Attachment B

Import of the datasets

```
df_pre <- read.csv("df_pre.csv")
df_post <- read.csv("df_post.csv")
```

```
df_pre = query(
" SELECT *
  FROM df_pre
  ORDER BY Year, Month;"
)
df_pre_m = query(
" SELECT Month,Year, COUNT(*)
  FROM df_pre
  GROUP BY Year,Month;"
)

df_post = query(
" SELECT *
  FROM df_post
  ORDER BY Year, Month;"
)
df_post_m = query(
" SELECT Month,Year, COUNT(*)
  FROM df_post
  GROUP BY Year,Month;"
)

df_h = as.data.frame(cbind(df_pre_m[,3],df_post_m[,3]))
colnames(df_h) = c("before","after") # columns names
```

horizontal monthly dataframe

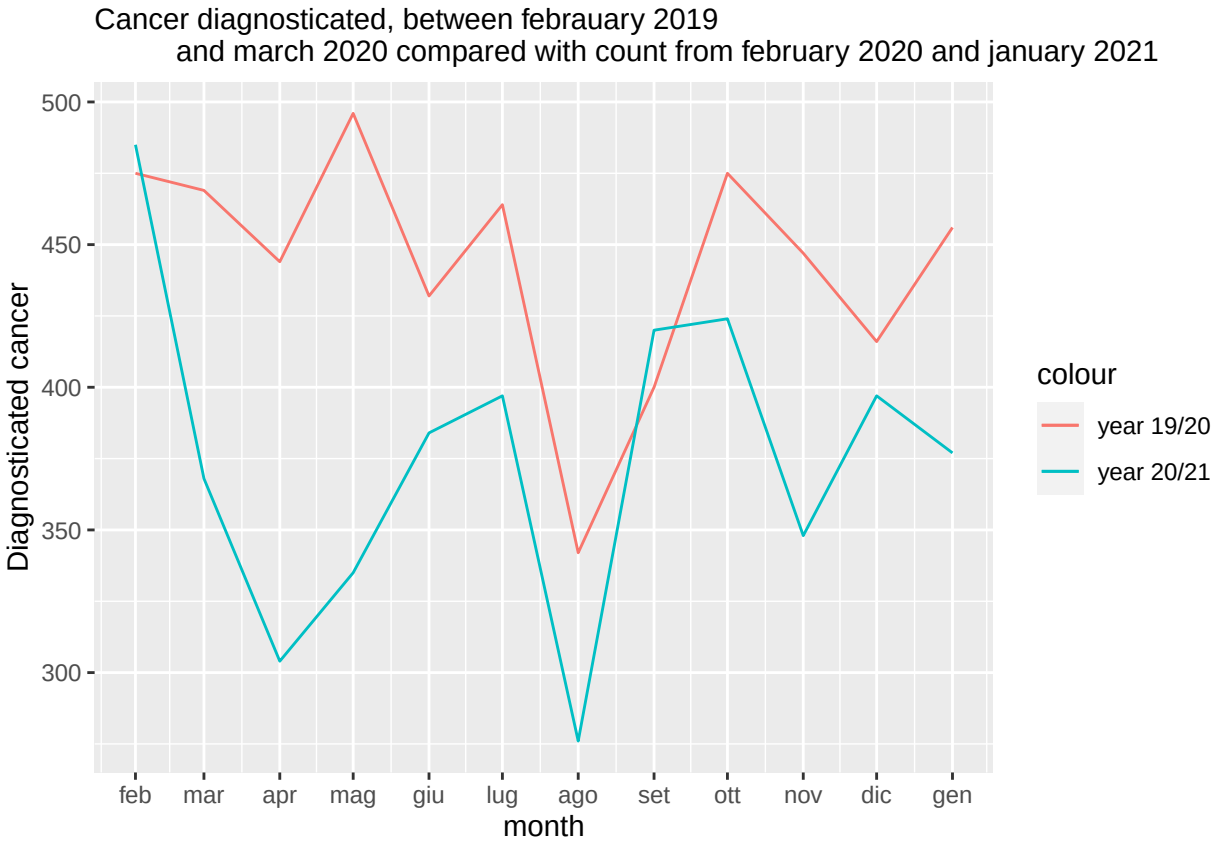
```
#### Vertical Dataframe
df_v = as.data.frame(matrix(data=c(df_h[,1],df_h[,2],rep(0,12),rep(1,12)),
                                  nrow = 24,ncol = 2, byrow =F))
colnames(df_v) =c("count","covid")
df_v$covid = as.factor(df_v$covid)
```

