MATH 7343: Applied Statistics

Homework -3

Sai Nikhil
NUID: 001564864
T. Sai Nikhil

1

8.5.4

Standard error of Sample mean:

From Section 8.2 (Textbook)

Let, the distribution of a population has mean (μ) and Standard deviation (σ). Let us take samples of size "n" from the population. The mean for one such Sample is called a sample mean. The Standard deviation of distribution of Sample means is called standard error of the Sample mean.

Relation between Standard error and Standard deviation of population $= \frac{\sigma}{\sqrt{n}}$

8.5.15

mean of population → $\mu = 211$ mg/100ml

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

$n = 25$ ← size of population

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

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

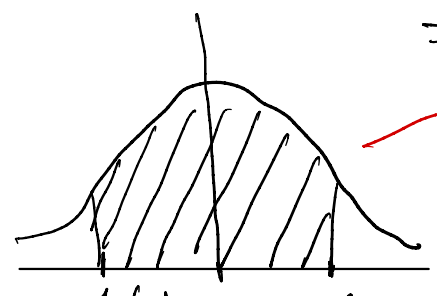
$$P(195.9 < \mu < 226.1) = 0.95$$

$$P(195.9 < \bar{x} < 226.1)$$

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

$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$= P(-1.64 < Z < +1.64)$$



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

$$\approx \boxed{0.899}$$

9.5.1

Point estimation

1) It involves using sample data to calculate a single number to estimate the parameter of interest. For eg. we might use sample mean $\bar{x}$ to estimate the population mean μ.

2) It is easy to compute but in general it is not a good metric as different samples can produce different sample means and there is some degree of uncertainity.

Interval estimation

1) It involves range of reasonable values that are intended to contain the parameter of interest, for eg. the population mean μ.

2) Because we are are evaluating parameter in a range of values we can report it with a certain degree of confidence. The range of values is called a confidence interval.

(2)

8.5.14

$\mu = 172.2$ pounds, $\sigma = 29.8$ pounds

(a) Since population follows $N(\mu, \sigma)$,

Sample follows $N(\mu_{\bar{x}}, \sigma_{\bar{x}})$,

where $\mu_{\bar{x}} = \mu$

$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ ← Size of sample

$\therefore$ Sample mean $(\mu_{\bar{x}}) = \mu =$ **172.2 pounds**

Sample standard deviation $(\sigma_{\bar{x}}) = \dfrac{29.8}{\sqrt{25}}$ pounds

$=$ **5.96 pounds**

(b) upper bound

$$P(\bar{x} < \bar{x}_0) = 0.9$$

$$P\left(Z < \frac{\bar{x}_0 - \mu}{\sigma/\sqrt{n}}\right) = 0.9$$

$$\bar{x}_0 = \mu + \frac{\sigma}{\sqrt{n}} \times \text{invNorm}(0.9)$$

$$\simeq 172.2 + \frac{29.8}{\sqrt{25}} \times 1.282$$

$$\simeq \boxed{179.841}$$

(c)

$$P(\bar{X} > \bar{x}_0) = 0.8$$

$$P(\bar{X} < \bar{x}_0) = 0.2$$

$$P\left(Z < \frac{\bar{x}_0 - \mu}{\sigma/\sqrt{n}}\right) = 0.2$$

$$\Rightarrow \bar{x}_0 = \mu + \frac{\sigma}{\sqrt{n}} \cdot \text{invNorm}(0.2)$$

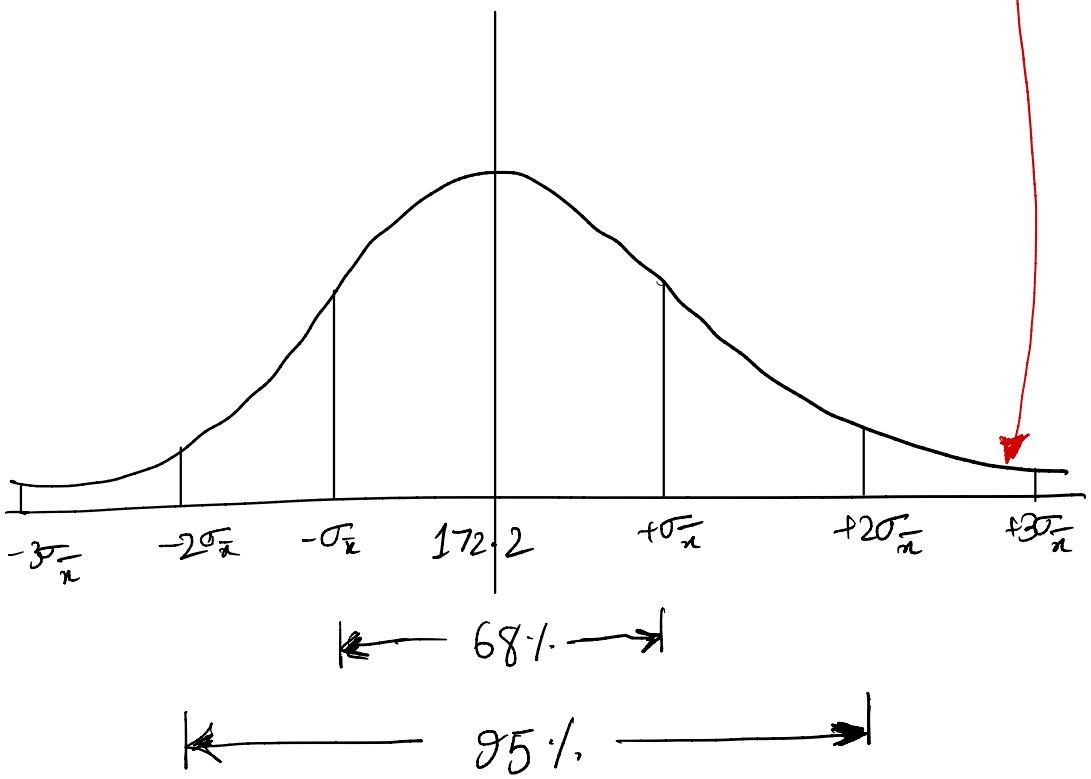
$$\simeq 172.2 + \frac{29.8}{\sqrt{25}} \cdot (-0.842)$$

$$\simeq \boxed{167.182}$$

(d)

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

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



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.

9.5.8

$n = 12$

$\bar{x}_1 = 4.49$ litres ; $S_1 = 0.83$ litres

$\bar{x}_2 = 3.11$ litres ; $S_2 = 0.62$ litres

(a) $\alpha = 0.05$

95% confidence interval for μ_1 (FVC) is

$$\left(\bar{x}_1 - t_{\alpha/2, n-1} \times \frac{S}{\sqrt{n}} , \bar{x}_1 + t_{\alpha/2, n-1} \times \frac{S}{\sqrt{n}} \right)$$

$t_{0.025, 11} = 2.201$

$$\therefore 95\%\ CI = \left(4.49 - 2.201 \times \frac{0.83}{\sqrt{12}} , 4.49 + 2.201 \times \frac{0.83}{\sqrt{12}} \right)$$

$$= \boxed{(3.96, 5.02)}$$

(b) $t_{0.05, 11} = 1.796$

$$\therefore 90\%\ CI = \left(4.49 - 1.796 \times \frac{0.83}{\sqrt{12}} , 4.49 + 1.796 \times \frac{0.83}{\sqrt{12}} \right)$$

$$= \boxed{(4.06, 4.92)}$$

95% CI interval length = 1.06

90% CI interval length = 0.86

The interval length is decreased

% decrease in length $= \left(1 - \frac{0.86}{1.06}\right) \times 100$

$\simeq \boxed{18\%}$

(c)

95% CI for μ_2 (FEV_1)

$= \left(3.71 - 2.201 \times \frac{0.62}{\sqrt{12}}, 3.71 + 2.201 \times \frac{0.62}{\sqrt{12}}\right)$

$= \boxed{(3.32, 4.10)}$

(d) When constructing confidence intervals, it is assumed that the underlying distributions for FVC and FEV_1 are approximately normal.

9.5.9

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

$= \boxed{(27.5, 31.7)}$

(b) length of CI $= 31.7 - 27.5$ weeks

$= \boxed{4.2 \text{ weeks}}$

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

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

$\therefore \boxed{n = 23}$

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

$\boxed{n = 50}$

9.5.13

Input:

```
library(magrittr)
library(dplyr)

lowbwt <- read.csv(file = './lowbwt.txt')

lowbwt %>%
  group_by(sex) %>%
  summarise(mu = mean(sbp), sigma = sd(sbp))

males_sbp <- subset(lowbwt, sex == 1)$sbp
females_sbp <- subset(lowbwt, sex == 0)$sbp

t.test(males_sbp)
t.test(females_sbp)
```

```
> library(magrittr)
> library(dplyr)
>
> lowbwt <- read.csv(file = './lowbwt.txt')
>
> lowbwt %>%
+   group_by(sex) %>%
+   summarise(mu = mean(sbp), sigma = sd(sbp))
# A tibble: 2 x 3
    sex    mu sigma
* <int> <dbl> <dbl>
1     0  46.5  11.1
2     1  47.9  11.8
>
> males_sbp <- subset(lowbwt, sex == 1)$sbp
> females_sbp <- subset(lowbwt, sex == 0)$sbp
>
> t.test(males_sbp)

   One Sample t-test

data:  males_sbp
t = 26.893, df = 43, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 44.27435 51.45292
sample estimates:
mean of x
 47.86364

> t.test(females_sbp)

   One Sample t-test

data:  females_sbp
t = 31.198, df = 55, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 43.47956 49.44901
sample estimates:
mean of x
 46.46429
```

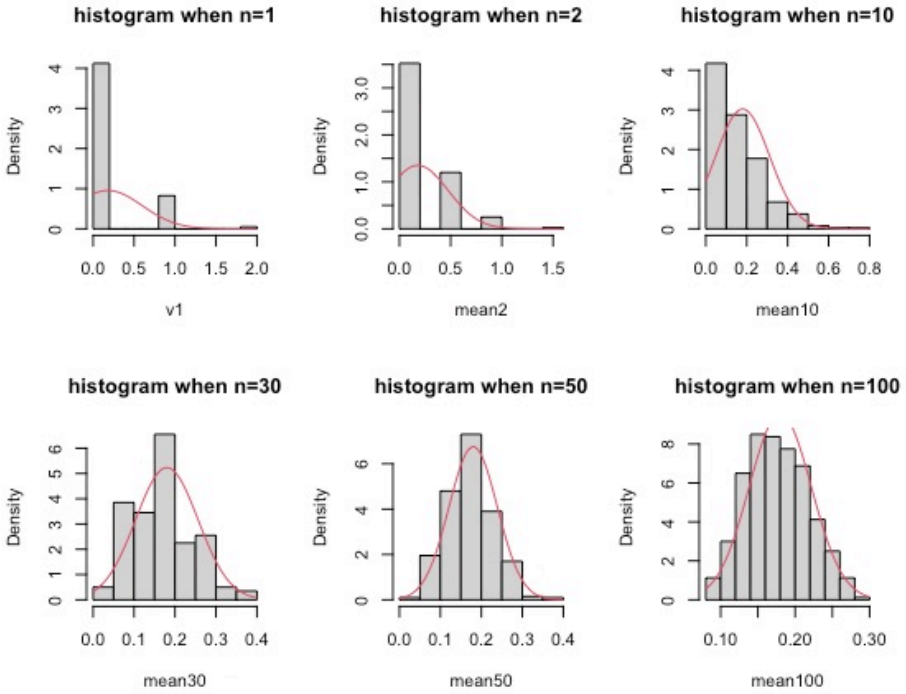
part (a)

part (b)

(c) It is unlikely that male and female population have same mean. As the sample size is already very high it is almost a true representative of population and $\bar{x}_m \neq \bar{x}_f$ (for sample) $\Rightarrow$ $\mu_m \neq \mu_f$ (for population)

3 Binomial (6, 0.03)

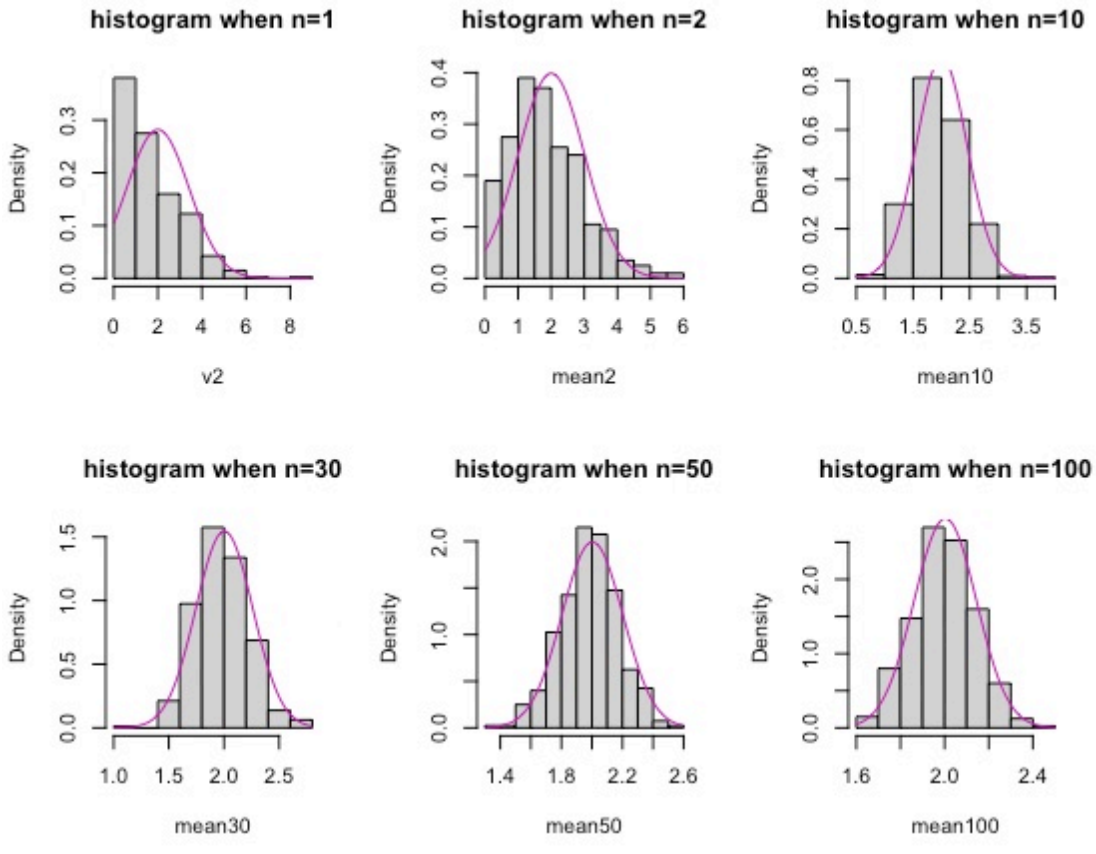
```
## Generate n=100 random variables from a Binomial(6,0.03),distribution
## Have n.obs=400 observations of each variable.
n <- 100
n.obs <-400
rand.data <- matrix(rbinom(n.obs*n, size=6,prob=0.03), nrow=n.obs)
## Create summary variables;
# v1--the first variable, mean2 -- average of the first twovariables
# mean10 -- average of the first 10 variables, similarly mean30,etc.
v1 <- rand.data[,1] #The first column (variable)
#Average of the first two columns(variables),
# apply function 'mean' on margin 1 (row-wise function application)
mean2 <- apply(rand.data[,1:2], FUN=mean, MARGIN=1)
#Average of the first ten columns(variables)
mean10<- apply(rand.data[,1:10], FUN=mean, MARGIN=1)
#Average of the first 30, 50, 100 columns (variables)
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)
##Plot the histograms of the summary variables above, overlay with
# the density plot of the normal distribution in the CLT.
mu <- 6*0.03 #Binomial mean
sigma <- sqrt(6*0.03*(1-0.03)) #Binomial standard deviation
## Arrange 6 plots in one page 2 by 3
par(mfrow=c(2,3))
# Histogram of v1,
# then add the normal density curve in red color (col=2)
hist(v1, freq = FALSE, main="histogram when n=1")
curve(dnorm(x, mean=mu, sd=sigma),col=2,add=T)
# Histograms of mean2, mean10, ....
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 $n \geq 100$, the distribution of means is close to normal distribution.

Poisson (2)

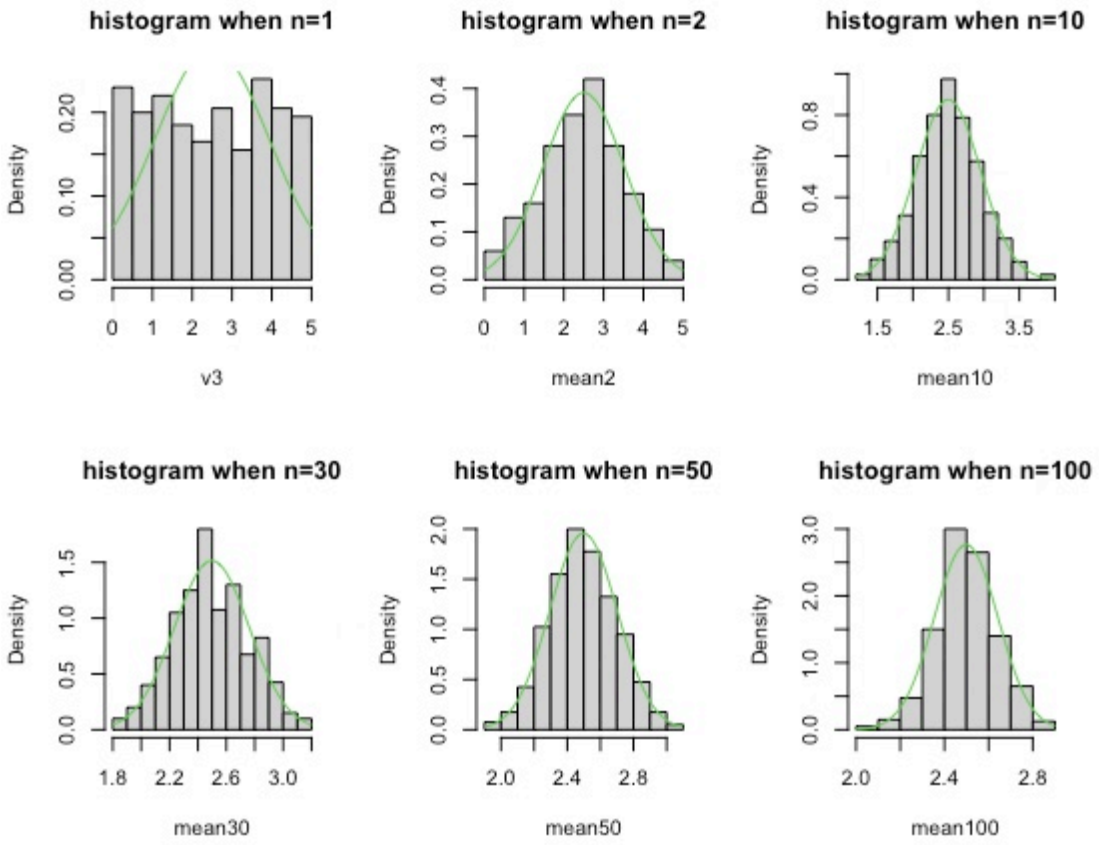
```
##Generate n=100 random variables from a Poisson(2) distribution
## Have n.obs=400 observations of each variable
n <- 100
n.obs <- 400
rand.data2 <- matrix(rpois(n.obs*n,2), nrow=n.obs)
v2 <- rand.data2[,1]
mean2 <- apply(rand.data2[,1:2], FUN=mean, MARGIN=1)
mean10<- apply(rand.data2[,1:10], FUN=mean, 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) #Poission standard deviation
## Arrange 6 plots in one page 2 by 3
par(mfrow=c(2,3))
# Histogram of v2,
# then add the normal density curve in red color (col=6)
hist(v2, freq = FALSE, main="histogram when n=1")
curve(dnorm(x, mean=mu, sd=sigma),col=6,add=T)
# Histograms of mean2, mean10, ....
hist(mean2, freq = FALSE, main="histogram when n=2")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(2)),col=6,add=T)
hist(mean10, freq = FALSE, main="histogram when n=10")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(10)),col=6,add=T)
hist(mean30, freq = FALSE, main="histogram when n=30")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(30)),col=6,add=T)
hist(mean50, freq = FALSE, main="histogram when n=50")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(50)),col=6,add=T)
hist(mean100, freq = FALSE, main="histogram when n=100")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(100)),col=6,add=T)
```



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

Uniform(0,5)

```
##Generate n=100 random variables from a Uniform(0,5)distribution
## Have n.obs=450 observations of each variable
n <- 100
n.obs <-400
rand.data3 <- matrix(runif(n.obs*n,min=0, max=5), nrow=n.obs)
v3 <- rand.data3[,1]
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) #Uniform mean
sigma <-sd(rand.data3) #Uniform standard deviation
## Arrange 6 plots in one page 2 by 3
par(mfrow=c(2,3))
# Histogram of v3,
# then add the normal density curve in red color (col=4)
hist(v3, freq = FALSE, main="histogram when n=1")
curve(dnorm(x, mean=mu, sd=sigma),col=3,add=T)
# Histograms of mean2, mean20, ....
hist(mean2, freq = FALSE, main="histogram when n=2")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(2)),col=3,add=T)
hist(mean10, freq = FALSE, main="histogram when n=10")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(10)),col=3,add=T)
hist(mean30, freq = FALSE, main="histogram when n=30")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(30)),col=3,add=T)
hist(mean50, freq = FALSE, main="histogram when n=50")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(50)),col=3,add=T)
hist(mean100, freq = FALSE, main="histogram when n=100")
curve(dnorm(x, mean=mu, sd=sigma/sqrt(100)),col=3,add=T)
```



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

4

```
> nobs <- 4000
> X <- rnorm(nobs)
> Y <- rf(nobs, 4, 9)
>
> print(mean(X^2 / Y))
[1] 1.818311
```