פרויקט מדעי נתונים: קורונה בישראל

אוראל חגג, נויה שוקר, עומרי וקנין, מני בוזגלו, דור חטואל

2021 1 7

נגיף הקורונה החל בסין והגיע לישראל בסוף פברואר 2020 וממשיך להתפשט במדינה עד היום.

הנגיף נגע בכל קבוצות האוכלוסייה בישראל גברים ונשים, מגזרים שונים, איזורים גיאורפיים וכן כל קבוצות הגילים.

חולי הקורונה נמצאים בשלושה דרגות קושי קל, בינוני וקשה. בנוסף קיימים מצבים קריטים יותר, מונשמים ומאושפזים בעקבות המחלה.

במחקר אנו נתייחס בין היתר למצבים השונים של החולים ולקשר בניהם.

נתונים מעניינים אותם אנו שומעים יום יום ונרצה לחקור הם מבחינת הדמוגרפיה של החולים, ההשפעה הגבוהה על אוכלוסיית המבוגרים, גברים, אזורים בארץ ובינהם גם מגזרים שונים.

מקורות המחקר

את נתוני המחקר אספנו ממאגר המידע הממשלתי:

- https://data.gov.il/dataset/covid-19/resource/e4bf0ab8-ec88-4f9b-8669-f2cc78273edd

בסיס הנתונים מכיל פרטים אודות כמות חולים, גיל ממוצע, ואחוז נשים נדבקות בכל אחת מדרגות הקושי וכן במצבים הקריטים של המחלה עבור כל יום מחודש מרץ ועד דצמבר.

- https://github.com/idandrd/israel-covid19-data/blob/master/CityData.csv

בסיס הנתונים מכיל פרטים אודות ערים שונות בישראל, גודל אוכלוסייה וכמות נדבקים בה עבור חודשים אפריל ומאי.

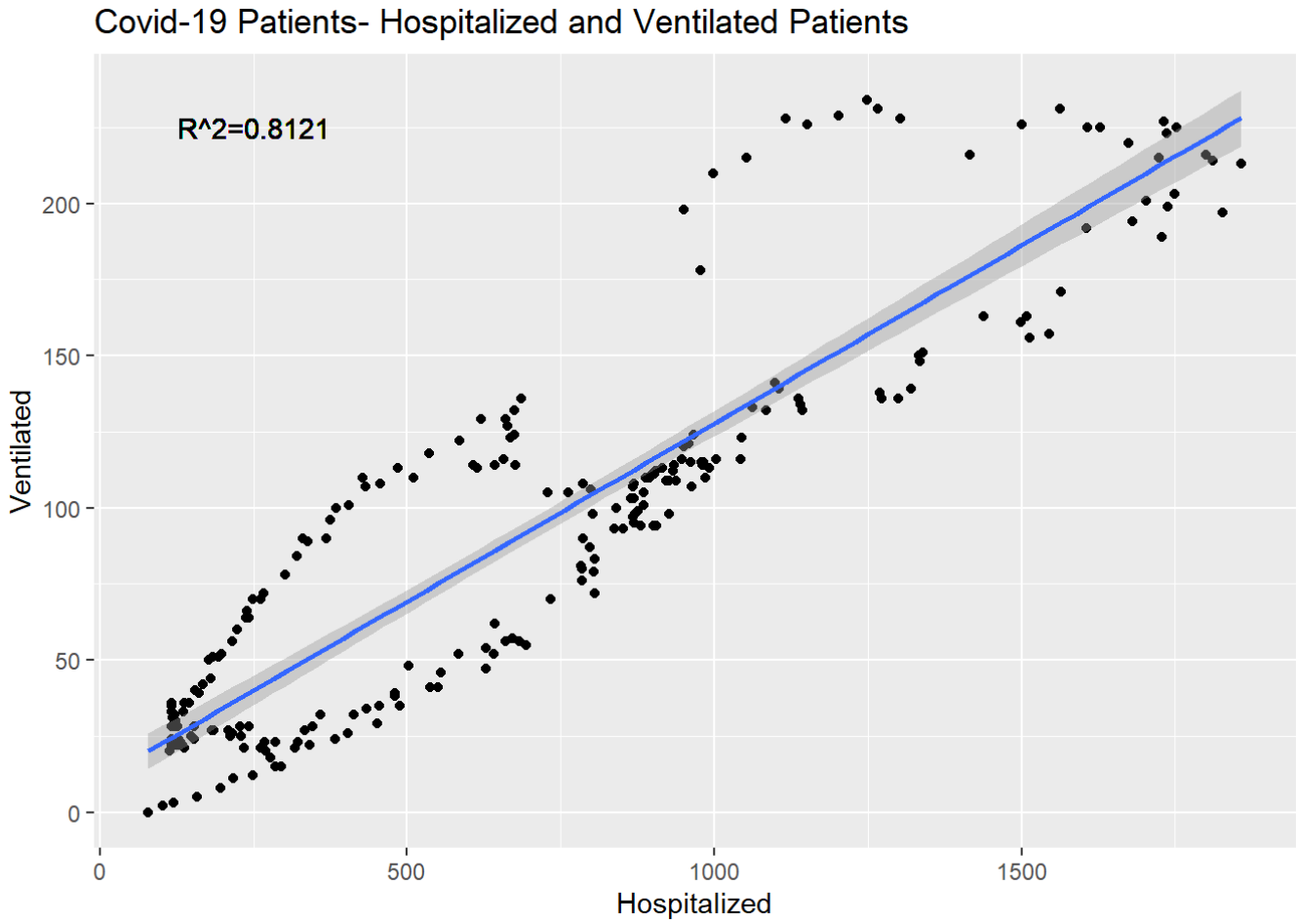
*על בסיסי הנתונים נערוך מניפולציות בכל שאלת מחקר ע"י שליפת עמודות רלוונטיות והוספת עמודות חישוב של הנתונים.

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(tidyverse)
library(ggalt)
library(reshape2)
library(readxl)
```

```
library(car)
library(agricolae)
```

שאלת מחקר: האם קיים קשר לינארי בין כמות חולים מאושפזים לכמות המונשמים?

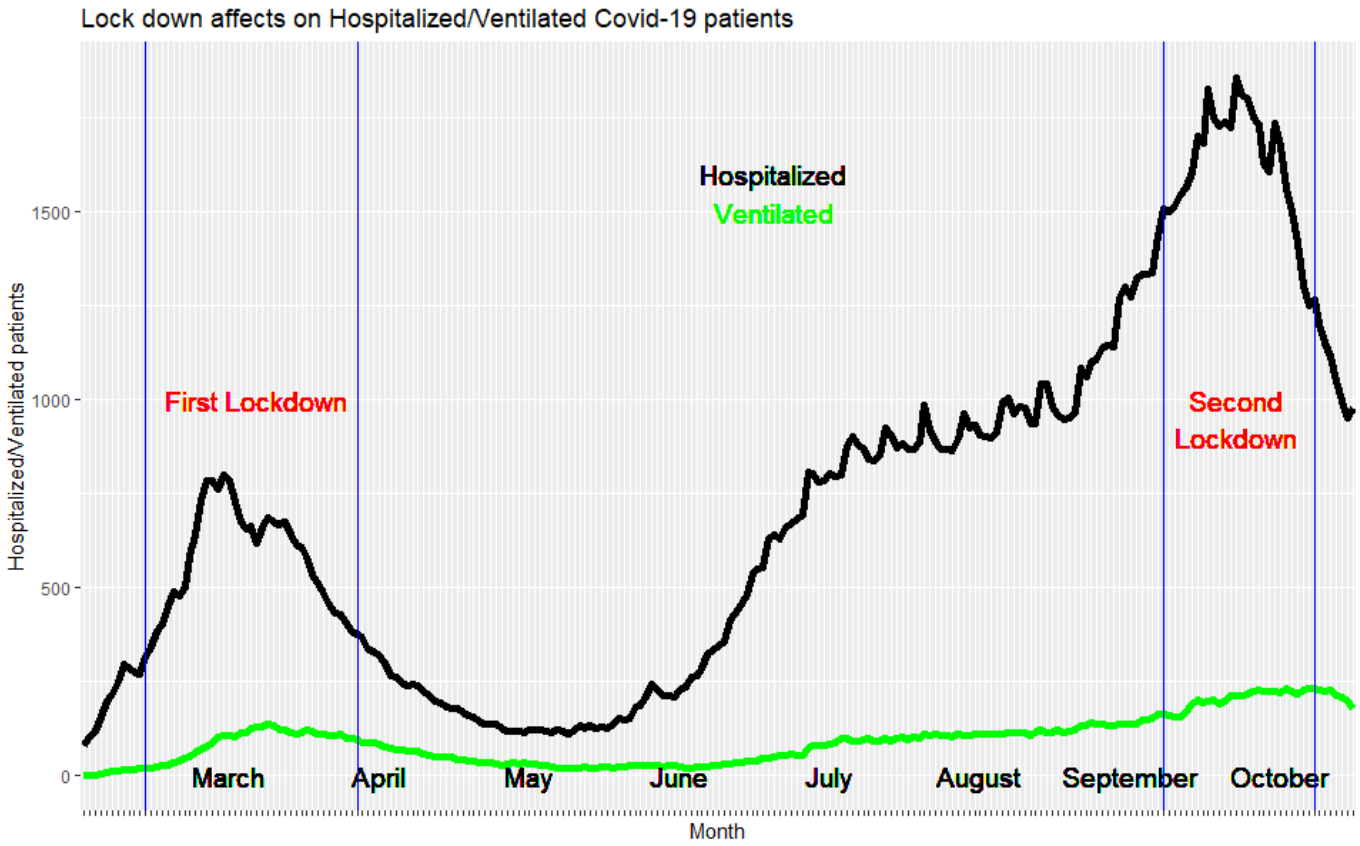
```
q1_data <- read.csv("question1.csv", header = TRUE)
colnames(q1_data) <- c("date", "hospitalized", "ventilated")

lm1<-lm(formula = q1_data$ventilated~q1_data$hospitalized)

g1<-ggplot(q1_data, aes(x=hospitalized, y=ventilated))+
  geom_point()+
  geom_smooth(method="lm")+labs(title = 'Covid-19 Patients- Hospitalized and Ventilated Patients',x='Hospitalize
d',y='Ventilated')+
  geom_text(x=250,y=225,label='R^2=0.8121')
g1
```



בתרשים הבא בחרנו להציג את מספר המונשמים כתלות במספר המאושפזים השתמשנו בתרשים נקודות ובקו מגמה . ציר ה-X מייצג את מספר המאושפזים וציר ה Y מייצג את מספר המונשמים.

