第二章

概率与常用概率分布

第一节　随机变量的概率

一、随机事件：随机试验中的各种可能的结果称为随机事件。

研究的主要目的是通过随机试验来研究随机现象的统计规律性，研究随机事件发生的可能性大小即随机事件发生的概率问题。

二、随机事件的概率及其性质

1、频率f：如果在n次重复试验中，事件A发生了m次，则称比值 m/n是事件A在这n次试验中发生的频率，记为$f(A)$，即$f(A)=m/n$。表示某现象在样本中出现的比率，是样本特征，样本指标。

二、随机事件的概率及其性质

2、概率P:

$$n \to \infty,\ f(A) = m/n \to P(A) = p$$

定义： 概率是度量随机事件发生可能性大小的一个数值。用P表示，范围$0<P<1$，是总体特征，总体指标。

二、随机事件的概率及其性质

3、小概率事件：特指发生概率$P\leq 0.05$，

或者发生概率$P\leq 0.01$的事件。

4、小概率事件原理：把小概率事件当成

不可能事件

三、随机变量的概率分布

1、离散性随机变量：指变量的取值是有限个或可列个（如自然数列），常见的分布有二项分布和泊松分布。

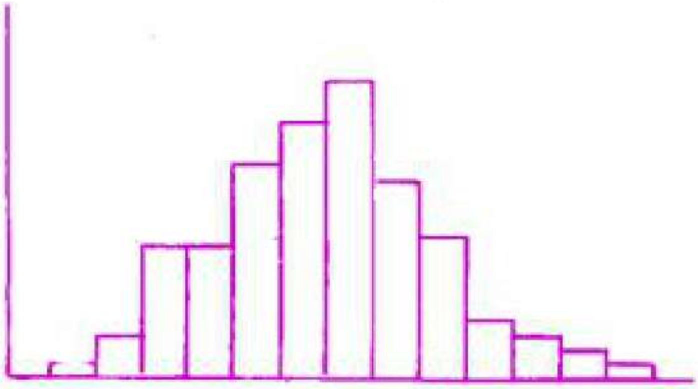
2、连续型随机变量：其取值充满某一区间，不能一一列举，常见的分布有正态分布。

第六节 正态分布

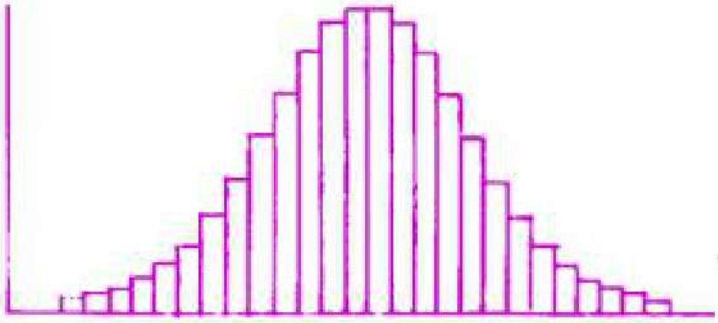
正态分布是最常见、最重要的一种连续型分布，为对称分布。

实际频数分布：中间频数多，两端频数越来越少，且左右大致对称。

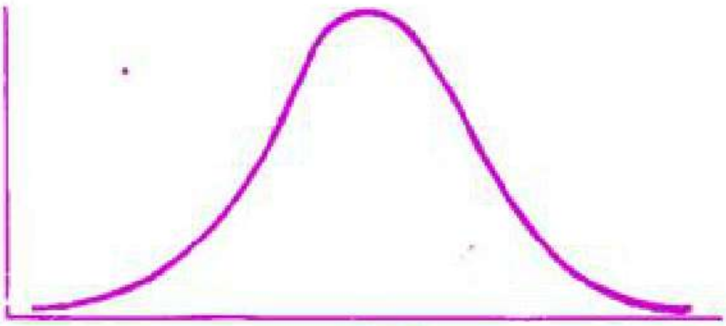
理论频数分布：正态曲线（高斯曲线）。



(1)



(2)



(3)

