

例题1



某地108名正常成年女子的血清总蛋白(g/L)如表所示，
试估计该地正常女子血清总蛋白 $<68.0\text{g/L}$ ， $<78.0\text{g/L}$ ，
 $\geq 78.0\text{g/L}$ 所占正常女子总人数的百分比。

例题1



表1 某地108名正常成年女子血清总蛋白(g/L)含量

67.3	75.4	73.1	70.9	75.1	72.6	78.2	68.8	73.8	71.5	66.5	75.1
70.7	68.9	73.3	72.3	76.5	74.3	75.9	75.4	67.2	71.8	76.2	70.6
70.7	75.6	73.3	72.4	76.6	67.3	80.8	74.3	73.9	71.6	79.9	69.3
80.3	75.7	73.5	81.2	74.4	72.5	77.1	67.3	74.1	68.0	76.4	70.4
71.0	75.8	73.6	78.1	68.7	72.6	77.6	72.2	74.2	72.1	76.3	69.7
71.1	75.7	73.5	72.7	78.3	72.5	77.2	68.2	74.2	72.3	76.5	70.5
71.2	83.7	73.7	75.8	74.7	72.6	69.5	66.0	76.1	77.7	80.5	83.1
64.1	75.1	76.3	77.8	65.2	75.0	72.7	78.8	71.1	71.8	72.9	76.1
71.2	75.2	72.9	79.5	73.9	75.2	73.1	79.5	81.8	74.5	81.6	74.5



编制频率表

1. 求极差 (range) : $R=83.1-64.1=19(\text{g/L})$

2. 决定组数、组段和组距:

组数=10

组距=极差/组数 $i= R/10=19/10=1.9\approx 2$

编制频率表



表2 108名正常成年女子血清总蛋白(g/L)频数分布

组段 (1)	频数, f (2)	组中数, X (3)	$f \cdot X$ (4)=(2) (3)	$f \cdot X^2$ (5)= (2) (3) ²
64.0~	2	65.0	130.0	8450.0
66.0~	6	67.0	402.0	26934.0
68.0~	8	69.0	552.0	38088.0
70.0~	15	71.0	1065.0	75615.0
72.0~	25	73.0	1825.0	133225.0
74.0~	23	75.0	1725.0	129375.0
76.0~	14	77.0	1078.0	83006.0
78.0~	7	79.0	553.0	43687.0
80.0~	6	81.0	486.0	39366.0
82.0~84.0	2	83.0	166.0	13778.0
合计	108	—	7982.0	591524.0

频率表

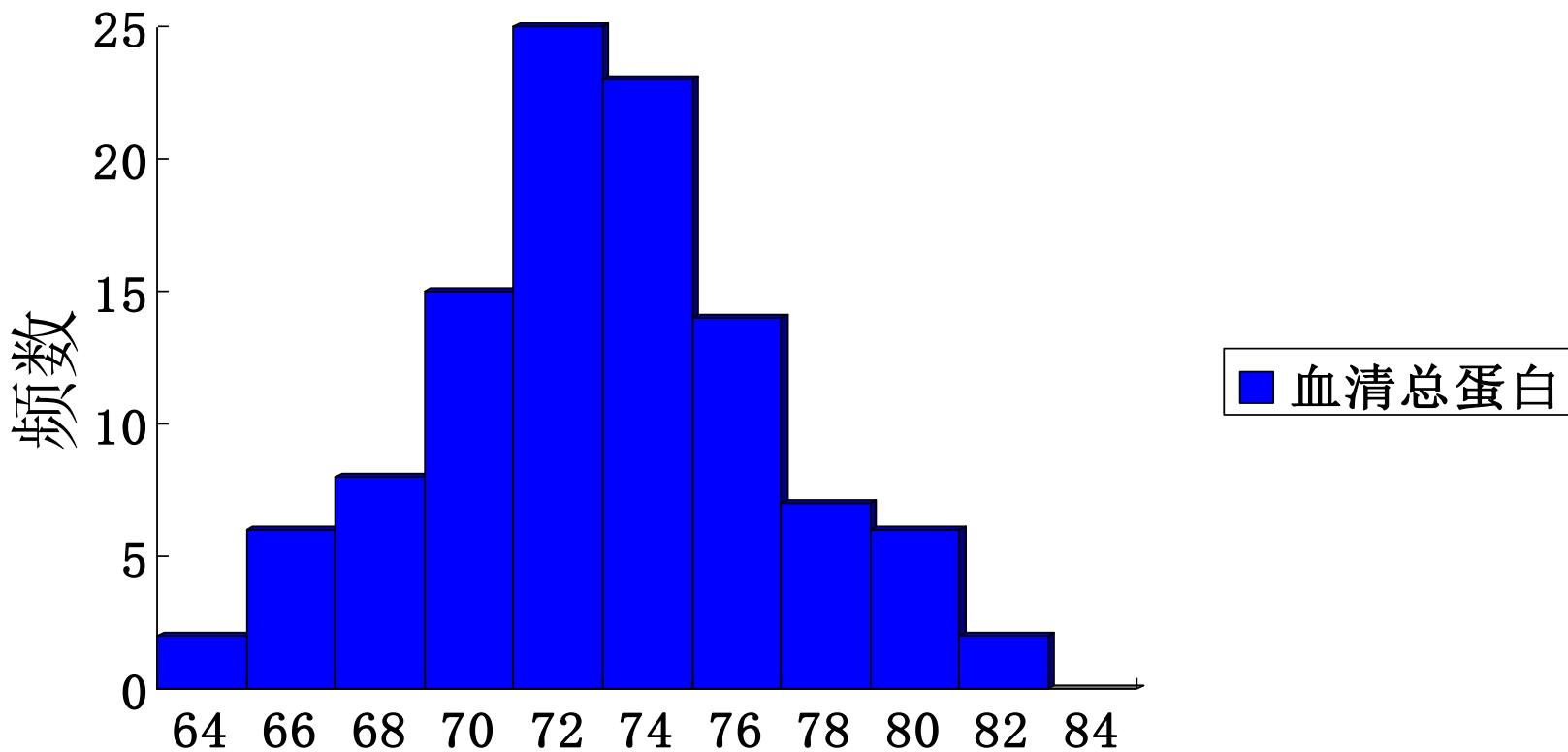


图1 某地108名正常成年女子血清总蛋白 (g/L) 含量

正态分布



正态分布（normal distribution）也叫高斯分布（Gaussian distribution），是最常见、最重要的一种连续型分布。

一、正态分布的数学形式

二、标准正态分布

三、曲线下面积

四、正态性检验

五、正态分布的应用

一. 正态分布的数学形式



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

☺ $f(X)$ = 随机变量 X 的频数，

称为概率密度函数

(probability density function)