```
Res<-residuals(lm1)

months1<-c('March','April','May','June','July','August','September','October')

y<-0
```

```
g2<-ggplot(data = q1_data,aes(x=date))+geom_line(data=q1_data,aes(y=hospitalized),group=1,size=2)+
  geom_line(data=q1_data,aes(y=ventilated),group=1,color='Green',size=2)+
  scale_x_discrete(labels = NULL)+geom_text(x=27,y=y,label='March',size=5)+geom_text(x=54,y=y,label='April',size=
5)+
  geom_text(x=81,y=y,label='May',size=5)+geom_text(x=108,y=y,label='June',size=5)+geom_text(x=135,y=y,label='Jul
y',size=5)+
  geom_text(x=162,y=y,label='August',size=5)+geom_text(x=189,y=y,label='September',size=5)+geom_text(x=216,y=y,la
bel='October',size=5)+
  geom_text(x=32,y=1000,label='First Lockdown',size=5,color='Red')+geom_text(x=208,y=1000,label='Second',size=5,c
olor='Red')+
  geom_text(x=208,y=900,label='Lockdown',size=5,color='Red')+ geom_text(x=125,y=1500,label='Ventilated',size=5,co
lor='Green')+
  geom_text(x=125,y=1600,label='Hospitalized',size=5,color='Black')+
  geom_vline(xintercept = 12,color='Blue')+geom_vline(xintercept = 50,color='Blue')+geom_vline(xintercept = 222,c
olor='Blue')+
  geom_vline(xintercept = 195,color='Blue')+
  labs(title = 'Lock down affects on Hospitalized/Ventilated Covid-19 patients',x='Month',y='Hospitalized/Ventila
ted patients')
```



בעזרת הגרפים ניסינו להמחיש את הקשר הלינארי בין המשתנים, ניתן לראות כי כאשר יש מגמות ירידה/עליה במתשנה המאושפזים הן חלות גם על משתנה המונשמים.

ניתוח סטטיסטי

נבחר במבחן מקדם המתאם של פירסון מכיוון שזה המדד הנפוץ ביותר לבדיקת קורלאציה בין שני משתנים מספריים, המדד ישקף את מידת הקשר בין המשתנים בסקלה שבין 1- ל-1, כאשר 0 מציין חוסר קשר, 1 קשר חיובי מושלם ו- 1- קשר שלילי מושלם.

מקור: https://www.tau.ac.il/~ricardo/mekuvan/lessons/b7/b7.2.htm

Pearson test:

הגדרת משתנים

$Y_i=\beta_0+\beta_1*X_{1i}+\varepsilon_i$

Y_i-המשתנה המוסבר, מספר המונשמים

x_1-המשתנה המסביר,מספר המאושפזים

השערות:

$H_0:\beta_1=0$

H_1:else

```
cor.test(q1_data$hospitalized, q1_data$ventilated, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  q1_data$hospitalized and q1_data$ventilated
## t = 31.32, df = 227, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8735675 0.9229670
## sample estimates:
##       cor
## 0.9011528
```

מהפלט התקבל P-value קטן מ0.05 ולכן, נדחה את H0 ברמת מובהקות 5% ונסיק כי קיים קשר לינארי בין המשתנים.

המודל הלינארי-הסופי

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	20.165970	3.365174	5.993	6.34e-09	***
q1_data$hospitalized	0.110919	0.004105	27.020	< 2e-16	***

ניתן לראות כי ערך ה P-Value של המשתנה המסביר מובהק. ולכן, זהו המודל הלינארי הסופי:

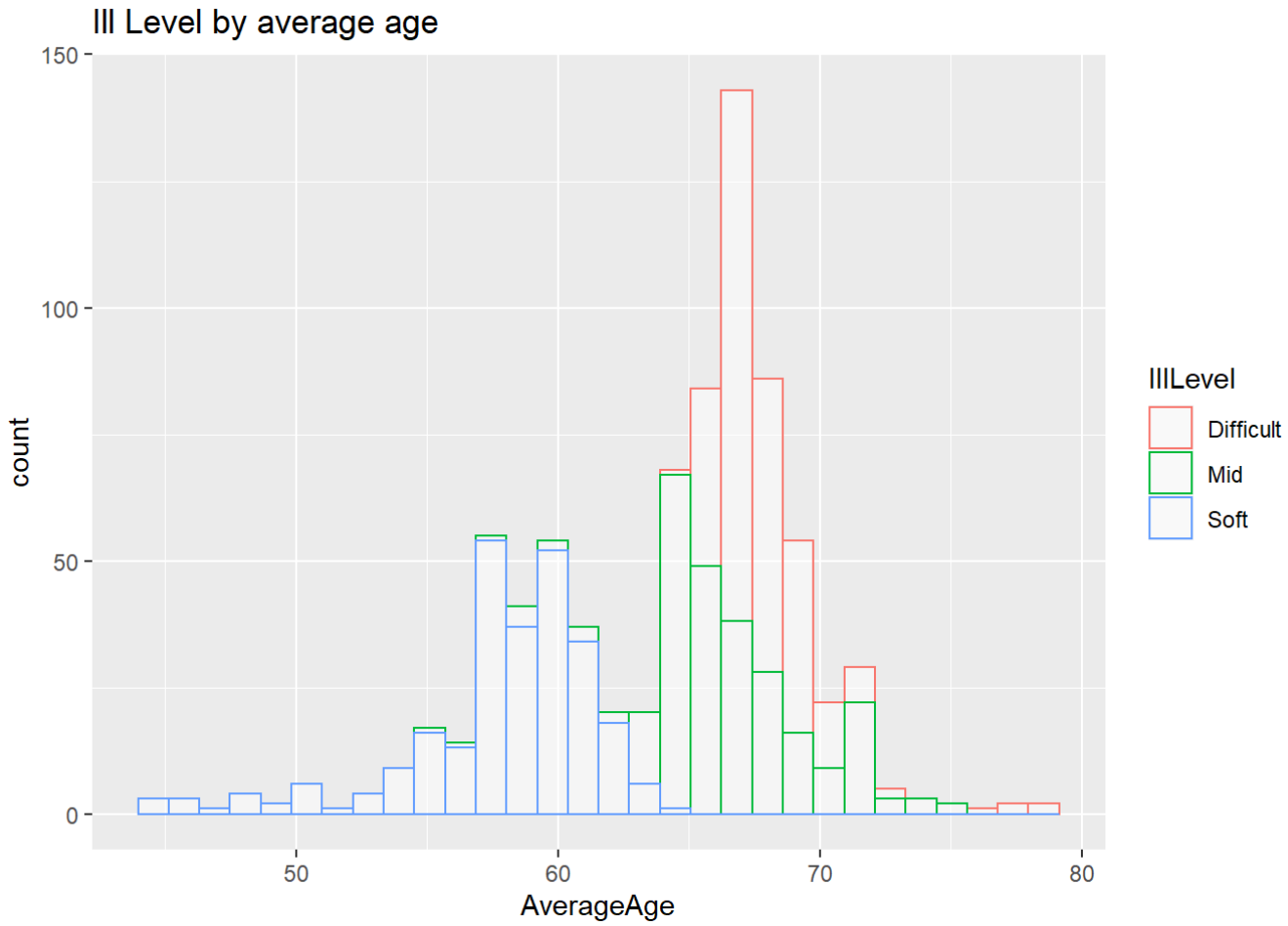
$Y_i=0.1109*X_{1i}+20.1659+\varepsilon_i$

```
summary(Res)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -37.235 -21.100  -6.153   0.000  11.667  86.670
```

שאלת מחקר: האם קיים קשר בין ממוצע גיל החולים לדרגת קושי תחלואתם (קל, בינוני, קשה)?

```
histdata<-read.csv('Datahist.csv',header = T)
colnames(histdata)<-c("Date","IllLevel", "AverageAge")
hist1<-ggplot(histdata,aes(x=AverageAge,color=IllLevel))+geom_histogram(fill="white",alpha=0.5)+
  labs(title='Ill Level by average age')+theme_gray()
hist1
```



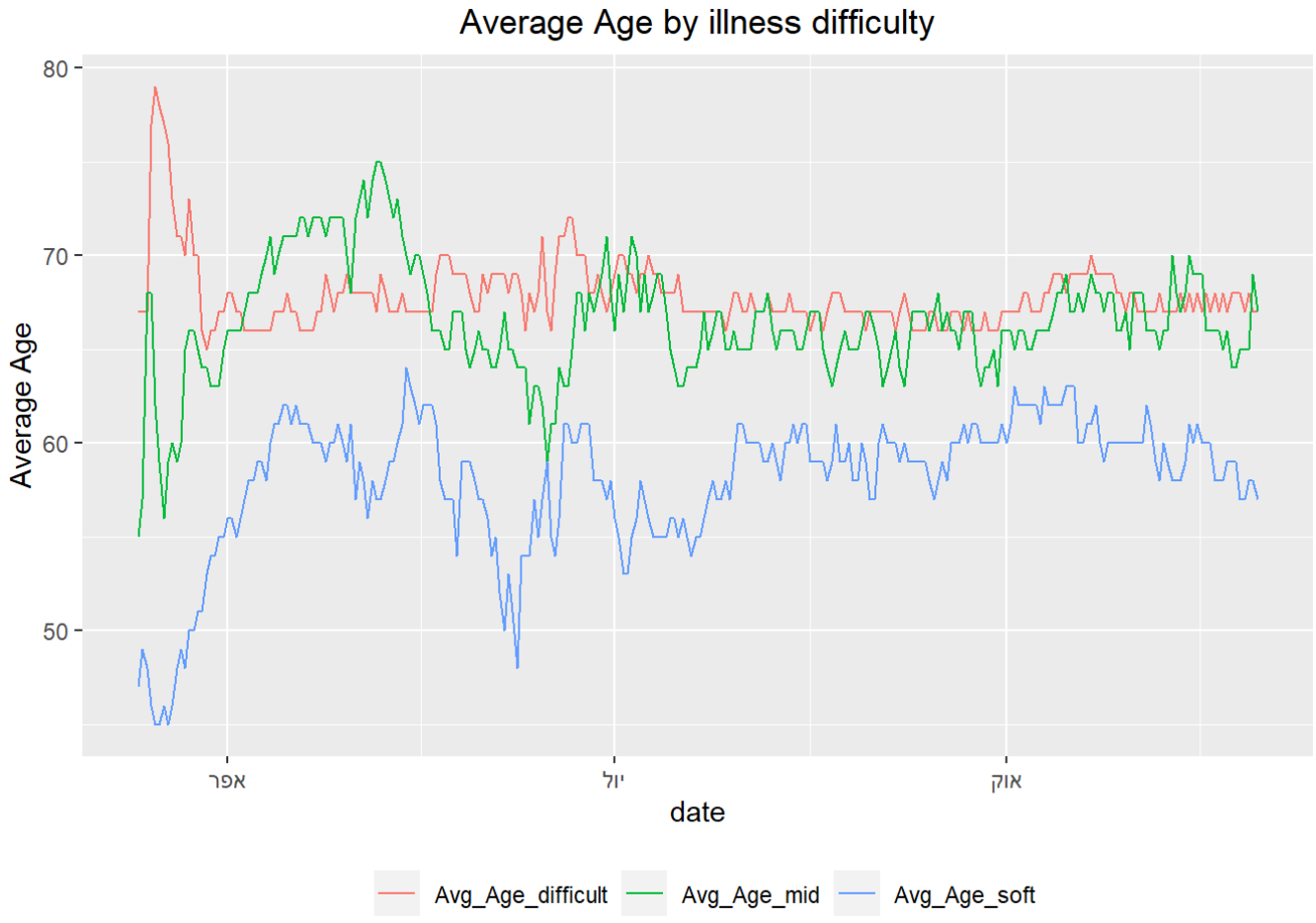
ההיסטוגרמה הבאה מציגה את התפלגות גיל החולים כתלות בדרגת התחלואה, בצבע כתום מוצגים חולים בדרגת תחלואה קשה, בירוק חולים בדרגת תחלואה בינונית ובכחול חולים בדרגת תחלואה קלה.