频数（频率）分布逐渐接近正态分布示意图

一、正态分布的概率密度函数

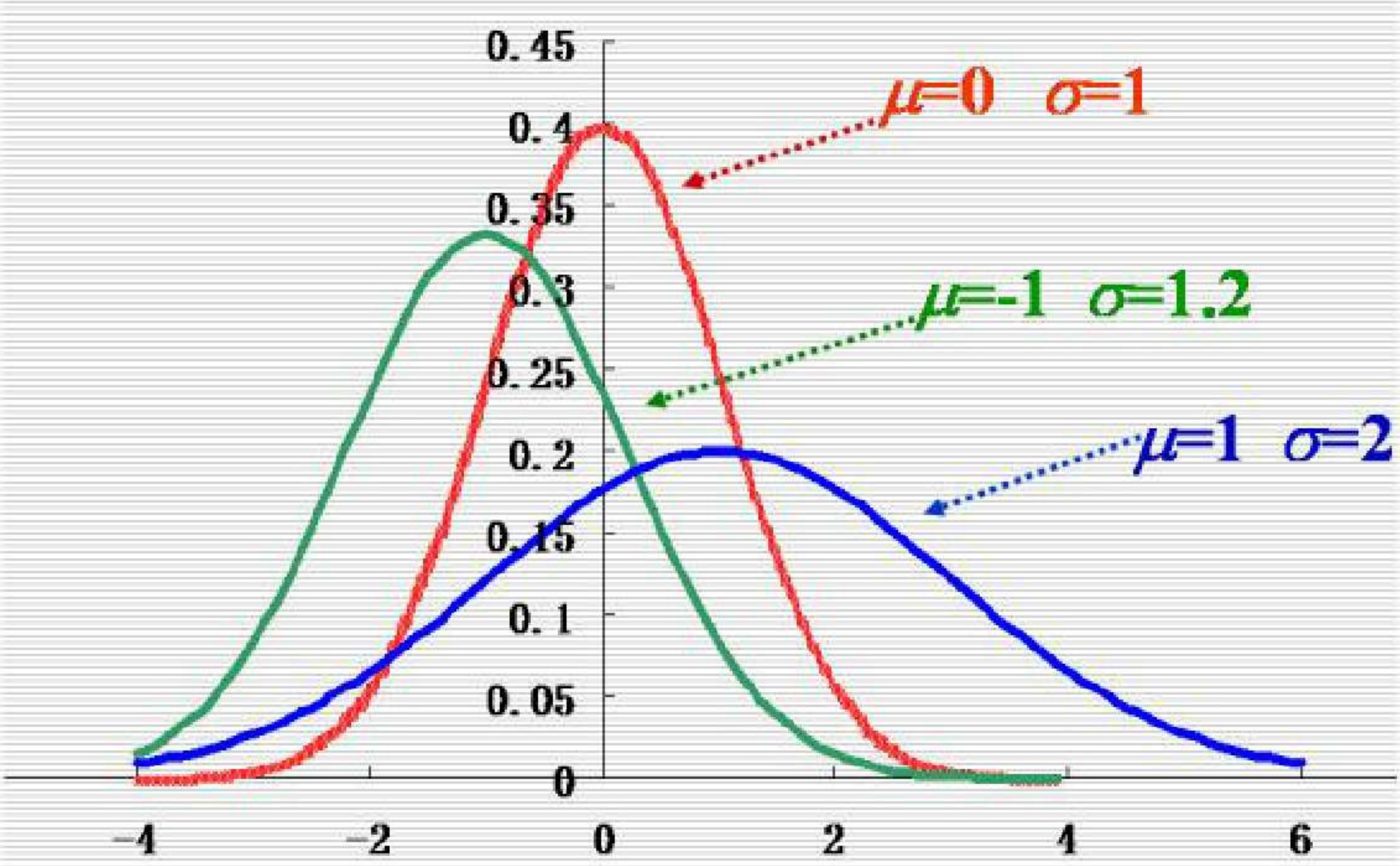
1.正态分布曲线的数学函数表达式

$$f(X)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(X-\mu)^2}{2\sigma^2}} \qquad -\infty < X < +\infty$$

则称X服从正态分布，记作$X \sim N(\mu,\sigma^2)$，μ为X的总体均数，σ^2为总体方差。

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

X	F(x)1	F(x)2	F(x)3		
-4	1.33830E-04	0.014606923	0.008764154		
-3.9	1.98656E-04	0.017927874	0.009918681	μ=0	σ =1
-3.8	2.91947E-04	0.021851583	0.01119727		
-3.7	4.24780E-04	0.026449721	0.012609115		
-3.6	6.11902E-04	0.031793866	0.014163525	μ=-1	σ=1.2
-3.5	8.72683E-04	0.03795331	0.015869833		
-3.4	0.00123222	0.044992491	0.017737304		
-3.3	0.00172257	0.052968111	0.019775029	μ=1	σ =2
...	...	...	...		
3.8	2.91947E-04	1.11525E-04	0.074863764		
4.1			0.060004526		


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4
−2
0
2
4
6
μ=0 σ=1
μ=-1 σ=1.2
μ=1 σ=2

