

© Copyright 2008, AcaStat Software. All rights Reserved.

<http://www.acastat.com>

Version 2.2

Table of Contents

THE BASICS	5
WORKBOOK FORMAT.....	5
ABOUT ACASTAT SOFTWARE.....	5
HAND CALCULATIONS.....	5
DATA TERMINOLOGY	5
ANALYZING RAW DATA WITH THE DATA GRID MODULE.....	5
ANALYZING SUMMARY DATA WITH THE SUMSTATS MODULE	6
LESSON 1: VARIABLES.....	7
<i>Problem 1.1 Classifying Characteristics.....</i>	<i>7</i>
LESSON 2: DATA FILE BASICS	8
<i>Problem 2.1 How to Create a Data File</i>	<i>8</i>
<i>Problem 2.2 Formatting the Data File</i>	<i>10</i>
<i>Problem 2.3 Recoding Data</i>	<i>12</i>
LESSON 3: DESCRIBING COUNTS AND PROPORTIONS.....	13
<i>Problem 3.1 Hand Calculating Univariate Descriptions.....</i>	<i>13</i>
<i>Problem 3.2 Hand Calculating Bivariate Descriptions</i>	<i>14</i>
<i>Problem 3.3 Univariate Analysis of Raw Data</i>	<i>15</i>
<i>Problem 3.4 Bivariate Analysis of Raw Data.....</i>	<i>15</i>
<i>Problem 3.5 Creating a Summary Table.....</i>	<i>15</i>
LESSON 4: DESCRIBING CONTINUOUS DATA	16
<i>Problem 4.1 Hand Calculations.....</i>	<i>17</i>
<i>Problem 4.2 Creating Summary Statistics with Raw Data</i>	<i>18</i>
LESSON 5: STANDARDIZED Z-SCORES.....	19
<i>Problem 5.1 Hand Calculations.....</i>	<i>19</i>
<i>Problem 5.2 Confirming Hand Calculations with SumStats</i>	<i>20</i>
<i>Problem 5.3 Creating Standardized Scores with Raw Data</i>	<i>21</i>
LESSON 6: HYPOTHESES	23
<i>Problem 6.1 Developing Hypotheses.....</i>	<i>23</i>
LESSON 7: RANDOM SAMPLING	24
<i>Problem 7.1 Comparing Random Samples to a Population</i>	<i>24</i>
<i>Problem 7.2 Sample v. Simulated Sampling Distributions.....</i>	<i>25</i>
LESSON 8: HYPOTHESIS TESTING.....	28
<i>Problem 8.1 Interpreting p-values</i>	<i>28</i>
NOMINAL AND ORDINAL DATA	29
LESSON 9: INTERVAL ESTIMATION FOR PROPORTIONS.....	30
<i>Problem 9.1 Hand Calculations.....</i>	<i>30</i>
<i>Problem 9.2 Using Summary Statistics in SumStats</i>	<i>31</i>
LESSON 10: COMPARING A POPULATION TO A SAMPLE PROPORTION (Z-TEST)	32
<i>Problem 10.1 Hand Calculations.....</i>	<i>32</i>
<i>Problem 10.2 Creating Summary Statistics with Raw Data</i>	<i>33</i>
<i>Problem 10.3 Using Summary Statistics in SumStats</i>	<i>34</i>
LESSON 11: COMPARING PROPORTIONS - TWO INDEPENDENT SAMPLES.....	35
<i>Problem 11.1 Hand Calculations.....</i>	<i>35</i>
<i>Problem 11.2 Using Summary Statistics in SumStats</i>	<i>36</i>
LESSON 12: CHI-SQUARE TEST OF INDEPENDENCE	37
<i>Problem 12.1 Hand Calculations.....</i>	<i>37</i>
<i>Problem 12.2 Calculating Chi-Square with Raw Data</i>	<i>39</i>
<i>Problem 12.3 Calculating Chi-Square from Summary Statistics.....</i>	<i>40</i>
LESSON 13: MEASURING ASSOCIATION FOR CHI SQUARE	41
<i>Problem 13.1 Hand Calculations.....</i>	<i>41</i>
<i>Problem 13.2 Interpreting Multiple Comparisons.....</i>	<i>42</i>
<i>Problem 13.3 Review the Basics</i>	<i>42</i>
INTERVAL AND RATIO DATA	43
LESSON 14: INTERVAL ESTIMATION FOR MEANS	44

<i>Problem 14.1 Hand Calculations</i>	44
<i>Problem 14.2 Creating Summary Statistics with Raw Data</i>	45
<i>Problem 14.3 Creating Confidence Intervals with SumStats</i>	46
<i>Problem 14.4 Interpreting Confidence Intervals</i>	46
LESSON 15: COMPARING A POPULATION MEAN TO A SAMPLE MEAN (T-TEST)	47
<i>Problem 15.1 Hand Calculations</i>	47
<i>Problem 15.2 Creating Summary Statistics with Raw Data</i>	48
<i>Problem 15.3 Calculating One-Sample t-tests with SumStats</i>	49
LESSON 16: COMPARING TWO INDEPENDENT SAMPLE MEANS (T-TEST).....	50
<i>Problem 16.1 Hand Calculations</i>	50
<i>Problem 16.2 Calculating t-tests from Raw Data</i>	52
<i>Problem 16.3 Using Summary Statistics in SumStats</i>	53
LESSON 17: COMPUTING F-RATIO	54
<i>Problem 17.1 Hand Calculations</i>	54
<i>Problem 17.2 Interpreting the F-ratio in SumStats t-tests</i>	55
LESSON 18: ONE-WAY ANALYSIS OF VARIANCE (ANOVA)	56
<i>Problem 18.1 Hand Calculations</i>	56
<i>Problem 18.2 ANOVA with Raw Data</i>	58
<i>Problem 18.3 Using Summary Statistics in SumStats</i>	59
LESSON 19: CORRELATION	60
<i>Problem 19.1 Hand Calculations</i>	60
<i>Problem 19.2 Evaluating Correlations with Raw Data</i>	61
LESSON 20: HYPOTHESIS TESTING FOR PEARSON R	62
<i>Problem 20.1 Hand Calculations</i>	62
<i>Problem 20.2 Evaluating Correlation Significance</i>	63
LESSON 21: SIMPLE LINEAR REGRESSION	64
<i>Problem 21.1 Hand Calculations</i>	64
<i>Problem 21.2 Simple Regression with Raw Data</i>	66
LESSON 22: DETERMINE THE STANDARD ERROR OF THE ESTIMATE	67
<i>Problem 22.1 Hand Calculations (Data from example 21.1)</i>	67
<i>Problem 22.2 Predicting Y with Raw Data</i>	68
APPENDIX	69
ADDITIONAL REVIEW QUESTIONS	70
<i>Requires Lesson 1</i>	70
<i>Requires Lesson 3</i>	71
<i>Requires Lesson 5</i>	72
<i>Requires Lesson 9</i>	72
<i>Requires Lesson 10</i>	72
<i>Requires Lesson 11</i>	73
<i>Requires Lesson 14</i>	75
<i>Requires Lesson 15</i>	75
<i>Requires Lesson 16</i>	75
<i>Requires Lesson 14 and 16</i>	77
<i>Requires Lesson 18</i>	77
<i>Requires Lesson 19 and 20</i>	78
TABLES.....	79
<i>Z Distribution Critical Values</i>	79
<i>T Distribution Critical Values</i>	80
<i>Chi-square Distribution Critical Values</i>	81
<i>F Distribution Critical Values</i>	82
SPSS INSTRUCTIONS.....	83
<i>Problem 2.1 How to Create a Data File</i>	83
<i>Problem 2.2 Formatting the Data File</i>	83
<i>Problem 2.3 Recoding Data</i>	84
<i>Problem 3.3 Univariate Analysis of Raw Data</i>	84
<i>Problem 3.4 Bivariate Analysis of Raw Data</i>	84
<i>Problem 4.2 Creating Summary Statistics with Raw Data</i>	84
<i>Problem 5.2 Confirming Hand Calculations with SumStats</i>	84
<i>Problem 5.3 Creating Standardized Scores with Raw Data</i>	85

<i>Problem 7.1 Comparing Random Samples to a Population</i>	<i>85</i>
<i>Problem 9.2 Using Summary Statistics in SumStats</i>	<i>85</i>
<i>Problem 10.2 Creating Summary Statistics with Raw Data</i>	<i>85</i>
<i>Problem 10.3 Using Summary Statistics in SumStats</i>	<i>86</i>
<i>Problem 11.2 Using Summary Statistics in SumStats</i>	<i>86</i>
<i>Problem 12.2 Calculating Chi-Square with Raw Data</i>	<i>86</i>
<i>Problem 12.3 Calculating Chi-Square from Summary Statistics.....</i>	<i>86</i>
<i>Problem 13.2 Interpreting Multiple Comparisons.....</i>	<i>86</i>
<i>Problem 14.2 Creating Summary Statistics with Raw Data</i>	<i>87</i>
<i>Problem 14.3 Creating Confidence Intervals with SumStats</i>	<i>87</i>
<i>Problem 15.2 Creating Summary Statistics with Raw Data</i>	<i>87</i>
<i>Problem 15.3 Calculating One-Sample t-tests with SumStats.....</i>	<i>87</i>
<i>Problem 16.2 Calculating t-tests from Raw Data</i>	<i>87</i>
<i>Problem 16.3 Using Summary Statistics in SumStats</i>	<i>88</i>
<i>Problem 18.2 ANOVA with Raw Data</i>	<i>88</i>
<i>Problem 18.3 Using Summary Statistics in SumStats</i>	<i>88</i>
<i>Problem 19.2 Evaluating Correlations with Raw Data</i>	<i>88</i>
<i>Problem 20.2 Evaluating Correlation Significance.....</i>	<i>88</i>
<i>Problem 21.2 Simple Regression with Raw Data</i>	<i>88</i>
<i>Problem 22.2 Predicting Y with Raw Data</i>	<i>89</i>

The Basics

Workbook Format

The Workbook is designed to supplement a basic statistics course. It uses a very applied approach to learning statistics through a series of lessons that connect hand calculations to raw data, presentation tables, and interpretation. Workbook lessons start with an example and hand calculation exercise. In addition, problems are provided to introduce how to apply the statistical technique to raw data and summary statistics. These problems require you to create and interpret tables. Each lesson provides space for instructors to reference readings and additional computational and interpretation exercises that are available in any of several excellent statistics textbooks. Additional exercises are included in the appendix. More background material is provided in the Research Methods Handbook that accompanies AcaStat software. All Workbook problems use alpha .05 as the benchmark for statistical significance.

About AcaStat Software

The Student Workbook is designed to be used with AcaStat or SPSS. AcaStat is a flexible low-cost alternative for producing statistics at school or work. Many of the basic skills learned while using AcaStat are directly transferable to other statistical software. For more information on purchasing AcaStat or downloading a free evaluation, please visit <http://www.acastat.com/prod07.htm>.

If using SPSS instead of AcaStat, Workbook exercises using summary data cannot be completed in SPSS. For these assignments, AcaStat provides an Excel spreadsheet that will complete all of the assignments normally completed in the SumStats module of AcaStat (see next page for details).

Hand Calculations

The objective of hand calculations is to introduce the basic underlying mathematics and concepts of an analytical technique. Unless instructed otherwise, round all calculations to three decimal places. You should also note that you can check your calculations with SumStats. There will be some variations since SumStats does not round summary estimates during the calculation process (SumStats is more accurate).

Data Terminology

Raw data (sometimes called source data) is data that has not been processed for use. Raw data that has undergone processing is referred to as *summary data*.

Analyzing Raw Data with the Data Grid Module


The objective of the Data Grid spreadsheet problems is to connect hand calculations using summary statistics to the more common practice of analyzing raw data with statistical software. If you do not have AcaStat, these exercises can be used with other statistical software to include SAS, SPSS, STATA, etc.

Analyzing Summary Data with the SumStats Module

The objective of SumStats problems is to show you how to conduct statistical comparisons when the raw data are not available. This is often the case if you only have summary statistics from a report produced by someone else. SumStats simply automates the hand calculations taught in the Workbook and Research Methods Handbook.

If using SPSS instead of AcaStat, please visit <http://www.acastat.com/spreadsheets.htm> for more details on downloading SumStats.xls. Although the macros will not work in non-Excel spreadsheet software, the spreadsheet will likely be compatible with other software as long as the spreadsheet software you are using supports Excel formulas and functions.

Lesson 1: Variables


	Reading Assignment:	
	Additional Exercises:	

A variable is defined as a characteristic that can form different values from one observation to another. If a characteristic is not a variable, it is a constant. As an example, if a study is conducted only of males, then sex cannot be a variable. Constants may also be referred to as controls. There are four levels of measurement: nominal, ordinal, interval, and ratio. Objects classified by type or characteristic with no specific order are a nominal level of measurement. Objects classified by type or characteristic with some logical order are ordinal. Objects classified by type or characteristic, with logical order and equal differences between levels of data are interval. Interval data with a zero starting point are a ratio level of measurement. The unit of analysis is the object under study. This could be people, schools, cities, countries, organizations, etc.

Problem 1.1 Classifying Characteristics

Fill in the blank boxes for the following characteristics. If a characteristic is a constant, provide the level of measurement for the underlying variable.

Characteristic	Constant/Variable	Level	Unit of Analysis
Hispanic			
Murder rate per 100,000 people	Variable	Ratio	cities, states, countries
Personal Income			
Intelligence score	Variable	Interval	people
Letter grade	Variable	Ordinal	student
Military Rank			
Religion			
Temperature			
Federal Agencies			
Blood pressure			
Kilometers	Variable	Ratio	cars, people who travel, cities
Male	Constant	Nominal	people
Sex (male or female)			

Lesson 2: Data File Basics		
	Reading Assignment:	
	Additional Exercises:	

The best way to envision a data file is to use the analogy of the common spreadsheet. In spreadsheets, you have columns and rows. In a rectangular data file, columns represent variables and rows represent observations. Variables are commonly measured as either numerical values or character strings. A numerical variable is best used whenever you wish to manipulate the data mathematically. Examples would be age, income, temperature, and job satisfaction rating scale. A string variable is used whenever you wish to treat the data entries as words. Examples would be names, cities, case identifiers, and race. Many times variables that could be considered string are coded as numeric. As an example, data for the variable "sex" might be coded 1 for male and 2 for female instead of using a string variable that would require characters (e.g., "Male" and "Female"). This has two benefits. First, numerical entries are easier and quicker to enter. Second, manipulation of numerical data with statistical software is generally much easier than using string variables. It is strongly recommended that string variables only be used when it is not possible to code entries as numbers (e.g. addresses, names, etc.). This is the standard practice for research data files. The following problems in this lesson will walk you through creating and formatting a data file in the Data Grid spreadsheet.

