Parametric vs Nonparametric methods: When is it usefull?

Filip Wilhelm Sjostrand

December 02, 2022

Abstract

The report covers approaches to identify the appropriateness of parametric vs. nonparametric methods in statistics. The methods discuss essential aspects of the underlying assumptions used in parametric statistics and how to overcome certain violations. The results derived in the report answer six questions regarding health and diet.

Introduction

This study aims to understand, contrast, and compare different parametric and nonparametric statistical methods from a practical viewpoint. The report will incorporate dietary and health data from 315 people sampled in a cross-sectional survey. The experimental units are people who have had non-cancerous lesions removed through surgery. The investigation will analyze the data such that when violations of parametric assumptions, a justified nonparametric method is implemented. Following that, we shall devise proper tests to answer the following questions:

1. Is there an association between beta carotene and retinol in diet and beta carotene and retinol in plasma?
2. Is there a relationship between beta carotine in plasma and cholesterol?
3. Is there an association between fat in diet and cholesterol and also between fiber in diet and cholesterol?
4. How does calories consumed relate to Quetelet index?
5. Are there differences between current smokers, former smokers and never smokers with respect to any of the dietary variables (what is in the diet), cholesterol, calories consumed, beta carotine and retinol in plasma and quetelet?

Methods/Results

Question 1

Diet and plasma levels for beta-carotene and retinol are bivariate sampled continuous variables, assuming that the amount in the diet is not assigned ton each experimental unit. Therefore, we could use Pearson Moment Correlation to estimate their association. However, this requires the data to be bivariate normal. If a bivariate dataset is univariate normal in any direction and the scatterplot is an ellipse-shaped data point

cloud, it indicates bivariate normality. Considering the output of *Graph 1*, we find that both BetaDiet and BetaPlasma appear to have heavier right tails, and similarly for RetinolDiet, but RetionlPlasma appears approximately normal. Thus, one could assume that RetinolDiet and RetinolPlasma are bivariate normal. However, observing *Graph 2* and *Graph 3*, we find neither contains a well-defined elliptical shape.

Therefore, it is reasonable to utilize a nonparametric method. Since we can see in *Graph 1* that there are heavy tails for both BetaDiet and BetaPlasma and RetinolDiet and RetinolPlasma, we should consider using a ranked correlation test. The ranks will offset any effects the outliers might have on the association. Each sample contains ties in the ranks. Thus, we have to apply the average rank for each tie. Since we have a large data set, we will utilize the normal approximation (instead of a permutation test) of Spearman's Rank Correlation.

Table 1: Association Between Diet and Plasma

	H0	Ha	Correlation	Test.Statistic	p.value
Beta-carotene	$\rho = 0$	$\rho \neq 0$	r_s = 0.1786	Z = 3.165	0.0016
Retinol	$\rho = 0$	$\rho \neq 0$	r_s = -0.0381	Z = -0.6745	0.5000

From *Table 1* we know that there is a significant association between beta-carotene in diet and plasma but not between retinol in diet and plasma. The relationship for beta carotene is weakly positive, indicating that an increase in dietary consumption slightly increases plasma levels.

Question 2

In *Graph 4*, we find that Cholesterol (continuous variable, bivariate sampled assuming not assigned to the experimental unit) is approximately normally distributed and, therefore, potentially applicable to use Pearson's correlation. However, looking at *Graph 5*, we find that the relationship between Cholesterol and BetaDiet appears to have an inverse exponential nature. Once again, using the ranks (with consideration to any ties) allows us to analyze the association. Since we have many observations we will use the normal approximation, by the central limit theorem.

Table 2: Relationship of Beta-Carotene in plasma and cholesterol

H0	Ha	Correlation	Test.Statistic	p.value
$\rho = 0$	$\rho \neq 0$	r_s = -0.1425	Z = -2.5256	0.0116

Table 2 indicates an insignificant relationship between Beta-Carotene in plasma and cholesterol levels. Hence, given no association, we need more evidence to support the data's gesture to have an inverse exponential relationship.

