第7章　分析化学中的数据处理

教学目的：用数理统计的方法处理实验数据，将会更好地表达结果，既能显示出测量的精密度，又能表达出结果的准确度；介绍显著性检验的方法，用于检验样本值与标准值的比较、两个平均值的比较和可疑值的取舍。

教学重点：总体平均值的估计；t 检验法

教学难点：对随机变量正态分布的理解；各种检验法的正确使用，双侧和单侧检验如何查表。

1. 总体与样本

总体：在统计学中，对于所考察的对象的全体，称为总体（或母体）。

个体：组成总体的每个单元。

样本（子样）：自总体中随机抽取的一组测量值（自总体中随机抽取的一部分个体）。

样本容量：样品中所包含个体的数目，用 n 表示。例题：

分析延河水总硬度，依照取样规则，从延河取来供分析用 2000ml 样品水，这 2000ml 样品水是供分析用的**总体**，如果从样品水中取出 20 个试样进行平行分析，得到 20 个分析结果，则这组分析结果就是延河样品水的一个**随机样本**，**样本容量**为 20。

2. 随机变量　来自同一总体的无限多个测量值都是随机出现的，叫随机变量。

$\bar{x}=\frac{1}{n}\sum x_i$，$\mu=\lim_{n\to\infty}\frac{1}{n}\sum x_i$（总体平均值），$\delta=\frac{\sum|x-\mu|}{n}$（单次测量的平均偏差）

7.1　标准偏差

7.1.1 总体标准偏差（无限次测量）

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$　n－测量次数

7.1.2 样本标准偏差（有限次测量）

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$　（n－1）－自由度

7.1.3 相对标准偏差

相对标准偏差（变异系数）$CV=\frac{s}{\bar{x}}\times100\%$

相对平均偏差 $=\frac{\bar{d}}{\bar{x}}\times100\%$

7.1.4 标准偏差与平均偏差

当测定次数非常多（n 大于 20）时，$\delta=0.797\sigma\approx0.8\sigma$，但是 $\bar{d}\neq00.8S$

7.1.5 平均值的标准偏差

统计学可证明 平均值的标准偏差与单次测量结果的标准偏差存在下列关系：

$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$，$\delta_{\bar{x}}=\frac{\delta}{\sqrt{n}}$（无限次测量）

$s_{\bar{x}}=\frac{s}{\sqrt{n}}$，$\bar{d}_{\bar{x}}=\frac{\bar{d}}{\sqrt{n}}$（有限次测量）

增加测定次数，可使平均值的标准偏差减少，但测定次数增加到一定程度时，这种减少作用不明显，因此在实际工作中，一般平行测定 3-4 次即可；当要求较高时，可适当增加平行测量次数

<例>

7.2 随机误差的正态分布

7.2.1 频数分布

频数：每组中数据的个数。

相对频数：频数在总测定次数中所占的分数。

频数分布直方图：以各组分区间为底，相对频数为高做成的一排矩形。

特点：

1. **离散特性**：测定值在平均值周围波动。波动的程度用总体标准偏差σ表示。
2. **集中趋势**：向平均值集中。用总体平均值μ表示。在确认消除了系统误差的前提下，总体平均值就是真值。

7.2.2 正态分布（无限次测量）

1．正态分布曲线：如果以 x-μ（随机误差）为横坐标，曲线最高点横坐标为 0，这时表示的是随机误差的正态分布曲线。 $y=f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$， 记为：N（μ，$\sigma^2$），

μ一决定曲线在 X 轴的位置

σ一决定曲线的形状，σ小→曲线高、陡峭，精密度好；σ→曲线低、平坦，精密度差。

随机误差符合正态分布：(1) 大误差出现的几率小，小误差出现的几率大；

（2） 绝对值相等的正负误差出现的几率相等；

（3） 误差为零的测量值出现的几率最大。

（4） x=μ时的概率密度为 $y_{x=\mu}=\dfrac{1}{\sigma\sqrt{2\pi}}$

2．标准正态分布 N（0，1）

令 $u=\dfrac{x-\mu}{\sigma}$， $y=f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{u^2}{2}} \Rightarrow y=\Phi(u)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$

7.2.3 随机误差的区间概率

所有测量值出现的概率总和应为 1，即 $P(-\infty,+\infty)=\dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{u^2}{2\sigma^2}}dx=1$

求变量在某区间出现的概率， $P(a,b)=\dfrac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{u^2}{2\sigma^2}}dx$

