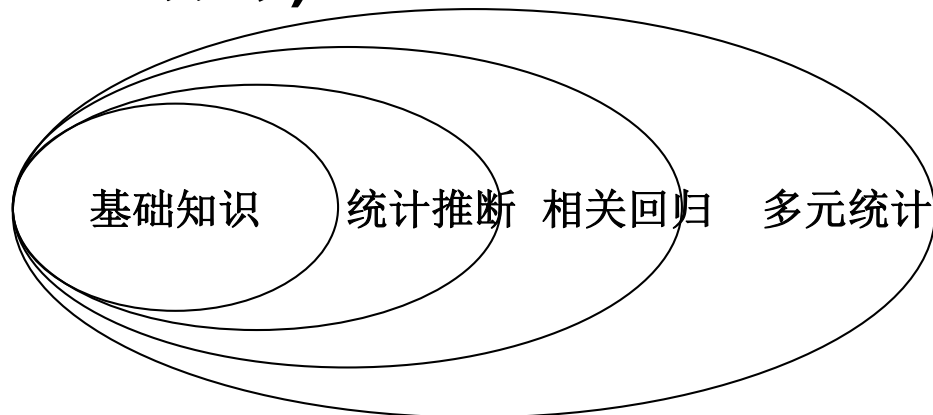


第五章 统计推断：参数估计

基础知识——特征：描述统计 ----- 经验分布
其它基础：概率、理论分布、抽样分布

统计推断——特征：参数估计（点估计、区间估计）-- 描述统计
(估计：平均数、方差、偏度、峭度)
差异：假设检验（参数假设检验、非参数假设检验）
(检验：平均数、方差、偏度、峭度)
(检验：分布)

相关回归——关系：



参数估计： 点估计 --- 描述统计
区间估计

统计推断——特征：参数估计（点估计、区间估计）-- 描述统计
(估计：平均数、方差、偏度、峭度)

估计参数 估计总体参数

第五章 统计推断：参数估计

- ❖ 离散变量----概率函数-----概率计算----统计推断
 - 二项分布 参数估计
 - Poisson**分布 假设检验
- ❖ 连续变量----概率密度函数----累积函数--概率计算----统计推断
 - 正态分布 参数估计
 - 标准正态分布 假设检验
 - U** 分布
 - t** 分布
 - F** 分布
 - χ^2** 分布

第五章 统计推断：参数估计

统计推断：

参数估计（点估计、区间估计）

假设检验

教材 P65—70 (第四版)

