COSC6323_Exercise_2

Rajat Gupta

2/06/2021

Loading the Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 3.6.3

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'readr' was built under R version 3.6.3

## Warning: package 'purrr' was built under R version 3.6.3

## Warning: package 'dplyr' was built under R version 3.6.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Loading the dataset and renaming the columns

```
btwt<-as_tibble(MASS::birthwt)
colnames(btwt) <- c("birthwt.below.2500",
                    "mother.age",
                    "mother.weight",
                    "race",
                    "mother.smokes",
                    "previous.prem.labor",
                    "hypertension",
                    "uterine.irr",
                    "physician.visits",
                    "birthwt.grams")
```

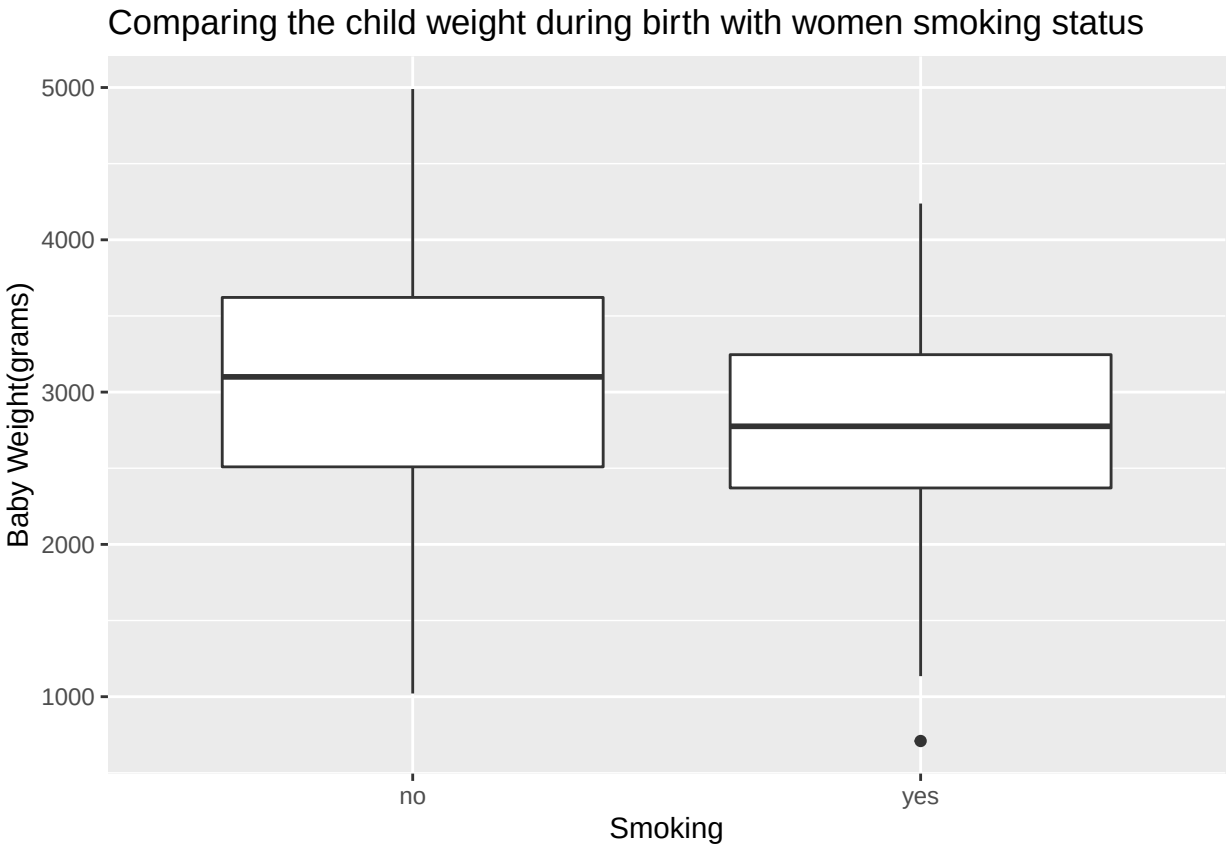
Changing into Categorical factors

```
btwt <- btwt %>%
  mutate(race = recode_factor(race, `1` = "white", `2` = "black", `3` = "other")) %>%
  mutate_at(c("mother.smokes", "hypertension", "uterine.irr", "birthwt.below.2500"),
            ~ recode_factor(.x, `0` = "no", `1` = "yes"))
```

