STA 2002, Summer 2024
Homework 2 Solution – Sampling distribution and Point estimator

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 $\mu = 3$?

Answer:

(a)
$$Bias(\widehat{\mu}_1) = \mathbb{E}(\widehat{\mu}_1) - \mu = \mathbb{E}(X_1/3) + \mathbb{E}(X_2/3) + \mathbb{E}(X_3/3) - \mu = 0,$$
$$Bias(\widehat{\mu}_2) = \mathbb{E}(\widehat{\mu}_2) - \mu = \mathbb{E}(X_1/4) + \mathbb{E}(X_2/3) + \mathbb{E}(X_3/5) - \mu = -\frac{13\mu}{60},$$

$\widehat{\mu}_1$ is always unbiased, $\widehat{\mu}_2$ is biased unless $\mu = 0$.

(b)
$$Var(\widehat{\mu}_1) = Var(X_1/3) + Var(X_2/3) + Var(X_3/3) = \frac{7+13+20}{9} = 4.44,$$
$$Var(\widehat{\mu}_2) = Var(X_1/4) + Var(X_2/3) + Var(X_3/5) = \frac{7}{16} + \frac{13}{9} + \frac{20}{25} = 2.68,$$

$\widehat{\mu}_2$ has the smallest variance.

(c) When $\mu = 3$,
$$MSE(\widehat{\mu}_1) = Bias(\widehat{\mu}_1)^2 + Var(\widehat{\mu}_1) = 4.44,$$
$$MSE(\widehat{\mu}_2) = Bias(\widehat{\mu}_2)^2 + Var(\widehat{\mu}_2) = \frac{169\mu^2}{3600} + 2.68 = 3.10,$$

$\widehat{\mu}_2$ 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

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

Answer:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} e^{-(x_i-\theta)} \cdot 1\left\{x_i \geqslant \theta\right\} \\ &= e^{-\sum_{i=1}^{n} x_i + n\theta} 1\left\{x_{(1)} \geqslant \theta\right\} \end{aligned}$$

where $x_{(1)} = \min\{x_1, x_2, \cdots x_n\}$.
Obviously $L(\theta)$ is an increasing function of θ for $\theta \leq X_{(1)}$ and the maximum is obtained at $\widehat{\theta} = X(1) \Rightarrow MLE$ is $\widehat{\theta} = X(1)$

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.

Answer:

(a) Denote the joint density of (X, K) by $f_{(X,K)}(x, k)$. Also, denote the conditional density of $X \mid K = k$ by $f_{X|K}(x \mid k)$ for $k \in \{0, 1\}$. Clearly, k can only take values on $\{0, 1\}$. By the law of total probability,

$$\begin{aligned} f_{(X,K)}(x, k) &= f_{X|K}(x \mid k)\mathbb{P}(K = k) \\ &= \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \pi_k \\ &= \begin{cases} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} \pi_0, & \text{if } x \in \mathbb{R}, k = 0, \\ \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \pi_1, & \text{if } x \in \mathbb{R}, k = 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

(b) 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. Then $n = n_0 + n_1$. Also, denote I_0 and I_1 to be respectively

$$I_0 = \{i; K_i = 0\}, \quad I_1 = \{i; K_i = 1\}$$

The likelihood function is

$$\begin{aligned}
L\left(\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2\right) &= \prod_{i=1}^{n} f_{(X,K)}\left(x_i, k_i\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^{n_0} \exp\left\{-\sum_{i\in I_0} \frac{\left(x_i-\mu_0\right)^2}{2\sigma_0^2}\right\} \pi_0^{n_0} \times \\
&\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n_1} \exp\left\{-\sum_{i\in I_1} \frac{\left(x_i-\mu_1\right)^2}{2\sigma_1^2}\right\} \left(1-\pi_0\right)^{n_1}
\end{aligned}$$

The log-likelihood function is

$$\begin{aligned}
l\left(\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2\right) =& n_0 \log \pi_0 + n_1 \log\left(1-\pi_0\right) - \sum_{i\in I_0} \frac{\left(x_i-\mu_0\right)^2}{2\sigma_0^2} - \sum_{i\in I_1} \frac{\left(x_i-\mu_1\right)^2}{2\sigma_1^2} - (n_0/2)\log\sigma_0^2 \\
&- (n_1/2)\log\sigma_1^2 - (n/2)\log(2\pi)
\end{aligned}$$

We now take the partial derivative of l with respect to $\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2$. Note that for μ_0, σ_0^2 their partial derivatives in the log-likelihood function are exactly the same as covered in class (see LN4 note) or in the book Example 6.4-3 as if there are only n_0 samples. For μ_1, σ_1^2 their partial derivatives in the log-likelihood function is the same as covered in class (see LN4 note) or in the book Example 6.4-3 as if there are only n_1 samples. Therefore, the maximum likelihood estimators are

$$\begin{aligned}
\widehat{\pi}_0 &= \frac{n_0}{n}, \\
\widehat{\mu}_0 &= \frac{1}{n_0} \sum_{i\in I_0} x_i, \\
\widehat{\sigma}_0^2 &= \frac{1}{n_0} \sum_{i\in I_0} \left(x_i - \widehat{\mu}_0\right)^2, \\
\widehat{\mu}_1 &= \frac{1}{n_1} \sum_{i\in I_1} x_i, \\
\widehat{\sigma}_1^2 &= \frac{1}{n_1} \sum_{i\in I_1} \left(x_i - \widehat{\mu}_1\right)^2.
\end{aligned}$$

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 that the standard deviation is 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?