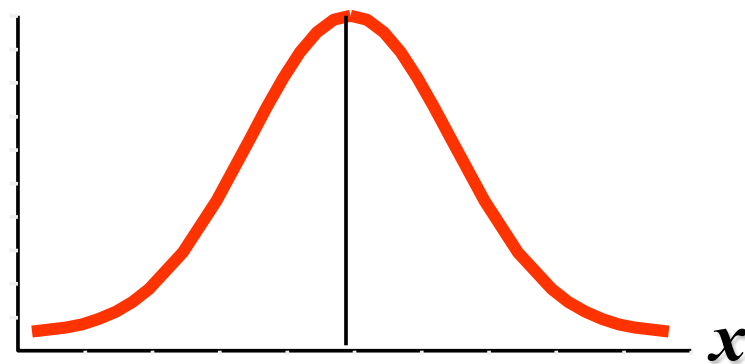
☺ σ^2 = 总体方差, μ = 总体均值

☺ $X \sim N(\mu, \sigma^2)$

☺ 以 X 为横坐标, $f(X)$ 为纵坐标,

绘制的曲线就是 正态曲线(normal curve)

$f(x)$



μ

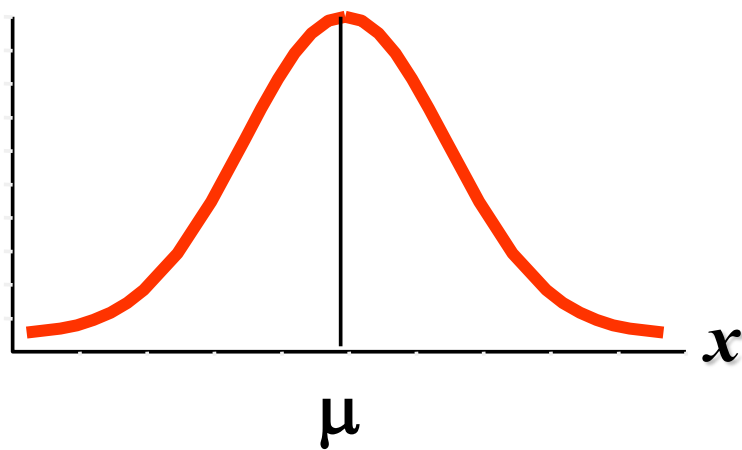
$$-\infty < x < +\infty$$

正态分布的特征



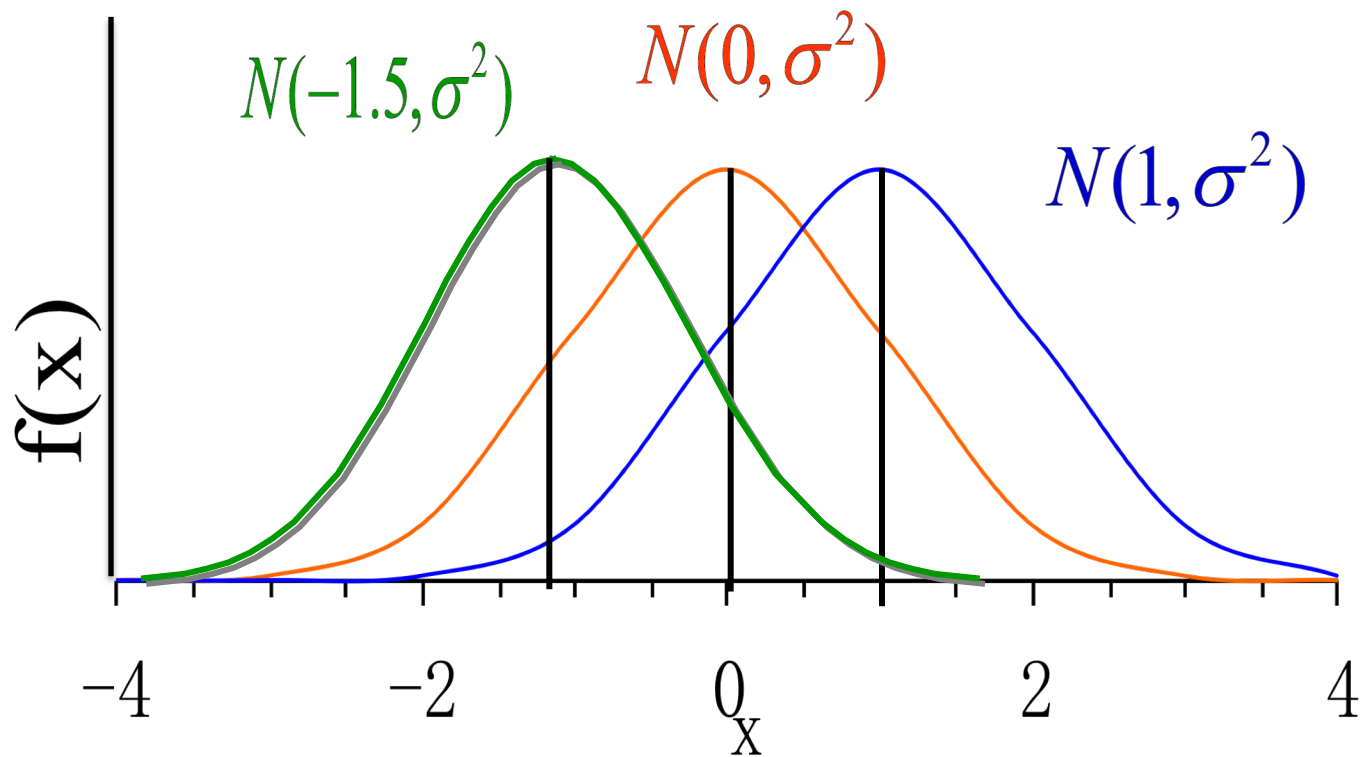
$$N(\mu, \sigma^2) = f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$f(x)$