2. 正态分布的特征

❶一个高峰:位于中央均数（μ）处

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

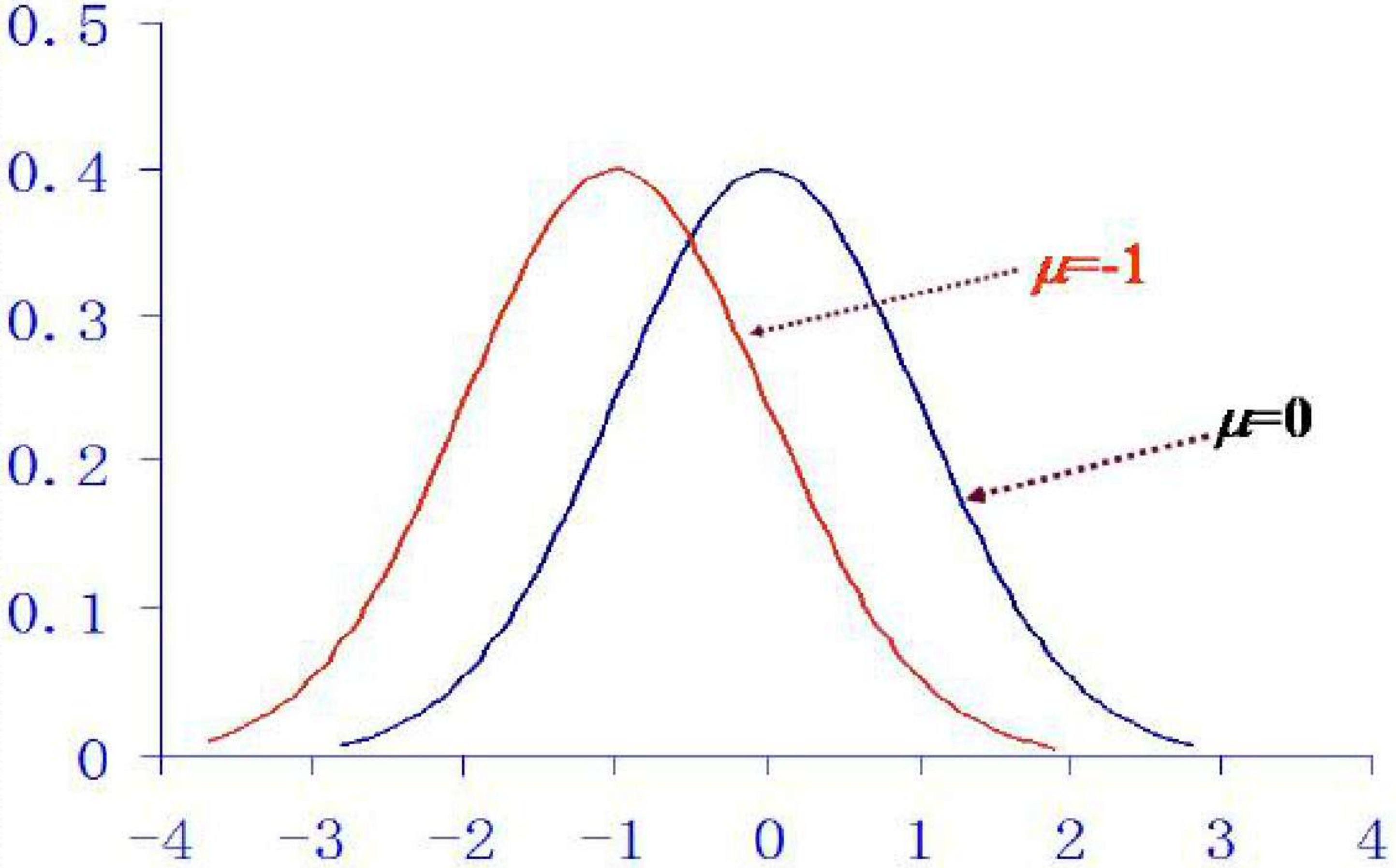


图2-5 正态分布位置变换示意图

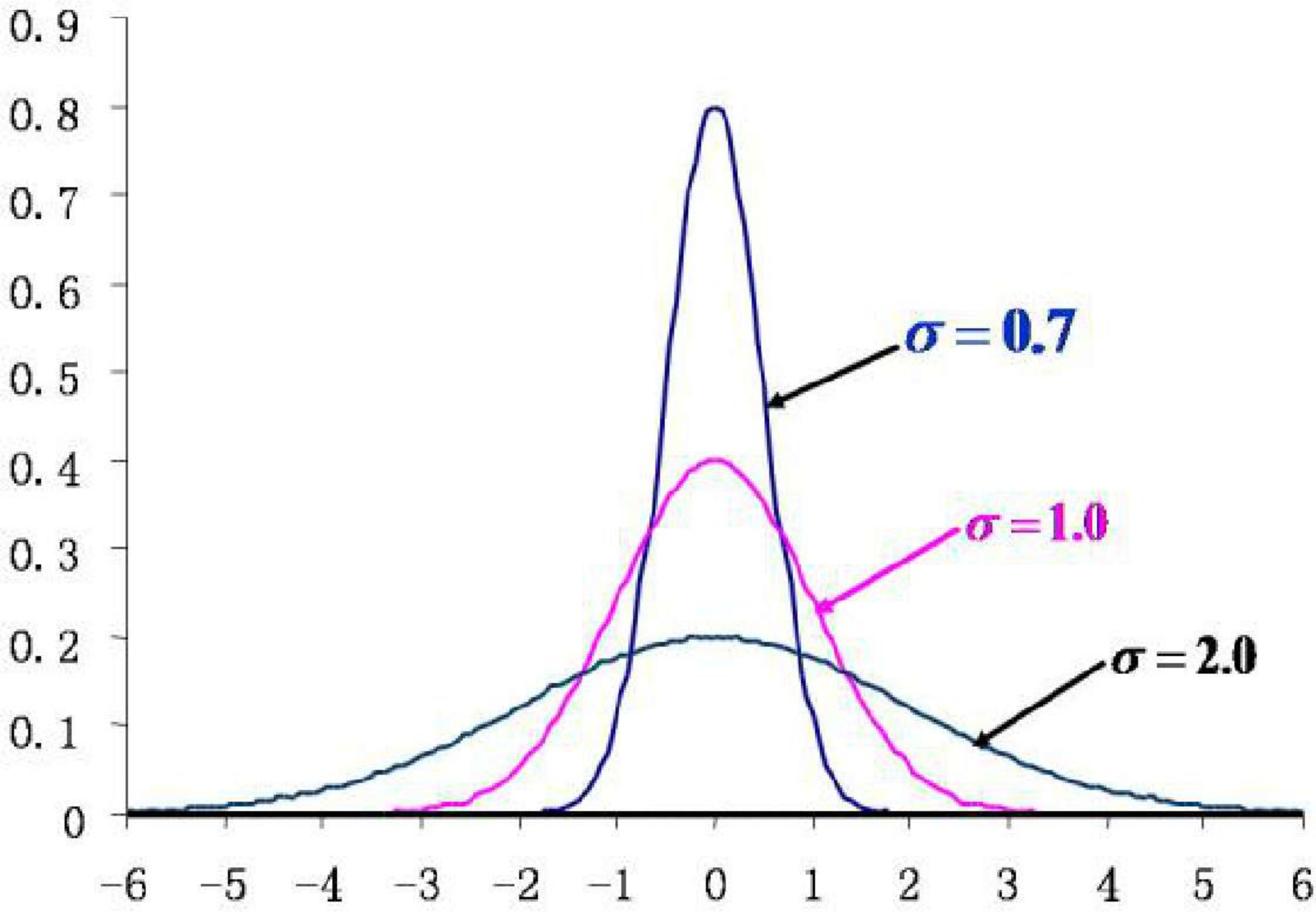


图2-6 正态分布形态变换示意图

4 正态曲线下的面积分布有一定的规律

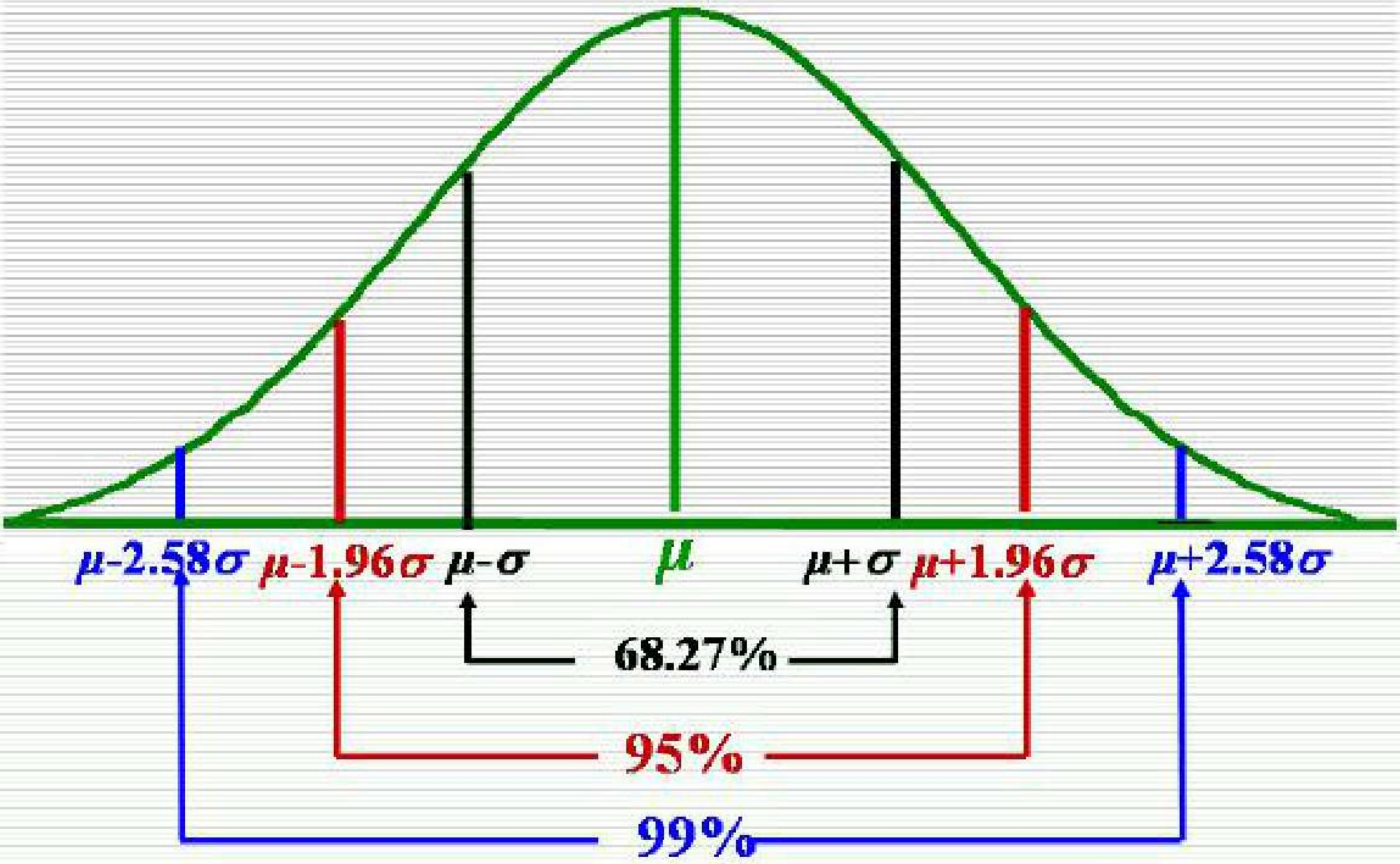
面积通过积分得来：$F(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{X}e^{\frac{-(X-\mu)^2}{2\sigma^2}}dX$

