1. $X_1, X_2, ..., X_n$ are observations of a random sample of size n from the geometric distribution with probability distribution $f(x,\theta) = \theta(1-\theta)^x$ for $x = 0, 1, \ldots$.

   (a) **(1 mark)** Find the estimator of θ by the method of moment.

   (b) **(1 mark)** Find the estimator of θ by the method of maximum likelihood.

   (c) **(2 marks)** Find the Cramer-Rao Lower Bound for the variance of an unbiased estimator for θ.

   (d) **(2 marks)** Does the variance of any unbiased estimator for θ achieve this bound? Why? Explain in details.

   (e) **(3 marks)** Find the limiting distribution of the maximum likelihood estimator for θ by Central Limit Theorem and Delta method. What phenomenon do you observe?

   (f) **(1 mark)** Is the geometric distribution a member of exponential family? Hence or otherwise, find the complete and minimal sufficient statistic.

   (g) **(1 mark)** Find the distribution of the minimal complete and sufficient statistic.

   (h) **(1 mark)** Find the UMVUE of $\frac{1-\theta}{\theta}$.

   (i) **(5 marks)** Find the UMVUE of θ.

Solutions:

(a) For MME:

$$\begin{aligned} M_1' &= \widetilde{E(X)} \\ \frac{1}{n}\sum_{i=1}^{n} x_i &= \frac{\widetilde{1-\theta}}{\theta} \\ \tilde{\theta} &= \frac{1}{1+\bar{X}} \end{aligned}$$

(b)

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \theta^n (1-\theta)^{\sum_{i=1}^n x_i} \\ l(\theta) &= log f_{\mathbf{X}}(\mathbf{x}) = nlog\theta + (\sum_{i=1}^{n} x_i) log(1-\theta) \\ l'(\theta) &= \frac{n}{\theta} + \frac{\sum_{i=1}^n x_i}{\theta - 1} \end{aligned}$$

Taking $l'(\theta) = 0$, we have $\theta = \frac{1}{1+\bar{x}}$ and $l''(\theta) < 0$ at $\theta = \frac{1}{1+\bar{x}}$.Therefore, we get the MLE for θ, which is $\hat{\theta} = \frac{1}{1+\bar{X}}$.

(c)

$$\begin{aligned} log f_{X_i}(x_i;\theta) &= log\theta + x_i log(1-\theta)\\ \frac{\partial^2}{\partial\theta^2} log f_{X_i}(x_i;\theta) &= -\frac{1}{\theta^2} - \frac{x_i}{(1-\theta)^2}\\ E(\frac{\partial^2}{\partial\theta^2} log f_{X_i}(x_i;\theta)) &= -\frac{1}{\theta^2} - \frac{\frac{1-\theta}{\theta}}{(1-\theta)^2} = -\frac{1}{\theta^2(1-\theta)}\\ g(\theta) &= \theta\\ g'(\theta) &= 1\\ CR\ Lower\ Bound &= -\frac{g'(\theta)^2}{nE(\frac{\partial^2}{\partial\theta^2} log f_{X_i}(x_i;\theta))}\\ &= \frac{\theta^2(1-\theta)}{n} \end{aligned}$$

(d)

$$\frac{\partial}{\partial\theta} log f_{\mathbf{X}}(\mathbf{x};\theta) = \frac{n}{\theta} - \frac{\sum_{i=1}^n x_i}{1-\theta} = \frac{n}{\theta-1}(\frac{\sum_{i=1}^n x_i}{n} - \frac{1-\theta}{\theta})$$

There is no function $A(n,\theta)$ s.t. $\frac{\partial}{\partial\theta} log f_{\mathbf{X}}(\mathbf{x};\theta) = A(n,\theta)[h(\mathbf{x}) - g(\theta)]$, where $g(\theta) = \theta$ and $h(\mathbf{x})$ is an unbiased estimator of $g(\theta)$. Therefore, no unbiased estimator of θ achieve the CR lower bound.

(e) By C.L.T., we have $\bar{X} \to \mathrm{N}(E(X_i), \frac{Var(X_i)}{n}) = \mathrm{N}(\frac{1-\theta}{\theta}, \frac{1}{n}\frac{1-\theta}{\theta^2})$.
Take

$$\begin{aligned} g(t) &= \frac{1}{1+t}\\ g'(t) &= -\frac{1}{(1+t)^2} \end{aligned}$$

So MLE $\hat{\theta} = g(\bar{X})$ Then by Delta Method, we have g($\hat{\theta}$) $\to$ $\mathrm{N}(g(\frac{1-\theta}{\theta}), g'(\frac{1-\theta}{\theta})^2 \frac{1}{n}\frac{1-\theta}{\theta^2}) = \mathrm{N}(\theta, \frac{(1-\theta)\theta^2}{n})$. So the limiting variance of MLE can achieve the CR lower bound.

(f)

$$f(x,\theta) = \theta(1-\theta)^x = exp(log(\theta) + xlog(1-\theta))$$

Therefore, geometric distribution belongs to exponential family. Hence, the complete and minimal sufficient statistic is $\sum_{i=1}^n d(X_i) = \sum_{i=1}^n X_i$.

(g) Since $X_i \overset{\text{i.i.d.}}{\sim}$ Geometric(θ), we have

$$\begin{aligned} M_{X_i}(t) &= \frac{\theta}{1-(1-\theta)e^t}\\ M_{\sum_{i=1}^n X_i}(t) &= \Pi_{i=1}^n M_{X_i}(t) = (\frac{\theta}{1-(1-\theta)e^t})^n \end{aligned}$$

Therefore, $\sum_{i=1}^n X_i \sim$ NB(n,θ).

(h) Denote $S = \sum_{i=1}^n X_i \sim$ NB(n,θ). Therefore,

$$\begin{aligned} E(S) &= \frac{n(1-\theta)}{\theta}\\ \Rightarrow E(\frac{S}{n}) &= \frac{1-\theta}{\theta} \end{aligned}$$

Since S is complete and sufficient and $E(\frac{S}{n}) = \frac{1-\theta}{\theta}$, we have $\frac{S}{n} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$ is the UMVUE for $\frac{1-\theta}{\theta}$

(i) Assume E(h(S)) = θ, where S is as defined in part (h). Then

$$\begin{aligned}
&\sum_{s=0}^{\infty} h(s)\frac{(n+s-1)!}{s!(n-1)!}\theta^n(1-\theta)^s = \theta \\
\Rightarrow\ &\sum_{s=0}^{\infty} h(s)\frac{(n+s-1)!}{s!(n-1)!}\theta^{n-1}(1-\theta)^s = 1 \\
\Rightarrow\ &\sum_{s=0}^{\infty} h(s)\frac{(n+s-1)!(n-2)!}{(n-1)!(n+s-2)!}\frac{(n-1+s-1)!}{(n-2)!s!}\theta^{n-1}(1-\theta)^s = 1 \\
\Rightarrow\ &h(s)\frac{(n+s-1)!(n-2)!}{(n-1)!(n+s-2)!} = 1 \\
\Rightarrow\ &h(s) = \frac{n-1}{n+s-1}
\end{aligned}$$

Therefore, h(S) = $\frac{n-1}{n+S-1} = \frac{n-1}{n-1+\sum_{i=1}^n X_i}$ is the UMVUE for θ.

2. Two independent random samples, X_{ij} for $i = 1, 2; j = 1, .., n_i$, are normally distributed with mean μ_i and variance σ_i^2. Define

$$\begin{aligned}
\bar{X}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \qquad \text{for } \ i = 1, 2 \\
S_i^2 &= \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \qquad \text{for } \ i = 1, 2 \\
\bar{\bar{X}} &= \frac{1}{\sum_{i=1}^2 n_i}\sum_{i=1}^{2}\sum_{j=1}^{n_i} X_{ij} \\
S_p^2 &= \frac{1}{\sum_{i=1}^2 (n_i - 1)}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \ .
\end{aligned}$$

(a) **(4 marks)** Assume that $\mu_1 = \mu_2 = \mu \in \mathcal{R}$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$. Find the distributions of $\bar{X}_1$, $\bar{X}_2$ and $\bar{\bar{X}}$. Which of $\bar{X}_1$, $\bar{X}_2$ and $\bar{\bar{X}}$ is the most efficient estimator of μ? Why?

(b) **(4 marks)** Assume that $\mu_1 = \mu_2 = \mu \in \mathcal{R}$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$. Find the distributions of S_1^2, S_2^2 and S_p^2. Which of S_1^2, S_2^2 and S_p^2 is the most efficient estimator of σ^2? Why?

(c) Assume that $\mu_1 \in \mathcal{R}$, $\mu_2 \in \mathcal{R}$, $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$.

i. **(1 mark)** Find the set of minimal complete and sufficient statistics for the unknown parameters of μ_1, μ_2, σ_1^2 and σ_2^2.

ii. **(6 marks)** Find the UMVUE of σ_1/σ_2.

(d) Assume that $\mu_1 \in \mathcal{R}$, $\mu_2 \in \mathcal{R}$, $\sigma_1^2 = \sigma_2^2 = \sigma^2 > 0$.

i. **(1 mark)** Find the set of minimal complete and sufficient statistics for the unknown parameters of μ_1, μ_2 and σ^2.

ii. **(4 marks)** Using the fact that sample mean and sample variance are independent, find the UMVUE of $(\mu_1 - \mu_2)/\sigma$.

Solutions:

(a)

$$\begin{aligned}
\bar{X}_1 &\sim N(\mu, \frac{\sigma^2}{n_1}) \\
\bar{X}_2 &\sim N(\mu, \frac{\sigma^2}{n_2}) \\
\bar{\bar{X}} &\sim N(\mu, \frac{\sigma^2}{n_1+n_2})
\end{aligned}$$

Since $\frac{\sigma^2}{n_1+n_2} < \frac{\sigma^2}{n_1}$ and $\frac{\sigma^2}{n_1+n_2} < \frac{\sigma^2}{n_1}$, we have $\bar{\bar{X}}$ is the most efficient estimator of μ.

(b)

$$\begin{aligned}
\frac{(n_1-1)S_1^2}{\sigma^2} &\sim \chi^2_{n_1-1} \\
\frac{(n_2-1)S_2^2}{\sigma^2} &\sim \chi^2_{n_2-1} \\
\frac{(n_1+n_2-2)S_p^2}{\sigma^2} &\sim \chi^2_{n_1+n_2-2}
\end{aligned}$$

$$\begin{aligned}
Var(\frac{(n_1-1)S_1^2}{\sigma^2} &= 2(n_1-1) \\
Var(S_1^2) &= \frac{2(\sigma^2)^2}{(n_1-1)^2}(n_1-1) = \frac{2\sigma^4}{n_1-1}
\end{aligned}$$

Similarly,

$$\begin{aligned}
Var(S_2^2) &= \frac{2\sigma^4}{n_2-1} \\
Var(S_p^2) &= \frac{2\sigma^4}{n_1+n_2-2}
\end{aligned}$$

Since $Var(S_p^2)$ is the smallest, S_p^2 is the most efficient one.

(c) i.

$$\begin{aligned}
f(\mathbf{X_1}, \mathbf{X_2}; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) &= exp(-\frac{n_1}{2}log(2\pi) - \frac{n_1}{2}log(\sigma_1^2) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1} x_{i1}^2 + \frac{\mu_1}{2\sigma_1^2}\sum_{i=1}^{n_1} x_{i1} \\
&\quad -\frac{n_2}{2}log(2\pi) - \frac{n_2}{2}log(\sigma_2^2) - \frac{1}{2\sigma_2^2}\sum_{i=1}^{n_2} x_{i2}^2 + \frac{\mu_2}{2\sigma_2^2}\sum_{i=1}^{n_1} x_{i2})
\end{aligned}$$

Therefore it belongs to the exponential family. ⇒ $(\sum_{i=1}^{n_1} X_{i1}^2, \sum_{i=1}^{n_1} X_{i1}, \sum_{i=1}^{n_2} X_{i2}^2, \sum_{i=1}^{n_2} X_{i2})$ is complete and sufficient. ⇒ $(\bar{X}_1, \bar{X}_2, S_1^2, S_2^2)$ is complete and sufficient.

ii. The independence of the two random samples implies that S_1^2 and S_2^2 are independent. By the distributions we get in part(b), the following holds:

$$\begin{aligned}
E(\sqrt{\frac{(n_1-1)S_1^2}{\sigma_1^2}}) &= \int_0^\infty \sqrt{t}\frac{t^{(n_1-1)/2-1}e^{-t/2}}{2^{(n_1-1)/2}\Gamma((n_1-1)/2)}dt \\
&= \int_0^\infty \frac{t^{(n_1)/2-1}e^{-t/2}}{2^{(n_1-1)/2}\Gamma((n_1-1)/2)}dt \\
&= \int_0^\infty \frac{2^{\frac{n_1}{2}}\Gamma(\frac{n_1}{2})}{\Gamma(\frac{n_1-1}{2})2^{\frac{n_1-1}{2}}}\frac{t^{\frac{n_1}{2}-1}e^{-t/2}}{2^{\frac{n_1}{2}}\Gamma(\frac{n_1}{2})}dt \\
&= \frac{2^{\frac{1}{2}}\Gamma(\frac{n_1}{2})}{\Gamma(\frac{n_1-1}{2})} \\
\Rightarrow E(\frac{\Gamma(\frac{n_1-1}{2})}{2^{\frac{1}{2}}\Gamma(\frac{n_1}{2})}\sqrt{(n_1-1)S_1^2}) &= \sigma_1 \\
E(\sqrt{\frac{\sigma_2^2}{(n_2-1)S_2^2}}) &= \int_0^\infty \frac{1}{\sqrt{t}}\frac{t^{(n_2-1)/2-1}e^{-t/2}}{2^{(n_2-1)/2}\Gamma((n_2-1)/2)}dt \\
&= \int_0^\infty \frac{t^{(n_2-2)/2-1}e^{-t/2}}{2^{(n_2-1)/2}\Gamma((n_2-1)/2)}dt \\
&= \int_0^\infty \frac{2^{\frac{n_2-2}{2}}\Gamma(\frac{n_2-2}{2})}{\Gamma(\frac{n_2-1}{2})2^{\frac{n_2-1}{2}}}\frac{t^{\frac{n_2-2}{2}-1}e^{-t/2}}{2^{\frac{n_2-2}{2}}\Gamma(\frac{n_2-2}{2})}dt \\
&= \frac{\Gamma(\frac{n_2-2}{2})}{2^{\frac{1}{2}}\Gamma(\frac{n_2-1}{2})} \\
\Rightarrow E(\frac{2^{\frac{1}{2}}\Gamma(\frac{n_2-1}{2})}{\Gamma(\frac{n_2-2}{2})}\sqrt{\frac{1}{(n_2-1)S_2^2}}) &= \frac{1}{\sigma_2} \\
\Rightarrow E(\frac{\Gamma(\frac{n_1-1}{2})\Gamma(\frac{n_2-1}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2-2}{2})}\sqrt{\frac{n_1-1}{n_2-1}\frac{S_1^2}{S_2^2}}) &= \frac{\sigma_1}{\sigma_2}
\end{aligned}$$

Therefore,$\frac{\Gamma(\frac{n_1-1}{2})\Gamma(\frac{n_2-1}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2-2}{2})}\sqrt{\frac{n_1-1}{n_2-1}\frac{S_1^2}{S_2^2}}$ is the UMVUE for $\frac{\sigma_1}{\sigma_2}$.

(d) i.

$$\begin{aligned}
f(\mathbf{X_1},\mathbf{X_2};\mu_1,\mu_2,\sigma^2) &= exp(-\frac{n_1}{2}log(2\pi)-\frac{n_1}{2}log(\sigma^2)-\frac{1}{2\sigma^2}(\sum_{i=1}^{n_1}x_{i1}^2+\sum_{i=1}^{n_2}x_{i2}^2)+\frac{\mu_1}{2\sigma^2}\sum_{i=1}^{n_1}x_{i1} \\
&\quad +\frac{\mu_2}{2\sigma^2}\sum_{i=1}^{n_1}x_{i2}-\frac{n_2}{2}log(2\pi)-\frac{n_2}{2}log(\sigma^2))
\end{aligned}$$

Therefore it belongs to the exponential family. ⇒ $(\sum_{i=1}^{n_1}X_{i1}^2+\sum_{i=1}^{n_2}X_{i2}^2, \sum_{i=1}^{n_1}X_{i1}, \sum_{i=1}^{n_2}X_{i2})$ is complete and sufficient. ⇒ $(\bar{X}_1, \bar{X}_2, S_p^2)$ is complete and sufficient.

ii. According to the theorem 7.3 in Chapter 1 of lecture notes, we have $X_1 - X_2$ and S_p^2

are independent. By the distributions we get in part(a) and (b), the following holds:

$$\begin{aligned}
E(\sqrt{\frac{\sigma^2}{(n_1+n_2-2)S_p^2}}) &= \int_0^\infty \frac{1}{\sqrt{t}} \frac{t^{\frac{n_1+n_2-2}{2}-1} e^{-\frac{t}{2}}}{2^{\frac{n_1+n_2-2}{2}} \Gamma(\frac{n_1+n_2-2}{2})} dt \\
&= \int_0^\infty \frac{t^{\frac{n_1+n_2-3}{2}-1} e^{-\frac{t}{2}}}{2^{\frac{n_1+n_2-2}{2}} \Gamma(\frac{n_1+n_2-2}{2})} dt \\
&= \int_0^\infty \frac{\Gamma(\frac{n_1+n_2-3}{2})}{\Gamma(\frac{n_1+n_2-2}{2}) 2^{\frac{1}{2}}} \frac{t^{\frac{n_1+n_2-3}{2}-1} e^{-t/2}}{2^{\frac{n_1+n_2-3}{2}} \Gamma(\frac{n_1+n_2-3}{2})} dt \\
&= \frac{\Gamma(\frac{n_1+n_2-3}{2})}{\Gamma(\frac{n_1+n_2-2}{2}) 2^{\frac{1}{2}}} \\
\Rightarrow E(\frac{\Gamma(\frac{n_1+n_2-2}{2}) 2^{\frac{1}{2}}}{\Gamma(\frac{n_1+n_2-3}{2})} \sqrt{\frac{1}{(n_1+n_2-2)S_p^2}}) &= \frac{1}{\sigma} \\
E(\bar{X}_1 - \bar{X}_2) &= \mu_1 - \mu_2 \\
\Rightarrow E((\bar{X}_1 - \bar{X}_2) \frac{\Gamma(\frac{n_1+n_2-2}{2}) 2^{\frac{1}{2}}}{\Gamma(\frac{n_1+n_2-3}{2})} \sqrt{\frac{1}{(n_1+n_2-2)S_p^2}}) &= \frac{\mu_1 - \mu_2}{\sigma}
\end{aligned}$$

Therefore, $(\bar{X}_1 - \bar{X}_2) \frac{\Gamma(\frac{n_1+n_2-2}{2}) 2^{\frac{1}{2}}}{\Gamma(\frac{n_1+n_2-3}{2})} \sqrt{\frac{1}{(n_1+n_2-2)S_p^2}}$ is the UMVUE of $\frac{\mu_1-\mu_2}{\sigma}$.

3. Suppose that we have two independent random samples: X_1, ..., X_n are exponential(θ), with density

$$f(x|\theta) = \theta e^{-\theta x}, \quad x > 0$$

and Y_1, ..., Y_m are exponential (μ).

(a) **(4 marks)** Find the UMP test for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ at $\alpha = 0.05$.

(b) **(2 marks)** Based on the test derived in part (a), determine the minimum sample size n required to obtain a power of at least 0.95 when $\theta_0 = 10$, $\theta_1 = 25$ and $\alpha = 0.05$.

(c) **(4 marks)** Find the expression of likelihood ratio, $\lambda(X_1, \dots, X_n, Y_1, \dots, Y_m)$, for testing $H_0 : \theta = \mu$ against $H_1 : \theta \neq \mu$.

(d) **(4 marks)** Hence or otherwise, find the likelihood ratio test for testing $H_0 : \theta = \mu$ against $H_1 : \theta \neq \mu$ at $\alpha = 0.1$, $n = 10$ and $m = 15$.

(e) **(4 marks)** Derive the approximate large sample likelihood ratio test for testing $H_0 : \theta = \mu$ against $H_1 : \theta \neq \mu$ at the significance level of α and a large values of n and m. Make your conclusion at $\alpha = 0.05$ if $\sum x_i = 100$, $\sum y_i = 50$ and $n = m = 50$. Write down the value of test statistic and critical value clearly.

Solutions:

(a)

$$f(x;\theta) = \theta e^{-\theta x} = exp(log(\theta) - \theta x)$$

Therefore, f belongs to exponential family with $a(\theta) = log(\theta), b(x) = 0, c(\theta) = -\theta, d(x) = x$. Since $c(\theta) = -\theta$ is decreasing, the UMP test for $\left\{ \begin{array}{l} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right.$ has critical region $C_1 = \{x : \sum_{i=1}^n d(X_i) \leq k\} = \{x : \sum_{i=1}^n X_i \leq k\}$ for some k s.t. $P(x \in C_1 | \theta \in \Theta_0) = \alpha$. By MGF technique, $2\theta \sum_{i=1}^n X_i \sim \chi^2_{(2n)}$.

$$\begin{aligned} C_1 = \{x : \sum_{i=1}^n X_i \leq k\} &= \{x : 2\theta \sum_{i=1}^n X_i \leq 2\theta k\} \\ P(2\theta_0 \sum_{i=1}^n X_i \leq 2\theta_0 k | \theta = \theta_0) &= \alpha \\ \Rightarrow 2\theta_0 k = \chi^2_{2n,1-\alpha} \qquad k &= \frac{1}{2\theta_0} \chi^2_{2n,1-\alpha} \\ \Rightarrow C_1 &= \{x : \sum_{i=1}^n X_i \leq \frac{1}{2\theta_0} \chi^2_{2n,1-\alpha}\} \end{aligned}$$

Then for $\left\{ \begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right.$. Consider C_1 as indicated above. Is $\sup\{P(x \in C_1 | \theta \in \Theta_0\} = \alpha$?

$$\sup_{\theta \leq \theta_0} \{P(\sum_{i=1}^n X_i \leq \frac{1}{2\theta_0} \chi^2_{2n,1-\alpha}) | \theta)\} = \sup_{\theta \leq \theta_0} \{P(2\theta \sum_{i=1}^n X_i \leq \frac{\theta}{\theta_0} \chi^2_{2n,1-\alpha}) | \theta)\} = \alpha$$

where $\frac{\theta}{\theta_0} \leq 1$. Therefore, $C_1 = \{x : \sum_{i=1}^n X_i \leq \frac{1}{2\theta_0} \chi^2_{2n,1-\alpha}\}$ is also the critical region for the UMP test for $\left\{ \begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right.$.

(b) $\theta_0 = 10, \theta_1 = 25, \alpha = 0.05, 1 - \beta \geq 0.95$.

$$\begin{aligned} 1 - \beta \geq 0.95 \quad &\Leftrightarrow \quad P(\sum_{i=1}^n X_i \leq \frac{1}{2\theta_0} \chi^2_{2n,1-\alpha}) | \theta = \theta_1) \geq 0.95 \\ & \quad\quad P(2\theta_1 \sum_{i=1}^n X_i \leq \frac{\theta_1}{\theta_0} \chi^2_{2n,1-\alpha}) | \theta = \theta_1) \geq 0.95 \\ &\Rightarrow \quad P(\chi^2_{(2n)} \leq 2.5 \chi^2_{2n,0.95}) \geq 0.95 \\ &\Rightarrow \quad 2.5 \chi^2_{2n,0.95} \geq \chi^2_{2n,0.05} \Rightarrow 2n = 27 \Rightarrow n = 14 \end{aligned}$$

The minimum sample size n is 14.

(c)

$$f(\mathbf{x}, \mathbf{y}; \theta, \mu) = \theta^n e^{-\theta \sum_{i=1}^n x_i} \mu^m e^{-\mu \sum_{i=1}^m y_i} = L(\theta, \mu; \mathbf{x}, \mathbf{y})$$

For numerator: define $\theta = \mu = \lambda$.

$$\begin{aligned} L(\lambda; \mathbf{X}, \mathbf{Y}) &= \lambda^{m+n} e^{-\lambda(\sum_{i=1}^n x_i + \sum_{i=1}^m y_i)} \\ l = logL &= (m+n) log(\lambda) - \lambda(\sum_{i=1}^n x_i + \sum_{i=1}^m y_i) \\ \frac{\partial l}{\partial \lambda} &= \frac{m+n}{\lambda} - (\sum_{i=1}^n x_i + \sum_{i=1}^m y_i) \\ \frac{\partial l}{\partial \lambda} = 0 \quad &\Rightarrow \quad \hat{\lambda} = \frac{m+n}{\sum_{i=1}^n x_i + \sum_{i=1}^m y_i} \\ L(\hat{\lambda}; \mathbf{X}, \mathbf{Y}) &= (\frac{m+n}{\sum_{i=1}^n x_i + \sum_{i=1}^m y_i})^{m+n} exp(-m-n) \end{aligned}$$

For denominator:

$$\begin{aligned}
l = logL &= nlog(\theta) - \theta\sum_{i=1}^{n} x_i + mlog(\mu) - \mu\sum_{i=1}^{m} y_i \\
\left\{\begin{array}{l} \frac{\partial l}{\partial\theta} = 0 \\ \frac{\partial l}{\partial\mu} = 0 \end{array}\right. &\Rightarrow \left\{\begin{array}{l} \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \\ \frac{m}{\mu} - \sum_{i=1}^{m} y_i = 0 \end{array}\right. \Rightarrow \left\{\begin{array}{l} \hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} \\ \hat{\mu} = \frac{m}{\sum_{i=1}^{m} y_i} \end{array}\right. \\
L(\hat{\theta}, \hat{\mu}; \mathbf{X}, \mathbf{Y}) &= (\frac{n}{\sum_{i=1}^{n} x_i})^n(\frac{m}{\sum_{i=1}^{m} y_i})^m exp(-m-n)
\end{aligned}$$

Therefore, the expression of Likelihood ratio test statistic is

$$\begin{aligned}
\lambda(\mathbf{X}, \mathbf{Y}) = \frac{L(\hat{\lambda}; \mathbf{X}, \mathbf{Y})}{L(\hat{\theta}, \hat{\mu}; \mathbf{X}, \mathbf{Y})} &= \frac{(\frac{m+n}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i})^{m+n}}{(\frac{n}{\sum_{i=1}^{n} x_i})^n(\frac{m}{\sum_{i=1}^{m} y_i})^m} \\
&= \frac{(m+n)^{m+n}}{m^m n^n}\frac{(\sum_{i=1}^{n} x_i)^n(\sum_{i=1}^{m} y_i)^m}{(\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i)^{m+n}} \\
&= \frac{(m+n)^{m+n}}{m^m n^n}\frac{(\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{m} y_i})^n}{(\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{m} y_i} + 1)^{m+n}}
\end{aligned}$$

(d) $\lambda(\mathbf{X}, \mathbf{Y}) \leq$ k $\Leftrightarrow \frac{m}{n}\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{m} y_i} \leq k_1$ or $\frac{m}{n}\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{m} y_i} \geq k_2$ for some constants k_1 and k_2. Under $H_0 : \mu = \theta = \lambda$,

$$\frac{(\sum_{i=1}^{n} x_i)/2n}{(\sum_{i=1}^{m} y_i)/2m} = \frac{(2\lambda\sum_{i=1}^{n} x_i)/2n}{(2\lambda\sum_{i=1}^{m} y_i)/2m} \sim F_{2n,2m}$$

So take

$$\begin{aligned}
k_1 &= F_{2n,2m,1-\frac{\alpha}{2}} = F_{20,30,1-0.05} = \frac{1}{F_{30,20,0.05}} = \frac{1}{2.04} = 0.4902 \\
k_2 &= F_{2n,2m,\frac{\alpha}{2}} = F_{20,30,0.05} = 1.93
\end{aligned}$$

Therefore, the likelihood ratio test for $\left\{\begin{array}{l} H_0 : \mu = \theta \\ H_1 : \mu \neq \theta \end{array}\right.$ has critical region C_1=$\{(\mathbf{X}, \mathbf{Y} : 1.5\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{m} y_i} \leq 0.4902 \; or \; 1.5\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{m} y_i} \geq 1.93\}$,i.e.$C_1$=$\{(\mathbf{X}, \mathbf{Y} : \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{m} y_i} \leq 0.3268 \; or \; \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{m} y_i} \geq 1.2867\}$

(e)

$$\begin{aligned}
-2log\lambda(\mathbf{X}, \mathbf{Y}) &= -2\{(m+n)log(m+n) - mlog(m) - nlog(n) \\
&\quad +nlog(\sum_{i=1}^{n} x_i) + mlog(\sum_{i=1}^{m} y_i) - (m+n)log(\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i)\} \sim \chi_1^2 \\
\Rightarrow C_1 &= \{\mathbf{x} : -2log\lambda(\mathbf{X}, \mathbf{Y}) \geq \chi_{1,\alpha}^2\}
\end{aligned}$$

$\chi_{1,0.05}^2 = 3.841, n = m = 50, \sum_{i=1}^{n} x_i = 100, \sum_{i=1}^{m} y_i = 50.$

$-2log\lambda(\mathbf{X}, \mathbf{Y})$ = -2(100log(100)-50log(50)-50log(50) + 50log(100)+50log(50) - (100)log(150))= 11.7783 > $\chi_{1,0.05}^2 \Rightarrow$ reject H_0.

4. (a) Individuals were classified according to their answers of the question: “Did you get married before you were 25?” and according to which ethnic group they are, i.e.,

	Group A	Group B
Yes	x_{11}	x_{12}
No	x_{21}	x_{22}

Let $X = (X_{11}, X_{12}, X_{21}, X_{22}) \sim$ multinomial $(n, P_{11}, P_{12}, P_{21}, P_{22})$.

Suppose that 100 females sampled from each of two ethnic groups. 62 females and 29 females said "Yes" for group A and group B, respectively. Perform the following tests at $\alpha = 0.05$. State clearly the hypothesis statements, value of test statistic, critical value and your conclusion for each test. There is no need to make the continuity correction for the Pearson's goodness of fit test.

i. (**6 marks**) Test whether the null hypothesis $H_0 : P_{11} = p^2, P_{12} = p(1-p), P_{21} = p(1-p), P_{22} = (1-p)^2$ is true by

A. Approximate large sample likelihood ratio test;

B. Pearson's goodness of fit test.

ii. (**6 marks**) Test whether the proportions of answering "Yes" for the two groups are equal at 0.05 level of significance by

A. z test;

B. Approximate large sample likelihood ratio test;

C. Pearson's goodness of fit test.

(b) Let $(X_1, ..., X_n)$ be a random sample from $U(0, \theta)$ with $\theta > 0$.

i. (**4 marks**) Find UMP test at the level of significance α for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

ii. **(2 marks)** Based on the test derived in part (i), calculate the minimum sample size n such that the test for testing $H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$ has a power of at least 0.98 at $\theta_1 = \frac{3}{4}$, where $\theta_1 \in \Theta_1$, when $\alpha = 0.05$.

iii. **(2 marks)** Based on the test derived in part (i), calculate the power at $\theta_1 = \frac{2}{3}$, where $\theta_1 \in \Theta_1$, for testing $H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$ when $\alpha = 0.05$ and $n = 10$.

Solutions:

(a) In order to handle this question, first we need to find the MLE under H_0.

$$\begin{aligned}
L(p) &= constant(p^2)^{x_{11}}(p(1-p))^{x_{12}}(p(1-p))^{x_{12}}((1-p)^2)^{x_{22}} \\
l = logL &= constant + (2x_{11} + x_{12} + x_{21})logp + (x_{12} + x_{21} + 2x_{22})log(1-p) \\
\frac{dl}{dp} = 0 &\Rightarrow \hat{p} = \frac{2x_{11} + x_{12} + x_{21}}{2(x_{11} + x_{12} + x_{21} + x_{22})} = \frac{191}{400} = 0.4775
\end{aligned}$$

i. A. hypothesis statement: $\begin{cases} H_0 & : \quad P_{11} = p^2, \quad P_{12} = p(1-p), P_{21} = p(1-p), P_{22} = (1-p)^2 \\ H_1 & : \quad otherwise \end{cases}$

Test statistic:

$$T = 2\sum_i \sum_j x_{ij} log \frac{x_{ij}}{n\hat{p_{ij}}} = 23.2054$$

Critical Value: $\chi^2_{4-1-1,0.05} = \chi^2_{2,0.05} = 5.991$. Since T = 23.2054 $> \chi^2_{2,0.05} = 5.991$, we reject H_0.

B. hypothesis statement: $\begin{cases} H_0 & : \quad P_{11} = p^2, \quad P_{12} = p(1-p), P_{21} = p(1-p), P_{22} = (1-p)^2 \\ H_1 & : \quad otherwise \end{cases}$

Test statistic:

$$T = \sum_i \sum_j \frac{(x_{ij} - n\hat{p_{ij}})^2}{n\hat{p_{ij}}} = 22.4126$$

Critical Value: $\chi^2_{4-1-1,0.05} = \chi^2_{2,0.05} = 5.991$. Since T = 22.4126 $> \chi^2_{2,0.05} = 5.991$, we reject H_0.

ii. The hypothesis for the following three parts art the same,

i.e. $\begin{cases} H_0 & : \quad \text{the proportions of ansering "Yes" for the two groups are equal} \\ H_1 & : \quad \text{the proportions of ansering "Yes" for the two groups are not equal} \end{cases}$.

A. For the z-test:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{62 + 29}{100 + 100} = 0.455$$
$$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = 4.6859$$

Critical value is $z_{0.025} = 1.96 < 4.6859$. ⇒ reject H_0.

B.

$$T = 2\sum_i \sum_j x_{ij} log \frac{x_{ij}}{\frac{a_i b_j}{n}} = 22.3935$$

Critical Value is $\chi^2_{(2-1)(2-1),0.05} = \chi^2_{1,0.05} = 3.841 <$ T=22.3935 ⇒ reject H_0.

C.

$$T = n(\sum_i \sum_j \frac{x_{ij}^2}{a_i b_j} - 1) = 21.9579$$

Critical Value is $\chi^2_{(2-1)(2-1),0.05} = \chi^2_{1,0.05} = 3.841 <$ T=21.9579 ⇒ reject H_0.

(b)

$$f(x;\theta) = \frac{1}{\theta} I_{(0 \le x \le \theta)}$$
$$f(\mathbf{x};\theta) = \frac{1}{\theta^n} I_{(x_{(n)} \le \theta)}$$

i. For testing $\begin{cases} H_0 & : & \theta = \theta_0 \\ H_1 & : & \theta = \theta_1 \quad (\theta_1 > \theta_0) \end{cases}$.
By N-P theorem, the critical region for this hypothesis is

$$\begin{aligned} C_1 &= \{\mathbf{x} : \frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_1)} \leq k\} \\ \frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_1)} &= \frac{\theta_1^n I_{(x_{(n)} \leq \theta_0)}}{\theta_0^n I_{(x_{(n)} \leq \theta_1)}} \end{aligned}$$

where $\frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_1)}$ is an decreasing function of $x_{(n)}$. Therefore,

$$\begin{aligned} \frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_1)} \leq k &\Leftrightarrow X_{(n)} \geq k' \\ P(X_{(n)} \geq k' | \theta_0) = \alpha &\Rightarrow 1 - (\frac{k'}{\theta_0})^n = \alpha \\ &\Rightarrow k' = \theta_0 (1-\alpha)^{\frac{1}{n}} \end{aligned}$$

Therefore, the critical region for the UMP test for $\begin{cases} H_0 & : & \theta = \theta_0 \\ H_1 & : & \theta = \theta_1 \quad (\theta_1 > \theta_0) \end{cases}$ is $C_1 = \{\mathbf{x} : X_{(n)} \geq \theta_0(1-\alpha)^{\frac{1}{n}}\}$.

Since C_1 is independent of θ_1, C_1 is also the critical region for the UMP test for $\begin{cases} H_0 & : & \theta = \theta_0 \\ H_1 & : & \theta > \theta_0 \end{cases}$.
Then consider $\begin{cases} H_0 & : & \theta \leq \theta_0 \\ H_1 & : & \theta > \theta_0 \end{cases}$, where $\Theta_0 = \{\theta : \theta \leq \theta_0\}$ and C_1 we achieve above. The following holds:

$$\begin{aligned} \sup\{P(\mathbf{x} \in C_1 | \theta \in \Theta_0)\} &= \sup_{\theta \leq \theta_0} \{P(X_{(n)} \geq \theta_0(1-\alpha)^{\frac{1}{n}} | \theta \leq \theta_0)\} \\ &= \sup_{\theta \leq \theta_0} \{1 - P(X_{(n)} \leq \theta_0(1-\alpha)^{\frac{1}{n}} | \theta \leq \theta_0)\} \\ &= \sup_{\theta \leq \theta_0} \{1 - (\frac{\theta_0(1-\alpha)^{\frac{1}{n}}}{\theta})^n\} \\ &= \sup_{\theta \leq \theta_0} \{1 - (\frac{\theta_0}{\theta_1})^n (1-\alpha\} = \alpha \end{aligned}$$

Thus, C_1 is also the critical region for the UMP test for $\begin{cases} H_0 & : & \theta \leq \theta_0 \\ H_1 & : & \theta > \theta_0 \end{cases}$.

ii. $\theta_0 = 0.5, \alpha = 0.05, C_1 = \{\mathbf{x} : X_{(n)} \geq 0.5 * 0.95^{\frac{1}{n}}\}$.

$$\begin{aligned} 1 - \beta = 0.98 \quad at\ \theta_1 = 0.75 &\Leftrightarrow P(\mathbf{x} \in C_1 | \theta_1 = 0.75) = 0.98 \\ &\Leftrightarrow 1 - (\frac{0.5(1-0.05)^{\frac{1}{n}}}{0.75})^n = 0.98 \\ &\Leftrightarrow 1 - (\frac{2}{3})^n 0.95 = 0.98 \Rightarrow n \approx 10 \end{aligned}$$

Therefore, the minimum sample size should be 10.

iii. power at $\theta_1 = \frac{2}{3}$ is given by

$$\begin{aligned} P(\mathbf{x} \in C_1 | \theta_1 = \frac{2}{3}) &= P(X_{(n)} \geq \theta_0 (1-\alpha)^{\frac{1}{n}} | \theta_1 = \frac{2}{3}) \\ &= 1 - (\frac{\theta_0}{\theta_1})^n (1-\alpha) \\ &= 1 - (\frac{\frac{1}{2}}{\frac{2}{3}})^{10} (1 - 0.05) = 0.9465 \end{aligned}$$

5. (**Bonus**) Let $(X_1, ..., X_n)$ be a random sample from a location distribution family

$$f(x;\theta) = \frac{1}{\theta} \exp\left(-\frac{x-\delta}{\theta} \right) I(x \geq \delta) \ .$$

(a) (**5 marks**) Suppose that θ is known. Find an exact likelihood ratio test at the level of significance α for testing $H_0 : \delta = \delta_0$ versus $H_1 : \delta \neq \delta_0$.

(b) (**5 marks**) When both θ and δ are unknown, find an exact likelihood ratio test at the level of significance α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

Solutions:

$$f(\mathbf{x}; \theta, \delta) = \frac{1}{\theta^n} exp(-\sum \frac{x_i - \delta}{\theta}) I_{(x_{(1)} \geq \delta)}$$

(a) When θ is known,

$$L(\delta) = \frac{1}{\theta^n} exp(-\frac{1}{\theta} \sum x_i + \frac{n}{\theta} \delta) I_{(\delta \leq x_{(1)})}$$

Draw the graph of $L(\delta)$, it is easy to see $L(\delta)$ is increasing and achieves maximum at $\delta = x_{(1)}$,i.e. $\hat{\delta} = x_{(1)}$. Therefore, the exact likelihood ratio test statistic is

$$\begin{aligned} \lambda(\mathbf{X}) = \frac{L(\delta_0)}{L(\hat{\delta})} &= \frac{\frac{1}{\theta^n} exp(-\frac{1}{\theta} \sum x_i + \frac{n}{\theta} \delta_0) I_{(\delta_0 \leq x_{(1)})}}{\frac{1}{\theta^n} exp(-\frac{1}{\theta} \sum x_i + \frac{n}{\theta} \hat{\delta}) I_{(\hat{\delta} \leq x_{(1)})}} \\ &= exp(\frac{n}{\theta} (\delta_0 - x_{(1)}) I_{(\delta_0 \leq x_{(1)})} \end{aligned}$$

Since $\lambda(\mathbf{X})$ is decreasing in $x_{(1)}$, the rejection region is $C_1 = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\} = \{\mathbf{x} : x_{(1)} \geq k'\}$, where k' is determined by

$$\begin{aligned} & \sup P(\mathbf{X} \in C_1 | \delta = \delta_0) = \alpha \\ \Rightarrow \ & P(x_{(1)} \geq k' | \delta = \delta_0)) = \alpha \\ \Rightarrow \ & exp(-\frac{n}{\theta}(k' - \delta_0)) = \alpha \\ \Rightarrow \ & k' = \delta_0 - \frac{\theta}{n} log\alpha \end{aligned}$$

Therefore, the exact likelihood ratio test at the level of significance α has critical region $C_1 = \{\mathbf{x} : x_{(1)} \geq \delta_0 - \frac{\theta}{n} log\alpha\}$.

(b)

$$L(\theta, \delta) \;=\; \frac{1}{\theta^n} exp(-\frac{1}{\theta}\sum x_i + \frac{n}{\theta}\delta) I_{(\delta \leq x_{(1)})}$$

Numerator: $\theta = \theta_0$ Applying the results in part(a), we have

$$\begin{aligned} \hat{\delta}_0 &= x_{(1)} \\ L(\theta_0, \hat{\delta}_0) &= \frac{1}{\theta_0^n} exp(-\frac{1}{\theta_0}\sum x_i + \frac{n}{\theta_0} x_{(1)}) = \frac{1}{\theta_0^n} exp(-\frac{n}{\theta_0}(\bar{x} - x_{(1)})) \end{aligned}$$

Denominator: For fixed θ, $L(\theta, \delta)$ is increasing in δ, and

$$\frac{\partial L}{\partial \theta} = -\frac{n}{\theta^{n+2}}(\theta - (\bar{x} - \delta)) exp(-\frac{n}{\theta}(\bar{x} - \delta))$$

Therefore the MLE in whole parameter space Θ is

$$\begin{aligned} \hat{\delta} &= x_{(1)} \\ \hat{\theta} &= \bar{x} - x_{(1)} \end{aligned}$$

So the likelihood test statistic is

$$\begin{aligned} \lambda(\mathbf{X}) &= \frac{\hat{\theta}^n}{\theta_0^n} \frac{exp(-\frac{1}{\theta_0}\sum x_i + \frac{n}{\theta_0} x_{(1)})}{exp(-\frac{1}{\hat{\theta}}\sum x_i + \frac{n}{\hat{\theta}} x_{(1)})} \\ &= (\frac{\bar{x} - x_{(1)}}{\theta_0})^n exp(-\frac{n}{\theta_0}(\bar{x} - x_{(1)}) + n) \end{aligned}$$

The critical region is $C_1 = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\} = \{\mathbf{x} : \frac{n}{\theta_0}(\bar{x} - x_{(1)}) \leq k_1 \; or \; \frac{n}{\theta_0}(\bar{x} - x_{(1)}) \geq k_2\}$. Under H_0, $2\frac{n}{\theta_0}(\bar{x} - x_{(1)})$ follows χ^2_{2n-2}(The proof is similar to the proof of theorem 7.3. And notice that $\frac{2}{\theta_0}\sum(X_i - \delta)$ follows χ^2_{2n} and $\sum(X_i - X_{(1)}) \perp (X_{(1)} - \delta)$). Therefore, reject H_0 if $2\frac{n}{\theta_0}(\bar{x} - x_{(1)}) < \chi^2_{2n-2,1-\alpha/2}$ or $2\frac{n}{\theta_0}(\bar{x} - x_{(1)}) > \chi^2_{2n-2,\alpha/2}$.