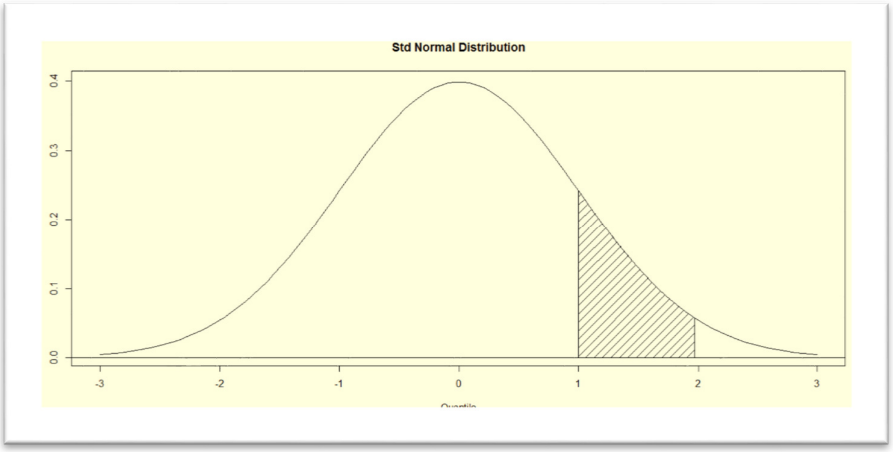
GLP Consulting

http://consultglp.com

基本统计学应用于化学分析系列（II）

适用于化学分析的概率分布

作者：杨元华



目录

2.0 总体(Population)与样本(Sample)之区别

定义： 数理统计中把研究对象的全体称为总体，构成总体的每一个对象称为个体或样本。

实验室的分析结果是来自样本，往往样本的部分是来源于一个大量的母体。许多有代表性的个体，称样本可从总体中采集，但分析的最终目标是总体（母体），既通过样本的分析和统计手段来推断总体的实况。统计的手段涉及到样本本身观察值的随机变化和样本与样本之间的数据概率分布。

图 2.1 表示总体与样本的关系

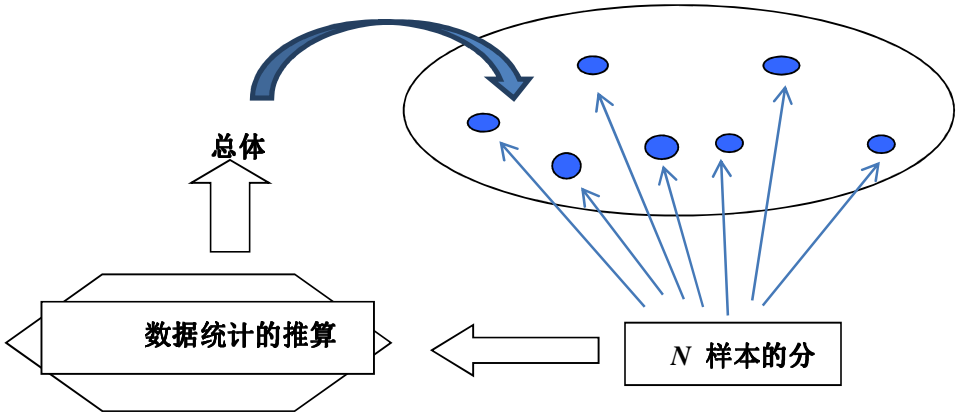


表 2.1 总体与样本统计值之比较

总体统计值	值符号	样本统计值	值符号
样品量取自总体	N	样本重复测试量	n
总体平均值	μ	样本平均值	$\bar{x}$
总体标准差	$\sigma=\sqrt{\frac{\sum x_i-\mu)^2}{N}}$	样本标准差	$s=\sqrt{\frac{\sum x_i-\bar{x})^2}{n-1}}$
总体方差	σ^2	样本方差	s^2

显然，一个样本测试所得的平均值 $\bar{x}$ 与总体的平均值是有差异的，

当有代表性样本的测试次数无限增加时，或测试无限数量的样本时，所得的平均值会逐渐接近总体平均值 μ。应用于大量测量数据的情况下，总体标准差则用 σ 表示，而 σ^2 则称总体方差。 (注：所谓有代表性的样本，是指用随机抽样方法获得的样本。)

