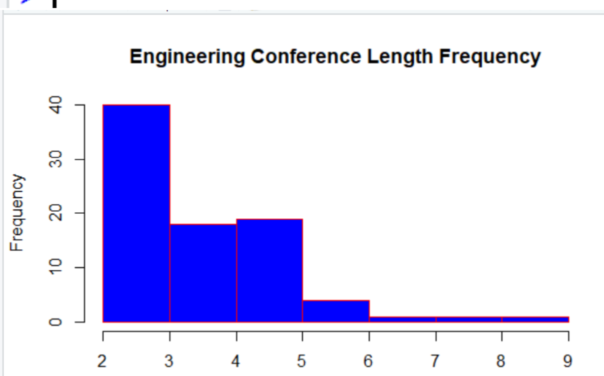
MSIS 5503 – Statistics for Data Science – Fall 2021 - Assignment 5

1. (**1 point**) In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

   Do everything in R and check with a calculator where appropriate:

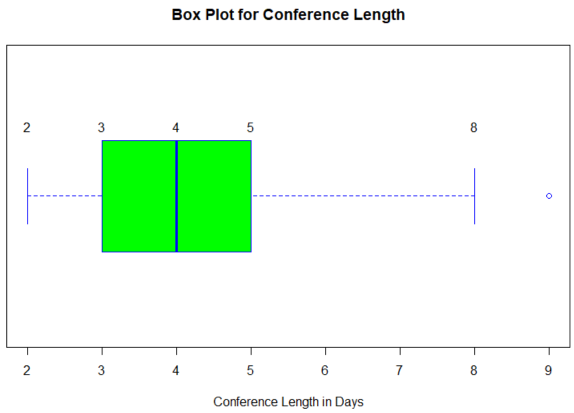
   a. Construct a Histogram with appropriate title and labels for the X and Y axes.
   b. Find the median, the first quartile, and the third quartile.
   c. Find the 10 and 65th percentiles
   d. Calculate the IQR
   e. Construct a box plot of the data and identify any outliers.
   f. The middle 50% of the conferences last from ________ days to ________ days.
   g. Calculate the sample mean of days of engineering conferences.
   h. Calculate the sample standard deviation of days of engineering conferences.
   i. Find the mode.
   j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
   k. Calculate the skewness and kurtosis for the data and *interpret* them.

```
> vec_days <- c(rep(2,4),rep(3,36),rep(4,18),rep(5,19),rep(6,4),7,8,9)
> hist(vec_days, breaks=8,border="red", col="blue", main="Engineering Conferenc
e Length Frequency",xlab="Length in Days",ylab="Frequency")
>
```




```
> print(paste("10th percentile is ", quantile(vec_days, 0.10, type=2)))
[1] "10th percentile is  3"
> print(paste("65th percentile is ", quantile(vec_days, 0.65, type=2)))
[1] "65th percentile is  4"
> print(paste("IQR is ",IQR(vec_days, type=2)))
[1] "IQR is  2"
```

```
> boxplot(vec_days, horizontal = TRUE, main="Box Plot for Conference Length",
+         xlab="Conference Length in Days",
+         border="blue",
+         col="green")
> text(x=boxplot.stats(vec_days)$stats, labels = boxplot.stats(vec_days)$stats, y = 1.25)
```




```
> print(paste("Outlier is ",9," days"))
[1] "Outlier is  9  days"
> print(paste("The middle 50% of the conferences last from ",quantile(vec_days, 0.25, type=2) ,"days to",
+             quantile(vec_days, 0.75, type=2), "days."))
[1] "The middle 50% of the conferences last from  3 days to 5 days."
> print(paste("The mean number of days for conference length is: ",round(mean(vec_days),4)))
[1] "The mean number of days for conference length is:  3.9405 "
> print(paste("The standard deviation for conference length is: ",round(sd(vec_days),4)," days"))
[1] "The standard deviation for conference length is:  1.2836   days"
```

```
> df <- data.frame(days_tbl <- table(vec_days))
> df
  vec_days Freq
1        2    4
2        3   36
3        4   18
4        5   19
5        6    4
6        7    1
7        8    1
8        9    1
> print(" The mode is : 3 days")
[1] " The mode is : 3 days"
> print(paste("The skewness of the data is:", round(skewness(vec_days),4)))
[1] "The skewness of the data is:  1.2812
> print(paste("The kurtosis of the data is:", round(kurtosis(vec_days),4)))
[1] "The kurtosis of the data is:  5.3728 "
> print(paste("The data is right skewed because the skewness is positive"))
[1] "The data is right skewed because the skewness is positive"
> print(paste("The data is lepto_kurtic because the kurtosis is greater than 3.",
+             " This means that there are many observations in the tail, including potential outliers"))
```

j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.