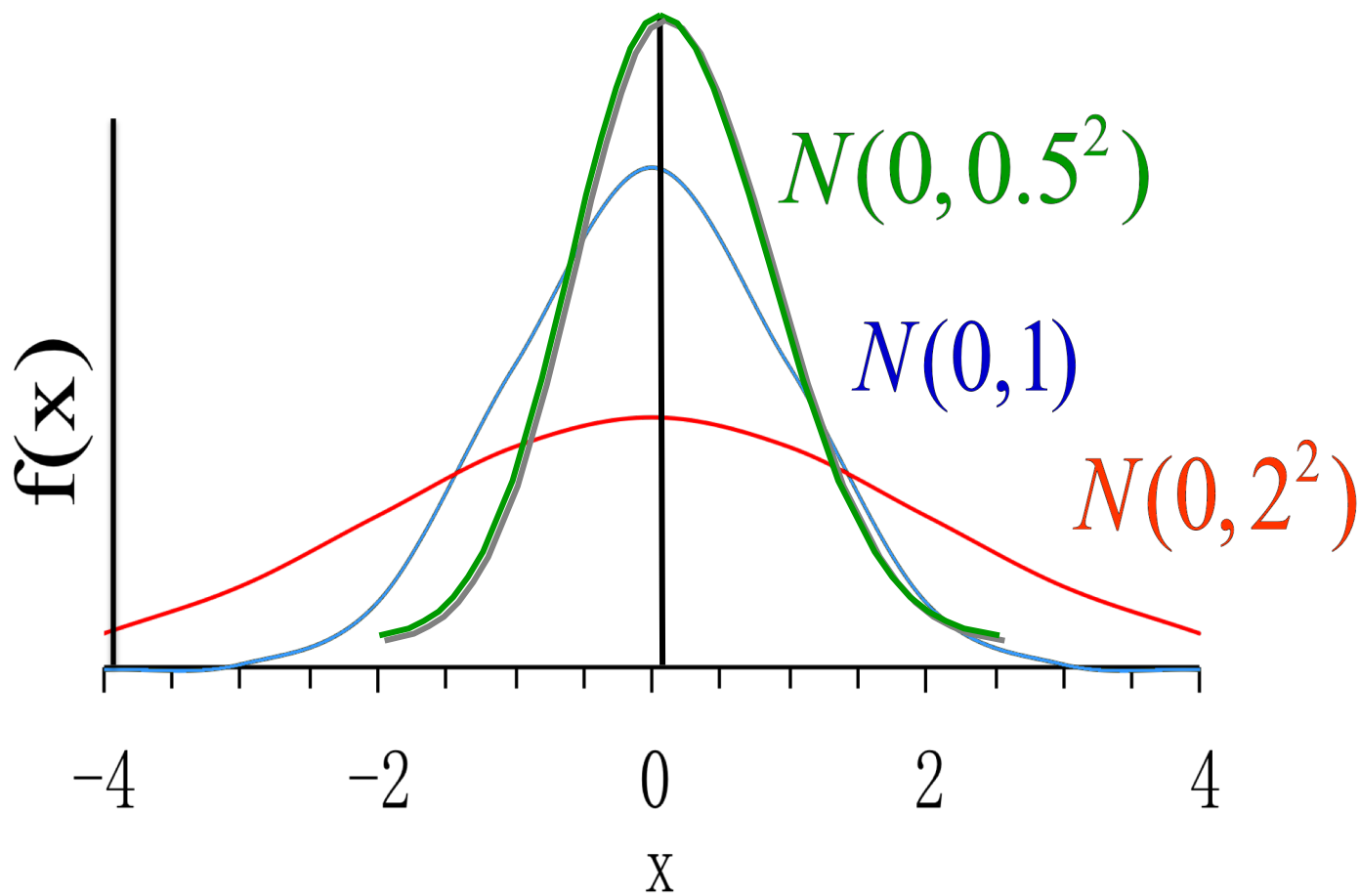


1. 正态曲线位于直角坐标系上方，以 $X=\mu$ 为中心，左右对称，两端以 x 轴为渐进线。
2. 在 $X=\mu$ 处 $f(x)$ 取最大值， $f(\mu) = 1/\sigma\sqrt{2\pi}$ ； x 越远离 μ ， $f(x)$ 值越小。
3. 位置参数 μ ，
形态参数 σ

正态分布的位置参数



正态分布的形态参数



二. 标准正态分布

(standard normal distribution)

两个参数: $\mu=0, \sigma=1$, 记为 $N(0,1)$

经 u 变换: 一般正态分布 $N(\mu, \sigma^2)$

转化为标准正态分布 $N(0,1)$; 其中 $u = \frac{X - \mu}{\sigma}$

$$f(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right), \quad -\infty < X < \infty$$

