Assignment- PBSR

Himanshu, MDS202327

2023-10-15

Question 1

Consider a continuous random variable X with probability density function

$$f(x) = \frac{ax^{a-1}}{c^a}; 0 \leq x \leq c$$

(i) Obtain the Cumulative Distribution function for the given pdf.

The cumulative distribution function is given by

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} f_X(x)\ dx \\
&= \int_0^x \frac{at^{a-1}}{c^a} dt \\
&= \frac{a}{c^a} \int_0^x t^{a-1} dt \\
&= \frac{a}{c^a} \left[\frac{t^a}{a}\right]_0^x \\
&= \frac{a}{c^a} \left[\frac{x^a}{a} - 0\right] \\
&= \frac{x^a}{c^a}
\end{aligned}$$

(ii) Obtain the inverted CDF

The inverted CDF is obtained by solving for $x = \frac{y^a}{c^a}$

$$\begin{aligned}
\frac{y^a}{c^a} &= x \\
y^a &= x\ c^a \\
y &= c\ x^{\frac{1}{a}}
\end{aligned}$$

(iii) Write a function in R to generate a sample from the given distribution using the inverse transform method which takes in parameters a and c.

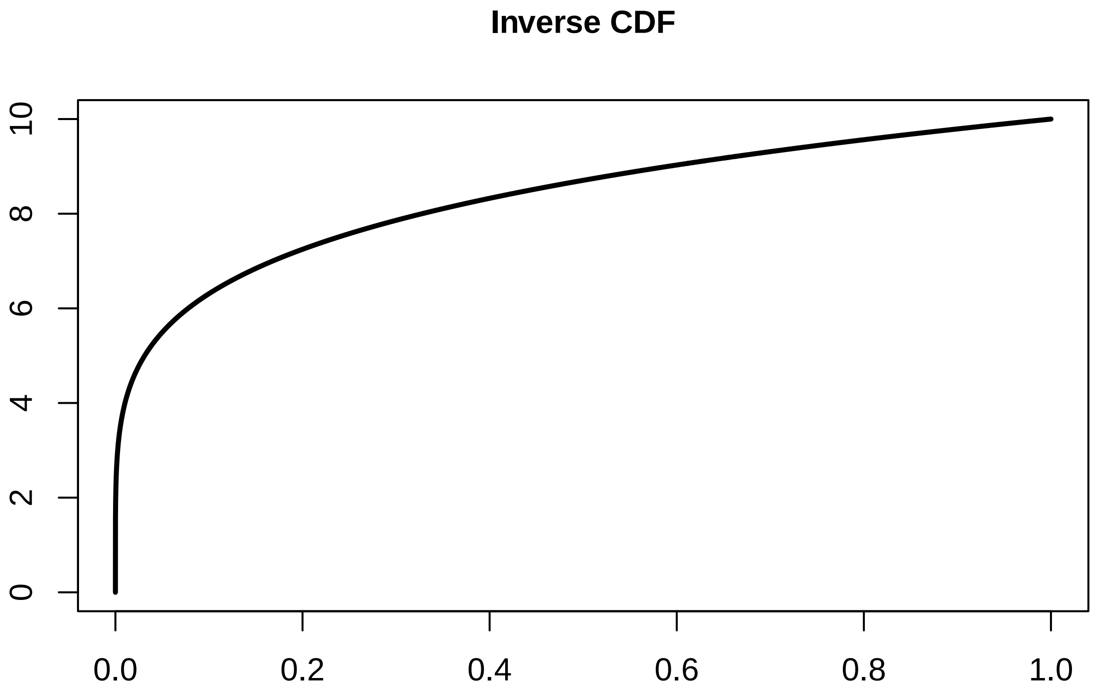
```r
# Inverse CDF
inverse_cdf = function(x, a, c) {
    c*(x^(1/a))
}
```

```r
# Generate 10000 random sample form uniform(0,1)
u = runif(10000)

# Plug in u in place of x to generate realization from f
f_samples = inverse_cdf(x=u, a=5, c=10)
```

(iv) Using R, plot the inverse CDF when a = 5 and c = 10 and determine the range of values with the least probability of being generated

