MATH 254 - Statistical Modeling and Applications - Lab 2

Tural Sadigov

NAME: ______________

Who are you working with? ______________

In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. In this mini-project, we will investigate a random sample of 1000 cases from this data using various tools we have been learning from both descriptive and inferential statistics. The data is from Textbook 1 (https://www.openintro.org/data/index.php?data=ncbirths) and available on GitHub (https://github.com/turalsadigov/MATH_254/blob/main/data/ncbirths.csv).

```r
library(tidyverse)

url_remote  <- "https://raw.githubusercontent.com/"
path_github <- "turalsadigov/MATH_254/main/data/"
file_name   <- "ncbirths.csv"
url = paste0(url_remote, path_github, file_name)
nc <- read_csv(url)
nc %>%
  head(5)
```

```
# A tibble: 5 x 13
   fage  mage mature    weeks premie visits marital gained weight lowbi~1 gender
  <dbl> <dbl> <chr>     <dbl> <chr>   <dbl> <chr>    <dbl>  <dbl> <chr>   <chr>
1    NA    13 younger ~    39 full ~     10 not ma~     38   7.63 not low male
2    NA    14 younger ~    42 full ~     15 not ma~     20   7.88 not low male
3    19    15 younger ~    37 full ~     11 not ma~     38   6.63 not low female
```

```
4    21    15 younger ~    41 full ~      6 not ma~     34   8    not low male
5    NA    15 younger ~    39 full ~      9 not ma~     27   6.38 not low female
# ... with 2 more variables: habit <chr>, whitemom <chr>, and abbreviated
#   variable name 1: lowbirthweight
# i Use `colnames()` to see all variable names
```

Descriptive statistics

1. What are the cases? Categorize variables as numerical variable (discrete or continuous) or categorical variable (regular or ordinal). Mention the unit of each numerical variable. Using command of the data frame in R, also mention the ranges of numerical variables and levels of categorical variables. Example is provided.

   ```
   ... %>%
     glimpse()
   ```

   - gender: A regular categorical variable with two levels: female and male

2. Let's look at weights of the babies with different genders and compare them. To get the weights of female babies, one can filter the data with the verb `filter` as below.

   ```
   nc %>%
     filter(gender == 'female') %>%
     select(weight) %>%
     head(5)
   ```

   ```
   # A tibble: 5 x 1
     weight
      <dbl>
   1   6.63
   2   6.38
   3   6.94
   4   8.81
   5   6
   ```

   ```
   nc %>%
     filter(gender == 'female') %>%
     select(weight) %>%
     summary()
   ```

   ```
        weight
    Min.   : 1.000
   ```

```
1st Qu.: 6.250
Median : 7.130
Mean   : 6.903
3rd Qu.: 7.750
Max.   :11.630
```

Obtain the male weights by modifying the code above in the R chunk below. Create side by side boxplots, and comment on the distribution. Note that, every time you create some graphical summary, you need to have correct labels (with units), titles, etc.

```
# extract weights of male babies (add your code below)
...

# get counts for each gender (add your code below)
...

# obtain side by side boxplots (add your code below)
...
```

3. Now, let's check if it is plausible that the weights in each category (female, male) are sampled from some hypothetical normal distributions. There are many ways of doing this, but one way is to produce Normal QQ plots which also known as Quantile-Quantile plots. Here, we expect to get a linear pattern if the sample is actually from Normal distribution. If the linearity is not plausible, then one could conclude that normality is not plausible. To get Normal QQ-plot, one can use base R function `qqnorm` or use function `stat_qq` from ggplot. Create side by side QQ plots for the weights of female and male babies. Comment on the scatterplots and poatterns you get. Do you believe that weights of each group are 'sampled' from normal distributions? Why or why not?

```
# edit code below
nc %>%
  ggplot(aes(sample = ...)) +
  stat_qq() +
  stat_qq_line()+
  facet_wrap(~...)
```

4. Do you think that each sample could be considered as simple random sample? Do you think the samples are independent from each other? Why or why not?

