مقایسه چند گروه (تحلیل واریانس)

ANOVA , Analysis of variance

- طرح تک عاملی متعادل

فرض کنید $k$ جامعه وجود داشته باشد و از هر جامعه نمونه تصادفی $n$ تایی انتخاب کنیم
(گروه یا تیمار)

| 1 | 2 | .... | i | .... | k | |
|---|---|---|---|---|---|---|
| $Y_{11}$ | $Y_{21}$ | ... | $Y_{i1}$ | ... | $Y_{k1}$ | $i=1\ldots k$ |
| ⋮ | | | | | ⋮ | $j=1\ldots n$ |
| $Y_{1n}$ | $Y_{2n}$ | ... | $Y_{in}$ | ... | $Y_{kn}$ | |
| $T_{1\circ}$ | $T_{2\circ}$ | ... | $T_{i\circ}$ | .... | $T_{k\circ}$ | $T_{\circ\circ}$ |
| $\bar{Y}_{1\circ}$ | $\bar{Y}_{2\circ}$ | ... | $\bar{Y}_{i\circ}$ | .... | $\bar{Y}_{k\circ}$ | $\bar{Y}_{\circ\circ}$ |

مجموع ... مجموع کل

میانگین ... میانگین کل

$\alpha_i$ را تعریف می کنیم اثر $i$ امین جامعه ($i$ امین تیمار (گروه))

$Y_{ij}$ صفت مورد بررسی

مدل $$Y_{ij} = \underbrace{\mu + \alpha_i}_{\mu_i} + \varepsilon_{ij}$$

$i=1,\ldots k$ , $j=1\ldots n$

فرض $\sum_{i=1}^{k}\alpha_i = 0$

$\varepsilon_{ij} \sim N(0,\sigma^2)$ و خطاها مستقلند

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: \mu_i \neq \mu_j \quad i \neq j \text{ برای حداقل یک} \end{cases}$$

معادل است با

$$\equiv \begin{cases} H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \\ H_1: \alpha_i \neq 0 \quad i \text{ برای حداقل یک} \end{cases}$$

سه تا مجموع بدست می آورند :

مجموع کل توان های دوم $$SST=\sum_{i=1}^{K}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{..})^2=\sum\sum Y_{ij}^2-\frac{T_{..}^2}{nK}$$

مجموع توان دوم تیمارها $$SSA=n\sum_{i=1}^{K}(\bar{Y}_{i.}-\bar{Y}_{..})^2=\frac{1}{n}\sum T_{i.}^2-\frac{T_{..}^2}{nK}$$

مجموع توان دوم خطاها $$SSE=\sum_{i=1}^{K}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i.})^2$$

به طوری که $$SST=SSA+SSE$$

برآورد واریانس خطا $\hat{\sigma}^2=\frac{SSE}{K(n-1)}$ و ثابت می شود.

جدول ANOVA به صورت زیر تشکیل می شود

| منابع تغییر | SS | df | MS | F |
|---|---|---|---|---|
| تیمارها | SSA | $K-1$ | $MSA=\frac{SSA}{K-1}$ | $F=\frac{MSA}{MSE}$ |
| خطاها | SSE | $K(n-1)$ | $MSE=\frac{SSE}{K(n-1)}$ | |
| کل | SST | $nK-1$ | | |

$$\text{if } F>F_{K-1,\,K(n-1),\,1-\alpha}\Rightarrow RH_0$$

مثال ۲۷: جدول نتایج بدست آمده از یک تحقیق درباره ی پایداری یک معرف فلئورسنت نگهداری شده در شرایط متفاوت را نشان می دهد. مقادیر داده شده، علائم فلئورسانس (در واحد دلخواه) از محلول های رقیق با غلظت های برابر می باشند. سه اندازه گیری تکراری روی هر نمونه انجام گرفته است

| شرایط | اندازه گیری های تکرار شده | مجموع $T_{i0}$ | میانگین |
|---|---|---|---|
| A تازه تهیه شده | 102 ، 100 ، 101 | 303 | $\bar{Y}_{10} = 101$ |
| B نگهداری شده برای یک ساعت در تاریکی | 101 ، 101 ، 104 | 306 | $\bar{Y}_{20} = 102$ |
| C نگهداری شده برای یک ساعت در نور کاهش یافته | 97 ، 95 ، 99 | 291 | $\bar{Y}_{30} = 97$ |
| D نگهداری شده برای یک ساعت در نور شدید | 90 ، 92 ، 94 | 279 | $\bar{Y}_{40} = 92$ |

$\bar{Y}_{00} = 98$ $\quad T_{00} = 1176$

آیا شرایط نگهداری روی اندازه گیری ها تاثیر داشته است

گام اول

$$\mu_1 = \mu_2 = \mu_3 \equiv \alpha_1 = \alpha_2 = \alpha_3 = 0$$

$$SST = (101-98)^2 + \cdots + (90-98)^2 = 210$$

یا $101^2 + 100^2 + \cdots + 90^2 - \frac{1176^2}{4\times 3} = 210$

$$SSA = 3\left[(101-98)^2 + (102-98)^2 + (97-98)^2 + (92-98)^2\right] = 24$$

یا $\frac{1}{3}(303^2 + 306^2 + 291^2 + 279^2) - \frac{1176^2}{4\times 3} = 24$

$$SSE = SST - SSA = 210 - 24 = 186$$

$n = 3 \quad k = 4$

پس جدول آنالیز واریانس:

| منابع تغییر | SS | df | MS | |
|---|---|---|---|---|
| تیمارها (بین نمونه ای) | 186 | 3 | 62 | $F = \frac{62}{3} = 20.7$ |
| خطا (درون نمونه ای) | 24 | 8 | 3 | |
| کل | 210 | 11 | | |

$$F_{3,8,0.95} = 4.066$$

$20.7 > 4.066 \Rightarrow RH.$ بنابراین شرایط روی میانگین پایداری تاثیر گذارند.

- طرح یک عاملی نامتعادل

مثال ۲۸.

در این طرح تعداد اندازه‌گیری‌ها برابر نیست مثلاً در مثال قبل

| میانگین | جمع | | | | |
|---|---|---|---|---|---|
| $\bar{Y}_{1\cdot} = 100.5$ | 201 | | 100 | 101 | A $n_1=2$ |
| $\bar{Y}_{2\cdot} = 102$ | 306 | 101 | 101 | 104 | B $n_2=3$ |
| $\bar{Y}_{3\cdot} = 97$ | 194 | | 95 | 99 | C $n_3=2$ |
| $\bar{Y}_{4\cdot} = 92$ | 276 | 90 | 92 | 94 | D $n_4=3$ |
| | $\bar{Y}_{\cdot\cdot}=97.7$ | | | | $N=10$ |

$T_{\cdot\cdot}=977$ $\quad N=2+3+2+3=10$ $\quad$ در این صورت

$$SST=\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2=(101-100.5)^2+\dots+(90-100.5)^2=192.1$$

$$SST=\sum\sum Y_{ij}^2-\frac{T_{\cdot\cdot}^2}{N}=101^2+\dots+90^2-\frac{977^2}{10}=192.1$$

$$SSA=\sum_{i=1}^{K}n_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2=2(100.5-97.7)^2+3(102-97.7)^2+2(97-97.7)^2+3(92-97.7)^2=169.6$$

$$SSA=\sum_{i=1}^{K}\frac{T_{i\cdot}^2}{n_i}-\frac{T_{\cdot\cdot}^2}{N}=\frac{101^2}{2}+\dots-\frac{977^2}{10}$$

$$SSE=\sum\sum(Y_{ij}-\bar{Y}_{i\cdot})^2=SST-SSA=192.1-169.6=22.5$$

$$F_{3,6,0.95}=4.76$$

| منبع تغییر | SS | df | MS | F |
|---|---|---|---|---|
| تیمار | 169.6 | $K-1=4-1=3$ | $169.6/3=56.533$ | $\frac{56.533}{3.75}=15.076$ |
| خطا | 22.5 | $N-K=10-4=6$ | $22.5/6=3.75$ | |
| کل | 192.1 | $N=10$ | | |

