Homework 1

Rolando Santos

2023-09-08

Question 1

The pmf of the amount of memory X (GB) in a purchased flash drive is.

x	1.00	2.0	4.00	8.0	16.0
p(x)	0.05	0.1	0.35	0.4	0.1

a. Compute E(X)

$$E(X) = \sum_{x \in D} x * p(x)$$

$$E(X) = 1 * 0.05 + 2 * 0.1 + 4 * 0.35 + 8 * 0.4 + 16 * 0.1$$

$$E(X) = 6.45$$

b. Compute V(X)

$$V(X) = \sum_{D} (x - \mu)^2 * p(x)$$

$$V(X) = (1 - 6.45)^2 * 0.05 + (2 - 6.45)^2 * 0.1 + (4 - 6.45)^2 * 0.35 + (8 - 6.45)^2 * 0.4 + (16 - 6.45)^2 * 0.1$$

$$V(X) = 15.6475$$

c. The standard deviation of X

$$SD(X) = \sqrt{V(X)}$$

$$SD(X) = \sqrt{15.6475}$$

$$SD(X) = 3.955692$$

Question 2

Consider the following sample of observations on coating thickness for low-viscosity paint:

0.83	0.88	0.88	1.04	1.09	1.12	1.29	1.31
1.48	1.49	1.59	1.62	1.65	1.71	1.76	1.83

a. Calculate a point estimate of the mean value of coating thickness, and state which estimator you used.

To find the point estimate of the mean, we use the sample mean.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{X} = \frac{1}{16} * (.83 + .88 + .88 + 1.04 + 1.09 + 1.12 + 1.29 + 1.31 + 1.48 + 1.49 + 1.59 + 1.62 + 1.65 + 1.71 + 1.76 + 1.83)$$

$$\bar{X} = 1.348125$$

The sample mean is 1.348125.

b. Calculate a point estimate of the variance of coating thickness, and state which estimator you used.

To find the point estimate of the variance, we use the sample variance.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$$

$$\begin{aligned} S^2 = \frac{1}{16-1} * ((.83 - 1.348125)^2 + (.88 - 1.348125)^2 + \\ (.88 - 1.348125)^2 + (1.04 - 1.348125)^2 + (1.09 - 1.348125)^2 \\ + (1.12 - 1.348125)^2 + (1.29 - 1.348125)^2 + (1.31 - 1.348125)^2 + \\ (1.48 - 1.348125)^2 + (1.49 - 1.348125)^2 + (1.59 - 1.348125)^2 + \\ (1.62 - 1.348125)^2 + (1.65 - 1.348125)^2 + (1.71 - 1.348125)^2 + \\ (1.76 - 1.348125)^2 + (1.83 - 1.348125)^2) \end{aligned}$$

$$S^2 = 0.1146029$$

The sample variance is 0.1146029.

Question 3

A confidence interval is desired for the true average stray-load loss μ(watts) for a certain type of induction motor. Assume that stray-load loss is normally distributed with $\sigma = 3$

a. Compute a 95% CI for μ when n = 25 and $\bar{x}$ = 58.3

$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = .95$$
$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = .95$$
$$(58.3 - 1.96\frac{3}{\sqrt{25}}, 58.3 + 1.96\frac{3}{\sqrt{25}})$$

$$(57.124, 59.476)$$

The 95% CI is (57.124, 59.476).

b. Compute a 95% CI for μ when n = 100 and $\bar{x}$ = 58.3

$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = .95$$
$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = .95$$
$$(58.3 - 1.96\frac{3}{\sqrt{100}}, 58.3 + 1.96\frac{3}{\sqrt{100}})$$
$$(57.712, 58.888)$$

The 95% CI is (57.712, 58.888).

c. Compute a 99% CI for μ when n = 25 and $\bar{x}$ = 58.3

$$P(-2.58 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 2.58) = .99$$
$$P(\bar{X} - 2.58\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2.58\frac{\sigma}{\sqrt{n}}) = .95$$
$$(58.3 - 2.58\frac{3}{\sqrt{25}}, 58.3 + 2.58\frac{3}{\sqrt{25}})$$

$$(56.752, 59.848)$$

The 99% CI is (56.752, 59.848).

