Ben Goodwin

DS6371 HW6

1. Bonferroni method to construct simultaneous confidence intervals
   1. Problem statement: How compelling is the evidence that there are differences in attitude toward the mobility types of handicaps.
   2. First a few plots to explore the possibility of differences:

Chart, box and whisker chart

Description automatically generated

Diagram, histogram

Description automatically generated

From these plots, there is some visual evidence to support evidence of a difference in some of the groups. We can investigate further.

Lets do a hypothesis test (ANOVA):

Step 1: H0: Mu\_2=Mu\_3=Mu\_5

Step 2: HA: At least one Mu\_i differs

Step 3: F-stat: 2.86

Step 4: P-value: 0.03

Step 5: Decision, Reject H\_0!

Step 6: Conclusion, we have evidence to suggest at least one of the mu\_i’s differs.

Step 7: Scope: The scope in this problem is limited as there are drawbacks to the researcher’s experimental design. I think this study provides a good starting point for further investigation into this problem.

Bonferroni method:

R-input:

Graphical user interface, text, application

Description automatically generated

R-output:

A close up of text on a white background

Description automatically generated

Text, table

Description automatically generated

I will use K=3, for 3 comparisons. Yielding a multiplier of 2.46. Common Standard error of 0.617 for the comparison pairs. DF=65, I=5, s\_p=1.633

X\_bar\_2: 4.429, X\_bar\_3:5.921,X\_bar\_5=5.343

Bonferroni: 1.517

95% CI’s

Mu\_2-Mu\_3: [-3.009,0.025]

Mu\_2-Mu\_5: [-2.431,0.603]

Mu\_3-Mu\_5: [-0.939,2.095]

In conclusion, from the confidence intervals above zero is included and we can conclude there is insufficient evidence to conclude any of the means for handicaps are different at the familywise rate of 0.05.

Additionally, it would seem that all of the assumptions are met for this data.

Normality of data, homogeneity of SD. The QQ-plot looks roughly normal, standard deviations seem to violate the assumption, but not severely so, and We will assume the data are independent, both between and within groups, and proceed with the ANOVA.

We can verify these below from the following plots:

A picture containing calendar

Description automatically generated

Chart

Description automatically generated

A picture containing chart

Description automatically generated

Question 2:

Problem statement: We are concerned with the comparison procedures available within one-way ANOVA. We would like to locate the half-widths of the 95% confidence intervals and the values of each.

Assumptions for ANOVA: Assumptions: The Assumptions of the ANOVA are: the incomes in each educational group come from a normal distribution, the variances of these normal distributions are equal, the data are independent within each group, and the data are independent between each group.

A picture containing chart

Description automatically generated

Chart

Description automatically generated

A picture containing calendar

Description automatically generated

From the plots above, it would appear the assumptions are met. The and QQ plots appear to be normally distributed. However, each group has a sample size greater than 30, thus allowing the CLT to enable the ANOVA to be robust to this assumption.

R input:

Text

Description automatically generated

R output:

Graphical user interface, text, application

Description automatically generated

Each of the half-widths can be directly found in the output:

The Bonferroni can be found by accessing the statistics variable at the sixth position

Output = 1.79, table =1.974, they match

The Tukey can be found by accessing the statistics variable at the fifth position

Output = 1.73, table =1.735, they match

The LSD can be found by accessing the statistics variable at the sixth position

Output = 1.23, table = 1.233

The Scheffe can be found by accessing the statistics variable at the seventh position

Output =1.96, table =1.957

The Dunnett half width is less easy to (for me) to determine from output, however the Dunnett Multiplier is 2.50316 and that is multiplied by .61718 to obtain 1.545

Output =1.545, table=1.545 (for comparisons with control only)

SAS input:

data handicapdata;

input score handicap $;

datalines;

