תכנון ניסויים

וניתוח שונות

ניתוח שונות

השוואת קבוצות רבות – ניתוח שונות

מחקרים רבים עורכים השוואות בין 3 קבוצות או יותר. בדרך כלל יש עניין לערוך בדיקה כללית – האם בכלל יש עדות לכך שהקבוצות שונות – וגם השוואות פרטניות בין זוגות של קבוצות.

כאן נציג ניתוח שונות, שיטה המכלילה את מבחן t כדי לבדוק האם יש בכלל הבדלים בין הקבוצות. בלועזית, Analysis of Variance או פשוט ANOVA.

תחילה, נסתכל על נתונים ממחקר שבדק את היעילות של 3 סוגים של אנטיביוטיקה ושל טיפול פלצבו על חיידק שהודבק לחיות מעבדה.

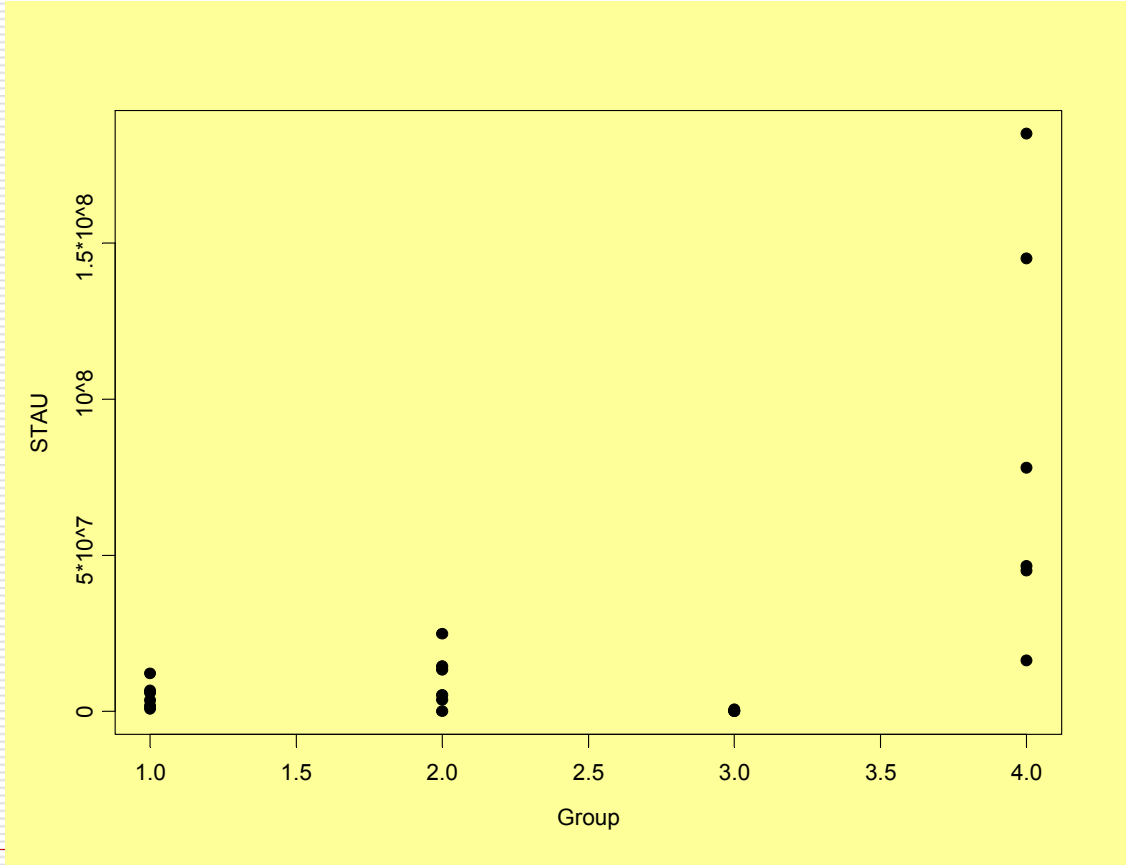
לכן יש $r = 4$ קבוצות בניסוי זה.

