it is purely random in nature. Let $\{\varepsilon_t\}$ denote such a series then it has zero mean [$E(\varepsilon_t)=0$], has a constant variance [$V(\varepsilon_t)=\sigma^2$] and is an uncorrelated [$E(\varepsilon_t \varepsilon_s)=0$] random variable. The plot of such series across time will indicate no pattern and hence forecasting the future values of such series is not possible.
- Auto Regressive Model: An AR model is one in which Y_t depends only on its own past values $Y_{t-1}, Y_{t-2}, Y_{t-3}$, etc. Thus:

$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, \dots, \varepsilon_t). \quad (3)$$

A common representation of an autoregressive model where it depends on p of its past values called AR(p) model is:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \dots + \beta_p Y_{t-p} + \varepsilon_t \quad (4)$$

- Moving Average Model – is one when Y_t depends only on the random error terms which follow a white noise process:

$$Y_t = f\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \dots, \varepsilon_{t-n}\}. \quad (5)$$

A common representation of a moving average model where it depends on a q of its past values is called MA(q) model and represented below:

$$Y_t = \beta_0 + \varepsilon_t + \phi \varepsilon_{t-1} + \phi \varepsilon_{t-2} + \phi \varepsilon_{t-3} + \dots + \phi_q \varepsilon_{t-q} \quad (6)$$

The error terms ε_t are assumed to be white noise processes with mean zero and variance σ^2 .

The simplest time series model is certainly the white noise. We use in our research MA(q) processes to linear processes and time series analysis of quarterly data for number of tourists in the Republic of Macedonia using actual past data for five years period (2012–2016) in order to provide extrapolation of number of guest that will visit the Republic of Macedonia in next year. In contrast to modeling in terms of mathematical equation, the moving average merely smooths the fluctuations in the data. A moving average model works well when the data have a fairly linear trend and a definite rhythmic pattern of fluctuations.

2. RESEARCH FINDINGS

We analyze quarterly data for total number of guests for the five years period from 2012–2016 that have visited the Republic of Macedonia. Official state statistics regularly provide monthly data for number of guests in the country. We summarize official data and calculate quarterly data for actual number of guests that have visited the Republic of Macedonia in five years period (2012–2016) as shown on next Table:

Table 1. Quarterly data for number of tourists in the Republic of Macedonia 2012–2016

Year	Quarter	Num. of guests
2012	Q ₁	89.383
	Q ₂	164.691
	Q ₃	291.895
	Q ₄	117.664
2013	Q ₁	91.858
	Q ₂	177.790
	Q ₃	303.054
	Q ₄	129.092
2014	Q ₁	94.404
	Q ₂	191.570
	Q ₃	318.709
	Q ₄	130.967
2015	Q ₁	98.709
	Q ₂	205.766
	Q ₃	367.811
	Q ₄	143.781
2016	Q ₁	115.222
	Q ₂	215.509
	Q ₃	370.437
	Q ₄	155.675

Source: State Statistical Office

The task of our research is to use a time series models for extrapolation number of guests, quarterly for year 2017. Forecasting number of guests enables appropriate planning in tourist sector (government, ministries, agencies, tour-operators, travel agencies, hotels and other relevant stakeholders).

In order to visualize data series, we plot them and create line chart:

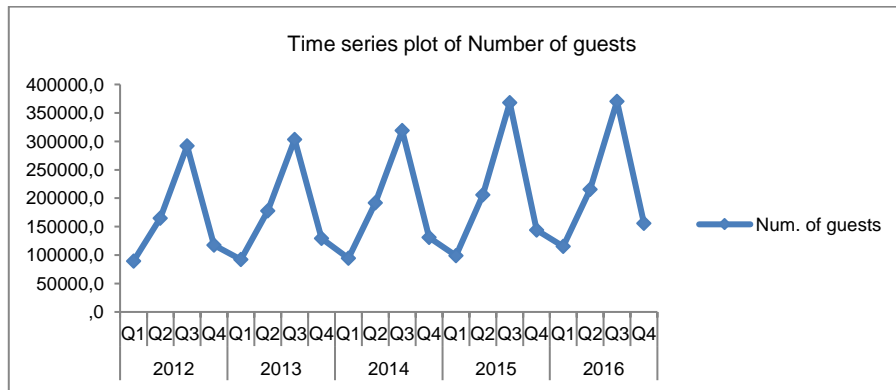


Figure 1. Time Series of number of tourist in the Republic of Macedonia 2012–2016

Analysis usually starts with charts where it is easier to visualize movement and behaviour of actual numbers. We can see from the chart that there is “pattern” that commonly repeats itself almost every year. There are “depths” and a “climaxes” in every period (years), in fact cycles repeats itself every year and this is clear example of seasonality. Beside this “up and down” movements, we can conclude that the overall direction of this plot is increasing, and that is trend component of actual data.

This mean that trend was clearly identified as well as the seasonal components in the actual numbers of the guests that have visited the Republic of Macedonia. Beside trend and seasonal component we need to check “irregularity” of actual data. The final component in the model that we are using here is going to be irregular component. That actually means that quarter to quarter variations of number of guests that exist does not follow any pattern. We can conclude that actual numbers have irregular aspect, also called “random aspect”. That component is always present in data, no matter whether they are time series or not. So we have to deal with variability of time series data. This is usually a movement of variables that do not have predictable pattern in real world. In fact we need to deal with all three components: trend, seasonality and irregularity. In order to use time series analysis, next step is to “smooth out” our data and for that we will use moving averages above four periods MA(4). It is necessity to use MA(4) because we have identified cycles with four periods (quarters).

Moving averages calculation MA(4) are presented in final right column on the following Table 2:

Table 2. Moving averages calculation MA(4)

T	Year	Quarter	Num. of guests	MA (4)
1	2012	Q ₁	89.383	
2		Q ₂	164.691	
3		Q ₃	291.895	165.908
4		Q ₄	117.664	166.527
5	2013	Q ₁	91.858	169.802
6		Q ₂	177.790	172.592
7		Q ₃	303.054	175.449
8		Q ₄	129.092	176.085
9	2014	Q ₁	94.404	179.530
10		Q ₂	191.570	183.444
11		Q ₃	318.709	183.913
12		Q ₄	130.967	184.989
13	2015	Q ₁	98.709	188.538
14		Q ₂	205.766	200.813
15		Q ₃	367.811	204.017
16		Q ₄	143.781	208.145
17	2016	Q ₁	115.222	210.581
18		Q ₂	215.509	211.237
19		Q ₃	370.437	214.211
20		Q ₄	155.675	216.186

Next step of our time series data analysis is the centered moving averages calculation CMA(4). Results are presented and added to previous data in last right column, as shown on Table 3:

Table 3. Centered moving averages calculation CMA(4)

t	Year	Quarter	Num. of guests	MA (4)	CMA (4)
1	2012	Q ₁	89.383		
2		Q ₂	164.691		
3		Q ₃	291.895	165.908	166.218
4		Q ₄	117.664	166.527	168.164
5	2013	Q ₁	91.858	169.802	171.197
6		Q ₂	177.790	172.592	174.020
7		Q ₃	303.054	175.449	175.767
8		Q ₄	129.092	176.085	177.808
.					

Table 4. (continued)

t	Year	Quarter	Num. of guests	MA (4)	CMA (4)
9	2014	Q ₁	94.404	179.530	181.487
10		Q ₂	191.570	183.444	183.678
11		Q ₃	318.709	183.913	184.451
12		Q ₄	130.967	184.989	186.763
13	2015	Q ₁	98.709	188.538	194.676
14		Q ₂	205.766	200.813	202.415
15		Q ₃	367.811	204.017	206.081
16		Q ₄	143.781	208.145	209.363
17	2016	Q ₁	115.222	210.581	210.909
18		Q ₂	215.509	211.237	212.724
19		Q ₃	370.437	214.211	215.198
20		Q ₄	155.675	216.186	

This step was necessary to “smooth” the time series data and to “clean” the data from the seasonality and irregularity. Now we plot the CMA(4) data on the chart:

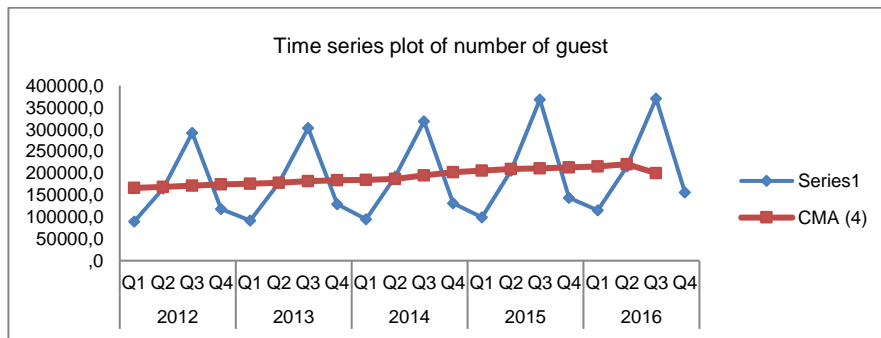


Figure 2. Time Series of number of tourist in the Republic of Macedonia 2012–2016

CMA(4) line is plotted on the chart with red color. This line in fact extract seasonal and irregular components and now we can clearly see the difference between original data and CMA(4) data. We proceed with our analysis and create new column S_t, I_t for seasonal and irregular components. This is necessary in order to “follow the model” requirements Table 4. As already explained in theoretical approach in first part of this paper, classical time series multiplicative model is:

$$Y_t = S_t \times I_t \times T_t \quad (7)$$

Table 5. S_t, I_t calculation

t	Year	Quarter	Num. of guests	MA (4)	CMA (4)	S_t, I_t
1	2012	Q ₁	89.383			
2		Q ₂	164.691			
3		Q ₃	291.895	165.908	166.218	1,76
4		Q ₄	117.664	166.527	168.164	0,70
5	2013	Q ₁	91.858	169.802	171.197	0,54
6		Q ₂	177.790	172.592	174.020	1,02
7		Q ₃	303.054	175.449	175.767	1,72
8		Q ₄	129.092	176.085	177.808	0,73
.						

