PRACTICE PROBLEMS for MIDTERM 3 in 2006
22S:30/105, Statistical Methods and Computing
Spring 2005, Instructor: Cowles
Midterm 3

Show your work on any problems that involve calculations.

There are 30 total points on this midterm. Point values for each question are shown in parentheses. I will grade on a curve and will give partial credit wherever possible.

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

1. (3) According to `http://www.city-data.com/city/Iowa-City-Iowa.html`, 34.6% of Iowa City residents are of German ancestry. This information was obtained from a census, and can be considered to be a correct value for the *population*.

   You wish to interview a simple random sample of 500 residents of Iowa City. What is the probability that your sample will contain at least 35% people of German ancestry? (Use Normal probability calculations.)

2. A Gallup Poll on energy use asked 512 randomly-selected adults in the U.S. whether they favored "increasing the use of nuclear power as a major source of energy." Gallup reported that 225 said "Yes."

   (a) (1/2) What is the population of interest? (circle one)
       i. all adults in the U.S.
       ii. the 512 adults surveyed
       iii. the 225 who said "Yes"
       iv. the proportion of all U.S. adults who favor increasing the use of nuclear power as a major source of energy
       v. the proportion of people surveyed who favor increasing the use of nuclear power as a major source of energy
       vi. none of the above

   (b) (1/2) What is the parameter of interest? (circle one)
       i. all adults in the U.S.
       ii. the 512 adults surveyed
       iii. the 225 who said "Yes"

       iv. the proportion of all U.S. adults who favor increasing the use of nuclear power as a major source of energy
       v. the proportion of people surveyed who favor increasing the use of nuclear power as a major source of energy
       vi. none of the above

   (c) (1) Use the data given to calculate the point estimate of the proportion of all U.S. adults who favor increasing the use of nuclear power as a major source of energy. (numeric answer)

   (d) (1/2) What is the conventional symbol for the number you calculated in the previous problem? (circle one)
       i. μ
       ii. $\hat{\mu}$
       iii. $\bar{x}$
       iv. p
       v. $\hat{p}$
       vi. s

   (e) (2) Are the assumptions regarding population size and sample size met so that you could use the normal approximation to calculate a 95% confidence interval for the proportion of interest? (yes/no) Briefly state each assumption and verify whether it is met in this problem.

   (f) (2) Regardless of how you answered the previous question, use the data given to calculate the 95% confidence interval.

(g) (1/2) What quantity are we 95% confident is in the interval you calculated? (circle one)

    i. μ

    ii. $\hat{\mu}$

    iii. $\bar{x}$

    iv. p

    v. $\hat{p}$

    vi. s

3. To see how much difference time of day made on the speed at which he could download files, a college sophomore performed an experiment. He placed a file on a remote server and then proceeded to download it at three different time periods of the day – 16 times in the morning at 7:00 a.m., 16 times in the evening at 5:00 p.m. and 16 times in the late night at midnight. To choose his days and times of measurement, he randomly selected 48 different days over a 120-day period, and for each selected day, he randomly determined whether he would do the download in the morning, evening, or late night. For each download, he measured the time in seconds.

  (a) (1) The variable of interest is (circle one):

    i. time of day

    ii. download time in seconds

    iii. the size of the file

    iv. the number of times downloaded

    v. none of the above

  (b) (1.5) The null hypothesis for the student's experiment could be stated informally as "Time of day has no effect on download time." Write this null hypothesis as a statement about population parameters, using conventional symbols.

  (c) (1) The type of hypothesis test that is most likely to be applicable for this problem is (circle one):

    i. one-sample t test

    ii. Chi-square test

    iii. ANOVA

    iv. linear regression

  (d) (1) What are the degrees of freedom for the test statistic for using the dataset described and the test you chose in the previous question? Give either one or two numbers, depending on which test you chose.

4. NOTE: There will not be any regression questions on midterm 3 in 2006. A question of this type could appear on the Final.

