Name: _____

Practice First Midterm Exam

Statistics 1000 Spring 2007 (Pfenning)

This is a closed book exam worth 150 points. You are allowed to use a calculator and a two-sided sheet of notes. There are 9 problems, with point values as shown. If you want to receive partial credit for wrong answers, show your work. Don't spend too much time on any one problem.

1. (5 pts.) Suppose we are interested in finding out if students tend to sleep less, the older they are.

   (a) What would be an appropriate display? (i) bar graph (ii) histogram (iii)side-by-side boxplots (iv) scatterplot

   (b) Which of these would provide the best summary? (i) compare percentages (ii) compare means and standard deviations (iii) compare Five Number Summaries (iv) report the correlation

2. (5 pts.) Suppose we are interested in finding out if smokers exercise less than non-smokers. Data values for exercise times include some high outliers.

   (a) What would be an appropriate display? (i) bar graph (ii) histogram (iii)side-by-side boxplots (iv) scatterplot

   (b) Which of these would provide the best summary? (i) compare percentages (ii) compare means and standard deviations (iii) compare Five Number Summaries (iv) report the correlation

3. (5 pts.) Suppose we are interested in finding out if males are just as likely as females to prefer the color black.

   (a) What would be an appropriate display? (i) bar graph (ii) histogram (iii)side-by-side boxplots (iv) scatterplot

   (b) Which of these would provide the best summary? (i) compare percentages (ii) compare means and standard deviations (iii) compare Five Number Summaries (iv) report the correlation

4. (10 pts.) Words per minute typed by experienced typists follows a normal distribution with mean 60 and standard deviation 15.

   (a) According to the 68-95-99.7 Rule, 95% of experienced typists type between _____and _____words per minute.

   (b) Suppose an experienced typist can type 78 words per minute. What is his standard (z) score?_____

5. (20 pts.) Two banks each have three tellers helping customers. One bank requires customers to stand in separate lines for the three tellers, the other has customers stand in a single line and be called to the next available teller. Below are a back-to-back stemplot and side-by-side boxplot for waiting times (stems are minutes) of 10 customers at the bank with separate lines and 11 customers at the bank with a single line.

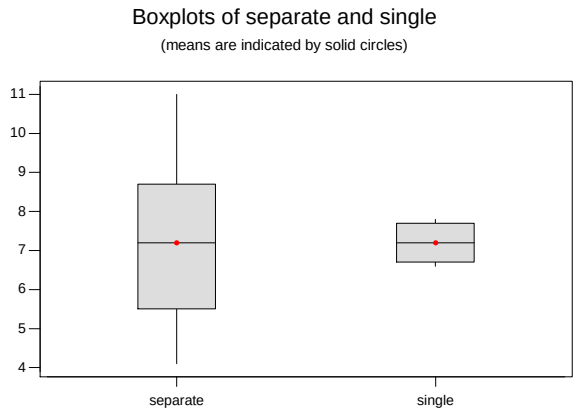
Separate			Single
	1	4	
6	2	5	
7	2	6	6 7 7 9
7	7	7	1 2 3 4 7 8 8
	5	8	
	3	9	
		10	
	0	11	

(a) Judging from the looks of the stemplot, which arrangement seems to be faster?
(i) separate lines (ii) single line (iii) both about the same

(b) For which arrangement do the waiting times have more spread?
(i) separate lines (ii) single line (iii) both about the same

(c) One fourth of the customers in the bank with separate lines waited _____minutes or less. (Find Q1.)