5—1 点估计

一、矩估计

二、极大似然估计

三、估计量的评选标准

5—2 区间估计

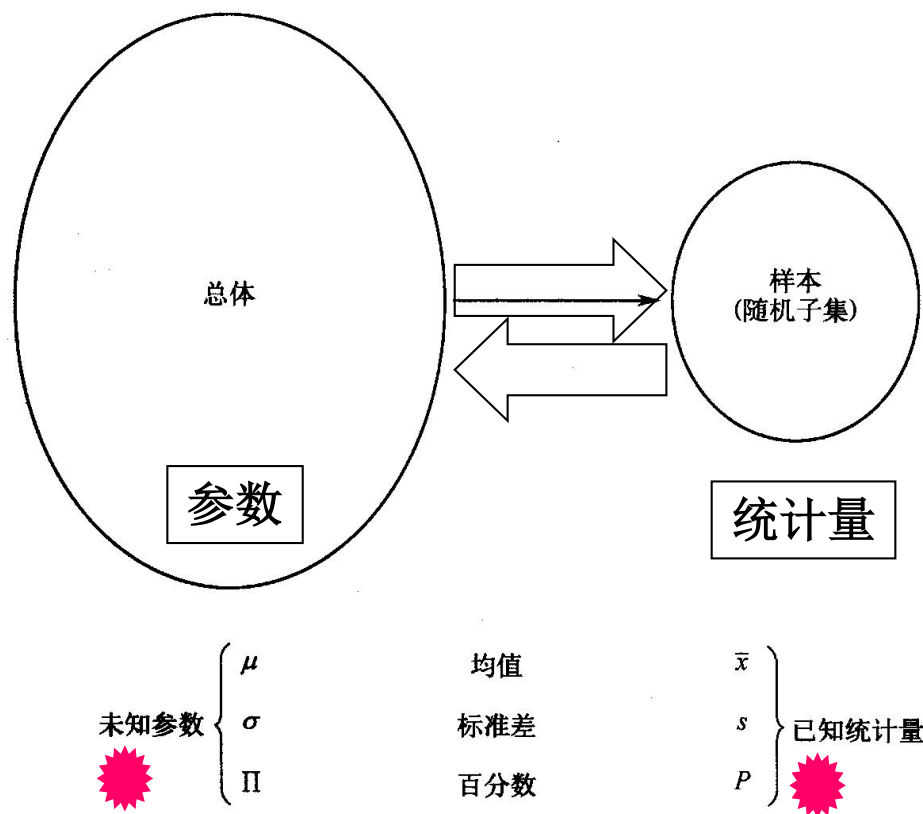


图 6.1 总体和样本,未知的总体参数和已知的样本统计量。

❖ 点估计: **point estimation**

明道绪教授: 将样本统计数直接作为相应总体参数的估计值叫点估计

❖ 区间估计: **interval estimation**

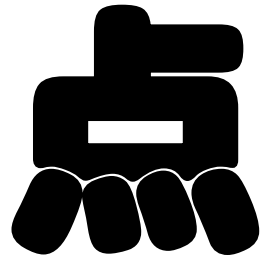
❖ 矩估计

矩是什么: 中心矩、原点矩

❖ 极大似然估计

5-1 点估计

教材中定义点估计 ($L = \bar{x} \pm u\sigma$) 不恰当



一、矩估计

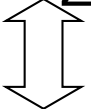
◆ 总体的参数（总体矩：平均数、方差、偏度、峭度）未知，用相应的样本矩估计总体矩的估计方法就是矩方法

◆ 无论总体分布类型是已知或未知，矩方法都是可行的

◆ 此内容已讲过（描述统计、理论分布）

总体中心矩

K 阶 $\mu_k = \sum p (x - \mu)^k = E[(X - \mu)^k]$



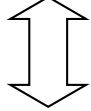
样本中心矩

K 阶 $m_k = \sum (x - \bar{x})^k / n$

二阶中心距是方差

$$S^2 = \sum (x_i - \bar{x})^2 / (n-1)$$

总体原点矩

$$\mu'_k = \sum p x^k = E(X^k)$$


样本原点矩

$$m'_k = \sum x^k / n$$

一阶原点距是平均数

$$\bar{x} = \sum x / n$$

理论分布

描述统计

一、矩估计

$$\overline{x}=(\sum x_i) / n$$

$$S^2=\sum (x_i-\overline{x})^2 / (n-1)$$

例： 男女大学生特征的数量估计

抽样：大学生样本 ←— 大学生总体
 估计：大学生样本 →→ 大学生总体

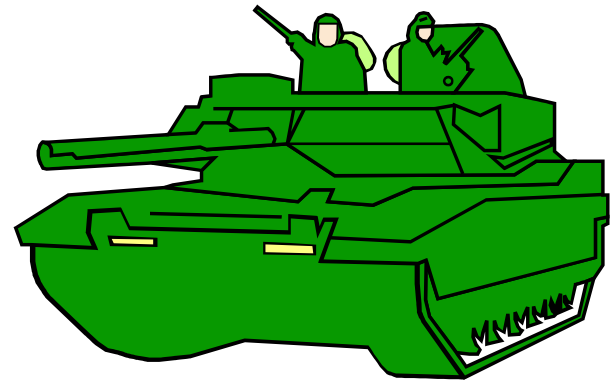
我知道你们。。。每月花多少钱！
 估计

Total		月生活费		月零花钱	
平均	458.421	平均	291.579	平均	166.842
标准误差	32.5007	标准误差	27.6704	标准误差	20.1319
中位数	400	中位数	240	中位数	200
众数	400	众数	200	众数	100
标准差	141.667	标准差	120.612	标准差	87.753
方差	20069.6	方差	14547.4	方差	7700.58
峰度	-0.3116	峰度	-0.5373	峰度	1.0923
偏度	0.86857	偏度	1.0945	偏度	0.89476
区域	450	区域	320	区域	350
最小值	300	最小值	180	最小值	50
最大值	750	最大值	500	最大值	400
求和	8710	求和	5540	求和	3170
观测数	19	观测数	19	观测数	19
Total		月生活费		月零花钱	
平均	361.111	平均	244.444	平均	116.667
标准误差	16.4474	标准误差	16.1128	标准误差	10.6948
中位数	400	中位数	200	中位数	100
众数	400	众数	200	众数	100
标准差	69.7802	标准差	68.3608	标准差	45.3743
方差	4869.28	方差	4673.2	方差	2058.82
峰度	-0.6873	峰度	-0.2506	峰度	0.09541
偏度	-0.0066	偏度	0.84458	偏度	0.82646
区域	250	区域	250	区域	150
最小值	250	最小值	150	最小值	50
最大值	500	最大值	400	最大值	200
求和	6500	求和	4400	求和	2100
观测数	18	观测数	18	观测数	18

	身高	体重	年龄	容貌	吃饭	花钱	生活费	零花钱
男生	172.4	63.9	20.9			458	292	167
女生	160.8	49.1	20.7			361	244	117

