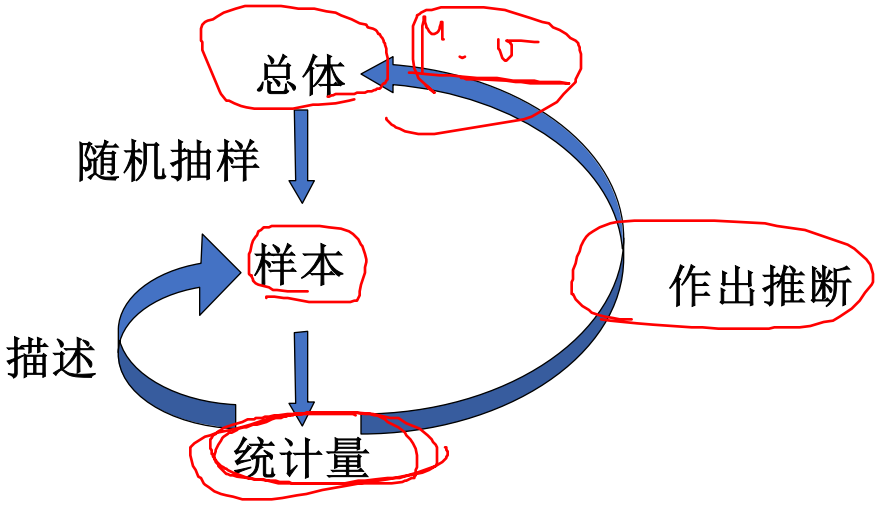


研究统计量的性质和评价一个统计推断的优良性，取决于其抽样分布的性质.

参数估计问题的一般提法

设有一个统计总体，总体的分布函数为

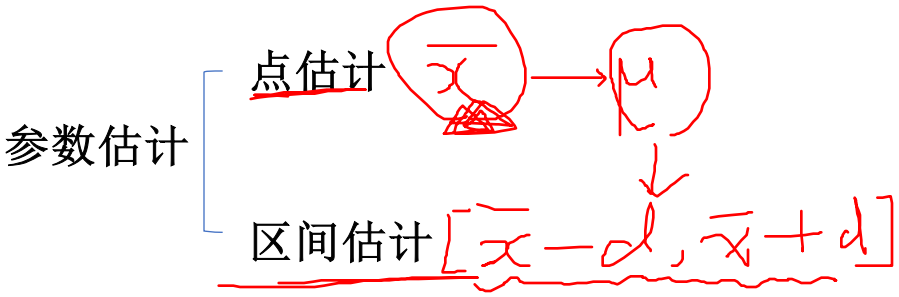
$F(x,\theta)$，其中θ为未知参数(θ 可以是向量).

