BALANGA Q2

2022-06-08

#setwd("~/Desktop/PH23- BIOSTATISTIC/report/BIOS Report/R markdown files")

library(readxl)  
library(data.table)  
library(knitr)  
library(tidyverse)  
library(stringr)  
library(flextable)  
  
neonatal <- read\_excel("neonatal.xlsx")  
format(head(neonatal))

## [1] "# A tibble: 6 × 7"   
## [2] " studyid bmicat birthweight lga gestationalage hypoglycaemia admission"  
## [3] " <chr> <chr> <dbl> <chr> <dbl> <chr> <chr> "  
## [4] "1 B4001 30-34.9(obes… 3420 norm… 40.2 No No "  
## [5] "2 B4002 30-34.9(obes… 2955 norm… 38.1 No No "  
## [6] "3 B4003 35-39.9 (sev… 5054 LGA 42 No No "  
## [7] "4 B4004 30-34.9(obes… 3455 norm… 39.6 No No "  
## [8] "5 B4005 35-39.9 (sev… 4510 LGA 39.6 No No "  
## [9] "6 B4006 30-34.9(obes… 3210 norm… 37.3 No Yes "

set\_flextable\_defaults(  
 font.size = 10, font.family = "Helvetica",  
 font.color = "#333333",  
 table.layout = "fixed",  
 border.color = "gray",  
 padding.top = 3, padding.bottom = 3,  
 padding.left = 4, padding.right = 4)

## Loading dataset

number of observations are given by which is total size of the population

nrow(neonatal)

## [1] 2466

## Knowing the data types in the dataframe

format(glimpse(neonatal))

## Rows: 2,466  
## Columns: 7  
## $ studyid <chr> "B4001", "B4002", "B4003", "B4004", "B4005", "B4006", "…  
## $ bmicat <chr> "30-34.9(obese)", "30-34.9(obese)", "35-39.9 (severly o…  
## $ birthweight <dbl> 3420, 2955, 5054, 3455, 4510, 3210, 3385, 3110, 3320, 3…  
## $ lga <chr> "normalsize", "normalsize", "LGA", "normalsize", "LGA",…  
## $ gestationalage <dbl> 40.2, 38.1, 42.0, 39.6, 39.6, 37.3, 40.0, 37.2, 40.5, 4…  
## $ hypoglycaemia <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "…  
## $ admission <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", …

## [1] "# A tibble: 2,466 × 7"   
## [2] " studyid bmicat birthweight lga gestationalage hypoglycaemia admission"  
## [3] " <chr> <chr> <dbl> <chr> <dbl> <chr> <chr> "  
## [4] " 1 B4001 30-34.9(obe… 3420 norm… 40.2 No No "  
## [5] " 2 B4002 30-34.9(obe… 2955 norm… 38.1 No No "  
## [6] " 3 B4003 35-39.9 (se… 5054 LGA 42 No No "  
## [7] " 4 B4004 30-34.9(obe… 3455 norm… 39.6 No No "  
## [8] " 5 B4005 35-39.9 (se… 4510 LGA 39.6 No No "  
## [9] " 6 B4006 30-34.9(obe… 3210 norm… 37.3 No Yes "  
## [10] " 7 B4007 30-34.9(obe… 3385 norm… 40 No No "  
## [11] " 8 B4008 30-34.9(obe… 3110 norm… 37.2 No No "  
## [12] " 9 B4009 30-34.9(obe… 3320 norm… 40.5 No No "  
## [13] "10 B4010 30-34.9(obe… 3418 norm… 40.5 No No "  
## [14] "# … with 2,456 more rows"

so all the columns are character variables except the birthweight We need to characterize the data accordinf to three groups in bmicat which can be done bu diving the bmi column and filtering rows with condition.

## identifying the unique values in the bmicat column

unique(neonatal$bmicat)

## [1] "30-34.9(obese)" "35-39.9 (severly obese)"  
## [3] ">=40 (morbidly obese)"

There are three unique values which means we have 3 groups in the bmicat column. Now we are going to classify the three groups.

