第三章 理论分布与抽样分布

统计学在中国的传播

作业：简述统计学在中国的传播。

基本概念

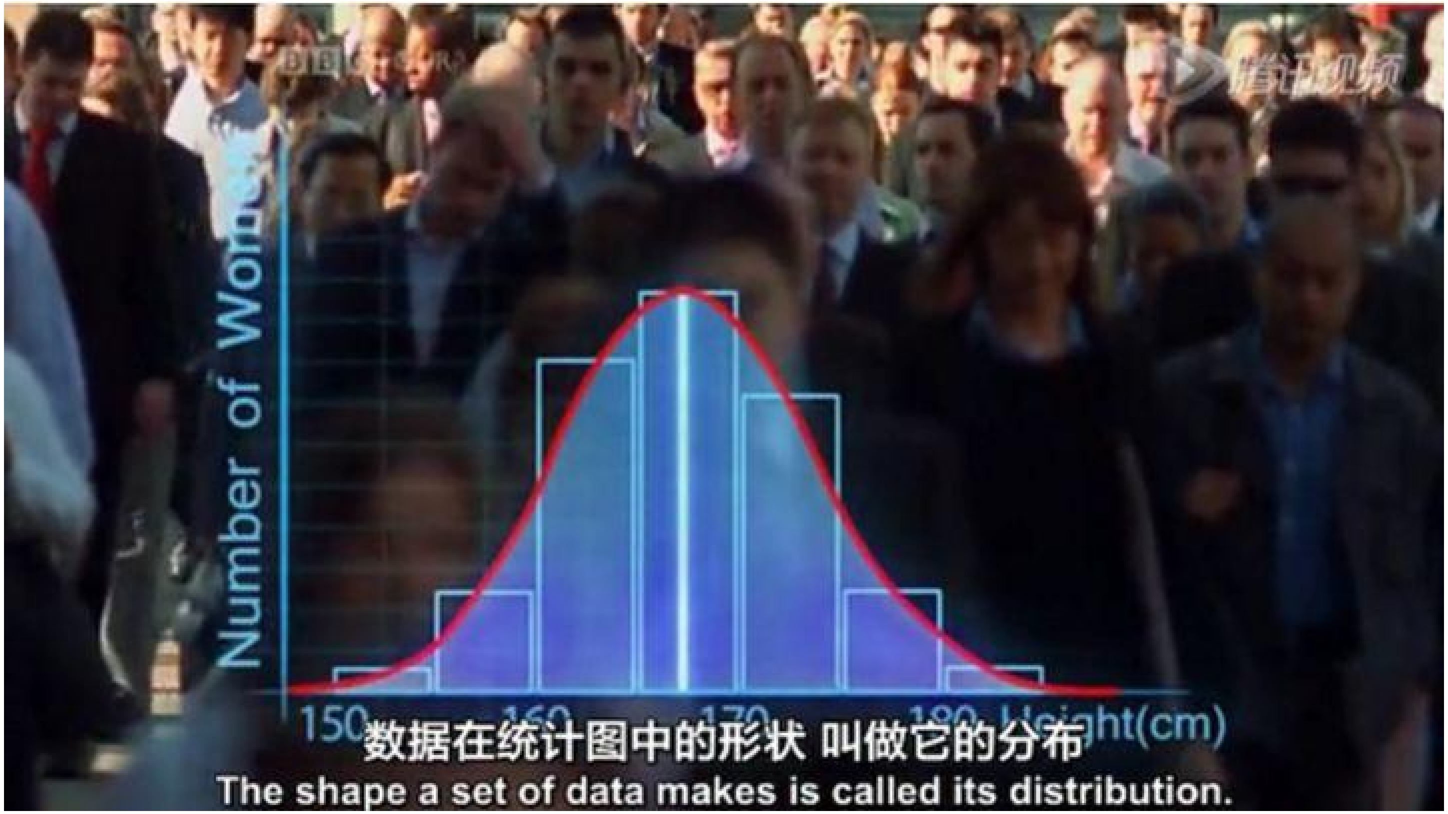
什么是随机变量？

是一个量化随机事件的函数，它将随机事件每一个可能出现的试验结果赋予一个数字；

分离散随机变量（数值间有间隔）和连续随机变量（有无数个结果）；

一般用 X 表示。（视频）

什么是分布？

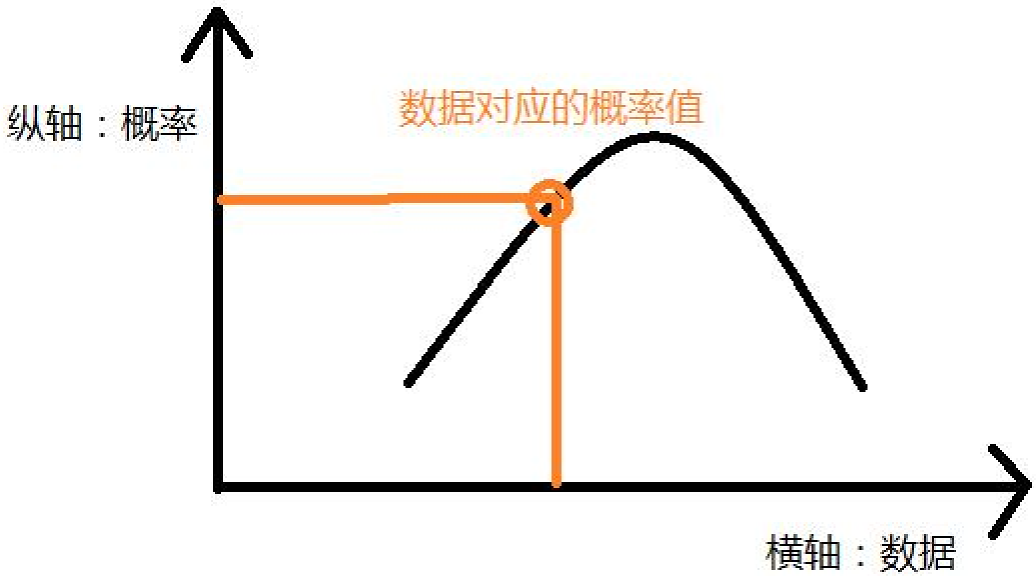

Number of Women
150
Height(cm)
数据在统计图中的形状 叫做它的分布
The shape a set of data makes is called its distribution.

