Analysis of variance

Meidan Greenberg

January 2020

קורס תכנון ניסויים 5000571, תש"פ סמסטר א'

המחלקה לתעשיה וניהול, מכללת סמי שמעון באר שבע

לידי: ד"ר יצחק מינצ'וק
מגיש: מידן גרינברג, ת.ז 203547500

הקדמה

במסגרת פרוייקט זה, ננתח ע"י שיטות שנלמדו בקורס את שאלת המחקר:
האם ישנם הבדלים באחוזי ההשמה בין אזורי ההתיישבות במדינה?
כלומר במילים אחרות, נרצה להשוות בין יעילות לשכות התעסוקה באזורים השונים בארץ.

לשם כך, נעשה שימוש בבסיס נתונים ממשלתי (https://data.gov.il/dataset/e-data-gov-il-dataset-yeshuvmoatzadata) מטעם שירות התעסוקה.

סטטיסטיקה תיאורית

לצורך הניסוי, נשווה בין אחוזי ההשמה של לשכות התעסוקה ב7 אזורים שונים בארץ שנדגמו במהלך השנים 2010-2019. המודל שאיתו נעבוד הוא בלוקים, כאשר:

b = {2010-2019} = 10
k = {Central, Haifa, Jerusalem, Judea and Samaria, North, South, Tel Aviv} = 7

Block model:
$yij = \mu + \alpha j + \beta i + \epsilon ij$
$(i = 1, \ldots, b | j = 1, \ldots, k)$

Assumptions:
(1) $\varepsilon ij \sim N(\mu, \sigma\varepsilon^2)$ ב"ת
(2) $\sum_{j=1}^{k} \hat{\alpha j} = 0$
(3) $\sum_{i=1}^{b} \hat{\beta i} = 0$

