

**University of California, Davis
Department of Statistics**

Statistics 100

Project

Fall 2017

Applied Statistical Analysis

The purpose of this project is to use statistical techniques including *descriptive statistics*, *comparison of two populations* and *analysis variance* in a real-life data analysis. You can use the **lead data** on the effects of lead exposure on human health, in particular its' sever adverse effects on children, *student survey data* or **any other dataset** in your own field of study.

The project will be done in two stages by **a group of maximum 6 students**. It would include objectives of the study, a short descriptive presentation of the data, results of the statistical analyses, final conclusions and discussion.

You can use **R, JMP, MINITAB** or any other statistical package to perform the statistical analysis.

The two parts of the project are as follow:

Part I:

Descriptive Analysis and Comparison of Two Populations: Due Monday, November 13

The first objective of Part I, is on *descriptive analysis* of data. Select any data from any source other than our textbook or use the lead data (only a few selected variables) for descriptive analysis. Give a detailed explanation of your data and report the source of the data. Prepare graphical representations and numerical summaries and explain their meanings.

The second objective of Part I, relates to the topics that are covered in chapters 6, 7 and 8 of the text. Use the same data that you used in your first project and compare two populations. You can have comparison of independent samples or paired observations. Give a short description of your data, construct confidence intervals, and perform hypothesis testing. Give a summary of your findings in a two-page report.

Part II:

Analysis of Variance: Due Friday, December 1

This part of the project is on applying the topics that you have learned in chapter 11 of the text to a real-life dataset. Again, you can select a new dataset or use the same data that you used in your previous project. Give a short description of your data and state null and alternative hypotheses. Perform the global F-test, write down the analysis of variance table and explain the results in detail. Provide more analysis of data by using multiple comparison techniques. Check the model assumptions and state your conclusions. Give a summary of your findings in a three-page report.

The Effect of Lead Exposure on Mental and Neurological Functions in Children

The effects of lead exposure on human health, in particular its severe adverse effects on children have been well documented in the literature. The following data were collected to study the effects of lead exposure on mental and neurological functions of children. The data is stored in the text file **lead.txt** and is available from the course website. The missing values are denoted by 'NA'. The data description is as follows:

VARIABLE DESCRIPTION:

- 1 IDENTIFICATION NUMBER
 - 2 AREA - RESIDENCE ON AUG'72
Area1=0-1 MILES FROM SMELTER
Area2=1-2.5 MILES
Area3=2.5-4.1 MILES
 - 3 AGE IN YEARS
 - 4 SEX MALE or FEMALE
- VARIABLES 5-18 CONTAIN IQ TEST RESULTS
- 5 INF - INFORMATION SUBTEST IN WISC AND WPPSI
 - 6 COMP - COMPREHENSION SUBTEST IN WISC AND WPPSI
 - 7 AR - ARITHMETIC SUBTEST IN WISC AND WPPSI
 - 8 DS - DIGIT SPAN SUBTEST(WISC) AND SENTENCE COMPLETION(WPPSI)
 - 9 V/RAW - RAW SCORE/VERBAL IQ (THIS IS TOTAL OF SCORES 5-8)
 - 10 PC - PICTURE COMPLETION SUBTEST IN WISC AND WPPSI
 - 11 BD - BLOCK DESIGN SUBTEST IN WISC AND WPPSI
 - 12 OA - OBJECT ASSEMBLY SUBTEST(WISC), ANIMAL HOUSE SUBTEST(WPPSI)
 - 13 COD - CODING SUBTEST(WISC), GEOMETRIC DESIGN SUBTEST(WPPSI)
 - 14 P/RAW - RAW SCORE/PERFORMANCE IQ (TOTAL OF SCORES IN 10-13)
 - 15 HH/INDEX - HOLLINGSHEAD INDEX OF SOCIAL STATUS
 - 16 IQV - VERBAL IQ
 - 17 IQP - PERFORMANCE IQ
 - 18 IQF - FULL SCALE IQ (NOT SUM OR AVERAGE OF IQV AND IQP)
- 19 TYPE OF IQ TEST WISC or WPPSI
(WISC USUALLY GIVEN TO CHILDREN GE 5 YRS 1 MONTH OF AGE
WPPSI USUALLY GIVEN TO CHILDREN LE 5YRS OF AGE)
 - 20 GROUP - BLOOD LEAD LEVEL GROUP
Group1=BLOOD LEAD LEVELS BELOW 40 MICROGRAMS/100ML IN BOTH 1972 & 1973
Group2=BLOOD LEAD LEVELS GREATER THAN OR EQUAL TO 40 MICROGRAMS/100ML IN BOTH 72 & 73 OR A LEVEL GREATER THAN OR EQUAL TO 40 IN 73 ALONE (3 CASES ONLY)
Group3=BLOOD LEAD LEVELS GREATER THAN OR EQUAL TO 40 MICROGRAMS/100M IN 72 AND LESS THAN 40 IN 73
 - 21 LD72 - BLOOD LEAD VALUES (MICROGRAMS/100ML) IN72 MISSING=NA
 - 22 LD73 - BLOOD LEAD VALUES (MICROGRAMS/100ML) IN 73
 - 23 FST2YRS - DID CHILD LIVE FOR 1ST 2 YRS WITHIN1 MILE OF SMELTER? YES or NO
 - 24 TOTYRS - TOTAL NUMBER OF YEARS SPENT WITHIN 4.1 MILES OF SMELTER