Table 6. (continued)

t	Year	Quarter	Num. of guests	MA (4)	CMA (4)	$S_{t,l}$
9	2014	Q ₁	94.404	179.530	181.487	0,52
10		Q ₂	191.570	183.444	183.678	1,04
11		Q ₃	318.709	183.913	184.451	1,73
12		Q ₄	130.967	184.989	186.763	0,70
13	2015	Q ₁	98.709	188.538	194.676	0,51
14		Q ₂	205.766	200.813	202.415	1,02
15		Q ₃	367.811	204.017	206.081	1,78
16		Q ₄	143.781	208.145	209.363	0,69
17	2016	Q ₁	115.222	210.581	210.909	0,55
18		Q ₂	215.509	211.237	212.724	1,01
19		Q ₃	370.437	214.211	215.198	1,72
20		Q ₄	155.675	216.186	220.387	0,71

We calculate $S_{t,l}$, by dividing actual number of guest (Y_t) with CMA(4) data. In fact we have got coefficients that explains how much seasonal and irregular components are above or below “smooth” line of CMA(4). We continue our analysis with quantification of seasonal component S_t . In order to do that we calculate average of each seasonal irregular components (indexes) for each quarter in order to avoid irregularity. S_t calculated values are presented on the Table 5 shown below:

Table 7. Seasonal component S_t

Quarter	S_t
1	0,53
2	1,02
3	1,74
4	0,70

We add calculated data for S_t in Table 6 for entire length of our time series data:

Table 8. Time series analysis with S_t

t	Year	Quarter	Num. of guests	MA (4)	CMA (4)	$S_{t,l}$	S_t
1	2012	Q ₁	89.383				0,53
2		Q ₂	164.691				1,02
3		Q ₃	291.895	165.908	166.218	1,76	1,74
4		Q ₄	117.664	166.527	168.164	0,70	0,70
05	2013	Q ₁	91.858	169.802	171.197	0,54	0,53
6		Q ₂	177.790	172.592	174.020	1,02	1,02
7		Q ₃	303.054	175.449	175.767	1,72	1,74
8		Q ₄	129.092	176.085	177.808	0,73	0,70
9	2014	Q ₁	94.404	179.530	181.487	0,52	0,53
10		Q ₂	191.570	183.444	183.678	1,04	1,02
11		Q ₃	318.709	183.913	184.451	1,73	1,74
12		Q ₄	130.967	184.989	186.763	0,70	0,70
13	2015	Q ₁	98.709	188.538	194.676	0,51	0,53
14		Q ₂	205.766	200.813	202.415	1,02	1,02
15		Q ₃	367.811	204.017	206.081	1,78	1,74
16		Q ₄	143.781	208.145	209.363	0,69	0,70
17	2016	Q ₁	115.222	210.581	210.909	0,55	0,53
18		Q ₂	215.509	211.237	212.724	1,01	1,02
19		Q ₃	370.437	214.211	215.198	1,72	1,74
20		Q ₄	155.675	216.186	220.387	0,71	0,70

Our next step is to “clean” the data from seasonal component by dividing Y_t/S_t . Calculated values are presented in last column of the Table 7:

Table 9. Desseasonalized data

T	Year	Quarter	Num. of guests	MA (4)	CMA (4)	$S_{t,t}$	S_t	Desseasonalized
1	2012	Q ₁	89.383				0,53	168.647
2		Q ₂	164.691				1,02	161.462
3		Q ₃	291.895	165.908	166.218	1,76	1,74	167.756
4		Q ₄	117.664	166.527	168.164	0,70	0,7	168.091
5	2013	Q ₁	91.858	169.802	171.197	0,54	0,53	173.317
6		Q ₂	177.790	172.592	174.020	1,02	1,02	174.304
7		Q ₃	303.054	175.449	175.767	1,72	1,74	174.169
8		Q ₄	129.092	176.085	177.808	0,73	0,7	184.417
9	2014	Q ₁	94.404	179.530	181.487	0,52	0,53	178.121
10		Q ₂	191.570	183.444	183.678	1,04	1,02	187.814
11		Q ₃	318.709	183.913	184.451	1,73	1,74	183.166
12		Q ₄	130.967	184.989	186.763	0,70	0,7	187.096
13	2015	Q ₁	98.709	188.538	194.676	0,51	0,53	186.243
14		Q ₂	205.766	200.813	202.415	1,02	1,02	201.731
15		Q ₃	367.811	204.017	206.081	1,78	1,74	211.386
16		Q ₄	143.781	208.145	209.363	0,69	0,7	205.401
17	2016	Q ₁	115.222	210.581	210.909	0,55	0,53	217.400
18		Q ₂	215.509	211.237	212.724	1,01	1,02	211.283
19		Q ₃	370.437	214.211	215.198	1,72	1,74	212.895
20		Q ₄	155.675	216.186	220.387	0,71	0,7	222.393

In order to make prediction of number of guest that will visit the Republic of Macedonia in 2017 we create next column named T_t – “Trend component in time t ” (t - referred to the first column). We run simple linear regression using the deseasonalized variables as Y and t variable as X in our regression model where Anova software is used. Summary output table 8 of our regression analysis is presented bellow:

Table 10. Regression analysis

Summary Output						
<i>Regression Statistics</i>						
Multiple R	0,96					
R Square	0,92					
Adjusted R Square	0,92					
Standard Error	5420,9276					
Observations	20					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	6,321E+09	6,32E+09	215,0984978	1,87725E-11	
Residual	18	528956205	29386456			
Total	19	6,85E+09				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	156482,52	2518,1914	62,14084	1,85213E-22	151191,9967	161773
t	3083,0564	210,21455	14,66624	1,87725E-11	2641,412021	3524,701

We use regression results to create column T_t as trend component, as shown on next table 9:

Table 11. Trend Component calculation

t	Year	Quarter	Num. of guests	MA (4)	CMA (4)	$S_{i,t}$	S_t	Dessesonalized	T_t
1	2012	Q ₁	89.383				0,53	168.647	159.566
2		Q ₂	164.691				1,02	161.462	162.649
3		Q ₃	291.895	165.908	166.218	1,76	1,74	167.756	165.732
4		Q ₄	117.664	166.527	168.164	0,70	0,70	168.091	168.815
5	2013	Q ₁	91.858	169.802	171.197	0,54	0,53	173.317	171.898
6		Q ₂	177.790	172.592	174.020	1,02	1,02	174.304	174.981
7		Q ₃	303.054	175.449	175.767	1,72	1,74	174.169	178.064
8		Q ₄	129.092	176.085	177.808	0,73	0,70	184.417	181.147
9	2014	Q ₁	94.404	179.530	181.487	0,52	0,53	178.121	184.230
10		Q ₂	191.570	183.444	183.678	1,04	1,02	187.814	187.313
11		Q ₃	318.709	183.913	184.451	1,73	1,74	183.166	190.396
12		Q ₄	130.967	184.989	186.763	0,70	0,70	187.096	193.479
13	2015	Q ₁	98.709	188.538	194.676	0,51	0,53	186.243	196.562
14		Q ₂	205.766	200.813	202.415	1,02	1,02	201.731	199.645
15		Q ₃	367.811	204.017	206.081	1,78	1,74	211.386	202.728
16		Q ₄	143.781	208.145	209.363	0,69	0,70	205.401	205.811
17	2016	Q ₁	115.222	210.581	210.909	0,55	0,53	217.400	208.894
18		Q ₂	215.509	211.237	212.724	1,01	1,02	211.283	211.978
19		Q ₃	370.437	214.211	215.198	1,72	1,74	212.895	215.061
20		Q ₄	155.675	216.186	220.387	0,71	0,70	222.393	218.144