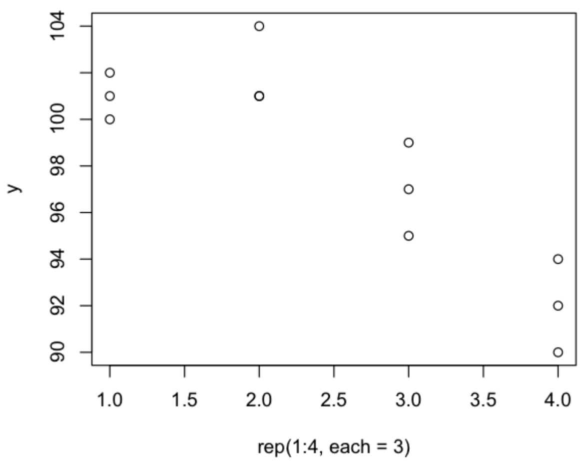
$$15.076 > F_{3,6,0.95}=4.76 \Rightarrow RH_0$$

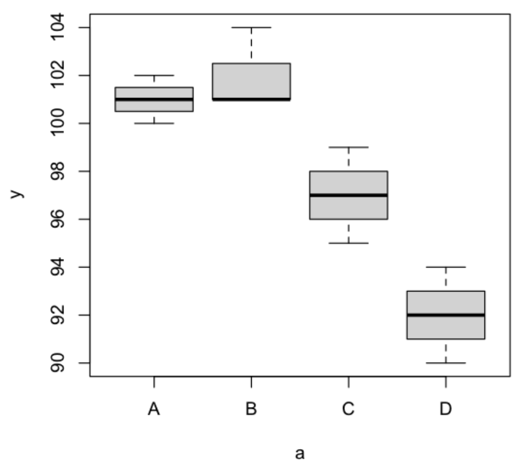
اثر A، B و C و D متفاوت عمل می‌کنند

## مثال ۲۷

```
> #################
> ## Example 27
>
> y = c(101,100,102,104,101,101,99,95,97,94,92,90)
> a<-rep(c("A","B","C","D"),each=3)
> a
 [1] "A" "A" "A" "B" "B" "B" "C" "C" "C" "D" "D" "D"
> aggregate(y~a,FUN=mean)
  a   y
1 A 101
2 B 102
3 C  97
4 D  92
> aggregate(y~a,FUN=sum)
  a   y
1 A 303
2 B 306
3 C 291
4 D 276
> mean(y)
[1] 98
> sum(y)
[1] 1176
> boxplot(y~a)
> plot(rep(1:4,each=3),y) ## point plots
> M<-aov(y~a)
> anova(M)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
a          3    186      62  20.667 0.0004002 ***
Residuals  8     24       3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

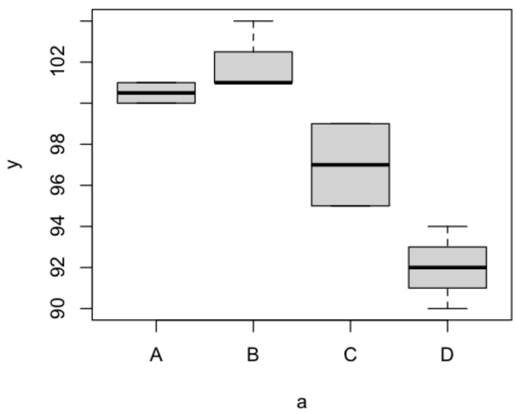
```
> summary(M)
            Df Sum Sq Mean Sq F value Pr(>F)    
a            3    186      62   20.67  4e-04 ***
Residuals    8     24       3                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(M)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ a)

$a
    diff       lwr        upr     p adj
B-A    1  -3.52881  5.5288096 0.8915309
C-A   -4  -8.52881  0.5288096 0.0847586
D-A   -9 -13.52881 -4.4711904 0.0009833
C-B   -5  -9.52881 -0.4711904 0.0313874
D-B  -10 -14.52881 -5.4711904 0.0004793
D-C   -5  -9.52881 -0.4711904 0.0313874

> 
> model.tables(M)
Tables of effects

 a
a
 A  B  C  D 
 3  4 -1 -6 
> 
> ### f Table
> alpha=0.05
> qf(1-alpha,3,8)
[1] 4.066181
```

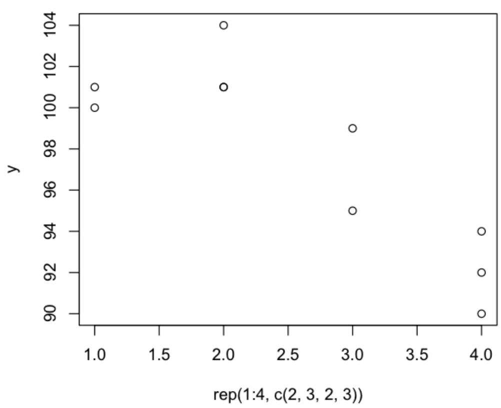


# مثال ۲۸

```
> ## Example 28
> y = c(101,100,104,101,101,99,95,94,92,90)
> a<-rep(c("A","B","C","D"),c(2,3,2,3))
> a
 [1] "A" "A" "B" "B" "B" "C" "C" "D" "D" "D"
> aggregate(y~a,FUN=mean)
  a     y
1 A 100.5
2 B 102.0
3 C  97.0
4 D  92.0
> aggregate(y~a,FUN=sum)
  a   y
1 A 201
2 B 306
3 C 194
4 D 276
> mean(y)
[1] 97.7
> sum(y)
[1] 977
> boxplot(y~a)
> plot(rep(1:4,c(2,3,2,3)),y) ## point plots
>
> M<-aov(y~a)
> anova(M)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
a          3  169.6  56.533  15.076 0.003357 **
Residuals  6   22.5   3.750
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> ### f Table
> alpha=0.05
> qf(1-alpha,3,6)
[1] 4.757063
> |
```



Subject:
Date

## ۳- آزمون کای دو پیرسون

آزمون کای دو پیرسون یا آزمون مربع خی برای بررسی متفاوت بودن معنی دار فراوانی های مشاهده از آنچه که در حالت قبول فرض صفر انتظار می رفت به کار می رود. یک مثال ساده: سکه ۱۲۰ بار پرتاب کنید اگر سکه سالم باشه انتظار داریم ۶۰ بار شیر ۶۰ بار خط بیاید اما در عمل ممکن است مثلاً ۴۵ بار شیر و ۷۵ بار خط آمده باشه و اکنون می خواهیم بررسی کنیم سکه سالم است.

آماره ی کای دو پیرسون (مربع خی) به صورت

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{k-1} \qquad k = \text{تعداد طبقه بندی ها}$$

بدست می آید که برای نمونه ی بزرگ دارای توزیع $\chi^2$ است

**مثال ۲۹** نتایج حاصل از خواندن حجم بورت و خواندن وزن در ترازو به صورت زیر است. آیا در خواندن حجم تعصبی وجود داشته است.

| عدد آخر | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | جمع |
|---|---|---|---|---|---|---|---|---|---|---|---|
| خواندن حجم بورت | 212 | 213 | 229 | 166 | 124 | 107 | 110 | 81 | 184 | 125 | 1500 |
| خواندن ترازو | 126 | 92 | 95 | 85 | 112 | 132 | 77 | 95 | 107 | 85 | 1000 |
| موارد انتظار بورت | 150 | 150 | 150 | 150 | 150 | 150 | 150 | 150 | 150 | 150 | 1500 |
| موارد انتظار ترازو | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1000 |

PAPCO

Subject:
Date

خواندن بورت $\chi^2 = \dfrac{(۲۱۲-۱۵۰)^۲ + \cdots + (۱۲۵-۱۵۰)^۲}{۱۵۰} = ۱۵۹,۳۶$

$K = 10$ $\chi^2_{9, 0.95} = 16.9$ $159.36 > 16.9 \Rightarrow$ تعصب وجود داشته

خواندن ترازو $\chi^2 = \dfrac{(126-100)^2}{100} + \cdots + \dfrac{(85-100)^2}{100} = 27.1$

$27.1 > 16.9 \Rightarrow$ تعصب وجود داشته

مثال ۲۰: تعداد شیشه آلات شکسته شده گزارش شده توسط کارکنان چهار آزمایشگاه در طول دوره ای مشخص به صورت: 9 11 17 24 : تعداد شکست ها شده است. آیا دلیلی بر متفاوت بودن اعتماد پذیری کارکنان وجود دارد.

$\begin{cases} H_0: \text{کارکنان یکسانند} \\ H_1: \text{کارکنان متفاوتند} \end{cases}$

$9 + 11 + 17 + 24 = 61 \qquad \dfrac{61}{4} = 15.25$

اگر کارکنان یکسان عمل می کردند انتظار داشتیم هر کدام ۱۵ تا شکست داشتند.

| $o_i$ | $e_i$ | $(o_i - e_i)^2 / e_i$ |
|---|---|---|
| 9 | 15.25 | $(9-15.25)^2/15.25 = 5.02$ |
| 11 | 15.25 | $(11-15.25)^2/15.25 = 0.201$ |
| 17 | 15.25 | $(17-15.25)^2/15.25 = 1.184$ |
| 24 | 15.25 | $(24-15.25)^2/15.25 = 2.561$ |

$\Sigma = 8.966$

$K = 4 \qquad \chi^2_{3, 0.95} = 7.81$

$8.966 > 7.81 \Rightarrow RH_0$

کارکنان در اعتماد پذیری یکسان نیستند

PAPCO