Question 3

We see in *Graph 6* that both variables have a Gaussian curve with some right tails, and the elliptical shape appears in *Graph 7* and *Graph 8*. Hence, the parametric Pearson's moment correlation is a practical choice here.

Table 3: Fiber/Fat consumption and Cholestrol

	H0	Ha	Correlation	Test.Statistic	p.value
Fat	$\rho = 0$	$\rho \neq 0$	r = 0.7098	t = 17.8298	0.0000
Fiber	$\rho = 0$	$\rho \neq 0$	r = 0.154	t = 2.7569	0.0062

From *Table 3* we find that Fat and Fiber have a significant positive relationship with Cholesterol. Fiber has a weak positive relationship, while Fat has a medium positive relationship with Cholesterol. This suggest that Fat has a bigger association with higher Cholesterol than Fiber does.

Question 4

Since the construction of the Quetelet index (BMI) is $\frac{Weight}{Height^2}$, we know that calories consumed are not independent of the variable since, on average, a person with a larger BMI should consume more calories than someone with a smaller BMI. Therefore we consider this as "Fixed X-sampling" and should use regression to assess the relationship.

Considering *Graph 9*, we find that Calories are right-skewed and that BMI has many observations on its left tail. Hence, the data does not follow bivariate normality. *Graph 10* also confirms the same, given a lack of elliptic shape. Hence, the t-test is invalid now, and we should use a permutation test for the least square estimate of β_1.

Table 4: Permuation Test of Slope

H0	Ha	Test.Statistic	p.value
$\beta_1 = 0$	$\beta_1 \neq 0$	$\hat{\beta}_1 = 0.399$	0.9479

From our permutation test we observe in *Table 4* that we fail to reject H_0. Thus, concluding that there is insufficient evidence of a significant slope between BMI and calories consumed.

Question 5

We could interpret each smoking category as a factor and therefore analyze using k-sample methods. The ANOVA model assumes that the errors have a normal distribution, equal variance, and independence. We know from previous analyzes that Fat, Fiber, Cholesterol, and RetinolPlasma, have some slight deviation from normality and thus should be further looked into if the assumption of equal variance holds. BMI, BetaDiet, BetaPlasma, and RetinolDiet need nonparametric methods.

Parametric

Table 5: Levene Test for Equal Variance

	Result	Reject
Fat	0.1883182	FALSE

	Result	Reject
Fiber	0.6578468	FALSE
Cholesterol	0.5182365	FALSE
RetinolPlasma	0.2688455	FALSE

When testing for equal variance we use the Levene Test. The hypothesis is: $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ vs. $H_a : \sigma_i^2 \neq \sigma_j^2$ for at least one pair (i, j). The output of the test is found in *Table 5* and thus we can confirm that all of the following variables have equal variance.

Table 6: ANOVA Result

	F	p-value
Fat	3.420329	0.0339328
Fiber	4.308052	0.0142694
Cholesterol	2.234710	0.1087336
RetinolPlasma	3.790757	0.0236254

Table 6 Informs that only Cholesterol did not have a mean that is significantly different from the rest. We shall further the analysis by using familywise comparisons to find which mean differs significantly from the other. Since we have not made any prespecified comparisons, we shall refer to Tukey's HSD to preserve the Type-I error.

Table 7: Parametric Familiy of Comparison (significant results)

Comparisons	diff	lwr	upr	p.adj	variable
3-1	-2.563	-4.701	-0.4251	0.01398	fiber
3-2	-2.508	-4.729	-0.2877	0.02232	fiber
2-1	60.940	1.087	120.8000	0.04490	RetinolPlasma

From *Table 7* we derive that there is a significant difference between current smokers (3) and former smokers (2) and also current smokers (3) and never smokers (1) in terms of fiber consumption. There is also a significant difference in Retinol Plasma levels between former and never smokers.

Nonparametric

The two options we have here is either a permutation test based on the F statistic or the Kruskal-Wallis which is a permutation ANOVA based on the ranks. We will utilize the Permutation F-test.

