Homework #1_FA2024_BMEN7340

First_Last_7340_HW1_FA24

2024-09-11

Question #1:

In a given vector: birthweight <- c(3600,1700,4000,3900,3100,3800,2200,3000)
Use a single R syntax to find the mean, median, min, max, 1st quartile, and 3rd quartile.
Use a single R syntax to find the standard deviation of "birthweight".
Use a single R syntax to create a histogram of "birthweight".

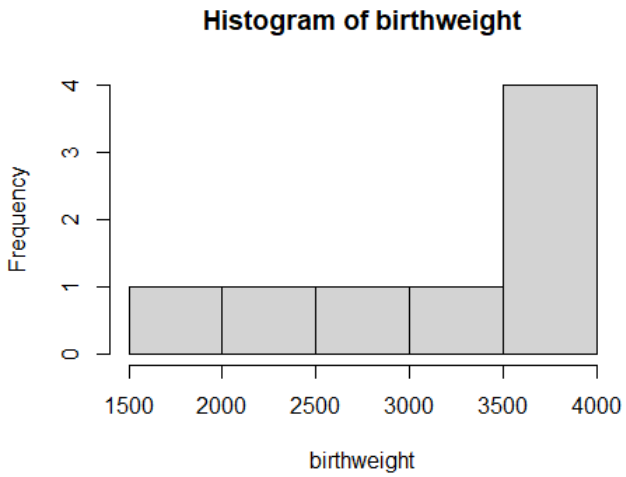
```
birthweight <- c(3600,1700,4000,3900,3100,3800,2200,3000)
summary(birthweight)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1700    2800    3350    3162    3825    4000

sd(birthweight)

## [1] 839.9617

hist(birthweight)
```

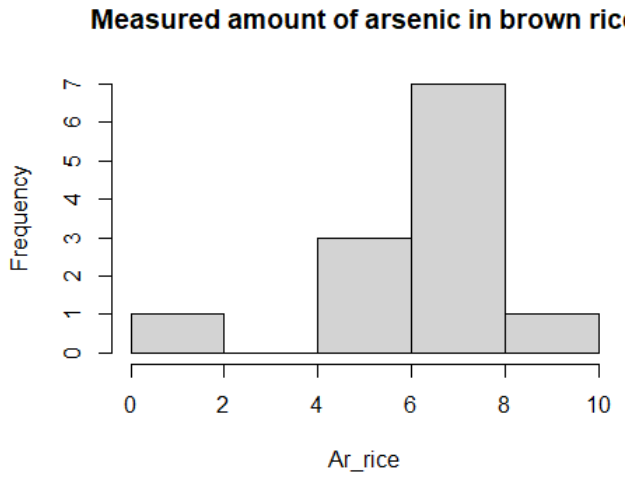


Question #2 is about arsenic in rice with a unit of ug/serving.

Ar_rice <- c(6.1, 5.4, 6.9, 4.9, 6.6, 6.3, 6.7, 8.2, 7.8, 1.5, 5.4, 7.3)
Construct a histogram with 2 units of bin size.

Give the title “Measured amount of arsenic in brown rice”.
Give a y label “Frequency”.

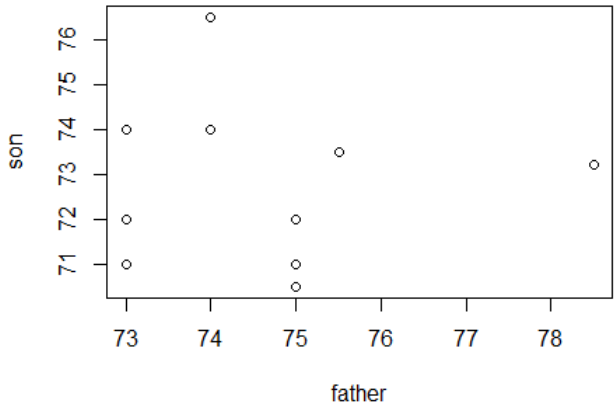
```
Ar_rice <- c(6.1, 5.4, 6.9, 4.9, 6.6, 6.3, 6.7, 8.2, 7.8, 1.5, 5.4, 7.3)
hist(Ar_rice, breaks=seq(from=0, to=10, by=2),
     main="Measured amount of arsenic in brown rice",
     ylab="Frequency")
```