```
> #######example 29
> x <- c(212, 213, 229,166,124,107,110,81,184,125)
> res <- chisq.test(x)
> res

	Chi-squared test for given probabilities

data:  x
X-squared = 159.36, df = 9, p-value < 2.2e-16

> qchisq(0.95,9)
[1] 16.91898
> #######example 30
> x <- c(9,11,17,24)
> res <- chisq.test(x)
> res

	Chi-squared test for given probabilities

data:  x
X-squared = 8.9672, df = 3, p-value = 0.02973

> qchisq(0.95,3)
[1] 7.814728
>
```

Calibration Correlation Regression

- رگرسیون، ضریب همبستگی، کالیبراسیون

هرگاه بین دو متغیر رابطه‌ی خطی وجود داشته باشد

وابسته

X Y

قد پدر | قد فرزند

متغیر مستقل (ثابت) | متغیر وابسته (تصادفی)

در تجزیه دستگاهی هم همیشه به دنبال سیگنالی

متناسب با غلظت هستیم و ایده‌آل این است

که بین سیگنال و غلظت رابطه خطی وجود داشته باشد

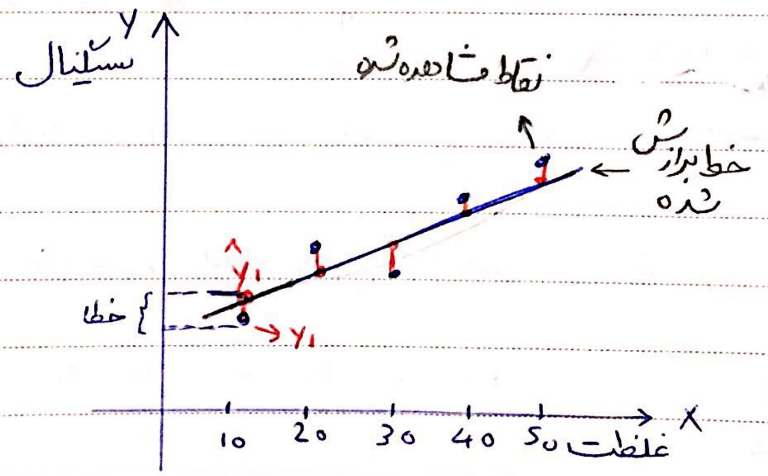
وابسته

رگرسیون: سیگنال ($Y$) غلظت ($X$)

برای اینکه بفهمید آیا بین دو متغیر ارتباط وجود دارد یا خیر اولین

گام رسم نمودار پراکنش (scatter plot) و سپس رسم خط برازش شده

به نقاط.




خطا: $Y_1-\hat{Y}_1$

پس هدف پیدا کردن این ارتباط است. پس رابطه خط به صورت

$$Y_i = a + bx_i + \varepsilon_i \qquad i=1\dots n$$

$a$: عرض از مبدأ، $b$: شیب، $n$: تعداد نمونه



به a و b ضرایب رگرسیونی هم می‌گویند

regression parameters

Subject:
Date

در نظر گرفته می شود که در رگرسیون فرض بر این است که خطاها ناهمبسته و دارای واریانس ثابت هستند و برای بدست آوردن آزمون فرضها فرض بر نرمال بودن خطاها است.

$$\varepsilon_i \sim (0, \sigma^2)$$

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$$

تحت فرض ناهمبستگی واریانس ها و ثابت بودن واریانس ها با مینیمم کردن مربعات خطا $\sum \varepsilon_i^2$ $\left(\sum (y_i - \hat{y}_i)^2\right)$ برآورد $a$ و $b$ بدست آورده می شود

$$\begin{cases} \hat{a} = \bar{y} - \hat{b}\bar{x} \\ \hat{b} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{cases}$$

نمونه

$$\left.\begin{matrix} (x_1, & y_1) \\ \vdots & \vdots \\ (x_n, & y_n) \\ \bar{x} & \bar{y} \end{matrix}\right\}$$

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

هدف بدست آوردن این خط هست

$$\hat{y} = \hat{a} + \hat{b}x \leftarrow \hat{y} = \hat{E}(Y \mid x)$$

میانگین خطا صفر فرض می کنیم $E(\varepsilon) = 0$

مثال ۳۱. محلولهای آبی استاندارد فلئورسئین در یک طیف سنج فلئورسانس آزمایش شده و شدت فلئورسانس زیر بدست داده اند

$\bar{y} = 13.1$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\sum y_i = 91.7 \leftarrow$ | 24.7 | 21 | 17.3 | 12.6 | 9 | 5 | 2.1 | شدت فلئورسانس |
| $\sum x_i = 42 \leftarrow$ | 12 | 10 | 8 | 6 | 4 | 2 | 0 | غلظت |

$\bar{x} = 6$

می خواهیم $\hat{y} = \hat{a} + \hat{b}x$ را بدست آوریم

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \hat{b} = \frac{S_{xy}}{S_{xx}}$$

Subject:
Date

$S_{xy}$ 216.2

$$\hat{b} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{(24.7\times 12 + \dots + 0\times 2.1) - \frac{(91.7\times 42)}{7}}{\underbrace{(0^2+2^2 \dots + 12^2) - \frac{1}{7}(42)^2}_{112}}$$

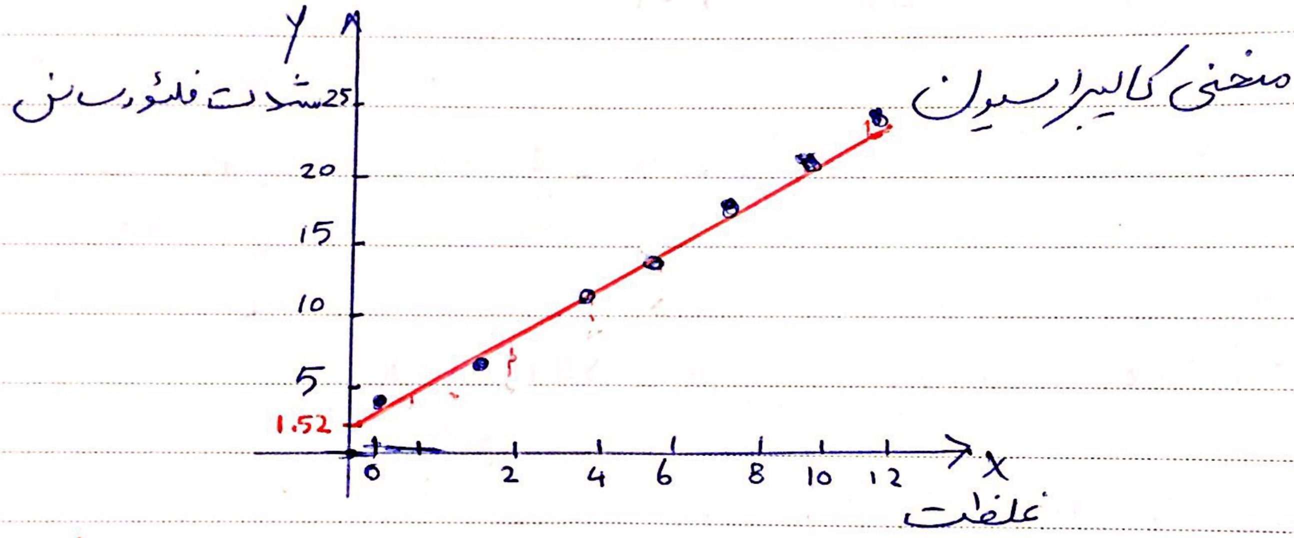
$$= 1.93$$

112 $S_{xx}$

$$\hat{a} = 13.1 - 1.93\times 6 = 1.52$$