```
q2_data <- read_excel("Q2B.xlsx", sheet=2)
```

```
## New names:
## * 1...תאריך <- תאריך
```

```
## * 5...תאריך <- תאריך
```

```
colnames(q2_data) <- c("date", "Avg_Age_difficult", "Avg_Age_mid", "Avg_Age_soft", "date1", "difficult_ill_amoun
t", "mid_ill_amount", "soft_ill_amount")

relevant_data <- q2_data[,c(1:4)]

#transform the data with like pivot function
a <- melt(relevant_data, id=c("date"), variable.name = "ill_Group")

graph2 <- ggplot(a, aes(x=date, y=value, color=ill_Group))+geom_line()+
  labs(title = "Average Age by illness difficulty", y = "Average Age") +
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_blank(), legend.position = "bottom")
graph2
```



התרשים הבא מציג את גיל החולים הממוצע בכל דרגת תחלואה בחודשים מרץ עד דצמבר.

ניתוח סטטיסטי

על מנת שנוכל לענות על שאלת המחקר "האם קיים קשר בין ממוצע גיל החולים לדרגת קושי תחלואת (קל,בינוני,קשה)". יש צורך בבדיקה האם הקבוצות (דרגות תחלואה) ההומוגניות שונות, לכן המבחן הסטטיסטי המתאים לנתוני השאלה הוא מבחן Duncan, ובנוסף בעזרתו נוכל לקטלג את הקבוצות לפי תוחלות.

מקור: https://www.statisticshowto.com/duncans-multiple-range-test/

:Duncan test

הגדרת משתנים

μ_1 -מייצג את תוחלת גיל החולים בדרגת תחלואה קשה

μ_2 -מייצג את תוחלת גיל החולים בדרגת תחלואה בנונית

μ_3 -מייצג את תוחלת גיל החולים בדרגת תחלואה קלה

השערות

H_0: $\mu_1=\mu_2=\mu_3$

H_1: else

```
duncan1<-duncan.test(histdata$AverageAge,histdata$IllLevel,789,8.7)
duncan1
```

```
## $statistics
##   MSerror  Df     Mean       CV
##       8.7 789 64.13763 4.598824
##
## $parameters
##     test            name.t ntr alpha
##   Duncan histdata$IllLevel   3  0.05
##
## $duncan
##      Table CriticalRange
## 2 2.776066     0.5039500
## 3 2.922726     0.5305738
##
## $means
##           histdata$AverageAge      std   r Min Max Q25 Q50 Q75
## Difficult            67.87121 1.863825 264  65  79  67  67  68
## Mid                  66.53030 3.071863 264  55  75  65  66  68
## Soft                 58.01136 3.640168 264  45  64  57  59  60
##
```

```
## $comparison
## NULL
##
## $groups
##           histdata$AverageAge groups
## Difficult            67.87121      a
## Mid                  66.53030      b
## Soft                 58.01136      c
##
## attr(,"class")
## [1] "group"
```

נדחה את את H0 ברמת מובהקות 5% כלומר הקבוצות אינן הומוגניות ומכאן ניתן לראות מהפלט שהקבוצות מתחלקות ל3.

שאלת מחקר: האם גברים מושפעים ממחלת הקורונה יותר מנשים?

על מנת לבדוק את שאלת המחקר ביצענו השוואה בין נשים לגברים ע"י ממוצע אחוז החולים בשלוש דרגות הקושי של מחלת הקורונה, קל, בינוני וקשה.

```
#data upload
corona<-read.csv("corona.csv",header = T,stringsAsFactors = FALSE)


#Retrieve columns and rows relevant to data frame
vars <- c(1,11,15,19)
womanANDman <- corona[ , vars]
colnames(womanANDman) <- c("date","mild_sick_women","medium_sick_women","hard_sick_women")
rows <- c(21,51,82,112,143,174,204,235,265)
womanANDman <-womanANDman[ rows, ]

#Adjustment varies by type to numeric
womanANDman[3]<-as.numeric(womanANDman$medium_sick_women)


#Adding complementary columns for men
womanANDman[["mild_sick_men"]] <- 100- womanANDman$mild_sick_women
```

```
womanANDman[["medium_sick_men"]] <- 100-womanANDman$medium_sick_women
womanANDman[["hard_sick_men"]] <- 100-womanANDman$hard_sick_women



#change wide format to Long for women
w<-select(womanANDman,date,ends_with("women"))
w1<-melt(w,
         id.vars=c("date"),
         measure.vars=c("mild_sick_women","medium_sick_women","hard_sick_women"),
         variable.name="group",
         value.name="p")


#Calculating the average number of sick women at all levels
w2<-group_by(w1,date)
w3 <-summarise(w2,p=mean(p))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

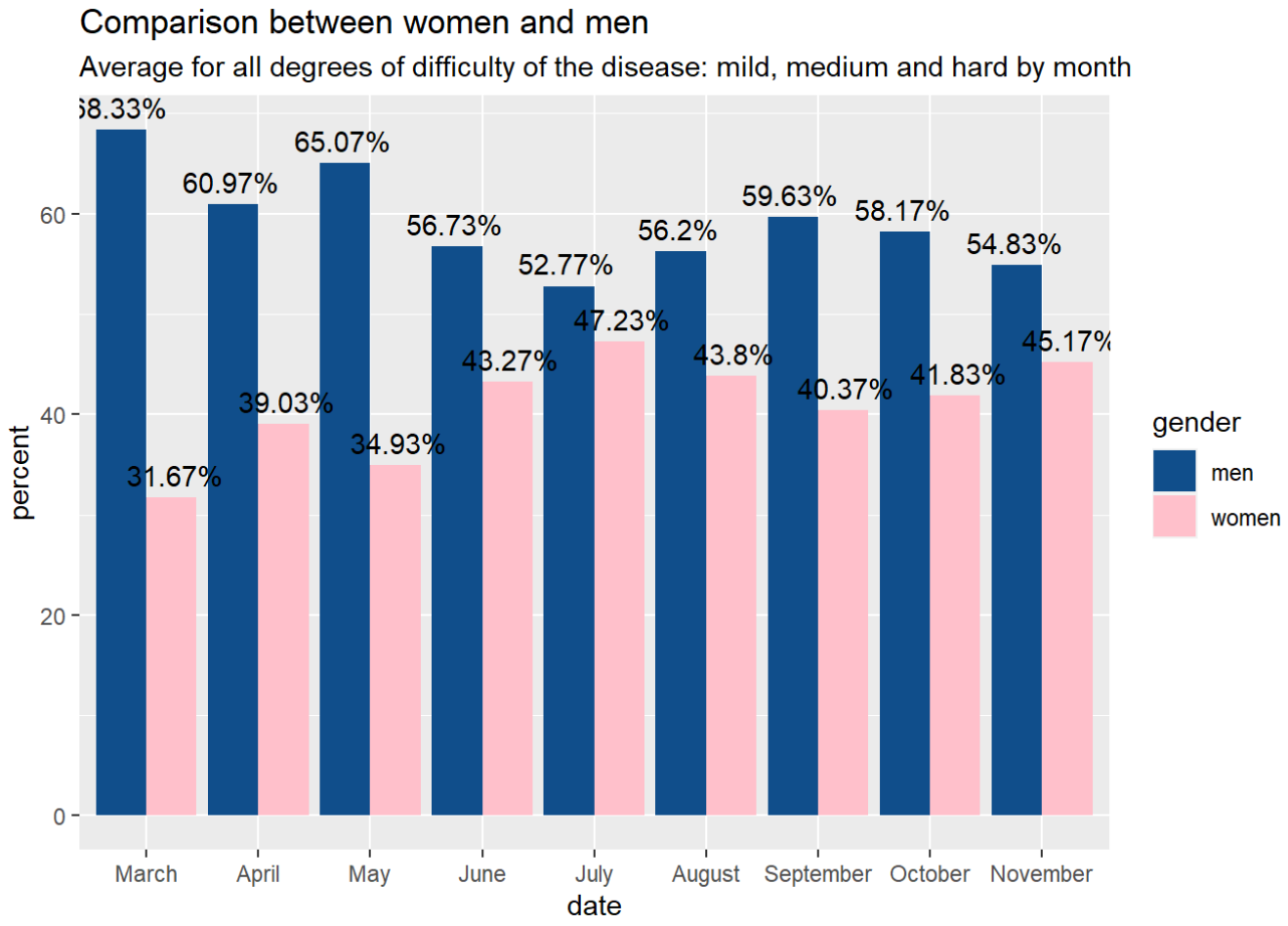
```
#Create a new column
w3$gender <- "women"


#change wide format to Long for men
m<-select(womanANDman,date,ends_with("_men"))
m1<-melt(m,
         id.vars=c("date"),
         measure.vars=c("mild_sick_men","medium_sick_men","hard_sick_men"),
         variable.name="group",
         value.name="p")

#Calculating the average number of sick men at all levels
m2<-group_by(m1,date)
m3 <-summarise(m2,p=mean(p))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
#Create a new column
m3$gender <- "men"


#Vertical linkage of a data frame of women to men
MandW<-rbind(w3,m3)

#Format change, arrangement and setting of factors for the date
MandW$date= as.Date(MandW$date, "%d/%m/%Y")
MandW<-arrange(MandW,date)
MandW$date <- as.factor(MandW$date)
```

גרף העמודות הבא מציג השוואה של ממוצע אחוז החולים ב3 דרגות הקושי בין גברים לנשים בחודשים מרץ ועד דצמבר.

```
#Create a graph of type geom bar
ggplot(data=MandW, aes(x=date, y=p,fill = gender))+
  geom_bar(stat="identity",position='dodge')+
  labs(title="Comparison between women and men",
       subtitle="Average for all degrees of difficulty of the disease: mild, medium and hard by month",
       y="percent")+
  geom_text(aes(label=paste0(round(p,2),"%")),position = position_dodge(width = 1),vjust=-0.55)+
  scale_x_discrete(labels=c("March","April","May","June","July","August","September","October","November","Decemb
er"))+
  scale_fill_manual("gender", values = c("women"="pink","men"="dodgerblue4"))
```



מהגרף ניתן לראות **באופן גורף** כי ממוצע אחוז גברים החולים בכל חודש ממרץ ועד דצמבר גבוהה יותר מממוצע אחוז הנשים.