   Mode would be a good choice because it tells you the most common conference length that allows for planning.
   Other answers acceptable depending on your explanation.

k. Calculate the skewness and kurtosis for the data and *interpret* them.
   Skewness = 1.2812 tells us that the random variable has more large values (than small values), each with smaller and smaller probabilities. It has a longer right tail relative to the normal distribution
   Kurtosis = 5.3728 (leptokurtic) tells us that there are many observations in the tails (relative to a normal distribution), including potential outliers.

2. (**1 point**) The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the Table below:

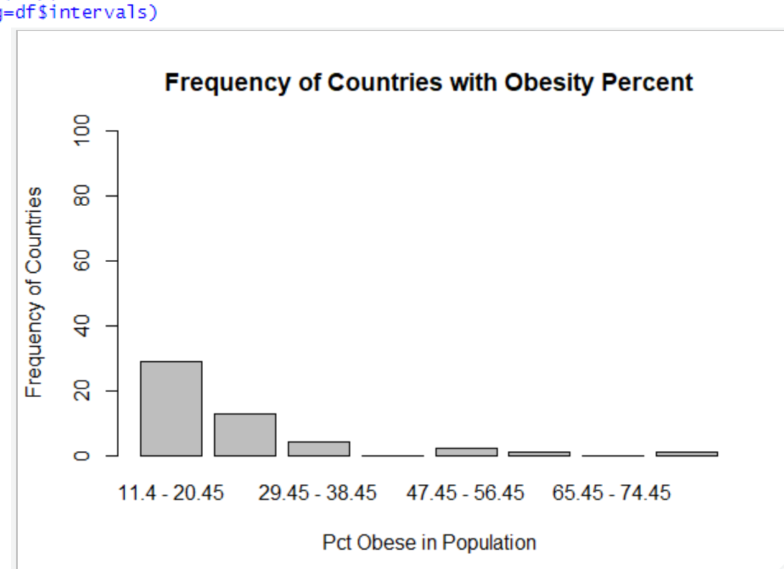
Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

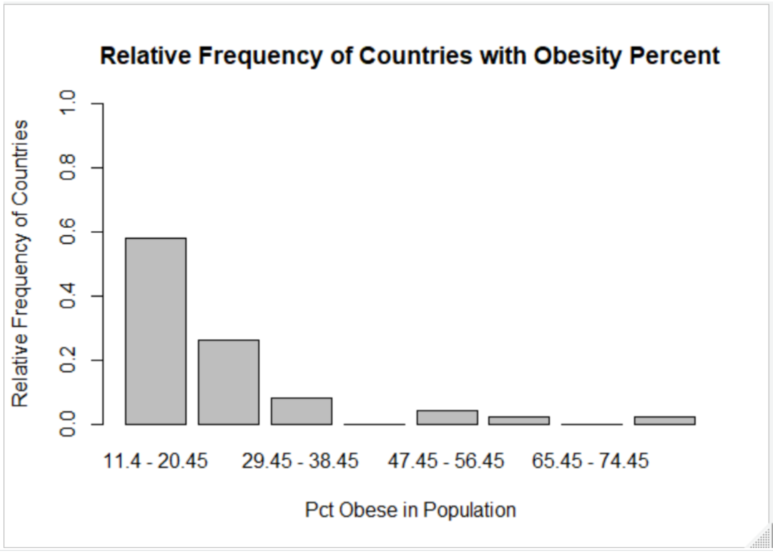
```
> low_pct <- c(11.4, 20.45, 29.45, 38.45, 47.45, 56.45, 65.45, 74.45)
> hi_pct <- c(20.45, 29.45, 38.45, 47.45, 56.45, 65.45, 74.45, 83.45)
> df <- data.frame(low_pct, hi_pct)
> df$intervals <- c(paste(low_pct,"-",hi_pct))
> df$xi <- (df$low_pct +df$hi_pct)/2
> df$fi <- c(29, 13, 4, 0, 2,1,0,1)
> df$rel_fi <- df$fi/sum(df$fi)
> df$cum_rel_fi = cumsum(df$rel_fi)
> df
  low_pct hi_pct     intervals     xi fi rel_fi cum_rel_fi
1   11.40  20.45  11.4 - 20.45 15.925 29   0.58       0.58
2   20.45  29.45 20.45 - 29.45 24.950 13   0.26       0.84
3   29.45  38.45 29.45 - 38.45 33.950  4   0.08       0.92
4   38.45  47.45 38.45 - 47.45 42.950  0   0.00       0.92
5   47.45  56.45 47.45 - 56.45 51.950  2   0.04       0.96
6   56.45  65.45 56.45 - 65.45 60.950  1   0.02       0.98
7   65.45  74.45 65.45 - 74.45 69.950  0   0.00       0.98
8   74.45  83.45 74.45 - 83.45 78.950  1   0.02       1.00
```

