Statistics in Analytical Chemistry

آمار در شیمی تجزیه

هر کمیت تجربی با عدم قطعیت همراه است (وجود خطاها)

هدف استفاده از آمار و تعریف چند اصطلاح

X1, X2, ...Xn
Mean
median

- گزارش بهترین نتیجه

ـ استفاده از میانگین حسابی (متداول‌ترین روش)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

ـ میانه: باید اعداد را به ترتیب مرتب کنیم. اگر تعداد اعداد فرد باشد، عدد وسط و اگر تعداد اعداد زوج باشد میانگین دو عدد وسط میانه نام دارد.

ـ مُد (شیوه): عددی که بیشترین تکرار را داشته باشد.

- بررسی داده ها از نظر صحت و دقت:

دقت: نزدیکی و تکرارپذیری داده ها به هم

صحت: نزدیکی داده ها به مقدار حقیقی

روش‌های بیان دقت:

۱. رنج (گستره): تفاوت بزرگ‌ترین عدد از کوچک‌ترین عدد. هرچه رنج کم‌تر، دقت بیش‌تر

w

۲. متوسط انحراف از میانگین $(\bar{d})$: هرچه $\bar{d}$ کوچک‌تر، دقت بیش‌تر

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

deviation=di= Xi-X̄

W= biggest value – smallest value

۳. انحراف معیار ← انحراف معیار نمونه، هرچه انحراف معیار کوچک‌تر، دقت بیش‌تر
← انحراف معیار جمعیت

standard deviation= SD

جمعیت: آنالیزهایی که بی‌نهایت‌بار تکرار شده باشد جمعیت نام دارد (حالت ایده‌آل)

متوسط آنالیزهایی که بی‌نهایت‌بار انجام شده باشند برابر با مقدار واقعی آنالیت (μ) است. $\bar{x} = \mu$

نمونه: به آنالیزهایی که به تعداد محدود تکرار می‌شوند نمونه یا Sample می‌گوییم (حالت واقعی)

انحراف معیار جمعیت $\delta = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$

انحراف معیار نمونه $S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

تعداد آنالیزهای تکراری انجام شده

sample
population

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N-1}} \qquad (6\text{-}4)$$

An Alternative Expression for Sample Standard Deviation

To find s with a calculator that does not have a standard deviation key, the following rearrangement is easier to use than directly applying Equation 6-4:

$$s = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N-1}} \qquad (6\text{-}5)$$

Example 6-1 illustrates the use of Equation 6-5 to calculate s.

EXAMPLE 6-1

The following results were obtained in the replicate determination of the lead content of a blood sample: 0.752, 0.756, 0.752, 0.751, and 0.760 ppm Pb. Find the mean and the standard deviation of this set of data.

Solution

To apply Equation 6-5, we calculate $\sum x_i^2$ and $(\sum x_i)^2/N$.

Sample	x_i	x_i^2
1	0.752	0.565504
2	0.756	0.571536
3	0.752	0.565504
4	0.751	0.564001
5	0.760	0.577600
	$\sum x_i = 3.771$	$\sum x_i^2 = 2.844145$

$$\bar{x} = \frac{\sum x_i}{N} = \frac{3.771}{5} = 0.7542 \approx 0.754 \text{ ppm Pb}$$

$$\frac{(\sum x_i)^2}{N} = \frac{(3.771)^2}{5} = \frac{14.220441}{5} = 2.8440882$$

Substituting into Equation 6-5 leads to

$$s = \sqrt{\frac{2.844145 - 2.8440882}{5 - 1}} = \sqrt{\frac{0.0000568}{4}} = 0.00377 \approx 0.004 \text{ ppm Pb}$$

5-12. Find the mean and median of each of the following sets of data. Determine the deviation from the mean for each data point within the sets and find the mean deviation for each set. Use a spreadsheet if it is convenient.

*(a) 0.0110 0.0104 0.0105

(b) 24.53 24.68 24.77 24.81 24.73

*(c) 188 190 194 187

(d) 4.52×10^{-3} 4.47×10^{-3}
4.63×10^{-3} 4.48×10^{-3}
4.53×10^{-3} 4.58×10^{-3}

*(e) 39.83 39.61 39.25 39.68

(f) 850 862 849 869 865

6-7. Consider the following sets of replicate measurements:

*A	B	*C	D	*E	F
9.5	55.35	0.612	5.7	20.63	0.972
8.5	55.32	0.592	4.2	20.65	0.943
9.1	55.20	0.694	5.6	20.64	0.986
9.3		0.700	4.8	20.51	0.937
9.1			5.0		0.954

For each set, calculate the (a) mean; (b) median; (c) spread, or range; (d) standard deviation; and (e) coefficient of variation.