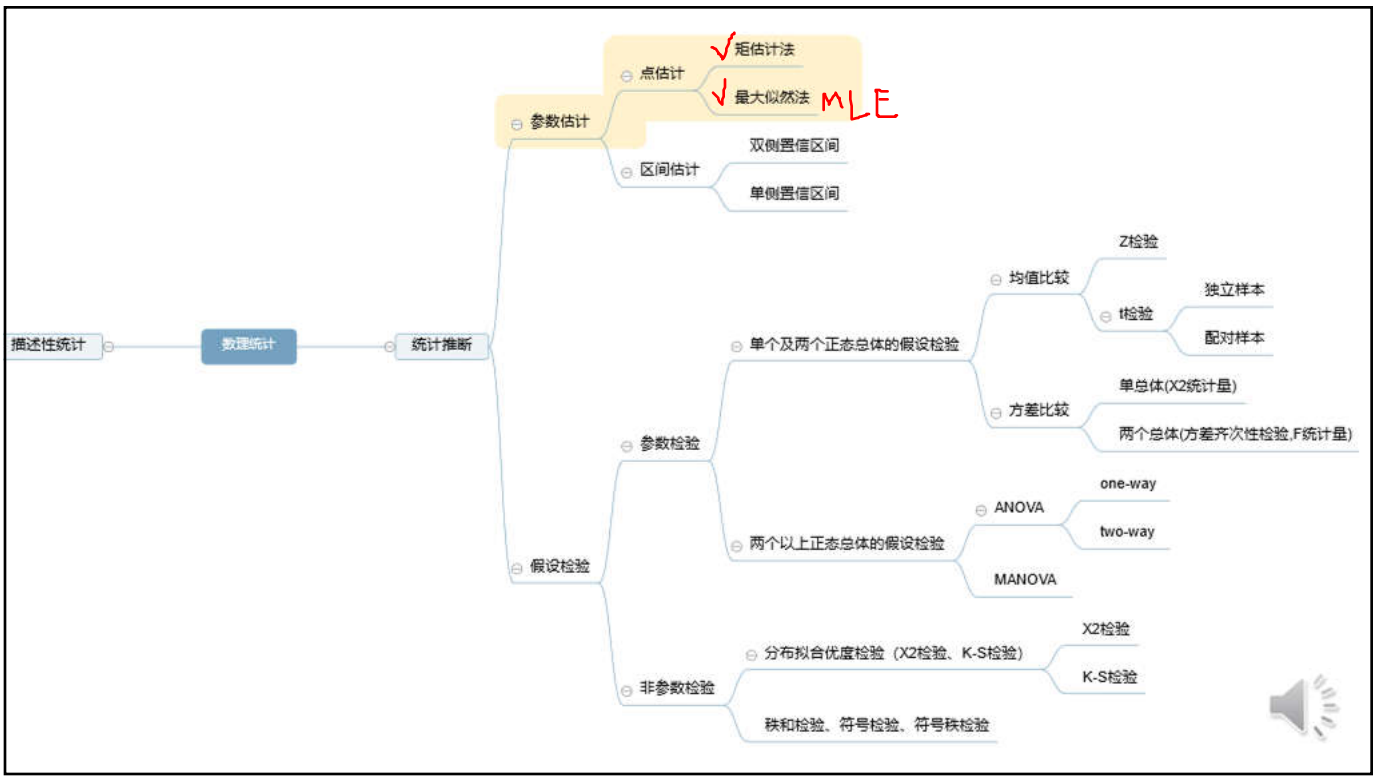
现从该总体抽样，得样本

$$X_1, X_2, \ldots, X_n$$

要依据该样本对产生该样本的参数 θ 作出估计,或估计 θ 的某个已知函数 $g(\theta)$.

这类问题称为参数估计.





举例：为估计总体均值 μ

我们需要构造出适当的样本的函数/统计量 $T(X_1,X_2,\ldots X_n)$，每当有了样本，就代入该函数中算出一个值，用来作为 μ 的估计值.

$T(X_1,X_2,\ldots X_n)$ 称为参数 μ 的点估计量，

把样本值代入 $T(X_1,X_2,\ldots X_n)$ 中，得到 μ 的一个点估计值.

我们知道, 若 $X\sim N(\mu,\sigma^2)$ 则 $E(X)=\mu$

由大数定律,

样本体重的平均值

$$\lim_{n\to\infty} P\{|\frac{1}{n}\sum_{i=1}^{n} X_i-\mu|<\varepsilon\}=1$$

自然想到把样本体重的平均值作为总体平均体重的一个估计.

用样本体重的均值 $\bar{X}$ 估计 μ .

类似地，用样本体重的方差 S^2 估计 σ^2.

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

二、寻求估计量的方法

1. 矩估计法
2. 极大似然法
3. 最小二乘法
4. 贝叶斯方法 ……

这里我们主要介绍前面两种方法.

1. 矩估计法

矩估计法是英国统计学家*K.*皮尔逊最早提出来的.由辛钦定理,

若总体 X 的数学期望 $E(X)=\mu$ 有限,则有

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

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \xrightarrow{P} E(X^k)=\mu_k \ (k=1,2,\cdots)$$

$$g(A_1,A_2,\cdots,A_k) \xrightarrow{P} g(\mu_1,\mu_2,\cdots,\mu_k)$$

其中**g** 为连续函数.

这表明,当样本容量很大时,在统计上,可以用样本矩去估计总体矩.这一事实导出矩估计法.

定义 用样本原点矩估计相应的总体原点矩,又用样本原点矩的连续函数估计相应的总体原点矩的连续函数,这种参数点估计法称为矩估计法.

理论依据: 大数定律

矩估计法的具体做法如下

设总体的分布函数中含有***k***个未知参数 $\theta_1,\theta_2,\cdots,\theta_k$,那么它的前***k***阶矩 $\mu_1,\mu_2,\cdots,\mu_k$,

都是这 k 个参数的函数,记为:

$$\mu_i = \mu_i(\theta_1, \theta_2, \cdots, \theta_k) \qquad i=1,2, \ldots, k$$

从这 k 个方程中解出

$$\theta_j = \theta_j(\mu_1, \mu_2, \cdots, \mu_k) \qquad j=1,2,\ldots,k$$

那么用诸 μ_i 的估计量 A_i 分别代替上式中的诸 μ_i，即可得诸 θ_j 的矩估计量：

$$\hat{\theta}_j = \theta_j(A_1, A_2, \cdots, A_k) \qquad j=1,2,\ldots,k$$

矩估计量的观察值称为矩估计值.