Every spring, Nenana, Alaska, hosts a contest in which participants try to guess the exact minute that a wooden stand placed on the frozen Tanana River will fall through the breaking ice. The contest started in 1917 as entertainment for railroad engineers. It has grown into an event in which hundreds of thousands of entrants enter their guesses on the Internet and compete for prizes of more than $300,000. Because so much money depends on the time of ice breakup, it has been recorded to the nearest minutie with great accuracy ever since 1917. An article in *Science* ("Climate Change in Nontraditional Datasets," Oct. 2001, p. 811) used the data to investigate global warming by asking the question whether ice breakup had tended to occur earlier over time.

The dataset available to us contains two variables:

- `year`
- `julian` – the number of days from midnight on Jan 1 until the time of ice breakup

Refer to the SAS output provided to answer the following questions.

  (a) (1) The null hypothesis is that there is no linear relationship between year and time of ice breakup. Write this null hypothesis as a statement about a population parameter. Use conventional symbols.

  (b) (1) The alternative hypothesis is that time of ice breakup decreases linearly over time. Write this alternative hypothesis as a statement about a population parameter. Use conventional symbols.

  (c) (1) Give a point estimate and a 95% confidence interval for the population slope (numeric answers).

  (d) (2) Does your answer to the preceding question provide evidence in favor of the alternative hypothesis? (yes/no) Explain briefly. (If you could not answer the previous question, pretend that the point estimate is -0.11 and the confidence interval is (-0.21,-0.01) and answer this question accordingly.)

(e) (1.5) What is the p-value for the one-sided test of no linear relationship between year and time of breakup?

(f) (1.5) Use the estimated regression equation to predict the time of breakup for this year (2005). Show your calculation.

(g) (1) On the SAS output, circle the numbers that provide the endpoints of the interval in which you are 95% confident that this year's breakup time will lie. Be sure to put your name on the SAS output.

(h) (1) Does the residual plot show evidence of severe outliers, nonlinearity,or inequality of variance? (yes/no)

(i) (1) What proportion of the variability in time of breakup is explained by year? (numeric answer)

(j) (0.5) What is the estimated value of the standard deviation of points around the regression line? (numeric answer)

```
options linesize = 72 ;

data nenana ;
infile '/group/ftp/pub/kcowles/datasets/nenana.dat'  delimiter = '09'x
firstobs = 2 ;
input year julian ;
run ;

proc print ;

proc reg lp ;
model julian = year  / clb clm cli ;
id year ;
run ;
plot residual. * predicted. = '.' / vplots = 2 ;
run ;
```

5. Many believe in the "homecourt advantage" in college basketball – the idea that a team plays better and is more likely to win when playing on their own court (home games) than when playing at the opponent's locationi (away games). THe following table shows the counts of wins and losses in home games and away games for a sample of college basketball games.

	Win	Lose	Total
Home	8	1	9
Away	4	5	9
Total	12	6	18

(a) (3) Write the table of expected counts that would be expected under the null hypothesis that the proportion of wins is the same for home games as away games.

	Win	Lose	Total
Home			9
Away			9
Total	12	6	18

(b) (1) Are the sample sizes large enough that a Chi square test could be used to test the null hypothesis? Briefly justify your answer.

6. This dataset contains the labor force participation rate (LBFP) of women in 19 cities in the United States in each of two years (1968 and 1972). The data help to measure the presence of women in the labor force over this period. The variables are:

1. city: City in the United States
2. lbfp1972: Labor Force Participation rate of women in 1972
3. lbfp1968: Labor Force Participation rate of women in 1968

Labor force participation is a quantitative variable.

Suppose that these 19 cities could be considered a simple random sample of cities in the U.S. We wish to use these data to draw conclusions as to whether the mean labor force participation of women in all U.S. cities stayed the same or changed between 1968 and 1972. We do not know in advance in which direction a change might have gone.

