Final Project

STA 104 Non-Parametric Statistics

Justin Luong and Ryandeep Chawla

December 2, 2022

I. Abstract

Nutrition and personal characteristics have been linked to individual health in various ways. It has been discovered that there is a disparity in the body fat of males and females, negative health effects with smokers, and health benefits with vitamin supplements. We use data from a cross-sectional study of 315 patients who had an elective surgical procedure during a three-year period to assess the association between nutrition and personal characteristics with individual health in detail. The methods used to analyze the data were QQ plots, Shapiro-Wilk Tests, Spearman Rank Correlation Tests, Kruskal-Wallis Tests, and Wilcoxon Rank Sum Tests. For individual health, we find that gender plays a small role, smoking history plays a significant role, and dietary factors play a significant role in certain cases.

II. Introduction

In this report, we will analyze a nutrition and health data set of 315 individuals to determine if there is an association between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene, and other carotenoids. We are curious about these relationships to advance the knowledge of diets and improve the health of our society.

In specific, we will observe dietary factors in consideration of the Quetelet index, smoking history, gender, and vitamin use. The Quetelet index is obtained by dividing the weight by the height squared. The dietary variables involved are calories, fat, fiber, alcohol, cholesterol, dietary beta-carotene, dietary retinol, plasma beta-carotene, and plasma retinol.

III. Methods/Results

Association

We first used QQ plots and conducted the Shapiro-Wilk Test to determine if the relevant data passes the normality assumption. All of the data were not normally distributed since the p-values were less than the significance level of 0.05, so we used the Spearman Rank Correlation Test to assess the association. With this test, we obtained the Pearson product-moment correlation. We determined if the correlation was significant if the p-value was smaller than the significance level of 0.05. The results are placed below.

QQ Plots

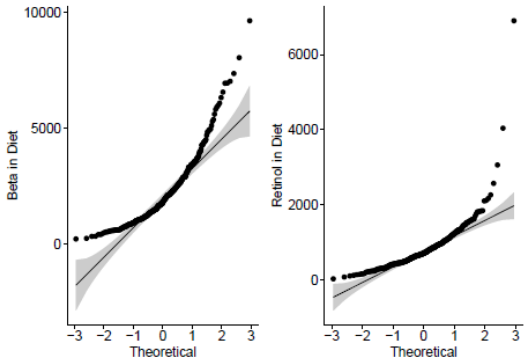


Figure 1: Beta Carotene and Retinol in Diet

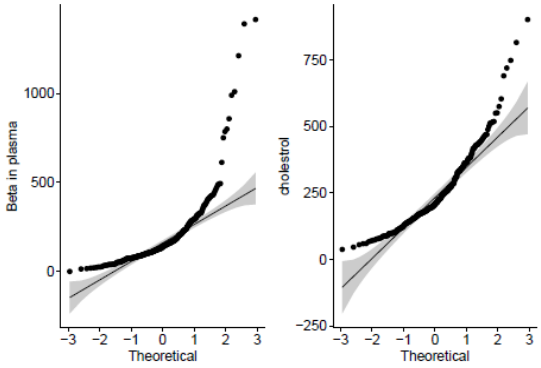


Figure 2: Beta Carotene in Plasma and Cholesterol

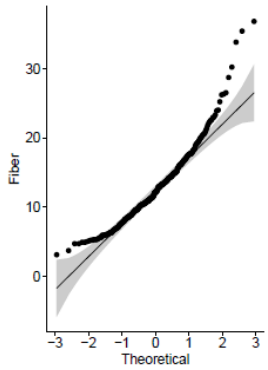


Figure 3. Fiber

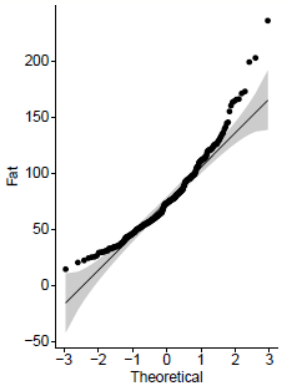


Figure 4: Fat and Cholesterol

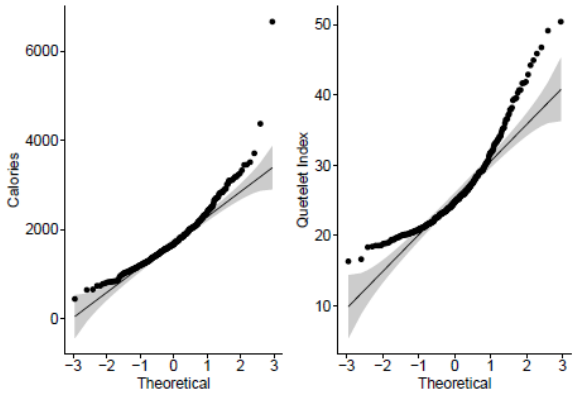


Figure 5. Calories and Quetelet Index

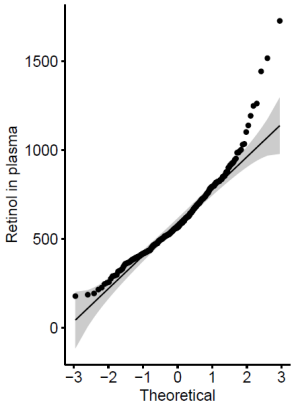


Figure 6: Retinol in Plasma

The QQ plots displayed values that did not stay close enough to the curves for all associations discussed below, which indicates that the data is not normally distributed.

Shapiro-Wilk Test Results

Variable	Shapiro-Wilk Test P-Value	Reject/Fail to Reject	Normality
Dietary Beta Carotene	1.262e-15	Reject	No
Dietary Retinol	< 2.2e-16	Reject	No
Plasma Beta Carotene	< 2.2e-16	Reject	No
Plasma Retinol	4.251e-11	Reject	No
Cholesterol	8.467e-14	Reject	No
Fat	3.899e-10	Reject	No
Fiber	2.248e-10	Reject	No
Calories	2.092e-13	Reject	No
Quetelet Index	2.312e-14	Reject	No

Spearman Rank Correlation Test Results

Variables	Correlation Coefficient	P-Value	Reject/Fail to Reject Null	Significant Association	Strength	Direction
Beta Carotene and Retinol in Diet	0.1962858	0.0004583	Reject	Yes	Weak	Positive
Beta Carotene and Retinol in Plasma	0.1306213	0.02039	Reject	Yes	Weak	Positive

Plasma Beta Carotene and Cholesterol	-0.142528	0.01133	Reject	Yes	Weak	Negative
Fat and Cholesterol	0.7556184	< 2.2e-16	Reject	Yes	Strong	Positive
Fiber and Cholesterol	0.2252991	5.466e-05	Reject	Yes	Weak	Positive
Calories and Quetelet Index	-0.01309881	0.8169	Fail to Reject	No	Very Weak, almost insignificant	Negative

Differences

Within diet, we have fat, fiber, alcohol, dietary beta-carotene, and dietary retinol. We also observe cholesterol, calories consumed, beta-carotene, and retinol in plasma and Quetelet Index. To observe the differences in the diet for individuals with different smoking backgrounds, we conducted the Kruskal-Wallis Test using the permutation method and adjustment of ties shown in the appendix. To observe the differences in the diet for people of different genders (male or female), we conducted the Wilcoxon Rank Sum Test. We will conclude that there is a significant difference if the p-value is less than the significance level of 0.05 for both tests.

