Applied Statistics I **Winter 2009**

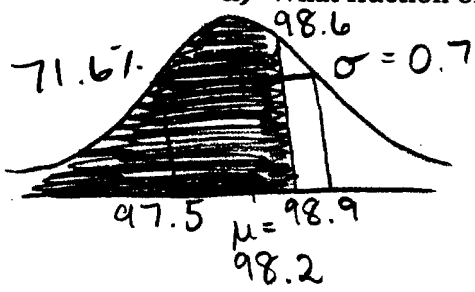
MIDTERM EXAM I
(80 points)

Name_ Sample Solutions

Read each question carefully. Write your work and answers on this test paper or, if you prefer, you may attach printed sheets to this test paper. If you need more space to write, you may continue your answer on the reverse but please indicate that you have done so. You may use resources including SPSS, applets, and your books and notes.

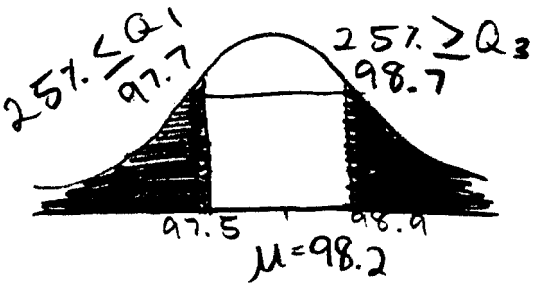
1) **(15 points)** Most people think that the "normal" adult body temperature is 98.6°F. That figure was recently challenged in a research article that said that a more accurate figure for the mean is 98.2°F with a standard deviation of 0.7°F. Assume these new findings are correct and that body temperatures have a Normal distribution to answer the questions below:

a) What fraction of people would be expected to have body temperatures below 98.6°F?



71.6% have body temperature lower than 98.6°F

b) Find the lower and upper quartiles of body temperatures using the information provided. (Hint: What proportion of all temperatures fall below the lower quartile? above the upper quartile?)



$\mu = \bar{x}$ $\bar{x} = 50\%$

$\bar{x} = M$ in Normal Distribution

$Q_1 = 25\%$ of data

$Q_3 = 75\%$ of data

$Q_1 = 25\%$ temperatures are lower than 99.2°F

$Q_3 = 75\%$ temperatures are lower than 98.2°F

25% of temperatures are higher than 98.2°F

$Q_1 = 97.7°F$

$Q_3 = 98.7°F$

Question 1) continued...

c) A feverish K student on study abroad goes to a rural clinic to have his temperature checked. Knowing that the K student might not know the Celsius scale but would certainly know statistics (☺), the doctor tells him that his standardized temperature is Z=3.22. What is his temperature in degrees Fahrenheit?

$Z = \frac{x - \mu \text{ (mean)}}{\sigma \text{ (sd)}}$ $\quad \mu = 98.2°F$ $\quad \sigma = .7°F$

$3.22 = \frac{x - \mu}{\sigma}$

$(.7)\ 3.22 = \frac{x - 98.2}{.7}(.7)$ $\quad 2.254 = x - 98.2\ (+98.2)$

$+(98.2)$

$100.454 = x$ $\quad$ 100.454°F

2) (20 points) Read the following brief report of a research study and answer the questions that follow:

> After menopause many women take supplemental estrogen. There is some concern that if these women also drink alcohol, their estrogen levels will rise too high. Twenty-four volunteers, 12 who were receiving supplemental estrogen and 12 who were not, were randomly divided into two groups. One group drank an alcoholic beverage and the other group drank a nonalcoholic beverage. An hour later everyone's estrogen level was checked. Only those on supplemental estrogen who drank alcohol showed a marked increase.

a) Explain briefly but specifically why this study fits the definition of an experiment.

An experiment requires "deliberately imposing some treatment on individuals in order to observe their responses" while an observational study only requires "observing individuals & measuring variables without attempting to influence responses." This is an experiment because the treatment being deliberately imposed on the women is whether or not they have an alcoholic beverage - we're not just observing their casual daily behavior, we are giving them a certain type of beverage

b) Identify the factor(s) and factor levels in this experiment.

Factors = supplemental estrogen & alcoholic beverage (alcohol)

Factor levels = supplemental estrogen or no supp. estrogen + alcoholic or non-alcoholic beverage

c) Identify the treatment(s) in this experiment.

