第三讲　常用概率分布
(PROBABILITY DISTRIBUTIONS)

一般数学表示方法

- 概率数学表示方法
  - X —— 符合某种概率分布的随机变量
  - x —— 随机变量X的一个实例, rnorm, …
  - f —— 随机变量X的概率密度分布函数（pdf）, dnorm, ….
  - F —— 累积密度分布函数（cdf）, pnorm, …
  - P(X=k) —— 概率群分布函数（probability mass distribution）

常用概率分布

- 离散概率分布（Discrete）
  - 离散均匀分布（discrete uniform）
  - 伯努利分布（Bernoulli）
  - 二项式分布（Binomial）
  - 泊松分布（Poisson）
  - 超几何分布（Hypergeometric）
- 连续概率分布（continuous）
  - 均匀分布（Uniform）
  - 正态分布（Normal/Gaussian）
  - 指数分布（Exponential）
  - Gamma分布
  - Beta分布
  - Gumbel分布

离散概率分布

离散均匀分布（discrete uniform）

- 概念：每次抽样存在多种可能结果，每种结果出现的概率完全一致，也就是X的取值空间为一个有穷的集合S={k1, k2, …, kn}且对于任意i€1…n, P(X=ki)=1/n
- R代码

```
>>> x <- sample(c(1,2,6,8,9), 10000, prob=rep(0.2,5),
                replace=TRUE)
>>> table(x)
## 另一种方法：用index
>>> idx <- runif(10000,0,5); idx <- ceiling(idx)
>>> x <- c(1,2,6,8,9)[idx]
>>> table(x)
```

伯努利试验（Bernoulli）

- 概念：仅存在两种可能结果的一次试验
- 举例：扔硬币，正面朝上（“H”），反面朝上（“T”）
  - P(X=“H”) = π, P(X=“T”) = 1- π
- R代码：进行1000次伯努利试验

```
>>> outcome <- sample(c(“T”,”F”), 1000,
                      prob=c(0.8, 0.2), replace=TRUE)
>>> ot <- table(outcome)
>>> ot <- ot/sum(ot)
```

二项式分布（Binomial）

- 概念

  重复进行n次独立的伯努利试验(可能结果为{1, 0}，出现1的概率是p)，n次结果中有X次1的概率分布就是二项式分布，X的可能取值范围是{0,1,…,n}，记作X ~ B(n,p)

- 二项式分布的概率函数

$$P(X) = C_n^X \pi^X (1-\pi)^{n-X}$$

$$C_n^X = \frac{n!}{X!(n-X)!}$$

二项式分布示例1

一个袋子里有5个乒乓球，其中2个黄球，3个白球，我们进行有放回的摸球游戏。因此每一次摸到黄球的概率是0.4，摸到白球的概率是0.6。

这个实验有三个特点：

1. 各次摸球是彼此独立的；
2. 每次摸球只有二种可能的结果，或黄或白；
3. 每次摸到黄球（或摸到白球）的概率是固定的。

具备这三点后，*n*次中有*X*次摸到黄球（或白球）的概率分布就是二项分布。

二项分布在医学研究中的应用

医学研究中很多现象观察结果是以两分类变量来表示的，如阳性与阴性、治愈与未愈、生存与死亡等等。如果每个观察对象阳性结果的发生概率均为π，阴性结果的发生概率均为（1—π）；而且各个观察对象的结果是相互独立的，那么，重复观察n个人，发生阳性结果的次人数X的概率分布为二项分布，记作B（X；n，π）。

二项式分布示例2

用针灸治疗头痛，假定结果不是有效就是无效，每一例有效的概率为$\pi=0.6$。某医生用此方法治疗头痛患者3例，2例有效的概率是多少？至少1例有效的概率呢？

$$C_3^2 0.6^2 (1-0.6)^{(3-2)} = 0.432$$

```
>>> n <- 3; pi <- 0.6
### p2 <- choose(3,2)*0.6^2*(1-0.6)^(3-2)
>>> p2 <- dbinom(2, n, pi)
>>> p123 <- sum(dbinom(c(1,2,3), n, pi))
### p123 <- 1-pbinom(0, n, pi)
### p123 <- pbinom(0, n, p, lower.tail=F)
```

