FIT5197 2025 S1 Assignment - Covers the lecture and tutorial materials up to, and including, week 8

SPECIAL NOTE: Please refer to the assessment page for rules, general guidelines and marking rubrics of the assessment (the marking rubric for the kaggle competition part will be released near the deadline in the same page). Failure to comply with the provided information will result in a deduction of mark (e.g., late penalties) or breach of academic integrity.

Part 1 Point Estimation (30 marks)

WARNING: you should strictly follow the 3-steps strategy as detailed in question 2 of week 5 tutorial (or any answer formats presented in the Week 5 quiz) to answer for the questions that are related to MLE estimators presented in this part. Any deviations from the answer format might result in a loss of marks!

Question 1 (7.5 marks)

Let $X \sim \mathcal{IG}\left(\theta : (\mu, \lambda)\right), \ \forall \mu > 0 \text{ and } \lambda > 0$. This means the random varible X follows the **inverse Gaussian distribution** with the set $(\theta : (\mu, \lambda))$ acting as the parameters of said distribution. Given that we observe a sample of size n that is independently and identically distributed from this distribution (i.i.d), $\mathbf{x} = (x_1, \ldots, x_n)$, please find the maximum likelihood estimate for μ and λ, that is μ_{MLE} and λ_{MLE}. The probability density function (**PDF**) is as follows:

$$f(x \mid \mu, \lambda) = \begin{cases} \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} e^{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

ANSWER

Step 1. Likelihood Function

The likelihood function $f(x \mid \mu, \lambda)$ is:

$$f(x \mid \mu, \lambda) = \prod_{i=1}^{n} \left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2} \exp\left(\frac{-\lambda(x_i - \mu)^2}{2\mu^2 x_i}\right)$$

$$f(x \mid \mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{n/2} \left(\prod_{i=1}^{n} x_i^{-3/2}\right) \exp\left(-\frac{\lambda}{2\mu^2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{x_i}\right)$$

Step 2. Negative Log-Likelihood Function

The negative log-likelihood function $\ell(x;\mu,\lambda) = -\ln(f(x \mid \mu,\lambda))$ is:

$$\ell(x;\mu,\lambda) = -\frac{n}{2}\ln\left(\frac{\lambda}{2\pi}\right) + \frac{3}{2}\sum_{i=1}^{n}\ln(x_i) + \frac{\lambda}{2\mu^2}\sum_{i=1}^{n}\left(x_i - 2\mu + \frac{\mu^2}{x_i}\right)$$

$$\ell(x;\mu,\lambda) = -\frac{n}{2}\ln(\lambda) + \frac{n}{2}\ln(2\pi) + \frac{3}{2}\sum_{i=1}^{n}\ln(x_i) + \frac{\lambda}{2}\left(\mu^{-2}\sum_{i=1}^{n}x_i - 2\mu^{-1}\sum_{i=1}^{n}1 + \sum_{i=1}^{n}\right.$$

or alternatively:

$$\ell(x;\mu,\lambda) = -\frac{n}{2}\ln(\lambda) + \frac{n}{2}\ln(2\pi) + \frac{3}{2}\sum_{i=1}^{n}\ln(x_i) + \frac{\lambda}{2\mu^2}\left(n\bar{x} - 2n\mu + n\mu^2\overline{1/x}\right)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ and $\overline{1/x} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$.

Step 3. Finding the MLE

Derivative with respect to λ:

$$\frac{\partial\ell(x;\lambda)}{\partial\lambda} = -\frac{n}{2\lambda} + \frac{1}{2\mu^2}\left(n\bar{x} - 2n\mu + n\mu^2\overline{1/x}\right)$$

Setting $\frac{\partial\ell(x;\lambda)}{\partial\lambda} = 0$:

$$\frac{n}{\lambda} = \frac{1}{\mu^2}\left(n\bar{x} - 2n\mu + n\mu^2\overline{1/x}\right)$$

$$\hat{\lambda} = \frac{n\mu^2}{n\bar{x} - 2n\mu + n\mu^2\overline{1/x}} = \frac{\mu^2}{\bar{x} - 2\mu + \mu^2\overline{1/x}}$$

Derivative with respect to μ:

$$\ell(x;\mu,\lambda) = -\frac{n}{2}\ln(\lambda) + \frac{n}{2}\ln(2\pi) + \frac{3}{2}\sum_{i=1}^{n}\ln(x_i) + \frac{\lambda}{2}\left(\mu^{-2}\sum_{i=1}^{n}x_i - 2\mu^{-1}\sum_{i=1}^{n}1 + \sum_{i=1}^{n}\right.$$

$$\frac{\partial\ell(x;\mu)}{\partial\mu} = -\frac{\lambda}{2}\left(-2\mu^{-3}\sum x_i + 2\mu^{-2}\sum 1\right)(-1) = \frac{\lambda}{2}\left(-2\mu^{-3}\sum x_i + 2\mu^{-2}\sum 1\right)$$

Setting $\frac{\partial\ell(x;\mu)}{\partial\mu}$ = 0:

$$-2\mu^{-3}(n\bar{x}) + 2\mu^{-2}(n) = 0$$

$$-\frac{2n\bar{x}}{\mu^3} + \frac{2n}{\mu^2} = 0$$

$$-\bar{x} + \mu = 0 \implies \hat{\mu} = \bar{x}$$

Substituting $\hat{\mu} = \bar{x}$ back into the equation for $\hat{\lambda}$:

$$\hat{\lambda} = \frac{\bar{x}^2}{\bar{x} - 2\bar{x} + \bar{x}^2\overline{1/x}} = \frac{\bar{x}^2}{-\bar{x} + \bar{x}^2\overline{1/x}} = \frac{\bar{x}^2}{\bar{x}(\bar{x}\overline{1/x} - 1)} = \frac{\bar{x}}{\bar{x}\overline{1/x} - 1}$$

$$\hat{\lambda} = \frac{1}{\overline{1/x} - 1/\bar{x}}$$

Question 2 (7.5 marks)

Suppose that we know that the random variable $X \sim \mathcal{D}ist(\mu = \theta, \sigma^2 = \theta^2)$ follows the PDF given below:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}\exp(-\frac{x}{\theta}) & x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Given a sample of n i.i.d observations $\mathbf{x} = (x_1, \ldots, x_n)$ from this distribution, please answer the following questions:

(a) Derive the MLE estimator for θ, i.e., $\hat{\theta}_{\text{MLE}}$, and show that it is unbiased. [2.5 Marks]

(b) Find an estimator with better MSE (i.e smaller MSE) compared to the $\hat{\theta}_{\text{MLE}}$ obtained from (a). [5 Marks]

ANSWER

Q2(a) MLE Estimator $\hat{\theta}_{\text{MLE}}$

Step 1. Likelihood function, $f(x|\theta)$

$$f(x|\theta) = \prod_{i=1}^{n}\left(\frac{1}{\theta}\exp(-\frac{x_i}{\theta})\right)$$

$$f(x|\theta) = \left(\frac{1}{\theta}\right)^n \exp\left(-\frac{\sum_{i=1}^n x_i}{\theta}\right)$$

$$f(x|\theta) = \frac{1}{\theta^n}\exp\left(-\frac{\sum_{i=1}^n x_i}{\theta}\right)$$

Step 2. Negative log likelihood function, $\ell(x;\theta)$

$$\ell(x;\theta) = \ln\left(\frac{1}{\theta^n}\exp\left(-\frac{\sum_{i=1}^n x_i}{\theta}\right)\right)$$

$$\ell(x;\theta) = -n\ln(\theta) - \frac{\sum_{i=1}^n x_i}{\theta}$$

Step 3. Finding the MLE

Derivative of negative log likelihood function with respect to θ:

$$\frac{d\ell(x;\theta)}{d\theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}$$

Set derivative = 0 and solve for θ:

$$-\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0$$

$$-n\theta + \sum_{i=1}^{n} x_i = 0$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \text{sample mean, } \bar{x}$$

For MLE = $\frac{\sum_{i=1}^{n} x_i}{n}$ to be unbiased, the expected value of the MLE must be equal to $f(x|\theta)$

$E[f(x|\theta)] = \theta$ as it is an exponential distribution with parameter $1/\theta$
$E[\hat{\theta}_{\text{MLE}}]$:

$$E[\hat{\theta}_{\text{MLE}}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

Since $E[X_i] = \theta$ for all i:

$$E[\hat{\theta}] = \frac{1}{n}\sum_{i=1}^{n} \theta = \frac{1}{n}(n\theta) = \theta$$

Therefore, the MLE $\hat{\theta} = \bar{X}$ is an unbiased estimator for θ.

