1. **(5 points each)**

Exercise 8.5.7 What is consistency?

Solution: Consistency is a property of estimators in statistics, which measures the accuracy of an estimator as the sample size increases. An estimator is considered consistent if its estimate approaches the true value of the parameter being estimated as the sample size increases. In other words, as the sample size increases, the difference between the estimated value and the true value decreases, approaching zero.

For example, the sample mean is a consistent estimator for the population mean if the sample size increases, the sample mean approaches the population mean. This property is desirable because it implies that the estimate of the parameter will get closer to the true value as the sample size increases, providing more reliable and accurate results.

Consistency is a desirable property for estimators because it provides an assurance that the estimates will be as close as possible to the true value, as the sample size increases. It is an important aspect to consider when choosing an estimator for a given problem, as it helps to ensure that the results are reliable and accurate.

Exercise 8.5.15 At the end of Section 8.3, it was noted that for samples of serum cholesterol levels of size 25 drawn from a population with μ = 211mg / 100 ml and standard deviation σ = 46 mg/100 ml, the probability that a sample mean x lies within the interval (193.0, 229.0) is 0.95. Furthermore, the probability that the mean lies below 226.1 mg/100 ml is 0.95, and the probability that it is above 195.9 mg/100 ml is 0.95. For all three of these events to happen simultaneously, the sample mean $\bar{x}$ would have to lie in the interval (195.9, 226.1). What is the probability that this occurs?

Solution: mean of population, $\mu = 211mg/100ml$

Standard deviation of population, $\sigma = 46\ mg/100ml$

Size of population, $n = 25$

$Let\ \bar{x}\ be\ sample\ mean$

$P(193.0 < \bar{x} < 229.0) = 0.95$

$P(195.9 < \mu < 226.1) = 0.95$

$$P(195.9 < \bar{x} < 226.1) = P\left(\frac{195.9 - 211}{\frac{46}{\sqrt{25}}} < Z < \frac{226.1 - 211}{\frac{46}{\sqrt{25}}}\right)$$

$$= P(-1.64 < Z < 1.64) = 2P(0 < Z < 1.64) = 0.899$$

Exercise 9.5.1 Explain the difference between point and interval estimation.

Solution: Point estimation and interval estimation are two methods used to estimate a parameter in statistics.

Point estimation is a method of estimating a parameter by calculating a single value that is believed to be the most likely value for that parameter. This value is referred to as a point estimate. Point estimation is used when the goal is to determine the best single estimate for a population parameter based on the sample data. For example, the sample mean is a point estimate of the population mean.

Interval estimation, on the other hand, is a method of estimating a parameter by calculating a range of values that is believed to contain the true value of the parameter with a specified level of confidence. This range of values is referred to as a confidence interval. Interval estimation is used when the goal is to determine a range of values that is likely to contain the true value of the parameter, rather than a single best estimate. The confidence interval takes into account the variability in the sample data, providing a more robust and reliable estimate of the parameter.

In summary, point estimation provides a single value estimate of a population parameter, while interval estimation provides a range of values that is believed to contain the true value of the parameter with a specified level of confidence.

2. **(10 points each)**
Exercise 8.5.14 For the population of adult males in the United States, the distribution of weights is approximately normal with mean μ = 172.2 pounds and standard deviation σ = 29.8 pounds [165].
(a) Describe the distribution of means of samples of size 25 which are drawn from this population.
(b) What is the upper bound for 90% of the mean weights of samples of size 25?
(c) What is the lower bound for 80% of the mean weights?
(d) Suppose that you select a single random sample of size 25 and find that the mean weight for the men in the sample is $\bar{x}$ = 190 pounds. How likely is this result? What would you conclude?

Solution: $\mu = 172.2\, pounds\,;\ \sigma = 29.8\, pounds\,; n = size\ of\ sample = 25$

a. Since population follows $N(\mu, \sigma)$
Sample follows $N(\mu_x, \sigma_x)$
Where $\mu_x = \mu\,;\ \sigma_x = \frac{\sigma}{\sqrt{n}}$
$\therefore Sample\ mean\ , \mu_x = \mu = 172.2\, pounds$
$$Sample\ standard\ deviation = \sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{29.8}{\sqrt{25}} = 5.96\, pounds$$

b. $P(\bar{x} < \overline{x_0}) = 0.9$
$$P\left(Z < \frac{\overline{x_0} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0.9$$