from horizontal to vertical monthly dataframe

Plot of comparasion of the two years of observations

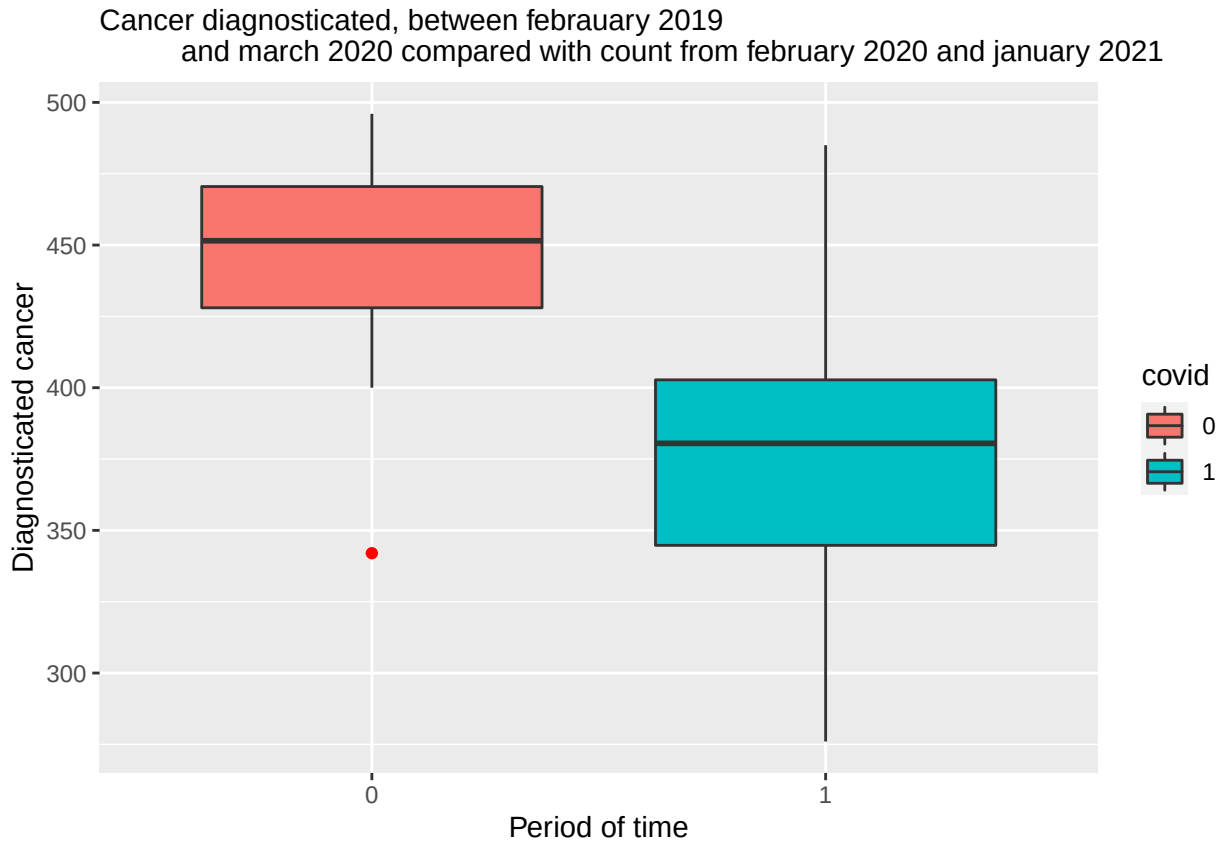
```
time_mesi=seq(as.Date("2019-2-1"), as.Date("2020-1-1"), by = "months")
ggplot(df_h, aes(time_mesi)) +
```

```
  ggtitle("Cancer diagnosticated, between febrauary 2019
          and march 2020 compared with count from february 2020 and january 2021") +
  geom_line(aes(y = before, colour = "year 19/20")) +
  geom_line(aes(y = after, colour = "year 20/21")) +
  xlab("month") +
  ylab("Diagnosticated cancer") +
  scale_x_date(breaks = '1 month',
               date_labels = "%b") +
  theme(plot.title = element_text(size = 11))
```