同样地，在有限次测量中可得样本标准差 s，但当测量次数增加时，$\bar{x}$ 越来越接近 μ，此时 s 也越来越接近 σ。

2.1 平均值标准误差 （Standard error of mean）

设有 m 组随机抽出的样本经过 n 次的测试给出 m 个平均值：

$$x_{11}, x_{12}, x_{13}, \ldots\ldots x_{1n} \rightarrow \bar{x}_1; s_{x1}$$
$$x_{21}, x_{22}, x_{23}, \ldots\ldots x_{2n} \rightarrow \bar{x}_2; s_{x2}$$
$$\ldots\ldots$$
$$x_{m1}, x_{m2}, x_{m3}, \ldots\ldots x_{mn} \rightarrow \bar{x}_m; s_{xm}$$

m 个样本的总平均值 $\bar{\bar{x}}$ 为

$$\bar{\bar{x}} = \frac{1}{m}\sum \bar{x}_i^2 \qquad (2 - 1)$$

我们如何从 m 个样本测定值得总的标准差呢？

各随机样本的测定平均值会有些不同，但测定精密度可视为相同时，每个样本的方差都可作为总体离散特性的量度。

因此 m 个样本均数 $\bar{\bar{x}}$ 的变异则以标准差表示，称均值标准误差 $s_{\bar{x}}$ 。根据随机误差的传递公式的方差为

$$s_{\bar{x}}^2 = \frac{1}{m}\left(s_{x1}^2 + s_{x2}^2 + \ldots + s_{xm}^2\right)$$

式中 Sx_i^2 为每个样本的方差，在相同条件下测量同一物理量，则可认为各次测量具有相同的精密度，即

$$s_{x1} = s_{x2} = s_{x3} = \ldots = s_{xm} = s$$

于是 $$s_{\bar{x}}^2 = \frac{s^2}{m} \quad 或 \quad s_{\bar{x}} = \frac{s}{\sqrt{m}}$$

数理统计学可以证明：用 m 个样本，每个样本作 n 次测量的平均值的标准差 $s_{\bar{x}}$ 与单次测量结果的标准差 s 的关系为：

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad (2\text{-}2)$$

由此可见平均值的精密度（标准差）$s_{\bar{x}}$ 是单次测量精密度 s 的 $\frac{1}{\sqrt{n}}$；当测量次数增加时，平均值得标准差减小，这也说明平均值的精密度会随着测定次数的增加而提高。在分析化学领域里，一般平行测定 3 至 4 次即可。

对总体标准差，同样有

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (2\text{-}3)$$

在后续“中心极限定理”题目时将有更进一步的讨论

2.2 自由度（Degree of freedom）

在统计学里，自由度 (degree of freedom, v) 是指在计算某一统计量时，取值不受限制的变量个数。

例 2. 1

如一组 3 个数据：10，11，12 的平均值为 11。当均值 11 被确定后， 如知道了 3 减 1 或 2 个数的值，第三个数的值就可确定。因此这里的自由度为 3 减 1，或 2。

例 2. 2

估计样本的方差是利用离差平方和 $\sum_{i=1}^{n}(x_i-\bar{x})^2$ 的，于是只要有（*n-1*）个数的离差平方和确定了，第 *n* 个数的值就不能变了，因为它受到均值的约束条件所限制。因此样本方差的自由度为 *n-1*.

例 2.3

设对于每一个自变量 x_i 都有一个因变量 y_i。若共有 *n* 个数据，则其一元线性回归方程为 $y=a+bx$，有 *a* 为截距，*b* 为斜率。由于有 *a* 和 *b* 的两个约束条件，一元线性回归的自由度为 *n-2*。

2.3 概率分布 – 正态，矩形，三角形

检测分析是在于样本，对象却是总体。统计学利用测试的观察值来估计总体的表现。由于随机误差的正负和大小在测定中难以预料，对于总体表现的结论就会不能绝对确定，即使取得大量的数据。不过从大量的数据中可找到某种统计性规律，也称概率分布。这里讨论的是在化学分析中比较常见的正态，矩形和三角形概率分布的理论。

2.3.1 正态概率分布 (Normal probability distribution)