$$\overline{x_0} = \ \mu + \frac{\sigma}{\sqrt{n}} \times invNorm(0.9)$$

$$\approx 172.2 + \frac{29.8}{\sqrt{25}} \times 1.282 = 179.841$$

c. $P(\bar{x} > \ \overline{x_0}) = 0.8$

$P(\bar{x} < \ \overline{x_0}) = 0.2$

$$P\left(Z < \frac{\overline{x_0} - \ \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0.2$$

$$\overline{x_0} = \ \mu + \frac{\sigma}{\sqrt{n}} \times invNorm(0.2)$$

$$\approx 172.2 + \frac{29.8}{\sqrt{25}} \times (-0.842) = 167.182$$

d. $Z = \frac{\bar{x} - \ \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

$$Z = \frac{190 - 172.2}{29.8/\sqrt{25}} \ \approx 2.99$$

It is obtained that “Z” is around 3 standard deviation times away from the mean. This is highly unlikely, as it has probability around 0.001.

Either there is some error in the measurement or the sample is too biased with outliers.

Exercise 9.5.9 9. For the population of infants undergoing fetal surgery for congenital anomalies, the distribution of gestational ages at birth is approximately normal with unknown mean μ and standard deviation σ. A random sample of 14 such infants has mean gestational age overline $\bar{x}$ = 29.6 weeks and standard deviation s = 3.6 weeks [186].

(a) Construct a 95% confidence interval for the true population mean μ.

(b) What is the length of this interval?

(c) How large a sample would be required so that the 95% confidence interval has length 3 weeks? Assume that the population standard deviation σ is known and that σ = 3.6 weeks.

(d) How large a sample would be needed for the 95% confidence interval to have length 2 weeks? Again, assume that σ is known.

Solution: $\bar{x} = \ 29.6\, weeks\, ; s = 3.6\, weeks$

a. $95\%\ CI\ for\ \mu = \left(29.6 - 2.160\ \times \frac{3.6}{\sqrt{14}}\ , 29.6 + 2.160\ \times \frac{3.6}{\sqrt{14}}\right) = (27.5, 31.7)$

b. $length\, of\ CI = 31.7 - 27.5\, weeks = 4.2\, weeks$

c. $(29.6 \pm 1.5) = \left(29.6 \pm 1.96 \times \frac{3.6}{\sqrt{n}}\right)$

$$=> n = \left\lceil \left(\frac{1.96 \times 3.6}{1.5}\right)^2 \right\rceil = \lceil 22.13 \rceil$$

$\therefore n = 23$

d. $n = \left\lceil \left(\frac{1.96 \times 3.6}{1}\right)^2 \right\rceil = \lceil 49.79 \rceil$

$\therefore n = 50$

Exercise 9.5.13 Serum zinc levels for 462 males between the ages of 15 and 17 are saved under the variable name zinc in the data set serum_zinc [75]. The units of measurement for serum zinc level are micrograms per deciliter.

(a) Find a two-sided 95% confidence interval for μ, the true mean serum zinc level for the population of males between the ages of 15 and 17 years.
(b) Interpret this confidence interval.
(c) Calculate a 90% confidence interval for μ.
(d) How does the 90% confidence interval compare to the 95% interval?

Solution: a. The 95% confidence interval for μ is (86.47783, 89.39663)
b. This means that if we repeated this process many times, 95% of the intervals generated would contain the true population mean.
c. The 90% confidence interval for μ is (86.71246, 89.16200)
d. The 95% interval is wider than the 90% interval, which means that it has a higher degree of uncertainty.

R Code:

```
library(haven)
serum_zinc <- read_dta("C:/Users/abhil/OneDrive/Desktop/MSAM-Northeastern/MATH7343/Homeworks/3/serum_zinc.dta")

# Confidence Interval for the mean
mean_zinc <- mean(serum_zinc$zinc)
std_dev <- sd(serum_zinc$zinc) / sqrt(nrow(serum_zinc))

# 95% Confidence Interval
alpha <- 0.05
z_critical <- qnorm(1 - alpha/2)
CI_95 <- mean_zinc + c(-z_critical, z_critical) * std_dev
CI_95

# 90% Confidence Interval
alpha <- 0.1
z_critical <- qnorm(1 - alpha/2)
```

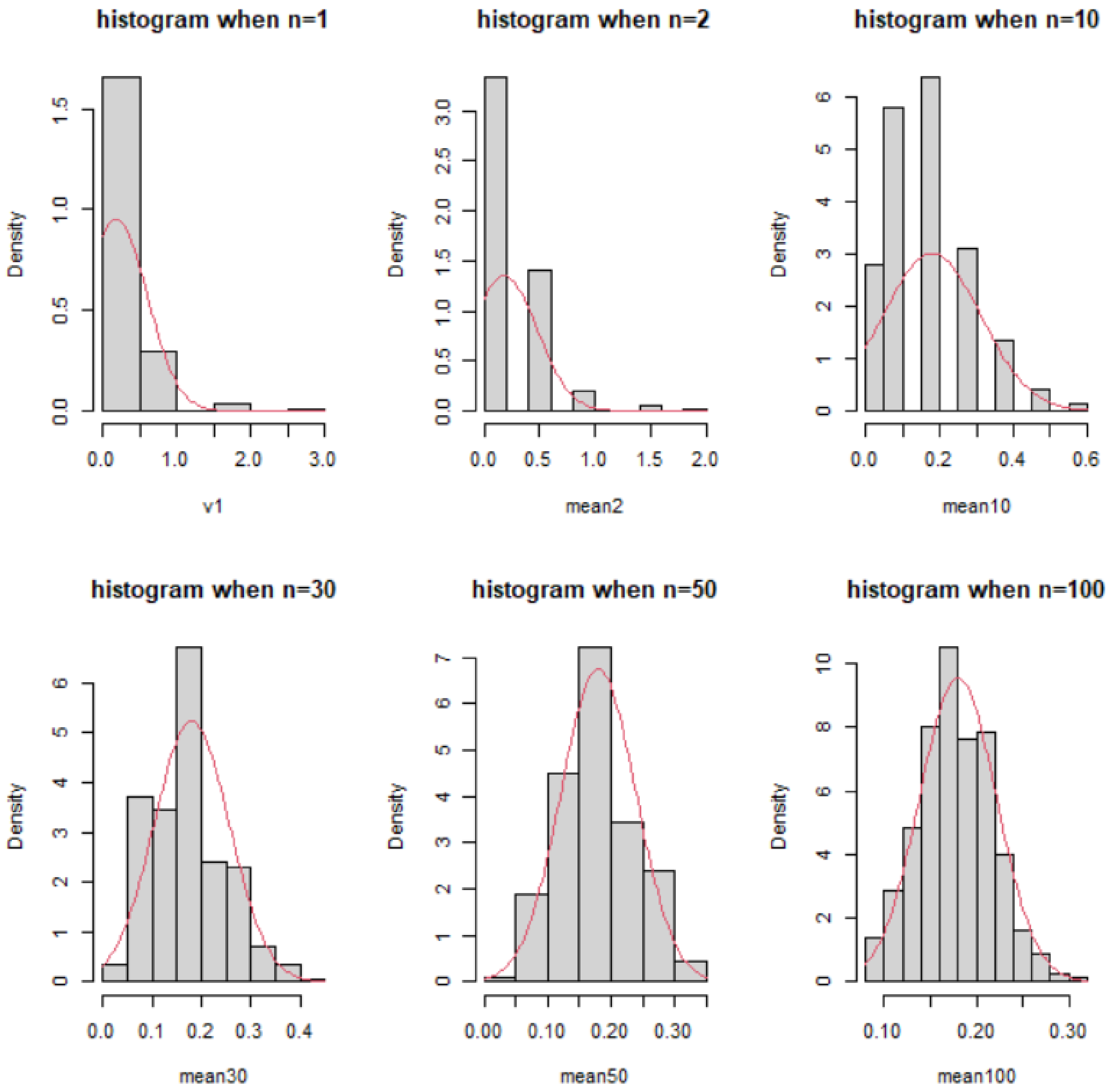
```
CI_90 <- mean_zinc + c(-z_critical, z_critical) * std_dev
CI_90
```

3. **(10 points) Run the simulation program for CLT in the following page. Try several types of random variables: Binomial (6,0.03), Poisson (1) and Uniform (0,5). At what sample size does the distribution of means in your simulation come close to normal distribution?**

   Solution:

   Binomial Distribution

```
n <- 100
n.obs <- 400
rand.data <- matrix(rbinom(n.obs*n, size=6, prob=0.03),
nrow=n.obs)
v1 <- rand.data[,1] #The first column (variable)
mean2 <- apply(rand.data[,1:2], FUN=mean, MARGIN=1)
mean10<- apply(rand.data[,1:10], FUN=mean, MARGIN=1)
mean30 <- apply(rand.data[,1:30], FUN=mean, MARGIN=1)
mean50 <- apply(rand.data[,1:50], FUN=mean, MARGIN=1)
mean100 <- apply(rand.data[,1:100], FUN=mean, MARGIN=1)
mu <- 6*0.03 #Binomial mean
sigma <- sqrt(6*0.03*(1-0.03)) #Binomial standard deviation
par(mfrow=c(2,3))
hist(v1, freq = FALSE, main="histogram when n=1")
curve(dnorm(x, mean=mu, sd=sigma),col=2,add=T)
hist(mean2, freq = FALSE, main="histogram when n=2")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(2)),col=2,add=T)
hist(mean10, freq = FALSE, main="histogram when n=10")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(10)),col=2,add=T)
hist(mean30, freq = FALSE, main="histogram when n=30")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(30)),col=2,add=T)
hist(mean50, freq = FALSE, main="histogram when n=50")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(50)),col=2,add=T)
hist(mean100, freq = FALSE, main="histogram when n=100")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(100)),col=2,add=T)
```

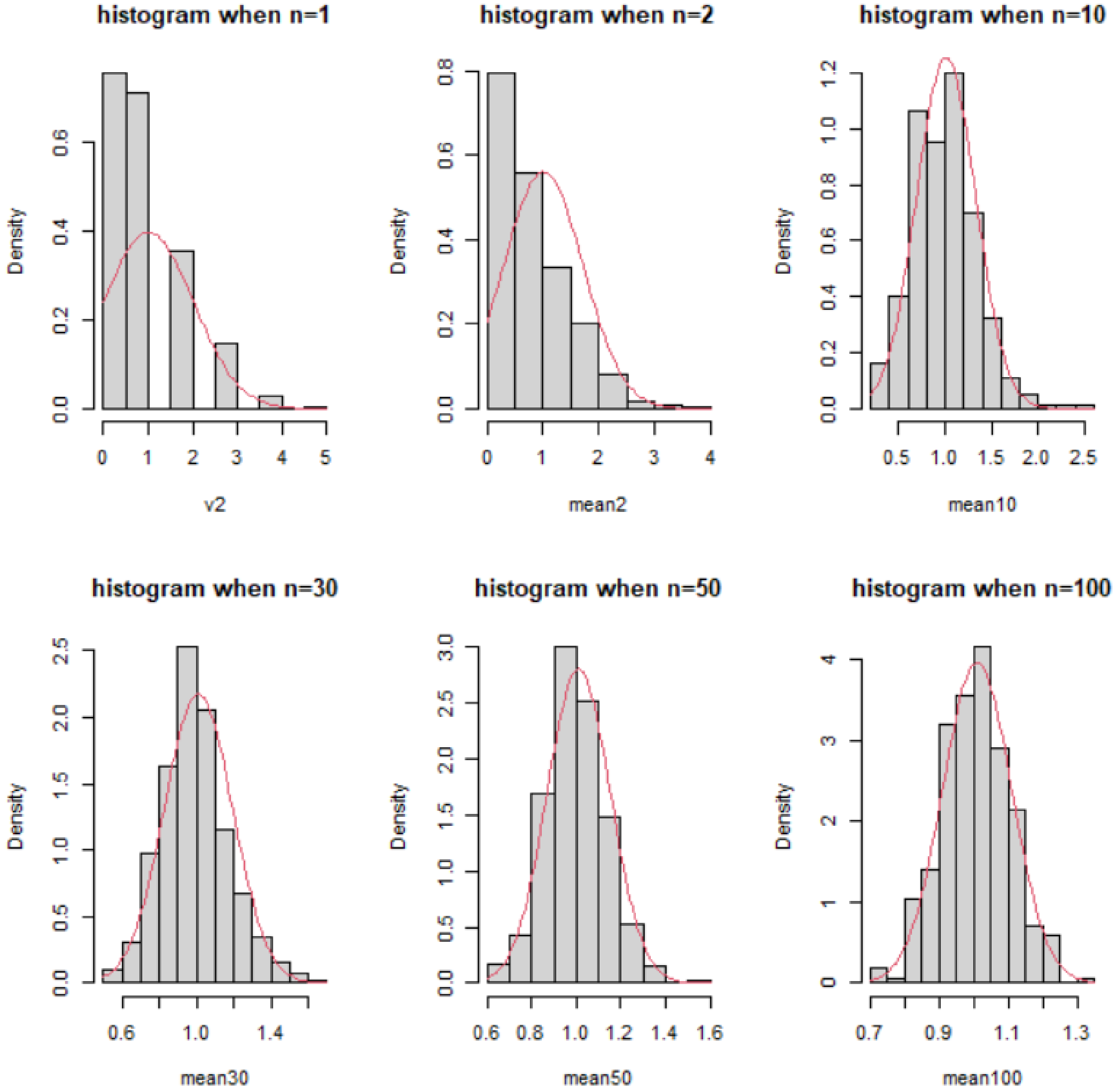


At $n \geq 50$, the distribution of means is close to normal distribution.

Poisson Distribution

```
n <- 100
n.obs <- 400
rand.data2 <- matrix(rpois(n.obs*n, 1), nrow=n.obs)
v2 <- rand.data2[,1] #The first column (variable)
mean2 <- apply(rand.data2[,1:2], FUN=mean, MARGIN=1)
mean10<- apply(rand.data2[,1:10], FUN=mean, MARGIN=1)
mean30 <- apply(rand.data2[,1:30], FUN=mean, MARGIN=1)
mean50 <- apply(rand.data2[,1:50], FUN=mean, MARGIN=1)
mean100 <- apply(rand.data2[,1:100], FUN=mean, MARGIN=1)
mu <- mean(rand.data2) #Poisson mean
sigma <- sd(rand.data2) #Poisson standard deviation
```

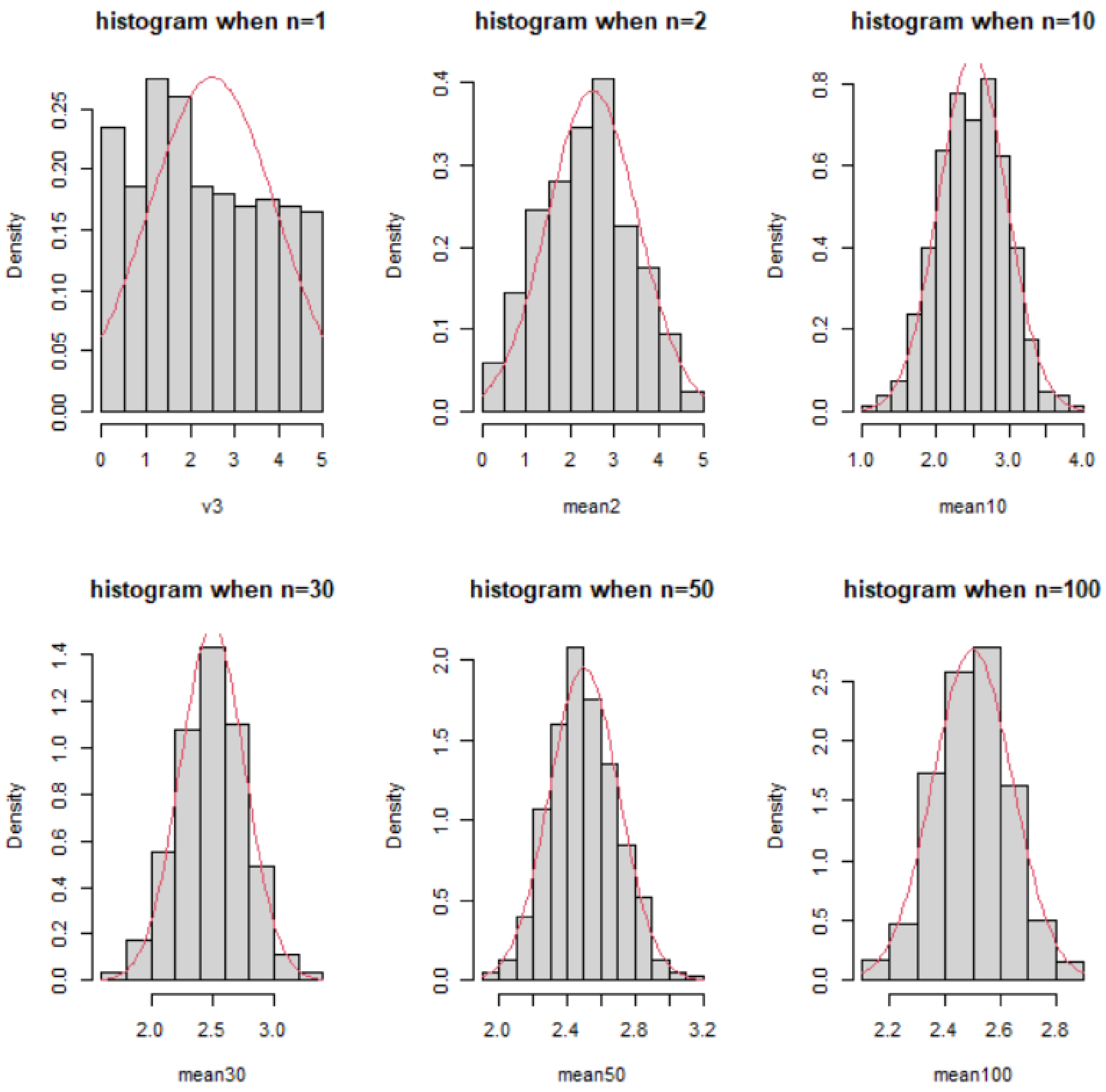
```
par(mfrow=c(2,3))
hist(v2, freq = FALSE, main="histogram when n=1")
curve(dnorm(x, mean=mu, sd=sigma),col=2,add=T)
hist(mean2, freq = FALSE, main="histogram when n=2")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(2)),col=2,add=T)
hist(mean10, freq = FALSE, main="histogram when n=10")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(10)),col=2,add=T)
hist(mean30, freq = FALSE, main="histogram when n=30")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(30)),col=2,add=T)
hist(mean50, freq = FALSE, main="histogram when n=50")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(50)),col=2,add=T)
hist(mean100, freq = FALSE, main="histogram when n=100")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(100)),col=2,add=T)
```



At $n \geq 10$, the distribution of means is close to normal distribution.

Uniform Distribution

```
n <- 100
n.obs <- 400
rand.data3 <- matrix(runif(n.obs*n, min = 0, max = 5),
nrow=n.obs)
v3 <- rand.data3[,1] #The first column (variable)
mean2 <- apply(rand.data3[,1:2], FUN=mean, MARGIN=1)
mean10<- apply(rand.data3[,1:10], FUN=mean, MARGIN=1)
mean30 <- apply(rand.data3[,1:30], FUN=mean, MARGIN=1)
mean50 <- apply(rand.data3[,1:50], FUN=mean, MARGIN=1)
mean100 <- apply(rand.data3[,1:100], FUN=mean, MARGIN=1)
mu <- mean(rand.data3) #Poisson mean
sigma <- sd(rand.data3) #Poisson standard deviation
par(mfrow=c(2,3))
hist(v3, freq = FALSE, main="histogram when n=1")
curve(dnorm(x, mean=mu, sd=sigma),col=2,add=T)
hist(mean2, freq = FALSE, main="histogram when n=2")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(2)),col=2,add=T)
hist(mean10, freq = FALSE, main="histogram when n=10")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(10)),col=2,add=T)
hist(mean30, freq = FALSE, main="histogram when n=30")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(30)),col=2,add=T)
hist(mean50, freq = FALSE, main="histogram when n=50")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(50)),col=2,add=T)
hist(mean100, freq = FALSE, main="histogram when n=100")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(100)),col=2,add=T)
```



At $n \geq 2$, the distribution of means is close to normal distribution.

4. **(5 points) Generate independent random variable X and Y, respectively, from Poisson(3) and a F-distribution with degrees of freedoms 5 and 3. Use nobs=8000 simulated observations of each variable to approximate the mean E(X^3 /Y). Submit the R codes with your answer.**

Solution:

```
> nobs <- 8000
> lambda <- 3
> df1 <- 5
> df2 <- 3
> X <- rpois(nobs, lambda)
> Y <- rf(nobs, df1, df2)
```

```
>
> mean(X^3 / Y)
[1] 101.7588
```