Question #3:

The two vectors “father” and “son” lists heights (in.). Use a descriptive statistic method to describe the two variables. father <- c(73.0, 75.5, 75.0, 75.0, 75.0, 74.0, 74.0, 73.0, 73.0, 78.5)
son <- c(74.0, 73.5, 71.0, 70.5, 72.0, 76.5, 74.0, 71.0, 72.0, 73.2)
write the command of a scatter plot
plot(father,son)

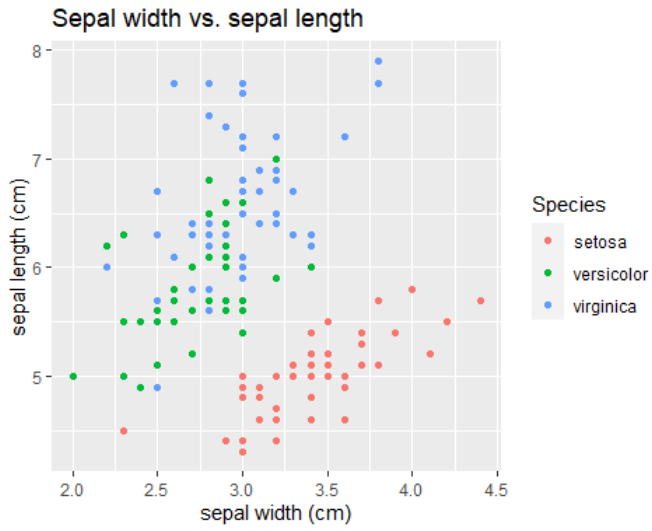
```
father <- c(73.0, 75.5, 75.0, 75.0, 75.0, 74.0, 74.0, 73.0, 73.0, 78.5)
son <- c(74.0, 73.5, 71.0, 70.5, 72.0, 76.5, 74.0, 71.0, 72.0, 73.2)
plot(father,son)
```



Question #4

Using the ggplot2 package, create a scatter plot of the iris dataset (included in ggplot2), where the x-axis represents the sepal width and the y-axis represents the sepal length. Color the points by the class of the species. The title is "Sepal width vs. sepal length".

```
# library(ggplot2)
# Scatter plot of sepal width vs. sepal length colored by species
ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
  geom_point() +
  labs(title = "Sepal width vs. sepal length",
       x = "sepal width (cm)",
       y = "sepal length (cm)")
```



Question #5, qnorm

qnorm: Assume that a randomly selected subjects is given a bone density test. The test scores are normally distributed with a mean of 0 and a standard deviation of 1. Find the bone density test score corresponding to the given information. 1. Find the bone density score separating the bottom 99% from the top 1% (Z0.01). 2. If the bone density in the bottom 2% and the top 2% are used as cutoff points for levels that are too low or too high, find the two readings that are cutoff values.

```
qnorm(0.01, mean=0, sd=1, lower.tail=F) # 2.33
```

```
## [1] 2.326348
```

```
qnorm(0.025, mean=0, sd=1, lower.tail=T) # -1.95
```

```
## [1] -1.959964
```

```
qnorm(0.025, mean=0, sd=1, lower.tail=F) # 1.95
```

```
## [1] 1.959964
```