מכאן ניתן לשער כי קיים הבדל בהשפעת הקורונה על המגדרים השונים.

על מנת לבסס את ההשערה נבצע ניתוח בגרף נוסף ובו נתייחס לכמויות נשים וגברים בשני המצבים הקריטים של המחלה, מונשמים ומאושפזים.

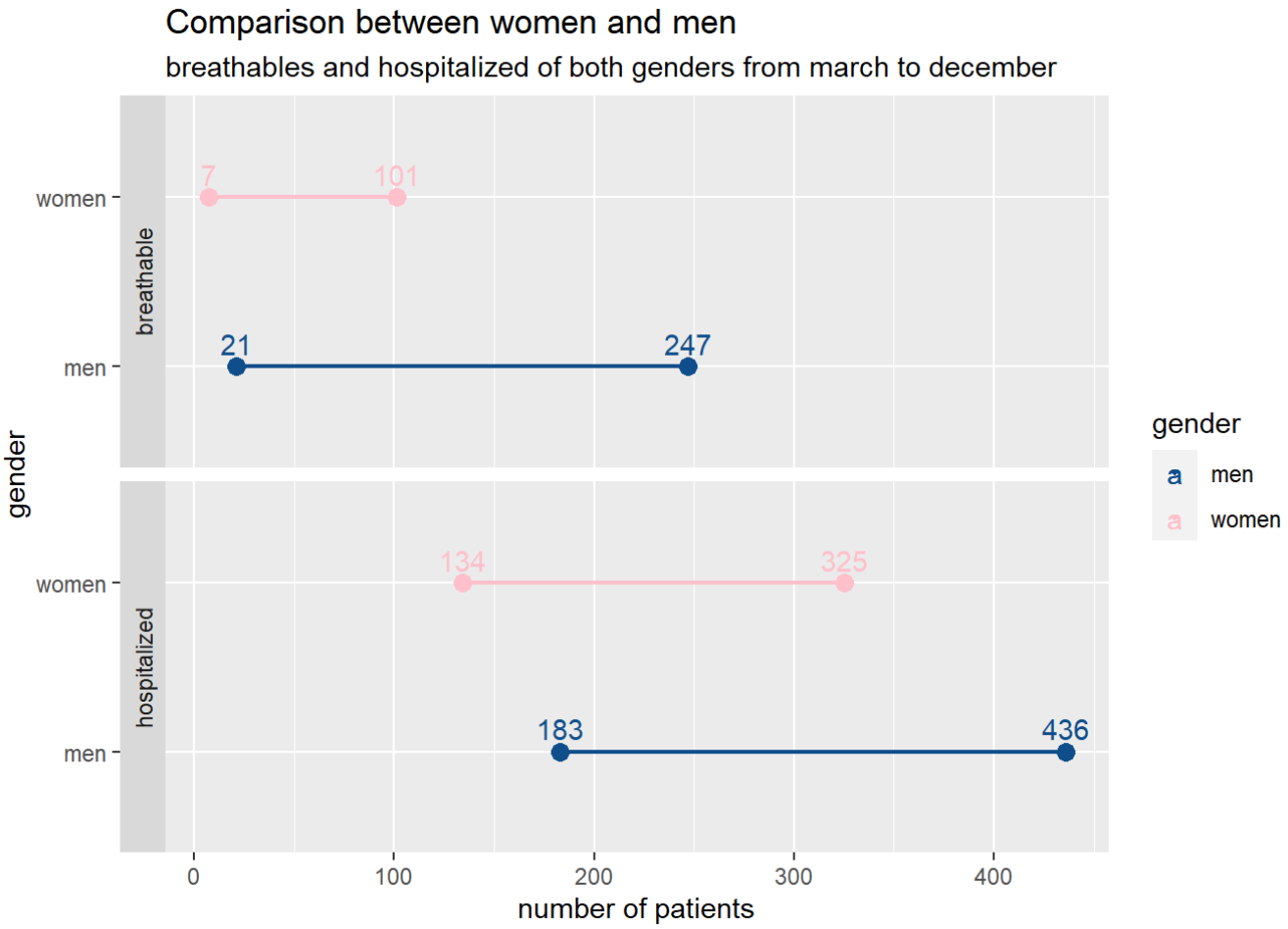
```
#Retrieve columns and rows relevant to data frame
vars <- c(1:3,6,7)
rows <- c(12,282)
```

```
hospANDbreat<- corona[ rows, vars]
colnames(hospANDbreat) <- c("date","hospitalized","hospitalized_women","breathable","breathable_women")


#Adjustment varies by type to numeric
hospANDbreat[4]<-as.numeric(hospANDbreat$breathable)

#Convert percentages to quantities and round whole nearest
for(i in 2:4){
  hospANDbreat[i+1] <- round((hospANDbreat[i]*(hospANDbreat[i+1]/100)))
}

#Adding complementary columns for men
hospANDbreat[["hospitalized_men"]] <-hospANDbreat$hospitalized - hospANDbreat$hospitalized_women
hospANDbreat[["breathable_men"]] <- hospANDbreat$breathable- hospANDbreat$breathable_women

#Change values to column Date
hospANDbreat$date<-c("start","end")

#change wide format to Long
hospANDbreat1<-melt(hospANDbreat,
        id.vars="date",
        measure.vars=c("breathable_women","breathable_men","hospitalized_women","hospitalized_men"),
        variable.name="category",
        value.name="amount")

#Add gender and category columns for each type
m<-filter(hospANDbreat1,category=="breathable_men")
m$gender<-"men"
m$category<-"breathable"
m1<-filter(hospANDbreat1,category=="hospitalized_men")
m1$gender<-"men"
m1$category<-"hospitalized"
w1<-filter(hospANDbreat1,category=="breathable_women")
w1$gender<-"women"
w1$category<-"breathable"
w2<-filter(hospANDbreat1,category=="hospitalized_women")
```

```
w2$gender<-"women"
w2$category<-"hospitalized"

#Vertical linkage of a data frame of all type
hosp_breat<-rbind(m,m1,w1,w2)

#change Long format to wide
hosp_breat1<-dcast(hosp_breat,category+gender~
                date,value.var = "amount")
```

הגרף הבא מציג כמות מונשמים ומאושפזים מתאריך התחלה במרץ לתאריך סיום בדצמבר ובכך מתאר את ההשוואה בין גברים לנשים בשני המצבים.

```
#Create a graph of type geom dumbbell
ggplot(hosp_breat1, aes(x = start, y = gender, xend = end, color = gender))+
  geom_dumbbell(size=0.75, size_x = 3, size_xend = 3)+
  facet_grid(category~., switch = "y")+
  scale_color_manual(values=c("dodgerblue4","pink"))+
 labs(title="Comparison between women and men",
     subtitle="breathables and hospitalized of both genders from march to december",
     x="number of patients")+
geom_text(aes(x = start,label=start),position = position_dodge(width = 1),vjust=-0.55)+
  geom_text(aes(x = end,label=end),position = position_dodge(width = 1),vjust=-0.55)
```



תוצאת הגרף מחזקת את ההשערה , ניתן לראות שהתקבלה כמות גברים **גבוהה** מנשים בשני הקריטריונים הן מבחינת מאושפזים והן מבחינת מונשמים.

משני הגרפים ניתן לשער שקיים הבדל בין המגדרים השונים.

ניתוח סטטיסטי

מבחן ווילקוקסון משמש כבדיקה האם אוכלוסיית x גדולה סטוכסטית מאוכלוסיית y , כלומר רוב הנבדקים באוכלוסיית x הם בעלי ערכים גבוהים יותר מאשר רוב הנבדקים באוכלוסיית y. במחקר שלנו נרצה לבצע השוואה בין שני אוכלוסיות בלתי תלויות, נשים וגברים ולכן בחרנו במבחן ווילקוקסון למדגמים בלתי תלויים כמבחן המתאים ביותר. במבחן נבדוק האם השפעת הקורונה על גברים גבוהה מהשפעתה על נשים.

מקור: https://statsandr.com/blog/wilcoxon-test-in-r-how-to-compare-2-groups-under-the-non-normality-assumption/

Wilcox test:

השערות:

H_0: $\theta_X=\theta_Y$

H_1: $\theta_X>\theta_Y$

```
#Retrieve relevant columns and rows
vars <- c(1:3,6,7,10,11,14,15,18,19)
rows<-c(12:282)
womanANDman<- corona[rows, vars]
colnames(womanANDman) <- c("date","hospitalized","hospitalized_women","breathable","breathable_women","mild","mil
d_sick_women","medium","medium_sick_women","hard","hard_sick_women")


#Adjustment varies by type to numeric
womanANDman[4]<-as.numeric(womanANDman$breathable)
womanANDman[8]<-as.numeric(womanANDman$medium)
womanANDman[9]<-as.numeric(womanANDman$medium_sick_women)
womanANDman[10]<-as.numeric(womanANDman$hard)

# calculates the amount of women in each rank
for(i in 2:10){
  womanANDman[i+1] <- round((womanANDman[i]*(womanANDman[i+1]/100)))
}

#Adding complementary columns for men
womanANDman[["hospitalized_men"]] <- womanANDman$hospitalized- womanANDman$hospitalized_women
womanANDman[["breathable_men"]] <- womanANDman$breathable- womanANDman$breathable_women
womanANDman[["mild_sick_men"]] <- womanANDman$mild- womanANDman$mild_sick_women
womanANDman[["medium_sick_men"]] <- womanANDman$medium-womanANDman$medium_sick_women
womanANDman[["hard_sick_men"]] <- womanANDman$hard-womanANDman$hard_sick_women

#Add columns that summarize the amount of infection from each gender
```

```r
allsicks<-mutate(womanANDman,
       woman=(hospitalized_women+breathable_women+mild_sick_women+medium_sick_women+hard_sick_women),
       man=(hospitalized_men+breathable_men+mild_sick_men+medium_sick_men+hard_sick_men))

#Change format from long to wide for woman
womanSicks<-melt(allsicks,
         id.vars="date",
         measure.vars="woman",
         variable.name="gender",
         value.name="sicks")

#Change format from long to wide for man
manSicks<-melt(allsicks,
                 id.vars="date",
                 measure.vars="man",
                 variable.name="gender",
                 value.name="sicks")

#Vertical linkage of a data frame of all type
SicksWM<-rbind(womanSicks,manSicks)

#wilcox test
wilcox.test(sicks~gender, data = SicksWM,alternative = "less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sicks by gender
## W = 30584, p-value = 0.0003814
## alternative hypothesis: true location shift is less than 0
```

