EXPLORATORY DATA ANALYSIS

Leonard Kiti Alii, Pwani University

Introduction

In this document we are going to discuss how to perfom T- tests and ANOVA

Boxplot

Boxplots are a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"

One sample T-test

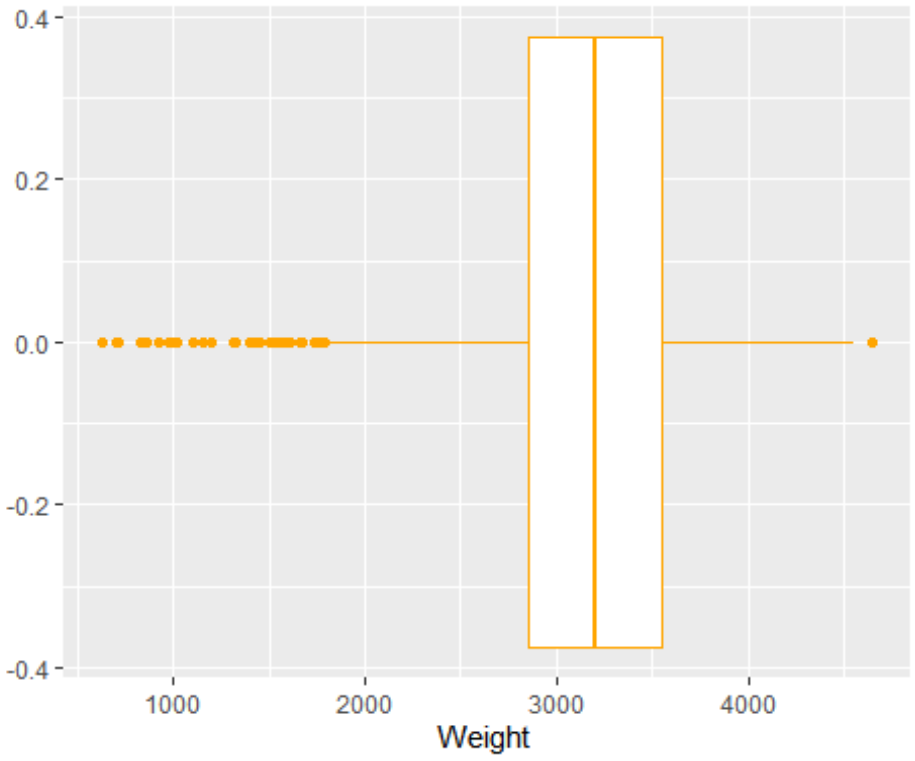
Question One: Is the **mean birthweight** significantly different from Zero ?

```
birthweight2 <- read.csv("C:/Users/Dr. Kiti/Desktop/DAAD
WORKSHOP/birthweight2.csv")

str(birthweight2)
```

```
## 'data.frame':    641 obs. of  11 variables:
##  $ id     : int  107 579 438 570 569 210 105 528 382 403 ...
##  $ matage : int  23 23 24 24 25 25 25 25 25 25 ...
##  $ ht     : int  2 2 1 2 1 1 2 2 1 2 ...
##  $ gestwks: int  39 41 36 39 31 38 38 39 39 40 ...
##  $ sex    : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 2 1 2 1 1 ...
##  $ bweight: int  3680 3120 2720 2550 1320 3260 3340 3040 3210 3380 ...
##  $ ethnic : int  1 4 3 4 4 1 1 4 3 3 ...
##  $ lbw    : Factor w/ 2 levels "Normal 2500+",..: 1 1 1 1 2 1 1 1 1 1 ...
##  $ agegrp : Factor w/ 4 levels "20-29 yrs","30-34 yrs",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ lbw2   : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ agegrp1: int  1 1 1 1 1 1 1 1 1 1 ...
```

```
library(ggplot2)
bxp<-ggplot(data = birthweight2,aes(bweight))+geom_boxplot(col =
"orange")+labs(x = "Weight ",y = " ")
bxp
```



In this section we wish to perform the One Sample T test

```
summary(birthweight2$bweight)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     630    2850    3200    3129    3550    4650

s1 =t.test(birthweight2$bweight)
s1

##
##  One Sample t-test
##
## data:  birthweight2$bweight
## t = 121.36, df = 640, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3078.507 3179.767
## sample estimates:
## mean of x
##  3129.137
```

Question Two: Is the **mean birthweight** significantly different from a 2500 units ?

```
s1_1=t.test(birthweight2$bweight,mu=2500)
s1_1
```