1.9 None

2.5 None

3 None

3.6 None

4.1 None

4.2 None

4.9 None

5.1 None

5.4 None

5.9 None

6.1 None

6.7 None

7.4 None

7.8 None

1.9 Amputee

2.5 Amputee

2.6 Amputee

3.2 Amputee

3.6 Amputee

3.8 Amputee

4 Amputee

4.6 Amputee

5.3 Amputee

5.5 Amputee

5.8 Amputee

5.9 Amputee

6.1 Amputee

7.2 Amputee

3.7 Crutches

4 Crutches

4.3 Crutches

4.3 Crutches

5.1 Crutches

5.8 Crutches

6 Crutches

6.2 Crutches

6.3 Crutches

6.4 Crutches

7.4 Crutches

7.4 Crutches

7.5 Crutches

8.5 Crutches

1.4 Hearing

2.1 Hearing

2.4 Hearing

2.9 Hearing

3.4 Hearing

3.7 Hearing

3.9 Hearing

4.2 Hearing

4.3 Hearing

4.7 Hearing

5.5 Hearing

5.8 Hearing

5.9 Hearing

6.5 Hearing

1.7 Wheelchair

2.8 Wheelchair

3.5 Wheelchair

4.7 Wheelchair

4.8 Wheelchair

5 Wheelchair

5.3 Wheelchair

6.1 Wheelchair

6.1 Wheelchair

6.2 Wheelchair

6.4 Wheelchair

7.2 Wheelchair

7.4 Wheelchair

7.6 Wheelchair

;

run;

proc glm data=handicapdata;

class handicap;

model score=handicap;

means handicap/dunnett("None");

means handicap/tukey;

means handicap/bon;

means handicap/scheffe;

means handicap/t;

run;

SAS output:

Graphical user interface, chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedGraphical user interface, application

Description automatically generatedChart, box and whisker chart

Description automatically generatedGraphical user interface, application

Description automatically generatedChart, box and whisker chart

Description automatically generatedGraphical user interface, application

Description automatically generatedChart, box and whisker chart

Description automatically generatedTable

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedGraphical user interface, application

Description automatically generated

Each of the half-widths can be directly found in the output:

The Bonferroni can be found by viewing the output

Output = 1.79, table =1.974, they match

Bonferroni: Multiplier = 2.90602, 95% CI half-width =

The Tukey can be found by viewing the output

Output = 1.73, table =1.735, they match

Tukey: Multiplier = , 95% CI half-width =

The LSD can be found by viewing the output

Output = 1.23, table = 1.233

LSD (t): Multiplier = 1.99714, 95% CI half-width =

The Scheffe can be found by viewing the output

Output =1.96, table =1.957

Scheffe: Multiplier = , 95% CI half-width =

The Dunnett half width can be found by viewing the output

Output=1.5449

Dunnett: Multiplier = 2.50316, 95% CI half-width =

Additionally we can once again view the hypothesis test for ANOVA on this data:

Lets do a hypothesis test (ANOVA):

Step 1: H0: Mu\_2=Mu\_3=Mu\_5

Step 2: HA: At least one Mu\_i differs

Step 3: F-stat: 2.86

Step 4: P-value: 0.03

Step 5: Decision, Reject H\_0!

Step 6: Conclusion, we have evidence to suggest at least one of the mu\_i’s differs.

Step 7: Scope: The scope in this problem is limited as there are drawbacks to the researcher’s experimental design. I think this study provides a good starting point for further investigation into this problem.

Question 3

Step 1: Discussion of assumptions:

Problem statement:

How strong is the evidence that at least one of the five population distributions of education level has a different mean income than any of the others?

Assumptions: The Assumptions of the ANOVA are: the incomes in each educational group come from a normal distribution, the variances of these normal distributions are equal, the data are independent within each group, and the data are independent between each group. From the previous homework’s, I have determined that log transformed data better meets model assumptions. The plots of the assumptions are below:

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generated

Chart, line chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

The and QQ plots of log transformed data appears to be slightly less skewed (in the other direction), but only slightly. However, each group has a sample size greater than 130, thus allowing the CLT to enable the ANOVA to be robust to this assumption.

Additionally, it would seem that all of the assumptions are met for this data.

Normality of data, homogeneity of SD. The QQ-plot looks roughly normal, standard deviations seem to violate the assumption, but not severely so, and We will assume the data are independent, both between and within groups, and proceed with the ANOVA.

Hypothesis test (ANOVA on Log transformed data)

Step 1:

H\_0:All median incomes are the same across education levels.

H\_A: At least one pair of income medians are different between education levels

Step 2: We can skip identification of p-value in ANOVA steps.

Step 3: P-value: P<0.0001

Step 4: F-value: 62.87

Step 5: Decision, reject H\_0.

Step 6: Conclusion: There is strong evidence to suggest that at least one of the median incomes for a education level is different from the others.

Step 7: Scope: This is an observational study, and thus, we cannot assign causal inference to this relationship. The NLSY is a random sample of households and, thus, is a random sample but not a simple random sample of subjects in the desired population. Inference can be generalized to the population of areas sampled in the United States, although one should be wary of the standard deviations and standard errors estimated here. Cluster sampling of households was employed, which introduces dependency/correlation at the cluster (household) level.

Step 2: Selection and execution of tests.