Inferential Statistics

Assuming the independence within the sample, and independence of samples from each other, and since each data is large enough (sample sizes are greater than 40), we could go ahead with the machinery needed for confidence intervals and hypothesis testing.

Confidence Intervals

5. Define parameter of interest (true population means) in the context of the problem. Create 99% confidence interval for difference of true means. Interpret the confidence interval. Is it still possible that, the confidence interval you created does not capture difference of true means (or true difference of population means)? Looking at your confidence interval, is it plausible that there is no difference in true mean weights? Why or why not?

```
# load library
library(infer)

# calculate and assign observed difference (edit code below)
observed_difference <-
  nc %>%
  specify(weight ~ gender) %>%
  calculate(stat = '...')

# print the observed difference
observed_difference

# create and assign theoretical sampling distribution of the difference of sample mea
diff_dist <-
  nc %>%
  specify(weight ~ gender) %>%
  assume(...)

# display the distribution
diff_dist

# create and assign confidence interval (edit code below)
ci <-
  diff_dist %>%
  get_confidence_interval(point_estimate = observed_difference,
                          # at the 99% confidence level
                          level = ...,
                          # using the standard error
                          type = "se")
```

```
# print confidence interval
ci

# shade the confidence interval (edit code below)
diff_dist %>%
  visualize() +
  shade_...(endpoints = ci)
```

Hypothesis testing

6. Let's test if there is a difference between true means (or equivalently, lets check if weights of the babies depends on the gender). State null and alternative hypothesis in the context of the problem and choose significance level.

   H_0 :

   H_a :

   $\alpha = ?$

Since we have checked the conditions for two-sample t test, we go ahead and find test statistic and corresponding p-value (in two different ways – 'manually' in 7 and 8, and using `infer` in 9 below).

7. Calculate

   - Sample mean of female baby weights (let's call it $\bar{x}$)
   - Sample mean of male baby weights (let's call it $\bar{y}$)
   - Difference of sample means ($\bar{x} - \bar{y}$)
   - Sample standard deviation of female baby weights (let's call it s_x)
   - Sample standard deviation of male baby weights (let's call it s_y)
   - Sample size of female baby weights (let's call it n)
   - Sample size of male baby weights (let's call it m)
   - Approximate standard error (deviation) of difference of sample means ($\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$)

```
# obtain summary statistics: mean, standard deviation and count for each group (edit
sum_stats <-
  nc %>%
  group_by(gender) %>%
  summarise(ave = ...,
            st_dev = ...,
            counts = ...)
```

```
sum_stats

x_bar = sum_stats[1,2]
x_bar
y_bar = ...
y_bar
s_x = ...
s_x
s_y = ...
s_y
n = ...
n
m = ...
m
se_difference = sqrt(...)
se_difference
```

8. Using your results from part (7) above, calculate the test statistic:

$$t = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

Calculate approximate degrees of freedom $df = n + m - 2$, and p-value, $2Pr[t_{df} < t]$.

```
# edit code below
df = ...
df
t = ...
t
p_value = 2*pt(t[1,1], df[1,1])
p_value
```

9. Use the following command to find the test statistic, degrees of freedom and p-value again. Check if they matched with above. Interpret the p-value (very carefully).

```
# calculate and assign observed t statistic (also test statistic) (edit code below)
observed_t_stat <-
  nc %>%
  specify(...) %>%
  hypothesise(null = ...) %>%
  calculate(stat = ...)
```

```
# print the observed t statistic
observed_t_stat

# create and assign theoretical sampling distribution of the difference of sample mea
diff_dist <-
  nc %>%
  specify(...) %>%
  assume(...)

# display the distribution
diff_dist

# get p-value (edit code below)
diff_dist %>%
  get_p_value(obs_stat = ...,
              direction = 'two-sided')

# shade p-value (edit code below)
diff_dist %>%
  visualize() +
  shade_p_value(obs_stat = ...,
                direction = ...)
```

10. Make a decision in the context of the problem according to the significance level you had chosen above. Would it be still possible that you made an error in the decision? If yes, what kind of error have you made? If an error has been made, then what could be the reason for that?