22S:30/105, Statistical Methods and Computing
Spring 2015, Instructor: Cowles
Midterm 3

Show your work on any problems that involve calculations.

Name: _____ Course no. (30 or 105) _____

1. This question uses some of the data from the following dataset:

```
Personal income and demographic data from the March, 2011 supplement to
the Current Population Survey.  Data on all 80,976 respondents aged 25
to 64 years who were currently in the labor force and who listed their
race as Asian, black, or white.  This is a random sample from all such
residents of the United States.
variables and coding:
Sex     1=male, 2=female
Income  Total personal income, dollars
Race    Person's race, 1=white, 2=black, 4=Asian
Age     Person's age in years
Educ    Educational attainment,
                1=less than high school
                2=some high school but no diploma
                3=high school graduate
                4=some college but less than bachelor's degree
                5=bachelor's degree
                6=master's, professional, or doctoral degree
        (Educational attainment is condensed from 16 levels in
         the CPS data.)
```

We will use 92 observations from the larger dataset. This sample began as a simple random sample from the larger dataset. There were too few people (6) with educational levels below high school graduate to draw any conclusions about those categories, so I deleted those observations. I deleted two additional observations with unlikely values of income.

We wish to use these data to determine whether mean income is different among U.S. adults with different levels of education: high school graduate, some college, bachelor's degree, and graduate degree. Refer to the attached SAS output to answer some of the following questions.

(a) ANOVA will be our first choice of statistical method with which to address our question. Why is ANOVA more appropriate than a Chi square test? (Answer in one or two sentences.)

(b) Write the null hypothesis to be tested. Use standard statistical symbols.

(c) The data used for this analysis actually are a random sample from the populations of interest. There are two other assumptions that must be met in order for the results of ANOVA to be trustworthy. List both assumptions, and for each one, refer to SAS output to tell whether it is likely met in this data.

(d) At the .05 significance level, can we reject the null hypothesis that mean income is the same in all 4 educational levels? State your conclusion, citing the relevant test statistic and p-value from the SAS output.

(e) At the .05 significace level, which pairs of population means are unequal?

(f) Do these results prove that getting more education causes people to have higher incomes? Why or why not?

(g) In the SAS output on page 8, the following confidence interval is given in the first row of a list: (-27779, 37807). What quantity are we 95% confident lies in that interval? Explain in words, and give appropriate statistical sympols.

2. We could use the same data to test whether mean income is the same for men as for women.

   (a) Which test procedure would be most appropriate for this purpose (circle one):

      i. paired t-test
      ii. two independent sample t-test
      iii. Chi square test
      iv. z test
      v. sign test

   (b) Is there anything in the attached SAS output that suggests that we should not use the procedure that you circled? Explain.

3. Do mothers of 6th grade girls think that there should be a dress code at their daughters' school? An elementary school principal selected a simple random sample of size 10 from among the mothers of 6th grade girls at his school. He contacted each of the mothers and asked her if she thought the school should institute a dress code. Four mothers said "yes" and six mothers said "no."

   (a) The population most likely of interest to the principal is (circle one):

      i. all mothers of current 6th grade girls
      ii. the 10 mothers whom he contacts
      iii. the mothers who say yes
      iv. the proportion who think there should be a dress code

   (b) Use the plus-four method to calculate a 95% confidence interval. (Numeric answer; show your work.)

   (c) From the SAS output below, find the following quantities and write them in.