二项式分布

治疗3例可能的有效例数及其概率

有效人数(x)	C_3^X	π^x	$(1-\pi)^{n-x}$	出现该结果概率$P(x)$
0	1	$0.6^0=1$	$0.4\times0.4\times0.4$	0.064
1	3	0.6	0.4×0.4	0.288
2	3	0.6×0.6	0.4	0.432
3	1	$0.6\times0.6\times0.6$	0.4^0	0.216

```
>>> x <- 0:3; n<- 3; pi <- 0.6
>>> p <- dbinom(x, n, pi)
>>> cp1 <- cumsum(p)
>>> cp2 <- pbinom(x, n, pi)
```

二项式分布的特征

二项分布的图形特征

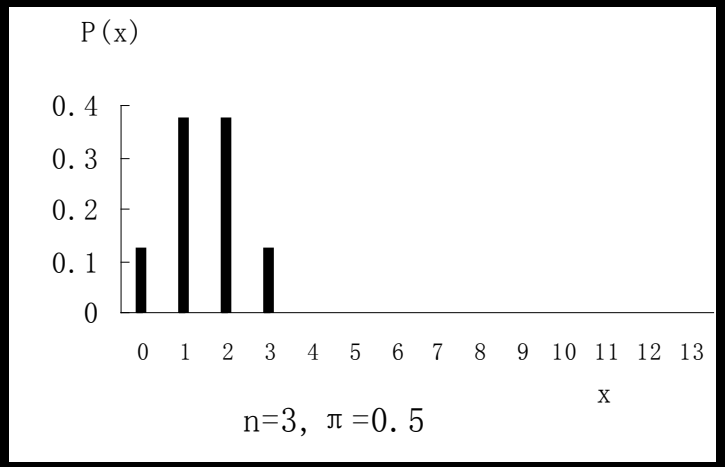
π接近0.5时，图形是对称的；

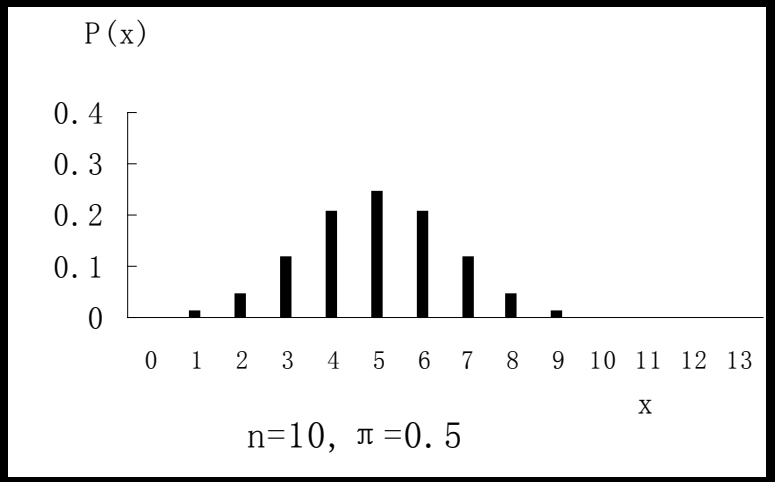
π离0.5愈远，对称性愈差，但随着n的增大，分布趋于对称

当$n\to\infty$时，只要π不太靠近0或1，当nP和$n(1-P)$都大于5时，二项分布近似于正态分布。

二项分布图形取决于π与n，高峰$\mu=n\pi$处

二项式分布

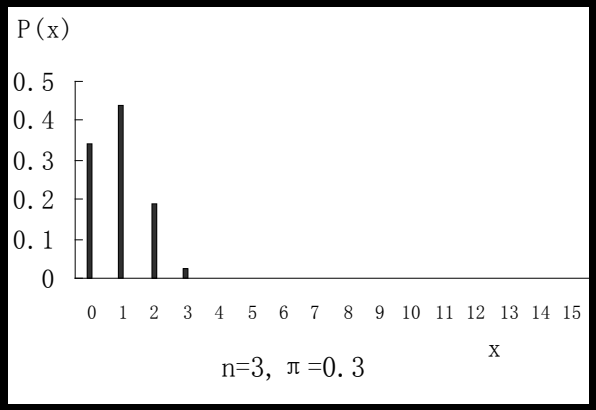


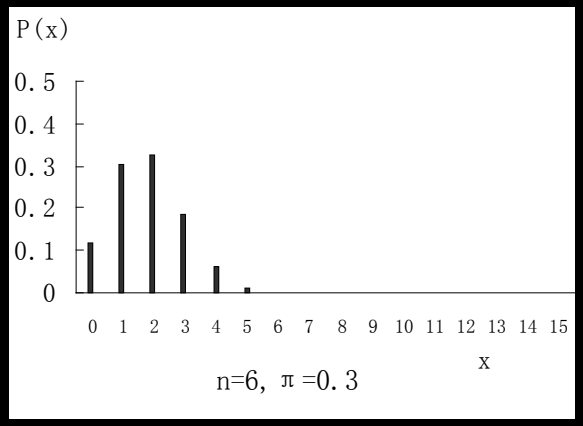


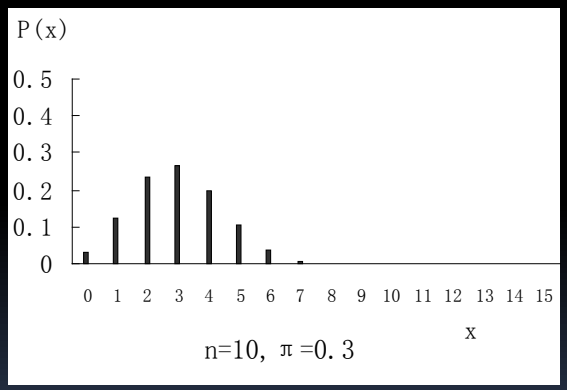
π=0.5时,不同n值对应的二项分布

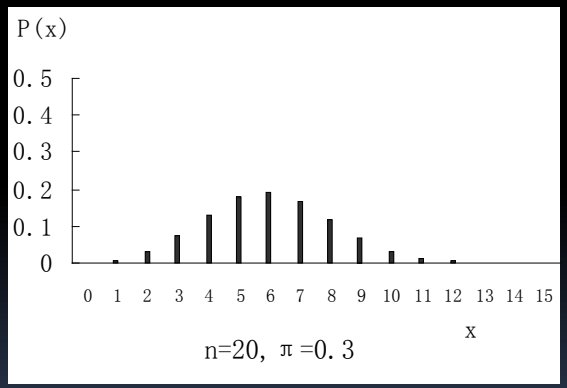
```
>>> x1 <-  0:3; x2 <- 0:10
>>> p1 <- dbinom(x1, 3, .5); p2 <- dbinom(x2, 10, .5)