1.Compare infants birth weight between smoking and non-smoking mothers

Creating a box plot to futher analyse the features

```
ggplot(btwt,aes(x=mother.smokes,y=birthwt.grams))+geom_boxplot()+labs(title ="Comparing the child weigh
       y="Baby Weight(grams)",x="Smoking")
```



From the above boxplot we can say that the median weight of non smoking is higher than 3000 and that of smoking is less than 3000. Also the boxplot of non smoking is more dispersed than the smoking. Now we have to check weather the difference in mean of the non smoking and smoking is statistically significant or not.

Summary wrt smoking

```
btwt %>%
  group_by(mother.smokes) %>%
  summarize(num.obs = n(),
            mean.birthwt = round(mean(birthwt.grams), 0),
            sd.birthwt = round(sd(birthwt.grams), 0),
            se.birthwt = round(sd(birthwt.grams) / sqrt(num.obs), 0))
```

```
## # A tibble: 2 x 5
##   mother.smokes num.obs mean.birthwt sd.birthwt se.birthwt
## * <fct>           <int>        <dbl>      <dbl>      <dbl>
## 1 no                115         3056        753         70
## 2 yes                74         2772        660         77
```

Performing t test

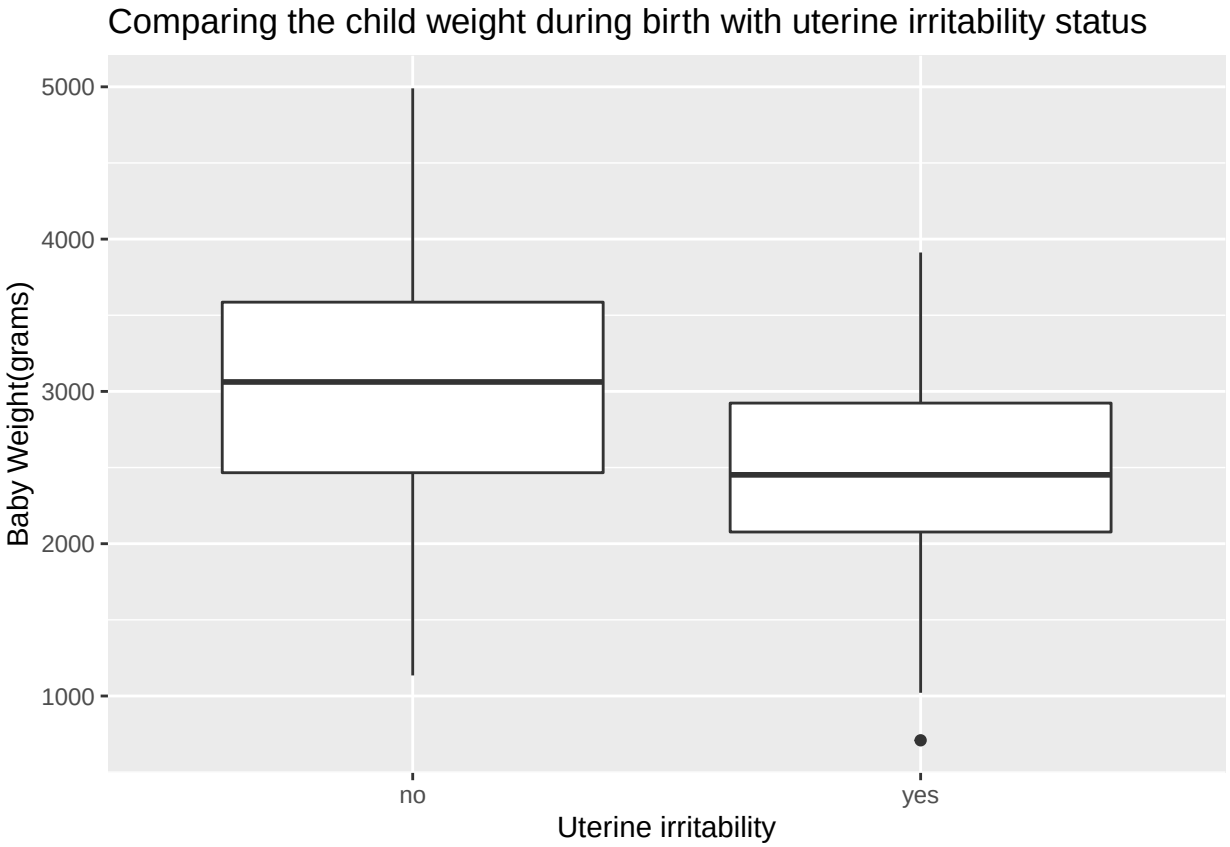
```
t.test(birthwt.grams ~ mother.smokes, data = btwt)
```

```
##
##  Welch Two Sample t-test
##
## data:  birthwt.grams by mother.smokes
## t = 2.7299, df = 170.1, p-value = 0.007003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   78.57486 488.97860
## sample estimates:
##  mean in group no mean in group yes
##          3055.696          2771.919
```