(d) The boxplots indicate that both distributions are
(i) very left-skewed (ii) fairly symmetric (iii) very right-skewed



6. (35 pts.) Researchers wanted to determine if seat position on a bus played a role in whether or not passengers experienced nausea:

	Nausea	No Nausea	Total
Front	60	900	960
Rear	190	850	1040
Total	250	1750	2000

(a) What is the best choice for explanatory variable?_____

(b) What is the overall probability of a passenger experiencing nausea?_____

(c) Find the probability of nausea, given that a passenger sat in the rear. _____

(d) Sitting in the front and sitting in the rear were (i) mutually exclusive (ii) independent (iii) both (iv) neither

(e) Experiencing nausea and sitting in the rear were (i) mutually exclusive (ii) independent (iii) both (iv) neither

(f) Find the probability of sitting in the rear or experiencing nausea.

(g) Find the probability of sitting in the rear and experiencing nausea.

(h) Compute a table of counts expected if nausea were independent of seat position.

(i) Calculate the chi-square statistic (round to the nearest tenth).

(j) The p-value is (i) much less than .05 (ii) close to .05 (iii) much more than .05

(k) What can we conclude?

i. seat position has no significant effect

ii. passengers in the rear are significantly more likely to experience nausea
iii. passengers in the rear are significantly less likely to experience nausea

7. (30 pts.) Cereal manufacturers looked at the relationship between number of days x that 14 cereal boxes spent on the supermarket shelf, and moisture content y. Scatterplot and regression output are given below.

   (a) What is the response variable?_____

   (b) Sitting on the shelf tends to make cereal (i) dryer (ii) soggier (iii) neither

   (c) Which of the following is the best guess for r?
   (i) -.95 (ii) -.55 (iii) -.15 (iv) .15 (v) .65 (vi) .95

   (d) If we switched the roles of x and y, then which of the following would change?
   (i) the value of r (ii) the equation of the regression line (iii) both (iv) neither

   (e) Predict the moisture content of a cereal box that sat on the shelf for 10 days.
   _____

   (f) What is the residual for a shelf time of 10 days, if the actual moisture content was 3.40? _____

   (g) Suppose a supermarket accidentally kept a cereal box on the shelf for 100 days. What can we say about its moisture content?

   i. It should equal 7.29.

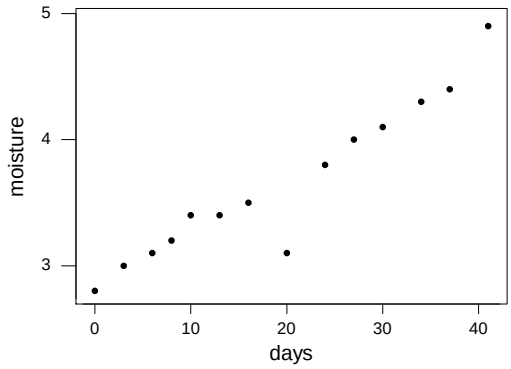
   ii. It should be very close to the predicted value because of the high x value.

   iii. It could be far from the predicted value because of extrapolation.

   (h) The box which spent 20 days on the shelf is an
   (i) outlier (ii) influential observation (iii) both (iv) neither

   (i) Taste tests indicated that the cereal is unacceptably soggy when the moisture content exceeds 4.1. Judging from the scatterplot, what would be a good time to remove unsold cereal from the shelf? After (i) a day (ii) a week (iii) a month (iv) a year

```
Regression Analysis: moisture versus days
The regression equation is
moisture = 2.79 + 0.045 days
Predictor        Coef     SE Coef          T        P
Constant      2.78551     0.09485      29.37    0.000
days         0.044620    0.004113      10.85    0.000
S = 0.1962      R-Sq = 90.7%     R-Sq(adj) = 90.0%
Unusual Observations
Obs       days   moisture         Fit      SE Fit    Residual    St Resid
  8       20.0     3.1000      3.6779      0.0525     -0.5779       -3.06R
R denotes an observation with a large standardized residual
```


moisture
5
4
3
0
10
20
30
40
days

8. (20 pts.) 350 students at 18 Seattle schools in high-crime areas participated in a study during the 1980's. About half of the students took part in a program throughout elementary school which trained them how to earn good grades and get along with others; the other half did not take part in the program. The pregnancy rate for young women in the program, by the time they reached the age of 21, was only 38 percent, compared with 56 percent for the women who had gotten no training.

