Data 602 - Assignment Three

Ensure you justify all computation and data visualizations with accompanying code.

1. The data set **NCBirths2004** consists of the Weight (in grams) of $n = 1009$ babies born in the state of North Carolina in 2004. All babies appearing in this sample had a gesteration periods of at least 37 weeks and were single births. Other variables in this data set include the Age of the birth mother, whether or not the birth mother was a Smoker during the gestation period, used Alcohol during the gestation period, the Gender of the baby. To access these data, install the `resampledata` package. For example,

```
install.packages("resampledata", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/camil/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'resampledata' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\camil\AppData\Local\Temp\RtmpovJWw7\downloaded_packages

library(resampledata)

##
## Attaching package: 'resampledata'

## The following object is masked from 'package:datasets':
##
##     Titanic
```

Note: This package has been installed in the R Studio “cloud” through the Data Science Hub. You will not be able to “install” this package on your version of R Studio through the datasciencehub.ucalgary.ca, as the packages you see in the “packages” pane are fixed for the moment.

```
head(NCBirths2004, 4)

##   ID MothersAge Tobacco Alcohol Gender Weight Gestation Smoker
## 1  1      30-34      No      No   Male   3827        40     No
## 2  2      30-34      No      No   Male   3629        38     No
## 3  3      35-39      No      No Female   3062        37     No
## 4  4      20-24      No      No Female   3430        39     No

library(dplyr)

##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order
to add
## additional features.  The original behavior of these functions should not
be affected by this.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```r
library(ggplot2)
library(resampledata)
library(boot)
```

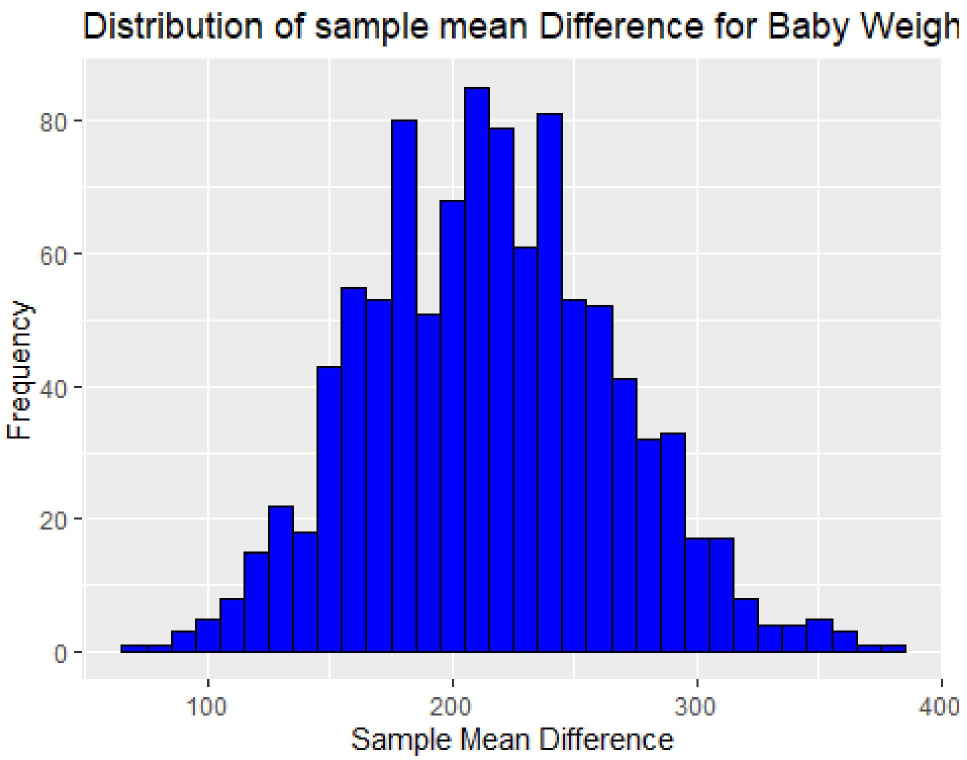
```
##
## Attaching package: 'boot'
## The following object is masked from 'package:mosaic':
##
##     logit
```

```
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
set.seed(435)

#(a) Create the bootstrap distribution for $\overline{X}_{NonSmoker} -
\overline{X}_{Smoker}$.

mydat <- data.frame(Weight = NCBirths2004$Weight, Smoker =
NCBirths2004$Smoker)
head(mydat, 5)
```

