

# רגסיה לニアרית ותכון ניסויים

## למערכות מידע

*תרגול חזרה, הכרות עם שפת R*

- **צוות הקורס**
  - מרצה: ד"ר מירב טיב-מימון
  - מתרגם: דור אמיר ([amire@post.bgu.ac.il](mailto:amire@post.bgu.ac.il))
- שעת קבלה: יומ שני לפני התרגול הראשון בתיאום מראש.
  
- **דרישות הקורס**
  - תרגילי בית – 10% (הגשה לתא 43)
  - חובת ההגשה של כלל התרגילים
  - ההגשה בזוגות בלבד
  - **בדיקה מדגמית !**
- בחינה סופית – 90%

## • כללי תוחלת:

$$E(b \cdot x + a) = b \cdot E(x) + a$$

$$E(x + y) = E(x) + E(y)$$

a, b constants

## • כללי שנות:

$$V(x) = E[(x - E[x])^2] = E[x^2] - (E[x])^2$$

$$V(b \cdot x + a) = b^2 \cdot V(x) + 0$$

$$V(x + y) = V(x) + V(y) + 2 \cdot COV(x, y)$$

$$COV(x, y) = E(x \cdot y) - E(x) \cdot E(y)$$

a, b constants

## • אם x ו-y הם משתנים בלתי תלויים (ב"ת) אז:

$$COV(x, y) = 0 \Rightarrow V(x + y) = V(x) + V(y)$$

$$x_i \sim ?(\mu, \sigma^2)$$

- נניח ש-  $x$  הינו משתנה מקרי:

- איך מתפלג הממוצע של  $n$  משתנים מקרים מסוג  $x$ ?

$$E(\bar{x}) = E\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{1}{n} \cdot E\left(\sum_{i=1}^n x_i\right) = \frac{1}{n} \cdot \sum_{i=1}^n [E(x_i)] = \frac{1}{n} \cdot \sum_{i=1}^n \mu = \frac{n \cdot \mu}{n} = \mu$$

$$V(\bar{x}) = V\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{1}{n^2} \cdot V\left(\sum_{i=1}^n x_i\right) = \frac{1}{n^2} \cdot \sum_{i=1}^n [V(x_i)] = \frac{1}{n^2} \cdot \sum_{i=1}^n \sigma^2 = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\bar{x} \sim ?\left(\mu, \frac{\sigma^2}{n}\right)$$

❖ **הסקה סטטיסטית עוסקת בשני מושגים:**

I. **פרמטר** – ערך קבוע, בדרך כלל אינו ידוע, המאפיין את התפלגות האוכלוסייה. למשל, בהתפלגות נורמלית:  $(\mu, \sigma^2)$

II. **סטטיסטי** – הוא פונקציה על מרחב המדגם (פונקציה של התוצאות במדגם בלבד והוא אינו תלוי בשום פרמטר לא ידוע), כלומר הוא משתנה מקרי

❖ **אמידה** - הערכת גודלו של פרמטר מסוים באוכלוסייה נתונה (למשל התוחלת) כשההערכה נעשית על סמך מדגם מקרי מתוך האוכלוסייה

- ❖ מבחןים בין אומד נקודתי עבור פרמטר, כלומר לאחר הוצאת המדגם מקבלים באמצעות שיטות שונות ערך מספרי האומד את הפרמטר, בין מציאות רוח "סביר" עבור הפרמטר, באופן שנוכל להיות כמעט בטוחים שהפרמטר אמנים כלול ברווח שהתקבל
- ❖ סטטיסטי טוב (בעל סיכון מינימאלי) מוגדר על פי שני קритריונים:
  - I. ההטיה שווה לאפס, כלומר  $E(\hat{\theta}) - \theta = 0$
  - II. הגדרה: אומד  $\hat{\theta}$  נקרא אומד חסר הטיה (אחס'ה) עבור פרמטר  $\theta$  אם קיים:  $E(\hat{\theta}) = \theta$  ו  $Var(\hat{\theta})$  מינימלית

## ❖ **משפט הגבול המרכזי**

אם נתונות תוצאות בלתי תלויות בעלות אותו חוק התפלגות  
כלשהו מאוכלוסייה בעלת תוחלת  $\mu$  ושונות<sup>2</sup>  $\sigma^2$ :

$$x_i \sim (\mu, \sigma^2)$$

$$\bar{X} \underset{n \rightarrow \infty}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

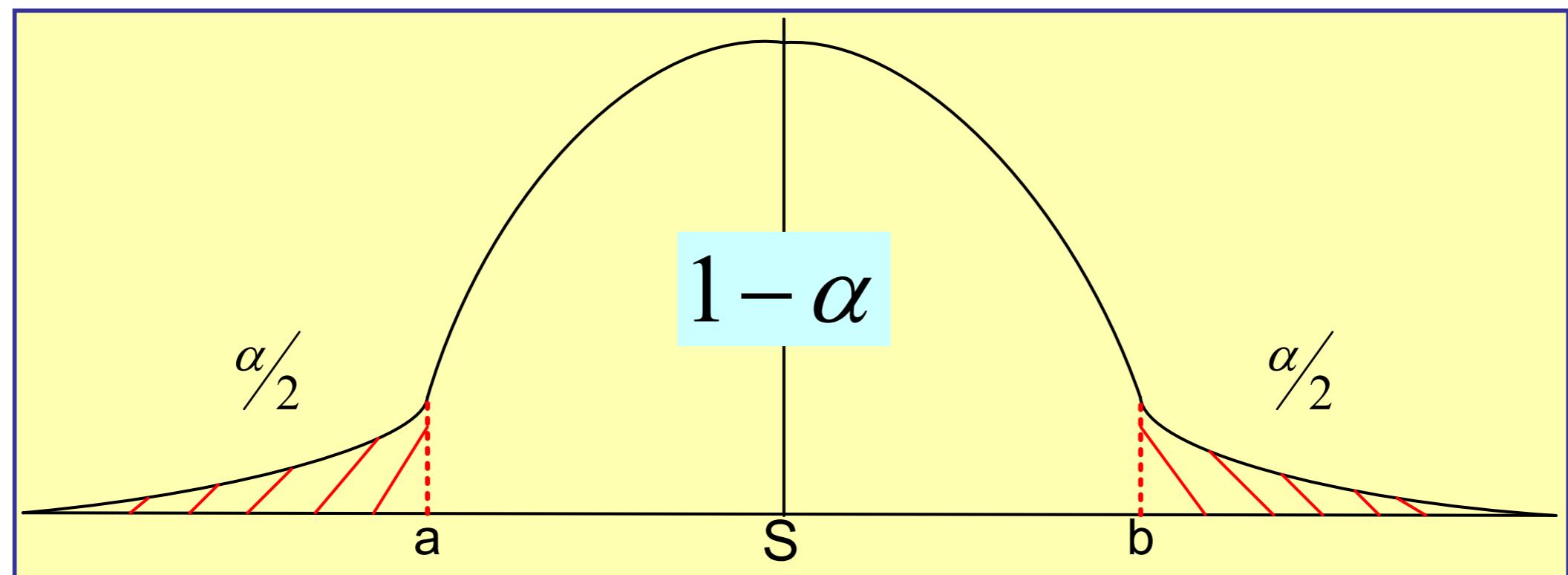
$$z_n = \frac{\bar{x}_n - \mu}{\sigma / \sqrt{n}} \underset{n \rightarrow \infty}{\sim} N(0, 1)$$

## רוח בר סמר:

- נקרא גם אומדן מרוח לפרמטר  $\theta$
- נניח שהסטטיסטי  $S$  הינו אח"ה לפרמטר  $\theta$
- יוצרים מרוח ביטחון,  $[a, b]$ , כך ש-  $a < \theta < b$
- רמת הסמר (או מקדם הביטחון) הינה ההסתברות,  $1-\alpha$ ,

רמת המובהקות

$$p(a < \theta < b) = 1 - \alpha$$



— **דוגמא:** רוח בר סמר עבר תוחלת של אוכלוסייה, כאשר השונות ידועה והמדגם גדול:

1. ממוצע המדגם הינו אמד חסר הטיה עבר התוחלת,

$$E(\bar{x}) = \mu \quad \text{כלומר:}$$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \quad \text{לפי משפט הגבול המרכזי:}$$