Question 4

To determine whether the pipe welds in a nuclear power plant meet specifications, a random sample of welds is selected, and tests are conducted on each weld in the sample. Suppose the specifications state that the mean strength of welds should exceed 100 lb/in^2

a. What hypotheses should be tested? Write down H_0 and H_a and explain your reason.

The test statistics for this are: $H_0 : \mu =< 100$ $H_a : \mu > 100$

The null hypothesis states that the weld strength does not meet the specifications, which in this case is that the mean strength of the welds are less than or equal to 0.

The alternate hypothesis states the opposite of the null hypthesis, which in our case is that the mean strength of welds should exceed 100.

b. Describe type I and II errors in the context of this problem situation.

The type 1 error here is that we reject H_0 when it is true, in this case is that we reject that the $mu <= 100$ even though it is, and we use these welds despite not meeting the specifications.

The type 2 error here is that we fail to reject H_0 when it is false, in this case is that we fail to reject that $mu <= 100$ even though it is not. In this case the welds do meet the specifications but are not used.

Question 5

a. Download the data set births.csv from Sakai, set your working directory, and import it into RStudio. Name the data frame as NCbirths

```
NCBirths <- read.csv("births.csv")
```

b. Extract the weight variable as a vector from the data frame and name it as weights. What units do you think the weights are in?

```
weights <- NCBirths$weight
weights[1:5] # The weight seems to be in ounces
```

```
## [1] 124 177 107 144 117
```

The weights appear to high to be pounds or kilograms, and too small to be in grams. The unit that makes the most sense is ounces.

c. Create a new vector named weights_in_pounds which are the weights of the babies in pounds. You can look up conversion factors on the internet.

```
weights_in_pounds <- weights / 16
```

d. Print the first 20 babies' weight in pounds.

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

```
##  [1]  7.7500 11.0625  6.6875  9.0000  7.3125  6.1250  9.1875  8.6250  6.5000
## [10]  7.6875  9.5625  8.0625  7.4375  6.7500  6.6250  7.8125  7.1875  8.0000
## [19]  8.2500  5.1875
```

e. What is the mean weight of all babies in pounds?

```
mean(weights_in_pounds)
```

```
## [1] 7.2532
```

f. The habit variable records the smoking status for mothers of each baby. What percentage of the mothers in the sample smoke? Hint: consider table() function.

```
table(NCBirths$Habit)
```

```
##
## NonSmoker    Smoker
##      1805       187
```

```
(187/(1805 + 187)) * 100
```

```
## [1] 9.38755
```

From the data set it appears that ~9.4% of mothers are smokers.

g. According to the Centers for Disease Control, approximately 14% of adult Americans are smokers. How far off is the percentage you found in (f) from the CDC's report?

```
14 - 9.4
```

```
## [1] 4.6
```

The data set is approximately 4.6% off from the the CDC estimate of 14%.

Question 6

a. Download the flint.csv from Sakai and read it into R. When you read in the data, name your object "flint"

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

b. The EPA states a water source is especially dangerous if the lead level (Pb) is 15 PPB or greater. What proportion of the locations tested were found to have dangerous lead levels?

```
flint <- flint %>% mutate(isDangerous = Pb >= 15)
mean(flint$isDangerous) * 100
```

```
## [1] 4.436229
```

```
# Another way
total <- nrow(flint)
dangerous <- nrow(flint %>% filter(Pb >= 15))
(dangerous/total) * 100
```

```
## [1] 4.436229
```

We can see that 4.436229% of the locations were found to have dangerous levels of lead.

c. Report the mean copper level for only test sites in the North region

```
flint_north <- flint %>% filter(Region == "North")
mean(flint_north$Cu)
```

```
## [1] 44.6424
```

The mean Cu level for test sites in the North region is 44.6424.

d. Report the mean copper level for only test sites with dangerous lead levels (at least 15PPB).

```
flint_dangerous <- flint %>% filter(Pb >= 15)
mean(flint_dangerous$Cu)
```

```
## [1] 305.8333
```

The mean Cu level for test sites with a PPB >= 15 is 305.8333.