由于有随机误差的存在，收集的大量的化学测试数据看似杂乱无章，但若把它们加以整理就不难看出这些数据的分散是有服从某种规律的。它们的特点是：

a. 数据分散但有一定性的波动

b. 靠近中间值得数据较多而过高和低的数据出现率则较少

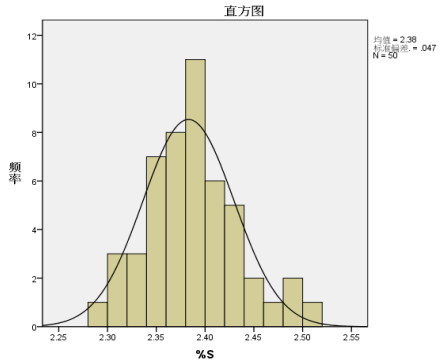
如把这些数据画出频率密度分布直方图(histogram)，可看出数据的分布规律.

例 2.4　　测定某燃油中硫磺的含量，重复测定共得 50 个数据，数据列在表 2.2 中。其频率分布直方图 如图 2.2。

表 2.2　燃油中硫磺含量的测定值*(%m/m)*

2.39	2.43	2.29	2.48	2.42	2.35	2.39	2.38	2.45	2.33
2.42	2.41	2.31	2.33	2.36	2.36	2.44	2.36	2.37	2.36
2.37	2.35	2.35	2.30	2.38	2.48	2.34	2.38	2.33	2.41
2.39	2.38	2.35	2.41	2.41	2.41	2.42	2.43	2.50	2.34
2.40	2.39	2.35	2.37	2.38	2.30	2.46	2.38	2.37	2.39

图 2.2　　50 组测量燃油中硫磺量的直方图和正态分布曲线



上述直方图 2.2 的横坐标 x 为测量值，纵坐标 y 为频率密度。y 也可以相对频率表示。这组全部数据给出平均值 2.38%及标准差 0.047%，而且测定值有明显的集中趋势，大多数测定值集中在平均值的附近。

如果测定次数更多，组分得更细，各组相对频数则趋向一个稳定值，称为概率。如图 2.2 所示，它反映了测定值随机误差分布的一般规律。这种分布特性，可用高斯分布的正态概率密度函数来表示：

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2\text{-}4)$$

式中，

y 表示概率密度（frequency density）；

x 表示样本测量值；

μ 是总体平均值 (population mean)；

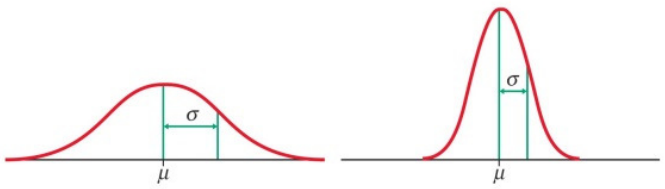
σ 为总体标准差；代表从总体平均值μ到正态分布曲线上两个拐点中任何一个的距离，表示样本值的离散特性；

e 是自然对数的底，等于 2.718。

注 1：μ 是正态分布曲线最高点的横坐标值，称为正态分布的极限平均值，在没有系统误差的情况下就是真值，它也表示样本值得集中趋势 σ 是从总体平均值 μ 到曲线拐点间的距离。 简便起见，这种正态分布曲线是以 $N(\mu,\sigma)$ 记着，（ $x-\mu$） 表示随机误差。

注 2: σ 决定曲线的形状，σ 小，数据的精密度好，曲线瘦高；σ大则表示数据分散，曲线较偏平。如图 2.3 所示。

图 2.3　两组平均值相同精密度不同的测试值的正态分布曲线



无论平均值和标准差为何值， 分布曲线和横坐标之间所夹的总面积就是概率密度函数在 $-\infty < x < \infty$ 区间的积分值，它代表具有各种大小偏差的样本值出现概率的总和，其值一定为 1。既

$$P(-\infty < x < \infty) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1 \qquad (2\text{-}5)$$