① X 轴与正态曲线所夹面积恒等于 1 或 100%；

② 区间 $\mu\pm\sigma$ 的面积为 68.27%；

区间 $\mu\pm1.96\sigma$ 的面积为 95.00%；

区间 $\mu\pm2.58\sigma$ 的面积为 99.00%。



正态分布曲线下的面积示意图

小　　结

❶正态曲线有无数条,因μ与σ的不同而不同,每条曲线均以μ为中心呈对称性分布。

❷正态曲线下的面积总和为100%或1，并且有三个常用的面积规律。

❸每条正态曲线下的面积都需通过对该曲线的函数式求积分来获得，但由于函数式各不相同，因而十分困难。

二、标准正态分布及分布函数表

1、标准正态转换

如果 $X_1, X_2, X_3 \ldots X_N$　$X \sim N(\mu, \sigma^2)$

即 X 服从正态分布，X为正态变量

令 $z = \dfrac{X-\mu}{\sigma}$ ⟹ 则 $z \sim N(0,1)$

即 z 服从标准正态分布，z 为标准正态变量

推导：$X_1, X_2, \ \ldots \ , X_{N-1}, X_N \qquad X \sim N(\mu, \sigma^2)$

$$\mu = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum X}{N}$$

$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$$

令$z = \dfrac{X-\mu}{\sigma}$ 则得：

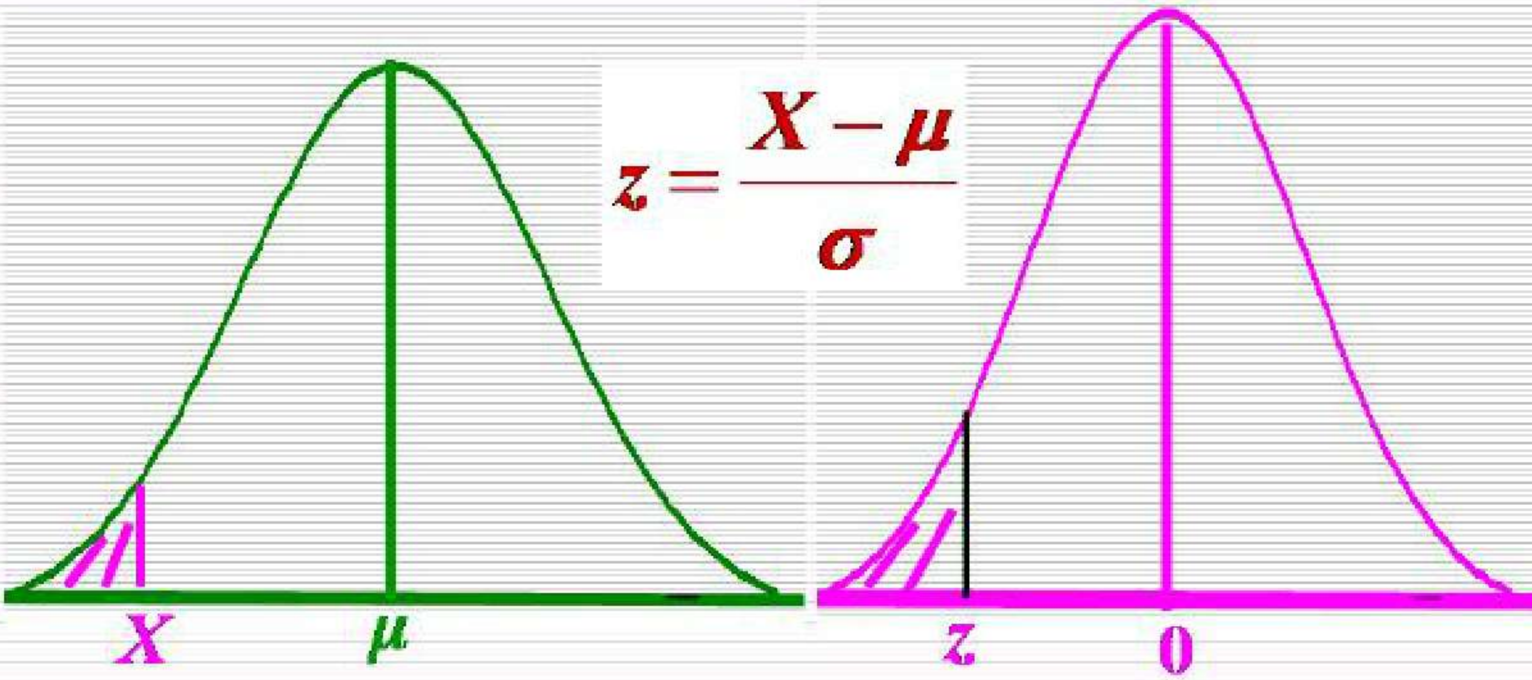
$z_1, z_2, z_3, \cdots \quad z_{N-1}, z_N$

$$\mu_z=\frac{z_1+z_2+\dots+z_N}{N}=\frac{\frac{X_1-\mu}{\sigma}+\frac{X_2-\mu}{\sigma}+\dots\frac{X_N-\mu}{\sigma}}{N}$$

$$=\frac{(X_1+X_2+\dots+X_N)-N\mu}{\sigma N}=\frac{\frac{\sum X}{N}-\mu}{\sigma}=0$$

$$\sigma_z=\sqrt{\frac{\sum(z-0)^2}{N}}=\sqrt{\frac{\sum\left(\frac{X-\mu}{\sigma}\right)^2}{N}}=\frac{1}{\sigma}\sqrt{\frac{\sum(X-\mu)^2}{N}}=1$$