概率积分表，p248。注意：表中列出的是单侧概率，求±u 间的概率，需乘以 2。

随机误差出现的区间	测量值出现的区间	概率
u＝±1	x＝μ±1σ	0.3413×2＝68.26％
u＝±2	x＝μ±2σ	0.4773×2＝95.46％
u＝±3	x＝μ±3σ	0.4987×2＝99.74％

结论：

1.随机误差超过 3σ的测量值出现的概率仅占 0.3%。

2.当实际工作中，如果重复测量中，个别数据误差的绝对值大于 3σ，则这些测量值可舍去。

<例>例：已知某试样中 Fe 的标准值为 3.78%，σ=0.10，又已知测量时没有系统误差，求 1）分析结果落在（3.78±0.20）%范围内的概率；2）分析结果大于 4.0%的概率。

解：1） $|u|=\dfrac{|x-u|}{\sigma}=\dfrac{0.20}{0.10}=2.0$ 查表，求得概率为 2*0.4773=0.9546 =95.46%

2）分析结果大于 4.0%的概率， $|u|=\dfrac{|x-u|}{\sigma}=\dfrac{4.00-3.78}{0.10}=2.2$，查表求得分析结果落在 3.78-4.00%以内的概率为 0.4861，那么分析结果大于 4.00%的概率为 0.5000-0.4861=1.39%

7.3　少量数据的统计处理

7.3.1 t 分布曲线（有限次测量中随机误差服从 t 分布）

有限次测量，用 S 代替σ，用 t 代替 u

$$t=\frac{\bar{x}-\mu}{s_{\bar{x}}}=\frac{\bar{x}-\mu}{s}\sqrt{n}$$

置信度（P）：表示的是测定值落在 $\mu \pm tS_{\bar{x}}$ 范围内的概率，当 f→∞，t 即为 u

显著性水平（α）=1-P：表示测定值落在 $\mu \pm tS_{\bar{x}}$ 范围之外的概率。

t 值与置信度及自由度有关，一般表示为 $t_{\alpha,f}$ ，见 p250，表 7－3（双侧表）

7.3.2 平均值的置信区间　　$\mu=\bar{x}\pm t\frac{S}{\sqrt{n}}$

意义：表示在一定的置信度下，以<u>平均值</u>为中心，包括<u>总体平均值μ</u>的范围。

从公式可知只要选定置信度 P，根据 P（或α）与 f 即可从表中查出 $t_{\alpha,f}$ 值，从测定的 $\bar{x}$ ，s，n 值就可以求出相应的置信区间。

<例>分析某固体废物中铁含量得如下结果：$\bar{x}$ =15.78%，s=0.03%，n=4，求

1）置信度为 95%时平均值的置信区间；2）置信度为 99%时平均值的置信区间解：置信度为 95%，查表得 $t_{0.05,3}=3.18$，那么 $\mu=\bar{x}\pm t\frac{S}{\sqrt{n}}=15.78\pm3.18\times\frac{0.03}{\sqrt{4}}=15.78\pm0.05\%$

置信度为 99%，查表得 $t_{0.05,3}=5.84$，那么 $\mu=\bar{x}\pm t\frac{S}{\sqrt{n}}=15.78\pm5.84\times\frac{0.03}{\sqrt{4}}=15.78\pm0.09\%$

对上例结果的理解:

1. **正确的理解**：在 15.78±0.05%的区间内，包括总体平均值的μ的概率为 95%。
2. **错误的理解**：a. 未来测定的实验平均值有 95%落入 15.78±0.05%区间内
   b. 真值落在 15.78±0.05%区间内的概率为 95%
   从该例可以看出，置信度越高，置信区间越大。

例 1 下列有关置信区间的定义中，正确的是：

a. 以真值为中心的某一区间包括测定结果的平均值的几率；
√b. 在一定置信度时，以测量值的平均值为中心的包括总体平均值的范围
c. 真值落在某一可靠区间的几率；d. 在一定置信度时，以真值为中心的可靠范围。

例 2 某试样含 Cl^-的质量分数的平均值的置信区间为 36.45%±0.10%（置信区间 90%），对此结果应理解为：

a. 有 90%的测量结果落在 36.45%±0.10%范围内；b. 总体平均值μ落在此区间的概率为 90%；
c. 若再作一次测定，落在此区间的概率为 90%；√d. 在此区间内，包括总体平均值μ的把握为 90%

7.3.3 显著性检验

判断是否存在系统误差。

1。t **检验**:不知道σ，检验 $\bar{x}$与μ，$\overline{x_1}$与$\overline{x_2}$

（1）比较平均值与标准值，统计量$t=\frac{|\overline{x}-\mu|}{S}\sqrt{n}$ （s＝s小）　t>t表，有显著差异，否则无。

（2）比较$\overline{x_1}$与$\overline{x_2}$　统计量　$t=\frac{|\overline{x_1}-\overline{x_2}|}{\overline{S}}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$　$\overline{S}^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