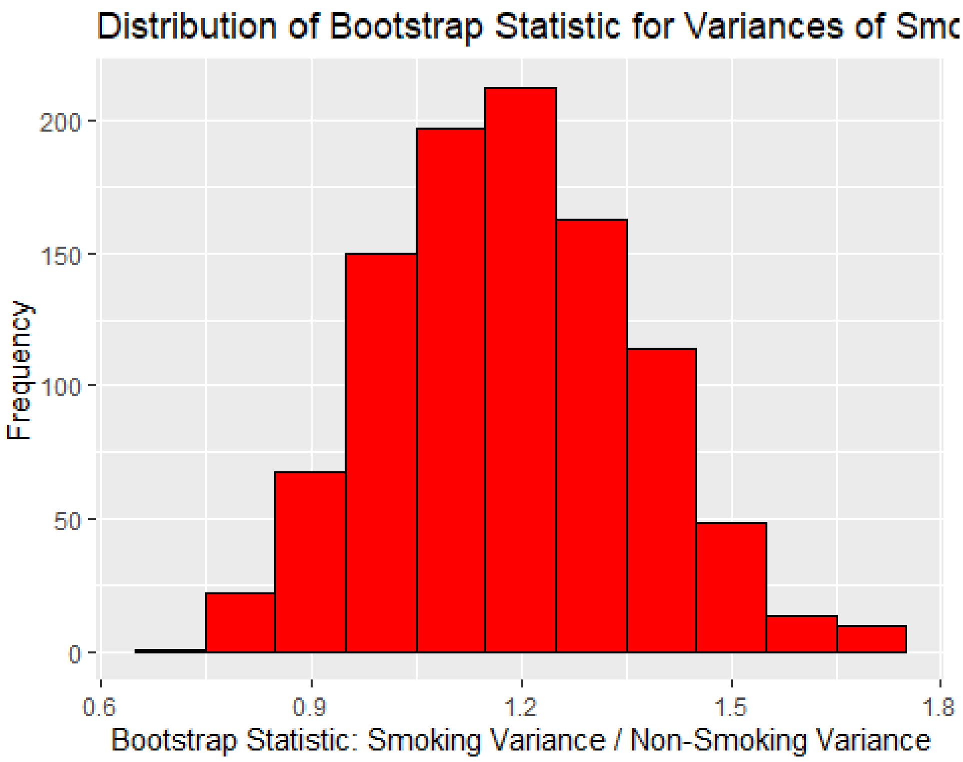
```
##   Weight Smoker
## 1   3827     No
## 2   3629     No
## 3   3062     No
## 4   3430     No
## 5   3827     No
```

```
non_smoking_weights <- filter(NCBirths2004, Smoker == "No")$Weight
smoking_weights <- filter(NCBirths2004, Smoker == "Yes")$Weight

non_smoking_weights <- do(1000) * mean(resample(non_smoking_weights, replace
= TRUE, na.rm = TRUE))
smoking_weights <- do(1000) * mean(resample(smoking_weights, replace = TRUE,
na.rm = TRUE))

mean_diff = abs(non_smoking_weights - smoking_weights)
ggplot(mean_diff, aes(x=mean)) +
  geom_histogram(col='black', fill='blue', binwidth=10, na.rm=TRUE) +
  xlab("Sample Mean Difference") + ylab("Frequency") +
  ggtitle("Distribution of sample mean Difference for Baby Weights of
Smoking/Non-smoking Mothers")
```