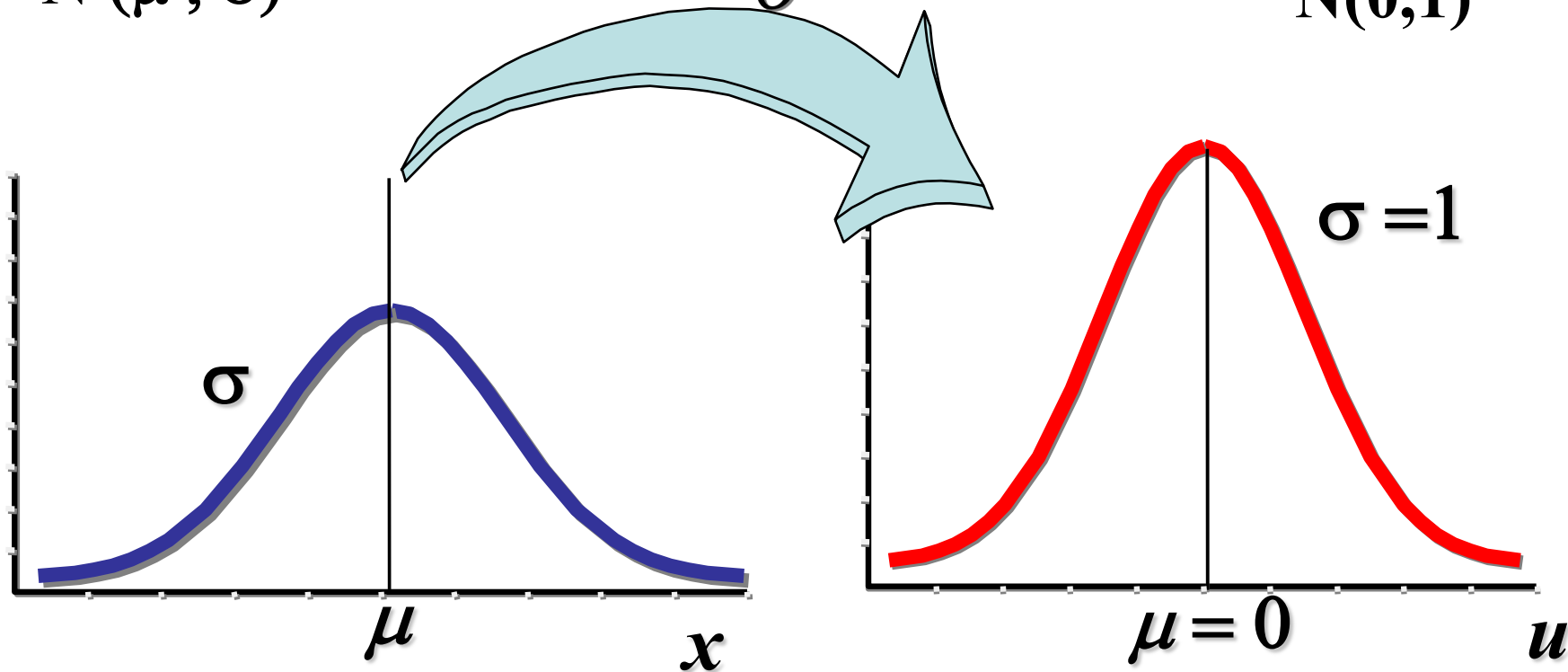
标准正态分布的转换



一般正态分布
 $N(\mu, \sigma)$

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

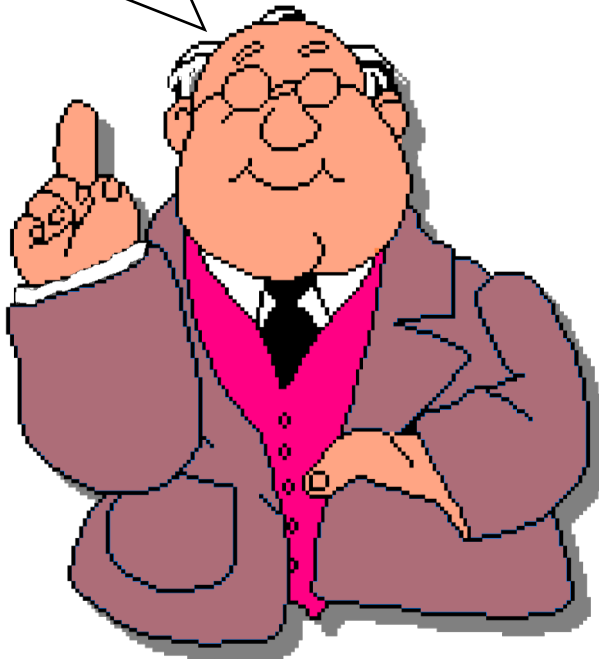
标准正态分布
 $N(0,1)$