2．F检验：比较精密度，即方差S_1和S_2，F表为单侧表

统计量　$F=\frac{s_{大}^2}{s_{小}^2}$　F>F表，有显著差异，否则无。

<例>一碱灰试样，用两种方法测得其中Na_2CO_3 结果如下 方法 1：

$\overline{x_1}=42.34$，$s_1=0.10$，$n_1=5$　方法 2：$\overline{x_2}=42.44$，$s_2=0.12$，$n_2=4$

解：先用 F 检验 s_1 与 s_2 有无显著差异：$F_{计算}=\frac{s_{大}^2}{s_{小}^2}=\frac{(0.12)^2}{(0.10)^2}=1.44$

查表 7-4，得 F 表=6.59，因 F 计算< F 表，因此 s_1 与 s_2 无显著差异用 t 检验法检验$\overline{x_1}$与$\overline{x_2}$

$$t_{计算}=\frac{|\overline{x_1}-\overline{x_2}|}{s}\sqrt{\frac{n_1 n_2}{n_1+n_2}}\ (s=s_{小})=\frac{|42.34-42.44|}{0.10}\sqrt{\frac{5\times 4}{5+4}}=1.49$$

查表 7-3，f=5+4-2=7，P=95%，得：t 表=2.36　，则 t 计算< t 表，因此，无显著差异。

7.3.4 异常值的取舍

1. $4\overline{d}$ 法（简单，但误差大）

依据：随机误差超过 3σ的测量值出现的概率是很小的，仅占 0.3%。δ＝0.80σ，3σ≈4δ。偏差超过 4δ的个别测定值可以舍去。

方法：a. 求出$\overline{x}$与平均偏差$\overline{d}$。$|x-\overline{x}|>4\overline{d}$，则测定值 x 可以舍去。

2.格鲁布斯（Grubbs）法

步骤：（1）数据由小到大排列，求出$\overline{x}$与s。x_1，x_2……x_n

(2) 统计量T　$T=\frac{\overline{x}-x_1}{s}$ (x_1为可疑值)　$T=\frac{x_n-\overline{x}}{s}$ （x_n为可疑值）

(3)将 T 与表值 $T_{a,n}$ 比较，$T>T_{a,n}$，舍去。

3．Q 检验法

步骤：（1）数据由小到大排列。

（2）计算统计量$Q=\frac{x_n-x_{n-1}}{x_n-x_1}$（$x_n$为可疑值）$Q=\frac{x_2-x_1}{x_n-x_1}$（$x_1$为可疑值）（$Q_{计算}=\frac{|x_{可疑}-x_{邻近}|}{x_{max}-x_{min}}$）

（3）比较 Q 计算和 Q 表（$Q_{P,n}$），若 Q 计算>Q 表，舍去，反之保留。 <例 10、11>

分别用三种检验法来判断 1.40 这个数据是否应该保留。

7.4 误差的传递

分析结果通常是经过一系列测量步骤之后获得的，其中每一步骤的测量误差都会反映到分析结果中去。设分析结果 Y 由测量值 A、B、C 计算获得，测量值的系统误差分别为 DA、DB、DC，标准偏差分别为 s_A、s_B、s_C。k_i 为常数。

7.4.1 系统误差的传递

1. 加减法

① $Y=k+k_aA+k_bB+k_cC$,
$\Delta Y=k_a\Delta A+k_b\Delta B+k_c\Delta C$

2. 乘除法

② $Y=m\dfrac{AB}{C}$, $\dfrac{\Delta Y}{Y}=\dfrac{\Delta A}{A}+\dfrac{\Delta B}{B}-\dfrac{\Delta C}{C}$

3. 指数关系

③ $Y=mA^n$, $\dfrac{\Delta Y}{Y}=n\dfrac{\Delta A}{A}$

4. 对数关系

④ $Y=m\lg A$, $\Delta Y=0.434m\dfrac{\Delta A}{A}$

7.4.2 随机误差的传递

1. 加减法

① $Y=k+k_aA+k_bB-k_cC$,
$s_Y^2=k_a^2s_A^2+k_b^2s_B^2+k_c^2s_C^2$

2. 乘除法

② $Y=m\dfrac{AB}{C}$, $\dfrac{s_Y^2}{Y^2}=\dfrac{s_A^2}{A^2}+\dfrac{s_B^2}{B^2}+\dfrac{s_C^2}{C^2}$