由于（2-5）式的积分计算与μ 和σ 有关，计算麻烦，为此可应用一个变量转换，即令

$$z=\frac{x-\mu}{\sigma} \qquad (2\text{-}6)$$

$$dz=\frac{dx}{\sigma} \quad 或 \quad dx=\sigma\times dz$$

因此式（2-5）被转为

$$f(x)dx=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz=\phi(z)dz$$

故，在把 z 代入式 (2-4)变成标准正态概率密度函数

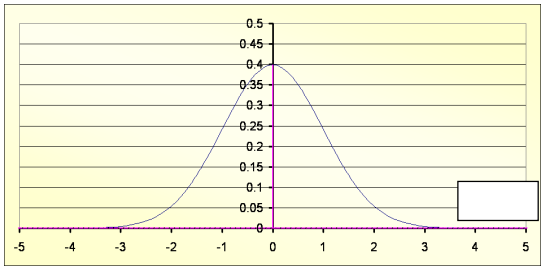
$$y=\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \qquad (2\text{-}6)$$

于是 $N(\mu,\sigma)$的正态概率密度函数变成 $N(\mu=0,\ \sigma=1)$的标准正态概率密度函数，记为 $N(0,1)$, 式（2-5）也就标准化为

$$P(-\infty<x<\infty)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{z^2}{2}}dx=1 \qquad (2\text{-}7)$$

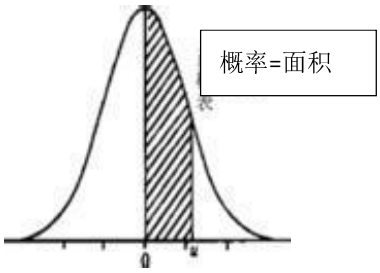
式(2-6)所得的曲线称为标准正态分布曲线，如图 2.4 所示。

图 2.4　标准正态分布曲线



如将不同 z 值对应的积分值（面积）做成表，称为正态分布概率积分表或简称 z 表。由 z 值可查表得到面积，也即某一区间的测量值或某一范围随机误差出现的概率。表 2.3 列出了$|z|$值的单侧积分表，当考虑 $\pm z$，既双侧问题时，需将表值乘 2。

表 2.3　正态概率积分表或 z 表



$\|z\|$	概率 P	$\|z\|$	概率 P	$\|z\|$	概率 P
0.00	0.0000	1.00	0.3413	2.00	0.4773
0.10	0.0398	1.10	0.3643	2.10	0.4821
0.20	0.0793	1.20	0.3849	2.20	0.4861
0.30	0.1179	1.30	0.4032	2.30	0.4893
0.40	0.1554	1.40	0.4192	2.40	0.4918
0.50	0.1915	1.50	0.4332	2.50	0.4938
0.60	0.2258	1.60	0.4452	2.60	0.4953
0.70	0.2580	1.70	0.4554	2.70	0.4965
0.80	0.2881	1.80	0.4641	2.80	0.4974
0.90	0.3159	1.90	0.4713	3.00	0.4987

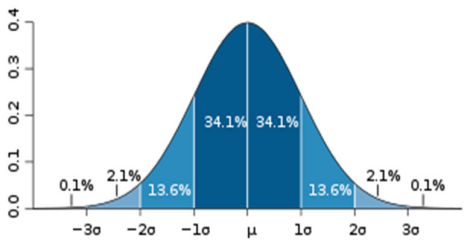
$\|z\|$	概率 P	$\|z\|$	概率 P	$\|z\|$	概率 P	$\|z\|$	概率 P
1.90	0.4713	1.94	0.4738	1.94	0.4738	1.98	0.4761
1.91	0.4719	1.95	0.4744	1.95	0.4744	1.99	0.4767
1.92	0.4726	1.96	0.4750	1.96	0.4750	2.00	0.4772
1.93	0.4732	1.97	0.4756	1.97	0.4756		

由表 2.3 可查出下列指定范围分析结果出现的概率：

分析结果 x 范围	$\|z\|$	双侧面积	概率（%）
$\mu \pm \sigma$	1	0.6826	68.26
$\mu \pm 2\sigma$	2	0.9544	95.44
$\mu \pm 1.96\sigma$	1.96	0.9500	95.00
$\mu \pm 3\sigma$	3	0.9974	99.74