Q2(b) Estimator with smaller MSE

The MSE of an estimator $\hat{\theta}$ is $MSE[\hat{\theta}] = bias_\theta(\hat{\theta})^2 + V_\theta[\hat{\theta}]$.
For the MLE $\hat{\theta}_{MLE} = \bar{X}$, $bias_\theta[\hat{\theta}_{MLE}] = 0$ and $V_\theta[\bar{X}] = \frac{\theta^2}{n}$, so $MSE[\hat{\theta}_{MLE}] = \frac{\theta^2}{n}$.

If we modify the estimator to be $\hat{\theta}_b = c\bar{X}$:

- The bias is $bias_\theta(\hat{\theta}_b) = (c-1)\theta$.
- The variance is $V_\theta[\hat{\theta}_b] = c^2\frac{\theta^2}{n}$.
- The MSE is $MSE[\hat{\theta}_b] = (c-1)^2\theta^2 + c^2\frac{\theta^2}{n} = \theta^2\left(\frac{c^2}{n} + c^2 - 2c + 1\right)$.

To minimize the MSE with respect to c, we take the derivative of $MSE[\hat{\theta}_b] = \theta^2\left(\frac{c^2}{n} + c^2 - 2c + 1\right)$ with respect to c and set it to zero:

$$\frac{d}{dc}MSE[\hat{\theta}_b] = \theta^2\left(\frac{2c}{n} + 2c - 2\right) = 0$$

$$\frac{2c}{n} + 2c - 2 = 0$$

$$c\left(\frac{1}{n}+1\right)=1$$

$$c=\frac{1}{\frac{1+n}{n}}=\frac{n}{n+1}$$

This yields $c=\frac{n}{n+1}$, and so, the optimal biased estimator is $\hat{\theta}^{*}=\frac{n}{n+1}\bar{X}$.

Substituting c back, the MSE of this estimator is:

$$MSE[\hat{\theta}^{*}]=\left(\frac{n}{n+1}\right)^{2}\frac{\theta^{2}}{n}+\left(\frac{n}{n+1}-1\right)^{2}\theta^{2}$$

$$MSE[\hat{\theta}^{*}]=\frac{n^{2}}{(n+1)^{2}}\frac{\theta^{2}}{n}+\left(-\frac{1}{n+1}\right)^{2}\theta^{2}$$

$$MSE[\hat{\theta}^{*}]=\frac{n\theta^{2}}{(n+1)^{2}}+\frac{\theta^{2}}{(n+1)^{2}}=\frac{(n+1)\theta^{2}}{(n+1)^{2}}=\frac{\theta^{2}}{n+1}$$

Since $\frac{\theta^{2}}{n+1}<\frac{\theta^{2}}{n}$ for $n>0$, the biased estimator $\hat{\theta}^{*}=\frac{n}{n+1}\bar{X}$ has a smaller MSE than the unbiased MLE $\bar{X}$.

Question 3 (7.5 marks)

Suppose that we know that a random variable X follows the distribution given below:

$$f\left(x|\theta\right)=\frac{\binom{2}{x}\theta^{x}(1-\theta)^{2-x}}{1-(1-\theta)^{2}},\ x=\{1,2\}$$

Imagine that we observe a sample of $\mathbf{n}$ i.i.d random varaibles $\mathbf{x}=(x_{1},\ldots,x_{n})$ and want to model them using this distribution. Please use the concept of maximum likelihood to estimate for the parameter θ.

ANSWER

Step 1. Likelihood Function

Likelihood function $f(x|\theta)=\prod_{i=1}^{n}f(x_{i}|\theta)$:

$$f(x|\theta)=\prod_{i=1}^{n}\frac{\binom{2}{x_{i}}\theta^{x_{i}}(1-\theta)^{2-x_{i}}}{1-(1-\theta)^{2}}$$

$$f(x|\theta)=\frac{\prod_{i=1}^{n}\binom{2}{x_{i}}\cdot\theta^{\sum_{i=1}^{n}x_{i}}\cdot(1-\theta)^{\sum_{i=1}^{n}(2-x_{i})}}{\left(1-(1-\theta)^{2}\right)^{n}}$$

Let $T=\sum_{i=1}^{n}x_{i}$. Substituting this, we get:

$$f(x|\theta) = \frac{\left(\prod_{i=1}^{n} \binom{2}{x_i}\right) \theta^T (1-\theta)^{2n-T}}{\left(1-(1-\theta)^2\right)^n}$$

Expanding the denominator, $1-(1-\theta)^2 = 1-(1-2\theta+\theta^2) = 2\theta-\theta^2 = \theta(2-\theta)$. This becomes:

$$f(x|\theta) = \frac{\left(\prod_{i=1}^{n} \binom{2}{x_i}\right) \theta^T (1-\theta)^{2n-T}}{(\theta(2-\theta))^n} = \frac{\left(\prod_{i=1}^{n} \binom{2}{x_i}\right) \theta^{T-n} (1-\theta)^{2n-T}}{(2-\theta)^n}$$

Step 2. Negative Log-Likelihood Function

The negative log-likelihood function $\ell(x;\theta) = -\ln(f(x|\theta))$ is:

$$\ell(x;\theta) = -\ln\left(\prod_{i=1}^{n} \binom{2}{x_i}\right) - (T-n)\ln(\theta) - (2n-T)\ln(1-\theta) + n\ln(2-\theta)$$

Let $-C = -\ln\left(\prod_{i=1}^{n} \binom{2}{x_i}\right)$, which is a constant with respect to θ. Then,

$$-\ell(\theta) = -C - (T-n)\ln(\theta) - (2n-T)\ln(1-\theta) + n\ln(2-\theta)$$

Step 3. Finding the MLE

Derivative of $\ell(x;\theta)$ with respect to θ:

$$\frac{d\ell(x;\theta)}{d\theta} = -\frac{T-n}{\theta} - \frac{2n-T}{1-\theta} - \frac{n(-1)}{2-\theta}$$

$$\frac{d\ell(x;\theta)}{d\theta} = -\frac{T-n}{\theta} - \frac{2n-T}{1-\theta} + \frac{n}{2-\theta}$$

Setting the derivative to zero and solving for θ:

$$-\frac{T-n}{\theta} - \frac{2n-T}{1-\theta} + \frac{n}{2-\theta} = 0$$

$$-(T-n)(1-\theta)(2-\theta) - (2n-T)\theta(2-\theta) + n\theta(1-\theta) = 0$$

$$-(T-n)(2-3\theta+\theta^2) - (4n\theta - 2n\theta^2 - 2T\theta + T\theta^2) + (n\theta - n\theta^2) = 0$$

$$-2(T-n) + 3(T-n)\theta - (T-n)\theta^2 - 4n\theta + 2n\theta^2 + 2T\theta - T\theta^2 + n\theta - n\theta^2 = 0$$