From the study we find that average birth weigh of baby for non smoking mother's is 3056 gms and for smoking mother's it is 2772 gms. The hypotheses test p-value of 2-sample t test is around 0.007 which is less than 0.05 significance level. Hence we can reject the null hypotheses and state that their is a significant difference between the means. The value of statistics is 2.7299 and 95% confidence interval is (78.575,488.979).

2.compare infants birth weight between agroup of mothers with a presence of uterine irritability and a group of mothers without it

```
ggplot(btwt,aes(x=uterine.irr,y=birthwt.grams))+geom_boxplot()+labs(title ="Comparing the child weight
       y="Baby Weight(grams)",x="Uterine irritability")
```



From the above boxplot we can see that the median baby weight in non uterine irritability is above 3000 and in case of uterine irritability is around 2500. Also in case of non uterine irritability the data dispersion is more when compare to the uterine irritability. Now using hypotheses tests we have to check weather their is a difference in mean is statically significant or not.

Summary wrt uterine irritability

```
btwt %>%
  group_by(uterine.irr) %>%
  summarize(num.obs = n(),
            mean.birthwt = round(mean(birthwt.grams), 0),
            sd.birthwt = round(sd(birthwt.grams), 0),
            se.birthwt = round(sd(birthwt.grams) / sqrt(num.obs), 0))
```

```
## # A tibble: 2 x 5
##   uterine.irr num.obs mean.birthwt sd.birthwt se.birthwt
## * <fct>         <int>        <dbl>      <dbl>      <dbl>
## 1 no              161         3031        694         55
## 2 yes              28         2449        742        140
```

Performing t test

```
t.test(birthwt.grams ~ uterine.irr, data = btwt)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  birthwt.grams by uterine.irr
## t = 3.8615, df = 35.696, p-value = 0.000455
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  275.8913 886.6553
## sample estimates:
##  mean in group no mean in group yes
##          3030.702          2449.429
```

From the study we find that average birth weigh of baby for no uterine irritability mother's is 3031 gms and for uterine irritability mother's it is 2449 gms. Also from the above hypotheses test we can say that p-value of t-test is around 0.0004 which is less than the 0.05 significance level. Hence we can say that we reject the null hypotheses and their is a significant difference in the mean of the birth weight wrt uterine irritability. The value of t-statistics is 3.8615 and the confidence interval is (275.89,886.65).