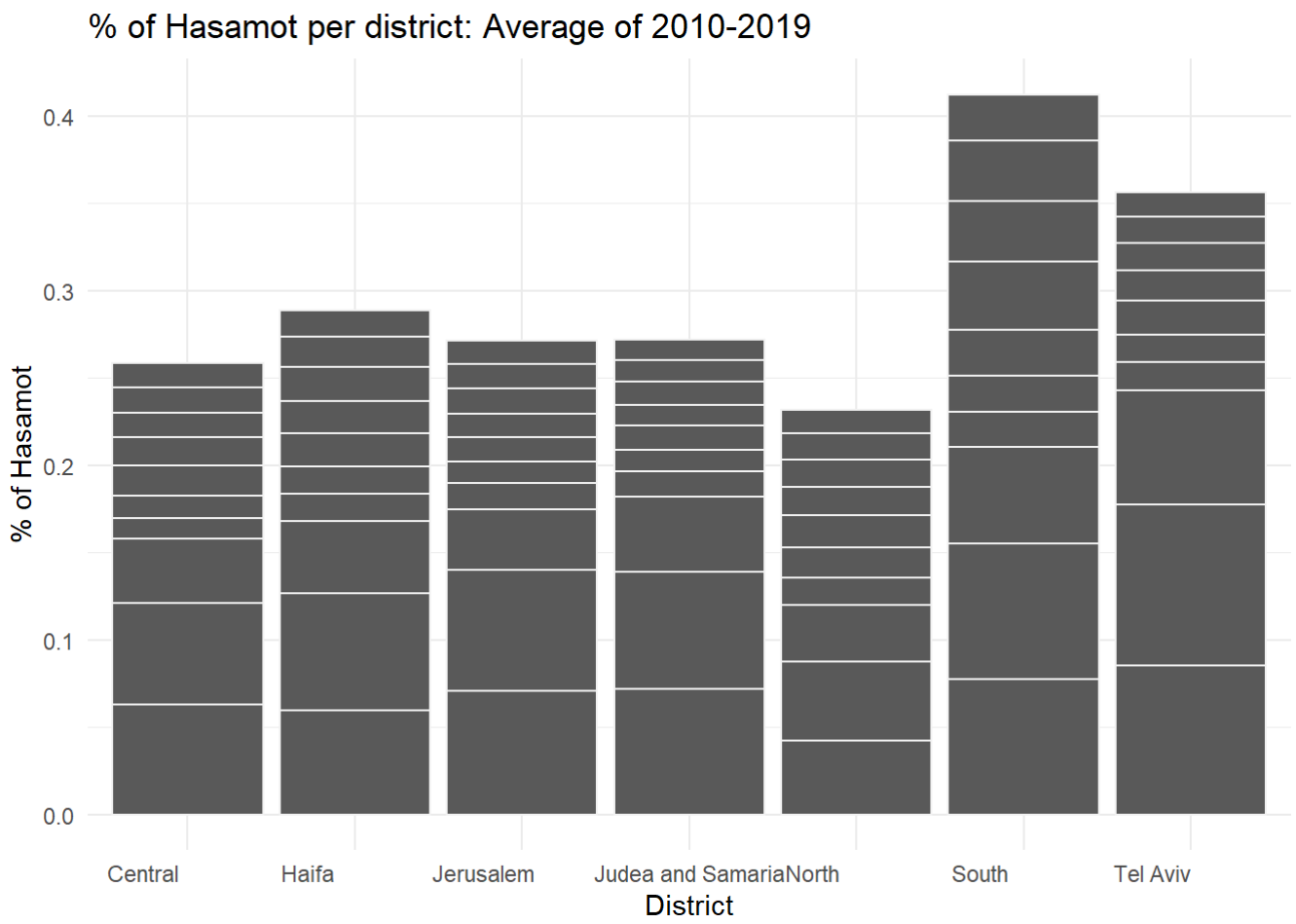
את הנחות 2,3 נבדוק בהמשך.
בשלב הראשוני, נשלוף את הנתונים הרלוונטיים לניסוי.

```
infodata <- read.csv("infoyeshuv.csv",sep=",", header=TRUE, stringsAsFactors = FALSE)
infodata$OnlyYear <-  substr(infodata$Month,1,4)  #Adding Only Year new column
infodata$HasamotPerJob <- infodata$Placement.from.reference / infodata$Total.jobseekers #Addi
ng % of Hasamot column

#Selecting values
totaldata <- infodata %>%
  group_by(Cbs.district, OnlyYear) %>%
  summarise(Hasamot_Mean = mean(HasamotPerJob)) %>%
  arrange(Cbs.district)
```

ננסה לחזות את התוצאות ע"י גרף.

```
ggplot(totaldata, aes(x=totaldata$Cbs.district, y=totaldata$Hasamot_Mean)) +
  geom_bar(stat = "identity", aes(x=totaldata$Cbs.district), color="gray95") + theme_minimal
() +
  xlab("District")+ylab("% of Hasamot") +ggtitle("% of Hasamot per district: Average of 2010-
2019")
```



ניתן לראות שאחוזי ההשמה הגבוהים ביותר שייכים ללשכות התעסוקה בדרום הארץ ובתל אביב.
נמשיך לאמת זאת ע"י ניתוח שונות.
נסדר את הנתונים בטבלה.

```
#Creating table
datatable<-xtabs(totaldata$Hasamot_Mean ~ totaldata$OnlyYear + totaldata$Cbs.district)
names(dimnames(datatable)) <- c("Year", "District")
print.table(datatable)
```

```
##       District
## Year   Central                Haifa                  Jerusalem
##   2010             0.06261178            0.05931714             0.07050701
##   2011             0.05812649            0.06714543             0.06925599
##   2012             0.03696544            0.04148120             0.03512852
##   2013             0.01187121            0.01556941             0.01502485
##   2014             0.01296354            0.01590219             0.01186527
##   2015             0.01747340            0.01899923             0.01431266
##   2016             0.01607261            0.01803400             0.01355608
##   2017             0.01388688            0.01950510             0.01435571
##   2018             0.01435760            0.01762976             0.01383057
##   2019             0.01385422            0.01496878             0.01355407
##       District
## Year   Judea and Samaria North                South
##   2010        0.07166440           0.04210307           0.07750330
##   2011        0.06699475           0.04550277           0.07760772
##   2012        0.04331700           0.03236457           0.05556887
##   2013        0.01468749           0.01544218           0.01964096
##   2014        0.01199578           0.01732308           0.02071914
##   2015        0.01388761           0.01874606           0.02649701
##   2016        0.01198107           0.01588801           0.03906142
##   2017        0.01331490           0.01569967           0.03459395
##   2018        0.01233664           0.01503809           0.03477294
##   2019        0.01193105           0.01378549           0.02637729
##       District
## Year   Tel Aviv
##   2010            0.08548725
##   2011            0.09219846
##   2012            0.06517631
##   2013            0.01615196
##   2014            0.01561775
##   2015            0.01939989
##   2016            0.01729611
##   2017            0.01559937
##   2018            0.01531068
##   2019            0.01425509
```

