R Assignment 04

Digvijay Jondhale 0862899

Loading Data ...

```
prison <- read.csv(file='canada_incarceration.csv');
```

```
str(prison);
```

```
'data.frame':	6 obs. of  3 variables:
 $ Group            : chr  "Indigenous" "Asian" "Black" "Caucasian" ...
 $ Incarcerated     : int  5009 1318 1895 13870 250 619
 $ Proportion_of_Pop: num  0.0486 0.1457 0.0348 0.7473 0.013 ...
```

```
summary(prison);
```

```
    Group            Incarcerated     Proportion_of_Pop
 Length:6           Min.   :  250.0   Min.   :0.01060  
 Class :character   1st Qu.:  793.8   1st Qu.:0.01845  
 Mode  :character   Median : 1606.5   Median :0.04170  
                    Mean   : 3826.8   Mean   :0.16667  
                    3rd Qu.: 4230.5   3rd Qu.:0.12143  
                    Max.   :13870.0   Max.   :0.74730  
```

Question 1– Canada's Prison System

a) [1 mark] What test should you use and why?

Ans :
We should use **Chi Squared Goodness of Fit Test**, as it allows us to compare the observed data with expected data, and see if the observed data deviates from the expected data significantly or not.

b) Create a barplot to visualize the data in the table. Create a side-by-side barplot with observed and expected counts next to each other.

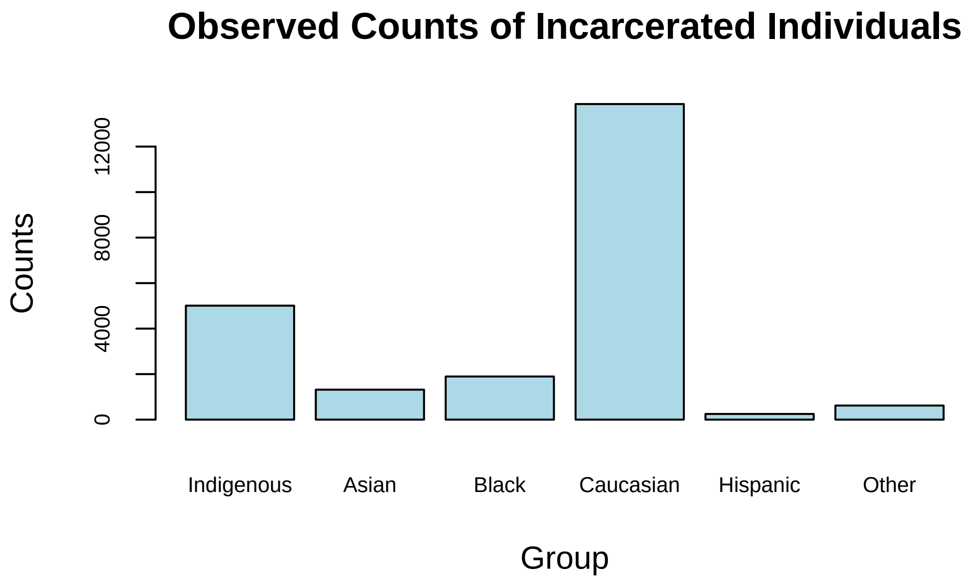
Ans :