<u>Do everything in R and check with a calculator where appropriate:</u>

a. Draw bar plots of frequency and relative frequency with appropriate title and labels for the X and Y axes.
b. Characterize the skewness of the data.
c. What percentage of countries have an obesity percentage greater than or equal to 47.45%
d. Calculate the (approximate) sample mean of obesity percentage
e. Calculate the (approximate) sample standard deviation of obesity percentage.

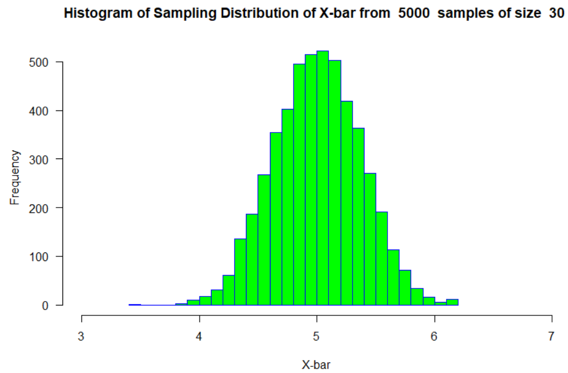
```
> barplot(df$fi,
+         main = "Frequency of Countries with Obesity Percent",
+         horiz=FALSE,
+         xlab = "Pct Obese in Population",
+         ylab = "Frequency of Countries",
+         ylim =c(0, 100),
+         names.arg=df$intervals)
> #
> barplot(df$rel_fi,
+         main = "Relative Frequency of Countries with Obesity Percent",
+         horiz=FALSE,
+         xlab = "Pct Obese in Population",
+         ylab = "Relative Frequency of Countries",
+         ylim =c(0, 1),
+         names.arg=df$intervals)
```

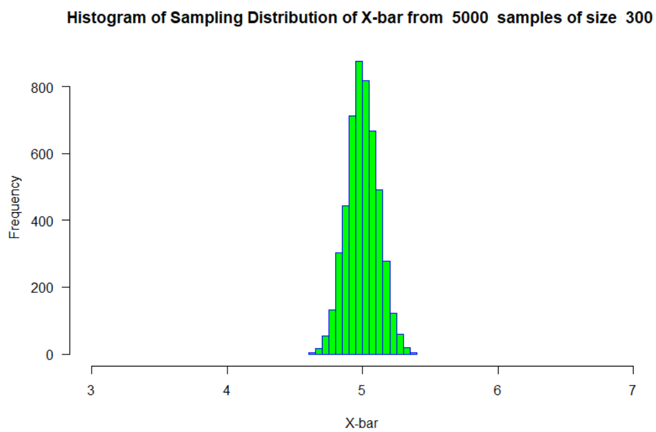





```
> print("The bar-graph shows that the pecentage of obesity is right-tailed (positively skewed)")
[1] "The bar-graph shows that the pecentage of obesity is right-tailed (positively skewed)"
> print(paste("Based on the cumulative relative frequencey graph, the percentage of countries with obsesity percentage greater than or equal to 47.45% is",
8," percent"))
[1] "Based on the cumulative relative frequencey graph, the percentage of countries with obsesity percentage greater than or equal to 47.45% is 8  percent"
> df$xifi <- df$xi*df$fi
> expected_val <- sum(df$xifi)/sum(df$fi)
> expected_val
[1] 23.3155
> print(paste("The approximate mean percent of population that is obsese - all countries ",round(expected_val, 4)))
[1] "The approximate mean percent of population that is obsese - all countries  23.3155"
> df$deviation <- df$xi - expected_val
> df$sq_dev <- df$deviation*df$deviation
> varian <- sum(df$fi*df$sq_dev)/(sum(df$fi)-1)
> st_dev <-sqrt(varian)
> print(paste("The approximate standard deviation of the percent of population that is obsese - all countries ",round(st_dev, 4)))
[1] "The approximate standard deviation of the percent of population that is obsese - all countries  12.954"
```

