Statistical Reasoning in Public Health: Homework #1

1. (7 points total, 1 point for each of a-g)

   The duration of time from first exposure to HIV infection to AIDS diagnosis is called the *incubation period.* The incubation periods of a random sample of 7 HIV infected individuals is given below (in years):

12.0	10.5	13.5	12.5
9.5	6.3	7.2	

   a. Calculate the sample mean. (1 pt)
   b. Calculate the sample median. (1 pt)
   c. Calculate the sample standard deviation (1pt).
   d. If the number 6.3 above were changed to 1.5, what would happen to the sample mean, median, and standard deviation? State whether each would increase, decrease, or remain the same. (1 pt)
   e. Assume that these data a seven random observations taken from a larger population whose values are normally distributed. (even if this assumption makes little sense) Using this assumption, coupled with prior computations, estimate an interval that contains 95% of incubation time values in the population of HIV patients from which the sample was taken. (1 pt)
   f. Suppose another random sample of 100 persons is taken from the same population. How will the sample mean, median and standard deviation values compare to those you reported in sections a- c? (1 pt)
   g. Suppose the distribution of these incubation period values is left-skewed in the population of persons with HIV. If you were take single random samples of each of the following size, what will likely be the shape of the distribution of sample values?
      1. n=75
      2. n=200
      3. n=3,200

2. (4 points total: 1 point for each of a – d)
   In a random survey of 3,015 boys age 11, the average height was 146 cm, and the standard deviation (SD) was 8 cm. A histogram suggested the heights were approximately normally distributed. Fill in the blanks.
   a. One boy was 170 cm tall. He was above average by __________ SDs.
   b. Another boy was 148 cm tall. He was above average by __________ SDs.
   c. A third boy was 1.5 SDs below the average height. He was __________ cm tall.
   d. If a boy was within 2.25 SDs of average height, the shortest he could have been is __________ cm and the tallest is __________ cm.

3. (8 points total: 1 point each for parts 1 and 2 of sections a-d)

Assume blood-glucose levels in a population of adult women are normally distributed with mean 90 mg/dL and standard deviation 13 mg/dL.

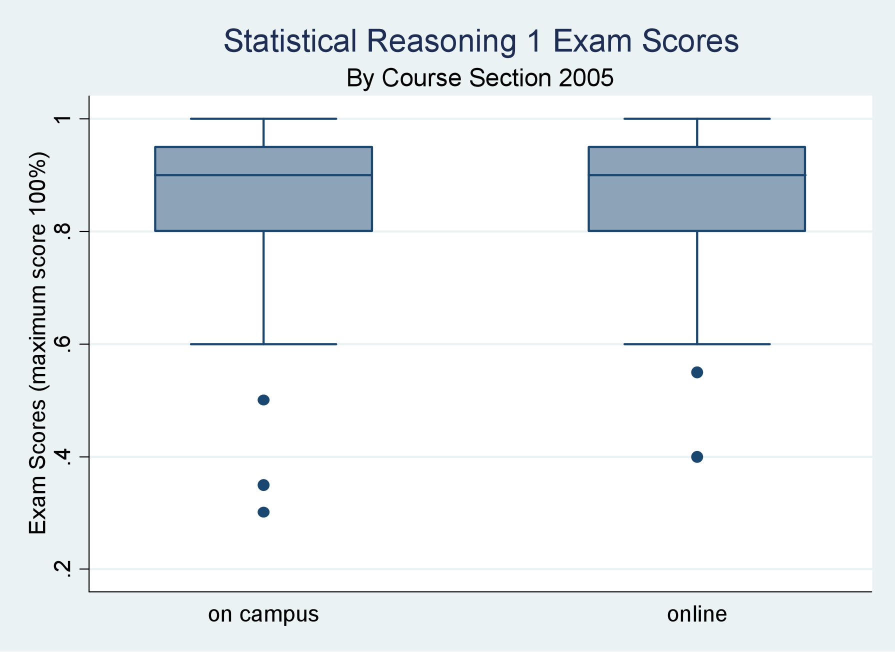
For parts a-d, answer each of the following:

1. What percentage of individuals would be called "abnormal" and need to be retested?
2. What is the normal range of glucose levels in units of mg/dL?

a. Suppose the "abnormal range" were defined to be glucose levels outside of 1 standard deviation of the mean (i.e., either at least 1 standard deviation above the mean, or at least 1 standard deviation below mean). Individuals with abnormal levels will be retested.
b. Suppose the "abnormal range" were defined to be glucose levels outside of 1.5 standard deviations of the mean.
c. Suppose the "abnormal range" were defined to be glucose levels outside of 2 standard deviations of the mean.
d. Suppose the "abnormal range" were defined to be glucose levels outside of 2.5 standard deviations of the mean.

