Chapter 2

Errors and Data Handling in Analytical Chemistry

Lecture 1

Lectured by Shouguo Wu

第二章 误差与数据处理

Errors and data processing

误差的分类：Classification

系统误差—固定原因造成的,具有单向性 System error

方法误差 Methodic error

仪器误差 Instrumental error

试剂误差 Regent's error

操作误差 Operational error

主观误差 Subjective error

随机误差—不定原因造成的，不易察觉 Random error

过失—粗心（责任心）造成的 Fault

准确度与误差
Accuracy and errors

实验值与真实值之间相符合的程度，用误差表示。

误差的表示：**Absolute error and Relative error**

绝对误差(E)=测得值（X）– 真实值（T）

$$\text{相对误差(RE)} = \frac{\text{测得值(X)} - \text{真实值(T)}}{\text{真实值(T)}} \times 1000‰$$

绝对误差： 表示测定值与真实值之差值。

相对误差： 是指误差在真实值(结果)中所占比率。

精密度与偏差
Precision and deviations

几次平行测定结果相互接近的程度，用偏差表示。

偏差的表示
- 绝对偏差 **Absolute deviation**
- 相对偏差 **Relative deviation**
- 平均偏差 **Mean deviation**
- 相对平均偏差 **Relative mean deviation**
- 总体标准偏差 **Population standard deviation**
- 样本标准偏差 **Sample standard deviation**
- 变异系数 **Coefficient of variation**

绝对偏差：单次测定值与平均值的差值。

相对偏差：绝对偏差在平均值所占百分率或千分 率。

$$\text{绝对偏差}(d) = x - \bar{x}$$

$$\text{相对偏差}(d\%) = \frac{x - \bar{x}}{\bar{x}} \times 100\%$$

平均偏差：是指单项测定值与平均值的偏差（取绝对值）之和，除以测定次数。

$$\text{平均偏差}\ \bar{d} = \frac{\sum |d_i|}{n} = \frac{\sum |x_i - \bar{x}|}{n} \quad (i = 1, 2, \ldots, n)$$

$$\text{相对平均偏差}\ (\bar{d}\%) = (\bar{d} / \bar{x}) \times 100\%$$

例：55.51，55.50，55.46，55.49，55.51

求：$\bar{x}$，$\bar{d}$，$\bar{d}$%

解：$\bar{X}=55.49$

$$\bar{d} = \frac{\sum |x_i - \bar{x}|}{n} = 0.016$$

$$\overline{d}(\%) = (\overline{d} / \overline{x}) \times 100\%$$

$$= 0.016/55.49 = 0.028\%$$

标准偏差

测定次数趋于无穷大时

总体标准偏差 ：

$$\sigma = \sqrt{\sum (Xi - \mu)^2 / n}$$

μ 为无限多次测定 的平均值（总体平均值）；即当消除系统误差时， μ即为真实值。

$$\lim_{n \to \infty} \overline{X} = \mu$$

有限测定次数

样本标准偏差 ：

$$S = \sqrt{\sum (Xi - \overline{X})^2 / n - 1}$$

相对标准偏差 （变异系数） ：$CV\% = S / \overline{X}$

例：甲：0.3，0.2，0.4，-0.2，0.4，0.0，0.1，0.3，0.2，-0.3

乙：0.0，0.1，0.7，0.2，0.1，0.2，0.6，0.1，0.3，0.1

求：甲组和乙组的$\bar{d}$和S。

甲组：$\bar{d}_1 = \dfrac{\Sigma \mid d_i \mid}{n} = 0.24$

乙组：$\bar{d}_2 = \dfrac{\Sigma \mid d_i \mid}{n} = 0.24$

由此说明
甲组的精密度好。

甲组：$S_1 = 0.28$

乙组：$S_2 = 0.34$

测定某水样中Fe的含量,五次测量结果分别为:7.48, 7.37, 7.47, 7.43和7.40,计算其**平均偏差、相对平均偏差、标准偏差和相对标准偏差。**

解：计算结果列于下表：

x mg•L^{-1}	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$
7.48	0.05	0.0025
7.37	0.06	0.0036
7.47	0.04	0.0016
7.43	0.00	0.0000
7.40	0.03	0.0009
$\bar{x} = 7.43$	$\Sigma \lvert d_i \rvert = 0.18$	$\Sigma (d_i) = 0.0086$