例2　设总体 X 在 $[a, b]$ 上服从均匀分布，a，b 未知．$X_1, \ldots, X_n$ 是来自 X 的样本，试求 a，b 的矩估计量.

解　$\mu_1 = E(X) = \dfrac{a+b}{2}$

$$\mu_2 = E(X^2) = D(X) + [E(X)]^2$$

$$= \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4}$$

即 $\begin{cases} a+b=2\mu_1 \\ b-a=\sqrt{12(\mu_2-\mu_1^2)} \end{cases}$

解得 $\begin{cases} a=\mu_1-\sqrt{3(\mu_2-\mu_1^2)} \\ b=\mu_1+\sqrt{3(\mu_2-\mu_1^2)} \end{cases}$

总体矩

于是 a , b 的矩估计量为

$$\hat{a}=\bar{X}-\sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\ ,\qquad \hat{b}=\bar{X}+\sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

样本矩

2. 最大似然法（MLE）

它是在总体类型已知条件下使用的一种参数估计方法.

它首先是由德国数学家高斯在1821年提出的. 然而,这个方法常归功于英国统计学家费歇.

Gauss

Fisher

费歇在1922年重新发现了这一方法，并首先研究了这种方法的一些性质.

似然 (likelihood)

$$L(\theta|x)=P(X=x|\theta)$$

未知量

已知

StatQuest:
Probability vs Likelihood

最大似然估计原理：

设$X_1,X_2,\ldots X_n$是取自总体X的一个样本，样本的联合密度(连续型）或联合分布律 (离散型)为$f(x_1,x_2,\ldots,x_n;\theta)$.

当给定样本$X_1,X_2,\ldots X_n$时，定义似然函数为：

$$L(\theta)=f(x_1,x_2,\ldots,x_n;\theta)$$

这里 $x_1,x_2,\ldots,x_n$ 是样本的观察值.

似然函数：

$$L(\theta)=f(x_1,x_2,\ldots,x_n;\theta)$$

$L(\theta)$ 看作参数θ 的函数，它可作为θ 将以多大可能产生样本值$x_1,x_2,\ldots,x_n$的一种度量.

最大似然估计法就是用使 $L(\theta)$ 达到最大值的 $\hat{\theta}$ 去估计 θ .

$$L(\hat{\theta})=\max_{\theta}L(\theta)$$

称 $\hat{\theta}$为 θ 的最大似然估计值. 而相应的统计量 $\hat{\theta}(X_1,\ldots,X_n)$ 称为 θ 的最大似然估计量.

两点说明：

1、求似然函数$L(\theta)$的最大值点，可以应用微积分中的技巧。由于$\ln(x)$是x的增函数，$\ln L(\theta)$与$L(\theta)$在θ的同一值处达到它的最大值，假定θ是一实数，且$\ln L(\theta)$是θ的一个可微函数。通过求解方程：

$$\frac{d \ln L(\theta)}{d\theta} = 0$$

可以得到θ的MLE (Maximum Likelihood Estimate)

若θ是向量，上述方程须用方程组代替.

两点说明：

1、求似然函数$L(\theta)$的最大值点，可以应用微积分中的技巧。由于$\ln(x)$是x的增函数，$\ln L(\theta)$与$L(\theta)$在θ的同一值处达到它的最大值，假定θ是一实数，且$\ln L(\theta)$是θ的一个可微函数。通过求解方程：

$$\frac{d \ln L(\theta)}{d\theta} = 0$$

可以得到θ的MLE (Maximum Likelihood Estimate)

若θ是向量，上述方程须用方程组代替.

第二节 估计量的评选标准

问题： 使用什么样的统计量去估计整体均值 μ？

可以用样本均值;

也可以用样本中位数或众数;

还可以用别的统计量.

常用的几条标准是：

1. 无偏性
2. 有效性
3. 相合性

一、无偏性

估计量是随机变量，对于不同的样本值会得到不同的估计值．我们希望估计值在未知参数真值附近摆动，而它的期望值等于未知参数的真值.这就导致无偏性这个标准．

设 $\hat{\theta}(X_1,\cdots,X_n)$ 是未知参数θ 的估计量，若

$$E(\hat{\theta}) = \theta$$

则称$\hat{\theta}$ 为 θ 的无偏估计．

二、有效性

设 $\hat{\theta}_1 = \hat{\theta}_1(X_1,\cdots,X_n)$ 和 $\hat{\theta}_2 = \hat{\theta}_2(X_1,\cdots,X_n)$

都是参数θ 的无偏估计量，若对任意 $\theta \in \Theta$，

$$D(\hat{\theta}_1) \le D(\hat{\theta}_2)$$

且至少对于某个$\theta \in \Theta$上式中的不等号成立，

则称 $\hat{\theta}_1$ 较 $\hat{\theta}_2$有效．

三、相合性

设 $\hat{\theta}(X_1,\ldots,X_n)$ 是参数 θ 的估计量，若对于任意 $\theta\in\Theta$，当 $n\to\infty$ 时 $\hat{\theta}(X_1,\ldots,X_n)$ 依概率收敛于 θ，则称 $\hat{\theta}$ 为 θ 的相合估计量.