Barplot's

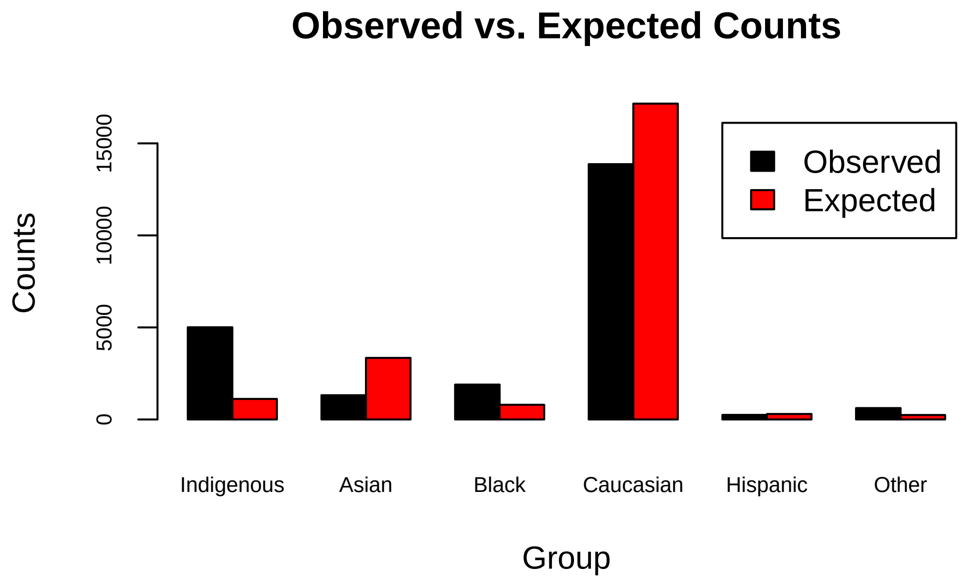
```
incarcerated <- prison$Incarcerated;
population_proportion <- prison$Proportion_of_Pop
population_total <- sum(incarcerated) / sum(population_proportion)


expected <- population_proportion * population_total

# barplot
barplot(incarcerated,
        names.arg = c("Indigenous", "Asian", "Black", "Caucasian", "Hispanic", "Other"),
        col = "lightblue", main = "Observed Counts of Incarcerated Individuals",
        ylab = "Counts", xlab = "Group",cex.axis = 0.7,cex.names = 0.7)
```



```r
counts <- rbind(incarcerated, expected)
barplot(counts,
        beside = TRUE,
        names.arg = c("Indigenous", "Asian", "Black", "Caucasian", "Hispanic", "Other"),
        col = c("black", "red"),
        legend.text = c("Observed", "Expected"),
        main = "Observed vs. Expected Counts",
        ylab = "Counts", xlab = "Group",cex.axis = 0.7,cex.names = 0.7)
```



c) [6 marks] Determine if there is any difference between the proportion of the population and the proportion of individuals incarcerated in federal prisons. (Use = 0.05.)

Ans :

Hypothesis * Ho : Observed Proportions of incarcerated individuals are equal to the population proportion. * Ha : Observed Proportions of incarcerated individuals are not equal to the population proportion.

Chi Sqared Test

```r
chisq_test <- chisq.test(x = prison$Incarcerated, p = prison$Proportion_of_Pop)
chisq_test
```

```
    Chi-squared test for given probabilities

data:  prison$Incarcerated
X-squared = 17532, df = 5, p-value < 2.2e-16
```

As the p-value (2.2e-16) is smaller than 0.05, we reject the null hypothesis and state that observed proportions of incarcerated individuals are not equal to the population. This indicates that certain groups are over or under represented in Canadian federal prisons.

Loading Data ...

```
chocolate <- read.csv(file='chocolate_antioxidants.csv');
```

```
str(chocolate);
```

```
'data.frame':	12 obs. of  3 variables:
 $ DC   : num  119 123 116 114 120 ...
 $ DC_MK: num  105.4 101.1 102.7 97.1 101.9 ...
 $ MC   : num  102.1 105.8 99.6 102.7 98.8 ...
```

```
summary(chocolate);
```

```
       DC              DC_MK              MC        
 Min.   :107.9   Min.   : 93.50   Min.   : 94.70  
 1st Qu.:115.3   1st Qu.: 99.58   1st Qu.: 98.67  
 Median :115.7   Median :101.00   Median : 99.65  
 Mean   :116.1   Mean   :100.70   Mean   :100.18  
 3rd Qu.:117.4   3rd Qu.:102.62   3rd Qu.:102.25  
 Max.   :122.6   Max.   :105.40   Max.   :105.80  
```

Question 2– Get Outta Here Oxidants!

a) [1 mark] What test should you use and why?

Ans : We should use one way ANOVA test because we have more than 2 levels of the factor(chocolate_type has 3 levels) , the dependent variable antioxidant capacity of blood plasma is continious, and we want to compare the mean of groups to determine if they are statistically significant or not.