Expressing in terms of θ:

$$\theta^2(-T+n+2n-T+n) + \theta(3T-3n-4n+2T+n) + (-2T+2n) = 0$$

$$\theta^2(-2T+4n) + \theta(5T-6n) + (2n-2T) = 0$$

$$2(T-2n)\theta^2 - (5T-6n)\theta + 2(T-n) = 0$$

Solve for θ as a quadratic equation:

$$\hat{\theta} = \frac{(5T-6n) \pm \sqrt{(5T-6n)^2 - 4(2(T-2n))(2(T-n))}}{2(2(T-2n))}$$

$$\hat{\theta} = \frac{(5T - 6n) \pm \sqrt{(5T - 6n)^2 - 16(T - 2n)(T - n)}}{4(T - 2n)}$$

where $T = \sum_{i=1}^{n} x_i$ and θ is in the range $(0, 1)$

To confirm this is a minimum point of the negative log-likelihood, compute the second derivative:

$$\frac{d\ell(x;\theta)}{d\theta} = -\frac{T - n}{\theta} - \frac{2n - T}{1 - \theta} + \frac{n}{2 - \theta}$$

$$\frac{d^2\ell(x;\theta)}{d\theta^2} = \frac{T - n}{\theta^2} + \frac{2n - T}{(1 - \theta)^2} + \frac{n}{(2 - \theta)^2}$$

Since $n > 0$ and $x \in \{1, 2\}$, we have $n \leq T \leq 2n$. Each term in the second derivative is positive, and at least one term will be non-zero for $\theta \in (0, 1)$.
This proves the second derivative is positive, thus, it is a minimum point of the negative log-likelihood.

The final MLE $\hat{\theta}$ will be one of the roots of the quadratic equation that lies in the interval $(0, 1)$.

Question 4 (7.5 marks)

Suppose that we know that the random variable X follows the PDF given below:

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Given a sample of n i.i.d observations $\mathbf{x} = (x_1, \ldots, x_n)$ from this distribution, please answer the following questions:

(a) Derive the MLE estimator for θ, i.e., $\hat{\theta}_{\text{MLE}}$. [4.5 Marks]

(b) Show that the estimator $\hat{\theta} = \overline{X} - 1$ (where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$) is an unbiased and consistent estimator for the given distribution. [3 Marks]

ANSWER

Q4(a) MLE Estimator

Step 1. Likelihood Function

The likelihood function $f(x|\theta)$ is:

$$f(x|\theta) = \prod_{i=1}^{n} e^{-(x_i - \theta)} = \exp\left(-\sum_{i=1}^{n}(x_i - \theta)\right) = \exp\left(-\left(\sum_{i=1}^{n} x_i - n\theta\right)\right)$$

$$f(x|\theta) = \exp\big(-n\overline{x} + n\theta\big)$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.
This has the constraint $\theta \leq x_{(1)} = \min(x_1, \ldots, x_n)$, as $f(x_i|\theta)$ is non-zero only when $x_i \geq \theta$.

Step 2. Negative Log-Likelihood Function

The negative log-likelihood function $\ell(x;\theta)$ is:

$$\ell(x;\theta) = n\overline{x} - n\theta$$

Step 3. Finding the MLE

To find the value of θ that minimizes $\ell(x;\theta) = n\overline{x} - n\theta$, subject to the constraint $\theta \leq x_{(1)}$.
The first derivative of $\ell(x;\theta)$ with respect to θ is:

$$\frac{d\ell(x;\theta)}{d\theta} = -n$$

Since $n > 0$, the derivative is always negative, indicating that $\ell(x;\theta)$ is a strictly decreasing function of θ.

To minimize a strictly decreasing function over an interval, we need to choose the largest possible value of θ within that interval. The constraint is $\theta \leq x_{(1)}$, so the largest possible value for θ is $x_{(1)}$.

Therefore, the value of θ that minimizes the negative log-likelihood is:

$$\hat{\theta}_{\text{MLE}} = x_{(1)} = \min(X_1, \ldots, X_n)$$

Q4(b) Unbiased and consistent estimator

Expected Value of X

The expected value of X is $E[X|\theta] = \int_{\theta}^{\infty} xe^{-(x-\theta)}dx$.
Let $u = x - \theta$, so $x = u + \theta$ and $dx = du$. When $x = \theta$, $u = 0$. When $x \to \infty$, $u \to \infty$.

$$E[X|\theta] = \int_0^{\infty} (u+\theta)e^{-u}du = \int_0^{\infty} ue^{-u}du + \theta\int_0^{\infty} e^{-u}du$$

$\int_0^{\infty} e^{-u}du = [-e^{-u}]_0^{\infty} = 0 - (-1) = 1.$

And $\int_0^{\infty} ue^{-u}du = [-ue^{-u}]_0^{\infty} - \int_0^{\infty}(-e^{-u})du = 0 + \int_0^{\infty} e^{-u}du = 1$ (using integration by parts).

Therefore, $E[X|\theta] = 1 + \theta$.

Unbiasedness of $\hat{\theta} = \overline{X} - 1$

$$E[\hat{\theta}] = E[\overline{X} - 1] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i - 1\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] - 1$$

$$E[\hat{\theta}] = \frac{1}{n}\sum_{i=1}^{n}(\theta+1) - 1 = \frac{1}{n}(n(\theta+1)) - 1 = \theta + 1 - 1 = \theta$$

Since $E[\hat{\theta}] = \theta$, the estimator $\hat{\theta} = \overline{X} - 1$ is unbiased for θ.

Variance of X

To find the variance, we first need $E[X^2|\theta]$.

$$E[X^2|\theta] = \int_{\theta}^{\infty} x^2 e^{-(x-\theta)} dx$$

Let $u = x - \theta$, so $x = u + \theta$ and $dx = du$.

$$E[X^2|\theta] = \int_{0}^{\infty} (u+\theta)^2 e^{-u} du = \int_{0}^{\infty} (u^2 + 2u\theta + \theta^2) e^{-u} du$$

$$E[X^2|\theta] = \int_{0}^{\infty} u^2 e^{-u} du + 2\theta \int_{0}^{\infty} u e^{-u} du + \theta^2 \int_{0}^{\infty} e^{-u} du$$

We found $\int_0^\infty u e^{-u} du = 1$ and $\int_0^\infty e^{-u} du = 1$.
Using integration by parts for $\int_0^\infty u^2 e^{-u} du$:

Let $a = u^2$, $db = e^{-u} du$.
Then, $da = 2u\, du$, $b = -e^{-u}$.

$$\int_{0}^{\infty} u^2 e^{-u} du = [-u^2 e^{-u}]_0^\infty - \int_{0}^{\infty} (-e^{-u})(2u) du = 0 + 2\int_{0}^{\infty} u e^{-u} du = 2(1) = 2$$

So, $E[X^2|\theta] = 2 + 2\theta(1) + \theta^2(1) = 2 + 2\theta + \theta^2$. The variance is
$Var(X|\theta) = E[X^2|\theta] - (E[X|\theta])^2 = (2 + 2\theta + \theta^2) - (\theta+1)^2 = 1$.

Consistency of $\hat{\theta} = \overline{X} - 1$

The Mean Squared Error (MSE) is $MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2$. Since $\hat{\theta}$ is unbiased, $Bias(\hat{\theta}) = 0$, so $MSE(\hat{\theta}) = Var(\hat{\theta})$.

$$Var(\hat{\theta}) = Var(\overline{X} - 1) = Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$Var(\hat{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n} 1 = \frac{n}{n^2} = \frac{1}{n}$$

As $n \to \infty$, $MSE(\hat{\theta}) = Var(\hat{\theta}) = \frac{1}{n} \to 0$. Therefore, the estimator $\hat{\theta} = \overline{X} - 1$ is consistent for θ.

Part 2 Confidence Interval Estimation & Central Limit Theorem (30 marks)

WARNING: If it is not explicitly stated, please assume the 95% confidence or 5%

significant level.

Question 1 (5 marks)

The SETU score of FIT units is known to follow a $\mathcal{N}(\mu = 4, \sigma^2 = 0.25)$ distribution. You take a sample of the units and check their last semester's SETU. How many units do you have to sample to have a 95% confidence interval for μ with width 0.1?

ANSWER

When the population standard deviation is known, the required sample size n is found by the formula for the width of a confidence interval for the population mean:

$$W = 2 \times z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

For a 95% confidence level, $\alpha = 0.05$, so $z_{\alpha/2} = z_{0.025} \approx 1.96$.
The population standard deviation is $\sigma = \sqrt{0.25} = 0.5$, and the desired width $W = 0.1$.
Plugging these values into the formula:

$$0.1 = 2 \times 1.96 \times \frac{0.5}{\sqrt{n}}$$

$$0.1 = \frac{1.96}{\sqrt{n}}$$

$$\sqrt{n} = 19.6$$

$$n = (19.6)^2$$

$$n = 384.16$$

Rounding it up, the required sample size is 385.

Question 2 (5 marks)

You do a poll to see what fraction p of the students participated in the FIT5197 SETU survey. You then take the average frequency of all surveyed people as an estimate $\hat{p}$ for p . Now it is necessary to ensure that there is at least 99% certainty that the difference between the surveyed rate $\hat{p}$ and the actual rate p is not more than 5%. At least how many people should take the survey?

ANSWER

The margin of error E is given by:

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For a 99% confidence level, $\alpha = 0.01$, so $z_{\alpha/2} = z_{0.005} \approx 2.576$. The desired margin of error is $E = 0.05$. Since we don't have a prior estimate for p, we use the most conservative estimate $\hat{p} = 0.5$ to maximize the sample size:

$$0.05 = 2.576\sqrt{\frac{0.5(1-0.5)}{n}}$$

$$0.05 = 2.576\sqrt{\frac{0.25}{n}}$$

$$\frac{0.05}{2.576} = \sqrt{\frac{0.25}{n}}$$

$$0.019417 \approx \sqrt{\frac{0.25}{n}}$$

$$(0.019417)^2 \approx \frac{0.25}{n}$$

$$n \approx \frac{0.25}{0.0003769}$$

$$n \approx 663.4$$

Rounding up to the nearest whole number, at least 664 people should take the survey.

Question 3 (5 marks)

Suppose you repeated the above polling process multiple times and obtained 100 confidence intervals, each with confidence level of 99%. About how many of them would you expect to be "wrong"? That is, how many of them would not actually contain the parameter being estimated? Should you be surprised if 4 of them are wrong?

ANSWER

When constructing 100 confidence intervals, each with a 99% confidence level, this means that for each interval, there is a 99% probability that it contains the true population parameter p.
Therefore, there is a $1\% = 0.01$ probability that any given interval will not contain the true parameter.

The expected number of "wrong" intervals out of 100 would be:

$$100 \times (1 - 0.99) = 100 \times 0.01 = 1$$

To determine if observing 4 wrong intervals is surprising, we use the binomial distribution $B(n = 100, p = 0.01)$, where n is the number of intervals and p is the probability of an interval being wrong.

The expected number of wrong intervals is $\mu = np = 100 \times 0.01 = 1$, and the standard deviation is $\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.01 \times 0.99} = \sqrt{0.99} \approx 0.995$.

Observing 4 wrong intervals is a deviation of $4 - 1 = 3$ from the expected value. In terms of standard deviations, this is approximately $\frac{3}{0.995} \approx 3.015$ standard deviations above the mean.

A deviation of more than 2 or 3 standard deviations from the mean is usually considered statistically significant or surprising.
Therefore, observing 4 wrong intervals can be considered a surprising result, as it is more than 3 standard deviations away from the expected number of wrong intervals.

Question 4 (5 marks)

Consider the random variable X following the Bernoulli distribution with a parameter θ, i.e., $X \sim \text{Be}(\theta)$, where $\theta = 0.9$. Given that you collect n random variable $X_1, X_2, \ldots, X_n$. Calculate the smallest sample size, n, you have to observe to guarantee that

$$P\left(\left|\frac{\sum_1^n X_i}{n} - \theta\right| > 0.01\right) < 0.1.$$

ANSWER

Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed Bernoulli random variables with parameter $\theta = 0.9$. The sample mean is $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$.
By definition, $E[X_i] = 0.9$ and $\text{Var}(X_i) = 0.9(1 - 0.9) = 0.09$. By the Central Limit Theorem, $\bar{X}_n \approx \mathcal{N}\left(0.9, \frac{0.09}{n}\right)$. We want to find the smallest n such that:

$$P\left(\left|\frac{\sum_{i=1}^n X_i}{n} - 0.9\right| > 0.01\right) < 0.1$$

Standardizing the sample mean, $Z = \frac{\bar{X}_n - 0.9}{\sqrt{0.09/n}} = \frac{\bar{X}_n - 0.9}{0.3/\sqrt{n}} \approx \mathcal{N}(0, 1)$. The inequality becomes:

$$P\left(|Z| > \frac{0.01}{0.3/\sqrt{n}}\right) < 0.1$$

$$P\left(|Z| > \frac{0.01\sqrt{n}}{0.3}\right) < 0.1$$

$$P\left(|Z| > \frac{\sqrt{n}}{30}\right) < 0.1$$

Let $z_{\alpha/2}$ be the critical value for $\alpha = 0.1$, so $\alpha/2 = 0.05$. We have $z_{0.05} \approx 1.645$. Thus, we need:

$$\frac{\sqrt{n}}{30} > 1.645$$

$$\sqrt{n} > 30 \times 1.645$$

$$\sqrt{n} > 49.35$$

$$n > 2435.4225$$

The smallest integer n satisfying this condition is $n = 2436$.

Question 5 (5 Marks)

The error for the production of a machine is uniformly distribute over [-0.75, 0.75] unit. Assuming that there are 100 machines working at the same time, approximate the probability that the final production differ from the exact production by more than 4.5 unit?

ANSWER

Let X_i be the error for the i-th machine, $X_i \sim U[-0.75, 0.75]$.
The mean and variance of X_i are:
$E[X_i] = 0$
$\text{Var}(X_i) = \frac{(0.75-(-0.75))^2}{12} = \frac{(1.5)^2}{12} = 0.1875$

For $n = 100$ machines, the total error is $S_{100} = \sum_{i=1}^{100} X_i$.
By the Central Limit Theorem, S_{100} is approximately normally distributed with mean $E[S_{100}] = 100 \times 0 = 0$ and variance $\text{Var}(S_{100}) = 100 \times 0.1875 = 18.75$.
The standard deviation is $\sigma_{S_{100}} = \sqrt{18.75} \approx 4.3301$.