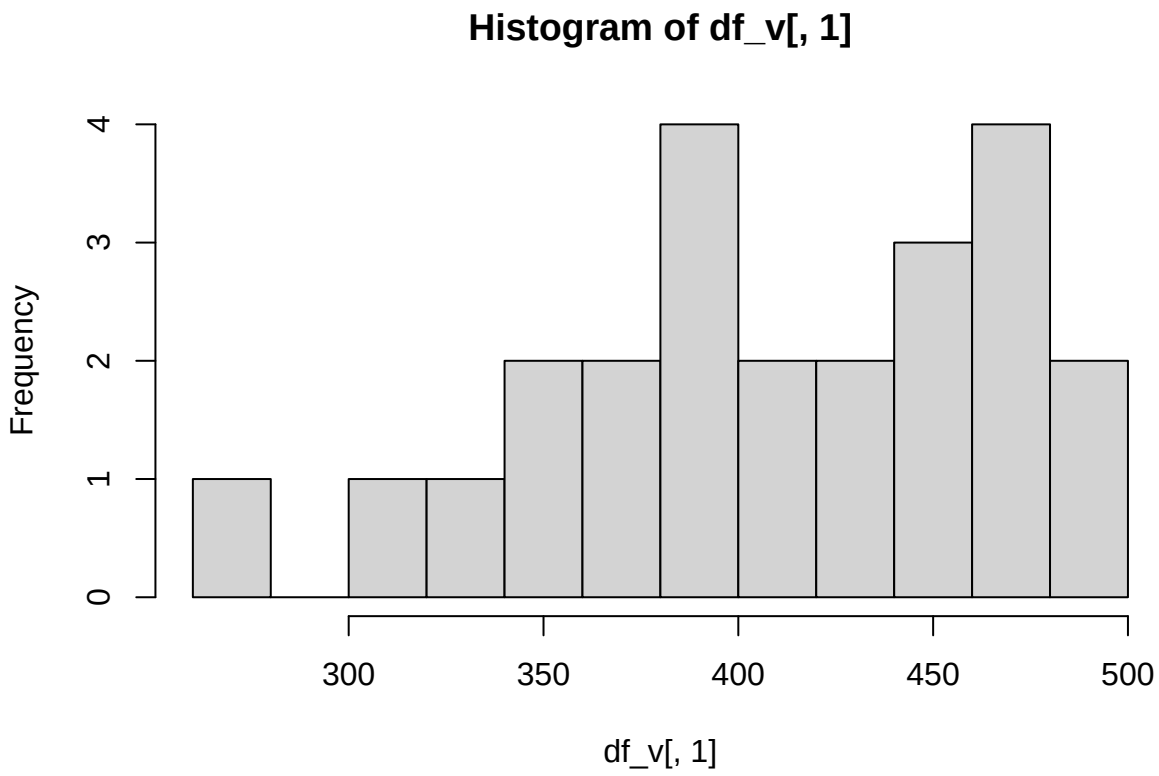
Boxplot of the two periods

```
ggplot(df_v, aes(x=covid,y=count,fill=covid)) + # boxplot dei 12 mesi pre covid - ROSSO
         geom_boxplot(outlier.colour = "red",) +
         ggtitle("Cancer diagnosticated, between febrauary 2019
          and march 2020 compared with count from february 2020 and january 2021") +
         xlab("Period of time") +
         ylab("Diagnosticated cancer") +
         theme(plot.title = element_text(size = 11))
```

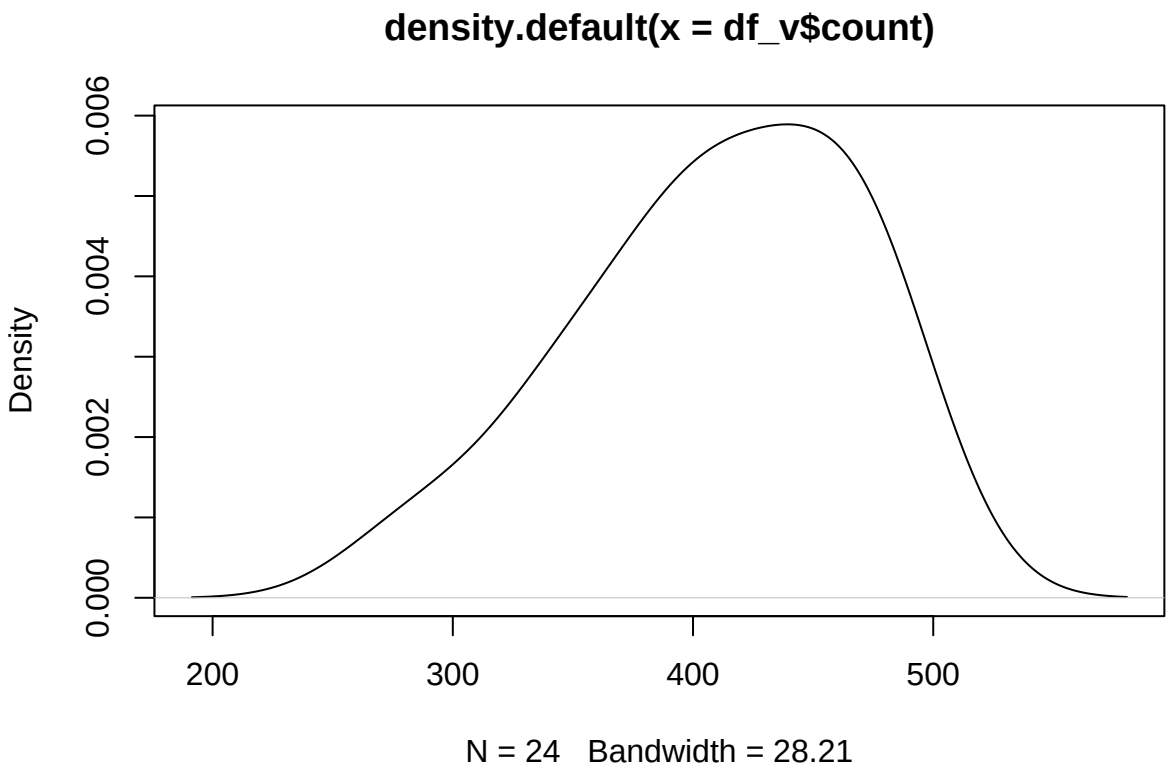


Data distribution analysis

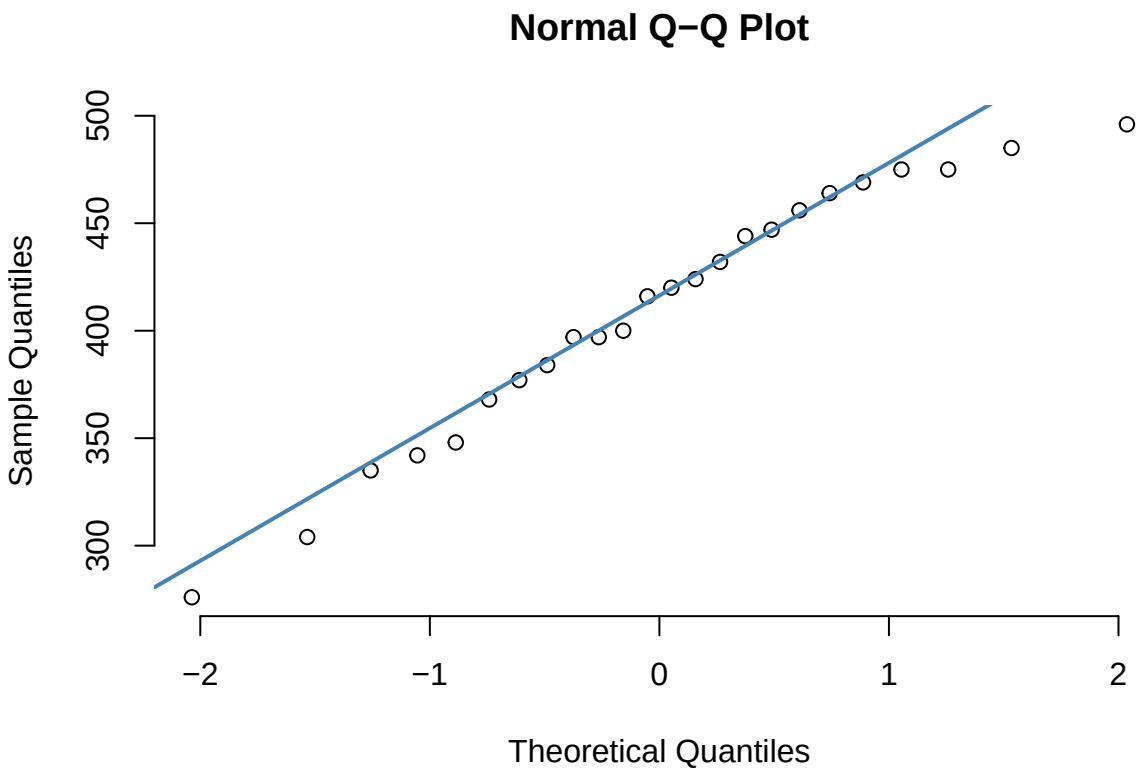
```
hist(df_v[,1],breaks = 8)
```




```
# Kernel Density Plot
d <- density(df_v$count) # returns the density data
plot(d) # plots the results
```




```
qqnorm(df_v$count, pch = 1, frame = FALSE)
qqline(df_v$count, col = "steelblue", lwd = 2)
```




```
# qqplot(df_1920v$conteggio)
shapiro.test(df_v$count)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_v$count
## W = 0.95976, p-value = 0.4335
```

```
shapiro.test(df_h$before)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_h$before
## W = 0.90154, p-value = 0.1661
```

```
shapiro.test(df_h$after)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_h$after
## W = 0.98335, p-value = 0.9937
```

```
# Test for equal variance in both the sample

bartlett.test(df_v$count~df_v$covid)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  df_v$count by df_v$covid
## Bartlett's K-squared = 0.93756, df = 1, p-value = 0.3329
```

There is empirical evidence about violation of normality, the nature of the data and the dimension of the sample suggest the try a distribution for positive counting data like Poisson or Negative Binomial.