微软 MS Excel ®的计算函数“=NORMDIST(z,0,1,TRUE)”也可给出单尾的 P概率。

图 2.5 双侧面积的正态概率分布曲线



注 3： 在化学测试领域里，经常采用（$\mu \pm 1.96\sigma$ ）的数据范围。重复测试数据出现在这区间的概率为 95%。

例 2.5 已知某次污水试样中总镉含量为 15.88mg/L, 测定的标准差为 0.63mg/L。设本测定中无系统误差，问

(1) 分析结果落在已确定的 $(15.50 \pm 2\text{x}0.63)$ 或（15.50 ± 1.26）mg/L 范围内的概率为多少？

(2) 若某测试人员得到一重复检测数据 16.82mg/L,问大于 16.82mg/L 的出现概率又为多少？

答：(1) 某次分析结果 x = 15.88mg/L; μ = 15.50mg/L; σ = 1.26mg/L

按 (2-6)式 $$z = \frac{x-\mu}{\sigma} = \frac{15.88-15.50}{0.63} = 0.6$$

由表 2.3 查得面积为 0.2258, 考虑到 $\pm z$,其概率为（2x0.2258）= 0.4516, 即 45.16%。

(2) 若 x = 16.80mg/L, $|x - \mu| = |16.82 - 15.50| = 1.30$

$$z = \frac{x-\mu}{\sigma} = \frac{16.82-15.50}{0.63} = 2.1$$

由于只考虑大于 16.82mg/L 的数据出现的概率，需求 $z > 2.1$ 时的概率。表 2.3 给出曲线左侧面积为 0.4821 (则 $z < 2.1$)，因此当 $z > 2.1$ 时，面积为 0.5000 – 0.4821 = 0.0179, 而这面积既代表概率 1.79%。

结论是大于 16.82mg/L 的数据出现率为 1.79%。

2.3.1.1 数据随机性与独立性的检验

一组长期累计的总体数据是否是随机性，符合正态概率分布和独立性的，可用 Anderson Darling（AD）统计法来检验。其公式为：

$$A^2 = -\frac{\sum_{i=1}^{n}(2i-1)[\ln(p_i)+\ln(1-p_{n+1-i})]}{n} - n \qquad (2\text{-}8)$$

$$A^{2^*} = A^2\left(1+\frac{0.75}{n}+\frac{2.25}{n^2}\right) \qquad (2\text{-}9)$$

式中：

n = 总数据点

p_i = 正态概率

A^2 为正态统计量，

A^{2*}为修正值。

按 s 式计算时表示 $A^{2*}(s)$，按移动极差 MR 式计算时表示 $A^{2*}(MR)$；

$|MR_i| = |x_{i+1} - x_i|$。

根据 $A^{2*}(s)$和 $A^{2*}(MR)$数值，可作如下判定（99%包含概率）：

a）$A^{2*}(s)$＜1.0 和 $A^{2*}(MR)$＜1.0，接受数据的正态性和独立性的假定；

b）$A^{2*}(s)$＞1.0 和 $A^{2*}(MR)$＞1.0，表明测量系统失控；

c）$A^{2*}(s)$＜1.0 和 $A^{2*}(MR)$＞1.0，表明系列结果呈非独立性

注 4： 计算 p_i 的概率可用微软 Excel 的函数 “$=NORMDIST(x_i, \bar{x}, s, TRUE)$”。

例 2.6 在制作某一个质量控制图时，收集了以下 20 个日常测试标准样的测定值：

序号 i	1	2	3	4	5	6	7	8	9	10
测定值	48.4	49.6	48.8	49.3	50.9	51.5	48.4	47.7	49.5	50.9
序号 i	11	12	13	14	15	16	17	18	19	20
测定值	50.8	49.4	48.8	50.3	50.9	50.4	49.0	48.7	50.3	49.7

其各自的移动极差$|MR_i|$为：

序号 i	1	2	3	4	5	6	7	8	9	10
极差移动值	-	1.2	0.8	0.5	1.6	0.6	3.1	0.7	1.8	1.4
序号 i	11	12	13	14	15	16	17	18	19	20
极差移动值	0.1	1.4	0.6	1.5	0.6	0.5	1.4	0.3	1.6	0.6

从这组数据可计算得出：

$\bar{x} = 49.67 mg/L; s = 1.05 mg/L; |\overline{MR}| = 1.07 mg/L; s_{MR} = |\overline{MR}|/1.128 = 0.95 mg/L$