什么是概率分布？

用统计图来表示随机变量所有结果和对应结果发生的概率；

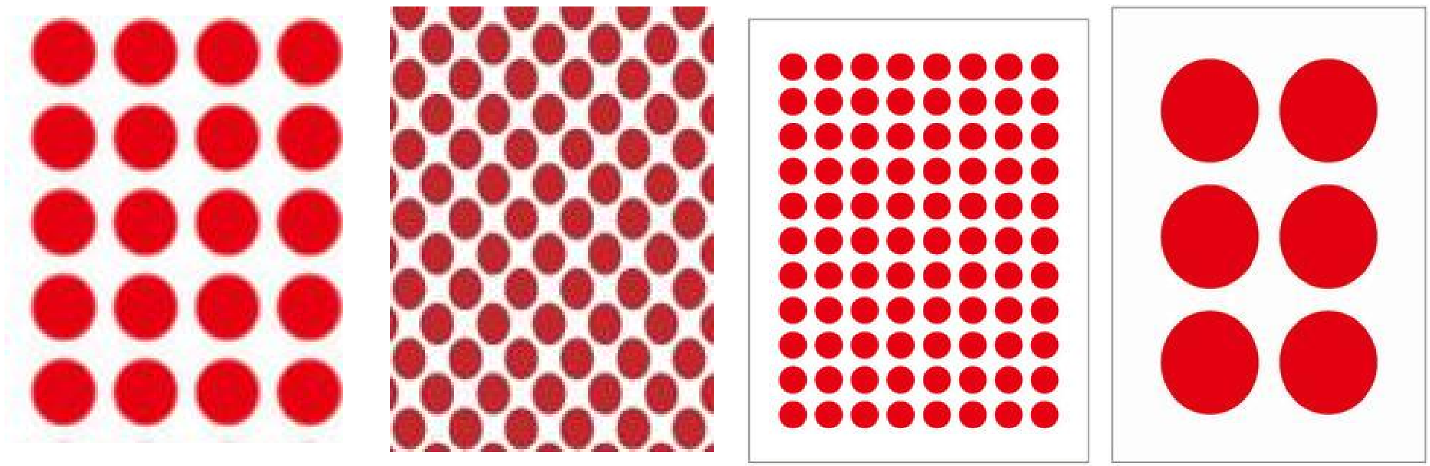
概率分布=随机变量+概率+分布（在统计图中的形状）
（视频）

概率分布就是在统计图中表示概率，横轴是数据的值，纵轴是横轴上对应数据值的概率。

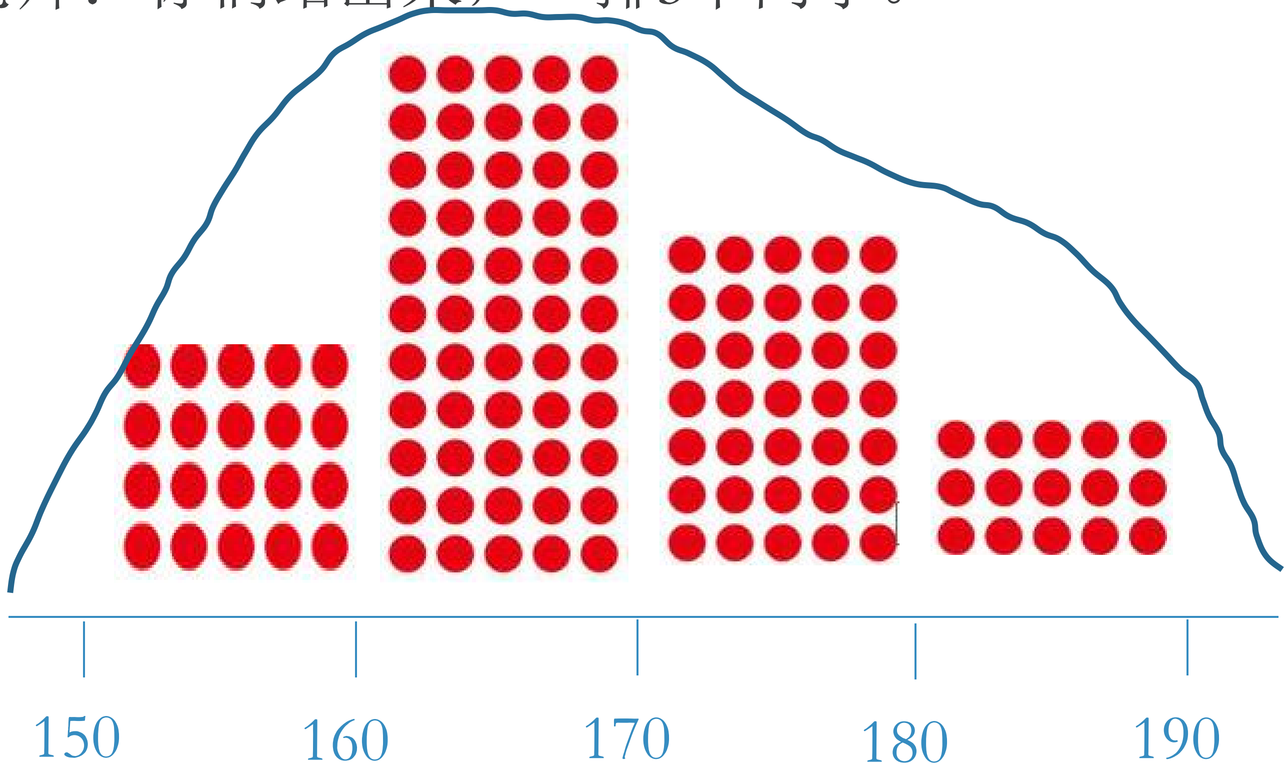


什么是概率密度函数？

现在有大小相同的四个房间，分别写着身高为150-160，160-170,170-180,180-190之间。

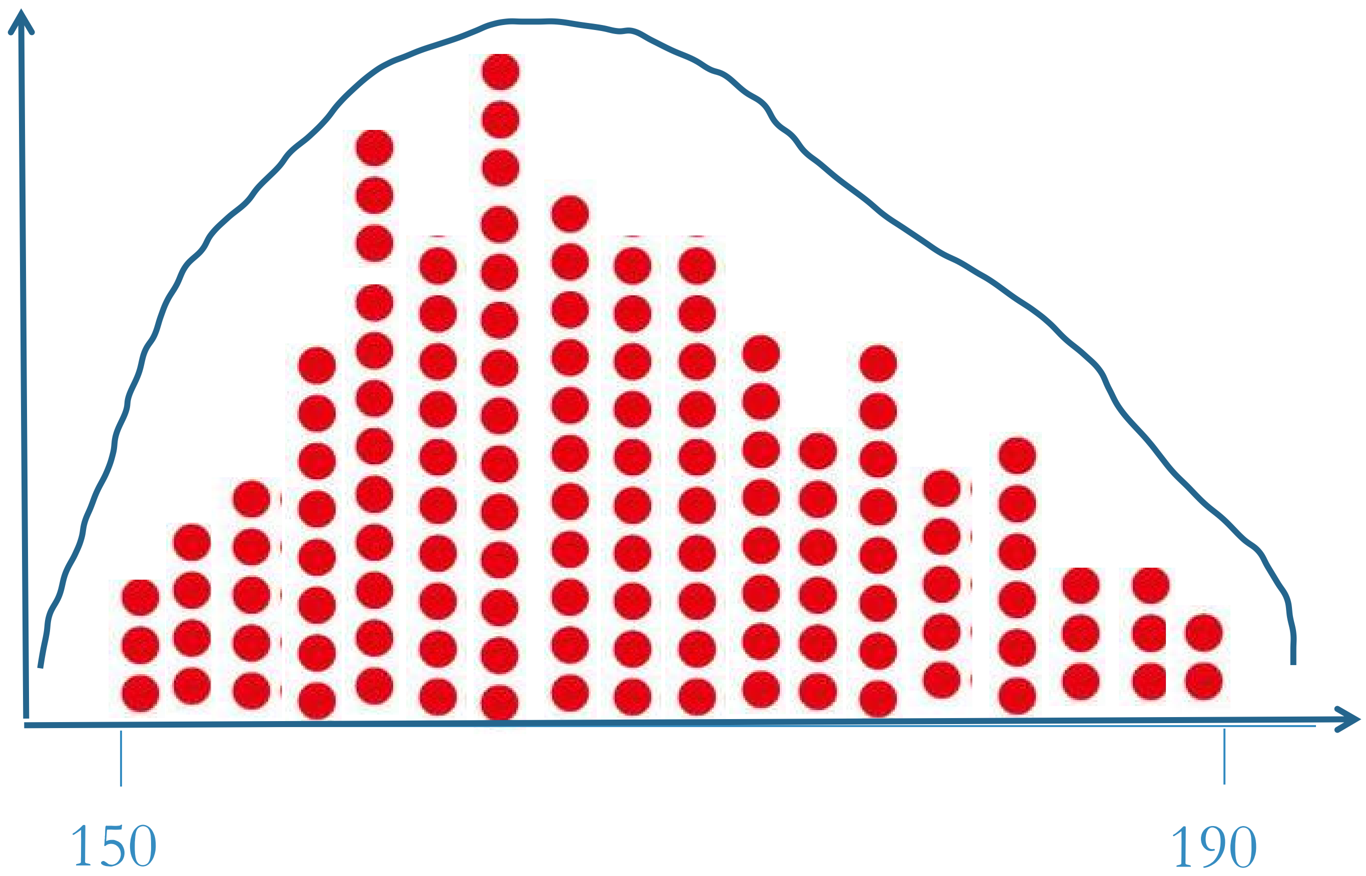


老师：你们站出来，一排5个同学。



这个区间内的概率，就是这个长方形的面积。

老师说：我觉得这个区间有点大了，我现在分40组，150-161,161-162,162-163......179-190,一排只站一个人。



变成这个样子，可以经过无限的缩小区间，我们就得到了一个曲线，就是概率密度函数了。也就是说小长条的面积，就是概率，就是里面站了多少个人的百分比。

3.1.3 正态分布

视频

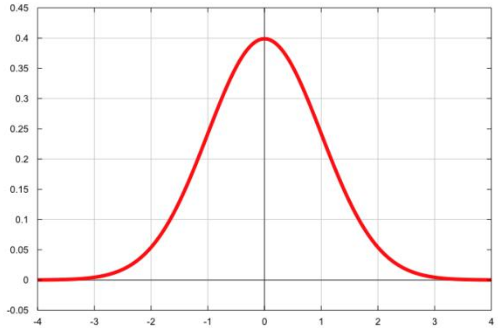
正态概率分布是连续型随机变量中最重要的分布。世界上绝大部分的分布都属于正态分布，人的身高体重、考试成绩、降雨量等都近似服从。