Kruskal-Wallis Tests Results

Difference	**P-Value**	**Kruskal-Wallis Statistic**	**Reject/Fail to Reject Null**	**Presence/ Absence of Difference**
Fat and Smoke	0.0335	6.8536	Reject	Present
Fiber and Smoke	0.006	10.233	Reject	Present
Alcohol and Smoke	5e-04	15.513	Reject	Present
Dietary Beta-Carotene and Smoke	0.0125	8.319	Reject	Present

Dietary Retinol and Smoke	0.853	0.31461	Fail to Reject	Absent
Cholesterol and Smoke	0.0615	5.6491	Fail to Reject	Absent
Calories and Smoke	0.1895	3.295	Fail to Reject	Absent
Beta In Plasma and Smoke	0.0015	12.606	Reject	Present
Plasma Retinol and Smoke	0.0336	6.7176	Reject	Present
Quetelet Index and Smoke	0.3022	2.3975	Fail to Reject	Absent

Wilcoxon Rank Sum Tests Results

Difference	**P-Value**	**Wilcoxon Statistic**	**Reject/Fail to Reject Null**	**Presence/ Absence of Difference**
Fat and Gender	0.0001249	3624.5	Reject	Present
Fiber and Gender	0.3213	5187.5	Fail to Reject	Absent
Alcohol and Gender	0.05114	4685	Fail to Reject	Absent
Dietary Beta Carotene and Gender	0.3017	5165	Fail to Reject	Absent
Dietary Retinol and Gender	0.1482	4938	Fail to Reject	Absent
Cholesterol and Gender	4.616e-06	3215	Reject	Present
Calories and Gender	0.000402	3788	Reject	Present
Beta in Plasma and Gender	0.01377	7087	Reject	Present
Plasma Retinol and Gender	0.02158	4470	Reject	Present

Quetelet Index and Gender	0.1769	4990	Fail to Reject	Absent

IV. Discussion

From the Shapiro-Wilk Tests, we found that all of the variables reject the null hypothesis and conclude that the data is not normally distributed. This suggests that we are working with nonparametric statistics.

From the Spearman Rank Correlation Tests, we found that all of the relationships were significant associations except for calories and Quetelet Index. All were weak correlations besides the relation between fat and cholesterol (strong) and the relation between calories and Quetelet Index can be considered insignificant. All were positive correlations besides the relation between plasma beta carotene and cholesterol and the relation between calories and Quetelet Index. This suggests that for relationships with a positive correlation, an increase in one of the factors will lead to an increase in the other. Negative correlation relationships have the opposite effect.

From the Kruskal-Wallis Tests, we found that all of the differences were significant except four: dietary retinol and smoke, cholesterol and smoke, calories and smoke, and Quetelet Index and smoke. This suggests that smoking history does not affect diet and health differently for those four.

From the Wilcoxon Rank Sum Tests, we found that half of the differences were significant for dietary factors and gender. This suggests that gender plays a slight role in an individual's diet and health, which makes sense as males and females have different body fat percentages.

Appendix Rcode

Ryandeep Chawla & Justin Luong

12/2/2022

```
data<- read.csv("nutritionstudy.csv")
```

```
dat1<- data.frame("Beta_diet"= data$BetaDiet,
                  "Retinol_diet"= data$RetinolDiet,
                  "Beta_plasma"= data$BetaPlasma,
                  "Retinol_plasma"= data$RetinolPlasma)
dat2<- data.frame("Beta_plasma"= data$BetaPlasma,
                  "Cholestrol"= data$Cholesterol)
dat3<- data.frame("Fat"= data$Fat,
                  "Fiber"= data$Fiber, "Cholestrol"= data$Cholesterol)
dat4<- data.frame("Calories"= data$Calories,"Quetelet_index"= data$Quetelet)
dat5<- data.frame("PriorSmoke"= data$PriorSmoke, "Fat"= data$Fat,
                  "Fiber"= data$Fiber, "Alcohol"= data$Alcohol,
                  "Cholestrol"= data$Cholesterol, "Calories"= data$Calories,
                  "Beta_diet"= data$BetaDiet, "Retinol_diet"= data$RetinolDiet,
                  "Beta_plasma"= data$BetaPlasma,
                  "Retinol_plasma"= data$RetinolPlasma,
                  "Quetelet_index"= data$Quetelet)
dat6<- data.frame("Gender"= data$Gender, "Fat"= data$Fat,
                  "Fiber"= data$Fiber, "Alcohol"= data$Alcohol,
                  "Cholestrol"= data$Cholesterol, "Calories"= data$Calories,
                  "Beta_diet"= data$BetaDiet, "Retinol_diet"= data$RetinolDiet,
                  "Beta_plasma"= data$BetaPlasma,
                  "Retinol_plasma"= data$RetinolPlasma,
                  "Quetelet_index"= data$Quetelet)
```

