2
15
12
4
3 3

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

Show your work on any problems that involve calculations.

Name: Solutions 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 too 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.)

Income is a quantitative variable, and we are interested in population means. The Chi square test is for inference about proportions.

2

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

$H_0: \mu_{hs} = \mu_{sc} = \mu_{bach} = \mu_{grad}$

where each mu is a population mean income for people at a given educational level

(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.

5

① All populations of interest are normal. Outliers in boxplots for ~~college degree and~~ grad calls this one into question. Maybe ok

② All population standard deviations are equal. In proc means output, the largest sample standard deviation (~~37486~~ 48026) is not more than twice as large as the smallest one (25407).

(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.

2 Yes. F Test statistic is 4.27, and the p-value is 0.0073. The p-value is much smaller than the significance level, so reject H_0.

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

2 $\mu_6 \neq \mu_3$ High school + graduate

$\mu_5 \neq \mu_3$ College degree and high school

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

2 No. They show a relationship but cannot prove causation. Perhaps people who are smarter and work harder get higher incomes and more education.

(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 95appropriate statistical sympols.

2 $\mu_6 - \mu_5$

↑ grad ↑ bachelor's

2

15

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

2 ii. two independent sample t-test

iii. Chi square test

iv. z test

v. sign test

1 pt for paired

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

2 Outlier for gender 2.

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."

2 (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.)

3 $\frac{4+2}{10+4} = \frac{6}{14} = .428$

$(0.169, 0.688)$

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

i. point estimate of population proportion

0.40

1

ii. 95% confidence interval from normal approximation

1 (0.0964, 0.7036)

iii. exact 95% confidence interval

1 (0.1216, 0.7376)

4 12

(d) The three confidence intervals are fairly different. Why would that happen with these data? The sample size (10) is very small.

2 The rules of thumb for the normal approximation are not met

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

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

What would you conclude from the confidence intervals provided?

2 Reject H_0. 0 is not in any of the intervals.

4