Development of Statistical Analysis Guidelines for Highway Materials and Research Activities

Final Report

No. FHWA-SC-06-05

by

James L. Burati, Jr.

and

Wei Lin

Civil Engineering Department
Clemson, SC 29634-0911

May 2006

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-SC-06-05		

4. Title and Subtitle	5. Report Date
Development of Statistical Analysis Guidelines for Highway Materials and Research Activities	May 2006
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
James L. Burati, Jr. and Wei Lin	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Clemson University Civil Engineering Department Clemson, SC 29634-0911	11. Contract or Grant No. 649
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
South Carolina Department of Transportation P O Box 191 Columbia, SC 29202-0191	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes

This study was conducted in cooperation with the U. S. Department of Transportation, Federal Highway Administration.

16. Abstract

Materials and research engineers within the South Carolina Department of Transportation (SCDOT) routinely must perform statistical analyses related to, but not limited to, such items as testing data, forensic investigation of problems, review of proposals and reports, determining population parameters for specification development, verification of contractor test results, developing precision and bias statements, and design of internal experiments and evaluation of experimental designs of externally funded projects. It is essential that decisions they make be based on sound statistical principles.

SCDOT personnel were interviewed to identify a prioritized list of the statistical concepts that must be understood to assist them in making sound decisions regarding materials and research issues. Guidebooks were developed for the identified concepts. The guidebooks serve as more than just "cookbooks" or procedures manuals, but also incorporate the basic underlying statistical concepts to support the procedures included in the guidebooks. Training modules were then developed for each of the statistical guidebooks and were offered to SCDOT personnel in five day-long workshops.

Guidelines that were developed included: Basic Concepts, Graphical Presentation of Data, Statistical Inferences on One-Sample Data, Statistical Inferences on Two-Sample Data, Goodness of Fit Tests, Identifying Outliers, Time Series Data and Control Charts, Regression Analysis, Experimental Design, PWL Specifications: Limits & Payments, and PWL Specifications: Risks.

17 Key Words	18. Distribution Statement
Statistics, Statistical Methods, Probability, Data Analysis, Regression Analysis, Design of Experiments, PWL Specifications, Risks	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19 Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	83	

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
NOTE: volumes greater than 1000 L shall be shown in m^3				
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
^oF	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^oC
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
^oC	Celsius	1.8C+32	Fahrenheit	^oF
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

TABLE OF CONTENTS

LIST OF TABLES

LIST OF EXHIBITS

CHAPTER 1 — INTRODUCTION

Background

As have many state highway agencies, the South Carolina Department of Transportation (SCDOT) has recently implemented new specifications and quality programs for a number of materials, including hot mixed asphalt concrete, portland cement concrete, and aggregates. To a great extent statistical concepts form the basis of these new specifications and quality programs.

Materials and research engineers within the SCDOT routinely must perform statistical analyses related to, but not limited to, such items as testing data, forensic investigation of problems, review of proposals and reports, determining population parameters for specification development, verification of contractor test results, developing precision and bias statements, and design of internal experiments and evaluation of experimental designs of externally funded projects.

SCDOT materials and research engineers deal daily with test data analysis, population comparisons, materials and procedural variability issues, and related statistical matters. Decisions affecting contractor payment, materials acceptance criteria, and specification development are part of their routine activities.

It is essential that these decisions be based on sound statistical principles. It is therefore necessary for SCDOT materials and research engineers to have a sound understanding of the fundamental statistical principles on which such decisions are based.

To facilitate this understanding of statistical principles, and to ensure the proper and consistent application of these principles, a set of statistical analysis guidelines needed to be established. These guidelines would provide a more structured methodology for the types of critical decisions that materials and research engineers must frequently make.

To ensure that these statistical guidelines are understood, a series of training manuals and workshops needed to be developed to transfer these statistical concepts to SCDOT personnel.

It is necessary to have a fundamental understanding of a variety of statistical concepts to be able to make correct decisions when faced with experimental or field data. The ability of materials and research engineers to make correct decisions will be enhanced by the development of statistical guidelines and associated training workshops.

Objectives

The objectives of this study were:

- To identify a prioritized list of the statistical concepts that must be understood by SCDOT materials and research engineers to assist them in making sound decisions regarding materials and research issues.

- To use new or existing documents or textbooks to develop for materials and research engineers a series of statistical guidebooks that present the topics identified in objective 1.
- To develop and offer a series of workshops to train SCDOT personnel in the statistical topics presented in the guidebooks developed in objective 2.

Methodology

The major items that needed to be accomplished to achieve the project objectives are discussed in each of the following sections. These major work tasks include:

- Identify SCDOT's training needs regarding statistical concepts.
- Prepare statistical guidebooks for the topics identified in step 1.
- Develop and offer training workshops on each of the statistical guidebooks.
- Evaluate statistical software packages and present a summary of their advantages and disadvantages to the SCDOT.

Identify SCDOT's Training Needs Regarding Statistical Concepts. The first step that was taken was to establish a SCDOT "Statistics Team." This team was composed of SCDOT personnel selected by the Research and Materials Engineer in consultation with the principal investigator (PI). The Statistics Team was charged to oversee the project on behalf of the SCDOT. The team served as a primary resource for determining for which statistical concepts guidebooks needed to be developed and training needed to be offered. The PI served as the facilitator during meetings at which the Statistics Team identified and prioritized SCDOT's statistics training needs. These meetings were held in Columbia to minimize travel costs for SCDOT personnel. The members of the Statistics Team are shown in Table 1.

Prepare Statistical Guidebooks. Guidebooks were developed for the statistical topics selected by the Statistics Team. For each guidebook the PI and Wei Lin, a graduate research assistant (GRA) from the Clemson University Mathematical Sciences Department, prepared draft outlines for review and approval by the Statistics Team. The PI and GRA then prepared the guidebooks that were used in the training sessions. Input from the training sessions was then incorporated into the final statistical guidebooks.

The guidebooks serve as more than just "cookbooks" or procedures manuals, but also incorporate the basic underlying statistical concepts to support the procedures included in the guidebooks. The guidebooks contain numerous examples that are similar to those types of data analyses and decisions that SCDOT materials and research engineers typically encounter.

Table 1. Statistics Team Members

Name	Position	Organization
Merrill Zwanka (Chair)	State Materials Engineer	SCDOT
Andy Johnson	State Pavement Design Engineer	SCDOT
Mike Sanders	Research Engineer	SCDOT
Bill Smyer	Quality Assurance Manager	SCDOT
Patti Gambill	Asst to Research & Materials Eng	SCDOT
Milt Fletcher	Research and Materials Engineer	SCDOT
David Law	Pavement and Materials Engineer	FHWA
Jim Burati	Principal Investigator	Clemson University

Offer Statistical Training. The PI developed and taught several statistical training workshops that covered each of the statistical guidebooks that were developed. The workshops were presented in a practical manner, with "hands on" workshop problems that were solved by the participants. In addition to "everyday" examples, a number of the examples and workshop problems were similar to the situations that the engineers encounter in their jobs.

Evaluate Statistical Software. It was originally intended that the PI and GRA would identify current statistical software packages for potential purchase by SCDOT. The Statistics Team expressed the desire to minimize the need to purchase software, and, if possible, to rely only on the use of Microsoft® Excel for any statistical calculations. After the Statistics Team decided on the topics that were to be addressed in the guidebooks and training, it was possible to accomplish all of the necessary tasks using only Excel. It was therefore not necessary to purchase any software for evaluation.

(*This page is intentionally blank.*)

CHAPTER 2 — DEVELOPMENT OF THE GUIDELINES

Introduction

The PI met with three different groups of SCDOT personnel to identify the types of statistical analyses that they typically encountered in their activities. These groups included the members of the Statistics Team as well as additional materials and research engineers and staff that were selected by the Statistics Team members. The discussions from these meetings identified the types of statistical guidelines to develop along with a general order of priority.

Summaries of Group Meetings

The PI traveled to Columbia on July 6, 2004 to interview selected SCDOT personnel to determine the statistical guidelines that needed to be developed for the project. Meetings were held with three different groups to allow for a wide range of input into the types of guidelines to develop.

Summaries of the topics presented to the PI during these meetings are presented in Exhibits 1-3. These summaries include the general types of analyses that the members of each group conducted, along with suggestions for possible items to include in the statistical guidelines.

Selection of Topics to Develop

The summaries of the interview meetings were used to identify major topics of common interest among the groups. These major topics were then used to develop a list of topics for which statistical guidelines would be developed.

For example, all three groups mentioned the need to compare two sets of data, while two groups identified other topics, including: PWL specifications, identifying outliers, basic experimental design concepts, time series analysis, and goodness of fit testing for normality. While they were each only mentioned by one group, regression analysis and graphical presentation of data were considered important to address in the guidelines.

After reviewing the summaries of all of the interviews, an outline of topics for which statistical guidelines would be developed was prepared. The topics selected for the development of guidelines are shown in Table 2.

Exhibit 1. Summary of Meeting with Group 1

Members: Merrill Zwanka, Milt Fletcher, Chad Hawkins, Toya Scipio, Aly Hussein, David Law

- Comparing two different processes or chemicals
  - How different is different?
  - Alpha levels—what they mean and how to select them
  - How to determine what sample size (*n*) to use
    - How to use OC/Power curves to determine appropriate *n*
  - Must compare both paired and independent samples

- Comparing supplier QC data to SCDOT filed tests
  - Have small samples sizes for SCDOT and large sample sizes for supplier data (in some cases SCDOT may be $n = 50$ and supplier maybe $n = 500+$)

- Comparing a set of data to a target value

- Theory and process for estimating PWL
  - Setting specification limits
  - Sample sizes
  - Risks

- Comparing nuclear gage results with core results
  - Comparing means and variances, correlation coefficient

- Regression analysis
  - Determining if the coefficients are different than zero
  - Possibly multiple linear regression analysis

- Some basic experimental design concepts
  - Primarily for in-house research projects

- How to evaluate the assumption of a normal distribution
  - Goodness of fit tests
  - Skewness

- Whenever possible, stick with Excel for computer calculations

- How to determine outliers
  - ASTM E-178

Exhibit 2. Summary of Meeting with Group 2

Members: Mike Sanders, Patti Gambill, Milt Fletcher, Terry Swygert, Leah Brigman, Temple Short

- How to develop precision and bias statements for SC-T test procedures
- How to evaluate data from AASHTO proficiency tests
  - Trends over time
    - Time series analysis
    - Plots
    - Control charts
- How to evaluate survey results
- How to best depict data in a graphical format
  - Best way to present data in a report
- Confidence intervals and level of significance
- How to develop OC curves and evaluate risks for PWL specifications
- How to compare two sets of data
- Basic experimental design
- Guidance on how to develop test and control sections

Exhibit 3. Summary of Meeting with Group 3

Members: Andy Johnson, Mike Lockman, Melissa Campbell, Bart Dominick, Wei Hong, Brian Dix

- How to consider data on FWD, structural number and subgrade modulus with respect to pavement design
- Need a test for outliers
- Currently use SigmaPlot for analysis
  - Consider frequency distribution
  - Use normal probability plot to check for normality
- How to analyze data sets for means and standard deviations for concrete flexural strength beams
- How to compare two sets of data
- How to handle aggregate sieve analyses
  - Multiple quarries
  - Willful variation vs. random variation
  - Time series analysis
- Rideability and friction (skid resistance)
  - Switching to a new laser profiler, and probably will use the mean and standard deviation of 3 runs

Table 2. Topics for which Statistical Guidelines Were Developed

Section Number	Guideline Topic
1	Basic Concepts
2	Graphical Presentation of Data
3	Statistical Inferences on One-Sample Data
4	Statistical Inferences on Two-Sample Data
5	Goodness of Fit Tests
6	Identifying Outliers
7	Time Series Data and Control Charts
8	Regression Analysis
9	Experimental Design
10	PWL Specifications: Limits and Payment
11	PWL Specifications: Risks

Preparation of Guidelines

For each guideline the PI and the GRA first prepared outlines of the topics to be covered. For some of the guideline topics the GRA developed the initial draft for review by the PI. The PI then prepared the final draft and added examples specific to SCDOT applications. For other guideline topics the PI prepared the first draft, which was reviewed by the GRA before the PI prepared the final draft.

In a number of the guidelines dealing with hypothesis testing concepts, extensive use of both computational analysis and computer simulation was used to develop the power curves and operating characteristic (OC) curves that are presented.

Throughout the guideline preparation process, current drafts of the guidelines were submitted with the quarterly progress reports for the project for review by the Statistics Team.

The guidebooks were prepared to serve as more than just “cookbooks” or procedures manuals, but also incorporated the basic underlying statistical concepts to support the procedures included in the guidebooks. The guidebooks include numerous everyday examples in addition to specific examples related to research and materials functions.

The statistical guidelines were then distributed and used with the statistical training that was conducted as part of this project. Outlines for each of the statistical guidelines that were developed for the project are included in Appendix A.

A summary of the statistical training that was conducted appears in Chapter 3.

(*This page is intentionally blank.*)

CHAPTER 3 — STATISTICAL TRAINING

Introduction

The project called for the PI to offer SCDOT research and materials engineers training on each of the statistical guidelines that were developed. A meeting was held on April 19, 2005 to update the Statistics Team on the status of the project and to decide on a delivery strategy for the training aspects of the project.

At the meeting it was decided that at that point it would be best to delay the statistical training until after the busiest portion of the summer construction season. This was done to make it easier for the selected participants to attend the training.

At the conclusion of the meeting it was agreed that the PI and Merrill Zwanka, the chair of the Statistics Team, would determine training dates that best fit everyone's schedules. October 4 & 5 was selected for initial training in statistics basics, with specific topics to be covered in one day sessions on the following dates: October 11, 13, 18, 20, 25.

Training Sessions

The initial basic statistical training was conducted October 4 and 5, 2005 as originally scheduled. The dates of subsequent training were modified based upon the number of topics that were covered at each training session. As a result, additional one-day training sessions were held on October 11, 18 and 25. Titles, attendance rosters, and the statistical guidelines covered for each of the training sessions are presented in Tables 3 through 7.

The training sessions were developed and presented by the PI. For the various training topics each participant received notebooks that included two types of handouts—the full text of the statistical concepts guideline and copies of the presentation graphics that were used during the training session. This allowed the participants to follow along with the presentation and also allowed the PI to refer to the underlying principles that are presented in the guidelines. Numerous examples and workshop problems were used during the training sessions.

Copies of the handouts of the presentation graphics for the first day of training are presented in Appendix B to illustrate the format for a typical training presentation.

The initial two-day training session on "Basic Statistical Concepts" (see Tables 3 and 4) covered the first four sections of the statistical guidelines, including

1. Basic Concepts.
2. Graphical Presentation of Data.
3. Statistical Inferences on One-Sample Data.
4. Statistical Inferences on Two-Sample Data.

This initial training provided the background and foundation necessary to understand the topics covered in subsequent training sessions.

