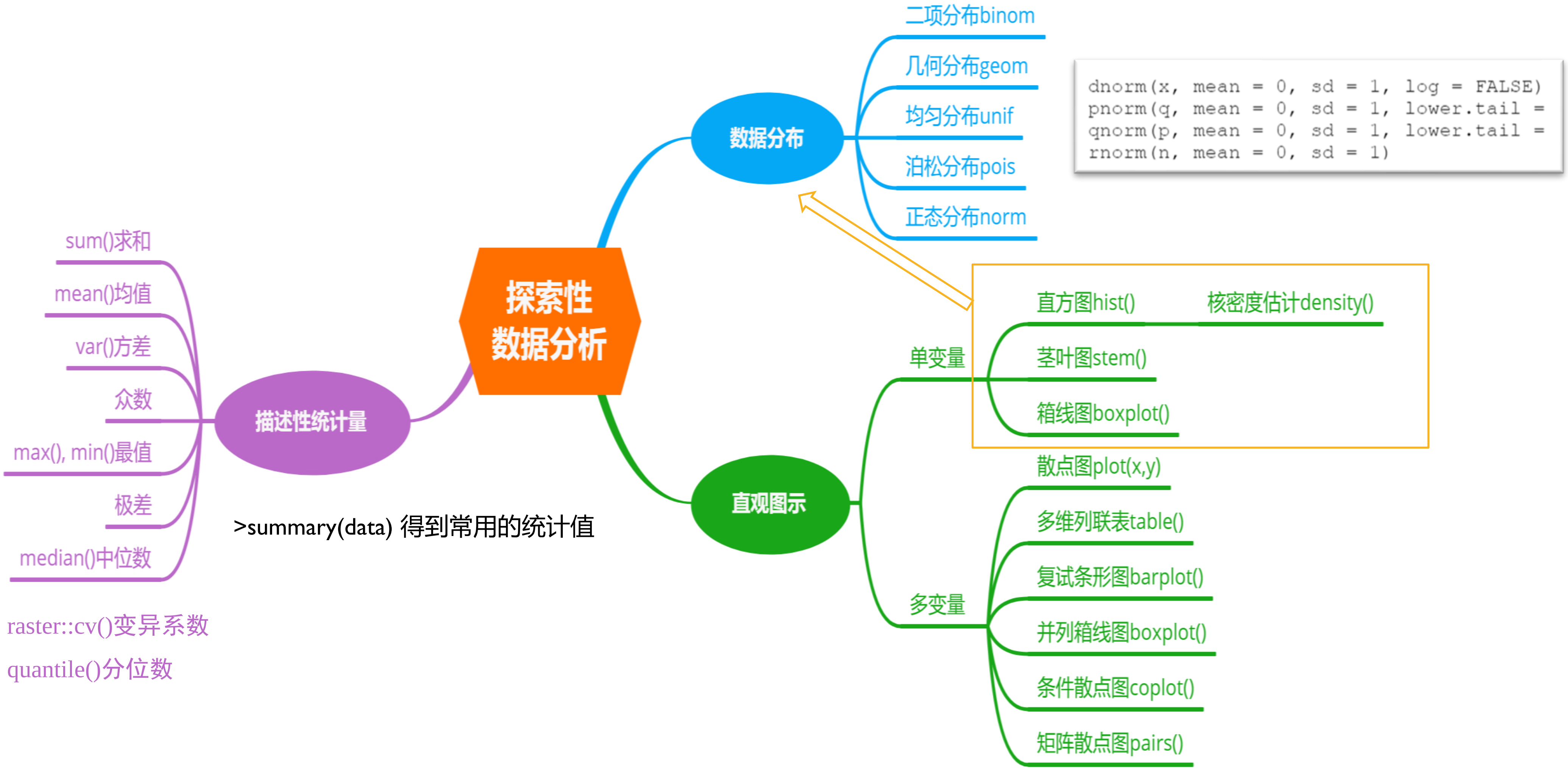
描述性统计思维导图

上节课回顾



第五章:统计推断 Ⅰ

- ➢统计推断
- ➢抽样分布
- ➢参数估计的基本原理
- ➢一个总体参数的区间估计
- ➢两个总体参数的区间估计

统计方法

- 统计方法
  - 统计描述
  - 统计推断
    - 抽样分布
    - 参数估计
      - 点估计
      - 区间估计
    - 假设检验

以统计量对总体参数做出推断：

期望、方差、概率分布

根据样本计算统计量的取值：

均值

方差

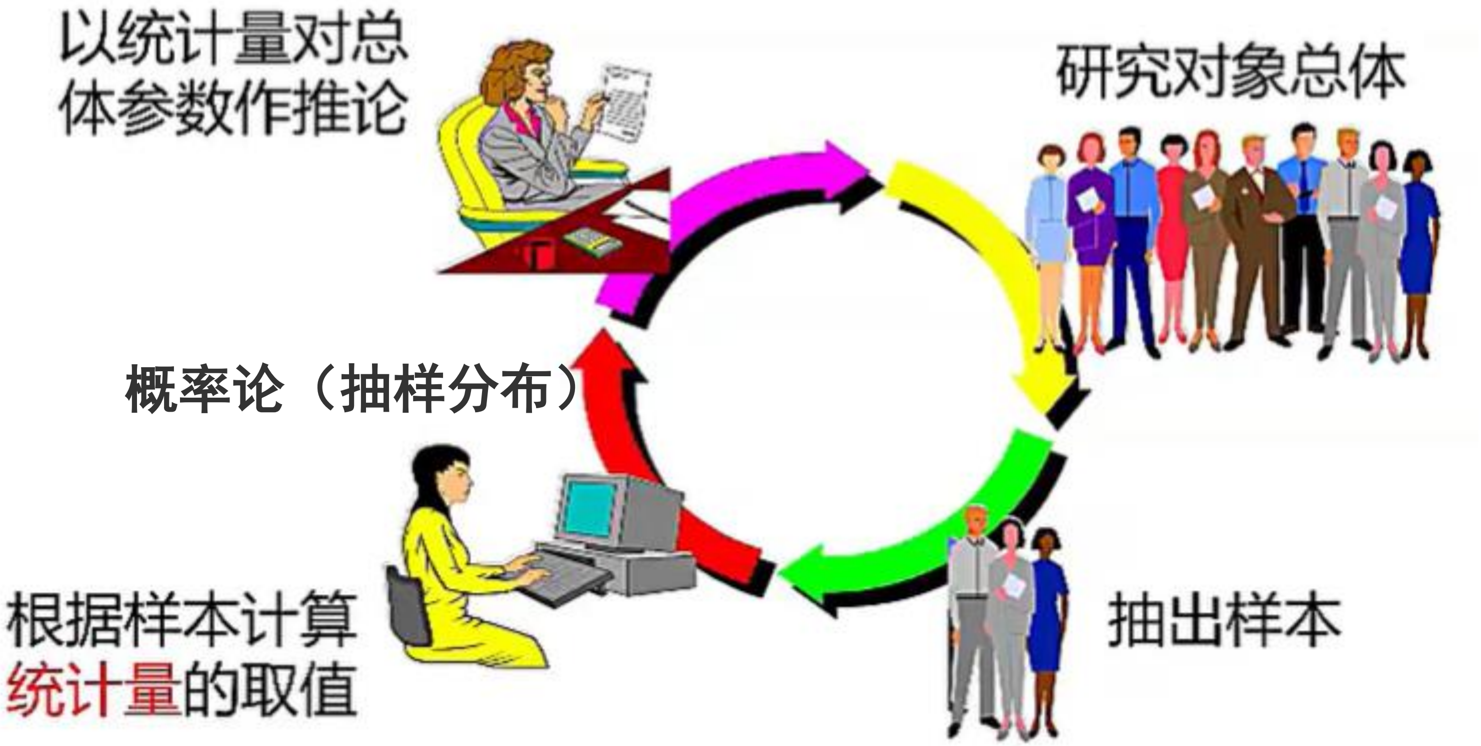
最值

中位数

数据关系

数据分布

统计推断



统计量

设$X_1, X_2, \dots, X_n$是总体X中抽取的容量为n的一个样本，如果由此样本构造的一个函数$T(X_1, X_2, \dots, X_n)$，不依赖于任何未知参数，则称函数$T(X_1, X_2, \dots, X_n)$是一个统计量。

- ✓ 统计量是一个随机变量；
- ✓ 当获得特定样本具体的观察值$x_1, x_2, \dots, x_n$时，计算出的$T(x_1, x_2, \dots, x_n)$的具体数值，就是获得一个具体统计量的值。

常用的统计量

样本均值： $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

样本方差： $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$

变异系数/离散系数： $CV=\frac{S}{\bar{X}}$

样本k阶原点矩： $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$

样本k阶中心矩： $B_k=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^k$

样本偏度： $SK=\frac{\sqrt{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^3}{\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)^{3/2}}$

样本峰度： $K=\frac{(n-1)\sum_{i=1}^{n}(X_i-\bar{X})^3}{\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)^4}-3$

统计推断-小结

- 统计推断(statistical inference)，是指从总体中抽取部分样本，通过对抽取部分所得到的带有随机性的数据进行统计量的合理分析，进而对总体作出科学的判断，它是伴随着一定概率的推测。
- 统计推断的一个基本特点是：其所依据的条件中包含有带随机性的观测数据。以随机现象为研究对象的概率论，是统计推断的理论基础。
- 统计推断的三个中心内容：抽样分布、参数估计和假设检验。

第五章：统计推断 I

- ➢统计推断
- ➢抽样分布
- ➢参数估计的基本原理
- ➢一个总体参数的区间估计
- ➢两个总体参数的区间估计

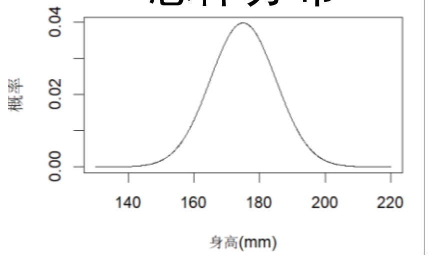
抽样分布（Sample distribution)

- **样本统计量的分布即抽样分布。**
  - ✓ 当我们要对某一总体的参数进行估计时，就要研究来自该总体的所有可能的样本统计量的分布问题。
  - ✓ 其结果来自容量相同的所有可能样本。

总体分布与样本分布

总体（身高）

总体分布

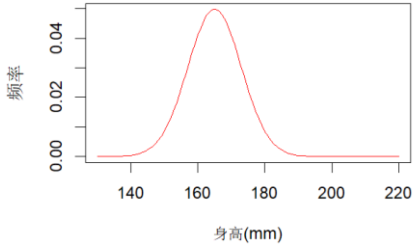


抽样

样本分布

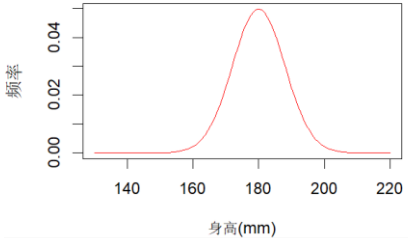
$(x_1, x_2, \dots, x_{200})$

样本1 



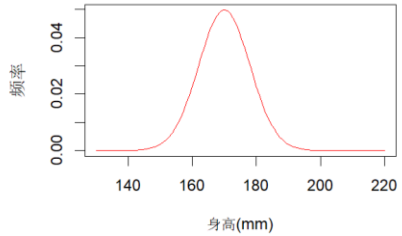
$(x_1', x_2', \dots, x_{200}')$

样本2 



$(x_1'', x_2'', \dots, x_{200}'')$

样本3 



- 总体分布是唯一的，但样本分布有多种可能；
- 样本容量是确定的，但可能的样本组合会很多；
- 一个样本的统计量（如样本均值）是样本的函数，需要区分统计量与特定样本统计量的取值。

样本均值： $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

重复抽样条件下

由正态分布导出的三大分布

Karl Pearson

卡方分布

William Gosset

***t*分布**

R.A. Fisher

***F*分布**

统计学“三剑客”与抽样三大分布

抽样分布-χ^2分布

CK. Pearson
统计学之父

✓ **卡方分布（χ^2分布）是由标准正态分布的平方和而来的。**

设 $X \sim N(\mu, \sigma^2)$, 则 $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