عرض از مبدا      شیب

$$\hat{Y} = \underbrace{1.52}_{\text{عرض از مبدا}} + \underbrace{1.93}_{\text{شیب}} X$$




$\hat{Y}_1 = 1.52 + 1.93\times 0 = 1.52$, $\hat{\varepsilon}_1 = 0.58$

$\hat{Y}_2 = 1.52 + 1.93\times 1 \cong 5.38$, $\hat{\varepsilon}_2 = -0.38$

$\hat{Y}_3 = 9.24$, $\hat{\varepsilon}_3 = -0.24$

$\hat{Y}_4 = 13.1$, $\hat{\varepsilon}_4 = -0.5$

$\hat{Y}_5 = 16.96$, $\hat{\varepsilon}_5 = 0.34$

$\hat{Y}_6 = 20.82$, $\hat{\varepsilon}_6 = 0.18$

$\hat{Y}_7 = 24.68$, $\hat{\varepsilon}_7 = 0.018$

$X = 14 \quad 16$      پیشگویی در نقاط جدید

$$\hat{Y}_{14} = 1.52 + 1.93\times 14 = 24.68$$

$$\hat{Y}_{16} = 28.5$$

Subject:
Date

در رگرسیون مشابه آنالیز واریانس، جدول زیر تشکیل می‌شود و با توجه

به آماره‌ی $F$ محاسبه شده معنی‌داری رگرسیون را بررسی می‌کنند

| منبع تغییرات | SS | df (تعداد متغیر ثابت) | MS | F |
|---|---|---|---|---|
| رگرسیون | SSR | 1 | $SSR \to MSR$ | $\frac{MSR}{MSE} = F_0$ |
| خطا | SSE | $n-2$ | $\underbrace{SSE/n-2}_{MSE}$ | |
| کل | SST | $n-1$ | | |

$$\text{if } F_0 > F_{1, n-2, 1-\alpha} \Rightarrow RH_0$$

$$\begin{cases} H_0 : b = 0 & \text{رگرسیون معنی‌دار نیست} \\ H_1 : b \neq 0 & \text{رگرسیون معنی‌دار است} \end{cases}$$

مجموع مربعات کل

$$S_{yy} = SST = \sum (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i\right)^2$$

مجموع مربعات خطا

$$\begin{cases} SSE = SST - SSR \\ SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SST - SSR \end{cases}$$

مجموع مربعات رگرسیون

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = \hat{b}\, S_{xy}$$

مثال ۳۲. در مثال ۳۱ معنی‌داری رگرسیون را آزمون کنید

$$SST = (2.1)^2 + (5)^2 + \dots + (24.7)^2 - \frac{1}{7}(91.7)^2 = 418.24$$

$$SSR = 1.093 \times 216.2 \simeq 417.3 \to MSR = 417.3$$

$$SSE \simeq 418.24 - 417.3 \simeq 0.94$$

| منبع | SS | df | MS | F |
|---|---|---|---|---|
| رگرسیون | 417.3 | 1 | 417.3 | |
| خطا | 0.94 | 5 | $0.94/5 \simeq 0.19$ | |
| کل | 418.24 | 6 | | |

$$F_0 = \frac{417.3}{0.19} = 2227.9 > F_{1, 5, 0.95} = 6.6$$

رگرسیون قویاً معنی‌دار است.

نکته ۱: معمولاً برای بدست آوردن آزمون فرض ها و فواصل اطمینان فرض می شود خطاها دارای توزیع نرمال هستند $\varepsilon \sim N(0, \sigma^2)$

$$y = \underbrace{a + bx}_{\text{ثابت}} + \varepsilon \Rightarrow y \sim \varepsilon$$

چون ترکیب خطی از $\varepsilon$ پس $y$ نرمال

$$y/x \sim N(a + bx, \sigma^2)$$

ثابت می شود یک برآورد مناسب برای $\sigma^2$:

$$\boxed{\hat{\sigma}^2 = \frac{SSE}{n-2} = MSE}$$ واریانس نمونه

نکته ۲: از طرفی چون $\hat{a}$ و $\hat{b}$ هم ترکیبی از $y$ هاست پس دارای توزیع نرمال هستند و ثابت شده است

$$\hat{a} \sim N(a, \frac{\sigma^2}{S_{xx}})$$

$$\hat{b} \sim N(b, \frac{\sigma^2 \sum x_i^2}{n S_{xx}}) \qquad S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

بنابراین فواصل اطمینان برای $a$ و $b$:

$$\boxed{\begin{aligned} a &: \hat{a} \pm t_{n-2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \\ b &: \hat{b} \pm t_{n-1} \cdot \hat{\sigma} \sqrt{\frac{\sum x_i^2}{n S_{xx}}} \end{aligned}}$$