שש חיות שובצו באקראי לכל טיפול. הנתונים הם ספירות של חיידקים לאחר תקופת הטיפול.

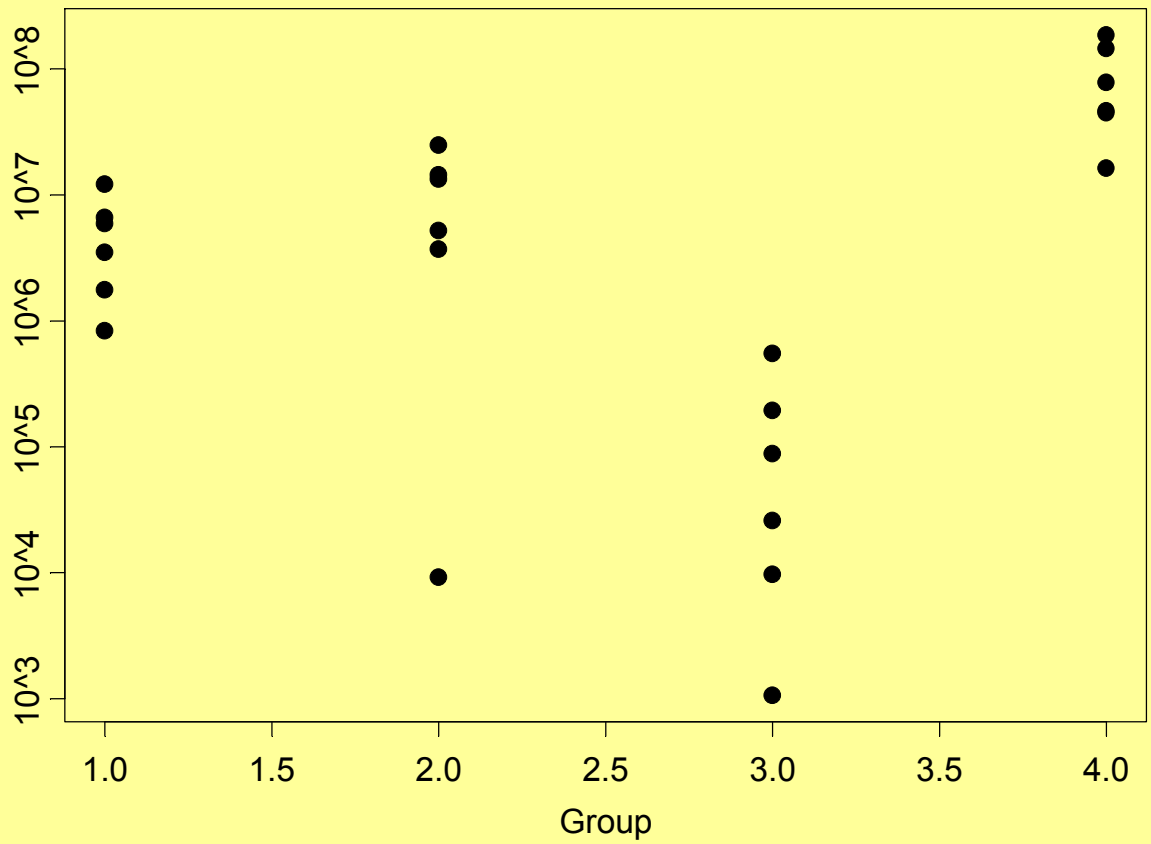
להלן הספירה של אחד החיידקים, ביחידות של 10^6.

	Drug 1	Drug 2	Drug 3	Placebo
	3.5	14.4	0.194	145
	5.9	24.8	0.00106	16.25
	1.76	0.0092	0.55	46.5
	6.6	13.3	0.0259	78
	12.1	5.2	0.0097	185
	0.83	3.7	0.088	45
Average	5.12	10.23	0.14	85.96
SD	4.09	9.07	0.21	65.50

ניתן לראות ששלושת הטיפולים עדיפים על הפלצבו (מספר 4). נראה שפיזור הנתונים עולה ביחד עם הספירה עצמה.