>>> par(mfrow=c(1,2));
>>> plot(x1, p1, type='h');  plot(x2, p2, type='h')
```

二项式分布









π=0.3时, 不同*n*值对应的二项分布

二项式分布

- 二项分布的均数和标准差

总体均数：$\mu = n\pi$

方差：$\sigma^2 = n\pi(1-\pi)$

标准差：$\sigma = \sqrt{n\pi(1-\pi)}$

```
>>> n <- 100; pi <- 0.6
>>> x <- rbinom(1000, n, pi)
>>> mu <- mean(x); sigma2 <- var(x); sigma <- sd(x)
```

Poisson分布

概念

Poisson分布也是一种离散型分布，用以描述罕见事件发生次数的概率分布。医学上人群中出生缺陷、多胞胎、染色体异常等事件等都是罕见的，可能发生这些事件的观察例数*n*常常很大，但实际上发生类似事件的数目却很小很小。

Poisson分布

- Poisson分布可以看作是发生的概率π（或未发生的概率$1-\pi$）很小，而观察例数n很大时的二项分布。

- 除二项式分布的三个基本条件以外，Poisson分布还要求π或（$1-\pi$）接近于0或1（例如<0.001或>0.999）。

Poisson分布

Poisson分布的概率函数为

$$P(X)=e^{-\lambda}\frac{\lambda^{X}}{X!}$$

式中，$\lambda=n\pi$ 为Poisson分布的总体均数，X为观察单位内某稀有事件的发生次数。

Poisson分布

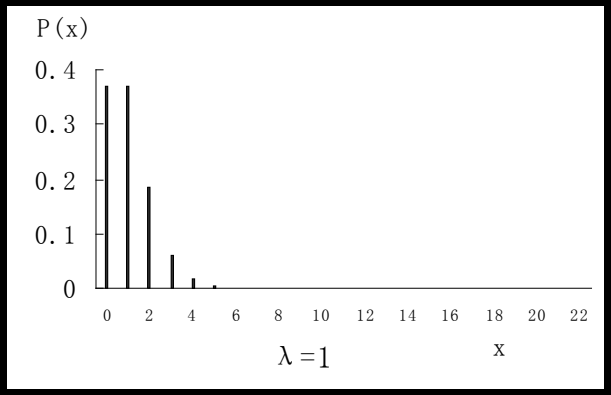
Poisson分布当总体均数λ值小于5时为偏峰，λ愈小分布愈偏，随着λ增大，分布趋向对称。