正态分布曲线下的面积



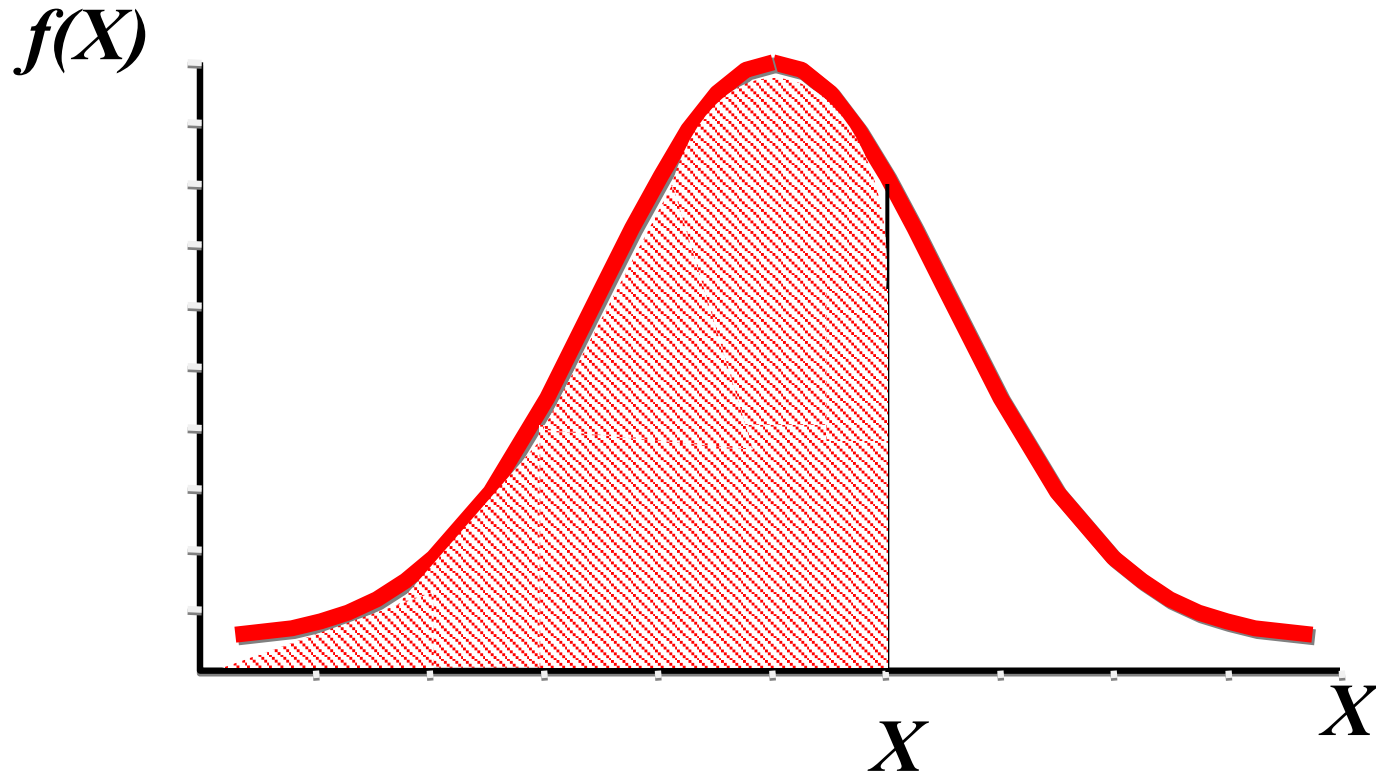
概率是曲线下的面积!



正态曲线下的面积分布有一定的规律。
求其一区间的面积，可通过下面积分公式得到。

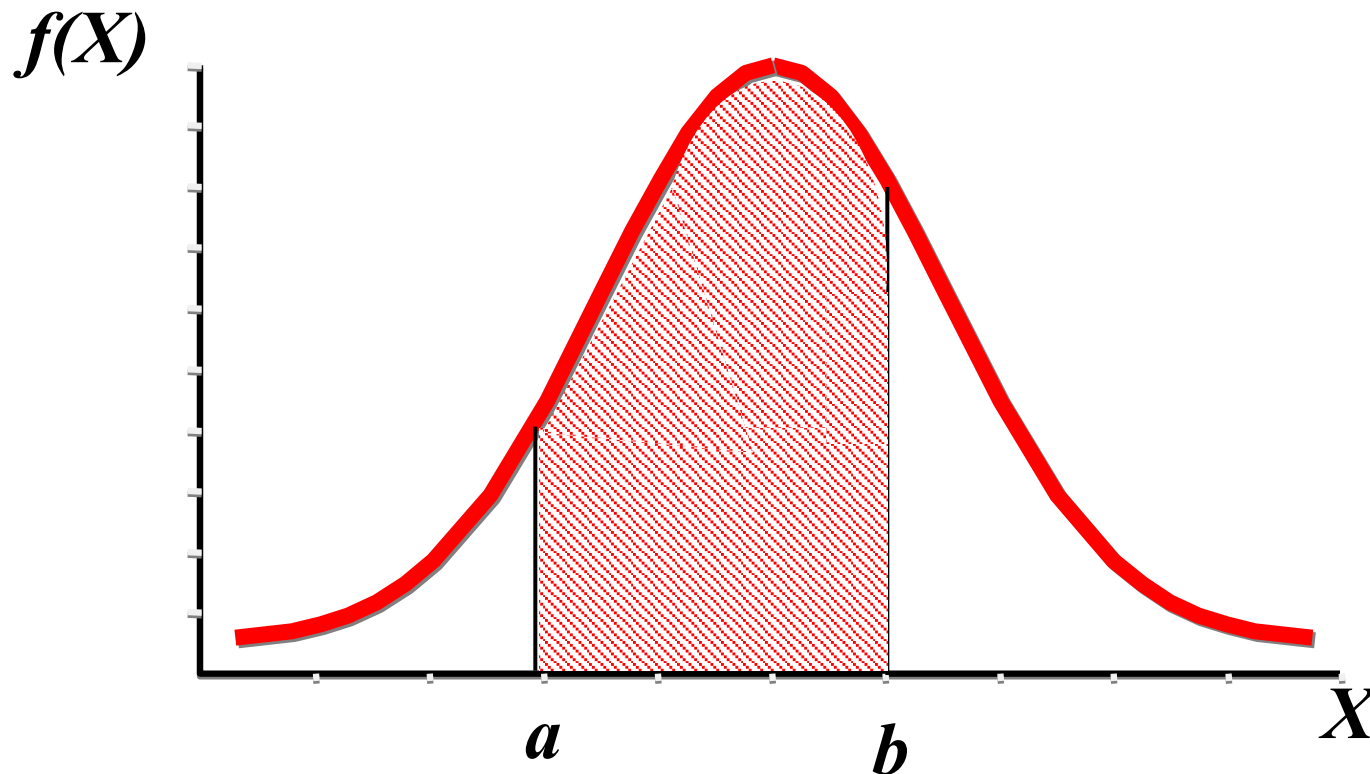
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