$A^2(s)$=0.327; $A^{2*}(s)$=0.341; $A^2(MR)$=0.497; $A^{2*}(MR)$=0.518

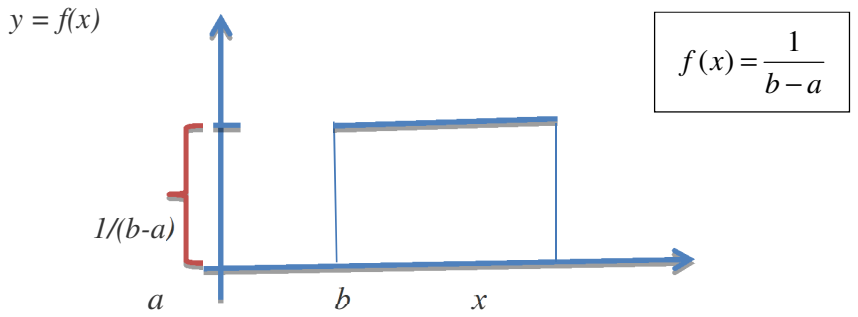
由于 $A^{2*}(s)$和 $A^{2*}(MR)$的得出值<1，则证明这组数据是随机性和独立性的。

2.3.2 矩形（均匀）概率分布 (Rectangular probability distribution)

若连续抛一个公平的硬币多次，则以等可能（概率 $p = 1/2$）出现正面或反面。若连续掷一只均质，有六个对称面的骰子多次，任何 1 至 6 点中之一点的出现概率为 1/6. 这种随机概率分布的情况表示它的概率是相等可能的。

当随机测量值(x)非常平均地散布在固定的最大值(b)和最低值(a)之间时，其概率密度函数 $f(x)$为个常数，既称矩形分布或均匀分布，如图 2.6 示。

图 2.6 矩形概率分布曲线图



注 5：概率在分布图内为 1。在 $(b-a)$ 的区间时，矩形的面积 $1 = y \times (b-a)$ 或

$$y = f(x) = \frac{1}{b-a} \qquad b \le x \le a \qquad (2\text{-}10)$$

测量期待值或平均值 $E(X)$ 的定义为：

$$E(X) = \int_{-\infty}^{+\infty} x.f(x)dx = \int_a^b x\frac{1}{b-a}dx = \frac{1}{2(b-a)}\left[x^2\right]_a^b$$

$$= \frac{b^2-a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}$$

因此， 平均值 $\mu = E(X) = \frac{b+a}{2}$ (2-11)

基于方差的定义： $V(X) = E(X^2) - [E(X)]^2$

于是

$$\sigma^2 = V(X) = \int_a^b x^2 . \frac{1}{b-a} dx - \left(\frac{b+a}{2}\right)^2 = \frac{1}{3(b-a)} \left[x^3\right]_a^b - \left(\frac{b+a}{2}\right)^2$$

$$= \frac{b^3 - a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2$$

$$= \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4}$$

$$= \frac{(b-a)^2}{12} \qquad (2\text{-}12)$$

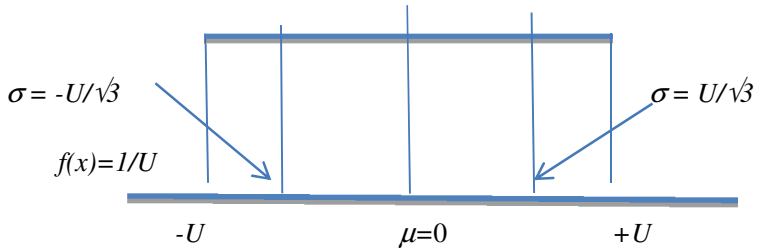
因此　　标准差　$\sigma = \frac{b-a}{2\sqrt{3}}$　　(2-13)