Question #6, pnorm

Assume that a randomly selected subjects is given a bone density test. The test scores are normally distributed with a mean of 0 and a standard deviation of 1.
1. Find the probability of the given bone density test score between -2.75 and 2.75. 2. Find the probability of a given bone density test score between -2.0 and 2.0

```
# Pr(-2.75 < Z < -0.75)
pnorm(2.75) - pnorm(-2.75) # 99.4%
```

```
## [1] 0.9940405
```

```
pnorm(2.75, mean=0, sd=1, lower.tail=T) - pnorm(-2.75, mean=0, sd=1,
lower.tail=T) # 99.4%
```

```
## [1] 0.9940405
```

```
# Pr(-2 < Z < 2)
pnorm(2) - pnorm(-2) # 95.4%
```

```
## [1] 0.9544997
```

Question #7, IQR

Finding probability of something being an outlier from a normal distribution.

```
# Start with finding Z score of third quartile
Quartile_3rd <- qnorm(0.75, mean=0, sd=1, lower.tail=T)  # 0.67449
Quartile_1st <- qnorm(0.25, mean=0, sd=1, lower.tail=T)  # -0.67449
```

```r
#Find the IQR
IQR <- (Quartile_3rd - Quartile_1st) #1.34

#Multiply by 1.5 to get to the outliers Z-score
Z_outlier_upper <- Quartile_3rd + IQR * 1.5   # 2.69
Z_outlier_lower <- Quartile_1st - IQR * 1.5   # -2.69

#Find the probability of outliers
pnorm(Z_outlier_upper, mean=0, sd=1, lower.tail=F) + pnorm(Z_outlier_lower,
mean=0, sd=1, lower.tail=T)
```

```
## [1] 0.006976603
```

```r
#0.006976603
```

Question #8, Critical values

Find the indicated critical values. Round results to two decimal places. Z0.10, Z0.97, and Z0.025

```r
qnorm(0.1, mean=0, sd=1, lower.tail=F) # 1.28
```

```
## [1] 1.281552
```

```r
qnorm(0.97, mean=0, sd=1, lower.tail=F) #-1.88
```

```
## [1] -1.880794
```

```r
qnorm(0.025, mean=0, sd=1, lower.tail=F) #1.95
```

```
## [1] 1.959964
```

Question #9, subsetting

Pulse Rates of Females: Refer to Data "Body Data" and use the pulse rates (beats per minute) of females (0=F, 1=M), construct a histogram of females' pulse rates Do the pulse rates of females appear to have a normal distribution?

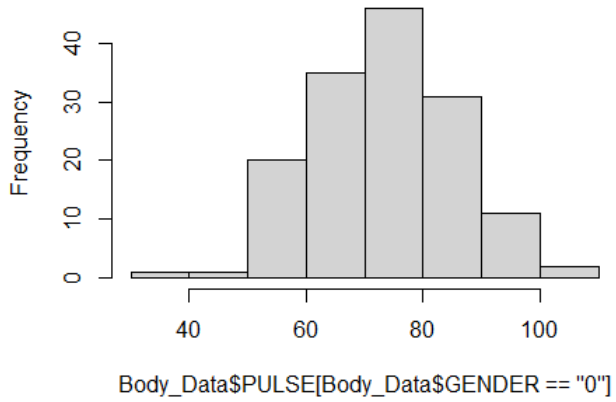
```r
# library(readxl)
Body_Data <- read_excel("C:/Users/Body Data.xlsx")
str(Body_Data)
```

```
## tibble [300 × 15] (S3: tbl_df/tbl/data.frame)
##  $ AGE      : num [1:300] 43 57 38 80 34 77 29 69 44 35 ...
##  $ GENDER   : num [1:300] 0 1 0 1 1 1 1 0 0 1 ...
##  $ PULSE    : num [1:300] 80 84 94 74 50 60 52 58 66 62 ...
##  $ SYSTOLIC : num [1:300] 100 112 134 126 114 134 118 138 114 124 ...
##  $ DIASTOLIC: num [1:300] 70 70 94 64 68 60 64 80 66 70 ...
##  $ HDL      : num [1:300] 73 35 36 37 50 55 53 40 45 62 ...
```

```
##  $ LDL      : num [1:300] 68 116 223 83 104 75 128 140 136 110 ...
##  $ WHITE    : num [1:300] 8.7 4.9 6.9 7.5 6.1 5.7 4.1 8.1 8 5.6 ...
##  $ RED      : num [1:300] 4.8 4.73 4.47 4.32 4.95 3.95 4.68 4.6 4.09 5.47
...
##  $ PLATE    : num [1:300] 319 187 297 170 140 192 191 286 263 193 ...
##  $ WEIGHT   : num [1:300] 98.6 96.9 108.2 73.1 83.1 ...
##  $ HEIGHT   : num [1:300] 172 186 154 160 179 ...
##  $ WAIST    : num [1:300] 120.4 107.8 120.3 97.2 95.1 ...
##  $ ARM CIRC : num [1:300] 40.7 37 44.3 30.3 34 31.4 27.4 34.2 32.5 40 ...
##  $ BMI      : num [1:300] 33.3 28 45.4 28.4 25.9 31.1 20.1 32.7 25.8 36.5
...
```

```
hist(Body_Data$PULSE[Body_Data$GENDER=="0"])
```