计算得：$\bar{d} = 0.036$, $\bar{d}$ (‰) = 4.8

$s = 0.046$, CV (‰) = 6.2

准确度与精密度的关系

Relations between accuracy and precision

例：现有三组各分析四次结果的数据如表所示

(真实值=0.31)

	I	II	III	IV	平均值
第一组	0.20	0.20	0.18	0.17	0.19
第二组	0.40	0.30	0.25	0.23	0.30
第三组	0.36	0.35	0.34	0.33	0.35

实验数据分析结果：

第一组：精密度很高，但平均值与标准样品数值相差很大，说明准确度低。

第二组：精密度不高，准确度也不高。

第三组：精密度高，准确度也高。

随机误差的正态分布

Gaussian distribution of random errors

随机误差的特点：**正负不定、大小不定**

大量测试数据，存在一个统计规律：**正态分布或称高斯分布**

$$y=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

y —概率密度；x — 测量值

μ — 总体平均值；σ — 标准偏差

$$y=\frac{n_i}{n}$$

（1）极大处在$x=\mu$, 表明大多数测量值在μ附近，即平均值能够较好地反映数据的集中趋势；

（2）$y_{max}=\frac{1}{\sigma\sqrt{2\pi}}$，概率密度的极大值取决于$\sigma$；

$N(\mu,\sigma^2)$

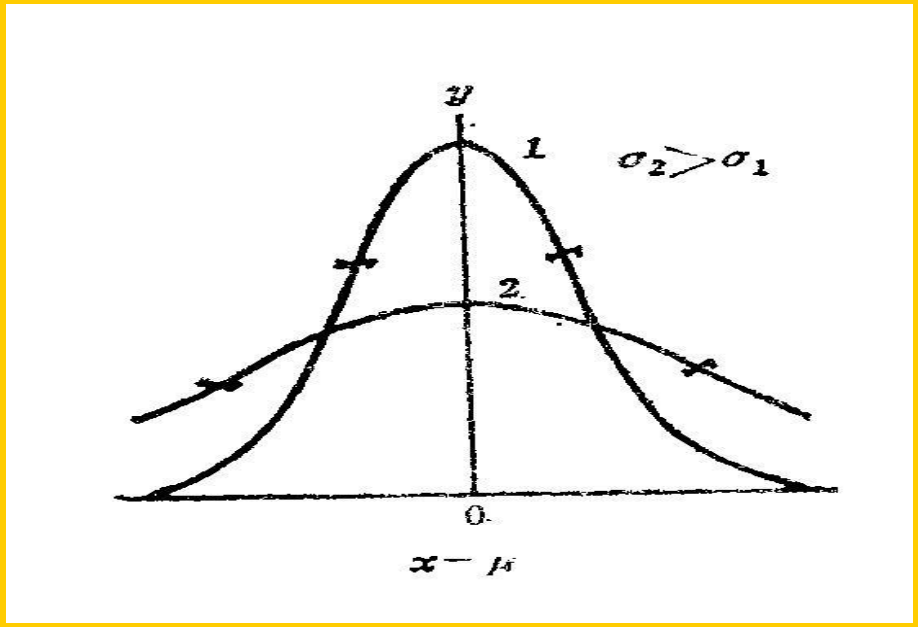
（3）以$x=\mu$成镜面对称，说明正负误差出现的概率相等

（4）$x\to\pm\infty$时， y→0,说明极大误差出现的概率为0。

正态分布（x分布）

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$N(\mu, \sigma^2)$$



标准正态分布（u分布）

$$u = \frac{x - \mu}{\sigma}$$

$$dx = \sigma du$$

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