نکته: S و δ جمع‌پذیر نیستند.

اما در تمام مراحل تجزیه از جمله Analysis, Sample Preparation, Sampling و... باید دقت را بررسی کنیم اما چون S و δ جمع‌پذیر نیستند برای جمع کردن خطا در تمام قسمت‌ها باید از یک روش جمع‌پذیر استفاده کنیم.

۴. واریانس، که جمع‌پذیر است، هرچه واریانس V کوچک‌تر، دقت بیش‌تر.

$$\text{واریانس } V = S^2 \quad \text{or} \quad \delta^2$$

$$S_T = \sqrt{v_1 + v_2 + \ldots + v_n}$$

نکته: چون واریانس، واحد اندازه‌گیری را به توان ۲ می‌رساند، خیلی متداول نیست.

۵. ضریب تغییر (C_v): هرچه C_v کوچک‌تر، دقت بیش‌تر

واقعی ↘ ایده‌آل ↙

$$C_v = \frac{S}{\bar{x}} \quad \text{or} \quad \frac{\delta}{\mu}$$

coefficient of variation=CV
ضریب تغییر

۶. RSD (Relative Standard Diviation)، هرچه RS'D کوچک‌تر، دقت بیش‌تر

$$RSD = \frac{S}{\bar{x}} \times ۱۰۰ \quad \text{or} \quad \frac{\delta}{\mu} \times ۱۰۰$$

نکته: متداول‌ترین روش بیان دقت RSD و بعد از آن S است.

نکته: عدد گزارش شده به‌عنوان نتیجه آزمایش فاکتور دقت $\pm \bar{x}$ است.

روش‌های بیان صحت:

۱. خطای مطلق:

$$\text{خطای مطلق} = \bar{x} - \mu$$

Xt
Error

واحد خطای مطلق همان واحد داده مورد اندازه‌گیری است. مثبت یا منفی بودن نشان‌دهنده جهت خطاست.

۲. خطای نسبی:

$$\text{خطای نسبی} = \frac{\bar{x} - \mu}{\mu} \times ۱۰۰ \Rightarrow \%$$

$$\times ۱۰۰۰ \Rightarrow \text{ppt}$$

$$\times ۱۰^{۶} \Rightarrow \text{ppm}$$

خطای نسبی واحد ندارد، پس میتوانیم خطای نسبی چند آنالیز را با هم مقایسه کنیم.

انواع خطا در تجزیه:

Completely fresh start

Gross errors

۱-**خطای فاحش (بزرگ):** اصلاً قابل قبول نیست، به‌راحتی قابل تغییر است و باید حذف شود. مقدار خطا بزرگ بوده و مشخص می‌باشد که مربوط به تغییرات و یا اشتباهات بزرگ است.

Determinate errors

۲-**خطای سیستماتیک:** این خطا هم اصلاً قابل قبول نیست. این خطا جهت مثبت و منفی دارد. سه منبع خطای سیستماتیک را ایجاد می‌کند (منبع روش، منبع دستگاهی، منبع شخصی)

5B-1 Sources of Systematic Errors

There are three types of systematic errors:

- **Instrumental errors** are caused by nonideal instrument behavior, by faulty calibrations, or by use under inappropriate conditions.
- **Method errors** arise from nonideal chemical or physical behavior of analytical systems.
- **Personal errors** result from the carelessness, inattention, or personal limitations of the experimenter.

Random errors
(Indeterminate)

3- **خطای تصادفی**

constant
proportional

انواع خطاهای سیستماتیک

ثابت: با تغییر میزان آنالیت و نمونه تغییر نمی‌کند، مانند خطای مصرف تیترانت توسط معرف در تیتراسیون.

متغیر: مانند خطای جذب رطوبت در هنگام توزین یک نمک جاذب رطوبت.

متناسب: با تغییر میزان آنالیت و نمونه، تغییر می‌کند. مانند خطای ۲درصدی حضور هافونیم در سنگ معدن زیرکونیوم.

خطای سیستماتیک را باید تشخیص داد، اندازه‌گیری کرد و آن را حذف نمود.

1 g ore 0.02 g
10 g 0.2 g

روش‌های اندازه‌گیری خطای سیستماتیک:

۱. استفاده از نمونه استاندارد.

۲. استفاده از نمونه شاهد (Blank). نمونه‌ای که تمام اجزا به جز آنالیت را دارد.

۳. استفاده از یک روش دیگر که روشی استاندارد باشد.

