Homework 1

1. Daily sales for 21 shops are shown as follows:

   8408 1374 1872 8879 2459 11413 608

   14138 6452 1850 2818 1356 10498

   7478 4019 4341 739 2127 3653 5794 8305

   (a). Compute the skewness and coefficient of variation.

   (b). Provide a five-number summary.

   (c). Show a boxplot.

   Option: You can use data analytics tools (Excel/R/Python) to answer the questions above and paste the screenshot.

2. Consider a finite population with five elements labeled A, B, C, D, and E. **Simple random sampling** is used to select 2 elements from population.

   (a). What is the probability that each sample of size 2 is selected?

   (b). Assume random number 1 corresponds to A, random number 2 corresponds to B, and so on. List the sample that will be selected by using the following random digits sequence:

   8 0 5 7 5 3 2

3. Consider a population with 100 students. From the aspect of gender, the population consists of 40 female students and 60 male students. From the aspect of classes, the population consists of 5 classes with 20 students in each class. Suppose we want to sample 40 students from the population.

   (a). If the sampling method is stratified sampling, how do we divide the population into several strata and how many students should we sample in each strata?

   (b). If the sampling method is cluster sampling, how do we divide the population into several clusters and how many clusters should we sample?

   (Hint: From aspect of gender or classes, the population can be divided)

4. A population has a mean of 200 and a standard deviation of 50. A sample of size 100 will be taken and the sample mean $\bar{x}$ will be used to estimate the population

mean.

(a). What is the expected value of $\bar{x}$?

(b). What is the standard deviation of $\bar{x}$?

(c). Show the sampling distribution of $\bar{x}$.

(d). What is the probability that $\bar{x}$ will be within ± 5 of population mean?

5. Forty-two percent of primary care doctors think their patients receive unnecessary medical care. Suppose the population of primary care doctors is sufficiently large.

   (a). A sample of 300 primary care doctors were taken. Show the sampling distribution of the proportion of the doctors who think their patients receive unnecessary medical care.

   (b). What is the probability that the sample proportion will be within ± 0.05 of the population proportion?

   (c). What would be the effect of taking a larger sample on the probability in part (b)? Why?

6. Multiple Choice Questions.

   (a). A 95% confidence interval for the population mean μ is computed from a random sample and found to be (6, 12). Which of the following is a correct statement?

   A. There is a 95% probability that μ is between 6 and 12.

   B. 95% of values sampled are between 6 and 12.

   C. If we took many, many additional random samples and from each computed a 95% confidence interval for μ, approximately 95% of these intervals would contain μ.

   D. There is a 95% probability that the true mean is 9 and a 95% chance that the true margin of error is 3.

   (b). A professor sampled 46 students from a large university to obtain a 95% confidence interval for the proportion of students in favor of raising ASB fees. The interval was (0.356, 0.397). If the professor had used a 90% confidence interval instead, the confidence interval would have been

   A. Wider and would have a smaller chance of missing the true proportion.

B. Narrower and would have a larger chance of missing the true proportion.

C. Narrower and would have a smaller chance of missing the true proportion.

D. Wider and would have a larger chance of missing the true proportion.

E. Wider, but the chance of missing the true proportion cannot be determined.

(c). The following is a 95% confidence interval for p: (0.28,0.52). How large was the sample used to construct this interval?

A. n = 28

B. n = 77

C. n = 64

D. n = 34

E. n = 91

7. The following annual premiums for automobile insurance are from a representative sample in this city.

   1905 3112 2312 2725 2545 2981 2677 2525 2627 2600 2370 2857 2962 2545 2675 2184 2529 2115 2332 2442

   Assume the population is approximately normal.

   (a). Provide a point estimate of the mean annual automobile insurance premium in this city.

   (b). If the standard deviation of the annual premiums in this city is 200, develop a 95% confidence interval for the mean annual automobile insurance premium in this city.

   (c). If the standard deviation of the annual premiums in this city is unknown, develop a 95% confidence interval for the mean annual automobile insurance premium in this city.

   (d). If you want to conduct a new survey to estimate the mean annual automobile insurance premium in this city, how large the new sample should be taken with a margin of error of 150? Use 95% confidence level.

   Option: You can use data analytics tools (Excel/R/Python) to answer the questions above and paste the screenshot.

8. The Centers for Disease Control reported the percentage of smoked people among

people which is 18 years old or older. Suppose that a study designed to collect new data on smokers and nonsmokers uses a preliminary estimate of the proportion who smoke of 0.30.

(a). How large a sample should be taken to estimate the proportion of smokers in the population with a margin of error of 0.02? Use 95% confidence level.

(b). Assume that the study uses your sample size recommendation in part (a) and finds 520 smokers. What is the point estimate of the proportion of smokers in the population?

(c). What is the 95% confidence interval for the proportion of smokers in the population?

9. One of the questions in a survey asked adults if they used the Internet at least occasionally. The results showed that 454 out of 478 adults aged 18–29 answered Yes; 741 out of 833 adults aged 30–49 answered Yes; 1058 out of 1644 adults aged 50 and over answered Yes.

   (a). Develop a point estimate of the proportion of adults aged 18–29 who use the Internet.

   (b). Develop a point estimate of the proportion of adults aged 30–49 who use the Internet.

   (c). Develop a point estimate of the proportion of adults aged 50 and over who use the Internet.

   (d). Comment on any relationship between age and Internet use that seems apparent.

   (e). Suppose your target population of interest is that of all adults (18 years of age and over). Develop an estimate of the proportion of that population who use the Internet.

10. Consider a sample of random variables: X_1, X_2, ..., X_n, where $n > 10$, $E[X_i] = \mu$, $Var[X_i] = \sigma^2 > 0$ and the estimator: $\widehat{\mu_n} = \frac{1}{n-10} \sum_{i=11}^{n} X_i$. [Hint: note that the sum is over $(n - 10)$ random variables].

    (a). Calculate the bias of $\widehat{\mu_n}$.

    (b). Calculate the variance of $\widehat{\mu_n}$.

    (c). Is $\widehat{\mu_n}$ efficient in a finite sample? If not, find another unbiased estimator of μ

which has greater relative efficiency than $\widehat{\mu_n}$.

11. Let X_1, X_2, ..., X_n be n independent observations from a population with mean μ and variance σ^2. Prove that the sample variance S^2 is an unbiased estimator of the population variance σ^2.
(Hint: $Var(X) = E(X^2) - [E(X)]^2$)

12. **Bonus**
Consider the case of infinite population, prove that when the sampling method is simple random sampling and the sample size is large, sample proportion $\bar{p} = \dfrac{x}{n}$ can be approximated well by normal distribution, where x is the number of elements in the sample that possess the characteristic of interest and n is the sample size.
(Hint: central limit theorem for random variable x)