TONGJI UNIVERSITY

生物统计学

biostatistics

张敬 教授

zhangjing@tongji.edu.cn

正态分布与医学参考值范围

正态分布

- 正态分布（**normal distribution**）又称为高斯分布。首先由德国数学家和天文学家德·莫阿弗尔提出，高斯虽然发现稍晚，但他迅速将正态分布应用于天文学，并对其性质作了进一步的研究，使正态分布的应用价值广为人知。

卡尔·弗里德里希·高斯

（C.F.Gauss，1777-1855）

一、正态曲线

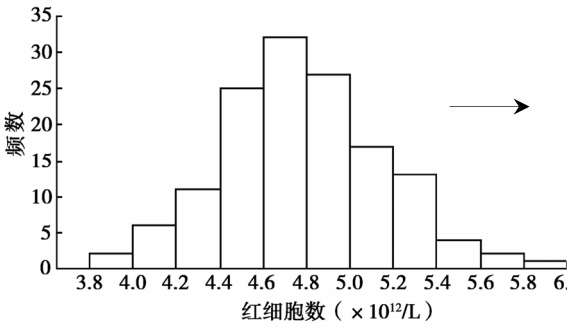


图2-1

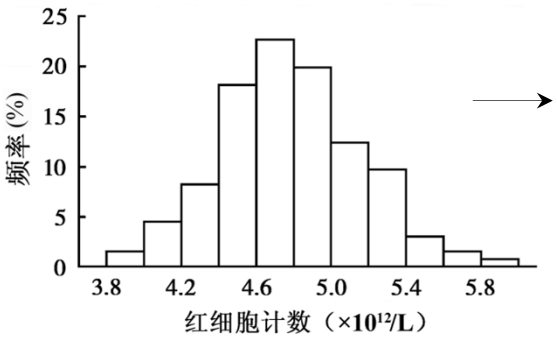


图3-1

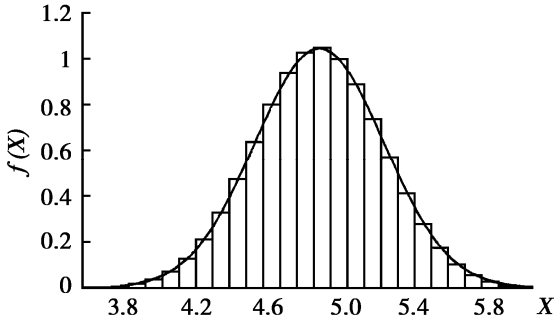
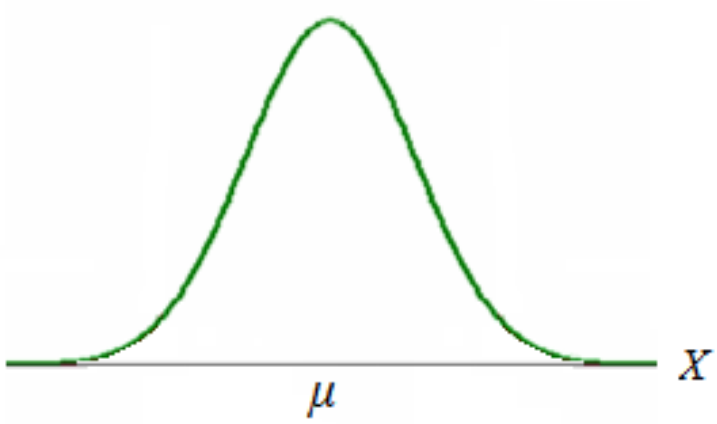


图3-2

某地正常成年男子红细胞数的分布情况

二、正态分布的特征

连续型随机变量X服从正态分布，记为X~ $N(\mu,\sigma^2)$



概率密度函数

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

概率分布函数

$$F(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{X}e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}dx$$