正态分布曲线下的面积例题1



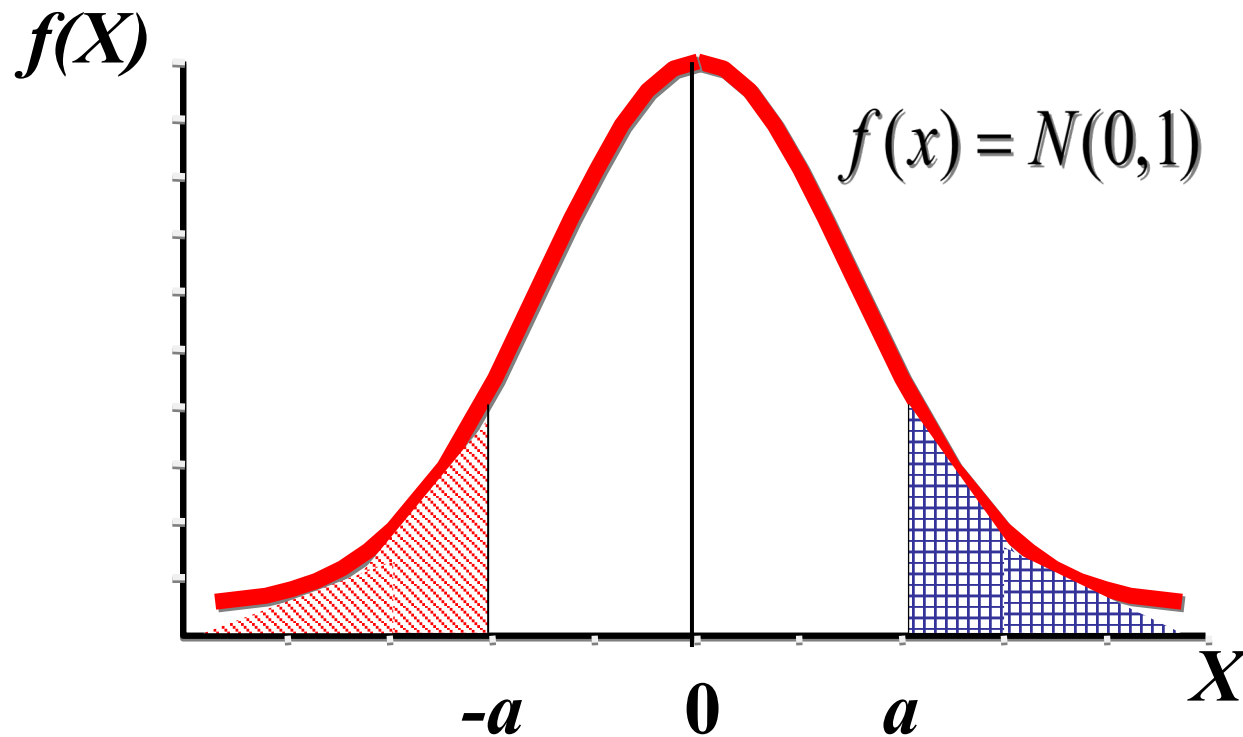
$$F(X) = \int_{-\infty}^X f(x) dx$$

正态分布曲线下的面积例题2



$$P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a)$$

正态分布曲线下的面积例题3



$$F(-a) = \int_{-\infty}^{-a} f(x) dx$$

$$F(a) = \int_{-\infty}^a f(x) dx = 1 - F(-a) \quad 16$$

查表求曲线下面积



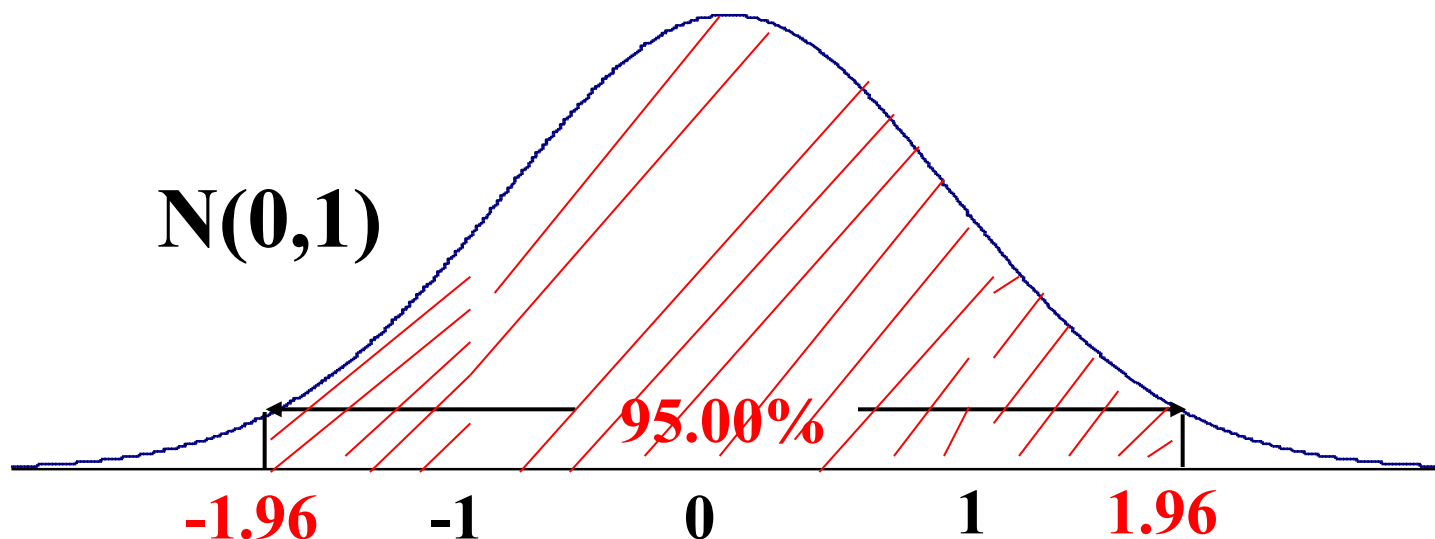
附表 1 标准正态曲线下面积 $\varphi(-u)$

u	0.00	0.01	0.02	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0012	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0015	0.0015	0.0014	0.0014
.....
-2.5	0.0062	0.0060	0.0059	0.0052	0.0051	0.0049	0.0048
.....
-1.9	0.0287	0.0281	0.0274	0.0250	0.0244	0.0239	0.0233
.....
-0.1	0.4602	0.4562	0.4522	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4761	0.4721	0.4681	0.4641