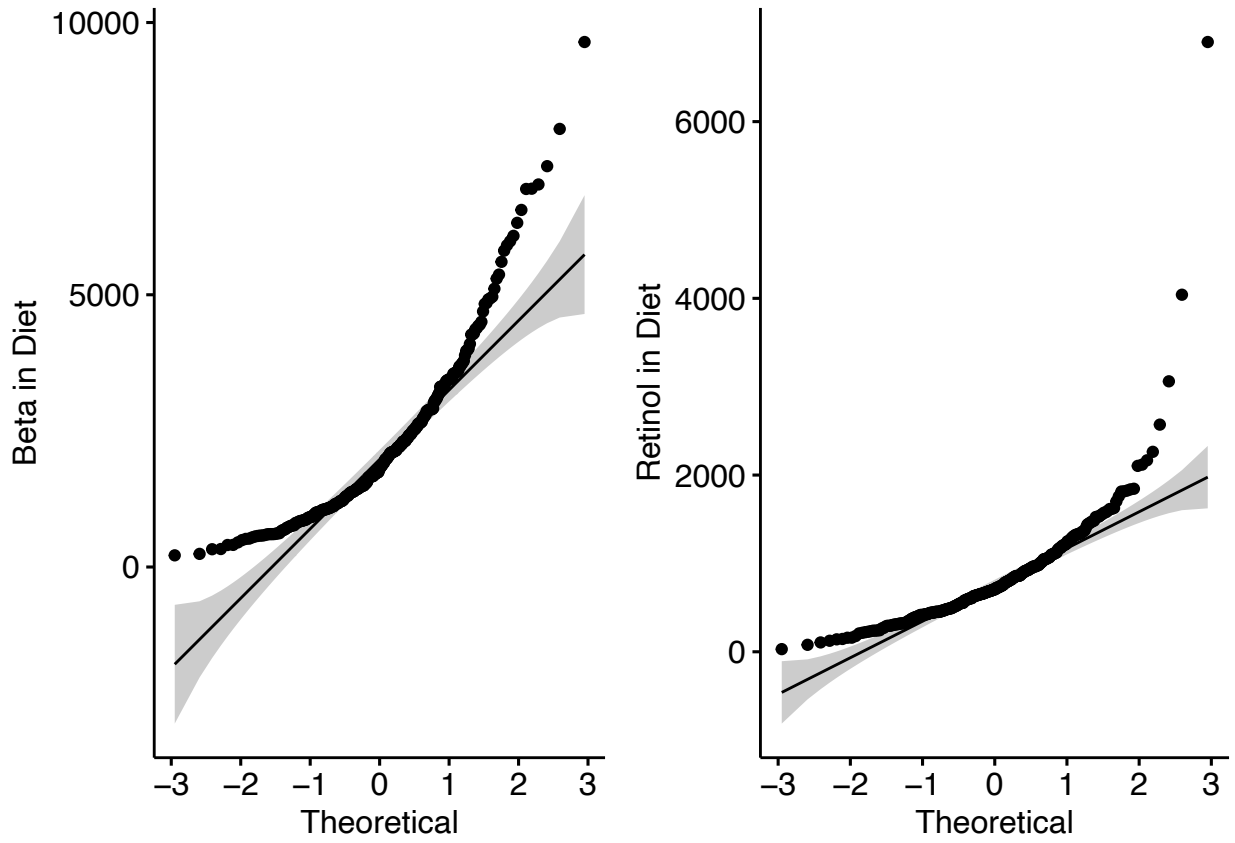
Question 1:

Checking for normality for beta in diet and retinol in diet

```
library("ggpubr")
```

```
## Loading required package: ggplot2
```

```
library("ggplot2")
plot1<- ggqqplot(dat1$Beta_diet, ylab = "Beta in Diet")
plot2<- ggqqplot(dat1$Retinol_diet, ylab = "Retinol in Diet")
ggarrange(plot1,plot2)
```




```
shapiro.test(dat1$Beta_diet)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat1$Beta_diet
## W = 0.87022, p-value = 1.262e-15
```

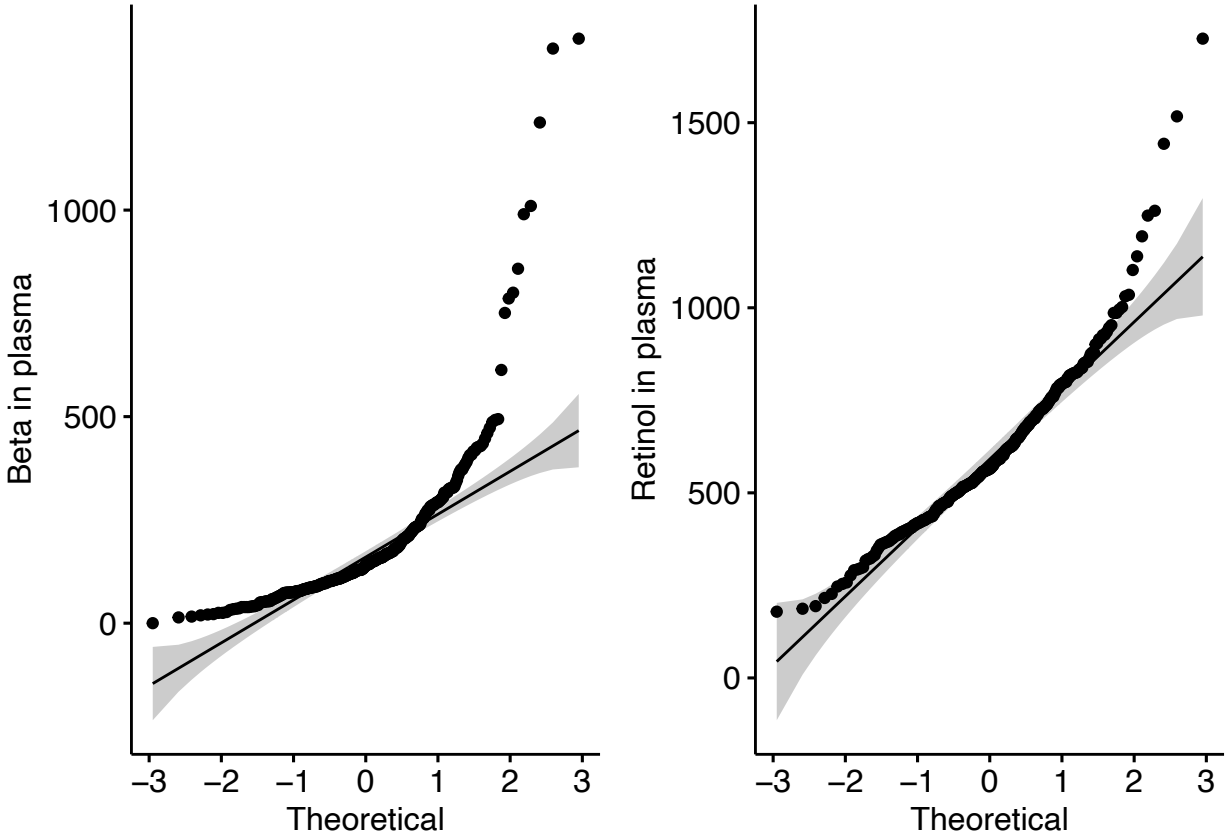
```
shapiro.test(dat1$Retinol_diet)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat1$Retinol_diet
## W = 0.71194, p-value < 2.2e-16
```

We reject the null from the shapiro wilks test and conclude that the data is not normal, similarly from the qqplot we can observe that the assumption for normality is not used.

