

# Medical Statistics 2<sup>nd</sup> Semester Take Home Final Exam

**Due Date: Dec 19 2023 (13:00) to Dec 23 2023 (23:59)**

Name: \_\_\_\_\_

## Notice

- Please **DO SOLVE ANSWERS BY YOURSELVES!!**
- You can use materials from other textbooks, lecture notes, and websites but you have to provide proper **CITATIONS**.
- Write down your answers in the **MS Word Document** and save your word filename like [student number-name.docx, i.e. 202015015-boncho-ku.docx]. it is admittable to submit your answers by converting your word file to pdf.
- If you make up your mind to submit your answers, send an E-mail to Dr. Boncho Ku and Dr. Mimi Ko with the attachment of your answer file.

## Questions

1. Which one of the following statements is False?
  - a) The probability of a type I error is the probability that you reject the null hypothesis when it is true.
  - b) Subjects are enrolled or grouped on the basis of their exposure, then are followed to document occurrence of disease in prospective cohort study.
  - c) Especially when more than 20% of cells have expected frequencies  $< 5$ , we need to use Fisher's exact test to determine if there are associations between two categorical variables.
  - d) To use the two-sample t-test, we need to assume that the data from both samples are normally distributed and they have the same variances.
  - e) In test for heterogeneity of meta-analysis, if Higgins  $I^2 > 75\%$ , studies are regarded homogeneous and the fixed effect model of meta-analysis can generally be used.

2. Solve following problems.

- a. Let  $X$  be a random variable having binomial distribution with parameters  $n = 25$  and  $p = 0.2$ . Evaluate  $P(X, \mu_x - 2\sigma_X)$ .
- b. If  $X$  is a random variable with satisfying  $P(X = 0) = P(X = 1)$ , what is  $E(X)$ ?
- c. If  $X$  is normally distributed mean 2 and variance 1, find  $P(|X - 2| < 1)$ .
- d. Suppose  $X$  has a binomial distribution with parameters  $n$  and  $p$ . For what  $p$  is  $\text{var}(X)$  maximized if we assumed  $n$  is fixed?

3. Let  $X$  be the maximum usage time (hour) of an A-manufactured cellular phone after a full charge. Assume that  $X$  is normally distributed with mean 70 (hours) and variance  $\sigma^2$ . If a purchaser of such cellular phones requires that at least 90 percent of the cellular phones are available to use exceeding 60 hours, what is the largest value of  $\sigma$  can be and still have the purchaser satisfied?

5. Solve the following problems:
  - a. Let  $X_1$ ,  $X_2$  and  $X_3$  be uncorrelated random variables with common variance  $\sigma^2$ . Find the correlation coefficient between  $X_1 + X_2$  and  $X_2 + X_3$
  - b. Let  $X_1$  and  $X_2$  be uncorrelated random variables. Find the correlation coefficient between  $X_1 + X_2$  and  $X_2 - X_1$  in terms of  $\text{Var}(X_1)$  and  $\text{Var}(X_2)$ .
  - c. Let  $X_1$ ,  $X_2$ , and  $X_3$  be independently distributed random variables with common mean  $\mu$  and common variance  $\sigma^2$ . Find the correlation coefficient between  $X_2 - X_1$  and  $X_3 - X_1$ .
  
6. When you start R with Rstudio, there is an example dataset named with `mtcars`. The `mtcars` dataset was extracted from the 1974 *Motor Trend* US magazine, and comprise fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973 - 74 models). The detailed description of variables in `mtcars` dataset can be checked by typing `help(mtcars)` in the prompt of R console window. Write R scripts and confirm the results for the following questions.
  - a. Extract `mpg` and `disp` variables from `mtcars` dataset and restore it in an object `x` and `y`, respectively.
  - b. Calculate mean, standard deviation, coefficient of variation, minimum, maximum, median, 25<sup>th</sup> and 75<sup>th</sup> quantiles, and interquartile range of `x` and `y`.
  - c. Make scatterplot of `x` and `y` and interpret in terms of the correlation coefficient between `x` and `y`.
  - d. Assume that `x` is the population of a mile per gallon of all automobiles of US from 1973 to 1974. Suppose a sample of size 2 automobiles are drawn from the population with replacement and calculate sample mean. Then repeat the same procedure 10,000 times (Hint: check the function `sample()`).
    - Make histogram of 10,000 sample means.
    - Calculate the mean and standard deviation of 10,000 sample means.
    - Compare the above results to the population in terms of mean and standard deviation: is the mean of 10,000 sample mean is approximate to the population mean? In what proportion did the standard deviation of the sample mean decrease compared to the standard deviation of the population?
  
7. There are five urns, and they are numbered 1 to 5. Each urn contains 10 balls. Urn  $i$  has  $i$  defective balls and  $10 - i$  non-defective balls,  $i = 1, 2, \dots, 5$ . For example, urn 3 has three defective balls and seven non-defective balls. Consider the following random experiment: First an urn is selected at random, and then a ball is selected at random from the selected urn. Suppose that the experimenter does not know which urn was selected. Let's ask two questions.
  - a) What is the probability that a defective ball will be selected?
  - b) If we have already selected the ball and noted that it is defective, what is the probability that it came from urn 5?

8. The distribution of grades in a large statistics course is as follows:

Grade:	A	B	C	D	F
Probability	0.1	0.4	0.3	0.1	0.1

To calculate student grade point averages, grades are expressed in a numerical scale with A=4, B=3, and so on down to F=0.

- Find the expected value.
  - Describe your strategy to **simulate** choosing students at random and recording their grades.
  - Based on your strategy described in b), perform the simulation with a sample size of size 30 and calculate the mean of their 30 grades **using R**.
  - Repeat c) 10,000 times and calculate the average of 10,000 means.
  - Make a histogram of 10,000 means.
  - Describe your conclusion based on the results of a) to e).
4. Describe the relationship between Pearson's correlation coefficients and the regression coefficient (slope) of univariate regression analysis. You're definitely able to refer to the Internet or other textbooks but you have to give descriptions or formulas in your own words and cite appropriately.