We want to find $P(|S_{100}| > 4.5) = P(S_{100} > 4.5) + P(S_{100} < -4.5)$. This is equivalent to:

$$Z_1 = \frac{4.5 - 0}{4.3301} \approx 1.0392$$

Since a normal distribution is symmetrical about 0, this is 2*$P(Z < -1.0392)$, where $Z \sim \mathcal{N}(0, 1)$.

Using the standard normal distribution table:

$$P(Z < -1.0392) \approx 0.1495$$

Therefore, the approximate probability is:

$$P(|S_{100}| > 4.5) \approx 0.1495 * 2 = 0.2990$$

Question 6 (5 Marks)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with mean λ. Thus, $Y = \sum_{i=1}^{n} X_i$ has a Poisson distribution with mean $n\lambda$. Moreover, by the Central limit Theorem, $\bar{X} = Y/n$ has, approximately, a Normal $(\lambda, \lambda/n)$ distribution for large n.

Show that for large values of n, the distribution of

$$2\sqrt{n}\left(\sqrt{\frac{Y}{n}}-\sqrt{\lambda}\right)$$

is independent of λ.

ANSWER

We are given that $\bar{X}=Y/n$ has an approximate normal distribution $\mathcal{N}(\lambda,\lambda/n)$ for large n. Substituting that,

$$W=2\sqrt{n}\left(\sqrt{\frac{Y}{n}}-\sqrt{\lambda}\right)=2\sqrt{n}\left(\sqrt{\bar{X}}-\sqrt{\lambda}\right).$$

To estimate the value of $\sqrt{x}$, we use a first-order Taylor series expansion of $g(x)=\sqrt{x}$ around $x=\lambda$:

$$g(x)\approx g(\lambda)+g'(\lambda)(x-\lambda)$$

Here, $g(x)=\sqrt{x}$ and $g'(x)=\frac{1}{2\sqrt{x}}$, so $g(\lambda)=\sqrt{\lambda}$ and $g'(\lambda)=\frac{1}{2\sqrt{\lambda}}$. This leads to:

$$\sqrt{\bar{X}}\approx\sqrt{\lambda}+\frac{1}{2\sqrt{\lambda}}(\bar{X}-\lambda).$$

Substituting this into the expression for W:

$$W=2\sqrt{n}\left(\left(\sqrt{\lambda}+\frac{1}{2\sqrt{\lambda}}(\bar{X}-\lambda)\right)-\sqrt{\lambda}\right)=2\sqrt{n}\left(\frac{1}{2\sqrt{\lambda}}(\bar{X}-\lambda)\right)=\frac{\sqrt{n}}{\sqrt{\lambda}}(\bar{X}-\lambda)$$

We know that $\bar{X}\sim\mathcal{N}(\lambda,\lambda/n)$, which implies that $\bar{X}-\lambda\sim\mathcal{N}(0,\lambda/n)$.
Let $Z=\bar{X}-\lambda$. Then $Z\sim\mathcal{N}(0,\lambda/n)$.
The statistic W can be written as $W=\frac{\sqrt{n}}{\sqrt{\lambda}}Z$.
The mean of W is $E[W]=\frac{\sqrt{n}}{\sqrt{\lambda}}E[Z]=\frac{\sqrt{n}}{\sqrt{\lambda}}\times 0=0$.
The variance of W is $\mathrm{Var}(W)=\left(\frac{\sqrt{n}}{\sqrt{\lambda}}\right)^2\mathrm{Var}(Z)=\frac{n}{\lambda}\times\frac{\lambda}{n}=1$.
As such, for large n, the distribution of W is approximately $\mathcal{N}(0,1)$, which is independent of λ.

Part 3 Hypothesis Testing (15 marks)

Question 1 (7.5 marks)

As a motivation for students to attend the tutorial, Levin is offering a lot of hampers this semester. He has designed a spinning wheel (This is an example https://spinnerwheel.com/) where there are four choices on it: "Hamper A", "Hamper B",

"Hamper C", and "Better Luck Next Time". These choices are evenly distributed on the wheel. If a student completes the attendance form for one of the tutorials, they will get a chance to spin the wheel.

As a hard-working student yourself, you have earned 12 chances at the end of the semester. When you finished your spins, the result showed {"N", "A", "N", "N", "B", "C", "N", "N", "N", "A", "A", "N"} ("A","B" and "C" denote three hampers respectively, while "N" denotes "Better Luck Next Time"). You are shocked by the result and feel the game might be faulty. Before questioning Levin, you would like to perform a hypothesis test to check whether you are really unlucky or has Levin secretly done something that had influenced the probability of winning or not. State your hypothesis, perform the test and interpret the result.

ANSWER

Hypothesis

Null Hypothesis (H_0): The spinning wheel is fair, and each outcome ("Hamper A", "Hamper B", "Hamper C", "Better Luck Next Time") has a probability of 0.25.
Alternative Hypothesis (H_1): The spinning wheel is not fair, and the probabilities of the outcomes are not all equal to 0.25.

Hypothesis test

We use a chi-square goodness-of-fit test.

Observed Frequencies (O_i):

- Hamper A: 3
- Hamper B: 1
- Hamper C: 1
- Better Luck Next Time (N): 7

Total Number of Spins (n): 12 Number of Categories (k): 4 Expected Frequencies (E_i) under H_0: $E_i = n \times p_i = 12 \times 0.25 = 3$ for each category.

The chi-square test statistic (χ^2) is calculated as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Plugging in the values:

$$\chi^2 = \frac{(3-3)^2}{3} + \frac{(1-3)^2}{3} + \frac{(1-3)^2}{3} + \frac{(7-3)^2}{3}$$

$$\chi^2 = \frac{0}{3} + \frac{(-2)^2}{3} + \frac{(-2)^2}{3} + \frac{(4)^2}{3}$$

$$\chi^2 = 0 + \frac{4}{3} + \frac{4}{3} + \frac{16}{3} = \frac{24}{3} = 8$$

The degrees of freedom are $df = k - 1 = 4 - 1 = 3$. Assuming a significance level (α) of 0.05, the critical value from the chi-square distribution is approximately $\chi^2_{crit} \approx 7.815$.

Conclusion

Since our calculated chi-square statistic ($\chi^2 = 8$) is greater than the critical value ($\chi^2_{crit} \approx 7.815$), we reject the null hypothesis.
There is statistically significant evidence at the 0.05 significance level to suggest that the spinning wheel is not fair. The observed frequencies of the outcomes are significantly different from what would be expected if the wheel were evenly distributed.

Question 2 (7.5 marks)

The operation team of a retailer is about to report the performance of year 2022. As the data analyst, your job entails reviewing the reports provided by the team. One of the reports regarding membership subscription looks suspicous to you. In this report, they compared the amount of money spent by the members against the non-members over the year. The methodology is that they randomly selected 20 customers and compared their spending before and after becoming a member.

The average spending before becoming a member is $\$88.5$ per week with a standard deviation of $\$11.2$. The average after becoming a member is $\$105$ per week with a standard deviation of $\$15$. In the report, the retailer claimed that after becoming a member, customers tend to spend 10% more than before on average.

As a statistician, you decide to perform a hypothesis test to verify the veracity of this claim. State your hypothesis, perform the test and interpret the result. Additionally, please suggest another methodology to compare member vs non-member.

ANSWER

Hypothesis

We want to determine if the average spending after becoming a member has increased significantly. Let μ_1 be the average spending before membership and μ_2 be the average spending after membership. The retailer claims a 10% increase, so we test:

Null Hypothesis (H_0): The average spending after becoming a member has not increased by more than 10%. Mathematically, $\mu_2 - \mu_1 \leq 0.10\mu_1$, which simplifies to $\mu_2 \leq 1.10\mu_1$.

Alternative Hypothesis (H_1): The average spending after becoming a member has increased by more than 10%. Mathematically, $\mu_2 - \mu_1 > 0.10\mu_1$, which simplifies to $\mu_2 > 1.10\mu_1$.

Given Data

- Sample size (n): 20 customers
- Average spending before ($\bar{x}_1$): $\$88.5$
- Standard deviation before (s_1): $\$11.2$
- Average spending after ($\bar{x}_2$): $\$105$
- Standard deviation after (s_2): $\$15$

Test Statistic

The methodology compares the spending of the same 20 customers before and after membership, indicating a paired t-test is appropriate. Let d_i be the difference in spending for the i-th customer (after - before), with an average difference of $\bar{d} = \bar{x}_2 - \bar{x}_1 = 105 - 88.5 = 16.5$.

To perform the paired t-test, we need the standard deviation of these differences (s_d), which is not provided in the summary statistics. Without the raw data or s_d, we cannot compute the exact t-statistic and p-value for the correct paired test.

If we were to proceed with an approximate independent samples t-test (which is not statistically sound given the study design), we could calculate a t-statistic. However, this approach is not recommended here due to the dependent nature of the samples.

The claimed 10% increase corresponds to $0.10 \times 88.5 = \$8.85$. The observed average increase is $\$16.5$, which is greater than the claim.

Interpretation

Based on the sample means, the observed increase in spending ($\$\,16.5$) exceeds the claimed 10% increase ($\$\,8.85$). However, without performing the correct paired t-test using the standard deviation of the differences, we cannot definitively conclude if this observed difference is statistically significant at a chosen significance level (e.g., $\alpha = 0.05$).

Alternative Methodology

A more appropriate way to directly compare the spending of members versus non-members would be to use independent samples:

- Randomly select two independent groups: one of members and one of non-members, over a specific period.
- Collect the spending data for each group.
- Perform an independent two-sample t-test, or Welch's t-test, if variances are unequal to compare the average spending between the two groups.

This approach avoids the dependency issues of the before-and-after comparison and directly assesses the difference in spending between the member and non-member

populations.

Part 4 Simulation (25 marks)

Suppose you are involved in a scientific research project. Your lab mates are struggling with a sampling problem. They have a probability density function as shown below, but none of them knows how to generate random numbers from this probability distribution. As a member with a background in data science in this lab, you want to help them solve the sampling problem.

$$f(x) = \begin{cases} 4x + 1 & -\frac{1}{4} \leq x < 0 \\ -\frac{4}{7}x + 1 & 0 \leq x < \frac{7}{4} \\ 0 & \text{otherwise} \end{cases}$$

(a) First of all, you want to calculate the cumulative density function $F(x)$ and the quantile function $Q(p)$ for $f(x)$.

(b) You can get random numbers distributed as per $f(x)$ by generating uniformly distributed numbers p from 0 to 1 and plug them into $Q(p)$. You know computer simulation helps a lot so you want to write a function to generate random numbers distributed as per $f(x)$. You call this function `samplingHelper` and it takes a single input **n** to be the number of realizations you want to generate. Besides, you want to use the following function template. The better your function is (errors handling, comments, variable names, etc) the higher the score you will get for this part.

```
{r}
samplingHelper <- function(n) {
    # Put down your own code here

    return(numbers) # numbers is an array of random numbers you
generated as per f(x)
}
```

(c) You want to call `samplingHelper` to generate 99,999 random numbers as per $f(x)$ and plot a histogram of the sample with 100 bins as well as overlay a theoretical curve on top of it.

(d) You know sharing knowledge is a good practise. You want to summarize the key steps of your sampling method. More importantly, you want to justify why this sampling method works. (less than 250 words)

(e) Your lab mates all appreciate your help and they get stuck on another sampling problem. The probability density function is given below

$$f(x) = e^{-x^2\pi} \text{ for } \mathrm{x} \in [-\infty, +\infty]$$

They need your help to generate random numbers as per this distribution. You decide to use the same sampling strategy as you discussed above. Now you want to derive its

cumulative density function $F(x)$ and the Quantile function $Q(p)$.

(f) You want to implement it as another function called `newSamplingHelper` . It takes a single input **n** to be the number of realizations you want to generate. Besides, you want to use the following function template. The better your function is (errors handling, comments, variable names, etc) the higher the score you will get for this part.

```
{r}
newSamplingHelper <- function(n) {
    # Put down your own code here

    return(numbers) # numbers is an array of random numbers you
generated as per f(x)
}
```

(g) You want to call `newSamplingHelper` to generate 99,999 random numbers as per $f(x)$ and plot a histogram of the sample with 100 bins as well as overlay a theoretical curve on top of it. What's your findings by comparing it with Gaussian distribution? (less than 100 words)

ANSWER

Part 4 (a) Calculate the cumulative density function F(x) and the quantile function Q(p) for f(x).

$$f(x) = \begin{cases} 4x + 1 & -\frac{1}{4} \leq x < 0 \\ -\frac{4}{7}x + 1 & 0 \leq x < \frac{7}{4} \\ 0 & \text{otherwise} \end{cases}$$

1. Cumulative Density Function (CDF) F(x)

CDF $F(x)$ is the integral of $F(x)$.
For $x < -\frac{1}{4}$:

$$F(x) = \int_{-\infty}^{x} 0dt = 0$$

For $-\frac{1}{4} \leq x < 0$:

$$F(x) = \int_{-\frac{1}{4}}^{x} (4t + 1)dt$$

$$F(x) = [2t^2 + t]_{-\frac{1}{4}}^{x}$$

$$F(x) = (2x^2 + x) - (2(-\frac{1}{4})^2 + (-\frac{1}{4}))$$

$$F(x) = 2x^2 + x - (\frac{2}{16} - \frac{1}{4})$$

$$F(x) = 2x^2 + x + \frac{1}{8}$$

For $0 \leq x < \frac{7}{4}$:

$$F(x) = F(0) + \int_0^x (-\frac{4}{7}t + 1)dt$$

$$F(x) = \frac{1}{8} + \left[-\frac{4}{7}\frac{t^2}{2} + t\right]_0^x$$

$$F(x) = \frac{1}{8} + \left[-\frac{2}{7}t^2 + t\right]_0^x$$

$$F(x) = \frac{1}{8} + (-\frac{2}{7}x^2 + x) - 0$$

$$F(x) = -\frac{2}{7}x^2 + x + \frac{1}{8}$$

For $x \geq \frac{7}{4}$:

$$F(x) = 1$$

This leads to a CDF $F(x)$ of:

$$F(x) = \begin{cases} 0 & x < -\frac{1}{4} \\ 2x^2 + x + \frac{1}{8} & -\frac{1}{4} \leq x < 0 \\ -\frac{2}{7}x^2 + x + \frac{1}{8} & 0 \leq x < \frac{7}{4} \\ 1 & x \geq \frac{7}{4} \end{cases}$$

2. Quantile Function (Inverse CDF) Q(p)

The quantile function $Q(p)$ is the inverse of CDF $F(x)$.

