Xiao Zhehao A0235422N

1. The MGF of $\sum X_i$ is $M(t)=E[e^{t\sum X_i}]=E[\prod e^{tX_i}]=\prod E[e^{tX_i}]=\prod\frac{1}{(1-2t)^{r_i/2}}=(1-2t)^{-\frac{\sum r_i}{2}}$. Hence, $E[\sum X_i]=\frac{d}{dt}(1-2t)^{-\frac{\sum r_i}{2}}\big|_{t=0}=-\frac{\sum r_i}{2}(1-2t)^{-\frac{\sum r_i}{2}-1}(-2)\big|_{t=0}$
$=(\sum r_i)(1)^{-\frac{\sum r_i}{2}-1}=\sum r_i$

and $E[(\sum X_i)^2]=\frac{d}{dt}\left[(\sum r_i)(1-2t)^{-\frac{\sum r_i}{2}-1}\right]\big|_{t=0}=\sum r_i\frac{d}{dt}\left[(1-2t)^{-\frac{\sum r_i}{2}-1}\right]_{t=0}=(\sum r_i)\left[(-\frac{\sum r_i}{2}-1)(1-2t)^{-\frac{\sum r_i}{2}-2}(-2)\right]_{t=0}=(\sum r_i)(\sum r_i+2)$ and $Var(\sum X_i)=E[(\sum X_i)^2]-E[\sum X_i]^2=(\sum r_i)^2+2\sum r_i-(\sum r_i)^2=2\sum r_i$

Hence, by MGF method, $\sum X_i$ has mean $\sum r_i$ and variance $2\sum r_i$

2. a) We need $P(\sum_{i=1}^{20}X_i\geq 19)$ where $X_i\sim Poisson(1.8), i\in[1,20]$. Since X_i are i.i.d. Poisson with $E(X_i)=Var(X_i)=1.8$, and $n=20$, by CLT $\frac{\sum X_i-20(1.8)}{\sqrt{20}1.8}\sim N(0,1)$.

Hence, $P(\sum X_i\geq 19)=P(\frac{\sum X_i-36}{6}\geq\frac{19-36}{6})=P(\frac{\sum X_i-36}{6}\geq-\frac{17}{6})=0.9977$.

b) We need $P(\sum_{i=1}^{120}Y_i<365)$ where $Y_i\sim Exp(3), i\in[1,120]$. Since Y_i are i.i.d. Exp with $E(Y_i)=\frac{1}{\lambda_i}=3$ and $Var(Y_i)=\frac{1}{\lambda_i^2}=9$ and $n=120$, by CLT $\frac{\sum Y_i-120(3)}{\sqrt{120(9)}}\sim N(0,1)$

Hence, $P(\sum Y_i<365)=P(\frac{\sum Y_i-360}{6\sqrt{30}}<\frac{365-360}{6\sqrt{30}})=P(\frac{\sum Y_i-360}{6\sqrt{30}}<\frac{5}{6\sqrt{30}})=0.5605$

3. a) We know that $E(X)=\theta$ and $f_X(x)=\frac{1}{\theta}e^{-x/\theta}$. Hence, $\lambda=\frac{1}{\theta}=E(X)^{-1}$. Since $E(X)\approx\bar{X}=\frac{1}{n}\sum X_i$ and $\hat{\lambda}_{mom}=\frac{n}{\sum X_i}=\frac{1}{\bar{X}}$, where $\bar{X}$ is the sample mean.

b) Since $X_{1\sim n}$ are i.i.d. exponential variables, the likelihood function is $L(\lambda;X_1,\dots,X_n)=\prod_{i=1}^{n}f_X(X_i)=\lambda^n e^{-\lambda\sum X_i}$. Taking the log we get $\ln(L)=n\ln\lambda-\lambda\sum X_i$

Since $f_X(x)$ is concave(exp), the maximum likelihood estimator is $\hat{\lambda}_{mle}=\arg\max_{\lambda}\ln(L)$ so at $\frac{d}{d\lambda}(n\ln\lambda-\lambda\sum X_i)=\frac{n}{\lambda}-\sum X_i=0$ so $\hat{\lambda}_{mle}=\frac{n}{\sum X_i}=\frac{1}{\bar{X}}$.