In order to make prediction we have to combine all components that we have calculated separately S_t and T_t . Results are in presented in the column “prediction”, calculated as multiplied seasonal and trend components. We finally proceed with forecasting by adding new periods (four quarters for 2017 in our next Table 10 and calculate forecasted values:

Table 12. Forecasting

t	Year	Quarter	Num. of guests	MA (4)	CMA (4)	$S_{i,t}$	S_t	Dessesonalized	T_t	Forecast
1	2012	Q ₁	89.383				0,53	168.647	159.566	84.570
2		Q ₂	164.691				1,02	161.462	162.649	165.902
3		Q ₃	291.895	165.908	166.218	1,76	1,74	167.756	165.732	288.373
4		Q ₄	117.664	166.527	168.164	0,70	0,70	168.091	168.815	118.170
5	2013	Q ₁	91.858	169.802	171.197	0,54	0,53	173.317	171.898	91.106
6		Q ₂	177.790	172.592	174.020	1,02	1,02	174.304	174.981	178.480
7		Q ₃	303.054	175.449	175.767	1,72	1,74	174.169	178.064	309.831
8		Q ₄	129.092	176.085	177.808	0,73	0,70	184.417	181.147	126.803
9	2014	Q ₁	94.404	179.530	181.487	0,52	0,53	178.121	184.230	97.642
10		Q ₂	191.570	183.444	183.678	1,04	1,02	187.814	187.313	191.059
11		Q ₃	318.709	183.913	184.451	1,73	1,74	183.166	190.396	331.289
12		Q ₄	130.967	184.989	186.763	0,70	0,70	187.096	193.479	135.435
13	2015	Q ₁	98.709	188.538	194.676	0,51	0,53	186.243	196.562	104.178
14		Q ₂	205.766	200.813	202.415	1,02	1,02	201.731	199.645	203.638
15		Q ₃	367.811	204.017	206.081	1,78	1,74	211.386	202.728	352.747
16		Q ₄	143.781	208.145	209.363	0,69	0,70	205.401	205.811	144.068
17	2016	Q ₁	115.222	210.581	210.909	0,55	0,53	217.400	208.894	110.714
18		Q ₂	215.509	211.237	212.724	1,01	1,02	211.283	211.978	216.217
19		Q ₃	370.437	214.211	215.198	1,72	1,74	212.895	215.061	374.205
20		Q ₄	155.675	216.186	220.387	0,71	0,70	222.393	218.144	152.701
21	2017	Q ₁	123.122	224.589	200.281	0,61	0,53		221.227	117.250
22		Q ₂	249.120	175.972			1,02		224.310	228.796
23		Q ₃					1,74		227.393	395.664
24		Q ₄					0,70		230.476	161.333

After finished calculations now we can plot forecasting series on the chart in order to visualize accuracy of the forecasting, as follows:

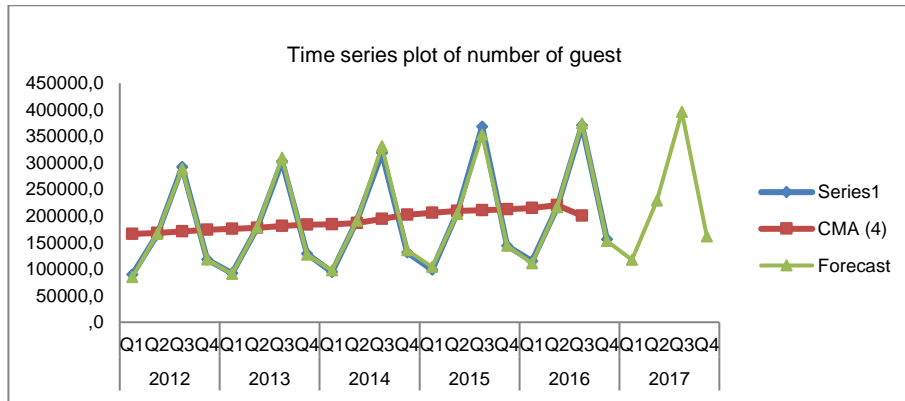


Figure 3. Time series of number of guest - forecasts

We can conclude from the chart that our forecasting line (green) – general flows is really well forecasted. This plot with 3 lines shows that actual data and predicted data for year 2017 are well approximated. Our forecasting values are appropriate with movement and behaviour of actual values. Finally we check our forecasting accuracy with actual data for 2017 and conclude that forecasting values are very close as the actual data for the first two quarters of 2017.

Table 13. Actual and forecasted data for number of tourist in the Republic of Macedonia 2017

Year	Quarter	Actual	Forecasted	Difference %
2017	Q ₁	123.122	117.250	-4,8%
	Q ₂	249.120	228.796	-8,1%

Source: State Statistical Office

Comparing actual and forecasted data we can conclude that there is relatively small difference between them and that time series forecasting moving average model provides accurate and reliable prediction of number of tourists with acceptable estimation. This finding confirms our research hypothesis that moving average model for the time series data provide accurate forecasting for the number of tourist guests in next year.

CONCLUSION

Time series modeling and forecasting has fundamental importance to various practical domains having in mind that time series forecasting enables predicting the future by understanding the past. In tourist industry forecasting future revenues depends on capability to make prediction of number of tourists, having in mind that there are data for average tourist spending per day. If all stakeholders in tourist industry have relevant and accurate forecast of number of tourists that will visit the country they can provide reliable input for their financial planning. In order to do that it is important to use quantitative methods and statistics tools that can help to deal with past data where seasonality, irregularity and sometimes random variables appears.

There are many comprehensive theoretical studies in literature that suggest the necessity to use quantitative models and techniques for various types of data. In this paper we focus on a time series analysis. A time series is a collection of observations made sequentially in time. Time series analysis refers to problems in which observations are collected at regular time intervals and there are correlations among successive observations.

Analysis of the data for the number of guest shows strong seasonal component, obvious pattern in data variations and trend. Having in mind that the moving average merely smooths the fluctuations in the data we chose this time series forecasting model for our research. A moving average model works well when the data have a fairly linear trend and a definite rhythmic pattern of fluctuations.

Research finding confirms our research hypothesis that moving average model for the time series data provide accurate forecasting for the number of tourist guests in next year.

This study outlines directions for future researches that could be investigated to improve the forecasting of other tourist and economic indicators for the Macedonian economy. Due to the fact that we use limited data and time series of number of guest (2012–2016) and compare with actual number of tourist for first two quarters of 2017, longer time series would allow estimation with greater precision.

REFERENCES

- Armstrong, Scott, J. 2006. Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting* 22: 583–589.
- Adikhari, Ratnadip, and R. K. Agrawal. 2013. *An Introductory Study on Time Series Modeling and Forecasting*. Berlin: LAP Lambert Academic Publishing, Germany.
- Bartholomew, David J. 1971. Time series analysis forecasting and control. *Journal of the Operational Research Society* 22 (2): 199–201
- Brockwell, Peter J., and Richard A. Davis. 2013. *Time Series: Theory and Methods*. New York: Springer Science & Business Media.
- Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* 31 (3): 307–327.
- Box, George E. P., and Gwilym M. Jenkins. 1971. Time Series Analysis: Forecasting and Control. *Operational Research Quarterly* 22 (2): 199–201.
- Chatfield, Chris. 1996. Model uncertainty and forecast accuracy. *Journal of Forecasting* 15 (7): 495–508.
- Chernick, Michael R., 1994. Time Series: Forecasting, Simulation, Applications by Gareth, Janacek and Louise Swift, Reviewed Work. *The American Statistician* 48 (1): 58–59.
- Cochrane, John H. 1997. *Time Series for Macroeconomic and Finance*. Chicago: Graduate School of Business, University of Chicago.
- Cottrell, Marie, Bernand Girard, Yvone Girard, Morgan Mangeas, and Corinne Muller. 1995. Neural modeling for time series: A statistical stepwise method for weight elimination. *IEEE Trans. Neural Networks* 6 (6): 1355–1364.
- Diggle, Peter J. 1990. *Time Series: A Biostatistical Introduction*. London: Oxford University Press.
- Fildes, Robert, and Spyros Makridakis. 1993. *The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting*. France, Fontainebleau: INSEAD.
- Gardner, Everette S. Jr., and Ed McKenzie. 1985. Forecasting Trends in Time Series. *Management Science* 31 (10): 1237–1246.
- Janacek, Gareth. 2009. Time series analysis forecasting and control. *Journal of Time Series Analysis* 31 (4): 303.
- Montgomery, Douglas C., Cheryl L. Jennings, and Murat Kulachi. 2008. *Introduction to Time Series Analysis and Forecasting*. London: Wiley.
- Nelson, Charles R., and Charles R. Plosser. 1982. Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics* 10 (2): 139–162.