togram of Body_Data$PULSE[Body_Data$GENDER




```
shapiro.test(Body_Data$PULSE[Body_Data$GENDER=="0"]) # p = 0.31 > 0.05, FTR
Ho, normally
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Body_Data$PULSE[Body_Data$GENDER == "0"]
## W = 0.98912, p-value = 0.3102
```

Question #10, Confidence Interval -1

List below are the measured radiation emissions (W/kg) corresponding to these randomly selected cell phones: S1, B1, B2, M1, T1, A1, P1, P2, N1, A2, K1. The data are from the Environment Working Group. The media often presents reports about the dangers of cell phone radiation as a cause of disease. Construct a 90% confidence interval estimate of the population mean. What does the result suggest about the Federal Communications Commission (FCC) standard that cell phone radiation must be less than 1.6 W/kg? 0.38; 0.55; 1.54; 1.55; 0.50; 0.6; 0.92; 0.96;1.00; 0.86; 1.46

```r
# Ho: The radiation is 1.6w/kg
# Ha: The radiation is less than 1.6w/kg
phoneRadiation <- c(0.38, 0.55, 1.54, 1.55, 0.5, 0.6, 0.92, 0.96, 1, 0.86,
1.46) #W/kg
## construct a 90% CI, alpha = 0.1 ##
shapiro.test(phoneRadiation)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  phoneRadiation
## W = 0.90192, p-value = 0.1951
```

```r
# p (0.1951) > 0.05, data is normally distributed
n <- length(phoneRadiation)
# check assumption, 1. SRS, 2. sample size is large
tcritical <- qt(0.05, df=n-1, lower.tail=T) #  t.critical is corresponding to
area alpha=0.05
se <- sd(phoneRadiation)/sqrt(n)
upper <- mean(phoneRadiation) + tcritical*se
lower <- mean(phoneRadiation) - tcritical*se
upper; lower
```

```
## [1] 0.7070946
```

```
## [1] 1.169269
```

```r
# we are 90% confidence that the true population mean radiation of cell phone
is between
# 0.71 to 1.17 W/kg, since 1.6w/kg the Ho is not in the range of 90% CI, we
reject Ho.
# Conclusion: The FCC suggested a less than 1.6W/kg, therefore the cell
phones meet the standard.
```