Table 8: Permutation ANOVA

	Statistic	Permutation p-value
Quetelet	2.0725404	0.133
BetaDiet	3.1048310	0.044

	Statistic	Permutation p-value
BetaPlasma	3.7154508	0.027
RetinolDiet	0.1060849	0.892

Table 9: Permutation Familywise Comparison (Only Significant)

Observed Difference	Permutation p-value	variable	Means
-651.6912	0.011	Diet	2-3

Table 8 informs us that Beta-carotene in diet and Beta-carotene in Plasma are the only one who has a significantly different mean. Analyzing which one we find in *Table 9* that only former smokers are significantly different from current smokers in terms of dietary Beta-Carotene.

Discussion

The importance of diet's effect on our health is always at the top of the agenda for national health. The results in the report can convey a message on how our consumptions affect our health. First, it is interesting that Beta carotene and retinol consumption affect the plasma values differently. Wagle (2020) explains that beta-carotene is a precursor to retinol (vitamin A) which occurs in plant sources instead of fat animal products as for the latter. The author also claims that retinol is much more potent than beta-carotene. Hence, intuitively one should expect that higher consumption of retinol should increase the values found in plasma. Therefore, looking further into why our observations occurred could be interesting.

High Cholesterol levels arise due to unhealthy diets, particularly those rich in unsaturated fats (Megan, 2021). Since we know that plant-based foods (low in saturated fats) are good sources of beta-carotene, we could see that the association we found in the data might be explainable. If a person has higher levels of beta carotene, they are more likely to consume a better diet and, therefore, have lower cholesterol levels.

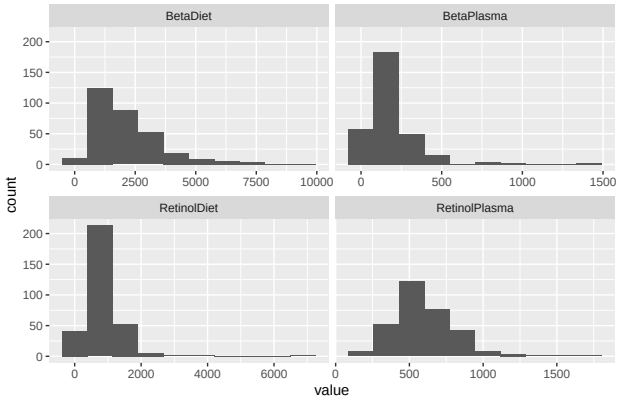
As mentioned, the higher cholesterol associated with increased fat consumption should not be unpredictable. However, it contradicts the previous paragraph that an increase in fiber positively affects cholesterol. Warwick (2022) confirms that increasing fiber intake reduces cholesterol. Hence, further research should look into, e.g., if the post-operation has harmed the body's ability to decrease cholesterol. The circumstances might also explain why we do not see a positive linear relationship between BMI and Calories consumed.

Lastly, it does not come as a surprise that smokers tend to consume fewer fibers. However, that contradicts that they consume more beta-carotene (plant-based) than former smokers. Nevertheless, the former smokers in our data consume more retinol (animal-based) than nonsmokers; therefore, we might see this discrepancy. A hypothetical cause is that former smokers indulge more in food due to a lack of dopamine previously gained from smoking.