设 矩形分布的μ 及最大与最小值分别为 $+U$ 和 $-U$ ，则 $\mu = 0$，及

则 (2-10)式为　　$f(x) = \frac{1}{U-(-U)} = \frac{1}{2U}$

$$\sigma = \frac{U-(-U)}{2\sqrt{3}} = \frac{U}{\sqrt{3}}$$

图 2.7　标准化矩形分布曲线



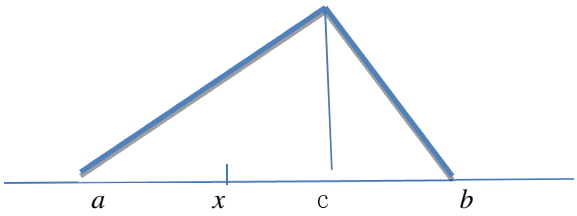
注 6：在正态概率分布下，倍成因子或称包含因子 k 等于 2 或 1.96 于 95%置信区间。在矩形概率分布下，包含因子为 **√3**。

2.3.3 三角概率分布 （Triangular probability distribution）

在应用估计不确定度的过程中，另一个常用到的概率分布为三角概率分布。在此，三角形分布是个基于低限为 a、众数为 c、上限为 b 的连续概率分布概念。

在理论上，通过三角型分布，可从中预测最大、最小及最可能的值 x，靠近最大值 b 和最小值 a 的值出现的可能性要小于靠近最可能值 c 的值。同样的，三角形里的面积，也是总概率为 1。见图 2.8。

图 2.8 三角概率分布曲线



三角概率分布函数为：

$$f(x)=\frac{2(x-a)}{(b-a)(c-a)} \quad ; \quad a \leq x \leq c \qquad (2\text{-}14a)$$

$$f(x)=\frac{2(b-x)}{(b-a)(b-c)} \quad ; \quad c \leq x \leq b \qquad (2\text{-}14b)$$

于 $x < a$ 和 $x > b$ 时，$f(x) = 0$

平均值给出 $$\mu=\frac{a+b+c}{3} \qquad (2\text{-}15)$$

方差 $$\sigma^2=\frac{a^2+b^2+c^2-ab-ac-bc}{18} \qquad (2\text{-}16)$$

当 a = -U, b = +U, $c = \mu$ =0 时，三角分布函数就标准化为：

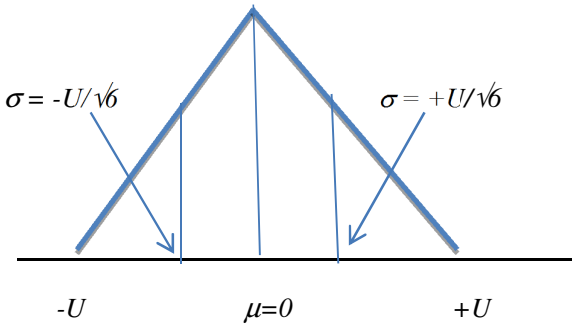
$$f(x)=\frac{2(x+U)}{(U+U)((0+U)}=\frac{(x+U)}{U^2} \quad ; \quad -U\le x\le 0 \qquad (2\text{-}17)$$

$$f(x)=\frac{2(U-x)}{(U+U)(U-0)}=\frac{(U-x)}{U^2} \quad ; \quad 0\le x\le +U \qquad (2\text{-}18)$$

$$\sigma^2=\frac{U^2+U^2-(U)(-U)}{18}=\frac{3U^2}{18}=\frac{U^2}{6} \qquad (2\text{-}19)$$

因此标准差 $$\sigma=\frac{U}{\sqrt{6}} \qquad (2\text{-}10)$$

图 2.9　标准三角形概率分布曲线



2.4　总结已知分散区间半宽 *U* 情况下的评定

若已知信息表明 X_i 之均值 $\bar{x}$ 分散区间的半宽为不确定度 *U*, 且 $\bar{x}$ 落于 $(\bar{x}-U)\sim(\bar{x}+U)$ 区间的概率 *p* 为 100%，即全部的数据落在此范围中，通过对其分布的估计，可以得出标准不确定度 $u(\bar{x})=U/k$, 因为包含因子 *k* 与分布状态有关，见表 2.4。

表 2.4　常用概率分布与 *k*,　$u(\bar{x})$ 的关系

概率分布类别	*p*(%)	*k*	标准不确定度 $u(\bar{x})$
正态	95	2 (1.96)	*U*/2
矩形（均匀）	100	√3	*U*/√3
三角	100	√6	*U*/√6

GLP Consulting

http://consultglp.com