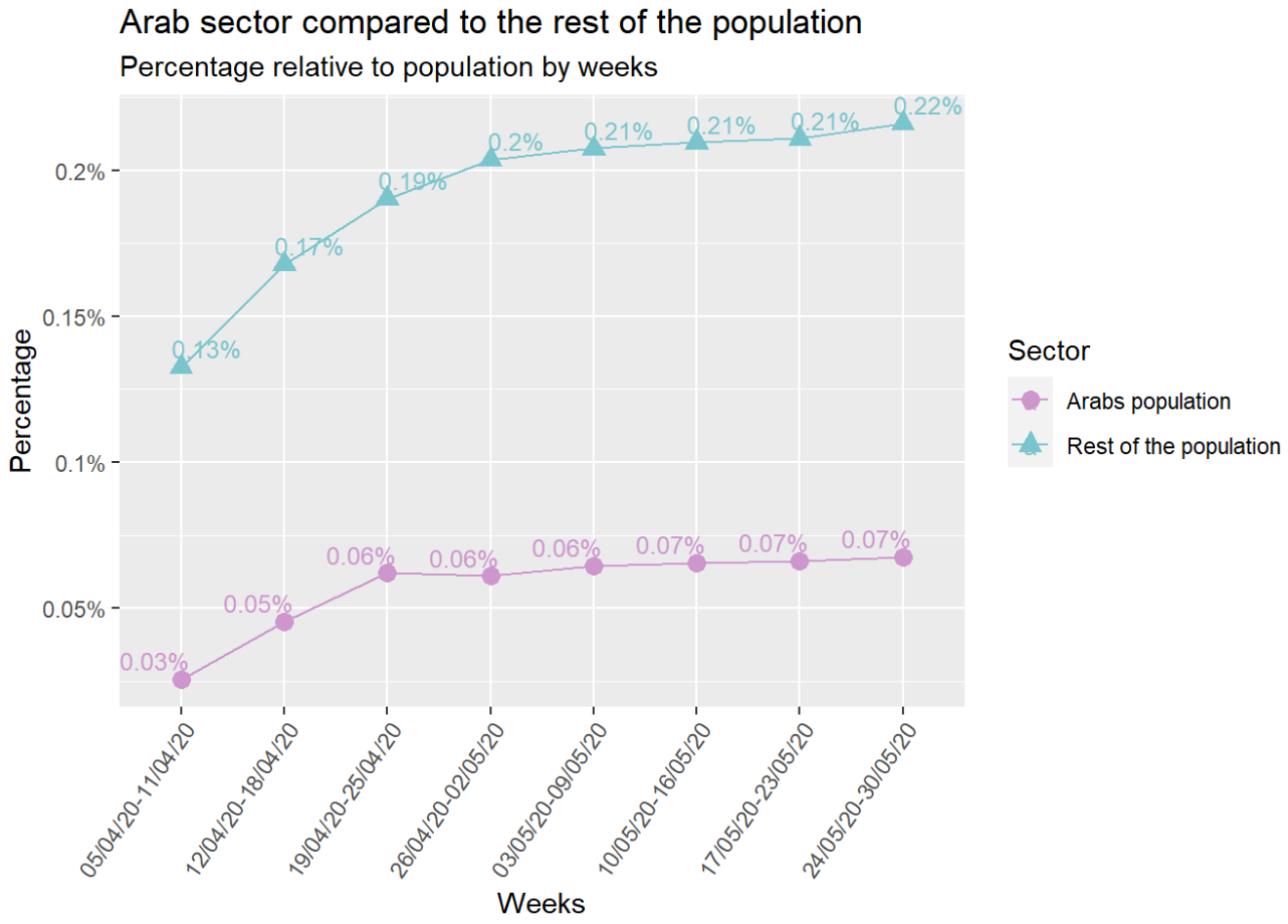
בהתחשב בגרפים המוצגים לעיל ניתן לראות כי כמות הגברים הנדבקים גבוהה מכמות הנשים לכן נשתמש ב alternative = less על מנת לבדוק את ההשערה האם כמות גברים גבוהה מכמות נשים.

מהפלט התקבל P-value קטן מ-0.05 כלומר, **נדחה** H_0 ברמת מובהקות של 5% ונסיק כי כמות הגברים הנדבקים גבוהה מכמות הנשים הנדבקות. מכאן ניתן לומר כי אכן השפעת הקורונה על גברים גבוהה מהשפעתה על נשים כפי ששיערנו על סמך תוצאות הגרפים שהתקבלו.

שאלת מחקר: האם האוכלוסייה במגזר הערבי מושפעת יותר ממחלת הקורונה?

```
#upload data
citys <- read.csv("citys.csv",header = T)

#Retrieving columns and rows for the rest of the population
vars <- c(2,9,15,19,24,29,31,33,36)
restP <- c(1:19,21,23,24,26:32,34:44,46,48:55,58,60:62,65:69,71:73,75,78:83,87,89:91,94:95,102,107,111,112,116,11
9,121,125,127,133,136,140,142:146,153,155,157,159:161,163,165,166,169,170,175,176,178,179,183:191)
restPop <- citys[restP,vars]
colnames(restPop) <- c("Population","05/04/20-11/04/20","12/04/20-18/04/20","19/04/20-25/04/20","26/04/20-02/05/2
0","03/05/20-09/05/20","10/05/20-16/05/20","17/05/20-23/05/20","24/05/20-30/05/20")



#Change data type and place 0 instead of NULL
restPop[5:9][restPop[5:9] =="-"]<- 0
restPop[is.na(restPop)]<- 0
restPop[] <- lapply(restPop, function(x) as.numeric(as.character(x)))
restPop[is.na(restPop)]<- 0

#Schema of each column in data frame
newRestPop <- as.data.frame(lapply(restPop, sum))
colnames(newRestPop) <- c("Population","05/04/20-11/04/20","12/04/20-18/04/20","19/04/20-25/04/20","26/04/20-02/0
5/20","03/05/20-09/05/20","10/05/20-16/05/20","17/05/20-23/05/20","24/05/20-30/05/20")


#Calculation of the percentage of infections in relation to the population
for(i in (2:9)){
  newRestPop[i] <- (newRestPop[i]/newRestPop[1])
}

#Add a column
newRestPop$Sector <- "Rest of the population"
```

```
#Creation data frame for the Arab sector by complements
arabPop <- citys[-restP,vars]
colnames(arabPop) <- c("Population","05/04/20-11/04/20","12/04/20-18/04/20","19/04/20-25/04/20","26/04/20-02/05/2
0","03/05/20-09/05/20","10/05/20-16/05/20","17/05/20-23/05/20","24/05/20-30/05/20")

#Change data type and place 0 instead of NULL
arabPop[5:9][arabPop[5:9] =="-"]<- 0
arabPop[is.na(arabPop)]<- 0
arabPop[] <- lapply(arabPop, function(x) as.numeric(as.character(x)))
arabPop[is.na(arabPop)]<- 0


#Schema of each column in data frame
newArabPop <- as.data.frame(lapply(arabPop, sum))
colnames(newArabPop) <- c("Population","05/04/20-11/04/20","12/04/20-18/04/20","19/04/20-25/04/20","26/04/20-02/0
5/20","03/05/20-09/05/20","10/05/20-16/05/20","17/05/20-23/05/20","24/05/20-30/05/20")


#Calculation of the percentage of infections in relation to the population
for(i in (2:9)){
  newArabPop[i] <- (newArabPop[i]/newArabPop[1])
}

#Add a column
newArabPop$Sector <- "Arabs population"

#Vertical linkage of data frames
allPop <- rbind(newRestPop, newArabPop)

#Change data frame format from Wide To long
newAllPop <- melt(allPop,
        id.vars="Sector",
        measure.vars=c("05/04/20-11/04/20","12/04/20-18/04/20","19/04/20-25/04/20","26/04/20-02/05/20","03/05/20-
09/05/20","10/05/20-16/05/20","17/05/20-23/05/20","24/05/20-30/05/20"),
        variable.name="dates",
        value.name="Proportion")
```



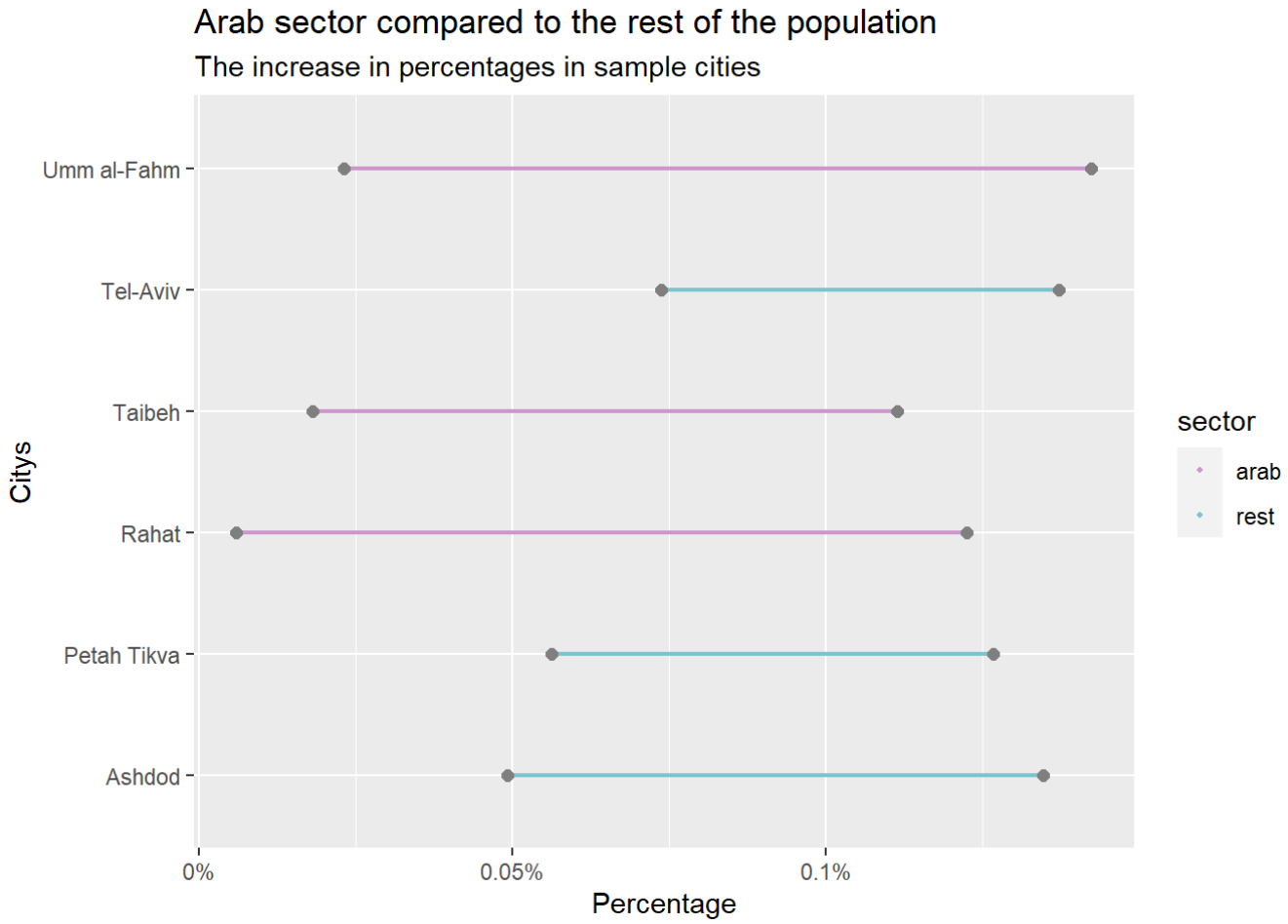
הגרף מתאר את אחוז הנדבקים ביחס לאוכלוסייה בכל שבוע בין החודשים אפריל ומאי עבור המגזר הערבי ויתר האוכלוסייה בישראל.