e. Report the mean lead and copper levels for all locations

```
mean(flint$Pb)
```

```
## [1] 3.383272
```

```
mean(flint$Cu)
```

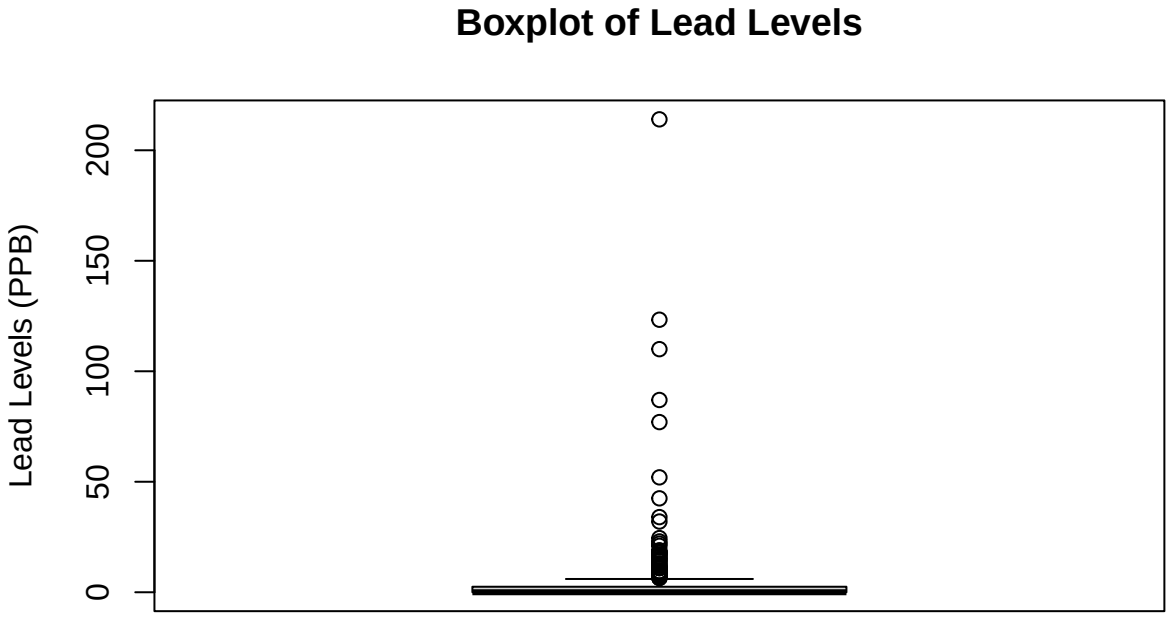
```
## [1] 54.58102
```

The mean lead levels is 3.383272.

The mean copper levels is 54.58102.

f. Create a box plot with a good title for the lead levels. Hint: consider boxplot() function.

```
boxplot(flint$Pb, main = "Boxplot of Lead Levels", ylab = "Lead Levels (PPB)")
```



g. Based on what you see in part (f), does the mean seem to be a good measure of center for the data? Report a more useful statistic for this data.

According to the box plot there is are several outliers that are causing the mean to be way higher than most of the data points which appear to be close to 0. A more useful statistic that could be used is the median.

```
median(flint$Pb)
```

```
## [1] 0
```

The median is 0, which is more in line with the boxplot is showing us.