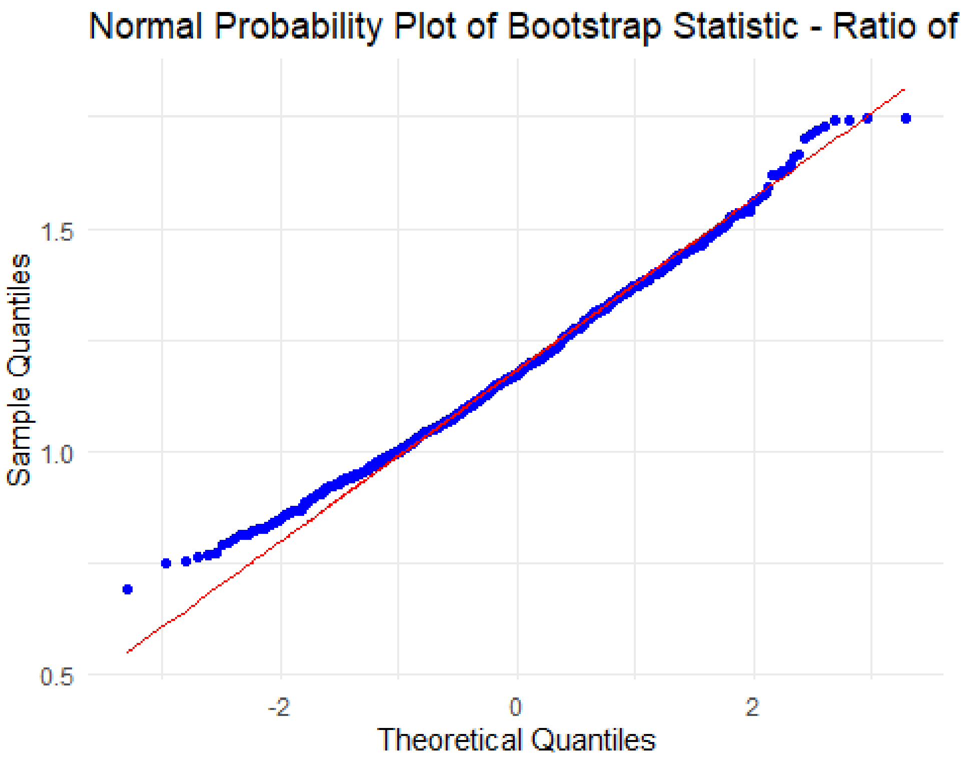
```
#(b) From your result in (a), compute the 95% confidence interval for
$\mu_{NonSmoker} - \mu_{Smoker}$.

#Confidence intervals
UL = quantile(mean_diff$mean, 0.975)
LL = quantile(mean_diff$mean, 0.025)
cat("95% bootstrap confidence interval for difference in
     mean is (", LL, ",", UL, ")")

## 95% bootstrap confidence interval for difference in
##      mean is ( 120.9777 , 317.4058 )

#(c) Compute the 95% confidence interval for $\mu_{NonSmoker} -
\mu_{Smoker}$, using the $t$-version.

mydat <- data.frame(Weight = NCBirths2004$Weight, Smoker =
NCBirths2004$Smoker)
t.test(~ Weight | Smoker, data = mydat, conf.level = 0.95, var.equal =
FALSE)$conf

## [1] 112.3161 317.6881
## attr(,"conf.level")
## [1] 0.95

#(d) Consider your result in both (b) and (c). What can you infer from these
data? Do children born to birth mother who did not smoke during pregnancy
weigh more on average than babies born to birth mothers who did smoke during
```

```
pregnancy?


#The Bootstrap Confidence Interval, (120.9777, 317.4058), indicates that we
can be 95% confident that the true difference in mean birth weights between
non-smokers and smokers falls within this range. The T-Version Confidence
Interval, (112.3161, 317.6881), has a slightly broader range. Importantly,
both intervals do not include zero, supporting the conclusion of a
statistically significant difference. Together, these analyses suggest that
children born to birth mothers who did not smoke during pregnancy tend to
weigh more on average.

#**2.** Refer to Question 1.

#(a) Create a distribution of the bootstrap statistic
$\frac{S_{Smoker}}{S_{NonSmoker}}$. USe 1000 as the number of
iterations/replications and provide a visualization of this distribution.

non_smoking_weights <- filter(NCBirths2004, Smoker == "No")$Weight
smoking_weights <- filter(NCBirths2004, Smoker == "Yes")$Weight

bootstrap_ratios <- replicate(1000, {
  non_smoking_var <- var(resample(non_smoking_weights, replace = TRUE, na.rm
= TRUE))
  smoking_var <- var(resample(smoking_weights, replace = TRUE, na.rm = TRUE))
  return(smoking_var / non_smoking_var)
})

ggplot(data.frame(V1 = bootstrap_ratios), aes(x = V1)) +
  geom_histogram(col = 'black', fill = 'red', binwidth = 0.1, na.rm = TRUE) +
  xlab("Bootstrap Statistic: Smoking Variance / Non-Smoking Variance") +
ylab("Frequency") +
  ggtitle("Distribution of Bootstrap Statistic for Variances of Smoker and
Non-Smoker Birth Weights")
```