VARIABLES 25-29 CONTAIN SYMPTOM DATA (AS REPORTED BY PARENTS)

- 25 PICA YES or NO
- 26 COLIC YES or NO
- 27 CLUMSINESS YES or NO
- 28 IRRITABILITY YES or NO
- 29 CONVULSIONS YES or NO

VARIABLES 30-37 CONTAIN NEUROLOGICAL TEST DATA

- 30 # OF TAPS FOR RIGHT HAND IN THE 2-PLATE TAPPING TEST (#TAPS IN ONE 10 SECOND TRIAL) MISSING=NA
- 31 # OF TAPS FOR LEFT HAND IN THE 2-PLATE TAPPING TEST (#TAPS IN ONE 10 SECOND TRIAL) MISSING=NA
- 32 VISUAL REACTION TIME RIGHT HAND (MILLISECONDS) MISSING=NA
- 33 VISUAL REACTION TIME LEFT HAND (MILLISECONDS) MISSING=NA
- 34 AUDITORY REACTION TIME RIGHT HAND (MILLISECONDS) MISSING=NA
- 35 AUDITORY REACTION TIME LEFT HAND (MILLISECONDS) MISSING=NA
- 36 FINGER-WRIST TAPPING TEST RIGHT HAND (# TAPS IN ONE 10 SECOND TRIAL) MISSING=NA
- 37 FINGER-WRIST TAPPING TEST LEFT HAND (#TAPS IN ONE 10 SECOND TRIAL) MISSING=NA
- 38 WWPS - WERRY-WEISS-PETERS SCALE FOR HYPERACTIVITY
None, Mild, Moderate, High

The Student Survey Data

Description: Students at a University in Washington State who were enrolled in a statistics course were surveyed in the first week of class. Their answers to a number of questions were recorded.

The Data: The data can be found through a link on the course webpage (not smartsite). The name of the data set is student.csv, it has 812 rows, and the following columns:

Column 1: semester: 9 indicates Fall 2007, 10 indicates Spring 2008, 11 indicates Fall 2008, 12 indicates Spring 2009.

Column 2: gender: M or F, denoting Male or Female.

Column 3: collegeYear: How many years the student has been attending college.

Column 4: height: The students' height in inches.

Column 5: weight: The students' weight in pounds.

Column 6: hsGPA: The students' high school GPA.

Column 7: collegeGPA: The students' current college GPA.

Column 8: financialAid: Y indicates the student has financial aid, N indicates they do not.

Column 9: tvHours: The average number of hours of TV the student watches per week.