$$f(x)dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \Phi(u)du$$

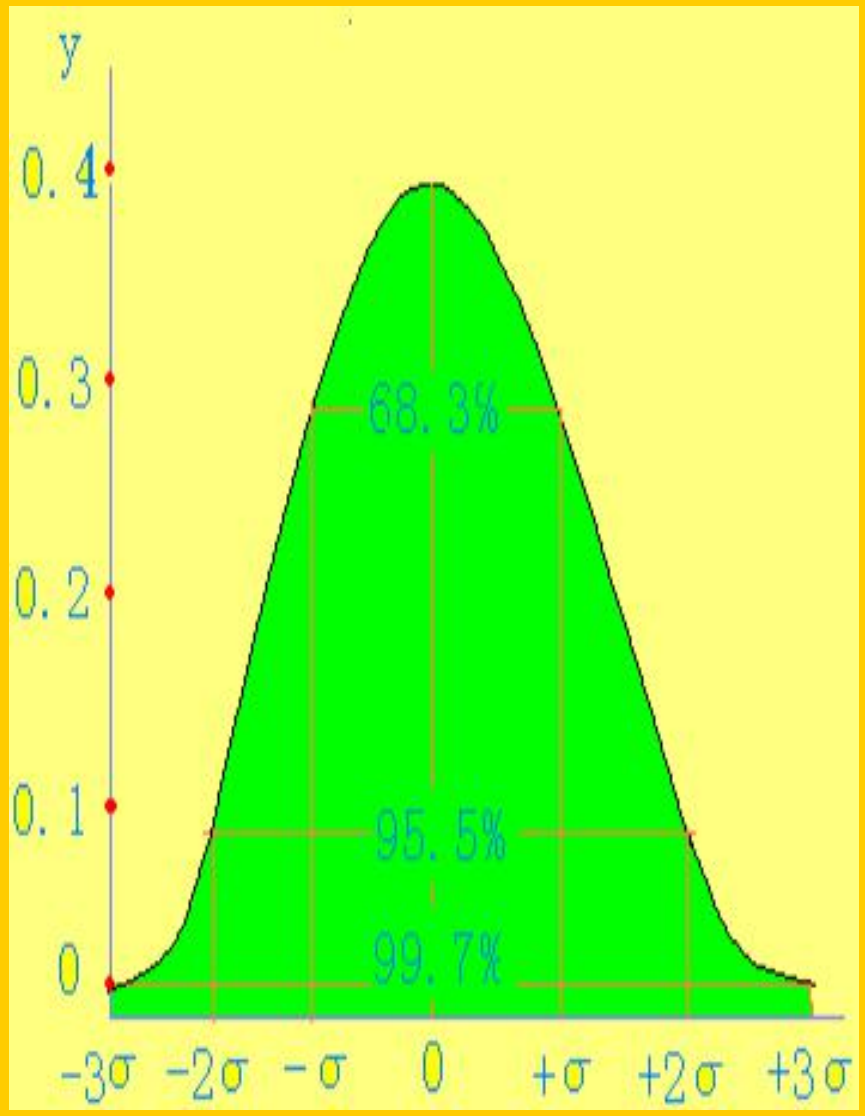
归一化的正态分布曲线

Normalized Gaussian distribution curve

标准正态分布曲线：$N(0,1)$

$$y = \Phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

$$y_{max} = \frac{1}{\sqrt{2\pi}}$$

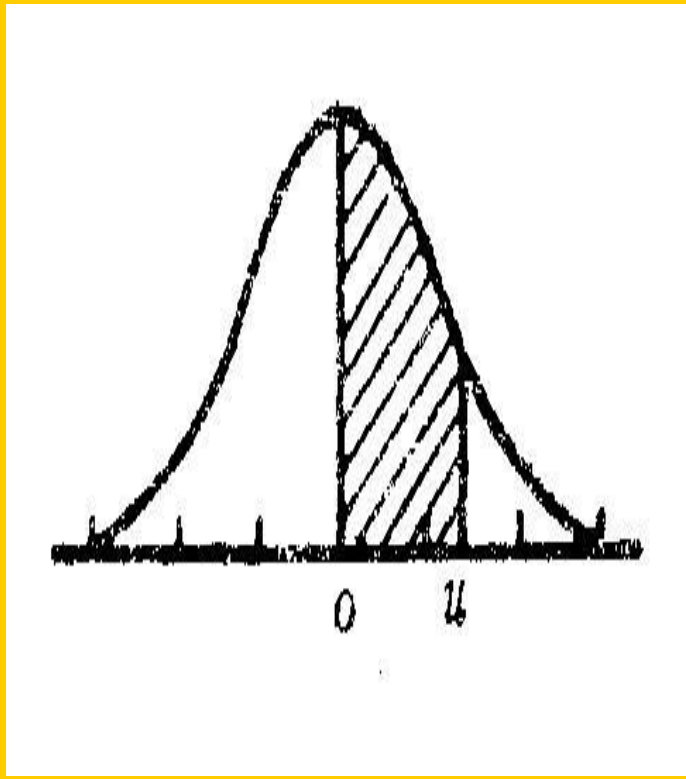


曲线的形状与 μ 和σ的大小无关

随机误差的区间概率
Interval probability of random errors

$$P=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du=1$$

$$P=\int_{-u}^{u}\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du$$



测定值X落在 ：

μ−σ ~ μ+σ　　概率 68.3%

μ−2σ ~ μ+2σ　　概率 95.5%

μ−3σ ~ μ+3σ　　概率 99.7%

例 某样品测量数据基本遵守正态分布$N[66.62，(0.21)^2]$，求测量值x落在（66.18～67.04）范围的概率。

解： 因$\mu=66.62$，$\sigma=0.21$

当$x=67.04$时，$u=\dfrac{x-\mu}{\sigma}=\dfrac{67.04-66.62}{0.21}=2.0$

当$x=67.04$时，$u=\dfrac{x-\mu}{\sigma}=\dfrac{66.18-66.62}{0.21}=-2.1$

查表得：P（2.0）=0.4773，P（−2.1）=−0.4821

所以 $P(66.18\leqslant x\leqslant 67.04)=P(2.0)-P(-2.1)$

$=0.4773+0.4821=95.94\%$