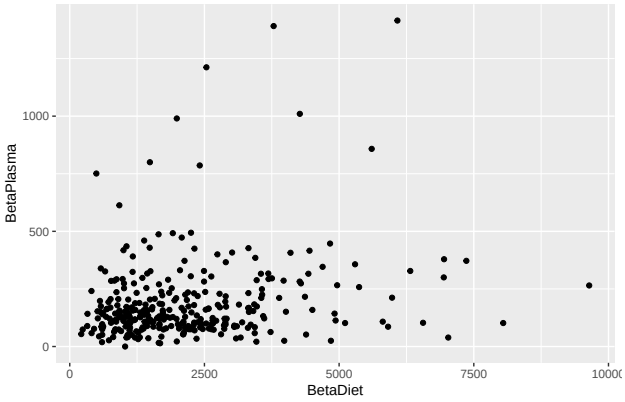
Appendix

Graphs

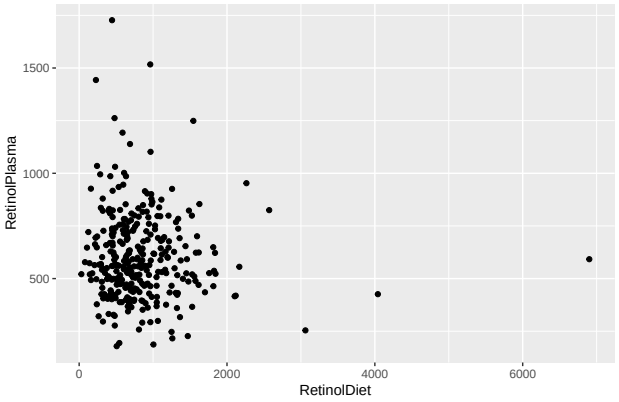
Graph 1: Distribution of Beta−Carotene and Retinol Variables

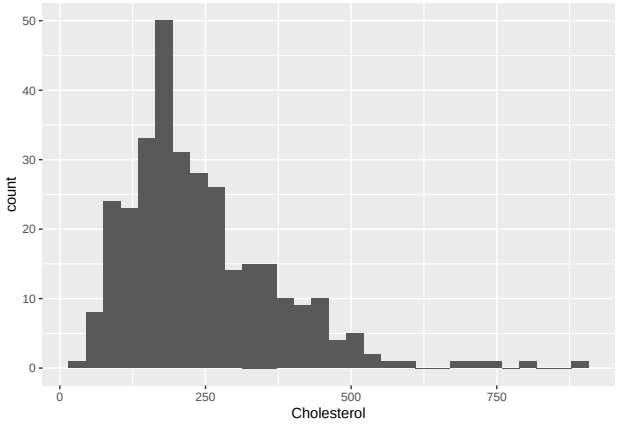


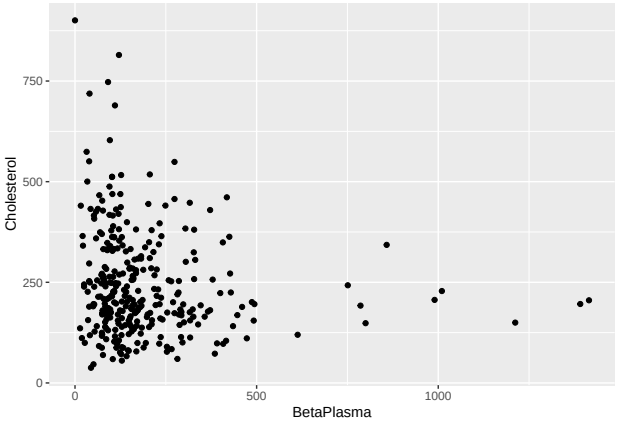
Graph 2: Beta−Carotene in Plasma vs. Diet