$\hat{\theta}$ 为 θ 的相合估计量

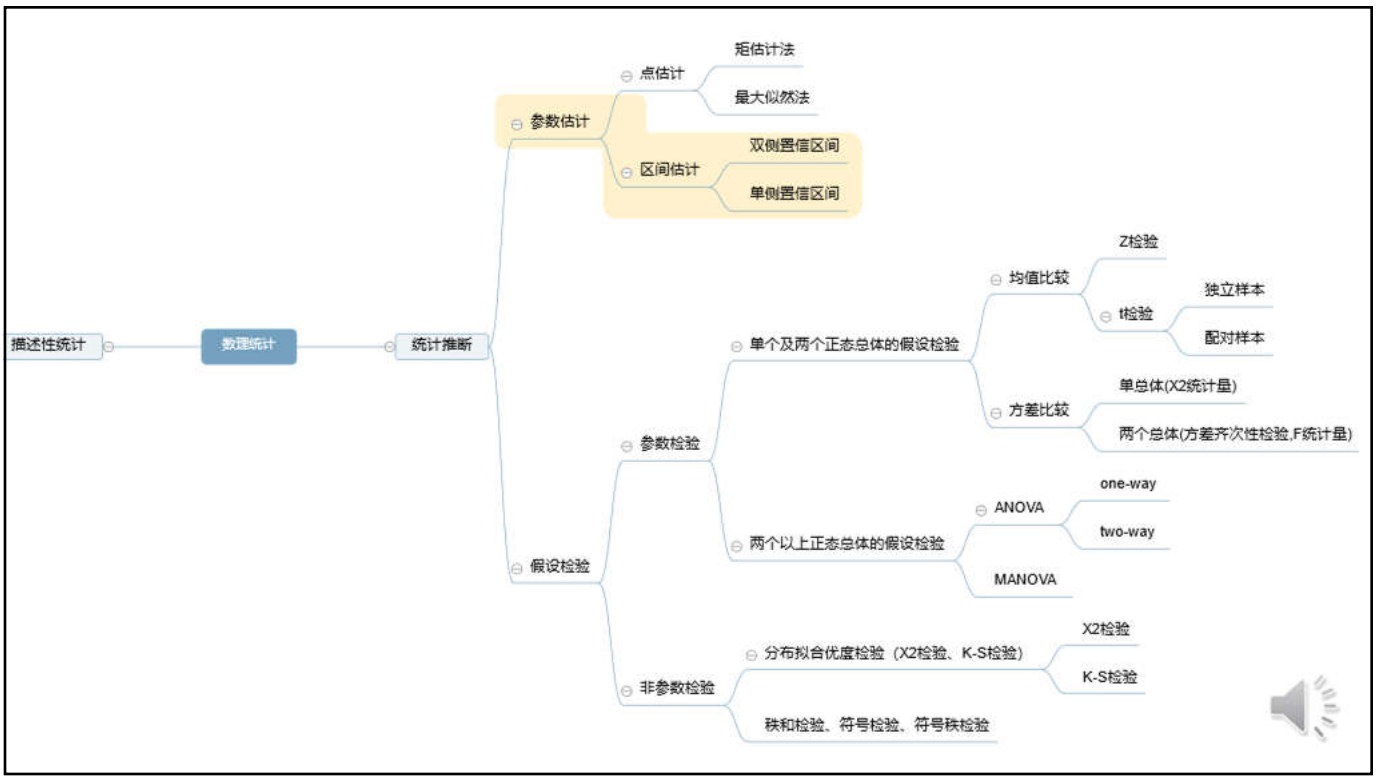
$\Leftrightarrow$ 对于任意 $\varepsilon>0$，有

$$\lim_{n\to\infty} P\{|\hat{\theta}-\theta|<\varepsilon\}=1,\quad \theta\in\Theta$$

A consistent estimator（相合估计量） in statistics is such an estimate which **hones in on the true value of the parameter being estimated more and more accurately as the sample size increases.** 会受益于样本量的增加