c) $\hat{\lambda}_{mom}=\hat{\lambda}_{mle}=\frac{1}{\bar{X}}=\frac{6}{3.8+3.24+1.4+1.22+4.5+4.6}=0.3198$

d) The bias of $\hat{\lambda}_{mle}$ is $E[\hat{\lambda}_{mle}]-\lambda=E[\frac{1}{\bar{X}}]-\lambda$. Since X is exp, $E[\sum X_i]=\frac{n-1}{\lambda}$ (* sum of exp distribution) and so $E[\frac{1}{\bar{X}}]=\frac{n\lambda}{n-1}$ so the bias is $\frac{n\lambda}{n-1}-\lambda=\frac{\lambda}{n-1}$ and $\hat{\lambda}_{mle}$ is biased.

4. Based on the p.d.f., the first moment $E[x]=\int_{-\infty}^{\infty}x f(x;\theta)=\int_{-1}^{1}x\frac{1}{2}(1+\theta x)dx=\int_{-1}^{1}\frac{x}{2}+\frac{x^2}{2}\theta\,dx=\left[\frac{x^2}{4}+\frac{x^3}{6}\theta\right]_{-1}^{1}=\frac{1}{4}+\frac{1}{6}\theta-\left(\frac{1}{4}-\frac{1}{6}\theta\right)=\frac{\theta}{3}$
The sample first moment is $\bar{X}=\frac{1}{n}\sum X_i$. Assuming $E[x]=\bar{X}$, we have $\frac{\theta}{3}=\frac{1}{n}\sum X_i$ so $\hat{\theta}_{mom}=\frac{3}{n}\sum X_i=3\bar{X}$

The bias is $E[\hat{\theta}_{mom}]-\theta=E[3\bar{X}]-\theta=3E[\bar{X}]-\theta=3\bar{X}-\theta=0$. Hence, it is unbiased.

5. The likelihood function of $f_X(x)$ is $L(\theta;x_1\dots x_n)=\prod_{i=1}^{n}e^{-(x_i-\theta)}$ for $x_i\geq\theta,\forall i\in[1,n]$. The log likelihood is $\ln L(\theta;x)=\sum_{i=1}^{n}\theta-x_i=n\theta-\sum x_i$
The derivative is $\frac{d}{d\theta}\sum\theta-x_i=n$, which is a constant. This implies the likelihood increases monotonically with θ. As the pdf support is for $x\geq\theta$, $\hat{\theta}_{mle}=\min(X_1,\dots,X_n)$

6. We know $\hat{\theta}_{mle}$ is the maximum order statistic $X_{(n)}$. Hence, the pdf and cdf of $\hat{\theta}_{mle}$ is that of $X_{(n)}$, which is $F_{X(n)}(x)=[F_X(x)]^n=(\frac{x}{\theta})^n$ and $f_{X(n)}(x)=\frac{d}{dx}F_{X(n)}(x)=n\frac{x^{n-1}}{\theta^n}$. (by definition of order statistics)

Hence, $E[\hat{\theta}_{mle}]=E[X_{(n)}]=\int_{-\infty}^{\infty}x f_{X(n)}(x)dx=n\int_0^{\theta}\frac{x^n}{\theta^n}=n\left[\frac{x^{n+1}}{(n+1)\theta^n}\right]_0^{\theta}=\frac{n}{n+1}\theta\neq\theta$. Hence, $\hat{\theta}_{mle}$ is a biased estimator of θ. To make it unbiased, we simply do $\hat{\theta}_{ub}=\frac{n+1}{n}\hat{\theta}_{mle}=\theta$.