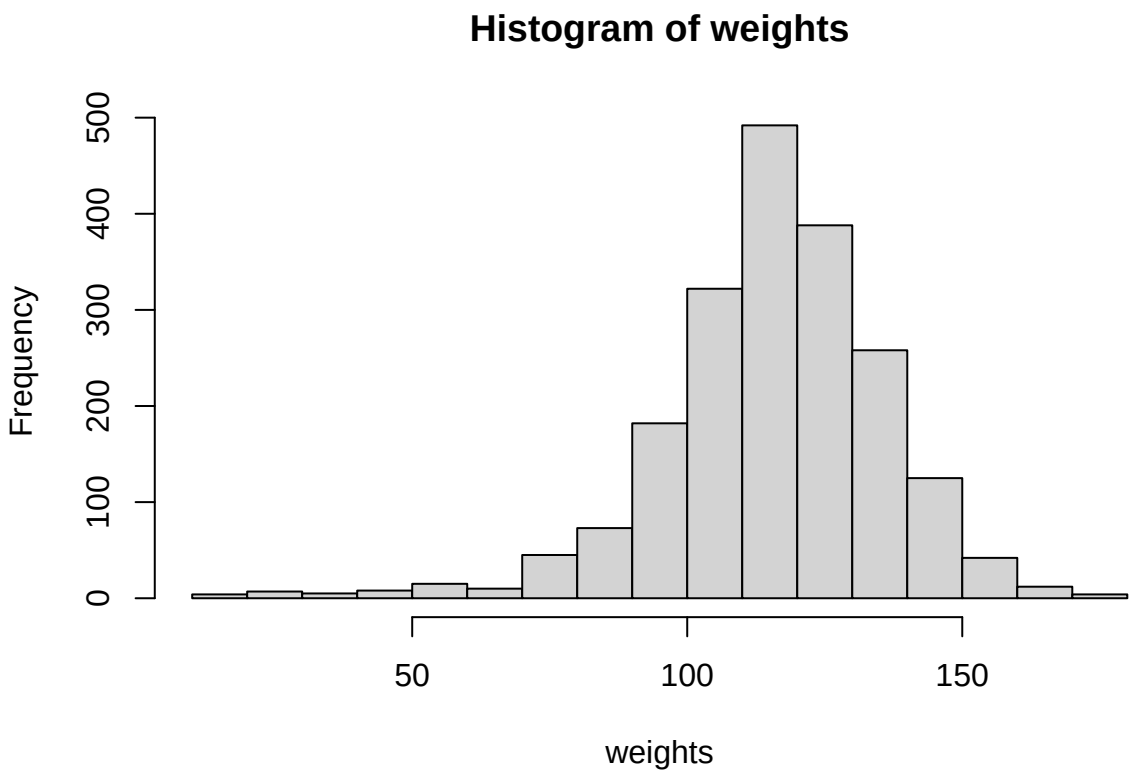
Question 7

a. Use hist() function to plot a histogram on the weight variable in the NCbirths. Do you think weight follows a normal distribution? Why?

```
weights <- NCBirths$weight
mean(weights)
```

```
## [1] 116.0512
```

```
hist(weights)
```



The weight follows a normal distribution, however it appears to be slightly left-skewed. Most of the weights seem to occur close to the mean which is ~116, and weights that are further away from the mean occur less.

b. Use sample() function to randomly select 10 observations from weight. Show the mean of these 10 observations.

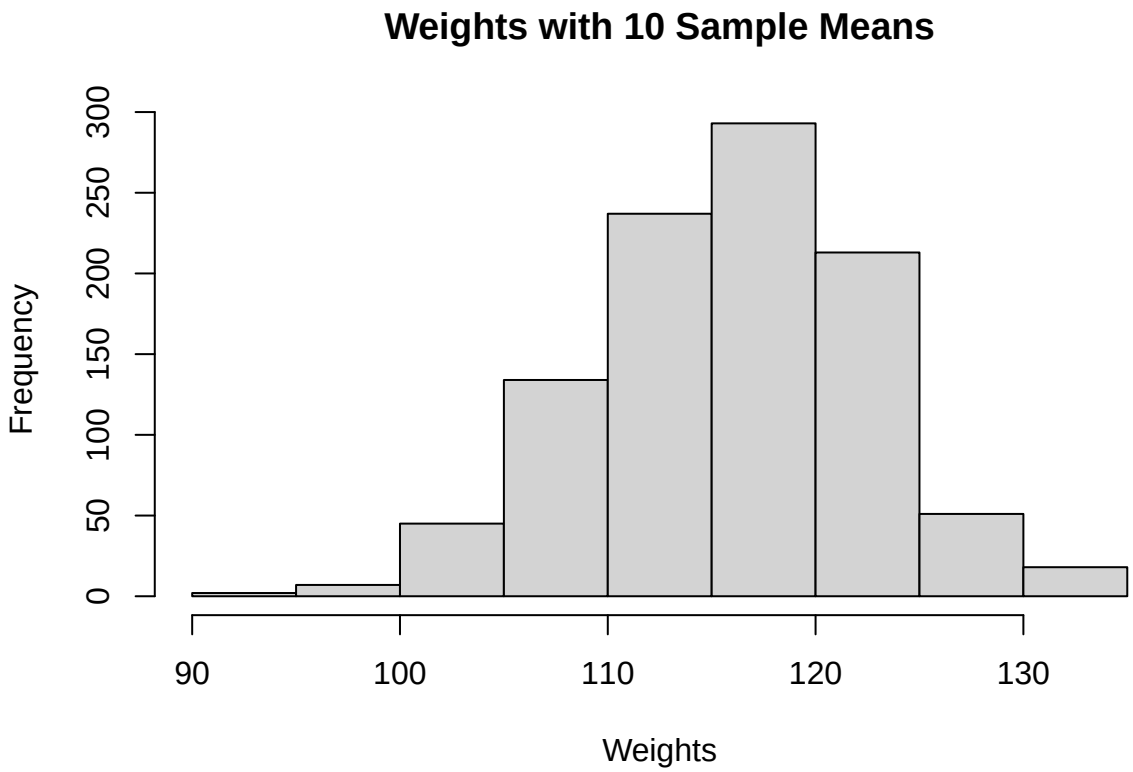
```
set.seed(400) # Setting a seed to get consistent results every run.
mean(sample(weights, size = 10))
```

```
## [1] 117.9
```