第三节　区间估计

- 置信区间定义
- 置信区间的求法
- 单侧置信区间


描述性统计
数理统计
统计推断
参数估计
点估计
矩估计法
最大似然法
区间估计
双侧置信区间
单侧置信区间
假设检验
参数检验
单个及两个正态总体的假设检验
均值比较
Z检验
t检验
独立样本
配对样本
方差比较
单总体(X2统计量)
两个总体(方差齐次性检验,F统计量)
两个以上正态总体的假设检验
ANOVA
one-way
two-way
MANOVA
非参数检验
分布拟合优度检验（X2检验、K-S检验）
X2检验
K-S检验
秩和检验、符号检验、符号秩检验

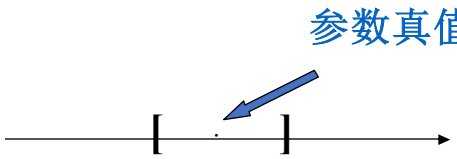

JBstatistics

Examples:

- We may be 95% confident that μ lies in the interval $(-0.2, 3.1)$
- We may be 99% confident that σ lies in the interval $(2.5, 13.4)$

置信区间是针对“参数”的，而不是“统计量”的

我们希望确定一个区间，使我们能以比较高的可靠程度相信它包含真参数值.



The confidence level is designated before examining the data. Most commonly, a **95%** confidence level is used.
通常采用95%的CI（置信区间）

这里所说的“可靠程度”是用概率来度量的，称为置信度或置信水平(confidence level).

习惯上把置信水平记作 $1-\alpha$ ，这里α是一个很小的正数.

置信水平的大小是根据实际需要选定的.

例如，通常可取置信水平 $1-\alpha$ =0.95或0.9等.

根据一个实际样本，由给定的置信水平，我们求出一个尽可能小的区间 $(\underline{\theta},\overline{\theta})$，使

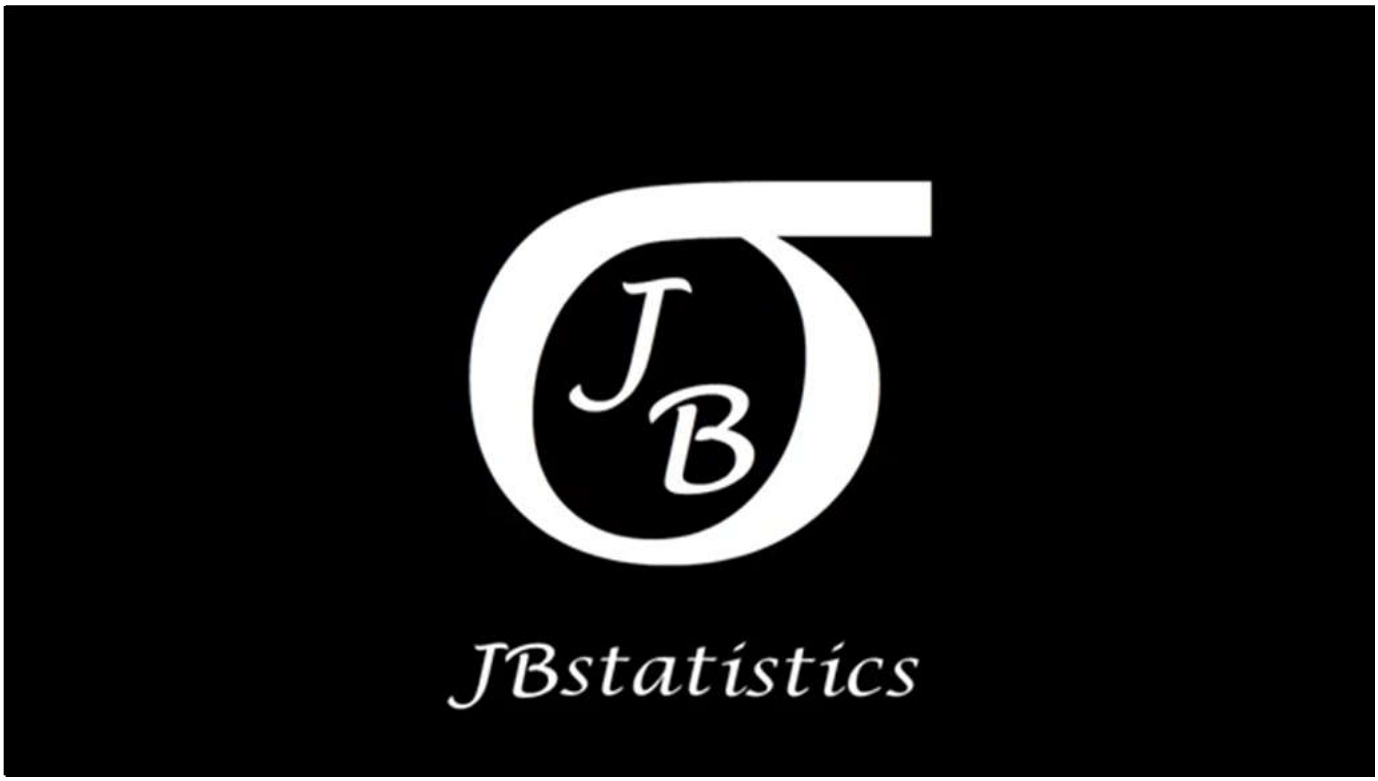
$$P\{\underline{\theta}<\theta<\overline{\theta}\}=1-\alpha$$

称区间 $(\underline{\theta},\overline{\theta})$ 为 θ 的置信水平为 $1-\alpha$ 的置信区间.

$\underline{\theta}$ 和 $\overline{\theta}$ 分别称为置信下限和置信上限.

1. 置信区间（CI）的第一种求解方法：利用抽样分布

*需要事先知道抽样分布



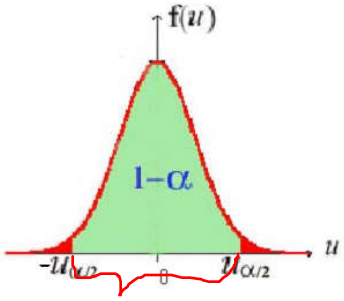
对于给定的置信水平，根据Z的分布，确定一个区间，使得Z取值于该区间的概率为置信水平.

已知 $Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ ~N（0，1）

对给定的置信水平 $1-\alpha$，

查正态分布表得 $u_{\alpha/2}$，使

$$P\{|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}| \le u_{\alpha/2}\}=1-\alpha$$



从中解得 $P\{\bar{X}-\dfrac{\sigma}{\sqrt{n}}u_{\alpha/2} \le \mu \le \bar{X}+\dfrac{\sigma}{\sqrt{n}}u_{\alpha/2}\}=1-\alpha$

从例1解题的过程，我们归纳出求置信区间的一般步骤如下：

1. 明确问题，是求什么参数的置信区间？置信水平$1-\alpha$是多少？

2. 寻找参数θ的一个良好的点估计

$$T(X_1, X_2, \ldots X_n)$$

3. 寻找一个待估参数θ和估计量T的函数$U(T,\theta)$，且其分布（抽样分布）为已知.

4. 对于给定的置信水平$1-\alpha$，根据$U(T,\theta)$的分布，确定常数a,b，使得

$$P(a<U(T,\theta)<b) = 1-\alpha$$

5. 对“$a<U(T,\theta)<b$”作等价变形，得到如下形式：

$$\underline{\theta} < \theta < \overline{\theta}$$

即

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

于是$(\underline{\theta},\overline{\theta})$就是$\theta$的$100(1-\alpha)\%$的置信区间.

置信水平

单侧置信区间

上述置信区间中置信限都是双侧的，但对于有些实际问题，人们关心的只是参数在一个方向的界限.

例如对于设备、元件的使用寿命来说，平均寿命过长没什么问题，过短就有问题了.

这时,可将置信上限取为+∞，而只着眼于置信下限，这样求得的置信区间叫单侧置信区间.

于是引入单侧置信区间和置信限的定义：

定义 设 θ 是一个待估参数，给定 $\alpha > 0$, 若由样本$X_1, X_2, \ldots X_n$确定的统计量

$$\underline{\theta} = \underline{\theta}(X_1, X_2, \cdots, X_n)$$

对于任意 $\theta \in \Theta$,满足

$$P\{\theta \geq \underline{\theta}\} = 1 - \alpha$$

则称区间 $[\underline{\theta}, +\infty)$是 θ 的置信水平为$1-\alpha$的单侧置信区间. $\underline{\theta}$ 称为 θ 的置信水平为 $1-\alpha$ 的单侧置信下限.

正态总体均值与方差的区间估计

- 单个总体 $N(\mu,\sigma^2)$ 的情况
- 两个总体 $N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$ 的情况

一、单个总体$N(\mu,\sigma^2)$的情况

$X\sim N(\mu,\sigma^2)$, 并设 $X_1,\dots,X_n$ 为来自总体的样本， $\bar{X},S^2$ 分别为样本均值和样本方差.