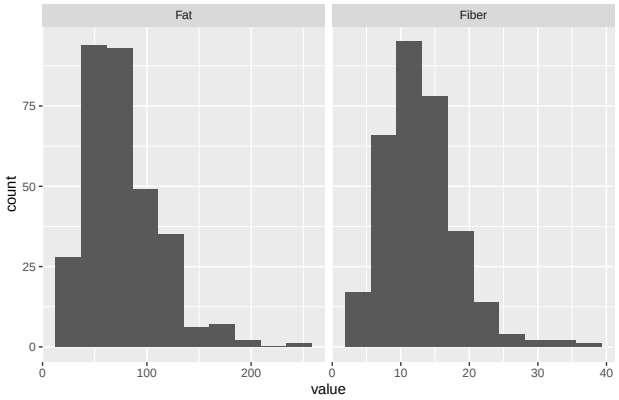
Graph 3: Retinol in Plasma vs. Diet

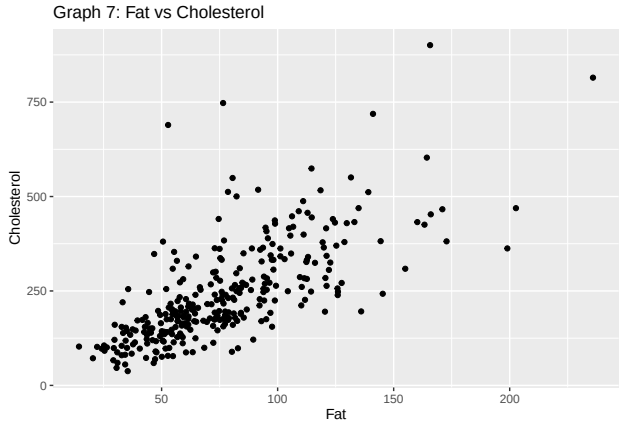


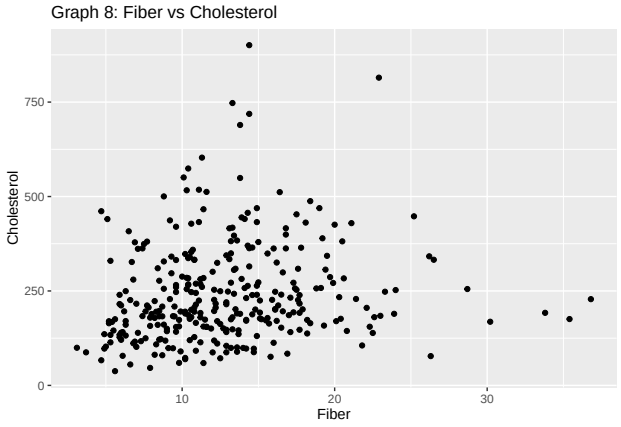


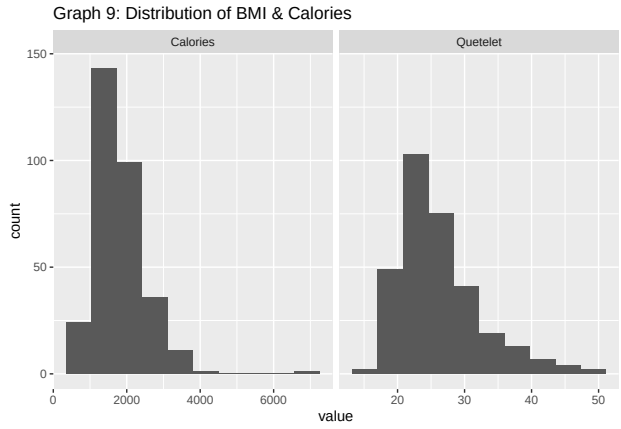


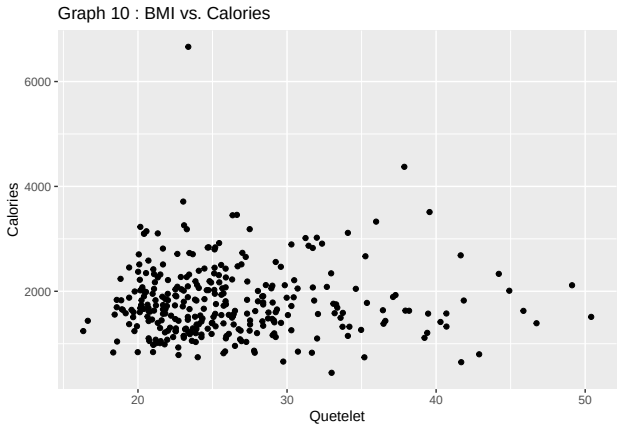
Graph 6: Distribution of Fat & Fiber











Sources

- Megan, S. (2021). Everything You Need to Know About High Cholesterol, Healthline, Available Online: https://www.healthline.com/health/high-cholesterol [Accessed 1 December 2022].
- Wagle, K. (2020). 14 Differences Between Retinol and β -Carotene, Public Health Notes, Available Online: https://www.publichealthnotes.com/ [Accessed 1 December 2022].
- Warwick, W. (2022). Fiber and Cholesterol: Is There a Link?, Healthline, Available Online: https://www.healthline.com/nutrition/fiber-and-cholesterol [Accessed 1 December 2022].