故z服从均数为0，标准差为1的正态分布，即 $z\sim N(0,1)$,又称z分布或称标准正态分布



$$z = \frac{X-\mu}{\sigma}$$

$X \sim N(\mu,\sigma^2)$的正态分布，则

$z \sim N(0,1)$,的标准正态分布。

2、标准正态分布的概率密度函数

$$\varphi(z)=\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}} \qquad -\infty<z<+\infty$$

标准正态分布只有一条曲线，其

$\mu_z=0$,$\sigma_z=1$

3、标准正态分布的特征

①一个高峰:位于中央均数（0）处

②对称性：以0为中心,左右完全对称

③两个参数：

$\mu=0$：位置参数

$\sigma=1$：形态参数

④ 标准正态曲线下的面积分布规律

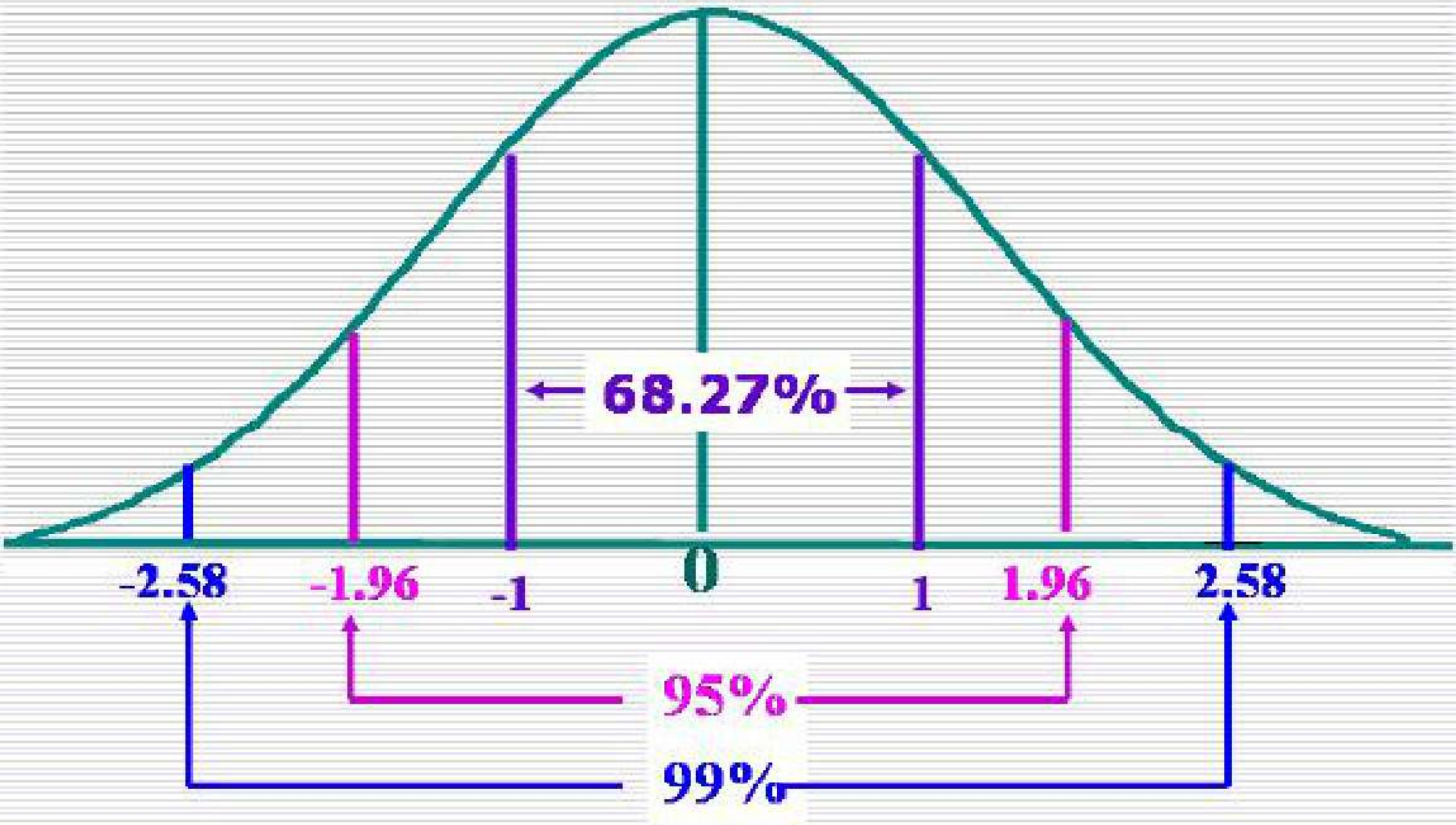
面积通过积分得来: $\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{z^2}{2}}dz$

❶ **标准**正态曲线下面积恒等于 1 或 100%;

❷ 区间 ± 1 的面积为 68.27%;

区间 ± 1.96 的面积为 95.00%;

