Andrew Abbott

MSDS 6371

Unit 5 Live Session HW

* How strong is the evidence that at least one of the five population distributions is different from the others?
  + At least two of the means are not equal.
* The data contains annual incomes in 2005 of a random sample of 2,584 Americans who were selected for the national Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005.
  + Data created by SAS code following has 4 variables.
    - Subject – Subject ID.
    - Educ – The years of education received groups.
    - Income2005 – Theannual income in 2005.
    - Logincome – the log of Income2005.
* Using histograms and qqplots to check for normality and variance difference I see that all of the groups are positively skewed with some variances differing significantly. I decide to use a log transformation.
* Descriptive statistics for each group before and after transformation.

| **Analysis Variable : Income2005** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Educ** | **N Obs** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| 12 | 1020 | 1020 | 36864.90 | 29369.73 | 300.0000000 | 410008.00 |
| 13-15 | 648 | 648 | 44875.96 | 33913.54 | 429.0000000 | 257286.00 |
| 16 | 406 | 406 | 69996.97 | 64256.80 | 200.0000000 | 519340.00 |
| <12 | 136 | 136 | 28301.45 | 21021.90 | 350.0000000 | 100000.00 |
| >16 | 374 | 374 | 76855.46 | 65428.29 | 63.0000000 | 703637.00 |

| **Analysis Variable : logincome** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Educ** | **N Obs** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| 12 | 1020 | 1020 | 10.2272149 | 0.8539854 | 5.7037825 | 12.9239320 |
| 13-15 | 648 | 648 | 10.3912107 | 0.9288173 | 6.0614569 | 12.4579436 |
| 16 | 406 | 406 | 10.7970859 | 0.9581051 | 5.2983174 | 13.1603141 |
| <12 | 136 | 136 | 9.8993404 | 0.9988809 | 5.8579332 | 11.5129255 |
| >16 | 374 | 374 | 10.8979022 | 1.0665910 | 4.1431347 | 13.4640179 |

* Since I am testing the difference across more than 2 groups I will use ANOVA.
* ANOVA table, F-statistic, df, and p-value

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 4 | 217.653784 | 54.413446 | 62.87 | <.0001 |
| **Error** | 2579 | 2232.120383 | 0.865498 |  |  |
| **Corrected Total** | 2583 | 2449.774168 |  |  |  |

* Based on the results of the ANOVA test of the log transformed Incomes with an F value of 62.87 and a p-value less than .0001, we reject the null hypothesis that all populations means are equal. There is ample evidence that at least 2 of the population groups have different means.
* Rather than using an ANOVA on a transformation, if I use a permutation test on the untransformed data, there is still strong evidence that the null hypothesis should be rejected with a p-value less than .0001.
* I trust the permutation test more because it does not do any type of transformation but rather uses 10000 permutations.

data incomes;

infile '/folders/myfolders/sasuser.v94/ex0525.csv' dlm=',' firstobs = 2;

input Subject Educ $ Income2005;

RUN;

data logs;

set incomes;

logincome = log(Income2005);

run;

proc univariate data=logs;

var logincome;

class Educ;

histogram logincome;

qqplot logincome;

run;

proc GLM data=logs;

class Educ;

model logincome=Educ;

run;

proc means data=incomes;

var Income2005;

class Educ;

run;

proc means data=logs;

var logincome;

class Educ;

run;

PROC NPAR1WAY SCORES=DATA DATA=incomes;

VAR Income2005;

CLASS educ;

EXACT SCORES=DATA / MC seed = 25071945;

TITLE 'Permutation Test (all groups have separate means)';

RUN;