۴. روش افزودن مقدار مشخصی استاندارد به نمونه حقیقی آنالیز شده و محاسبه بازیابی (recovery). به این روش spike می‌گویند.

5- تغییر دراندازه نمونه برای تشخیص خطاهای ثابت

EXAMPLE 5-2

Suppose that 0.50 mg of precipitate is lost as a result of being washed with 200 mL of wash liquid. If the precipitate weighs 500 mg, the relative error due to solubility loss is $-(0.50/500) \times 100\% = -0.1\%$. Loss of the same quantity from 50 mg of precipitate results in a relative error of -1.0%.

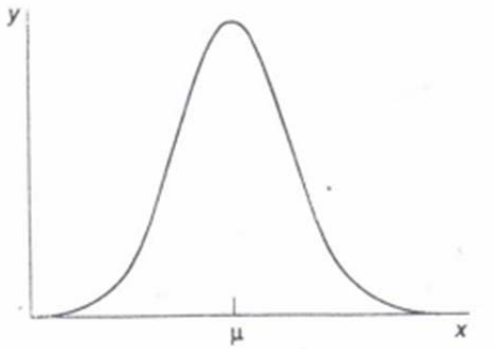
5-11. A loss of 0.4 mg of Zn occurs in the course of an analysis for that element. Calculate the percent relative error due to this loss if the mass of Zn in the sample is

*(a) 30 mg. (b) 150 mg.
*(c) 300 mg. (d) 500 mg.

5-10. The color change of a chemical indicator requires an overtitration of 0.03 mL. Calculate the percent relative error if the total volume of titrant is

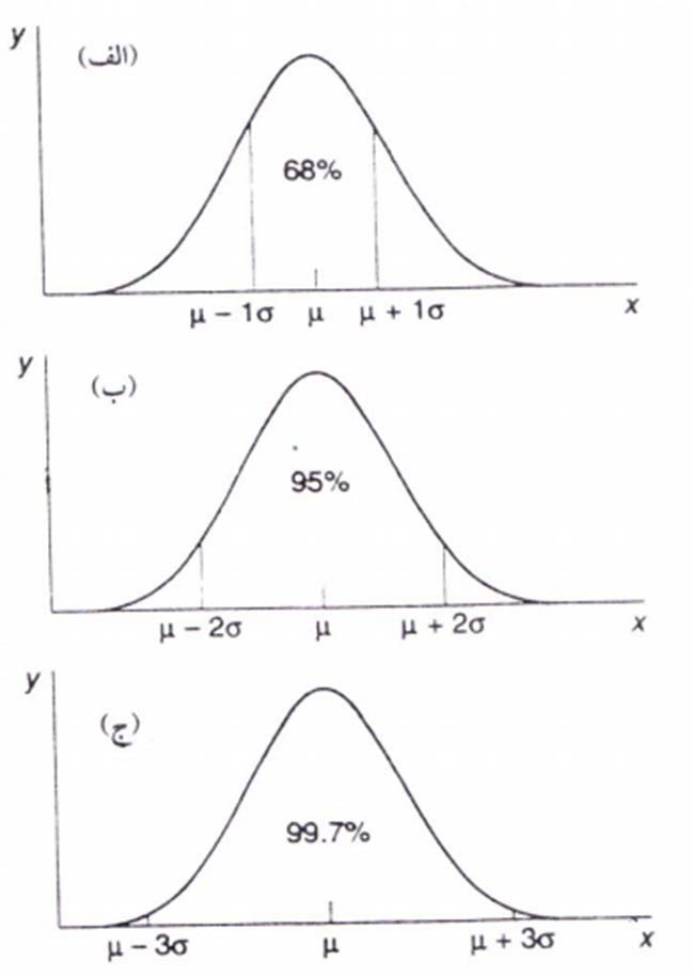
*(a) 50.00 mL. *(b) 10.0 mL.
*(c) 25.0 mL. (d) 30.0 mL.

خطای تصادفی (راندوم / نامعین): این نوع خطاها منبع و علت مشخصی ندارند و عواملی باعث ایجاد این خطا می‌شوند که در کنترل ما نیستند، پس نه قابل اندازه‌گیری و نه قابل حذف شدن هستند پس تنها می‌توانیم مقدار خطای راندوم را تخمین بزنیم. برای این‌کار از منحنی خطای نرمال استفاده می‌کنیم.



The equation for a normalized Gaussian curve has the form

$$y = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$



نتیجه‌گیری از منحنی خطای نرمال:

احتمال وقوع خطای صفر، از تمام خطاهای مثبت یا منفی بیش‌تر است.