查表求曲线下面积例题



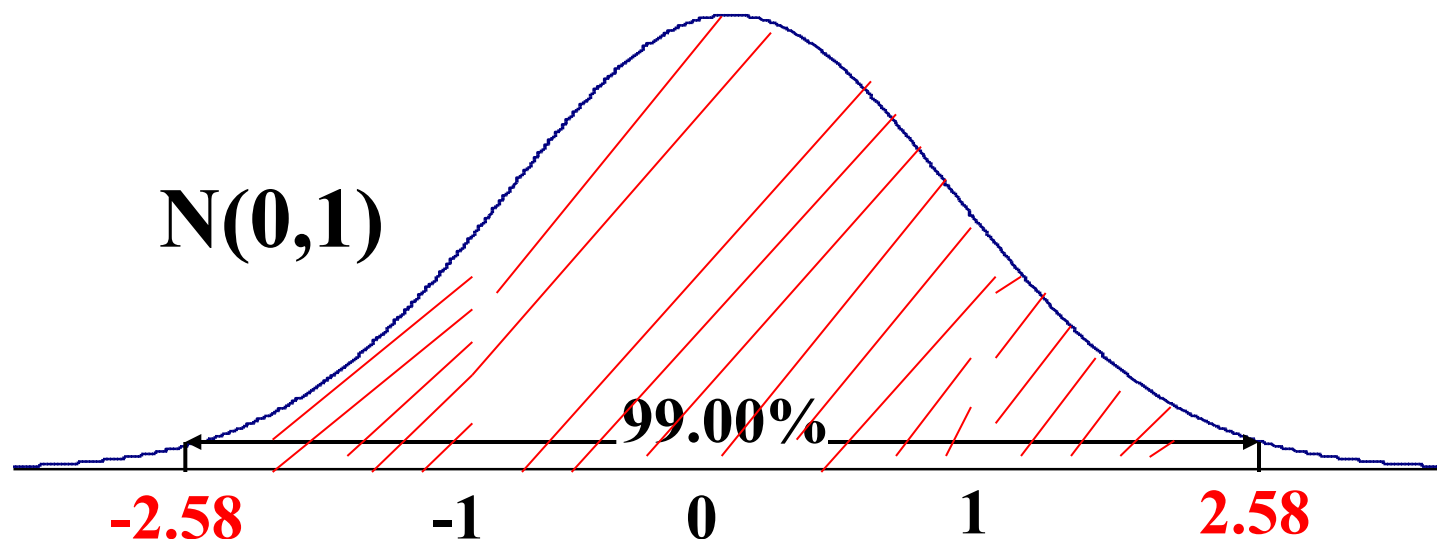
$$\begin{aligned} P(-1.96 < u < 1.96) &= \Phi(1.96) - \Phi(-1.96) \\ &= [1 - \Phi(-1.96)] - \Phi(-1.96) = 1 - 2\Phi(-1.96) \\ &= 1 - 2 \times 0.025 = 0.95 \end{aligned}$$



查表求曲线下面积例题



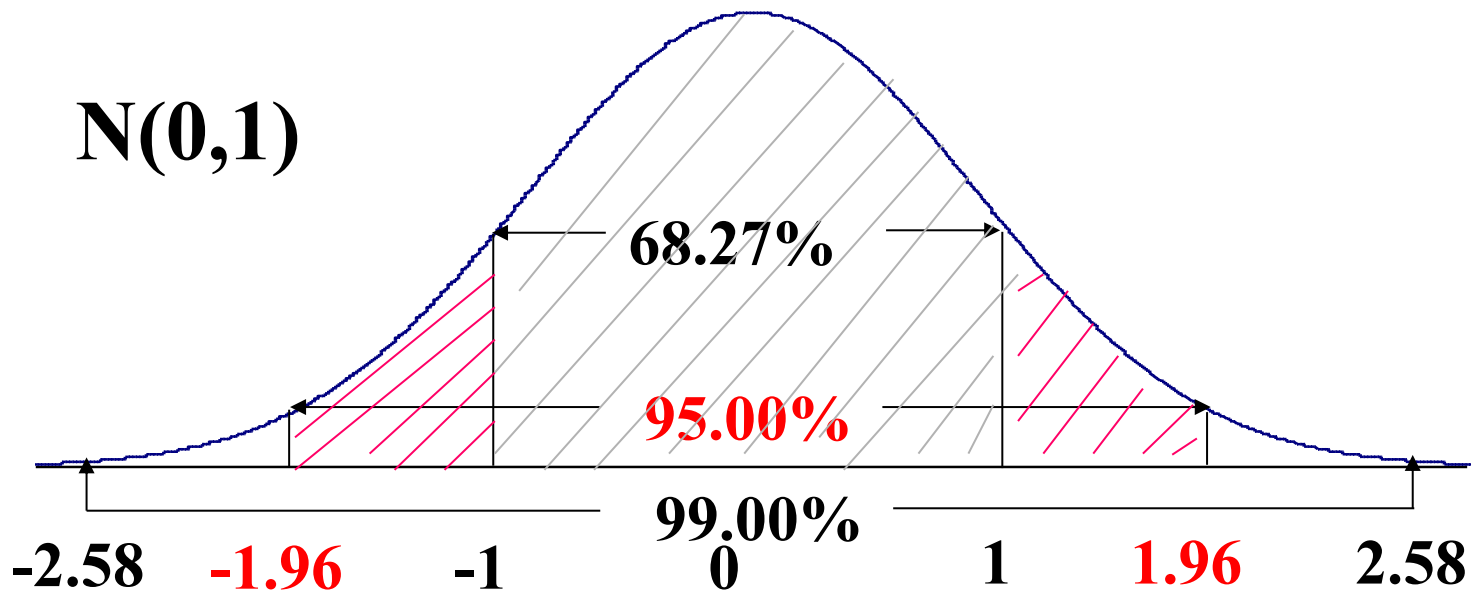
$$\begin{aligned} P(-2.58 < u < 2.58) &= \Phi(2.58) - \Phi(-2.58) \\ &= [1 - \Phi(-2.58)] - \Phi(-2.58) = 1 - 2\Phi(-2.58) \\ &= 1 - 2 \times 0.0049 \approx 0.99 \end{aligned}$$