3. (**0.6 points** each for a and b, **0.3 points** for each part of c):
   a. Using **R,** for 5000 samples of sizes 30 and 300 from a population X ~ N(**μ** = 5, **σ** = 2), construct a histogram of the sampling distribution of $\overline{X}$.
   b. Using R, calculate the *empirical* mean and standard deviation of $\overline{X}$ in each case. Compare them with the *theoretically* predicted mean and standard deviation of $\overline{X}$.
   c. **In your own words**:
      i. What does the sampling distribution of the estimator $\overline{X}$ represent?
      ii. What does the **expected value** of the estimator $\overline{X}$ tell us about the relationship between the *sample means* $\overline{X}$ and the *population mean* **μ**?
      iii. What does the **standard error** of the estimator $\overline{X}$ tell us about the relationship between the *sample mean* $\overline{X}$ from each sample and the *population mean* **μ**?
      iv. Using the *theoretical sampling distribution* of $\overline{X}$ for this example, 99% of sample means lie between which values? Do this for each sample size. Do the same using the *empirical sampling distribution for each sample size*.
      v. Based on your answer to the previous question, would you be surprised if you took a sample of 30, to find a sample mean greater than 5.5? How about with a sample of size 300?
      vi. Why does the $\overline{X}$, the sample mean, from any sample get closer to **μ** as you increase sample size?






```
> # Generating an empirical sampling distribution of sample mean - X-bar
> # Define the x_bar vector
> num_samp = 5000
> samp_size = 30
> x_bar <- vector("numeric", num_samp)
> #
> # Each x-bar is the mean of a random sample of size 50 drawn from a Normal(10, 5)
> #
> # We are generating 1000 X-bars (from 1000 samples) and storing them in the x-bar vector
> #
> for (i in 1:num_samp) {
+   x_bar[i] = mean(rnorm(samp_size, 5, 2))
+ }
> #
> # Calculate the mean and standard deviation (called standard error) of x-bar
> # from the empirical sampling distribution formed by 1000 samples of size 50
> #
> Exp_x_bar <- mean(x_bar)
> stderr <- sd(x_bar)
> print(paste("The Expected value of X-bar is: ", round(Exp_x_bar,4)))
[1] "The Expected value of X-bar is:  4.9932"
> print("The theoretical mean of the sampling distribution is 5")
[1] "The theoretical mean of the sampling distribution is 5"
> print(paste("The standard error or standard deviation of X-bar is: ", round(stderr,4),
+              " versus predicted std error ",round(2/sqrt(samp_size),4)))
[1] "The standard error or standard deviation of X-bar is:  0.3625  versus predicted std error  0.3651"
> hist(x_bar,
+      main=paste("Histogram of Sampling Distribution of X-bar from ",num_samp,
+                 " samples of size ", samp_size, ""),
+      xlab="X-bar",
+      border="blue",
+      col="green",
+      xlim=c(3, 7),
+      las=1,
+      breaks=20)
> #
```

```
> # Generating an empirical sampling distribution of sample mean - X-bar
> # Define the x_bar vector
> num_samp = 5000
> samp_size = 300
> x_bar <- vector("numeric", num_samp)
> #
> # Each x-bar is the mean of a random sample of size 50 drawn from a Normal(10, 5)
> #
> # We are generating 1000 X-bars (from 1000 samples) and storing them in the x-bar vector
> #
> for (i in 1:num_samp) {
+   x_bar[i] = mean(rnorm(samp_size, 5, 2))
+ }
> #
> # Calculate the mean and standard deviation (called standard error) of x-bar
> # from the empirical sampling distribution formed by 1000 samples of size 50
> #
> Exp_x_bar <- mean(x_bar)
> stderr <- sd(x_bar)
> print(paste("The Expected value of X-bar is: ", round(Exp_x_bar,4)))
[1] "The Expected value of X-bar is:  4.9996"
> print("The theoretical mean of the sampling distribution is 5")
[1] "The theoretical mean of the sampling distribution is 5"
> print(paste("The standard error or standard deviation of X-bar is: ", round(stderr,4),
+             " versus predicted std error ",round(2/sqrt(samp_size),4)))
[1] "The standard error or standard deviation of X-bar is:  0.1154  versus predicted std error  0.1155"
> hist(x_bar,
+      main=paste("Histogram of Sampling Distribution of X-bar from ",num_samp,
+                 " samples of size ", samp_size, ""),
+      xlab="X-bar",
+      border="blue",
+      col="green",
+      xlim=c(3, 7),
+      las=1,
+      breaks=20)
```