区间 ± 2.58 的面积为 99.00%。



标准正态分布曲线下的面积示意图

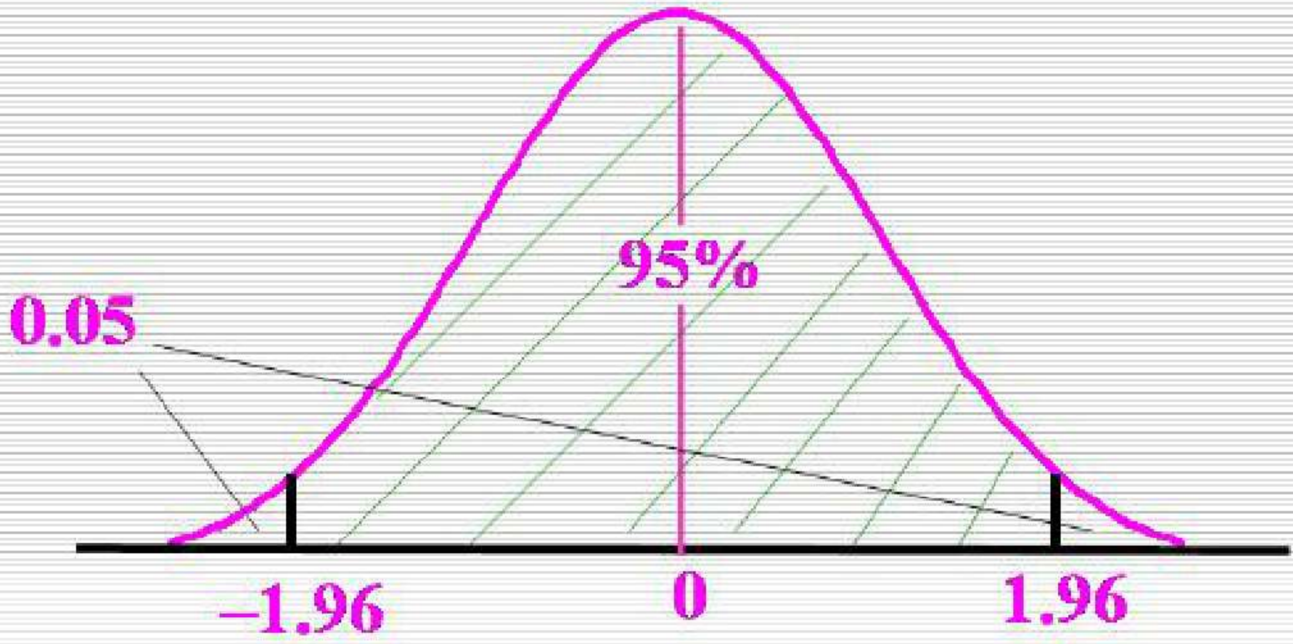
4、标准正态分布表（z值表） P_{274}

$z_{\alpha/2}$:标准正态分布曲线下双尾面积之和为α时所对应的z值的简记。

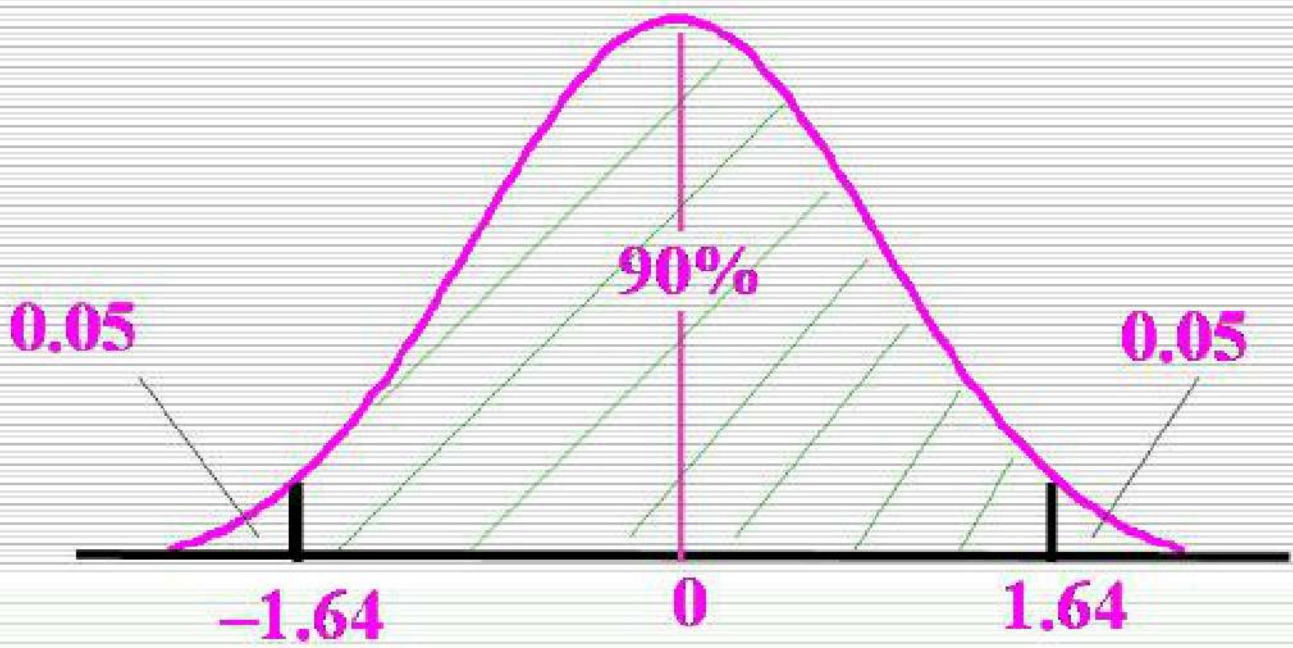
z_{α}:标准正态分布曲线下单尾面积（左侧或右侧）为α时所对应的z值的简记。

$$z_{0.05/2} = \pm 1.96 \qquad z_{0.01/2} = \pm 2.58$$

$$z_{0.05} = \pm 1.64 \qquad z_{0.01} = \pm 2.32$$



$$z_{0.05/2} = \pm 1.96(z_{0.025})$$



$$z_{0.05} = \pm 1.64 (z_{0.10/2})$$

常用 z 界值表

参考范围1-a（%）	单侧z_a	双侧$z_{a/2}$
80	0.84	1.28
90	1.28	1.64
95	1.64	1.96
99	2.33	2.58

小　　结

❶标准正态分布曲线仅一条$\mu=0$，$\sigma=1$。

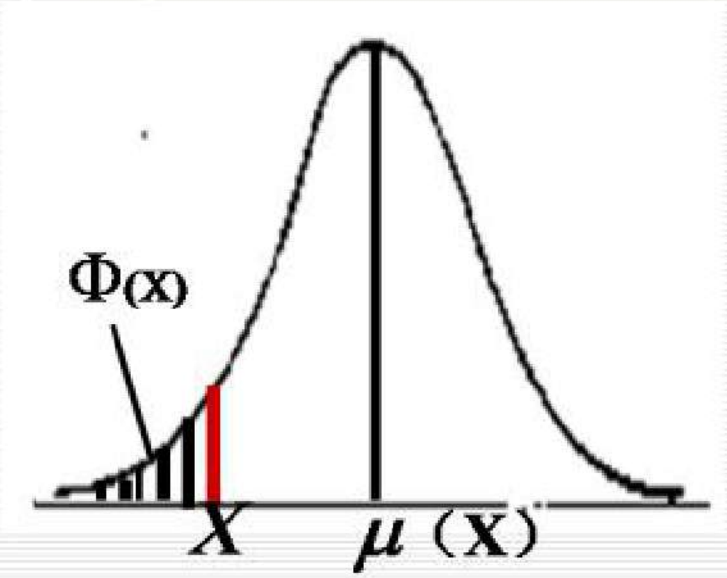
❷标准正态分布曲线的面积可通过查 z 值表获得，而且有三个常用的面积规律。

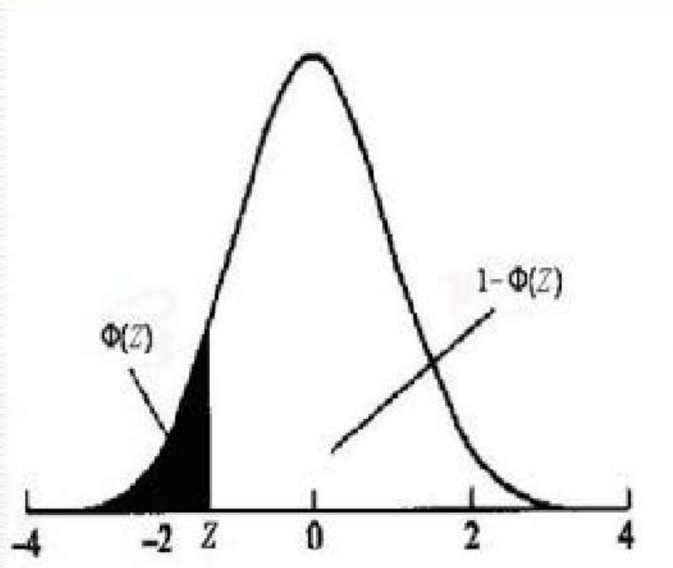
❸普通的正态分布曲线下某区间的面积可通过 z 转换，查 z 值表获得。

第四节 正态分布及其应用

正态分布是很多分布的基础，应用及其广泛。正态分布的理论除了可估计频数分布外，并可应用于质量控制及制定医学参考值范围。