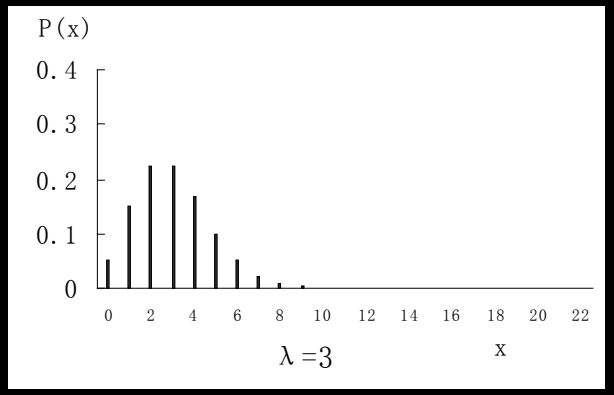
- Poisson分布有以下特性：

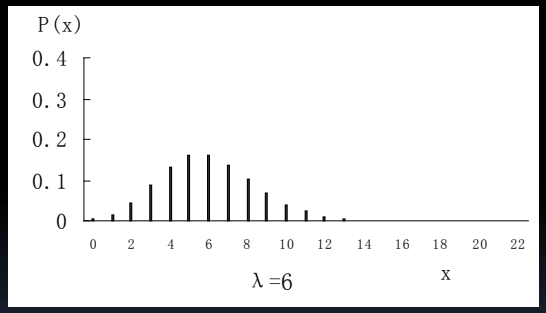
（1）Poisson分布的总体均数与总体方差相等，均为λ

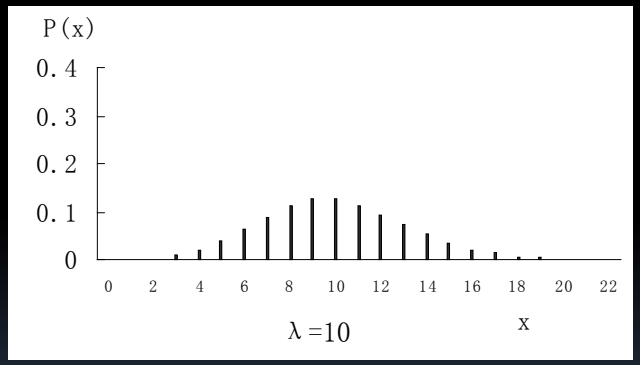
（2）Poisson分布的观察结果有可加性

Poisson分布









λ取不同值时的Poisson分布图

Poisson分布

- 如果某地新生儿先天性心脏病的发病概率为8‰，那么该地120名新生儿中有4人患先天性心脏病的概率有多大？

$$\lambda = n\pi = 120 \times 0.008 = 0.96$$

$$P(X=4) = \frac{e^{-0.96}0.96^4}{4!} = 0.014$$

```
>>> n <- 120; pi <- 0.008
>>> x <- 4
>>> p1 <- dpois(x, n*pi)
>>> p2 <- dbinom(x, n, pi)
```

Poisson分布

单侧累计概率计算

如果稀有事件发生次数的总体均数为λ，那么该稀有事件发生次数至多为*k*次的概率

$$P(X \le k) = \sum_{X=0}^{k} P(X) = \sum_{X=0}^{k} e^{-\lambda} \frac{\lambda^X}{X!}$$

发生次数至少为*k*次的概率

$$P(X \ge k) = 1 - P(X \le k-1)$$

Poisson分布

- 上例中，至多有4人患先天性心脏病的概率有多大？至少有 5 人患先天性心脏病的概率有多大？

至多有4人患先天性心脏病的概率

$$P(X \le 4) = \sum_{X=0}^{4} P(X) = \sum_{X=0}^{4} \frac{e^{-0.96} 0.96^{X}}{X!}$$

$$= \frac{e^{-0.96} 0.96^{0}}{0!} + \frac{e^{-0.96} 0.96^{1}}{1!} + \frac{e^{-0.96} 0.96^{2}}{2!} + \frac{e^{-0.96} 0.96^{3}}{3!} + \frac{e^{-0.96} 0.96^{4}}{4!} = 0.997$$

至少有 5 人患先天性心脏病的概率为

$$P(X \ge 5) = 1 - P(X \le 4) = 1 - 0.997 = 0.003$$

Poisson分布

- 实验显示某100cm^2的培养皿平均菌落数为6个，试估计该培养皿菌落数小于3个的概率，大于1个的概率。

该培养皿菌落数小于3个的概率

$$P(X<3)=\sum_{X=0}^{2}P(X)=\sum_{X=0}^{2}\frac{e^{-6}6^{X}}{X!}=\frac{e^{-6}6^{0}}{0!}+\frac{e^{-6}6^{1}}{1!}+\frac{e^{-6}6^{2}}{2!}=0.062$$