Code

```
# Packages ---------
library(tidyr)
library(readr)
library(dplyr)
library(ggplot2)
library(knitr)

# Changing the default theme
```

```r
theme_set(theme_grey())

# Get & tidy data ----------
df <- read_delim(
  file = "~/Documents/UC Davis/Courses/STA 104/Project/nutritionstudy.csv",
  delim = ";",
  escape_double = FALSE,
  col_types = cols(
    Smoke = col_factor(levels = c("No", "Yes")),
    Vitamin = col_factor(levels = c("1", "2", "3")),
    Gender = col_factor(levels = c("Male", "Female")),
    VitaminUse = col_factor(levels = c("Regular", "Occasional", "No")),
    PriorSmoke = col_factor(levels = c("1", "2", "3"))),
  trim_ws = TRUE
  )


df$Alcohol <- as.numeric(gsub(",", ".", df$Alcohol))
df <- df %>% select(-ID)

# Distribution of Beta & Retinol variables ----------
g1 <- df %>%
  select(BetaDiet, BetaPlasma, RetinolDiet, RetinolPlasma) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x') +
  labs(title="Graph 1: Distribution of Beta-Carotene and Retinol Variables")

# Bivariate Normal Beta -----------
g2 <- df %>%
  ggplot(aes(BetaDiet, BetaPlasma)) +
  geom_point() +
  labs(title="Graph 2: Beta-Carotene in Plasma vs. Diet")

# Bivariate Normal Retinol -----------
g3 <- df %>%
  ggplot(aes(RetinolDiet, RetinolPlasma)) +
  geom_point() +
  labs(title="Graph 3: Retinol in Plasma vs. Diet")

# Result Beta ----------
diet <- rank(df$BetaDiet, ties.method="average")
```

```r
plasma <- rank(df$BetaPlasma, ties.method="average")
meand <- mean(diet)
meanp <- mean(plasma)

r <- sum((diet-meand)*(plasma-meanp))/sqrt(sum((diet-meand)**2)*sum((plasma-meanp)**2))
z <- r*sqrt(nrow(df)-1)
p <- round(2*pnorm(z, lower.tail = FALSE),4)

null <- "$\\rho = 0$"
alt <-"$\\rho \\ne 0$"

result1 <- data.frame(
  H0 = null,
  Ha =  alt,
  Correlation = paste("$r_s$ = ", as.character(round(r,4))),
  Test.Statistic = paste("Z = ", as.character(round(z,4))),
  p.value = p
  )

# Result Retinol ----------
diet <- rank(df$RetinolDiet, ties.method="average")
plasma <- rank(df$RetinolPlasma, ties.method="average")
meand <- mean(diet)
meanp <- mean(plasma)

r <- sum((diet-meand)*(plasma-meanp))/sqrt(sum((diet-meand)**2)*sum((plasma-meanp)**2))
z <- r*sqrt(nrow(df)-1)
p <- round(2*pnorm(z),4)

result1 <- result1 %>%
  rbind(list(
    null,
    alt,
    paste("$r_s$ = ", as.character(round(r,4))),
    paste("Z = ", as.character(round(z,4))),
    p
    )
  )

rownames(result1) <- c("Beta-carotene", "Retinol")

# Cholesterol distribution ----------
g4 <- df %>%
```

```r
  ggplot(aes(Cholesterol)) +
  geom_histogram() +
  labs(titel="Graph 4: Cholesterol Distribution")

# Cholesterol BetaDiet scatterplot ----------
g5 <- df %>%
  ggplot(aes(BetaPlasma, Cholesterol)) +
  geom_point() +
  labs(titel="Graph 5: Dietary Beta-Carotene vs. Cholesterol")

# Result BetaDiet Cholesterol ----------
plasma <- rank(df$BetaPlasma, ties.method="average")
chol <- rank(df$Cholesterol, ties.method="average")
meanp <- mean(plasma)