      i. point estimate of population proportion

      ii. 95% confidence interval from normal approximation

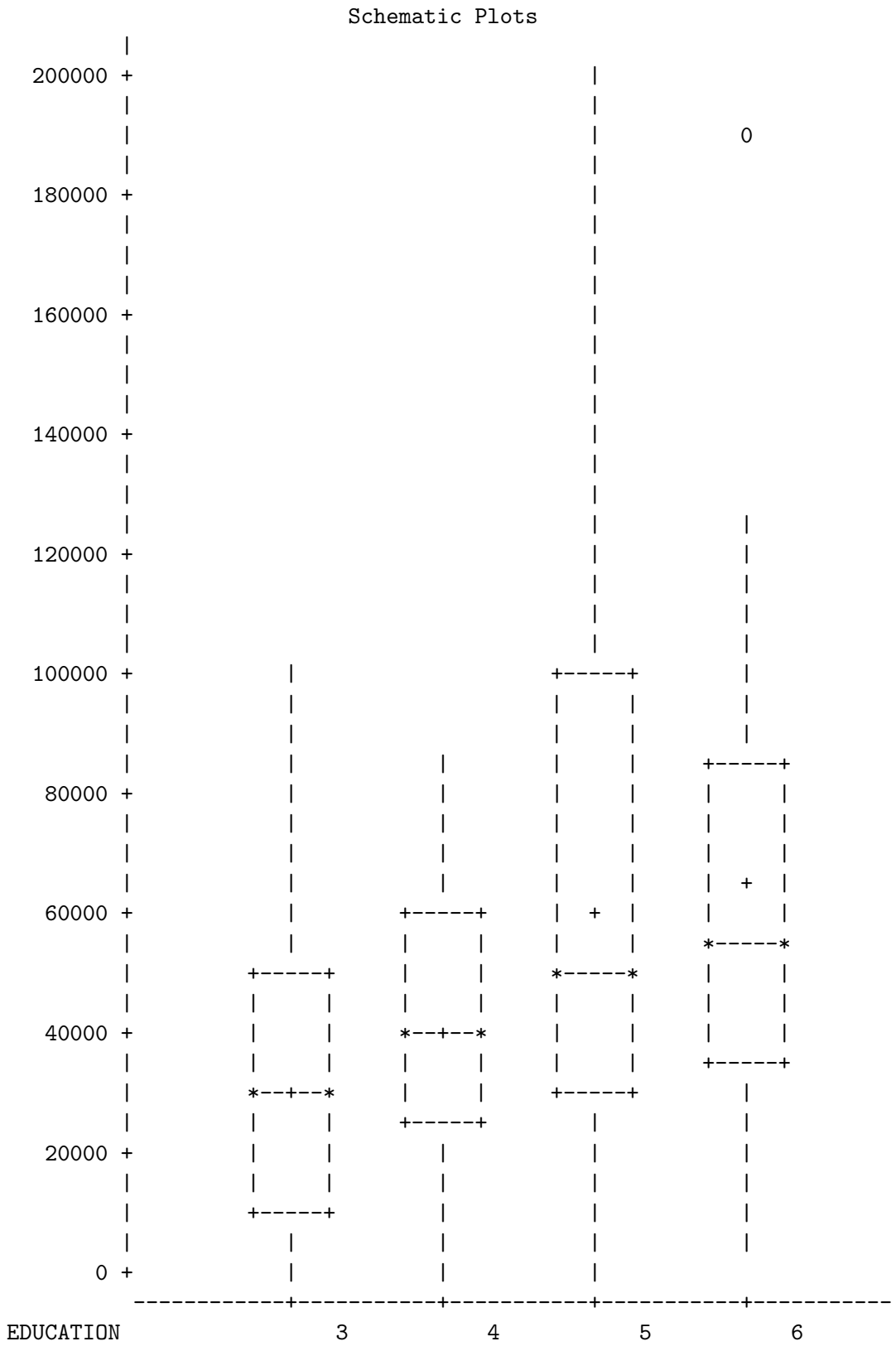
      iii. exact 95% confidence interval

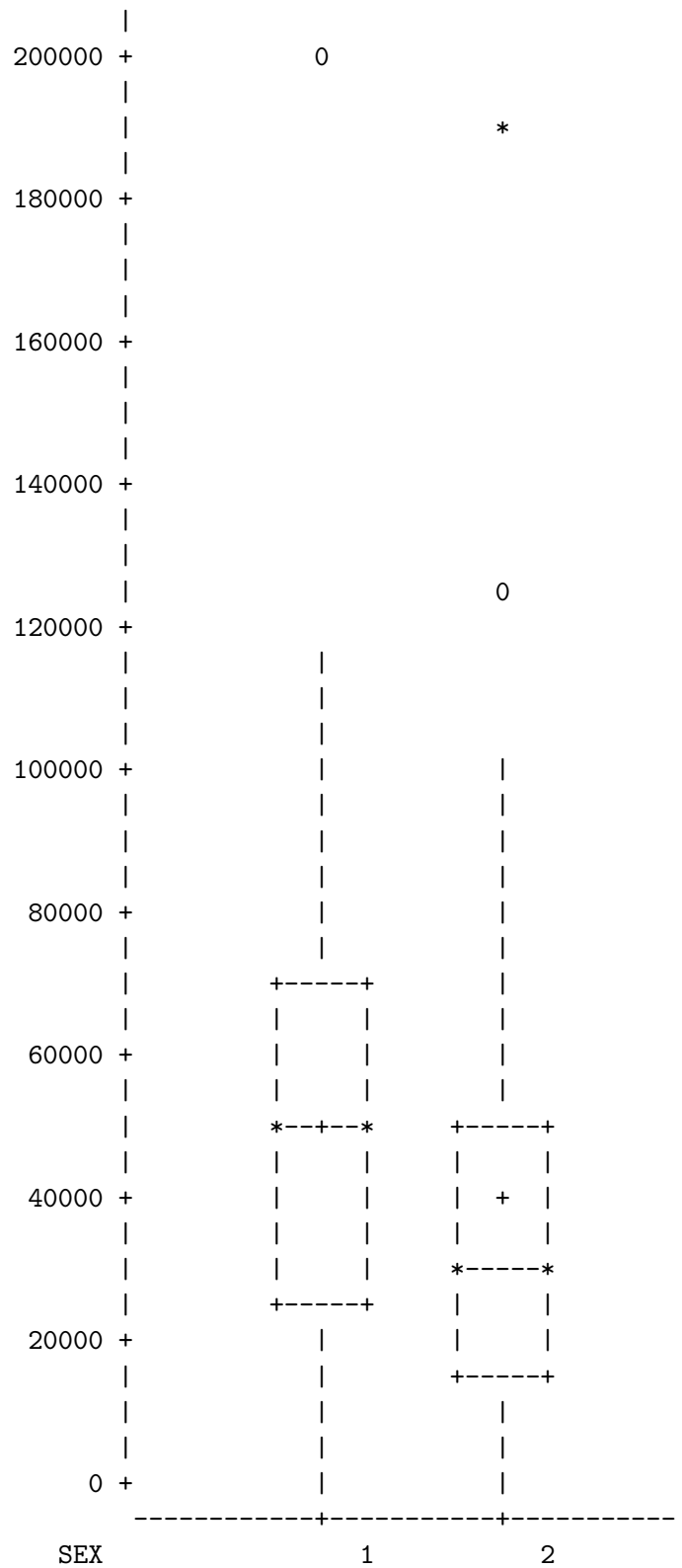
(d) The three confidence intervals are fairly different. Why would that happen with these data?