均值 μ 的置信区间

1° σ^2为已知

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

可得到 μ的置信水平为 $1-\alpha$ 的置信区间为

$$(\bar{X}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}},\bar{X}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}) \quad 或 \quad (\overline{X}\pm\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}})$$

2° σ^2 为未知

$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

此分布不依赖于任何未知参数

由 $P\{|\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}| < t_{\frac{\alpha}{2}}(n-1)\} = 1-\alpha$

可得到 μ 的置信水平为 $1-\alpha$ 的置信区间为

$$(\bar{X} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1), \quad \bar{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1))$$

或 $$(\overline{X} \pm \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1))$$

2. 方差 σ^2 的置信区间

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

由

$$P\{\chi^2_{1-\frac{\alpha}{2}}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}(n-1)\} = 1-\alpha$$

可得到 σ^2 的置信水平为 $1-\alpha$ 的置信区间为

$$(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)})$$

由 $P\{\sqrt{\chi^2_{1-\frac{\alpha}{2}}(n-1)} < \frac{(n-1)S}{\sigma} < \sqrt{\chi^2_{\frac{\alpha}{2}}(n-1)}\} = 1-\alpha$

可得到标准差σ的置信水平为 $1-\alpha$ 的置信区间为

$$\left(\frac{\sqrt{n-1}S}{\sqrt{\chi^2_{\frac{\alpha}{2}}(n-1)}}, \frac{\sqrt{n-1}S}{\sqrt{\chi^2_{1-\frac{\alpha}{2}}(n-1)}}\right)$$

二、两个总体 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ 的情况

设已给定置信水平为 $1-\alpha$，并设 $X_1, X_2 \ldots, X_{n_1}$ 是来自第一个总体的样本，$Y_1, Y_2 \ldots, Y_{n_2}$ 是来自第二个总体的样本,这两个样本相互独立.且设 $\bar{X}, \bar{Y}$ 分别为第一、二个总体的样本均值，S_1^2, S_2^2 为第一、二个总体的样本方差.

1. 两个总体均值差 $\mu_1 - \mu_2$ 的置信区间

1° σ_1^2, σ_2^2 为已知

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

因为 X, Y 相互独立,所以 $\bar{X}, \bar{Y}$ 相互独立.

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

或
$$\frac{(\bar{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

于是得到 $\mu_1 - \mu_2$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$(\overline{X} - \overline{Y} \pm u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

2° $\sigma_1^2 = \sigma_2^2 = \sigma^2$, σ^2为未知

加权标准差

The weighted standard deviation

$$\frac{(\bar{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_\omega \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

其中 $S_\omega = \sqrt{S_\omega^2}$,

$$S_\omega^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

于是得到 $\mu_1-\mu_2$ 的置信水平为 $1-\alpha$ 的置信区间为

$$(\overline{X}-\overline{Y}\pm t_{\frac{\alpha}{2}}(n_1+n_2-2)S_\omega\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$$

其中 $S_\omega=\sqrt{S_\omega^2}$，$S_\omega^2=\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$.

t统计量仅取决于两个样本的均值和方差。根据中心极限定理（CLT），即使底层分布不符合正态，累加或平均后也会呈现出正态性。因此，t检验对（大多数）正态性的偏离是相当健壮/容忍的。但它对于方差齐次性的偏离仍是敏感的（使用时推荐先做方差齐次性检验）。

In the end, the t statistic depends only on the mean and variance of the two samples. The CLT says that (under most circumstances) those rapidly become normal even when the underlying population distribution is not. So the t-test is quite robust to (most) departures from normality. This has been verified by many simulation studies. Note, by the way, that it is not at all robust to departures from homogeneity of variance.

3 $\sigma_1^2\neq\sigma_2^2$ σ_1^2,σ_2^2 未知

$$t=\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}}\sim t(k)$$

$$k=\frac{\left(\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x-1}+\frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y-1}}$$

2. 两个总体方差比 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信区间

(μ_1, μ_2 为已知)

由 $\frac{\frac{S_1^2}{S_2^2}}{\frac{\sigma_1^2}{\sigma_2^2}} \sim F_{\frac{\alpha}{2}}(n_1-1, n_2-1)$

即 $$P\{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1) < \frac{\frac{S_1^2}{S_2^2}}{\frac{\sigma_1^2}{\sigma_2^2}} < F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\} = 1-\alpha$$