מעבר לסקלה לוגריתמית מקטינה באופן משמעותי את השוני של הפיזור על פני הקבוצות.



להלן טבלת הנתונים לאחר מעבר ללוג בסיס 10.

	Drug 1	Drug 2	Drug 3	Placebo
	0.544068	1.158362	-0.7122	2.161368
	0.770852	1.394452	-2.97469	1.210853
	0.245513	-2.03621	-0.25964	1.667453
	0.819544	1.123852	-1.5867	1.892095
	1.082785	0.716003	-2.01323	2.267172
	-0.08092	0.568202	-1.05552	1.653213
Average	0.56	0.49	-1.43	1.81
SD	0.42	1.27	0.98	0.39

ניתוח שונות יאפשר לנו לבדוק האם ההבדלים בין התרופות אמיתיות או, אולי, תוצאה רק של הפיזור המקרי של הנתונים.

$Y_{i,j}$ מסמן את התצפית ה- j בקבוצה i. $i=1,\ldots,r,\ \ j=1,\ldots,n_i$

$$Y_{i,j} \sim N(\mu_i, \sigma^2)$$

המודל מבטא את ההבדל בין הקבוצות על פי התוחלות שלהן – לכל קבוצה תוחלת משלה.

לעומת זאת, המודל מניח שבכל קבוצה אותה סטיית תקן.

הנחות אלו נראות סבירות לאחר המעבר ללוגים, למעט, אולי, תצפית אחת חריגה בתרופה 2.

דרך חלופית לכתוב את המודל מדגיש את ההבדל, בכל קבוצה, בין התוחלת שלה לבין "התוחלת הממוצעת"

$$\mu = (1/r)\sum_{i=1}^{r}\mu_i$$

בהצגה זו של המודל, נגדיר $\alpha_i = \mu_i - \mu$

למעשה α_i משקף את התרומה לתוחלת (מעבר לממוצע) מהשתייכות בקבוצה i.

בהצגה זו של המודל $Y_{i,j} \sim N(\mu + \alpha_i, \sigma^2)$

ועוד דרך חלופית לכתוב את המודל מדגיש את ההבדל בין החלק המבני בו (התוחלות השונות מקבוצה לקבוצה) לבין החלק הסטוכסטי (הפיזור של התצפית סביב התוחלת).

כתיב זה מציג את המודל לניתוח שונות בצורה הבאה:

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$

כאשר

$$\varepsilon_{i,j} \sim N(0, \sigma^2)$$

וכאשר מתקיים האילוץ

$$\sum_{i=1}^{r} \alpha_i = 0$$

ניתוח השונות מתמקד בבדיקת השערת האפס:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$

כנגד האלטרנטיבה H_1 שלא כל התוחלות שוות.

לפי ההצגה החלופית של המודל, הצורה של השערת האפס היא

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

נסמן ב- $\bar{Y}_i$ את הממוצע בקבוצה i וב- $\bar{Y}$ את הממוצע של כל התצפיות.

אם נרצה לחשב את השונות של Y בכל המדגם, תחילה נחשב את סה"כ סכום הריבועים

$$SST = \sum_{i=1}^{r} \sum_{j=1}^{n(i)} (Y_{i,j} - \bar{Y})^2$$

SST הוא קיצור ל- Sum of Squares Total.

נחשב גם סכומי ריבועים בכל קבוצה בנפרד:

$$SS_i = \sum_{j=1}^{n(i)} (Y_{i,j} - \overline{Y}_i)^2$$

הנחה סטנדרטית של ניתוח שונות היא שבכל קבוצה אותה שונות.
לכן יש הגיון לחבר ביחד את סכומי הריבועים מכל הקבוצות.
מקבלים ביטוי הנקרא "סכום הריבועים בתוך הקבוצות" והמסומן SSW כקיצור ל- Sum of Squares Within.

$$SSW = \sum_{i=1}^{r} SS_i$$

עם מעט אלגברה, ניתן לראות שקיים הקשר הבא בין SST לבין SSW:

$$SST = SSW + \sum_{i=1}^{r} n_i (\bar{Y}_i - \bar{Y})^2$$

הביטוי האחרון בצד ימין יהיה גדול עם הממוצעים של הקבוצות שונים זה מזה. הוא נקרא "סכום הריבועים בין הקבוצות" ומסומן SSB כקיצור ל- Sum of Squares Between.

ניתוח בנתונים:

אם השערת האפס נכונה, לכל הקבוצות בדיוק אותה תוחלת.

תחת הנחות אלה, מסתבר שכל המנות הבאות הן אומדים חסרי הטיה של σ^2.

$$\frac{SST}{n-1}, \quad \frac{SSW}{n-r}, \quad \frac{SSB}{r-1}$$

כאן $\sum_i n_i = n$ מסמן את סה"כ מספר התצפיות.

המנה $SSW/(n-r)$ נשארת אומד חסר הטיה ל- σ^2 גם אם יש תוחלות שונות בקבוצות.

אבל המנה $SSB/(r-1)$ הופכת להיות גדולה מדי אם יש תוחלות שונות. לכן, נוכל לבדוק את ההשערה של תוחלות שוות תוך השואה של שתי המנות האלה.

סטטיסטי המבחן הוא

$$F = \frac{SSB/(r-1)}{SSW/(n-r)} = \frac{MSB}{MSW}$$

תחת השערת האפס של תוחלות שוות, לסטטיסטי F יש התפלגות הנקראת F (על שם Fisher) עם $r-1$ ו- $n-r$ דרגות חופש.

ככל שיש הבדלים גדולים יותר בין התוחלות, הסטטיסטי F יהיה גדול יותר. לכן, חישוב של p-value בניתוח שונות תמיד יהיה ההסתברות לקבל סטטיסטי F גדול או שווה מזה שהתקבל מן הנתונים שלנו.

הערה: אם נבצע ניתוח שונות למערך נתונים עם 2 קבוצות בלבד, נקבל p-value זהה לזה המתקבל ממבחן t דו-צדדי תחת ההנחה של סטיות תקן שוות.

כל הביטויים שהרכבנו כאן מסוכמים בטבלה הנקראת "טבלת ניתוח שונות". להלן הטבלה לניסוי על הטיפולים לחיידקים בעיניים.

הטבלה מראה שיש עדות חזקה נגד השערת האפס, הנדחית עם p-value הקטן מ- 0.0001.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
factor(Group)	3	32.24158	10.74719	14.80391	0.00002610836
Residuals	20	14.51940	0.72597		

ב- R, הפקודה aov מחשבת את סכומי הריבועים, דרגות החופש, הסטטיסטי F, וה- p-value התאים.

```
fit1 <- aov(log10(STAU) ~ factor(Group))
summary(fit1)
```

```
                Df Sum of Sq  Mean Sq  F Value          Pr(F)
factor(Group)    3  32.24158 10.74719 14.80391 0.00002610836
    Residuals 20  14.51940  0.72597
```

ניתן לקבל מספר גרפים שימושיים לצורך בדיקת הנחות המודל מן הפקודה

plot(fit1)

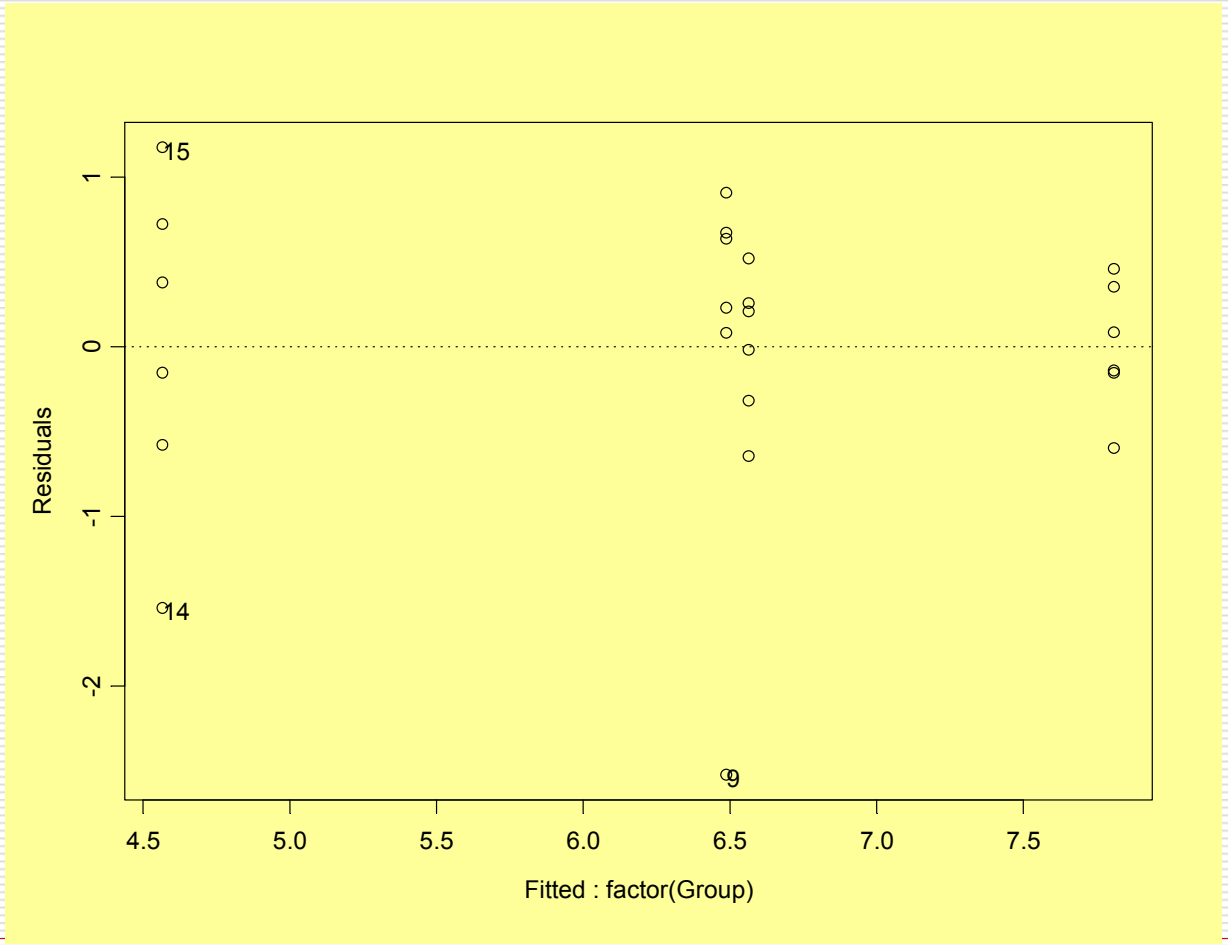
התרשימים מתבססים על השאריות

$$e_{i,j} = Y_{i,j} - \hat{Y}_{i,j} = Y_{i,j} - \bar{Y}_i$$

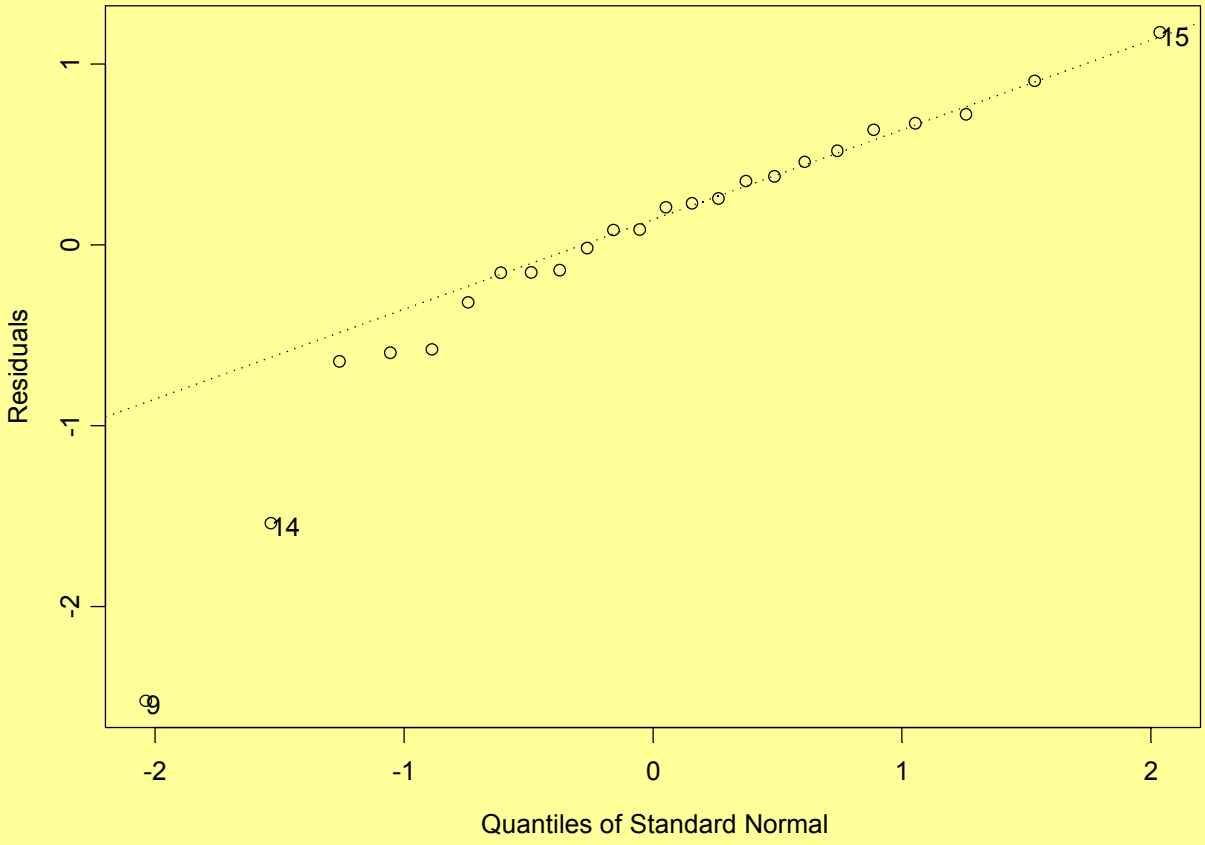
בין היתר:

- שאריות מול ערכים מנובאים.
- תרשים הסתברות נורמלית של השאריות.

שאריות מול ערכים מנובאים



תרשים הסתברות נורמלית של השאריות



רווחי סמך לכל ההשוואות הזוגיות, עם התחשבות להשוואות מרובות לפי שיטת Bonferroni.

ב- R, להפעיל

multicomp(fit1,method="bonferroni")

להלן טבלת הנתונים לאחר מעבר ללוג בסיס 10.

	Drug 1	Drug 2	Drug 3	Placebo
	0.544068	1.158362	-0.7122	2.161368
	0.770852	1.394452	-2.97469	1.210853
	0.245513	-2.03621	-0.25964	1.667453
	0.819544	1.123852	-1.5867	1.892095
	1.082785	0.716003	-2.01323	2.267172
	-0.08092	0.568202	-1.05552	1.653213
Average	0.56	0.49	-1.43	1.81
SD	0.42	1.27	0.98	0.39