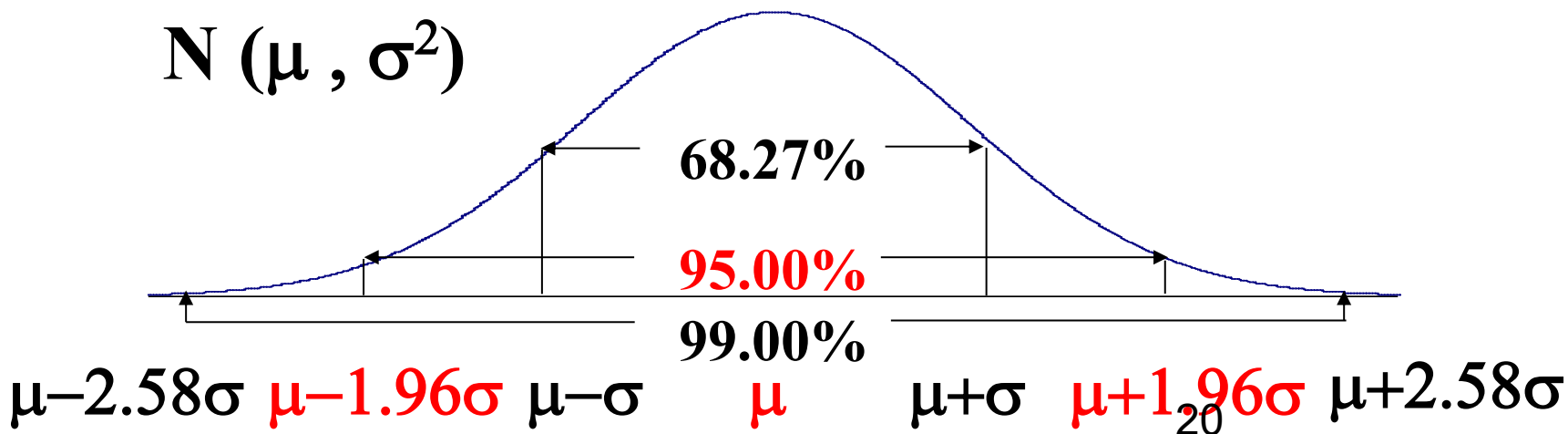
曲线下面积分布规律



$N(0,1)$



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



曲线下面积分布规律



N(0,1)

N (μ , σ^2)

标准正态分布

正态分布

面积或概率

-1~1

$\mu \pm \sigma$

68.27%

-1.96~1.96

$\mu \pm 1.96\sigma$

95.00%

-2.58~2.58

$\mu \pm 2.58\sigma$

99.00%



计算

1. 由频数分布判断，基本符合正态分布规律。
2. 计算均数、标准差，

$$\bar{X} = \frac{f_1 X_1 + f_2 X_2 + f_3 X_3 + \cdots + f_k X_k}{f_1 + f_2 + f_3 + \cdots + f_k} = \frac{\sum f X_i}{\sum f_i}$$

$$\text{频数表样本标准差 } S = \sqrt{\frac{\sum f X^2 - (\sum f X)^2 / \sum f}{\sum f - 1}}$$

$$\bar{X} = \frac{7982.0}{108} = 73.9(\text{g/L}),$$

$$S = \sqrt{\frac{591524.0 - 7982.0^2 / 108}{108 - 1}} = 3.9(\text{g/L})$$



3. 标准正态分布的转换。样本量较大, 故用样本均值代替 μ , S 代替 σ 。

$$u = \frac{X - \mu}{\sigma} \approx \begin{cases} \frac{68.0 - 73.9}{3.9} = -1.51 = u_1 \\ \frac{78.0 - 73.9}{3.9} = 1.05 = u_2 \end{cases}$$



4. 估计 u_1 和 u_2 的分布函数，查附表1，得

$\phi(-1.51) = 0.0655$ ，故 $P(X < 68.0) = 0.0655$ ，

$\phi(1.05) = 1 - \phi(-1.05) = 1 - 0.1469 = 0.8531$ ，

故 $P(X < 78.0) = 0.8531$ ， $P(X \geq 78.0) = 0.1469$

5. 下结论。

108例血红蛋白.sav [数据集1] - IBM SPSS Statistics 数据编辑器

文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助

报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(O)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(E)
降维
度量(S)
非参数检验
预测(D)
生存函数
多重响应
缺失值分析
多重回归
复杂抽样
质量控制
ROC曲线

频率(F)...
描述(D)...
探索(E)...
交叉表(C)...
比率(R)...
P-P图(P)...
Q-Q图(Q)...

名称	数值	小数	标签	值	缺失	列	右
1	血清蛋白			无	无	15	
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							

探索

因变量列表(D): 血清蛋白

因子列表(F):

标注个案

输出
 两者都(B) 统计量 图

探索: 图
 箱图
 按因子水平分组(F) 不分组(D) 无
 带检验的正态图(O)

描述性(D)
 茎叶图(S) 直方图(H)

统计量(S)...
绘制(T)...
选项(O)...
Bootstrap(B)...

确定 粘贴(P) 重置(R)

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
血清蛋白	.057	108	.200 [*]	.993	108	.856

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction