FIT5197 2024 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.

No external R libraries allowed. Only base packages

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 $X \sim IG(\theta : (\mu, \lambda))$, $\forall\, \mu > 0$ 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), $x = (x_1, \ldots, x_n)$, 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

Given: $X \sim IG(\theta : (\mu, \lambda))$ follows a inverse gaussian distribution.
The pdf is given by:

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

We need to find maximum likelihood estimators for μ and λ.

Step 1: Finding likelihood function for $f(x \mid \mu, \lambda)$:
We know that likelihood function is given by:

$$L(x|\mu,\lambda) = \prod_{i=1}^{n} f(x|\mu,\lambda)$$

Writing Likelihood with respect to PDF,

$$L(x|\mu,\lambda) = \prod_{i=1}^{n} \left(\frac{\lambda}{2\pi x^3}\right)^{\left(\frac{1}{2}\right)} e^{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}}$$

Step 2: Taking Negative Log-likelihood function for the above likelihood equation:

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

Considering property of logarithms, $\log(a \cdot b) = \log(a) + \log(b)$ we can simply as follows:

$$= -\sum_{i=1}^{n} \left(\frac{1}{2}\log\left(\frac{\lambda}{2\pi x_i^3}\right) - \log\left(e^{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}}\right)\right)$$

$$= -\sum_{i=1}^{n} \left(\frac{1}{2}\log\left(\frac{\lambda}{2\pi x_i^3}\right) - \frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i}\right)$$

$$= -\sum_{i=1}^{n} \left(\frac{1}{2}\log(\lambda) - \frac{1}{2}\log(2\pi) - \frac{3}{2}\log(x_i) - \frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i}\right)$$

$$-L(x|\mu,\lambda) = -\frac{n}{2}\log(\lambda) + \frac{n}{2}\log(2\pi) + \frac{3}{2}\sum_{i=1}^{n}\log(x_i) + \sum_{i=1}^{n}\frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i} \text{ -------- (D)}$$

Now in order to find the maximum estimators for μ and λ we partially derrive with respect to it by substituitng the above equation to zero.

Step 3: First we partially derrive with respect to μ, Considering equation D, since we are partially deriving with respect to μ other elements in the equation which do consists μ will be zero while partially deriving and below is the resultant.

$$\frac{\partial}{\partial\mu} - L(x|\mu,\lambda) = \frac{\partial}{\partial\mu}\left(\sum_{i=1}^{n} \frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i}\right)$$

$$\frac{\partial}{\partial\mu}(-L(x|\mu,\lambda)) = \frac{\partial}{\partial\mu}\left(\sum_{i=1}^{n} \frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i}\right)$$

Applying Chain Rule, We get:

$$\frac{\partial}{\partial \mu}\left(\frac{(x_i-\mu)^2}{\mu^2}\right)=\frac{\partial}{\partial \mu}\left(\frac{x_i^2-2\mu x_i+\mu^2}{\mu^2}\right)$$

$$=\frac{\partial}{\partial \mu}\left(\frac{x_i^2}{\mu^2}-\frac{2x_i}{\mu}+1\right)$$

$$=\frac{-2x_i^2}{\mu^3}+\frac{2x_i}{\mu^2}$$

$$\frac{\partial}{\partial \mu}\left(\frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i}\right)=\frac{\lambda}{2x_i}\left(\frac{-2x_i^2+2\mu x_i}{\mu^3}\right)$$

$$=\frac{\lambda}{2x_i}\left(\frac{2x_i(\mu-x_i)}{\mu^3}\right)$$

$$=\frac{\lambda(\mu-x_i)}{\mu^3}$$

Now considering for all values of n we get as follows:

$$\frac{\partial}{\partial \mu}(-L(x|\mu,\lambda))=\sum_{i=1}^{n}\frac{\lambda(\mu-x_i)}{\mu^3}=0$$

Simplifying:

$$\frac{\lambda}{\mu^3}\sum_{i=1}^{n}(\mu-x_i)=0$$

$$\sum_{i=1}^{n}(\mu-x_i)=0$$

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

Therefore Maximum likelihood estimator for μ is sample mean.

$$\mu_{MLE}=\frac{\sum_{i=1}^{n}x_i}{n}$$

Step 4: Now we find maximum likelihood estimator for λ by partially deriving with respect to it.

$$\frac{\partial}{\partial \lambda}(-L(x|\mu,\lambda))=\frac{\partial}{\partial \lambda}\left(-\frac{n}{2}\ln(\lambda)+\sum_{i=1}^{n}\frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i}\right)$$

Let's compute the derivative step by step:

$$\frac{\partial}{\partial \lambda}\left(-\frac{n}{2}\ln(\lambda)\right)=-\frac{n}{2\lambda}$$

$$\frac{\partial}{\partial\lambda}(\sum_{i=1}^{n}\frac{\lambda(x_i-\mu)^2}{2\mu^2x_i})=\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\mu^2x_i}$$

Summing these results:

$$-\frac{n}{2\lambda}+\sum_{i=1}^{n}\frac{(x_i-\mu_{MLE})^2}{2\mu_{MLE}^2x_i}=0$$

Solving for λ:

$$\frac{n}{2\lambda}=\sum_{i=1}^{n}\frac{(x_i-\mu_{MLE})^2}{2\mu_{MLE}^2x_i}$$

$$\lambda_{MLE}=\frac{n}{\sum_{i=1}^{n}\frac{(x_i-\mu_{MLE})^2}{\mu_{MLE}^2x_i}}$$

Therefore the MLE estimators are

$$\mu_{MLE}=\frac{\sum_{i=1}^{n}x_i}{n}$$

and

$$\lambda_{MLE}=\frac{n}{\sum_{i=1}^{n}\frac{(x_i-\mu_{MLE})^2}{\mu_{MLE}^2x_i}}$$

Question 2 (7.5 marks)

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

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

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

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

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

ANSWER

(a) To Derive MLE estimator for θ, i.e., $\hat{\theta}_{MLE}$, and show that it is unbiased:

Given: The Probability Density Function is as follows:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}\exp(-\frac{x}{\theta}) & x > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Step 1: Writing likelihood Function for the PDF:

We know that likelihood function is given by $L(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ substituting this to the pdf we get:

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

$$= (\frac{1}{\theta})^{n} \exp(-\frac{\sum_{i=1}^{n} x_i}{\theta})$$

Step 2: Now we find negative log-likelihood:

$$-L(x|\theta) = -\log L(x|\theta)$$

$$-L(x|\theta) = -\log((\frac{1}{\theta})^{n} \exp(-\frac{\sum_{i=1}^{n} x_i}{\theta}))$$

by applying property of logarithms $\ln(a \cdot b) = \ln(a) + \ln(b)$, we get:

$$-L(x|\theta) = -(n\log(\frac{1}{\theta}) - \frac{\sum_{i=1}^{n} x_i}{\theta})$$

$$-L(x|\theta) = -n\ln(\frac{1}{\theta}) + \frac{\sum_{i=1}^{n} x_i}{\theta}$$

$$-L(x|\theta) = n\ln(\theta) + \frac{\sum_{i=1}^{n} x_i}{\theta}$$

Step 3: Now we partially derive with respect to Theta

To find the MLE, we take the partial derivative of the negative log-likelihood function with respect to θ and set it to zero:

$$\frac{\partial}{\partial\theta}(-L(x|\theta)) = \frac{\partial}{\partial\theta}(n\ln(\theta) + \frac{\sum_{i=1}^{n} x_i}{\theta})$$

Let's compute the derivative step by step:

$$\frac{\partial}{\partial\theta}(n\ln(\theta)) = \frac{n}{\theta}$$

$$\frac{\partial}{\partial\theta}\left(\frac{\sum_{i=1}^{n} x_i}{\theta}\right) = -\frac{\sum_{i=1}^{n} x_i}{\theta^2}$$

Summing these results:

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

Multiplying through by θ^2:

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

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

$$\theta_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Now we Show that θ_{MLE} is Unbiased

An estimator $\hat{\theta}$ is unbiased if $E[\hat{\theta}] = \theta$.

Here, $\theta_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Since $x_i \sim \text{Dist}(\mu = \theta, \sigma^2 = \theta^2)$, we have:

$$E[x_i] = \theta$$

Therefore:

$$E[\theta_{MLE}] = E\left[\frac{\sum_{i=1}^{n} x_i}{n}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[x_i]$$

$$= \frac{1}{n} \cdot n \cdot \theta$$

$$= \theta$$

Thus, θ_{MLE} is an unbiased estimator of θ.

Question 3 (7.5 marks)

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

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

Imagine that we observe a sample of n i.i.d random varaibles $x = (x_1, \ldots, x_n)$ and want to model them using this distribution. Please use the concept of maximum likelihood to estimate for the parameter θ.

ANSWER

Given the pdf:

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

Step 1: Likelihood Function:

We first find the likelihood function for it. W.K.T likelihood fucntion is given by:

$$L(x\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

pluging in the pdf:

$$L(x\theta) = \prod_{i=1}^{n} \frac{(2/x_i)\theta^{x_i}(1-\theta)^{2-x_i}}{1-(1-\theta)^2}$$

Step 2: Applying -ve log-Likelihood:

Now we find the negative log-likelihood.

$$L(x\theta) = -\log L(\theta) = -\sum_{i=1}^{n} \log\left(\frac{(2/x_i)\theta^{x_i}(1-\theta)^{2-x_i}}{1-(1-\theta)^2}\right)$$

Applying the properties of logarithms we get:

$$= -\sum_{i=1}^{n} \left(\log\left(\frac{2}{x_i}\right) + x_i \log(\theta) + (2-x_i)\log(1-\theta) - \log(1-(1-\theta)^2)\right)$$

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

Step 3: Partially deriving with respect to theta:

Now in order to find MLE with respect to θ we consider the partial dervative with respect to it by susbtiuting the partial dervative equation to zero.

$$\frac{\partial(-L(x|\theta))}{\partial\theta} = -\sum_{i=1}^{n} (\frac{x_i}{\theta} - \frac{(2 - x_i)}{1 - \theta} - \frac{1 - \theta - \theta}{\theta(2 - \theta)})$$

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

setting partial dervative equation to zero we get:

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

Question 4 (7.5 marks)

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

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

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

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

(a) Given PDF function:

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

We need to derrive the MLE estimator for theta.

Step 1: Applying Likelihood Function for the PDF:

W.K.T likelihood function is given by

$$L(x\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

applying for PDF:

$$L(\theta) = \prod_{i=1}^{n} e^{-(x_i-\theta)} \quad \text{for } x_i \geq \theta$$

Since the likelihood is zero in case of any $(x_i < \theta)$, and hence we only consider the case where all $(x_i \geq \theta)$

$$L(\theta) = e^{-\sum_{i=1}^{n}(x_i-\theta)} = e^{-\sum_{i=1}^{n} x_i + n\theta}$$

Step 2: Taking Negative Log Likelihood:

$$L(\theta) = -\log L(\theta) = -\log(e^{-\sum_{i=1}^{n} x_i + n\theta}) = \sum_{i=1}^{n} x_i - n\theta$$

Step 3: Partially deriving Negative Log Likelihood with respect to Theta:

$$\frac{\partial(-L(\theta))}{\partial\theta} = \frac{\partial}{\partial\theta}(\sum_{i=1}^{n} x_i - n\theta) = -n$$

Since the dervative of -n is a constant which does not depend on θ, the negative log likelihood function is only minimized when value of θ is large for all $(x_i \geq \theta)$ therefore

$$\hat{\theta}_{MLE} = \min(x_i)$$

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 $N(\mu = 4, \sigma^2 = 0.25)$ 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

It is given that the setu score of FIT units follow a normal distribution ->(*)

1.And the variance of these setu score is $\sigma^2 = 0.25$.
2.Mean μ of the scores is given by 4.
3.We also know that the 95% confidence interval is given for μ at the width of 0.1.
4.Given variance the standard deviation will be $\sigma = \sqrt{0.25} = 0.5$

From equation(*) we know that the score follows a normal distribution and σ^2 & μ are given and hence we consider

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Therefore CI can be expanded as

$$CI = (\bar{X} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}, \bar{X} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}})$$

From number 3 in given we write the above equation as follows:

$$(\bar{X} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} - (\bar{X} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}})) = 0.1$$

Subsituting the values we get

$$(4 + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} - (4 - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}})) = 0.1$$

$$(4 + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} - 4 + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}) = 0.1$$

$$2 * (z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}) = 0.1$$

Since we need to find the sample space we re-arrange the above equation with respect to $\sqrt{n}$

$$2 * \frac{Z_{\frac{\alpha}{2}} * \sigma}{0.1} = \sqrt{n}$$

Simplifying the above equation we get

$$n = 20 * Z_{\frac{\alpha}{2}} * \sigma^2 ---> (a)$$

Now we find the value for $Z_{\frac{\alpha}{2}}$
The sgnificance value is given by

$$\alpha = 1 - \frac{95}{100} = 0.05$$

This remaning 5% is split between 2 tails of the distribution and therefore

$$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025}$$

finding cumulative probability

$$p = 1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$$

z-table value for p=0.975 for $Z_{0.025}$ is = 1.96. Substituiting all the values to equation (a) we get

$$n = 20 * 1.96 * 0.5^2$$

$$n = 384.16$$

rounding

$$n = 385$$

Thus, We need to smaple 385 units to have 95% confidence interval forμ with width of 0.1.

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

Given:
The Question states that there is 99% confidence interval
then the significance value i.e α is 1% (0.01) and this is split between 2 tails given by $\frac{\alpha}{2}$ =0.01/2
Considering standard normal distribution table the value for $Z_{\frac{0.01}{2}} = Z_{0.005} = 2.576$
In the survey process there are 2 possible outcomes, i.e. a student takes a survey or a student doesn't take a survey,
considering this nature we can relate this to bernouli distribution. The interval is given by

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The maximum diference between the surveyed rate and actual rate is not more than 5%, and we can consider that as,

$$M = Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

We need to find the number people that needs to take the survey, which is the sample space. Hence we arrange the above as followed.

$$M = Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

By taking square root on both side

$$M^2 = (Z_{\alpha/2})^2 \frac{\hat{p}(1-\hat{p})}{n}$$

to solve for sample space,

$$n = \frac{(Z_{\alpha/2})^2 \hat{p}(1-\hat{p})}{M^2}$$

substituting the values from given

$$n = \frac{(2.576)^2 * 0.5 * (1-0.5)}{0.05}^2$$

$$n = \frac{6.635 * 0.25}{0.0025}$$

$$n = 663.5$$

rounding up

$$n = 664$$

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 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

Question 4 (5 marks)

Consider the random variable X following the Bernoulli distribution with a parameter θ, i.e., $X \sim Be(\theta)$, 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

Given: For a Bernoulli distribution we can consider the following:

1. $X \sim Be(\theta) = 0.9$
2. $\mu\ of X \sim: \mu = \theta = 0.9$
3. $Variance\ \sigma^2 of X \sim: \sigma^2 = \theta(1-\theta) = 0.9(1-0.9) = 0.09$

By considering central limit theorem, we know that the sample mean of n independent and identically distributed random variables is normally distributed, hence we consider the below.

$$\sum x \sim N(\theta, \sigma/n)$$

$$= N(0.9, 0.09/n)$$

The Standard deviation is given by

$$\sigma = \sqrt{\frac{\sigma^2}{n}}$$

$$= \sqrt{\frac{0.09}{n}}$$

$$\sigma = \frac{0.3}{\sqrt{n}}$$

We need to find the below probability:

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

$P(\frac{\sum_1^n X_i}{n})$ can be considered as the sample mean $\bar{X}$ sustituting the values we get the following,

$$P(|X - 0.9| <= 0.01) > 0.9$$

Now we find the Z score for the sample mean Wkt $Z = \frac{(\bar{X}-\theta)}{\sigma}$

We need to evaulate

$$P\left(\left|\frac{(\bar{X} - 0.9)}{\sigma}\right| <= \frac{0.01}{\sigma}\right) > 0.9$$

$$P\left(|Z| <= \frac{0.01}{\frac{0.3}{\sqrt{n}}}\right) > 0.9 ----> (B)$$

We now need to find the critical value of Z. for a 90% confidence interval the data is within the interval and rest 10% lies out which is divided between the 2 tails i.e. 5% (0.05) in each tail. We need to find Z score corresponding to 0.95(CI 0.90 + 0.5 tail).

From the Z table we can find the cumulative probability for the given Z-scores $Z_{0.95}$ the value is 1.645. Considering equation B now we solve for n:

$$\frac{0.01\sqrt{n}}{0.3} > 1.645$$

$$\frac{0.01\sqrt{n}}{>} 0.4935$$

$$\sqrt{n} > 49.35$$

applying square root on both side

$$(\sqrt{n})^2 > (49.35)^2$$

$$n > 2435.42$$

$$n = 2436$$

Therefore the smallest sample size to guarantee $P\left(\left|\frac{\sum_1^n X_i}{n} - \theta\right| > 0.01\right) < 0.1$ is 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

Given:

1.The error for the prodcution is uniformly distributed -> (*)

2.Error distribution range $[-0.75, 0.75]$

3.Number of machines working at the same time n = 100

4.We need to find what is the PROBABILITY that the final production differ from the exact production by more than 4.5 units.

We know that the error distribution range is given by $[-0.75, 0.75]$, For a Uniform distribution of the form U(a,b) where a = -0.75 and b=0.75 the mean μ and σ^2 is given as follows

$$\mu = \frac{a + b}{2}$$

&

$$\sigma^2 = \frac{b - a^2}{12}$$

substituting the values for a and b we get

$$\mu \frac{-0.75 + 0.75}{2} = 0$$

&

$$\sigma^2 = \frac{1.5^2}{12} = 0.1875$$

$$\sigma = \sqrt{0.1875}$$

For independent and identically distributed random variables with mean μ and standard deviation σ^2 the sum of errors according to CLT is given by:

$$\sum x \approx N(n\mu, n\sigma^2)$$

Substituting the values we get

$$\sum x \approx N(100(0), 100(0.1875))$$

$$\sum x \approx N(0, 18.75)$$

from the above standard deviation can derived as

$$\sigma = \sqrt{18.75} \approx 4.33$$

Now let's find the probability of the difference in the production that deviates more than 4.5 units, For 100 machines and in the distribution considering the 2 tails we can define the probability as follows:

$$P(S_{100} > 4.5) = P(S_{100} > 4.5) + P(S_{100} < -4.5) - - - - > (A)$$

First we find the Z score given by

$$Z = \frac{S - \sigma}{\mu}$$

$$Z = \frac{4.5 - 0}{4.33} = \frac{4.5}{4.33}$$

$$Z \approx 1.04$$

&

$$Z = \frac{-4.5 - 0}{4.33} = \frac{-4.5}{4.33}$$

$$Z \approx 1.04$$

Using standard normal distribution tables, we find probability for the Z scores
i)$\text{for} Z > 1.04 : P(Z > 1.04) \approx 0.1492$

ii)$\text{for} Z < -1.04 : P(Z > 1.04) \approx 0.1492$

Substituting and combining values to equation A we get:

$$P(S_{100} > 4.5) = 0.1492 + 0.1492$$

$$P(S_{100} > 4.5) = 0.2984$$

Therefore the probability that the final production differ from the exact production by more than 4.5 unit is 0.2984 or 29.84%

Question 6 (5 Marks)

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

Given:

X1, X2,....Xn are i.i.d random variables of a poisson distribution with mean λ

$Y = \sum_{i=1}^{n} X_i$ has a Poisson distribution with mean λ

by CLT $X = Y/n$ has, approximately, a Normal $(\lambda, \lambda/n)$

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

to check is independent of λ.

First we consider the CLT for $\bar{X}$

We know that Y is the sum of n i.i.d. in Poisson λ random variables, Y follows a Poisson distribution with mean λ. Considering the sample mean $\bar{X} = \frac{Y}{n}$ is normally distributed:

$$\bar{X} \sim N\,(\lambda, \frac{\lambda}{n}) ----> (*)$$

Hence, $Y \sim N\,(n\lambda, n\lambda)$ for all large n.

Consider $W = \sqrt{\frac{Y}{n}}$. so that we can show the distribution of $(2\sqrt{n}(W - \sqrt{\lambda})$ is independent of λ

Now in order to approximate the distribution we apply Delta Method which helps to find distribution fucntion of a random variable.

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

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

Now we consider $(2\sqrt{n}(W - \sqrt{\lambda})$
substituting the W to the above we get

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

$$= 2\sqrt{n} \cdot \frac{1}{2\sqrt{\lambda}}(\bar{X} - \lambda)$$

$$= \sqrt{n} \cdot \frac{1}{\sqrt{\lambda}}(\bar{X} - \lambda)$$

$$= \frac{\sqrt{n}}{\sqrt{\lambda}}(\bar{X} - \lambda) ----> (1)$$

W.k.t from equation (*)

$$\bar{X} \sim N(\lambda, \frac{\lambda}{n})$$

, thus:

$$\frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \sim N(0, 1) ----> (2)$$

By considering equation 1,2 and (*) we can conclude

$$\frac{\sqrt{n}}{\sqrt{\lambda}}(\bar{X}-\lambda)=\frac{\bar{X}-\lambda}{\sqrt{\frac{\lambda}{n}}}\sim N(0,1)$$

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

The goal here will be to check if the results of the spining wheel is fair or not. First we will set up our hypothesis:

1. Our Null Hypothesis H0 is given as: "The spining wheel has an equal probability of 0.25 and is fair"
2. Our Alternate Hypothesis H1 is given as:"The spining wheel has unequal probability and in unfair"

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

Here we need check whether a cutomer tends to spend 10% more on average after becoming a member

1. Our Null Hypothesis(H0): There is no difference in typical spending before and after membership.
2. Our Alter Hypothesis(H1): The average spending after becoming a member exceeds the average spending 10% before becoming a member.

therefore

$$H0 = \mu_d = 0$$

and

$$H1 = \mu_d > 0$$

Part 4 Simulation (25 marks) - no external libraries R allowed. Only base packages

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

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

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

(a) calculating the cumulative density function $F(x)$ and the quantile function $Q(p)$ for $f(x)$.
Given:

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

1. Considering the First Interval $(-\frac{1}{4} \le x < 0)$

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

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

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

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

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

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

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

1. Considering the Second Interval $(0 \le x < \frac{7}{4})$ First let's consider $(F(x))at(x = 0)$
   $F(0) = 2(0)^2 + 0 + \frac{1}{8} = \frac{1}{8}$

Next, we integrate from 0 to x :

$$F(x) = \frac{1}{8} + \int_{0}^{x} (-\frac{4}{7}t + 1)\, dt$$

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

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

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

Therefore the cumulative distribution function (F(x)) is given as follows :

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

Now We calculate the quantile function Q(p):

Considering CDF has derived from above we can write:

For the First interval from the CDF:$(-\frac{1}{4} \leq x < 0)$ The CDF is given by:

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

We need to solve for x in terms of p :

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

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

Since above equation D is of the quadratic form, we can consider the quadratic equation formula.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a})$$

Plugging in the values we get

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

$$x = \frac{-1 \pm \sqrt{1 - 1 + 8p}}{4}$$

$$x = \frac{-1 \pm \sqrt{8p}}{4}$$

$$x = \frac{-1 \pm 2\sqrt{2p}}{4}$$

$$x = \frac{-1 + 2\sqrt{2p}}{4}$$

Considering the interval where x lies we only take +ve part of equation.

$$x = -\frac{1}{4} + \frac{\sqrt{2p}}{2}$$

For the interval $(0 \leq x < \frac{7}{4})$ The CDF for this interval is given as follows:

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

solving for x in terms of p:

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

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

Above is another equation in quadratic form and hence we consider the same quadratic equation.

$$x = \frac{-1 \pm \sqrt{1 - 4 \cdot (-\frac{2}{7}) \cdot (\frac{1}{8} - p)}}{2 \cdot -\frac{2}{7}}$$

$$x = \frac{-1 \pm \sqrt{1 + \frac{8}{7} \cdot (\frac{1}{8} - p)}}{-\frac{4}{7}}$$

$$x = \frac{-1 \pm \sqrt{1 + \frac{1}{7} - \frac{8p}{7}}}{-\frac{4}{7}}$$

$$x = \frac{-1 \pm \sqrt{\frac{8}{7} - \frac{8p}{7}}}{-\frac{4}{7}}$$

$$x = \frac{-1 \pm \sqrt{\frac{8(1-p)}{7}}}{-\frac{4}{7}}$$

Multiplying the numerator and denominator by 7:

$$x = \frac{7(-1 \pm \sqrt{\frac{8(1-p)}{7}})}{-4}$$

$$x = \frac{7}{4}\left(1 \pm \sqrt{\frac{8(1-p)}{7}}\right)$$

Considering the interval where x lies we only take +ve part of equation

$$x = \frac{7}{4}\left(1 - \sqrt{\frac{8(1-p)}{7}}\right)$$

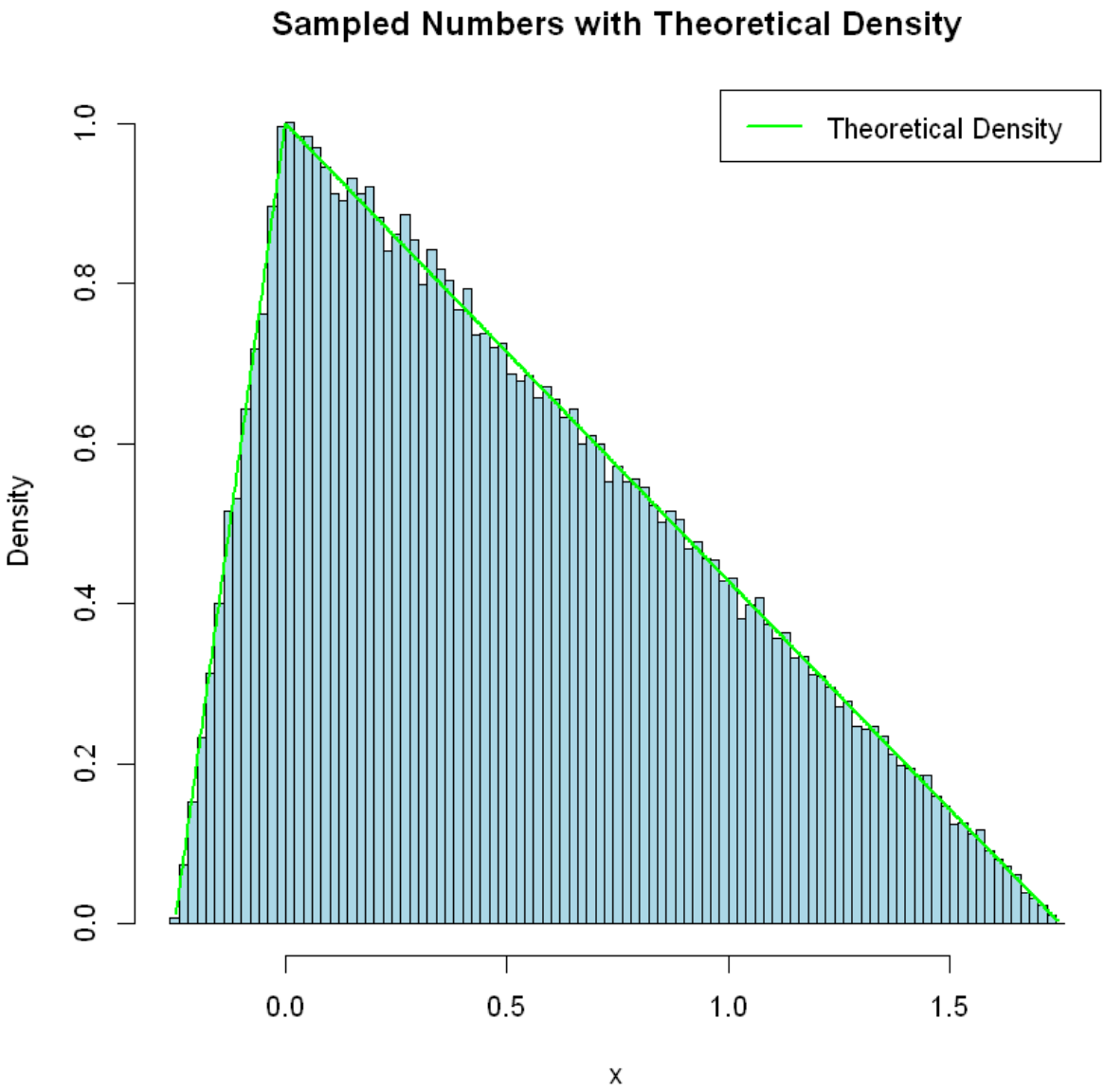
Therefore the Final CDF and Quantile Function is given by:

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

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

In [3]:
```r
samplingHelper <- function(n) {
     p <- runif(n) #Uniform Random Number generation
    numbers <- numeric(n) #Vector storage intialisation
    #Quantile Function
    Q <- function(p) {
    if (p < 1/8) {
      return(-1/4 + sqrt(2*p)/2)
    } else {
      return(7/4 * (1 - sqrt((8 * (1 - p))/7)))
    }
  }
    #Applying Quantile Function for uniform random number
    for (i in 1:n) {
    numbers[i] <- Q(p[i])
  }
     return(numbers) #Generated number returned.
}
```

In [11]:
```r
set.seed(123) # reproducibility
sampled_numbers <- samplingHelper(99999)

# Histogram plot with 100 bins
hist(sampled_numbers, breaks=100, probability=TRUE,
     main="Sampled Numbers with Theoretical Density",
     xlab="x", col="lightblue", border="black")

#theoretical density function f(x)
f <- function(x) {
  ifelse(x < -1/4, 0,
         ifelse(x < 0, 4 * x + 1,
                ifelse(x < 7/4, -4/7 * x + 1, 0)))
}

# x values sequence for plotting the theoretical density
x_vals <- seq(min(sampled_numbers), max(sampled_numbers), length.out=1000)
y_vals <- f(x_vals)

# theoretical density curve
lines(x_vals, y_vals, col="green", lwd=2)

#Legend
legend("topright", legend=c("Theoretical Density"), col="green", lwd=2)
```



Explanation

(d): While sampling for a given pdf we first derrive the cumulative distribution fucntion at each respective interval, and for the derived CDF we find the inverse of it which is known as the Quantile function. By using the quantile function we can transform the uniform random number to obtain samples from the desired distribution.

This sampling method is known as inverse transform sampling as it implements properties of the CDF and it's inverse. The idea here is that, for a uniform random numbers p are transformed using the quantile function Q(p) which ensure the generated numbers follow the specific distribution. Here the quantile function Q(p) accuratley maps the unifrom distribution into desired distribution by taking inverse of the CDF.

Answer (e) :

Given: The Pdf is given as follows:

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

Step1: Let's Find the CDF for f(x):
The CDF is defined as $f(x) = \int_{-\infty}^{x} f(t)\,dt$ Considering the pdf we solve as follows:

$$F(x) = \int_{-\infty}^{x} e^{-t^2\pi}\,dt$$

substituting $u = t\sqrt{\pi})$, we get:

$$F(x) = \sqrt{\frac{\pi}{4}} \int_{-\infty}^{x\sqrt{\pi}} e^{-u^2}\,du$$

The integral is related to the error function $(\text{erf}(x))$:

$$F(x) = \frac{1}{2}(1 + \text{erf}(x\sqrt{\pi}))$$

where $\text{erf}(x)$ is defined as: $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,dt$

Therefore, the cumulative distribution function F(x) is:

$$F(x) = \frac{1}{2}(1 + \text{erf}(x\sqrt{\pi}))$$

Now We find the Quantile Function Q(p): Given : Given:

$$p = F(x) = \frac{1}{2}(1 + \text{erf}(x\sqrt{\pi}))$$

We solve for x in terms of p :

$$2p = 1 + \text{erf}(x\sqrt{\pi})$$

$$\text{erf}(x\sqrt{\pi}) = 2p - 1$$

$$x\sqrt{\pi} = \text{erf}^{-1}(2p - 1)$$

$$x = \frac{\text{erf}^{-1}(2p - 1)}{\sqrt{\pi}}$$

Therefore, the quantile function Q(p) is:

$$Q(p) = \frac{\text{erf}^{-1}(2p - 1)}{\sqrt{\pi}}$$

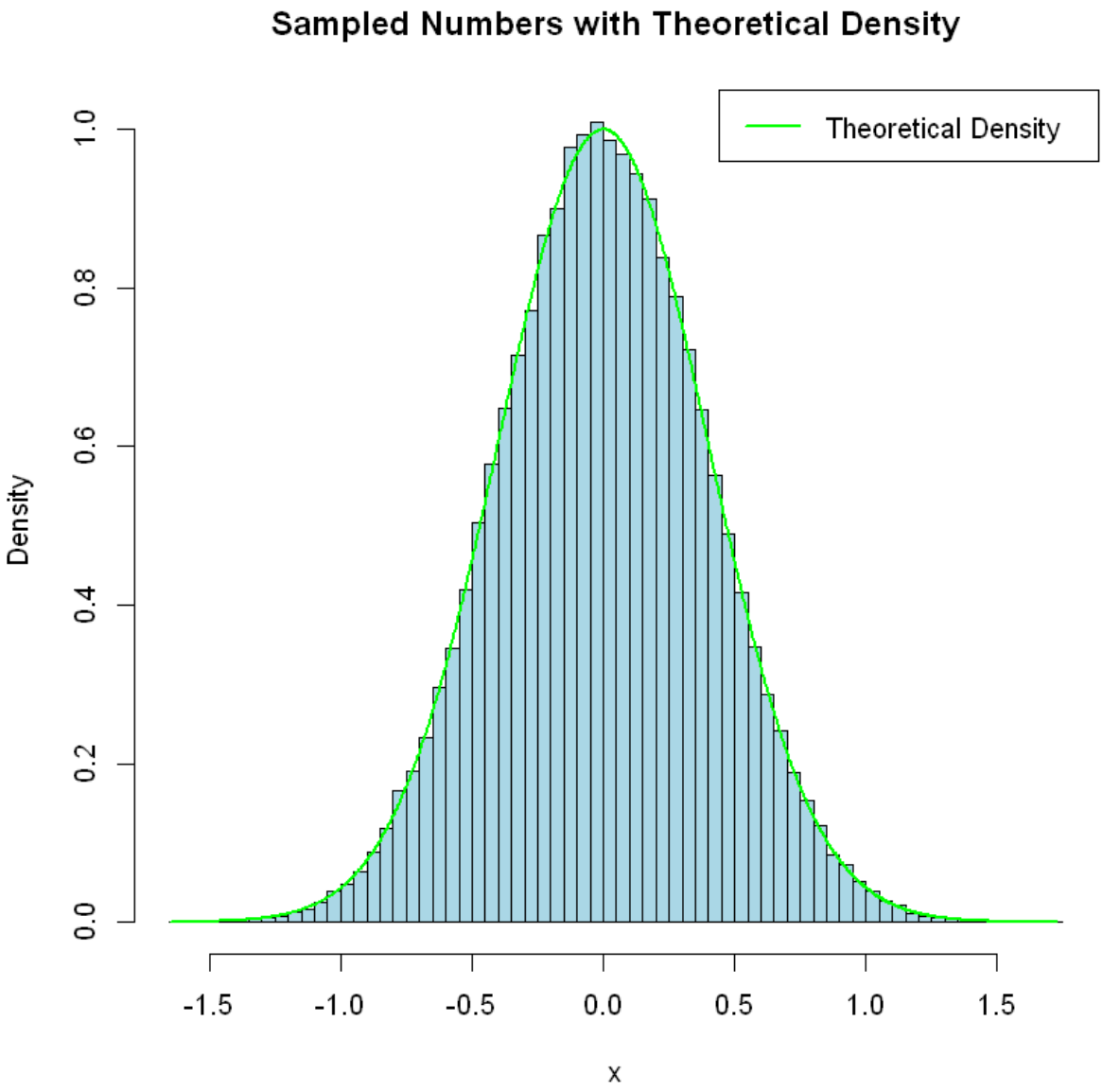
Therefore the CDF and Qunatile function is given as follows:

$$F(x) = \frac{1}{2}(1 + \text{erf}(x\sqrt{\pi}))$$

$$Q(p) = \frac{\text{erf}^{-1}(2p - 1)}{\sqrt{\pi}}$$

In [6]:
```r
newSamplingHelper <- function(n){
 p <- runif(n)

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

  #  erf function
  erf <- function(x) 2 * pnorm(x * sqrt(2)) - 1

  # Defining approximate inverse erf function
  erfinv <- function(y) qnorm((y + 1) / 2) / sqrt(2)

  # quantile function Q(p)
  Q <- function(p) {
    return(erfinv(2 * p - 1) / sqrt(pi))
  }

  # random numbers generation using the quantile function Q(p)
  for (i in 1:n) {
    numbers[i] <- Q(p[i])
  }

  # array of generated numbers
  return(numbers)
}
```

In [12]:
```r
set.seed(123) # eproducibility
sampled_numbers <- newSamplingHelper(99999)

# Histogram plot with 100 bins
hist(sampled_numbers, breaks=100, probability=TRUE,
     main="Sampled Numbers with Theoretical Density",
     xlab="x", col="lightblue", border="black")

# theoretical density function f(x)
f <- function(x) {
  exp(-x^2 * pi)
}

# x values sequence for plotting the theoretical density
x_vals <- seq(min(sampled_numbers), max(sampled_numbers), length.out=1000)
y_vals <- f(x_vals)

# theoretical density curve
lines(x_vals, y_vals, col="green", lwd=2)

# legend
legend("topright", legend=c("Theoretical Density"), col="green", lwd=2)
```



Answer: (g)

The histogram of the sampled numbers resembles the theoretical density function $f(x) = e^{-x^2\pi}$ above, which is a Gaussian-like curve. The curve is similar to that of a standard Gaussian distribution with a peak at x=0 and a tail that diminish symmetrically on both sides. The distribution generated using equation above, however, has the factor of π, which influences the speed at which the tails diminish. The higher the value of π, the steeper the distribution with less thick tails, and the lower the value, the flatter the distribution with fatter tails.

In [ ]: