

# בדיקת הנחות המודל, אבחון המודל וטרנספורמציות.

נעזר ב Slide 7.pdf של ד"ר ניצה ברקון

# הנחות המודל

המודל

$$y_{ij} = \underbrace{\mu + \tau_i}_{\mu_i} + \tilde{\varepsilon}_{ij}$$

$$i = 1, \dots, a$$

$$j = 1, \dots, n_i$$

$$\sum_{i=1}^a \tau_i = 0$$

$$\tilde{\varepsilon}_{ij} \sim NID(0, \sigma^2)$$

מקורם של כל התצפיות הוא מהמודל (אין תצפיות חריגות)

השגיאה מתפלגת נורמאלית

השונות קבועה עבור כל השגיאות (בכל  $a$  האוכלוסיות הנבדקות).

השגיאות בלתי תלויות

תחת הנחות המודל, הניתוח הסטטיסטי אשר ביצענו הוא מדויק. אחרת ייתכן והוא שגוי (בקצת... בהרבה)?

## הנחות המודל...

- לפעמים נדע מראש שניתן להניח שכל הנחות המודל תקפות, לרוב על פי ניסיון קודם.
- לפעמים נדע מראש שהנחות המודל לא מתקיימות באופן חריף. לדוגמא: אנו יודעים שההתפלגות אשר ממנה אנו דוגמים היא אסימטרית ימנית.
- לפעמים (לרוב) לא נדע, ולכן עלינו לבחון את הנחות המודל במקביל לביצוע הניתוח הסטטיסטי....

# בחינת הנחות המודל...

האמד של תצפית  $j$ ,  $i$  על פי המודל.  $\hat{y}_{ij}$

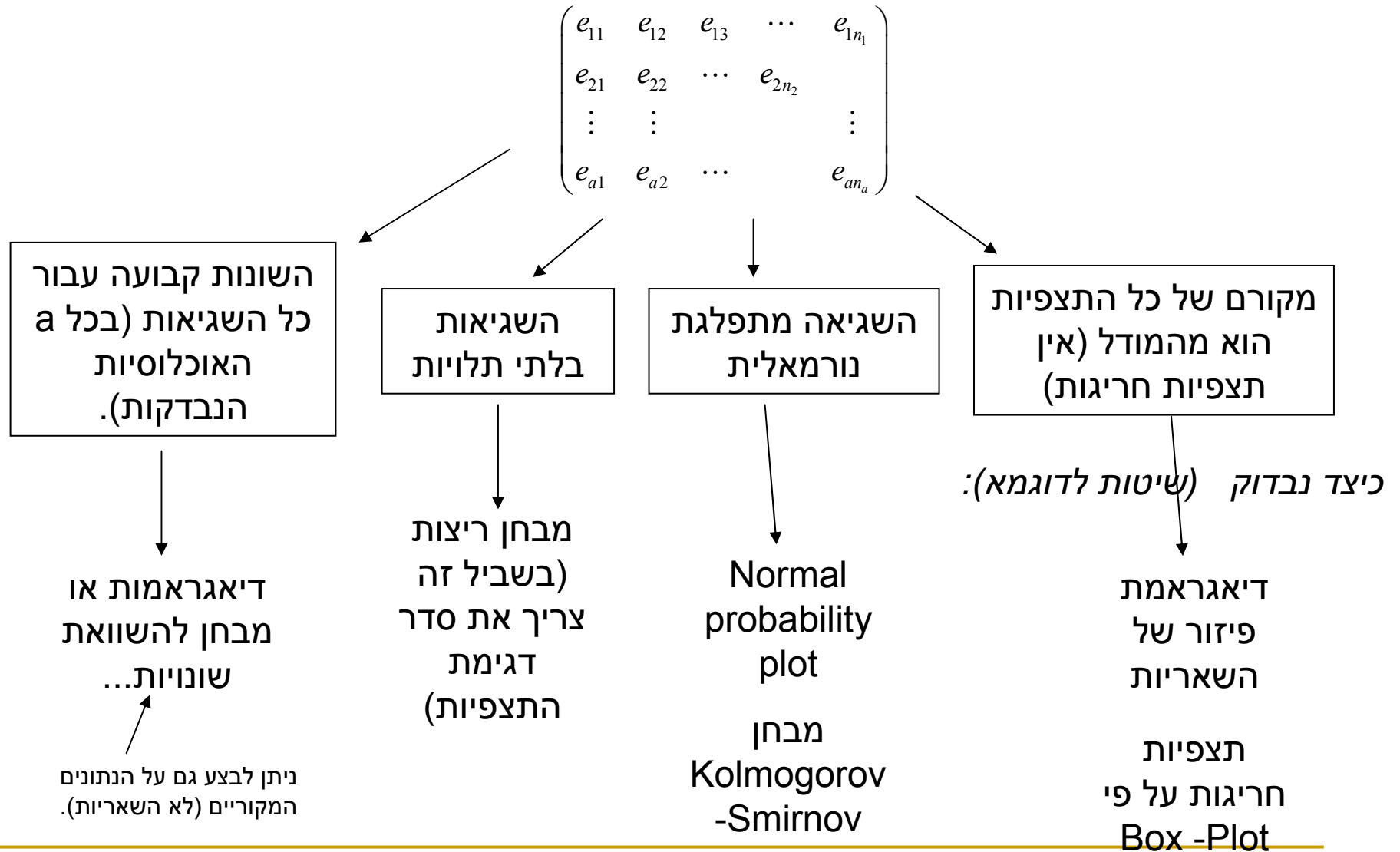
השארית של תצפית  $j$ ,  $i$  על פי המודל.  $e_{ij} = y_{ij} - \hat{y}_{ij}$

כיצד אומדים את התצפית ה  $j$ ,  $i$  על פי המודל?  $\hat{y}_{ij} = \hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) = \bar{y}_{i.}$

שאריות:

$$\begin{pmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1n_1} \\ y_{21} & y_{22} & \cdots & y_{2n_2} & \\ \vdots & \vdots & & & \vdots \\ y_{a1} & y_{a2} & \cdots & & y_{an_a} \end{pmatrix} - \begin{pmatrix} \bar{y}_{1.} & \bar{y}_{1.} & \bar{y}_{1.} & \cdots & \bar{y}_{1.} \\ \bar{y}_{2.} & \bar{y}_{2.} & \cdots & \bar{y}_{2.} & \\ \vdots & \vdots & & & \vdots \\ \bar{y}_{a.} & \bar{y}_{a.} & \cdots & & \bar{y}_{a.} \end{pmatrix} = \begin{pmatrix} e_{11} & e_{12} & e_{13} & \cdots & e_{1n_1} \\ e_{21} & e_{22} & \cdots & e_{2n_2} & \\ \vdots & \vdots & & & \vdots \\ e_{a1} & e_{a2} & \cdots & & e_{an_a} \end{pmatrix}$$

# מה ניתן ללמוד מהשאריות...



---

# זיהוי תצפיות חריגות

# תצפיות חריגות: דוגמת Time Trial

השיטה של סימון תצפיות חריגות ב Box-Plot:

$$IQ = Q3 - Q1$$

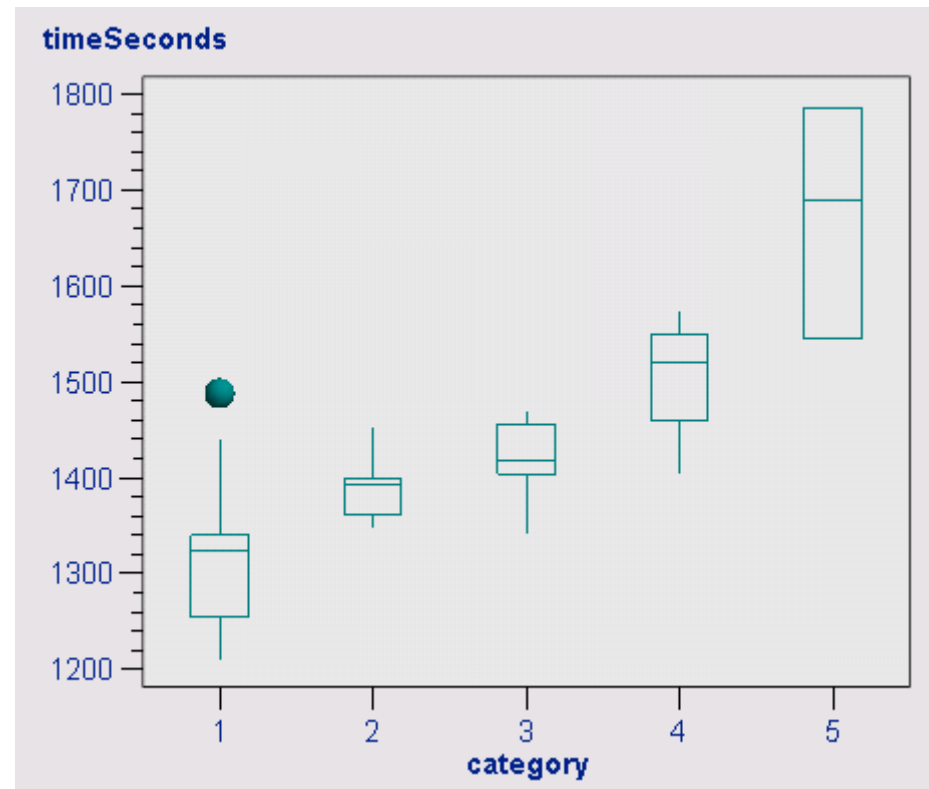
$$Low = Q1 - 1.5IQ$$

$$High = Q3 + 1.5IQ$$

$$Outlier < Low \text{ or } High < Outlier$$

תצפית חריגה

(outlier)



---

# בחינת נורמאליות של השאריות



## בחינת נורמאליות של נתונים

- יש לבחון את השאריות של כל אוכלוסיה בנפרד. מדוע?
- שימוש ב-Normal Probability Plot (pp – plot), (qq-plot).
- מבחן קולמוגורוב-סמירנוב (Kolmogorov-Smirnov) לבדיקת טיב התאמה.
- דוגמאות ניתן לראות בקובץ ההרצאות של ניצה ברקן "Slide 7.pdf".

---

# בחינת שוויון שונויות

# מבחן Bartlett לשוויון שוניות

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_a^2$$

$H_1 : otherwise$

$$S_1^2, \dots, S_a^2 \quad S_p^2 = \frac{\sum_{i=1}^a (n_i - 1) S_i^2}{N - a}$$

$$S_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2}{n_i - 1}$$

סטטיסטי המבחן

$$q = (N - a) \log_{10} S_p^2 - \sum_{i=1}^a (n_i - 1) \log_{10} S_i^2$$

q קטן

Si דומים

q גדול

Si שונים

$$\chi_0^2 = 2.3026 \frac{q}{c}$$

$$c = 1 + \frac{1}{3(a-1)} \left( \sum_{i=1}^a (n_i - 1)^{-1} - (N - a)^{-1} \right)$$

Bartlett's Test for Homogeneity of timeSeconds Variance			
Source	DF	Chi-Square	Pr > ChiSq
category	4	9.1406	0.0577

$$\chi_0^2 > \chi_{\alpha, a-1}^2 \quad \text{דחה } H_0 \text{ אם:}$$

## רגישות מבחן Bartlett ומבחנים נוספים.

- באופן כללי (ללא הוכחה), הסקה לגבי שונויות רגישה הרבה יותר להנחות הנורמאליות.
- כנ"ל לגבי הסקה לגבי שונות של אוכלוסיה רגילה המסתמכת על התפלגות חי-בריבוע.
- מבחן נוסף הוא מבחן Levine. למרות שעוצמת מבחן זה היא באופן כללי יותר נמוכה מעוצמת מבחן Bartlett, מבחן זה פחות רגיש להנחת הנורמאליות.

Levene's Test for Homogeneity of timeSeconds Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
category	4	3.164E8	79100523	2.56	0.0597
Error	29	8.9616E8	30902130		

---

# בדיקת אי-תלות של השגיאות

# בדיקת אי-תלות של השגיאות

- מקורות תלות בין שגיאות:
  - דגימה לא מקרית.
  - תהליך הניסוי משתנה לאורך זמן (לאורך הדגימה).
- הדרך הטיפוסית לבדוק היא לבצע מבחן ריצות (לא נבצע באופן מפורש).

---

# טיפול בנתונים אשר לא מתאימים למודל

## טרנספורמציות

- המוטיבציה בביצוע טרנספורמציות היא כפולה:
  - להפוך את הנתונים לנורמאליים.
  - לייצב את השונות.
- הרבה פעמים אותה טרנספורמציה מטפלת בשתי הבעיות במקביל.
- לפעמים ניתן לבחור טרנספורמציה על פי הנתונים ולפעמים ניתן להשתמש בידע נוסף (מוקדם).



## דוגמא:

■ הרבה פעמים נתונים כלכליים הקשורים לצמיחה/דעיכה דורשים טרנספורמצית Log.

■ ניקח לדוגמא ערך בסיס (Base) אשר עובר הרבה שינויים (m שינויים), כל פעם בשיעור  $(1 + r_k)$ .

■ אז הערך העכשווי הוא CurrentValue:

$$Base \cdot (1 + r_1) \cdot (1 + r_2) \cdot \dots \cdot (1 + r_m) = CurrentValue$$

■ אם כך, במידה ושיעור השינויים הוא i.i.d אז על פי משפט הגבול המרכזי בקרוב מתקיים:

$$\log(CurrentValue) \sim Normal$$

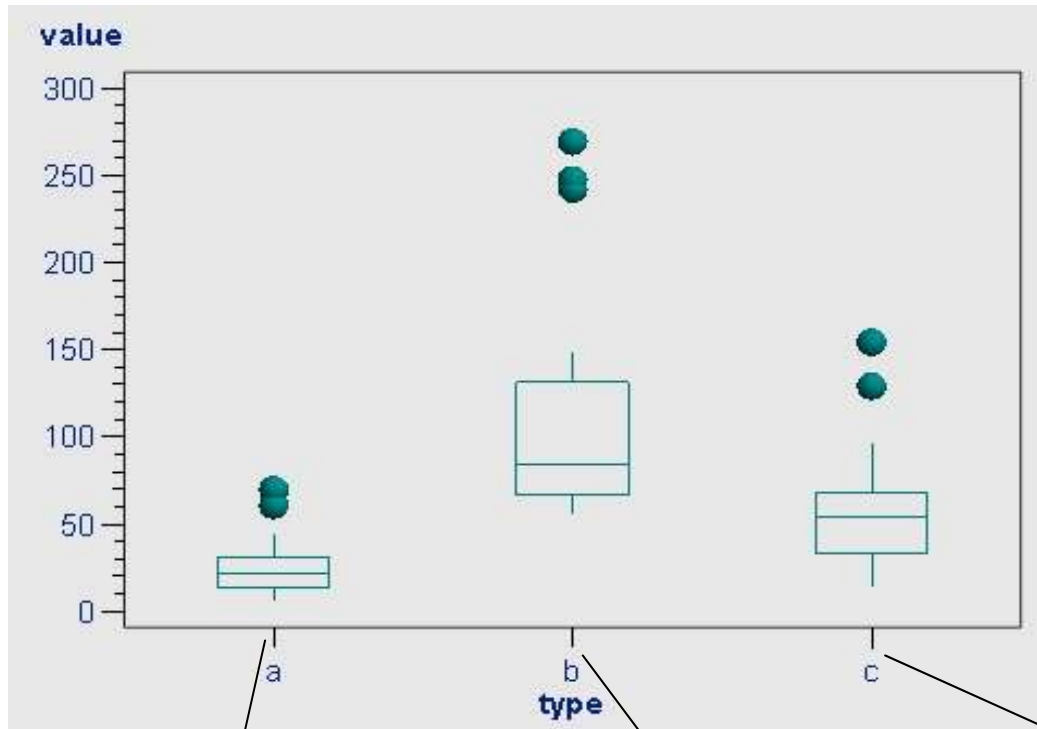
■ כי זהו סכום של הרבה משתנים מקריים i.i.d

$$\log(Base) + \log(1 + r_1) + \log(1 + r_2) + \dots + \log(1 + r_m) = \log(CurrentValue)$$

## המשך דוגמא:

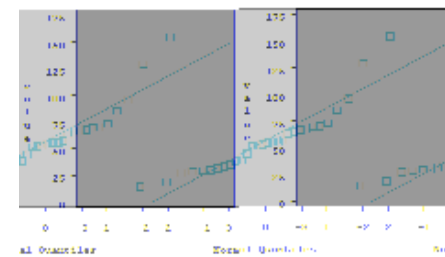
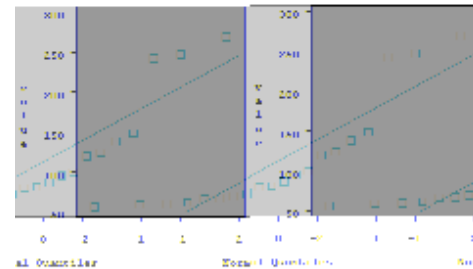
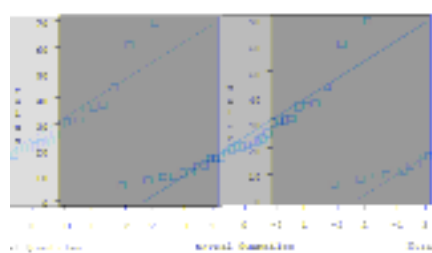
- אנו מעוניינים להשוות את שיעור הצמיחה (צמצום) של חברות משלושה מגזרים:
  - a – תעשייה כבדה.
  - b – הי-טק.
  - c – תרופות.
- נלקח מדגם של חברות גדולות מכל מגזר ונמדד השינוי בהיקף המכירות של החברה באחוזים במהלך השנים 1990 עד 2000 במשק.

# המשך דוגמא, הנתונים.

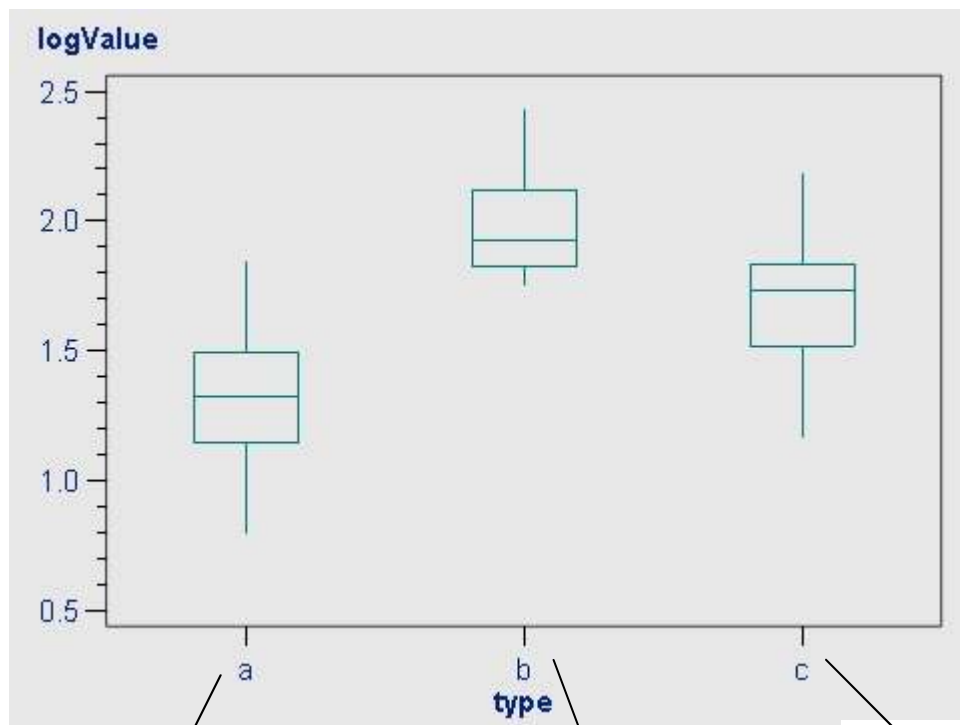


Levene's Test for Homogeneity of value Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
type	2	2.0627E8	1.0314E8	7.66	0.0009
Error	77	1.037E9	13467060		

Bartlett's Test for Homogeneity of value Variance			
Source	DF	Chi-Square	Pr > ChiSq
type	2	49.5367	<.0001

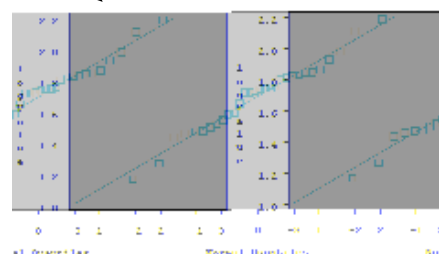
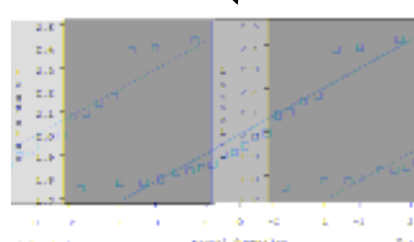
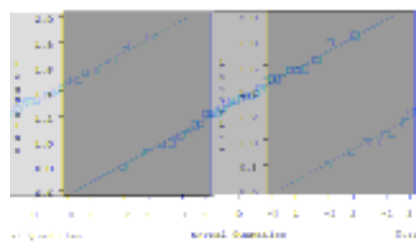


# נבצע טרנספורמציה $\log$ על הנתונים



Levene's Test for Homogeneity of logValue Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
type	2	0.00379	0.00190	0.37	0.6918
Error	77	0.3945	0.00512		

Bartlett's Test for Homogeneity of logValue Variance			
Source	DF	Chi-Square	Pr > ChiSq
type	2	0.5835	0.7470



# ביצוע ניתוח שונות

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.59441496	2.79720748	50.82	<.0001
Error	77	4.23828762	0.05504270		
Corrected Total	79	9.83270258			

R-Square	Coeff Var	Root MSE	logValue Mean
0.568960	14.39618	0.234612	1.629680

Source	DF	Anova SS	Mean Square	F Value	Pr > F
type	2	5.59441496	2.79720748	50.82	<.0001

# מבחני Bonferroni כ Post-Hoc

הנתונים לאחר הטרנספורמציה

Alpha	0.001
Error Degrees of Freedom	77
Error Mean Square	0.055043
Critical Value of t	3.75627

Comparisons significant at the 0.001 level are indicated by ***.				
type Comparison	Difference Between Means	Simultaneous 99.9% Confidence Limits		
b - c	0.29798	0.04358	0.55237	***
b - a	0.67026	0.41586	0.92466	***
c - b	-0.29798	-0.55237	-0.04358	***
c - a	0.37228	0.14474	0.59982	***
a - b	-0.67026	-0.92466	-0.41586	***
a - c	-0.37228	-0.59982	-0.14474	***

הנתונים המקוריים

Alpha	0.001
Error Degrees of Freedom	77
Error Mean Square	1523.029
Critical Value of t	3.75627

Comparisons significant at the 0.001 level are indicated by ***.				
type Comparison	Difference Between Means	Simultaneous 99.9% Confidence Limits		
b - c	55.81	13.49	98.13	***
b - a	87.55	45.23	129.87	***
c - b	-55.81	-98.13	-13.49	***
c - a	31.74	-6.11	69.59	
a - b	-87.55	-129.87	-45.23	***
a - c	-31.74	-69.59	6.11	