STAT 408 Homework 1 Solution

1.

a. $E(X) = \sum_{\text{all } x} xp(x) = 1(.05) + 2(.10) + 4(.35) + 8(.40) + 16(.10) = 6.45$ GB.

b. $V(X) = \sum_{\text{all } x} (x-\mu)^2 p(x) = (1-6.45)^2(.05) + (2-6.45)^2(.10) + \ldots + (16-6.45)^2(.10) = 15.6475.$

c. $\sigma = \sqrt{V(X)} = \sqrt{15.6475} = 3.956$ GB.

2.

a. We use the sample mean, $\bar{x} = 1.3481.$

b. We will use sample variance to estimate the population variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 0.1146$$

3.

a. $58.3 \pm \frac{1.96(3)}{\sqrt{25}} = 58.3 \pm 1.18 = (57.1, 59.5).$

b. $58.3 \pm \frac{1.96(3)}{\sqrt{100}} = 58.3 \pm .59 = (57.7, 58.9).$

c. $58.3 \pm \frac{2.58(3)}{\sqrt{25}} = 58.3 \pm 1.548 = (56.752, 59.848)$

4.

a.

Let μ be the mean strength of welds in the population

H_0: μ = 100 versus H_a: μ > 100.

b.

Type I error: The population mean is 100 but we reject it.

Type II error: The population mean is not 100 but we fail to reject it.

5.

a.

```
setwd("C:/Users/mxi1/Dropbox/Loyola/STAT 437")
NCbirths <- read.csv('births.csv')
```

b.

```
weights <- NCbirths$weight
```

The unit of weights vectors is ounce

c.

```
weights_in_pounds <- weights * 0.0625
```

d.

```
weights_in_pounds[1:20]
```

e.

```
mean(weights_in_pounds)
```

The mean weights of all babies is 7.25 pounds

f.

```
table(NCbirths$Habit)/dim(NCbirths)[1]
```

About 9.4% mothers smoke in the dataset.

g.

```
0.14 - 0.0938755
```

The percentage of smoking mother is 4.6% lower than the national average of adult smokers.

6.

a.

```
setwd("C:/Users/mxi1/Dropbox/Loyola/STAT 437")

flint <- read.csv('flint.csv')
```

b.

```
mean(flint$Pb >= 15)
```

4.4% of the locations tested were found to have dangerous lead levels

c.

```
mean(flint$Cu[flint$Region=='North'])
```

The mean copper level for only test sites in the North region is 44.64.

d.

```
mean(flint$Cu[flint$Pb >= 15])
```

The mean copper level for only test sites with dangerous lead levels is 305.83.

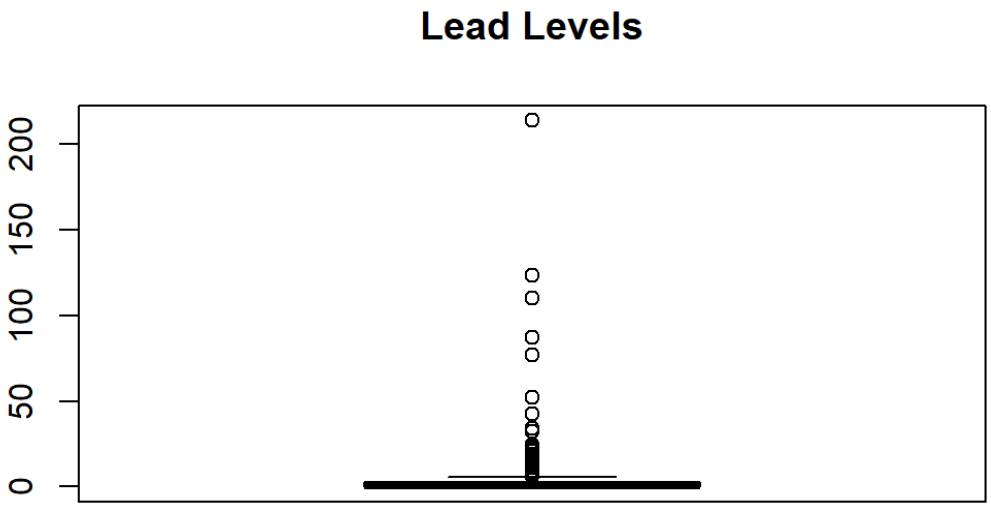
e.

```
mean(flint$Pb)

mean(flint$Cu)
```

The mean lead and copper levels for all locations are 3.38 and 54.58.

f.

```
boxplot(flint$Pb, main="Lead Levels")
```



g.

median(flint$Pb)

summary(flint$Pb)