Tukey-Kramer procedure:

R-input:

A picture containing graphical user interface

Description automatically generated

R-output:

Chart, box and whisker chart

Description automatically generated

A picture containing table

Description automatically generatedText, table

Description automatically generated

SAS-input:

Graphical user interface, text, application

Description automatically generated

SAS-output:

Table

Description automatically generated

Based on the output from above, we have each of the group comparisons, which are below

<12-12 -0.328 -0.5597 -0.0961 0.001

>16-12 0.671 0.5172 0.8242 0.000

13-15-12 0.164 0.0364 0.2916 0.004

16-12 0.570 0.4209 0.7189 0.000

>16-<12 0.999 0.7443 1.2529 0.000

13-15-<12 0.492 0.2523 0.7314 0.000

16-<12 0.898 0.6461 1.1493 0.000

13-15->16 -0.507 -0.6716 -0.3418 0.000

16->16 -0.101 -0.2828 0.0812 0.555

16-13-15 0.406 0.2451 0.5666 0.000

We have a difference in log transformed median incomes between <12 and 12 by -0.328, with a 95% CI falling between -0.5597 and -0.0961. With a P-Value of 0.001, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS) Since the data is log transformed we have a median difference of -0.328 between the two levels of education.

We have a difference in log transformed median incomes between >16 and 12 by 0.671, with a 95% CI falling between -0.5172 and 0.8242. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS)

Since the data is log transformed we have a median difference of 0.671 between the two levels of education.

We have a difference in log transformed median incomes between 13-15 and 12 by 0.164, with a 95% CI falling between 0.0364 and 0.2916. With a P-Value of 0.004, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS)

Since the data is log transformed we have a median difference of 0.164 between the two levels of education.

We have a difference in log transformed median incomes between 16 and 12 by 0.57, with a 95% CI falling between 0.4209 and 0.7189. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS) Since the data is log transformed we have a median difference of 0.57 between the two levels of education.

We have a difference in log transformed median incomes between >16 and <12 by 0.999, with a 95% CI falling between 0.7443 and 1.2529. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS)

Since the data is log transformed we have a median difference of 0.999 between the two levels of education.

We have a difference in log transformed median incomes between 13-15 and <12 by 0.492, with a 95% CI falling between 0.2523 and 0.7314. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS)

Since the data is log transformed we have a median difference of 0.492 between the two levels of education.

We have a difference in log transformed median incomes between 16 and <12 by 0.898, with a 95% CI falling between 0.6461 and 1.1493. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS)

Since the data is log transformed we have a median difference of .898 between the two levels of education.

We have a difference in log transformed median incomes between 13-15 and >16 by -0.507, with a 95% CI falling between -0.6716 and -0.3418. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS) Since the data is log transformed we have a median difference of -0.507 between the two levels of education.

We have a difference in log transformed median incomes between 16 and >16 by -0.101, with a 95% CI falling between -0.2828 and -0.0812. With a P-Value of 0.555, thus a non-significant result indicating a difference in log transformed median income. (This is verified in R and SAS)

Since the data is log transformed we have a median difference of -0.101 between the two levels of education.

We have a difference in log transformed median incomes between 16 and 13-15 by 0.406, with a 95% CI falling between 0.2451 and 0.5666. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS) Since the data is log transformed we have a median difference of 0.406 between the two levels of education.

Dunnett’s Procedure:

R-input:

R-output:

SAS input:

Graphical user interface, text, application

Description automatically generated

SAS output:

Table

Description automatically generated

We have a difference in log transformed median incomes between >16 and 12 by 0.67, with a 95% CI falling between 0.53 and 0.81. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS) Since the data is log transformed we have a median difference of 0.67 between the greater than 16 education and 12.

We have a difference in log transformed median incomes between 16 and 12 by 0.57, with a 95% CI falling between 0.434 and 0.71. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS) Since the data is log transformed we have a median difference of 0.57 between 16 education and 12.

We have a difference in log transformed median incomes between 13-15 and 12 by 0.164, with a 95% CI falling between 0.048 and 0.28. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS) Since the data is log transformed we have a median difference of 0.164 between the greater than 16 education and 12.

We have a difference in log transformed median incomes between <12 and 12 by -0.33, with a 95% CI falling between -0.54 and -0.117. With a P-Value of 0.000, thus a significant result indicating a difference in log transformed median income. (This is verified in R and SAS) Since the data is log transformed we have a median difference of -0.33 between the greater than <12 education and 12.

Interpretation and conclusion:

From the above tests, we can see that there are difference in the median income between the groups above. In short we can detect differences in the educational level group comparisons in each comparison except between >16 and 16. The data is log transformed so we are to interpret conclusions on a log scale.