PHP 2500 Introduction to Biostatistics

Problem Set Four

Due: Tuesday November 13th, at the beginning of class

Please show all your work.

1. Pagano #20, (Chapter 7, p194)
2. Pagano #8, (Chapter 8, p211)
3. Pagano #9, (Chapter 8, p211)
4. Pagano #11, (Chapter 8, p212)
5. Pagano #12, (Chapter 8, p212)
6. Pagano #15, (Chapter 8, p212)

7. Sixty percent of the employees of a large heath care system were absent due to sickness for three or more days last year. Assume that the absentee rate due to illness will be the same this year and that we select a random sample of 150 employees.

   (a) What is the probability that the sample proportion of employees absent three or more days will be between 0.5 and 0.65?

   (b) What is the probability that the sample proportion will exceed 0.7?

   (c) What is the expected number in the sample who will be absent three or more days due to illness?

8. The average rate of gun deaths in providence is 8.7 per month. Suppose we track the number of gun deaths by month over a two-year period.

    (a) What is the probability that the average gun death rate over the two-year period will be between 7.5 and 10 per month?

    (b) What is the probability that the sample average will exceed 12 deaths per month in just one year?

    (d) What is the total expected number of gun deaths over the two year period?

9. The Study habits portion of the survey of Study Habits and Attitudes (SSHA) psychological test consists of two sets of questions. One set of questions measures “delay avoidance” (procrastination) and the other measures “work methods”. A subject’s study habits score is the sum X+Y of the delay avoidance score X and the work methods score Y. The distribution of X in a broad population of students at Schools of Public Health is N(25,10) and the distribution of Y in the same population is N(30,9).

    (a) If a subject’s X and Y scores are independent, what is the distribution of the study habits score X+Y?

    (b) Assuming the scores are independent, what percentage of the population has a study habits score of 60 or higher?

    (c) In fact, the X and Y score are strongly correlated with a correlation of 0.7. In this case what is the effect on the expected value and variance you found in (a)?

    (d) Assuming the study habits score has a normal distribution with the mean and variance found in (c), what is the proportion in the population with a study habits score exceeding 60?

    (e) Repeat (d) with a correlation of –0.7.

10. Two investigators independently estimate the mean of a population. There estimates are $\overline{X}$ (std.err. =0.3) and $\overline{Y}$ (se=1.2) respectively. As the project statistician you are asked to combine the two estimates to obtain an overall estimate (this is one of the problems addressed by Meta-Analysis which is more appropriately called combining information). Three alternatives proposed by colleagues are:
    (1) Simply use $\overline{X}$ because it has the smallest std.err.
    (2) Average the two estimates ($\overline{X}+\overline{Y}$)/2
    (3) Use 0.94 $\overline{X}$ + 0.06 $\overline{Y}$ because it came to you in a dream:

$$\frac{(1.2)^2}{(0.3)^2+(1.2)^2}\overline{X}+\frac{(0.3)^2}{(0.3)^2+(1.2)^2}\overline{Y}=0.94\overline{X}+0.06\overline{Y}$$

(We will discuss this estimate in more detail later.)

    (a) Show that each of the three proposed estimates has expected value equal to μ, the mean of the population.

    (b) Find the variance of each proposed estimate. Which is preferred in the sense of having the smallest variance?

11. Anatomy of a box plot: suppose that X is normally distributed, having mean μ and variance σ^2.
    (a) Show that the inter-quartile range of X is 1.34σ.
    The inter-quartile range is the middle 50% of the distribution, defined in class as 75 centile – 25 centile. That is IQR=B-A where B and A are defined by:

$$P(X \le A) = 0.25 \quad \text{and} \quad P(X \le B) = 0.75$$

    (b) Using the result in (a) find the probability of exceeding the “upper fence” defined by

$$upf = B + \frac{3}{2}(1.34\,\sigma)$$

    (c) Using the result in (a) find the probability of exceeding the “lower fence” defined by

$$lof = A - \frac{3}{2}(1.34\,\sigma)$$

    (d) Using the result of (b) and (c) explain why observation beyond the “fences” may be considered as exceptional or unusual values.