3.1.3.1 正态分布的定义及其特征

1.定义

- 若连续型随机变量x的概率密度函数为：

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



- μ为平均数，σ^2为方差，则称随机变量x服从参数为μ与σ^2的正态分布，记为x~N（μ，σ^2），其概率分布函数F（x）为：

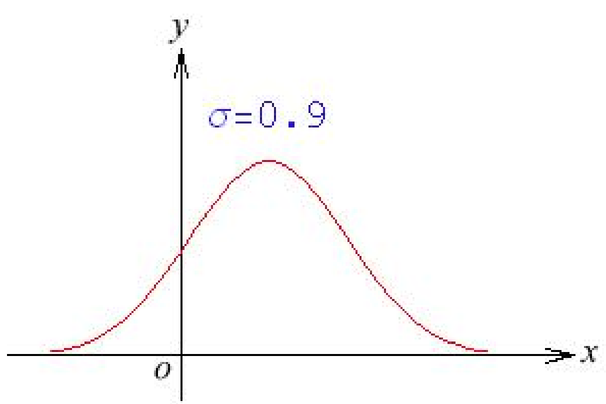
$$\mathbf{F}(\mathbf{x}) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp(-\frac{(x-\mu)^2}{2\sigma^2}) dx$$

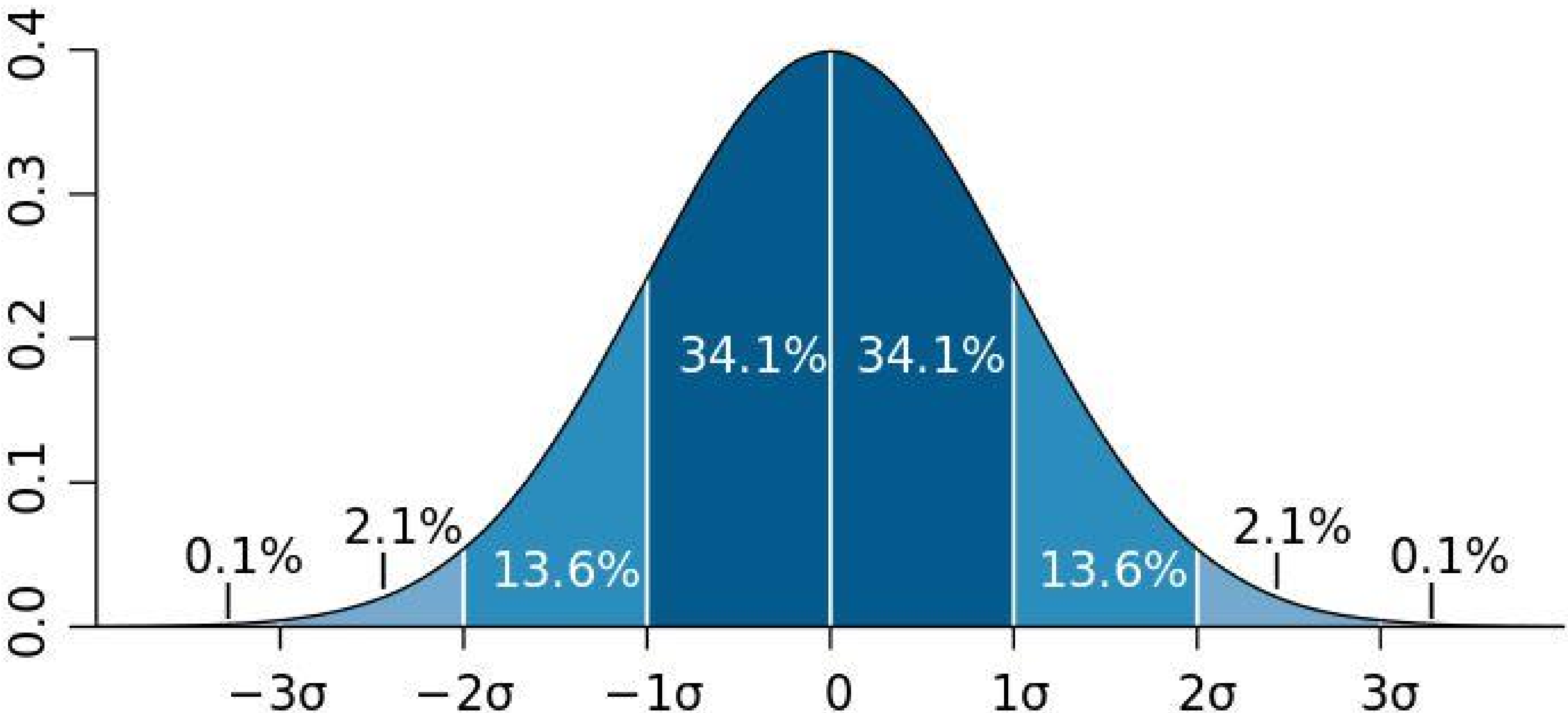
2.特征

- 正态分布曲线是以均数μ为中心左右对称分布的单峰悬钟型曲线。在平均数的左右两侧，只有（x-μ）的绝对值相等，f(x)值就相等。

- f(x)在x=μ处达到最大值，且$f(\mu)=1/\sigma\sqrt{2\pi}$。

- f(x)是非负函数，以横轴为渐近线，分布为-∞到+∞，且曲线在μ±σ处各有一个拐点。



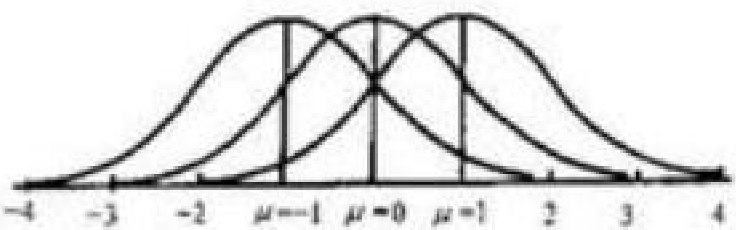


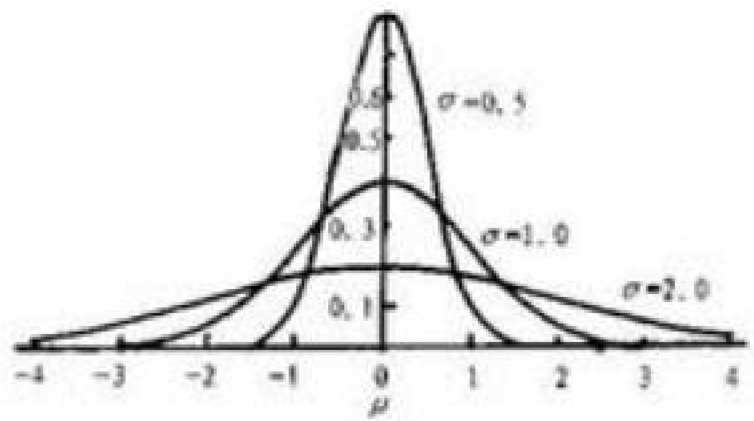