(e) Suppose you wanted to test the following hypotheses regarding the population proportion of moms of 6th grade girls who want a dress code.

$$H_0: \quad p = 0$$
$$H_A: \quad p \neq 0$$

What would you conclude from the confidence intervals provided?


Schematic Plots
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
0
EDUCATION
3
4
5
6


200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
SEX
1
2

The ANOVA Procedure

Class Level Information

Class	Levels	Values
EDUCATION	4	3 4 5 6

Number of Observations Read	92
Number of Observations Used	92

The ANOVA Procedure

Dependent Variable: INCOME

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	16169794816	5389931605.5	4.27
Error	88	111125611362	1262791038.2	
Corrected Total	91	127295406178		

Source	Pr > F
Model	0.0073
Error	
Corrected Total	

R-Square	Coeff Var	Root MSE	INCOME Mean
0.127026	76.00893	35535.77	46752.10

Source	DF	Anova SS	Mean Square	F Value
EDUCATION	3	16169794816	5389931605	4.27

Source	Pr > F
EDUCATION	0.0073

The ANOVA Procedure

Bonferroni (Dunn) t Tests for INCOME

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	88
Error Mean Square	1.2628E9
Critical Value of t	2.69921

Comparisons significant at the 0.05 level are indicated by ***.

EDUCATION Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
6 - 5	5014	-27779	37807	
6 - 4	22948	-8849	54745	
6 - 3	33627	2581	64673	***
5 - 6	-5014	-37807	27779	
5 - 4	17934	-9852	45720	
5 - 3	28613	1689	55536	***
4 - 6	-22948	-54745	8849	
4 - 5	-17934	-45720	9852	
4 - 3	10679	-15022	36380	
3 - 6	-33627	-64673	-2581	***
3 - 5	-28613	-55536	-1689	***
3 - 4	-10679	-36380	15022	

The MEANS Procedure

Analysis Variable : INCOME

EDUCATION	N Obs	N	Mean	Std Dev	Minimum
3	30	30	31774.80	25406.65	0
4	26	26	42453.62	25799.73	0
5	22	22	60387.77	46691.04	0
6	14	14	65401.71	48025.95	5000.00

Analysis Variable : INCOME

EDUCATION	N Obs	Maximum
3	30	100000.00
4	26	85300.00
5	22	200000.00
6	14	191100.00

The MEANS Procedure

Analysis Variable : INCOME

SEX	N Obs	N	Mean	Std Dev	Minimum
1	55	55	50878.15	35960.33	0
2	37	37	40618.78	39135.59	0

Analysis Variable : INCOME

SEX	N Obs	Maximum
1	55	200000.00
2	37	191100.00

The FREQ Procedure

resp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	4	40.00	4	40.00
N	6	60.00	10	100.00

Binomial Proportion
resp = A

Proportion	0.4000
ASE	0.1549
95% Lower Conf Limit	0.0964
95% Upper Conf Limit	0.7036
Exact Conf Limits	
95% Lower Conf Limit	0.1216
95% Upper Conf Limit	0.7376