## Classifying in 3 groups

obese <- neonatal %>% filter(bmicat %in% c("30-34.9(obese)"))  
  
severely\_obese <- neonatal %>% filter(bmicat %in% c("35-39.9 (severly obese)"))  
  
morbidly\_obese <- neonatal %>% filter(bmicat %in% c(">=40 (morbidly obese)"))

## Finding mean and standard deviation of the each group alongwith overall population

For the class of obese patients, we need to find the number of patients who were admitted to the NICU at birth and their mean and standard deviation. In the below chunk we are calculating the median and the IQR of the birthweight of the class of all 3 groups of patients as well as the mean median of all groups in the form of a table

#  
  
df1 <- data.frame(  
 "Groups" = c("Obese", "Severly Obese", "Morbidly Obese","All"),  
 c(mean(obese$birthweight),mean(severely\_obese$birthweight),mean(morbidly\_obese$birthweight),mean(neonatal$birthweight)),  
 c(sd(obese$birthweight),sd(severely\_obese$birthweight),sd(morbidly\_obese$birthweight),sd(neonatal$birthweight)),   
 c(IQR(obese$gestationalage), IQR(severely\_obese$gestationalage), IQR(morbidly\_obese$gestationalage),IQR(neonatal$gestationalage)),  
 c(median(obese$gestationalage), median(severely\_obese$gestationalage), median(morbidly\_obese$gestationalage),median(neonatal$gestationalage)))   
  
  
 df1 <- setNames(df1,c("Groups","mean","SD","IQR","median"))   
kable(df1)

| Groups | mean | SD | IQR | median |
| --- | --- | --- | --- | --- |
| Obese | 3447.420 | 652.9939 | 1.9 | 39.3 |
| Severly Obese | 3445.350 | 658.7284 | 1.8 | 39.2 |
| Morbidly Obese | 3450.111 | 783.3571 | 1.9 | 39.1 |
| All | 3447.216 | 666.9211 | 1.9 | 39.3 |

With regards to the LGA we can calculate the percentage by classifying into 2 groups. before hand we can identify if the groups are more than 2 in LGA by using unique command.

## percentage of LGA

unique(neonatal$lga)

## [1] "normalsize" "LGA"

df2 <- data.frame(  
 "Groups" = c("Obese", "Severly Obese", "Morbidly Obese","All"),  
  
c(obese %>% filter(lga %in% c("normalsize")) %>% nrow()/nrow(obese)\*100,  
severely\_obese %>% filter(lga %in% c("normalsize")) %>% nrow()/nrow(severely\_obese)\*100,  
 morbidly\_obese %>% filter(lga %in% c("normalsize")) %>% nrow()/nrow(morbidly\_obese)\*100,  
 neonatal %>% filter(lga %in% c("normalsize")) %>% nrow()/nrow(neonatal)\*100),  
  
  
c(obese %>% filter(lga %in% c("LGA")) %>% nrow()/nrow(obese)\*100,  
severely\_obese %>% filter(lga %in% c("LGA")) %>% nrow()/nrow(severely\_obese)\*100,  
morbidly\_obese %>% filter(lga %in% c("LGA")) %>% nrow()/nrow(morbidly\_obese)\*100,  
neonatal %>% filter(lga %in% c("LGA")) %>% nrow()/nrow(neonatal)\*100),  
  
c(obese %>% filter(lga %in% c("normalsize")) %>% nrow(),  
severely\_obese %>% filter(lga %in% c("normalsize")) %>% nrow(),  
 morbidly\_obese %>% filter(lga %in% c("normalsize")) %>% nrow(),  
 neonatal %>% filter(lga %in% c("normalsize")) %>% nrow()),  
  
c(obese %>% filter(lga %in% c("LGA")) %>% nrow(),  
severely\_obese %>% filter(lga %in% c("LGA")) %>% nrow(),  
morbidly\_obese %>% filter(lga %in% c("LGA")) %>% nrow(),  
neonatal %>% filter(lga %in% c("LGA")) %>% nrow())  
)  
  
df2 <- setNames(df2,c("Groups","normalsize %","lga %","number of normalsize","number of lga"))  
flextable(df2)

| Groups | normalsize % | lga % | number of normalsize | number of lga |
| --- | --- | --- | --- | --- |
| Obese | 73.63476 | 26.36524 | 1,254 | 449 |
| Severly Obese | 71.88082 | 28.11918 | 386 | 151 |
| Morbidly Obese | 67.25664 | 32.74336 | 152 | 74 |
| All | 72.66829 | 27.33171 | 1,792 | 674 |

# percentage of Hypoglycaemia

Now for the Hypoglycaemia we can calculate the percentage.

df3 <- data.frame(  
 "Groups" = c("Obese", "Severly Obese", "Morbidly Obese","All"),  
 c(obese %>% filter(hypoglycaemia %in% c("Yes")) %>% nrow()/nrow(obese)\*100,  
 severely\_obese %>% filter(hypoglycaemia %in% c("Yes")) %>% nrow()/nrow(severely\_obese)\*100,  
 morbidly\_obese %>% filter(hypoglycaemia %in% c("Yes")) %>% nrow()/nrow(morbidly\_obese)\*100,  
 neonatal %>% filter(hypoglycaemia %in% c("Yes")) %>% nrow()/nrow(neonatal)\*100),  
  
 c(obese %>% filter(hypoglycaemia %in% c("No")) %>% nrow()/nrow(obese)\*100,  
 severely\_obese %>% filter(hypoglycaemia %in% c("No")) %>% nrow()/nrow(severely\_obese)\*100,  
 morbidly\_obese %>% filter(hypoglycaemia %in% c("No")) %>% nrow()/nrow(morbidly\_obese)\*100,  
 neonatal %>% filter(hypoglycaemia %in% c("No")) %>% nrow()/nrow(neonatal)\*100),  
   
 c(obese %>% filter(hypoglycaemia %in% c("Yes")) %>% nrow(),  
 severely\_obese %>% filter(hypoglycaemia %in% c("Yes")) %>% nrow(),  
 morbidly\_obese %>% filter(hypoglycaemia %in% c("Yes")) %>% nrow(),  
 neonatal %>% filter(hypoglycaemia %in% c("Yes")) %>% nrow()),  
  
 c(obese %>% filter(hypoglycaemia %in% c("No")) %>% nrow(),  
 severely\_obese %>% filter(hypoglycaemia %in% c("No")) %>% nrow(),  
 morbidly\_obese %>% filter(hypoglycaemia %in% c("No")) %>% nrow(),  
 neonatal %>% filter(hypoglycaemia %in% c("No")) %>% nrow())  
   
 )  
  
 df3 <- setNames(df3,c("Groups","yes %","no %","number for yes","number for no"))  
  
kable(df3)

| Groups | yes % | no % | number for yes | number for no |
| --- | --- | --- | --- | --- |
| Obese | 7.516148 | 92.48385 | 128 | 1575 |
| Severly Obese | 6.703911 | 93.29609 | 36 | 501 |
| Morbidly Obese | 10.619469 | 89.38053 | 24 | 202 |
| All | 7.623682 | 92.37632 | 188 | 2278 |

For the NICU admission at birth we can calculate it’s percentage by the code below

# percentage of Admissions

df4 <- data.frame(  
  
 "Groups" = c("Obese", "Severly Obese", "Morbidly Obese","All"),  
 c(obese %>% filter(admission %in% c("Yes")) %>% nrow()/nrow(obese)\*100,  
 severely\_obese %>% filter(admission %in% c("Yes")) %>% nrow()/nrow(severely\_obese)\*100,  
 morbidly\_obese %>% filter(admission %in% c("Yes")) %>% nrow()/nrow(morbidly\_obese)\*100,  
 neonatal %>% filter(admission %in% c("Yes")) %>% nrow()/nrow(neonatal)\*100),  
  
 c(obese %>% filter(admission %in% c("No")) %>% nrow()/nrow(obese)\*100,  
 severely\_obese %>% filter(admission %in% c("No")) %>% nrow()/nrow(severely\_obese)\*100,  
 morbidly\_obese %>% filter(admission %in% c("No")) %>% nrow()/nrow(morbidly\_obese)\*100,  
 neonatal %>% filter(admission %in% c("No")) %>% nrow()/nrow(neonatal)\*100),  
   
 c(obese %>% filter(admission %in% c("Yes")) %>% nrow(),  
 severely\_obese %>% filter(admission %in% c("Yes")) %>% nrow(),  
 morbidly\_obese %>% filter(admission %in% c("Yes")) %>% nrow(),  
 neonatal %>% filter(admission %in% c("Yes")) %>% nrow()),  
  
 c(obese %>% filter(admission %in% c("No")) %>% nrow(),  
 severely\_obese %>% filter(admission %in% c("No")) %>% nrow(),  
 morbidly\_obese %>% filter(admission %in% c("No")) %>% nrow(),  
 neonatal %>% filter(admission %in% c("No")) %>% nrow()/nrow(neonatal)\*100)  
   
 )  
  
 df4 <- setNames(df4,c("Groups","yes %","no %","number for yes","number for no"))  
  
kable(df4)

| Groups | yes % | no % | number for yes | number for no |
| --- | --- | --- | --- | --- |
| Obese | 20.02349 | 79.97651 | 341 | 1362.00000 |
| Severly Obese | 18.99441 | 81.00559 | 102 | 435.00000 |
| Morbidly Obese | 18.58407 | 81.41593 | 42 | 184.00000 |
| All | 19.66748 | 80.33252 | 485 | 80.33252 |

## P-value calculation

There are 3 groups in here so we can do an 3 groups Analysis of variance test statistical analysis test to see if the groups are different. before doing the test we need to define the hypothesis at 5% significance level which will help us to determine if the groups there exists a significant difference between them.

Null Hypothesis: H0: Birthweight of the newborns is the same in all groups. Alternate Hypothesis: H1: Birthweight of the newborns is different in the groups.

### For birthweight

anova <- aov(birthweight~bmicat,data=neonatal)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)  
## bmicat 2 3.834e+03 1917 0.004 0.996  
## Residuals 2463 1.096e+09 445143

TukeyHSD(anova)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = birthweight ~ bmicat, data = neonatal)  
##   
## $bmicat  
## diff lwr upr  
## 30-34.9(obese)->=40 (morbidly obese) -2.690185 -113.45959 108.0792  
## 35-39.9 (severly obese)->=40 (morbidly obese) -4.760526 -128.82178 119.3007  
## 35-39.9 (severly obese)-30-34.9(obese) -2.070341 -79.50668 75.3660  
## p adj  
## 30-34.9(obese)->=40 (morbidly obese) 0.9982132  
## 35-39.9 (severly obese)->=40 (morbidly obese) 0.9955455  
## 35-39.9 (severly obese)-30-34.9(obese) 0.9978350

Our result show that p-value is much large than 0.05 so we can reject the Alternate hypothesis. TukeyHSD test is used to determine how much the groups differ from each other. Our results show that p-value for all the 3 groups is very large so we conclude that all of the 3 groups are similar to each other. Moreoever the F-stats value is very small which also supports our hypothesis.