菌落数大于1个的概率为

$$P(X>1)=1-P(X=0)-P(X=1)=1-\frac{e^{-6}6^{0}}{0!}-\frac{e^{-6}6^{1}}{1!}=0.983$$

超几何分布（Hypergeometric）

- 概念
  - 超几何分布是一种不放回的抽样
  - 已知1000个产品中不合格产品为2%，现抽取50个样品，出现3个不合格的几率是多少？

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}},$$

连续概率分布

连续均匀分布（Uniform）

- 概念
  - 与离散均匀分布类似，但变量的取值范围是一个闭合区间[a, b]

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

正态分布

正态分布，也称为高斯分布（Gaussian）

正态曲线(normal curve)是一条高峰位于中央，两侧逐渐下降并完全对称，曲线两端永远不与横轴相交的钟型曲线该曲线表现为中间高，两边低，左右对称，略显钟形，类似于数学上的正态分布曲线。因为频率的总和等于1，故曲线下的面积等于1。

正态分布R函数

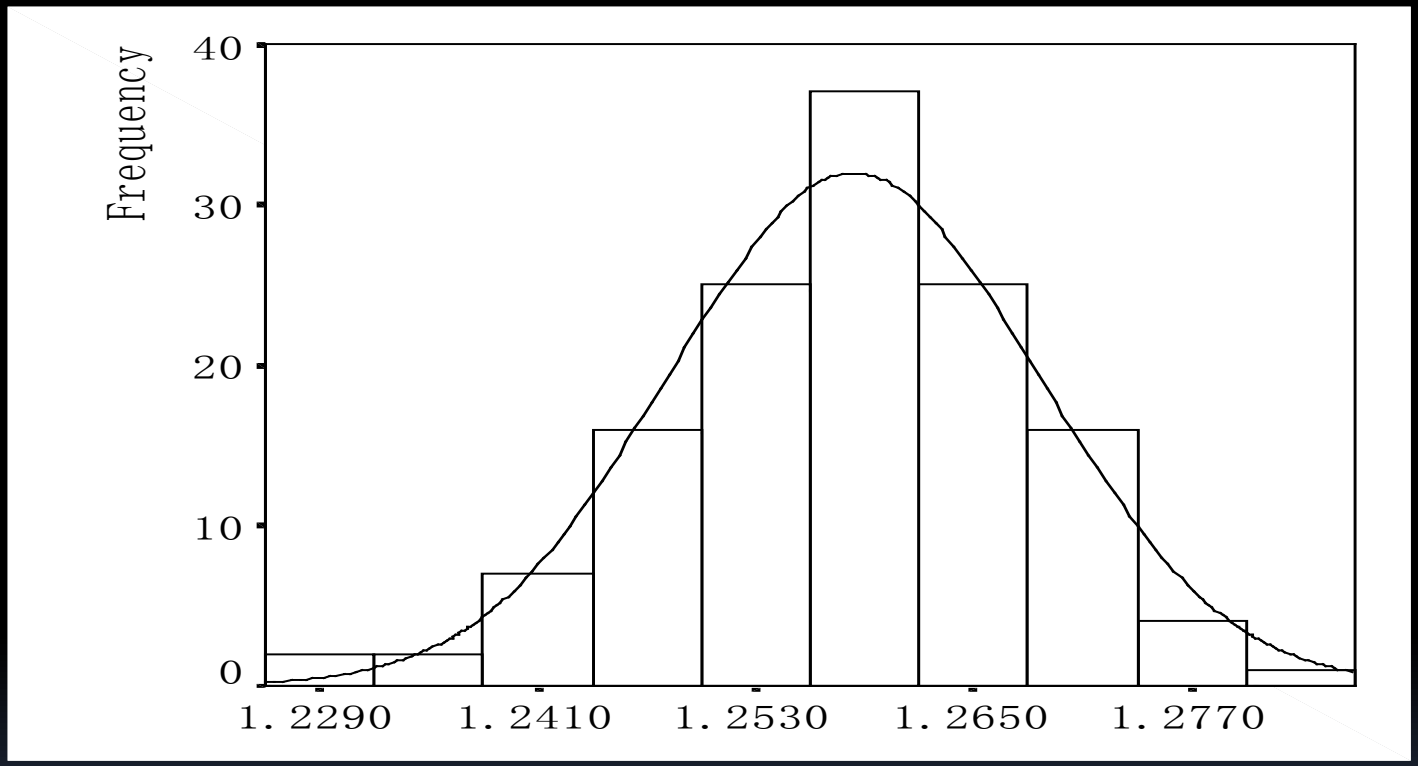
>>> rnorm(n, mu, sd) # 注意是sd而不是var

>>> dnorm(x, mu, sd) # 概率密度

>>> pnorm(x, mu, sd) # 累积概率密度

>>> qnorm(q, mu, sd) # 0<=q<=1,求分位点

正态分布



体模“骨密度”测量值的分布接近正态分布示意图（频率密度=频率/组距）

正态分布

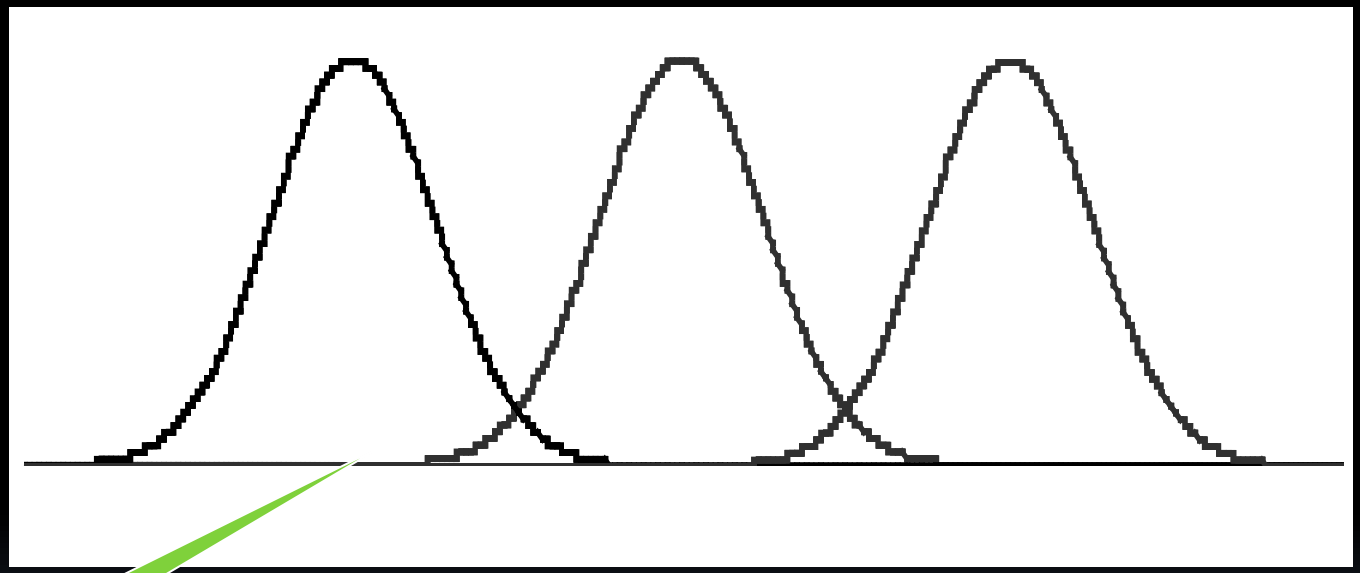
- 正态概率密度曲线的位置与形状具有如下特点
- （1）关于$x=\mu$ 对称。
- （2）在$x=\mu$处取得该概率密度函数的最大值，在 $x=\mu\pm\sigma$ 处有拐点，表现为钟形曲线。