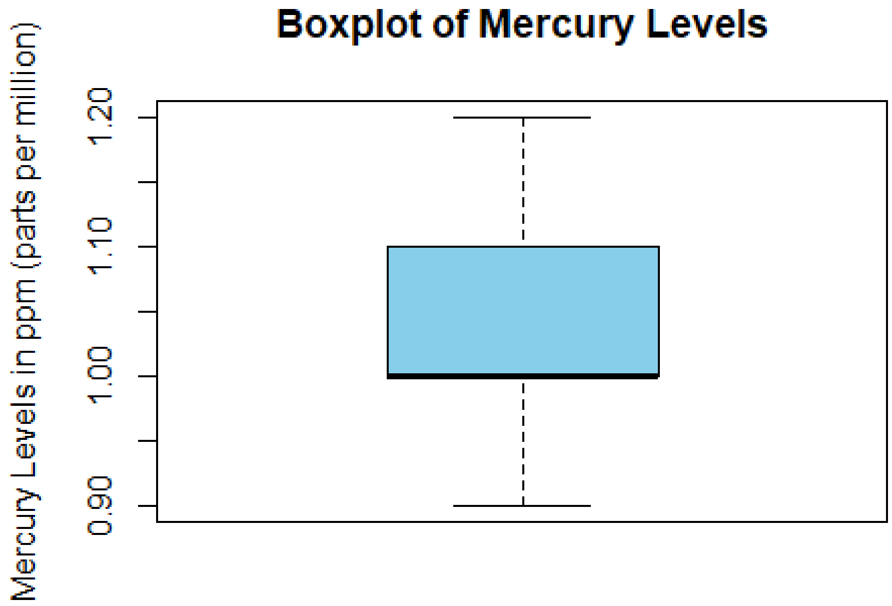
```
#(b) Create a Normal Probablity Plot of this bootstrap statistic. Does the
ratio of the sample standard deviations appear to follow a Normal
distribution? Explain.

ggplot(data.frame(V1 = bootstrap_ratios), aes(sample = V1)) +
  stat_qq(col = 'blue') +
  stat_qqline(col = 'red') +
  ggtitle("Normal Probability Plot of Bootstrap Statistic - Ratio of Standard
Deviations") +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles") +
  theme_minimal()
```

```
## Warning: The following aesthetics were dropped during statistical
transformation: sample
## ℹ This can happen when ggplot fails to infer the correct grouping
structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



```
#In a QQ plot, if the points fall approximately along a straight line, it
suggests that the data follows a normal distribution. In our case, this
appears to be the case, signifying a normal distribution.


#(c) Compute the 95% bootstrap percentile interval for
$\frac{\sigma_{Smoker}}{\sigma_{NonSmoker}}$.

bootstrap_interval <- quantile(bootstrap_ratios, c(0.025, 0.975))
cat("95% Bootstrap Percentile Interval for the ratio of standard deviations:
(", round(bootstrap_interval[1], 4), ",", round(bootstrap_interval[2], 4),
")\n")

## 95% Bootstrap Percentile Interval for the ratio of standard deviations: (
0.861 , 1.5393 )

#(d) Consider the result you obtained in part (c). Explain the practical
meaning of this result with respect to the variable **Weight**.


#95% Bootstrap Percentile Interval for the ratio of standard deviations is
(0.861, 1.5393) As #this interval includes 1, then variability in birth
weights between babies born to smoking #mothers and non-smoking mothers is
relatively similar.

#**3.**  Health Canada sets an action level for mercury in fish at 1 ppm
(part per million). If mercury levels are higher than this value, then this
```

```
value in commercial fish then Health Canada will take action to impose a
moritorium on fishing in the area where the fish are harvested. Recently,
there have been concerns about mercury levels in walleye fish populating the
portion of the Athabasca River that is down stream from Whitecourt, where
local First Nations harvest walleye as part of a commercial fishing
operation. A biologist randomly picked $n = 31$ walleye from a recent
commercial fishing catch downstream from Whitecourt, and measured the mercury
(in ppm) from each walleye. The ppms, are provided below.


#(a) Establish a statistical hypothesis that allows the biologist to see if
mercury levels in walleye fish harvested from the Athabasca River (downstream
of Whitecourt) exceed Heath Canada's action level.