על כן נקבל את רוח הסמר:

$$P\left(z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha ; z_{\alpha/2} = -z_{1-\alpha/2}$$

$$P\left(\bar{x} - z_{1-\alpha/2} \cdot \sigma / \sqrt{n} \leq \mu \leq \bar{x} + z_{1-\alpha/2} \cdot \sigma / \sqrt{n}\right) = 1 - \alpha$$

• התפלגות "хи בריבוע":

— עבור סדרת משתנים מקריים  $z_1, z_2, \dots, z_n$  בלתי מתואמים,

$$\text{כך: } \forall i \quad z_i \sim N(0, 1)$$

— גדייר משתנה חדש:

$$X = \sum_{i=1}^n z_i^2 \sim \chi_n^2$$

$$X + Y \sim \chi_{n+m}^2 \quad \text{אזי:} \quad Y \sim \chi_m^2 \quad \text{ו} \quad X \sim \chi_n^2 \quad \text{אם}$$

— לדוגמה: עבור סדרת משתנים מקריים  $x_1, \dots, x_n$  בלתי תלויים, כך:  $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$  —  $\forall i \quad x_i \sim N\left(\mu, \sigma^2\right)$   
השונות המדגמית היא:

השונות המדגמית הינה  
אמד חסר הטיה ל-  $\sigma^2$

$$\tilde{s}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$E(\tilde{s}^2) = \sigma^2$$

$$Y = \frac{(n-1) \cdot \tilde{S}^2}{\sigma^2} \sim ?$$

— נגדיר את הסטטיסטי:

$$Y = \frac{(n-1) \cdot \sum_{i=1}^n (x_i - \bar{x})^2}{(n-1) \cdot \sigma^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} = \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma} \right)^2$$

$$= \sum_{i=1}^n \left( \frac{x_i - \mu}{\sigma} \right)^2 - \left( \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right)^2$$

$\chi_n^2$                        $\chi_1^2$

הוכחה בעזרת  
משפט הפיצול

על כן:

$$Y \sim \chi_{n-1}^2$$

- **התפלגות  $t$**  (נקראת גם התפלגות "סטודנט"):  
— עבור  $X \sim \chi_n^2$  ו-  $Z \sim N(0,1)$ , אזי:

$$\frac{Z}{\sqrt{\frac{X}{n}}} \sim t_n$$

- התפלגות  $t$  סימטרית סביב האפס וכאשר  $n$  שואף לאינסוף, התפלגות  $t$  שואפת לתפלגות הנורמלית
- דוגמא: נגדיר את הסטטיסטי הבא לתחלת האוכלוסייה

**כאשר השונות לא ידועה:**

$$T = \frac{\bar{X}_n - \mu}{\tilde{s} / \sqrt{n}} \sim ?$$

**נזכיר -**

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

— נמצא את התפלגות הסטטיסטי:

$$T = \frac{\left( \frac{\bar{X}_n - \mu}{\sigma} \right)}{\left( \frac{\tilde{s}}{\sqrt{n}} \right)} = \frac{\left( \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \right)}{\left( \frac{\tilde{s}}{\sigma} \right)}$$

Z ~ (0,1)

?

דוגמה למשתנה  
хи בריבוע

$$Y = \frac{(n-1) \cdot \tilde{s}^2}{\sigma^2} \sim \chi_{n-1}^2$$

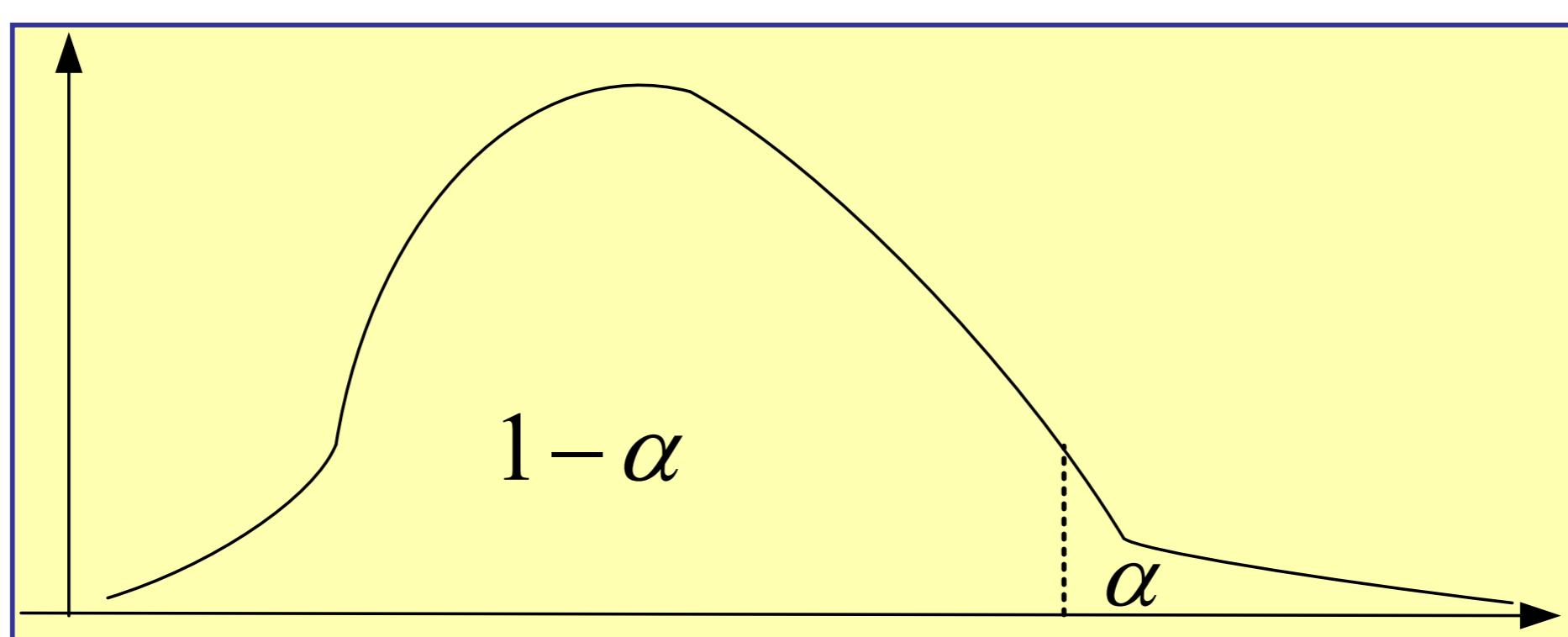


$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}} \sim t_{n-1}$$

$$\Rightarrow \frac{\tilde{s}}{\sigma} = \sqrt{\frac{(n-1) \cdot \tilde{s}^2}{\sigma^2 \cdot (n-1)}} = \sqrt{\frac{Y}{n-1}}$$

$$\frac{X/n}{Y/m} \sim F_{n,m}$$

— התפלגות F: אם  $Y \sim \chi_m^2$  ו-  $X \sim \chi_n^2$  אז:



— במהלך הקורס ניעזר בטבלת ניתוח שונות המשמשת בסטטיסטי שמתפלג F

## נושאים:

- **חזרה**
- **רגסיה ליניארית פשוטה**

- **Y – המשתנה תלוי**
- $x_n, \dots, x_1$  – המשתנים הבלתי תלויים  
(המשתנים המסבירים)
- מודל גרסיה ליניארית מגדיר קשר סטטיסטי בין המשתנה תלוי לקבוצת המשתנים הבלתי תלויים
- **לדוגמה:**
- Y – כמות יבול שבועי בחלוקת מסויימת בחממה
- $x_1$  – כמות השקיה יומית
- $x_2$  – כמות חומר דשן בחלוקת
- $x_3$  – אחוז לחות בחממה עליו ניתן לשלוט
- קשר מתמטי (פונקציונלי, דטרמיניסטי) -  $(x_1, x_2, x_3) = f \rightarrow Y$
- אולם קשר זה לא קיים למציאות כי סביר להניח שקיים גורמים נוספים שלא נלקחים בחשבון. גורמים אלו נקראים גורמי רעש. במקרה שלנו למשל טמפרטורה, טיפולים, שגיאות מדידה ועוד'

## רגסיה ליניארית פשוטה - הנחות המודל:

(1) ליניאריות - מהנחה זו נובע ש:

$$E[\epsilon_i] = 0$$

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i$$

(2) סופג את כל הרעש, לכל  $i$ ,  $V[x_i] = 0$

$$V[\epsilon_i] = \sigma^2$$

$$\epsilon_i \sim N(0, \sigma^2)$$

$$(5) \text{ הטעויות ב"ת} - Cov[\epsilon_i, \epsilon_j] = 0 \quad \text{לכל } i \neq j$$

## מציאת אומדיים ל- $\beta_1, \beta_0$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{s_{xy}}{s_{xx}}$$

$$s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

לצרכי חישוב נוח להשתמש:

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

## דוגמה

רוצים לבנות מודל גרסיה ליניארית עבור הקשר בין מספר שורות קוד בתוכנה לבין הזמן הפיתוח שלה.

X	שורות קוד	560	420	790	1100	640	1350
Y	זמן פיתוח (דק.)	185	160	210	320	190	350

- בנו את משוואת הגרסיה.
- מה ההשפעה של הוספת 10 שורות קוד על זמן הפיתוח?
- כמה זמן אTEM צופים שייקח לפתח תוכנה בהיקף של 2000 שורות קוד?

# דוגמה - פתרון

שורות קוד	560	420	790	1100	640	1350
זמן פיתוח (דק.)	185	160	210	320	190	350

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}; b_0 = \bar{y} - b_1 \bar{x}$$

$$\sum_{i=1}^n x_i = 4860$$

$$\sum_{i=1}^n y_i = 1415$$

$$\sum_{i=1}^n x_i^2 = 4556200$$

$$\sum_{i=1}^n x_i y_i = 1282800$$

$$b_1 = 0.22$$

$$b_0 = 57.2$$

$$\hat{y}_i = 57.2 + 0.22 * x_i$$

## נושאים:

- **חזרה**
- **גרסיה ליניארית פשוטה**
- **הכרות עם שפת R**

# שפת R

## ▪ התקינה

<http://cran.r-project.org/bin/windows/base/> - שפת R

<http://www.rstudio.com/ide/download/desktop> - R-Studio

## ▪ מדריכים

<http://www.cookbook-r.com/>

<http://www.statmethods.net/>

<http://quanttrader.info/public/gettingStartedWithR.html>

<http://www.cyclismo.org/tutorial/R/>

<http://cran.r-project.org/doc/manuals/R-intro.html>

# למה ? R

LinkedIn Job Search Results for "R"

**SEARCH**

Advanced >

All  
Jobs  
More...

Keywords: R

Company:

Title:

Location: Located in or near: Israel

Country: Israel

Search Reset

Location: All (19)  
+ Add

Company: All (4)  
Facebook (4)  
Microsoft (2)  
SundaySky (2)  
IBM (1)  
AOL (1)  
+ Add

19 results for R

Sort by Relevance ▾

Job Title	Company	Location	Date	Action
Account Executive: Education Customers- Intern	Microsoft	Ra'anana, Israel	Oct 20, 2014	<a href="#">View</a>
Senior Software Developer and Researcher	Outbrain	Netanya, Israel	Oct 26, 2014	<a href="#">View</a>
ACCOUNT EXECUTIVE	Microsoft	Ra'anana, Israel	Oct 13, 2014	<a href="#">View</a>
Data Analyst	AVG Technologies	Tel Aviv	Oct 22, 2014	<a href="#">View</a>
Data Scientist	AGT IoTA	Israel	Oct 21, 2014	<a href="#">View</a>
Data Scientist- for the Innovation Lab TLV	Citi	Israel	Oct 19, 2014	<a href="#">View</a>
Data Scientist	IBM	Petach Tikva - Israel	Oct 7, 2014	<a href="#">View</a>
Senior Physical Design Flow Support Engineer	Apple	Herzliya -Israel	Oct 15, 2014	<a href="#">View</a>
Data Scientist	Facebook	Tel Aviv -Israel	Oct 15, 2014	<a href="#">View</a>
Clinical Research Associate	ClinTec International	Tel Aviv	Oct 15, 2014	<a href="#">View</a>

# שפט R – דוגמה

- נתונים: השפעה של מהירות הרכב על מרחק העזירה.

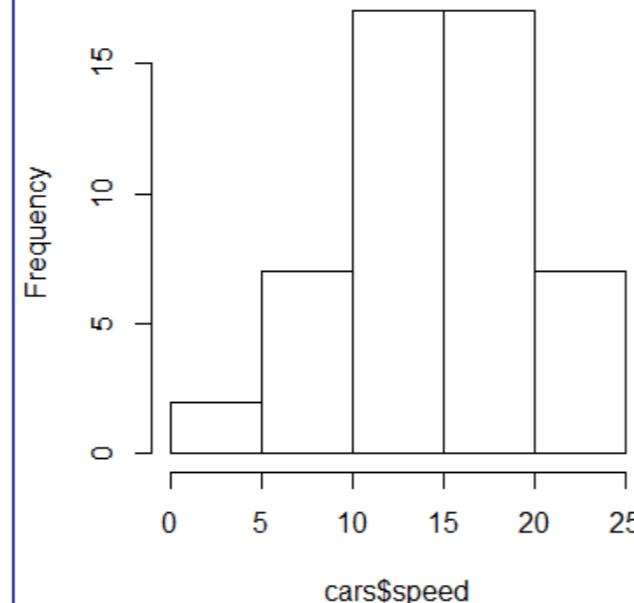
- סטטיסטיקה תיאורית:

```
summary(cars)
hist(cars$speed)
hist(cars$dist)
par(mfrow=c(1,2))
hist(cars$speed)
hist(cars$dist)
```

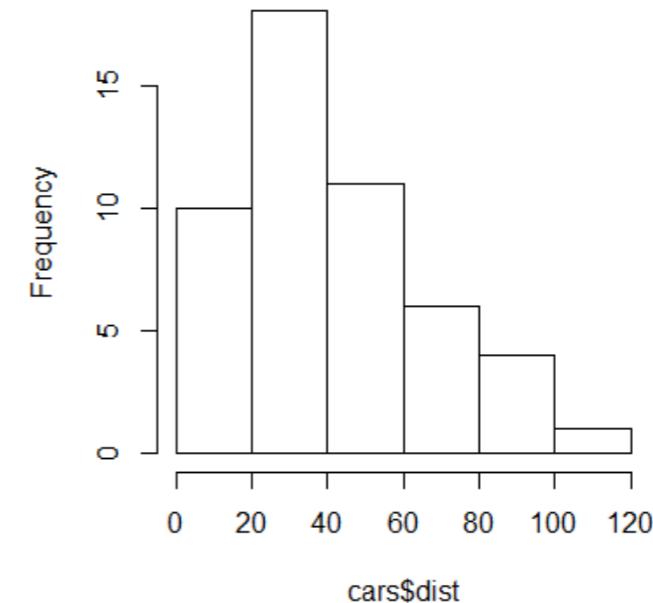
> **summary (cars)**

	speed	dist
Min. :	4.0	2.00
1st Qu.:	12.0	26.00
Median :	15.0	36.00
Mean   :	15.4	42.98
3rd Qu.:	19.0	56.00
Max.   :	25.0	120.00

Histogram of cars\$speed



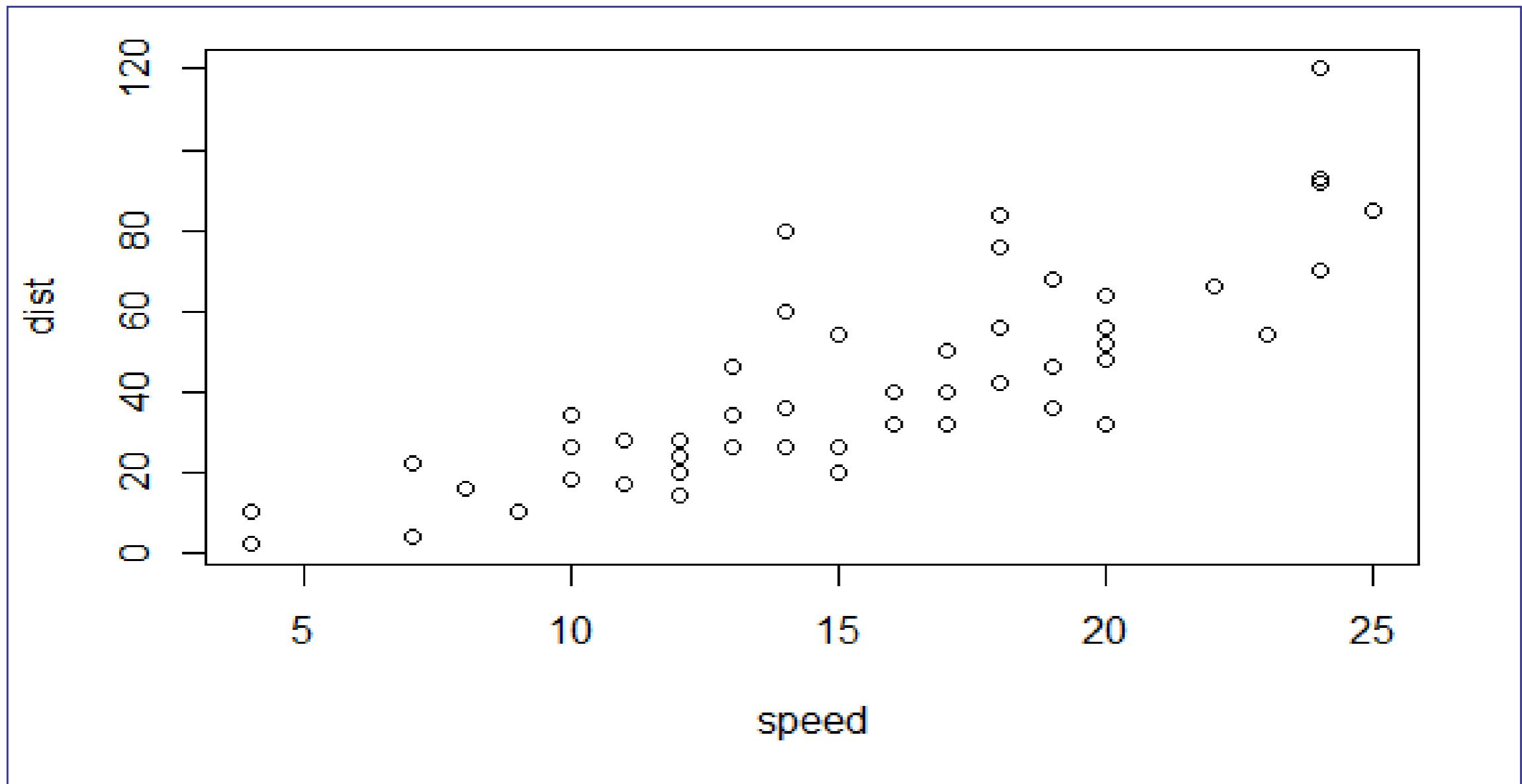
Histogram of cars\$dist



# שפת R – דוגמה

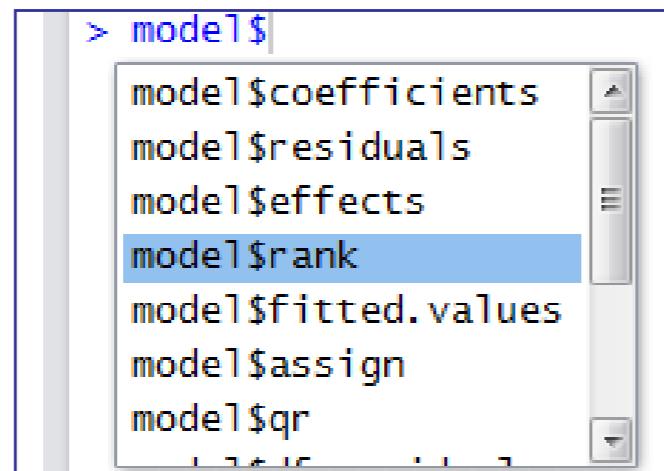
- מה הקשר בין מהירות למרחק עצירה?

```
plot (dist~speed, data=cars)
```



# גרסיה לינארית ב-R

- `lm(response_variable ~ explanatory_variable)`
- יחזיר את מקדמי הגרסיה
- `model <- lm(y ~ x)`
- אם מעוניינים בעוד נתונים, יש לשמר את התוצאה למשנה
- `summary(model)`
- פקודה שמצוירה את כל הנתונים עבור המודל
- `model$...`
- גישה לנตอน מסוים של משתנה



# גרסיה לינארית ב-R - פלט

```
Console ~/R/ ⌂
> summary ( model)

call:
lm(formula = cars$dist ~ cars$speed)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069 -9.525 -2.272  9.215 43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791   6.7584  -2.601  0.0123 *  
cars$speed    3.9324   0.4155   9.464 1.49e-12 *** 
---
signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

> |
```

**מודל**

$e_i = y_i - \hat{y}_i$

$b_0$

$b_1$