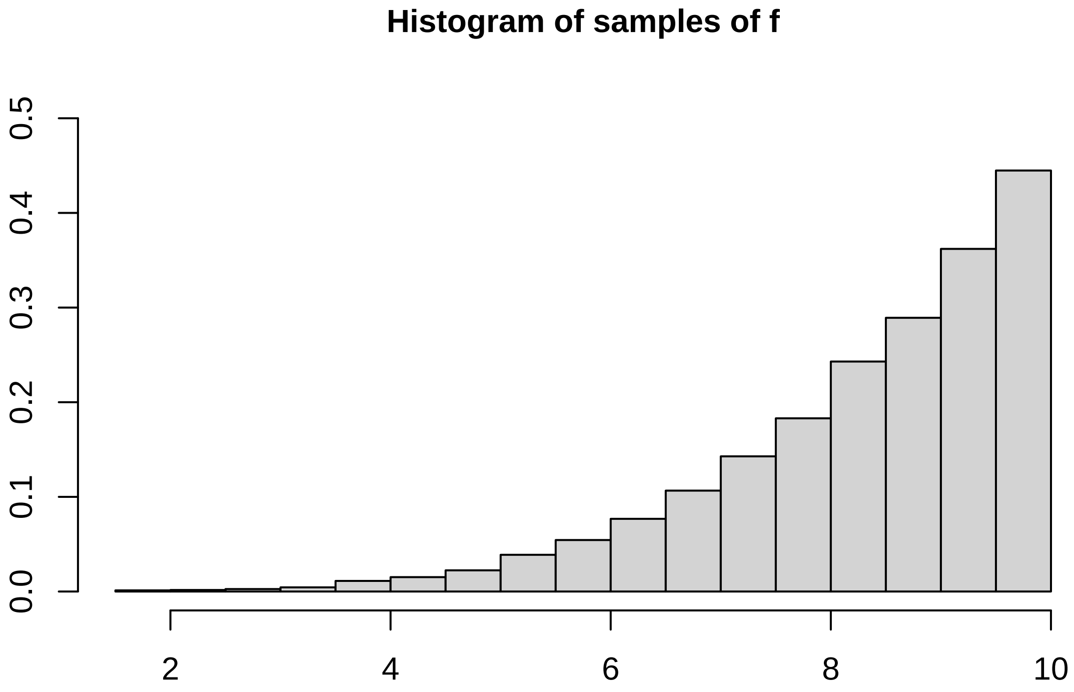
```r
x_val = seq(0,1,0.0001)
par(cex.main=1)
plot(x_val,inverse_cdf(x_val,5,10), type="l",
     lwd=2.5, main="Inverse CDF", xlab="", ylab="")
```



The range of values in the interval [0,4] has a probability of being generated ~1%

(v) Use the above function to draw a sample of size 1000 from the given distribution with parameter values a = 5 and c = 10

```r
par(cex.main=1)
hist(f_samples, main="Histogram of samples of f",
     xlab = "", ylab = "", probability = TRUE, ylim = c(0,0.5))
```



(vi) Using R, check if the empirical mean is approximately equal to the theoretical mean.

```
empirical_mean = mean(f_samples)
print(paste("Empirical Mean", round(empirical_mean,3)))
```

```
## [1] "Empirical Mean 8.326"
```

```
a =5
c = 10
actual_mean = a*c/(a+1)
print(paste("Actual Mean is", round(actual_mean,3)))
```

```
## [1] "Actual Mean is 8.333"
```

Question 2

You are conducting a study to understand the distribution of the average heights of people in a large city. You have access to the heights of 10,000 individuals, but collecting data from such a large population is time-consuming and costly. Instead, you decide to use a random sample of 100 individuals to estimate the population mean height.

(i) Simulate the heights of 10,000 individuals in the city. Assume that the heights follow a) Uniform, b) Poisson, c) Normal distribution. Assume the mean and variance for each distribution is 160 cm and 160 cm2 respectively.

```
# Samples from Uniform Distribution with mean = 160, variance = 160
unif = runif(10000, min = 160-4*sqrt(30), max = 160+4*sqrt(30))
# Samples from Poisson Distribution with actual mean = 60, variance = 160
pois = rpois(10000,lambda = 160)
# Samples from Normal Distribution with actual mean = 160, variance = 160
norm = rnorm(10000, mean = 160, sd = 4*sqrt(10))
```

(ii) Randomly select 100 individuals from all three (a, b, c above) simulated population and calculate the sample mean height.

```
unif_sample = sample(unif, size = 100)
print(paste("Sample mean height from Uniform Distribution is",
            round(mean(unif_sample),3)))
```

```
## [1] "Sample mean height from Uniform Distribution is 158.113"
```

```
pois_sample = sample(pois, size = 100)
print(paste("Sample mean height from Poisson Distribution is",
            round(mean(pois_sample),3)))
```

```
## [1] "Sample mean height from Poisson Distribution is 160.35"
```

```
norm_sample = sample(norm, size = 100)
print(paste("Sample mean height from Normal Distribution is",
            round(mean(norm_sample),3)))
```

```
## [1] "Sample mean height from Normal Distribution is 160.721"
```

(iii) Repeat step 2 a large number of times (e.g., 1,000 or more) to create a distribution of sample means.

```
unif_dist = c()
pois_dist = c()
norm_dist = c()

for (i in 1:1000){
  u = sample(unif, size = 100)
  p = sample(pois, size = 100)
  n = sample(norm, size = 100)

  unif_dist = append(unif_dist, mean(u))
  pois_dist = append(pois_dist, mean(p))
  norm_dist = append(norm_dist, mean(n))
}
```

(iv) Calculate the population mean height and the standard error of the sample mean from the simulated data.

```
se_unif = sd(unif_dist)/sqrt(1000)
se_pois = sd(pois_dist)/sqrt(1000)
```

```
se_norm = sd(norm_dist)/sqrt(1000)

print(paste("Population mean height from Uniform is",
            round(mean(unif_dist),3), "and standard error is", round(se_unif,4)))
```

```
## [1] "Population mean height from Uniform is 159.868 and standard error is 0.0394"
```

```
print(paste("Population mean height from Poisson is",
            round(mean(pois_dist),3), "and standard error is", round(se_pois,4)))
```

```
## [1] "Population mean height from Poisson is 159.978 and standard error is 0.0391"
```

```
print(paste("Population mean height from Normal is",
            round(mean(norm_dist),3), "and standard error is", round(se_norm,4)))
```

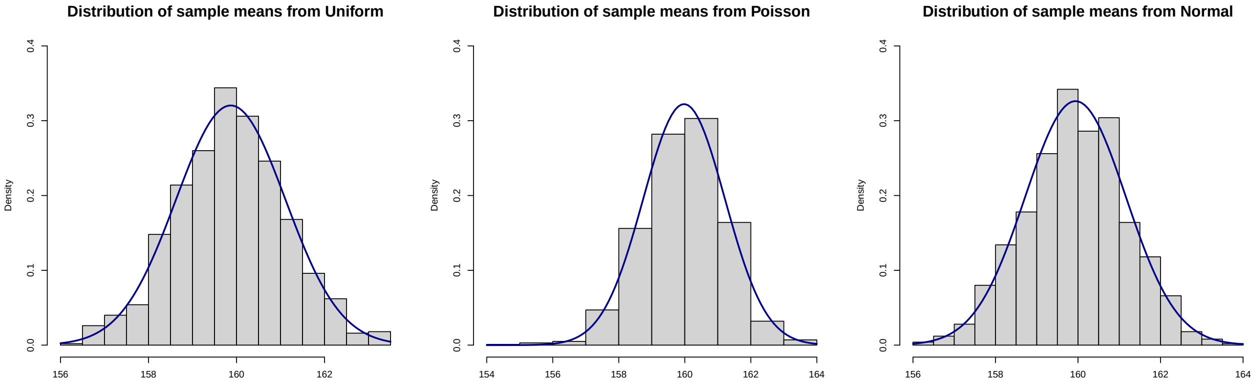
```
## [1] "Population mean height from Normal is 159.939 and standard error is 0.0387"
```

(v) Plot a histogram of the distribution of sample means and overlay a normal density plot with mean and standard deviation calculated from previous question.

```
par(mfrow = c(1, 3))
par(cex.main=1.7)
hist(unif_dist, main = "Distribution of sample means from Uniform",
     xlab = "", y_lab = "", probability = TRUE, ylim=c(0,0.4))
curve(dnorm(x, mean=mean(unif_dist), sd=sd(unif_dist)),
      col="darkblue", lwd=2, add=TRUE)

hist(pois_dist, main = "Distribution of sample means from Poisson",
     xlab = "", y_lab = "", probability = TRUE, ylim=c(0,0.4))
curve(dnorm(x, mean=mean(pois_dist), sd=sd(pois_dist)),
      col="darkblue", lwd=2, add=TRUE)

hist(norm_dist, main = "Distribution of sample means from Normal",
     xlab = "", y_lab = "", probability = TRUE, ylim=c(0,0.4))
curve(dnorm(x, mean=mean(norm_dist), sd=sd(norm_dist)),
      col="darkblue", lwd=2, add=TRUE)
```



(vi) Discuss your findings and any insights regarding the sample mean distribution compared to the population distribution.

Although the population distribution followed Uniform, Poisson and Normal with mean = 160 and variance = 160 but the sample mean distribution seems to fit right like a Normal Distribution with mean ~160 and standard deviation ~1.25

Question 3

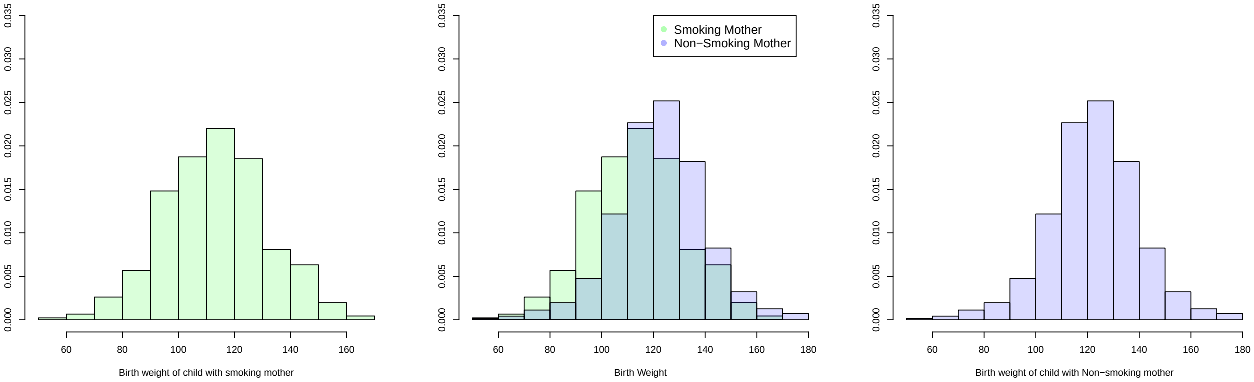
In this question, you will determine whether pregnancy smoking is harmful to a child's health. You will use the ”Birth Weight” and ”Maternal Smoker” columns from the baby dataset.

(i) Plot overlaying histograms of birth weights of babies born to non-smoking mothers and smoking mothers. Write down your observations about the mean of each distribution.

```r
par(mfrow=c(1,3))
par(cex.main=1.7)
hist(df[df$Maternal.Smoker=="True",]$Birth.Weight, col = rgb(0, 1, 0, 0.15),
     main = "", xlab="Birth weight of child with smoking mother", ylab="",
     probability = TRUE, ylim=c(0,0.035))

hist(df[df$Maternal.Smoker=="True",]$Birth.Weight, col = rgb(0, 1, 0, 0.15),
     main = "", xlab="Birth Weight", ylab="", probability = TRUE,
     ylim=c(0,0.035), xlim=c(50,180))
hist(df[df$Maternal.Smoker=="False",]$Birth.Weight, col = rgb(0, 0, 1, 0.15),
     probability = TRUE, add=TRUE)
legend(120, 0.035, legend=c("Smoking Mother", "Non-Smoking Mother"),
       col=c(rgb(0, 1, 0, 0.3), rgb(0, 0, 1, 0.3)), pch = 19, cex=1.2)

hist(df[df$Maternal.Smoker=="False",]$Birth.Weight, col = rgb(0, 0, 1, 0.15),
     main = "", xlab="Birth weight of child with Non-smoking mother",
     ylab="", probability = TRUE, ylim=c(0,0.035))
```



The mean of birth weight of child with smoking mother seems to be a bit less than the mean of birth weight of child with non-smoking mother as can be seen in the histogram of both.

(ii) Find the difference between the average weight of the smoking group and the average weight of the non-smoking group. This is the observed difference.

```r
obs_diff = abs(mean(df[df$Maternal.Smoker=="True",]$Birth.Weight) -
                 mean(df[df$Maternal.Smoker=="False",]$Birth.Weight))
cat(paste("The mean of birth weight of child with smoking mother and
mean of birth weight of child with non-smoking mother",
round(mean(df[df$Maternal.Smoker=="True",]$Birth.Weight),3), "and
",round(mean(df[df$Maternal.Smoker=="False",]$Birth.Weight),3),
"respectively Hence the difference between them is", round(obs_diff,3)))
```

```
## The mean of birth weight of child with smoking mother and
## mean of birth weight of child with non-smoking mother 113.819 and
##  123.085 respectively Hence the difference between them is 9.266
```

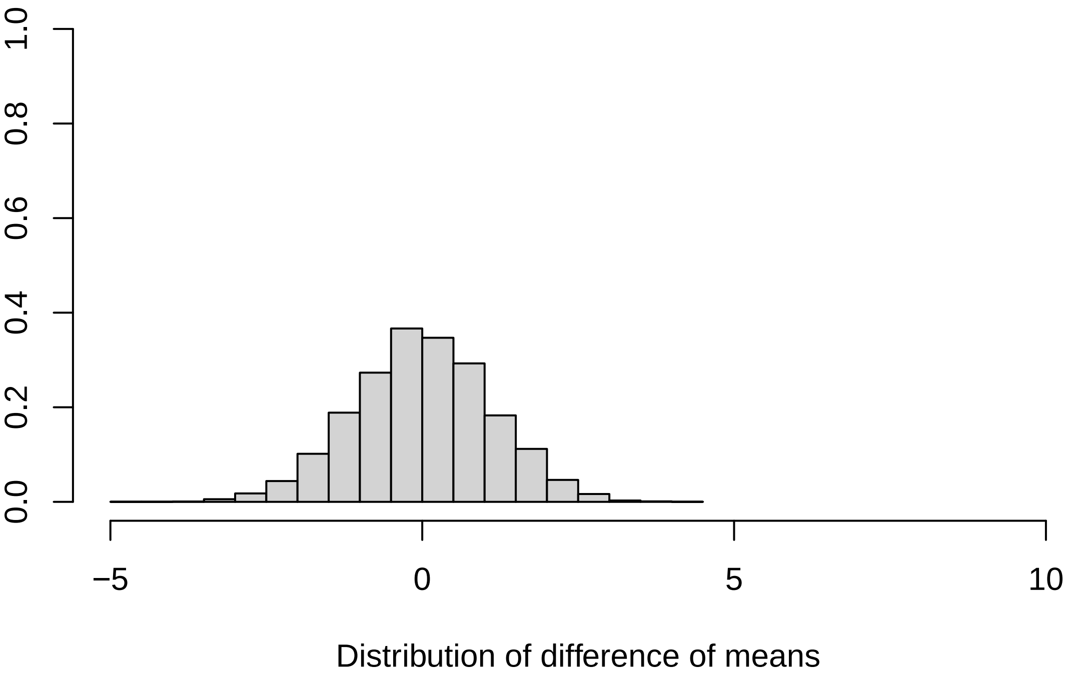
The mean of birth weight of child with smoking mother and mean of birth weight of child with non-smoking mother is 113.8192 and 123.0853 respectively, hence the difference between them is 9.2661

(iii) If there were no difference between the two distributions in the underlying population, then whether a birth weight has the label True or False with respect to maternal smoking should make no difference to the average. The idea, then, is to randomly shuffle all the labels among the mothers. This process is called random permutation. Using this idea, shuffle the labels and find the difference you found in (ii) for this new labelling. Repeat this step a large number of times (1,000 or more), calculate the mean and variance of the differences, and plot a histogram.

```r
diff = c()
for (i in 1:10000) {
  df$Maternal.Smoker.New = sample(df$Maternal.Smoker)
  diff_val = (mean(df[df$Maternal.Smoker.New=="True",]$Birth.Weight)-
            mean(df[df$Maternal.Smoker.New=="False",]$Birth.Weight))
  diff = append(diff, diff_val)
}
print(paste("The mean and variance of differences are",
            round(mean(diff),3), round(var(diff),4)))
```

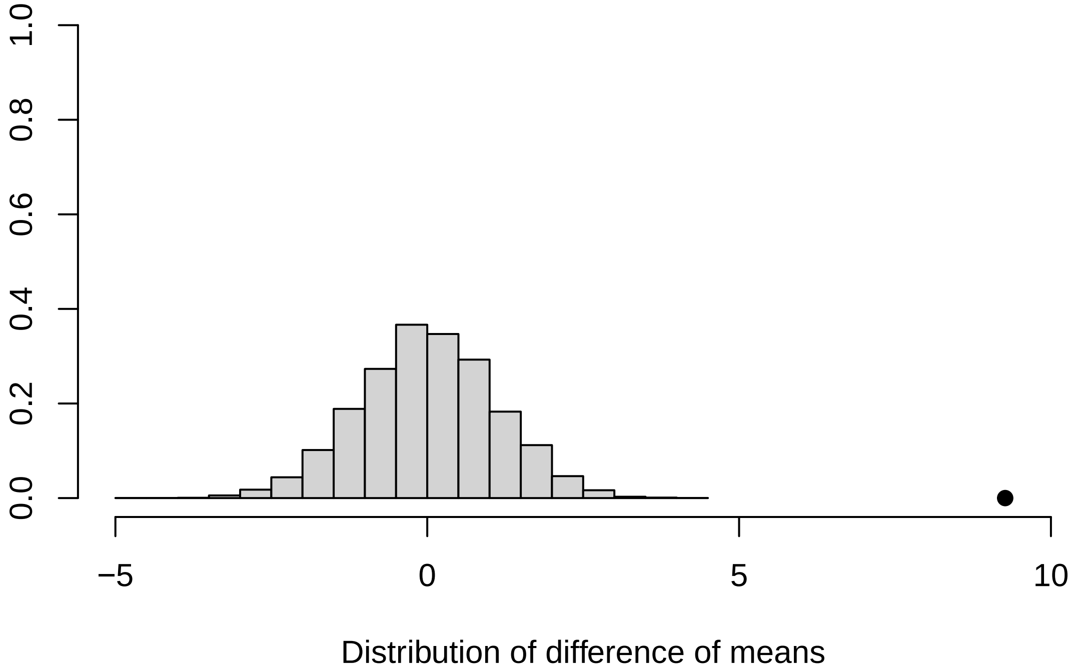
```
## [1] "The mean and variance of differences are 0.006 1.1957"
```

```r
hist(diff, main = "", xlab="Distribution of difference of means",
     ylab="", probability = TRUE, ylim = c(0,1), xlim = c(-5,10))
```



(iv)Plot the observed difference you obtained in (ii) and overlay this point to the histogram mentioned in (iii).

```r
hist(diff, main = "", xlab="Distribution of difference of means",
     ylab="", probability = TRUE, ylim = c(0,1), xlim=c(-5,10))
points(x=obs_diff,y=0,pch=19)
```



(v) Calculate how many standard deviations the observed difference is from the mean of the simulated variables. What conclusions can you draw from this?

```
k = abs(mean(diff)-obs_diff) / sd(diff)
print(paste("The observed difference is", round(k,2),
            "s.d. away from simulated mean of differences"))
```

```
## [1] "The observed difference is 8.47 s.d. away from simulated mean of differences"
```

Using the Chebychev's Inequality, that says, for any $k > 0$ we have

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Since, we got k = 8, from Chebychev's Inequality, we get $P(|X - \mu| \geq 8\sigma) \leq \frac{1}{8^2} = 0.015625$. This means that the probability to observe a value as large as 9.2661 is less than 1.56 %. Hence we can conclude that the two distributions in the underlying population are different.