Case 1: $0 \leq p < \frac{1}{8}$ (This corresponds to $-\frac{1}{4} \leq x < 0$)

$$p = 2x^2 + x + \frac{1}{8}$$

$$2x^2 + x + (\frac{1}{8} - p) = 0$$

Using quadratic formula,

$$x = \frac{-1 \pm \sqrt{1^2 - 4(2)(\frac{1}{8} - p)}}{2(2)} = \frac{-1 \pm \sqrt{1 - (1 - 8p)}}{4} = \frac{-1 \pm \sqrt{8p}}{4}$$

Since x must be in $[-\frac{1}{4}, 0)$, we choose the '+' sign:

$$Q(p) = \frac{-1 + \sqrt{8p}}{4} \quad \text{for } 0 \leq p < \frac{1}{8}$$

Case 2: $\frac{1}{8} \leq p < 1$ (This corresponds to $0 \leq x < \frac{7}{4}$)

$$p = -\frac{2}{7}x^2 + x + \frac{1}{8}$$

$$\frac{2}{7}x^2 - x + (p - \frac{1}{8}) = 0$$

$$16x^2 - 56x + (56p - 7) = 0$$

Using quadratic formula,

$$x = \frac{56 \pm \sqrt{(-56)^2 - 4(16)(56p - 7)}}{2(16)}$$

$$x = \frac{56 \pm \sqrt{3136 - 64(56p - 7)}}{32}$$

$$x = \frac{56 \pm \sqrt{3136 - 3584p + 448}}{32}$$

$$x = \frac{56 \pm \sqrt{3584(1 - p)}}{32}$$

$$x = \frac{56 \pm 16\sqrt{14(1 - p)}}{32}$$

$$x = \frac{7 \pm 2\sqrt{14(1 - p)}}{4}$$

Since x must be in $[0, 47)$, we choose the '−' sign:

$$Q(p) = \frac{7 - 2\sqrt{14(1 - p)}}{4} \quad \text{for } \frac{1}{8} \leq p < 1$$

The quantile function Q(p) is:

$$Q(p) = \begin{cases} \frac{-1+\sqrt{8p}}{4} & 0 \leq p < \frac{1}{8} \\ \frac{7-2\sqrt{14(1-p)}}{4} & \frac{1}{8} \leq p < 1 \end{cases}$$

In [ ]:
```r
# (b) Write a function samplingHelper to generate random numbers
# distributed as per f(x).
samplingHelper <- function(n) {
  # Input:
  #   n: The number of realizations (random numbers) to generate.
  # Output:
  #   numbers: An array of n random numbers distributed as per f(x).

  # Error handling for input n
  if (!is.numeric(n) || length(n) != 1 || n <= 0 || floor(n) != n) {
    stop("Input 'n' must be a single positive integer.")
  }

  # Generate n uniform random numbers from U(0,1)
```

```r
  uniform_samples <- runif(n)

  # Initialize an empty vector to store the generated numbers
  numbers <- numeric(n)

  # Apply the quantile function Q(p)
  for (i in 1:n) {
    p <- uniform_samples[i]
    if (p < 0 || p >= 1) { # Shouldn't happen, just for robustness
      if (p == 1) { # Handle edge case p=1 for Q(p) if runif could produce it
        numbers[i] <- 7 / 4
      } else {
        # This case should ideally not be reached if p is from runif(n)
        # which generates in (0,1) or [0,1)
        warning(paste(
          "Uniform sample p =", p,
          "is outside [0,1). Check runif behavior or input p."
        ))
        # Error handling: Assign NA value if runif returned value outside [0,1)
        numbers[i] <- NA # Or handle as error
      }
    } else if (p < 1 / 8) {
      numbers[i] <- (-1 + sqrt(8 * p)) / 4
    } else { # p >= 1/8 and p < 1
      numbers[i] <- (7 - 2 * sqrt(14 * (1 - p))) / 4
    }
  }
  return(numbers)
}
```

In [ ]:

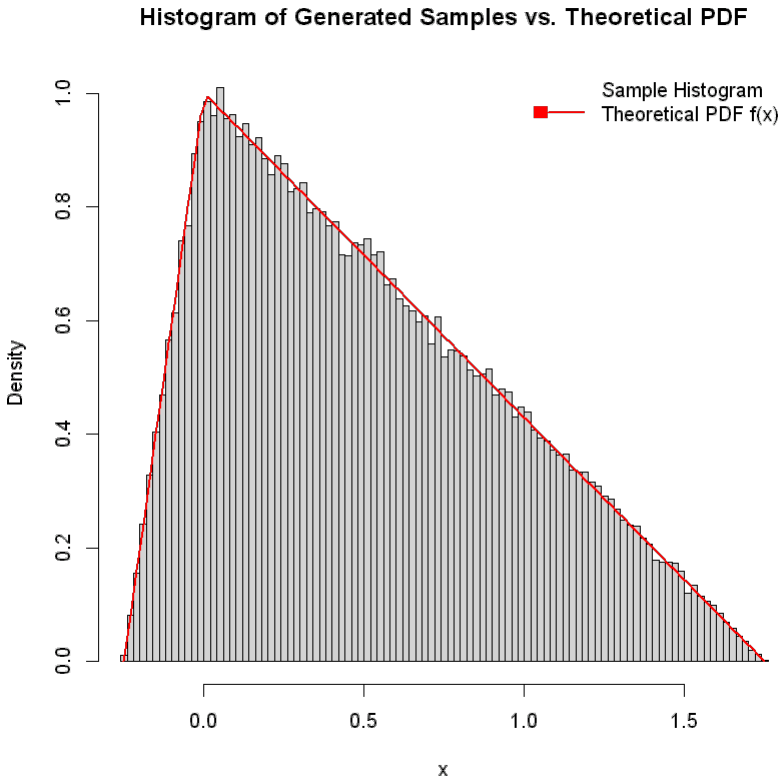
```r
# (c) Call samplingHelper to generate 99,999 random numbers and plot a
# histogram with a theoretical curve.
# Generate 99,999 random numbers
num_samples <- 99999
generated_samples1 <- samplingHelper(num_samples)

# Define the theoretical PDF f(x) for plotting
theoretical_pdf1 <- function(x) {
  # Vectorize the function to handle vector inputs for 'x'
  sapply(x, function(val) {
    if (val >= -1 / 4 && val < 0) {
      return(4 * val + 1)
    } else if (val >= 0 && val < 7 / 4) {
      return(-4 / 7 * val + 1)
    } else {
      return(0)
    }
  })
}

# Plot the histogram and overlay the theoretical PDF
# Calculate reasonable x-limits for the plot based on the distribution range
x_limits <- c(-0.25, 1.75)
hist_data <- hist(generated_samples1,
  breaks = 100, freq = FALSE,
  main = "Histogram of Generated Samples vs. Theoretical PDF",
  xlab = "x", ylab = "Density",
  xlim = x_limits
)
```

```r
# Overlay the theoretical PDF
# curve() is convenient for plotting functions
curve(theoretical_pdf1,
  from = x_limits[1], to = x_limits[2],
  add = TRUE, col = "red", lwd = 2
)

legend("topright",
  legend = c("Sample Histogram", "Theoretical PDF f(x)"),
  fill = c(NA, "red"), border = c(NA, NA), lty = c(NA, 1), lwd = c(NA, 2),
  col = c(NA, "red"), bty = "n"
)
```



Part 4(d) Summarize the key steps of your sampling method and justify why it works.

The sampling method used is called Inverse Transform Sampling. The steps are:

1. Derive the Cumulative Distribution Function (CDF), $F(x)$: Integrate the probability density function (PDF), $f(x)$, to obtain $F(x) = P(X \leq x)$.
2. Derive the Quantile Function, $Q(p)$: Find the inverse of the CDF, $Q(p) = F^{-1}(p)$. This is done by setting $F(x) = p$ and solving for x in terms of p.
3. Generate Uniform Samples: Generate a random number U from a standard uniform distribution on the interval $[0, 1)$, i.e., $U \sim U(0, 1)$.
4. Transform Uniform Samples: Compute $X = Q(U)$. The resulting random variable X will have the probability distribution defined by $f(x)$.