（3）曲线下面积为1。

（4）μ决定曲线在横轴上的位置，μ增大，曲线沿横轴向右移；反之,μ减小，曲线沿横轴向左移。

（5）σ决定曲线的形状，当μ恒定时，σ越大，数据越分散，曲线越“矮胖”；σ越小，数据越集中，曲线越‘瘦高’。见图4-5。

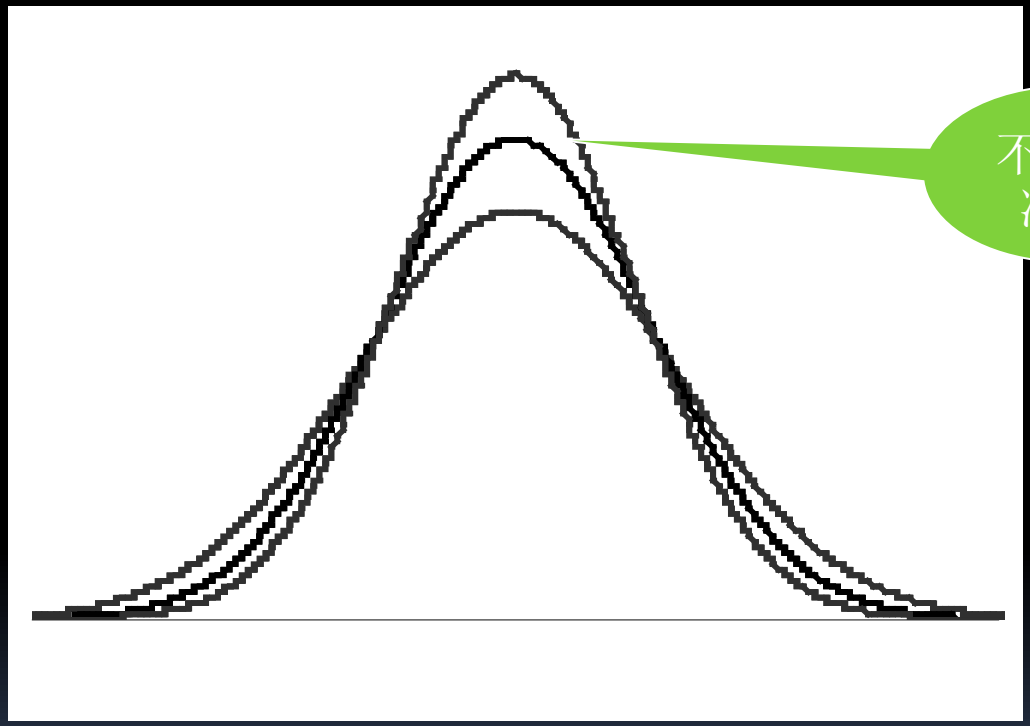
正态分布



不同
均数

U1　　U2　　U3

正态分布



正态分布

- 对任意一个服从正态分布 $N(\mu,\sigma^2)$ 的随机变量，可作如下的标准化变换，也称Z变换,

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

- Z服从总体均数为0、总体标准差为1的正态分布。我们称此正态分布为标准正态分布(standard normal distribution)，用$N(0.1)$表示。

正态分布

已知X服从均数为μ、标准差为σ的正态分布，试估计：

X取值在区间 $\mu \pm 1.96\sigma$ 上的概率：

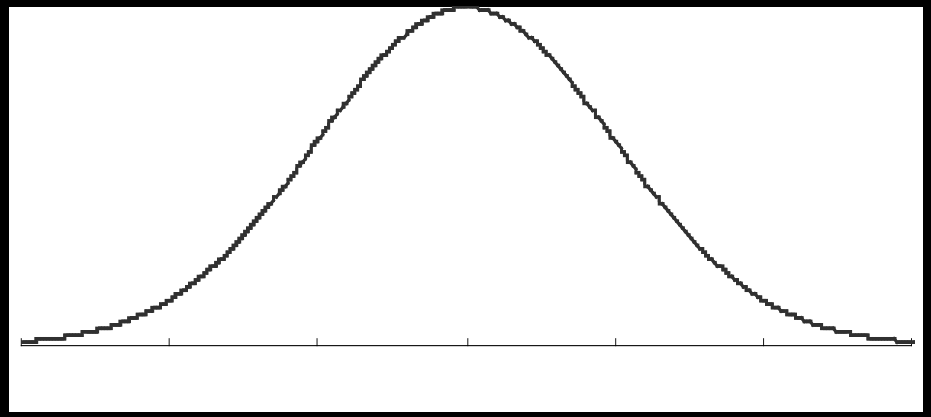
X取值在区间 $\mu \pm 2.58\sigma$ 上的概率。

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{(\mu - 1.96\sigma) - \mu}{\sigma} = -1.96$$