בכדי לוודא שהנתונים שבטבלה אכן נכונים למודל, נשווה בין ממוצע ממוצעי העמודות, לבין ממוצע ממוצעי השורות, הרי שנקבל בשני המקרים את:

$\overline{y..} = \mu$

```
mean(rowMeans(datatable))   #Yi.
```

```
## [1] 0.02987093
```

```
mean(colMeans(datatable))   #Y.j
```

```
## [1] 0.02987093
```

```
mean(rowMeans(datatable)) == mean(colMeans(datatable)) #Yi.=Y.j = Y..
```

```
## [1] TRUE
```

נבדוק את הנחה (2) לעיל, כלומר האם מתקיים:

$$\sum_{j=1}^{k} \hat{\alpha j} = 0$$

כאשר:

$$\hat{\alpha j} = \overline{y.j} - \overline{y..}$$

```
sumj<-0
for(j in 1:7) { #k=7
sumj <- sumj+ (mean(datatable[,j] - mean(colMeans(datatable))))
}
sumj
```

```
## [1] -1.387779e-17
```

הסכום שהתקבל שואף ל0. ומכאן נסיק שההנחה מתקיימת.
נמשיך לבדיקת הנחה (3) לעיל, כלומר האם מתקיים

$$\sum_{i=1}^{b} \hat{\beta i} = 0$$

כאשר:

$$\hat{\beta i} = \overline{yi.} - \overline{y..}$$

```
sumi<-0
for(i in 1:10) { #b=10
sumi <- sumi+ (mean(datatable[i,] - mean(rowMeans(datatable))))
}
sumi
```

```
## [1] -1.040834e-17
```

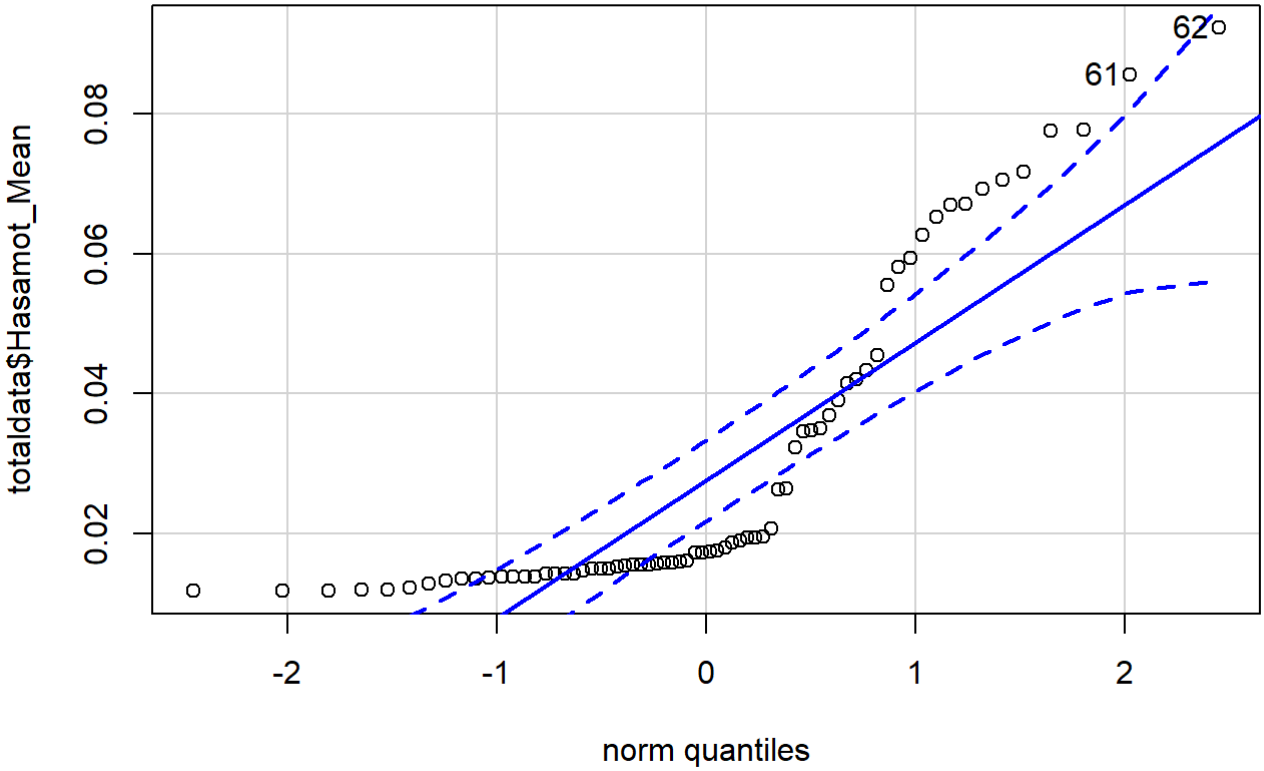
גם כאן קיבלנו שהסכום שואף ל0.

בדיקת הנחת התפלגות נורמלית

בכדי לבדוק הנחת התפלגות נורמלית, נשתמש ב2 כלים עיקריים:
1. Q-Q Plot
2. Kolmogorov Smirnof test

Q-Q Plot

```
qqPlot(totaldata$Hasamot_Mean)
```




```
## [1] 62 61
```

ניתן לראות שהנקודות שואפות להיצמד לקו המגמה, אך אי אפשר להגיד שהן נמצאות בטווחו.
נמשיך לבדיקת קולמוגורוב-סמירנוף.

Kolmogorov - Snirnof Test

$when:$
Xi= אחוזי ההשמה הכוללים של אזור i
$H0: Xi \sim N(\mu, \sigma^2)$
$H1: else$

```
ks.test(totaldata$Hasamot_Mean, pnorm)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  totaldata$Hasamot_Mean
## D = 0.50473, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

ערך הסטטיסטי שקיבלנו (0.50473), אינו אופטימלי להחלטה כיוון שלא שואף ל0 או ל1. הנתון תואם לתצוגת הqqPlot.
היות וערך pValue שואף ל0, נוכל להסיק כי ייתכן ואחוזי ההשמה מתפלגים בצורה נורמלית כלשהי, אך לא בהכרח תחת התפלגות נורמלית סטנדרטית.

בדיקת שיוויון שונויות

לצורך בדיקת שיוויון שונויות בין הטיפולים (אזורים), נשתמש במבחן קוחרן היות וגודלי המדגם שווים.

$$H0 : \sigma 1^2 = \sigma 2^2 = ... = \sigma k^2 = \sigma$$
$$H1 : else$$

```
#Cochran test
varj.vector <- c(var(datatable[,1]), var(datatable[,2]), var(datatable[,3]), var(datatable[,4
]),
                var(datatable[,5]),var(datatable[,6]),var(datatable[,7]))

#Logical testing
max(varj.vector)/sum(varj.vector) >  0.3751 # G0.05,7,9
```

```
## [1] FALSE
```

במבחן הדחייה קיבלנו שליליות, כלומר לא נדחה H0.
לפיכך נסיק בר"מ 0.05 כי השונויות של שנות הדגימה שוות בינהן.

טבלת ANOVA

ניצור טבלת ANOVA וננתח את תוצאותה.