7. a) Since each normal distribution has support over $(-\infty,\infty)$, the GMM's is the union with support $(-\infty,\infty)$ and joint density is $f(x)=\pi_0 N(x|\mu_0,\sigma_0^2)+\pi_1 N(x|\mu_1,\sigma_1^2)$

*note log here represents ln

b) The likelihood func is $L=\prod_{i=1}^{n}[\pi_{K_i}N(X_i|\mu_{K_i},\sigma_{K_i}^2)]$ and $\log L=\sum[\log\pi_{K_i}+\log N(X_i|\mu_{K_i},\sigma_{K_i}^2)]=n_0\log\pi_0+n_1\log(1-\pi_0)+\sum\log N(X_i|\mu_{K_i},\sigma_{K_i}^2)$. Taking the derivative wrt each param, and solving

Now, we take the derivative wrt each param and solve when the derivative is 0 to maximise the likelihood of the param

$\frac{d}{d\pi_0}\log L=\frac{d}{d\pi_0}(n_0\log\pi_0+n_1\log(1-\pi_0))=\frac{n_0}{\pi_0}-\frac{n_1}{1-\pi_0}=0,\ \frac{n_0}{\pi_0}=\frac{n_1}{1-\pi_0},\ n_0-\pi_0n_0=n_1\pi_0,\ n_0=N\pi_0,\ \hat{\pi}_0=\frac{n_0}{N}$ and $\hat{\pi}_1=\frac{n_1}{N}$ analagously

$\frac{d}{d\mu_0}\log L=\frac{d}{d\mu_0}\sum\log\left(\frac{1}{\sqrt{2\pi\sigma_{K_i}^2}}\exp\left(-\frac{(X_i-\mu_{K_i})^2}{2\sigma_{K_i}^2}\right)\right)=\frac{d}{d\mu_0}\sum\left[\log\left(\frac{1}{\sqrt{2\pi\sigma_{K_i}^2}}\right)-\frac{(X_i-\mu_{K_i})^2}{2\sigma_{K_i}^2}\right]=-\frac{d}{d\mu_0}\sum\frac{(X_i-\mu_{K_i})^2}{2\sigma_{K_i}^2}=-\sum\mathbb{1}_{K_i=0}\frac{d}{d\mu_0}\frac{(X_i-\mu_0)^2}{2\sigma_0^2}=\sum\mathbb{1}_{K_i=0}\frac{X_i-\mu_0}{\sigma_0^2}=0$

$\sum\mathbb{1}_{K_i=0}(X_i-\mu_0)=0\quad n_0\mu_0=\sum\mathbb{1}_{K_i=0}X_i\quad\hat{\mu}_0=\bar{X}_{K_i=0}$ and $\hat{\mu}_1=\bar{X}_{K_i=1}$ analagously

$\frac{d}{d\sigma_0^2}\log L=\frac{d}{d\sigma_0^2}\sum\log\left(\frac{1}{\sqrt{2\pi\sigma_{K_i}^2}}\exp\left(-\frac{(X_i-\mu_{K_i})^2}{2\sigma_{K_i}^2}\right)\right)=\frac{d}{d\sigma_0^2}\sum\left[\log\left(\frac{1}{\sqrt{2\pi\sigma_{K_i}^2}}\right)-\frac{(X_i-\mu_{K_i})^2}{2\sigma_{K_i}^2}\right]=\sum\mathbb{1}_{K_i=0}\frac{d}{d\sigma_0^2}\left(\log\frac{1}{\sqrt{2\pi\sigma_0^2}}-\frac{(X_i-\mu_0)^2}{2\sigma_0^2}\right)=\sum\mathbb{1}_{K_i=0}\left(-\frac{1}{2\sigma_0^2}+\frac{(X_i-\mu_0)^2}{2\sigma_0^4}\right)=-\frac{n_0}{2\sigma_0^2}+\frac{1}{2\sigma_0^4}\sum\mathbb{1}_{K_i=0}(X_i-\mu_0)^2=0$

$\frac{1}{2\sigma_0^4}\sum\mathbb{1}_{K_i=0}(X_i-\mu_0)^2=\frac{n_0}{2\sigma_0^2}\quad\frac{1}{n_0}\sum\mathbb{1}_{K_i=0}(X_i-\mu_0)^2=\sigma_0^2\quad\sigma_0^2=\mathrm{Var}(X_{K_i=0})$ and $\sigma_1^{-2}=\mathrm{Var}(X_{K_i=1})$ analagously. Var is pop variance

c) $\pi_0=0.9491,\ \mu_0=49.9577,\ \sigma_0^2=99.8605,\ \pi_1=0.0509,\ \mu_1=60.8127,\ \sigma_1^2=101.6141$

ST2132 Mathematical Statistics
Assignment 1
Due date: 7 February 2025 23:59 on Canvas
Please submit your writeup in pdf format. Submit your code in zip format.

Please answer all questions. Please work on the assignment by yourself only. We follow a strict rule on academic honesty.

Generative AI policy: The use of generative AI tools (e.g., ChatGPT, Gemini, etc.) is strictly prohibited for structural or conceptual questions in this course. However, these tools may be used to assist with coding tasks, provided their use is clearly documented.

1. Let $X_1, X_2, \ldots, X_n$ be independent chi-square distribution with degrees of freedom $r_1, r_2, \ldots, r_n$, that is, $X_i \sim \chi^2(r_i)$, where $r_i > 0$, for $i = 1, \ldots, n$. Identify the distribution of $\sum_{i=1}^n X_i$ using the moment generating function technique.

2. The number of floods that occur in a certain region over a given year is a random variable having a Poisson distribution with mean 1.8, independently from one year to the other. Moreover, the time period (in days) during which the ground is flooded, at the time of an arbitrary flood, is an exponential random variable with mean 3. We assume that the durations of the floods are independent. Using the central limit theorem, calculate (approximately)

   (a) the probability that over the course of the next 20 years, there will be at least 19 floods in this region. Assume that we do not need to apply half-unit correction for this question.

   (b) the probability that the total time during which the ground will be flooded over the course of the next 120 floods will be smaller than 365 days.

3. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. exponential random variables with pdf given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0\,, \\ 0 & \text{otherwise.} \end{cases}$$

   We are interested in estimating the parameter λ using a number of methods.

   (a) What is the method of moment estimator $\hat{\lambda}_{mom}$ of λ?

   (b) What is the maximum likelihood estimator $\hat{\lambda}_{mle}$ of λ?

   (c) Suppose the following $n = 6$ samples was generated from a exponential distribution with parameter λ as described in this question:

$$3.8 \quad 3.24 \quad 1.4 \quad 1.22 \quad 4.5 \quad 4.6$$

   Compute $\hat{\lambda}_{mom}$ and $\hat{\lambda}_{mle}$.

   (d) Suppose that $n > 1$. Is $\hat{\lambda}_{mle}$ an unbiased estimator of λ? Why?

4. Consider random samples $X_1, \ldots, X_n$ drawn i.i.d. from the probability density function given by

$$f(x;\theta) = \begin{cases} \frac{1}{2}(1+\theta x), & \text{if } -1 \le x \le 1\,, \\ 0, & \text{otherwise.} \end{cases}$$

   where $-1 \le \theta \le 1$. Find the method of moment estimator of θ. Is it an unbiased estimator of θ?

5. Find the maximum likelihood estimator of the unknown parameter θ where $X_1, X_2, \ldots, X_n$ is a sample from the distribution whose density function is

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

6. Let $X_1, \ldots, X_n$ be i.i.d. random samples from Uniform$[0, \theta]$. We saw in class that the mle $\hat{\theta}$ of θ is $\hat{\theta} = \max\{X_1, \ldots, X_n\}$. Show that this is a biased estimator of θ. Use the result to propose an unbiased estimator of θ. Hint: First find the cdf and then the pdf of $\max\{X_1, \ldots, X_n\}$. Then use it to calculate its mean.

7. (Gaussian Mixture Model (GMM)) 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. For example, it is used in image processing and various clustering algorithms. 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 $K = k$ with $k \in \{0, 1\}$, 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 $(X, K) \sim \text{GMM}(\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$. 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.

(c) On Canvas, there is a dataset called "GMM.csv", which is a comma separated file. This dataset contains $n = 100,000$ rows and 2 columns. The first column is the realizations of K, and the second column is the realizations of X. Using any computing language of your choice, e.g. Python/R/Julia/MATLAB/Excel, compute the maximum likelihood estimators derived in part (b) using this dataset. Attach your code at the end of your assignment.