	Estimate	Std.Error	Lower Bound	Upper Bound
1-2	0.0762	0.492	-1.360	1.520
1-3	2.0000	0.492	0.557	3.440 ****
1-4	-1.2500	0.492	-2.680	0.195
2-3	1.9200	0.492	0.481	3.360 ****
2-4	-1.3200	0.492	-2.760	0.119
3-4	-3.2400	0.492	-4.680	-1.800 ****

Machine

1	2	3	4	5
.125	.118	.123	.126	.118
.127	.122	.125	.128	.129
.125	.120	.125	.126	.127
.126	.124	.124	.127	.120
.128	.119	.126	.129	.121

Remember, each observation is the sum of a common value, a level effect and a residual value. Value-splitting just breaks each observation into its component parts. The first step in value-splitting is to calculate the mean values (rounding to the nearest thousandth) within each machine to get the level means.

Machine	1	2	3	4	5
	.1262	.1206	.1246	.1272	.123

We can then *sweep* (subtract the level mean from each associated data value) the means through the original data table to get the residuals

1	2	3	4	5
-.0012	-.0026	-.0016	-.0012	-.005
.0008	.0014	.0004	.0008	.006
-.0012	-.0006	.0004	-.0012	.004
-.0002	.0034	-.0006	-.0002	-.003
.0018	-.0016	.0014	.0018	-.002

The next step is to calculate the grand mean from the individual machine means as:

Grand Mean

.12432

Finally, we can sweep the grand mean through the individual level means to obtain the level effects:

1	2	3	4	5
.00188	-.00372	.00028	.00288	-.00132

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-value
Factor levels	000137.	4	000034.	< 4.86 2.87
residuals	000132.	20	000007.	
corrected total	000269.	24		

By dividing the Factor-level mean square by the residual mean square, we obtain a F-value of 4.86 which is greater than the cut-off value of 2.87 for the F-distribution at 4 and 20 degrees of freedom and 95% confidence. Therefore, there is sufficient evidence to reject the hypothesis that the levels are all the same

Test	By dividing the Factor-level mean square by the residual mean square, we obtain a F-value of 4.86 which is greater than the cut-off value of 2.87 for the F-distribution at 4 and 20 degrees of freedom and 95% confidence. Therefore, there is sufficient evidence to reject the hypothesis that the .levels are all the same
Conclusion	From the analysis of these data we can conclude that the factor "machine" has an effect. There is a statistically significant difference in the pin diameters across the machines on which they were .manufactured

TABLE 6.6. Arithmetic breakup of data given in Table 6.1

	observations y_{ti}		grand average $\bar{y}$		treatment deviations $\bar{y}_t - \bar{y}$		residuals (within-treatment deviations) $y_{ti} - \bar{y}_t$
	$\begin{bmatrix} 62 & 63 & 68 & 56 \\ 60 & 67 & 66 & 62 \\ 63 & 71 & 71 & 60 \\ 59 & 64 & 67 & 61 \\ & 65 & 68 & 63 \\ & 66 & 68 & 64 \\ & & & 63 \\ & & & 59 \end{bmatrix}$	=	$\begin{bmatrix} 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ & 64 & 64 & 64 \\ & 64 & 64 & 64 \\ & & & 64 \\ & & & 64 \end{bmatrix}$	+	$\begin{bmatrix} -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ & 2 & 4 & -3 \\ & 2 & 4 & -3 \\ & & & -3 \\ & & & -3 \end{bmatrix}$	+	$\begin{bmatrix} 1 & -3 & 0 & -5 \\ -1 & 1 & -2 & 1 \\ 2 & 5 & 3 & -1 \\ -2 & -2 & -1 & 0 \\ & -1 & 0 & 2 \\ & 0 & 0 & 3 \\ & & & 2 \\ & & & -2 \end{bmatrix}$
vector*	Y	=	A	+	T	+	R
sum of squares**	98,644	=	98,304	+	228	+	112
degrees of freedom†	24	=	1	+	3	+	20

* In standard notation each vector would be represented by a column of twenty-four elements. A rearrangement of these elements is used above to emphasize the structure of the design.

** Squared length of vector.

† Number of dimensions in which the vector is constrained to lie.