(چون از $\hat{\sigma}$ استفاده می کنیم توزیع $t$ استودنت) اگر $\sigma$ را داشتیم توزیع $z$

Date

برای بررسی فرضیه $\begin{cases} H_0: b=0 \\ H_1: b\neq 0 \end{cases}$ که معادله همان آزمون معنی داری رگرسیون بود که از روی جدول آنالیز واریانس و F آزمون می شود می توان از آماره زیر هم استفاده کرد:

$$T_0 = \frac{\hat{b}}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2} \quad \text{if } |T_0| > t_{n-2, 1-\alpha/2} \Rightarrow RH_0$$

و برای $\begin{cases} H_0: b=b_0 \\ H_1: b\neq b_0 \end{cases}$

$$T_{0b} = \frac{\hat{b}-b_0}{\hat{\sigma}/\sqrt{S_{xx}}} \quad \text{if } |T_{0b}| > t_{n-2, 1-\alpha/2} \Rightarrow RH_0$$

برای $\begin{cases} H_0: a=a_0 \\ H_1: a\neq a_0 \end{cases}$

$$T_{0a} = \frac{\hat{a}-a_0}{\hat{\sigma}\sqrt{\frac{\sum x_i^2}{nS_{xx}}}} \quad \text{if } |T_{0a}| > t_{n-2, 1-\alpha/2} \Rightarrow RH_0$$

استفاده کنید.

مثال ۳۳ برای مثال ۳۱ ، فاصله اطمینان برای ضرایب رگرسیونی بدست آورید و آزمون فرض های $\begin{cases} H_0: b=0 \\ H_1: b\neq 0 \end{cases}$ و $\begin{cases} H_0: a=0 \\ H_1: a\neq 0 \end{cases}$ را انجام دهید

$$a: \hat{a} \pm t_{n-2,1-\alpha/2}\hat{\sigma}\sqrt{\frac{\sum x_i^2}{nS_{xx}}} = 1.52 \pm 2.57\sqrt{0.19}\sqrt{\frac{364}{7\times 112}} \simeq (0.76, 2.28)$$

جایگذاری کنید

$n=7$

$t_{5,0.975} = 2.57$

$\hat{\sigma}^2 = MSE \underset{\text{جدول بدست آوردیم}}{\simeq} 0.19$

$\hat{a} = 1.52$

$\sum x_i^2 = 0^2+2^2+4^2+6^2+8^2+10^2+12^2 = 364$

$S_{xx} = 112$

P4PCO

Scanned with CamScanner

Subject:
Date

$$b: \hat{b} \pm t_{n-2,1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = 1.93 \pm 2.57 \times \frac{\sqrt{0.19}}{\sqrt{112}}$$

$$= (1.825, 2.035)$$

$$T_0 = \frac{1.93}{\sqrt{0.19/112}} \cong 47 > 2.57 \Rightarrow RH_0$$

$$T_{0a} = \frac{1.52}{\sqrt{0.19(3.64)/7 \times 112}} \cong 5.1 > 2.57 \Rightarrow RH_0$$

(بدست آوردن مقادیر $x_0$ یا $y_0$ از روی خط رگرسیون برازش شده)

اکنون با داشتن $\hat{y} = \hat{a} + \hat{b}x$ به راحتی می توانید با داشتن

یک $y_0$، $x_0$ جدید یا با یک $x_0$، $y_0$ جدید را بدست آورید

مثال ۳۴ در مثال ۳۱ دیدیم خط $\hat{y} = 1.52 + 1.93x$ بدست آمد اکنون

اگر $y_0 =$ ۲٫۹ ، $y_0 =$ ۱۳٫۵ ، $y_0 =$ ۲۳ باشد $x_0$های

متناظر را بدست آورید

۲٫۹ = ۱٫۵۲ + ۱٫۹۳ $x_0 \Rightarrow \hat{x}_0 \simeq 0.72$

۱۳٫۵ = ۱٫۵۲ + ۱٫۹۳ $x_0 \Rightarrow \hat{x}_0 \simeq 6.21$

۲۳ = ۱٫۵۲ + ۱٫۹۳ $x_0 \Rightarrow \hat{x}_0 \simeq$ ۱۱٫۱۳

همچنین ثابت شده است که می توان خطای تقاطعی $x_0$ بدست آمده (انحراف استاندارد) را از رابطه زیر بدست آورد. انحراف استاندارد اندازه‌گیری غلظت نمونه مجهول

سیگنال نمونه مجهول

$$S_{x_0} = \frac{\hat{\sigma}}{\hat{b}} \sqrt{1 + \frac{1}{n} + \frac{(Y_0 - \bar{Y})^2}{\hat{b}^2 \underbrace{\sum (x_i - \bar{x})^2}_{S_{xx}}}}$$

برای مثال قبل: $Y_0 = 2.9$

$$S_{x_0} = \frac{\sqrt{0.19}}{1.93} \times \sqrt{1 + \frac{1}{7} + \frac{(2.9 - 13.1)^2}{1.93^2 . 112}}$$

$$\cong 0.26$$

$Y_0 = 13.5$ $\quad S_{x_0} = \frac{\sqrt{0.19}}{1.93} \times \sqrt{1 + \frac{1}{7} + \frac{(13.5 - 13.1)^2}{1.93^2 . 112}} \simeq 0.24$

$Y_0 = 23$ $\quad S_{x_0} = \frac{\sqrt{0.19}}{1.93} \times \sqrt{1 + \frac{1}{7} + \frac{(23 - 13.1)^2}{1.93^2 \times 112}} \simeq 0.26$

نکته: فاصله اطمینان برای $x_0$:

$$x_0 : \hat{x}_0 \pm t_{n-2, 1-\alpha/2} \, S_{x_0}$$

اولی $x_0 : 0.72 \pm 2.57 \, . \, 0.26 \cong (0.05 , 1.4)$

دومی $x_0 : 6.21 \pm 2.57 \, . \, 0.24 \simeq (5.59 , 6.8)$

سومی $x_0 : 11.13 \pm 2.57 \, . \, 0.26 \approx (10.46 , 11.8)$

دقت دومی بیشتر است طول فاصله اطمینان هم کوچکتر است می توان نتیجه گرفت

$S_{x_0} = 0.24$

هرچه $y_0$ به $\bar{y}$ نزدیکتر است چون جمله سوم زیر رادیکال در $S_{x_0}$ به صفر میل می‌کند پس $S_{x_0}$ کوچکتر می‌شود.

نکته: اگر برای بدست آوردن یک $y_0$ ، تعداد m قرائت (تکرار) انجام دهد

در این صورت

$$S_{x_0} = \frac{\hat{\sigma}}{\hat{b}} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{\hat{b}^2 \sum (x_i - \bar{x})^2}}$$

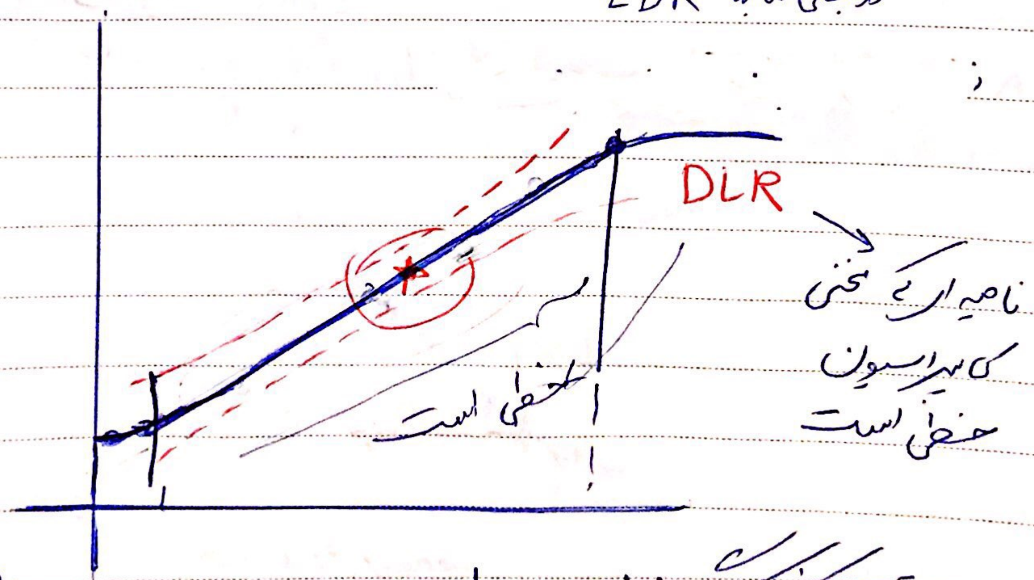
و باعث افزایش دقت (کم شدن $S_{x_0}$ می‌شود)

نکته: برای اینکه دقت اندازه‌گیری غلظت بیشتر باشد و در واقع مقدار $S_{x_0}$ کم‌تر باشد، m و n بزرگ باشد و $y_0$ به $\bar{y}$ نزدیک باشد یعنی

هرچه سیگنال نمونه مجهول به $\bar{y}$ نزدیکتر باشد یا در وسط DLR باشد

(Dynamic linear range)

در بعضی کتابها LDR




دقت کنید که اگر نمونه مجهول در محدوده‌ی DLR نباشد حق اندازه‌گیری نمونه مجهول را نداریم و برای سنجش باید نمونه مجهول را غلیظ تر یا رقیق تر کنیم.

Date

# Detection limit یا limit of detection حد تشخیص

(DL)

حد تشخیص یا حد پایینی تشخیص (Lod) و یا حد آشکارسازی در شیمی تجزیه
(lower limit of detection)
کمترین مقدار یک ماده است که می‌توان از محلول بلنک (Blank) که فاقد آن ماده است تشخیص داده شود. یا به عبارت دیگر کمترین غلظتی است که می‌تواند به وسیله یک روش خاص اندازه‌گیری شود. (کوچکترین غلظتی که یک روش می‌تواند با درجه اطمینان مشخصی آن را تشخیص دهد)

LOD →

اگر برای غلظت صفر تکرار داشتیم انحراف معیار غلظت صفر $S_B$

$$Y = a + bX \xrightarrow{X=0} Y = a = \underbrace{Y_B}_{\text{سیگنال شاهد}}$$

$$S_B = \sqrt{\hat{\sigma}^2}$$ از عدم اسفیوز

$$\left\{ \begin{array}{l} Y_0 = Y_B + 3S_B \\ Y_0 = Y_B + bX_0 \end{array} \right. \Rightarrow \boxed{X_0 = \frac{3S_B}{b}}$$

مقدار حد تشخیص LOD ↓
حداقل غلظت‌های قابل تشخیص

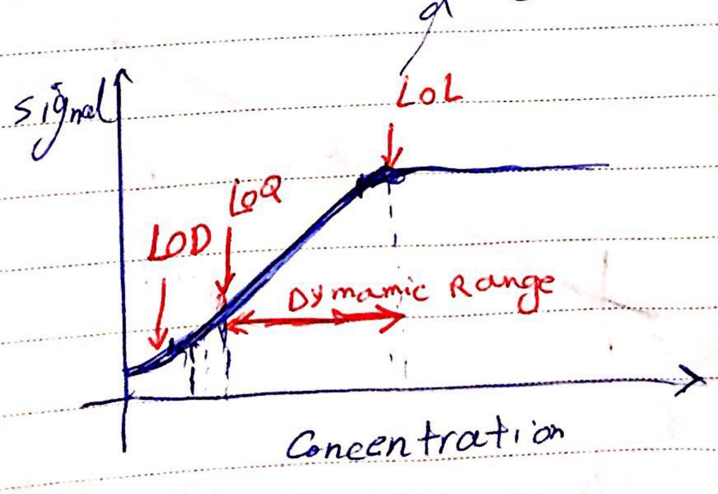
$Y_0 = Y_B + 10S_B$ → limit of quantitative or LOQ

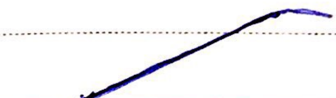
$$LOQ = \frac{10S_B}{b}$$

گستره DLR

LoL : limit of linearity

LOQ از LoL تا



LoQ — LoL
ناحیه خطی
(گستره غلظت)

Subject:
Date

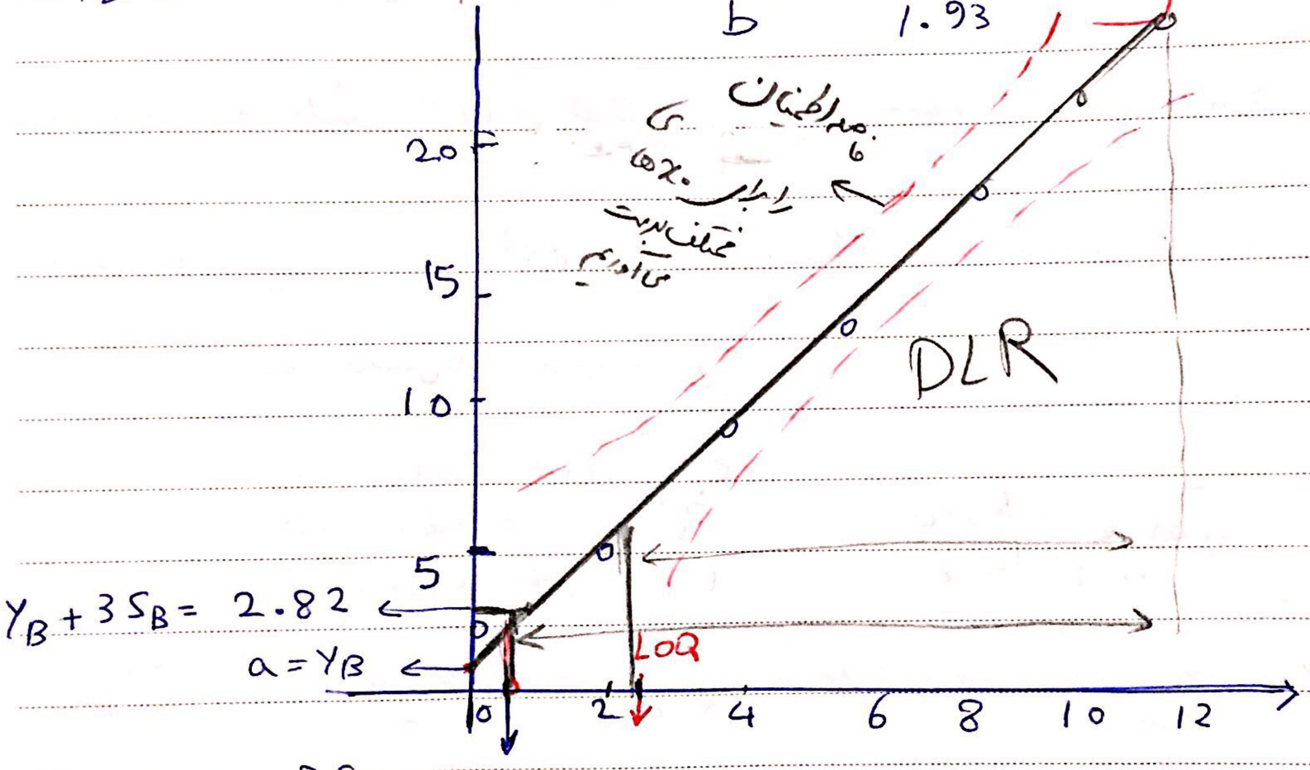
مثال ۲۵ حد تشخیص را برای تعیین مقدار فلئورسین مثال ۳۱ را برآورد نمایید

$$a = Y_B = 1.52 \qquad S_B = \sqrt{0.19} \simeq 0.433$$

$$Y_0 = 1.52 + 3 \times 0.433 = 2.82$$

$\longrightarrow X_0 = 0.67$

$$\text{DL or LoD: } X_0 = \frac{3S_B}{b} = \frac{3 \times 0.433}{1.93} = 0.67$$




$$\frac{3S_B}{b} = 0.67 = \text{LoD}$$

$$\text{LoQ} = \frac{10 \times 0.433}{1.93} = 2.24$$

# حساسیت (sensitivity) دستگاه

دو نوع حساسیت داریم

① حساسیت کالیبراسیونی : همان شیب منحنی کالیبراسیون است هرچه (b) بزرگتر باشد حساسیت روش بیشتر است و به عبارت دیگر غلظت های نزدیکتر بهم ، بهتر تشخیص داده می شود