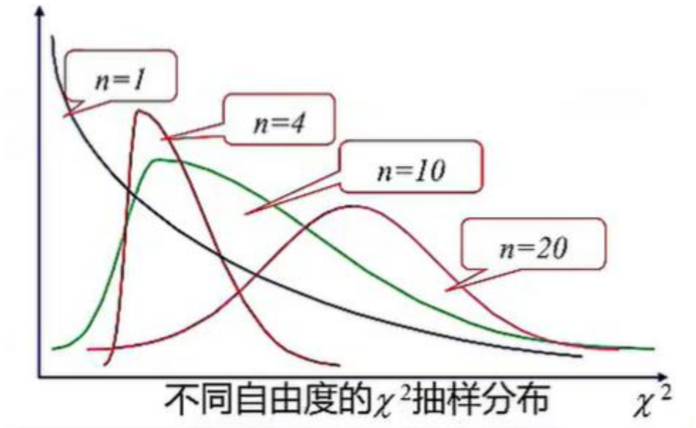
令 $Y = Z^2$, 则Y服从自由度为1的χ^2分布，即 $Y \sim X^2(1)$

若$Z_1, Z_2, \ldots Z_n$ 相互独立，且均服从$N(0,1)$，则${Z_1}^2 + {Z_2}^2 + \cdots {Z_n}^2$ 服从自由度为n的χ^2分布，$Z \sim \chi^2(n)$

当总体 $X \sim N(\mu, \sigma^2)$, 从中抽取容量为n的样本，则 $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim X^2(n-1)$

χ^2分布性质

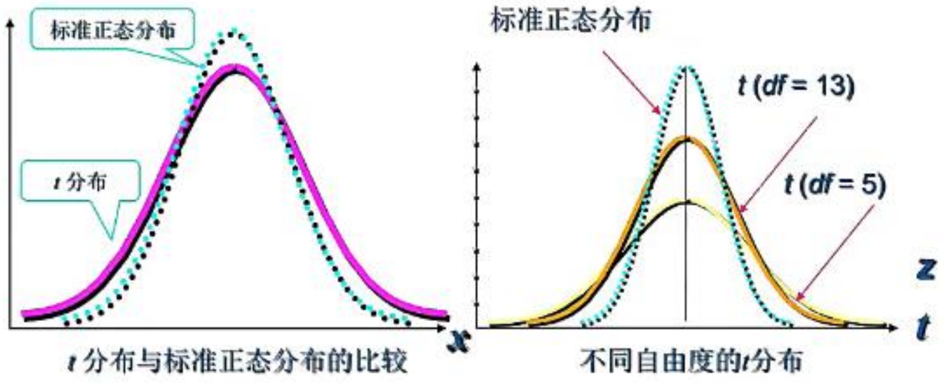
- 变量值始终为正。分布的形状取决于其自由度n的大小，通常为不对称的正偏分布，但随着自由度的增大逐渐区域对称；
- $X\sim\chi^2(n)$,则 $E(\chi^2)=n$, $D(X)=2n$ (n为自由度）
- 可加性：$X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$,且相互独立，则 $X+Y\sim\chi^2(n_1+n_2)$



抽样分布-*t*分布

William Sealy Gosset (笔名：Student)
小样本统计理论的开创者

- 定义：$X \sim N(0,1), Y \sim X^2(n)$，且 X 和 Y 相互独立，则$\frac{X}{\sqrt{Y/n}}$ 服从自由度为n的t分布，$\sim t(n)$
- t分布是类似正态分布的一种对称分布，它通常要比正态分布平坦和分散。分布依赖于称之为自由度(degree freedom)的参数。随着自由度n的增大，分布也逐渐趋于正态分布。



t分布

- 总体方差未知时，人们会很自然地想到用样本方差来代替

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad \Rightarrow \quad T=\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

- 统计量T服从什么分布呢？

定理 设$X_1,X_2,\cdots,X_n(n\geq 2)$是来自总体$N(\mu,\sigma^2)$的一个样本，则，

$$T=\frac{\bar{X}-\mu}{s/\sqrt{n}}\sim t(n-1)$$

即，T服从自由度为$n-1$的t分布。

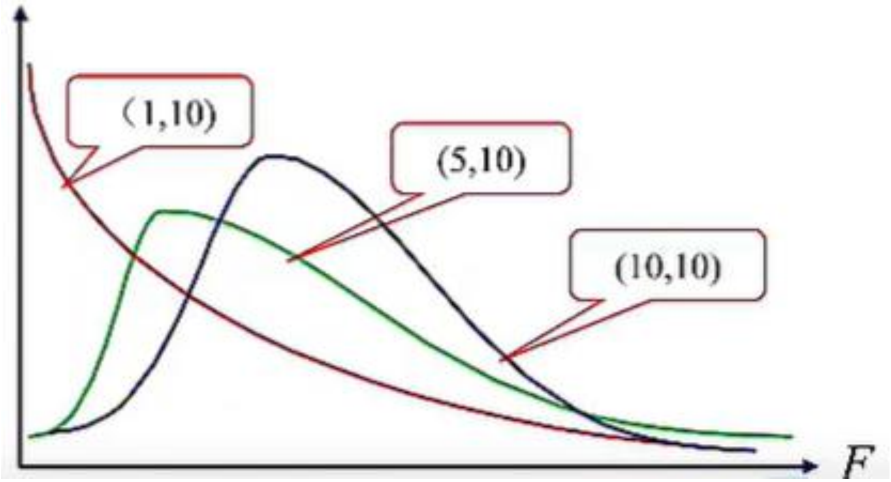
抽样分布-*F*分布

R.A. Fisher
推断统计学之父

- 定义：$X\sim X^2(m), Y\sim X^2(n)$, 且相互独立，则 $\frac{X/m}{Y/n}$ 服从自由度为 m 和 n 的F 分布， $\sim F(m,n)$；

$$\boldsymbol{F}=\frac{\boldsymbol{X}/\boldsymbol{m}}{\boldsymbol{Y}/\boldsymbol{n}}$$

$$\sim F(m,n)$$



F 分布概率密度曲线

样本均值的抽样分布

【例】设一个总体，含有4个元素（个体），即总体单位数N=4。4 个个体分别为$X_1=1$、$X_2=2$、$X_3=3$ 、$X_4=4$ 。总体的均值、方差及分布如下

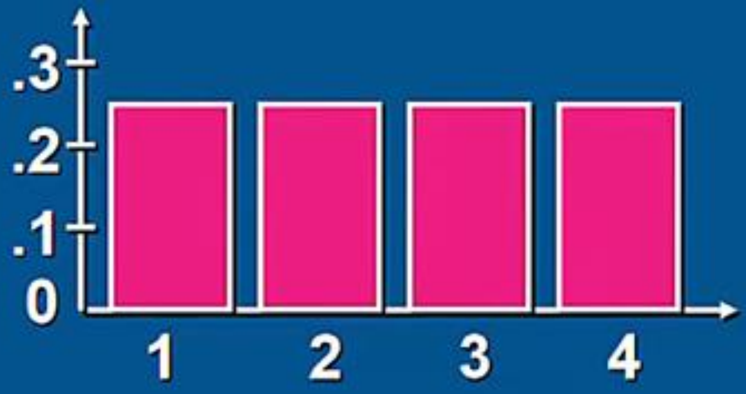
均值和方差

总体分布的数字特征

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = 2.5$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} = 1.25$$

总体分布



有放回随机抽样结果

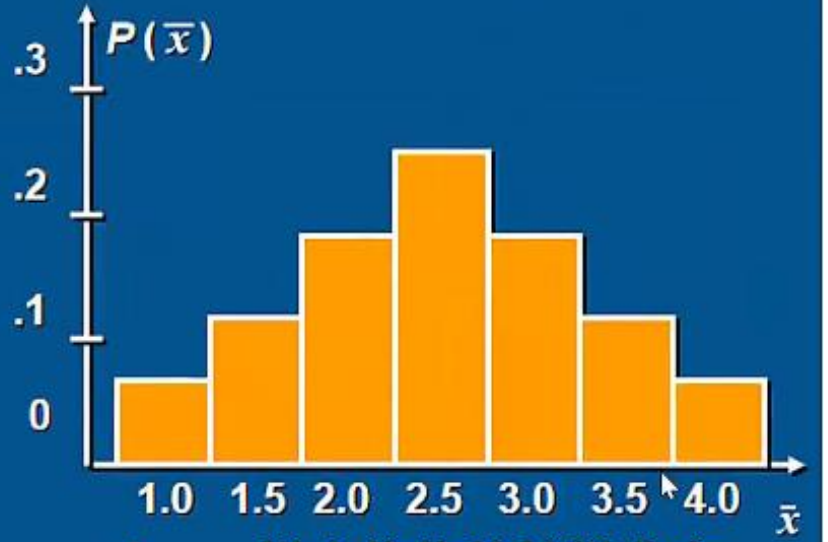
➡ 现从总体中抽取$n=2$的简单随机样本，在重复抽样条件下，共有$4^2=16$个样本。所有样本的结果如下表

所有可能的$n=2$ 的样本（共16个）				
第一个观察值	第二个观察值			
	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

样本均值的抽样分布

➡ 计算出各样本的均值，如下表。并给出样本均值的抽样分布

16个样本的均值（$\bar{x}$）				
个观察值	第二个观察值			
	1	2	3	4
	1.0	1.5	2.0	2.5
2	1.5	2.0	2.5	3.0
3	2.0	2.5	3.0	3.5
4	2.5	3.0	3.5	4.0



样本均值的抽样分布

样本均值抽样分布的数字特征

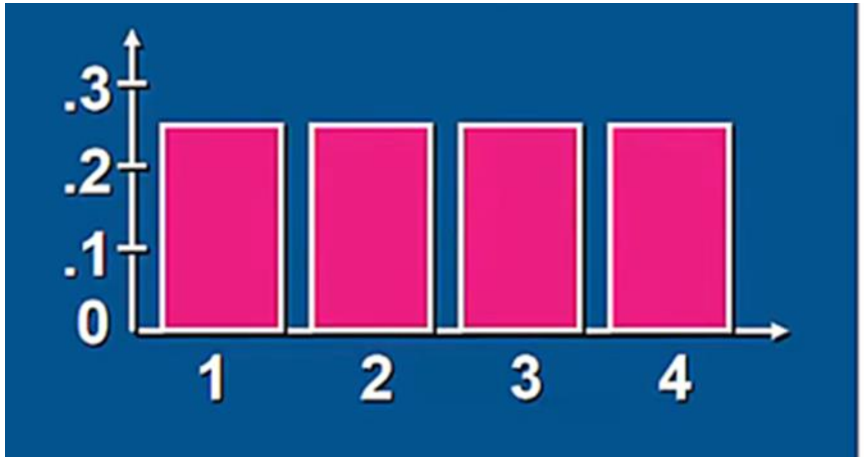
$$\mu_{\overline{x}} = \frac{\sum_{k=1}^{t} \overline{x}_k}{t} = \frac{1.0 + 1.5 + \cdots + 4}{16} = 2.5$$

$$\sigma_{\overline{x}} = \sqrt{\frac{\sum_{k=1}^{t} (\overline{x}_k - \mu_{\overline{x}})^2}{t-1}} = \sqrt{\frac{(1.0-2.5)^2 + (1.5-2.5)^2 + \cdots + (4.0-2.5)^2}{15}} = 0.72$$

- 样本均值与总体均值相等，但样本方差与总体方差不等；
- 这是为什么呢？

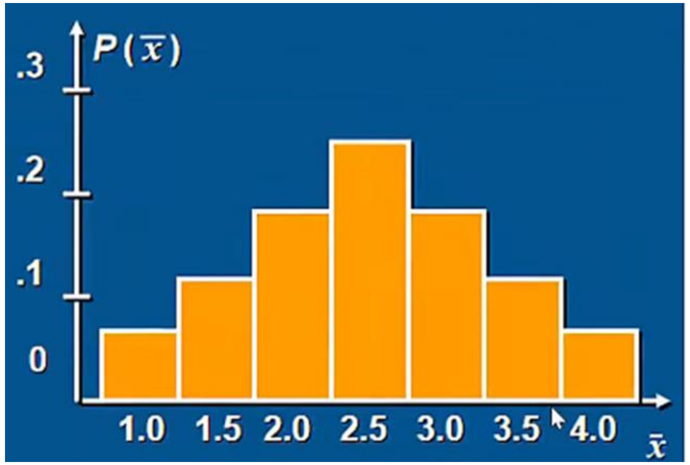
总体分布（N=4）

$\mu = 2.5, \sigma = 1.25$



均值的抽样分布(n=2)