احتمال وقوع خطای راندوم + و خطای راندوم ـ با هم برابرند.

احتمال وقوع خطاهای راندوم خیلی بزرگ، خیلی کم است.

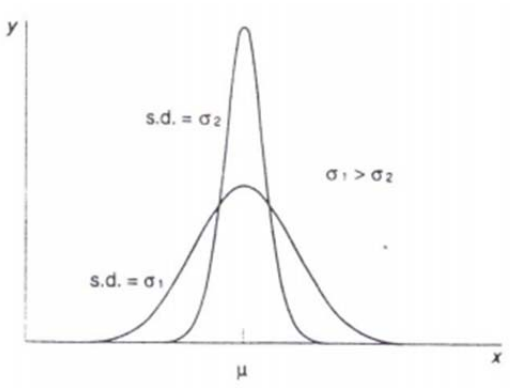
منحنی خطای نرمال را می‌توان با یکی از روش‌های بیان دقت درجه‌بندی کرد:

٪۶۸/۳ از منحنی بین ۱S+ و ۱S− است یعنی ٪۶۸/۳ از آنالیزها تنها خطای بین ۱S+ و ۱S− را دارند.

٪۹۵/۵ از منحنی بین ۲S+ و ۲S− است.

٪۹۹/۷ از منحنی بین ۳S و ۳S− است.

برطبق فرمول انحراف استاندارد، هرچه n (تعداد تکرار آزمایش‌ها) بیش‌تر شود، S (انحراف استاندارد) کم‌تر و در نتیجه خطای راندوم کم‌تر می‌شود. با افزایش تعداد تکرار آنالیزها s کوچک شده و منحنی خطای نرمال فشرده‌تر خواهد شد.



- تستهای آماری

Student t – test :۱

هدف از این تست، تعیین فاصله و حدود اطمینان است که با یک درصد اطمینان مشخص، μ در آن محدوده باشد.

درجه آزادی = N − ۱

degree of freedom

$$\mu = \bar{x} \pm \frac{tS}{\sqrt{N}}$$

N= تعداد آنالیزها

S= انحراف استاندارد

t = یک عدد آماری که از جداول با سطح اطمینان و درجه آزادی بدست می‌آید.

پیدا کردن محدوده اطمینان زمانی که σ معلوم نیست:

confidence interval

$$\text{CI for } \mu = \bar{x} \pm \frac{ts}{\sqrt{N}}$$

پیدا کردن محدوده اطمینان زمانی که σ معلوماست و یا S تقریب خوبی از σ است:

$$\text{CI for } \mu = \bar{x} \pm \frac{z\sigma}{\sqrt{N}}$$

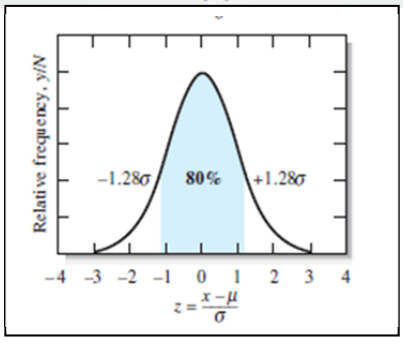
EXAMPLE 7-1

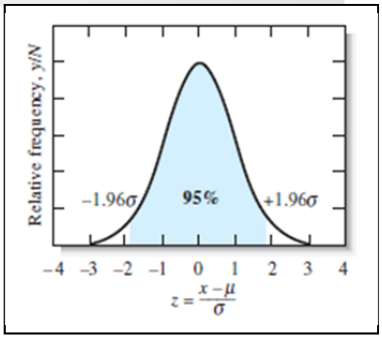
Determine the 80% and 95% confidence intervals for (a) the first entry (1108 mg/L glucose) in Example 6-2 (page 107) and (b) the mean value (1100.3 mg/L) for month 1 in the same example. Assume that in each part, $s = 19$ is a good estimate of σ.

(b) For the seven measurements, N = 7

$$80\%\ \text{CI} = 1100.3 \pm \frac{1.28 \times 19}{\sqrt{7}} = 1100.3 \pm 9.2\ \text{mg/L}$$

$$95\%\ \text{CI} = 1100.3 \pm \frac{1.96 \times 19}{\sqrt{7}} = 1100.3 \pm 14.1\ \text{mg/L}$$





EXAMPLE 7-3

A clinical chemist obtained the following data for the alcohol content of a sample of blood: % C_2H_5OH: 0.084, 0.089, and 0.079. Calculate the 95% confidence interval for the mean assuming that (a) the three results obtained are the only indication of the precision of the method and that (b), from previous experience on hundreds of samples, we know that the standard deviation of the method $s = 0.005\%\ C_2H_5OH$ and is a good estimate of σ.