מתוצאות הגרף ניתן לראות כי בכל אחד מן השבועות אחוז הנדבקים ביתר האוכלוסייה **עולה** על אחוז הנדבקים במגזר הערבי, מכאן ניתן להסיק כי **אין השפעה** גבוהה יותר של המגזר הערבי על מחלת הקורונה.

```
#Retrieving columns and rows for sample citys
vars <- c(1,2,4,37)
Population <- c(2,4,6,25,33,45)
```

```r
midgam <- citys[Population,vars]
colnames(midgam) <- c("city","Population","First_disease","Last_disease")

#Change data type
midgam[4]<-as.numeric(midgam$Last_disease)

#Rename the Cities column and add a sector column accordingly
midgam$city<-c("Tel-Aviv","Petah Tikva","Ashdod","Rahat","Umm al-Fahm","Taibeh")
midgam$sector<-c("rest","rest","rest","arab","arab","arab")

#Calculation of the percentage of infections in relation to the population in city
for(i in (3:4)){
  midgam[i] <- (midgam[i]/midgam[2])
}
```

```r
#Create a graph of type geom dumbbell
ggplot(midgam, aes(x=First_disease, xend=Last_disease, y=city, group=city)) +
    geom_dumbbell(aes(color=sector),
                  size=0.75, , size_x = 2, size_xend = 2,
                  colour_x = "gray50", colour_xend = "gray50")+
    scale_x_continuous(labels = function(x) paste0(x*100, "%"))+
    scale_color_manual(values=c("plum3","cadetblue3"))+
    labs(title="Arab sector compared to the rest of the population",
        subtitle="The increase in percentages in sample cities",
        y="Citys",x="Percentage")
```



נבחן את שאלת המחקר באופן שונה כך שנבדוק את העלייה באחוז הנדבקים, גרף זה מציג עבור 3 ערים מדגמיות מכל אוכלוסייה את העליה באחוז הנדבקים מתחילת חודש אפריל ועד סוף חודש מאי.

מהגרף נסיק מסקנה **שונה** כיוון שבכל אחת מהערים במגזר הערבי התקבלה עלייה גבוהה יותר, כלומר ייתכן ו**קיימת השפעה** של המגזר הערבי על הקורונה.

ניתוח סטטיסטי

מבחן t הינו מבחן הבודק האם קיים הבדל בין ממוצעי שני אוכלוסיות שונות. נרצה לבצע השוואה בין אוכלוסיית המגזר הערבי ליתר האוכלוסיה בישראל, שתי האוכלוסיות אינן תלויות ולכן בחרנו במבחן t למדגמים בלתי תלויים כמבחן המתאים ביותר. במבחן נבדוק האם ערים במגזר הערבי מושפעות יותר ממחלת הקורונה.

מקור: #/https://researchbasics.education.uconn.edu/t-test

:t test

השערות:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 > \mu_2$

```
#Retrieve relevant columns and rows
vars <- c(1,2,4:29)
restP <- c(1:19,21,23,24,26:32,34:44,46,48:55,58,60:62,65:69,71:73,75,78:83,87,89:91,94:95,102,107,111,112,116,11
9,121,125,127,133,136,140,142:146,153,155,157,159:161,163,165,166,169,170,175,176,178,179,183:191)
restPopulation <- citys[restP,vars]

#Add a column
restPopulation$Sector <- "Rest of the population"

#Creation data frame for the Arab sector by complements
arabPopulation <- citys[-restP,vars]
arabPopulation$Sector <- "arab sector"
allPopulation <- rbind(arabPopulation,restPopulation)


#Change data type and place 0 instead of NULL
allPopulation[20:28][allPopulation[20:28] =="-"]<- 0
allPopulation[is.na(allPopulation)]<- 0
allPopulation[20:28] <- lapply(allPopulation[20:28] ,function(x) as.numeric(as.character(x)))
allPopulation[is.na(allPopulation)]<- 0


#Change data frame format from Wide To long
alltest<- melt(allPopulation,
         id.vars="Sector",
         measure.vars=c(3:28),
         variable.name="Dates",
```

```
         value.name="Quantity")

#t test
t.test(Quantity~ Sector, data = alltest,var.equal=TRUE,alternative="less")
```

```
##
##  Two Sample t-test
##
## data:  Quantity by Sector
## t = -11.614, df = 4964, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -76.44921
## sample estimates:
##            mean in group arab sector mean in group Rest of the population
##                             7.932692                            96.997993
```

בהתחשב בנתונים ניתן לשער כי הכמות הנדבקים ביתר האוכלוסיה גבוהה מהכמות במגזר הערבי לכן נשתמש ב alternative = less על מנת לבדוק את ההשערה האם כמות נדבקי יתר האוכלוסיה גבוהה מכמות נדבקי המגזר הערבי.

מהפלט התקבל P-value קטן מ-0.05 כלומר, **נדחה** H_0 ברמת מובהקות של 5% ונסיק כי ממוצע יתר האוכלוסיה גבוה ממוצע האוכלוסיה במגזר הערבי.

ניתן לראות כי תוצאות מבחן t שוות ערך להשערתינו על כך כי אין השפעה גבוהה יותר של מחלת הקורונה על ערים במגזר הערבי, כפי שקיבלנו בתוצאות הגרף הראשון.

שאלת מחקר: האם יש קשר בין אזורים מגורים(דרום,מרכז וצפון) ושבוע קלנדרי לבין כמות הנדבקים בקורונה?

```
#קריאת בסיס הנתונים גיליון 1 באקסל
q5_data <- read_excel("q5.xlsx", sheet=1)
colnames(q5_data)<- c("aren_number", "area", "City", "Population","03/04/2020", "05/04/2020",   "06/04/2020",
"07/04/2020",   "09/04/2020",   "11/04/2020",   "12/04/2020",   "13/04/2020",   "14/04/2020",   "16/04/2020",
"17/04/2020",   "18/04/2020",   "19/04/2020",   "20/04/2020",   "23/04/2020",
                      "24/04/2020", "26/04/2020",   "29/04/2020",   "30/04/2020",   "01/05/2020",   "02/05/2020",
"04/05/2020",   "05/05/2020",   "06/05/2020",   "08/05/2020",   "09/05/2020",
```

```
                        "10/05/2020", "15/05/2020",   "18/05/2020",   "21/05/2020",   "24/05/2020",   "29/05/2020",
"30/05/2020")

#סידור הנתונים לפי שבועות
aa<- lapply(q5_data[,c(4:37)], function(x) as.numeric(as.character(x)))
clean_data <- cbind(q5_data[, c(1:3)], aa)

#0 כל מה שריק השמה
clean_data[is.na(clean_data)] <- 0

#השמה של כל השבועות בחודש הרביעי בטבלה לפי הסדר
april_culomns <- colnames(clean_data)[endsWith(colnames(clean_data), "/04/2020")]

a <- melt(data = cbind(clean_data[,c(1:4)], clean_data[,april_culomns]),
          id.vars = c("aren_number", "area", "City", "Population"),
          variable.name = "date")

a$month <- "April"

#השמה של כל השבועות בחודש החמישי בטבלה לפי הסדר
may_culomns <- colnames(clean_data)[endsWith(colnames(clean_data), "/05/2020")]


b <- melt(data = cbind(clean_data[,c(1:4)], clean_data[,may_culomns]),
          id.vars = c("aren_number", "area", "City", "Population"),
          variable.name = "date")

b$month <- "May"

both_month_data <- rbind(a,b)

completeFun <- function(data, desiredCols) {
  completeVec <- complete.cases(data[, desiredCols])
  return(data[completeVec, ])
}
```

```
clean_data <- cbind(q5_data[, c(1:3)], aa)
unique(clean_data$aren_number)
```

```
## [1]  2  3  1 NA
```

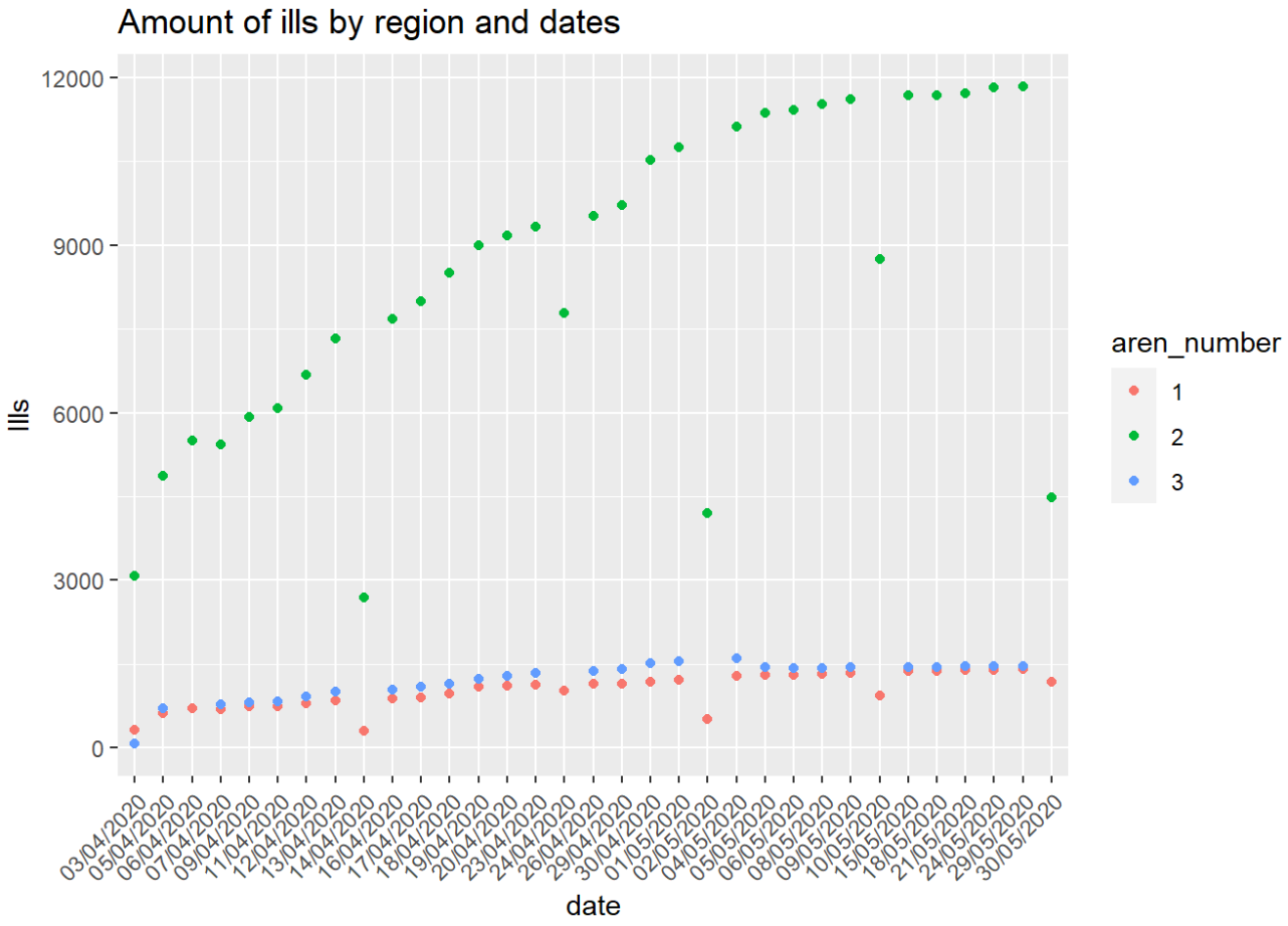
```
s<-completeFun(clean_data, 'aren_number')
unique(s$aren_number)
```

```
## [1] 2 3 1
```

```
c<- melt(data = s,
         id.vars = c("aren_number", "area", "City", "Population"),
         variable.name = "date", value.name = "Ills"
)

unique(c$aren_number)
```

```
## [1] 2 3 1
```