$\mu_{\bar{x}} = 2.5$，$\sigma_{\bar{x}} = 0.72$



样本均值的抽样分布

- 容量相同的所有可能样本的样本均值的概率分布。
- 进行推断总体均值μ的基础。
- 要分析容量相同的所有可能样本的样本均值的概率分布，必须了解总体的分布情况。

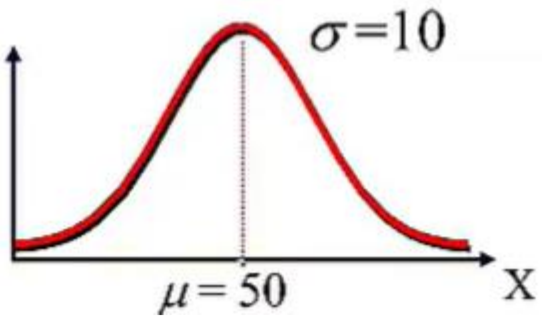
1. 总体服从正态分布

当总体服从正态分布$N(\mu,\sigma^2)$时，来自该总体的所有容量为n的样本的均值$\bar{x}$也服从正态分布。

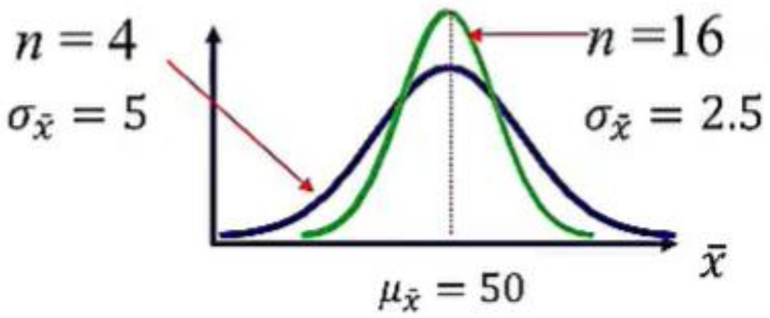
样本均值的抽样分布

重要结论

当总体服从正态分布$N(\mu,\sigma^2)$时，来自该总体的所有容量为n的样本的均值$\bar{x}$也服从正态分布，数学期望为μ，方差为${\sigma^2}/{n}$。即$\bar{x}\sim N(\mu,{\sigma^2}/{n})$。



总体分布



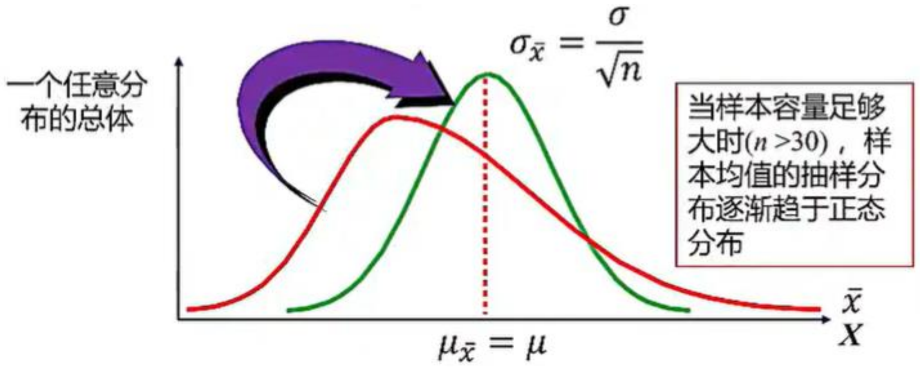
样本均值分布

样本均值的抽样分布

2. 总体不是正态分布

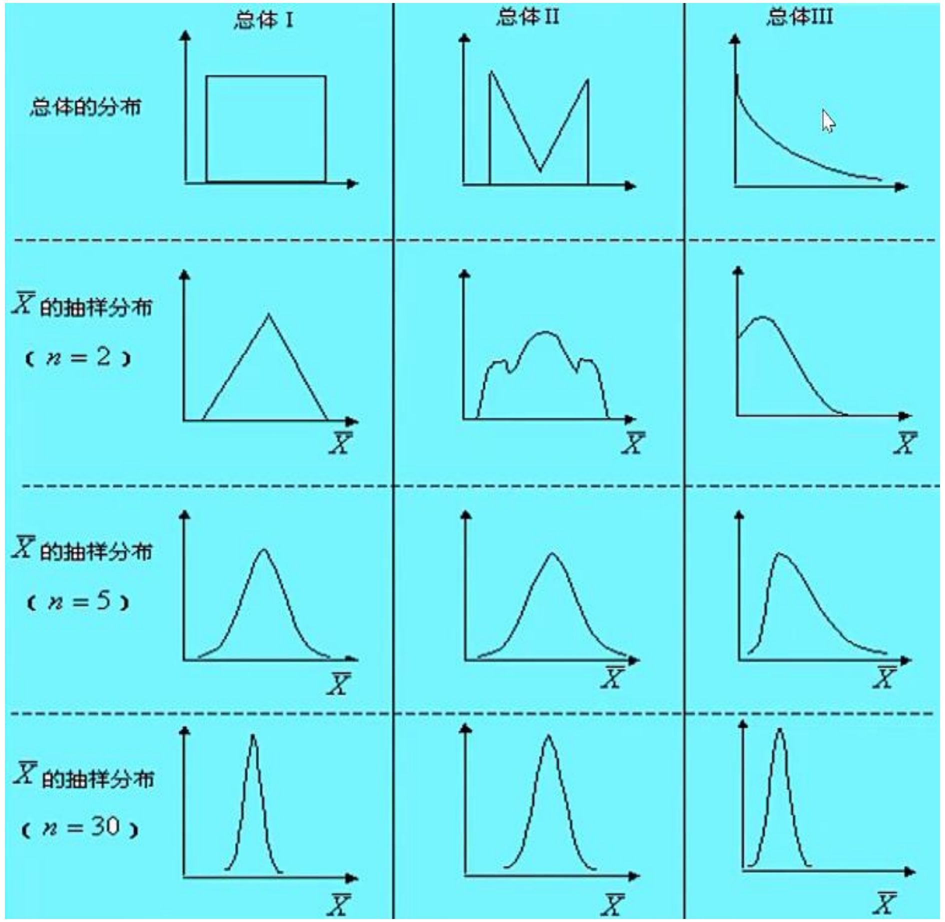
设从均值为μ，方差为σ^2的一个任意总体中抽取容量为n的样本，当n充分大时（$n > 30$），根据中心极限定理，样本均值的抽样分布近似服从均值为μ、方差为σ^2/n的正态分布。

独立同分布中心极限定理



中心极限定理 (Central Limit Theorem)

样本均值$\bar{X}$的抽样分布
趋于正态分布的过程

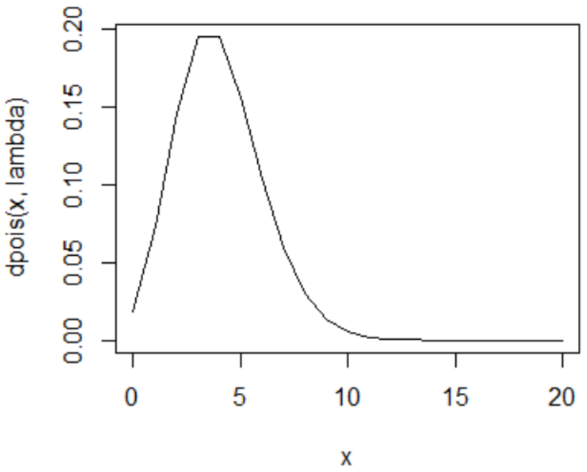


中心极限定理在R中的模拟

- **泊松分布模拟中心极限定理**

总体服从λ=4的泊松分布, $X \sim P(4)$

```
> x<-seq(0,20)
> lambda=4.0
> plot(x,dpois(x,lambda),type="l")
```



从总体$X \sim P(4)$中随机抽取容量 $n = 10$ 的样本$(x_1, x_2 \dots x_{10})$

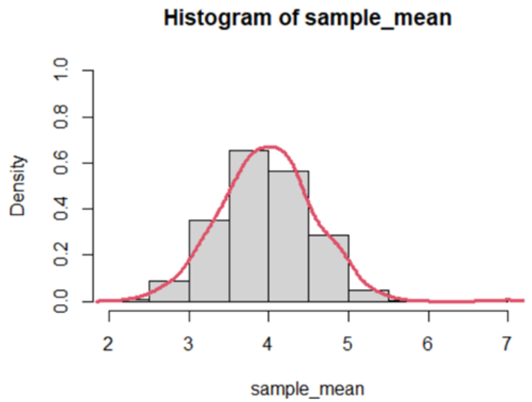
样本1： $x_1, x_2 \dots x_{10}$ $\longrightarrow$ $\overline{x_1}$

样本2： $x_1, x_2 \dots x_{10}$ $\longrightarrow$ $\overline{x_2}$

⋮

样本500： $x_1, x_2 \dots x_{10}$ $\longrightarrow$ $\overline{x_{500}}$

$\overline{x_1}, \overline{x_2} \dots \overline{x_{500}}$ $\longrightarrow$

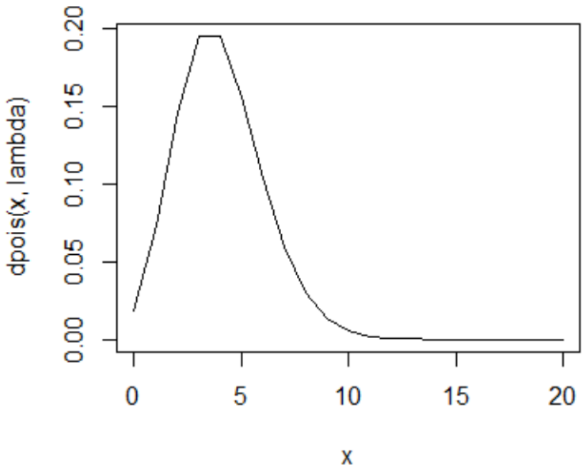


中心极限定理在R中的模拟

- **泊松分布模拟中心极限定理**

总体服从λ=4的泊松分布, $X \sim P(4)$

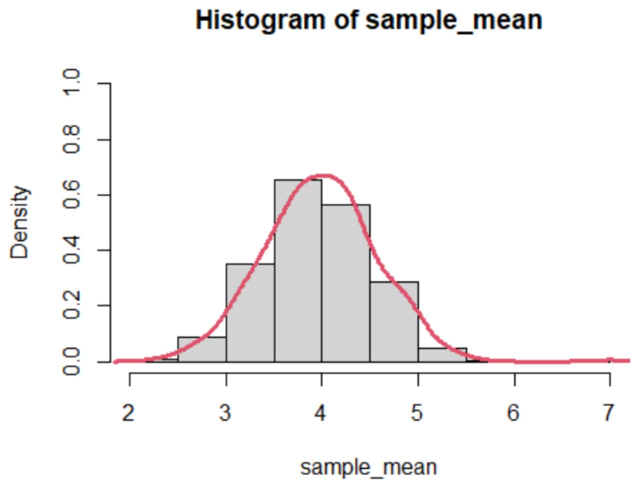
```
> x<-seq(0,20)
> lambda=4.0
> plot(x,dpois(x,lambda),type="l")
```




```
n=10 #样本容量
sample_mean<-array(0,dim=500) #存储样本均值
for (i in 1:500){
  x<-rpois(n,4)
  sample_mean[i]<-mean(x)
}
hist(sample_mean,prob=T)
lines(density(sample_mean),col=2,lwd=3)
```

每次从$X \sim P(4)$泊松分布总体中抽取n=10个样本

计算并存储每次抽样的样本均值

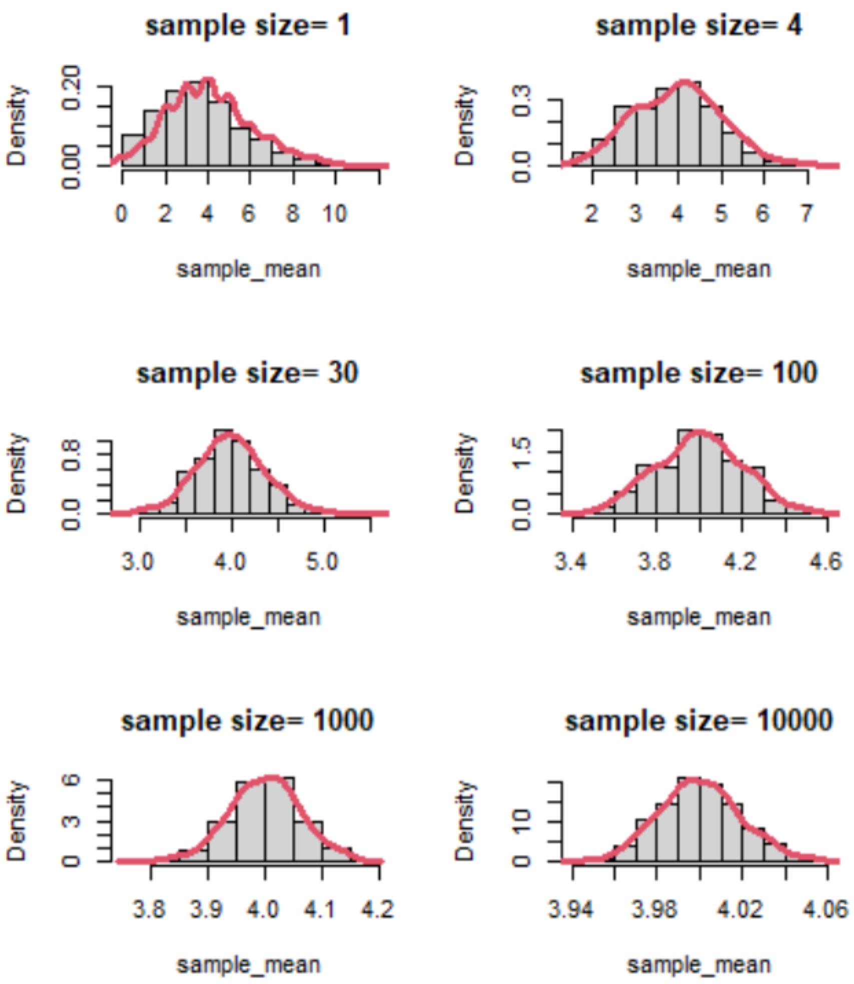


中心极限定理在R中的模拟

- 一起产生样本容量为1，4，30，100，1000的样本均值抽样分布

```
> par(mfrow=c(3,2))  定义出一个3x2的画布布局，可容纳6幅图
> sample_size<-c(1,4,30,100,1000,10000) 样本量为1，4，30，100，1000，10000
> for (j in 1:6) {
+   sample_mean<-array(0,dim=500)
+   for (i in 1:500){
+     x<-rpois(sample_size[j],4)  随机数生成器（服从λ=4的泊松分布）
+     sample_mean[i]<-mean(x)  计算样本均值
+   }
+   hist(sample_mean,prob=T,
+        main=paste("sample size=",sample_size[j]))
+   lines(density(sample_mean),col=2,lwd=3)
+ }
```

中心极限定理在R中的模拟

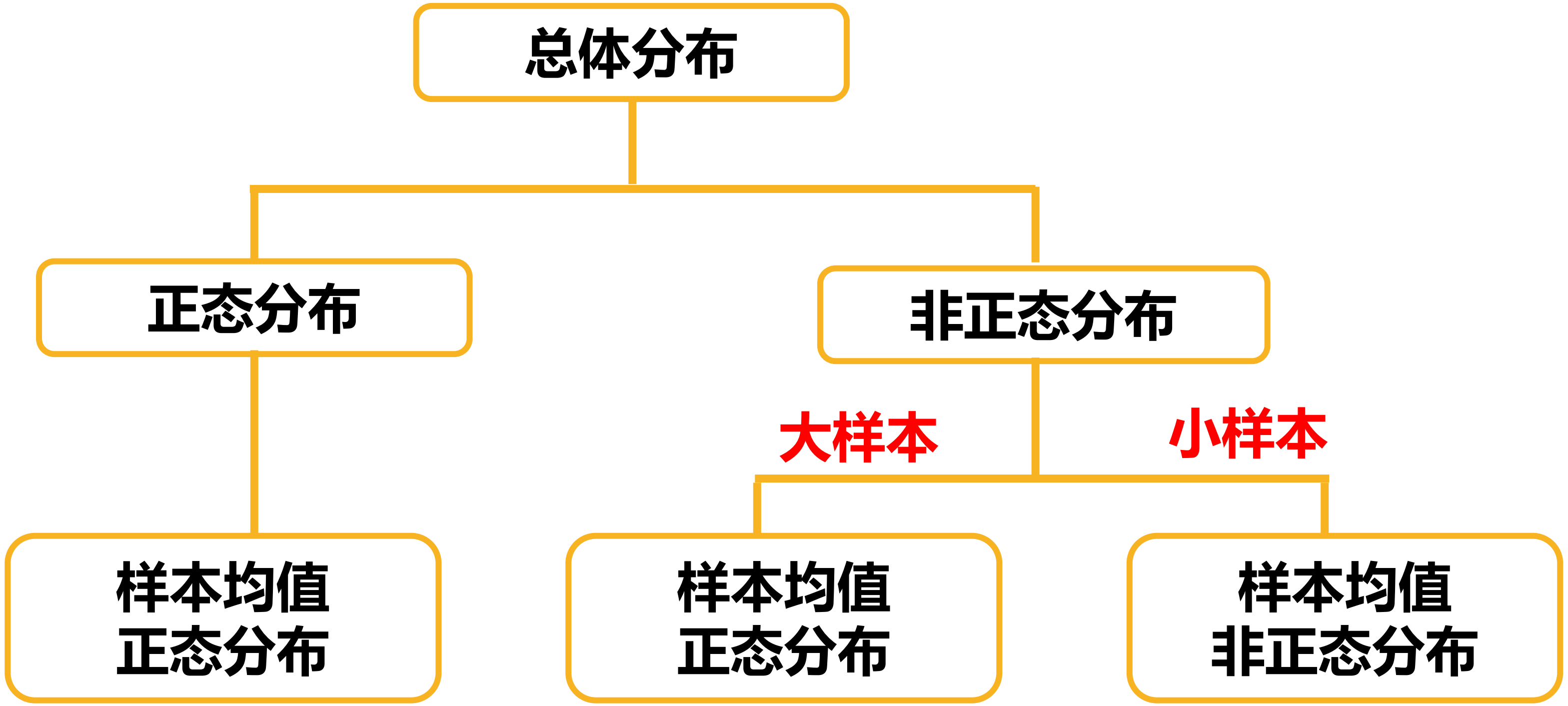


自由练习

中心极限定理应用

中心极限定理应用

- 样本均值的抽样分布与总体分布的关系



样本均值与方差

- 样本均值的均值（数学期望）等于总体均值

$$E(\bar{X}) = \mu$$

- 样本均值的方差

✓重复抽样 $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ 其中n是样本容量，N是总体容量

✓不重复抽样 $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

标准差与标准误

- **标准差**是对一次抽样的原始数据进行计算的；
- **标准误**（**standard error**）是对多次抽样的样本统计量进行计算的（这个统计量可以是均值，也可以是其他的），即样本均值的标准误差

✓重复抽样 $\boldsymbol{\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}}$ 其中$\boldsymbol{n}$是样本容量，$\boldsymbol{N}$是总体容量

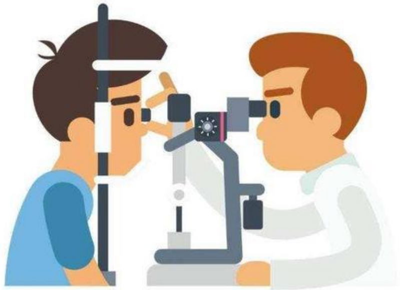
✓不重复抽样 $\boldsymbol{\sigma_{\bar{X}}=\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}}$

第五章：统计推断 Ⅰ

- ➢统计推断
- ➢抽样分布
- ➢参数估计的基本原理
- ➢一个总体参数的区间估计
- ➢两个总体参数的区间估计

案例

眼镜质量

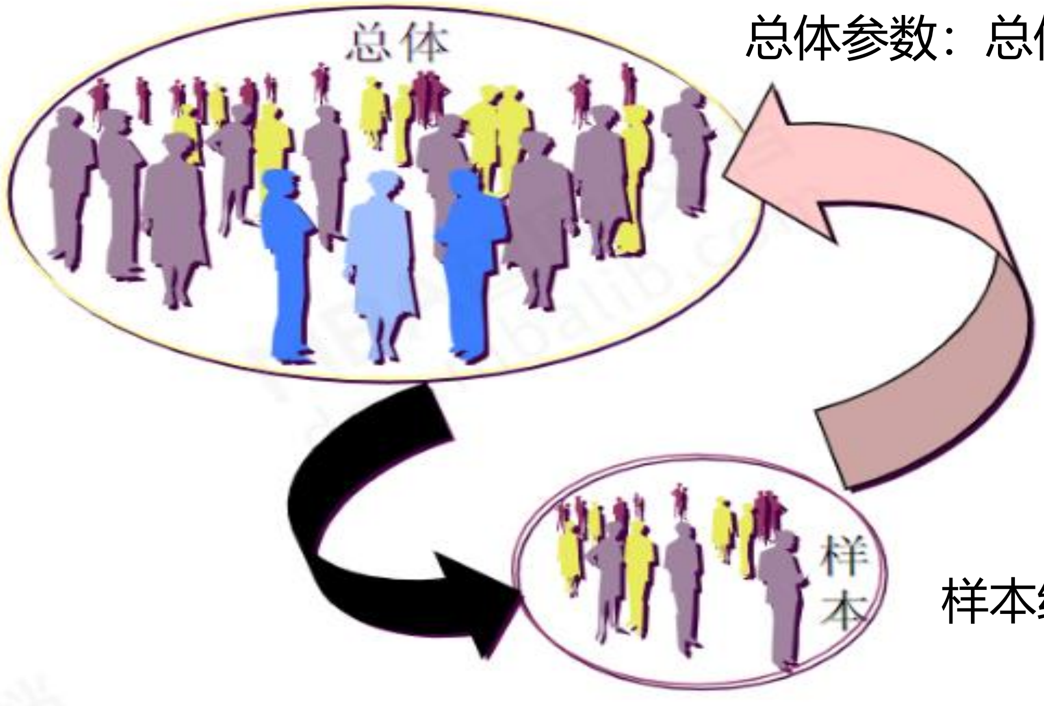


- 2024年，武汉市质量监督局对大明眼镜店的眼镜质量进行检查，分别抽取了5种规格的眼镜，每个规格抽取了3个批次，发现合格率均值为92.5%。

- 大明眼镜店所有眼镜（总体）的合格率是到多少呢?

参数估计

- 参数估计是用样本统计量来估计总体分布的未知参数。



总体参数：总体均值、比例和方差等

样本统计量：样本均值、比例和方差等

参数估计

- 参数估计有两种基本形式：**点估计**和**区间估计**。
- **点估计（point estimation)**: 如果构造一个统计量$\hat{\theta}(X_1, X_2, \dots, X_n)$作为参数$\theta$的估计量，则称为参数$\theta$的点估计。即，通过一个样本具体的数值来估计一个总体参数的取值。

例如：一家公司生产一种加湿器，再一次灌满水的情况下，加湿器加湿时间越长，说明越有效，为了知道这家公司总体加湿器的加湿持续时间，我们对样本进行了检测。

经过检测，样本均值为24h，用样本均值估计总体均值。我们期望总体和样本的均值大致相等，所以得到估计的总体均值也为24h，此时，样本均值24h称之为总体均值的点估计量。

点估计

- 点估计的方法有矩法估计、最大似然法估计等。

矩法估计（method of moment)

矩的定义：设X为随机变量，若$E(|X|^k)$存在，则称$E(X^k)$为X的k阶原点矩；

若$E(|X-EX|^k)$存在，则称$E(X-EX)^k$为X的k阶中心矩。

- 原点矩（离原点的距离）

样本(A_k)	总体
$\frac{1}{n}\sum_{i=1}^{n} X_i$	$E(X)$
$\frac{1}{n}\sum_{i=1}^{n} X_i^2$	$E(X^2)$
$\vdots$	$\vdots$
$\frac{1}{n}\sum_{i=1}^{n} X_i^k$	$E(X^k)$

- 中心矩（离中心的距离）

样本(B_k)	总体
$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})$	$E(X-EX)$
$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$	$E[(X-EX)^2]=DX$
$\vdots$	$\vdots$
$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$	$E[(X-EX)^k]$

矩法估计

- **理论依据**

大数定理：概率论中讨论随机变量序列的算术平均值向随机变量各数学期望的算术平均值收敛的定律。

矩法估计：

如果总体 X 的 k 阶矩存在，

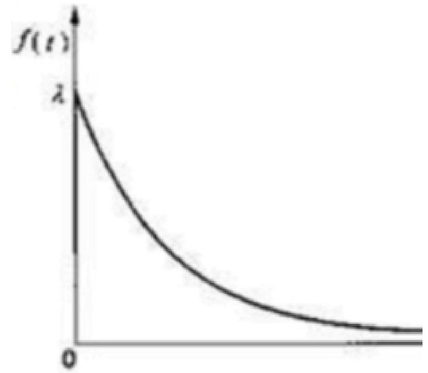
- 则样本的 k 阶矩以概率收敛到总体的 k 阶矩；
- 样本矩的连续函数收敛到总体矩的连续函数。

Karl Pearson(1857—1936)
公认为统计学之父，标准差、成分分析、卡方检验都是他提出的

实例1 矩法估计

已知

设总体X服从指数分布，密度函数是 $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\geq 0,\\ 0 & x<0,\end{cases}$



其中，λ是未知参数，若$X_1,X_2,X_3,\ldots X_n$来自总体X的一个样本，试用矩法估计估计参数λ

一阶原点矩：

$$\frac{1}{n}\sum_{i=1}^{n}X_i=E(X)$$

$$\frac{X_1+X_2+\cdots+X_N}{n}$$

$$E(X)=\int_{-\infty}^{\infty}|x|f(x)dx=\int_0^{\infty}xf(x)dx=\int_0^{\infty}x\cdot\lambda e^{-\lambda x}dx=\frac{1}{\lambda}\int_0^{\infty}\lambda xe^{-\lambda x}d\lambda x$$

令$u=\lambda x$，则：

$$E(X)=\frac{1}{\lambda}\int_0^{\infty}ue^{-u}du=\frac{1}{\lambda}[(-e^{-u}-ue^{-u})|(\infty,0)]=\frac{1}{\lambda}$$

结果：指数分布的一阶矩（均值）是$1/\lambda$，因此它的估计是 $\hat{\lambda}=\frac{1}{\bar{X}}$

实例2 矩法估计

下面的观测值为来自指数分布的一个样本：

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\geq 0,\\ 0 & x<0,\end{cases}$$

0.59132754	0.12854935	0.46900228	0.29835980	0.24341462
0.06566637	0.40085536	2.99687123	0.05278912	0.09898594

编写相应的R函数来估计其参数 λ

一阶矩法估计：$\hat{\lambda}=\dfrac{1}{\overline{X}}$

```
> X<-c(0.59132754,0.12854935,0.46900228,
+      0.29835980,0.24341462,0.06566637,
+      0.40085536,2.99687123,0.05278912,
+      0.0989859)
> lambda<- 1/mean(X)
> lambda
[1] 1.87062
```

实例3

设总体$\boldsymbol{X}$的均值为$\boldsymbol{\mu}$, 方差是$\boldsymbol{\sigma^2}$, $\boldsymbol{X_1, X_2, X_3, \dots X_n}$是来自总体$\boldsymbol{X}$的一个样本，试用矩估计法估计均值$\boldsymbol{\mu}$和方差$\boldsymbol{\sigma^2}$

正态分布 $\boldsymbol{N}$（$\boldsymbol{\mu, \sigma^2}$）

解：计算总体X的一阶原点矩、二阶中心矩

$$\alpha_1 = E(X) = \mu \text{ , } \alpha_2 = E[X - E(X)]^2 = \sigma^2$$

样本的一阶原点矩、二阶中心矩 $A_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ $\quad B_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

得到方程组

$$\begin{cases} \hat{\mu} = \bar{X}, \\ \hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \dfrac{n-1}{n}S^2 \end{cases}$$

$$\boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

矩法估计-小结

- 矩法估计：用样本k阶矩作为总体k阶矩的估计量，建立含有待估参数的方程，从而解出待估参数。

一般地，不论总体服从什么分布，总体期望μ与方差σ^2存在，则它们的矩估计量分别为：

$$\boldsymbol{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}} \qquad \boldsymbol{\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2}$$

矩法估计–小结

$$\boxed{E(X^k)=\frac{1}{n}\sum_{i=1}^{n}X_i^k=A_k, k=1,2,\ldots,m}$$

- 矩法估计是基于简单的替换思想建立起来的一种估计方法；
- 矩法估计的理论基础是大数定律；
- 一般常用的是用样本的一阶原点矩估计总体的均值，样本的二阶中心矩估计总体的方差；
- 矩法估计缺点：未能充分利用总体信息；要求总体有k阶原点矩存在；其结果可能是不唯一的。

统计推断

上节课回顾

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

统计推断

参数估计

假设检验

点估计

区间估计

矩法估计

大数定理

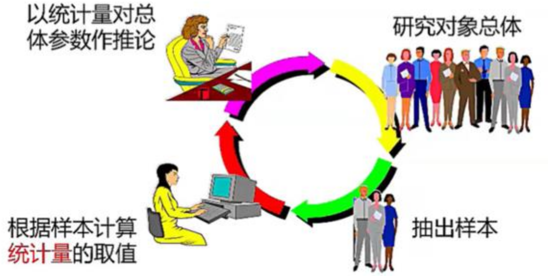
$E(X^k)=A^k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$

总体 样本

极大似然估计

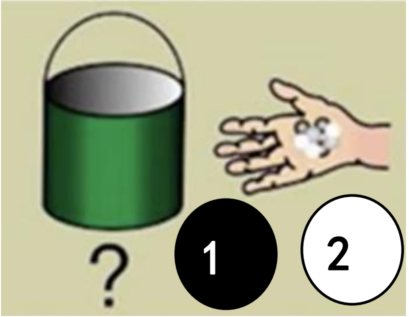
$L(\theta;x)=\prod_{i=1}^{n}f(x_i;\theta)$ 极大

样本 总体



极大似然估计

(maximum likelihood estimate)



事件A发生的概率与某一未知参数θ有关，θ取值不同，则事件A发生概率$P(A|\theta)$也不同，当我们在一次试验中事件A发生了，则认为此时的$\hat{\theta}$值应为θ的一切可能取值中使$P(A|\theta)$达到最大的那一个。

例：有两外形相同的箱子，各装100个球

一箱　99个白球　1个黑球

一箱　1个白球　99个黑球

现在从两箱中任取一箱，并从箱中任取一球，结果所取得的球是白球。

问：所取得的球来自哪一箱？

极大似然估计

定义：设总体X的概率密度或分布律为$f(x;\theta)$, $\theta\epsilon\Theta$是未知参数， x_1， $x_2,\dots,x_n$是来自总体X的样本，称 $L(\theta;x)=L(\theta;\ x_1,\ x_2,\dots,x_n)\ =\ sup\prod_{i=1}^{n}f(x_i;\theta)$

为θ的似然函数，若样本取值x固定时，$L(\theta;x)$是θ的函数.

称$L(\theta)$为似然函数. 极大似然估计法就是在参数θ的可能取值范围Θ内, 选取使$L(\theta)$ 达到最大的参数值$\hat{\theta}$作为参数θ的估计值. 即取$\hat{\theta}$, 使

$$L(\theta)=L(x_1,x_2,\cdots,x_n;\hat{\theta})=\max_{\theta\in\Theta}L(x_1,x_2,\cdots,x_n;\theta).$$

极大似然估计

实例4: 设总体X服从正态分布$N(\mu,\sigma^2)$，其中μ,σ^2为未知参数，x_1，$x_2,\dots,x_n$是来自总体X的一个样本，试用极大似然法估计参数(μ,σ^2)

$$\boxed{\boldsymbol{f(x)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}}$$

解：正态分布的似然函数为

$$L(\mu,\sigma^2;x)=\prod_{i=1}^{n}f(x_i;\mu,\sigma^2)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

相应的**对数似然函数**为

$$\ln L(\mu,\sigma^2;x)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

求偏导

$$\begin{cases}\dfrac{\partial lnL(\mu,\sigma^2;x)}{\partial\mu}=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2=0\\[2ex]\dfrac{\partial lnL(\mu,\sigma^2;x)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2\end{cases}\Rightarrow\begin{cases}\hat{\boldsymbol{\mu}}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i=\bar{x}\\[2ex]\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2\end{cases}$$

实例5 设总体X服从指数分布，密度函数是

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, x\geq 0\\ 0, x<0\end{cases}$$

其中λ是未知参数，若x_1，$x_2, \dots, x_n$是来自总体X的一个样本，试用极大似然法估计参数λ

解：只考虑$x_i \geq 0$的部分，指数分布的似然函数为

$$L(\lambda;x)=\prod_{i=1}^{\mathrm{n}} f(x_i;\lambda)=\lambda^{\mathrm{n}}\exp\left[-\lambda\sum_{i=1}^{n} x_i\right]$$

$$\boldsymbol{lnL(\lambda;x)=nln\,\lambda-\lambda\sum_{i=1}^{n} x_i} \qquad \boldsymbol{\frac{\partial lnL(\lambda;x)}{\partial\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n} x_i}$$

$\Rightarrow$ $\lambda=n/\sum_{i=1}^{n} x_i=1/\bar{X}$

实例6 某电子管的使用寿命X（单位:小时）服从指数分布X

$$p(x;\theta)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}}, x>0 \\ 0, other\end{cases}(\theta>0)$$

今取得一组样本X_k的数据如下，问如何用极大似然法估计θ？

16	29	50	68	100	130	140	270	280
340	410	450	520	620	190	210	800	1100

1. 构造似然函数

当$x_i>0$,（$i=1,2,\ldots,n$）时，似然函数为

$$L(\theta)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-\frac{x_i}{\theta}}=\theta^{-n}\cdot e^{-\frac{1}{\theta}\sum_{i=1}^{n}x_i}$$

2. 取对数 $$\ln L=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i$$

3. 建立似然方程 $$\frac{d\ln L}{d\theta}=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i=0$$

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$$

$$=\frac{5723}{18}=318\text{（小时）}$$

实例6 某电子管的使用寿命X（单位:小时）服从指数分布X

$$p(x;\theta)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}},x>0\\ 0,other\end{cases}(\theta>0)$$

今取得一组样本X_k的数据如下，问如何用极大似然法估计θ？

16	29	50	68	100	130	140	270	280
340	410	450	520	620	190	210	800	1100

1. 构造似然函数

当$x_i>0$,（$i=1,2,\ldots,n$）时，似然函数为

$$L(\theta)=\prod_{i=1}^{n}\frac{1}{\theta}e^{-\frac{x_i}{\theta}}=\theta^{-n}\cdot e^{-\frac{1}{\theta}\sum_{i=1}^{n}x_i}$$

2. 取对数 $\boxed{\ln L=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i}$

3. 建立似然方程 $\frac{d\ln L}{d\theta}=-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i=0.$

R实现：

```
> x<-c(16,29,50,68,100,130,140,270,280,340,
> n<-length(x)
> m<-sum(x)
> f<-function(theta) -n*log(theta)-m/theta
> optimize(f,c(0,1100),maximum=T)
$maximum
[1] 317.9444

$objective
[1] -121.7138
```

构建函数

求极值

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}=\frac{5723}{18}$$=318（小时）

optimize（）函数

在单参数场合, 我们可以使用**R**中的函数`optimize( )`求极大似然估计值. `optimize( )`的调用格式如下:

`optimize( )`的调用格式

```
optimiz(f = , interval = , lower = min(interval),
        upper = max(interval), maximum = TRUE,
        tol = .Machine$double.eps^0.25, ...)
```

说明: `f`是似然函数, `interval`是参数θ的取值范围, `lower`是θ的下界, `upper`是θ的上界, `maximum = TRUE`是求极大值, 否则(`maximum = FALSE`)表示求函数的极小值, `tol`是表示求值的精确度, `...` 是对`f`的附加说明.

矩法估计 VS 极大似然法估计

总体　　样本

$$E(X^k) = A_k = \frac{1}{n}\sum_{i=1}^{n} X_i{}^k$$

矩估计法：

- 优点：简单易行；不需知道总体的分布类型；
- 缺点：当总体类型已知时，没有充分利用分布提供的信息；一般情况下，矩估计量不具有唯一性。

极大似然法估计：

- 优点：极大似然估计把估计问题转化为求最大值点问题，极大似然估计有很好的统计性质；
- 缺点：需要知道总体的概率分布，应用范围受到限制；有些情况下，极大似然估计量不是唯一的

样本

$$L(x_i;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

总体

点估计小结

- **点估计：用样本的估计量的某个取值范围直接作为总体参数的估计值；**

  （例如：用样本均值直接作为总体均值的估计；用两个样本均值之差直接作为总体均值之差的估计）

- **点估计无法给出估计值接近总体参数程度的信息。**

  （虽然在重复抽样条件下，点估计的均值可望接近总体真值，但由于样本是随机的，抽出一个具体的样本得到的估计值等同于总体真值的可能性很小。）

  （一个点估计量的可靠性是由它的抽样标准误来衡量的，这表明一个具体的点估计值无法给出估计的可靠性的度量）。

区间估计 (interval estimate)　　给结论留有余地

- 虽然平均而言，$\bar{X}$能正确的代表μ，但每次观察到的$\bar{X}$不会刚好等于μ，而是随着抽到的样本不同有高有低：

$$\boldsymbol{\mu} = \bar{X} \pm \textbf{标准误}$$

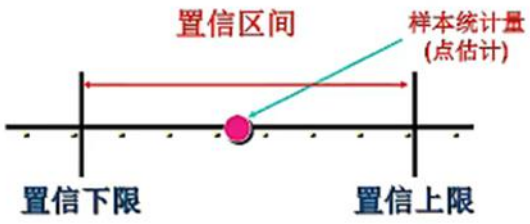
- 因此，除了点估计外，我们还想进一步知道从样本中得到的估计值有多可靠，由于样本的估计值本身也是一个随机变量，不一定刚好等于总体参数，因此我们问：估计值与总体参数有多接近？

区间估计 (interval estimate)

置信区间

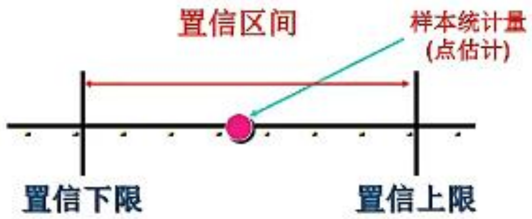
- 在点估计的基础上，给出总体参数估计的一个区间范围（该区间由样本统计量加减标准误而得到），并根据样本统计量的抽样分布能够对样本统计量与总体参数的接近程度给出一个概率度量。

置信水平



如：某个班级统计期末均分在75~85之间，置信水平是95%。

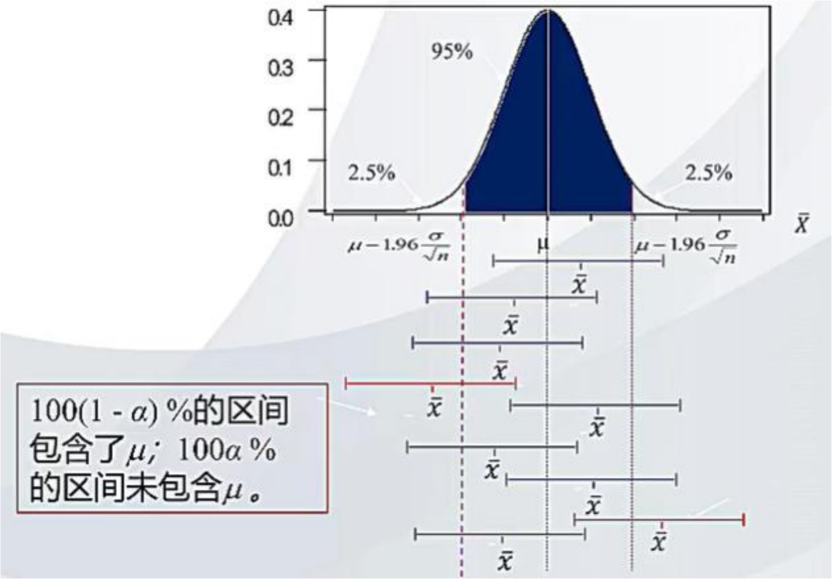
置信区间(confidence interval)



- 由样本估计量所构造的总体参数的估计区间成为**置信区间**，分别以统计量的置信上线和置信下限为上下界。
  - ✓ 在某种程度上可以确信这个区间会包含真正的总体参数，所以给它取名为置信区间；
  - ✓ 用一个具体的样本所构造的区间是一个特定的区间，我们无法知道这个样本所产生的区间是否包含总体参数的真值；
  - ✓ 只能是希望这个区间是大量包含总体参数真值的区间中的一个，但它也可能是少数几个不包含参数真值的区间中的一个。

置信水平 (confidence level)

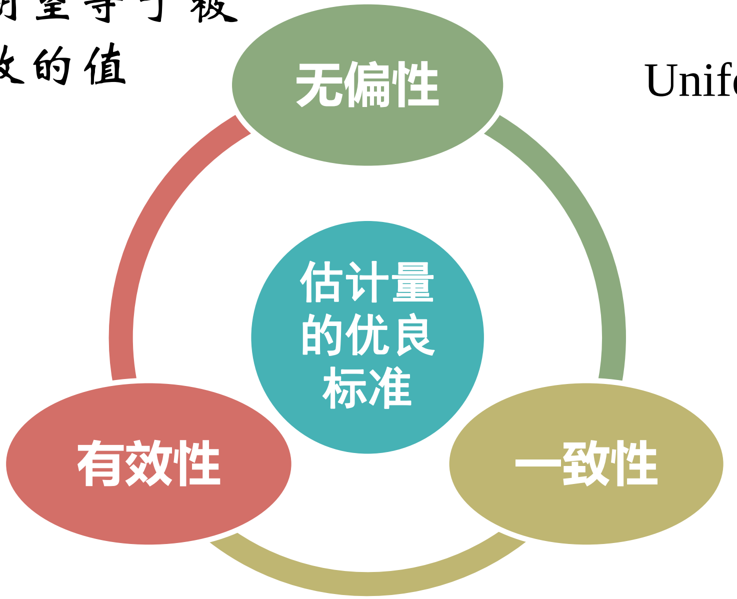
- 将构造置信区间的步骤重复若干次，得到的若干置信区间中，包含总体参数真值的次数所占的比例，称为**<u>置信水平</u>**。记为$1-\alpha$，α为未包含总体参数真值的次数所占的比例。
- 常用的置信水平（$1-\alpha$）为99%，95%，90%；相应的α为0.01，0.05，0.10。



评价估计量的标准

- 对于同一个未知参数，可以采用不同的估计量进行估计。
- 哪一种估计量是好的或者是有的估计量呢？用什么标准来评价一个估计量的好坏呢？

样本统计量的数学期望等于被估计的总体参数的值

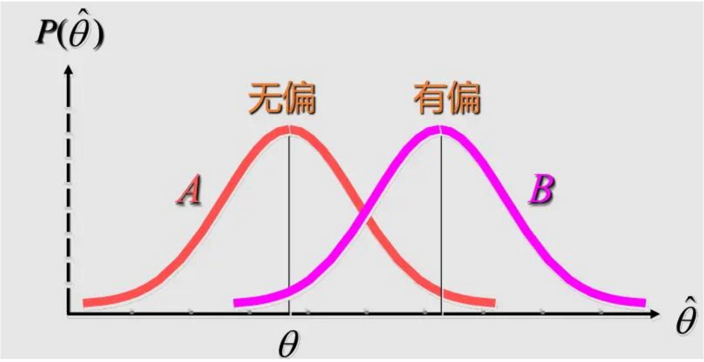


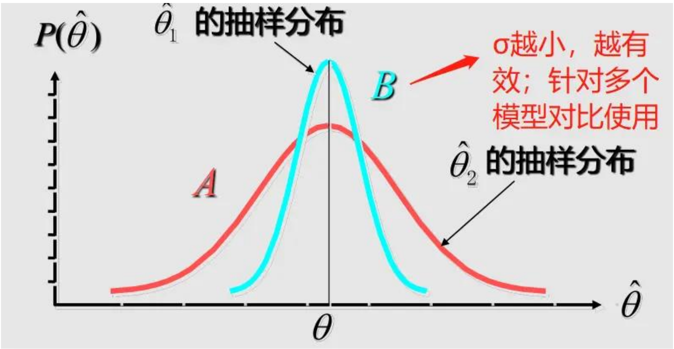
估计量的方差越小越有效

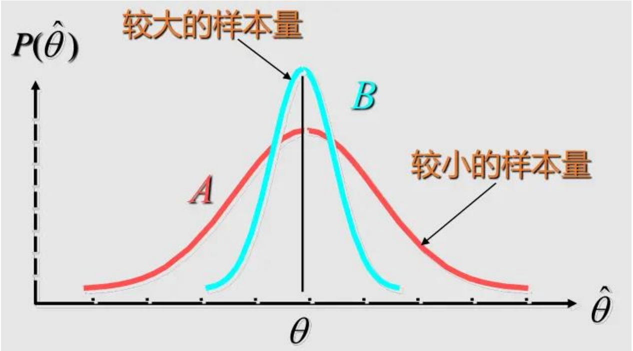
样本数目越大，估计量就越来越接近总体参数的真实值

一致方差最小无偏估计量
（UMVUE)
Uniformly Minimum Variance Unbiased Estimate

一致方差最小无偏估计量







第五章：统计推断 Ⅰ

- ➢统计推断
- ➢抽样分布
- ➢参数估计的基本原理
- ➢一个总体参数的区间估计
- ➢两个总体参数的区间估计

总体均值的区间估计

- 总体均值估计
  - 正态总体
    - σ^2已知
    - σ^2未知
  - 非正态总体
    - 大样本
    - 小样本

✓ 若总体服从正态分布，总体方差σ^2已知；

✓ 若总体不服从正态分布，总体方差σ^2已知, 在大样本条件下（$n \geq 30$）;

两者样本均值均可用正态分布近似

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

- 置信水平为（1−α）时，置信区间为 $\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$

例7. 某实验室测量铝的比重16次，得样本均值$\bar{x}$=2.705 ,设总体

$X \sim N(\mu，0.029^2)$. 求: μ的95%的置信区间.

$$\left(\bar{X}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\ \bar{X}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right)$$

解：已知$\bar{x}$=2.705，n=16, 1-α=0.95, $Z_{\alpha/2}$= 1.96

$$2.705-\frac{0.029}{\sqrt{16}}\bullet 1.96 \le \mu \le 2.705+\frac{0.029}{\sqrt{16}}\bullet 1.96$$

即用$\bar{x}=2.705$来估计μ值的可靠度达到95%的区间范围是（2.691，2.719）

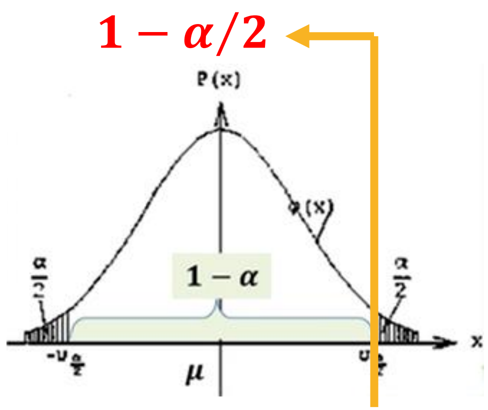
例 8 包糖机某日开工包了12包糖，称得重量(单位:克)分别为

506, 500, 495, 488, 504, 486, 505, 513, 521, 520, 512, 485.

假设重量服从正态分布，且标准差为 σ = 10, 试求糖包的平均重量μ的1−**α**置信区间(分别取**α** = 0. 10和**α** = 0. 05).

$$\left(\bar{X}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\ \bar{X}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right)$$

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```




```
qnorm(1-alpha/2,mean=0,sd=1,lower.tail = T)
> qnorm(alpha/2,mean=0,sd=1,lower.tail = FALSE)
```

例 8 包糖机某日开工包了12包糖，称得重量(单位:克)分别为

506, 500, 495, 488, 504, 486, 505, 513, 521, 520, 512, 485.

假设重量服从正态分布，且标准差为$\sigma = 10$, 试求糖包的平均重量μ的1−$\boldsymbol{\alpha}$置信区间(分别取$\boldsymbol{\alpha}$ = 0. 10和$\boldsymbol{\alpha}$ = 0. 05).

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2},\ \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$$

```
> CI.function<-function(x,sigma,alpha){
+   #构造函数计算置信区间
+   xbar=mean(x)
+   n=length(x)
+   z=qnorm(1-alpha/2,mean=0,sd=1,lower.tail = T)
+   return(c(xbar-sigma*z/sqrt(n),xbar+sigma*z/sqrt(n)))
+ }
```

```
> x<-c(506,500,495,488,504,486,505,513,521,520,512,485)
> CI.function(x,10,0.1)
[1] 498.1684 507.6650
> CI.function(x,10,0.05)
[1] 497.2587 508.5746
```

大样本：正态总体、 σ^2未知；或非正态总体， σ^2未知

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \qquad \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \qquad \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim N(0,1)$$

- 置信水平为（1-α）时，置信区间为

$$\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2} \le \mu \le \bar{X} + \frac{S}{\sqrt{n}} z_{\alpha/2}$$

例9 在一家大公司的49名员工的样本中，这些员工一年中平均有7天请病假，其标准差为2.5天。请给出整个大公司的员工一年中平均请病假的天数。

- 大样本、标准差未知：

$$\bar{x}-\frac{S}{\sqrt{n}}z_{\alpha/2}\leq\mu\leq\bar{x}+\frac{S}{\sqrt{n}}z_{\alpha/2}$$

$$7-1.96\frac{2.5}{\sqrt{49}}\leq\mu\leq7+1.96\frac{2.5}{\sqrt{49}}$$

$$6.3\leq\mu\leq7.7$$

小样本、正态总体、 σ^2未知

- 假定总体服从正态分布，总体方差未知，样本较小（$n < 30$）;
  - 需要使用t分布统计量 $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$
  - 置信水平为（1-α）时，置信区间为

$$\bar{X} - \frac{s}{\sqrt{n}} t_{\alpha/2}(n-1) \le \mu \le \bar{X} + \frac{s}{\sqrt{n}} t_{\alpha/2}(n-1)$$

◆ 如果是小样本、非正态总体、 σ^2未知，则无法计算

例 8 包糖机某日开工包了12包糖，称得重量（单位：克）分别为

506，500，495，488，504，486，505，513，521，520，512，485.

假设重量服从正态分布，~~且标准差为 σ = 10~~，试求糖包的平均重量 μ 的1-α置信区间（分别取α = 0. 10和α = 0. 05）.

$$\left[\bar{X}-\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1),\ \bar{X}+\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right]$$

- 标准差未知：

```
> x<-c(506,500,495,488,504,486,505,513,521,520,512,485)
> t.test(x,conf.level = 0.9)$conf.int
[1] 496.4360 509.3973
attr(,"conf.level")
[1] 0.9
```

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

```
> t.test(x,conf.level = 0.90)

        One Sample t-test

data:  x
t = 139.37, df = 11, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 496.4360 509.3973
sample estimates:
mean of x
 502.9167
```

statistic	the value of the t-statistic.
parameter	the degrees of freedom for the t-statistic.
p.value	the p-value for the test.
conf.int	a confidence interval for the mean appropriate to the specified alternative hypothesis.
estimate	the estimated mean or difference in means depending on whether it was a one-sample test or a two-sample test.

```
> t.test(x,conf.level = 0.9)$conf.int
[1] 496.4360 509.3973
attr(,"conf.level")
[1] 0.9
```

test or a

alternative	a character string describing the alternative hypothesis.
method	a character string indicating what type of t-test was performed.
data.name	a character string giving the name(s) of the data.

例 8 包糖机某日开工包了12包糖，称得重量(单位:克)分别为

506, 500, 495, 488, 504, 486, 505, 513, 521, 520, 512, 485.

假设重量服从正态分布，~~且标准差为 $\sigma = 10$~~, 试求糖包的平均重量 μ 的 $1-\alpha$ 置信区间(分别取 $\alpha = 0.10$ 和 $\alpha = 0.05$).

假设当标准差未知, 求 $\alpha = 0.05$ 的时候平均重量 μ 的 $1-\alpha$ 置信区间：

```
> x<-c(506,500,495,488,504,486,505,513,521,520,512,485)
> t.test(x,conf.level = 0.95)$conf.int
[1] 494.9742 510.8592
attr(,"conf.level")
[1] 0.95
```

小结

总体分布	样本容量	σ已知	σ未知
正态分布	大样本	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$
	小样本	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$
非正态分布	大样本	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$

第五章:统计推断 Ⅰ

- ➢统计推断
- ➢抽样分布
- ➢参数估计的基本原理
- ➢一个总体参数的区间估计
- ➢两个总体参数的区间估计

区间估计-两正态总体

假设有两个正态总体$X \sim N$（$\mu_1, \sigma_2{}^2$）, X_1, X_2, …, X_n, 为来自总体X的一个样本，Y_1, Y_2, …, Y_n, 为来自总体Y的一个样本，1-α为置信度，$\bar{X}$、$\bar{Y}$分别为第一、第二样本均值，S_1^2, S_2^2分别为第一、第二样本方差。

1. 两方差都已知时, 均值差$\mu_1-\mu_2$的区间估计

$$\bar{X}-\bar{Y} \sim N(\mu_1-\mu_2, \frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2})$$

构造统计量： $Z=\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$

置信区间: $\left[\bar{X}-\bar{Y}-Z_{\alpha/2}\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}},\ \bar{X}-\bar{Y}+Z_{\alpha/2}\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}}\right]$

区间估计-两正态总体

假设有两个正态总体$X \sim N$（$\mu_1, \sigma_2{}^2$），X_1，X_2，…，X_n，为来自总体X的一个样本，Y_1，Y_2，…，Y_n，为来自总体Y的一个样本，1-α为置信度，$\bar{X}$、$\bar{Y}$分别为第一、第二样本均值，S_1^2，S_2^2分别为第一、第二样本方差。

1. 两方差都已知时，均值差$\mu_1 - \mu_2$的区间估计

置信区间:$\left[\bar{X} - \bar{Y} - Z_{\alpha/2}\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}},\ \bar{X} - \bar{Y} + Z_{\alpha/2}\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}\right]$

```
> f.tmp.sample<-function(x,y,alpha,sigma1,sigma2){
+   n1=length(x);n2=length(y)
+   xbar=mean(x)-mean(y)
+   zstar=qnorm(1-alpha/2)
+   sigma=(sigma1/n1+sigma2/n2)^(1/2)
+   return(xbar+c(-zstar*sigma,zstar*sigma))
+ }
```

例9 为比较两个小麦品种的产量，选择18块条件相似的试验田，采用相同的耕作方法做实验，结果播种甲品种的8块实验田的单位面积产量和播种乙品种的10块实验田的单位面积产量分别为：

甲品种	628	583	510	554	612	523	530	615		
乙品种	535	433	398	470	567	480	498	560	503	426

假定每个品种的单位面积产量均服从正态分布，甲品种产量的方差为2140，乙品种产量的方差为3250，试求两个品种平均面积产量差的置信区间（取α=0.05）

方差已知还是未知？

方差已知：$[\bar{X}-\bar{Y}-Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}}, \bar{X}-\bar{Y}+Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}}]$

```
> f.tmp.sample<-function(x,y,alpha,sigma1,sigma2){
+   n1=length(x);n2=length(y)
+   xbar=mean(x)-mean(y)
+   zstar=qnorm(1-alpha/2)
+   sigma=(sigma1/n1+sigma2/n2)^(1/2)
+   return(xbar+c(-zstar*sigma,zstar*sigma))
+ }
```

```
> x<-c(628,583,510,554,612,523,530,615)
> y<-c(535,433,398,470,567,480,498,560,503,426)
```

```
> f.tmp.sample(x,y,0.05,2140,3250)
[1]  34.66688 130.08312
```

这两个品种平均面积产量差的置信区间为（34.69，130.08）

区间估计-两正态总体

（2）当两总体的方差相同，$\sigma_1{}^2=\sigma_2{}^2=\sigma^2$,且未知时

$$\Leftarrow \begin{cases} Z=\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim N(0,1) \\ \dfrac{(n_1-1)S_1^2}{\sigma^2}\sim x^2(n_1-1),\ \dfrac{(n_2-1)S_2^2}{\sigma^2}\sim x^2(n_2-1) \\ \dfrac{(n_1-1)S_1^2}{\sigma^2}+\dfrac{(n_2-1)S_2^2}{\sigma^2}\sim x^2(n_1+n_2-2) \end{cases}$$

构造统计量：$T=\dfrac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})S^2}}\sim t(n_1+n_2-2)$,

置信区间: $\bar{X}-\bar{Y}\pm\dfrac{S}{\sqrt{n}}t_{\alpha/2}(n_1+n_2-2)$

$X\sim N(0,1), Y\sim X^2(n)$，且 X 和 Y 相互独立，则 $\dfrac{X}{\sqrt{Y/n}}$ 服从自由度为n的t分布，$\sim t(n)$

区间估计-两正态总体

(3) 当两总体的方差 σ_1^2 和 σ_2^2 未知，且$\sigma_1^2 \neq \sigma_2^2$时

构造统计量： $T = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{{S_1}^2}{n_1} + \dfrac{{S_2}^2}{n_2}}} \sim t(v)$

自由度：$\hat{v} = \left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2 / (\dfrac{(S_1^2)^2}{n_1^2(n_1 - 1)} + \dfrac{(S_2^2)^2}{n_2^2(n_2 - 1)})$

置信区间: $[\bar{X} - \bar{Y} - t_{\frac{\alpha}{2}}(\hat{v})\sqrt{\dfrac{{S_1}^2}{n_1} + \dfrac{{S_2}^2}{n_2}}, \bar{Y} - \bar{X} + t_{\frac{\alpha}{2}}(\hat{v})\sqrt{\dfrac{{S_1}^2}{n_1} + \dfrac{{S_2}^2}{n_2}}]$

区间估计-两正态总体

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

说明: 若仅出现数据x, 则进行单样本t检验; 若出现数据x和y, 则进行二样本的t检验(见6.3节); `alternative=c("two.sided", "less", "greater")`用于指定所求置信区间的类型; `alternative="two.sided"`是缺省值, 表示求置信区间`alternative="less"`表示求置信上限; `alternative="greater"`表示求置信下限. `mu`表示均值, 它仅在假设检验中起作用, 默认值为零.

区间估计-两正态总体

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

（2）当两总体的方差相同且未知，

R语言：`> t.test(x,y,conf.level=1-alpha,var.equal = T)$conf.int`

(3) 当两总体的方差σ_1^2和σ_2^2未知，且$\sigma_1^2 \neq \sigma_2^2$时

R语言：`> t.test(x,y,conf.level=1-alpha,var.equal = F)$conf.int`

例9 为比较两个小麦品种的产量，选择18块条件相似的试验田，采用相同的耕作方法做实验，结果播种甲品种的8块实验田的单位面积产量和播种乙品种的10块实验田的单位面积产量分别为：

甲品种	628	583	510	554	612	523	530	615		
乙品种	535	433	398	470	567	480	498	560	503	426

假定每个品种的单位面积产量均服从正态分布，~~甲品种产量的方差为2140，乙品种产量的方差为3250~~，试求两个品种平均面积产量差的置信区间（取α=0.05）

若方差未知：

```
> t.test(x,y,var.equal = T)$conf.int
[1]  29.46961 135.28039
attr(,"conf.level")
[1] 0.95
> t.test(x,y,var.equal = F)$conf.int
[1]  30.74046 134.00954
attr(,"conf.level")
[1] 0.95
```

方差是否相等？

两方差比$\sigma_1{}^2/\sigma_2{}^2$的置信区间

方差之比的区间估计

$$\left.\begin{array}{l}\dfrac{(n_1-1)S_1{}^2}{\sigma_1{}^2}\sim\chi^2(n_1-1),\\ \dfrac{(n_2-1)S_2{}^2}{\sigma_2{}^2}\sim\chi^2(n_2-1)\end{array}\right\}$$

定义：X~$\chi^2(m)$, Y~$\chi^2(n)$,且相互独立，则$\frac{X/m}{Y/n}$服从自由度为m和n的*F*分布，~*F*(m,n)

构造统计量：$F=\dfrac{S_1{}^2/\sigma_1{}^2}{S_2{}^2/\sigma_2{}^2}\sim F(n_1-1,n_2-1)$

$$P\left(F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\le\frac{S_1{}^2}{S_2{}^2}\cdot\frac{\sigma_1{}^2}{\sigma_2{}^2}\le F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\right)=1-\alpha$$

置信区间：$\left(\dfrac{S_1{}^2}{S_2{}^2}\cdot F_{\frac{\alpha}{2}}(n_1-1,n_2-1),\dfrac{S_1{}^2}{S_2{}^2}\cdot F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\right)$

R语言：
```
> var.test(x,y,conf.level=1-alpha)
```

R语言：> `var.test(x,y,conf.level=1-alpha)`

Description

Performs an F test to compare the variances of two samples from normal populations.

Usage

```
var.test(x, ...)

## Default S3 method:
var.test(x, y, ratio = 1,
         alternative = c("two.sided", "less", "greater"),
         conf.level = 0.95, ...)
```

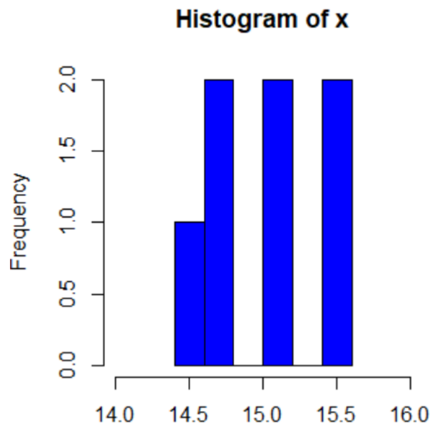
例10 有两台机床生产同一型号的滚珠，根据以往经验知，这两台机床生产的滚珠直径都服从正态分布，现分别从这两台机床生产的滚珠中随机地抽取7个和9个，测得它们的直径如下（单位:mm）

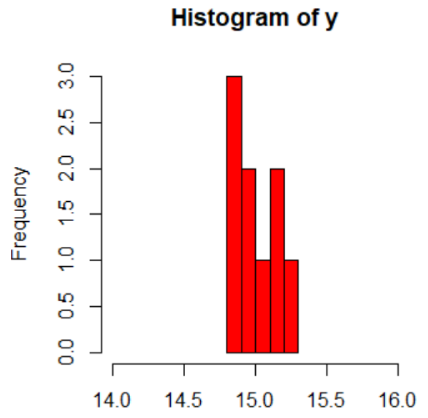
机床甲	15.2	14.5	15.5	14.8	15.1	15.6	14.7		
机床乙	15.2	15.0	14.8	15.2	15.0	14.9	15.1	14.8	15.3

试问机床甲生产的滚珠的方差是否比机床乙生产的滚珠直径的方差大？

$$\frac{\sigma_1{}^2}{\sigma_2{}^2} > 1$$

```
> x<-c(15.2,14.5,15.5,14.8,15.1,15.6,14.7)
> y<-c(15.2,15.0,14.8,15.2,15.0,14.9,15.1,14.8,15.3)
> var.test(x,y,conf.level=1-0.05)$conf.int
[1]  1.121337 29.208290
attr(,"conf.level")
[1] 0.95
```

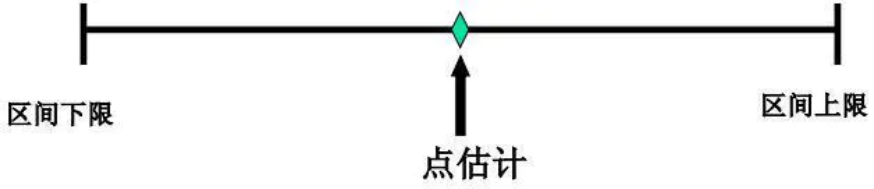




方差比值95%的置信区间为[1.21,29.2]，可以推断甲的方差应该比乙的要大

区间估计—定义

- 围绕点估计值构造总体参数的一个区间，这就是区间估计。



$$P(\underline{\theta} < \hat{\theta} < \overline{\theta}) = 1-\alpha$$

置信下限　置信上限　置信水平

精确性　可靠性

区间估计—基本步骤

(1)根据已知条件，构造合适的统计量

(2)根据统计量和给定显著水平，确定置信下限和置信上限

(1) 根据已知条件，构造合适的统计量

统计量

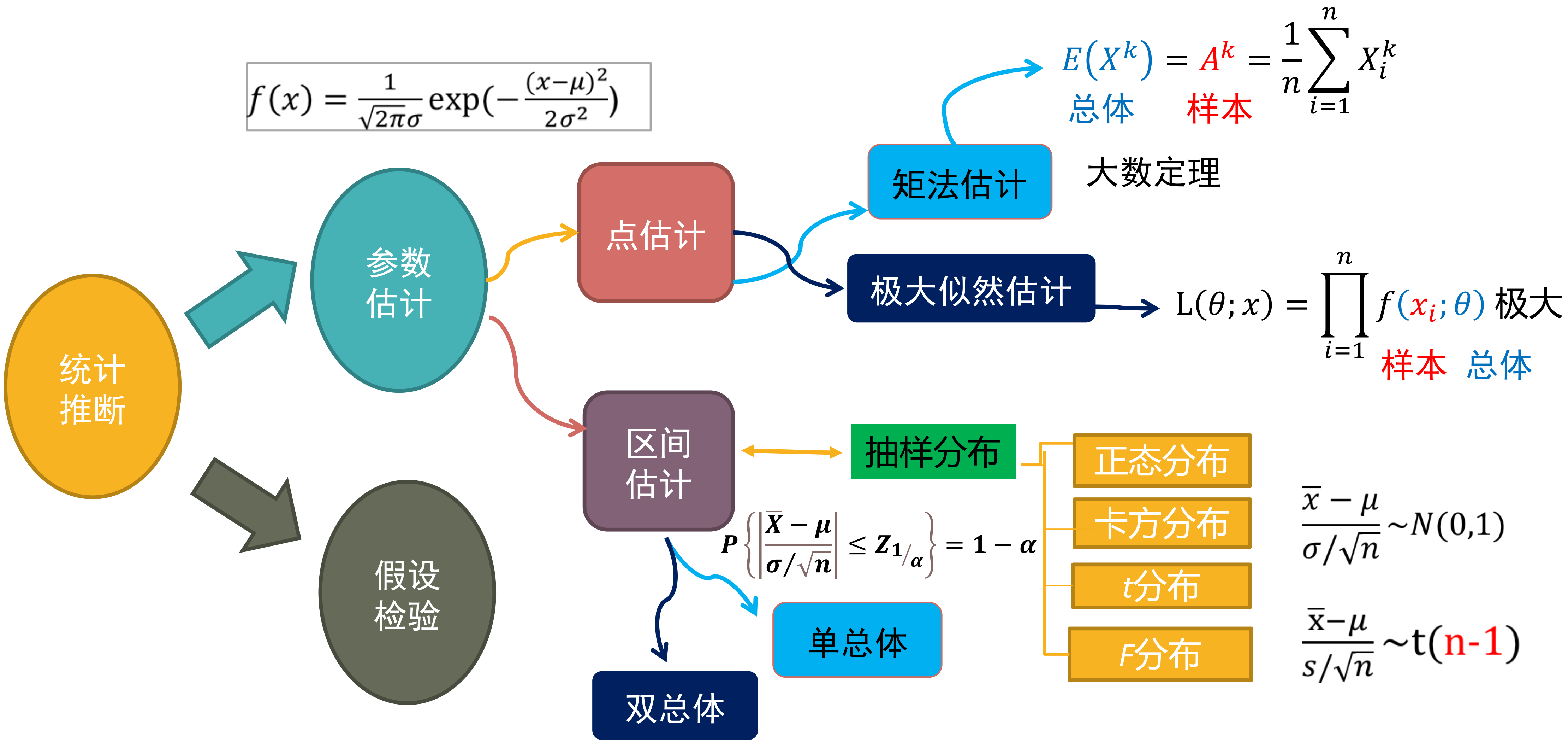
- 单正态总体 μ的置信区间
  - 方差已知：$\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
  - 方差未知：$\dfrac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$
- 双正态总体
  - $\mu_1-\mu_1$的置信区间
    - 方差已知：$Z=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
    - 方差未知
      - $\sigma_1^2=\sigma_2^2=\sigma^2$：$T=\dfrac{\overline{X}-\overline{Y}}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$
      - $\sigma_1^2\neq\sigma_2^2$：$T=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t(\widehat{\nu})$
        $\widehat{\nu}=\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2 \Big/ \left(\dfrac{(S_1^2)^2}{n_1^2(n_1-1)}+\dfrac{(S_2^2)^2}{n_2^2(n_2-1)}\right)$
  - ${\sigma_1}^2/{\sigma_2}^2$的置信区间
    - $F=\dfrac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_2}^2} \sim F(n_1-1,n_2-1)$

(2) 根据统计量和给定显著水平，确定置信下限和置信上限

置信区间

- 单正态总体 μ的置信区间
  - 方差已知：$\left(\bar{X}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\ \bar{X}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right)$　qnorm(p,...)
  - 方差未知：$\left[\bar{X}-\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1),\ \bar{X}+\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right]$　t.test(...)
- 双正态总体
  - $\mu_1-\mu_1$的置信区间
    - 方差已知：$\left[\bar{X}-\bar{Y}-Z_{\alpha/2}\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}},\ \bar{X}-\bar{Y}+Z_{\alpha/2}\sqrt{\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}}\right]$
    - 方差未知
      - $\sigma_1^2=\sigma_2^2=\sigma^2$：$\bar{X}-\bar{Y}\pm\frac{S}{\sqrt{n}}t_{\alpha/2}(n_1+n_2-2)$
      - t.test(x,y. var.equal=...)
      - $\sigma_1^2\neq\sigma_2^2$：$[\bar{X}-\bar{Y}\pm t_{\frac{\alpha}{2}}(\hat{v})\sqrt{\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}}$
  - $\sigma_1{}^2/\sigma_2{}^2$的置信区间：$\left(\frac{S_1{}^2}{S_2{}^2}\cdot F_{\frac{\alpha}{2}}(n_1-1,n_2-1),\frac{S_1{}^2}{S_2{}^2}\cdot F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\right)$　var.test(x,y. =...)

统计推断-小结


$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$
统计推断
参数估计
假设检验
点估计
区间估计
矩法估计
大数定理
$E(X^k)=A^k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$
总体
样本
极大似然估计
$\mathrm{L}(\theta;x)=\prod_{i=1}^{n}f(x_i;\theta)$ 极大
样本
总体
抽样分布
正态分布
卡方分布
t分布
F分布
$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$
$\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$
$P\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|\le Z_{1/\alpha}\right\}=1-\alpha$
单总体
双总体

练习题

1. 产生符合正态分布的100个随机数，然后用矩估计得到其期望和方差的估计值。
2. 设元件无故障工作时间X服从指数分布（λ未知），随机抽取样本寿命如下：
   518，612，713，388，424，试用极大似然估计求λ的点估计。
3. 正常人的脉搏平均每分钟72次，医生测得10例中毒患者的每分钟的脉搏次数如下：
   54，67，68，78，70，66，67，70，65，69
   已知脉搏次数服从正态分布。试计算这10名患者的平均脉搏次数的点估计和95%的区间估计，试分析这10名患者是否正常？
4. 甲乙两种水稻种在10块试验田中，每块田中甲、乙各一半，假设两种水稻产量X，Y服从正态分布，且方差相等，收获后产量如下（千克）。求：
   （1）两种水稻产量期望差的95%的置信区间；
   （2）方差比的区间估计，并判定方差是否相等，若不相等，试重新计算问题（1）。

甲种	140，137，136，140，145，148，140，135，144，141
乙种	135，118，115，140，128，131，130，115，131，125