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

可得到 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信水平为 $1-\alpha$ 的置信区间为

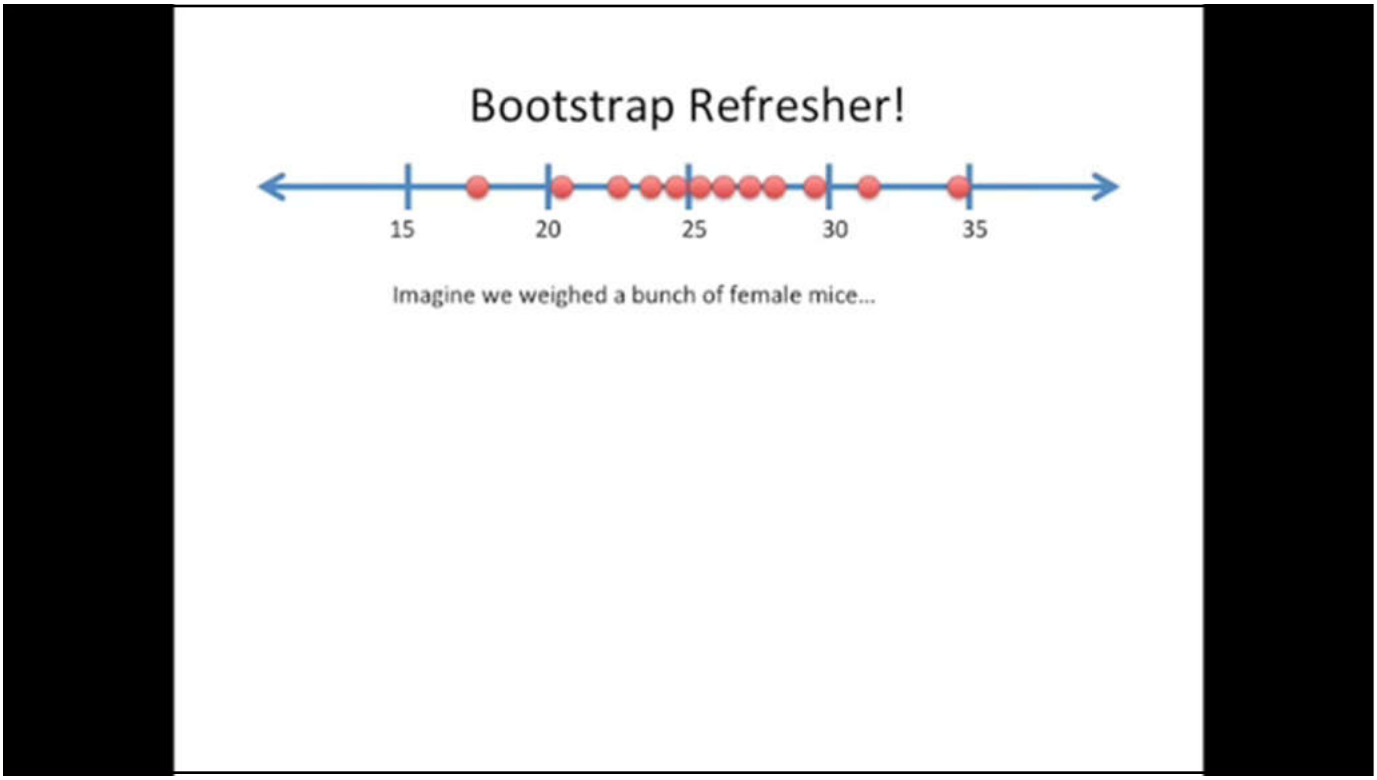
$$(\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{1-\alpha}{2}}(n_1-1,n_2-1)})$$

2. 置信区间（CI）的第二种求解方法：利用Bootstrapping

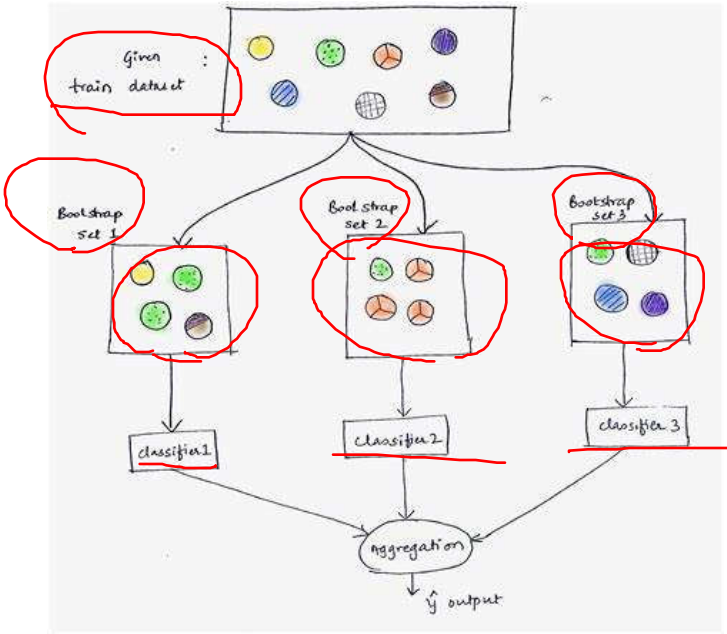
*不需要事先知道抽样分布

自助法

Bootstrapping是一种有放回的抽样



baggiry



Bootstrapping是一种有放回的抽样，在集成学习（ensemble learning）中也有广泛应用