STAT 135 CONCEPTS OF STATISTICS
HOMEWORK 1

Assigned June 22, 2021, due June 29, 2021

This homework pertains to materials covered in Lecture 1 and 2. The assignment can be typed or handwritten, with your name on the document, and with properly labeled computer output for those problems that require it. To obtain full credit, please write clearly and show your reasoning. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the Week 1 Assignment on bCourses.

Problem 1. Let $x_1, x_2, \ldots, x_N$ be a list of numbers with mean μ and standard deviation σ (the square root of the variance σ^2). Assume $a_1, \ldots, a_M$ are the distinct values of that list with frequencies $n_1, \ldots, n_M$ respectively. Show that

$$\sigma^2 = \sum_{i=1}^{M} \frac{n_i a_i^2}{N} - \mu^2.$$

Problem 2. During Lecture 1, we saw two samples randomly drawn from the entire population of Youtube users with two different mobile app designs. For each sample, two variables were collected - Session view duration (hours) and Internet speed (Mbps). The scatter plot of the samples is shown again in Figure 1.

(1) This data set may include multiple sessions from the same user during the testing period. Are the samples from Design A (blue points) i.i.d? If not, explain how the view hours of different sessions from the same user might be correlated with each other.

(2) A data analyst intern did not fully randomize the sampling of the users: Design A was mostly shown to accounts registered in rural areas, while Design B was mostly shown to accounts registered in the Bay area. As a result, he got more samples in red with higher Internet speed than samples in blue. Is it still appropriate to analyze this data set to decide which design is better?

Problem 3. A class has two sections. Students in Section 1 have an average score of 80 with an SD of 10. Students in section 2 have an average score of 87 with an SD of 10.

(1) If possible, say whether the SD of the scores of all the students in the class is
   (a) less than 10
   (b) equal to 10
   (c) greater than 10

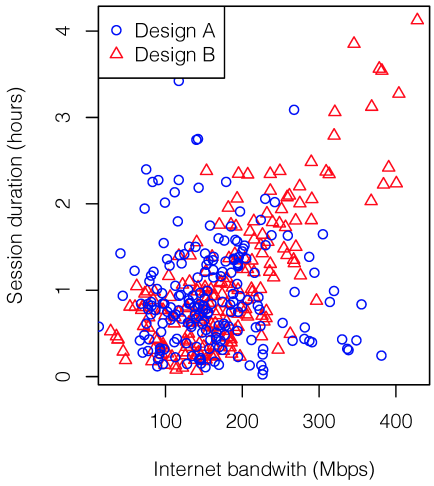

FIGURE 1. Samples from A/B testing

Explain your choice. If it is not possible to make the choice with the information given, explain why not.

(2) Suppose section 1 has 30 students and section 2 has 20. Find the SD of the scores of all the students in the class.

Problem 4. Let X have the distribution given below, in which $0 < \theta < 1/6$:

value	1	2	3	4
probability	2θ	θ	3θ	$1-6\theta$

Let $X_1, \ldots, X_n$ be i.i.d samples drawn from the distribution above, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean.

(1) Classify each of the following quantities as a random variable or a real number.
   (a) X
   (b) $E(X)$
   (c) $X_1 * X_n$
   (d) $\bar{X}_n$

(2) Decide whether each of the following statements is true or false. Justify your answer.
   (a) $\bar{X}_n = E(X)$ for any sample size $n \geq 1$.
   (b) $E(\bar{X}_n) = E(X)$ for any sample size $n \geq 1$.
   (c) If n is sufficiently large, $\bar{X}_n$ is likely to be exactly equal to $E(X)$.
   (d) If n is sufficiently large, $\bar{X}_n$ can be very close to $E(X)$.

(3) If $n = 2$, write out the joint distribution of the two independent samples (X_1, X_2). Also, calculate $P\left(\frac{X_1+X_2}{2} \geq 3\right)$.

(4) Can you come up with an unbiased estimator of the population parameter θ?

Problem 5. Suppose we play the following guessing game. You pick a number from the list $x_1, x_2, \ldots, x_N$, and ask me to guess what it is. Each number from the list has an equal chance of being picked. My strategy is to guess that the value is some constant c no matter what you pick. Thus if x_i is what you pick, the amount of error that I make with my strategy is $x_i - c$. Define the mean squared error of my strategy to be

$$mse_c = \frac{1}{N}\sum_{i=1}^{N}(x_i - c)^2.$$

Show that the value of c that minimizes mse_c is $c = \mu$, and that $mse_\mu = \sigma^2$.

Problem 6. Suppose $X_1, \ldots, X_n$ are i.i.d samples from a population $f(x)$ with mean μ and variance σ^2. Denote $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as the sample mean.

(1) Prove that

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

is an unbiased estimator of σ^2.

(2) Denote $\hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2}$. Then

$$\hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$$

Is $\hat{\sigma}_n$ an unbiased estimator of σ? If not, does $\hat{\sigma}_n$ tend to underestimate or overestimate σ?
(*Hint*: Use the formula $\mathrm{Var}(\hat{\sigma}_n) = E(\hat{\sigma}_n^2) - [E(\hat{\sigma}_n)]^2$.)

Problem 7. Two surveys were independently conducted to estimate a population mean μ. Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are the two samples obtained that are both i.i.d from the same population. Denote $\bar{X}_n$ and $\bar{Y}_m$ as the sample means. For some real numbers α and β, the two sample means can be combined to give a better estimator:

$$\hat{\mu}_{m+n} = \alpha\bar{X}_n + \beta\bar{Y}_m.$$

(1) Find the conditions on α and β that make the combined estimate unbiased.

(2) What choice of α and β minimizes the variance of $\hat{\mu}_{m+n}$, subject to the condition of unbiasedness?

Problem 8. A sample of size $n = 200$ is taken from a population that has a proportion $p = 1/2$. Denote $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as the sample proportion.

(1) Find δ such that $P(|\hat{p}_n - p| \geq \delta) = 0.025$.

(2) If, in the sample, $\hat{p}_n = 0.25$, will the 95% confidence interval for p contain the true value of p?

Problem 9. A coin lands heads with probability p. It is tossed 400 times. Compute the bootstrap 95% confidence interval for p and compare with the corresponding interval which uses a conservative estimate for the standard error, when the observed number of heads is

(1) 280,
(2) 150.

Problem 10. Given a standard Normal random variable $Z \sim N(0,1)$, define $T = Z^2$. The new T is called a χ^2 distributed random variable. The probability density function of T can be computed as

$$f(t) = \begin{cases} \frac{1}{\sqrt{2\pi t}} e^{-\frac{t}{2}}, & \text{if } t \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

If we compare this with the general form of an Gamma(α, β) density function (see this wikipedia page), we will recognize that the χ^2 probability density function is, in fact, a Gamma$(\frac{1}{2}, \frac{1}{2})$.

(1) Compute $E(T)$ and $\mathrm{Var}(T)$ using $f(t)$.
(2) If $Z_1, Z_2, \ldots, Z_n$ are i.i.d standard normal random variables, define $T_n = Z_1^2 + Z_2^2 + \ldots + Z_n^2$. Then we call T_n a χ^2_n random variable (*chi-square with n degrees of freedom*). What are $E(T_n)$ and $\mathrm{Var}(T_n)$?
(3) Prove that the probability density function of T_n is the same as Gamma$(\frac{n}{2}, \frac{1}{2})$, or equivalently $T_n \sim$ Gamma$(\frac{n}{2}, \frac{1}{2})$.
(*Hint*: If $Y_1 \sim$ Gamma(α_1, β), $Y_2 \sim$ Gamma(α_2, β), and Y_1 and Y_2 are independent, then $Y_1 + Y_2 \sim$ Gamma$(\alpha_1 + \alpha_2, \beta)$.)
(4) Given two independent random variables $T_n \sim \chi^2_n$ and $S_m \sim \chi^2_m$, prove $T_n + S_m \sim \chi^2_{m+n}$.
(5) Use the central limit theorem to derive a good approximate normal distribution for T_n.