Question #11, Confidence Interval -2

Bipolar depression treatment. In an experiment designed to test the effectiveness of paroxetine for treating bipolar depression, subjects were measured using the Hamilton depression scale, with the results given: placebo group, n=43, mean=21.57, s=3.87; Paroxetine treatment group: n=33, mean=20.38, s=3.91. Use a 0.05 significance level to test the claim that the treatment group and placebo group come from populations with the same mean. What does the result of the hypothesis test suggest about paroxetine as a treatment for bipolar depression?

```r
# Formulate Ho and Ha, based on the research question
# Ho: the mean difference of depression level is = 0
# Ha: the mean difference is not zero

# use confidence interval method to test Ho hypothesis
```

```r
mean.diff <- (21.57-20.38) # 1.19
df <- (33-1) # the smaller of n-1
tcrit <- qt(0.025, df,lower.tail=F) # 2.03
SEM <- sqrt(3.87^2/43+ 3.91^2/33) # standard error to the mean is calculated
based on unequal variance equation
ME <- tcrit*SEM
upper <- mean.diff + ME # 3.025
lower <- mean.diff - ME # -0.645
lower;upper
```

```
## [1] -0.6450223
```

```
## [1] 3.025022
```

```r
# Conclusion: we are 95% confidence that the population mean difference is
between
# -0.645 and 3.025 (unit), we failed to reject Ho.

# use t test statistic and p value  method to test the hypothesis
# assumption check, SRS, independent groups, large sample sizes

ttest.stat <- (mean.diff-0)/SEM # 1.32
pvalue <- 2*pt(1.32, df=32, lower.tail=F) # p=0.196, p > 0.05 (alpha), FTR Ho

# conclusion: If the Ho is true,  the probability of having a mean difference
of 1.19 unit
# or more extreme is 0.196 or 19.6%, merely by chance. Since 0.196 is more
than type I error (0.05),
# we failed to reject the Ho. We support the conclusion that there is no
difference in mean depression level.
```

Question #12, t.test with hypothesis test-1

Diastolic blood pressure for women. Use the diastolic blood pressure measurements for adult females in dataset “Body Data” to test the claim that the adult female population has a mean diastolic blood pressure less than 90 mmHg. A diastolic blood pressure above 90 is considered to be hypertension. Use a 0.05 significance level, based on the result, can we conclude that none of the adult females in the sample have hypertension?

```r
# Ho: mean >= 90 mmHg; Ha: mean < 90mmHg
t.critical <- qt(0.05, df=146, lower.tail=F) # t.critical = 1.65 at 95%
confidence
mean(Body_Data$DIASTOLIC[Body_Data$GENDER ==0]) # mean = 70.16
```

```
## [1] 70.16327
```

```r
length <- length(Body_Data$DIASTOLIC[Body_Data$GENDER ==0]) # n=147
SD <- sd(Body_Data$DIASTOLIC[Body_Data$GENDER ==0]) # sd=11.2
SEM <- SD/sqrt(length)
```

```r
upper <- (70.16 + t.critical*SEM)
upper
```

```
## [1] 71.6919
```

```r
lower <- (70.16 - t.critical*SEM)
lower
```

```
## [1] 68.6281
```

```r
# we are 95% confidence that the true mean is between 68 - 72 mmHg, since
90mmHg is not in the CI range, we reject the Ho, conclude that the mean is
less than 90mmHg. We can't comment on individual subject's status.
pnorm(90, mean=70.16, sd=SEM, lower.tail = F)
```

