Northeastern University
College of Science

Module 3 Homework

1) **(10 points)** A random variable X has pdf

$$f(x) = \frac{3^x e^{-3}}{x!}, \quad x = 0, 1, 2, \ldots \ldots.$$

Find P(X = 1). Then find $P(-3 < X < 5)$.

CODE:

```
compute_factorial <- function(n) {
  if (n == 0) return(1)
  else return(n * compute_factorial(n - 1))
}

"for random variable X"
calculate_pdf_X <- function(x) {
  return((3^x * exp(-3)) / compute_factorial(x))
}

# P(X = 1)
probability_X_equals_1 <- calculate_pdf_X(1)

# Calculate P(-3 < X < 5)
probability_minus_3_to_5 <- sum(sapply(seq(0, 5), calculate_pdf_X))

# Print results
cat("P(X = 1):", probability_X_equals_1, "\n")
cat("P(-3 < X < 5):", probability_minus_3_to_5, "\n")
```

OUTPUT:

P(X = 1): 0.1493612

P(-3 < X < 5): 0.9160821

2) **(6 points)** If two carriers of the gene for albinism marry and have children, then each of their children has a probability of 1/7 of being albino. Let the random variable Y denote the number of their children having the gene for albinism out of all 5 of their children. Then Y follows a binomial(n, p) distribution. Find the values for n and p.

n =_____ p = _____

ANSWER:

When two individuals who carry the gene for albinism get married and have children, each of their offspring has a 1/7 chance of inheriting the gene and being albino. Out of a total of 5 children, we establish a random variable, Y, to reflect the proportion of their offspring that have the albinism gene. Y has a binomial distribution, and the following are its parameters:
"n" refers to the total number of kids, which is 5.
"p" stands for the 1/7 possibility that any kid will carry the albinism gene.

n = 5
p = 1/7

3) **(9 points)** For Y following a binomial (n = 3, p = 0.25) distribution, compute the following:
   a) P(Y ≤ 2) =
   b) E(Y) =
   c) Var(Y) =

ANSWER:
the calculations for the Y B(3, 0.25), or binomial distribution, given:

a) P(Y $\leqslant$ 2):
   P(Y = 0) = (3 choose 0) * 0.25^0 * 0.75^3 = 1 * 1 * 0.421875 = 0.421875
   P(Y = 1) = (3 choose 1) * 0.25^1 * 0.75^2 = 3 * 0.25 * 0.5625 = 0.421875
   P(Y = 2) = (3 choose 2) * 0.25^2 * 0.75^1 = 3 * 0.0625 * 0.75 = 0.140625
   AFTER SUMMING UP THE PROBABLITIES:
   P(Y $\leqslant$ 2) = P(Y = 0) + P(Y = 1) + P(Y = 2) = 0.421875 + 0.421875 + 0.140625 = **0.984375**

b) E(Y) (Expected Value):

E(Y) = n * p = 3 * 0.25 = **0.75**

c) Var(Y) (Variance):

Var(Y) = n * p * (1 - p) = 3 * 0.25 * (1 - 0.25) = 3 * 0.25 * 0.75 = **0.5625**

4) For X following a Chi-square distribution with degree of freedom m = 5, compute the following:

CODE:

```
"degree of freedom"
dof <- 5

# a) Compute P(2 < X < 5)
probability <- pchisq(5, dof) - pchisq(2, dof)

# b) Compute E(X)
mean <- dof

# c) Compute Var(X)
var <- 2 * dof

# d) Monte Carlo simulation
set.seed(123)
n_simul <- 100000
simul_data <- rchisq(n_simul, dof)
d_probability <- sum(simul_data > 2 & simul_data < 5) / n_simul


cat("a) P(2 < X < 5) =", probability, "\n")
cat("b) E(X) =", mean, "\n")
cat("c) Var(X) =", var, "\n")  # Corrected variable name
cat("d) Monte Carlo estimate =", d_probability, "\n")  # Corrected variable name
```

```
# if Monte Carlo estimate agrees with a)
if (abs(d_probability - probability) < 0.001) {
  cat("Monte Carlo estimate agrees with a)\n")
} else {
  cat("Monte Carlo estimate does not agree with a)\n")
}
```

a) **(5 points)** P(2 < X < 5) =

OUTPUT:
P(2 < X < 5) = 0.4332648

b) **(2 points)** E(X) =

OUTPUT:
E(X) = 5

c) **(2 points)** Var(X) =

OUTPUT:
Var(X) = 10

d) **(6 points)** Also, use a Monte Carlo simulation with sample size n=100,000 to estimate P(2 < X < 5). What is your Monte Carlo estimate? Does it agree with the answer in a)?

OUTPUT:
Monte Carlo estimate = 0.4325

5) Suppose X follows a Chi-square distribution with degree of freedom m = 6 so that E(X) = 6 and Var(X) = 12. Also, let Y = 3X - 5.

a) **(8 points)** Find E(Y) and Var (Y).

ANSWER:
X has a Chi-square distribution with m = 6 degrees of freedom, and we know that Y = 3X - 5. We'll determine Y's variance (Var(Y)) and expected value (E(Y)).

E(Y):
E(Y) = E(3X-5), which is 3 * E(X) - 5, or 3 * 6 - 5 = 18 - 5 = 13.

Var(Y):
Var(Y) = Var(3X-5), which is 32 * Var(X), which is 9 * 12 = 108.

Therefore, a) E(Y) = 13 and b) Var(Y) = 108

b) **(2 points)** Does Y follow a Chi-square distribution with degree of freedom m=6?

ANSWER:
No, Y does not have a m = 6 degrees of freedom Chi-square distribution. The sum of squares of independent standard normal random variables or the sum of squares of independent normal random variables are commonly used for the Chi-square distribution, which is defined for non-negative values of a random variable.

Y = 3X - 5 is a linear transformation of X, not a sum of squares of independent variables, and it can take negative values. As a result, Y does not have a m = 6 degree of freedom Chi-square distribution.

6) **(30 points)** The distribution of the expression values of the patients with the Zyxin gene are distributed according to $N(\mu = 1.6, \sigma = 0.4)$.
   a. What is the probability that a randomly chosen patient have the Zyxin gene expression values between 1 and 1.6?
   b. Use a Monte Carlo simulation of sample size n=500,000 to estimate the probability in part (a). Give your R code, and show the value of your estimate.
   c. What is the probability that exactly 2 out of 5 patients have the Zyxin gene expression values between 1 and 1.6? Please show your work on how to arrive at the answer. Give your answer to at least four decimal places.

CODE:

```
#a)
set.seed(98765)

monte_simulations <- 500000


zyxin_gene_exp <- rnorm(monte_simulations, mean = 1.6, sd = 0.4)
```

```
prob_btw_1_and_1.6 <- mean(zyxin_gene_exp > 1 & zyxin_gene_exp < 1.6)


prob_btw_1_and_1.6
#b)
zyxin_gene_patient_trials <- 5
zyxin_gene_patient_successes <- 2

success_gene_prob <- prob_btw_1_and_1.6

#c)
prob_2_out_of_5 <- choose(zyxin_gene_patient_trials,
zyxin_gene_patient_successes) *
success_gene_prob^zyxin_gene_patient_successes * (1 -
success_gene_prob)^(zyxin_gene_patient_trials - zyxin_gene_patient_successes)


prob_2_out_of_5
```

OUTPUT:

```
[1] 0.433192
[1] 0.3417187
```

7) **(20 points) Consider two random variables X and Y following F distribution. Note that this is a continuous distribution. DO NOT use Monte Carlo for any part of this question. In other words, do NOT use "rf" to do this question.**
a) Hand in a R script that calculates the mean and variance of two random variables X~F(m=1,n=10) and Y~F(m=12,n=10) **from their density functions**.
(b) Use the formula in Table 3.4.1 to calculate the means and variances directly.
(c) Run your script in (a), and check that your answers agree with those from part (b).

CODE:

```
calculate_F_distribution <- function(m, n) {
 if (m <= 0 || n <= 0) {
  stop("Degrees of freedom for the F-distribution must be greater than zero.")
 }
```

```
  # Calculate mean using the formula
  mean <- n / (n - 2)

  # Calculate variance using the formula
  variance <- (2 * n^2 * (m + n - 2)) / (m * (n - 2)^2 * (n - 4))

  return(list(mean = mean, variance = variance))
}

# Parameters for X
m_X <- 1
n_X <- 10

# Parameters for Y
m_Y <- 12
n_Y <- 10

# Calculate mean and variance using the function
X <- calculate_F_distribution(m_X, n_X)
Y <- calculate_F_distribution(m_Y, n_Y)

cat("For X ~ F(m=1, n=10):\n")
cat("Mean:", X$mean, "\n")
cat("Variance:", X$variance, "\n\n")

cat("For Y ~ F(m=12, n=10):\n")
cat("Mean:", Y$mean, "\n")
cat("Variance:", Y$variance, "\n")

# Direct calculation using formulas
mean_X_direct <- n_X / (n_X - 2)  # Formula for mean
variance_X_direct <- (2 * n_X^2 * (m_X + n_X - 2)) / (m_X * (n_X - 2)^2 * (n_X - 4))
mean_Y_direct <- n_Y / (n_Y - 2)
variance_Y_direct <- (2 * n_Y^2 * (m_Y + n_Y - 2)) / (m_Y * (n_Y - 2)^2 * (n_Y - 4))

cat("\nDirect Calculation (Means and Variances):\n")
cat("For X ~ F(m=1, n=10):\n")
cat("Mean:", mean_X_direct, "\n")
```

```
cat("Variance:", variance_X_direct, "\n\n")

cat("For Y ~ F(m=12, n=10):\n")
cat("Mean:", mean_Y_direct, "\n")
cat("Variances:", variance_Y_direct, "\n")
```

Note: Make sure that you clearly understand the difference between discrete and continuous random variables and use the appropriate functions to find the expected value, variance etc. It is conceptually wrong to use “integrate” for discrete random variables and “sum” for continuous random variables and very few points will be given if you get this wrong.