**Introductory SAS Programming**

**Problem 1:**

For this problem you will create a data set, generate descriptive statistics, create histograms and a scatterplot, and perform a paired t-test.

Data Set:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | CLASS | GENDER | EXAM1 | EXAM2 |
| 101 | 1 | M | 80 | 80 |
| 102 | 2 | M | 90 | 97 |
| 103 | 1 | M | 66 | 44 |
| 104 | 2 | M | 56 | 59 |
| 105 | 1 | M | 100 | 75 |
| 106 | 2 | F | 99 | 68 |
| 107 | 1 | F | 87 | 78 |
| 108 | 2 | F | 78 | 77 |
| 109 | 1 | F | 90 | 72 |
| 110 | 2 | F | 83 | 67 |

Variable labels:

Class: class in school

Exam1: score on first exam

Exam2: score on second exam

Value labels:

CLASS 1= undergraduate

2 = graduate

1) *Create a data file, read in the data, and print the data. Note that a $ must follow a variable name that contains character data.*

**data** exams;

input id class gender $ exam1 exam2;

cards;

101 1 M 80 80

102 2 M 90 97

103 1 M 66 44

104 2 M 56 59

105 1 M 100 75

106 2 F 99 68

107 1 F 87 78

108 2 F 78 77

109 1 F 90 72

110 2 F 83 67

;

/\* Print data \*/

**proc** **print** data=exams;

title "Problem 1: Exam Scores";

**run**;

3) *Print only graduate students that received lower than an 80% on either exam.*

**proc** **print** data=exams;

title2 "Graduate Students Earning Less than an 80% On Either Exam";

where class=**2** AND (exam1<**80** OR exam2<**80**);

**run**;

4) *Get column totals for exam1 and exam2 by class. Note that the data must be sorted by class first. When sorting data, SAS overwrites the original data set, so it can be a good idea to save the sorted data to a new data file.*

**proc** **sort** data=exams out=sortedexams;

by class;

**run**;

**proc** **print** data=sortedexams;

title2 "Total Exam Scores by Class";

by class;

sum exam1 exam2;

**run**;

5) *Generate descriptive statistics for the scores on exams 1 and 2.*

**proc** **means** data=exams maxdec=**2**;

title2 "Descriptive Statistics for Exam Scores";

var exam1 exam2;

**run**;

6) *Create histograms and get additional descriptive statistics.*

**proc** **univariate** data=exams;

var exam1 exam2;

histogram;

**run**;

7) *Create a scatterplot of exam 1 scores vs exam 2 scores.*

**proc** **gplot** data=exams;

title2 "Plot of Exam 1 Scores vs Exam 2 Scores";

plot exam2\*exam1;

symbol1 value=dot;

**run**;

**quit**;

8) *Find the correlation between Exam1 and Exam 2.*

**proc** **corr** data=exams;

title2 "Correlation for Exam Scores";

var exam1 exam2;

**run**;

9) *Test for a significant difference in mean scores between exams with a paired t-test.*

**proc** **ttest** data=exams;

title2 "Paired t-test for Exam Scores";

paired exam1\*exam2;

**run**;

**quit**;

**Problem 2:**

For this problem you will read in an existing data set and perform a linear regression. The data set contains two numeric variables, cricket chirps per second and temperature in degrees Fahrenheit.

1) *Create a data file and read in the data.*

**data** crickets;

infile 'D:\Problem 2.txt';

input chirps temperature;

**run**;

2) *Check to make sure data were read in properly.*

**proc** **print** data=crickets;

title "Problem 2: Cricket Chirps and Temperature";

**run**;

3) *Perform a simple linear regression with chirps as the predictor and temperature as the response.*

**proc** **reg** data=crickets plots=diagnostics;

title2 "Regression Analysis for Cricket Chirps/Minute and Temperature";

model temperature = chirps / cli clm;

**run**;

**quit**;

**Problem 3:**

For this problem you will use two do loops to enter categorical data, create a format to label the variables, run a chi-square test of independence, and create a bar chart. The data are occurrences of eye color and hair color.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Eye Color | | | |
| Hair Color | Brown | Blue | Hazel | Green |
| Black | 68 | 29 | 15 | 5 |
| Brown | 119 | 84 | 54 | 29 |
| Red | 26 | 17 | 14 | 14 |
| Blonde | 7 | 94 | 10 | 16 |

1) *Create a data file and enter the data with hair color being the rows and eye color being the columns. You will need a third variable to represent the number of occurrences in each cell (in this case “count”). The “@@” tells SAS to read multiple observations for a variable on one line. Then, as always, print the data to make sure it was correctly entered.*

**data** color;

do row=**1** to **4**;

do col=**1** to **4**;

input count @@;

output;

end;

end;

cards;

68 29 15 5

119 84 54 29

26 17 14 14

7 94 10 16

;

**proc** **print** data=color;

title "Problem 3: Hair and Eye Color Data";