Normal distribution

正态分布是单峰分布，以 $X=\mu$ 为中心左右完全对称

正态曲线在 $X=\mu$ 处最大值， $X=\mu\pm\sigma$ 处有拐点，呈现为钟型

正态分布由两个参数 μ和 σ 决定

μ 是位置参数，决定着正态曲线在X轴上的位置

σ 是形状参数，决定着正态曲线的分布形状

正态曲线下的面积分布有一定的规律

Normal distribution

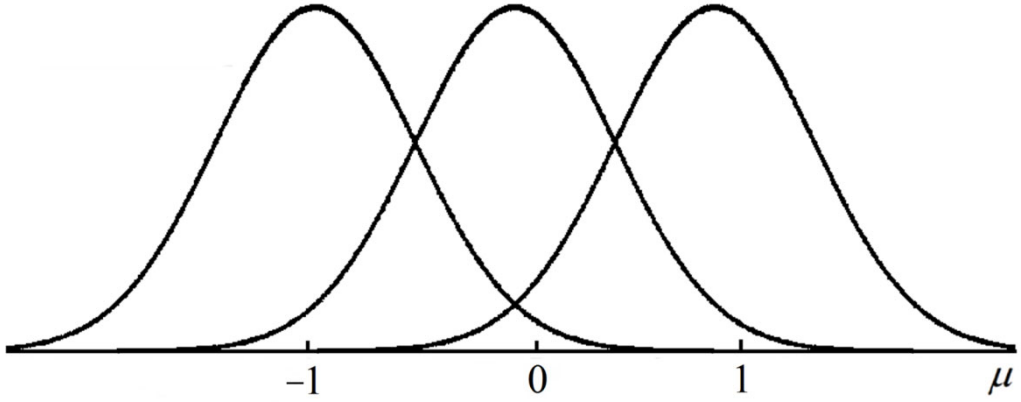


图3-3

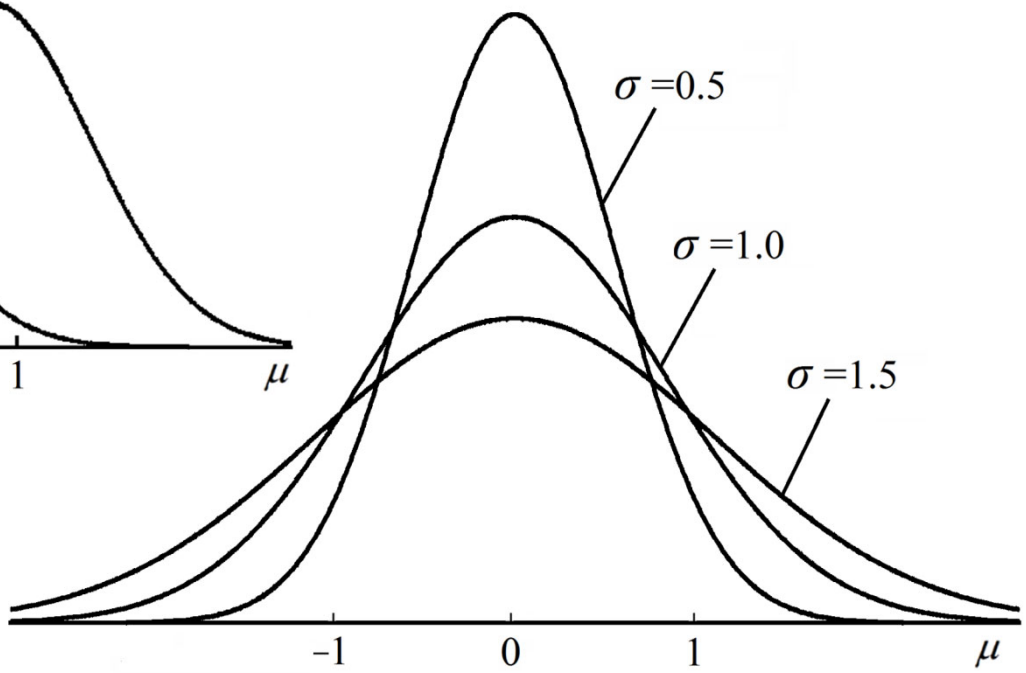


图3-4

Normal distribution

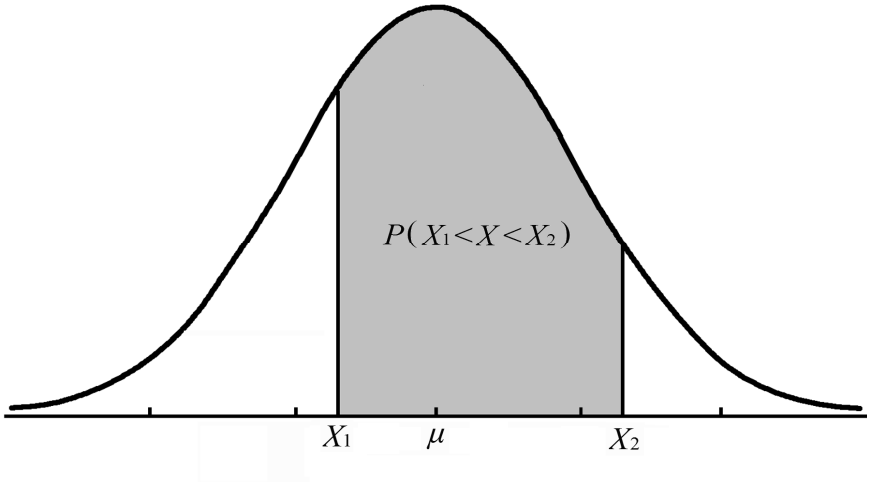


图3-5

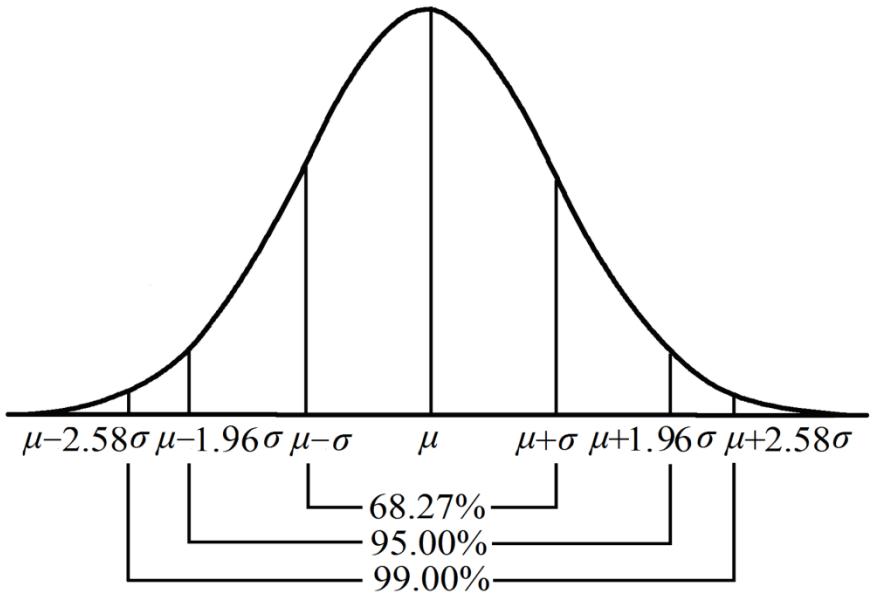