In our case Poisson can't be used because the variance is much bigger compared to the mean of the distribution. (violation of one of the assumption for the Poisson distribution)

Checking mean and variance difference

```
mean(df_h$before)
```

```
## [1] 443
```

```
mean(df_h$after)
```

```
## [1] 376.25
```

```
var(df_h$before)
```

```
## [1] 1732.727
```

```
var(df_h$after)
```

```
## [1] 3161.841
```

Dummy variables

Dummy variables for model the strong seasonality of august

```
# dummy
dm=make.dummy(length(df_v[,1]),start=2, freq=12)
# dm[1:6,]
# give names to the dummies
nomi=c("dm4","dm5","dm6","dm7","dm8","dm9","dm10","dm11","dm12","dm1","dm2","dm3")
dimnames(dm)=list(NULL,nomi)
colnames(dm)<-nomi
```

Test for difference between periods

Linear model with normal assumption, t-test and Wilconox-Mann test

```
# 12 month
# Linear model
summary(lm(df_v$count~df_v$covid+dm[,8]))
```

```
##
## Call:
## lm(formula = df_v$count ~ df_v$covid + dm[, 8])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81.40 -18.09   0.00  18.35  99.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   452.15      11.48  39.397  < 2e-16 ***
```

```
## df_v$covid1   -66.75      15.87  -4.205 0.000398 ***
## dm[, 8]      -109.77      28.72  -3.823 0.000992 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.88 on 21 degrees of freedom
## Multiple R-squared:  0.606,  Adjusted R-squared:  0.5684
## F-statistic: 16.15 on 2 and 21 DF,  p-value: 5.663e-05
```

```
# T test:
t.test(df_v$count~df_v$covid, var.equal =T)
```

```
##
##  Two Sample t-test
##
## data:  df_v$count by df_v$covid
## t = 3.3051, df = 22, p-value = 0.003223
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   24.86594 108.63406
## sample estimates:
## mean in group 0 mean in group 1
##          443.00          376.25
```

```
# Wilcoxon-Mann-Whitney test:
wilcox.test(df_v$count~df_v$covid)
```

```
## Warning in wilcox.test.default(x = c(475, 469, 444, 496, 432, 464, 342, : cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df_v$count by df_v$covid
## W = 121, p-value = 0.005089
## alternative hypothesis: true location shift is not equal to 0
```

Test with negative binomial regression model for overdispersed counting data

```
# 12 month
bn_covid = glm.nb(df_v$count~df_v$covid+dm[,8])
summary(bn_covid)
```

```
##
## Call:
## glm.nb(formula = df_v$count ~ df_v$covid + dm[, 8], init.theta = 173.8636152,
##     link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.44331  -0.48980   0.02804   0.42578   2.67190
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.11611    0.02639 231.716  < 2e-16 ***
## df_v$covid1 -0.16404    0.03701  -4.432 9.35e-06 ***
## dm[, 8]     -0.30542    0.06981  -4.375 1.21e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(173.8636) family taken to be 1)
##
##     Null deviance: 62.314  on 23  degrees of freedom
## Residual deviance: 24.176  on 21  degrees of freedom
## AIC: 249.35
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  173.9
##           Std. Err.:  72.2
##
##  2 x log-likelihood:  -241.352
```

```
# percentage difference after covid can be calculated with 1-exp(-0.18359)
1-exp(-0.18359)
```

```
## [1] 0.167723
```

Check for correct assumption of Negative binomial model instead of Poisson model.

```
poi_covid <- glm(df_v$count ~ df_v$covid, family = "poisson")
pchisq(2 * (logLik(bn_covid) - logLik(poi_covid)), df = 1, lower.tail = FALSE)
```

```
## 'log Lik.' 3.066798e-20 (df=4)
```

confirm the correct assumption of negative binomial instead of Poisson.

Both models confirm the mean difference between the two sample. Same conclusion.

Intepretation of the model parameter

```
(1-exp(-0.16404))*100
```

```
## [1] 15.12919
```

15.129 is the percentage difference between the two periods.