正态随机变量有69.3%的值在均值加减一个标准差的范围内，95.4%的值在两个标准差内，99.7%的值在三个标准差内。

工业控制3σ原理的依据

- **正态分布曲线因参数μ和σ²**的不同而表现出一系列曲线，所以正态分布曲线是一个曲线族，不是一条曲线。（讨论）

- **μ是位置参数，σ²是总体的变异度，是形状参数。σ²越大，曲线越‘胖’，说明数据比较分散，反之亦然。**

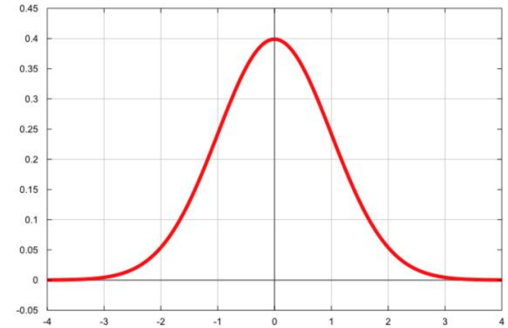




$$N(\mu,\sigma^2)$$

- 曲线f(x)与横轴之间所围成的面积等于 ？ (讨论)

3.1.3.2 标准正态分布



- 由于正态分布是依赖于参数**μ和σ²的一簇分布，正态曲线的位置和形态随其不同而变化。**
- **因此，需要将一般的N（μ，σ²）转换为μ=0，σ²=1的正态分布，也即通常称为标准正态分布。**
- **其概率密度函数和分布函数分别为：** $\varphi(\mathrm{u})$和 $\Phi(\mathrm{u})$

$$\varphi(\mathrm{u})=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \qquad \Phi(\mathrm{u})=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-\frac{u^2}{2}}du$$

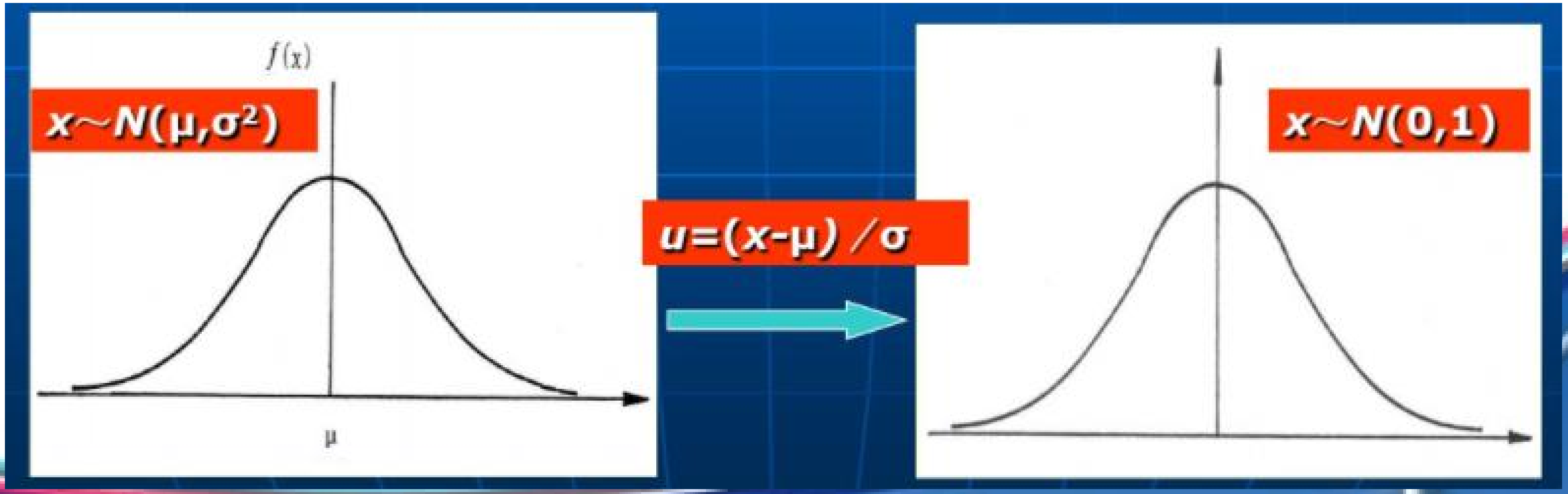
- **此时，称随机变量u服从标准正态分布，**

记做u~N(0,1)。

- 任何一个服从正态分布N(μ，σ²)的随机变量x都可以通过下式所示的标准化变换，将其转化为服从标准正态分布的随机变量u。

$$u=（x-\mu）/\sigma$$

- u称为标准正态变量或标准正态离差。



3.1.3.3 正态分布的概率计算

1.标准正态分布的概率计算

- 设u服从标准正态分布，则u在[u_1，u_2)内取值的概率为：

$$P(\mathrm{u}_1 \le \mathrm{u} < u_2) = \frac{1}{\sqrt{2\pi}}\int_{u_1}^{u_2} e^{-\frac{u^2}{2}} du = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u_2} e^{-\frac{u^2}{2}} du - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u_1} e^{-\frac{u^2}{2}} du$$

$$= \Phi(u_2) - \Phi(u_1)$$

- $\Phi(u_2)$和$\Phi(u_1)$ 可由表查得。视频

◈由上式及正态分布的对称性可推出下列关系，再用正态分布表能方便的计算出以下概率：

$$P(0 \le \mathrm{u} < u_1) = \Phi(u_1) - 0.5$$

$$P(u \ge u_1) = \Phi(\text{-}u_1)$$

$$P(|u| \ge u_1) = 2\Phi(\text{-}u_1)$$

$$P(|u| < u_1) = 1\text{-}2\Phi(\text{-}u_1)$$

$$P(u_1 \le \mathrm{u} < u_2) = \Phi(u_2) - \Phi(u_1)$$

板书

对于标准正态分布，特殊区间的概率为：

$P(-1 \le u < 1) = 0.6826$

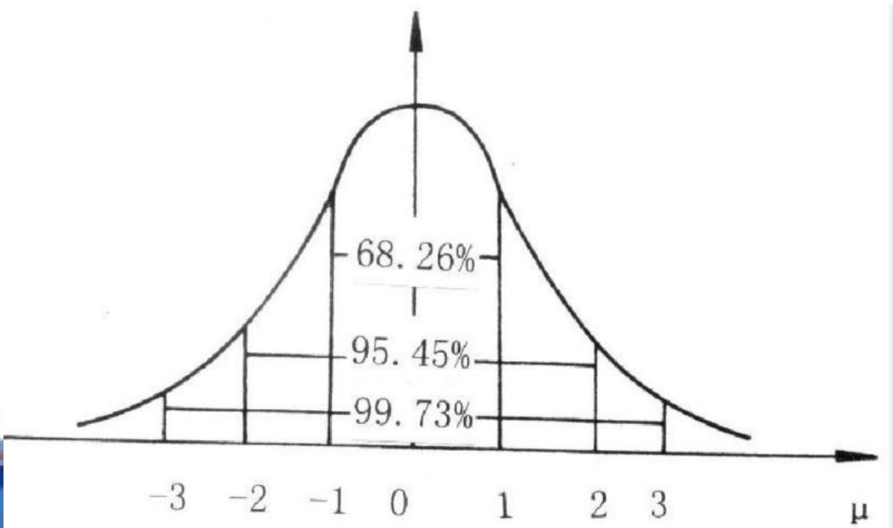
$P(-2 \le u < 2) = 0.9545$

$P(-3 \le u < 3) = 0.9973$

$P(-1.96 \le u < 1.96) = 0.95$

$P(-2.58 \le u < 2.58) = 0.99$

标准正态分布的三个常用概率如图示



u变量在上述区间以外取值的概率分别为：

$$P(|u| \geq 1) = 2\Phi(-1) = 1 - P(-1 \leq u < 1)$$

$$= 1 - 0.6826 = 0.3174$$

$$P(|u| \geq 2) = 2\Phi(-2)$$

$$= 1 - P(-2 \leq u < 2)$$

$$= 1 - 0.9545 = 0.0455$$

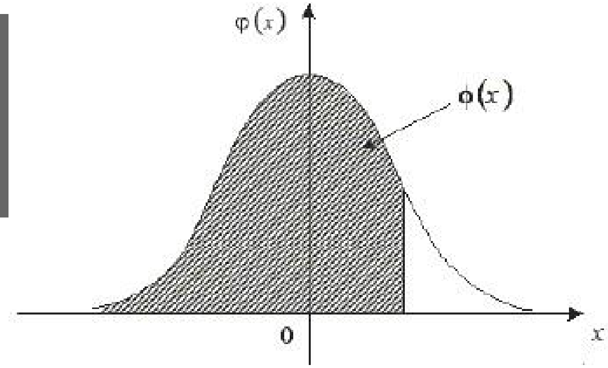
$$P(|u| \geq 3) = 1 - 0.9973 = 0.0027$$

$$P(|u| \geq \underline{1.96}) = 1 - 0.95 = 0.05$$

$$P(|u| \geq \underline{2.58}) = 1 - 0.99 = 0.01$$

统计检验中常用

2.一般正态分布的概率计算



- 正态分布密度曲线和横轴围成的一个区域，其面积为1。这实际上表明了随机变量x在（-∞，+∞）之间取值是一个必然事件，其概率为1。

- 若随机变量x服从正态分布N(μ，σ²)，则x 的取值落在任意区间[x_1，x_2）的概率，记为P（$x_1 \leq x < x_2$），即

$$P（x_1 \leq x \leq x_2） = \frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2}} dx$$

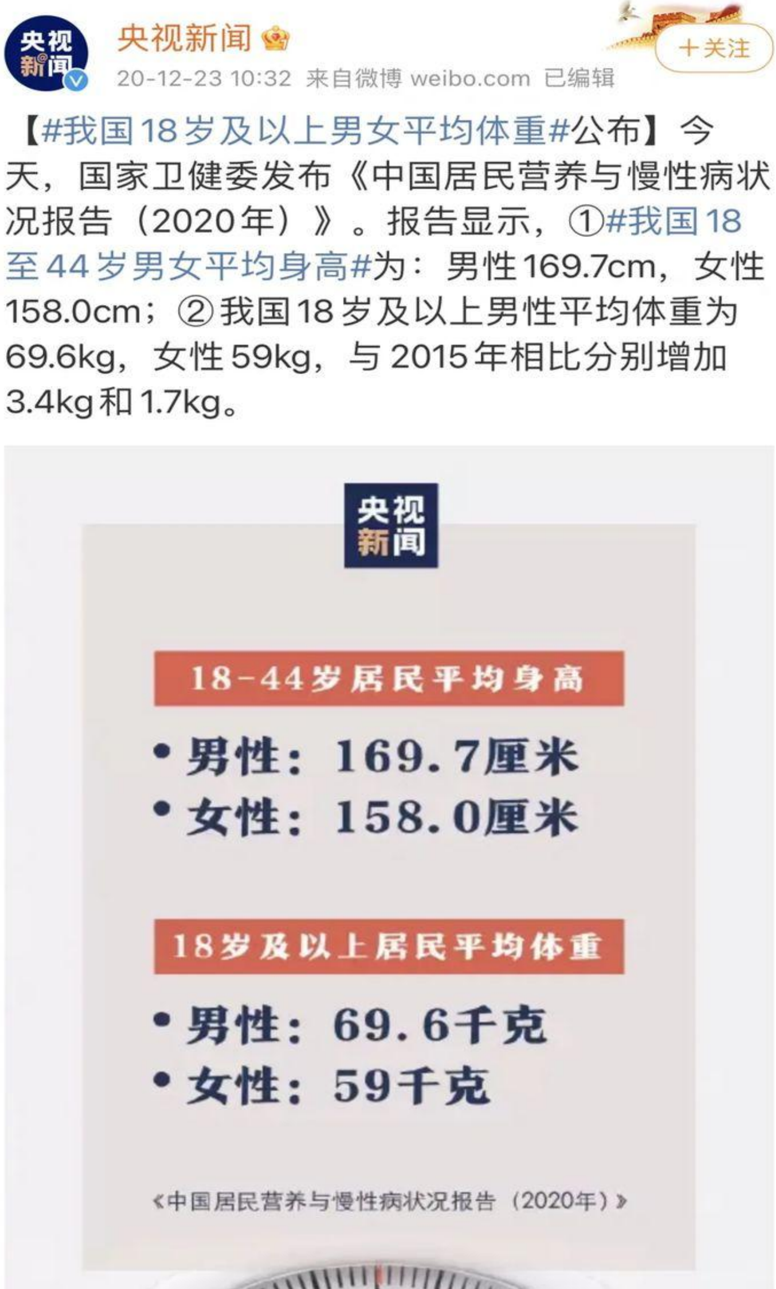
- **变换u=（x-μ）/σ**，得dx=σdu，得：

$$P(x_1 \le x < x_2) = \frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{(x_1-\mu)/\sigma}^{(x_2-\mu)/\sigma} \sigma e^{-\frac{u^2}{2}}du$$

$$= \frac{1}{\sqrt{2\pi}}\int_{u_1}^{u_2} e^{-\frac{u^2}{2}}du$$

- 这表明，服从正态分布N（μ，σ^2）的随机变量x落在**[x_1，x_2)内的概率等于**服从标准正态分布随机变量u落在[（x_1-μ）/σ，（x_2-μ）/σ]，即**[u_1，u_2)的概率**。
- 因此，计算一般正态分布的概率时，只要将区间的上下限标准化，就可用标准正态的方法计算。