   (a) What kind of study was this? (i) observational study (ii) experiment (iii) anecdotal evidence (iv) multistage sample

   (b) Which of these best describes the intended population of interest?

      i. 350 students at 18 Seattle schools in high crime areas

      ii. all students at Seattle schools

      iii. all students at schools in high crime areas

   (c) The treatment group's pregnancy rate was how much lower than the rate for the control group? _____

   (d) Which of the following could be a possible lurking (confounding) variable?

      i. if students in one group had a different Health and Sex Ed teacher than those in the other group

      ii. if students in one group were trained to get along with others and students in the other group were not

      iii. if female students in one group tended to get pregnant and those in the other group did not

   (e) What would be the best way for researchers to assign some students to attend the program, others not? (i) put males in one group and females in the other (ii) ask for volunteers (iii) make a random assignment

   (f) This problem involves (i) one quantitative and one categorical variable (ii) two quantitative variables (iii) two categorical variables

   (g) To summarize differences, we (i) compare percentages (ii) compare means (iii) report the correlation r

9. (20 pts.) Consider these two designs, both of which summarize actual studies conducted in recent years:

   - **Design I:** Researchers reviewed 200 births from the previous year, and found that breast-fed infants acquired significantly fewer infections.
   - **Design II:** Some hospitals in a region were assigned to implement a program to encourage new mothers to breast-feed; the rest offered no such program. Afterwards, infants in the breast-feeding program acquired significantly fewer infections.

   (a) Which of these is an experiment? (i) Design I (ii) Design II (iii) both (iv) neither

   (b) In Design I, 200 is the (i) sample size (ii) population size

   (c) Who constitutes the control group in Design II? (i) infants participating in the breast-feeding program (ii) infants not participating in the breast-feeding program

   (d) What is a possible confounding (lurking) variable in Design I?

   i. whether or not a baby is breast-fed
   ii. income/education of the mother
   iii. whether or not a baby develops an infection

   (e) Which of these would be the better way to decide which hospitals in Design II would implement the breast-feeding program?

   i. survey doctors and nurses to determine which hospitals have the best chance of implementing the program successfully
   ii. flip a coin for each hospital; if it lands on heads, the program will be implemented, and if it lands on tails the program will not be implemented

   (f) Would Design II be a double-blind study?

   i. Yes, because this is the safest way to avoid the placebo effect.
   ii. No, because mothers would be aware that the hospital is encouraging them to breast-feed.
   iii. No, because doctors have to know if a baby was breast-fed when they diagnose infections.

10. (for additional practice) Researchers are interested in whether people tend to snack more while they watch TV. Here are three possible designs:

    Design A: Recruit participants for a study. While they are presumably waiting to be interviewed, half of the individuals sit in a waiting room with snacks available and a TV on. The other half sit in a waiting room with snacks available and no TV, just magazines. Researchers determine if people consume more snacks in the TV setting.

    Design B: Recruit participants for a study. Ask them to recall, for each hour of the previous day, whether they were watching TV, and what food they consumed each hour. Determine if food consumption was higher during TV times.

    Design C: Recruit participants for a study. Give them journals to record hour by hour their activities the following day, including TV watched and food consumed. Determine if food consumption is higher during TV times.

    (a) Which of these designs is an observational study? (i) A (ii) B (iii) C
    (You may circle any number of these from none to all three.)

    (b) Which one of the three designs is most flawed because of people's faulty memories?
    (i) A (ii) B (iii) C

    (c) Which one of the three designs is most flawed because people may behave differently when they know they're being studied? (i) A (ii) B (iii) C

    (d) Which one of the three designs is most flawed because of lack of realism?
    (i) A (ii) B (iii) C

    (e) The explanatory variable is (i) quantitative (ii) categorical (iii) we have no way of knowing