Tutorial 7

ST2137-2420

Material

This tutorial offers practice with SAS. All the questions are centered around reproducing previous R/Python analysis in SAS. Please revise the material from chapter 6 of the course textbook. There are a few questions where you will need to read the SAS documentation. In particular, the `proc univariate` and `proc freq` will be useful.

Dataset: Student Performance

If necessary, the following code snippet can be used to convert `G1` from character to numeric, in a new dataset in SAS.

```
data st2137.stud_perf2;
  set st2137.stud_perf;
  G1_num = input(G1, 8.);
run;
```

1. Generate summary statistics for `G1` scores, conditioned on `Medu`.

Analysis Variable : G1_num							
Medu	N Obs	Mean	Std Dev	Minimum	Maximum	Median	N
0	3	12.0000000	4.5825757	7.0000000	16.0000000	13.0000000	3
1	59	9.7457627	3.0433263	5.0000000	18.0000000	9.0000000	59
2	103	10.5631068	2.9394874	5.0000000	18.0000000	10.0000000	103
3	99	10.6060606	3.5190971	3.0000000	18.0000000	11.0000000	99
4	131	11.9083969	3.3176692	5.0000000	19.0000000	12.0000000	131

Figure 1: G1 summary

2. Generate the following boxplots of `G1` scores, by `Medu`. Compare the distribution of `G1` scores with those of `G3` scores (we had used this variable throughout chapter 3 of the textbook).
3. Conduct a χ^2 test of independence of the variables `famrel` and `goout` at 5% significance level.

As we discussed in the tutorial, due to the expected cell counts being less than 5, we can turn to simulation to get the p-value. Here is the code that will do it:

```
proc freq data=ST2137.STUD_PERF;
    tables  (famrel) *(goout) / chisq nopercent norow nocol nocum
        plots(only)=(freqplot mosaicplot);
    exact chisq / mc;
run;
```

4. Obtain and interpret the 90% confidence interval for the odds ratio between variables `nursery` and `higher`.

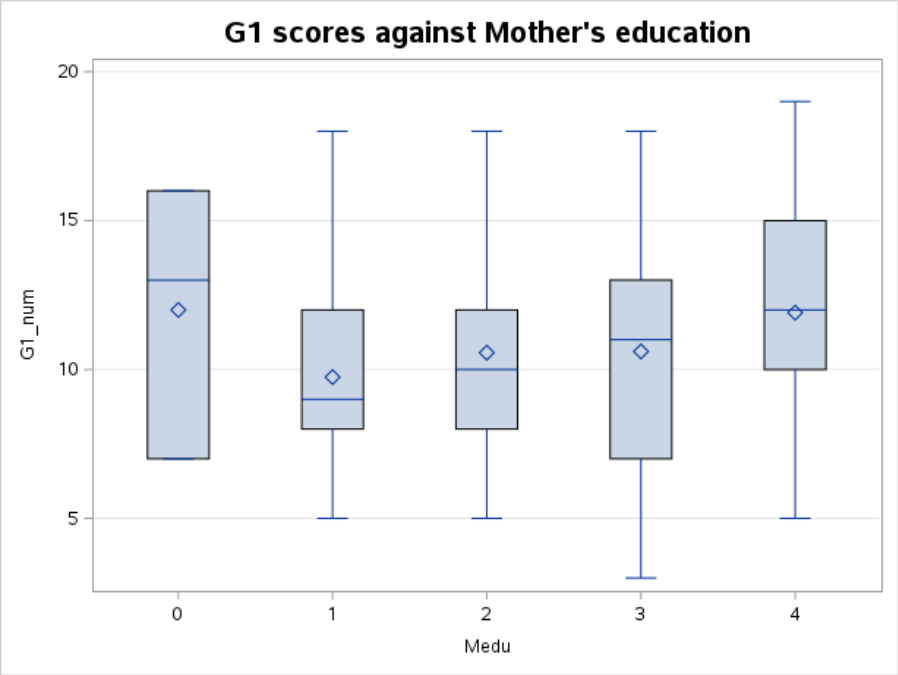

Figure 2: G1 boxplot

Frequency	Table of famrel by goout					
	goout					
famrel	1	2	3	4	5	Total
1	1	4	2	0	1	8
2	2	2	4	5	5	18
3	5	20	25	13	5	68
4	11	51	61	48	24	195
5	4	26	38	20	18	106
Total	23	103	130	86	53	395

Figure 3: Table

Statistics for Table of famrel by goout

Statistic	DF	Value	Prob
Chi-Square	16	16.9473	0.3890
Likelihood Ratio Chi-Square	16	18.1756	0.3137
Mantel-Haenszel Chi-Square	1	1.6426	0.2000
Phi Coefficient		0.2071	
Contingency Coefficient		0.2028	
Cramer's V		0.1036	

WARNING: 40% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 395

Figure 4: Statistics

Pearson Chi-Square Test	
Chi-Square	16.9473
DF	16
Asymptotic Pr > ChiSq	0.3890

Monte Carlo Estimate for the Exact Test	
Pr >= ChiSq	0.3854
99% Lower Conf Limit	0.3729
99% Upper Conf Limit	0.3979
Number of Samples	10000
Initial Seed	897020554

Figure 5: Monte Carlo estimate

The 90% confidence interval for the odds ratio is (0.6375, 4.6102), although the point estimate is 1.7143. It is a warning that even if the point estimate seems far from 1, we should still pay attention to the confidence intervals.

In order to obtain 90% (instead of 95%), it is necessary to add the `alpha` option in the code.

```
proc freq data=ST2137.STUD_PERF2;
    tables  (nursery) *(higher) / chisq relrisk alpha=0.1 expected deviation
    nopercent norow nocol nocum plots(only)=(freqplot mosaicplot);
run;
```

Working with contingency tables

5. Reproduce the χ^2 test that we performed on the political association data in Example 4.9 of the text.

The following code will work directly on the contingency table from the notes:

```
proc format;
   value partyfmt 1='Dem'
                  2='Ind'
                  3='Rep';
   value genderfmt 1='female'
                   2='male';
run;

data PoliticalPref;
   input party gender count;
   label party='Political Party Preference';
   datalines;
1 1 762
1 2 484
2 1 327
2 2 239
3 1 468
3 2 477
;

proc sort data=PoliticalPref;
   by descending gender descending party;
run;

proc freq data=PoliticalPref order=data;
   format gender genderfmt. party partyfmt.;
   tables gender*party / chisq relrisk;
   /*exact pchi or;*/
   weight Count;
   title 'Case-Control Study of High Fat/Cholesterol Diet';
run;
```

Robust statistics

6. Upload the `mass_chem` dataset from chapter 5 (robust statistics) to SAS. Use `PROC UNIVARIATE` to obtain the following robust estimates of location and scale:
   - Trimmed mean ($\gamma = 0.1$)
   - Winsorised mean ($\gamma = 0.1$)
   - MAD

   Can you explain the differences with the estimates we obtained from R/Python?

The following code will be able to generate the required output:

```
/* Exploring Data */
proc univariate data=ST2137.CHEM trimmed=.1 winsorized=.1  robustscale;
    /*ods select Histogram;*/
    var chem;
    /*histogram chem;*/
run;
```

Trimmed Means

Percent Trimmed in Tail	Number Trimmed in Tail	Trimmed Mean	Std Error Trimmed Mean	95% Confidence Limits		DF	t for H0: Mu0=0.00	Pr > |t|
12.50	3	3.218333	0.137117	2.929041	3.507626	17	23.47137	<.0001

Winsorized Means

Percent Winsorized in Tail	Number Winsorized in Tail	Winsorized Mean	Std Error Winsorized Mean	95% Confidence Limits		DF	t for H0: Mu0=0.00	Pr > |t|
12.50	3	3.176250	0.138122	2.884838	3.467662	17	22.99598	<.0001

Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	0.950000	0.704236
Gini's Mean Difference	2.830906	2.508825
MAD	0.355000	0.526323
Sn	0.799042	0.799042
Qn	0.733227	0.633002

Figure 6: Robust statistics

The reason for the difference is that SAS modifies the γ from 0.1 to 0.125.