### for Gestational age

Null hypothesis H0: Gestational age is the same in all groups. Alternate hypothesis H1: Gestational age is different in the groups.

anova <- aov(gestationalage~bmicat,data=neonatal)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## bmicat 2 76 38.04 5.803 0.00306 \*\*  
## Residuals 2463 16147 6.56   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(anova)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = gestationalage ~ bmicat, data = neonatal)  
##   
## $bmicat  
## diff lwr upr  
## 30-34.9(obese)->=40 (morbidly obese) 0.61703605 0.19194624 1.0421259  
## 35-39.9 (severly obese)->=40 (morbidly obese) 0.52766599 0.05156726 1.0037647  
## 35-39.9 (severly obese)-30-34.9(obese) -0.08937006 -0.38654053 0.2078004  
## p adj  
## 30-34.9(obese)->=40 (morbidly obese) 0.0019496  
## 35-39.9 (severly obese)->=40 (morbidly obese) 0.0254625  
## 35-39.9 (severly obese)-30-34.9(obese) 0.7603987

Our results show that p-value is less than 0.05 so we reject our null hypothesis and we can say that the groups are different in geostational age. Note that residuals are also less here as compared to the previous analysis. The asteric next to p-value also shows the significance of the test. Tukeytest here shows the difference between the groups. The p-value for the severaly obese and obese comes out to be more than 0.05 so we can say that there is a high significant difference between these 2 groups as compared to obese and morbidltly obese as well as as the several obese and morbdily obese.

### For LGA

here the data is categorical so we will use the chisquare test. For the test again we define our hypothesis as

Null hypothesis H0: LGA is the same in all groups. Alternate hypothesis H1: LGA is different in the groups.

chisq.test(table(neonatal$lga))

##   
## Chi-squared test for given probabilities  
##   
## data: table(neonatal$lga)  
## X-squared = 506.86, df = 1, p-value < 2.2e-16

The results show that p-value is less than 0.05 so we reject our null hypothesis and we can say that the groups are different in LGA. our degrees of freedom is 1 so we can say that the test is independent.

### For Hypoglycaemia

Null hypothesis H0: Hypoglycaemia is the same in all groups. Alternate hypothesis H1: Hypoglycaemia is different in the groups.

chisq.test(table(neonatal$hypoglycaemia))

##   
## Chi-squared test for given probabilities  
##   
## data: table(neonatal$hypoglycaemia)  
## X-squared = 1771.3, df = 1, p-value < 2.2e-16

In this case the groups are different in the hypoglycaemia because the p-value is less than 0.05. The degrees of freedom is 1 so we can say that the test is independent. X-squared value is 1771.8 so we can say that the test is not significant. Remember that actual degreess of freedom is 2 but R always takes value 1 less than size of population.

### For admission

Null hypothesis H0: Admission is the same in all groups. Alternate hypothesis H1: Admission is different in the groups.

chisq.test(table(neonatal$admission))

##   
## Chi-squared test for given probabilities  
##   
## data: table(neonatal$admission)  
## X-squared = 907.55, df = 1, p-value < 2.2e-16

Admission is different in the groups. The p-value is less than 0.05. X-squared value of chisquare test is 907.5 so we can say that the test is not significant.

## Paragraph about the Analysis

The analysis is based on the following questions: \* Is the BMI is siginficaltly different in the groups \* What’s the relation between the BMI and neonatal outcomes.

We have performed several statistical tests to determine the significance of the data such ANOVA, chi-square and Tukey test. The result for the tests with numerical data column such as birthweight and geostational have different p-values at 5% significance level with 95% confidence interval. For the birthweight the p-value is >0.05 which shows that birthweight for the 3 groups are almost identical. It is also verified from the mean values of the birth weight where all values are near to 3450. IQR is also very small so we can say that the birthweight is not significantly different in the groups. With regards to geostational age the p-value is <0.05 which shows that the geostational age is significantly different in the groups. Moreover the Tukey test show that the two groups of 35-39.9 (severly obese)-30-34.9(obese) are not siginficaltly different as compared to all other combination of groups.

There are 3 categorical variables in the dataset which required the chi-squared test of significance. In all the 3 categorical variables the p-value is <0.05 which shows that the categorical variables are significantly different in the groups. Overall analysis shows that LGA, Hypoglycaemia and Admissions for all the 3 groups are significantly different from BMI while there exists some corelation for birthweight and hypoglycaemia.