3. 指数关系

③ $Y=mA^n$, $\dfrac{s_Y^2}{Y^2}=n^2\dfrac{s_A^2}{A^2}$

4. 对数关系

④ $Y=m\lg A$, $s_Y=0.434m\dfrac{s_A}{A}$

7.4.3 极值误差

① $Y=k+k_aA+k_bB-k_cC$,
$\varepsilon_{Y_{\max}}=|k_a\varepsilon_A|+|k_b\varepsilon_B|+|k_c\varepsilon_C|$

② $Y=m\dfrac{AB}{C}$, $\dfrac{\varepsilon_Y}{Y}=\left|\dfrac{\varepsilon_A}{A}\right|+\left|\dfrac{\varepsilon_B}{B}\right|+\left|\dfrac{\varepsilon_C}{C}\right|$

7.5 回归分析法

7.5.1 一元线性回归方程

式中 x，y 分别为 x 和 y 的平均值，a 为直线的截距，b 为直线的斜率，它们的值确定之后，

$$y_i=a+bx_i+e_i$$

$$Q=\sum_{i=1}^{n}(y_i-a-bx_i)^2$$

$$\frac{\partial Q}{\partial b}=-2\sum_{i=1}^{n}x_i(y_i-a-bx_i)=0$$

$$\frac{\partial Q}{\partial a}=-2\sum_{i=1}^{n}(y_i-a-bx_i)=0$$

$$a=\frac{\sum_{i=1}^{n}y_i-b\sum_{i=1}^{n}x_i}{n}=\bar{y}-b\bar{x}$$

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

一元线性回归方程及回归直线就定了。

2 相关系数

相关系数的定义式如下：

$$r = b\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(y_i - y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

相关系数的物理意义如下：

a. 当所有的认值都在回归线上时，r= 1。

b. 当 y 与 x 之间完全不存在线性关系时，r=0。

c. 当 r 值在 0 至 1 之间时，表示例与 x 之间存在相关关系。r 值愈接近 1，线性关系就愈好。

7.6 提高分析结果准确度的方法

1 选择合适的分析方法

(1) 根据试样的中待测组分的含量选择分析方法。高含量组分用滴定分析或重量分析法；低含量用仪器分析法。

(2) 充分考虑试样中共存组分对测定的干扰， 采用适当的掩蔽或分离方法。

(3) 对于痕量组分，分析方法的灵敏度不能满足分析的要求，可先定量富集后再进行测定。

2 减小测量误差

→称量：分析天平的称量误差为±0.0002g，为了使测量时的相对误差在0.1%以下，试样质量必须在0.2 g以上。

→滴定管读数常有±0.01 mL的误差，在一次滴定中，读数两次，可能造成±0.02 mL的误差。为使测量时的相对误差小于0.1%，消耗滴定剂的体积必须在20 mL以上，最好使体积在25 mL左右，一般在20至30mL之间。

→微量组分的光度测定中，可将称量的准确度提高约一个数量级。

3 减小随机误差

在消除系统误差的前提下，平行测定次数愈多，平均值愈接近真值。因此，增加测定次数，可以提高平均值精密度。在化学分析中，对于同一试样，通常要求平行测定2--4次。

4 消除系统误差

由于系统误差是由某种固定的原因造成的，因而找出这一原因，就可以消除系统误差的来源。有下列几种方法：

(1) 对照试验-contrast test

(2) 空白试验- blank test

(3) 校准仪器 -calibration instrument

(4) 分析结果的校正-correction result

(1) 对照试验

→与标准试样的标准结果进行对照;标准试样、管理样、合成样、加入回收法。

→与其它成熟的分析方法进行对照;国家标准分析方法或公认的经典分析方法。

→由不同分析人员，不同实验室来进行对照试验。 内检、外检。

(2) 空白试验

- 空白实验：在不加待测组分的情况下，按照试样分析同样的操作手续和条件进行实验，所测定的结果为空白值，从试样测定结果中扣除空白值，来校正分析结果。
- 消除由试剂、蒸馏水、实验器皿和环境带入的杂质引起的系统误差，但空白值不可太大。

(3) 校准仪器

仪器不准确引起的系统误差，通过校准仪器来减小其影响。例如砝码、移液管和滴定管等，在精确的分析中，必须进行校准，并在计算结果时采用校正值。

(4) 分析结果的校正

校正分析过程的方法误差，例用重量法测定试样中高含量的SiO_2，因硅酸盐沉淀不完全而使测定结果偏低，可用光度法测定滤液中少量的硅，而后将分析结果相加。

本 章 作 业

P_{268} 2，4，7　P_{269} 8，12，13　P_{270} 17，18，22　P_{271} 29