t Tests

One Sample t-Tests

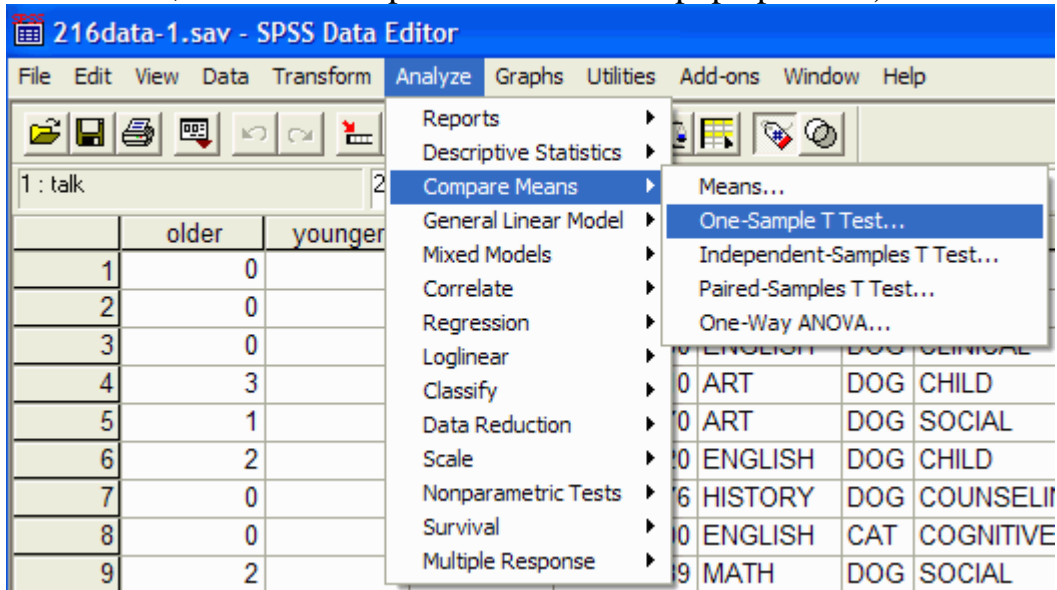
One sample t-tests can be used to determine if the mean of a sample is different from a particular value. In this example, we will determine if the mean number of older siblings that the PSY 216 students have is greater than 1.

We will follow our customary steps:

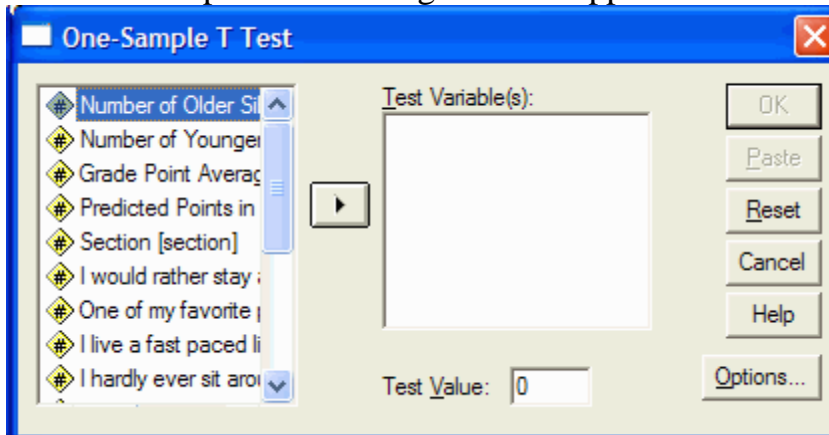
1. Write the null and alternative hypotheses first:
 $H_0: \mu_{216 \text{ Students}} \leq 1$
 $H_1: \mu_{216 \text{ Students}} > 1$
Where μ is the mean number of older siblings that the PSY 216 students have.
2. Determine if this is a one-tailed or a two-tailed test. Because the hypothesis involves the phrase "greater than", this must be a one tailed test.
3. Specify the α level: $\alpha = .05$
4. Determine the appropriate statistical test. The variable of interest, older, is on a ratio scale, so a z-score test or a t-test might be appropriate. Because the population standard deviation is not known, the z-test would be inappropriate. We will use the t-test instead.
5. Calculate the t value, or let SPSS do it for you!

The command for a one sample t tests is found at Analyze | Compare Means | One-Sample T Test (this is shorthand for clicking on the Analyze menu item at the top of the window, and then clicking on Compare Means from the drop

down menu, and One-Sample T Test from the pop up menu.):

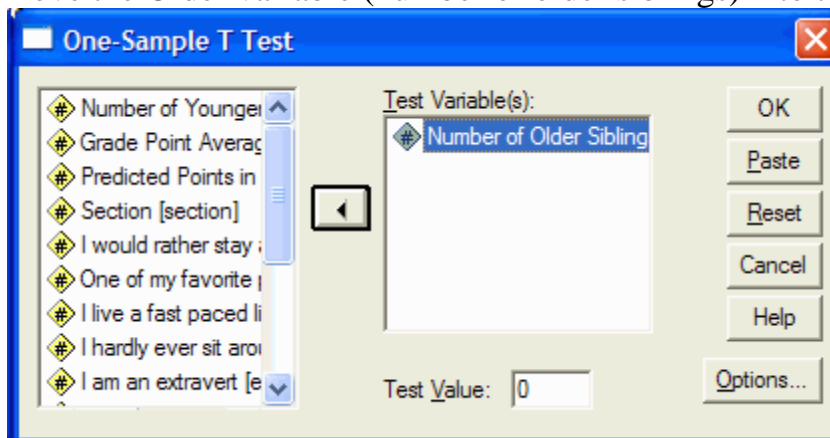


The One-Sample t Test dialog box will appear:

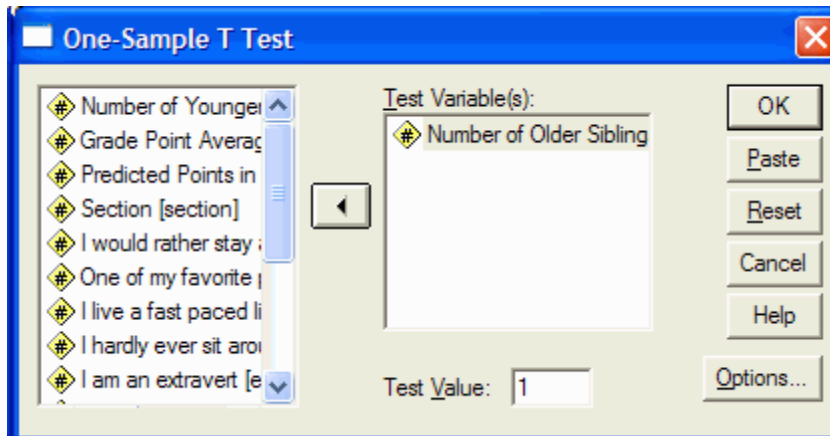


Select the dependent variable(s) that you want to test by clicking on it in the left hand pane of the One-Sample t Test dialog box. Then click on the arrow button to move the variable into the Test Variable(s) pane. In this example,

move the Older variable (number of older siblings) into the Test Variables box:



Click in the Test Value box and enter the value that you will compare to. In this example, we are comparing if the number of older siblings is greater than 1, so we should enter 1 into the Test Value box:



Click on the OK button to perform the one-sample t test. The output viewer will appear. There are two parts to the output. The first part gives descriptive statistics for the variables that you moved into the Test Variable(s) box on the One-Sample t Test dialog box. In this example, we get descriptive statistics for the Older variable:

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
Number of Older Siblings	46	1.26	1.255	.185

This output tells us that we have 46 observations (N), the mean number of older siblings is 1.26 and the standard deviation of the number of older siblings is 1.255. The standard error of the mean (the standard deviation of the sampling distribution of means) is 0.185 ($1.255 / \text{square root of } 46 = 0.185$).

The second part of the output gives the value of the statistical test:

One-Sample Test						
	Test Value = 1					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Number of Older Siblings	1.410	45	.165	.261	-.11	.63

The second column of the output gives us the t-test value: $(1.26 - 1) / (1.255 / \text{square root of } 46) = 1.410$ [if you do the calculation, the values will not match exactly because of round-off error). The third column tells us that this t test has 45 degrees of freedom ($46 - 1 = 45$). The fourth column tells us the *two-tailed* significance (the 2-tailed p value.) But we didn't want a two-tailed test; our hypothesis is one tailed and there is no option to specify a one-tailed test. Because this is a one-tailed test, look in a table of critical t values to determine the critical t. The critical t with 45 degrees of freedom, $\alpha = .05$ and one-tailed is 1.679.

- Determine if we can reject the null hypothesis or not. The decision rule is: if the one-tailed critical t value is less than the observed t AND the means are in the right order, then we can reject H_0 . In this example, the critical t is 1.679 (from the table of critical t values) and the observed t is 1.410, so we fail to reject H_0 . That is, there is insufficient evidence to conclude that the mean number of older siblings for the PSY 216 classes is larger than 1.

If we were writing this for publication in an APA journal, we would write it as:

A *t* test failed to reveal a statistically reliable difference between the mean number of older siblings that the PSY 216 class has ($M = 1.26, s = 1.26$) and 1, $t(45) = 1.410, p < .05, \alpha = .05$.