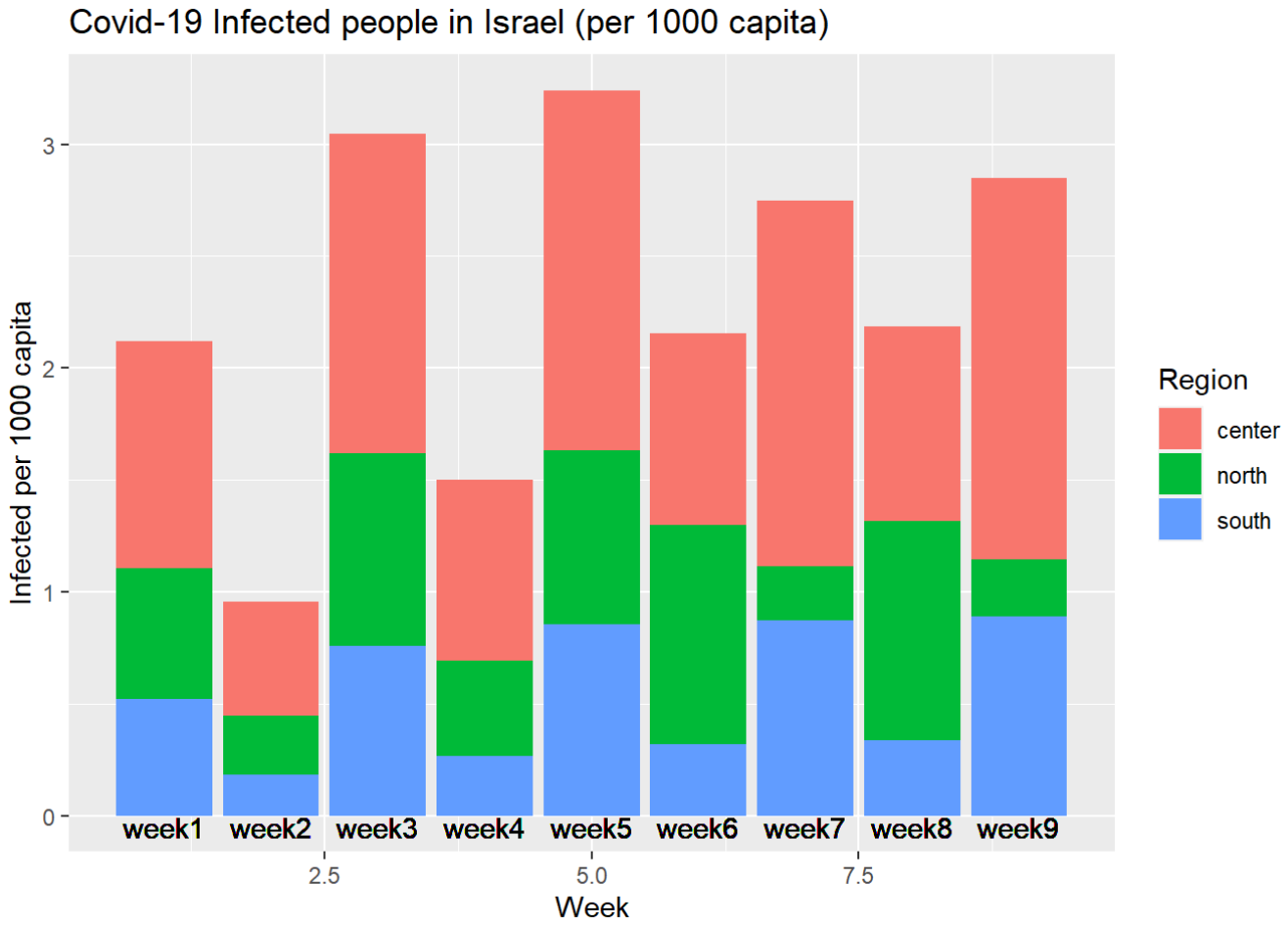
```
c$aren_number <- factor(c$aren_number)
c$date <- factor(c$date)

gd<- aggregate(Ills ~ aren_number + date, data = c, FUN = 'sum')

graphQ5<-ggplot(gd, aes(x=date, y=Ills, color=aren_number))+
  geom_point() + labs(title = "Amount of ills by region and dates")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
graphQ5
```



גרף 2: סך הכל נדבקים ל-1000 איש

```
Q5<-read.csv('Q5.csv',header = T)

q5data<- data.frame(Q5)
colnames(q5data)<-c('week','region','ills(per 1000 capita)')

graphQ52<-ggplot(data = q5data,aes(x=q5data$week,y=q5data$`ills(per 1000 capita)`,fill=q5data$region))+
```

```
  geom_bar(stat='identity')+labs(title = 'Covid-19 Infected people in Israel (per 1000 capita)',fill='Region')+
  xlab('Week')+ylab('Infected per 1000 capita')+geom_text(x=1,y=-0.05,label='week1')+geom_text(x=2,y=-0.05,label=
'week2')+
  geom_text(x=3,y=-0.05,label='week3')+geom_text(x=4,y=-0.05,label='week4')+geom_text(x=5,y=-0.05,label='week5')+
  geom_text(x=6,y=-0.05,label='week6')+geom_text(x=7,y=-0.05,label='week7')+geom_text(x=8,y=-0.05,label='week8')+
  geom_text(x=9,y=-0.05,label='week9')


graphQ52
```



ניתוח סטטיסטי

בסיס הנתונים מכיל פרטים על כמות חולים באזורים השונים, נקודת ההנחה היא שבמרכז כמות הנדבקים גדולה משמעותית מאזורי הצפון והדרום, בכדי שתהיה פרופורציה בין כמות הנדבקים לבין כמות נבדקת ביצענו חישובים ונבדקו כמות הנדבקים לכל אלף איש. המבחן שבחרנו כדי לבדוק את שאלת מחקר זו הוא מבחן Anova חד כיווני, כיוון שיש צורך לבדוק את שוויון תוחלות הכמויות באזורים השונים (דרום, מרכז וצפון). השערת האפס בניתוח היא שאכן מתקיים שוויון תוחלות, ההשערה האלטרנטיבית היא שאין שוויון. מבחן anova הוא מבחן סטטיסטי הבודק את ההבדלים בין תוחלות באזורים השונים, המבחן בודק האם נמצאו הבדלים בין קבוצות במדגם והאם הם אכן משקפות הבדלים אמיתיים באוכלוסיה.

מקור: http://www.p-value.co.il/category/%D7%A0%D7%99%D7%AA%D7%95%D7%97-%D7%A9%D7%95%D7%A0%D7%95%D7%AA-%D7%97%D7%93-%D7%9B%D7%99%D7%95%D7%95%D7%A0%D7%99.

Anova test:

הגדרת משתנים

μ_1 -מייצג את תוחלת כמות החולים בדרום הארץ

μ_2 -מייצג את תוחלת כמות החולים במרכז הארץ

μ_3 -מייצג את תוחלת כמות החולים בצפון הארץ

השערות:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1:else

```
#אנובה להשפעת אזור על חולים לכל 1000 איש
lm2<-lm(q5data$`ills(per 1000 capita)`~q5data$region,data =q5data )
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: q5data$`ills(per 1000 capita)`
##               Df Sum Sq Mean Sq F value   Pr(>F)
## q5data$region  2 2.0445 1.02226   8.202 0.001929 **
## Residuals     24 2.9912 0.12464
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

מהפלט התקבל ש-Pvalue שואף לאפס וקטן מ-5% לכן, נדחה H_0 ברמת מובהקות של 5% ונסיק כי התוחלות שונות כלומר כמות הנדבקים בכל אזור היא שונה.

סיכום ומסקנות

מניתוח השפעת מחלת הקורונה על האוכלוסייה בישראל ניתן להסיק כי קיימת השפעה מבחינה דמוגרפית.

מתקיים קשר בין ממוצע גיל החולים לדרגת קושי תחלואתם (קל, בינוני וקשה), כלומר שככל שדרגת קושי התחלואה חמורה יותר, ממוצע גיל החולים מבוגר יותר בהתאמה.

מבחינת מגדר החולים הגברים באוכלוסייה הישראלית הושפעו ממחלת הקורונה יותר מנשים, בדיקה זו נעשתה ביחס לכל מצבי המחלה, המצבים הקריטיים, מונשמים ומאושפזים וכן דרגות הקושי קל, בינוני וקשה.

מבדיקת השפעת המגזר הערבי על כמות הנדבקים התקבלה התוצאה כי ערים במגזר הערבי אינם מהווים השפעה, תוצאה זו מעניינת כיוון שאנו רגילים לשמוע בחדשות שההדבקה בערים אלו גבוהה יותר כיוון שיש משפחות יותר גדולות וקיימת צפיפות אוכלוסין.

מבחינת השפעת אזורים גאוגרפיים (דרום, מרכז, צפון) התקבל קשר מובהק בין אזור מגורים לכמות הנדבקים בקורונה, כמו כן ראינו כי באזור מרכז כמות הנדבקים היא הגדולה ביותר בהשוואה לדרום ולצפון.

מסקנה נוספת מבחינת הקשר של המצבים החמורים של המחלה, מונשמים ומאושפזים היא שאכן קיים קשר שהינו קשר לינארי עולה חזק בין המשתנים הנ"ל.

מודל כלכלי

ניתוח זמן כדאי לקניית כרטיס טיסה בקו תל-אביב ניו-יורק

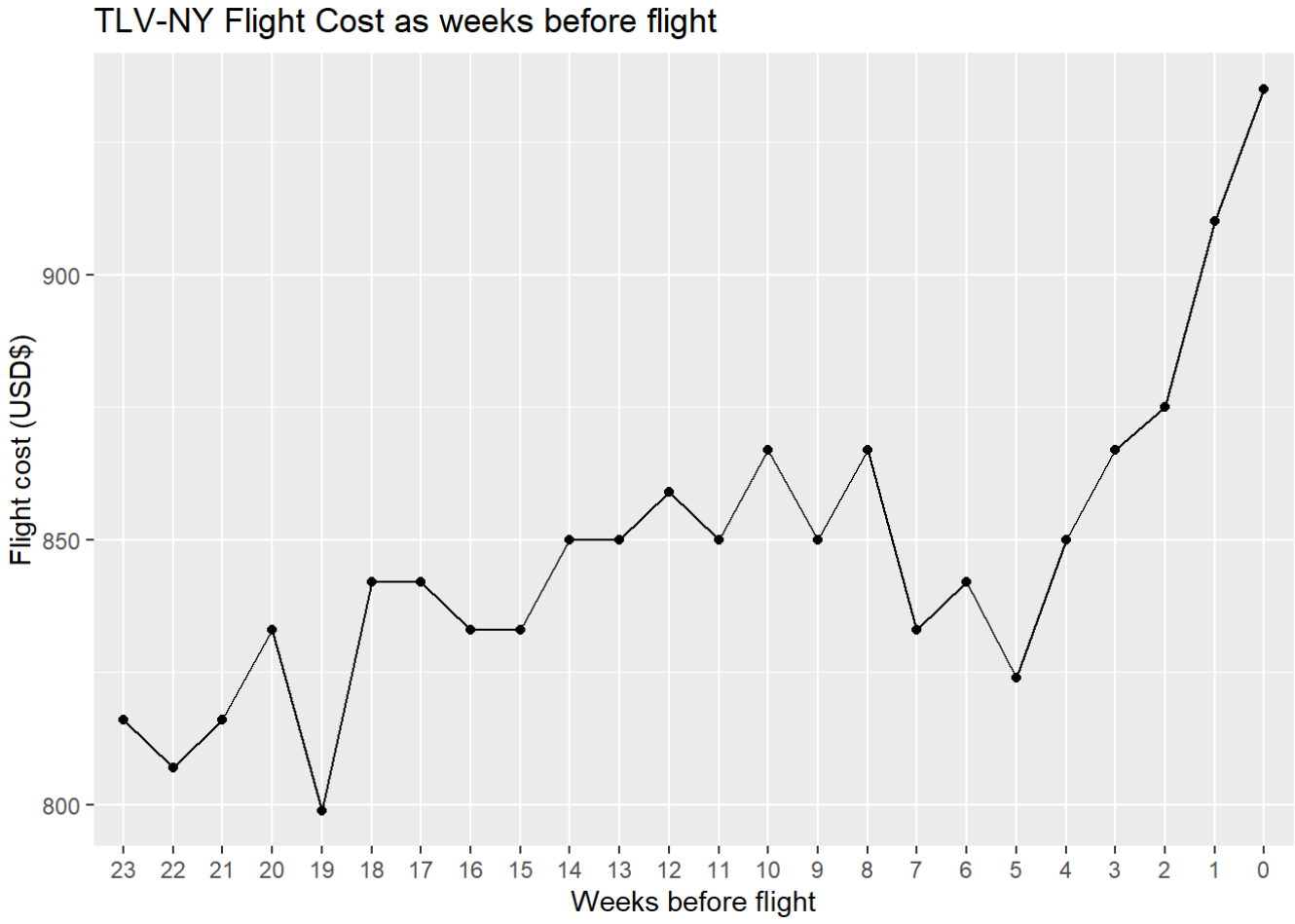
בניתוח הבא ננתח מהו הזמן המיטבי מבחינת עלות, לקניית כרטיס טיסה לניו-יורק מתל-אביב תוך כדי התייחסות לפרמטר הריבית שאותו עלינו לקחת בחשבון בתשלום מראש של תקופה.

בחרנו להשתמש בנתונים מאתר ,סקאי סקאנר (https://www.skyscanner.co.il/bttb/best-time-to-book-cheap-flights)

עלויות הטיסה בדולרים כתלות במספר השבועות מראש שהוזמן הכרטיס

```
flight<- read.csv('flight1.csv',header = T)
colnames(flight)<- c('Weeksbeforeflight','Flightcost','Totalflightcost')
```

```
flight$Weeksbeforeflight<-factor(flight$Weeksbeforeflight,levels=c(23:0))


eg1<-ggplot(data = flight,aes(x=Weeksbeforeflight,y=Flightcost,group=1))+
  geom_line()+
  geom_point()+
  labs(title = 'TLV-NY Flight Cost as weeks before flight',y='Flight cost (USD$)',x='Weeks before flight')
 eg1
```



$$FV_n = C_0 \cdot (1+r)^n = PV \cdot (1+r)^n$$

FV- מצייג ערך עתידי שנצטרך להחזיר בסוף תקופת ההלוואה

Co- מייצג את הערך ההתחלתי של ההלוואה

r- ריבית לתקופה n

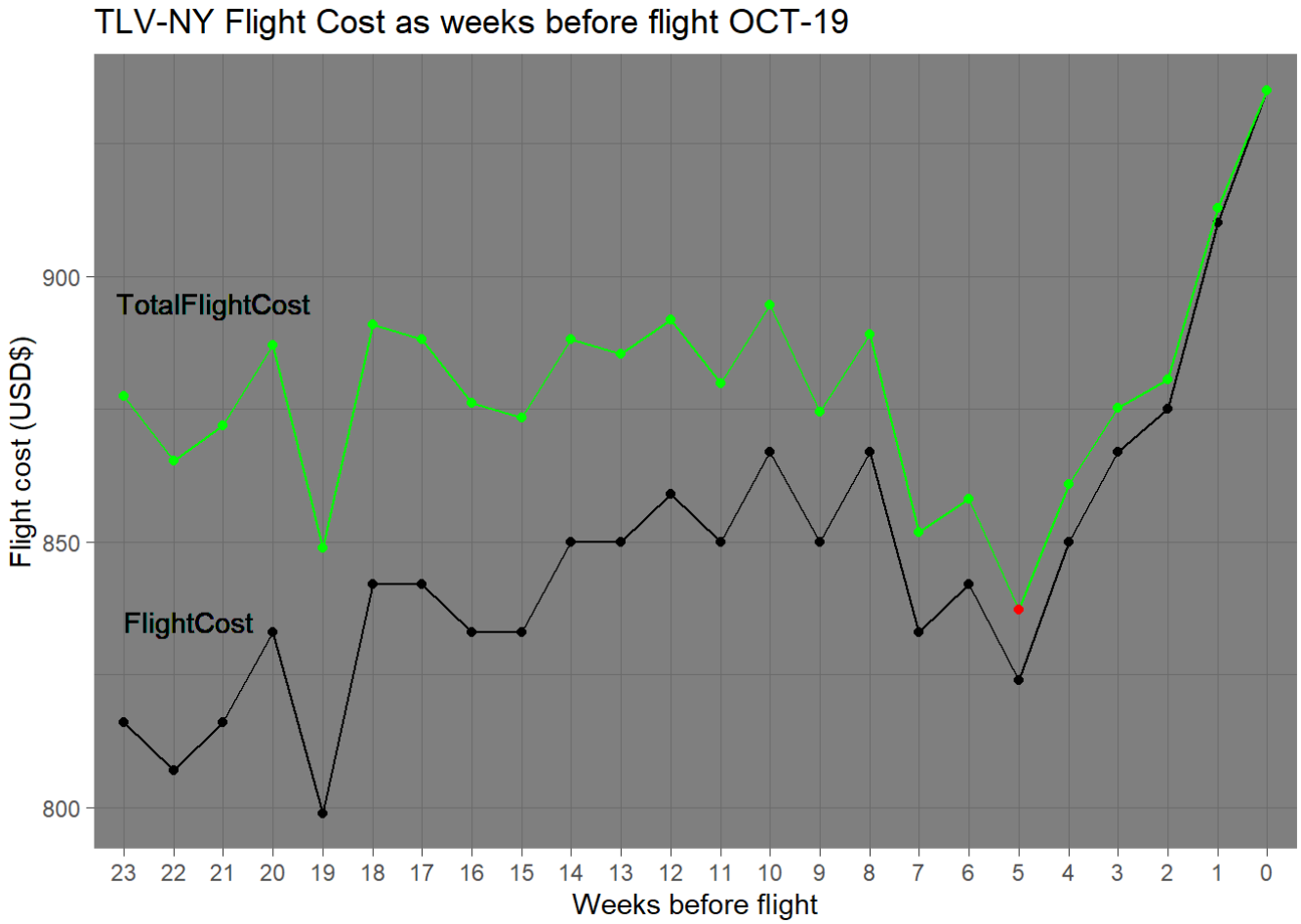
n-מספר התקופות שהריבית נצברת

נניח כי ההלוואה שהוצעה לנו תוחזר ביום הטיסה והיא נושאת ריבית שבועית של 0.2% בחישוב ריבית דריבית ונניח שעלות ביטוח ביטול הטיסה היא דולר לשבוע מראש

(ריבית אפקטיבית מגולמת לשנה 10.94%)

השוואה בין עלות הטיסה הנקובה למול עלות הטיסה הכוללת

```
eg2<-ggplot(data = flight,aes(x=Weeksbeforeflight,y=Flightcost,group=1))+
  geom_line()+
  geom_point()+
  geom_line(data = flight,aes(x=Weeksbeforeflight,y=Totalflightcost,group=1),color='Green')+
  geom_point(data = flight,aes(x=Weeksbeforeflight,y=Totalflightcost,group=1),color='Green')+
  labs(title = 'TLV-NY Flight Cost as weeks before flight OCT-19',y='Flight cost (USD$)',x='Weeks before flight')
+
  theme_dark()+
  geom_text(x=2.8,y=895,label='TotalFlightCost')+
  geom_text(x=2.3,y=835,label='FlightCost')+
  geom_point(x=19,y=837.27,color='Red')
eg2
```



ניתן לראות כי בשבוע החמישי לפני מועד הטיסה (הנקודה האדומה בגרף העליון) העלות הכוללת היא מינימלית ולכן הזמן המיטבי לאחר לקיחה בחשבון את משתני הריבית ועלות ביטוח ביטול נסיעה הוא השבוע החמישי לפני מועד הטיסה.