Answer:

(a) The sample mean

$$\bar{X} = \frac{15.2 + 14.2 + 14.0 + 12.2 + 14.4 + 12.5 + 14.3 + 14.2 + 13.5 + 11.8 + 15.2}{11} = 13.773.$$

Since

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

we have

$$\mathbb{P}(-Z_{0.005} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq Z_{0.005}) = 0.01,$$

and a 99% confidence interval is given by

$$\bar{X} - Z_{0.005}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{0.005}\frac{\sigma}{\sqrt{n}},$$

plug in $Z_{0.005} = 2.57$, we obtain

$$13.384 \leq \mu \leq 14.162.$$

(b) The 95% lower confidence bound is

$$\mu \geq \bar{X} - Z_{0.05}\sigma/\sqrt{n},$$

plug in $Z_{0.05} = 1.64$, we obtain

$$\mu \geq 13.526.$$

(c) Here we choose the value of n, such that

$$Z_{0.025}\frac{\sigma}{\sqrt{n}} \leq 2,$$

that is

$$n \geq \left(\frac{Z_{0.025}\sigma}{2}\right)^2 = \left(\frac{1.96\sigma}{2}\right)^2 = 0.24,$$

therefore, n is at least 1.

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.)

Answer: Sample size $n = 12$. The sample mean

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = 2.08,$$

the sample variance

$$s^2 = \frac{\sum_{i=1}^{12}(x_i - \bar{x})^2}{12 - 1} = 0.0245.$$

We know that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(12-1),$$

and

$$\mathbb{P}\left(\chi^2_{0.975,11} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{0.025,11}\right) = 0.95,$$

the 95% confidence interval for σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{0.025,11}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{0.975,11}},$$

which means that the 95% confidence interval for σ is

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{0.025,11}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{0.975,11}}}$$

plug in $\chi_{0.025,11} = 21.92$, $\chi_{0.975,11} = 3.82$, we obtain

$$0.111 < \sigma < 0.266.$$

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.)

Answer:

We combine the unsure votes with the ones who think legalization is a good idea, thus making the vote to have only 2 options: oppose the legalization, and others.

According to the sample proportion model, let $\widehat{p}$ be the sample proportion opposing the legalization.

$$\widehat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

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

By using $\widehat{p}(1-\widehat{p})$ as an estimation of $p(1-p)$, since $\widehat{p} = 0.32$, $n = 1346$, we have the 95% confidence interval to be:

$$\widehat{p} - Z_{0.025}\frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}} \leq p \leq \widehat{p} + Z_{0.025}\frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}},$$

which is

$$0.295 = 0.32 - 0.025 \leq p \leq 0.32 + 0.025 = 0.345.$$

Interpretation: It says that we are 95% confident that the actual percentage of people opposing the legalization is within 29.5% and 34.5%.

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.

   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?

Answer: For this problem, the data are randomly generated. Any reasonable output by R will be OK.

(a) R code and output:

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

> mean(x)

[1] 100.0021

> var(x)

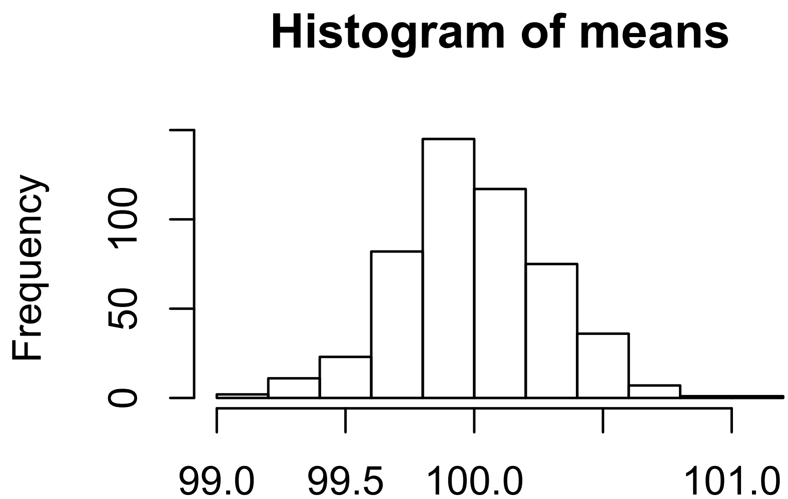
[1] 10.06762

The sample mean and variance are very close to the true value. (Any reasonable observations will do).

(b) Well there's no question in this part :)

(c) R code and output:

The histogram looks like



(d) The theoretical sampling distribution of sample mean should be a normal distribution with mean 100, standard deviation $3/\sqrt{100} = 0.3$, which is

$$\widehat{\mu} \sim \mathcal{N}(100, 0.3^2).$$

This histogram resembles this theoretical distribution, because the 500 sample means are independent and all follow this distribution.