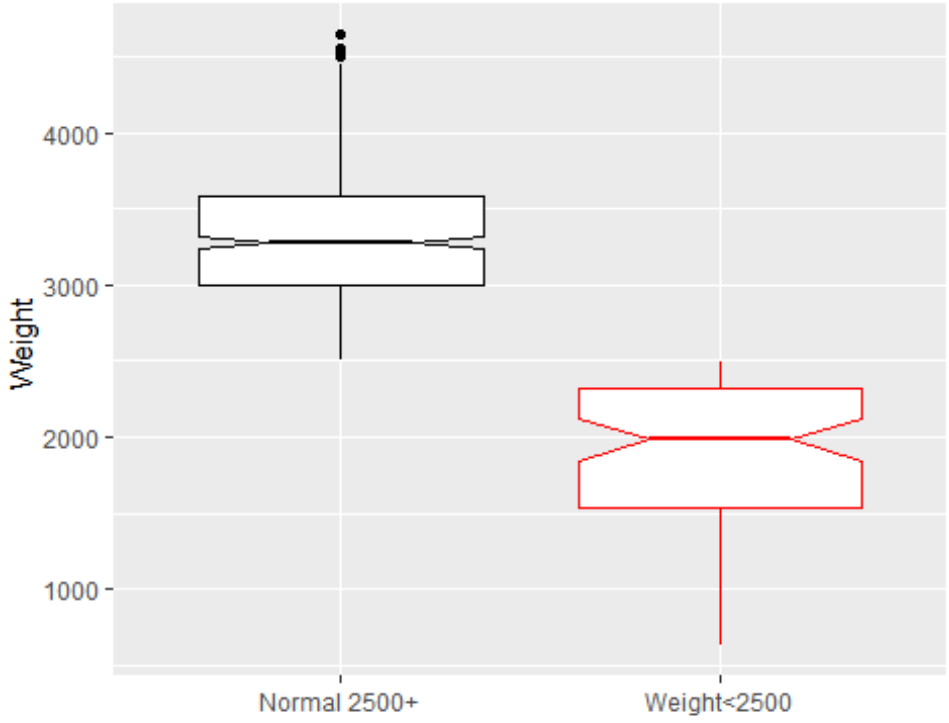
```
##
##  One Sample t-test
##
## data:  birthweight2$bweight
## t = 24.401, df = 640, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 2500
## 95 percent confidence interval:
##  3078.507 3179.767
## sample estimates:
## mean of x
##  3129.137
```

Two sample independent T test

Question Three: Is there any significance difference in the **mean birth weight** across the two categories in lbw

First we use the boxplot to visualise the relationship

```
bxp1<-ggplot(data = birthweight2,aes(lbw, bweight))+geom_boxplot(col =
c(1,2),notch = T)+labs(x = " ",y = "Weight ")
bxp1
```



We conduct the Two sample independent T-test

This is used to check whether there are significant differences across two independent groups

```
s2=t.test(birthweight2$bweight~birthweight2$lbw)
s2

##
##  Welch Two Sample t-test
##
## data:  birthweight2$bweight by birthweight2$lbw
## t = 23.736, df = 95.296, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1324.794 1566.618
## sample estimates:
## mean in group Normal 2500+  mean in group Weight<2500
##                   3309.569                   1863.862
```

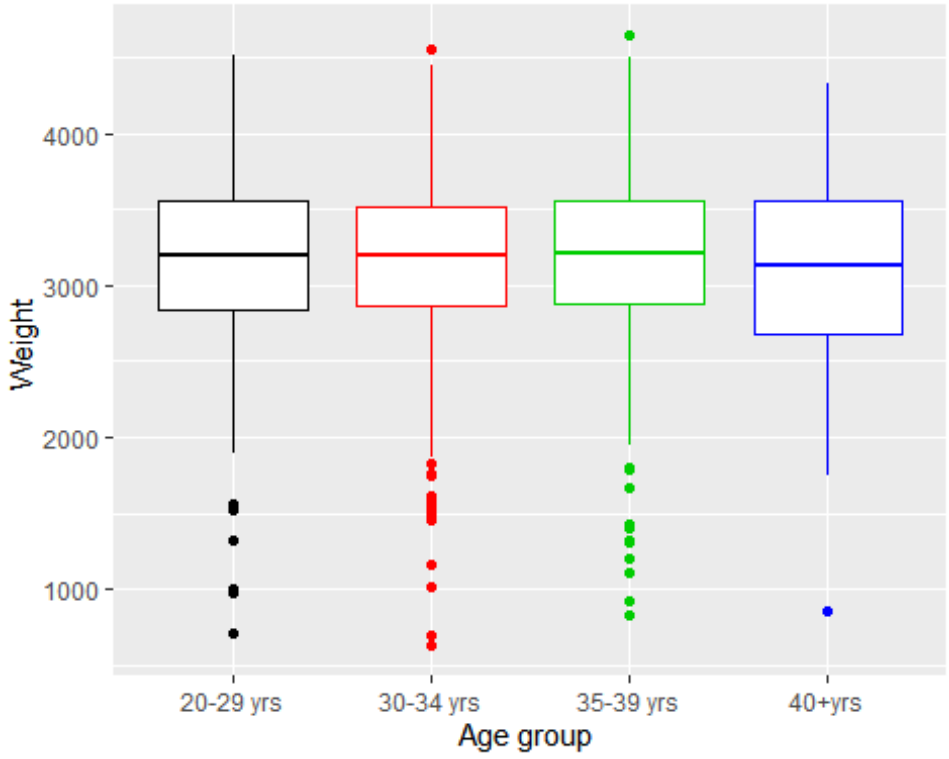
Differences across three or more groups; Use of ANOVA

Question Four: Is there any significant difference in the **mean birth weight** across the agegroups?

To check whether there are any differences in a continuous variable across three or more groups we use the ANOVA.

First we visualise the distribution

```
bxp2<-ggplot(data = birthweight2,aes(agegrp, bweight))+geom_boxplot(col =
c(1,2,3,4))+labs(x = "Age group ",y = "Weight ")
bxp2
```



We conduct the ANOVA test

```
tapply(birthweight2$bweight,birthweight2$agegrp,mean)

## 20-29 yrs 30-34 yrs 35-39 yrs    40+yrs
##  3102.326  3137.745  3132.884  3112.625

oneway<-aov(birthweight2$bweight~birthweight2$agegrp)
summary(oneway)

##                       Df    Sum Sq Mean Sq F value Pr(>F)
## birthweight2$agegrp    3     99258   33086   0.077  0.972
## Residuals            637 272620864  427976
```