#Null hypothesis: Mercury levels in walleye fish harvested from the Athabaska
River is equal to the Health Canada's action level
#Alternative hypothesis: Mercury levels in walleye fish harvested from the
Athabaska River exceed Health Canada's action level
#H0: μ ≤ 1
#H1: μ > 1


#(b) Refer to your hypotheses in (a). In the context of your statistical
hypotheses in part #(a), explain *both* a Type I Error and a Type II Error.

#Type I Error (False Positive): Falsely concluding there is a problem when
there isn't. This would involve incorrectly asserting that mercury levels in
the fish exceed Health Canada's action levels when, in reality, they are at
or below 1 ppm.
#Type II Error (False Negative): Failing to detect a problem that actually
exists. This would entail not recognizing that the average mercury level is
higher than 1 ppm when it is, indeed, above the specified threshold.

#(c) Visualize these data with either a violin plot or a boxplot, and comment
on the disribution of mercury levels on walleye harvested from the Athabaska
River downstream from Whitecourt.

m_levels <- c(1.2, 1.1, 1.0, 1.0, 1.1, 1.0, 1.0, 1.0, 0.9, 1.1, 1.1, 1.2,
1.0, 1.1, 1.0, 1.1, 1.0, 0.9, 1.0, 1.0, 1.1, 1.0, 1.0, 1.1, 1.2, 1.0, 1.1,
1.0, 1.0, 1.2, 1.1)

boxplot(m_levels,
        main = "Boxplot of Mercury Levels",
        ylab = "Mercury Levels in ppm (parts per million)",
        col = "skyblue",
        border = "black",
        horizontal = FALSE,
        names = c("Mercury Levels"))
```



```
#(d) Do these data suggest that Health Canada should place a moritorium on
commercial walleye fishing on the Athabaska River downstream of Whitecourt?
In your finding, interpret the meaning of the $P$-value you have computed.
*IF* you reject the null hypothesis, provide a 95% confidence interval for
the mean mercury (in ppm) of walleye found downstream from Whitecourt.

hypothesized_mean <- 1.0
t_test <- t.test(m_levels, mu = hypothesized_mean, alternative = "greater")
p_value_t_test <- t_test$p.value
cat("T-Test p-value:", p_value_t_test, "\n")

## T-Test p-value: 0.0006595483

conf_interval <- t.test(m_levels)$conf.int
cat("95% Confidence Interval:", conf_interval, "\n")

## 95% Confidence Interval: 1.021857 1.081368

#We take on a one-sided approach to calculate the T-Test p-value. If the
obtained p-value is below the conventional significance level of 0.05, we
would reject the null hypothesis. In this context, rejection implies that
mercury levels are likely higher than the specified action level. The
computed T-Test p-value in our study is 0.0006595483, providing substantial
support for this conclusion. Additionally, the 95% confidence interval,
[1.021857, 1.081368], indicates with 95% confidence that the true mean
mercury level in walleye fish from this location falls within this range.
Both statistical tests provide strong evidence of elevated mercury levels.
```

```
#**4.** Coffee markets that conform to organic standards focus on the
environmental aspect of coffee growing, such as the use of shade trees, and
reduced reliance on herbicides and pesticides. Researchers investigating
organic coffee growers in Southern Mexico took a representative, random
sample of $n = 845$ coffee growers, of which 475 were certified to sell
organic coffee and 75 were transitioning to sell organic coffee.

#*In the United States*, 60% of all coffee growers are organically certified.
Is there ample statistical evidence to confirm that the proportion of
certified coffee growers in Southern Mexico who are either certified or in
the process of being certified, is more than 60%?
#Ensure you completely justify your answer, using method(s) covered in DATA
602.



#H0: p=0.60 (proportion is equal to 60%)
#H1: p>0.60 (proportion is more than 60%)

#Total coffee growers sampled  N = 847
#Number of growers certified organic = 475
#Number of growers transitioning to organic = 75

#Calculate P
p <- (475 + 75) / 845  #Compute proportion of growers
print(p)
```

```
## [1] 0.6508876
```

```
p <- 0.6508876
sample_size <- 847
null_hypothesis_p <- 0.60

# Calculate the z-test statistic
z_test <- (p - null_hypothesis_p) / sqrt((null_hypothesis_p * (1 -
null_hypothesis_p)) / sample_size)
print(z_test)
```

```
## [1] 3.023069
```

```
# Significance level
alpha <- 0.05
# Calculate the critical z-value for a right-tailed test
critical_z_value <- qnorm(1 - alpha)
# Print the critical z-value
print(critical_z_value)
```