② حساسیت تجزیه‌ای : در یک نقطه تعریف می شود و عبارت است از

$$\text{حساسیت تجزیه‌ای} = \frac{b}{\hat{\sigma}_{sig}}$$

$b \rightarrow$ شیب منحنی کالیبراسیون

$\hat{\sigma}_{sig} \rightarrow$ انحراف استاندارد سیگنال در یک نقطه خاص (از روی تکرارها محاسبه می‌شود)

$$\hat{\sigma}_{sig} = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

مثال 36. در اندازه‌گیری سرب به روش نشر شعله‌ای معادله زیر بدست آمد

$$I = 0.0312 + 1.12\, C$$

$I$ : کاشدت نشر — $C \rightarrow$ غلظت

| C | تکرار | mean | $S = \hat{\sigma}$ |
|---|---|---|---|
| 10 | 10 | 11.62 | 0.15 $= \hat{\sigma}_{10}$ |
| 1 | 10 | 1.14 | 0.025 $= \hat{\sigma}_1$ |
| 0 | 29 | 0.0296 | 0.0082 $= \hat{\sigma}_0$ |

Subject:
Date

الف) حساسیت کالیبراسیونی را بدست آورید

ب) حساسیت تجزیه‌ای در غلظت های ۰ و ۱۰

ج) LOD و LOQ را بدست آورید

حساسیت کالیبراسیونی $= \hat{b} = 1.12$

حساسیت تجزیه‌ای در صفر $= \dfrac{1.12}{0.025}$ حساسیت تجزیه‌ای در ۱۰ $= \dfrac{1.12}{0.15}$

$$LOD = 3\frac{S_B}{b} = 3 \times \frac{0.0082}{1.12}$$

چون تکرار داریم $S_B = \hat{\sigma}_0$ غلظت صفر

$$LOQ = 10\frac{S_B}{b} = 10 \times \frac{0.0082}{1.12}$$

نکته:

در محاسبات رگرسیونی معیار ضریب تعیین نیز محاسبه می‌شود که

بیانگر درصد تغییرات غلظت و سیگنال است

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$\text{Adjusted } R^2 = 1 - \frac{SSE/n-K}{SST/n-1}$$

در مثال ۳۱:

$$R^2 = \frac{417.3}{418.24} = 0.9978$$

$$\text{Adjusted } R^2 = 1 - \frac{0.94/5}{418.24/6} = 0.9973$$

Subject:
Date

# Correlation coefficient — ضریب همبستگی

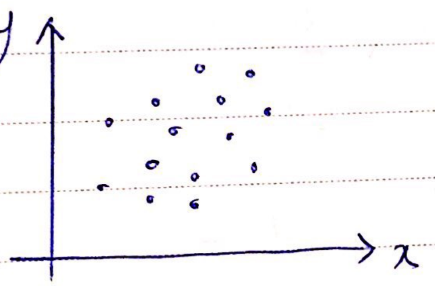
برای بررسی میزان همبستگی (خطی) دو متغیر x و y از معیار ضریب همبستگی خطی

که به صورت زیر محاسبه می شود

r ضریب همبستگی نمونه → ρ ضریب همبستگی جامعه

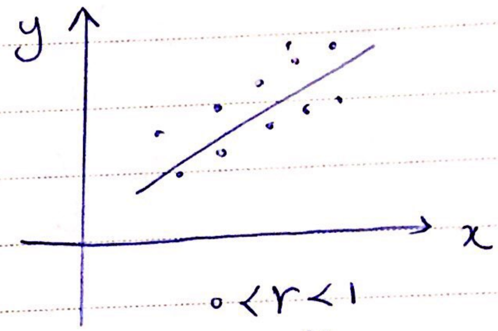
$$\hat{\rho} = r = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

r عددی در بازه‌ی ۱ و ۱- است.



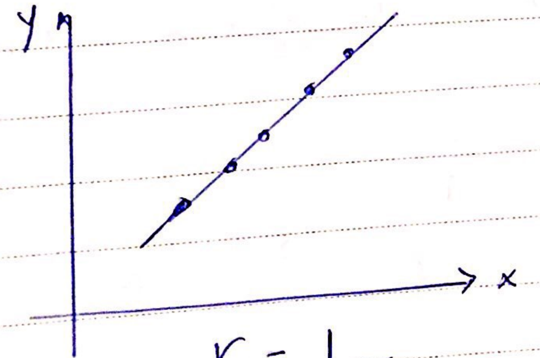

$r = 0$

ارتباط خطی بین x و y وجود ندارد



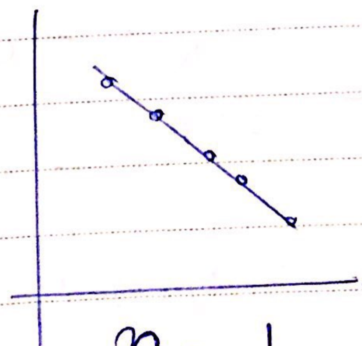

$0 < r < 1$

همبستگی مثبت بین x و y



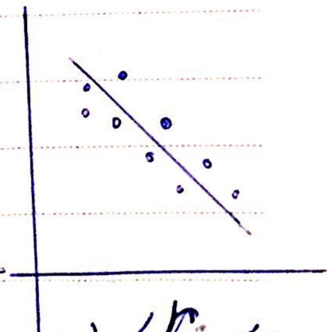

$r = 1$

همبستگی کامل مثبت



$r = -1$

همبستگی کامل منفی



$-1 < r < 0$

همبستگی منفی

Subject:
Date

آزمون فرض: برای بررسی فرض آماری
$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$$

$$T_0 = \frac{\sqrt{n-2}\; r}{\sqrt{1-r^2}}$$

استفاده می‌شود که

$$\text{if } |T_0| > t_{n-2,\,1-\alpha/2} \Rightarrow R\,H_0$$

مثال ۳۷: برای مثال ۳۱ ضریب همبستگی را بدست آورید و آزمون
$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$$
را انجام دهید.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{216.2}{\sqrt{112}\sqrt{418.24}} = 0.99888$$

$$t_0 = \frac{\sqrt{7-2}\;.\;0.99888}{\sqrt{1-0.99888^2}} \simeq 47.2$$

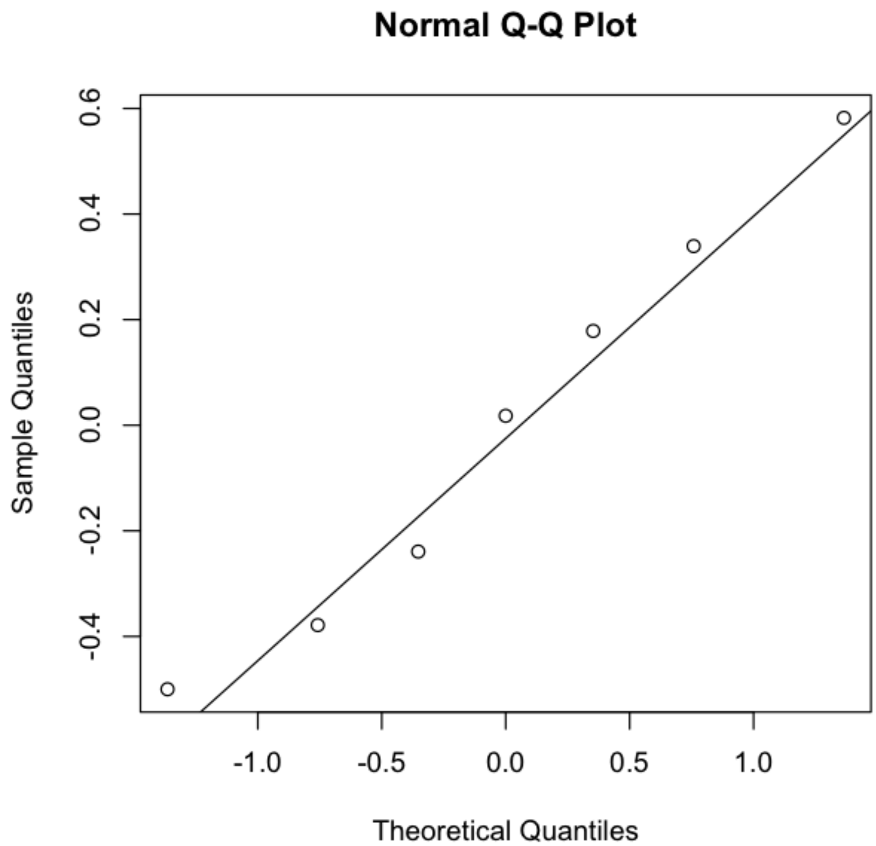
$$t_{5,\,0.975} = 2.57$$

$$47.2 > 2.57$$

$\Rightarrow$ فرض $H_0$ رد می‌شود.

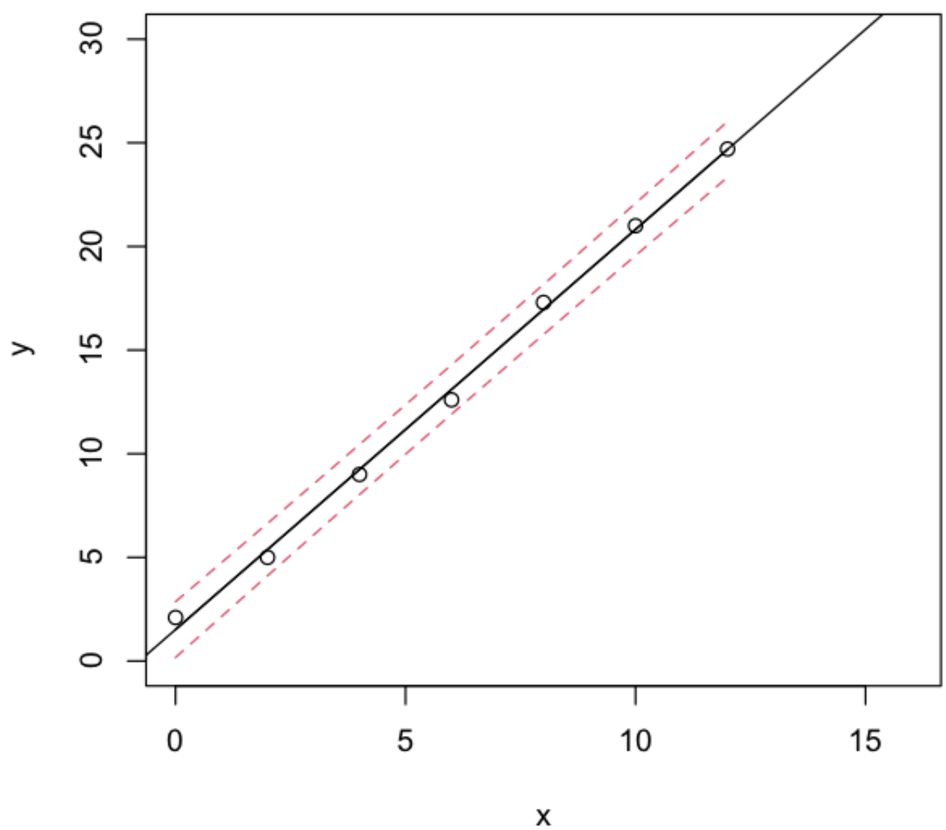
```
> ######example 31,31,33
> x<-c(0,2,4,6,8,10,12)
> y<-c(2.1,5,9,12.6,17.3,21,24.7)
>
> reg<-lm(y~x)
> summary(reg)