c. Use a for loop to repeat the (b) 1000 times. Save 1000 means in a vector. Show the histogram for 1000 means. Is this distribution close to normal?

```
# Writing the loop in a function since this code is used for the next few questions.
get_sample_means <- function(x){
  sample_means <- c()
  for (i in 1:1000){
    sample_means[i] <- mean(sample(weights, size = x))
  }
  return(sample_means)
}
```

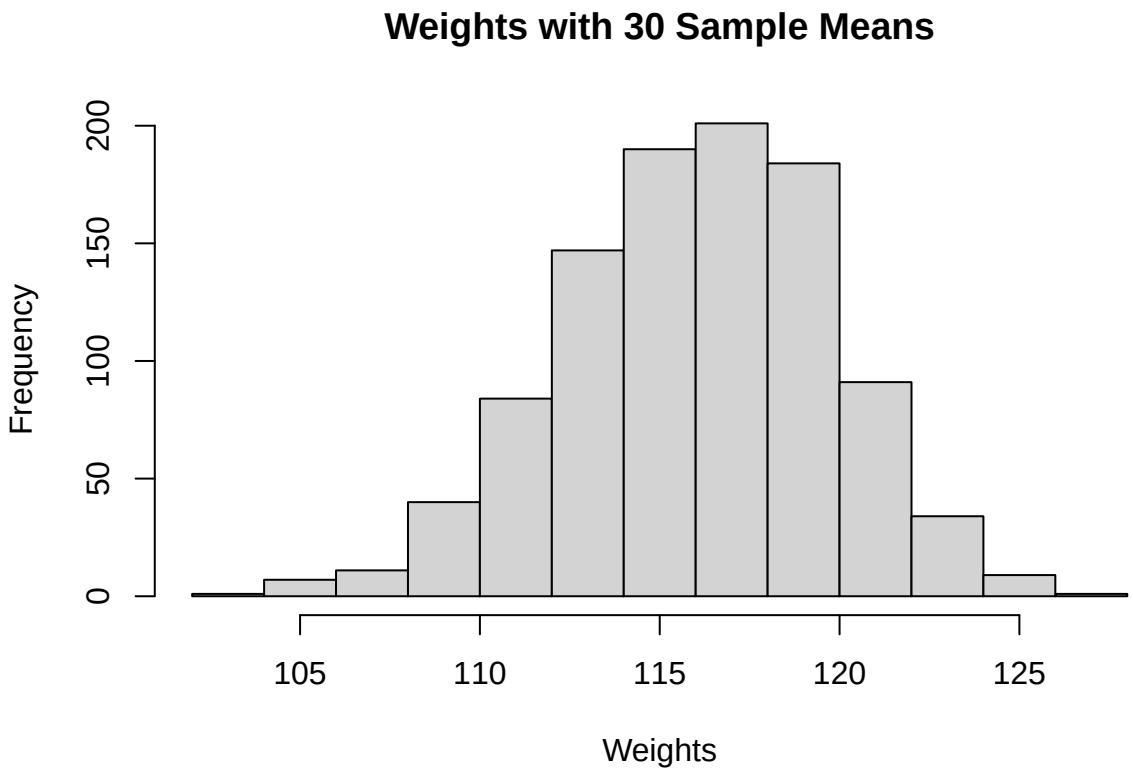
```
hist(get_sample_means(10), main = "Weights with 10 Sample Means", xlab = "Weights")
```



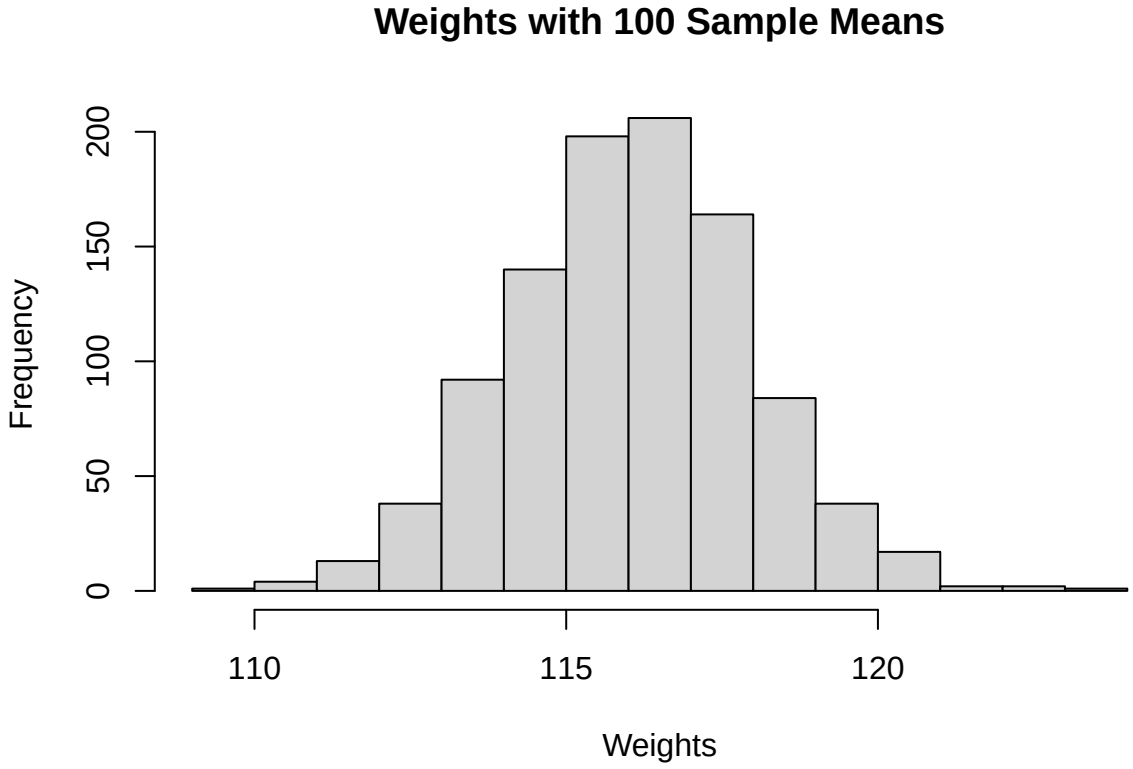
The distribution looks slightly normal, it is clear that there is not enough data points to show a proper distribution.

d. Change the sample size 10 in (b) to 30 and 100, Repeat (c) for these two sample sizes. Are these two distributions close to normal? Interpret your reason.

```
hist(get_sample_means(30), main = "Weights with 30 Sample Means", xlab = "Weights")
```




```
hist(get_sample_means(100), main = "Weights with 100 Sample Means", xlab = "Weights")
```



The distribution with 30 samples appears more normal compared to the distribution with 10 samples, and the distribution with a 100 samples appear the most normal, and is not as skewed as the the original distribution (more than likely due to the large outliers in the original weights vector not appearing in the sample).