Treatments: (4) supp estrogen + alcohol
supp estrogen + no alcohol
no supp estrogen + alcohol
no supp estrogen + no alcohol

alcohol	supp. estrogen: no	supp. estrogen: yes
yes	no	yes
	yes	yes
no	no	yes
	no	no

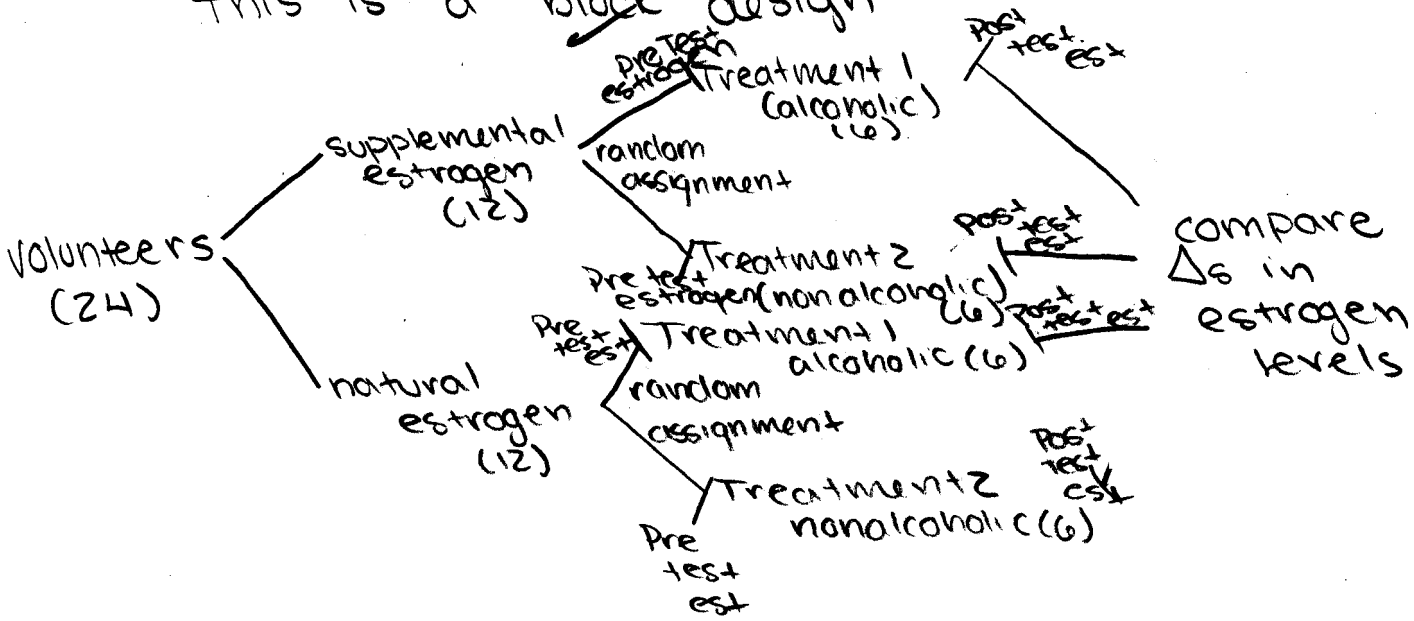
Question 2) continued...

d) What is the response variable measured in the experiment?

the response variable is change in estrogen level

e) Identify the experimental design by name and sketch a diagram of the experiment.

this is a block design



This is a block design because the random assignment of individuals to treatment is carried out separately within each block

3) (15 points) Read the following brief report of a statistical study and answer the questions that follow:

> A question posed on the Lycos Web site on June 18, 2000 asked visitors to the site to say whether they thought that marijuana should be legally available for medicinal purposes.

a) Identify the sampling method employed and say whether it is a probability or non-probability method.

- Voluntary response sampling method.
- This is a non-probability method

b) What is the sample?

The sample is the individuals who visit the Lycos web site and participate in the survey voluntarily.

c) Identify by name one potential source of bias and explain why it would make it difficult to generalize to the population of U.S. adults 18 and over.

