Statistics - Homework 5 (Due May 21, 2021)

1. Two normal random variables X and Y are jointly distributed with

   $Var(X) = 25$ and $Var(Y) = 1600$. It is known that $P(Y > 80 \mid X = 50) = 0.1$ and

   $P(Y > 22 \mid X = 40) = 0.7886$.

   (1) What is the correlation coefficient between *X* and *Y* ?

   (2) What is the expected value of *Y* given *X* = 50 ?

2. Let X be a random variable with a standard normal density and $Y=X^2$. Prove that Y is a random variable with a chi-squared distribution with degree of freedom 1.

3. Fifty balls which are of equal size and numbered in sequence from 1 to 50 are placed in an opaque box. A random experiment is conducted by repeating the following steps *three times* in a row:

   (i) Draw one ball from the opaque box,

   (ii) Record the number marked on the ball, and

   (iii) Return the ball to the box.

   Students A and B are asked to conduct the above random experiment independently.

   (1) What is the probability that the minimum of the three numbers recorded by student A is greater than the maximum of the three numbers recorded by student B?

   (2) Develop your own R code to estimate the above probability. (A reference R code is provided at the end of this homework file.)

   (3) In a gambling play, student A and student B represent the player (gambler) and the banker, respectively. A player needs to pay the banker 20 Taiwan dollars in advance to play the game. The player will then receive 200 Taiwan dollars from the banker, if the minimum of the three numbers recorded by the player is greater than the maximum of the three numbers recorded by the banker. An addicted player plays the game daily. What is the expected amount of money the player will win or lose (i.e. the amount of money received from the banker minus the amount of money paid by the player) in a one-week period?

4. A scientist is given an assignment of determining the average *soil moisture content* (SMC) within a farm. Assuming soil moisture contents at different locations in the farm are identically distributed with expected value of μ and standard deviation of σ. The scientist plans to take SMC measurements at *n* locations and all measurements ($x_i,\ i = 1, \cdots, n$) can be considered as independent of others. The average SMC in the farm (μ) is then estimated by $\bar{x}_n$, the average of all SMC measurements.

   (1) If SMC are measured at 25 locations, what is the probability that $\bar{x}_n$ will fall in the range of $(\mu - 0.1\sigma,\ \mu + 0.1\sigma)$?

   (2) At least how many measurements need to be taken in order for the probability that $\bar{x}_n$ falls in the range of $(\mu - 0.1\sigma,\ \mu + 0.1\sigma)$ to be higher than 0.95?

5. Assuming the soil moisture content in Problem 3 can be characterized by a gamma density with mean and standard deviation being 15% and 9%, respectively. The *n* SMC measurements taken by the scientist can be considered as a random sample of size *n*. Conduct the following simulations using R:

(1) Simulate 10,000 random samples, each of sample size *n*, for *n* = 100, 200, 300, 400, 500, 1000, and 2000, respectively.
(2) For any of the above sample sizes, calculate the average soil moisture content ($\bar{x}_n$) of each individual random sample and determine the percentage (*p*) of random samples satisfying $\mu - 0.1\sigma \le \bar{x}_n \le \mu + 0.1\sigma$.
(3) *Based on the Weak Law of Large Numbers*, calculate the probabilities that $\bar{x}_n$ falls in the range of $(\mu - 0.1\sigma,\ \mu + 0.1\sigma)$ for *n* = 100, 200, 300, 400, 500, 1000, and 2000, respectively.
(4) Show scatter plots of (*n*, *p*) based on the results of (2) and (3), respectively. [Show values of *n* and *p* in the x-axis and y-axis, respectively.]
(5) Derive *p versus n* relationships for the two scatter plots in (4). [**Note**: You can derive the relationships *empirically or theoretically*.]

Reference R code:

```
# 2015 STAT HW-6 Problem 2 (Coded by KSC, 11/15/2015)
M=50  # X ~ U[1,M] discrete unif.
n=3   # random sample of size n from X
CDF.Y=matrix(0,ncol=n,nrow=M) # CDF of order statistics
PDF.Y=matrix(0,ncol=n,nrow=M) # PDF of order statistics
Fy1=function(y,M) {
      y=y/M
      3*y*(1-y)^2+3*y^2*(1-y)+y^3
}
Fy2=function(y,M) {
      y=y/M
      3*y^2*(1-y)+y^3
}
Fy3=function(y,M) {
      (y/M)^3
}
CDF.Y[,1]=Fy1(1:M,M)
CDF.Y[,2]=Fy2(1:M,M)
CDF.Y[,3]=Fy3(1:M,M)
PDF.Y=CDF.Y
for (i in 2:M) PDF.Y[i,]=CDF.Y[i,]-CDF.Y[i-1,]
p1=1-sum(CDF.Y[,1]*PDF.Y[,3]) # Prob. of (Y1 > Y3)
p2=sum(PDF.Y[,1]*CDF.Y[,3])   # Prob. of (Y1 >= Y3)
p3=p2-p1                      # Prob. of (Y1 = Y3)
c(p1,p2,p3)
#
S=seq(1,M,1)
N=1000000
count=0
for (i in 1:N) {
      y1=min(sample(S,n,replace=TRUE))
      y3=max(sample(S,n,replace=TRUE))
      if(y1>=y3) count=count+1
}
count/N
```