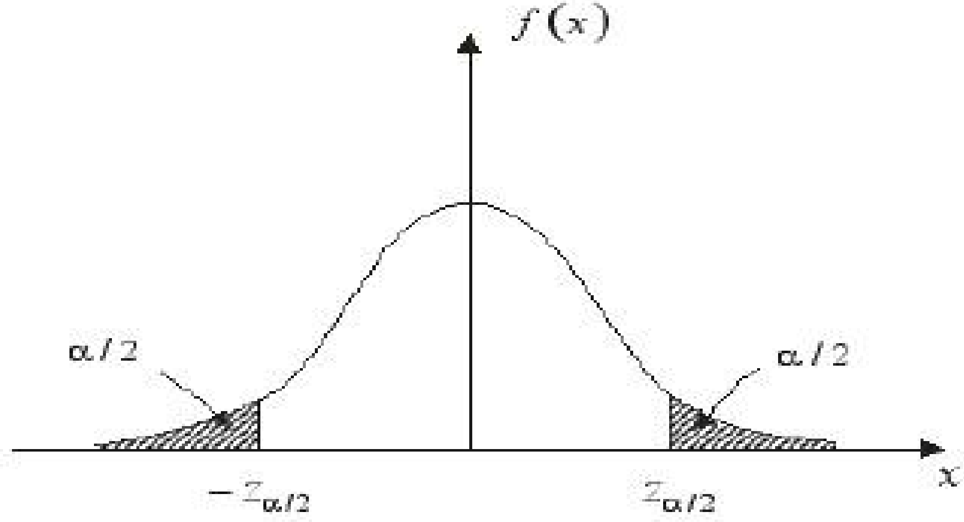
学会正态分布有什么用呢？



假设我们学校女生的平均身高为158.0cm,方差为25cm。

x~N（158，25）

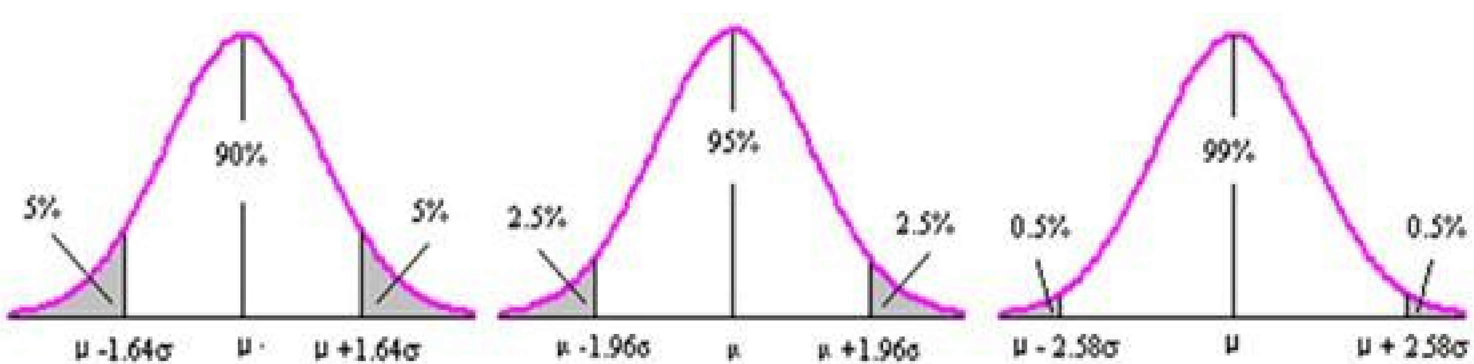
身高在xxxcm下的概率为多少呢？



- 在统计分析中，不仅注意随机变量x在平均数加减不同倍数便准差区间（μ-kσ，μ+kσ）内取值的概率，而且也很关心x在此区间外取值的概率。
- 我们把随机变量x在平均数μ加减不同倍数标准差区间之外取值的概率称作两尾（双侧）概率，记做a。
- 对应于双侧概率，可以求得随机变量x小于μ-kσ或大于μ+kσ的概率，称为单侧概率，记做a/2.

- 例如，x在（μ-1.96σ，μ+1.96σ）之外取值的两尾概率为0.05，而单尾概率为0.025，即P（x<μ-1.96σ）=P（μ+1.96σ)=0.025。

- x在（μ-2.58σ，μ+2.58σ）之外取值的双尾概率为0.01，单尾为0.005。



3.1 理论分布

3.1.1 二项式分布

- 二项式分布是最重要的离散型分布之一，它在理论和实践应用上都有重要的地位。
- 产生这种分布重要的实践源泉是贝努利试验。

3.1.1.1 贝努利试验及其概率公式

1.贝努利试验

- 很多实际问题中，我们感兴趣的是试验中某件事A是否发生。比如食品抽样检验中，样品是否合格。
- 只有两种结果或者说只有两个基本事件A与$\bar{A}$。
- 像这样只有两种可能结果的随机试验称为贝努利试验。

3.1.1.1 贝努利试验及其概率公式

- 我们常把贝努利试验中的两种结果分别称为“成功”和“失败”，即A（成功）与$\bar{A}$（失败）构成整个事件。
- 贝努利试验在完全相同的条件下独立重复n次，并作为一个随机试验称之为n重（次）贝努利试验。

3.1.1.1 贝努利试验及其概率公式

2.贝努利试验的概率公式

- 在贝努利试验中，事件A可能发生也可能不发生。用随机变量 x 表示贝努利试验的两种结果，并记当A发生时，x=1；当A不发生（即$\bar{A}$发生）时，x=0。前者概率为p，后者为q，则贝努利试验的概率公式为：

$$\begin{cases} \mathrm{P}(x=1)=p \\ \mathrm{P}(x=0)=q \end{cases} \text{其中 } x=\begin{cases} 1 \text{ 出现成功} \\ 0 \text{ 出现失败} \end{cases}$$

- 上式也称为两点分布。

3.1.1.2 二项式的定义及其特点

1.定义

- 在n重贝努利试验中，事件A可能发生的次数是0，1，…，n次，考虑n重贝努利试验中正好发生k（0≤k≤n）次的概率，记为$P_n(k)$。
- 事件A在n次试验中正好发生k次共有 种情况。
- 由贝努利试验的独立性可知，A在某k次试验中发生而在其余的n-k次试验中不发生的概率为 。
- 由概率论定理有：$P_n(k)=$ （k=0，1，…，n）
- 上式即为，n次贝努利试验中事件A正好发生k次的概率，也称为二项概率公式。

3.1.1.2 二项式的定义及其特点

- 根据二项概率公式，二项分布可定义为：
- $P(x=k)=P_n(k)=C_n^k p^k q^{n-k}$（k=0，1，…，n）
- 式中p>0，q>0，p+q=1，则称随机变量x服从参数为p和q的二项分布（binomial distribution），记为x~B(n，p）。