Checking for normality for beta in plasma and retinol in plasma

```
plot3<- ggqqplot(dat1$Beta_plasma, ylab = "Beta in plasma")
plot4<- ggqqplot(dat1$Retinol_plasma, ylab = "Retinol in plasma")
ggarrange(plot3,plot4)
```




```
shapiro.test(dat1$Beta_plasma)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat1$Beta_plasma
## W = 0.66071, p-value < 2.2e-16
```

```
shapiro.test(dat1$Retinol_plasma)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat1$Retinol_plasma
## W = 0.92905, p-value = 4.251e-11
```

Using Spearman rank test

adding adjusted ranks

```
rank_beta_diet<- rank(dat1$Beta_diet, ties.method = "average")
rank_retinol_diet<- rank(dat1$Retinol_diet, ties.method = "average")
rank_beta_plasma<- rank(dat1$Beta_plasma, ties.method = "average")
rank_retinol_plasma<- rank(dat1$Retinol_plasma, ties.method = "average")
```

for diet, finding the pearson product moment correlation on the adjusted rank

```
cor.test(rank_beta_diet,rank_retinol_diet)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rank_beta_diet and rank_retinol_diet
## t = 3.5415, df = 313, p-value = 0.0004583
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.08767971 0.30028031
## sample estimates:
##       cor
## 0.1962858
```

```
cor(rank_beta_diet,rank_retinol_diet)
```

```
## [1] 0.1962858
```

for Plasma

```
cor.test(rank_beta_plasma,rank_retinol_plasma)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rank_beta_plasma and rank_retinol_plasma
## t = 2.3309, df = 313, p-value = 0.02039
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02040797 0.23769818
## sample estimates:
##       cor
## 0.1306213
```

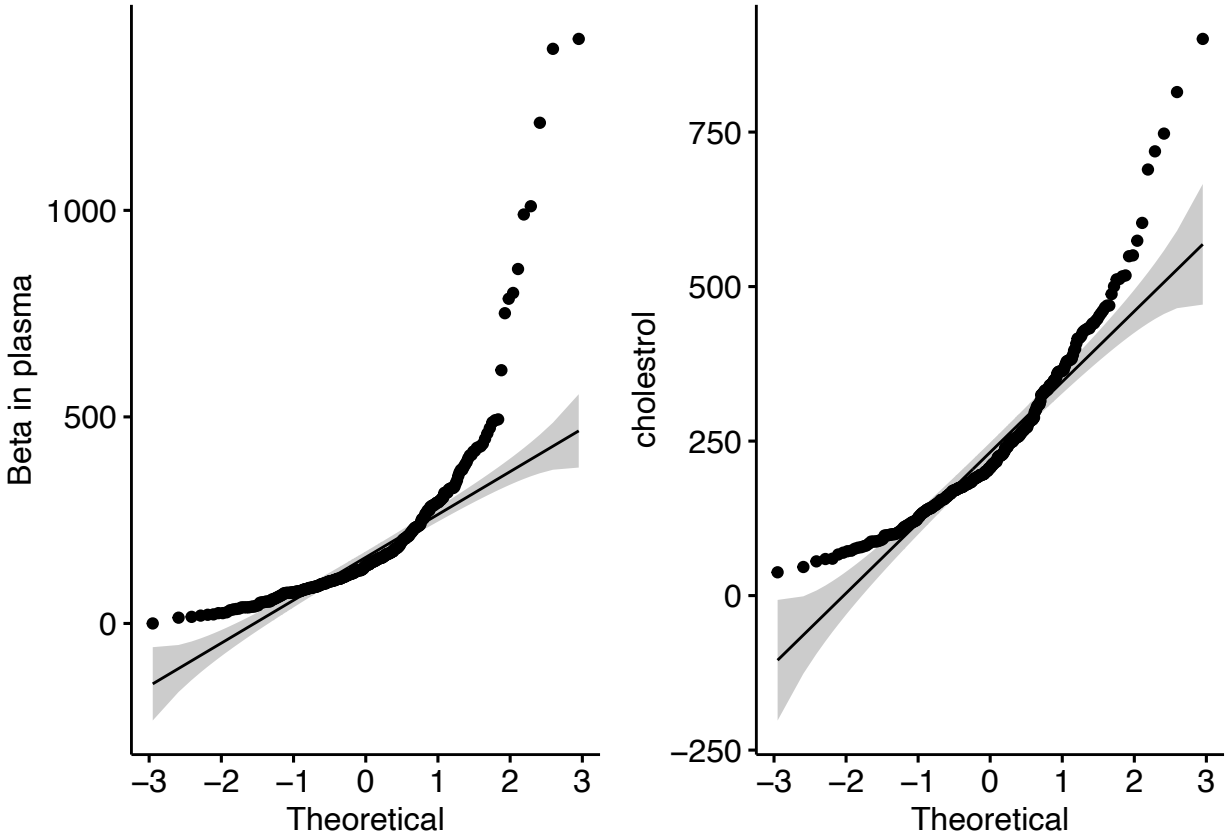
Weak positive correlation, but we reject the null hypothesis and decide that the correlation is not equal to 0, hence there is a correlation between the two.

–

Question 2 beta carotine in plasma and cholesterol

checking normality

```
plot5<- ggqqplot(dat2$Beta_plasma, ylab = "Beta in plasma")
plot6<- ggqqplot(dat2$Cholestrol, ylab = "cholestrol")
ggarrange(plot5,plot6)
```




```
shapiro.test(dat2$Beta_plasma)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat2$Beta_plasma
## W = 0.66071, p-value < 2.2e-16
```

```
shapiro.test(dat2$Cholestrol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat2$Cholestrol
## W = 0.89689, p-value = 8.467e-14
```

We fail the normality assumption and cannot use normal tests since the data is not normal

Using Spearman Rank Correlation test to see the relationship between beta in plasma and cholestrol

Creating adjusted ranks:

```
rank_beta_plasma<- rank(dat2$Beta_plasma, ties.method = "average")
rank_cholestrol<- rank(dat2$Cholestrol, ties.method = "average")
```

Calculating pearson product momemtum correlation

```
cor.test(rank_beta_plasma,rank_cholestrol)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rank_beta_plasma and rank_cholestrol
## t = -2.5476, df = 313, p-value = 0.01133
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.24911226 -0.03253241
## sample estimates:
##       cor
## -0.142528
```

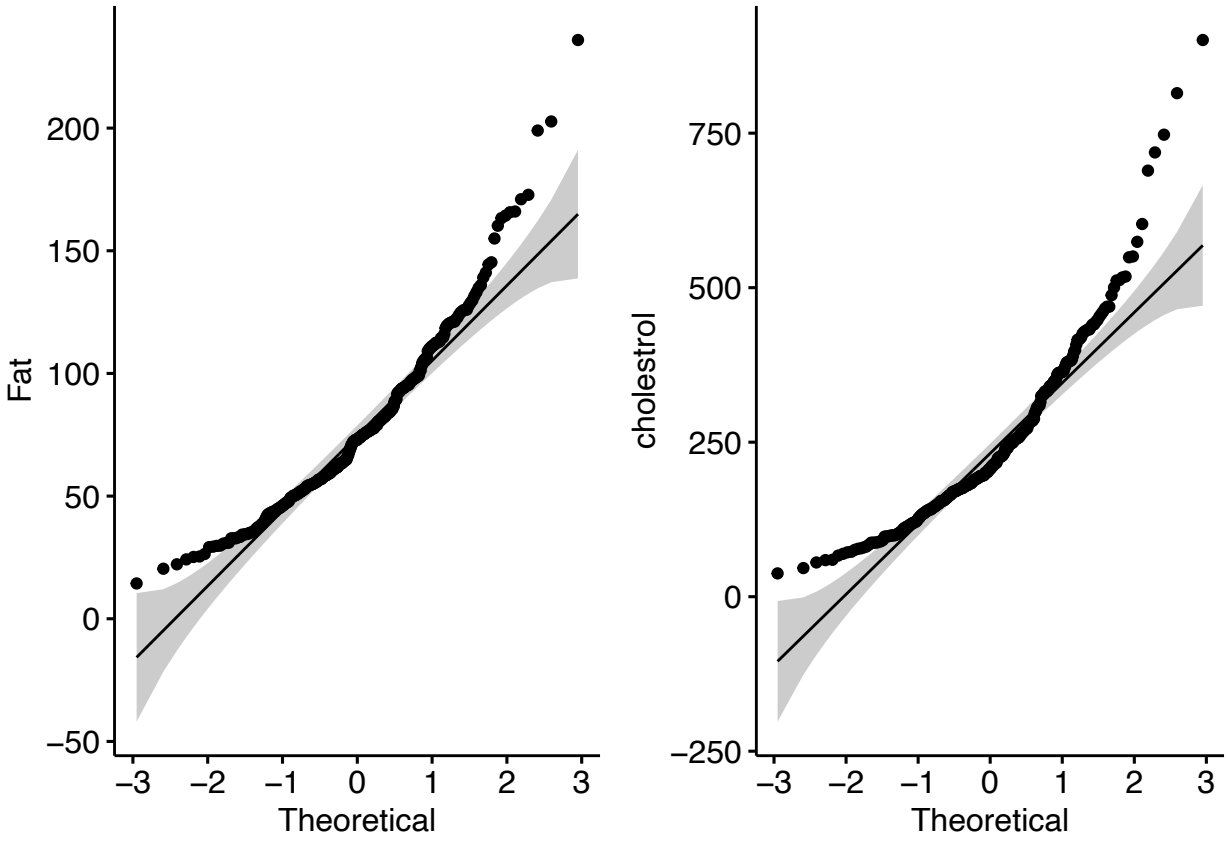
We reject null and conclude that there is a correlation and we can also say that there is a negative correlation between them as the correlation is negative

Question 3)

association between fat in diet and cholesterol and also between fiber in diet and cholesterol

Checking normality for fat in diet and cholesterol

```
plot7<- ggqqplot(dat3$Fat, ylab = "Fat")
plot8<- ggqqplot(dat3$Cholestrol, ylab = "cholestrol")
ggarrange(plot7,plot8)
```




```
shapiro.test(dat3$Fat)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat3$Fat
## W = 0.93866, p-value = 3.899e-10
```

```
shapiro.test(dat3$Cholestrol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat3$Cholestrol
## W = 0.89689, p-value = 8.467e-14
```

We reject the null, hence we can conclude that the data is not normal, qqplot also deviates from normality

Using Spearman Rank Correlation test to see the association between fat in diet and cholesterol

Creating adjusted ranks:

```
rank_fat<- rank(dat3$Fat, ties.method = "average")
rank_cholestrol<- rank(dat3$Cholestrol, ties.method = "average")
```

Calculating pearson product momemtum correlation

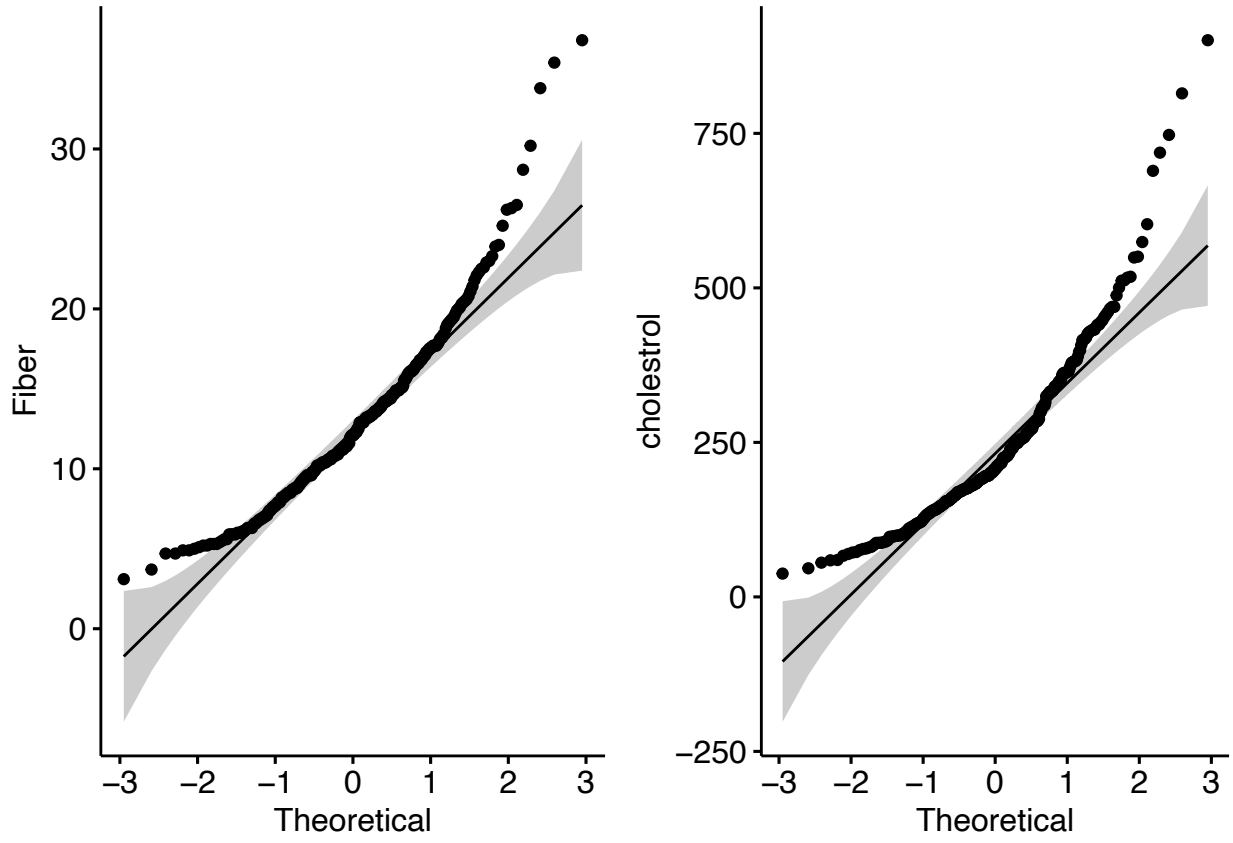
```
cor.test(rank_fat,rank_cholestrol)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rank_fat and rank_cholestrol
## t = 20.409, df = 313, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7038863 0.7993769
## sample estimates:
##       cor
## 0.7556184
```

The p-value is really is small, reject the null that correlation is 0, and corr is high positive correlation

Checking normality for fiber in diet

```
plot9<- ggqqplot(dat3$Fiber, ylab = "Fiber")
ggarrange(plot9,plot8)
```



```
shapiro.test(dat3$Fiber)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat3$Fiber
## W = 0.93636, p-value = 2.248e-10
```

We reject the null that the data is normal as the p-value is really small.

Calculating adjusted rank for Fiber in diet

```
rank_fiber<- rank(dat3$Fiber, ties.method = "average")
```

Finding pearson product moment correlation:

```
cor.test(rank_fiber,rank_cholestrol)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rank_fiber and rank_cholestrol
## t = 4.0911, df = 313, p-value = 5.466e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1177222 0.3276495
## sample estimates:
##       cor
## 0.2252991
```

We reject the null, and conclude that the correlation is not 0, the correlation is relatively weak at 0.2252

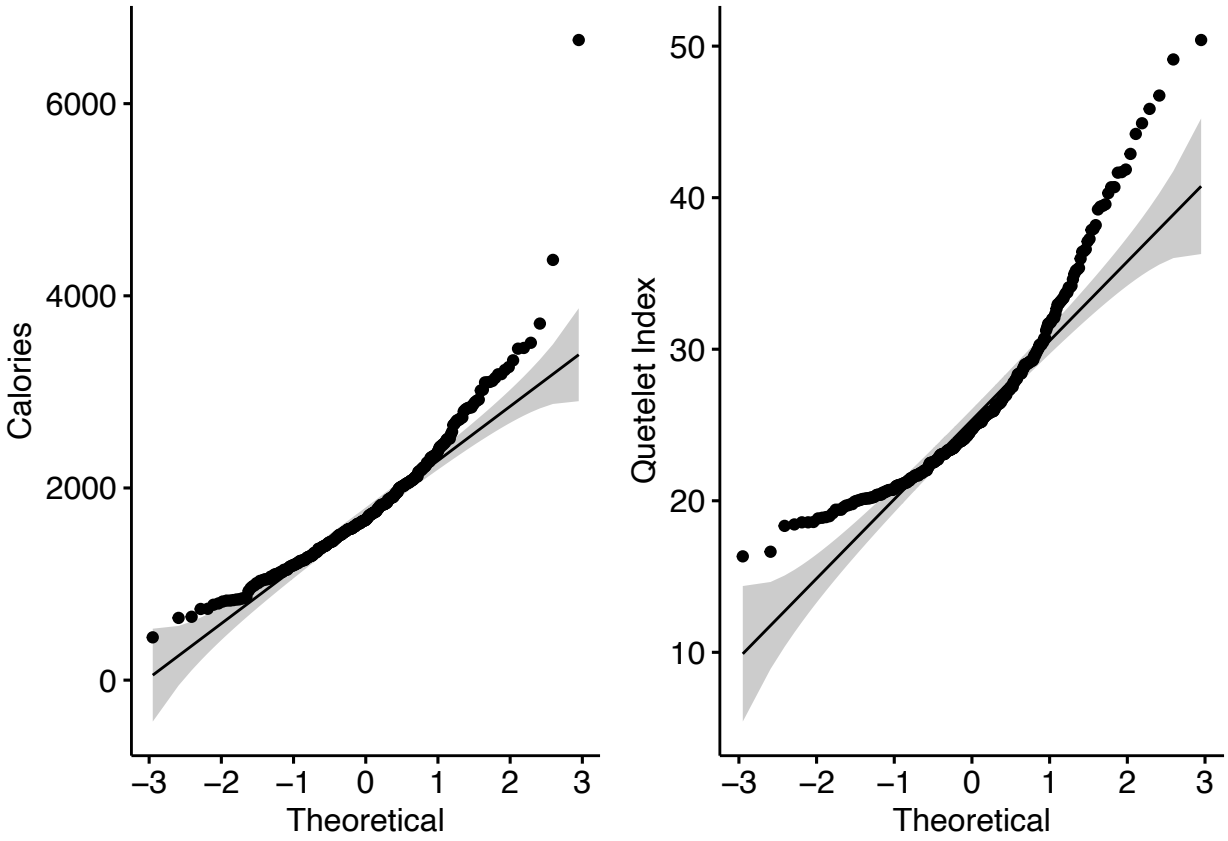
–

Question 4)

calories consumed relate to Quetelet index?

Checking normality:

```
plot10<- ggqqplot(dat4$Calories, ylab = "Calories")
plot11<- ggqqplot(dat4$Quetelet_index, ylab = "Quetelet Index")
ggarrange(plot10,plot11)
```




```
shapiro.test(dat4$Calories)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat4$Calories
## W = 0.90208, p-value = 2.092e-13
```

```
shapiro.test(dat4$Quetelet_index)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat4$Quetelet_index
## W = 0.88912, p-value = 2.312e-14
```

We reject the null as both have a very small p-value, which means we fail the normality assumption

Finding adjusted ranks

```
rank_calories<- rank(dat4$Calories, ties.method = "average")
rank_queteletindex<- rank(dat4$Quetelet_index, ties.method = "average")
```

Correlation test

```
cor.test(rank_calories, rank_queteletindex)
```

```
##
##  Pearson's product-moment correlation
##
## data:  rank_calories and rank_queteletindex
## t = -0.23176, df = 313, p-value = 0.8169
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.12342809  0.09755034
## sample estimates:
##         cor
## -0.01309881
```

We see a very high p-value, which means we fail to reject the null hypothesis, meaning the correlation could be 0. The correlation value we get is -0.01309 however the 95% confidence interval does contain 0 in it.

–

Question 5

In diet we have fat, fiber, alcohol, betadiet and retinol diet, then we also have cholesterol, calories consumed, beta carotine and retinol in plasma and quetelet

creating data frames for each with different smokers

```
df1<- data.frame("Fat"=dat5$Fat, "Prior Smoke"= dat5$PriorSmoke)
df2<- data.frame("Fiber"=dat5$Fiber, "Prior Smoke"= dat5$PriorSmoke)
df3<- data.frame("Alcohol"=dat5$Alcohol, "Prior Smoke"= dat5$PriorSmoke)
df4<- data.frame("Beta_Diet"=dat5$Beta_diet, "Prior Smoke"= dat5$PriorSmoke)
df5<- data.frame("Retinol_diet"=dat5$Retinol_diet, "Prior Smoke"= dat5$PriorSmoke)
df6<- data.frame("Cholestrol"=dat5$Cholestrol, "Prior Smoke"= dat5$PriorSmoke)
df7<- data.frame("Calories"=dat5$Calories, "Prior Smoke"= dat5$PriorSmoke)
df8<- data.frame("Beta_plasma"=dat5$Beta_plasma, "Prior Smoke"= dat5$PriorSmoke)
df9<- data.frame("Retinol_plasma"=dat5$Retinol_plasma, "Prior Smoke"= dat5$PriorSmoke)
df10<- data.frame("QueteletIndex"=dat5$Quetelet_index, "Prior Smoke"= dat5$PriorSmoke)
```

P-values for the KW test by using permutation method:

For Fat and Smoke as factors

```
df1$Prior.Smoke<- as.factor(df1$Prior.Smoke)
N<- length(df1$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df1$Fat)
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df1$Prior.Smoke==levels(df1$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df1$Prior.Smoke==levels(df1$Prior.Smoke)[i]])-b)^2
```

```
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df1$Fat)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df1$Prior.Smoke==levels(df1$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.0301
```

```
kw_obs
```

```
## [1] 6.853543
```

KW test in R

```
kruskal.test(df1$Fat~df1$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df1$Fat by df1$Prior.Smoke
## Kruskal-Wallis chi-squared = 6.8536, df = 2, p-value = 0.03249
```

p-value is less than 0.05 hence we reject the null, there significant differences present.

For Fiber and Smokes

```
df2$Prior.Smoke<- as.factor(df2$Prior.Smoke)
N<- length(df2$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df2$Fiber)
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df2$Prior.Smoke==levels(df2$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
```

```
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df2$Prior.Smoke==levels(df2$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df2$Fiber)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df2$Prior.Smoke==levels(df2$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.0056
```

```
kw_obs
```

```
## [1] 10.23197
```

KW test in R

```
kruskal.test(df2$Fiber~df2$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df2$Fiber by df2$Prior.Smoke
## Kruskal-Wallis chi-squared = 10.233, df = 2, p-value = 0.005997
```

p-value is less than 0.05 hence we reject the null < 0.05, there significant differences present.

For alcohol and Smoke as factors

```
df3$Prior.Smoke<- as.factor(df3$Prior.Smoke)
N<- length(df3$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df3$Alcohol)
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df3$Prior.Smoke==levels(df3$Prior.Smoke)[i])
}
```

```
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df3$Prior.Smoke==levels(df3$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df3$Alcohol)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df3$Prior.Smoke==levels(df3$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 3e-04
```

```
kw_obs
```

```
## [1] 14.82101
```

KW test in R

```
kruskal.test(df3$Alcohol~df3$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df3$Alcohol by df3$Prior.Smoke
## Kruskal-Wallis chi-squared = 15.513, df = 2, p-value = 0.0004279
```

p-value is less than 0.05 hence we reject the null, there significant differences present.

For beta diet and Smoke as factors

```
df4$Prior.Smoke<- as.factor(df4$Prior.Smoke)
N<- length(df4$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df4$Beta_Diet)
n<- rep(0,3)
for(i in 1:3)
{
```

```
  n[i]<-sum(df4$Prior.Smoke==levels(df4$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df4$Prior.Smoke==levels(df4$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df4$Beta_Diet)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df4$Prior.Smoke==levels(df4$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.0152
```

```
kw_obs
```

```
## [1] 8.318955
```

KW test in R

```
kruskal.test(df4$Beta_Diet~df4$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df4$Beta_Diet by df4$Prior.Smoke
## Kruskal-Wallis chi-squared = 8.319, df = 2, p-value = 0.01562
```

p-value is < 0.05 hence we reject the null, there significant differences present.

For retinol diet and Smoke as factors

```
df5$Prior.Smoke<- as.factor(df5$Prior.Smoke)
N<- length(df5$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df5$Retinol_diet)
```

```
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df5$Prior.Smoke==levels(df5$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df5$Prior.Smoke==levels(df5$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df5$Retinol_diet)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df5$Prior.Smoke==levels(df5$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.8585
```

```
kw_obs
```

```
## [1] 0.3146094
```

KW test in R

```
kruskal.test(df5$Retinol_diet~df5$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df5$Retinol_diet by df5$Prior.Smoke
## Kruskal-Wallis chi-squared = 0.31461, df = 2, p-value = 0.8544
```

p-value is > 0.05, we fail to reject the null and conclude that there are no significant differences.

For cholestrol and Smoke as factors

```
df6$Prior.Smoke<- as.factor(df6$Prior.Smoke)
N<- length(df6$Prior.Smoke)
```

```
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df6$Cholestrol)
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df6$Prior.Smoke==levels(df6$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df6$Prior.Smoke==levels(df6$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df6$Cholestrol)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df6$Prior.Smoke==levels(df6$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.057
```

```
kw_obs
```

```
## [1] 5.649126
```

KW test in R

```
kruskal.test(df6$Cholestrol~df6$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df6$Cholestrol by df6$Prior.Smoke
## Kruskal-Wallis chi-squared = 5.6491, df = 2, p-value = 0.05933
```

p-value is > 0.05 hence we fail to reject the null, there no significant differences present.

For calories and Smoke as factors

```
df7$Prior.Smoke<- as.factor(df7$Prior.Smoke)
N<- length(df7$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df7$Calories)
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df7$Prior.Smoke==levels(df7$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df7$Prior.Smoke==levels(df7$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df7$Calories)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df7$Prior.Smoke==levels(df7$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.1909
```

```
kw_obs
```

```
## [1] 3.295021
```

KW test in R

```
kruskal.test(df7$Calories~df7$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df7$Calories by df7$Prior.Smoke
## Kruskal-Wallis chi-squared = 3.295, df = 2, p-value = 0.1925
```

p-value is > 0.05 hence we fail to reject the null, and there are no significant differences present.

For Beta in plasma and Smoke as factors

```
df8$Prior.Smoke<- as.factor(df8$Prior.Smoke)
N<- length(df8$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df8$Beta_plasma)
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df8$Prior.Smoke==levels(df8$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df8$Prior.Smoke==levels(df8$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df8$Beta_plasma)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df8$Prior.Smoke==levels(df8$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.0018
```

```
kw_obs
```

```
## [1] 12.60545
```

KW test in R

```
kruskal.test(df8$Beta_plasma~df8$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df8$Beta_plasma by df8$Prior.Smoke
## Kruskal-Wallis chi-squared = 12.606, df = 2, p-value = 0.001831
```

p-value is < 0.05 hence we reject the null, there significant differences present.

For Retinol and Smoke as factors

```
df9$Prior.Smoke<- as.factor(df9$Prior.Smoke)
N<- length(df9$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df9$Retinol_plasma)
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df9$Prior.Smoke==levels(df9$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df9$Prior.Smoke==levels(df9$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df9$Retinol_plasma)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df9$Prior.Smoke==levels(df9$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.0323
```

```
kw_obs
```

```
## [1] 6.717488
```

KW test in R

```
kruskal.test(df9$Retinol_plasma~df9$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df9$Retinol_plasma by df9$Prior.Smoke
## Kruskal-Wallis chi-squared = 6.7176, df = 2, p-value = 0.03478
```

p-value is less than 0.05 hence we reject the null, there significant differences present.

For Quentelet Index and Smoke as factors

```
df10$Prior.Smoke<- as.factor(df10$Prior.Smoke)
N<- length(df10$Prior.Smoke)
a<- 12/(N*(N+1))
b<- (N+1)/2
rank_obs<- rank(df10$QueteletIndex)
n<- rep(0,3)
for(i in 1:3)
{
  n[i]<-sum(df10$Prior.Smoke==levels(df10$Prior.Smoke)[i])
}
kw_obs= 0
for(i in 1:3)
{
  kw_obs=kw_obs+n[i]*(mean(rank_obs[df10$Prior.Smoke==levels(df10$Prior.Smoke)[i]])-b)^2
}
kw_obs<-kw_obs*a
cnt=0
for(i in 1:10000)
{
  permut<-sample(df10$QueteletIndex)
  prank<-rank(permut)
  kw<-0
  for(j in 1:3)
  {
    kw=kw+n[j]*(mean(prank[df10$Prior.Smoke==levels(df10$Prior.Smoke)[j]])-b)^2
  }
  kw<-kw*a
  if(kw>kw_obs)
  {
    cnt=cnt+1
  }
}
cnt/10000
```

```
## [1] 0.3094
```

```
kw_obs
```

```
## [1] 2.397501
```

KW test in R

```
kruskal.test(df10$QueteletIndex~df10$Prior.Smoke)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df10$QueteletIndex by df10$Prior.Smoke
## Kruskal-Wallis chi-squared = 2.3975, df = 2, p-value = 0.3016
```

p-value is > 0.05 hence we fail to reject the null, and can conclude that there are no significant differences present.

–

Question 6)

creating the dataframes:

```
dfr1<- data.frame("Fat"=dat6$Fat,"Gender"=dat6$Gender)
dfr2<- data.frame("Fiber"=dat6$Fiber,"Gender"=dat6$Gender)
dfr3<- data.frame("Alcohol"=dat6$Alcohol,"Gender"=dat6$Gender)
dfr4<- data.frame("Beta_diet"=dat6$Beta_diet,"Gender"=dat6$Gender)
dfr5<- data.frame("Retinol_diet"=dat6$Retinol_diet,"Gender"=dat6$Gender)
dfr6<- data.frame("Cholestrol"=dat6$Cholestrol,"Gender"=dat6$Gender)
dfr7<- data.frame("Calories"=dat6$Calories,"Gender"=dat6$Gender)
dfr8<- data.frame("Beta_plasma"=dat6$Beta_plasma,"Gender"=dat6$Gender)
dfr9<- data.frame("Retinol_plasma"=dat6$Retinol_plasma,"Gender"=dat6$Gender)
dfr10<- data.frame("QueteletIndex"=dat6$Quetelet_index,"Gender"=dat6$Gender)
```

For Fat and Gender as factors

```
wilcox.test(dfr1$Fat[dfr1$Gender=="Female"],dfr1$Fat[dfr1$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr1$Fat[dfr1$Gender == "Female"] and dfr1$Fat[dfr1$Gender == "Male"]
## W = 3624.5, p-value = 0.0001249
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is < 0.05 we reject the null and hence there are differences in the two groups for fat

For Fiber and Gender as factors

```
wilcox.test(dfr2$Fiber[dfr2$Gender=="Female"],dfr2$Fiber[dfr2$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr2$Fiber[dfr2$Gender == "Female"] and dfr2$Fiber[dfr2$Gender == "Male"]
## W = 5187.5, p-value = 0.3213
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is > 0.05 we fail to reject the null and hence there are no significant differences in the two groups for fiber

For Alcohol and Gender as factors

```
wilcox.test(dfr3$Alcohol[dfr3$Gender=="Female"],dfr3$Alcohol[dfr3$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr3$Alcohol[dfr3$Gender == "Female"] and dfr3$Alcohol[dfr3$Gender == "Male"]
## W = 4685, p-value = 0.05114
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is > 0.05 we fail to reject the null and hence there are no significant differences in the two groups for alcohol

For beta in diet and Gender as factors

```
wilcox.test(dfr4$Beta_diet[dfr4$Gender=="Female"],dfr4$Beta_diet[dfr4$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr4$Beta_diet[dfr4$Gender == "Female"] and dfr4$Beta_diet[dfr4$Gender == "Male"]
## W = 5165, p-value = 0.3017
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is > 0.05 we fail to reject the null and hence there are no differences in the two groups for beta in diet

For retinol in diet and Gender as factors

```
wilcox.test(dfr5$Retinol_diet[dfr5$Gender=="Female"],dfr5$Retinol_diet[dfr5$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr5$Retinol_diet[dfr5$Gender == "Female"] and dfr5$Retinol_diet[dfr5$Gender == "Male"]
## W = 4938, p-value = 0.1482
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is > 0.05 we fail to reject the null and hence there are no differences in the two groups for retinol in diet

For cholestrol and Gender as factors

```
wilcox.test(dfr6$Cholestrol[dfr6$Gender=="Female"],dfr6$Cholestrol[dfr6$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr6$Cholestrol[dfr6$Gender == "Female"] and dfr6$Cholestrol[dfr6$Gender == "Male"]
## W = 3215, p-value = 4.616e-06
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is < 0.05 we reject the null and hence there are significant differences in the two groups for cholestrol

For calories and Gender as factors

```
wilcox.test(dfr7$Calories[dfr7$Gender=="Female"],dfr7$Calories[dfr7$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr7$Calories[dfr7$Gender == "Female"] and dfr7$Calories[dfr7$Gender == "Male"]
## W = 3788, p-value = 0.000402
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is < 0.05 we reject the null and hence there are significant differences in the two groups for calories

For beta in plasma and Gender as factors

```
wilcox.test(dfr8$Beta_plasma[dfr8$Gender=="Female"],dfr8$Beta_plasma[dfr8$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr8$Beta_plasma[dfr8$Gender == "Female"] and dfr8$Beta_plasma[dfr8$Gender == "Male"]
## W = 7087, p-value = 0.01377
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is < 0.05 we reject the null and hence there are significant differences in the two groups for beta in plasa

For retinol in plasma and Gender as factors

```
wilcox.test(dfr9$Retinol_plasma[dfr9$Gender=="Female"],dfr9$Retinol_plasma[dfr9$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr9$Retinol_plasma[dfr9$Gender == "Female"] and dfr9$Retinol_plasma[dfr9$Gender == "Male"]
## W = 4470, p-value = 0.02158
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is < 0.05 we reject the null and hence there are significant differences in the two groups for retinol in plasma

For Quetelet index and Gender as factors

```
wilcox.test(dfr10$QueteletIndex[dfr10$Gender=="Female"],dfr10$QueteletIndex[dfr10$Gender=="Male"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dfr10$QueteletIndex[dfr10$Gender == "Female"] and dfr10$QueteletIndex[dfr10$Gender == "Male"]
## W = 4990.5, p-value = 0.1769
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is > 0.05 we fail to reject the null and hence there are no significant differences in the two groups for beta in diet