例 对矿样进行150次全铁含量分析，已知结果符合正态分布（$0.4695,0.0020^2$）。求大于0.4735的测定值可能出现的次数。

解：

$$u=\frac{|x-\mu|}{\sigma}=\frac{0.4735-0.4695}{0.0020}=2$$

查表，P=0.4773，故在150次测定中大于0.4773的测定值出现的概率为：

0.5000-0.4773=0.0227

$150\times0.0227\approx3$

答：大于0.4735的测定值可能出现3次。

少量实验数据的统计处理 Statistical handling of experimental data

平均值的可靠性

一个样本的测量，$x_1, x_2, x_3, \ldots x_n$

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + x_3 + \ldots + x_n\right)$$

标准偏差s代表测量的精密度，根据随机误差的传递，有

$$s^2_{\bar{x}} = \frac{1}{n^2}\left(s^2_{\bar{x}_1} + s^2_{\bar{x}_2} + s^2_{\bar{x}_3} + \ldots + s^2_{\bar{x}_n}\right)$$

相同条件下测量，具有相同精密度，即

$$s_{\bar{x}_1} = s_{\bar{x}_2} = s_{\bar{x}_3} = \ldots = s_{\bar{x}_n} = s \qquad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

少量实验数据的统计处理Statistical handling of experimental data

*t*分布　*t*-distribution

实际工作中，通过有限次的测定是无法得知 μ 和 σ 的，测定值或随机误差也不呈正态分布，这就给少量测定数据的统计处理带来了困难。此时若用*s*代替 σ 从而对 μ 作出估计必然会引起偏离，而且测定次数越少，偏离就越大。如果根据测定次数的多少，引入新统计量 $t_{\alpha,f}$ 来取代*u*，可以补偿这一误差。但此时测定值或随机误差将遵从*t*分布而不是*u*分布。

