STAT 2002 - Probability and Statistics II, Summer 2024
Homework 2 - Point Estimation and Confidence Intervals
100 points total.

This homework is due **Beijing Time 11:59pm Tuesday, June 25, 2024** on BlackBoard. No late homework accepted.

Please make sure to **SHOW ALL WORK** in order to receive full credit.

1. (20 points) Suppose that the expectations of three random variables are equal ($\mathbb{E}(X_1) = \mathbb{E}(X_2) = \mathbb{E}(X_3) = \mu$), and their variances are $Var(X_1) = 7$, $Var(X_2) = 13$, and $Var(X_3) = 20$. Consider the point estimates for parameter μ:

$$\widehat{\mu}_1 = \frac{X_1}{3} + \frac{X_2}{3} + \frac{X_3}{3}, \qquad \widehat{\mu}_2 = \frac{X_1}{4} + \frac{X_2}{3} + \frac{X_3}{5}$$

   (a) Calculate the bias of each point estimate. Is any one of them unbiased?

   (b) Calculate the variance of each point estimate. Which one has the smallest variance?

   (c) Calculates the mean square error (MSE) of each point estimate. Which point estimate has the smallest mean square error for $\mu = 3$?

2. (10 points) Find the maximum likelihood estimator of the unknown parameter θ where $X_1, \ldots, X_n$ is a sample from the distribution whose density function is

$$f(x) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

3. (20 points) This question is about (a simplified version of) the *Gaussian Mixture Model (GMM)*, which is a popular model in statistics, data science and machine learning. Suppose that K is a discrete random variable that can either be 0 or 1 with probability π_0 and π_1 respectively, that is,

$$K = \begin{cases} 0 & \text{with probability } \pi_0 = P(K = 0) \\ 1 & \text{with probability } \pi_1 = P(K = 1) = 1 - \pi_0 \end{cases}$$

   Conditional on $K = k$ for $k \in \{0, 1\}$, the distribution of X is $N(\mu_k, \sigma_k^2)$, a normal distribution with mean μ_k and variance σ_k^2. That is

$$X|K = 0 \sim N(\mu_0, \sigma_0^2),$$

$$X|K = 1 \sim N(\mu_1, \sigma_1^2).$$

   (a) Derive the joint density of (X, K). State clearly the support of (X, K) in the joint density. Hint: consider conditional distribution and the law of total probability.

(b) Denote the distribution of $(X, K) \sim GMM(\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$. Note that π_1 can be omitted as a parameter since $\pi_1 = 1 - \pi_0$. Suppose that we have an i.i.d. random sample of size n of these n pairs $(X_1, K_1), (X_2, K_2), \ldots, (X_n, K_n)$. Each X_i belongs to either group 0 or group 1 depending on K_i. Using part (a), derive the maximum likelihood estimator for all the five parameters $\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2$. Hint: Let $n_0 = \sum_{i=1}^n \mathbf{1}\{K_i = 0\}$ and $n_1 = \sum_{i=1}^n \mathbf{1}\{K_i = 1\}$ be the number of X_i that belongs to group 0 and group 1 respectively. You may find expressing the likelihood function in terms of n_0 and n_1 useful.

4. (20 points) The data shown below describe temperatures (degrees Celsius) for wheat grown at Harper Adams Agricultural College in Junes between 1982 and 1993.

   15.2 14.2 14.3 14.2 14.0 13.5 12.2 11.8 14.4 12.5 15.2

   Assume the data are random samples from a normal distribution, with the standard deviation known to be $\sigma = 0.5$.

   (a) Construct a 99% two-sided confidence interval on the mean temperature.

   (b) Construct a 95% lower-confidence bound on the mean temperature.

   (c) Suppose that we wanted to be 95% confident that the error in estimating the mean temperature is less than 2 degrees Celsius. What sample size should be used?

5. (10 points) A healthcare provider monitors the number of CAT scans performed each month in each of its clinics. The most recent year of data for a particular clinics are as follows (the reported variable is the number of CT scans each month expressed as the number of CT scans per thousand members of the health plan):

   2.31, 2.09, 2.36, 1.95, 1.98, 2.25, 2.16, 2.07, 1.88, 1.94, 1.97, 2.02.

   Find a two sided 95% confidence interval for the standard deviation. (Hint: you may find the chi-square table in textbook or find the values you need using R.)

6. (10 points.) A *CNN/ORC Poll* (`http://www.pollingreport.com/drugs.htm`) conducted in Jan. 2014, asked the following question: "Do you think the use of marijuana should be made legal, or not?". Based on the poll's results (for July 15-17, 2019) shown below, calculate a 95% confidence interval for p, the proportion of all American adults who *oppose* the legalization, and interpret your interval in context. (Hint: One way to deal with the unsure votes is to combine them with the ones who think legalization is a good idea, thus making the vote to have only 2 options: oppose the legalization, and others.)

7. (10 points) R practice. Suppose that an experimenter observes a set of variables that are taken to be normally distributed with an unknown mean and variance. Using simulation methods, for given values of the mean and variance, we can simulate the data values that the experimenter might obtain. More interestingly, we can simulate lots of possible samples of which, in reality, the experimenter would observe only one. Performing this simulation allows us to check on sampling distributions of the parameter estimates.

NPR/PBS NewsHour/Marist Poll. July 15-17, 2019. N=1,346 adults nationwide. Margin of error ± 3.5.

"Do you think legalizing marijuana nationally is a good idea or a bad idea?"

	A good idea	A bad idea	Unsure
	%	%	%
7/15-17/19	63	32	5

Let us assume that $\mu = 100$ and $\sigma^2 = 9$, which, in fact, the experimenter does not know. In our simulation study, we assume that the experimenter will observe 100 observations, which are normally distributed. To simulate a sample of 100 observations from $N(100, 9)$, which the experimenter might observe, the R command is

```
x = rnorm(100,mean=100,sd=3)
```

The vector x will contain 100 values which are observations from a normal distribution $N(100, 9)$.

(a) What is the mean and the variance of this sample? How do the sample mean and sample variance compare to true values of the mean and variance?

Instructions. Use functions `mean` and `var` in R to find the mean and the variance.

```
mean(x)
var(x)
```

(b) Obtain random samples from the sampling distributions for the sample mean and the sample variance.

Instructions. In order to check the sampling distribution of the sample mean $\widehat{\mu}$ and of the sample variance $\widehat{\sigma}^2$, we will simulate 100 samples for several times (say 500 times). To simulate 500 times, we run the `rnorm` command within a for loop and create a matrix X with 500 rows and 100 columns, each row corresponding to one sample of 100 observations:

```
n = 100 #number of observations in one sample
S = 500 #number of simulations
X = matrix(0,nrow=S, ncol=n)
for(i in 1:S){
  X[i,] = rnorm(n,mean=100,sd=3)
}
```

To obtain the sample means of the 500 samples, we apply the function `apply` as follows:

```
means = apply(X,1,mean)
```

The vectors `means` will contain the 500 sample means and 500 sample variances of the 500 samples.

(c) Plot the sample means using a histogram.

Instructions. The R command for a histogram is `hist`.

```
hist(means)
```

(d) What is the (theoretical) sampling distribution of $\widehat{\mu}$ if we know that the 500 samples come from a normal distribution $N(100, 9)$? Does the histogram approximate the sampling distribution for the sample mean? Why?