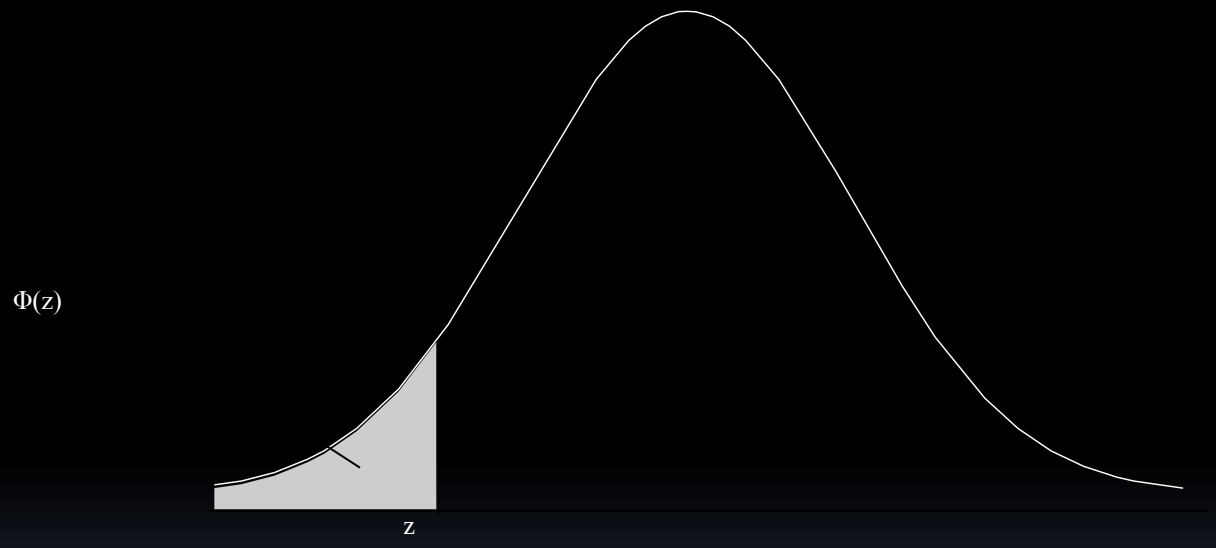
$$z_2 = \frac{x_1 - \mu}{\sigma} = \frac{(\mu + 1.96\sigma) - \mu}{\sigma} = 1.96$$

正态分布

- 正态曲线下面积的分布规律



正态分布



正态分布

- 某地1986年120名8岁男孩身高均数为 $\overline{X}$=123.02cm，标准差为S=4.79cm，试估计

(1)该地8岁男孩身高在130cm以上者占该地8岁男孩总数的百分比

(2) 身高在120cm～128cm者占该地8岁男孩总数的百分比;

(3) 该地80%的男孩身高集中在哪个范围？

正态分布

- 求Z值:

$$Z = \frac{130 - 123.02}{4.79} = 1.46$$

- 查表:

$$\Phi(-1.46) = 0.0721$$

- 理论上该地8岁男孩身高在130cm以上者占该地8岁男孩总数的7.21%。

正态分布

- 先计算120 和128所对应的Z值

$$z_1 = \frac{120 - \bar{X}}{S} = \frac{120 - 123.02}{4.79} = -0.63$$

$$z_2 = \frac{128 - \bar{X}}{S} = \frac{128 - 123.02}{4.79} = 1.04$$

$$\Phi(-0.63) = 0.2643$$

$$\Phi(1.04) = 1 - \Phi(-1.04) = 1 - 0.1492 = 0.8508$$

- 正态曲线下区间（－0.63，1.04）上的面积等于

$$\Phi(1.04) - \Phi(-0.63) = 0.8508 - 0.2643 = 0.5865$$

正态分布

- 80%的8岁男孩身高集中在$\overline{X} \pm 1.28S$区间内，即116.9cm与129.2cm之间。

```
>>> xbar <- mean(x)
>>> x80 <- xbar + qnorm(c(0.10,0.90))
```

正态分布

正态分布的应用

（一）确定医学参考值范围

医学参考值范围(reference ranges):是指特定的“正常”人群数据中大多数个体的取值所在的范围。人们习惯用该人群95%的个体某项医学指标的取值范围作为该指标的医学参考值范围。

正态分布

确定医学参考值范围的方法有两种：

（1）百分位数法:适用于任何分布型的资料。

双侧95%参考值范围: $(P_{2.5}, P_{97.5})$

单侧范围:P_{95}以下，（如血铅、发汞），

或P_5以上（如肺活量）。

（2）正态分布法

$$\overline{X} \pm 1.96S$$

正态分布

- 调查某地120名健康女性血红蛋白，直方图显示，其分布近似于正态分布，$\bar{X}=117.4$（g/L），$S=10.2$（g/L），试估计该地健康女性血红蛋白的95%参考值范围。

因血红蛋白过高、过低均为异常，所以按双侧估计95%医学参考值范围

$$\bar{X}+1.96S=117.4+1.96\times10.2=137.9(g/l)$$

$$\bar{X}-1.96S=117.4-1.96\times10.2=97.41(g/l)$$