```
## [1] 1.644854
```

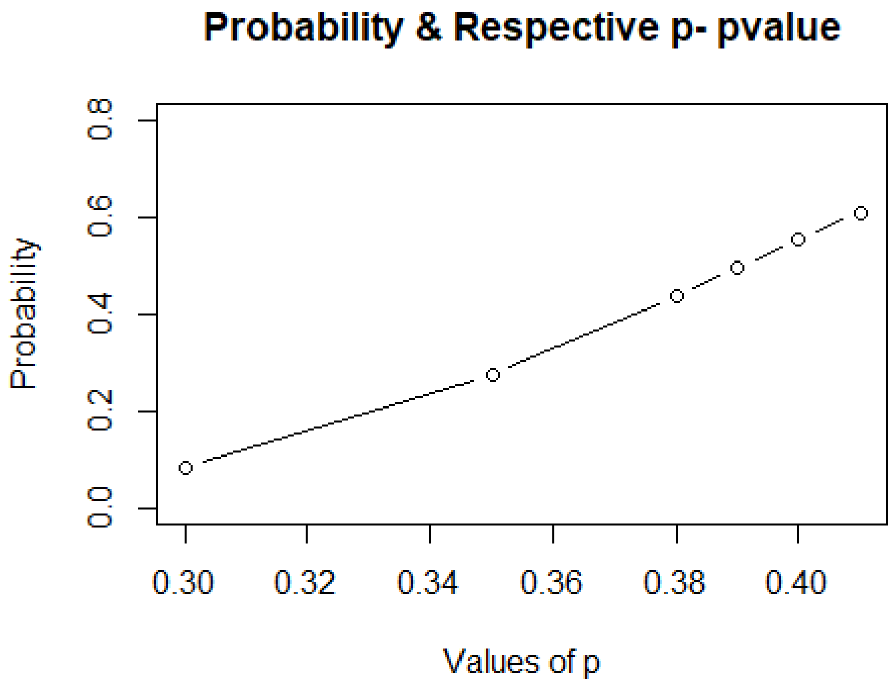
```
#Conclusion: Calculated z-score (3.023069) is greater than the critical z-
value (1.644854). In this case we would reject the null hypothesis. This
```

```
suggests that the observed proportion of certified or transitioning coffee
growers in Southern Mexico is statistically significantly greater than 60%.

#**5.** As a budding data scientist with much promise, a person who is
considering running as a Member of Parliament (MP) for a certain riding hires
you to conduct some polling. Due to the time investment and the cost (time
and finances) of a political campaign, you decide to take a random sample of
$n = 50$ voters who live within this particular riding. Each are to be asked
"if they would support this particular candidate if they ran as a
reprsentative for Party X in the next federal election".  If your
polling/sampling suggests that they will receive at least 45% of the vote,
then you will council this person to "run for office". In your preliminary
statistical work, you have decided that there is enough statistical evidence
to support the "mimimum of 45%"-claim if out of $n = 50$ randomly chosen
voters, at least 20 indicate they will vote for this candidate if they run.

#(a) State the statistical hypotheses.

#The candidate will not receive at least 45% of the vote if they run for
office.
#H0:p<0.45

#The candidate will receive at least 45% of the vote if they run for office.
#H1:p ≥ 0.45

#(b) Compute the value of $\alpha$ used in your derivation of the decision
rule.

sample_size <- 50
min_support <- 20
alpha <- 0.05
critical_value <- qbinom(1 - alpha, size = sample_size, prob = 0.45)
print(critical_value)

## [1] 28

#(c) What if the candidate were to receive 42% of the vote. Compute the
probability that you will conclude they should run for office. Interpret the
meaning of this probability.

p_42 <- 0.42
# Probability of observing at least 20 supporters
p_value <- pbinom(min_support - 1, size = sample_size, prob = p_42,
lower.tail = FALSE)
print(p_value)

## [1] 0.663807

#The calculated p-value is 0.663807, which exceeds the significance level
(alpha) of 0.05. Consequently, we fail to reject the null hypothesis. There
```

```
is insufficient evidence to support the candidate in running for office. As
determined earlier, a critical value of 28 indicates that one would only
recommend the candidate to run if there are 28 or more supporting voters.