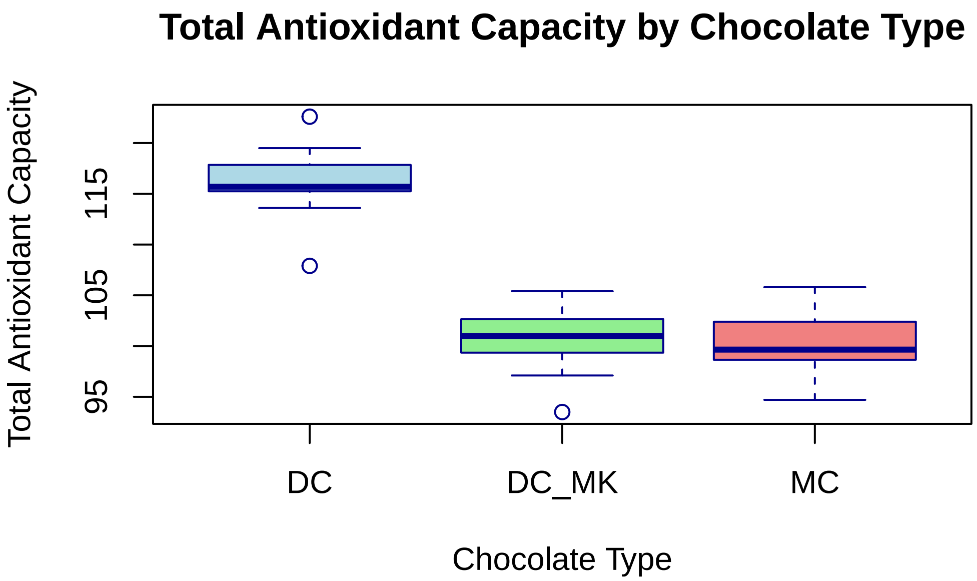
b) [2 marks] Create a boxplot to visualize the data in the table (properly labelled and with a title).

Ans :

Box-plot

```
library(tidyr)
chocolate_long <- pivot_longer(chocolate, cols = c("DC", "DC_MK", "MC"),
                               names_to = "chocolate_type", values_to = "antioxidant_capacity


boxplot(antioxidant_capacity ~ chocolate_type, data = chocolate_long,
        main = "Total Antioxidant Capacity by Chocolate Type",
        xlab = "Chocolate Type", ylab = "Total Antioxidant Capacity",
        col = c("lightblue", "lightgreen", "lightcoral"), border = "darkblue")
```



c) [6 marks] Determine if there is any difference between the mean total antioxidant capacity of blood plasma after consuming the different types of chocolate. (Use =0.05.)

Ans :

Performing ANOVA Test

```
anova_result <- aov(antioxidant_capacity ~ chocolate_type, data = chocolate_long)
summary(anova_result)
```

```
               Df Sum Sq Mean Sq F value   Pr(>F)    
chocolate_type  2 1952.6   976.3   93.58 2.52e-14 ***
Residuals      33  344.3    10.4                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p-value is extremely small , we reject the null hypothesis that all means are equal. This indicates that there is statistically significant difference between the antioxidant capacities in blood plasma of the three chocolate types.

d) [2 marks] You should find a difference in part c). Which group means are different from the others? Use Tukey’s HSD with a significance level of = 0.05. You just need to write a few sentences indicating which means are different and your reasoning.

Ans :
Tukey’s HSD Test:

```
tukey <- TukeyHSD(anova_result)
tukey$chocolate_type;
```

```
                diff        lwr        upr        p adj
DC_MK-DC -15.3583333 -18.594104 -12.122562 1.000200e-12
MC-DC    -15.8750000 -19.110771 -12.639229 4.489742e-13
MC-DC_MK  -0.5166667  -3.752438   2.719104 9.190724e-01
```

- For dark chocolate with milk (DC_MK) it is significantly lower than that of dark chocolate (DC) as p-value is significant, Similarly, the mean antioxidant capacity of milk chocolate (MC) is lower than that of dark chocolate (DC) as p-value is smaller than 0.05.
- For MC vs DC_MK. the p-value is indicates not significant.
- From the above results , we can see that Dark Chocolate (DC) has a significantly higher atioxidant capacity than both dark_chocolate with milk and milk chocolate.
    - There is no significant difference between dark chocolate with milk and milk chocolate.

- The above results indicates that adding milk in dark chocolate or using milk chocolate does not alter the antioxident levels in blood plasma compared to dark chocolate alone.