b) The empirical expected value of $\overline{X}$ for sample size 30 = 4.9932 and for sample size 300 = 4.9996 (against the theoretical value of 5) and the standard error (or standard deviation of $\overline{X}$) for sample size 30 = 0.3625 (versus the theoretical value 0.3651) and for sample size 300 = 0.1154 (versus predicted 0.1155).
(**Note:** Your empirical numbers will be different, but not your theoretical numbers)

C1) What the sampling distribution of the estimator $\bar{X}$ represents:
The sampling distribution of $\overline{X}$ is the distribution of the statistic (estimator) of $\overline{X}$ from all possible samples of a particular size.

*C2) What the **expected value** of the estimator $\bar{X}$ tell us about the relationship between the sample means $\bar{X}$ and the population mean $\boldsymbol{\mu}$:*
The average of the sample means ($\overline{X}$) or E($\overline{X}$) = $\boldsymbol{\mu}$ (when we consider all possible samples of a particular size).

*C3) What the **standard error** of the estimator $\bar{X}$ tell us about the relationship between the sample mean $\bar{X}$ from each sample and the population mean $\boldsymbol{\mu}$*:
The standard error *of the estimator* $\overline{X}$ tells us the variability or spread of the $\overline{X}$ from each sample away from $\boldsymbol{\mu}$ (when we consider all possible samples of a particular size)

C4) Using the theoretical sampling distribution of $\overline{X}$ for this example, 99% of sample means lie between which values? Do this for each sample size. Do the same using the *empirical sampling distribution for each sample size.*

Sample size = 30

```
> print(paste("Theoretically, 99% of X-bars are between: ", qnorm(0.005, pop_mean, pop_sd/sqrt(samp_size)), " and ", qnorm(0.995, pop_mean, p
op_sd/sqrt(samp_size))))
[1] "Theoretically, 99% of X-bars are between:  4.05944012410896  and  5.94055987589104"
```

```
> print(paste("Empirically, 99% of X-bars are between: ", quantile(x_bar, probs = 0.005, na.rm=FALSE, names = TRUE, type=2)," and ",quantile
(x_bar, probs = 0.995, na.rm=FALSE, names = TRUE, type=2) ))
[1] "Empirically, 99% of X-bars are between:  4.05312878825103  and  5.94734101239136"
```

Sample size = 300

```
> print(paste("Theoretically, 99% of X-bars are between: ", qnorm(0.005, pop_mean, pop_sd/sqrt(samp_size)), " and ", qnorm(0.995, pop_mean, p
op_sd/sqrt(samp_size))))
[1] "Theoretically, 99% of X-bars are between:  4.7025688516419  and  5.2974311483581"
> print(paste("Empirically, 99% of X-bars are between: ", quantile(x_bar, probs = 0.005, na.rm=FALSE, names = TRUE, type=2)," and ",quantile
(x_bar, probs = 0.995, na.rm=FALSE, names = TRUE, type=2) ))
[1] "Empirically, 99% of X-bars are between:  4.70743903033927  and  5.29424804392041"
```

C5) Based on your answer to the previous question, would you be surprised if you took a sample of 30, to find a sample mean greater than 5.5? How about with a sample of size 300?

If you use the above answer to be your guide, you would see that for a sample size of 30, $\overline{\mathrm{X}} = 5.5$ is well within the value of 99% of all possible sample means, whereas $\overline{\mathrm{X}} = 5.5$ is outside the bounds of 99% of the $\overline{\mathrm{X}}$s, for sample size = 300. Therefore, the probability of observing a sample mean of 5.5 is less than 0.01 and we would be surprised if it happened.

C6) Why does the $\bar{X}$, the sample mean, from any sample get closer to $\boldsymbol{\mu}$ as you increase sample size?

As the sample size increases, the standard error reduces (for example, based on the formula, the sample size is in the denominator of the formula for standard error). This means, that the standard deviation of the sampling distribution of $\overline{\mathrm{X}}$ is smaller. This means, that the $\overline{\mathrm{X}}$ from every sample is close to the population mean, as well as to other $\overline{\mathrm{X}}$s. Also, as sample size increases, the sample contains more information about the population characteristics such as the mean.