#(d) Repeat for (c) for these values of $p$: $p = 0.41, p = 0.40, p = 0.39, p
= 0.38, p = 0.35$ and $p = 0.30$. For each differing value of $p$, compute
the probability computed in part (c). THEN, create a plot with the differing
values of $p$ on the $x$-axis and the probabilties computed on the $y$-axis.


sample_size <- 50
min_support <- 20
alpha <- 0.05
p_values <- c(0.41, 0.40, 0.39, 0.38, 0.35, 0.30)
probabilities <- numeric(length(p_values))
for (i in seq_along(p_values)) {
  # Calculate probability
  probabilities[i] <- pbinom(min_support - 1, size = sample_size, prob =
p_values[i], lower.tail = FALSE)
}

results <- data.frame(p = p_values, probability = probabilities)
print(results)
```

```
##      p probability
## 1 0.41   0.6099048
## 2 0.40   0.5535236
## 3 0.39   0.4957191
## 4 0.38   0.4376490
## 5 0.35   0.2735637
## 6 0.30   0.0848026
```

```
plot(p_values, probabilities, type = "b", main = "Probability & Respective p-
pvalue",
     xlab = "Values of p", ylab = "Probability",
     ylim = c(0, 0.8))

points(0.45, alpha, col = "red", pch = 16)
text(0.45, alpha, "  Critical Value", pos = 2)
```



```
#(e) What does your plot/graph in part (d) tell you about your statistical
test? How can you improve your test? Provide some suggestion(s), reasoning
why each would make your statisical test better.

#From this plot we see the value of p decreases, the probability of observing
at least 20 supporters (the minimum for advising to run) also decreases. This
test might be improved my increasing the sample size which could lead to a
more precise estimate. We could also adjust the significance test depending
on the tolerance for type errors

#**6.** In 2012, an Angus Reid[^2] poll surveyed $n = 1010$ randomly chosen
Canadians from which 601 supported a ban on singe-use plastics. A more recent
survey in 2019 of $n= 1000$ Canadians[^3] found that 561 supported a ban on
single-use plastics.


#(a) Compute *a* 95% confidence interval for $p_{2019} - p_{2012}$, the
difference between the proportion of Canadians who currently support a ban on
single-use plastics and the proportion of Canadians who supported such a ban
in 2012.

#Calculate proportions
p2012 <- 601/1010
p2019 <- 561/1000
se <- sqrt((p2012*(1-p2012))/1010 + (p2019*(1-p2019))/1000)
```

```
confidence_int_95 <- p2019 - p2012 + c(-1.96,1.96)*se
print(confidence_int_95)

## [1] -0.077207743  0.009108733

#(b) From your result in (a), can you infer there is a statistically
significant difference between $p_{2019}$ and $p_{2012}$. Why or why not?

#The 95% confidence interval suggests that there is no statistically
significant distinction between P2019 and P2012. This is evidenced by the
interval encompassing 0, signifying that the observed difference in
proportions lacks statistical significance.




#Q7. What do these data suggest? Ensure you address any
conditions/assumptions you have made about these data. Also ensure you
provide the $P$-value and interpret its meaning in the context of these data.


cereal_w <- c(497.2, 499.9, 495.8, 514.2, 490.0, 498.3, 495.1, 486.7)
mean_cereal_weight <- mean(cereal_w)
print(mean_cereal_weight)

## [1] 497.15

sd_cereal_weight <- sd(cereal_w)
print(sd_cereal_weight)

## [1] 8.158606

#Comparing the stated weight vs the mean weight:
#Stated weight: 500grams
#Calculated Mean Weight: 497.15grams

#Based off this above calculation, we can see Usman is not getting the amount
of cereal stated on the box.

#Conditions/Assumptions

#1. The weights of cereal in these boxes are treated as independent
variables.
#2. 8 boxes represent a sample and not a population
#3. Assuming that the weights of cereal in each box are normally distributed

#H0:u=500
#H0:u≠500

observed_weights <- c(497.2, 499.9, 495.8, 514.2, 490.0, 498.3, 495.1, 486.7)
```

```
stated_weight <- 500
t_test <- t.test(observed_weights, mu = stated_weight, alternative =
"two.sided")
print(t_test$p.value)
```

```
## [1] 0.3560478
```

```
#As this is a two-tailed test, we compare the p-value to the 0.025
significance level (α/2). With a p-value of 0.3560478, we observe that it is
greater than both 0.025 and 0.05. Therefore, we do not reject the null
hypothesis. There is not sufficient evidence to claim that the mean weight of
cereal is significantly different from the stated weight of 500 grams based
on this sample.
```