(a) (1) The populations of interest are (circle one):

   i. all women in the labor force

   ii. all cities in the U.S. in 1968 and all cities in the U.S. in 1972

   iii. the 19 cities included in the study

   iv. labor force participation of women

   v. the average labor force participation of women in all U.S. cities in 1968 and the average labor force participation of women in all U.S. cities in 1972

   vi. the average labor force participation of women in the 19 cities in 1968 and the average labor force participation of women in the 19 cities in 1972

   vii. none of the above

(b) (1) The parameters of interest are (circle one):

   i. all women in the labor force

   ii. all cities in the U.S. in 1968 and all cities in the U.S. in 1972

   iii. the 19 cities included in the study

   iv. labor force participation of women

   v. the average labor force participation of women in all U.S. cities in 1968 and the average labor force participation of women in all U.S. cities in 1972

   vi. the average labor force participation of women in the 19 cities in 1968 and the average labor force participation of women in the 19 cities in 1972

   vii. none of the above

(c) (1) The variable of interest is (circle one):

   i. all women in the labor force

   ii. all cities in the U.S. in 1968 and all cities in the U.S. in 1972

   iii. the 19 cities included in the study

   iv. labor force participation of women

   v. the average labor force participation of women in all U.S. cities in 1968 and the average labor force participation of women in all U.S. cities in 1972

   vi. the average labor force participation of women in the 19 cities in 1968 and the average labor force participation of women in the 19 cities in 1972

   vii. none of the above

(d) (1) The study has been set up as a

   i. single-sample problem

   ii. paired-sample problem

   iii. two-independent-sample problem

   iv. none of the above

(e) (2) Write appropriate null and alternative hypotheses for a test that would address the nutritionists' research question. Use conventional symbols.

(f) (2) Use the SAS output to give the test statistic and p-value for the test (numeric answers)

(g) (2) Should we reject the null hypothesis at the .05 significance level? (yes/no) Briefly explain.

(h) (1) The appropriate interpretation of the result of the hypothesis test alone is:

   i. the data provided sufficient evidence to reject the null hypothesis that the average of women's labor force participation in U.S. cities did not change between 1968 and 1972

   ii. the data data did not provide sufficient evidence to reject the null hypothesis that the average of women's labor force participation in U.S. cities did not change between 1968 and 1972

   iii. the data prove that the average of women's labor force participation in U.S. cities did not change between 1968 and 1972

   iv. the data prove that the average of women's labor force participation in U.S. cities changed between 1968 and 1972

   v. the data prove that the average of women's labor force participation in U.S. cities increased between 1968 and 1972

(i) (2) Does any part of the SAS output give clear evidence of a change in one particular direction between 1968 and 1972? (yes/no) If so, cite the specific output and state the direction of change.

Obs	city	lbfp1972	lbfp1968	lbfp1972-lbfp1968 diff
1	N.Y.	0.45	0.42	0.03
2	L.A.	0.50	0.50	0.00
3	Chicago	0.52	0.52	0.00
4	Philadelphia	0.45	0.45	0.00
5	Detroit	0.46	0.43	0.03
6	San Francisco	0.55	0.55	0.00
7	Boston	0.60	0.45	0.15
8	Pitt.	0.49	0.34	0.15
9	St. Louis	0.35	0.45	-0.10

```
10      Connecticut      0.55      0.54      0.01
11      Wash., D.C.      0.52      0.42      0.10
12      Cinn.            0.53      0.51      0.02
13      Baltimore        0.57      0.49      0.08
14      Newark           0.53      0.54     -0.01
15      Minn/St. Paul    0.59      0.50      0.09
16      Buffalo          0.64      0.58      0.06
17      Houston          0.50      0.49      0.01
18      Patterson        0.57      0.56      0.01
19      Dallas           0.64      0.63      0.01
```

                The MEANS Procedure

```
                                      Lower 95%      Upper 95%
Variable    N        Mean     Std Dev  CL for Mean   CL for Mean
-------------------------------------------------------------------
lbfp1968   19    0.4931579  0.0679912   0.4603872     0.5259286
lbfp1972   19    0.5268421  0.0707933   0.4927208     0.5609634
diff       19    0.0336842  0.0597412   0.0048899     0.0624785
     (lbfp1972-lbfp1968)
-------------------------------------------------------------------
```

                The UNIVARIATE Procedure
                  Variable: diff (lblfp1972 - lbfp1968)

                Tests for Location: Mu0=0

```
        Test           -Statistic-    -----p Value------

        Student's t    t  2.457704   Pr > |t|    0.0244
        Sign           M       5.5   Pr >= |M|   0.0074
        Signed Rank    S        46   Pr >= |S|   0.0062
```