一、估计频数分布





$$X \sim N(\mu, \sigma^2) \longrightarrow z \sim N(0,1)$$

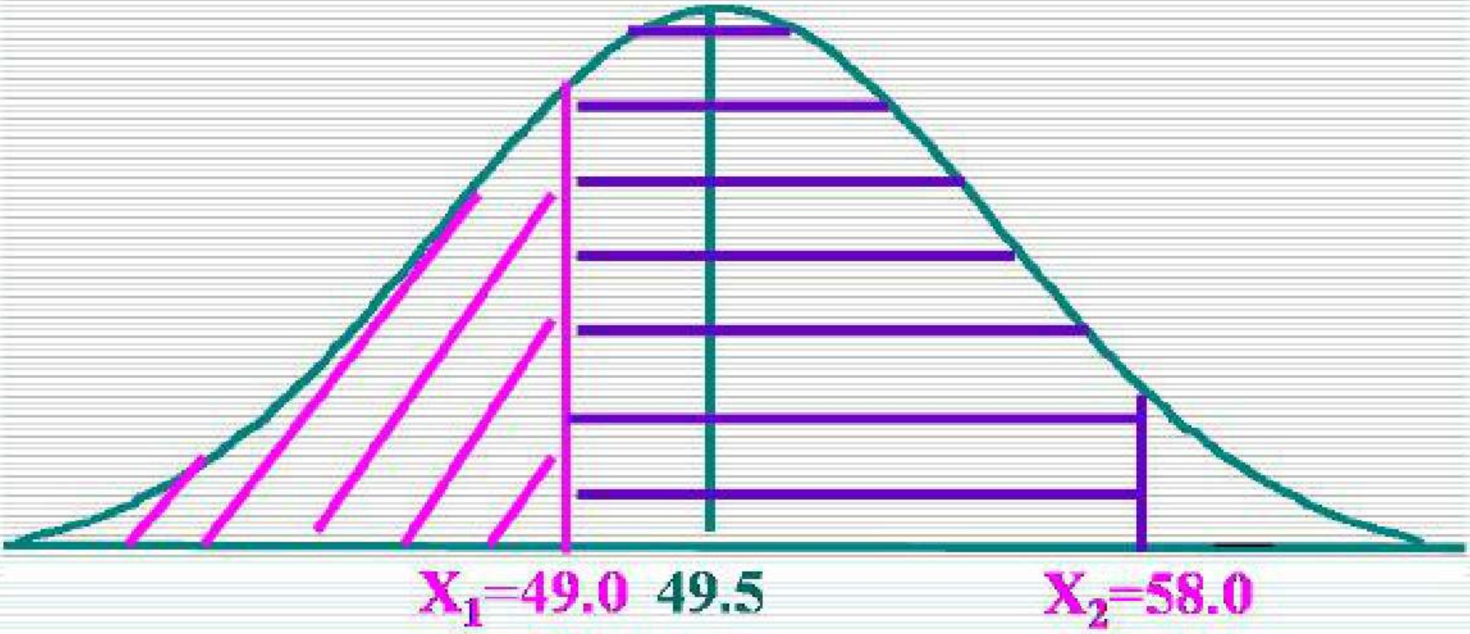
$$\Phi(X) = \Phi(z)$$

$$z = \frac{X - \mu}{\sigma} \quad 或 \quad z = \frac{X - \overline{X}}{S}$$

通过z变换，一般正态分布转换为标准正态分布，再查z值表可得某区间的频数分布比例。

例3-10

对例3-1资料，已计算出某药100片的含量均数$\bar{X}=49.5(mg)$，标准差 $S=4.88(mg)$，试估计：①含药量在49.0mg以下者占药片总数的百分比；②在49.0~58.0mg之间者占药片总数的百分比；③在58.0mg以上者占药片总数的百分比。



$$X \sim N(49.5, 4.88)mg$$

本例：$n = 100, \bar{X} = 49.5, S = 4.88, X_1 = 49.0$

$X_2 = 58.0$

$$Z_1 = \frac{49.0 - 49.5}{4.88} = -0.10 \quad \Phi(Z_1) = \Phi(-0.10) = 0.4602$$

$$Z_2 = \frac{58.0 - 49.5}{4.88} = 1.74 \quad \begin{aligned}\Phi(Z_2) &= \Phi(1.74) \\ &= 1 - \Phi(-1.74) \\ &= 1 - 0.0409 = 0.9591\end{aligned}$$

据上可得：

①含药量在49.0mg以下者占药片总数的百分比为46.02%[$\Phi(z_1)$];