图3-6

Standard normal distribution

三、标准正态分布

μ =0、 σ =1的正态分布即为标准正态分布

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

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

$$\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{z^2}{2}}\,dz$$

Standard normal distribution

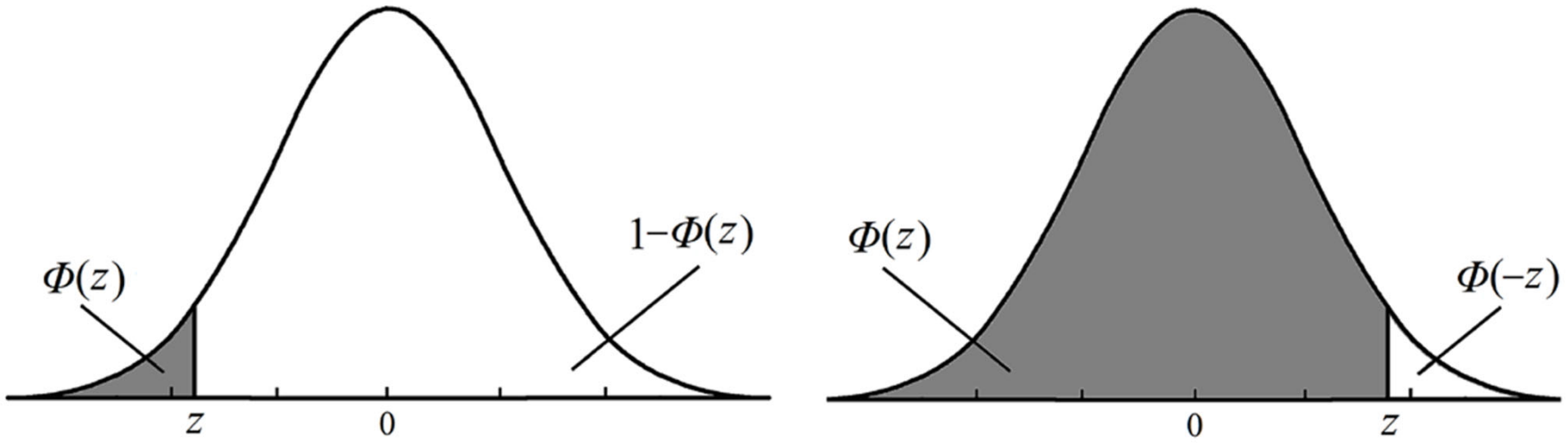


图3-7

Standard normal distribution

z 在区间（z_1, z_2 ）上的概率（曲线下的面积）

$$P\left(z_1 < z < z_2\right) = \Phi\left(z_2\right) - \Phi\left(z_1\right)$$

当 μ 和 σ 未知时，可以利用样本均数 $\bar{X}$ 和标准差 S 计算 z

$$z = \frac{X - \bar{X}}{S}$$

例3-1 若$X\sim N(\mu,\sigma^2)$，试计算X取值在区间$\mu\pm1.96\sigma$上的概率。

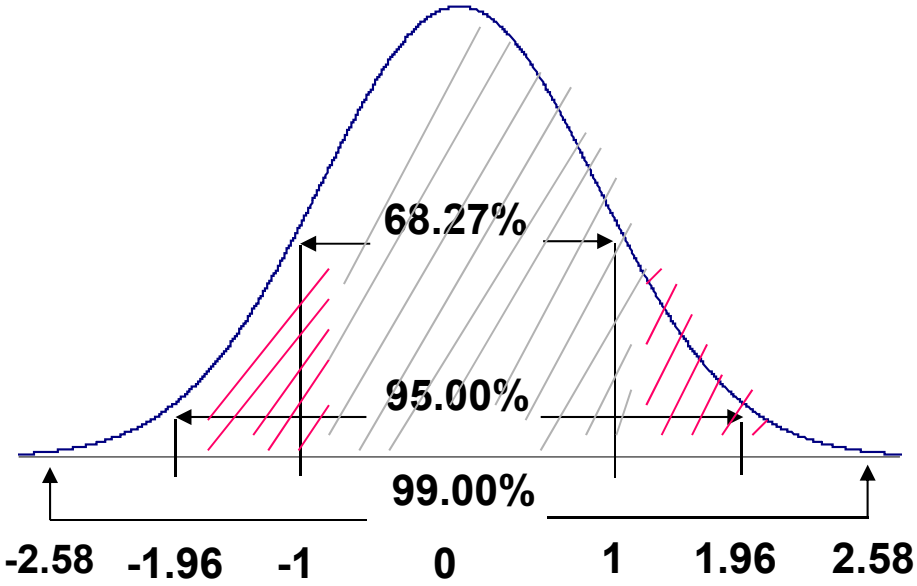
$$z_1=\frac{X_1-\mu}{\sigma}=\frac{(\mu-1.96\sigma)-\mu}{\sigma}=-1.96$$

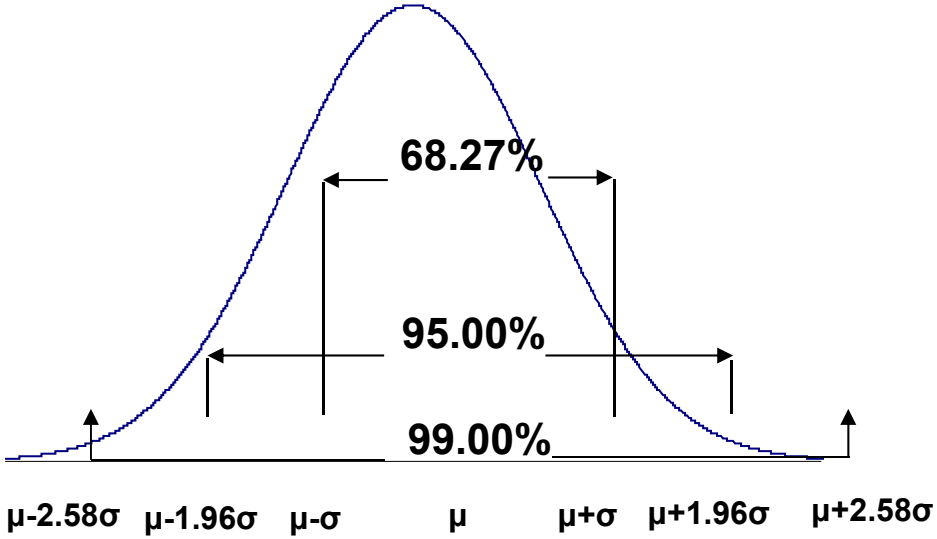
$$z_2=\frac{X_2-\mu}{\sigma}=\frac{(\mu+1.96\sigma)-\mu}{\sigma}=1.96$$

$$P(-1.96<z<1.96)=\Phi(1.96)-\Phi(-1.96)=(1-\Phi(-1.96))-\Phi(-1.96)$$

$$=1-2\Phi(-1.96)=1-2\times0.025=0.95$$

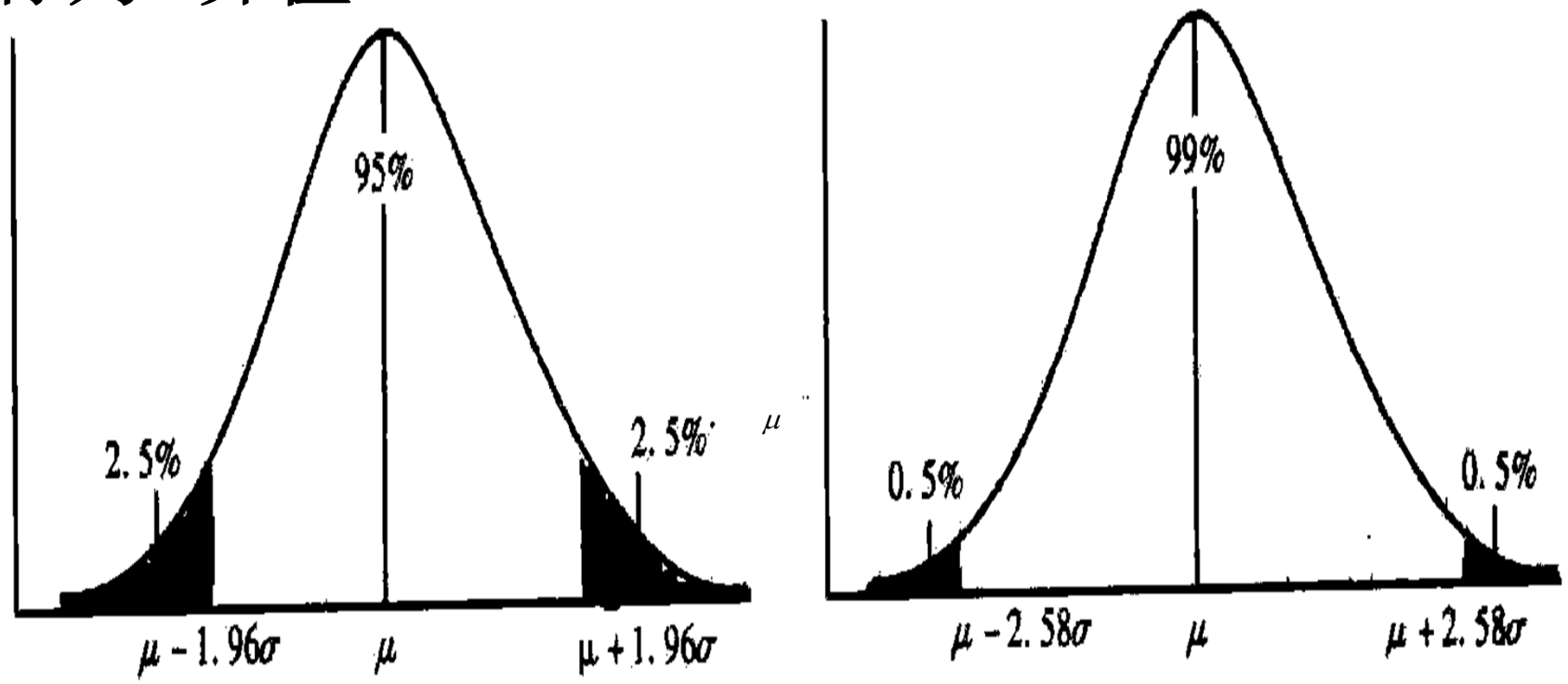
要求记住的三个区间面积





标准正态分布	正态分布	面积或概率
-1～1	μ ± σ	68.27%
-1.96～1.96	μ ±1.96 σ	95.00%
-2.58～2.58	μ ±2.58 σ	99.00%

（2）统计中常用尾部面积的u值，记 u_α ，称为u界值。



$u_{0.05/2}$=1.96 （双侧）　　$u_{0.01/2}$=2.58（双侧）

$u_{0.05}$=1.64 （单侧）　　$u_{0.01}$=2.33（单侧）

Standard normal distribution

例3-2 已知某地140名正常成年男子红细胞计数近似服从正态分布，$\bar{X}$ =4.78×10^{12}/L，S =0.38×10^{12}/L。①该地正常成年男子红细胞计数在4.0×10^{12}/L以下者占该地正常成年男子总数的百分比；

Standard normal distribution

$$z = \frac{X - \bar{X}}{S} = \frac{4.0 - 4.78}{0.38} = -2.05$$

查附表1 $\Phi(-2.05) = 0.0202$，表明该地成年男子红细胞计数低于 4×10^2/L 者约占该地正常成年男子总数的2.02%

Standard normal distribution

② 红细胞计数在4.0×10^{12}/L～5.5×10^{12}/L者占该地正常成年男子总数的百分比

Standard normal distribution

$$P(4.00 < X < 5.50) = P(\frac{4.00-4.78}{0.38} < \frac{X-\mu}{\sigma} < \frac{5.50-4.78}{0.38})$$

$$= P(-2.05 < z < 1.89)$$

$$\left(1-\Phi\left(-1.89\right)\right)-\Phi\left(-2.05\right)=\left(1-0.0294\right)-0.0202=0.9504$$

- 表明红细胞计数在 **4.0×10¹²/L～5.5×10¹²/L**者约占该地正常成年男子总数的**95.04%**。

四、正态分布的应用

1.估计正态分布**X**值在特定值范围内的分布比例。

2.制定某临床指标的的参考值范围

3.利用 $\overline{X}\pm 3S$ 估计变量值的范围或对极端值做取舍。

4.许多统计方法的统计推断建立在正态分布基础上。

第二节　医学参考值范围

一、医学参考值范围的概念

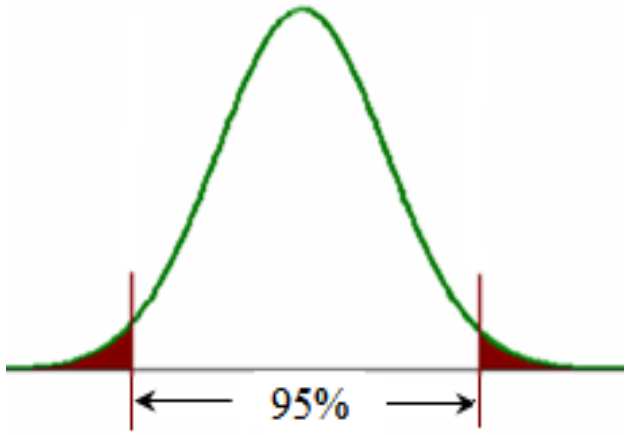
- **医学参考值范围，指“正常”人的解剖、生理、生化指标等数据大多数个体值的波动范围。**

确切含义是，从选择的参照总体中获得的所有个体观察值，用统计学方法建立百分位数界限，由此得到个体观察值的波动区间。通常情况使用的是95%参考值范围。

- 确定医学参考值范围的意义

1. 基于临床实践，从个体角度，作为临床上判定正常与异常的参考标准，即用于划分界限或分类。

2. 基于预防医学实践，从人群角度，可用来评价儿童的发育水平，如制订不同年龄、性别儿童某项发育指标的等级标准。



确定95%参考值范围示意图

二、制订医学参考值范围的注意事项

1. 确定同质的参照总体

- **一般选择“正常”人，主要是排除了对研究指标有影响的疾病或有关因素的同质人群。**

2. 选择足够例数的参照样本

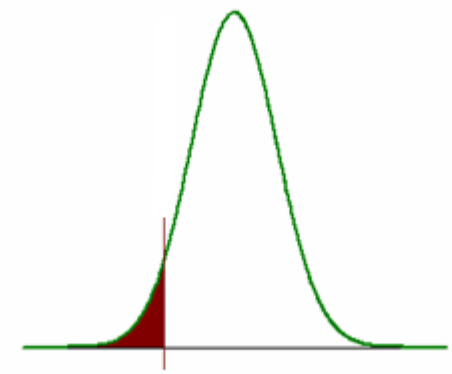
通常情况下，确定参考值范围需要大样本，如果例数过少，确定的参考值范围往往不够准确。

3. 控制检测误差

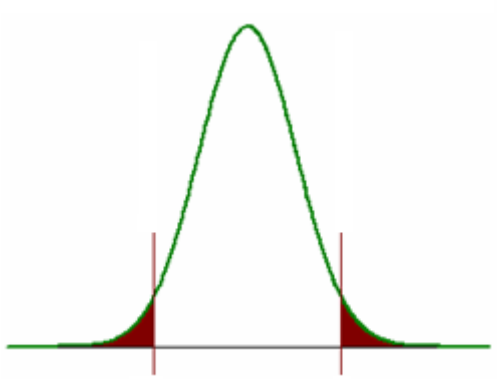
为保证原始数据可靠，检测过程中要严格控制随机误差，避免系统误差和过失误差。

4. 选择单、双侧界值

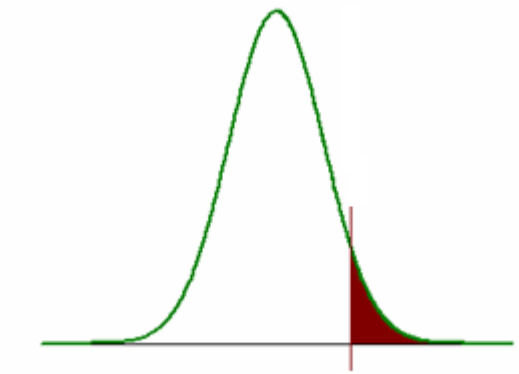
依据专业知识确定，研究指标无论过高或过低均属异常，采用双侧界值；有些指标仅过大或者过小为异常，采用单侧界值。



肺活量参考值范围



白细胞数参考值范围



血铅参考值范围

5. 选择适当的百分数范围

结合专业知识，根据研究目的、研究指标的性质、数据分布特征等情况综合考虑。百分数范围的不同将导致不同的假阳性率和假阴性率。

6. 选择计算参考值范围的方法

- 根据资料的分布类型，样本含量的多少和研究目的等，选用适当的方法确定参考值范围。

三、医学参考值范围的计算方法

- 百分位数法适合于任何分布类型的资料，在实际中最为常用。由于参考值范围所涉及的常常是波动较大的两端数据，使用百分位数法必须要有较大的样本含量，否则结果不稳定。
- 正态分布法要求资料服从或近似服从正态分布，优点是结果比较稳定，在样本含量不是很大的情况下仍然能够进行处理；若偏态分布资料经变量变换能转换为正态分布或近似正态分布，仍可用正态分布法。

表 3-1　医学参考值范围的正态分布法和百分位数法计算公式

概率（%）	正态分布法			百分位数法		
	双侧	单侧		双侧	单侧	
		下限	上限		下限	上限
90	$\bar{X}\pm1.64S$	$\bar{X}-1.28S$	$\bar{X}+1.28S$	$P_5 \sim P_{95}$	P_{10}	P_{90}
95	$\bar{X}\pm1.96S$	$\bar{X}-1.64S$	$\bar{X}+1.64S$	$P_{2.5}\sim P_{97.5}$	P_5	P_{95}
99	$\bar{X}\pm2.58S$	$\bar{X}-2.33S$	$\bar{X}+2.33S$	$P_{0.5}\sim P_{99.5}$	P_1	P_{99}

例3-3 已知某地140名正常成年男子红细胞计数近似服从正态分布，$\bar{X}$ =4.78×10^{12}/L，S =0.38×10^{12}/L，估计该地正常成年男子红细胞计数95%参考值范围。

- 近似正态分布资料可按正态分布法处理，因红细胞计数值过大或过小均为异常，故应估计双侧**95%**参考值范围：

$$\bar{X} \pm z_{0.05/2} S = 4.78 \pm 1.96 \times 0.38 = (4.04\ ,\ 5.52)$$

即该地正常成年男子红细胞计数的95%参考值范围为4.04×10^{12}/L～5.52×10^{12}/L。

- **例3-4 某年某地测得 100 名正常成年人的血铅含量值（μg/dl），试确定该地正常成年人血铅含量的95%参考值范围。**

根据经验已知正常成年人的血铅含量近似对数正态分布，因此首先对原始数据作对数变换，经正态性检验可知对数值服从正态分布（$P>0.50$），故编制对数值频数表，再利用正态分布法求95%参考值范围。

4	4	5	5	6	6	7	7	7	7	7	8	8	8	8	8	8	8	9	9
10	10	10	10	10	10	10	10	11	11	11	12	13	13	13	13	13	13	13	13
13	13	14	14	14	15	15	16	16	16	16	16	16	16	16	17	17	17	17	17
18	18	18	18	19	20	20	20	20	21	21	22	22	22	23	24	24	25	25	26
26	26	27	27	28	28	29	30	30	31	31	32	32	32	33	35	41	44	50	51

表3-2 某年某地100名正常成年人血铅含量（μg/dl）对数值频数表

对数组段	频数	累计频数
0.6～	4	4
0.7～	2	6
0.8～	5	11
0.9～	9	20
1.0～	12	32
1.1～	15	47
1.2～	18	65
1.3～	14	79
1.4～	12	91
1.5～	5	96
1.6～	3	99
1.7～1.8	1	100
合计	100	—

依据表3-2，设X为对数组段的组中值，$n=100$，$\sum fX=120$，$\sum fX^2=149.73$，则对数值的均数和标准差为：

$$\bar{X}=\frac{\sum fX}{n}=\frac{120}{100}=1.2(\mu g/dl)$$

$$S=\sqrt{\frac{\sum fX^2-(\sum fX)^2/n}{n-1}}=\sqrt{\frac{149.73-120^2/100}{100-1}}$$

$$S=0.2406(\mu g/dl)$$

因为血铅含量仅过大异常，故参考值范围应为单侧，求单侧95%上限值：

$$\lg^{-1}(\bar{X}+1.64S)=\lg^{-1}(1.2+1.64\times 0.2406)=39.3173(\mu g/dl)$$

即该地正常成年人血铅含量95%参考值范围为小于39.3173μg/dl 。

- **例3-5 依据表2-4某地630名50岁～60岁正常女性血清甘油三酯含量（mmol/L）的资料，估计其血清甘油三脂含量的单侧95%参考值范围，为该地50～60岁女性高血脂诊断与治疗提供参考依据。**

资料显现出血清甘油三脂含量数值偏小的人数较多，呈正偏态分布，故选用百分位数法计算参考值范围；依据专业知识，为该地50～60岁女性高血脂诊断与治疗提供参考依据应计算单侧95%界值P_{95}。

表2-4 某地630名正常女性血清甘油三酯含量(mmol/L)的频数表

甘油三脂	频数	累积频数	累积频率(%)
0.10～	27	27	4.3
0.40～	169	196	31.1
0.70～	167	363	57.6
1.00～	94	457	72.5
1.30～	81	538	85.4
1.60～	42	580	92.1
1.90～	28	608	96.5
2.20～	14	622	98.7
2.50～	4	626	99.4
2.80～	3	629	99.8
3.10～	1	630	100.0
合计	630	—	—

$$P_{95}=1.90+(630\times95\%-580)/28\times0.30=2.098$$（mmol/L）

即该地50～60岁正常女性血清甘油三脂含量的单侧95%参考值范围为小于2.098 mmol/L。

- 许多统计方法都要求资料服从正态分布或者近似正态分布，在使用这些方法之前需对资料进行正态性判定。如有充足的专业知识和经验得知某些医学指标服从正态分布，或样本含量足够大时，可不必再作正态性判定。正态性判定的方法有两类：一是图示法，二是计算法，图示法简单易行但比较粗糙，计算法检验效率较高，可利用统计软件获得计算结果。

小 结

1. 正态分布是许多统计分析方法的理论基础，是医学研究应用中重要的一种连续型分布。

2. 正态分布受到两个参数影响，总体均数 μ 是位置参数，决定着正态曲线在横轴上的位置；总体标准差 σ 是形状参数，决定着正态曲线的分布形状。不同的 μ 与 σ 对应不同的正态分布，记为$X\sim N(\mu,\sigma^2)$ 。正态曲线下的面积即为概率，利用其面积分布规律 可估计频数分布和确定医学参考值范围。

3. μ =0、 σ =1的正态分布称作标准正态分布，即 $z \sim N(0,1)$ 。对服从 $N(\mu,\sigma^2)$ 的任意随机变量X，都可经 z 变换转化成标准正态分布， $z=(X-\mu)/\sigma$ 。

4. 医学参考值范围指同质总体中某医学指标大多数个体值的波动范围。计算参考值范围常用的方法有正态近似法和百分位数法，当资料服从正态分布或转换值服从正态分布，可用正态近似法；若资料不服从正态分布或未知分布类型，可用百分位数法。

Thank You

同济大学生命科学与技术学院