*t*值的定义：

$$t_{\alpha,f} = \frac{\bar{x} - \mu}{s}$$

$$s_{\bar{x}}$$

式中：$t_{\alpha,f}$ 是随置信度*P*和自由度*f*而变化的统计量。

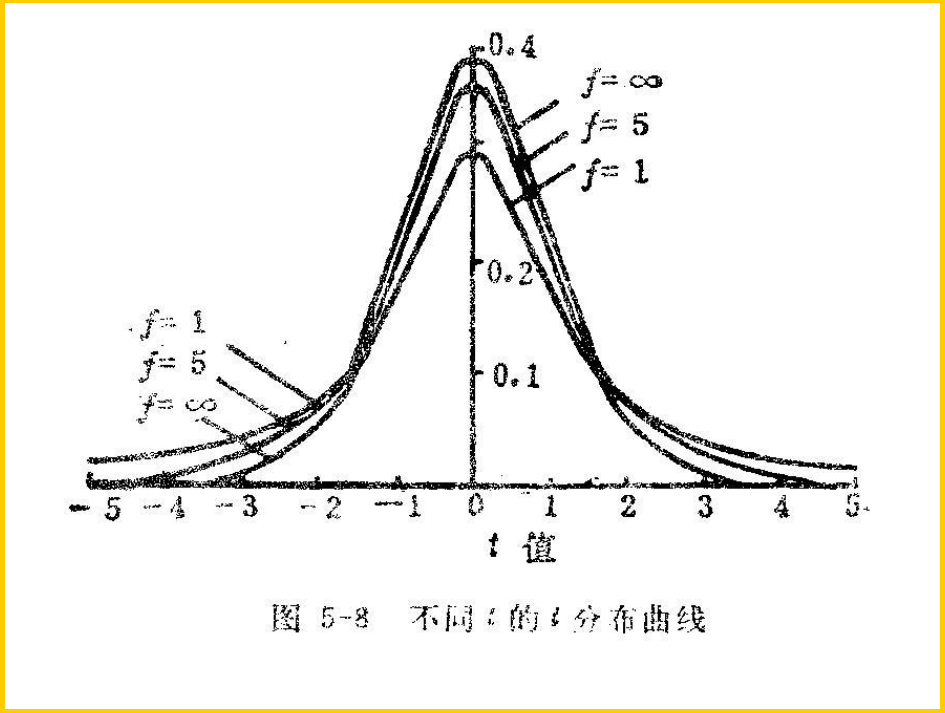


图 5-8 不同 f 的 t 分布曲线

随着测定次数增多，*t*分布曲线愈来愈尖锐，测定值的集中趋势亦更加明显。当$f \to \infty$时，*t*分布曲线就与*u*分布曲线合为一体，因此可以认为*u*分布就是*t*分布的极限。

如果用样本平均值 $\bar{x}$ 表达总体平均值 μ，这时必须用 $s_{\bar{x}}$代替s，则统计量$t_{\alpha,f}$用下式来表达。

$t_{\alpha,f}$ 表达式：

$$t_{\alpha,f} = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

式中：$t_{\alpha,f}$是随置信度P和自由度f而变化的统计量。

平均值的置信区间

Confidence interval of the mean value

$$\mu = \bar{x} \pm t_{\alpha,f} s_{\bar{x}}$$

$$= \bar{x} \pm \frac{s}{\sqrt{n}} t_{\alpha,f}$$

置信度（显著性水平 α）：真实值（总体平均值μ ）在某一范围内出现的概率。

$t_{\alpha,f}$：显著性水平为α ，自由度为n-1的 t值（临界值）

含义：在选定的置信度下，总体平均值将出现在以平均值为中心的某一范围内，此范围称为平均值的置信区间

例 分析纯明矾中Al的含量（%）为：10.74，10.77，10.77，10.77，10.81，10.82，10.73，10.86，10.81。已知纯明矾中Al量的真实值为10.77%，试加以统计处理并报结果。

解：

$$\bar{x}=\frac{\sum x_i}{n}=10.79 \qquad s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=0.042$$

查表，$t_{\alpha,f}=t_{0.95,\,8}=2.31$

$$\mu=\bar{x}\pm t_{\alpha,f}s_{\bar{x}}=\bar{x}\pm\frac{s}{\sqrt{n}}t_{\alpha,f}$$

$$=10.79\pm\frac{0.042}{\sqrt{9}}\times 2.31=10.79\pm 0.04(\%)$$

答：明矾中Al量理论值为10.77%在平均值的置信区间内，说明该测定方法准确，测定值与理论值的差别是由随机误差引起的。

影响置信区间的因素
Influencing factor of the confidence interval

- 显著性水平α越大，置信区间越大（t值增大）
- 测量次数n越多，置信区间越小（平均值更接近真值，当n>20时，t值变化不大，置信区间也变化不大）
- 试验的偶然误差越小，s值越小，置信区间减小

Home works

Pages 56 and 57 in textbook

Questions: 3,4,6,9,10