```
anova.final <- aov(formula = totaldata$Hasamot_Mean ~ totaldata$Cbs.district +
                     totaldata$OnlyYear, data=totaldata)
summary(anova.final)
```

```
##                        Df   Sum Sq  Mean Sq F value   Pr(>F)
## totaldata$Cbs.district  6 0.002392 0.000399   8.701 1.24e-06 ***
## totaldata$OnlyYear      9 0.029531 0.003281  71.632  < 2e-16 ***
## Residuals              54 0.002474 0.000046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ניתוח התוצאות:
ערכי F חושבו כך שהpValue שלהם נמוך יותר מ0.001, ולכן נשווה את הסטטיסטיים בטבלת F עבור ר"מ 0.001.
נבדוק האם הבלוקים מובהקים, כלומר:

$$H0 : \beta 1 = \beta 2 = ... = \beta 10 = 0$$
$$H1 : else$$

נבדוק האם מתקיים:

$$F1(FOnlyYear) > Fdfblock, dfMSE, \alpha = F9, 54, 0.001 = 3.6$$
$$F1 = 71.632 >> 3.6$$

נדחה H0 בר"מ 0.001 ונסיק כי הבלוקים מובהקים.
נמשיך לבדיקת הגורם.

$$H0 : \alpha 1 = \alpha 2 = ... = \alpha 7$$
$$H1 : else$$

נבדוק האם מתקיים:

$$F2(FDistrict) > FdfSSA, dfMSE, \alpha = F6, 54, 0.001 = 4.4$$
$$F2 = 8.701 > 4.4$$

נדחה H0 בר"מ 0.001 ונסיק כי יש הבדלים באחוזי ההשמות בין האזורים השונים בארץ.