4. (4 points total:1 point for each section a – d)
Below find side-by-side boxplots comparing the 140.611 final exam scores between the online and on-campus versions of the course in academic year (AY) 2005-06.

\



a. What is your overall impression of the relative performance of the two sections based on this boxplot presentation?
b. What (roughly) was the median exam score in each of the two sections?
c. What was the highest score obtained by students in the on-campus section?

d. What type of distribution shape do the exam scores have in each of the sections?

5. (4 points total: 1 point for each of sections a –d)
Below find information about the 1987 medical expenditures of adult (≥ 40 years) U.S. residents by sex (among those who spent at least 1.00 US dollar on medical related items or services). This data comes from the 1987 National Medical Expenditures Survey (joint project of NCHS and HCFA – now CMS).

mean	: $2640	mean	: $2544
sd	: $5160	sd	: $5243
median	: $787	median	: $673

a. Which of the following three terms best describes the basic shape of the expenditure distributions for both males and females – symmetric, right-skewed, left-skewed? Support your visual assessment with numerical justification.
b. Can you suggest a substantive reason for the resulting shapes of these expenditure distributions for both males and females?
c. Assume for the moment that these data are normally distributed (even if this assumption seems silly) for females. Using this assumption and the sample mean and standard deviation estimate an interval that contains 95% of the observed medical expenditure values among all U.S. females 40 years or older in 1987.
d. Is the interval estimate in part (c) reasonable given the nature of this data?

5. (12 points total: 1 point for sections a-c for each article)

Look through several research journals of *interest* to you. Choose **four articles** of your choice published in the past three years that report study results that involved at least some data analysis (some possible journals are *New England Journal of Medicine, Lancet American Journal of Public Health, British Medical Journal, American Journal of Epidemiology,* etc.). For each article, answer the following:

a. Give the citation (author, title, journal date, page, etc.)
b. In 1–2 sentences, what is the main study question being addressed?
c. Are the statistical methods described (this is usually in the methods section)? If so, briefly *list* the type of statistical methods used (e.g.: 2-sample t-test; paired t-test; Fisher's exact test; multiple linear regression, logistic regression; proportional hazards analysis; analysis of variance; chi-square test, etc.)

Sample Quiz Questions (NOTE: these are a required part of the homework): *Choose the correct answer for the following multiple-choice questions, and* ***justify (explain)*** *your choice: (6 points total: 1 point for providing a correct answer, 1 point for a correct justification)*

6. A sample of 5 body weights (in pounds) is as follows: 116, 168, 124, 132, 110. The *sample median* is:
   a. 124.
   b. 116.
   c. 132.
   d. 130.
   e. None of the above.

7. Suppose a random sample of 100 12-year-old boys were chosen and the heights of these 100 boys recorded. The sample mean height is 64 inches, and the sample standard deviation is 5 inches. You may assume heights of 12-year-old boys are normally distributed. Which interval below includes approximately 95% of the heights of 12-year-old boys?

   a. 63 to 65 inches.
   b. 39 to 89 inches.
   c. 54 to 74 inches.
   d. 59 to 69 inches.
   e. Cannot be determined from the information given.
   f. Can be determined from the information given, but none of the above choices is correct.

8. Cholesterol levels are measured on a random sample of 1,000 persons, and the sample standard deviation is calculated. Suppose a second survey were repeated in the same population, but the sample size tripled to 3,000. Then which of the following is true?

    a. The new sample standard deviation would tend to be smaller than the first and approximately about one-third the size.
    b. The new sample standard deviation would tend to be larger than the first and approximately about three times the size.
    c. The new sample standard deviation would tend to be larger than the first, but we cannot approximate by how much.
    d. None of the above is true because there is no reason to believe one sample standard deviation would tend to be systematically larger or smaller than the other.