Independent Samples t-Tests Single Value Groups

When two samples are involved, the samples can come from different individuals who are not matched (the samples are independent of each other.) Or the sample can come from the same individuals (the samples are paired with each other) and the samples are not independent of each other. A third alternative is that the samples can come from different individuals who have been matched on a variable of interest; this type of sample will not be independent. The form of the t-test is slightly different for the independent samples and dependent samples types of two sample tests, and SPSS has separate procedures for performing the two types of tests.

The Independent Samples t-test can be used to see if two means are different from each other when the two samples that the means are based on were taken from different individuals who have not been matched. In this example, we will determine if the students in sections one and two of PSY 216 have a different number of older siblings.

We will follow our customary steps:

1. Write the null and alternative hypotheses first:

$$H_0: \mu_{\text{Section 1}} = \mu_{\text{Section 2}}$$

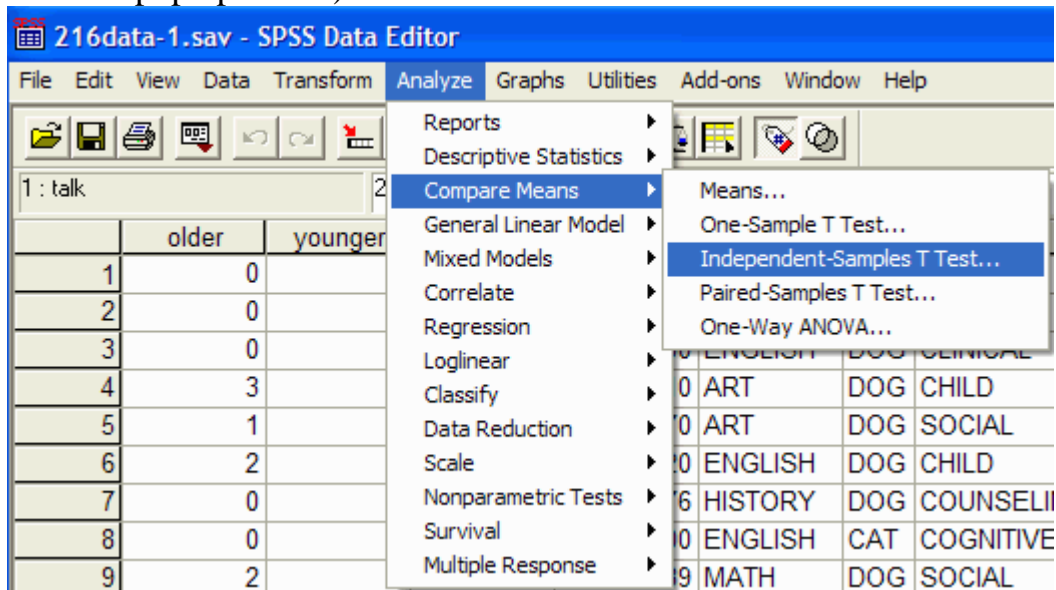
$$H_1: \mu_{\text{Section 1}} \neq \mu_{\text{Section 2}}$$

Where μ is the mean number of older siblings that the PSY 216 students have.

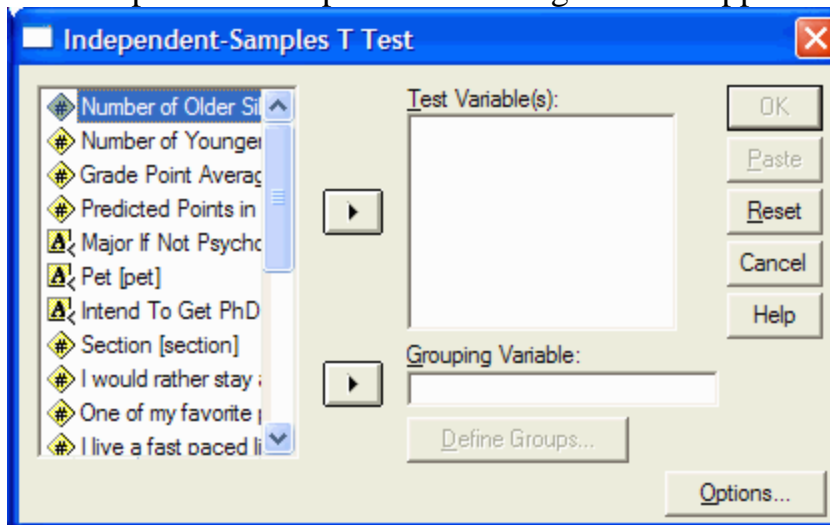
2. Determine if this is a one-tailed or a two-tailed test. Because the hypothesis involves the phrase "different" and no ordering of the means is specified, this must be a two-tailed test.
3. Specify the α level: $\alpha = .05$
4. Determine the appropriate statistical test. The variable of interest, older, is on a ratio scale, so a z-score test or a t-test might be appropriate. Because the population standard deviation is not known, the z-test would be inappropriate. Furthermore, there are different students in sections 1 and 2 of PSY 216, and they have not been matched. Because of these factors, we will use the independent samples t-test.
5. Calculate the t value, or let SPSS do it for you!

The command for the independent samples t tests is found at Analyze | Compare Means | Independent-Samples T Test (this is shorthand for clicking on the Analyze menu item at the top of the window, and then clicking on Compare Means from the drop down menu, and Independent-Samples T Test

from the pop up menu.):

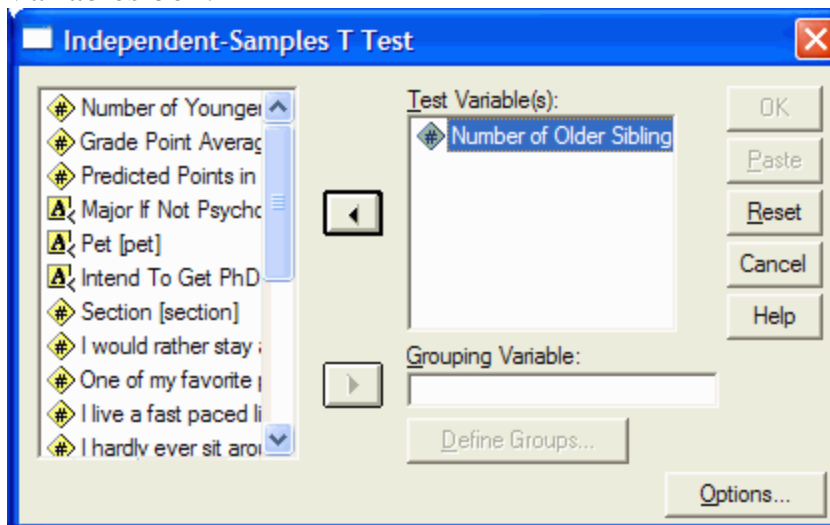


The Independent-Samples t Test dialog box will appear:

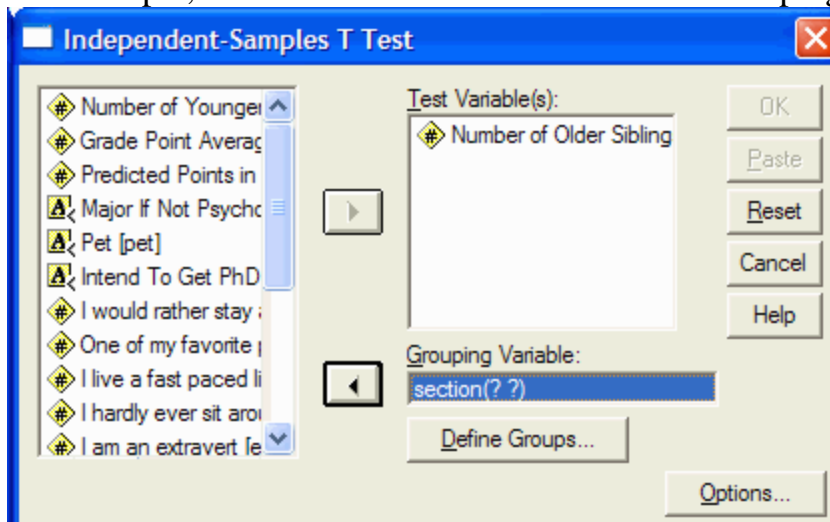


Select the dependent variable(s) that you want to test by clicking on it in the left hand pane of the Independent-Samples t Test dialog box. Then click on the upper arrow button to move the variable into the Test Variable(s) pane. In this example, move the Older variable (number of older siblings) into the Test

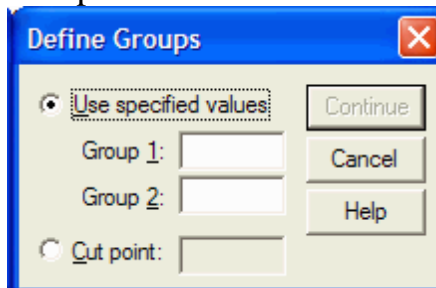
Variables box:



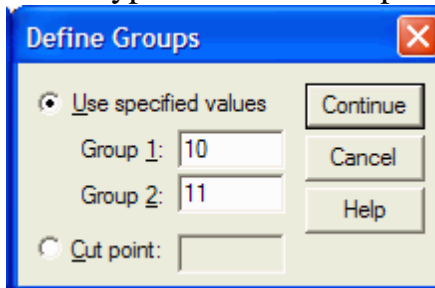
Click on the independent variable (the variable that defines the two groups) in the left hand pane of the Independent-Samples t Test dialog box. Then click on the lower arrow button to move the variable in the Grouping Variable box. In this example, move the Section variable into the Grouping Variable box:



You need to tell SPSS how to define the two groups. Click on the Define Groups button. The Define Groups dialog box appears:



In the Group 1 text box, type in the value that determines the first group. In this example, the value of the 10 AM section is 10. So you would type 10 in the Group 1 text box. In the Group 2 text box, type the value that determines the second group. In this example, the value of the 11 AM section is 11. So you would type 11 in the Group 2 text box:



Click on the Continue button to close the Define Groups dialog box. Click on the OK button in the Independent-Samples t Test dialog box to perform the t-test. The output viewer will appear with the results of the t test. The results have two main parts: descriptive statistics and inferential statistics. First, the descriptive statistics:

	Section	N	Mean	Std. Deviation	Std. Error Mean
Number of Older Siblings	10	14	.86	1.027	.275
	11	32	1.44	1.318	.233

This gives the descriptive statistics for each of the two groups (as defined by the grouping variable.) In this example, there are 14 people in the 10 AM section (N), and they have, on average, 0.86 older siblings, with a standard deviation of 1.027 older siblings. There are 32 people in the 11 AM section (N), and they have, on average, 1.44 older siblings, with a standard deviation of 1.318 older siblings. The last column gives the standard error of the mean for each of the two groups.

The second part of the output gives the inferential statistics:

Independent Samples Test							
		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Df
Number of Older Siblings	Equal variances assumed	1.669	.203	-1.461	44	.151	
	Equal variances not assumed			-1.612	31.607	.117	

The columns labeled "Levene's Test for Equality of Variances" tell us whether an assumption of the t-test has been met. The t-test assumes that the variability of each group is approximately equal. If that assumption isn't met, then a special form of the t-test should be used. Look at the column labeled "Sig." under the heading "Levene's Test for Equality of Variances". In this example, the significance (p value) of Levene's test is .203. If this value is less than or equal to your α level for the test (usually .05), then you can reject the null hypothesis that the variability of the two groups is equal, implying that the variances are unequal. If the p value is less than or equal to the α level, then you should use the bottom row of the output (the row labeled "Equal variances not assumed.") If the p value is greater than your α level, then you should use the middle row of the output (the row labeled "Equal variances assumed.") In this example, .203 is larger than α , so we will assume that the variances are equal and we will use the middle row of the output.

The column labeled "t" gives the observed or calculate t value. In this example, assuming equal variances, the t value is 1.461. (We can ignore the sign of t for a two tailed t-test.) The column labeled "df" gives the degrees of freedom associated with the t test. In this example, there are 44 degrees of freedom.

The column labeled "Sig. (2-tailed)" gives the two-tailed p value associated with the test. In this example, the p value is .151. If this had been a one-tailed test, we would need to look up the critical t in a table.

- Decide if we can reject H_0 : As before, the decision rule is given by: If $p \leq \alpha$, then reject H_0 . In this example, .151 is not less than or equal to .05, so we fail to reject H_0 . That implies that we failed to observe a difference in the number of older siblings between the two sections of this class.

If we were writing this for publication in an APA journal, we would write it as:
A t test failed to reveal a statistically reliable difference between the mean number of older siblings that the 10 AM section has ($M = 0.86$, $s = 1.027$) and that the 11 AM section has ($M = 1.44$, $s = 1.318$), $t(44) = 1.461$, $p = .151$, $\alpha = .05$.

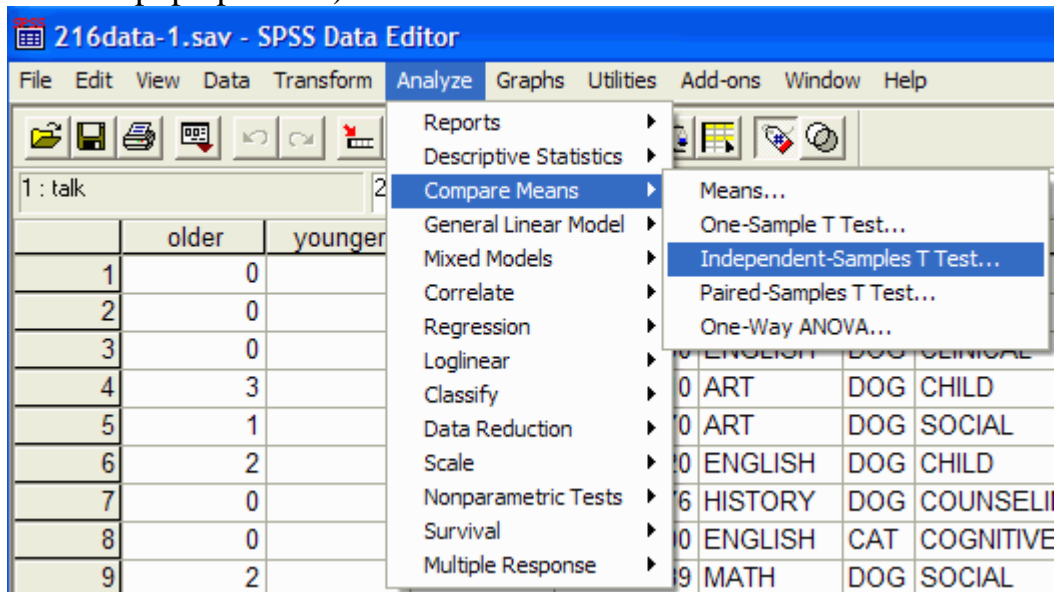
Independent Samples t-Tests Cut Point Groups

Sometimes you want to perform a t-test but the groups are defined by a variable that is not dichotomous (i.e., it has more than two values.) For example, you may want to see if the number of older siblings is different for students who have higher GPAs than for students who have lower GPAs. Since there is no single value of GPA that specifies "higher" or "lower", we cannot proceed exactly as we did before. Before proceeding, decide which value you will use to divide the GPAs into the higher and lower groups. The median would be a good value, since half of the scores are above the median and half are below. (If you do not remember how to calculate the median see the [frequency command in the descriptive statistics tutorial](#).)

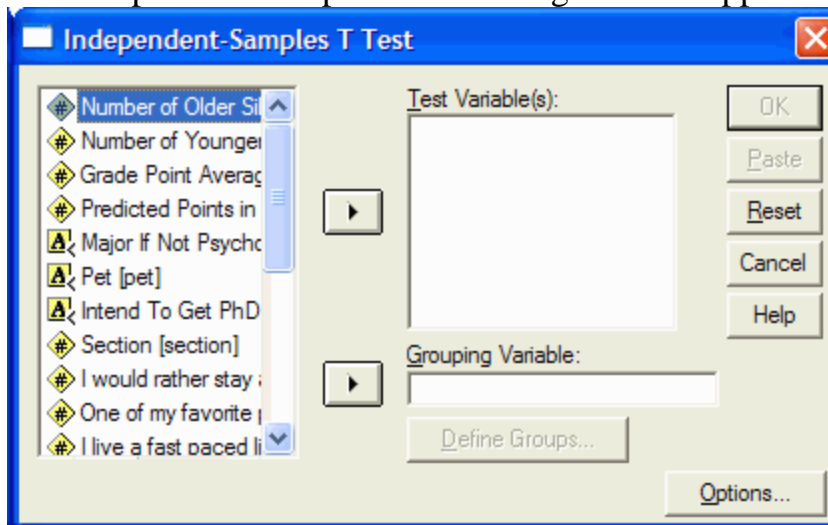
1. Write the null and alternative hypotheses first:
 $H_0: \mu_{\text{lower GPA}} = \mu_{\text{higher GPA}}$
 $H_1: \mu_{\text{lower GPA}} \neq \mu_{\text{higher GPA}}$
Where μ is the mean number of older siblings that the PSY 216 students have.
2. Determine if this is a one-tailed or a two-tailed test. Because the hypothesis involves the phrase "different" and no ordering of the means is specified, this must be a two tailed test.
3. Specify the α level: $\alpha = .05$
4. Determine the appropriate statistical test. The variable of interest, older, is on a ratio scale, so a z-score test or a t-test might be appropriate. Because the population standard deviation is not known, the z-test would be inappropriate. Furthermore, different students have higher and lower GPAs, so we have a between-subjects design. Because of these factors, we will use the independent samples t-test.
5. Calculate the t value, or let SPSS do it for you.

The command for the independent samples t tests is found at Analyze | Compare Means | Independent-Samples T Test (this is shorthand for clicking on the Analyze menu item at the top of the window, and then clicking on Compare Means from the drop down menu, and Independent-Samples T Test

from the pop up menu.):

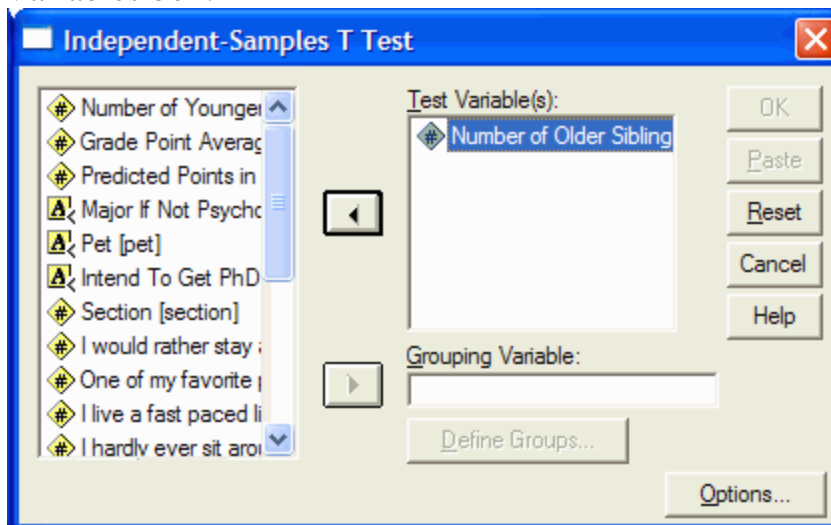


The Independent-Samples t Test dialog box will appear:

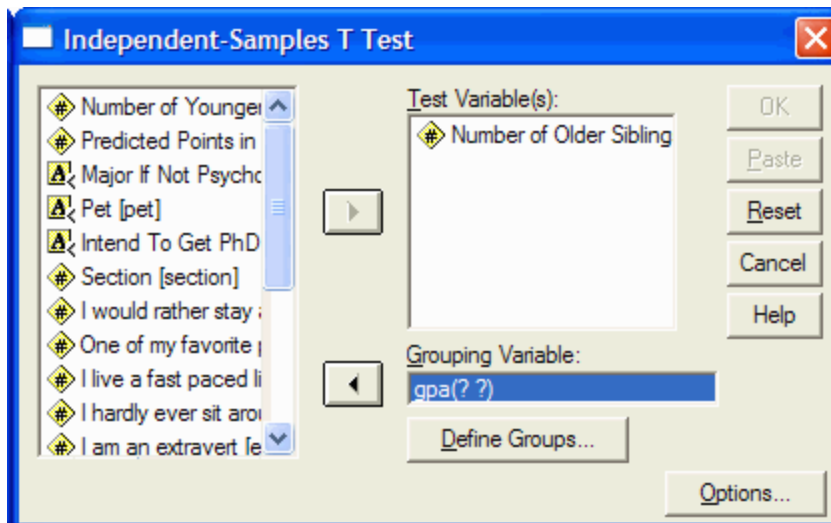


Select the dependent variable(s) that you want to test by clicking on it in the left hand pane of the Independent-Samples t Test dialog box. Then click on the upper arrow button to move the variable into the Test Variable(s) pane. In this example, move the Older variable (number of older siblings) into the Test

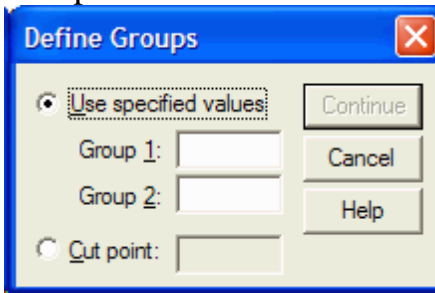
Variables box:



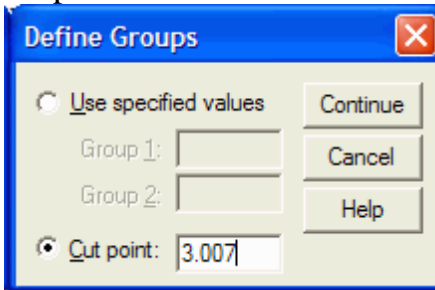
Click on the independent variable (the variable that defines the two groups) in the left hand pane of the Independent-Samples t Test dialog box. Then click on the lower arrow button to move the variable in the Grouping Variable box. (If there already is a variable in the Grouping Variable box, click on it if it is not already highlighted, and then click on the lower arrow which should be pointing to the left.) In this example, move the GPA variable into the Grouping Variable box:



You need to tell SPSS how to define the two groups. Click on the Define Groups button. The Define Groups dialog box appears:



Click in the circle to the left of "Cut Point:". Then type the value that splits the variable into two groups. Group one is defined as all scores that are greater than or equal to the cut point. Group two is defined as all scores that are less than the cut point. In this example, use 3.007 (the median of the GPA variable) as the cut point value:



Click on the Continue button to close the Define Groups dialog box. Click on the OK button in the Independent-Samples t Test dialog box to perform the t-test. The output viewer will appear with the results of the t test. The results have two main parts: descriptive statistics and inferential statistics. First, the descriptive statistics:

Group Statistics					
	Grade Point Average	N	Mean	Std. Deviation	Std. Error Mean
Number of Older Siblings	>= 3.01	23	1.04	1.186	.247
	< 3.01	23	1.48	1.310	.273

This gives the descriptive statistics for each of the two groups (as defined by the grouping variable.) In this example, there are 23 people with a GPA greater than or equal to 3.01 (N), and they have, on average, 1.04 older siblings, with a standard deviation of 1.186 older siblings. There are 23 people with a GPA less than 3.01 (N), and they have, on average, 1.48 older siblings, with a standard deviation of 1.310 older siblings. The last column gives the standard error of the mean for each of the two groups.

The second part of the output gives the inferential statistics:

		Independent Samples Test					
		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Df
Number of Older Siblings	Equal variances assumed	.776	.383	-1.180	44	.244	
	Equal variances not assumed			-1.180	43.575	.244	

As before, the columns labeled "Levene's Test for Equality of Variances" tell us whether an assumption of the t-test has been met. Look at the column labeled "Sig." under the heading "Levene's Test for Equality of Variances". In this example, the significance (p value) of Levene's test is .383. If this value is less than or equal to your α level for this test, then you can reject the null hypothesis that the variabilities of the two groups are equal, implying that the variances are unequal. In this example, .383 is larger than our α level of .05, so we will assume that the variances are equal and we will use the middle row of the output.

The column labeled "t" gives the observed or calculated t value. In this example, assuming equal variances, the t value is 1.180. (We can ignore the sign of t when using a two-tailed t-test.) The column labeled "df" gives the degrees of freedom associated with the t test. In this example, there are 44 degrees of freedom.

The column labeled "Sig. (2-tailed)" gives the two-tailed p value associated with the test. In this example, the p value is .244. If this had been a one-tailed test, we would need to look up the critical t in a table.

- Decide if we can reject H_0 : As before, the decision rule is given by: If $p \leq \alpha$, then reject H_0 . In this example, .244 is greater than .05, so we fail to reject H_0 . That implies that there is not sufficient evidence to conclude that people with higher or lower GPAs have different number of older siblings.

If we were writing this for publication in an APA journal, we would write it as:
 An equal variances *t* test failed to reveal a statistically reliable difference between the mean number of older siblings for people with higher ($M = 1.04$, $s = 1.186$) and lower GPAs ($M = 1.48$, $s = 1.310$), $t(44) = 1.18$, $p = .244$, $\alpha = .05$.

Paired Samples t-Tests

When two samples are involved and the values for each sample are collected from the same individuals (that is, each individual gives us two values, one for each of the two groups), or the samples come from matched pairs of individuals then a paired-samples t-test may be an appropriate statistic to use.

The paired samples t-test can be used to determine if two means are different from each other when the two samples that the means are based on were taken from the matched individuals or the same individuals. In this example, we will determine if the students have different numbers of younger and older siblings.

1. Write the null and alternative hypotheses:

$$H_0: \mu_{\text{older}} = \mu_{\text{younger}}$$

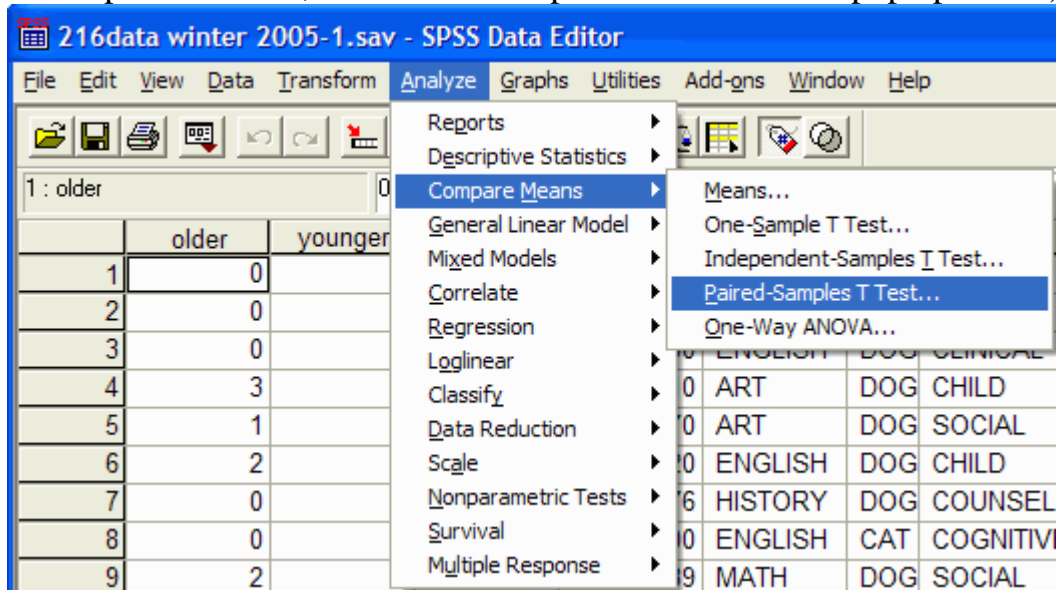
$$H_1: \mu_{\text{older}} \neq \mu_{\text{younger}}$$

Where μ is the mean number of siblings that the PSY 216 students have.

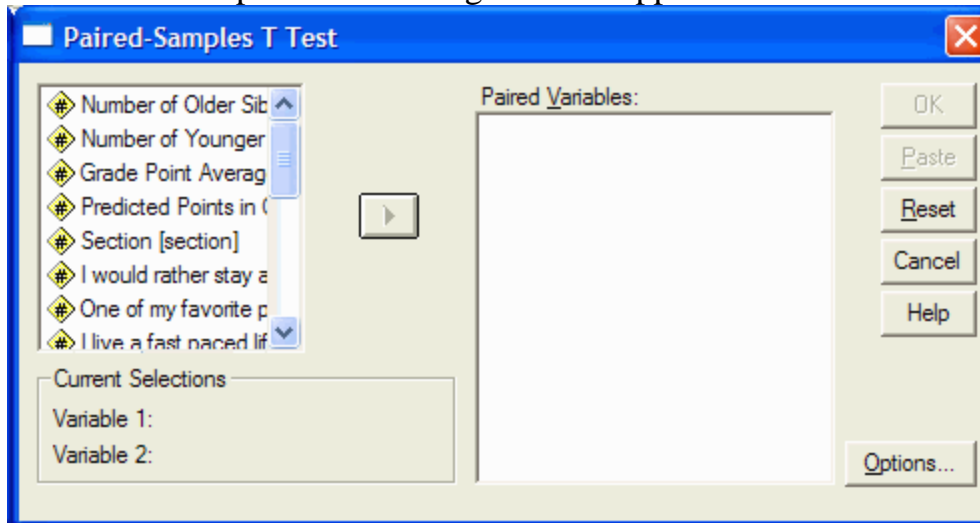
2. Determine if this is a one-tailed or a two-tailed test. Because the hypothesis involves the phrase "different" and no ordering of the means is specified, this must be a two tailed test.
3. Specify the α level: $\alpha = .05$
4. Determine the appropriate statistical test. The variables of interest, older and younger, are on a ratio scale, so a z-score test or a t-test might be appropriate. Because the population standard deviation is not known, the z-test would be inappropriate. Furthermore, the same students are reporting the number of older and younger siblings, we have a within-subjects design. Because of these factors, we will use the paired samples t-test.
5. Let SPSS calculate the value of t for you.

The command for the paired samples t tests is found at Analyze | Compare Means | Paired-Samples T Test (this is shorthand for clicking on the Analyze menu item at the top of the window, and then clicking on Compare Means from

the drop down menu, and Paired-Samples T Test from the pop up menu.):

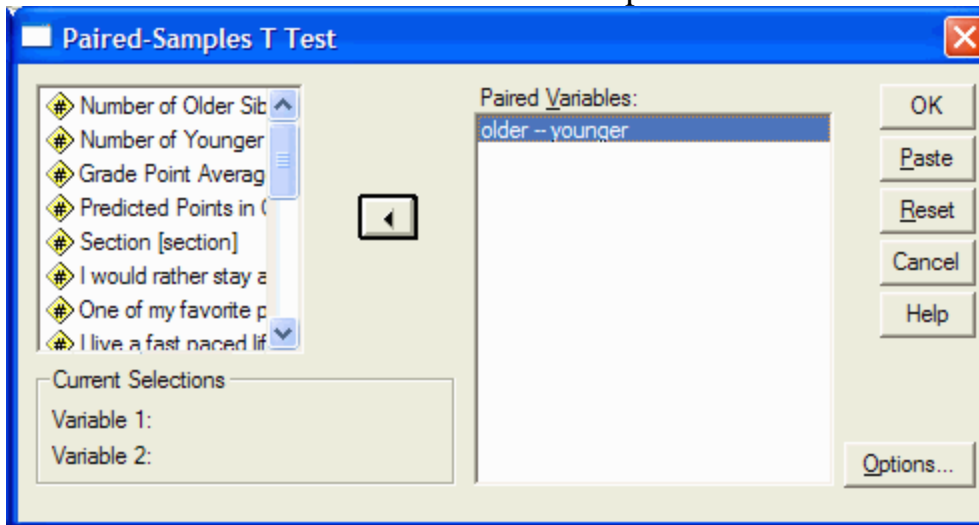


The Paired-Samples t Test dialog box will appear:



You must select a pair of variables that represent the two conditions. Click on one of the variables in the left hand pane of the Paired-Samples t Test dialog box. Then click on the other variable in the left hand pane. Click on the arrow button to move the variables into the Paired Variables pane. In this example, select Older and Younger variables (number of older and younger siblings) and

then click on the arrow button to move the pair into the Paired Variables box:



Click on the OK button in the Paired-Samples t Test dialog box to perform the t-test. The output viewer will appear with the results of the t test. The results have three main parts: descriptive statistics, the correlation between the pair of variables, and inferential statistics. First, the descriptive statistics:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Number of Older Siblings	1.24	45	1.264	.188
	Number of Younger Siblings	1.13	45	1.198	.179

This gives the descriptive statistics for each of the two groups (as defined by the pair of variables.) In this example, there are 45 people who responded to the Older siblings question (N), and they have, on average, 1.24 older siblings, with a standard deviation of 1.26 older siblings. These same 45 people also responded to the Younger siblings question (N), and they have, on average, 1.13 younger siblings, with a standard deviation of 1.20 younger siblings. The last column gives the standard error of the mean for each of the two variables.

The second part of the output gives the correlation between the pair of variables:

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Number of Older Siblings & Number of Younger Siblings	45	-.292	.052

This again shows that there are 45 pairs of observations (N). The correlation between the two variables is given in the third column. In this example $r = -.292$. The last column give the p value for the correlation coefficient. As always, if the p value is less than or equal to the alpha level, then you can reject the null hypothesis that the population correlation coefficient (ρ) is equal to 0. In this case, $p = .052$, so we fail to reject the null hypothesis. That is, there is insufficient evidence to conclude that the population correlation (ρ) is different from 0.

The third part of the output gives the inferential statistics:

		Paired Differences					t	df
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference			
					Lower	Upper		
Pair 1	Number of Older Siblings - Number of Younger Siblings	.111	1.980	.295	-.484	.706	.377	

The column labeled "Mean" is the difference of the two means ($1.24 - 1.13 = 0.11$ in this example (the difference is due to round off error).) The next column is the standard deviation of the difference between the two variables (1.98 in this example.)

The column labeled "t" gives the observed or calculated t value. In this example, the t value is 0.377 (you can ignore the sign.) The column labeled "df" gives the degrees of freedom associated with the t test. In this example, there are 44 degrees of freedom. The column labeled "Sig. (2-tailed)" gives the two-tailed p value associated with the test. In this example, the p value is .708. If this had been a one-tailed test, we would need to look up the critical value of t in a table.

- Decide if we can reject H_0 : As before, the decision rule is given by: If $p \leq \alpha$, then reject H_0 . In this example, .708 is not less than or equal to .05, so we fail to reject H_0 . That implies that there is insufficient evidence to conclude that the number of older and younger siblings is different.

If we were writing this for publication in an APA journal, we would write it as:

A paired samples *t* test failed to reveal a statistically reliable difference between the mean number of older ($M = 1.24$, $s = 1.26$) and younger ($M = 1.13$, $s = 1.20$) siblings that the students have, $t(44) = 0.377$, $p = .708$, $\alpha = .05$.