נמשיך לקבוצות הומוגניות

קבוצות הומוגניות

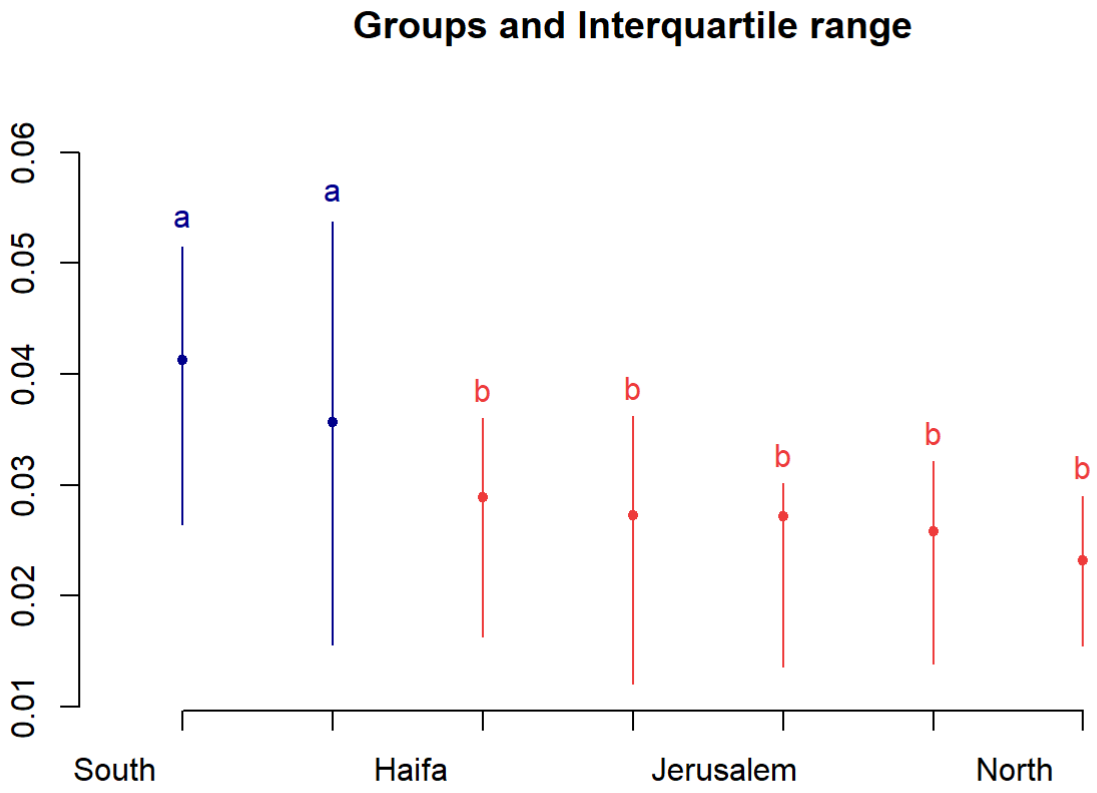
נבדוק את הומוגניות האזורים השונים בארץ בעזרת מבחן דנקן.

```
#Duncan test

out<- duncan.test(anova.final, "totaldata$Cbs.district", console =TRUE)
```

```
##
## Study: anova.final ~ "totaldata$Cbs.district"
##
## Duncan's new multiple range test
## for totaldata$Hasamot_Mean
##
## Mean Square Error:  4.580682e-05
##
## totaldata$Cbs.district,  means
##
##                        totaldata.Hasamot_Mean        std  r        Min
## Central                            0.02581832 0.01961594 10 0.01187121
## Haifa                              0.02885522 0.01977171 10 0.01496878
## Jerusalem                          0.02713907 0.02350552 10 0.01186527
## Judea and Samaria                  0.02721107 0.02418421 10 0.01193105
## North                              0.02318930 0.01210260 10 0.01378549
## South                              0.04123426 0.02175514 10 0.01964096
## Tel Aviv                           0.03564929 0.03198808 10 0.01425509
##                              Max
## Central               0.06261178
## Haifa                 0.06714543
## Jerusalem             0.07050701
## Judea and Samaria     0.07166440
## North                 0.04550277
## South                 0.07760772
## Tel Aviv              0.09219846
##
## Alpha: 0.05 ; DF Error: 54
##
## Critical Range
##           2           3           4           5           6           7
## 0.006068318 0.006383030 0.006590269 0.006740733 0.006856426 0.006948815
##
## Means with the same letter are not significantly different.
##
##                        totaldata$Hasamot_Mean groups
## South                              0.04123426      a
## Tel Aviv                           0.03564929      a
## Haifa                              0.02885522      b
## Judea and Samaria                  0.02721107      b
## Jerusalem                          0.02713907      b
## Central                            0.02581832      b
## North                              0.02318930      b
```

```
plot(out,variation="IQR")
```



בפלט קיבלנו את הקבוצות ההומוגניות:

a = {דרום, תל אביב}

b = {צפון, ירושלים, יהודה ושומרון, מרכז, חיפה}

כאשר עבור קבוצה a, אחוזי ההשמה הם הגבוהים ביותר.

מסקנות

במסגרת הניתוח שהתבצע עבור יעילות לשכות התעסוקה באזורים השונים בארץ, ניתן להגיע למסקנות הנ"ל:

1. אחוזי ההשמות בשנים 2010-2019 נמוכים מאד, כאשר הממוצע הכולל של אחוזי ההשמות עומד על כ2% בלבד.

במילים אחרות, 98% מהפונים ללשכות התעסוקה (בממוצע) לא שובצו למשרה כלשהי, ונותרו מובטלים.

2. מתוקף ניתוח השונות והקבוצות ההומוגניות, נסיק כי לשכות התעסוקה בדרום הארץ ובעיר תל אביב הן היעילות ביותר.