Solution

(a) $\sum x_i = 0.084 + 0.089 + 0.079 = 0.252$

$$\sum x_i^2 = 0.007056 + 0.007921 + 0.006241 = 0.021218$$

$$s = \sqrt{\frac{0.021218 - (0.252)^2/3}{3 - 1}} = 0.0050\%\ C_2H_5OH$$

In this instance, $\bar{x} = 0.252/3 = 0.084$. Table 7-3 indicates that $t = 4.30$ for two degrees of freedom and the 95% confidence level. Thus, using Equation 7-5,

$$95\%\ \text{CI} = \bar{x} \pm \frac{ts}{\sqrt{N}} = 0.084 \pm \frac{4.30 \times 0.0050}{\sqrt{3}}$$

$$= 0.084 \pm 0.012\%\ C_2H_5OH$$

(b) Because $s = 0.0050\%$ is a good estimate of σ, we can use z and Equation 7-2

$$95\%\ \text{CI} = \bar{x} \pm \frac{z\sigma}{\sqrt{N}} = 0.084 \pm \frac{1.96 \times 0.0050}{\sqrt{3}}$$

$$= 0.084 \pm 0.006\%\ C_2H_5OH$$

Note that a sure knowledge of σ decreases the confidence interval by a significant amount even though s and σ are identical.

TABLE 7-1

Confidence Levels for Various Values of z

Confidence Level, %	z
50	0.67
68	1.00
80	1.28
90	1.64
95	1.96
95.4	2.00
99	2.58
99.7	3.00
99.9	3.29

TABLE 7-2

Size of Confidence Interval as a Function of the Number of Measurements Averaged

Number of Measurements Averaged	Relative Size of Confidence Interval
1	1.00
2	0.71
3	0.58
4	0.50
5	0.45
6	0.41
10	0.32

پیدا کردن حدود اطمینان
مقایسه میانگین با مقدار واقعی آزمون T
مقایسه دو میانگین
مقایسه دقت دو روش آزمون F
آشکارسازی خطای فاحش، داده های پرت آزمون Q

2- حذف نتایج مشکوک: Q-test

می خواهیم بدانیم خطایی که باعث اختلاف نتایج شده ناشی از خطای سیستماتیک و یا فاحش است و باید آن را حذف کنیم و یا ناشی از خطای راندوم است و باید آن را نگه داریم.

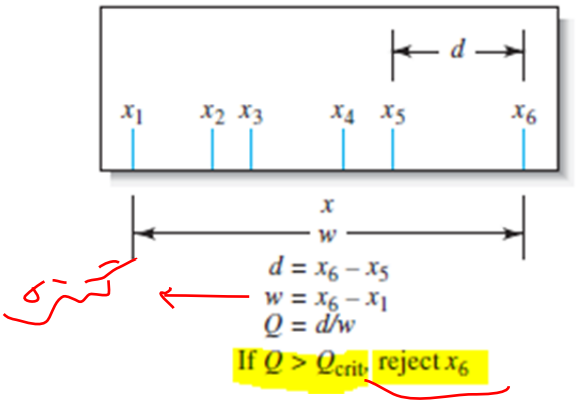


Figure 7-6 The Q test for outliers.

$$Q = \frac{|x_q - x_n|}{w} \qquad (7\text{-}17)$$

TABLE 7-5

Critical Values for the Rejection Quotient, Q^*

Number of Observations	Q_{crit} (Reject if $Q > Q_{crit}$)		
	90% Confidence	95% Confidence	99% Confidence
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

EXAMPLE 7-11

The analysis of a city drinking water for arsenic yielded values of 5.60. 5.64, 5.70, 5.69, and 5.81 ppm. The last value appears anomalous; should it be rejected at the 95% confidence level?

Solution

The difference between 5.81 and 5.70 is 0.11 ppm. The spread (5.81 − 5.60) is 0.21 ppm. Thus,

$$Q = \frac{0.11}{0.21} = 0.52$$

For five measurements, Q_{crit} at the 95% confidence level is 0.71. Because $0.52 < 0.71$, we must retain the outlier at the 95% confidence level.

با استفاده از آزمون Q، بزرگترین عدد (یعنی X) راکه بایستی در سری زیر باقی بماند حساب کنید. (سطح اطمینان 95 درصد را در نظر بگیرید)

X ،73 ،67 ،64 ،63