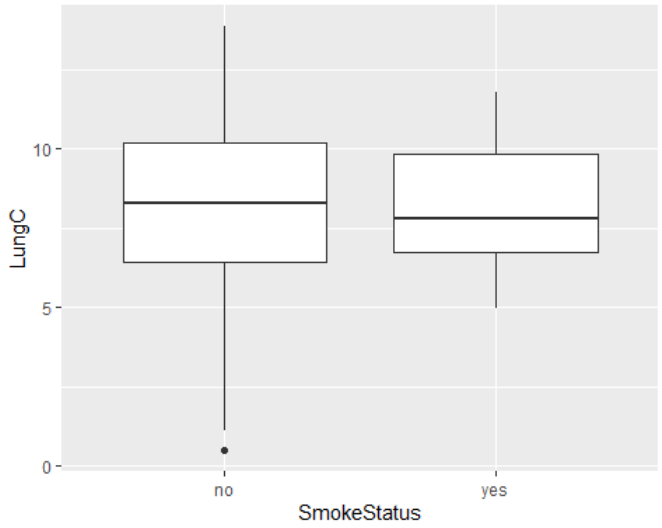
```
## [1] 2.896704e-102
```

Quesiton #13

LCData set. Research question: Do lung capacity (in liters) different between smokers and nonsmokers? a. Use boxplot, examine the relationship between a numeric outcome variable (Y) and a categorical variable (x). b. write Ho and Ha hypotheses, and assumption check. c. decide a left-sided, right-sided, or two-sided test, decide a significant level (or type I error). d. assume equal variances, use t.test command to find a p value. e. interpret the R output, identify the test.statistic, df, p value, 95% CI, and means. f. interpret the CI, and use CI to make a conclusion (reject Ho or FTR Ho?). g. manually calculate the mean difference's 95% CI. h. manually calculate the t.statistic and p value.

```r
#library(readxl)
LCData <- read_excel("C:/Users/LCData.xlsx")
View(LCData)
#LCData$Smoke <- as.factor(LCData$Smoke)
#LCData$Smoke <- factor(LCData$Smoke, levels=c('no','yes'))
#boxplot(LCData$'Smoke==no',LCData$'Smoke==yes')

#library(ggplot2)
SmokeStatus <- LCData$'Smoke'
LungC <- LCData$'LC'
ggplot(LCData, aes(x = SmokeStatus, y = LungC)) + geom_boxplot()
```




```
#b.
##Ho:The lung capacity of non-smokers is the same as the lung capacity of
smokers, the difference = 0.
##Ha: The lung capacity of non-smokers is not equal to the lung capacity of
smokers.

nonSm <- LCData$'LC'[LCData$'Smoke' == c("no")]
Smokers <- LCData$'LC'[LCData$'Smoke' == c("yes")]
shapiro.test(nonSm)

##
##  Shapiro-Wilk normality test
##
## data:  nonSm
## W = 0.98857, p-value = 0.1507

#p-value = 0.1507
shapiro.test(Smokers)

##
##  Shapiro-Wilk normality test
##
## data:  Smokers
## W = 0.95242, p-value = 0.4642

#p-value = 0.4642
##Since the p-values for both datasets are greater than 0.05 it indicates
that both are normally distributed

#c.
#using a 2-sided test with alpha =0.05

#d
```

```r
t.test(nonSm,Smokers, mu = 0, alt= "two.sided", var.eq = T, conf = 0.95,
paired=F)
```

```
##
##  Two Sample t-test
##
## data:  nonSm and Smokers
## t = -0.32277, df = 198, p-value = 0.7472
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.515432  1.089135
## sample estimates:
## mean of x mean of y
##  8.111852  8.325000
```

```r
#p-value = 0.7472

#e.
#95 percent confidence interval:(-1.515432,  1.08913)
#mean of NonSm,Smokers = 8.111852,  8.325000
#t = -0.32277, df = 198, p-value = 0.7472

#f.
#since the conf. interval contains the tstat value, we fail reject the Ho
since we are 95% confident that the interval from -1.515432 to 1.08913
actually contains the true value of the difference (mu=0).

#g.
mean.diff <- mean(nonSm)-mean(Smokers)
#-0.2131
df <- length(nonSm)-1
tcrit <- qt(0.025,df,lower.tail = F)
#1.9732
SEM <- sqrt(sd(nonSm)^2/182 +sd(Smokers)^2/18)#0.525
ME <- tcrit*SEM
upper <- mean.diff+ME
lower <- mean.diff-ME
lower;upper
```

```
## [1] -1.249744
```

```
## [1] 0.8234477
```

```r
#h.
ttest.stat<- (mean.diff-0)/SEM #-0.4057
pvalue<- 2*pt(-0.4057, df=18, lower.tail=F)
#Since pvalue > 0.05, FTR Ho.
```