Table 3. October 4, 2005 Training Session on "Basic Statistical Concepts I"

Statistical Guidelines Covered		
Section Number	**Title**	
1	Basic Concepts	
2	Graphical Presentation of Data	
Attendees		
Last Name	**First Name**	**Agency**
Campbell	Melissa	SCDOT
Dix	Brian	SCDOT
Duncan	Hamilton	FHWA
Eleazer	Charles	SCDOT
Fletcher	Milt	SCDOT
Gambill	Patti	SCDOT
Hallman	A Lindy	SCDOT
Hawkins	Chad	SCDOT
Hong	Wei	SCDOT
Hussein	Aly	SCDOT
Johnson	Andy	SCDOT
Law	David B	FHWA
Lockman	Michael	SCDOT
McGee	Mike	SCDOT
Miller	Paul	SCDOT
Sanders	Michael R	SCDOT
Scipio	Toya	SCDOT
Short	Temple	SCDOT
Smyer	William M	SCDOT
Swygert	Terry	SCDOT
Todd	Jerry	SCDOT
Zwanka	Merrill	SCDOT

Table 4. October 5, 2005 Training on "Basic Statistical Concepts II"

Statistical Guidelines Covered		
Section Number	**Title**	
3	Statistical Inferences on One-Sample Data	
4	Statistical Inferences on Two-Sample Data	
Attendees		
Last Name	**First Name**	**Agency**
Campbell	Melissa	SCDOT
Dix	Brian	SCDOT
Duncan	Hamilton	FHWA
Eleazer	Charles	SCDOT
Fletcher	Milt	SCDOT
Gambill	Patti	SCDOT
Hallman	A Lindy	SCDOT
Hawkins	Chad	SCDOT
Hong	Wei	SCDOT
Hussein	Aly	SCDOT
Johnson	Andy	SCDOT
Law	David B	FHWA
Lockman	Michael	SCDOT
McGee	Mike	SCDOT
Miller	Paul	SCDOT
Sanders	Michael R	SCDOT
Scipio	Toya	SCDOT
Short	Temple	SCDOT
Smyer	William M	SCDOT
Swygert	Terry	SCDOT
Todd	Jerry	SCDOT
Zwanka	Merrill	SCDOT

Table 5. October 11, 2005 Training on "Data Analysis"

Statistical Guidelines Covered		
Section Number	**Title**	
5	Goodness of Fit Tests	
6	Identifying Outliers	
7	Time Series Data and Control Charts	
Attendees		
Last Name	**First Name**	**Agency**
Brigman	Leah	SCDOT
Campbell	Melissa	SCDOT
Dix	Brian	SCDOT
Fletcher	Milt	SCDOT
Gambill	Patti	SCDOT
Hawkins	Chad	SCDOT
Hong	Wei	SCDOT
Hussein	Aly	SCDOT
Law	David B	FHWA
Lockman	Michael	SCDOT
Miller	Paul	SCDOT
Sanders	Michael R	SCDOT
Scipio	Toya	SCDOT
Short	Temple	SCDOT
Smyer	William M	SCDOT
Swygert	Terry	SCDOT
Zwanka	Merrill	SCDOT

Table 6. October 18, 2005 Training on "Regression Analysis and Experimental Design"

Statistical Guidelines Covered		
Section Number	**Title**	
8	Regression Analysis	
9	Experimental Design	
Attendees		
Last Name	**First Name**	**Agency**
Brigman	Leah	SCDOT
Campbell	Melissa	SCDOT
Dix	Brian	SCDOT
Fletcher	Milt	SCDOT
Gambill	Patti	SCDOT
Hawkins	Chad	SCDOT
Hong	Wei	SCDOT
Hussein	Aly	SCDOT
Law	David B	FHWA
Lockman	Michael	SCDOT
Miller	Paul	SCDOT
Sanders	Michael R	SCDOT
Scipio	Toya	SCDOT
Short	Temple	SCDOT
Smyer	William M	SCDOT
Swygert	Terry	SCDOT
Zwanka	Merrill	SCDOT

Table 7. October 25, 2005 Training on "Specification Development and Risk Analysis"

Statistical Guidelines Covered		
Section Number	**Title**	
10	PWL Specifications: Limits and Payment	
11	PWL Specifications: Risks	
Attendees		
Last Name	**First Name**	**Agency**
Brigman	Leah	SCDOT
Campbell	Melissa	SCDOT
Dix	Brian	SCDOT
Duncan	Hamilton	FHWA
Fletcher	Milt	SCDOT
Gambill	Patti	SCDOT
Hawkins	Chad	SCDOT
Hong	Wei	SCDOT
Hussein	Aly	SCDOT
Johnson	Andy	SCDOT
Law	David B	FHWA
Lockman	Michael	SCDOT
Miller	Paul	SCDOT
Sanders	Michael R	SCDOT
Scipio	Toya	SCDOT
Short	Temple	SCDOT
Smyer	William M	SCDOT
Swygert	Terry	SCDOT
Zwanka	Merrill	SCDOT

CHAPTER 4 — EXCEL PROCEDURES MANUAL

Introduction

The Statistics Team indicated that, if at all possible, they would prefer to use only Microsoft® Excel for all statistical calculations. This is because it is readily available within the SCDOT and would require no additional expense for software purchases. It was possible to develop all statistical guidelines using only Excel. Therefore, no evaluation software was purchased with project funds.

As noted in Chapter 1, the statistical guidebooks that were developed serve as more than just "cookbooks" or procedures manuals, but also incorporate the basic underlying statistical concepts to support the procedures included in the guidebooks. The SCDOT also desired to have more specific step-by-step instructions regarding how to use Excel to perform certain statistical analyses.

It was decided that in some cases it would be possible to develop such step-by-step procedures in Excel as long as they were supplemented by references to the more thorough coverage of underlying concepts presented in the statistical guidelines. For some cases, such as experimental design, it was decided that the statistical concepts in the guidelines were absolutely necessary and that a simplified step-by-step procedure was not an appropriate option. In other cases, such as the general coverage of the graphical presentation of data, it was decided that the coverage presented in the statistical guidelines was the best method to convey the necessary information.

An Excel procedures manual was developed to present the step-by-step procedures to perform selected statistical analyses. The analyses that were selected for inclusion in the Excel procedures manual met two criteria. First, they needed to be analyses that would be done on a fairly common basis; and second, they needed to be analyses that could be followed with only limited reference to the underlying statistical concepts in the guidelines.

The procedures manual included procedures for the following analyses:

- Comparing Two Independent Samples.
- Comparing Results from Split Samples.
- Comparing Sample Results to a Standard or Target Value.
- Fitting a Line to a Set of (X, Y) Data.

The full text of the procedures manual is presented in Appendix C.

(*This page is intentionally blank.*)

CHAPTER 5 — SUMMARY AND OBSERVATIONS

Summary

Recognizing the need for materials and research engineers to have a sound understanding of fundamental statistical principles, the SCDOT contracted with Clemson University to develop a set of statistical analysis guidelines. These guidelines would provide a more structured methodology for the types of critical decisions that materials and research engineers must frequently make.

Three major tasks were completed during the project. These tasks include:

1. Identify SCDOT's training needs regarding statistical concepts.
2. Prepare statistical guidebooks for the identified topics.
3. Develop and offer training workshops on each of the statistical guidebooks.

A SCDOT "Statistics Team," consisting of selected SCDOT and FHWA personnel was formed and was charged to oversee the project. Interviews were then conducted with three different groups of SCDOT personnel. The input from these interviews and the Statistics Team was used to determine for which statistical concepts guidebooks were needed.

After receiving approval from the Statistics Team of topic outlines, the PI and GRA then prepared the statistical guidebooks. The PI developed and taught several statistical training workshops that covered each of the statistical guidebooks that were developed. Input from the training sessions was then incorporated into the final statistical guidebooks. The guidebooks serve as more than just "cookbooks" or procedures manuals, but also incorporate the basic underlying statistical concepts to support the procedures included in the guidebooks.

Finally, at the request of the Statistics Team, the PI developed an Excel procedures manual that presents the step-by-step procedures to perform selected statistical analyses, including: Comparing Two Independent Samples, Comparing Results from Split Samples, Comparing Sample Results to a Standard or Target Value, and Fitting a Line to a Set of (X, Y) Data.

Observations

The process of using a Statistics Team and interviews of SCDOT personnel to determine the guidelines to develop seemed to be an effective and efficient way to make this determination. From the training sessions the guidebooks appeared to be written at an appropriate level for engineering and materials engineers. The training sessions covered the necessary concepts and the full-day format allowed for relatively easy scheduling for the training participants. However, given the nature the concepts presented, it might have been more effective to conduct the training sessions as a longer series of 2- to 4-hours workshops. It might have made it easier to comprehend the large amount of information included in the various statistical guidelines. Such a format, however, would be much more difficult for the nearly 20 participants to schedule.

(*This page is intentionally blank.*)

APPENDIX A — OUTLINES OF STATISTICAL GUIDELINES DEVELOPED

Section 1: Basic Concepts

1.1 Population And Sample
- 1.1.1 Definitions
- 1.1.2 Examples
- 1.1.3 Inferences

1.2 Random Variables
- 1.2.1 Examples

1.3 Probability
- 1.3.1 Definitions
- 1.3.2 Discrete Probability Distributions
- 1.3.3 Continuous Probability Distributions

1.4 Sampling
- 1.4.1 Types of Sampling Plans

1.5 Population Parameters And Sample Statistics
- 1.5.1 Mean
- 1.5.2 Median
- 1.5.3 Variance
- 1.5.4 Standard Deviation
- 1.5.5 Skewness
- 1.5.6 Characterizing Distributions

1.6 Simple Characteristics Of A Distribution
- 1.6.1 Location
- 1.6.2 Variability (or Spread)
- 1.6.3 Symmetry and Skewness
- 1.6.4 Mode

1.7 Estimation

1.8 Hypothesis Testing
- 1.8.1 Examples
- 1.8.2 Hypothesis Test Procedures
- 1.8.3 Types of Error
- 1.8.4 Decision Rules
- 1.8.5 Operating Characteristic (OC) Curves
- 1.8.6 p-value

APPENDIX A (continued)

Section 2: Graphical Presentation Of Data

APPENDIX A (continued)

Section 3: Statistical Inferences Of One-Sample Data

APPENDIX A (continued)

Section 4: Statistical Inferences Of Two-Sample Data

4.1 Comparing Means—Background

4.2 Comparing Means For Paired Data
- 4.2.1 Application Examples
- 4.2.2 Procedure
- 4.2.3 Illustrative Examples

4.3 Comparing Means For Independent Data
- 4.3.1 Application Examples
- 4.3.2 Hypothesis Testing Procedures
- 4.3.3 When Both Population Variances Are Known
- 4.3.4 Known Variances—A Special Case When n = 1 (the D2S Method)
- 4.3.5 When Both Variances Are Unknown But Equal
- 4.3.6 When Both Variances Are Entirely Unknown

4.4 Comparing Variances
- 4.4.1 Testing Procedures for Two-Sided Test
- 4.4.2 Illustrative Examples

4.5 Power Of The Tests
- 4.5.1 Power of the two-sample z-test
- 4.5.2 Power of the D2S Method
- 4.5.3 Power of the Paired t-test
- 4.5.4 Power of the Two Sample t-test with Equal Variances
- 4.5.5 Power of the F-test for Variances—Equal Sample Sizes
- 4.5.6 Power of the F-test for Variances—Unequal Sample Sizes

4.6 Selecting A Sample Size
- 4.6.1 Illustrative Examples

Section 5 Goodness-of-Fit Tests (Particularly for Normality)

5.1 Graphical Methods for Checking for Normality
- 5.1.1 Histograms
- 5.1.2 Box-Plot
- 5.1.3 Normal Probability Paper
- 5.1.4 Quantile-Quantile (qq) Plot (Normal Probability Plot)

5.2 Testing For Goodness of Fit (GOF)
- 5.2.1 Chi-Square Test
- 5.2.2 Kolmogorov-Smirnov (K-S) Test
- 5.2.3 Lilliefors Test for Normality
- 5.2.4 Anderson-Darling Test for Normality
- 5.2.5 Shapiro-Wilk Test for Normality
- 5.2.6 Recommended GOF Test for Normality

APPENDIX A (continued)

Section 6: Identifying Outliers

APPENDIX A (continued)

Section 7: Time Series Data and Control Charts

APPENDIX A (continued)

APPENDIX A (continued)

Section 10: PWL Specifications: Limits And Payment

APPENDIX A (continued)

Section 11: PWL Specifications: Risks

(*This page is intentionally blank.*)

APPENDIX B — Example Presentation Graphics: Basic Concepts

Basic Concepts

SCDOT

Statistical Training

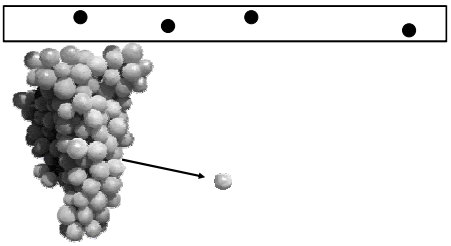
Topics

- Populations & Samples
- Random Variables – distributions
- Sampling
- Population Parameters & Sample Statistics

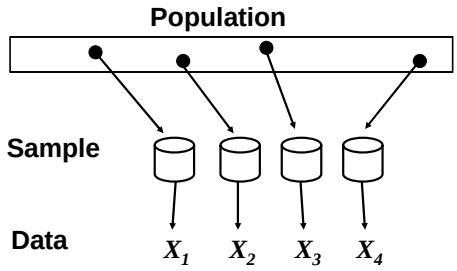
Topics – Basic Introductions

- Characteristics of a Distribution
  - Center, Spread, Skewness
- Estimation
  - Point & Interval
- Hypothesis Testing Concepts
  - Types of Errors

Populations & Samples



Definitions



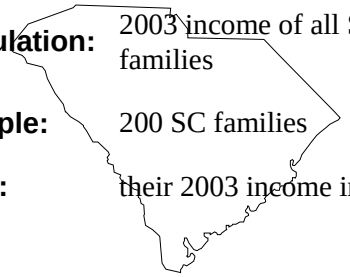
Example: Is a coin fair?

Population: all possible coin flips

Sample: 10 coin flips

Data: Number of heads/tails

Example: SC Family Income



Population: 2003 income of all SC families

Sample: 200 SC families

Data: their 2003 income in $

Example: Light Bulb Life

Population: all light bulbs produced by a manufacturer

Sample: 100 light bulbs

Data: 100 lives, in hours

Two Cases in Statistics

- We know the population & want to investigate its properties
- We have a sample & want to make inferences about the population

Random Variable

- A **random variable** can be thought of as the numeric result of operating a non-deterministic mechanism or performing a non-deterministic experiment to generate a random result.

A Random Variable (r.v.)

- cannot be assigned a value;
- does not describe the actual outcome of a particular experiment
- but rather describes the possible, as-yet-undetermined outcomes

Random Variable

- For example, a random variable can be used to describe the process of rolling a fair die and the possible outcomes { 1, 2, 3, 4, 5, 6 }.

Discrete *r.v.*

Random Variable

- Or, a random variable can be used to describe the possible outcomes for an aggregate gradation test for a particular aggregate.

Continuous *r.v.*

Probability

- Likelihood of occurrence of each possible outcome of a r.v.
- $0 \le P \le 1$.
- Three different definitions:
  - Based on logic.
  - Based on experience
  - Subjective probability

Definition of Probability

- **Based on Logic:**

$$\text{Probability } (A) = \frac{\text{number of ways event } (A) \text{ can occur}}{\text{total number of possible outcomes}}$$

Definition of Probability

- **Based on Experience:**
  (*Relative Frequency*)

$$\text{Probability}(A) = \lim_{n\to\infty} \frac{x}{n}$$

Definition of Probability

- **Subjective Probability:**
  quantifiable subjective level of belief ranging from 0 (complete disbelief) to 1 (complete belief)

Distributions of r.v.'s

- **Discrete:** probability distribution
- **Continuous:** probability density function, $f(x)$, or p.d.f.

Discrete Probability Distribution

- Example:
  - Rolling one die
  - Rolling two dice

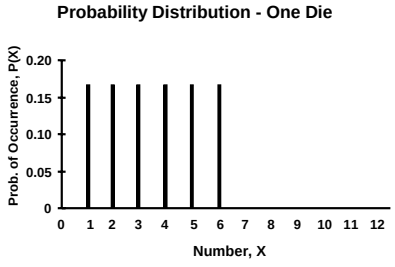
Rolling One Die

X	No. of ways	Pr(*X*)
1		
2		
3		
4		
5		
6		

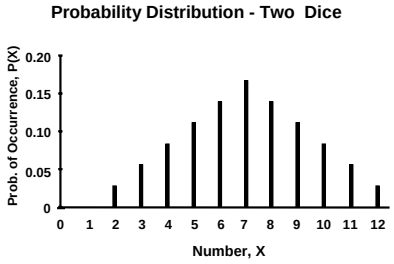
Rolling Two Dice

X	No. of ways	Pr(*X*)
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

Rolling One Die

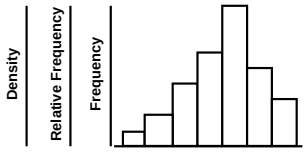


Rolling Two Die



Continuous Probability Distribution

- Frequency
- Relative Frequency
- Density

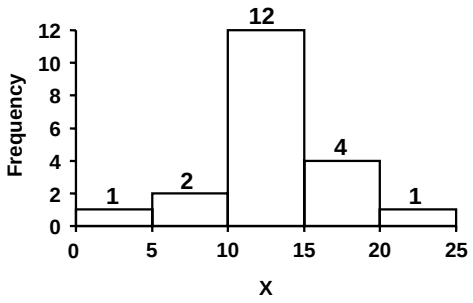


Example: Histograms

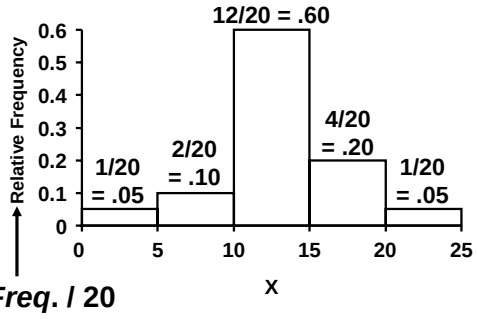
- **Data (20 values):**

  4.4, 7.3, 11.5, 10.5, 16.8, 14.1, 12.3, 20.5, 11.5, 13.2, 13.4, 8.5, 11.7, 15.8, 12.1, 18.1, 12.3, 14.5, 16.5, 13.9

Frequency

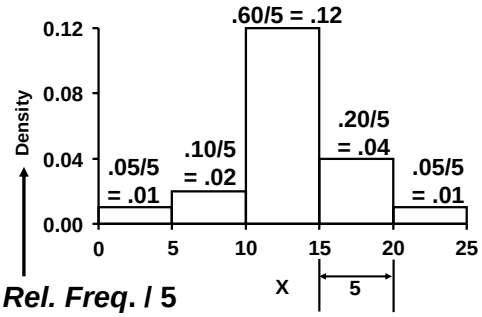


Relative Frequency



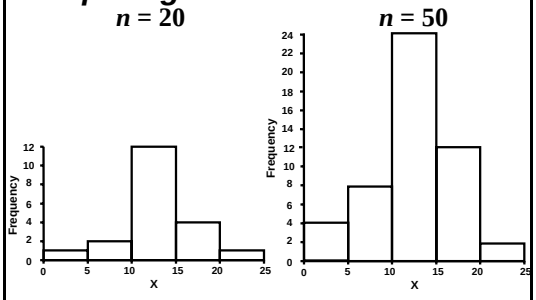
Freq. / 20

Density

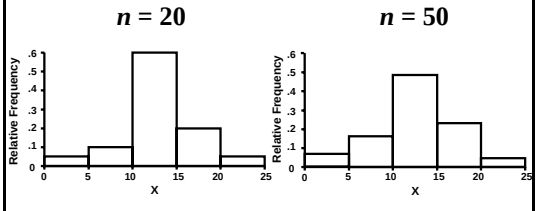


Rel. Freq. / 5

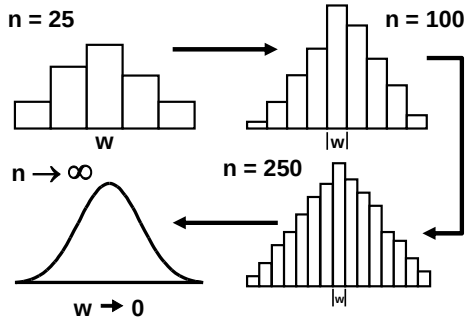
Comparing Distributions



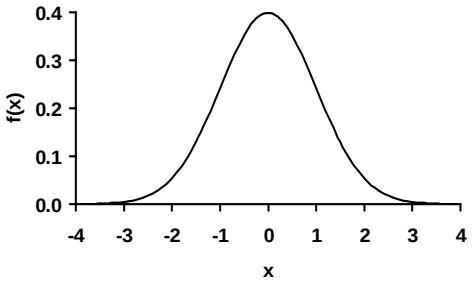
Comparing Distributions



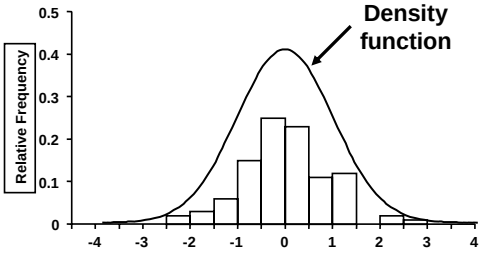
Limit—Smooth Curve



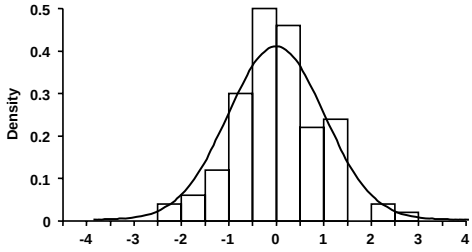
Density Function, $f(x)$, (pdf)



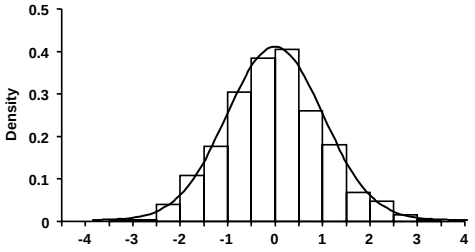
Rel Freq Histogram, n = 100



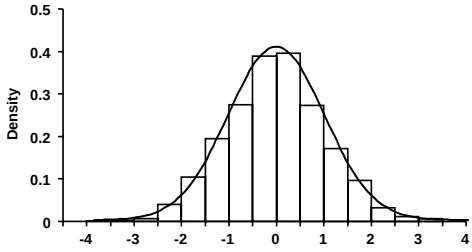
Density Histogram, n = 100



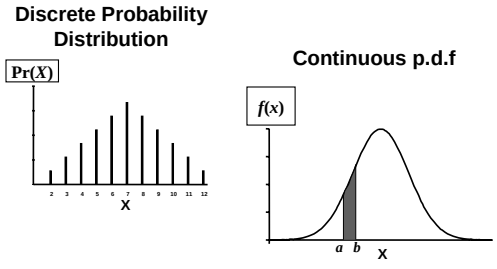
Density Histogram, n = 500



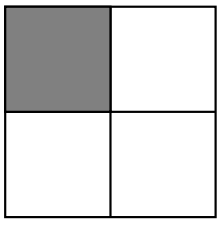
Density Histogram, n = 2000



Probabilities?



Relative Frequency (Area?)



Probabilities

$$\Pr(a \le X \le b) = \frac{\int_a^b f(x)dx}{\int_{-\infty}^{+\infty} f(x)dx} = \int_a^b f(x)dx$$

1.0

Statistical Training 1. Basic Concepts 39

Sampling

- (*Simple*) Random Sampling
- Stratified Random Sampling
- Systematic Sampling

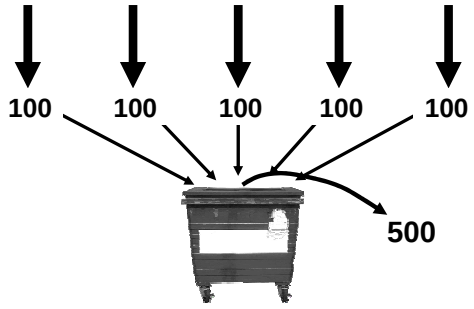
Statistical Training 1. Basic Concepts 40

Validity of Sample Data

- Random Sampling (i.e., unbiased)
- Same population (i.e., the same conditions)

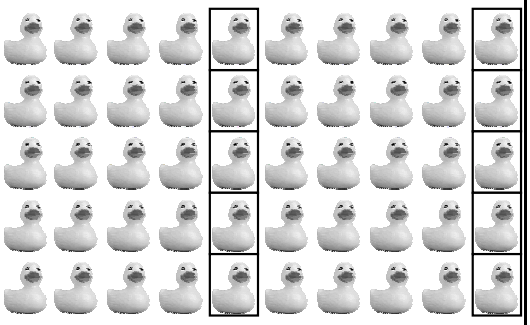
Statistical Training 1. Basic Concepts 41

Example: Sampling, n = 500



Statistical Training 1. Basic Concepts 42

Systematic Sampling

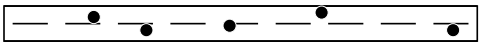


Statistical Training 1. Basic Concepts 43

Where to Select Samples?

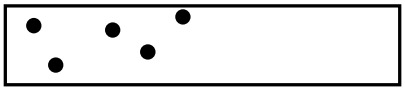
~~Systematically~~

Randomly



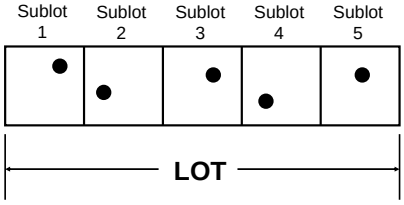
Statistical Training 1. Basic Concepts 44

Random Sampling



Statistical Training 1. Basic Concepts 45

Stratified Random Sampling



Statistical Training 1. Basic Concepts 46

Parameters & Statistics

- Numerical measures that describe the population are called **population parameters**
- The parameters are estimated by **sample statistics**, which are quantities based only on the sample

What Measures Are Needed?

- Location, or **center**
- Spread, or **variability**
- Presence or lack of symmetry, **skewness**

Center or Location

- **Mean**
  - **Parameter:** Population Mean, μ
  - **Statistic:** Sample Mean, $\overline{X}$
- **Median**, *M*

Population Mean

- The average of all of the measurements in the population
- The weighted average of all *distinct* values the population can take, with the weights being their probabilities of occurrence

Example: Population Mean

- **Option 1:** with a *finite* population, the mean is the simple average of all values in the population.
- Say the population is {90, 90, 90, 90, 10}
- What is the population mean?

Example: Population Mean

- Since we have a finite population, we can use:

$$\mu = \frac{\sum_{\text{all } X} X}{N}$$

$$\mu = \frac{90+90+90+90+10}{5} = 74.0$$

Example: Population Mean

- With either a finite population or a known probability distribution:

$$\mu = \sum_{\text{all distinct } X \text{ values}} P(X)X$$

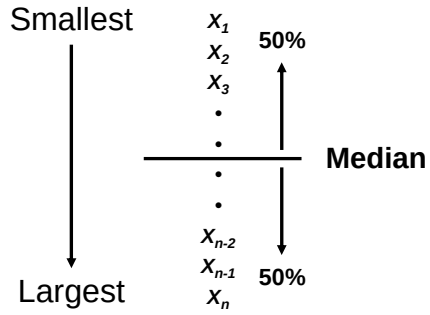
$$\mu = 0.8(90) + 0.2(10) = 74.0$$

Sample Mean

- The average of all of the measurements in the sample:

$$\overline{x} = \frac{\sum_{\text{all } i} x_i}{n}$$

Median



Example

Data Set A

1, 2, 3, 4, 5, 6, 7

Mean = Median =

Data Set B

1, 2, 3, 4, 5, 6, 84

Mean = Median =

Example

Data Set C

8, 9, 10, 11, 12, 13, 14

Median =

Data Set B

1, 2, 3, 4, 5, 6, 7, 8

Median =

Spread or Variability

- **Variance**
  - **Parameter:** Population Variance, σ^2
  - **Statistic:** Sample Variance, s^2
- **Standard Deviation**
  - **Parameter:** Population Standard Deviation, σ
  - **Statistic:** Sample Standard Deviation, s

Variance

$$\sigma^2 = \frac{\sum_{\text{all } X} (X - \mu)^2}{N}$$ **Population**

Sample $$s^2 = \frac{\sum_{\text{all } x_i} (x_i - \overline{x})^2}{n - 1}$$

Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{\text{all } X} (X - \mu)^2}{N}}$$ **Population**

Sample $$s = \sqrt{\frac{\sum_{\text{all } x_i} (x_i - \overline{x})^2}{n - 1}}$$

Degrees of Freedom

$$\sigma = \sqrt{\frac{\sum_{\text{all } X} (X - \mu)^2}{\boldsymbol{N}}}$$ **Population**

Sample $$s = \sqrt{\frac{\sum_{\text{all } x_i} (x_i - \overline{x})^2}{\boldsymbol{n - 1}}}$$

Degrees of Freedom

Suppose we determine the mean of four data points to be 5.00.

How many independent data points do we now have?

In other words, how many data points have freedom with regard to their numerical value?

Symmetry or Asymmetry

- **Skewness**
  - **Parameter:** Population Skewness, γ_1
  - **Statistic:** Skewness Coefficient, g_1

Skewness

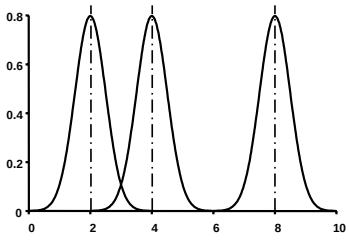
$$\gamma_1 = \frac{E[(X-\mu)^3]}{\sigma^3}$$ **Population**

Sample

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{s^3}$$

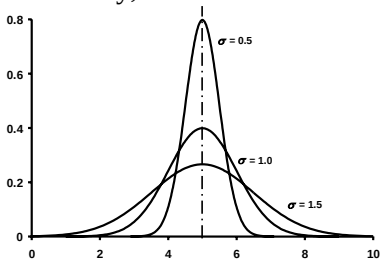
Distribution Characteristics

- Location or Center, **Mean**



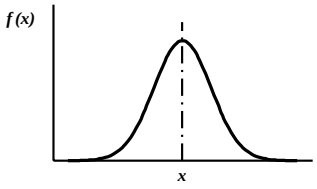
Distribution Characteristics

- Variability, **Standard Deviation**



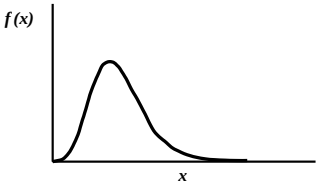
Distribution Characteristics

- Symmetry, **Skewness** = 0



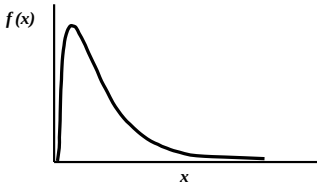
Distribution Characteristics

- Some Positive Skewness



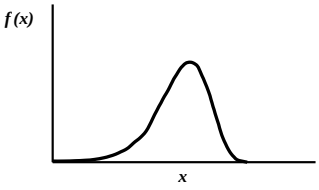
Distribution Characteristics

- Severe Positive Skewness



Distribution Characteristics

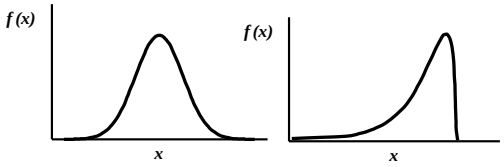
- Some Negative Skewness



Distribution Characteristics

- Mode: any local maximum or peak

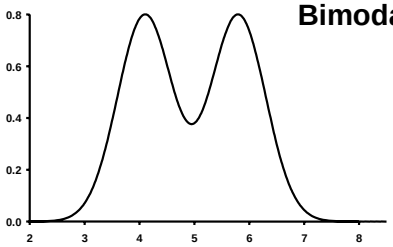
Unimodal



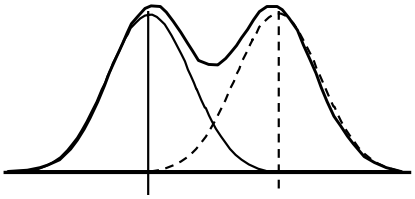
Distribution Characteristics

- Mode: any local maximum or peak

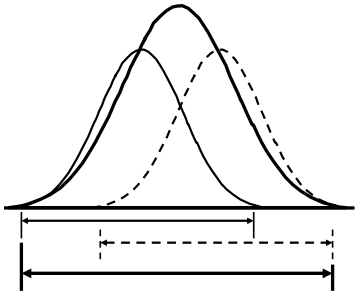
Bimodal



Obvious



Unimodal Distribution, but?



Statistical Inference

- **Estimation**
  - Point estimates
  - Interval estimates
- **Hypothesis Testing**

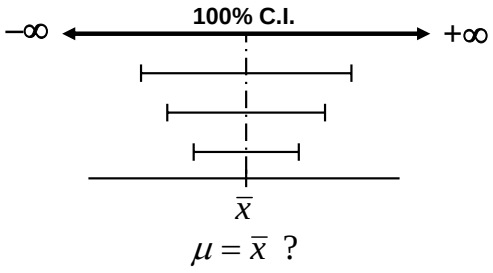
Point Estimate

- Single value that "best" approximates the target quantity
- For example, Parameters & Statistics:
  - σ^2 and s^2
  - μ and $\bar{x}$

Interval Estimate

- Point estimate is easy, but gives no indication of its accuracy
- Confidence Interval
- Level of Significance

Confidence Intervals



$\mu = \bar{x}$?

Hypothesis Testing

- Making Decisions, e.g., Coin Flips
- Is this coin fair?

Hypothesis Testing

- Is this coin fair?

Hypothesis Testing

- Is this coin fair?

Hypothesis Testing

- Is this coin fair?

Hypothesis Testing

- Is this coin fair?

Hypothesis Testing

- Is this coin fair?

Hypothesis Test Procedures

- Null Hypothesis, $\boldsymbol{H_0}$
- Alternate Hypothesis, $\boldsymbol{H_a}$
- Decision Rule
- Test Statistic
- Rejection Region

Hypotheses

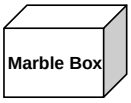
- Coin Flip Example:
  - H_0: $p = 0.5$
  - H_a: $p \neq 0.5$
- SC Family Income Example:
  - H_0: $\text{Income}_{2004} > \text{Income}_{2003}$
  - H_a: $\text{Income}_{2004} \leq \text{Income}_{2003}$

Decision Rule

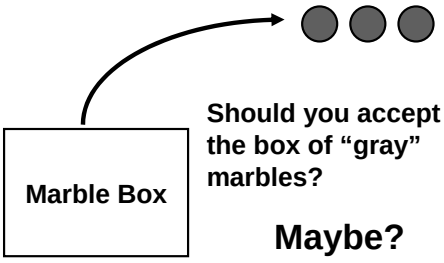
- Reject H_0 if the test statistic is in the rejection region
- Test Statistic: $$z = \frac{\bar{x} - 0.5}{s_{\bar{x}}}$$
- Reject if the |z| is *too big*.
- How do we decide what is too big?

Risks–Simple Illustration

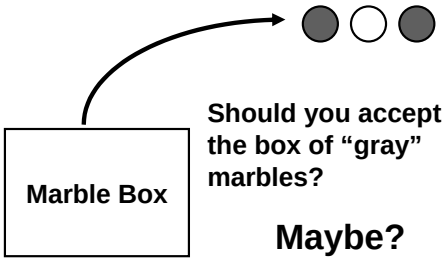
- We have a box of 10 marbles
- We want the box to have **<u>at least</u>** 8 **gray** marbles
- We take a **sample** of 3 marbles to determine whether to accept or reject the box



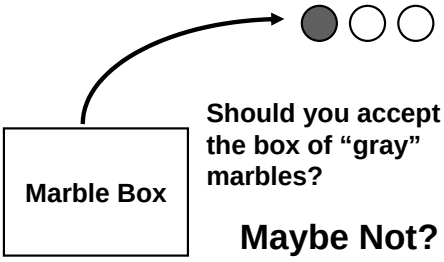
Risks (cont.)



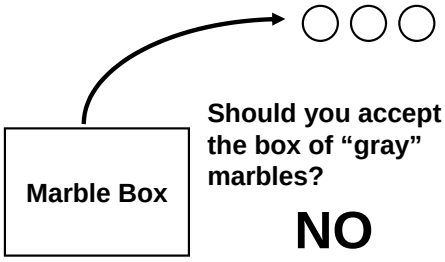
Risks -- Example (cont.)



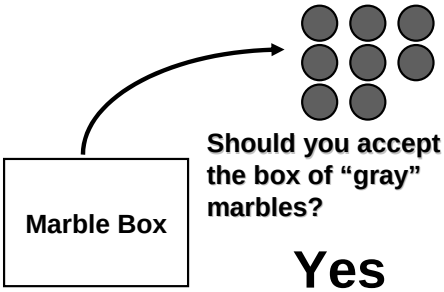
Risks -- Example (cont.)



Risks -- Example (cont.)



Risks -- Example (cont.)



Types of Errors

- **Type I Error:**
  - Rejecting H_0 when it is actually true
  - Probability of Type I Error is α
- **Type II Error**:
  - Accepting H_0 when it is actually false
  - Probability of Type II Error is β

Risks

	H_0 True	H_0 False
Accept	✓	Type II β
Reject	Type I α	✓

Importance of α and β

- α is the **Significance Level**
  - A smaller α makes a decision to reject H_0 more significant.
- $(1 - \beta)$ is the **Power** of the Test
  - Measures the ability to reject H_0 when it is actually false.

Decision Rule

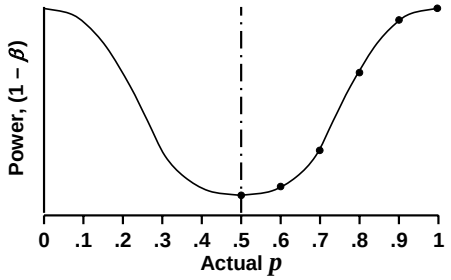
- Would like a rule that gives the lowest α and β.
- Can only control one, usually α
- The power varies depending upon the true parameter of the population.

Example: Power

- Coin Flip Example: suppose the coin is not fair, i.e., $p \neq 0.5$
- Which of the following actual p values should be easier to detect as different from 0.5?

$p = 0.55$ or $p = 0.85$?

Example: Power Curve



Key Points

- Population vs. Sample
- Random Variables
- Distributions & Density Functions
- Sampling
- Parameters & Sample Statistics

Key Points

- Distribution Center, Spread & Shape
- Estimation: Point & Interval
- Hypothesis Testing: α, β & Power

(*This page is intentionally blank.*)

APPENDIX C — Procedures Manual

Procedure: **Comparing Two Independent Samples**

Purpose: Compare two sets of data to see if it is reasonable to assume that they came from the same population; that is, if they can be assumed to come from normal populations with equal variances and equal means.

Background: Statistical Concepts Section 4: Inferences on Two-Sample Data

Steps Involved:

Step 1: Use an *F*-test to determine whether to assume the variances are equal or unequal.

Enter the data for the two independent samples into two columns in the Excel spreadsheet. Label each column appropriately. In Figure 1 we are comparing Contractor tests for air voids with independently obtained SCDOT tests.

	A	B
1	**Contractor**	**SCDOT**
2	6.42	7.52
3	7.18	11.38
4	5.04	9.2
5	4.56	5.32
6	7.12	3.18
7	7.98	
8	6.32	
9	6.08	
10	5.92	
11	5.78	

Figure 1. Air Voids Example

Step 2: Click on Tools at the top of the Excel screen, and then click on Data Analysis … to open the menu shown in Figure 2. (Note, if there is no Data Analysis … option shown, then it can be added to the menu by first clicking on Add Ins … and then checking Analysis ToolPack.)

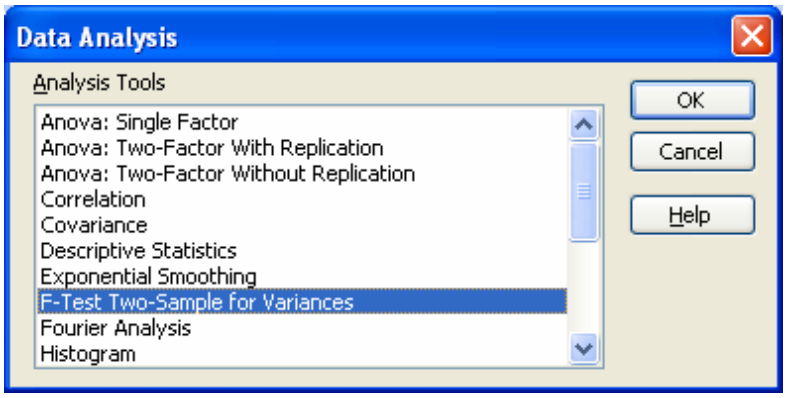


Figure 2. Data Analysis Menu

Step 3: Highlight F-Test: Two-Sample for Variances and click on OK to open the menu in Figure 3 and allow the *F*-test to be conducted.

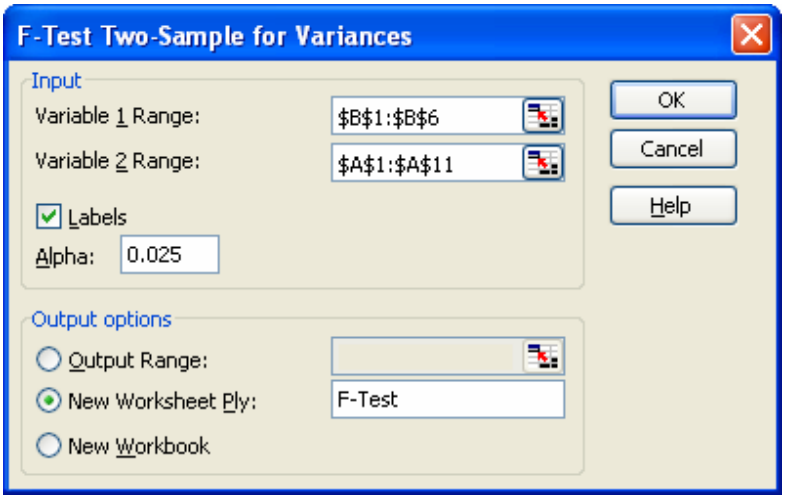


Figure 3. *F*-Test Menu

Enter the spreadsheet range (Variable 1 Range: and Variable 2 Range: in Figure 3) for the SCDOT and Contractor data sets. Check the Labels box if you have labels included in the variable ranges.

Step 4: Select a level of significance, α, for the comparison. This is shown as Alpha: in Figure 3. The level of significance is the probability that you will incorrectly declare the variances to be different when they are, in fact, equal. A typical α value would be 0.05. But, we are performing a two-sided test. That is, we are testing whether or not the variances are equal. Excel conducts a one-sided test. That is, it tests whether or not one variance is larger than the other variance. Therefore, we need to enter $\alpha/2$ rather than the α value that we selected. In Figure 3, $\alpha = 0.05$ has been selected for the comparison, and therefore 0.05/2 = 0.025 is entered for Alpha.

In Figure 3, the circle has been checked for New Worksheet Ply: and the name of the new worksheet (where the *F*-test results will be displayed) has been entered as F-Test.

Step 5: Click OK and Excel performs the *F*-test and creates the new worksheet F-Test that includes the results shown in Figure 4.

	A	B	C
1	F-Test Two-Sample for Variances		
2			
3		*SCDOT*	*Contractor*
4	Mean	7.32	6.24
5	Variance	10.2994	1.036266667
6	Observations	5	10
7	df	4	9
8	F	9.938947504	
9	P(F<=f) one-tail	0.002329313	
10	F Critical one-tail	4.718078458	

Figure 4. Excel Results for the Two Sample *F*-Test

Step 6: Interpret the *F*-test results. The Excel results in Figure 4 show both the SCDOT and Contractor sample means and sample variances along with the number of observations in each sample.

There are two ways to use Figure 4 to determine the results of the *F*-test. We can use either of the following to conclude that the sample variances are not equal:

1. Conclude the variances are not equal when the F value that is shown is greater than the F Critical one-tail value that is shown. In Figure 4, we would conclude that the variances are not equal since 9.939 > 4.718.

2. Conclude the variances are not equal when the P(F<=f) one-tail value shown is less than the $\alpha/2$ value that we had selected. In Figure 4, we would conclude that the variances are not equal since 0.0023 < 0.025.

If 1 and 2 are not true, then we conclude that there is no reason to believe the variances are not equal and, therefore, we assume that the variances are equal.

Step 7: Use a *t*-test to determine whether to assume the means are equal or unequal. Note that the manner in which the *t*-test is conducted depends upon whether the variances are assumed to be equal or unequal. If in *Step 6* we conclude that the variances were assumed to be EQUAL, then the *t*-test for equal variances must be used, and we will follow the steps beginning with *Step 8a* and proceed from there. If in *Step 6* we concluded that the variances of the two samples were assumed to be NOT EQUAL, then the *t*-test for unequal variances must be used, and we will skip to *Step 8b* and continue with the steps that follow it.

Step 8a: Use the *t*-test for equal variances to determine whether to assume that the samples means are equal or unequal. To illustrate this process, we will use the asphalt content data in Figure 5. The *F*-test has already been performed on these two samples and it was concluded that the variances can be assumed to be equal.

	A	B
1	**Contractor**	**SCDOT**
2	6.41	5.42
3	6.23	5.78
4	6.08	6.23
5	6.55	5.38
6	6.11	5.62
7	5.97	5.79
8	6.28	
9	6.07	
10	5.92	
11	5.76	
12	6.06	
13	5.71	

Figure 5. Asphalt Content Example

Click on Tools at the top of the Excel screen, and then click on Data Analysis … to open the menu shown in Figure 6.

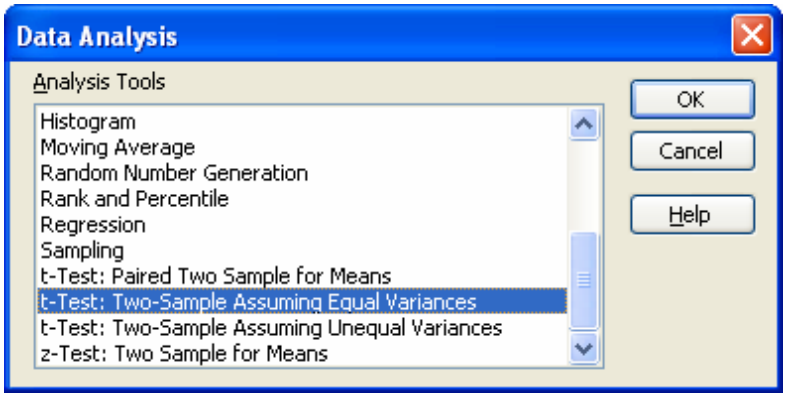


Figure 6. Data Analysis Menu

Step 9a: Highlight t-Test: Two-Sample Assuming Equal Variances and click on OK to open the menu in Figure 7 and allow the *t*-test to be conducted.

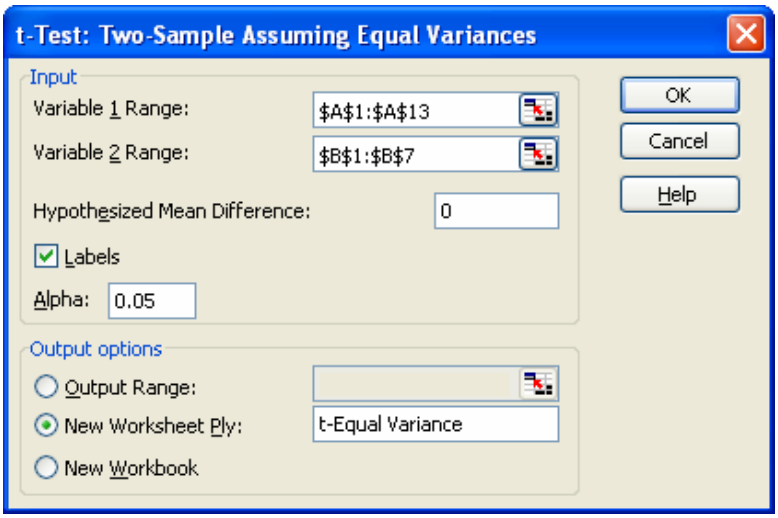


Figure 7. Two Sample Equal Variance *t*-Test Menu

Enter the spreadsheet range (Variable 1 Range: and Variable 2 Range: in Figure 7) for the Contractor and SCDOT data sets. Check the Labels box if you have labels included in the variable ranges.

Step 10a: Select a level of significance, α, for the comparison. This is shown as Alpha: in Figure 7. A typical α value would be 0.05, which has been entered in Figure 7.

In Figure 7, the circle has been checked for New Worksheet Ply: and the name of the new worksheet (where the *t*-test results will be displayed) has been entered as t-Equal Variance.

Step 11a: Click OK and Excel performs the *t*-test and creates the new worksheet t-Equal Variance that includes the results shown in Figure 8.

	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		*Contractor*	*SCDOT*
4	Mean	6.095833333	5.703333333
5	Variance	0.060699242	0.096506667
6	Observations	12	6
7	Pooled Variance	0.071889063	
8	Hypothesized Mean Difference	0	
9	df	16	
10	t Stat	2.927778699	
11	P(T<=t) one-tail	0.00492782	
12	t Critical one-tail	1.745883669	
13	P(T<=t) two-tail	0.00985564	
14	t Critical two-tail	2.119905285	

Figure 8. Excel Results for the Two Sample Equal Variance *t*-Test

Step 12a: Interpret the *t*-test results. The Excel results in Figure 8 show both the SCDOT and Contractor sample means and sample variances along with the number of observations in each sample. Figure 8 shows values for a one-tail test, i.e., testing whether or not one mean is larger than the other mean. It also shows values for a two-tail test, i.e., testing whether or not the means are equal. Since we are generally interested in whether or not the means are equal, we will usually use the values for the two-tailed test, and that is what we will consider here.

There are two ways to use Figure 8 to determine the results of the *t*-test. We can use either of the following to conclude that the sample means are not equal:

1. Conclude the means are not equal when the absolute value of the t Stat value that is shown is greater than the t Critical two-tail value that is shown. In Figure 8, we would conclude that the means are not equal since $|2.928| > 2.120$.

2. Conclude the means are not equal when the P(T<=t) two-tail value shown is less than the α value that we had selected. In Figure 8, we would conclude that the means are not equal since $0.0099 < 0.05$.

If 1 and 2 are not true, then we conclude that there is no reason to believe the means are not equal and, therefore, we assume that the means are equal.

Step 8b: Use the *t*-test for unequal variances to determine whether to assume that the samples means are equal or unequal. To illustrate this process, we will use the air voids data from Figure 1 that are repeated in Figure 9. In *Step 6* we concluded from the *F*-test performed on these two samples that the variances can be assumed to be unequal.

	A	B
1	**Contractor**	**SCDOT**
2	6.42	7.52
3	7.18	11.38
4	5.04	9.2
5	4.56	5.32
6	7.12	3.18
7	7.98	
8	6.32	
9	6.08	
10	5.92	
11	5.78	

Figure 9. Air Voids Example

Click on Tools at the top of the Excel screen, and then click on Data Analysis … to open the menu shown in Figure 10.

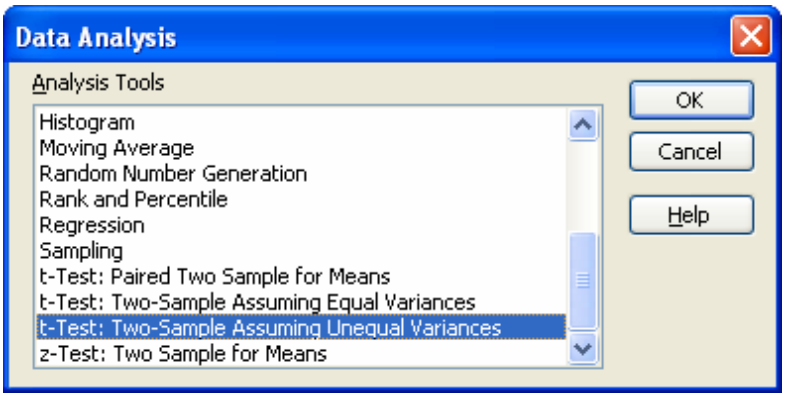


Figure 10. Data Analysis Menu

Step 9b: Highlight t-Test: Two-Sample Assuming Unequal Variances and click on OK to open the menu in Figure 11 and allow the *t*-test to be conducted.

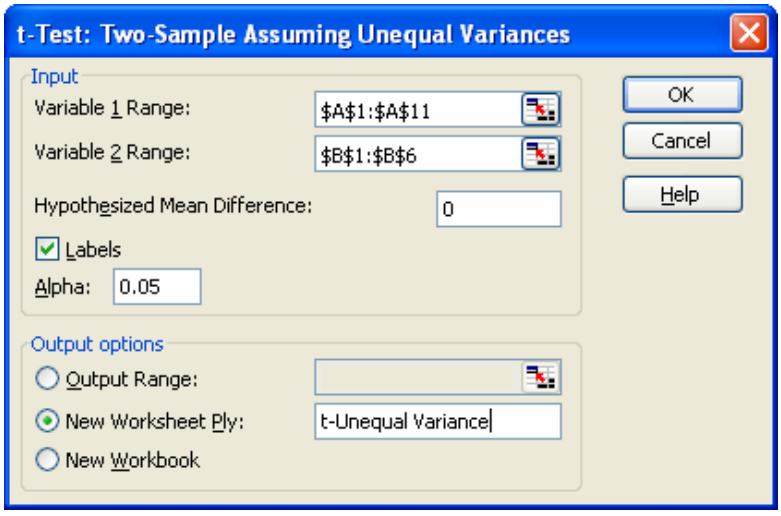


Figure 11. Two Sample Unequal Variance *t*-Test Menu

Enter the spreadsheet range (Variable 1 Range: and Variable 2 Range: in Figure 11) for the Contractor and SCDOT data sets. Check the Labels box if you have labels included in the variable ranges.

Step 10b: Select a level of significance, α, for the comparison. This is shown as Alpha: in Figure 11. A typical α value would be 0.05, which has been entered in Figure 11.

In Figure 11, the circle has been checked for New Worksheet Ply: and the name of the new worksheet (where the *t*-test results will be displayed) has been entered as t-Unequal Variance.

Step 11b: Click OK and Excel performs the *t*-test and creates the new worksheet t-Unequal Variance that includes the results shown in Figure 12.

	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		*Contractor*	*SCDOT*
4	Mean	6.24	7.32
5	Variance	1.036266667	10.2994
6	Observations	10	5
7	Hypothesized Mean Difference	0	
8	df	4	
9	t Stat	-0.734251152	
10	P(T<=t) one-tail	0.251757529	
11	t Critical one-tail	2.131846782	
12	P(T<=t) two-tail	0.503515058	
13	t Critical two-tail	2.776445105	

Figure 12. Excel Results for the Two Sample Unequal Variance *t*-Test

Step 12b: Interpret the *t*-test results. The Excel results in Figure 12 show both the SCDOT and Contractor sample means and sample variances along with the number of observations in each sample. Figure 12 shows values for a one-tail test, i.e., testing whether or not one mean is larger than the other mean. It also shows values for a two-tail test, i.e., testing whether or not the means are equal. Since we are generally interested in whether or not the means are equal, we will usually use the values for the two-tailed test, and that is what we will consider here.

There are two ways to use Figure 12 to determine the results of the *t*-test. We can use either of the following to conclude that the sample means are not equal:

1. Conclude the means are not equal when the absolute value of the t Stat value that is shown is greater than the t Critical two-tail value that is shown. In Figure 12, we would conclude that the means can be assumed to be equal since $|-0.734| \leq 2.776$.

2. Conclude the means are not equal when the P(T<=t) two-tail value shown is less than the α value that we had selected. In Figure 8, we would conclude that the means can be assumed to be equal since $0.5035 \geq 0.05$.

If 1 and 2 are not true, then we conclude that there is no reason to believe the means are not equal and, therefore, we assume that the means are equal.

Procedure: **Comparing Results from Split Samples**

Purpose: Compare the results from split samples (i.e., paired samples) to see if it is reasonable to assume that they came from normal populations with equal variances and equal means.

Background: Statistical Concepts Section 4: Inferences on Two-Sample Data

Steps Involved:

Step 1: Use a *t*-test to determine whether to assume the means are equal or unequal.

Enter the data for the split samples into two columns in the Excel spreadsheet. It is important that the Contractor and SCDOT results for each split sample appear in the same row. Label each column appropriately. In Figure 1 we are comparing Contractor's split test results for asphalt content with the SCDOT split test tests.

	A	B	C	D
1	**Sample Pair**	**SCDOT, x1**	**Contractor, x2**	**Diff., x1 -x2**
2	1	5.75	5.65	0.10
3	2	5.48	5.45	0.03
4	3	5.62	5.50	0.12
5	4	5.58	5.60	-0.02
6	5	5.60	5.53	0.07
7	6	5.55	5.51	0.04
8	7	5.86	5.78	0.08
9	8	5.49	5.40	0.09
10	9	5.67	5.68	-0.01
11	10	5.80	5.70	0.10

Figure 1. Asphalt Content Example

Step 2: Click on Tools at the top of the Excel screen, and then click on Data Analysis … to open the menu shown in Figure 2. (Note, if there is no Data Analysis … option shown, then it can be added to the menu by first clicking on Add Ins … and then checking Analysis ToolPack.)

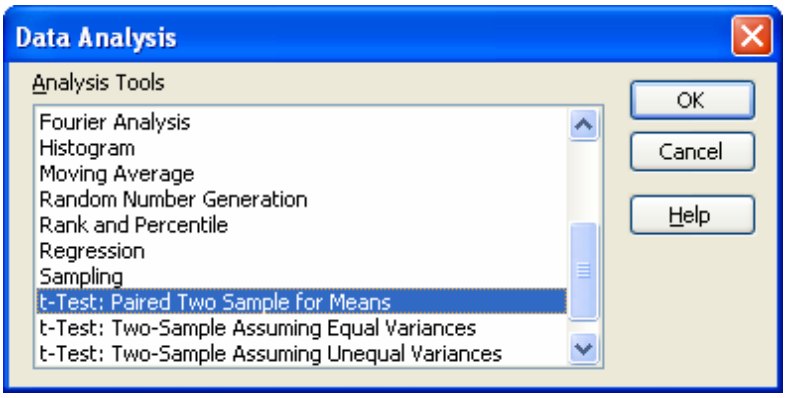


Figure 2. Data Analysis Menu

Step 3: Highlight t-Test: Paired Two Sample for Means and click on OK to open the menu in Figure 3 and allow the *t*-test to be conducted.

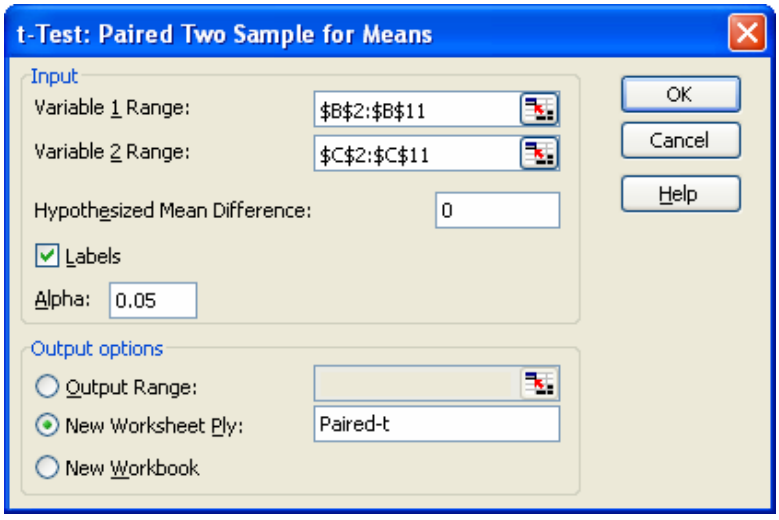


Figure 3. Paired *t*-Test Menu

Enter the spreadsheet range (Variable 1 Range: and Variable 2 Range: in Figure 3) for the SCDOT and Contractor data sets. Check the Labels box if you have labels included in the variable ranges.

Step 4: Select a level of significance, α, for the comparison. This is shown as Alpha: in Figure 3. The level of significance is the probability that you will incorrectly declare the variances to be different when they are, in fact, equal. A typical α value would be 0.05, which has been entered in Figure 3.

In Figure 3, the circle has been checked for New Worksheet Ply: and the name of the new worksheet (where the *t*-test results will be displayed) has been entered as Paired-t.

Step 5: Click OK and Excel performs the paired *t*-test and creates the new worksheet Paired-t that includes the results shown in Figure 4.

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		*SCDOT, x1*	*Contractor, x2*
4	Mean	5.64	5.58
5	Variance	0.016533333	0.014755556
6	Observations	10	10
7	Pearson Correlation	0.927634919	
8	Hypothesized Mean Difference	0	
9	df	9	
10	t Stat	3.946761087	
11	P(T<=t) one-tail	0.00168561	
12	t Critical one-tail	1.833112923	
13	P(T<=t) two-tail	0.00337122	
14	t Critical two-tail	2.262157158	

Figure 4. Excel Results for the Paired *t*-Test

Step 6: Interpret the paired *t*-test results. The Excel results in Figure 4 show both the SCDOT and Contractor sample means and sample variances along with the number of paired observations. Figure 4 shows values for a one-tail test, i.e., testing whether or not one mean is larger than the other mean. It also shows values for a two-tail test, i.e., testing whether or not the means are equal. Since we are generally interested in whether or not the means are equal, we will usually use the values for the two-tailed test, and that is what we will consider here.

There are two ways to use Figure 4 to determine the results of the paired *t*-test. We can use either of the following to conclude that the sample means are not equal:

1. Conclude the means are not equal when the absolute value of the t Stat value that is shown is greater than the t Critical two-tail value that is shown. In Figure 4, we would conclude that the means are not equal since $|3.947| > 2.262$.

2. Conclude the means are not equal when the P(T<=t) two-tail value shown is less than the α value that we had selected. In Figure 4, we would conclude that the means are not equal since $0.0034 < 0.05$.

If 1 and 2 are not true, then we conclude that there is no reason to believe the means are not equal and, therefore, we assume that the means are equal.

Step 7: Use an *F*-test to determine whether to assume the variances are equal or unequal.

For convenience the asphalt content data from Figure 1 are repeated in Figure 5.

	A	B	C	D
1	**Sample Pair**	**SCDOT, x1**	**Contractor, x2**	**Diff., x1 -x2**
2	1	5.75	5.65	0.10
3	2	5.48	5.45	0.03
4	3	5.62	5.50	0.12
5	4	5.58	5.60	-0.02
6	5	5.60	5.53	0.07
7	6	5.55	5.51	0.04
8	7	5.86	5.78	0.08
9	8	5.49	5.40	0.09
10	9	5.67	5.68	-0.01
11	10	5.80	5.70	0.10

Figure 5. Asphalt Content Example

Step 8: Click on Tools at the top of the Excel screen, and then click on Data Analysis … to open the menu shown in Figure 6.

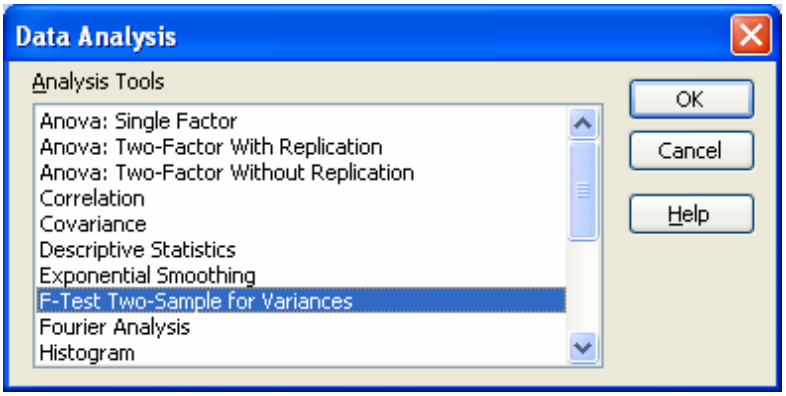


Figure 6. Data Analysis Menu

Step 9: Highlight F-Test: Two-Sample for Variances and click on OK to open the menu in Figure 7 and allow the *F*-test to be conducted.

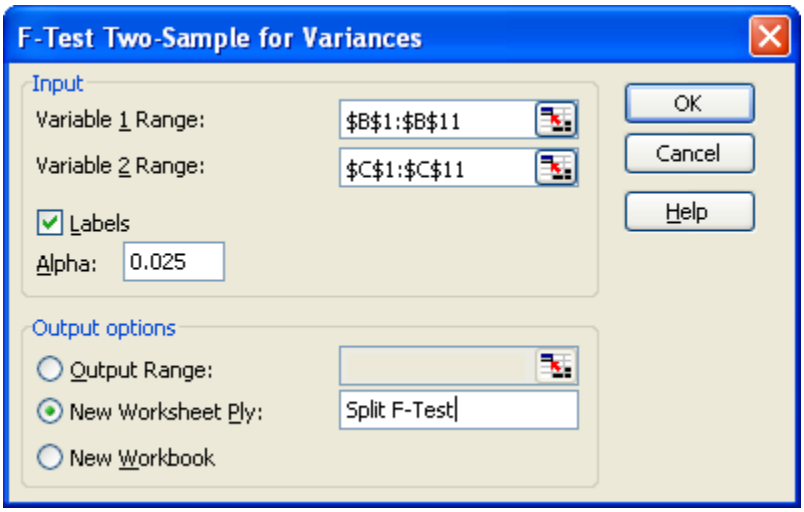


Figure 7. *F*-Test Menu

Enter the spreadsheet range (Variable 1 Range: and Variable 2 Range: in Figure 3) for the SCDOT and Contractor data sets. Check the Labels box if you have labels included in the variable ranges.

Step 10: Select a level of significance, α, for the comparison. This is shown as Alpha: in Figure 7. The level of significance is the probability that you will incorrectly declare the variances to be different when they are, in fact, equal. A typical α value would be 0.05. But, we are performing a two-sided test. That is, we are testing whether or not the variances are equal. Excel conducts a one-sided test. That is, it tests whether or not one variance is larger than the other variance. Therefore, we need to enter $\alpha/2$ rather than the α value that we selected. In Figure 7, $\alpha = 0.05$ has been selected for the comparison, and therefore 0.05/2 = 0.025 is entered for Alpha.

In Figure 7, the circle has been checked for New Worksheet Ply: and the name of the new worksheet (where the *F*-test results will be displayed) has been entered as Split F-Test.

Step 11: Click OK and Excel performs the *F*-test and creates the new worksheet F-Test that includes the results shown in Figure 8.

	A	B	C
1	F-Test Two-Sample for Variances		
2			
3		*SCDOT, x1*	*Contractor, x2*
4	Mean	5.64	5.58
5	Variance	0.016533333	0.014755556
6	Observations	10	10
7	df	9	9
8	F	1.120481928	
9	P(F<=f) one-tail	0.434106199	
10	F Critical one-tail	4.025994158	

Figure 8. Excel Results for the Two Sample *F*-Test

Step 12: Interpret the *F*-test results. The Excel results in Figure 8 show both the SCDOT and Contractor sample means and sample variances along with the number of observations in each sample.

There are two ways to use Figure 8 to determine the results of the *F*-test. We can use either of the following to conclude that the sample variances are not equal:

1. Conclude the variances are not equal when the F value that is shown is greater than the F Critical one-tail value that is shown. In Figure 8, we would conclude that the variances can be assumed to be equal since $1.120 \leq 4.026$.

2. Conclude the variances are not equal when the P(F<=f) one-tail value shown is less than the $\alpha/2$ value that we had selected. In Figure 8, we would conclude that the variances can be assumed to be equal since $0.434 \geq 0.025$.

If 1 and 2 are not true, then we conclude that there is no reason to believe the variances are not equal and, therefore, we assume that the variances are equal.

Procedure: Comparing Sample Results to a Standard or Target Value

Purpose: Compare the results of sample data to a given standard or target value to see if it is reasonable to assume that they came from a normal population with mean equal to the standard or target value.

Background: Statistical Concepts Section 3: Inferences on One-Sample Data

Steps Involved:

Note: There are two simple methods in which this comparison can be made using Excel, and they will give identical results. Both methods are presented below. The first method is based on using a *t*-test to make the comparison, and begins at *Step 1a*. The second method is identical in concept, but uses a confidence interval to make the comparison, and begins at *Step 1b*.

Step 1a: Use a *t*-test to determine whether or not to assume the sample data came from a normal populations with mean equal to the given target value.

Enter the sample data to be compared into a column in the Excel spreadsheet. Enter the target value into a second column, with the target value repeated for each row where there are sample results. Label each column appropriately. In Figure 1 we are comparing a Contractor's test results for asphalt content with a specified target value of 6.00.

	A	B
1	**Test Results**	**Target**
2	6.15	6.00
3	5.63	6.00
4	5.99	6.00
5	6.08	6.00
6	6.18	6.00
7	6.43	6.00
8	5.88	6.00
9	5.88	6.00
10	5.09	6.00
11	5.66	6.00
12	6.13	6.00
13	5.79	6.00
14	6.05	6.00
15	5.60	6.00
16	6.01	6.00
17	5.80	6.00
18	6.13	6.00
19	5.71	6.00
20	5.64	6.00
21	5.77	6.00

Figure 1. Asphalt Content Example

Step 2a: Click on Tools at the top of the Excel screen, and then click on Data Analysis … to open the menu shown in Figure 2. (Note, if there is no Data Analysis … option shown, then it can be added to the menu by first clicking on Add Ins … and then checking Analysis ToolPack.)

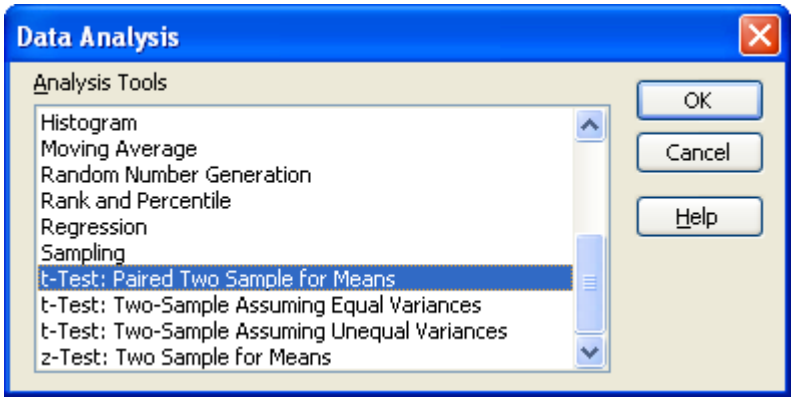


Figure 2. Data Analysis Menu

Step 3a: Highlight t-Test: Paired Two Sample for Means and click on OK to open the menu in Figure 3 and allow the *t*-test to be conducted.

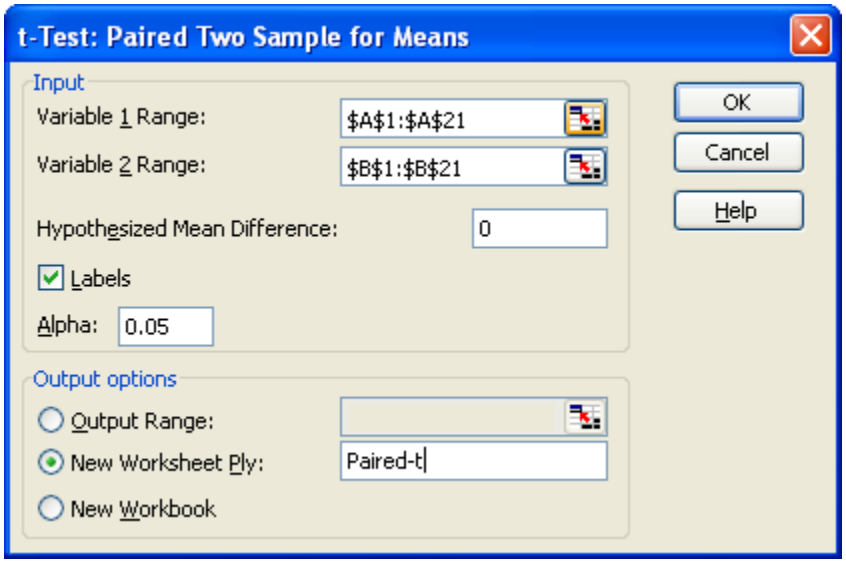


Figure 3. Paired *t*-Test Menu

Enter the spreadsheet range (Variable 1 Range: and Variable 2 Range: in Figure 3) for the SCDOT and Contractor data sets. You can either leave the Hypothesized Mean Difference: blank or enter 0. Check the Labels box if you have labels included in the variable ranges.

Step 4a: Select a level of significance, α, for the comparison. This is shown as Alpha: in Figure 3. The level of significance is the probability that you will incorrectly declare the sample to have come from a population with mean not equal to the target value when, in fact, it did. A typical α value would be 0.05, which has been entered in Figure 3.

In Figure 3, the circle has been checked for New Worksheet Ply: and the name of the new worksheet (where the *t*-test results will be displayed) has been entered as Paired-t.

Step 5a: Click OK and Excel performs the paired *t*-test and creates the new worksheet Paired-t that includes the results shown in Figure 4.

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		*Test Results*	*Target*
4	Mean	5.879944926	6
5	Variance	0.085076657	0
6	Observations	20	20
7	Pearson Correlation	#DIV/0!	
8	Hypothesized Mean Difference	0	
9	df	19	
10	t Stat	-1.840730929	
11	P(T<=t) one-tail	0.040666533	
12	t Critical one-tail	1.729132792	
13	P(T<=t) two-tail	0.081333067	
14	t Critical two-tail	2.09302405	

Figure 4. Excel Results for the Paired *t*-Test

Step 6a: Interpret the paired *t*-test results. The Excel results in Figure 4 show the sample mean and sample variance along with the number of observations. The Target mean and standard deviation are also shown. Since the same target value was entered for each observation, the mean equals the target value and the standard deviation is 0. Figure 4 shows values for a one-tail test, i.e., testing whether or not the mean is larger (or smaller) than the target value. It also shows values for a two-tail test, i.e., testing whether or not the mean equals the target value. Since we are generally interested in whether or not the mean equals the target value, we will usually use the values for the two-tailed test, and that is what we will consider here.

There are two ways to use Figure 4 to determine the results of the paired *t*-test. We can use either of the following to conclude that the sample is not from a population with mean equal to the target value:

1. Conclude the mean does not equal the target value when the absolute value of the t Stat value that is shown is greater than the t Critical two-tail value that is shown. In Figure 4, we would conclude that we can assume that the mean is equal to the target value since $|-1.841| \leq 2.093$.

2. Conclude the mean does not equal the target value when the P(T<=t) two-tail value shown is less than the α value that we had selected. In Figure 4, we would conclude that we can assume that the mean is equal to the target value since $0.081 \geq 0.05$.

If 1 and 2 are not true, then we conclude that there is no reason to believe the mean does not equal the target value, and, therefore, we assume that the mean equals the target value.

Step 1b: Use a *confidence interval* about the sample mean to determine whether or not to assume the sample data came from a normal populations with mean equal to the given target value.

Enter the sample data to be compared into a column in the Excel spreadsheet. Label the column appropriately. We will compare the Contractor's asphalt content test results from Figure 1 with the same specified target value of 6.00. For convenience, the test results are shown in Figure 5.

	A
1	**Test Results**
2	6.15
3	5.63
4	5.99
5	6.08
6	6.18
7	6.43
8	5.88
9	5.88
10	5.09
11	5.66
12	6.13
13	5.79
14	6.05
15	5.60
16	6.01
17	5.80
18	6.13
19	5.71
20	5.64
21	5.77

Figure 5. Asphalt Content Example

Step 2b: Click on Tools at the top of the Excel screen, and then click on Data Analysis … to open the menu shown in Figure 6. (Note, if there is no Data Analysis … option shown, then it can be added to the menu by first clicking on Add Ins … and then checking Analysis ToolPack.)

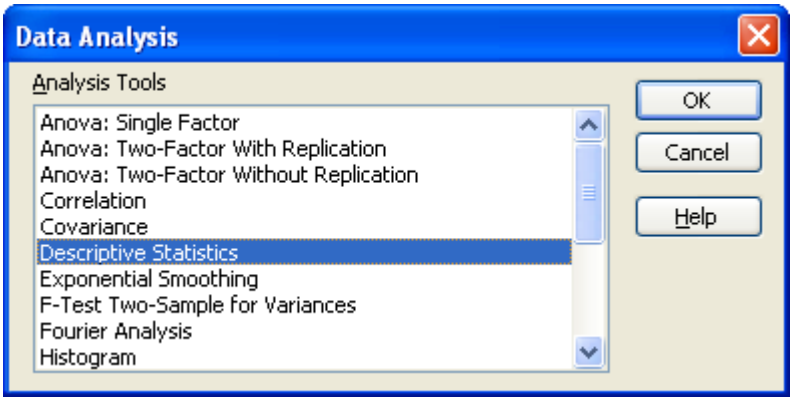


Figure 6. Data Analysis Menu

Step 3b: Highlight Descriptive Statistics and click on OK to open the menu in Figure 6 and allow the location of the data to be input.

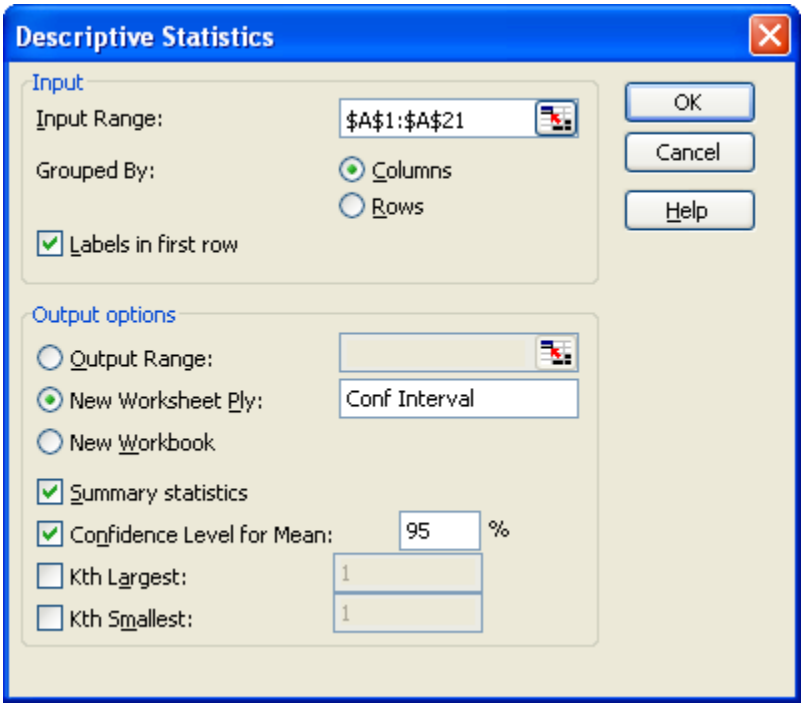


Figure 7. Paired *t*-Test Menu

Enter the spreadsheet range (Input Range:) for the Contractor test result data. Check the Labels box if you have labels included in the variable ranges.

Step 4b: Select a level of significance, α, for the comparison. The level of significance is the probability that you will incorrectly declare the sample to have come from a population with mean not equal to the target value when, in fact, it did. A typical α value would be 0.05, which is used in this example.

Check the Summary Statistics and Confidence Level for the Mean: boxes. Then calculate the Confidence Level for the Mean: as $[(1 - \alpha) \times 100\%]$. For this example, the confidence interval is $[(1 – 0.05) \times 100\%] = 95\%$, which is entered in Figure 7

In Figure 7, the circle has been checked for New Worksheet Ply: and the name of the new worksheet (where the *t*-test results will be displayed) has been entered as Conf Interval.

Step 5b: Click OK and Excel calculates the descriptive statistics and creates the new worksheet Conf Interval that includes the results shown in Figure 8.

	A	B
1	*Test Results*	
2		
3	Mean	5.879944926
4	Standard Error	0.065221414
5	Median	5.880781786
6	Mode	#N/A
7	Standard Deviation	0.291679031
8	Sample Variance	0.085076657
9	Kurtosis	1.726691287
10	Skewness	-0.731324717
11	Range	1.341849384
12	Minimum	5.086197931
13	Maximum	6.428047315
14	Sum	117.5988985
15	Count	20
16	Confidence Level(95.0%)	0.136509988

Figure 8. Excel Results for the Descriptive Statistics

Step 6b: Determine the confidence interval about the sample mean within which we are 95% confident that the true population mean falls. While there are many useful descriptive statistics shown in the output in Figure 8, only the Mean value, 5.8799, and the Confidence Level(95.0%), 0.1365, are needed to reach a decision regarding whether or not to assume that the sample results came from a population with mean equal to the target value.

The interval within which we are 95% confident the true population mean falls is then calculated as the Mean ± the Confidence Level(95.0%). From Figure 8, the confidence interval is 5.8799 ± 0.1365, or from 5.7434 to 6.0164. If the target

value falls within this confidence interval, we conclude that we can assume that the test results came from a population with mean equal to the target value. If the target value falls outside this confidence interval, we conclude that the test results came from a population with mean not equal to the target value. In this case, since the target value, 6.00, falls within the confidence interval, we assume that the test results did come from a population with mean equal to the target value.

(*This page is intentionally blank.*)

Procedure: Fitting a Line to a Set of (*X*, *Y*) Data

Purpose: To determine the equation of a line that can be used to represent the relationship between related pairs of two variables (*X* and *Y*). For example, the curing time, *X*, and the measured compressive strength, *Y*, are variables that are both related to the 'event' that is breaking a concrete cylinder.

Background: Statistical Concepts Section 2: Graphical Presentation of Data
Statistical Concepts Section 8: Regression Analysis

Steps Involved:

Step 1: Use Excel to develop a scatter plot of the (*X*, *Y*) data pairs.

Enter the data to be plotted into two columns in the Excel spreadsheet. Label each column appropriately. In Figure 1 we have asphalt content (AC) and VMA results from samples taken at an asphalt plant. We wish to plot these to determine if there is a relationship between these two characteristics.

Determine which variable will be on the horizontal axis (*X*) and which will be on the vertical axis (*Y*). If there is an implied causal relationship between the variables, the independent variable (the one that causes the other) should be on the horizontal axis (*X*-axis) and the dependent variable (the one that may be caused by the other) should be on the vertical axis (*Y*-axis).

Since we can control the amount of asphalt cement put into the mix, in our example, we select AC content as the independent (*X*) variable and VMA as the dependent (*Y*) variable.

	A	B
1	**AC**	**VMA**
2	6.030	17.680
3	6.020	17.650
4	6.010	17.080
5	6.010	17.490
6	5.960	17.430
7	6.200	17.910
8	5.960	17.260
9	6.040	17.530
10	6.040	17.380
11	5.930	17.110
12	5.870	17.000
13	5.780	16.910
14	6.240	17.240
15	6.980	19.480
16	6.670	18.420
17	7.250	20.210
18	6.510	19.860
19	4.880	15.240
20	4.810	14.490
21	4.950	14.480
22	5.310	15.270

Figure 1. Data for the AC-VMA Example

Step 2: Click on the Chart Wizard icon at the top of the Excel screen (Figure 2) to open the menu shown in Figure 3. Click on the XY (Scatter) chart type and then click on Next >.

Figure 2. Icon for the Chart Wizard in Excel

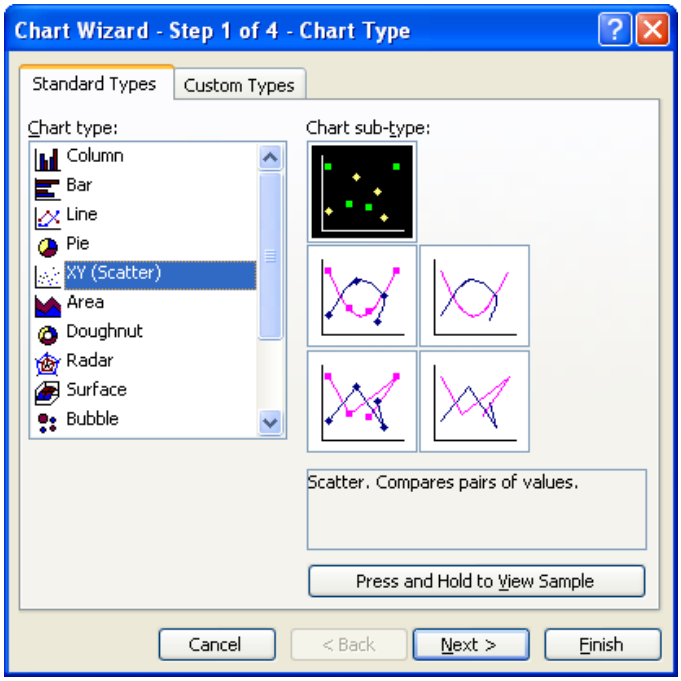


Figure 3. Selecting the XY (Scatter) Plot in the Excel Chart Wizard

Step 3: In the next menu, shown in Figure 4, click on the Series tab at the top of the menu and then enter the spreadsheet range for the X Values: and Y Values:, and then, click Next >.

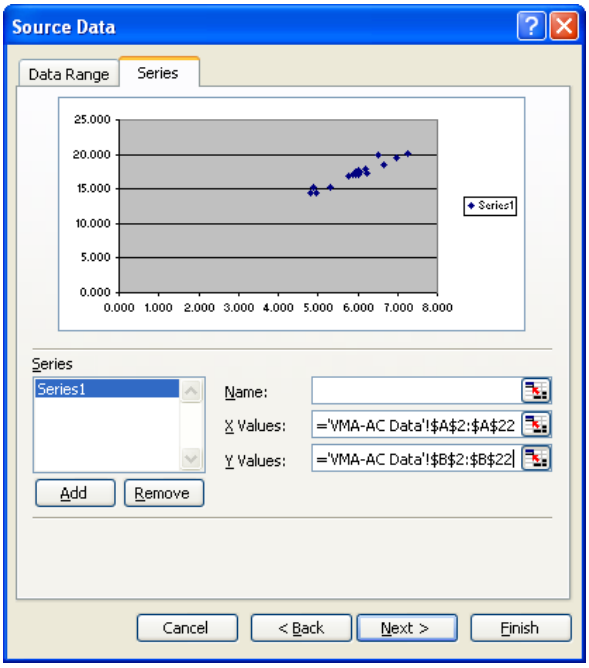


Figure 4. Source Data Input Menu for the AC-VMA Example

Step 4: In the next menu (Figure 5) for the chart you can choose to add or delete or Titles, Axes, Gridlines, a Legend, or Data Labels. In Figure 5, the box for Show legend has been unchecked to allow for the plot to fill the full page. Also, in this example no chart or axis titles have been entered. Click on Next > to proceed to the menu to select a location for the scatter plot.

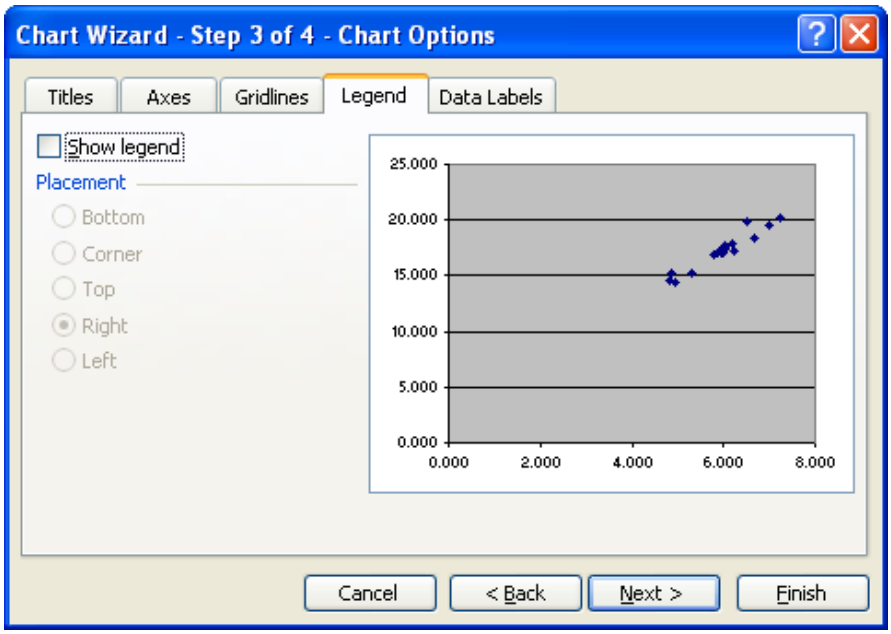


Figure 5. Chart Options Menu for Excel Scatter Plots

Step 5: In the next menu (Figure 6) select the location for the scatter plot. It can be included as an object in an existing worksheet by checking the As object in: button, or it can be created as a new worksheet by checking the As new sheet: button and inputting a name for the new worksheet (VMA-AC-Plot in Figure 6).

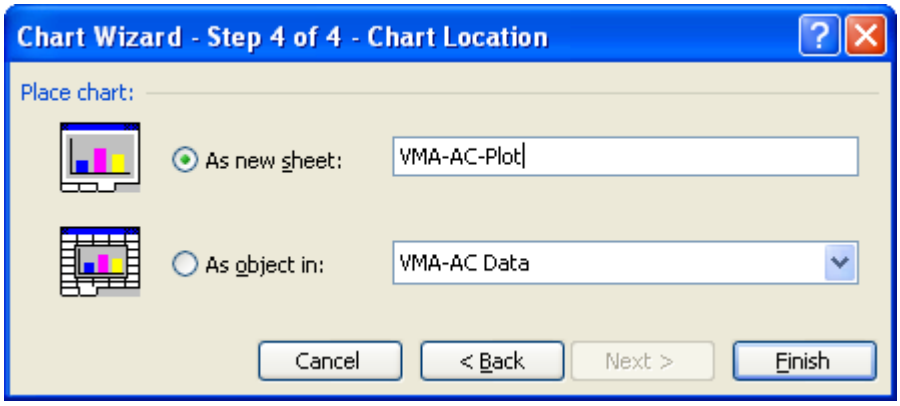


Figure 6. Chart Location Menu for Excel Scatter Plots

Step 6: Click on the VMA-AC-Plot worksheet tab to open the plot shown in Figure 7. Next, modify the format of the axes to allow for better evaluation of the (*X*, *Y*) data pairs. This can be done for the *X*-axis by double clicking on the axis to bring up the menu shown in Figure 8. This menu can be used to change the colors and styles of axes (Patterns tab), the axis scale (Scale tab), the font style and size (Font tab), and the format in which the axis values are displayed (Number tab).

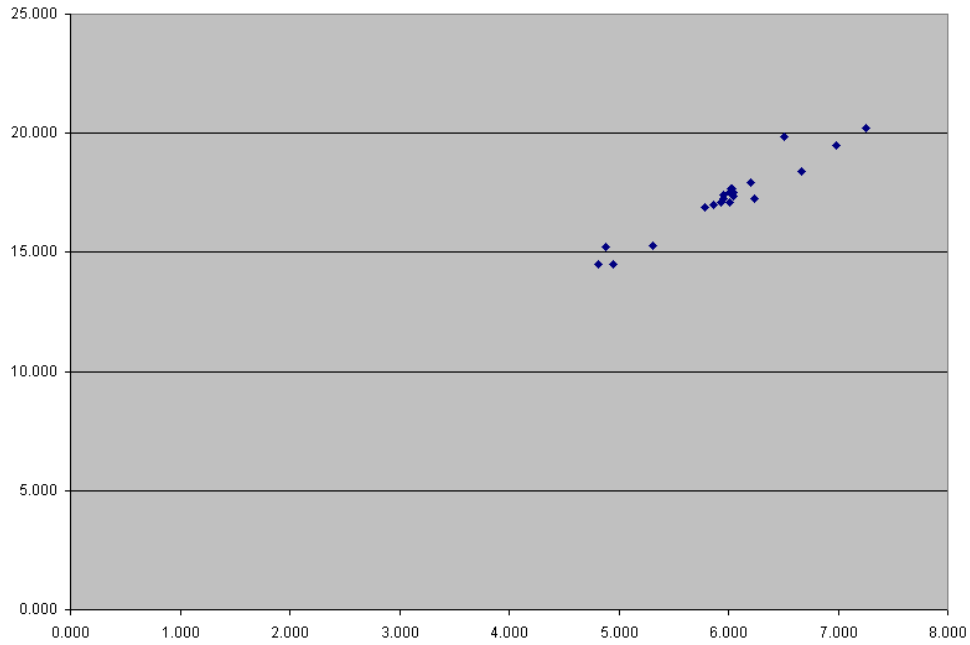


Figure 7. Excel Scatter Plot of the AC-VMA Sample Data

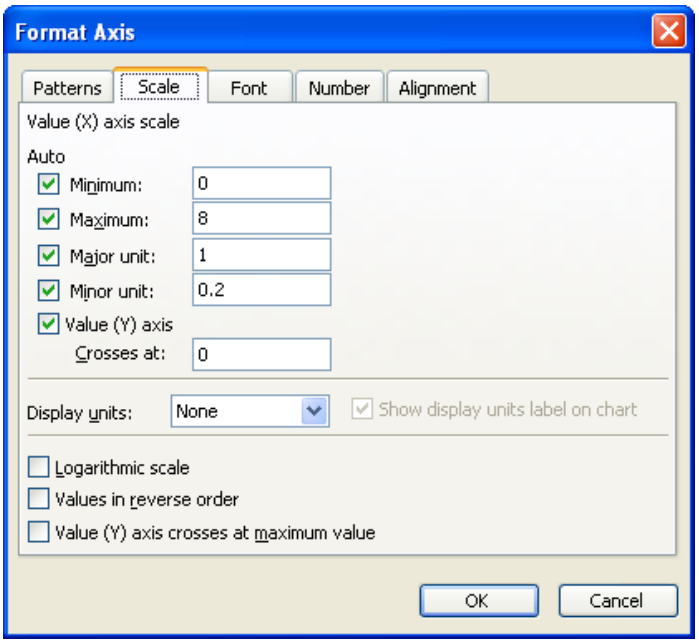


Figure 8. Format Axis Menu for the X-Axis for the AC-VMA Sample Data

Figure 8 shows the default, or Auto, values Excel selected for the axis scale. Figure 9 shows the values that were input to change the scale of the *X*-axis to begin at 4.5 and end at 7.5. The scale of the *Y*-axis can be modified in a similar fashion to yield the new scatter plot shown in Figure 10.

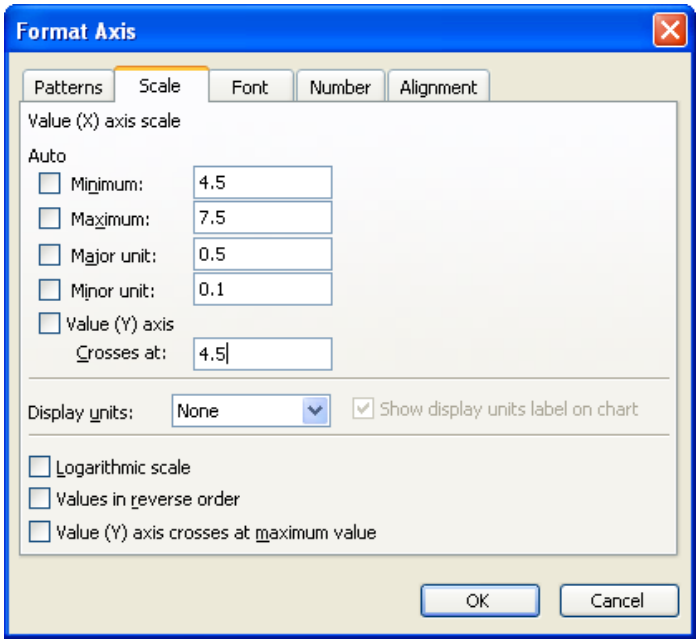


Figure 9. Format Axis Menu Showing the New Scale for the X-axis

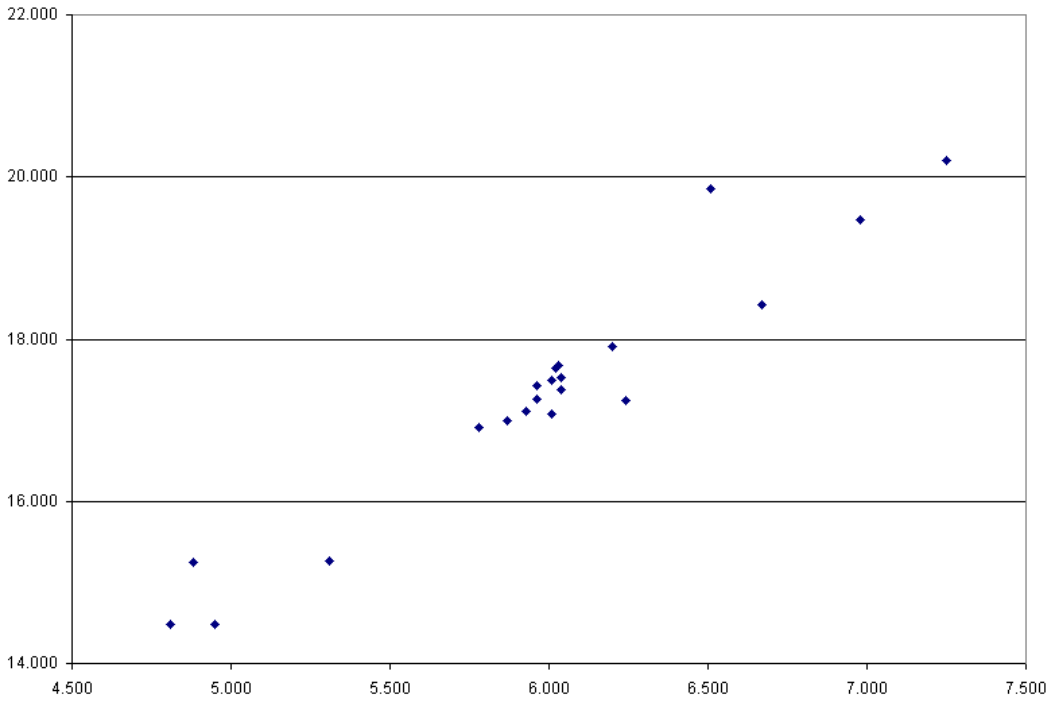


Figure 10. Scatter Plot with Axes Revised for the Spread of the Data

Step 7: Examine the (*X*, *Y*) data pairs to see if there is any obvious pattern. For example, do they appear to form a straight line, or does some sort of curved line better describe their relationship? If there appears to be a general pattern to the data, then the next step is to identify an equation that produces a line that approximately describes the pattern in the data. To do this, place the cursor on any one of the (*X*, *Y*) data pairs and then click the right mouse button to bring up the menu shown in Figure 11.

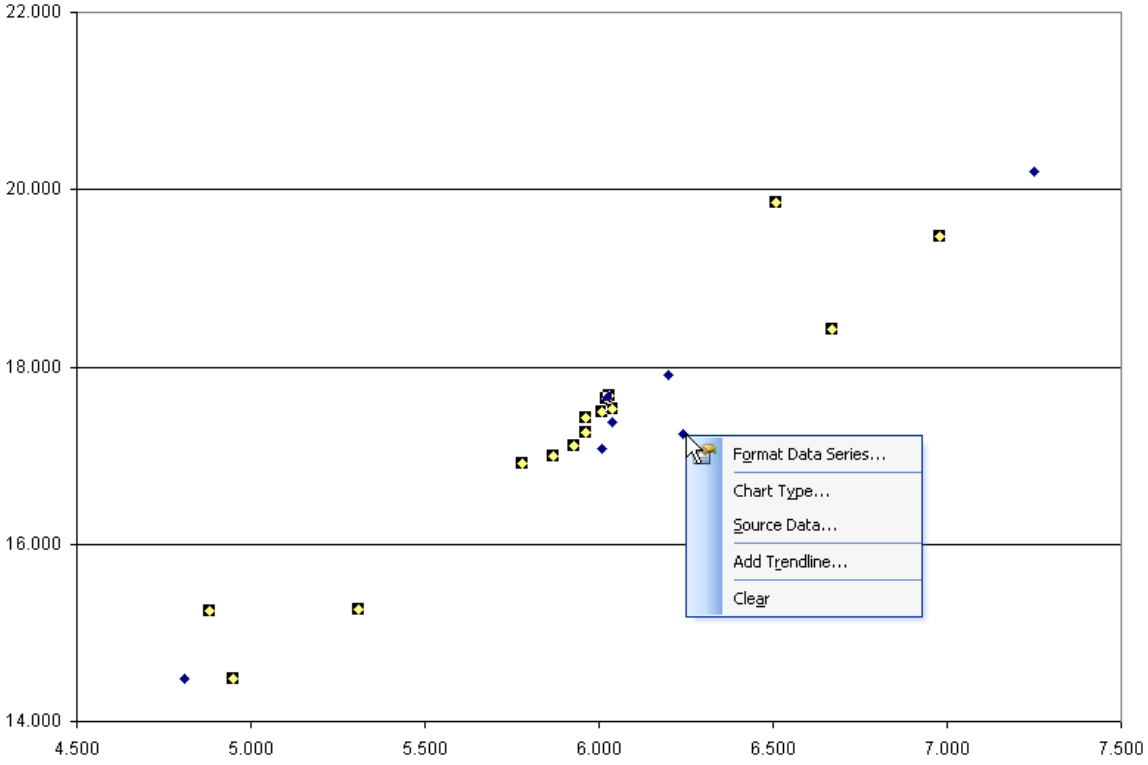


Figure 11. Excel Menu for Adding a Trendline to a Scatter Plot

This menu can be used to change the color and style of the plotted points (Format Data Series…), modify the type of plot (Chart type…), modify the data points to be plotted (Source Data…), and to fit a line to the (*X*, *Y*) data pairs (Add Trendline…). Click on Add Trendline… to bring up the menu shown in Figure 12.

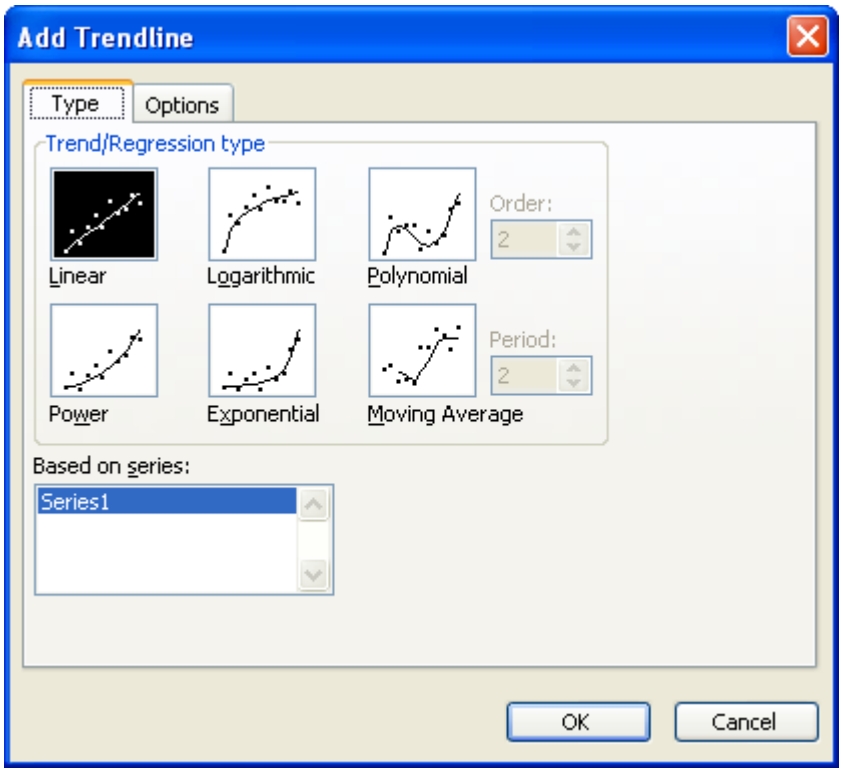


Figure 12. Excel Menu for Adding a Trendline to a Scatter Plot

Step 8: If the (*X*, *Y*) data pairs appear to exhibit a straight line relationship, select Linear as the Trend/Regression type. If the (*X*, *Y*) pairs exhibit more of a curved relationship, select Logarithmic if the curve appears to be concave downward and either Power or Exponential if the curve appears to be concave upward. Polynomial can be used to fit either concave or convex curves as well as more complicated relationships, but the equations can get complex for higher order plots. For simplicity, select Linear first if the data approximately form a straight line.

Once the trend type has been selected, click on the Options tab to bring up the menu shown in Figure 13.

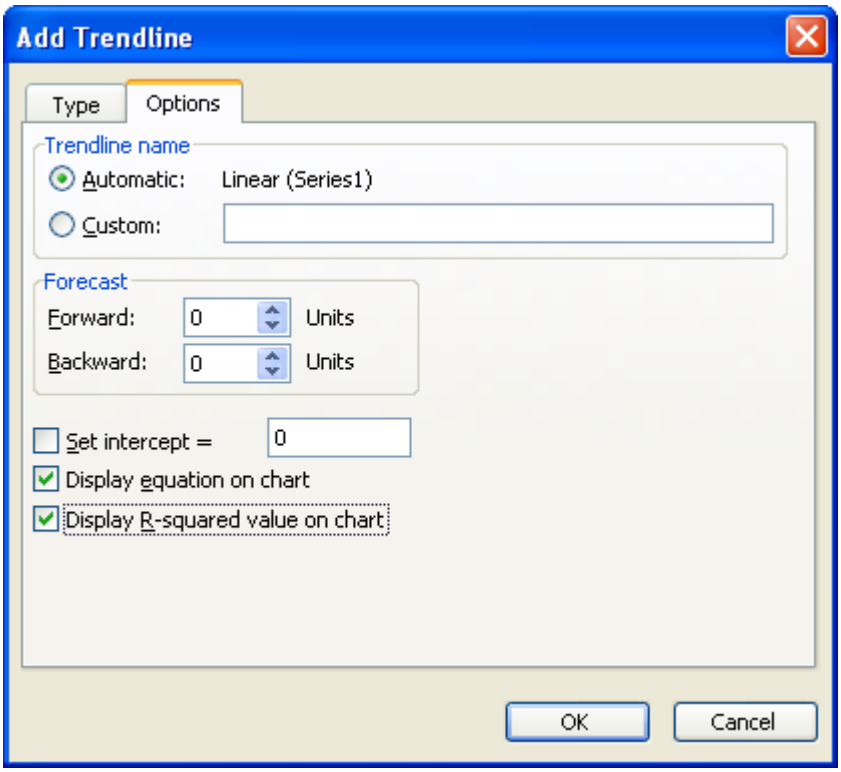


Figure 13. Excel Menu for Selecting Trendline Options

Check the boxes to Display equation on chart and to Display R-squared value on chart. The *R*-squared (or R^2) value is a relative measure of the proportion of the variability of the *Y*-values that is explained by the trendline (known as the regression line). As such, *R*-square must range between a minimum of 0, i.e., the trendline explains no more of the variability of the *Y*-values than is done by the average of the *Y*-values, and a maximum of 1.0, i.e., the trendline passes directly through all of the (*X*, *Y*) pairs on the scatter plot.

If there is some physical relationship that would indicate that the *Y* intercept should go through a particular point, most often 0, then you can check the Set intercept = box and enter a value to force the fitted line to go through the selected *Y* value when $X = 0$. In most cases you will leave the Set intercept = box unchecked. You can also check the Custom: button and give the trendline a name. Clicking on OK results in the trendline shown in Figure 14.

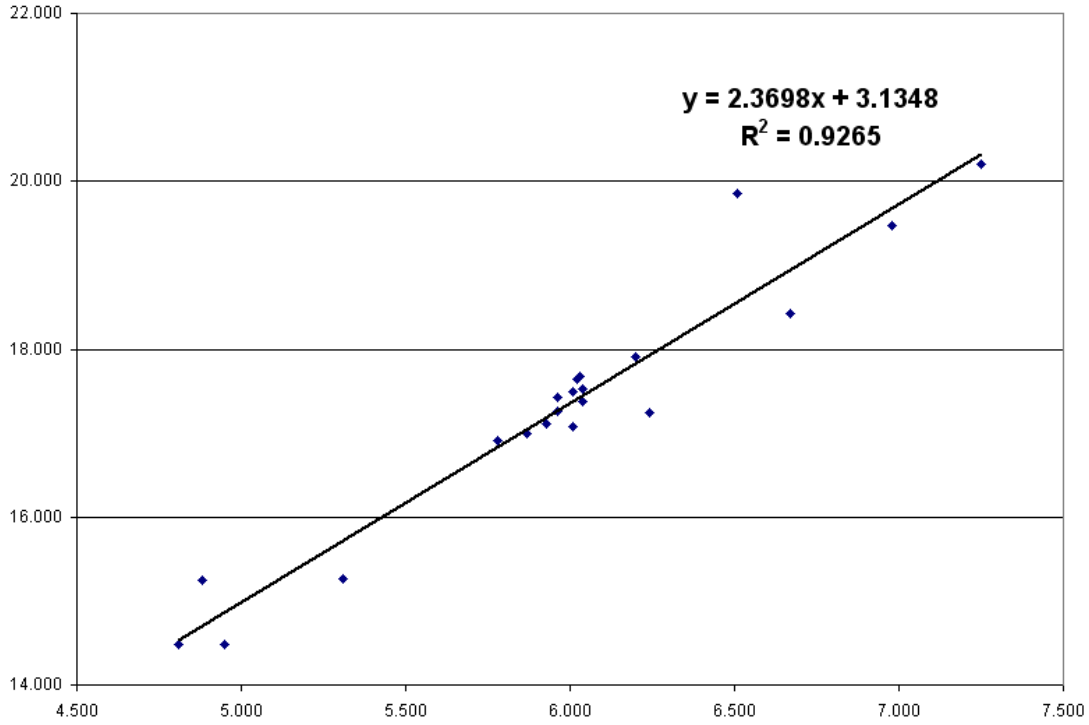


Figure 14. Linear Trendline Fit to the AC-VMA Sample Data

Step 9: Evaluate how well the trendline fits the (*X*, *Y*) data pairs. The trendline in Figure 14 is a reasonably good fit to the data since the R^2 value, 0.9265, is high and since the data points are randomly scattered about the trendline. The data points should be randomly spread above and below the trendline with no obvious patterns. An example of a pattern in the data points would be if the (*X*, *Y*) data pairs for small and large values of *X* were all above the trendline, while the (*X*, *Y*) pairs for medium values of *X* were all below the trendline. This would indicate that a curved line would be a better fit for the data than is the straight line.

Step 10: If in *Step 8* it was decided that the relationship between the *X* and *Y* values was curvilinear rather than linear, or if a pattern such as the one described in *Step 9* is detected, then Logarithmic (concave downward), either Power or Exponential (concave upward), or Polynomial (concave upward or downward) would be selected for the Trend/Regression type. This can be illustrated with the 28-day concrete compressive strength versus water/cement ratio example data shown in Figure 15.

	A	B
1	**W/C Ratio**	**28-Day, psi**
2	0.34	6350
3	0.35	6050
4	0.40	5165
5	0.45	4220
6	0.50	3800
7	0.55	3150
8	0.60	2740
9	0.63	2540

Figure 15. Data for the Compressive Strength vs. W/C Ratio Example

Step 11: Using the procedures outlined in *Step 2* through *Step 6*, develop the scatter plot shown in Figure 16. Since the data appear to approximately fit a straight line, use the procedures in Step 7 and Step 8 to fit a straight line to the data, as shown in Figure 17.

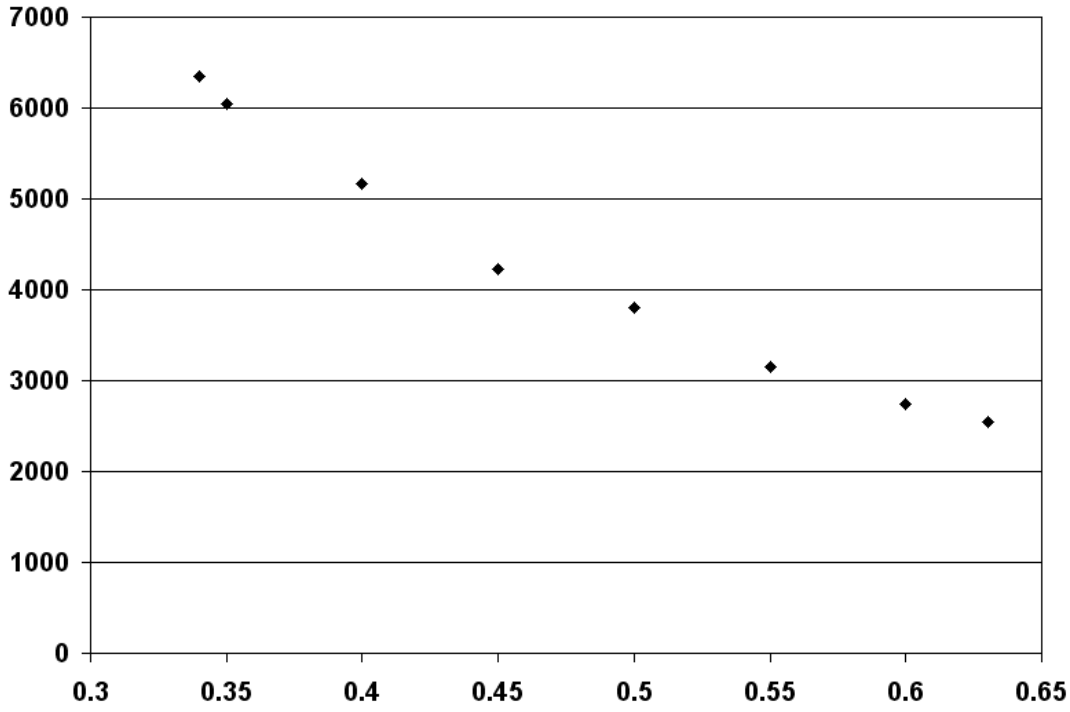


Figure 16. Scatter Plot of Compressive Strength vs. W/C Ratio

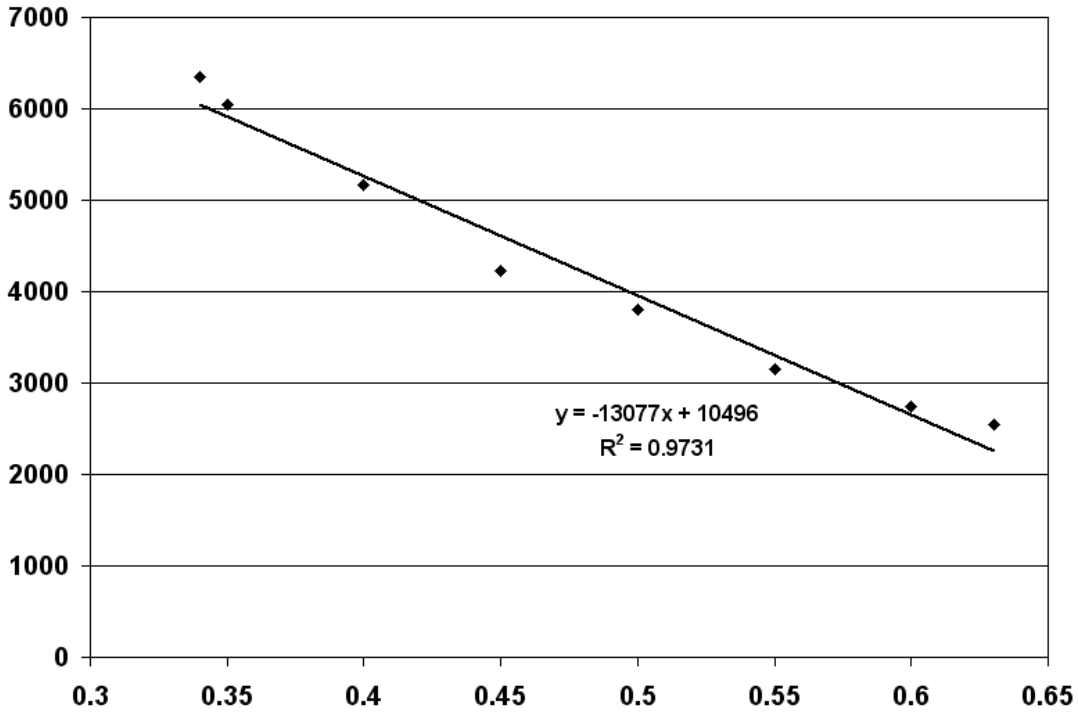


Figure 17. Linear Trendline Fit to the Compressive Strength vs. W/C Ratio Data

Step 12: Evaluate how well the trendline fits the (*X*, Y) data pairs. Initially, the trendline in Figure 17 appears to be a good fit to the data since the R^2 value, 0.9731, is high. However, further investigation reveals that the data points may not be randomly scattered about the trendline. Specifically, the two smallest and the two largest W/C ratios have strengths above the trendline, while the four middle W/C ratios all have strengths below the trendline. This pattern indicates that a concave upward curved line would be a better fit for the data than is the straight line. Figure 18 shows a polynomial line fit to the data. The initial linear trendline is also shown for comparison.

While the polynomial trendline has a slightly larger R^2 value than the linear trendline, 0.9976 compared to 0.9731, the major difference between the two lines is in the spread of the data points about the two lines. The curved trendline eliminates the pattern that was present in the manner in which the data points were distributed about the linear trendline. This indicates that the polynomial trendline is a better fit for the data. However, depending upon the intended use for the model, since the linear trendline still has a high R^2 value, it might still be selected because of its simplicity compared with the polynomial equation.

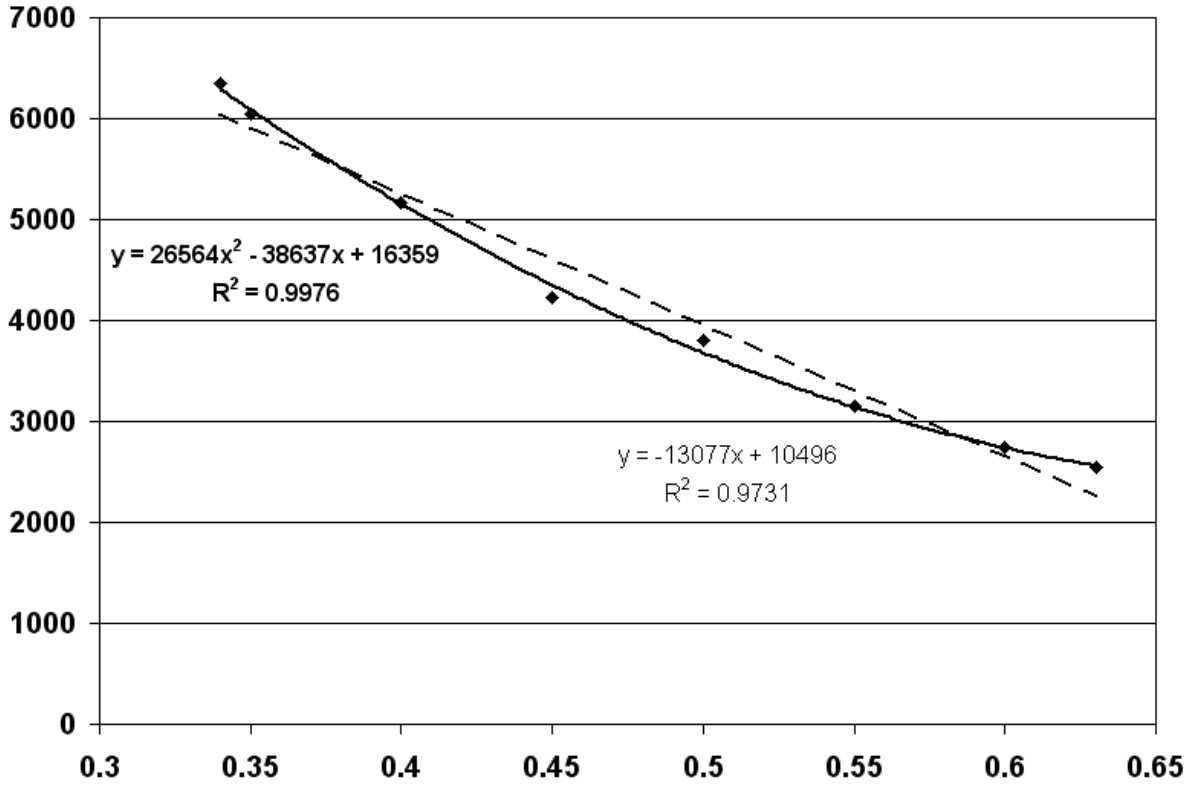


Figure 18. Comparison of Linear and Polynomial Trendlines Fit to the Compressive Strength vs. W/C Ratio Data

Total Printing Cost	$300.00
Total Number of Documents	30
Cost per Unit	$10.00