Or in simpler words, the sampling method uses the Quantile Function $Q(p)$ to map the domain of a standard uniform distribution to the probability distribution $f(x)$.

Justification: If U is a uniform random variable on $[0,1]$, then $F^{-1}(U)$ has F as its CDF.

Let $U \sim U(0,1)$ and let $X = Q(U) = F^{-1}(U)$.
We want to show that the CDF of X, denoted $F_X(x)$, is equal to $F(x)$.
$F_X(x) = P(X \leq x)$ Substitute $X = F^{-1}(U)$: $F_X(x) = P(F^{-1}(U) \leq x)$

Since $F(x)$ is a continuous and strictly increasing function, applying F to both sides of the inequality $F^{-1}(U) \leq x$ preserves the inequality direction: $F_X(x) = P(U \leq F(x))$

Because U is a standard uniform random variable, its CDF is $P(U \leq y) = y$ for any $y \in [0,1]$. Since $F(x)$ is a CDF, its value $F(x)$ is always in $[0,1]$.

Therefore, $P(U \leq F(x)) = F(x)$, and so, $F_X(x) = F(x)$.
This means that the random variable X generated by this method has the desired CDF $F(x)$, and consequently, the desired PDF $f(x)$.

Part 4(e) Derive the CDF $F(x)$ and the Quantile function $Q(p)$ for $f(x) = e^{-x^2\pi}$.

The PDF is given by $f(x) = e^{-x^2\pi}$ for $x \in [-\infty, +\infty]$.

Verify it's a valid PDF:

For a valid PDF, $\int_{-\infty}^{\infty} f(x)dx = 1$. The integral $\int_{-\infty}^{\infty} e^{-ax^2}dx = \sqrt{\frac{\pi}{a}}$. For $f(x) = e^{-x^2\pi}$, we have $a = \pi$. So, $\int_{-\infty}^{\infty} e^{-x^2\pi}dx = \sqrt{\frac{\pi}{\pi}} = \sqrt{1} = 1$. Therefore, $f(x)$ is a valid probability density function.

This PDF is a form of the Gaussian/Normal distribution.
The general form of a normal PDF is $N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Comparing $e^{-x^2\pi}$ with this form, we can see that the mean $\mu = 0$. The exponent is $-x^2\pi = -\frac{x^2}{2\sigma^2}$. So, $\pi = \frac{1}{2\sigma^2}$, which implies $\sigma^2 = \frac{1}{2\pi}$. The standard deviation is $\sigma = \sqrt{\frac{1}{2\pi}} = \frac{1}{\sqrt{2\pi}}$. The normalization constant: $\frac{1}{\sigma\sqrt{2\pi}} = \frac{1}{(\frac{1}{\sqrt{2\pi}})\sqrt{2\pi}} = 1$. This confirms that $f(x) = e^{-x^2\pi}$ is the PDF of a normal distribution $N(0, \sigma^2 = \frac{1}{2\pi})$.

Cumulative Density Function (CDF) $F(x)$

The CDF is $F(x) = \int_{-\infty}^{x} e^{-t^2\pi}dt$.

For a normal distribution $N(\mu, \sigma^2)$, the CDF is $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$, where $\Phi(z)$ is the CDF of the standard normal distribution $N(0,1)$.

Here, $\mu = 0$ and $\sigma = \frac{1}{\sqrt{2\pi}}$. So, $F(x) = \Phi\left(\frac{x-0}{1/\sqrt{2\pi}}\right) = \Phi(x\sqrt{2\pi})$.

The function $\Phi(z)$ is often expressed using the error function: $\text{erf}(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2} dt$, as $\Phi(z) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$.

Therefore, $F(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x\sqrt{2\pi}}{\sqrt{2}}\right)\right] = \frac{1}{2}\left[1 + \text{erf}(x\sqrt{\pi})\right]$.

Quantile Function (Inverse CDF) $Q(p)$

The quantile function $Q(p)$ is found by solving $F(x) = p$ for x. $p = \Phi(x\sqrt{2\pi})$

Apply the inverse standard normal CDF, $\Phi^{-1}(p)$ to both sides: $\Phi^{-1}(p) = x\sqrt{2\pi}$
$x = \frac{\Phi^{-1}(p)}{\sqrt{2\pi}}$. $Q(p) = \frac{\Phi^{-1}(p)}{\sqrt{2\pi}}$.

It should be noted that since $f(x)$ is the PDF of $N(0, \sigma^2 = \frac{1}{2\pi})$, the quantile function $Q(p)$ is directly given in R as `qnorm(p, mean = 0, sd = 1/sqrt(2*pi))` .

In [ ]:
```r
# (f) Implement newSamplingHelper.
newSamplingHelper <- function(n) {
    # Input:
    #   n: The number of realizations (random numbers) to generate.
    # Output:
    #   numbers: An array of n random numbers distributed as per
    #   f(x) = exp(-x^2*pi).

    # Error handling for input n
    if (!is.numeric(n) || length(n) != 1 || n <= 0 || floor(n) != n) {
        stop("Input 'n' must be a single positive integer.")
    }

    # Calculate the standard deviation for the target normal distribution
    sigma <- 1 / sqrt(2 * pi)

    # Generate n uniform random numbers from U(0,1)
    uniform_samples <- runif(n)

    # Apply the quantile function Q(p) for N(0, sigma^2)
    numbers <- qnorm(uniform_samples, mean = 0, sd = sigma)

    return(numbers)
}
```

In [ ]:
```r
# (g) Call newSamplingHelper to generate 99,999 random numbers, plot
# histogram, and compare with Gaussian.
# Generate 99,999 random numbers
num_samples2 <- 99999
generated_samples2 <- newSamplingHelper(num_samples2)

# Define the theoretical PDF f(x) = exp(-x^2*pi) for plotting
theoretical_pdf2 <- function(x) {
    exp(-x^2 * pi)
}

# Plot the histogram and overlay the theoretical PDF
# Determine reasonable x-limits for the plot, e.g., +/- 4 standard deviations
sigma_val <- 1 / sqrt(2 * pi)
```

```r
x_limits_new <- c(-4.5 * sigma_val, 4.5 * sigma_val) # Adjusted for viz

hist_data_new <- hist(generated_samples2,
    breaks = 100, freq = FALSE,
    main = "Histogram of New Samples vs. Theoretical PDF",
    xlab = "x", ylab = "Density",
    xlim = x_limits_new,
    ylim = c(0, theoretical_pdf2(0) * 1.1)
)
# Ensure peak is visible

# Overlay the theoretical PDF
curve(theoretical_pdf2,
    from = x_limits_new[1], to = x_limits_new[2],
    add = TRUE, col = "blue", lwd = 2
)

legend("topright",
    legend = c("Sample Histogram", "Theoretical PDF f(x)"),
    fill = c(NA, "blue"), border = c(NA, NA), lty = c(NA, 1), lwd = c(NA, 2),
    col = c(NA, "blue"), bty = "n"
)

# Findings:
# The histogram of the generated samples closely matches the theoretical
# PDF curve f(x) = exp(-x^2*pi).
# This theoretical PDF is a Gaussian (normal) distribution with mean 0
# and variance sigma^2 = 1/(2*pi).
# The standard deviation is sigma = 1/sqrt(2*pi) approx 0.3989.
# Compared to the standard normal distribution N(0,1), this distribution
# is more peaked around its mean (0), and has lighter tails (i.e., values
# fall off more quickly as they move away from the mean) because its
# standard deviation (approx 0.3989) is smaller than 1.
```