Assignment Week 12

STAT2011 Probability and Estimation Theory

Question 1

In a gambling game, a player wins the game if they roll 10 fair, six-sided dice, and get a sum of at least 40.

a. Approximate the probability of winning by simulating the game 104 times. Use set.seed(200) for this question.

b. Compute the Central Limit Theorem Approximation $P(Y \geq 40)$, where Y is the sum of 10 dice, and compare it to the Monte Carlo approximation obtained above. You can use the fact that $E(Y) = 35$, $Var(Y) = 175/6$.

```
#1.a. Monte Carlo approximation
set.seed(200)

wins = 0

for (i in 1:10000){
  rolls = sample(x=1:6, 10, replace=TRUE)
  if (sum(rolls) >= 40) {
    wins = wins + 1
  }
}

prob = wins/10000
prob

#1.b. central limit theorem
ans = 1-pnorm(40, mean = 35, sd = sqrt(175/6))
ans
```

Output:

```
[1] 0.1988

[1] 0.1772697
```

Question 2

The frequency table below summarises 320 counts,

Value	0	1	2	3	4
Freq	130	133	49	7	1

modelled as values taken by i.i.d. random variables with common Bin(4, p) distribution,

i.e. $P(X = x) = \binom{4}{x} p^x (1-p)^{4-x}$, for x = 0, 1, 2, 3, 4

for some unknown p.

a. Recall $E(X) = np$, estimate p using the method of moments.

b. Using (a), find expected frequencies (E) for each of the classes “0”, “1”, “2”, “3” and “4”. Round to the nearest integer.

c. Compute standardised residuals (SR) given by $SR = \frac{O-E}{\sqrt{E}}$ for each of the classes “0”, “1”, “2”, “3” and “4’, where O represents the observed frequencies. If $|SR| < 2$, then the fitted binomial model is said to be a good model for the data. Comment on the goodness of fit.

```
n = 4
x = 320
ex = (0*130 + 1*133 + 2*49 + 3*7 + 4*1)/x
p = ex/n
print(noquote(paste("p = ", p)))

#2.b
px <- function(a){
  b = (factorial(4)/(factorial(a)*factorial(4-a)) * p ^ a *
(1-p) ^ (4-a))
  return(b*x)
}
for (i in 0:4) {
  print(noquote(paste("E(",i,") =",trunc(px(i)))))
  #example: E(1) = 131
}
```

```
#2.c.
obs <- c(130, 133, 49, 7, 1)
SR <- function(a){
  (obs[a+1]-px(a))/sqrt(px(a))
}
for (i in 0:4) {
  print(noquote(paste("SR(",i,")=",SR(i))))
}
```

Output:

```
[1] p =  0.2

[1] E( 0 ) = 131

[1] E( 1 ) = 131

[1] E( 2 ) = 49

[1] E( 3 ) = 8

[1] E( 4 ) = 0

[1] SR( 0 )= -0.0936353465578064

[1] SR( 1 )= 0.16840386955545

[1] SR( 2 )= -0.021680684567916

[1] SR( 3 )= -0.416467660809336

[1] SR( 4 )= 0.682000733137436
```

Comments:

Since for all the standardised residuals the |SR| < 2, the fitted binomial model is a goodmodel for the data. The observed SR is quite far from 2 or -2, so the model is quite accurate.

Question 3

a. Generate a random sample of size 25 from a normal distribution with mean $\mu = 3$ and standard deviation $\sigma = 1.5$. Assume σ is known and we want to estimate μ. Using the sample generated, find a 95% confidence interval (CI) for μ.

b. Repeat the process in (a) 20 times. Using your 20 samples, calculate 20 CIs for μ. How many of these 20 intervals contain the true mean $\mu = 3$? Output this number from your code, but no need to print the 20 CIs themselves.

```
set.seed(100)
samples = rnorm(25, mean=3, sd=1.5)
SE = sd(samples)/sqrt(mean(samples))
lower = mean(samples)-SE
upper = mean(samples)+SE
print(noquote(paste("lower bound =",lower)))
print(noquote(paste("upper bound =",upper)))

ans = 0
for (i in 1:20) {
  ans = ans + 1
  samples = rnorm(25, mean=3, sd=1.5)
  SE = sd(samples)/sqrt(mean(samples))
  lower = mean(samples)-SE
  upper = mean(samples)+SE
  if ((lower <= 3) & (upper >= 3)) {
    ans = ans + 1
  }
}
print(ans)
```

Output:

```
[1] lower bound = 2.56938652543926

[1] upper bound = 3.75512951674226

[1] 40
```

Question 4

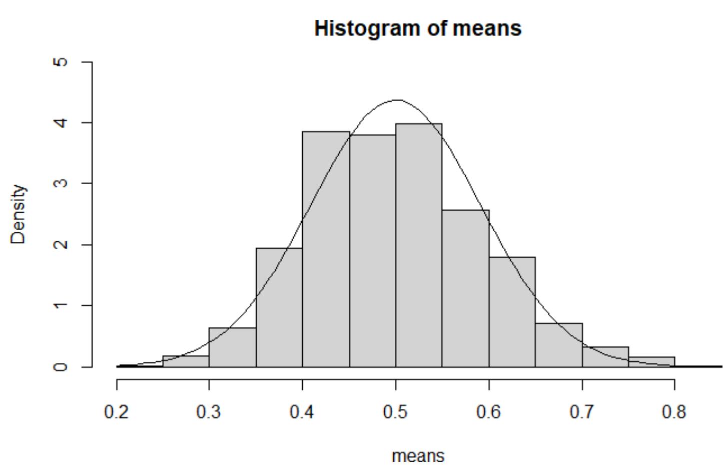
a. Generate a random sample of size 30 from the exponential distribution with parameter λ = 2 and find the mean of your sample. Repeat this process 1000 times and draw a histogram of these 1000 means (use prob=T in hist). (Do not print the 1000 means.)

b. Next we check whether the Central Limit Theorem gives a good approximation for the distribution of the means. Overlay the histogram with a normal density curve with appropriate mean and variance. (You will need to use the mean and variance of exponential distributions from lectures. No need to derive). Comment on the fit.

```
set.seed(100)
n = 30
means <- c()
lambda = 2
for (i in 1:1000){
  samples = rexp(n,lambda)
  means[i] = mean(samples)
}
hist(means, prob=T, ylim=c(0,5))

mean1 = 1/lambda
sd1 = sqrt(1/(lambda^2*n))
curve(dnorm(x, mean=mean1, sd=sd1), add=T)
```

Output:



Comments:

The Central Limit Theorem is a good approximation for the distribution of the means.The curve fits quite well with the histogram of the means, which means that the CLT is quiteaccurate.