Problem 2.1 How to Create a Data File

Open AcaStat and select the Data Grid tab to begin creating a data file. The data you will be entering are on the following page. These data represent a random sample of 50 cases from the GSS93 data and will be used in many of problems in the Workbook. The data contain seven variables (columns) of data. Each row of data represents one person (your unit of analysis) who responded to the General Social Survey of 1993. This survey represents the characteristics and attitudes of a random sample of adults in the United States in 1993.

Steps to Formatting Variables:

1. Variable names should be short (8 or less characters).
2. If you look at the data table on the following page, you will see seven variables: Idnum, Age, Edu, Sex, Manager, JobSat, Income. Each column in the Data Grid spreadsheet must be formatted to represent one of these variables.
3. To begin formatting columns, make sure the Data Grid is visible and select the Variable Format pull-down menu. Click the Format Variables option and use the variable name list to select "V1". Replace "V1" with the first variable name "Idnum" and click the "Save" button.
4. Use the variable name list to select "V2". Replace "V2" with the variable name "Age" and click the "Save" button. Continue until you have formatted the seven columns and then **save the data file** as "Workbook".
5. ENTER THE DATA FROM THE TABLE ON THE FOLLOWING PAGE. Click on a cell to begin entering data (start in the first column and first row). After completing one cell's entry, pressing an arrow key or the Enter key on your keyboard will move you to an adjoining cell. Double click on a cell to edit contents. Once you have completed data entry, **save the data file and then reopen the saved file**. AcaStat removes the extra columns and rows when saving a file, so reopening presents just the columns and rows of data you entered.

<i>Idnum</i>	<i>Age</i>	<i>Edu</i>	<i>Sex</i>	<i>Manager</i>	<i>JobSat</i>	<i>Income</i>
101	37	12	2	1	1	26872
102	36	12	1	1	1	44992
103	55	20	1	1	0	116920
104	57	12	1	1	0	31478
105	35	19	1	2	1	52228
106	52	09	2	1	0	24976
107	26	13	1	1	0	25311
108	48	12	2	1	1	28086
109	32	12	2	1	1	22710
110	38	14	2	2	1	43782
111	52	14	1	9	1	31840
112	32	12	2	1	0	31734
113	51	12	2	1	1	27381
114	50	12	2	1	1	23500
115	39	12	1	1	0	35366
116	39	19	2	2	0	57720
117	36	12	2	1	0	33198
118	43	12	1	1	1	24654
119	48	10	2	1	0	24064
120	41	12	1	1	1	25498
121	39	14	2	9	1	36536
122	38	16	1	2	1	66118
123	39	18	1	2	1	58578
124	29	12	2	2	1	22041
125	41	11	2	1	1	26462
126	50	16	2	2	1	40720
127	35	15	2	1	1	29373
128	26	12	2	9	0	20327
129	23	16	2	2	1	33699
130	34	16	2	2	0	37581
131	43	12	1	1	0	29247
132	19	11	1	1	1	19071
133	47	12	1	1	1	55120
134	38	08	1	1	0	18938
135	37	12	1	1	0	25185
136	40	12	2	1	1	40480
137	28	17	1	2	0	34942
138	33	16	1	1	1	25101
139	35	17	1	1	0	47783
140	21	15	2	1	1	23766
141	35	14	1	1	1	25827
142	40	12	1	1	0	25120
143	63	12	1	1	1	38199
144	21	13	1	1	0	19337
145	28	15	2	1	0	23608
146	62	09	2	1	1	19542
147	38	10	1	1	0	29266
148	53	12	2	9	1	34168
149	37	16	1	1	1	38298
150	33	17	1	1	1	29186

Problem 2.2 Formatting the Data File

Make sure the "Workbook.dcs" file is open and visible in the Data Grid and there are no empty rows and columns before proceeding. You will notice that the "Workbook" data file has numbers in some columns that appear to represent words. As an example, the sex variable contains the numbers 1 and 2 instead of the character strings "Male" and "Female". To make your analyses easier to read, you can format the data file to place in the analysis output what each of these values represent (value labels) and a longer description of the variable (variable label). To help you format the variables, a data dictionary is provided on the following page. The dictionary has the variable name (which you have already entered), a variable label, value labels (if any), and missing values (if any).

Steps:

1. Open the Format Variable screen and select the "Idnum" variable.
2. Enter "Respondent Number" in the Variable Label textbox. Since there are no value labels or missing values for this variable, click "Save". Continue formatting variable labels for the remaining columns.
3. The variable "Sex" is the first variable that will need value labels. To format the values, enter "1" in the value textbox and then "Male" in the adjoining label textbox. Repeat the process for Female (note: 2 = Female).
4. Click the File Information bar on the control panel and enter the following information in the Project Notes textbox:

"Employee Data

Based on a random sample of 50 cases. Used for Workbook exercises.

5. **Save the data file.**
6. Once the data have been formatted and saved, click the Data pull-down menu and select Create Data Dictionary to produce the data dictionary for the "Workbook" data file. It should match the data dictionary on the following page.

Filename: Workbook.dcs
Number of Variables: 8
Number of Records: 50
File Size: 1877 bytes
Last Modified: 3/15/2007 5:56:00 PM

Variable Name: Idnum
Variable Label: Respondent Number
Value Labels: None
Missing Values: None

Variable Name: Age
Variable Label: Respondent age in years
Value Labels: None
Missing Values: None

Variable Name: Edu
Variable Label: Years of education
Value Labels: None
Missing Values: None

Variable Name: Sex
Variable Label: Respondent sex
Value Labels:
1 = Male
2 = Female
Missing Values: None

Variable Name: Manager
Variable Label: In Management Position?
Value Labels:
1 = No
2 = Yes
Missing Values: 9

Variable Name: JobSat
Variable Label: Satisfied with job
Value Labels:
0 = No
1 = Yes
Missing Values: None

Variable Name: Income
Variable Label: Respondent individual income
Value Labels: None
Missing Values: None

Problem 2.3 Recoding Data

It is sometimes easier to describe, compare, and interpret data if it is recoded into more meaningful groupings. As an example, the Workbook data file contains education in years. This is very useful if you wish to describe education with a mean but it is not in a format that will easily tell us how many of our sample are high school graduates or how many have attended college, etc. Data Grid provides a module that allows you to create a new education variable that converts years of education into discrete categories. This will increase your data file from seven variables to eight and allows you to describe education using continuous data (means) or to describe education using discrete subgroups (percents). The following procedure will create a new education variable that uses years of education to categorize education as less than high school, high school, and some college.

Please Note

This problem must be successfully completed before continuing to the next lesson.

Steps:

1. Select the Variable pull-down menu and click Recode Variables. Use the variable name list to select the variable you wish to use to create a recode. In this case, select "Edu".
2. Click the "<" option button and enter "12" in the Value textbox.
3. Enter in the NewVar textbox "EduCat". This will be the variable name for the recoded variable. Enter "1" in the New Value textbox.
4. Click Run Recode. This procedure will create a variable named "EduCat" and all values in the "Edu" variable that are less than 12 years are represented by the value "1" in the "EduCat" variable.
5. To continue the recode, click the "=" option button and leave "12" in the Value textbox. Keep the same NewVar name but change the New Value to "2". Click Run Recode.
6. Next, click the ">" option button and leave "12" in the Value textbox. Keep the same NewVar name but change the New Value to "3". Click Run Recode and then close the module.
7. To complete the process, format the new variable "EduCat" to the following: Variable label = "Education Level", value 1= "<12 yrs" 2= "HS Grad" 3= "College".
8. Save the data file as "Workbook" and then use the Statistical Procedures panel to run a listing of the variables Idnum, Edu, and EduCat by selecting the List Variables option and moving the variables into the list variable box and clicking Run Procedure. Review the listing to verify that the coding was successful. If not successful, delete the EduCat column and do the recode operation again.

Problem 3.2 Hand Calculating Bivariate Descriptions

The following exercise involves creating and interpreting contingency tables, also called crosstabulations.

For the following table, calculate the missing statistics. An example is provided for adults with less than 12 years of education.

Column %: Of those who are female, 16.67% (4/24) have less than a high school education.

Of those who are male, 11.54% (3/26) have less than a high school education.

Row %: Of those who have less than a high school education, 57.14% (4/7) are female.

Of those who have less than a high school education, 42.86% (3/7) are male.

Total %: 8% (4/50) of the sample are females who have less than a high school education.

6% (3/50) of the sample are males who have less than a high school education.

Crosstabulation: EduCat (Rows) by Sex (Columns)

Column Variable Label: Respondent sex

Row Variable Label: Education Level

Count	Male	Female	Total
< 12 yrs	3	4	7
Row %	42.86	57.14	
Col %	11.54	16.67	14.00
Total %	6.00	8.00	
HS Grad	10	11	21
Row %		52.38	
Col %			22.00
College	13	9	22
Row %			44.00
Col %			
Total	26	24	50
Row %		48.00	100.00

Problem 3.3 Univariate Analysis of Raw Data

Open the Workbook data file in Data Grid and use the Frequency procedure in the Statistical Procedures panel to answer the following questions:

1. What percentage of the sample is 40 years of age?
2. What percentage of the sample is 40 years of age or younger?
3. What percentage of the sample completed more than 12 years of education?
4. What percentage of the sample is female?

Problem 3.4 Bivariate Analysis of Raw Data

Use the Workbook data file and Crosstabulation procedure in the Statistical Procedures panel to answer the following questions:

1. What percentage of females are managers?
2. Of those who managers, what percentage is male?
3. What percentage with some college education are managers?
4. Based on the three levels of education, who are more likely to be managers?

Problem 3.5 Creating a Summary Table


Use Frequencies (sex and EduCat) and Crosstabulation (columns=sex, rows=EduCat) to create summary statistics from the Workbook data file to complete the following table.

Education level for all subjects in the sample and by sex subgroups

	Total	Male	Female
Observations (n) =			
Education Level (column %)			
Less than high school			
High School			
College			

Interpretation (i.e., Based on this table, do men and women differ?): _____

Lesson 4: Describing Continuous Data

	Reading Assignment:	
	Additional Exercises:	

For interval/ratio level data, measures of central tendency and measures of variation are common descriptive statistics. Measures of central tendency describe a series of data with a single attribute. Measures of variation describe how widely the data elements vary.

Measures of Central Tendency

Mode: The most frequently occurring score.

Median: The point on a rank ordered list of scores below which 50% of the scores fall. If the number of scores is odd, the median is the score located in the position represented by $(n+1)/2$. If the number of scores is even, the median is the average of the two middle scores.

Mean: The sum of the scores is divided by the number of scores (n) to compute an arithmetic average of the scores in the distribution.

Measures of Variation

Range: The difference between the highest and lowest score (high-low).

Variance: The average of the squared deviations between the individual scores and the mean. The larger the variance the more variability there is among the scores.

Standard deviation: The square root of variance. It provides a representation of the variation among scores that is directly comparable to the raw scores.

Problem 4.1 Hand Calculations

Students in Class A were taught statistics by the conventional method using a textbook for course instruction. Students in Class B were taught with a combination of textbook instruction and hands-on computer problem sets. At the end of the courses, both classes were evaluated by testing the students on 15 statistical exercises. The data for the two classes are provided below. Assume the data represent random samples from two large classes. A higher score indicates more successfully completed exercises. (round to 2 decimals)

Class A - Conventional Instruction			
Student	Score X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
A	10	1.8	3.24
B	9	0.8	0.64
C	5	-3.2	10.24
D	7	-1.2	1.44
E	6	-2.2	4.84
F	7	-1.2	1.44
G	8	-0.2	0.04
H	7	-1.2	1.44
I	12	3.8	14.44
J	11	2.8	7.84

Class B - Alternative Instruction			
Student	Score X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
K	12		
L	10		
M	7		
N	9		
O	11		
P	9		
Q	9		
R	10		
S	14		
T	13		

n=	10	Sum of Sq=	
$\Sigma X_i =$	82	$S^2 =$	5.07
$\bar{X} =$	8.2	S =	2.25
Mode=	7		
Median=	7.5	use sample variance measures (n-1)	

n=		Sum of Sq=	
$\Sigma X_i =$		$S^2 =$	
$\bar{X} =$		S =	
Mode=			
Median=		use sample variance measures (n-1)	

On average, which class completed more exercises? _____

Which class had more variation among students in exercises completed? _____

Explain: _____

Problem 4.2 Creating Summary Statistics with Raw Data

Open the Workbook data file in Data Grid and use the Descriptives procedure in the Statistical Procedures panel to answer the following questions:

1. What is the mean and median income for our sample?

2. Describe the distribution of income shown in the histogram (mention skewness and kurtosis).

3. Set any income over 100,000 in Workbook data as missing and calculate descriptive statistics. Compare the mean and median income to that reported in question 1.

4. Compare the variation in the sample when income over 100,000 is included in the calculations to the variation in the sample when income over 100,000 is missing.

Lesson 5: Standardized Z-Scores

i	Reading Assignment:	
	Additional Exercises:	

A standardized z-score represents the relative position of an individual score in a distribution as compared to the mean and the variation of all scores in the distribution. A negative z-score indicates the score is below the distribution mean. A positive z-score indicates the score is above the distribution mean. The z-score value represents the number of standard deviations between the original score and the mean. To obtain a standardized score you must subtract the mean from the individual score and divide by the standard deviation. Standardized scores provide you with a score that is directly comparable within and between different groups of cases.

Problem 5.1 Hand Calculations

Using the same data as in Lesson 4, calculate z-scores for each student. (round to 2 decimals)

Class A - Conventional Instruction			
Student	Score X_i	$X_i - \bar{X}$	$Z = \frac{X_i - \bar{X}}{S}$
A	10	1.8	0.80
B	9	0.8	0.36
C	5	-3.2	-1.42
D	7	-1.2	-0.53
E	6	-2.2	-0.98
F	7	-1.2	-0.53
G	8	-0.2	-0.09
H	7	-1.2	-0.53
I	12	3.8	1.69
J	11	2.8	1.24

Class B - Alternative Instruction			
Student	Score X_i	$X_i - \bar{X}$	$Z = \frac{X_i - \bar{X}}{S}$
K	12		
L	10		
M	7		
N	9		
O	11		
P	9		
Q	9		
R	10		
S	14		
T	13		

$$\bar{X} = \boxed{8.2}$$

$$S = \boxed{2.25}$$

$$\bar{X} = \boxed{10.4}$$

$$S = \boxed{2.12}$$

Who did comparably better in their class, student B or student N? _____

Explain (defend) your answer: _____

Problem 5.2 Confirming Hand Calculations with SumStats

Instructions for starting SumStats
Open the "SumStats" panel.
Select the Standardized z-scores option.

Steps:

- ✓ Enter Student B's score from Problem 5.1 into score box in SumStats.
- ✓ Enter the mean for B's Class into the mean box in SumStats.
- ✓ Enter the standard deviation for B's Class into the Std Dev box in SumStats.
- ✓ Click Run Procedure
- ✓ Enter the results in the table below and repeat the process for Student N.
- ✓ The results should be similar except for slight variations due to rounding. If not, redo your hand calculations to find the error.

	Raw Score	Z-Score
Student B		
Student N		

Problem 5.3 Creating Standardized Scores with Raw Data

Enter the following data into Data Grid and follow the steps below to complete the table on the following page.

Data ⇒

Student	History	Math	English	Science
A	70	66	90	88
B	75	65	85	95
C	90	80	60	83
D	86	64	66	73
E	84	54	74	54
F	87	91	87	85
G	74	46	76	67
H	90	80	93	87
I	82	62	72	77
J	95	65	85	73
K	67	57	87	68
L	72	73	74	72
M	84	74	92	87
N	85	72	88	66
O	79	59	89	63
P	95	75	85	73
Q	88	68	78	91
R	85	95	75	77
S	87	67	55	82
T	91	81	87	94

Steps:

Create a new data file by entering the data into the spreadsheet. Make sure you format each column with the variable name indicated. Delete extra columns.

Save the data file as "zscore data".

Choose the Descriptives option and select the save z-scores option.


Select History, Math, English, and Science variables from the listbox (put them in the analysis list) and run the Descriptives procedure. This should create four additional z-score variables (view Data Grid to confirm).

The data entered from the previous page represents the mid-year exam scores for twenty students. Some subjects, such as math, generally have lower mean scores than others. In addition, since each subject is taught by a different teacher, the difficulty of the mid-year exams will vary. You must identify the weakest course for each student so that remedial instruction, if needed, can be arranged.

- a. Use the standardized scores (z-scores) to advise each student on where their skills are the weakest.
- b. If their weakest subject has a score more than two standard deviations below the mean, remedial instruction is mandatory. This is a z-score less than -2.0. (round to hundredths “.00”)

Student	Weakest Subject	Exam Score	Z-Score	Remedial Instruction?	
				Yes	No
A				Yes	No
B				Yes	No
C				Yes	No
D				Yes	No
E				Yes	No
F				Yes	No
G				Yes	No
H				Yes	No
I				Yes	No
J				Yes	No
K				Yes	No
L				Yes	No
M				Yes	No
N				Yes	No
O				Yes	No
P				Yes	No
Q				Yes	No
R				Yes	No
S				Yes	No
T				Yes	No

Lesson 6: Hypotheses

	Reading Assignment:	
	Additional Exercises:	


A hypothesis is a formal statement that presents the expected relationship between an independent and dependent variable. A dependent variable is a variable that contains variations for which we seek an explanation. An independent variable is a variable that is thought to affect (cause) variations in the dependent variable.

Problem 6.1 Developing Hypotheses

For each of the following pairs of variables, indicate which is likely to be the independent and dependent variable. Write a brief research hypothesis to describe the relationship.

Variables	DV	IV	Hypothesis
Person's Sex		X	The body height of men will generally be greater than the body height of women.
Body Height	X		
Child IQ			
Parent IQ			
Crime Rate	X		As the number of police officers per capita increases, the crime rate will decrease.
Police per capita		X	
Jobless Rate			
Sales Tax Revenue			
Race			
Political Party Affiliation			

Lesson 7: Random Sampling

	Reading Assignment:	
	Additional Exercises:	

Inferential measures represent an estimate of the true value of an unknown population characteristic. Since they are based on less than the entire population, their usefulness is dependent on the concept of random sampling. A random sample is one where each and every element in a population has an equal opportunity of being selected for the sample. As a general principle, a random sample is likely to be more representative of the true population parameter as the sample size increases.

Problem 7.1 Comparing Random Samples to a Population

Use the AFQT data file to complete the following table. Assume that the 1000 observations in the AFQT data are the entire population. Begin by calculating frequencies for the variable "INCLEVEL". These represent the population parameters and should match the entries in the table below. Proceed with the following steps to conduct several random samples of varying sizes.

Steps

1. Open the AFQT data file and select the Random Sample option in the Statistical Procedures panel. Enter a sample size (start with 15) and click Create Sample. Run Frequencies on INCLEVEL.
2. Record the percentages in the table below. (round to a whole number)
3. Repeat steps until the table is complete (start by reopening the AFQT data file). Please note that you will run three random samples of $n=15$ and three of $n=200$.

Income Level	Parameter	Random Sample Size (n)					
		15	15	15	200	200	200
Low Income	18%						
Moderate Income	78%						
High Income	4%						

Compare the percentages from the random samples to the known population parameters for each income level.

Use the AFQT data file to complete the table on the following page. Assume that the 1000 observations in the AFQT data represent the entire population. The purpose of this exercise is present some of the basic premises of inferential statistics by creating a simulation of a sampling distribution and comparing it to the distribution of one random sample and the distribution of the population (already provided with a reference line representing the population mean).

Steps

1. Open the AFQT data file and select the Repeated Sampling option in the Statistical Procedures panel.
2. For conducting a repeated random sampling, select AFQT, enter a sample size of 30 and enter 500 for the number of iterations. Click Run Iterations.
3. Click the Charts tab and chart the histogram on the next page and record the mean and sample standard deviation.
4. Open the AFQT data file again and select the Random Sample option in the Statistical Procedures panel.
5. Create one random sample for a sample size of 30. Chart the histogram for AFQT in the following table and record the mean, sample standard deviation, and standard error.
6. For the sampling and one-sample histograms, indicate on the x-axis the location of the AFQT mean and one standard deviation above and below the mean.

Problem 7.2 Sample v. Simulated Sampling Distributions

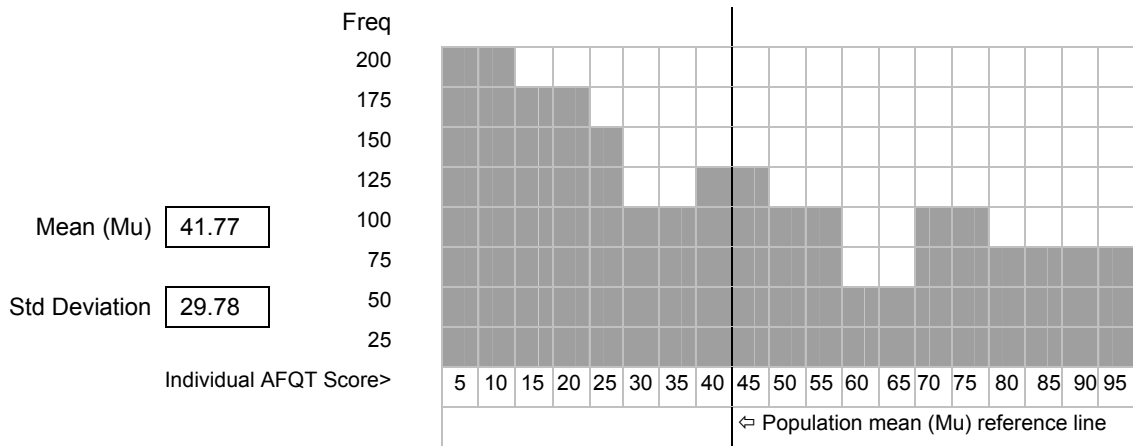
1. Compare the means of the population, sampling distribution, and one random sample of 30 cases (are they similar? different? explain).

2. If you did not know the true population mean for AFQT, do you believe the mean from the sampling distribution would serve as an accurate surrogate? Why?

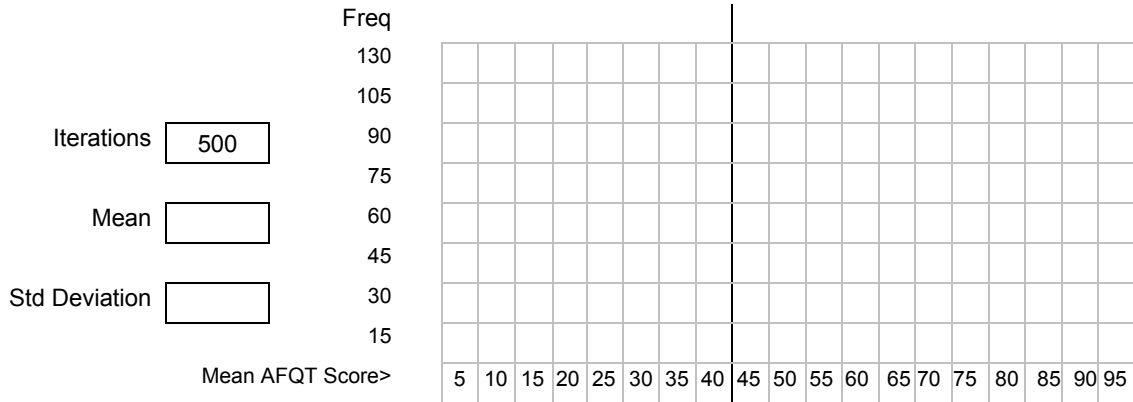
3. Compare the standard deviation of the sampling distribution and the one random sample of 30 cases. (are they similar? different? explain).

4. Compare the standard deviation of the sampling distribution to the standard error of the random sample of 30 cases. (are they similar? different? explain).

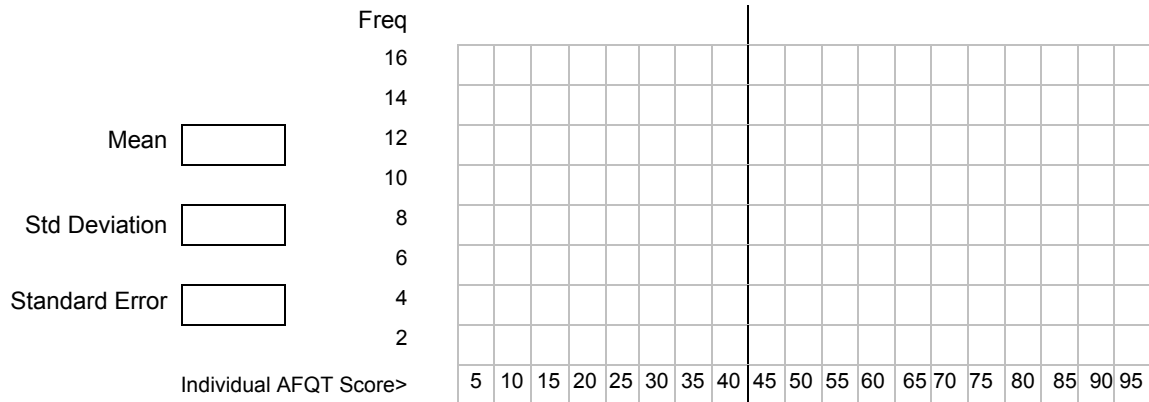
Population Distribution of Individual AFQT Scores




Sampling distribution based on 500 Random Samples of 30 Individual AFQT Scores



Distribution of one Random Sample of 30 Individual AFQT Scores



Lesson 8: Hypothesis Testing

	Reading Assignment:	
	Additional Exercises:	

When conducting statistical tests with computer software, the exact probability of a Type I error is calculated. It is presented in several formats but is most commonly reported as "p <" or "Sig." or "Signif." or "Significance." Using "p <" as an example, if a priori you established a threshold for statistical significance at alpha .05, any test statistic with significance at or less than .05 is considered statistically significant and you must reject the null hypothesis of no difference.

Problem 8.1 Interpreting p-values

Fill in the blanks for the following table.

P <	Alpha	Type I Error	Final Decision
.05	.05	5% chance difference is not significant	Statistically significant
.10	.05	10% chance difference is not significant	Not statistically significant
.02	.05	___% chance difference is not significant	
.96	.05	___% chance difference is not significant	
.001	.05	___% chance difference is not significant	
.09	.05	___% chance difference is not significant	
.12	.05	___% chance difference is not significant	
.50	.05	___% chance difference is not significant	

Indicate in the following table what is and is not statistically significant.

Mean scores from a survey of public and private sector employees on attitudes toward the work place (higher scores represent a greater level of each characteristic). Alpha = .05.					
	Public Employee	Private Employee	p (2-tailed)	Significant?	
Job Satisfaction	6.31	7.97	0.056	Yes	No
Amount of Red Tape	5.82	3.23	0.001	Yes	No
Level of Mgt Support	5.73	5.93	0.543	Yes	No
Job Security	7.85	6.34	0.036	Yes	No

Nominal and Ordinal Data

This section presents techniques for using counts and proportions to create inferential statistics. To use these techniques both the independent and dependent variable must be nominal or ordinal. In addition, the data are assumed to be from a random sample. To conduct significance tests, the test statistic is compared to a critical value established from a sampling distribution. Tables of critical values are attached at the end of the Workbook for hand calculation problems.

Indicate for the following comparisons whether the techniques presented in this section are applicable to the variables indicated. Remember both the independent and dependent variables must be nominal or ordinal.

Dependent Variable	Independent Variable	Fits this section?	
		Yes	No
individual annual income	sex	Yes	No
religious (yes/no)	race	Yes	No
supports nat'l health (yes/no)	income level (low, moderate, high)	Yes	No
neighborhood (urban/suburb)	have children (yes/no)	Yes	No
crime rate per capita	city population	Yes	No
body weight (kilograms)	body height (centimeters)	Yes	No
political party preference	race	Yes	No
physicians per capita	census region	Yes	No
voted in election (yes/no)	student status (yes/no)	Yes	No
education in years	parent's education level	Yes	No
savings in dollars	age in years	Yes	No

Problem 9.2 Using Summary Statistics in SumStats

Instructions for starting SumStats
Open the "SumStats" panel.
Select the Margin of Error for Proportions option.

- ✓ Enter the count (n) and proportion into the SumStats boxes and Click Run Procedure.
- ✓ Fill in the blank boxes below for 95% CI. (round to 1 decimal)

The promotion system is ...	Staff	+/- Margin of Error	95% CI
Fair	50%		
Sometimes Fair	30%		
Not Fair	20%		
Count (n)	100		

1. Interpret the 95% confidence interval for staff who views the promotion system as fair.

2. Interpret the 95% confidence interval for staff who views the promotion system as sometimes fair.

3. Interpret the 95% confidence interval for staff who views the promotion system as not fair.

<p>Decide Results of Null Hypothesis</p> <p>Since the test statistic of 1.724 did not meet or exceed the critical value of 1.96, you must conclude there is no statistically significant difference between the historical proportion of clients reporting poor service and the current proportion of clients reporting poor service.</p>	<p>Decide Results of Null Hypothesis</p>
--	---

Problem 10.2 Creating Summary Statistics with Raw Data

Open the GSS93 data file in Data Grid and use Statistical Procedures to run Frequencies on the variable "WrkStat". Determine how many total responses are available and what proportion of the sample are retired (don't forget to convert percent to a proportion: e.g. 20% = .20). You should note that the Statistical Procedures Panel does not conduct z-tests, so the purpose of this exercise is to create summary statistics to use in problem 10.3.

	Valid Responses (n)	Percent Retired	Proportion Retired
Work Status			

Problem 10.3 Using Summary Statistics in SumStats


Instructions for starting SumStats
Open the "SumStats" panel.
Select the Z-Test option.
Use the One sample Proportion section to complete the following assignment.

Historical records indicated that in 1980, 13% (.13) of all adults in the United States were retired. Use the information provided here and from the random sample in problem 10.2 to fill in the blanks in the following table. Since you are using summary data, you will need to use the SumStats z-test statistics module (one-sample proportion) to answer the question that follows. Note that historical data represent the population proportion.

Title:			
	Historical 1980	Sample 1993	p-value (2-tailed)
Proportion Retired			
	Count (n)		

Is there evidence that the proportion of retirees in the United States has increased or decreased since 1980? Explain.

Lesson 11: Comparing Proportions - Two Independent Samples

	Reading Assignment:	
	Additional Exercises:	

Commonly called a two-sample z-test, this technique compares two proportions from two random samples; such as males and females, Republicans and Democrats, Urban and Rural residents. If there are more than two comparable groups, chi-square is a more appropriate technique (e.g., Republican, Independent, Democrat or urban, suburban, rural comparisons).

Problem 11.1 Hand Calculations

<p>Example: A survey was conducted of students from the Princeton public school system to determine if the incidence of hungry children was consistent in two schools located in lower-income areas. A random sample of 80 elementary students from school A found that 23% did not have breakfast before coming to school. A random sample of 180 elementary students from school B found that 7% did not have breakfast before coming to school.</p>	<p>Problem: A survey was conducted of unemployed adults in Oregon and Ohio to determine if there are differences in the proportion who are chronically unemployed. A random sample of 110 unemployed adults in Oregon found that 15% were chronically unemployed. A random sample of 100 unemployed adults in Ohio found that 8% were chronically unemployed. Although there is a 7 percentage point difference, can we safely assume that the chronic unemployed rate is greater in Oregon?</p>
<p>State the Hypothesis</p> <p>Ho: There is no statistically significant difference between the proportion of students in school A not eating breakfast and the proportion of students in school B not eating breakfast.</p> <p>Ha: There is a statistically significant difference between the proportion of students in school A not eating breakfast and the proportion of students in school B not eating breakfast.</p>	<p>State the Hypothesis</p> <p>Ho:</p> <p>Ha:</p>
<p>Set the Rejection Criteria</p> <p>Alpha .05, Zcv = 1.96</p>	<p>Set the Rejection Criteria</p> <p>Alpha .05, Zcv =</p>
<p>Compute the Test Statistic</p> <p><i>Estimate of Standard Error</i></p> $\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \hat{p} = \frac{80(.23) + 180(.07)}{80 + 180}$ $\hat{p} = .119 \quad \hat{q} = .881$ $s_{p1-p2} = \sqrt{\hat{p}\hat{q} \frac{n_1 + n_2}{n_1 n_2}}$ $s_{p1-p2} = \sqrt{.119(.881) \frac{80 + 180}{80(180)}} \quad s_{p1-p2} = .043$	<p>Compute the Test Statistic</p> <p><i>Estimate of Standard Error</i></p>

<p><i>Test Statistic</i></p> $Z = \frac{p_1 - p_2}{s_{p_1-p_2}} \quad Z = \frac{.23 - .07}{.043} \quad Z = 3.721$	<p><i>Test Statistic</i></p>
<p>Decide Results of the Null Hypothesis</p> <p>Since the test statistic 3.721 exceeds the critical value of 1.96, you conclude there is a statistically significant difference between the proportion of students in school A not eating breakfast and the proportion of students in school B not eating breakfast.</p>	<p>Decide Results of the Null Hypothesis</p>

Problem 11.2 Using Summary Statistics in SumStats

Instructions for starting SumStats
Open the "SumStats" panel.
Select the Z-Test option.
Use the Two Sample Proportion section to complete the following assignment.

A random sample survey was conducted by a nonprofit organization of city managers in medium (population 25k to 60k) and small cities (population < 25k). The results were presented in the following table but without statistical analysis. Use the table summary data to compute p-values and interpret the results. Don't forget to convert percents to proportions.

Characteristics of city managers working for small and medium cities

City Manager Characteristics	Small Cities	Medium Cities	p < ^a
Graduate Degree	52%	56%	_____
Male	95%	88%	_____
White	98%	92%	_____
	(n=250)	(n=100)	

^a Two-tailed test for the difference between two sample proportions.

Interpret your results: _____

Lesson 12: Chi-square Test of Independence

i	Reading Assignment:	
	Additional Exercises:	

Chi-square compares two or more subgroups across two or more criteria. By convention, subgroups of the independent variable are placed in the columns and the subgroups of the dependent variable are represented by the rows. (round to 2 decimals)

Problem 12.1 Hand Calculations

Example: You wish to evaluate the association between a person's sex and their attitudes toward school spending on athletic programs. A random sample of adults in your school district produced the following table.			Problem: You wish to evaluate the association between a person's education level and their attitudes toward school spending on athletic programs. A random sample of adults in your school district produced the following table.		
(Counts)	Female	Male	(Counts)	High School	College
Spend more money	15	25	Spend more money	20	18
Spend the same	5	15	Spend the same	23	15
Spend less money	35	10	Spend less money	10	23
State the Hypothesis Ho: There is no association between a person's sex and their attitudes toward spending on athletic programs. Ha: There is an association between a person's sex and their attitudes toward spending on athletic programs.			State the Hypothesis Ho: Ha:		
Set the Rejection Criteria Determine degrees of freedom $df=(3 - 1)(2 - 1)$ or $df=2$ Alpha = .05 Chi-square distribution table, critical value = 5.991			Set the Rejection Criteria Determine degrees of freedom $df= \underline{\hspace{2cm}}$ Alpha = .05 Chi-square distribution table, critical value = $\underline{\hspace{2cm}}$		
Compute the Test Statistic $X^2 = \sum \left[\frac{(Fo - Fe)^2}{Fe} \right]$			Compute the Test Statistic $X^2 = \sum \left[\frac{(Fo - Fe)^2}{Fe} \right]$		

<i>Frequency Observed (Fo)</i>				<i>Frequency Observed (Fo)</i>			
	Female	Male	Row Total		High School	College	Row Total
Spend more money	15	25	40	Spend more money			
Spend the same	5	15	20	Spend the same			
Spend less money	35	10	45	Spend less money			
Column Total	55	50	105	Column Total			
<i>Frequency Expected (Fe)</i>				<i>Frequency Expected (Fe)</i>			
	Female	Male	Row Total		High School	College	Row Total
Spend more money	$55 \cdot 40 / 105 = 20.952$	$50 \cdot 40 / 105 = 19.048$	40	Spend more money			
Spend the same	$55 \cdot 20 / 105 = 10.476$	$50 \cdot 20 / 105 = 9.524$	20	Spend the same			
Spend less money	$55 \cdot 45 / 105 = 23.571$	$50 \cdot 45 / 105 = 21.429$	45	Spend less money			
Column Total	55	50	105	Column Total			
<i>Chi-square Calculations (Fo-Fe)²/Fe</i>				<i>Chi-square Calculations (Fo-Fe)²/Fe</i>			
	Female	Male			High School	College	
Spend more money	$\frac{(15-20.952)^2}{20.952}$	$\frac{(25-19.048)^2}{19.048}$		Spend more money			
Spend the same	$\frac{(5-10.476)^2}{10.476}$	$\frac{(15-9.524)^2}{9.524}$		Spend the same			
Spend less money	$\frac{(35-23.571)^2}{23.571}$	$\frac{(10-21.429)^2}{21.429}$		Spend less money			
Σ	21.200			Σ			
Decide Results of Null Hypothesis Since the chi-square test statistic 21.2 exceeds the critical value of 5.991, you may conclude there is a statistically significant association between a person's sex and their attitudes toward spending on athletic programs. As is apparent in the contingency table, males are more likely to support spending than females.				Decide Results of Null Hypothesis			

Problem 12.2 Calculating Chi-Square with Raw Data

Open the Workbook data file in Data Grid and use the Crosstabulation procedure in the Statistical Procedures panel to evaluate the association between sex, education level (EduCat), and job satisfaction (JobSat) and employee position (Manager). Record in the table below the chi-square results and indicate which relationships are and are not significant.

Round Chi-square to hundredths (".00")

Do not round p-values

Title:					
Relationship	Chi-square	n	p-value	Significant?	
Manager * Sex				Yes	No
Manager * EduCat				Yes	No
JobSat * Manager				Yes	No

Problem 12.3 Calculating Chi-Square from Summary Statistics

Instructions for starting SumStats
Open the "SumStats" panel.
Select the Chi-Square option.
Convert the percentages below into counts.
Enter the counts into SumStats starting at the top left white cell (represents Staff who said "Fair").
Click Run Procedure.

A random sample of company employee attitudes toward the promotion system resulted in the summary statistics given below.

Round Chi-square to hundredths (".00")

Do not round p-values

Attitudes of Staff and Management concerning the promotion system.

The promotion system is ...	<u>Staff</u>	<u>Managers</u>
Fair	50%	70%
Sometimes Fair	30%	13%
Not Fair	20%	17%
Count	100	30

Compute chi-square _____ and report the p-value _____.

Interpretation: _____

Lesson 13: Measuring Association for Chi square

i	Reading Assignment:	
	Additional Exercises:	

When using the chi-square statistic, Cramer's V coefficients can be helpful in interpreting the strength of a relationship between two variables once statistical significance has been established. Cramer's V is also useful for comparing multiple X^2 test statistics and is generalizable across contingency tables of varying sizes. It is not affected by sample size and therefore is very useful in situations where you suspect a statistically significant chi-square was the result of large sample size instead of any substantive relationship between the variables. It is interpreted as a measure of the relative (strength) of an association between two variables.

$$V = \sqrt{\frac{X^2}{n(q-1)}} \quad \text{where } q = \text{smaller \# of rows or columns}$$

Problem 13.1 Hand Calculations

Use the information in problem 12.2 to compute Cramer's V. Assuming that the assumptions for chi square are met, which characteristic has the strongest relationship with employee position?

Round Chi-square and Cramer's V to hundredths (".00")

Do not round p-values

Relationship	Chi-square	p-value	n	q	Cramer's V
Manager * Sex					
Manager * EduCat					
JobSat * Manager					

Problem 13.2 Interpreting Multiple Comparisons

Open the GSS93 data file in Data Grid and use the Chi-square option in the Statistical Procedures panel to examine the relationship between PartyID and Sex, Race, and AgeCat4. Complete the table and indicate which comparisons are significant and then rank order (highest=1) the relationships by the strength of the relationship.

Round Chi-square and Cramer's V to hundredths (".00")

Do not round p-values

Title:						
Relationship	Chi-square	p-value	Signif. ?		Cramer's V	Rank (1,2,3)
PartyID * Sex			Yes	No		
PartyID * Race			Yes	No		
PartyID * AgeCat4			Yes	No		

Problem 13.3 Review the Basics

Identify the independent and dependent variables, level of measurement for each, and unit of analysis (from Workbook data file).

Relationship	Dependent Var		Independent Var		Unit of Analysis
	DV	Level	IV	Level	
Manager * Sex					
Manager * EduCat					
JobSat * Manager					

Interval and Ratio Data

This section presents techniques for using continuous data to create inferential statistics. For the methods reviewed in this section, the dependent variable must be interval or ratio level data. The type of technique used will depend on the level of the independent variable (nominal/ordinal or interval/ratio).

Indicate for the following comparisons whether the techniques presented in this section are applicable to the variables indicated.

Dependent Variable	Independent Variable	Fits this section?	
		Yes	No
individual annual income	sex	Yes	No
religious (yes/no)	race	Yes	No
supports nat'l health (yes/no)	income level (low, moderate, high)	Yes	No
neighborhood (urban/suburb)	# of children	Yes	No
crime rate per capita	city population	Yes	No
body weight (kilograms)	body height (centimeters)	Yes	No
political party preference	race	Yes	No
physicians per capita	census region	Yes	No
voted in election (yes/no)	student status (yes/no)	Yes	No
education in years	parent's education level	Yes	No
savings in dollars	age in years	Yes	No

Problem 14.2 Creating Summary Statistics with Raw Data

Use Workbook data to calculate descriptive statistics for the variables age, edu, and income. Assume this is a random sample of employees in one organization. Enter the results below. They will be used for problem 14.3.

Round means and standard deviations to hundredths (".00")

Title: The mean age, education, and income of employees.			
Characteristics	Mean	Std Dev.	n
Age (years)			
Education (years)			
Annual Income (Dollars)			

Problem 14.3 Creating Confidence Intervals with SumStats

Instructions for starting SumStats
Open the "SumStats" panel.
Select the Margin of Error for Means option.
Use the results from problem 14.2 to complete the table below.
Enter the values for n, Mean, StdDev (Standard Deviation) for the age variable.
Click Run Procedure, record the results in the table (95% confidence interval), and repeat for each characteristic.

Round to hundredths (".00")


Title: The mean age, education, and income of employees.					
Characteristics	Mean	Std Dev.	n	SumStats Calculations	
				+/-	95% CI
Age (years)					
Education (years)					
Annual Income (Dollars)					

Problem 14.4 Interpreting Confidence Intervals

Interpret the 95% confidence interval for age.

Interpret the 95% confidence interval for education.

Interpret the 95% confidence interval for income.

Lesson 15: Comparing a Population Mean to a Sample Mean (T-test)		
	Reading Assignment:	
	Additional Exercises:	

Commonly called a one-sample t-test, this technique compares a mean obtained from a random sample to a population mean. It is generally used to determine if there is a difference between what was obtained in a random sample and an established benchmark that either originated from all members of a comparable population or is assumed to represent the population.

Problem 15.1 Hand Calculations

<p>Example: Compare the mean age of incoming students to the known mean age for all previous incoming students. A random sample of 30 incoming college freshmen revealed the following statistics: mean age 19.5 years, standard deviation 1 year. The college database shows the mean age for previous incoming students was 18.</p>	<p>Problem: The mean income for all working adults in the state of Indiana was \$35,000 in 1999. A random sample of 88 working adults in Richmond Indiana in 2001 revealed the following statistics: mean income \$39,000, standard deviation \$15,000. Can you conclude that, on average, Richmond workers earn more than other workers in Indiana (assume constant dollars).</p>
<p>State the Hypothesis</p> <p>Ho: There is no significant difference between the mean age of past college students and the mean age of current incoming college students.</p> <p>Ha: There is a significant difference between the mean age of past college students and the mean age of current incoming college students.</p>	<p>State the Hypothesis</p> <p>Ho:</p> <p>Ha:</p>
<p>Set the Rejection Criteria</p> <p>Significance level .05 alpha, 2-tailed test</p> <p>Degrees of Freedom = n-1 or 29</p> <p>Critical value from t-distribution = 2.045</p>	<p>Set the Rejection Criteria</p> <p>Significance level .05 alpha, 2-tailed test</p> <p>Degrees of Freedom _____</p> <p>Critical value from t-distribution =</p>
<p>Compute the Test Statistic</p> <p><i>Standard error of the sample mean</i></p> $s_x = \frac{s}{\sqrt{n}} \quad s_x = \frac{1}{\sqrt{30}} \quad s_x = .183$ <p><i>Test statistic</i></p> $t = \frac{\bar{x} - \mu}{s_x} \quad t = \frac{19.5 - 18}{.183} \quad t = 8.197$	<p>Compute the Test Statistic</p> <p><i>Standard error of the sample mean</i></p> <p><i>Test statistic</i></p>

<p>Decide Results of Null Hypothesis</p> <p>Given that the test statistic (8.197) exceeds the critical value (2.045), the null hypothesis is rejected in favor of the alternative. There is a statistically significant difference between the mean age of the current class of incoming students and the mean age of freshman students from past years. In other words, this year's freshman class is on average older than freshmen from prior years.</p>	<p>Decide Results of Null Hypothesis</p>
--	---

Problem 15.2 Creating Summary Statistics with Raw Data

Open the Workbook data file in Data Grid to run Descriptives on the variable "Income". Determine the summary statistics needed to conduct a one-sample t-test. You should note that the Statistical Procedures panel does not conduct one-sample t-tests, so the purpose of this exercise is to create summary statistics to use in problem 15.3.

Round means and standard deviations to hundredths (".00")

	Workbook Sample
Mean Income	
(Std Deviation)	()
n=	

Problem 15.3 Calculating One-Sample t-tests with SumStats


Instructions for starting SumStats
Open the "SumStats" panel.
Select the T-Test option.
Use the one sample means t-test section to complete the following assignment.

Historical records indicated that in 2000, the mean income of employees in your organization was \$31,000. Use the information developed from the random sample in problem 15.2 to determine if there is evidence that income has changed significantly from 2000 to 2007, the year of the survey. Please note that these are constant dollars (adjusted for inflation so the two measures are comparable). Since you are using summary data, you will need to use the SumStats t-test statistics module (one-sample mean) to answer this question.

Title:			
	2000 Benchmark	2007 Sample	p-value
Mean Income			
(Std Deviation)		()	
	n=		

Interpretation: _____

Lesson 16: Comparing Two Independent Sample Means (T-test)

	Reading Assignment:	
	Additional Exercises:	

Commonly called a two-sample t-test, this technique compares two means from two random samples; such as mean incomes of males and females, Republicans and Democrats, Urban and Rural residents. If there are more than two comparable groups, ANOVA is a more appropriate technique (e.g., Republican, Independent, Democrat or urban, suburban, rural comparisons). The following lesson assumes homogeneity of variance.

Problem 16.1 Hand Calculations

<p>Example: You obtained the number of years of education from one random sample of 38 police officers from City A and the number of years of education from a second random sample of 30 police officers from City B. The average years of education for the sample from City A is 15 years with a standard deviation of 2 years. The average years of education for the sample from City B is 14 years with a standard deviation of 2.5 years. Is there a statistically significant difference between the education levels of police officers in City A and City B?</p>	<p>Problem: The mean age of National Guard soldiers in two states was created from a random sample in each state. The Oregon (OR) sample of 33 soldiers resulted in a mean age of 38 with a standard deviation of 5.3. The Washington (WA) sample of 25 soldiers resulted in a mean age of 42 with a standard deviation of 6.1. Is there a statistically significant difference between the ages of National Guard soldiers in Oregon and Washington?.</p>
<p>Assumptions</p> <p>Random sampling</p> <p>Independent samples</p> <p>Interval/ratio level data</p>	<p>Assumptions</p> <p>Random sampling?</p> <p>Independent samples?</p> <p>Interval/ratio level data?</p>
<p>State Hypotheses</p> <p>Ho: There is no statistically significant difference between the mean education level of police officers working in City A and the mean education level of police officers working in City B.</p> <p>Ha: There is a statistically significant difference between the mean education level of police officers working in City A and the mean education level of police officers working in City B.</p>	<p>State Hypotheses</p> <p>Ho:</p> <p>Ha:</p>
<p>Set the Rejection Criteria</p> <p>Degrees of freedom = $38+30-2=66$</p> <p>Alpha.05, $t_{cv}= 2.000$</p>	<p>Set the Rejection Criteria</p> <p>Degrees of freedom = _____</p> <p>Alpha.05, $t_{cv}=$ _____</p>

Problem 16.2 Calculating t-tests from Raw Data

Open the Workbook data file in Data Grid and use the t-test option in the Statistical Procedures panel to answer the question of whether there is a significant difference between managers and non-managers in their years of education or income. The continuous variable is "Edu" and the categorical variable is "Manager". The Group A value for Manager is 1 and the Group B value is 2 (note that 1=No and 2=Yes in the Manager variable).

Round means and standard deviation to hundredths (".00")

Do not round p-values

Title:			
	Position		
	Non-Manager	Manager	p-value
Mean Education			
(Std Deviation)	()	()	
n=			
Mean Income			
(Std Deviation)	()	()	
n=			

Interpretation: _____

Problem 16.3 Using Summary Statistics in SumStats

Instructions for starting SumStats
Open the "SumStats" panel.
Select the T-Test option.
Use the two sample means t-test section to complete the following assignment.


Conduct significance tests with SumStats to complete the table. Assume this is a random sample of professional employees in a large federal agency and there is equal variance in the two samples.

Comparison between male and female professional employees in their age and incomes.

<u>Characteristics</u>	<u>Female</u> (Std. Dev.)	<u>Male</u> (Std. Dev.)	p < (2-tailed)
Mean Age	39.3 (10.6)	36.8 (10.2)	<input type="text"/>
Mean Income	30,513 (8,403)	37,446 (15,100)	<input type="text"/>
Cases (n)	24	26	

What can you conclude given the results in the above table?

Lesson 17: Computing F-ratio

	Reading Assignment:	
	Additional Exercises:	

The F-ratio is used to determine whether the variances in two independent samples are equal (homogeneous). If the F-ratio is not statistically significant, you may assume there is homogeneity of variance and employ the standard t-test for the difference of means. If the F-ratio is statistically significant, use an alternative t-test computation such as the Cochran and Cox method.

Compare the test statistic with the f critical value (F_{cv}) listed in the F distribution. If the f-ratio equals or exceeds the critical value, the null hypothesis (H_0) $\sigma_1^2 = \sigma_2^2$ (there is no difference between the sample variances) is rejected. If there is a difference in the sample variances, the comparison of two independent means should involve the use of the Cochran and Cox method.

Problem 17.1 Hand Calculations

<p>Example: Statistics from samples used in a t-test.</p> <p>Sample A $S^2 = 20$ $n = 10$</p> <p>Sample B $S^2 = 30$ $n = 30$</p>	<p>Problem: Statistics from samples used in a t-test.</p> <p>Sample A $S^2 = 1750$ $n = 30$</p> <p>Sample B $S^2 = 920$ $n = 50$</p>
<p>Set Rejection Criteria</p> <p>Note: numerator is sample with largest variance</p> <p>df for numerator (Sample B) = $B_n - 1$ or 29</p> <p>df for denominator (Sample A) = $A_n - 1$ or 9</p> <p>Consult F-Distribution table for df = (29,9), alpha.05</p> <p>$F_{cv} = 2.70$</p>	<p>Set Rejection Criteria</p> <p>df for numerator (Sample ___) = _____</p> <p>df for denominator (Sample ___) = _____</p> <p>Consult F-Distribution table for df = (__, __), alpha.05</p> <p>$F_{cv} =$ _____</p>
<p>Compute the Test Statistic</p> $F_{ratio} = \frac{s_b^2}{s_a^2} \quad F = \frac{30}{20} \quad F = 1.50$	<p>Compute the Test Statistic</p>
<p>Compare</p> <p>The test statistic (1.50) did not meet or exceed the critical value (2.70). Therefore, there is no statistically significant difference between the variance exhibited in Sample A and the variance exhibited in Sample B. Assume homogeneity of variance for tests of the difference between sample means.</p>	<p>Compare</p>


Problem 17.2 Interpreting the F-ratio in SumStats t-tests

Re-examine your results from problem 16.3 to determine if you should use the equal or unequal variance results. Put the best estimate of the p-value below and indicate whether the result is based on equal or unequal estimates.

Characteristics	Female	Male	P < (2-tailed)	Homogeneity of Variance	
Mean Age	39.3	36.8	<input type="text"/>	Equal	Unequal
Std. Deviation	10.6	10.2			
Mean Income	30,513	37,446	<input type="text"/>	Equal	Unequal
Std. Deviation	8,403	15,100			
Cases (n)	24	26			

Interpret the p-values. Have any of your previous conclusions stated in problem 16.3 changed?

Lesson 18: One-Way Analysis of Variance (ANOVA)

	Reading Assignment:	
	Additional Exercises:	

ANOVA is used to evaluate means from two or more subgroups. A statistically significant ANOVA indicates there is more variation between subgroups than would be expected by chance. It does not identify which subgroup pairs are significantly different from each other.

Problem 18.1 Hand Calculations

<p>Example: You obtained the number of years of education from one random sample of 38 police officers from City A, the number of years of education from a second random sample of 30 police officers from City B, and the number of years of education from a third random sample of 45 police officers from City C. The average years of education for the sample from City A is 15 years with a standard deviation of 2 years. The average years of education for the sample from City B is 14 years with a standard deviation of 2.5 years. The average years of education for the sample from City C is 16 years with a standard deviation of 1.2 years. Is there a statistically significant difference between the education levels of police officers in City A, City B, and City C?</p>	<p>Problem: The mean age of National Guard soldiers in three states was measured with a random sample in each state. The Oregon (OR) sample of 33 soldiers resulted in a mean age of 38 with a standard deviation of 5.3. The Washington (WA) sample of 25 soldiers resulted in a mean age of 42 with a standard deviation of 6.1. The Idaho (ID) sample of 30 soldiers resulted in a mean age of 29 with a standard deviation of 4.8. Is there a statistically significant difference between the ages of National Guard soldiers in Oregon, Washington, and Idaho? (Note: The sum of squares calculation assumes population variance and standard deviation were reported rather than sample variance and sample standard deviation.)</p>						
	City A	City B	City C		OR	WA	ID
Mean (years)	15	14	16	Mean (_____)			
Standard Deviation	2	2.5	1.2	Standard Deviation			
S ² (Variance)	4	6.25	1.44	S ² (Variance)			
N (number of cases)	38	30	45	N (number of cases)			
Sum of Squares (S ² *n)	152	187.5	64.8	Sum of Squares (S ² *n)			
Sum of Scores (mean*n)	570	420	720	Sum of Scores (mean*n)			
<p>State the Hypothesis</p> <p>Ho: There is no statistically significant difference among the three cities in the mean years of education for police officers.</p> <p>Ha: There is a statistically significant difference among the three cities in the mean years of education for police officers.</p>				<p>State the Hypothesis</p> <p>Ho:</p> <p>Ha:</p>			

<p>Set the Rejection Criteria</p> <p><i>Numerator Degrees of Freedom</i></p> <p>df=k-1 where k=3 (number of independent samples)</p> <p>df=2</p> <p><i>Denominator Degrees of Freedom</i></p> <p>df=n-k where n=113 (sum of n for all independent samples) df=110</p> <p><i>Establish Critical Value</i></p> <p>At alpha.05, df=(2,110)</p> <p>Consult f-distribution, Fcv = 3.08</p>	<p>Set the Rejection Criteria</p> <p><i>Numerator Degrees of Freedom</i></p> <p><i>Denominator Degrees of Freedom</i></p> <p><i>Establish Critical Value</i></p> <p>At alpha.05, df=(____ , ____)</p> <p>Consult f-distribution, Fcv = _____</p>
<p>Compute the Test Statistic</p> <p><i>Estimate Grand Mean</i></p> $\bar{X}_g = \frac{(\sum X_1 + \sum X_2 + \sum X_3)}{(n_1 + n_2 + n_3)}$ $\bar{X}_g = \frac{(570 + 420 + 720)}{(38 + 30 + 45)}$ $\bar{X}_g = \frac{1710}{113} \quad \bar{X}_g = 15.133$ <p><i>Estimate F Statistic</i></p> $F = \frac{\sum n_k (\bar{X}_i - \bar{X}_g)^2 / (K - 1)}{\left[\sum (X_i - \bar{X}_1)^2 + \sum (X_i - \bar{X}_2)^2 + \sum (X_i - \bar{X}_3)^2 \right] / (N - K)}$ $= \frac{38(15 - 15.133)^2 + 30(14 - 15.133)^2 + 45(16 - 15.133)^2 / (3 - 1)}{[152 + 187.5 + 64.8] / (113 - 3)}$ $= \frac{(.672 + 38.51 + 33.826) / 2}{3.676} \quad F = 9.931$	<p>Compute the Test Statistic</p> <p><i>Estimate Grand Mean</i></p> <p><i>Estimate F Statistic</i></p>
<p>Decide Results of Null Hypothesis</p> <p>Since the F-statistic (9.931) exceeds the F critical value (3.08), we reject the null hypothesis and conclude there is a statistically significant difference between the three cities in the mean years of education for police officers.</p>	<p>Decide Results of Null Hypothesis</p>

Problem 18.2 ANOVA with Raw Data

Open the Workbook data file in Data Grid and use the ANOVA option in the Statistical Procedures panel to complete the table and answer the question of whether there is a significant difference in income among those with less than high school, high school, or college education levels.

Round means and standard deviation to hundredths (".00")

Do not round the p-value or F statistic

Title:			
	Education Category		
	< High School	High School	College
Mean Income			
Std. Dev.			
n			
F-statistic		p-value	

Interpretation: _____

Problem 18.3 Using Summary Statistics in SumStats

Instructions for starting SumStats
Open the "SumStats" panel.
Select the ANOVA option.

Use SumStats to determine if there is a statistically significance difference among Census regions for individual income reported in the following summary table. (1993 data)


Comparison of mean individual incomes among four Census regions in the United States (1993).

	Census Regions (Std. Dev.)			
	<u>Northeast</u>	<u>Midwest</u>	<u>South</u>	<u>West</u>
Mean	20890 (21529)	16958 (18698)	13171 (17757)	18071 (20887)
n	126	210	233	149

F-statistic _____

p-value _____

Interpretation: _____

Lesson 19: Correlation		
	Reading Assignment:	
	Additional Exercises:	

Correlation is used to create a summary measure that reflects the covariation between two continuous variables. The Pearson Correlation Coefficient presented here can range from a -1.00 to 1.00. A positive coefficient indicates the values of variable A vary in the same direction as variable B. A negative coefficient indicates the values of variable A and variable B vary in opposite directions.


Problem 19.1 Hand Calculations

Example: The following data were collected to estimate the correlation between years of formal education and income at age 35.						Problem: The following data were collected from similar sized cities to estimate the correlation between fire deaths per 100,000 and fire department response times.					
Verify Conditions Data are paired interval observations. There is a linear relationship between Education (X) and Income (Y).						Verify Conditions Are Data paired interval observations? Yes / No Is there a linear relationship between response time (X) and fire deaths (Y)? Yes / No					
Compute Pearson's r						Compute Pearson's r					
	Educ Years	Income \$1000					Response Minutes	Death Rate			
Name	X	Y	XY	X²	Y²	City	X	Y	XY	X²	Y²
Susan	12	25	300	144	625	Elk	5	5			
Bill	14	27	378	196	729	LaPor	2	2			
Bob	16	32	512	256	1024	Mista	8	10			
Tracy	18	44	792	324	1936	Cini	4	2			
Joan	12	26	312	144	676	Ash	7	9			
$\Sigma =$	72	154	2294	1064	4990	$\Sigma =$					
n =	5					n =					
$r_{xy} = \frac{n \Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n \Sigma X^2 - (\Sigma X)^2] * [n \Sigma Y^2 - (\Sigma Y)^2]}}$ $r_{xy} = \frac{5(2294) - 72(154)}{\sqrt{[5(1064) - 72^2] * [5(4990) - 154^2]}}$ $r_{xy} = \frac{382}{\sqrt{167824}} \quad r_{xy} = .933$											

<p>Interpret</p> <p>There is a very high positive correlation between the variation of education and the variation of income. Individuals with higher levels of education earn more than those with comparably lower levels of education.</p>	<p>Interpret</p>
<p>Determine Coefficient of Determination</p> <p>$r^2 = .933^2 \quad r^2 = .871$</p> <p>Eighty-seven percent of the variance displayed in the income variable can be associated with the variance displayed in the education variable.</p>	<p>Determine Coefficient of Determination</p>

Problem 19.2 Evaluating Correlations with Raw Data

Open the Workbook data file in Data Grid and use the correlation procedure in the Statistical Procedures panel to describe the association, if any, between individual income and years of education. Is there a linear relationship?

Lesson 20: Hypothesis Testing for Pearson r		
	Reading Assignment:	
	Additional Exercises:	

Correlation coefficients summarize covariation within a sample. The following technique tests whether the coefficient (association) can be inferred to the underlying population. As with other significance tests, the sample data should originate from a random sample of the population.

Problem 20.1 Hand Calculations

<p>Example: Based on a Pearson r of .933 for annual income and education obtained from a national random sample of 20 employed adults.</p>	<p>Problem: Is a Pearson r of .953 for the correlation between fire deaths and fire department response times in five cities statistically significant?.</p>
<p>State the Hypothesis</p> <p>Ho: There is no association between annual income and education for employed adults.</p> <p>Ha: There is an association between annual income and education for employed adults.</p>	<p>State the Hypothesis</p> <p>Ho:</p> <p>Ha:</p>
<p>Set the Rejection Criteria</p> <p>df = 20-2 or df = 18</p> <p>tcv @ .05 alpha (2-tailed) = 2.101</p>	<p>Set the Rejection Criteria</p> <p>df = _____</p> <p>tcv @ .05 alpha (2-tailed) = _____</p>
<p>Compute Test Statistic</p> $t = r \sqrt{\frac{n-2}{1-r^2}}$ $t = .933 \sqrt{\frac{20-2}{1-.871}}$ $t = 11.022$	<p>Compute Test Statistic</p>
<p>Decision</p> <p>Since the test statistic 11.022 exceeds the critical value 2.101, there is a statistically significant association in the national population between an employed adult's education and their annual income.</p>	<p>Decision</p>

Problem 20.2 Evaluating Correlation Significance

Use the Workbook data file and Correlation procedure to complete the table and answer the question of whether there is a significant relationship between age and income. Between education in years and income.

Round Pearson R to hundredths (".00")

Do not round the p-value

Title:					
	Pearson R	Cases (n)	p < (2-tailed)	Significant?	
Income * Age				Yes	No
Income * Education				Yes	No

Describe the relationships evaluated in the above table and indicate what they mean for the population of agency employees.

Lesson 21: Simple Linear Regression

i	Reading Assignment:	
	Additional Exercises:	

Linear regression involves predicting the score for a dependent variable (Y) based on the score of an independent variable (X). Data are tabulated for two variables X and Y. The data are compared to determine a relationship between the variables with the use of Pearson's r. If a significant relationship exists between the variables, it is appropriate to use linear regression to base predictions of the Y variable on the relationship developed from the original data.

Problem 21.1 Hand Calculations

Example: The following data were collected to estimate the correlation between years of formal education and income at age 35 and are the same data used in an earlier example to estimate Pearson r.						Problem: The following data were collected from similar sized cities to estimate the correlation between fire deaths per 100,000 and average fire department response times.					
	Educ Years	Income \$1000					Response Minutes	Death Rate			
Name	X	Y	XY	X²	Y²	City	X	Y	XY	X²	Y²
Susan	12	25	300	144	625	Elk	5	5			
Bill	14	27	378	196	729	LaPor	2	2			
Bob	16	32	512	256	1024	Mista	8	10			
Tracy	18	44	792	324	1936	Cini	4	2			
Joan	12	26	312	144	676	Ash	7	9			
$\Sigma =$	72	154	2294	1064	4990	$\Sigma =$					
$n =$	5					$n =$					
$\bar{X} =$	14.4	30.8				$\bar{X} =$					

Problem 21.2 Simple Regression with Raw Data

Open the Workbook data file in Data Grid and use the Regression procedure in the Statistical Procedures panel to fill in the table below.

Round slope to hundredths (".00")

Do not round the p-value

Title:		
Comparison	Slope (b)	p < (2-tailed)
Income * Age		
Income * Education		

1. Evaluate the impact of Age on Income (slope and significance).

2. Evaluate the impact of Education on Income (slope and significance).

Lesson 22: Determine the Standard Error of the Estimate

i	Reading Assignment:	
	Additional Exercises:	

In problem 21.1, the value for the dependent variable was predicted based on a value of the independent variable. The following problem expands prediction from the sample to the entire underlying population. As with other estimation techniques, the predicted value is placed within an interval of possible population values.

Problem 22.1 Hand Calculations (Data from example 21.1)

Standard error of the estimate of Y $s_e = \sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n-2}}$

Name	Edu Years X	Income \$1000 Y	a	b	$\hat{Y} = a + bX$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
Susan	12	25	-9.650	2.809			
Bill	14	27	-9.650	2.809			
Bob	16	32	-9.650	2.809			
Tracy	18	44	-9.650	2.809			
Joan	12	26	-9.650	2.809			
	u = 5					$\Sigma =$	32.205
	Mean = 14.4	30.8					

$$s_e = \sqrt{\frac{32.205}{3}} \quad s_e = 3.276$$

Determine the Confidence interval for predicted Y

$$CI = \hat{Y} \pm t_{cv}(s_e)$$

where

Degrees of Freedom (df) = 5-2 or 3

Alpha .05, 2-tailed Based on the t-distribution (see table) $t_{cv} = 3.182$

Confidence interval for predicted Y

$$Y_c = 32.485 \pm 3.182(3.276) \quad Y_c = 32.485 \pm 10.424$$

Given a 5% chance of error, the estimated income for a person with 15 years of education will be \$32,485 plus or minus _____ or somewhere between _____ and _____.

Problem 22.2 Predicting Y with Raw Data

Enter the following data into the Data Grid, run Regression with Deaths as the dependent variable (Y) and answer the questions that follow. Assume these cities have similar characteristics (size, geography, age of housing stock, medical services, poverty and crime rates, etc.).

Round to hundredths (".00")

	Deaths	Fire Stations
	Per 10,000	Count
Cities (< 100,000 population)	Y	X
Elk	8	5
LaPor	2	6
Mista	1	10
Cini	4	5
Ash	4	7
Plant	1	8
Zana	9	2
Tempo	10	2
Bullford	6	4
Rocknell	2	9
Sylvana	6	7
Sleepy Hollow	10	4
Taylor	12	3
Nachos	10	1
Fredena	3	9

If Bullford closed one fire station, what is the predicted number of deaths? _____

Assume that these cities represent a random sample of cities from one state. If the state required all cities with less than 100,000 residents to have 10 fire stations, what is the predicted number of deaths in cities of similar size? _____

Appendix

Additional Review Questions

Requires Lesson 1

R1. Use the Intro and Methods section from the following condensed paper to identify the following:

- a) Research question
- b) Theory
- c) Two abstract concepts
- d) Two variables and level of measurement for each
- e) Identify two statistical controls used by the author to improve external validity

A Comparative Analysis of Public and Private Sector Entrant Quality

American Journal of Political Science, Vol. 39, No. 3, August 1995, pp. 628-639

Introduction:

Many commentators argue the federal government faces a quality crisis. Conventional wisdom suggests that poor pay, inadequate recruiting, and bureaucrat bashing have discouraged quality entrants from seeking employment with the public sector. However, public and private sector employees differ in ways that run counter to the prediction that poor monetary incentives or image battering will leave the public sector disadvantaged in hiring quality employees (Meier 1993). Comparative research between public and private sector employees has found that public managers have a higher need for achievement (Guyot 1962), service to society (Kilpatrick, Cummings, and Jennings 1964), serving the public interest (Rainey 1982; Perry and Wise 1990), and job security (Schuster 1974; Newstrom, Reif, and Monckza 1976; Bellante and Link 1981) than private sector employees. Adding to this list of differences is empirical evidence suggesting public employees value financial rewards less than their private sector counterparts (Kilpatrick, Cummings, and Jennings 1964; Schuster 1974; Rainey 1982, 1983).

Method:

The data used to compare public and private sector employee quality are provided by the National Longitudinal Survey of Youth (NLSY). If the federal government was unable to attract quality personnel during the 1980s, this representative sample of young labor force entrants should demonstrate that. Specifically, the NLSY will be used to test the premise that the aptitude of civil service employees hired during the 1980s is less than the aptitude of private sector employees hired during the same period.

Aptitude is measured by the Armed Forces Qualifications Test (AFQT), a cognitive abilities test battery. The assumption guiding this analysis is that aptitude, as measured by AFQT, is a valid measure of entrant quality, not unlike its purpose in the armed forces screening process. This measure is admittedly incomplete. As an example, the GAO has noted that three kinds of information are needed to assess employee quality: knowledge and ability, individual values and motivations, and match between individual and job (Volcker 1990, 139).

An ordinary least squares regression model is employed to answer the question of whether when controlling for sex, race, economic status, and occupation, federal sector employees have AFQT scores that are significantly different from those employed in the private sector. Sex (0=female, 1=male) and race (0=nonwhite, 1=white) are included in the model to control for their differing incidence of employment between the public and private sector (Blank 1985) and variance in AFQT scores (Herrnstein and Murray 1994). In addition, family income from 1978 is used as an estimate of economic status. This is added to the model to control for NLSY oversampling of economically disadvantaged whites. Occupation is confined to those who had occupations in accounting, engineering, computers, secretarial services, management, and guard/night watchman positions as specified by Bureau of Census occupation codes.

Requires Lesson 3

R2. Create row and column percents for the table below and answer the following questions.

Attitudes toward property tax increases and whether waste by city government officials is a problem

Property taxes	Government Waste			Total
	Serious	Moderate	No Problem	
Increase	15	10	20	
Keep same	40	35	10	
Decrease	45	10	5	
Total				

- What percentage of those who view government waste as a serious problem support a tax increase?
- What percentage of those who support a tax decrease also believe there is no problem in government waste?
- Assuming the table is representative of city voters, use percentages to support whether to ask or not ask for a tax increase.

R3. Read the synopsis of Table 2 and answer the questions that follow.

Data on promotion rates (averaged over the period 1988-1990) show that women and men were promoted at nearly the same rate at every grade level but with two important exceptions (Table 2). In GS-9 professional jobs, men were promoted at a rate 33 percent higher than women, and in GS-11 jobs, at a rate 40 percent higher than women. This is a significant finding. First, it means that the glass ceiling is probably not where conventional wisdom places it—at the level where people break into management jobs. It is, in fact, in the very early stages of a career. Three-quarters of employees in professional positions start at or below GS-11 and generally all of them must pass through those grades before they can even apply for a supervisory position. So while the promotion rate for men and women at higher grades is about the same, there is a small numerical base of women eligible for promotion. This partly accounts for the relatively slow progress of women in increasing their numbers in senior executive jobs.

Table2: Average Promotion Rates for Women and Men in Professional Occupations, 1988-1990.

Grade	Women	Men
GS 9	33%	44%
GS 11	15%	21%
GS 12	13%	10%
GS/GM 13	11%	8%
GS/GM 14	9%	7%
GS/GM 15	1%	1%

- Are the authors' conclusions correct? (why/why not)
- If we assume 100,000 women and 100,000 men both enter government service at the same time as a GS 9, how many men and women will be promoted to GS/GM 14? How many will be promoted to GS/GM 15?

Requires Lesson 5

R4. *Using Standardized Scores to Make Decisions:* Its your job as a program manager to conduct an annual review of the performance of ten external vendors working on your projects and make recommendations on whether their contract should be renewed for another year. You routinely collect data on the annual number of security violations but are aware that the security procedures and the strictness of enforcement can vary from year to year. The data for the past two years are provided below.

Vendor	# security violations 1998	# security violations 1999
A	5	7
B	4	5
C	10	12
D	2	4
E	1	6
F	2	4
G	3	4
H	2	5
I	12	10
J	7	9

Conduct an objective analysis of the vendors' performance. Be sure to include the following:

- Describe and compare each year's performance data using mean, median, mode, and standard deviation. Which year had more variation among the vendors?
- Use z-scores for vendor D to explain why a doubling of violations from 1998 to 1999 may not represent a serious increase.

Requires Lesson 9

R5. A random sample of 300 employed adults in Middletown found that 65% have health insurance. Based on these data, what is your estimate for the entire employed adult Middletown population? Complete the following table by calculating to three decimal places the standard error of the proportion for each cell using the relevant sample size (n) and proportion (p). Once done, identify three characteristics of the standard error of a proportion.

Proportion (p)	Sample Size (n)					
	50	300	500	1000	5000	10000
.90						
.80						
.70						
.60						
.50						
.40						
.30						
.20						
.10						

Requires Lesson 10

R6. Historical data indicate that about 20% of your agency's clients believe they were given poor service. Now under new management for six months, a random sample of 90 clients found that 15% believe they were given poor service. Has there been a significant change in service quality?

R7. A random sample survey was conducted of city managers. The following data were collected and compared to known population parameters for city managers (ICMA Baseline).

Characteristics of city managers			
Characteristics	Survey Results (percent)	ICMA Baseline ^a (percent)	Prob. ^b $P_s \neq P_u$
Graduate Degree	62	64	.46
Male	90	88	.28
White	97	97	.99
Population < 25,000	78	83	.03
	(n=314)	(n=7231)	

^a Source: ICMA. 1994 Municipal Year Book. Washington D.C.

^b Two-tailed test for the difference between a sample and population proportion.

- Verify statistical significance with SumStats and discuss the representativeness of the sample for the population of all city managers.
- For population <25,000, what is the minimum amount of change in the survey percentage necessary for $p > .05$ (i.e., not statistically significant)?

Requires Lesson 11

R8. In 1989, a national random sample of adults collected the following information on labor force characteristics for public and private sector employees. Table 1 represents only one of a series of comparisons that investigated the assumption by many that public and private sector employees are characteristically the same.

Table 1: Labor force characteristics of public and private sector employees
(standard deviation in parentheses)

	Pvt	Public	$p <^*$
Mean age (years)	37.6 (11.3)	40.9 (11.2)	.050
% male	50.9	46.8	.427
% with college	44.1	60.6	.002
Sample size	674	109	

* 2-tailed significance tests

- A report based on Table 1 claims the following: Public sector employees are on average older, proportionally higher educated and less likely to be men than their counterparts in the private sector. Is this true? Why?
- Outline a short paper that conducts a systematic exploration and statistical analysis of the relationship depicted in the box outline of Table 1.

R9 Critically review the following report and confirm the statistical calculations are correct for proportions (do not confirm calculations for Mean Age).

Introduction

The purpose of this study is to compare the characteristics of Agency X employees with those of the general labor pool. Agency X is faced with the problem of recruiting and retaining an employee workforce that is characteristically similar to the labor pool from which they draw their employees. The agency Inspector General hired an outside research firm to evaluate the level of disparity, if any, between agency employees and the labor pool. The following is a report of the findings.

Methods

In May 1999, a random sample survey of 1000 adults was conducted. The sample was drawn randomly from the local telephone directory where Agency X gets a large share of their employees. From this sample, 639 completed surveys were returned. A random sample survey was also conducted of 500 Agency X employees in professional and managerial positions. There were 193 completed surveys returned.

Results

The results of the two surveys are displayed in Table 1. There are statistically significant differences between Agency X and the labor pool in education, sex, and job position. Agency X employees are more educated (16 years v. 13 years) and more likely to be in management positions (40% v. 5%) than those in the local labor pool. In addition, females and minorities are under-represented in Agency X. As an example, 35% of Agency X employees are female compared to 52% in the labor pool.

Table 1: Characteristics of agency employees and the general labor pool

Characteristics	Agency X	Pool	P <
Mean years education	16	13	.001
Mean age	45	52	.035
% Female	35	50	.000
% Minority	20	22	.439
% Management	40	05	.000
n	193	639	

Conclusion

There is disparity between the characteristics of all Agency X employees and the characteristics of the labor pool. For all positions at Agency X, we recommend a concentrated effort be made to increase the proportion of females and minorities to reduce evidence of workforce unrepresentativeness.

Requires Lesson 14

R10. A random sample of 50 new households in Washington City revealed the mean number of personal vehicles was 2.2; standard deviation .7. Based on a 5% chance of error, estimate the range of personal vehicles new households will bring into Washington City.

R11. In 1997, a random sample of 600 employees of the state Department of Social Services were asked to complete survey questionnaires pertaining to organizational culture and attitudes toward work and family. There were 279 usable questionnaires returned. Table 1 represents only one of a series of comparisons that investigated the argument that working women have greater family responsibilities than working men.

Table 1: Level of satisfaction with the balance achieved between work and family life *

	Total	Sex	
		Male	Female
Mean satisfaction	275	302	256
Std Dev	135	105	100
Total cases (n)	279	126	149

Margin of error +/-

* The scale ranges from 100 (not satisfied) to 500 (very satisfied)

- a) Compute margin of error.
- b) Interpret the margin of error for female (i.e., What does it mean?).
- c) Based solely on the data presented in Table 1, is it safe to say that on average men who work in the department are more satisfied than women with the balance they achieve between work and family life?

Outline a short paper that conducts a systematic exploration and statistical analysis of the relationship depicted in the box outline of Table 1.

Requires Lesson 15

R12. The mean number of miles all employees travel between work and home in the metro area is 17.2. A random sample of 50 employees of Agency X revealed the following statistics: mean miles 14.5, standard deviation 5.5 miles. Is there a statistically significant difference between the number of miles traveled by Agency X employees and the population mean for the metro area?

Requires Lesson 16

R13. You obtained the number of public parks in a random sample of 17 similar size cities in State A and a random sample of 20 similar size cities in State B. The average number of parks in State A cities was 9 with a standard deviation of 3.2 parks. The average number of parks in State B cities was 11 with a standard deviation of 1.8 parks.

- a) Is there a statistically significant difference between the number of parks in State A and State B?
- b) Discuss what happens to the statistical significance when the standard deviation for State A is one park; when the standard deviation is four parks.
- c) Discuss what happens to the statistical significance when the sample sizes for States A and B are halved; when the sample sizes are doubled.

R14. In 1994, a national random sample of electrical engineers collected the following information on labor force characteristics for public and private sector engineers. Table 3 represents only one of a series of comparisons that investigated the assumption by many that public and private sector employees are characteristically the same.

Table 3: Mean annual income and seniority of electrical engineers employed in the public and private sector (standard deviation)

	Pvt	Public	p < *
Mean pay (U.S. \$)	70,094 (33,251)	60,858 (23,101)	.001
Mean seniority (years)	10.2 (8.5)	10.5 (9.1)	.728
Sample size	477	99	

* 2-tailed significance tests

- Interpret the results of the statistical test for the difference in mean years of seniority in Table 3. In other words, indicate what the statistical test indicates for the null hypothesis and describe in layman's terms what the results mean. (Hint: This is the step in hypothesis testing called "Decide Results")
- Interpret the results of the statistical test for the difference in mean annual income in Table 3. In other words, indicate what the statistical test indicates for the null hypothesis and describe in layman's terms what the results mean.

R15. Critically review the following report and verify the statistical test results.

You are responsible for the design and implementation of alternative means to reduce the number of chronically unemployed women. Your agency has funded a pilot study at an annual cost of \$50 million dollars. Your research department has just presented you with this report. Before passing the report on to your superiors, you must decide if it warrants further distribution. Since you were closely involved in the sample selection, you are confident the original two comparison samples were selected properly.

Introduction

The purpose of this study is to evaluate whether providing job training to unemployed females improves their ability to find and hold a job. A pilot program was conducted in the state of Florida with money from a federal grant to evaluate the effectiveness of job training before implementing this program nationally.

Methods

In July 1999, a random sample survey was conducted of 500 unemployed adults who participated in the job training program (293 responded). The sample was drawn randomly from the 4,500 who participated in the program. Another random sample survey was conducted of 500 unemployed adults (out of 14,500) who qualified for the job training program but did not participate (95 responded). A t-test for the difference between two independent means was used at alpha .05.

Results

The results of the two surveys are displayed in Table 1. With the exception of females with children under the age of ten, those with job training had weeks of employment that were significantly greater (at least $p < .05$) than those who did not receive training.

Table 1: Mean weeks employed during past 12 months *

Characteristics	Job Training	No Training	P <
All Females	22	18	.035
Female with child < 10 yrs age	24	20	.125
Female with no children	30	5	.001
n	95	293	

* For those participating in the program, the 12 months began after 3 months of training to be a clerical worker.

Conclusion

As shown in Table 1, those who participated in the job training program had greater employed weeks than those who were not provided training. The job training program was an obvious success and has broad implications for improving the employment rates for all chronically unemployed females. This pilot study suggests the training program can be implemented successfully nationally for all chronically unemployed females.

Requires Lesson 14 and 16

R16 In 1989, a national random sample of U.S. adults collected the following information on labor force characteristics for public and private sector employees. Table 2 represents only one of a series of comparisons that investigated the assumption by many that public and private sector employees are characteristically the same.

Table 2: Labor force characteristics of public and private sector employees (standard deviation in parentheses)

	Pvt	Public	p < *
Mean age (years)	37.6 (11.3)	40.9 (11.2)	.050
% male	50.9	46.8	.427
% with college	44.1	60.6	.002
Sample size	674	109	

* 2-tailed significance tests

- A report based on Table 2 claims the following: There is no difference between the average age of public and private sector employees. Is this true? Why?
- Given the descriptive statistics in Table 2, what is the mean age of all public sector employees in the United States?
- Outline a short paper that conducts a systematic exploration and statistical analysis of the relationship depicted in the box outline of Table 2.

Requires Lesson 18

R17. You obtained the number of public parks in a random sample of 15 similar size cities in State C. The average number of parks in State C cities was 7 (standard deviation of 5 parks).

- Using the data provided in question **R13** for States A and B, is there a statistically significant difference between the number of parks in State A, B, and C?
- For the above question, discuss what happens when the standard deviation for the sample from State C is doubled. What happens when you add 10 cities to each sample; when 10 cities are subtracted from each sample?

Requires Lesson 19 and 20

R18. The following data were collected to estimate the correlation between the number of restaurant violations per year and the number of annual inspections per restaurant mandated by city ordinance.

City	# Inspections Required (per restaurant)	Average # Violations (per restaurant)
A	5	2
B	3	3
C	1	4
D	2	3
E	6	0
F	4	2
G	4	1
H	5	1
I	2	3
J	1	4
K	3	2
L	0	8

- Estimate and interpret the Pearson's correlation coefficient.
- Determine if there is a statistically significant correlation.
- Prepare a one paragraph description for the Mayor of City M on the results of the study and your recommendation on whether City M should require inspections.

R19. Use the following table to write a short paper that explores the association between a person's education level (EDUC) and the number of brothers and sisters (SIBS) and the education levels of a person's spouse (SPEDUC), mother (MAEDUC), and father (PAEDUC). Be sure to include the following:

- State the hypotheses implicit in the table p values.
- Rank order and interpret the associations included in the table.
- Explore reasons to explain the associations you find in the table.

- - Correlation Coefficients - -

	EDUC	MAEDUC	PAEDUC	SIBS	SPEDUC
EDUC	1.0000 (1510) P= .	.4188 (1232) P= .000	.4634 (1065) P= .000	-.2640 (1501) P= .000	.6190 (789) P= .000
MAEDUC	.4188 (1232) P= .000	1.0000 (1233) P= .	.6723 (974) P= .000	-.2970 (1232) P= .000	.4274 (661) P= .000
PAEDUC	.4634 (1065) P= .000	.6723 (974) P= .000	1.0000 (1069) P= .	-.2751 (1066) P= .000	.4002 (581) P= .000
SIBS	-.2640 (1501) P= .000	-.2970 (1232) P= .000	-.2751 (1066) P= .000	1.0000 (1505) P= .	-.2234 (788) P= .000
SPEDUC	.6190 (789) P= .000	.4274 (661) P= .000	.4002 (581) P= .000	-.2234 (788) P= .000	1.0000 (790) P= .

(Coefficient / (Cases) / 2-tailed Significance)
 " . " is printed if a coefficient cannot be computed

Tables

Z Distribution Critical Values

Z	Area Beyond	Z	Area Beyond	Z	Area Beyond
0.150	0.440	1.300	0.097	2.450	0.007
0.200	0.421	1.350	0.089	2.500	0.006
0.250	0.401	1.400	0.081	2.550	0.005
0.300	0.382	1.450	0.074	2.580	0.005
0.350	0.363	1.500	0.067	2.650	0.004
0.400	0.345	1.550	0.061	2.700	0.004
0.450	0.326	1.600	0.055	2.750	0.003
0.500	0.309	1.650	0.050	2.800	0.003
0.550	0.291	1.700	0.045	2.850	0.002
0.600	0.274	1.750	0.040	2.900	0.002
0.650	0.258	1.800	0.036	2.950	0.002
0.700	0.242	1.850	0.032	3.000	0.001
0.750	0.227	1.900	0.029	3.050	0.001
0.800	0.212	1.960	0.025	3.100	0.001
0.850	0.198	2.000	0.023	3.150	0.001
0.900	0.184	2.050	0.020	3.200	0.001
0.950	0.171	2.100	0.018	3.250	0.001
1.000	0.159	2.150	0.016	3.300	0.001
1.050	0.147	2.200	0.014	3.350	0.000
1.100	0.136	2.250	0.012	3.400	0.000
1.150	0.125	2.300	0.011	3.450	0.000
1.200	0.115	2.350	0.009	3.500	0.000

This is only a portion of a much larger table.

The area beyond Z is the proportion of the distribution beyond the critical value (region of rejection). Example: If a test at .05 alpha is conducted, the area beyond Z for a two-tailed test is .025 (Z-value 1.96). For a one-tailed test the area beyond Z would be .05 (z-value 1.65).

T Distribution Critical Values

DF	Alpha						
	1-TAIL >	.10	.05	.025	.01	.005	.0005
	2-TAIL >	.20	.10	.05	.02	.01	.001
1		3.078	6.314	12.706	31.821	63.657	636.61
2		1.886	2.920	4.303	6.965	9.925	31.598
3		1.638	2.353	3.182	4.541	5.841	12.941
4		1.533	2.132	2.776	3.747	4.604	8.610
5		1.476	2.015	2.571	3.365	4.032	6.859
6		1.440	1.943	2.447	3.143	3.707	5.959
7		1.415	1.895	2.365	2.998	3.499	5.405
8		1.397	1.860	2.306	2.896	3.355	5.041
9		1.383	1.833	2.262	2.821	3.250	4.781
10		1.372	1.812	2.228	2.764	3.169	4.587
11		1.363	1.796	2.201	2.718	3.106	4.437
12		1.356	1.782	2.179	2.681	3.055	4.318
13		1.350	1.771	2.160	2.650	3.012	4.221
14		1.345	1.761	2.145	2.624	2.977	4.140
15		1.341	1.753	2.131	2.602	2.947	4.073
16		1.337	1.746	2.120	2.583	2.921	4.015
17		1.333	1.740	2.110	2.567	2.898	3.965
18		1.330	1.734	2.101	2.552	2.878	3.922
19		1.328	1.729	2.093	2.539	2.861	3.883
20		1.325	1.725	2.086	2.528	2.845	3.850
21		1.323	1.721	2.080	2.518	2.831	3.819
22		1.321	1.717	2.074	2.508	2.819	3.792
23		1.319	1.714	2.069	2.500	2.807	3.767
24		1.318	1.711	2.064	2.492	2.797	3.745
25		1.316	1.708	2.060	2.485	2.787	3.725
26		1.315	1.706	2.056	2.479	2.779	3.707
27		1.314	1.703	2.052	2.473	2.771	3.690
28		1.313	1.701	2.048	2.467	2.763	3.674
29		1.311	1.699	2.045	2.462	2.756	3.659
30		1.310	1.697	2.042	2.457	2.750	3.646
40		1.303	1.684	2.021	2.423	2.704	3.551
60		1.296	1.671	2.000	2.390	2.660	3.460
120		1.289	1.658	1.980	2.358	2.617	3.373
∞		1.282	1.645	1.960	2.326	2.576	3.291

This is only a portion of a much larger table

Chi-square Distribution Critical Values

df	Alpha				
	.10	.05	.02	.01	.001
1	2.706	3.841	5.412	6.635	10.827
2	4.605	5.991	7.824	9.210	13.815
3	6.251	7.815	9.837	11.345	16.266
4	7.779	9.488	11.688	13.277	18.467
5	9.236	11.070	13.388	15.086	20.515
6	10.645	12.592	15.033	16.812	22.457
7	12.017	14.067	16.622	18.475	24.322
8	13.362	15.507	18.168	20.090	26.125
9	14.684	16.919	19.679	21.666	27.877
10	15.987	18.307	21.161	23.209	29.588
11	17.275	19.675	22.618	24.725	31.264
12	18.549	21.026	24.054	26.217	32.909
13	19.812	22.362	25.472	27.688	34.528
14	21.064	23.685	26.873	29.141	36.123
15	22.307	24.996	28.259	30.578	37.697
16	23.542	26.296	29.633	32.000	39.252
17	24.769	27.587	30.995	33.409	40.790
18	25.989	28.869	32.346	34.805	42.312
19	27.204	30.144	33.687	36.191	43.820
20	28.412	31.410	35.020	37.566	45.315
21	29.615	32.671	36.343	38.932	46.797
22	30.813	33.924	37.656	40.289	48.268
23	32.007	35.172	38.968	41.638	49.728
24	33.196	36.415	40.270	42.980	51.179
25	34.382	37.652	41.566	44.314	52.620
26	35.563	38.885	42.856	45.642	54.052
27	36.741	40.113	44.140	46.963	55.476
28	37.916	41.337	45.419	48.278	56.893
29	39.087	42.557	46.693	49.588	58.302
30	40.256	43.773	47.962	50.892	59.703

This is only a portion of a much larger table.

F Distribution Critical Values

For Alpha .05

Denominator												
DF	Numerator DF											
	2	3	4	5	7	10	15	20	30	60	120	500
5	5.81	5.43	5.21	5.06	4.90	4.75	4.64	4.58	4.51	4.45	4.41	4.39
7	4.76	4.36	4.13	3.98	3.80	3.65	3.52	3.45	3.38	3.31	3.27	3.25
10	4.12	3.72	3.49	3.33	3.14	2.98	2.85	2.78	2.70	2.62	2.58	2.55
13	3.82	3.42	3.19	3.03	2.84	2.67	2.54	2.46	2.38	2.30	2.25	2.22
15	3.69	3.29	3.06	2.91	2.71	2.55	2.40	2.33	2.25	2.16	2.11	2.08
20	3.50	3.10	2.87	2.71	2.52	2.35	2.20	2.12	2.04	1.95	1.90	1.86
30	3.32	2.93	2.69	2.54	2.34	2.16	2.02	1.93	1.84	1.74	1.69	1.64
40	3.24	2.84	2.61	2.45	2.25	2.08	1.93	1.84	1.75	1.64	1.58	1.53
60	3.16	2.76	2.53	2.37	2.17	1.99	1.84	1.75	1.65	1.54	1.47	1.41
120	3.08	2.68	2.45	2.29	2.09	1.91	1.75	1.66	1.56	1.43	1.35	1.28
500	3.02	2.63	2.39	2.23	2.03	1.85	1.69	1.59	1.48	1.35	1.26	1.16

This is only a portion of a much larger table.

SPSS Instructions

The following instructions provide basic guidance on how to use SPSS and the SumStats.xls Excel file (provided by AcaStat) to complete the Workbook assignments. The instructions assume a basic understanding of SPSS.

Problem 2.1 How to Create a Data File

1. Open SPSS to begin creating a data file.
2. Variable names should be short (8 or less characters).
3. If you look at the data table, you will see seven variables: Idnum, Age, Edu, Sex, Manager, JobSat, Income. Each column in the Data Editor spreadsheet must be formatted to represent one of these variables.
4. To begin formatting columns, make sure the SPSS Data Editor is visible and you have selected a cell in the first column. Select the Data pull-down menu. Click the Define Variable option and replace "VAR00001" with the first variable name "Idnum" and click the "OK" button. Note: you may also double-click the column header to display the data format controls.
5. Select a cell in the second column and repeat the process for the variable name "Age". Continue until you have formatted the seven columns and then **save the SPSS data file** as "Workbook".
6. Click on a cell to begin entering data (start in the first column and first row). After completing one cell's entry, pressing an arrow key or the Enter key on your keyboard will move you to an adjoining cell. Double click on a cell to edit contents. Once you have completed data entry, **save the data file**.

Problem 2.2 Formatting the Data File

1. Make sure the "Workbook.sav" file is open and visible in the Data Editor before proceeding.
2. Select the "Idnum" column. Select the Data pull-down menu. Click Define Variable.
3. Click Labels and enter "Respondent Number" in the Variable Label textbox. Since there are no value labels or missing values for this variable, click "OK". Continue formatting variable labels for the remaining columns.
4. The variable "Sex" is the first variable that will need value labels. To format the values, enter "1" in the value textbox and then "Male" in the adjoining value label textbox. Click the add button. Repeat the process for Female (note: 2 = Female).
5. **Save the data file.**
6. Once the data have been formatted and saved, click the File pull-down menu and select Display Data Info and open the data file you just created ("Workbook"). An output window should appear displaying a data dictionary that should match the data dictionary shown in the Workbook.

Problem 2.3 Recoding Data

1. Click Transform, Recode, Into Different Variables. Select the variable you wish to use to create a recode. In this case, select "Edu" and click the arrow button. Click the Out Variable Name box and enter "EduCat". Click Change. Enter "Education Level" in the label box.
2. Click the Old and New Values button. Click the Range "Lowest Thru" option and enter "11" in the Value textbox. Click New Value, enter "1", and click add.
3. To continue the recode, click Old Value Value option and enter "12" in the textbox. Enter "2" for the new value and click add.
4. Click the Old and New Values button. Click the Range "Thru Highest" option and enter "13" in the Value textbox. Click New Value, enter "3", and click add.
5. Click Continue and Ok to create the new variable. To complete the process, format the new variable "EduCat" to the following: Variable label = "Education Level", value 1= "<12 yrs" 2= "HS Grad" 3= "College".
6. Save the data file as "Workbook" and then click Analyze/Reports/Case Summaries and run a listing of the variables Idnum, Edu, and EduCat. Review the listing to verify that the coding was successful. If not successful, delete the EduCat column and do the recode operation again.

Problem 3.3 Univariate Analysis of Raw Data

1. Open the Workbook data file in the Data Editor.
2. Use the Analyze/Descriptive Statistics/Frequencies procedure to answer the questions.

Problem 3.4 Bivariate Analysis of Raw Data

1. Open the Workbook data file in the Data Editor.
2. Use the Analyze/Descriptive Statistics/Crosstabs procedure to answer the questions.

Problem 4.2 Creating Summary Statistics with Raw Data

1. Open the Workbook data file in the Data Editor.
2. Use the Analyze/Descriptive Statistics/Descriptives procedure to answer the questions.

Problem 5.2 Confirming Hand Calculations with SumStats

1. Open the "SumStats" Excel file.
2. Select the Standardized z-scores option.

Problem 5.3 Creating Standardized Scores with Raw Data

1. Enter the data into Data Editor.
2. Format each column with the variable name indicated and save the data file as "zscore data".
3. Choose the Analyze/Descriptive Statistics/Descriptives procedure and select
4. "Save standardized values..".
5. Select History, Math, English, and Science variables from the listbox (put them in the analysis list) and run the Descriptives procedure. This should create four additional z-score variables (view Data Editor to confirm).

Problem 7.1 Comparing Random Samples to a Population

1. Download the AFQT SPSS data file from the following location:
2. <http://www.acastat.com/Pub/Data/AFQT.sav>
3. Open the AFQT data file and select Data/Select Cases in the Data Editor module.
4. Click Random Sample of Cases option and the sample button.
5. Enter a sample size (start with 15) and the total number of cases (1000) and click Continue.
6. Run Frequencies on INCLEVEL.

Note: SPSS does not have a procedure for repeated random samples.

Problem 9.2 Using Summary Statistics in SumStats

1. Open the "SumStats" Excel file.
2. Select the Margin of Error for Proportions option.

Problem 10.2 Creating Summary Statistics with Raw Data

1. Download the GSS93 SPSS data file from the following location:
2. <http://www.acastat.com/Pub/Data/GSS93.sav>
3. Open the 'GSS93.sav' data file in Data Editor and use Analyze/Descriptive Statistics/Frequencies on the variable "WrkStat".
4. Determine how many total responses are available and what proportion of the sample are retired (don't forget to convert percent to a proportion: e.g. 20% = .20). You should note that SPSS does not conduct z-tests on proportions, so the purpose of this exercise is to create summary statistics to use in problem 10.3.

Problem 10.3 Using Summary Statistics in SumStats

1. Open the “SumStats” Excel file.
2. Select the Z-Test option.
3. Use the One Sample Proportion section to complete the assignment.

Problem 11.2 Using Summary Statistics in SumStats

1. Open the “SumStats” Excel file.
2. Select the Z-Test option.
3. Use the Two Sample Proportion section to complete the assignment.

Problem 12.2 Calculating Chi-Square with Raw Data

1. Open the Workbook data file in Data Editor and use Analyze/Descriptive Statistics/Crosstabs to evaluate the association between sex, education level (EduCat), and job satisfaction (JobSat) and employee position (Manager).
2. The row variable is Manager. The column variables are Sex, EduCat, and JobSat.
3. Click the Statistics button and select chi-square and click the Continue button.
4. Click the Cells button and select count, row %, column %, and total %. Click the Continue button.

Problem 12.3 Calculating Chi-Square from Summary Statistics

1. Open the “SumStats” Excel file.
2. Select the Chi-Square option.
3. Convert the percentages into counts.
4. Enter the counts into SumStats starting at the top left white cell (represents Staff who said “Fair”).

Problem 13.2 Interpreting Multiple Comparisons

1. Open the GSS93 data file in Data Editor.
2. Select Analyze/Descriptive Statistics/Crosstabs and use the statistics button to select the Chi-square and Phi/Cramers V options to examine the relationship between PartyID and Sex, Race, and AgeCat4 (PartyID is the row variable).
3. Complete the table and indicate which comparisons are significant and then rank order (highest=1) the relationships by the strength of the relationship.

Problem 14.2 Creating Summary Statistics with Raw Data

1. Open Workbook data.
2. Select Analyze/Descriptive Statistics/Descriptives to calculate descriptive statistics for the variables age, edu, and income. Assume this is a random sample of U.S. Adults. Enter the results. They will be used for problem 14.3.

Problem 14.3 Creating Confidence Intervals with SumStats

1. Open the "SumStats" Excel file.
2. Select the Margin of Error for Means option.
3. Use the results from problem 14.2 to complete the table.
4. Enter the values for n, Mean, StdDev (Standard Deviation) for the age variable.
5. Record the results in the table (95% confidence interval), and repeat for each characteristic.

Problem 15.2 Creating Summary Statistics with Raw Data

1. Open the Workbook data file in Data Editor
2. Run Analyze/Descriptive Statistics/Descriptives on the variable "Income".
3. Determine the summary statistics needed to conduct a one-sample t-test. The purpose of this exercise is to create summary statistics to use in problem 15.3.

Problem 15.3 Calculating One-Sample t-tests with SumStats

1. Open the "SumStats" Excel file.
2. Select the T-Test option.
3. Use the one sample means t-test section to complete the assignment.

Problem 16.2 Calculating t-tests from Raw Data

1. Open the Workbook data file in Data Editor
2. Use Analyze/Compare Means/Independent Samples to answer the question of whether there is a significant difference between managers and non-managers in their years of education or income.
3. The continuous variable is "Edu" and the categorical variable is "Manager". The Group A value for Manager is 1 and the Group B value is 2 (note that 1=No and 2=Yes in the Manager variable).

Problem 16.3 Using Summary Statistics in SumStats

1. Open the “SumStats” Excel file.
2. Select the T-Test option.
3. Use the two sample means t-test section to complete the assignment.

Problem 18.2 ANOVA with Raw Data

1. Open the Workbook data file in Data Editor.
2. Select Analyze/Compare Means/One-Way ANOVA to complete the table and answer the question of whether there is a significant difference in income among those with less than high school, high school, or college education levels. Income is the dependent variable and education is the factor variable.

Problem 18.3 Using Summary Statistics in SumStats

1. Open the “SumStats” Excel file.
2. Select the ANOVA option.

Problem 19.2 Evaluating Correlations with Raw Data

1. Open the Workbook data file in Data Editor.
2. Use Analyze/Correlate/Bivariate to describe the association, if any, between individual income and years of education.
3. Use Graph/Scattergram/Simple to determine if there is a linear relationship.

Problem 20.2 Evaluating Correlation Significance

1. Open the Workbook data file.
2. Use Analyze/Correlate/Bivariate to complete the table and answer the question of whether there is a significant relationship between age and income. Between education in years and income.

Problem 21.2 Simple Regression with Raw Data

1. Open the Workbook data file in Data Editor
2. Use the Analyze/Regression/Linear procedure to fill in the table.

Problem 22.2 Predicting Y with Raw Data

1. Enter the data into the Data Editor.
2. Use the Analyze/Regression/Linear procedure with Deaths as the dependent variable (Y) to answer the questions.