Call:
lm(formula = y ~ x)

Residuals:
       1        2        3        4        5        6        7 
 0.58214 -0.37857 -0.23929 -0.50000  0.33929  0.17857  0.01786 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.5179     0.2949   5.146  0.00363 ** 
x             1.9304     0.0409  47.197 8.07e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4328 on 5 degrees of freedom
Multiple R-squared:  0.9978,	Adjusted R-squared:  0.9973 
F-statistic:  2228 on 1 and 5 DF,  p-value: 8.066e-08
```

```
> confint(reg,level=0.95)
              2.5 %   97.5 %
(Intercept) 0.75970 2.276014
x           1.82522 2.035495
> anova(reg)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x          1 417.34  417.34  2227.5 8.066e-08 ***
Residuals  5   0.94    0.19                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> qt(0.975,5)
[1] 2.570582
> qf(0.95,1,5)
[1] 6.607891
> fitted(reg)
        1         2         3         4         5         6         7 
 1.517857  5.378571  9.239286 13.100000 16.960714 20.821429 24.682143 
> confint(reg)
              2.5 %   97.5 %
(Intercept) 0.75970 2.276014
x           1.82522 2.035495
> #normality
> resid(reg)
          1           2           3           4           5           6           7 
 0.58214286 -0.37857143 -0.23928571 -0.50000000  0.33928571  0.17857143  0.01785714 
> qqnorm(resid(reg))
> qqline(resid(reg))
> shapiro.test(resid(reg))

	Shapiro-Wilk normality test

data:  resid(reg)
W = 0.965, p-value = 0.8603
```




```
> #(interval,var(E(y)hat))
> predict(reg,interval="confidence",se.fit=T)
$fit
        fit       lwr       upr
1  1.517857  0.759700  2.276014
2  5.378571  4.783824  5.973319
3  9.239286  8.769097  9.709475
4 13.100000 12.679450 13.520550
5 16.960714 16.490525 17.430903
6 20.821429 20.226681 21.416176
7 24.682143 23.923986 25.440300

$se.fit
[1] 0.2949360 0.2313668 0.1829115 0.1636011 0.1829115 0.2313668 0.2949360

$df
[1] 5

$residual.scale
[1] 0.4328477
```

```
> #plot
>
>
> require(graphics)
> plot(x,y)
> abline(coef(reg))
> pred.frame<-data.frame(x=seq(0,12,by=0.5))
> pp<-predict(reg,newdata=pred.frame,int="p")
> pred.size<-pred.frame$x
> plot(x,y,xlim=c(0,16),ylim=c(0,30))
> abline(coef(reg))
> matlines(pred.size,pp,col=c(1,2,2),lty=c(1,2,2))
>
```



```
> ##############
> # Example 34
> coef(reg)
(Intercept)           x
   1.517857    1.930357
> a<-coef(reg)[1]
> b<-coef(reg)[2]
> y0=c(2.9,13.5,23)
> x0=(y0-a)/b
> x0
[1]  0.7160037  6.2072155 11.1285846
> SS=anova(reg)
> MSE=SS$"Mean Sq"[2]
> n=length(x)
> Sx0=(sqrt(MSE)/b)*sqrt(1+1/n+((y0-mean(y))^2/(b^2*var(x)*(n-1))))
> Sx0
[1] 0.2645698 0.2397542 0.2631933
> alpha=0.05
> x0+qt(1-alpha/2,n-2)* Sx0
[1]  1.396102  6.823523 11.805144
> x0-qt(1-alpha/2,n-2)* Sx0
[1]  0.03590545  5.59090769 10.45202483
>
>
  ##############

> ##############
> # Example 37
> cor.test(x,y)

	Pearson's product-moment correlation

data:  x and y
t = 47.197, df = 5, p-value = 8.066e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9920730 0.9998421
sample estimates:
      cor
0.9988796
```