There is a very large undercoverage bias in this study because it only includes individuals who have access to the internet and visit Lycos. Additionally, individuals who tend to use the internet most frequently are of younger generations, and may be more open minded regarding marijuana use. This may lead to a misleading statistic that can't be accurately generalized to the greater population (18 yrs +) because they may not have been sufficiently represented in the sample, and may have a very different opinion of medicinal marijuana legalization that was not expressed.

4) (30 points) The SPSS data set `tbirths.sav` stored in the SPSS data sets directory (accessed from the course webpage) contains year 2002 teen birth rates for 21 states categorized by region: 1) Nine (9) Northeastern states (NE) and 2) Twelve (12) Southern states (SO) respectively. Below is a table containing information from the data set:

Teen Birth Rates (number of births per 1000 females age 15-19)			
Northeastern States (NE)		**Southern States (SO)**	
State	**Rate**	**State**	**Rate**
Connecticut	~~25.8~~ 26	Alabama	~~54.5~~ X
Maine.	~~25.4~~ 25	Delaware	~~46.3~~ X
Massachusetts	~~23.3~~	Florida	~~44.6~~ X
New Hampshire	~~20.0~~	Georgia	~~53.7~~ X
New Jersey	~~26.8~~	Kentucky	~~51.0~~ X
New York	~~29.5~~ 30	Maryland	~~35.4~~ X
Pennsylvania	31.6 32	Mississippi	~~64.7~~
Rhode Island	35.6 36	North Carolina	~~52.2~~ X
Vermont	~~24.2~~ 24	South Carolina	~~53.0~~ X
		Tennessee	~~54.3~~ X
		Virginia	~~37.6~~ X
		West Virginia	~~45.5~~ X

35.4
37.6
44.5
45.5
46.3
51.0
52.2
53.2
54.3
54.5
55.7
64.7

23: 23.3
24: 24.2

30: 29.5
32: 31.6

a) A portion of a stemplot comparing the two regions is given below. Please complete the table by adding the data representing the Northeastern states following the rules used for the Southern states.

Northeastern (NE)		Southern (SO)
430	2	
765	2	
20	3	
6	3	58
	4	
	4	566
	5	1234
	5	56
	6	
	6	5

b) Write down the 5-number summary for the Northeastern states by inspecting the data and stemplot (not using SPSS).

5-number summary:

$Q_1 = \frac{23.3 + 24.2}{2}$ $Q_3 = \frac{29.5 + 31.6}{2}$

NE:

min = 20.0 Q_1 = 23.75 median = 25.8 Q_3 = 30.55

max = 35.6

Question 4) continued...

c) For the Northeastern States, find the "upper fence" used in outlier detection (1.5xIQR rule) and say whether you find any states that are outliers because of extraordinarily high teen birth rates.

$IQR = 30.55 - 23.75 = 6.8$

$6.8 \times 1.5 = 10.2$

upper fence $= Q_3 + 10.2 = 30.55 + 10.05 = 40.75$

There are no outliers for NE because no values are above 40.75 births/1000 teen females.

d) Explain why a stemplot is preferable to a histogram for comparing the birth rates of the two regions for this data set.

A stemplot is preferable to a histogram with this data set because it is a small set and a stem plot shows each value and more detail in this case.

e) How are teen birth rates related to region? Give a complete answer as illustrated in class, using **the 4-step process** in reporting your work. You may, however, explicitly reference earlier parts of this problem, where appropriate, to support your answer.

(1.) State: How are teen birth rates related to region?

(2.) Formulate: We need to compare the distributions by examining graphical + numerical summaries of the data, including appropriate measures of center + spread.

(3.) Solve:

* see stemplots

rounded

	NE	SO
min	20	35
Q_1	23.5	45.5
$\tilde{x}$	26	51.5
Q_3	31	54.5
Max	36	65
$\bar{x}$	26.911	49.558
s	4.6887	8.156
IQR	7.5	9

(4.) Conclude

Both the stemplot graphs & the numerical summaries show that there is a higher instance of teen births in the southern states than there are in the northern states. The median # of births per 1000 females aged 15-19 was 26 in the NE states, a # that is $\times 50.5\%$ lower than the median of the 50 states, 54.5. The mean # of births in the NE states, 26.9, is also 54.2% lower than the mean # of births in the 50 states which was 249.6. Therefore there is a lower instance of teen births in NE states than in 50 states.