meanc <- mean(chol)

r <- sum((plasma-meanp)*(chol-meanc))/sqrt(sum((plasma-meanp)**2)*sum((chol-meanc)**2))
z <- r*sqrt(nrow(df)-1)
p <- round(2*pnorm(z),4)

null <- "$\\rho = 0$"
alt <-"$\\rho \\ne 0$"

result2 <- data.frame(
  H0 = null,
  Ha =  alt,
  Correlation = paste("$r_s$ = ", as.character(round(r,4))),
  Test.Statistic = paste("Z = ", as.character(round(z,4))),
  p.value = p
  )

# Distribution of Beta & Retinol variables ----------
g6 <- df %>%
  select(Fiber, Fat) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x') +
  labs(title="Graph 6: Distribution of Fat & Fiber")

# Result Fat ----------
res <- cor.test(df$Fat, df$Cholesterol)
```

```r
result3 <- data.frame(
  H0 = null,
  Ha =  alt,
  Correlation = paste("r = ", as.character(round(res$estimate,4))),
  Test.Statistic = paste("t = ", as.character(round(res$statistic,4))),
  p.value = round(res$p.value, 4)
  )

# Result Fiber ----------
res <- cor.test(df$Fiber, df$Cholesterol)

result3 <- result3 %>%
  rbind(list(
    null,
    alt,
    paste("r = ", as.character(round(res$estimate,4))),
    paste("t = ", as.character(round(res$statistic,4))),
    round(res$p.value, 4)
    )
  )

rownames(result3) <- c("Fat", "Fiber")

# Fat, Fiber v. Cholesterol scatterplot ----------
g7 <- df %>%
  ggplot(aes(Fat, Cholesterol)) +
  geom_point() +
  labs(title="Graph 7: Fat vs Cholesterol")

g8 <- df %>%
  ggplot(aes(Fiber, Cholesterol)) +
  geom_point() +
  labs(title="Graph 8: Fiber vs Cholesterol")

# Distribution of Beta & Retinol variables ----------
g9 <- df %>%
  select(Quetelet, Calories) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x') +
  labs(title="Graph 9: Distribution of BMI & Calories")
```

```r
g10 <- df %>%
  ggplot(aes(Quetelet, Calories)) +
  geom_point() +
  labs(title="Graph 10 : BMI vs. Calories")

# Permutation test ----------
cal <- df$Calories
bmi <- df$Quetelet
b_obs  <- lm(cal~bmi)$coefficients[2]

tot <- 10000
p <- c()
for(i in 1:tot){
  perm <- sample(bmi)
  b <- lm(cal~perm)$coefficients[2]
  p[i] <- (abs(b) >= abs(b_obs))+0
}
pval <- sum(p)/tot

result4 <- data.frame(
  H0 = "$\\beta_1=0$",
  Ha =  "$\\beta_1 \\ne 0$",
  Test.Statistic = paste("$\\hat{\\beta}_1$ = ", as.character(round(b_obs,4))),
  p.value = round(pval, 4)
  )


# Test for equal variance ----------
equal <- data.frame(Result = c(
  car::leveneTest(Fat~PriorSmoke, df)$`Pr(>F)`[1],
  car::leveneTest(Fiber~PriorSmoke, df)$`Pr(>F)`[1],
  car::leveneTest(Cholesterol~PriorSmoke, df)$`Pr(>F)`[1],
  car::leveneTest(RetinolPlasma~PriorSmoke, df)$`Pr(>F)`[1]
  )
)
rownames(equal) <- c("Fat", "Fiber", "Cholesterol", "RetinolPlasma")
equal <- equal %>% mutate(Reject = case_when(Result < 0.05 ~ TRUE, Result >= 0.05 ~ FALSE))
kable(equal, caption = "Levene Test for Equal Variance")

# ANOVA ----------
Fat <- c(
  anova(lm(Fat~PriorSmoke, df))$`F value`[1],
  anova(lm(Fat~PriorSmoke, df))$`Pr(>F)`[1]
```

```
  )
Fiber <- c(
  anova(lm(Fiber~PriorSmoke, df))$`F value`[1],
  anova(lm(Fiber~PriorSmoke, df))$`Pr(>F)`[1]
  )
Cholesterol <- c(
  anova(lm(Cholesterol~PriorSmoke, df))$`F value`[1],
  anova(lm(Cholesterol~PriorSmoke, df))$`Pr(>F)`[1]
  )
RetinolPlasma <- c(
  anova(lm(RetinolPlasma~PriorSmoke, df))$`F value`[1],
  anova(lm(RetinolPlasma~PriorSmoke, df))$`Pr(>F)`[1]
  )

res1 <- data.frame(Fat, Fiber, Cholesterol, RetinolPlasma)

res1 <- t(res1)
colnames(res1) <- c("F", "p-value")

# Pairwise comparison ----------
fat   <-  data.frame(signif(TukeyHSD(aov(Fat ~ PriorSmoke, df))$PriorSmoke,4))
fiber <-  data.frame(signif(TukeyHSD(aov(Fiber ~ PriorSmoke, df))$PriorSmoke, 4))
retpl <-  data.frame(signif(TukeyHSD(aov(RetinolPlasma ~ PriorSmoke, df))$PriorSmoke, 4))
fat <- tibble::rownames_to_column(fat, "Comparisons")
fiber <- tibble::rownames_to_column(fiber, "Comparisons")
retpl <- tibble::rownames_to_column(retpl, "Comparisons")

res <- fat %>%
  rbind(fiber, retpl) %>%
  cbind(variable = c(rep("fat", 3), rep("fiber", 3), rep("RetinolPlasma", 3))) %>%
  filter(p.adj < 0.05)

# Permutation F test ----------
permutation <- function(variable){
  tot=1000
  p <- c()
  fobs <- anova(lm(variable~df$PriorSmoke))$`F value`[1]
  for(i in 1:tot){
    perm <- sample(variable)
    f <- anova(lm(perm~df$PriorSmoke))$`F value`[1]
    p[i] <- (f>=fobs)+0
  }
  pval <- sum(p)/tot
```

```
  return(c(fobs, pval))
}

res2 <- data.frame(
  permutation(df$Quetelet),
  permutation(df$BetaDiet),
  permutation(df$BetaPlasma),
  permutation(df$RetinolDiet)
)

res2 <- data.frame(t(res2))
row.names(res2) <- c("Quetelet", "BetaDiet", "BetaPlasma", "RetinolDiet")
colnames(res2) <- c("Statistic", "Permutation p-value")

# Pairwise comparison permutation ----------
pair_permut <- function(variable){
  diff_obs <- abs(data.frame(TukeyHSD(aov(variable ~ PriorSmoke, df))$PriorSmoke)$diff)

  tot=1000
  Q <- c()
  p1 <- c()
  p2 <- c()
  p3 <- c()
  for(i in 1:tot){
    perm <- sample(df$BetaDiet)
    dfp <- data.frame(smoke=df$PriorSmoke, perm)
    dfp <- dfp %>%
      group_by(smoke) %>%
      summarise(mean=mean(perm))
    diff <- c(
      abs(dfp[1,2]-dfp[2,2]),
      abs(dfp[1,2]-dfp[3,2]),
      abs(dfp[2,2]-dfp[3,2])
      )
    Q[i] <- max(unlist(diff))
    p1[i] <- (diff[1]>=diff_obs[1])+0
    p2[i] <- (diff[2]>=diff_obs[2])+0
    p3[i] <- (diff[3]>=diff_obs[3])+0
  }
  critical <- quantile(Q, 0.95)
  pval1 <- sum(p1)/tot
  pval2 <- sum(p2)/tot
  pval3 <- sum(p3)/tot
```

```
  dfr <- data.frame(data.frame(TukeyHSD(aov(variable ~ PriorSmoke, df))$PriorSmoke)$diff,pval=c(pval1, p
  colnames(dfr) <- c("Observed Difference", "Permutation p-value")
  return(dfr)
}

res3 <- pair_permut(df$BetaDiet) %>%
  rbind(pair_permut(df$BetaPlasma)) %>%
  cbind(variable=c(rep("Diet", 3), rep("Plasma", 3))) %>%
  cbind(Means=rep(c("1-2", "1-3", "2-3"),2)) %>%
  filter(`Permutation p-value` < 0.05)

# Print Graphs ----------
g1
g2
g3
g4
g5
g6
g7
g8
g9
g10
```