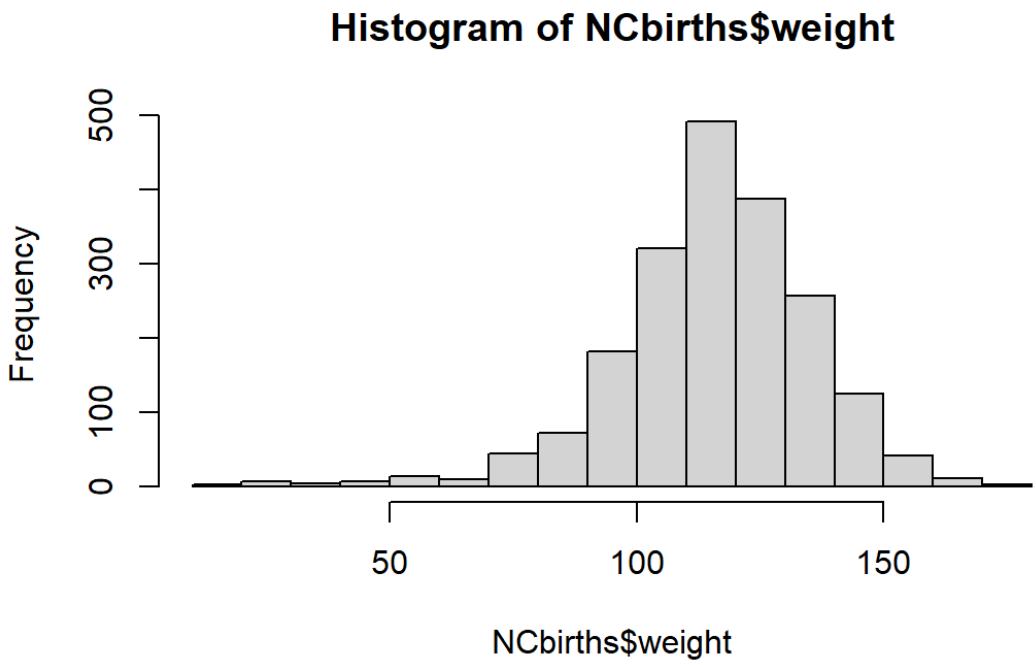
The mean is not a good measure of center for this dataset because the outliers would largely affect the value of mean. Instead, we can use median because it is resistant to outliers.

7.

a.

set.seed(2022)

hist(NCbirths$weight)



Weight is not a normal distribution because it has a long left tails, which is left-skewed.

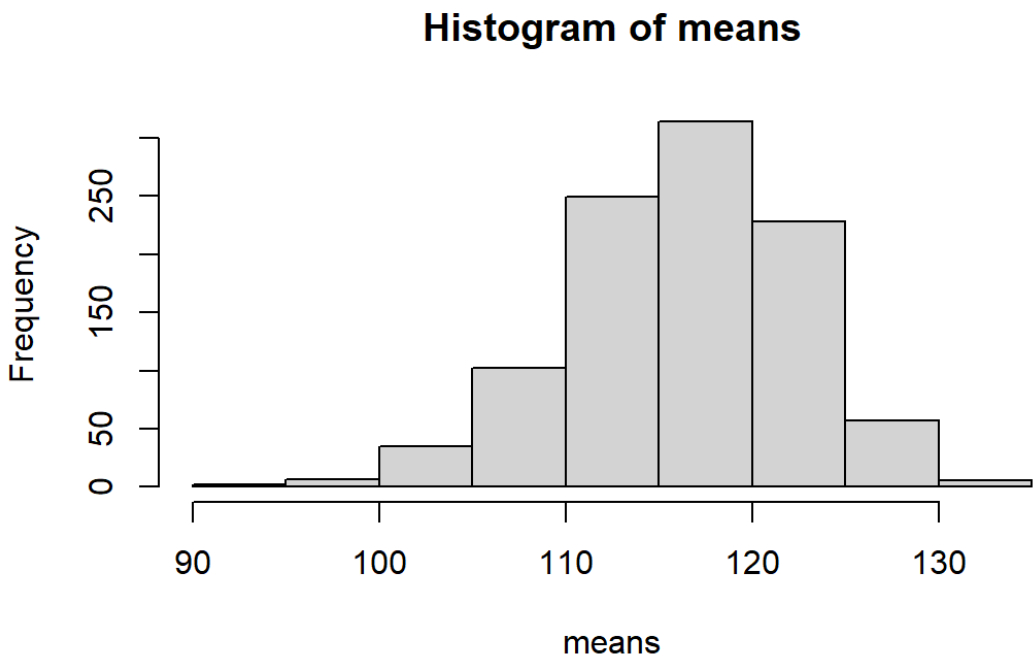
b.

```
index <- sample(1992, size = 10)

mean <- NCbirths$weight[index]

mean

136 113 155 113 106  95 124 120 115 113
```

c.

```
means <- c()

for(i in 1:1000){

  index <- sample(1992, size = 10)

  mean <- mean(NCbirths$weight[index])

  means[i] <- mean

}
```

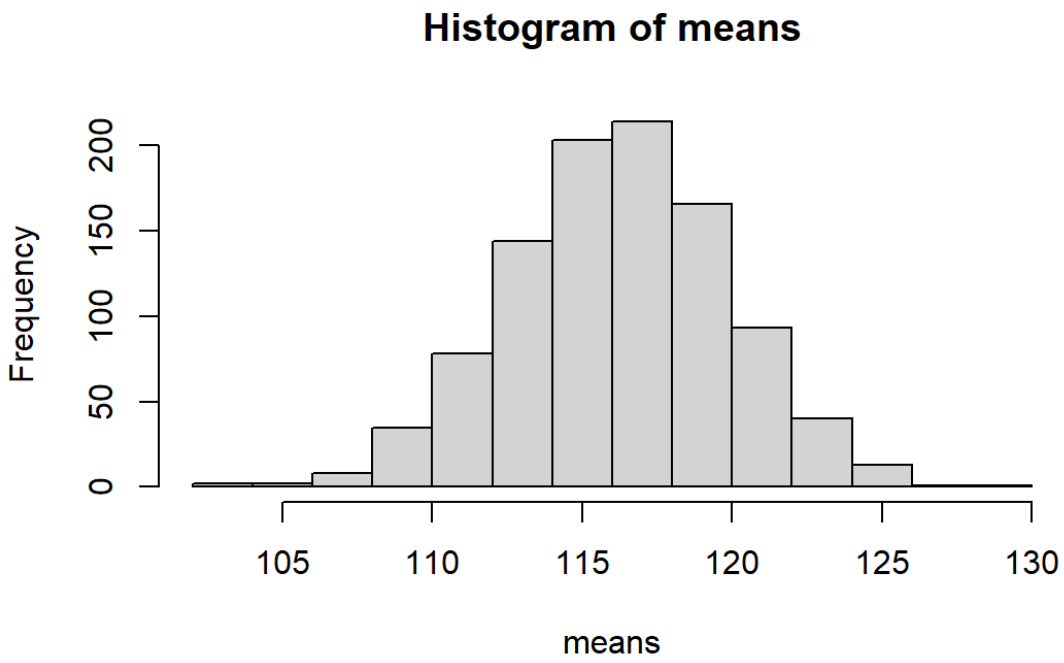
hist(means)



Now the distribution is closer to normal but still left-skewed.

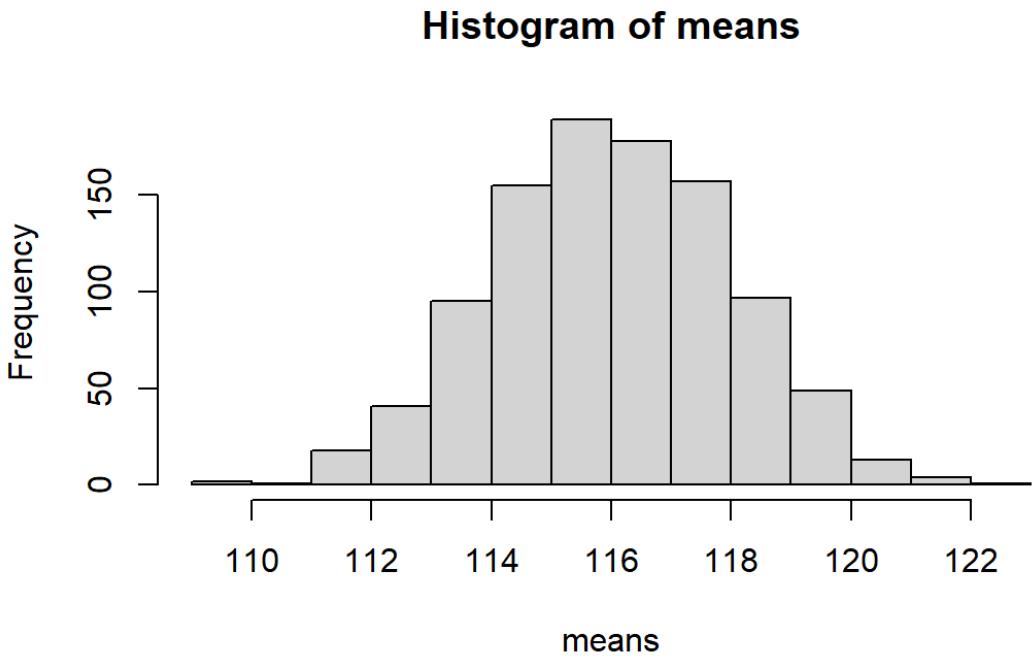
d.

```
means <- c()
for(i in 1:1000){
  index <- sample(1992, size = 30)
  mean <- mean(NCbirths$weight[index])
  means[i] <- mean
}
hist(means)
```



Sample size is 30 makes the distribution much closer to normal.

```
for(i in 1:1000){
  index <- sample(1992, size = 100)
  mean <- mean(NCbirths$weight[index])
  means[i] <- mean
}
hist(means)
```



Sample size is 100 makes the distribution almost same to normal.

Based on central limit theorem, when sample size is greater than 30, the distribution of sample mean can be approximated by a normal distribution.