**run**;

2) *Create a table.*

**proc** **freq** data=color;

weight count;

table row\*col;

**run**;

**quit**;

3) *Create a format to label the “row” and “col” variables a format to apply to the numeric variables for hair and eye color so that the rows and columns in your results are clearly identified. It’s a good idea to create the format at the beginning of your code. It’s also usually a good idea to assign the format in the data step. Assigning it in a specific procedure only makes it active for that procedure. Note that you have to rerun the entire data step.*

**proc** **format**;

value rowfmt **1**="Black"

**2**="Brown"

**3**="Red"

**4**="Blonde";

value colfmt **1**="Brown"

**2**="Blue"

**3**="Hazel"

**4**="Green";

**data** color;

do row=**1** to **4**;

do col=**1** to **4**;

input count @@;

output;

end;

end;

label row="Hair Color" col="Eye Color";

format row rowfmt. col colfmt.;

cards;

68 29 15 5

119 84 54 29

26 17 14 14

7 94 10 16

;

4) *Get rid of row and column percentages and perform a chi-square test of independence.*

**proc** **freq** data=color;

title2 "Chi-square Test for Hair and Eye Color";

weight count;

table row\*col / chisq norow nocol;

**run**;

**quit**;

5) *Prepare the SAS environment for the bar chart. Note that anything enclosed in* /\* \*/ *is a comment about the code and is not read by SAS.*

/\*Define the axis characteristics \*/

axis1 value=none label=none;

axis2 label=(angle=**90** "Frequency");

axis3 label=none;

/\* Define the legend options \*/

legend1 frame;

6) *Create a bar chart*

**proc** **gchart** data=color;

title2 "Frequenct of Eye Color by Hair Color";

vbar col / subgroup=col group=row sumvar=count

legend=legend1 gspace=**6** space=**0** discrete

maxis=axis1 raxis=axis2 gaxis=axis3;

pattern1 color=stbr;

pattern2 color=blue;

pattern3 color=libg;

pattern4 color=stolg;

**run**;

**quit**;

7) *Make the chart three dimensional.*

**proc** **gchart** data=color;

title2 "Frequenct of Eye Color by Hair Color";

vbar3d col / subgroup=col group=row sumvar=count

legend=legend1 gspace=**6** space=**0** discrete

maxis=axis1 raxis=axis2 gaxis=axis3;

pattern1 color=stbr;

pattern2 color=blue;

pattern3 color=libg;

pattern4 color=stolg;

**run**;

**quit**;

**Problem 4:**

For this problem you will read data in from an Excel file, create a permanent SAS data file, use the “by” statement to find means for different levels of a variable, perform a one-way ANOVA and Tukey’s Studentized Range Test, and then write the output to a document file. These data contain two variables that represent the alcohol content of wine from three different regions.

1. *Create a location for the permanent SAS data file and read in the data from the Excel file.*

*\*\*\*IMPORTANT: Make sure the Excel file is closed. This will not work if it is open.\*\*\**

Libname wine 'D:';

**PROC** **IMPORT** OUT=wine.content DATAFILE= "D:\Problem 4.xlsx"

DBMS=xlsx REPLACE;

SHEET="Alcohol";

GETNAMES=YES;

**run**;

**proc** **print** data=wine.content;

**run**;

1. *Find the mean alcohol content of the wines by region. To do this the data must first be sorted by region.*

**proc** **sort** data=wine.content out=wine.sortedcontent;

by Region;

**run**;

**proc** **means** data=wine.sortedcontent maxdec=**2**;

by Region;

var Alcohol\_Content;

**run**;

1. *Perform a one-way ANOVA to test for a mean difference in alcohol content in at least one of the regions and perform Tukey’s Studentized Range Test to see where the difference(s) may lie.*

**proc** **glm** data=wine.content plots=diagnostics;

class Region;

model Alcohol\_Content=Region;

lsmeans Region / adjust=tukey;

**run**;

**quit**;

1. *Write the results to a Word document. Note that if you change the code and re-run the analysis the Word document won’t be updated if it is still open. If the Word document is closed it will automatically be overwritten.*

Libname wine 'D:';

**PROC** **IMPORT** OUT=wine.content DATAFILE= "D:\Problem 4.xlsx"

DBMS=xlsx REPLACE;

SHEET="Alcohol";

GETNAMES=YES;

**run**;

**proc** **print** data=wine.content;

**run**;

/\* Enter this where you want the Word document to start\*/

ods rtf file="D:\Problem 4 Output.doc";

**proc** **sort** data=wine.content out=wine.sortedcontent;

by Region;

**run**;

**proc** **means** data=wine.sortedcontent maxdec=**2**;

by Region;

var Alcohol\_Content;

**run**;

**proc** **glm** data=wine.content plots=diagnostics;

class Region;

model Alcohol\_Content=Region;

lsmeans Region / adjust=tukey;

**run**;

**quit**;

/\* Enter this where you want the Word document to end \*/

ods rtf close;