②在49.0~58.0mg之间者占药片总数的百分比为49.89%［$\Phi(z_2)-\Phi(z_1)$］；

③在5.08mg以上者占药片总数的百分比为4.09%［$1-\Phi(z_2)=\Phi(-1.74)$］。

二、质量控制

质量控制常用 $\bar{X} \pm 2S$ 作为上、下警戒线，$\bar{X} \pm 3S$ 作为上下控制线。若进行某一次测量的指标超过上、下警戒线，甚至超过上、下控制线，则有理由认为其指标的波动不仅仅是随机误差引起，可能发生了质量的改变。

第五节　医学参考值范围的制定

人体内很多生理生化指标的频数分布呈正态分布或近似正态分布，还有少数指标近似对数正态分布 。故可用正态分布的原理来制定很多生理生化指标的参考值范围。

一、 医学参考值的概念

医学参考值（reference value）是指包括绝大多数正常人的人体形态、机能和代谢产物等各种生理及生化指标的常数，也称正常值。

由于存在个体差异，生物医学数据并非常数，而是在一定范围内波动，故采用医学参考值范围作为判定正常和异常的参考标准。

二、“正常人”的含义

并不是指机体任何器官，组织的形态和机能都正常的健康人，而是排除了影响所研究指标的疾病和异常的同质人群。

三、确定单、双侧（结合专业知识）

双侧：某指标无论过高过低均异常

血清总胆固醇无论过低或过高均属异常

白细胞数无论过低或过高均属异常

单侧：某指标仅过高或仅过低异常

血清转氨酶仅过高异常（单侧上限）

肺活量仅过低异常（单侧下限）

四、确定百分比（结合误诊与漏诊危害大小）

1、医学参考值范围有80%、90%、95%、99%等，最常用的为95%。

2、根据该指标所诊断疾病在临床中误诊与漏诊危害性大小来选择，若误诊危害大，则可考虑选99%，反之选80%。

五、选择计算方法（结合资料分布）

计算医学参考值范围的常用方法：

正态分布法：资料分布近似正态

百分位数法：资料分布呈偏态

1、正态分布法

要求：$n \geq 100$，资料呈正态分布

双侧参考值范围：$\overline{X} \pm z_{\alpha/2} S$

常用的双侧95%参考值范围：$\overline{X} \pm 1.96S$

单侧参考值范围：$>\overline{X} - z_{\alpha} S$或$<\overline{X} + z_{\alpha} S$

常用的单侧95%参考值范围：

$>\overline{X} - 1.64S$（下限）或 $<\overline{X} + 1.64S$（上限）

例 如

估计某地108名正常成年女子血清总蛋白（$\overline{X}=73.9g/L$，$S=3.9g/L$）的95%参考值范围。

本例解:

因血清总蛋白过多或过少均为异常，故按双侧估计正常成年女子血清总蛋白的95%参考值范围。已知血清总蛋白均数 $\overline{X}=73.9g/L, S=3.9g/L, z_{0.05/2}=1.96$，故

故 下限：$\overline{X}-z_{0.05/2}S=73.9-1.96*3.9=66.3(g/L)$

上限：$\overline{X}+z_{0.05/2}S=73.9+1.96*3.9=81.5(g/L)$

2、百分位数法

要求：偏态分布资料，样本含量较正态分布法要多（>100）

双侧参考值范围：

常用双侧95%参考值范围：$(P_{2.5},\ P_{97.5})$

单侧参考值范围：

常用单侧95%参考值范围：$>P_5$ 或 $<P_{95}$

例　测得某年某地282名正常人的尿汞值如下表，试制定正常人尿汞值的95%参考值范围。

某年某地正常人尿汞值（μg/L）测量结果

尿汞值	频数 f	累计频数 Σf	累计频率（%）
0～	45	45	16.0
8.0～	64	109	38.6
16.0～	96	205	72.7
24.0～	38	243	86.2
32.0～	20	263	93.3
40.0～	11	274	97.2 → P_{95}
48.0～	5	279	98.9
56.0～	2	281	99.6
64.0～72.0	1	282	100.0

鉴于正常人的尿汞值为偏态分布，且过高为异常，故用百分位数法计算上侧界值即第95百分位数

$$P_{95} = L + \frac{i}{f_{95}}(n \times 95\% - \sum f_L)$$

$$= 40.0 + \frac{8.0}{11}(282 \times 95\% - 263) = 43.6(\mu g/L)$$

故该地正常人的尿汞值的95%医学参考值范围为 $<43.6(\mu g/L)$。

本例：$M = 18.67, G = 15.96, M \neq G$，

属偏态分布，采用百分位数法

如果：$M \approx G$，即对数正态分布资料

其参考值范围应如何获得？

请用另一种方法计算例 3－5某地正常女性发汞值的95%参考值范围。并与用百分位数法计算结果 进行比较。

1.标准正态分布的均数与标准差分别为(　　)。

A. 0与1　B. 1与0　C. 0与0　D. 1与1

2.正态分布有两个参数μ与 σ，(　　)相应的正态曲线的形状越扁平。

A. μ 越大　B. μ 越小　C. σ越大　D. σ 越小

3.对数正态分布是一种(　　)分布。

A. 正态　B. 近似正态　C. 左偏态　D. 右偏态

4. 正态曲线下、横轴上，从均数到的面积为(　　　　)。

A. 95%　　B. 50%　　C. 97.5%　　D. 不能确定（与标准差的大小有关）

5. 若X服从以 ， 为均数和标准差的正态分布，则X的第95百分位数等于(　　)。

A. $\mu-1.64\sigma$　　B. $\mu+1.64\sigma$

C. $\mu+1.96\sigma$　　D. $\mu+2.58\sigma$

6. 若正常成人的血铅含量X 近似服从对数正态分布，拟用300名正常人血铅值确定99%参考值范围，最好采用公式(　　　)计算。（其中Y=logX）

A. $\overline{X} \pm 2.58S$　　B. $\overline{X} + 2.33S$

C. $\log^{-1}(\overline{Y} \pm 2.58S_Y)$　　D. $\log^{-1}(\overline{Y} + 2.33S_Y)$

7. 正态曲线下、横轴上，从均数-1.96倍标准差到均数的面积为(　　　)。

A. 95%　　B. 45%　　C. 97.5%　　D. 47.5%

8. 标准正态分布曲线下中间90%的面积所对应的横轴尺度的范围是（　　）。

A. −1.64到+1.64　　B. −∞ 到+1.64

C. −∞到+1.28　　D. −1.28到+1.28