一、矩估计

$$\bar{x} = (\sum x_i) / n$$



例：二战期间德军坦克数量的估计

总体：先求总体平均编号值 $=\mu$

$$\mu = (\sum X) / N, \quad X=1, 2, \dots, N$$

再求坦克数量 $N = 2\mu - 1$

$$N = 10, \quad \mu = 5.5, \quad 2\mu - 1 = 11 - 1 = 10 = N$$

$$N = 100, \quad \mu = 50.5, \quad 2\mu - 1 = 101 - 1 = 100 = N$$

$$N = 1000, \quad \mu = 500.5, \quad 2\mu - 1 = 1001 - 1 = 1000 = N$$

样本：样本平均编号值： $\bar{x} = (\sum x_i) / n$

估计坦克数量： $2\bar{x} - 1$

《生活中的概率趣事》，P184

123 和 150。那么如何求 N 值呢？有很多种方法，虽然没有放之四海皆准的答案，但目测 N 显然远远大于 150。这个可以通过从 1, 2, ..., N 中随机抽样三个观测

数据所得出的最大期望值是 $0.75 \times N$ （观测数据是均匀分布的），鉴于 $0.75 \times 200 = 150$ ，得出 $N = 200$ 。注意这种估值是如何基于概率计算得出，“高阶方法”由此

全面展开。当然，后续还有很多方法用于改善估值的过程，不过我们最好还是就此结束这个故事，不再探讨技术性的细节。

战争结束后，统计学家们才得到了答案。但真实的答案的情况并不为人所知，这在统计学界是很少见。事实证明这些戴着厚眼镜片的统计学家们表现出色，远胜于英美的情报部门。当时的理查德·拉格尔斯（Richard Ruggles）和亨利·布罗迪（Henry Brodie）1947 年在《美国统计协会杂志》上发表了《第二次世界大战时经济的实证研究》，你可以在里面找到更多关于第二次世界大战统计应用的例子。表 8-1 正是三个不同月份的产量数据。其中，统计学家和情报人员的估值对比都来源于 Speer 部门的官方数字。可以看出，情报估值被严重夸大。

表 8-1 第二次世界大战期间德国坦克月产量的预估值和实际值

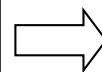
时 间	统 计 估 值	情 报 估 值	实 际 值
1940. 6	169	1000	122
1941. 6	244	1550	271
1942. 8	327	1550	342

德军坦克数 表8-1

Likelihood: 可能性

若有多个事件，往往总是具有最大概率(可能性)的事件发生；
反之，若某个事件发生了，你就可以认定它就具有最大概率

----- 你怎样让它具有最大概率？



求极大值

二、极大似然估计 (maximum likelihood estimation)

当总体分布类型已知，则估计其参数最好用极大似然法。

原理： 离散型变量

已知 概率函数 为 $f(x;\theta)$ ， θ 为未知参数

样本的一组已知的观察值记为 x_1 、 x_2 、 x_3 、...、 x_n

其中某个特定的 x_i 的概率就是 $f(x_i;\theta)$ ，

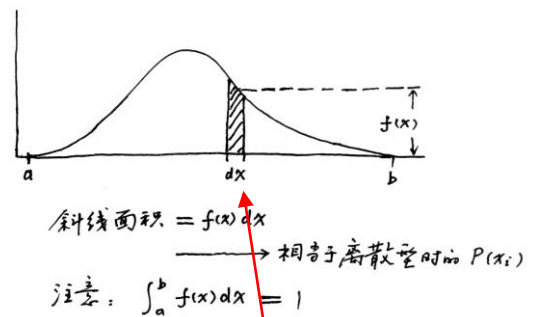
而所有的 x_i 同时发生的概率就是 $\prod f(x_i;\theta)$

独立事件同时发生的概率 = 各独立事件概率的乘积

即： $f(x_i;\theta) \rightarrow \prod f(x_i;\theta) = L(\theta)$

对其求极大值：先将 $\ln L(\theta)$ 对 θ 求导数，

后令其等于 0，从而得到 θ 的估计。



二、极大似然估计 (maximum likelihood estimation)

原理：连续型变量

已知概率密度函数为 $f(x;\theta)$ ， θ 为未知参数

样本的一组已知的观察值记为 $x_1, x_2, x_3, \dots, x_n$

其中某个特定的 x_i 的概率就是 $f(x_i;\theta) \Delta x_i$ ，

而所有的 x_i 同时发生的概率就是 $\prod f(x_i;\theta) \Delta x_i$

Δx_i 是不依赖于 θ 的增量

即: $f(x_i;\theta) \rightarrow \prod f(x_i;\theta) = L(\theta)$

对其求极大值：先将 $\ln L(\theta)$ 对 θ 求导数，

后令其等于 0，从而得到 θ 的估计。

RA Fisher (1890—1962)：现代统计学的另一位奠基人

22岁：极大似然估计

25岁：Pearson 相关系数的精确分布

主要贡献：利用随机化做实验
提出方差分析及 F 检验

建立：群体遗传学
生统遗传学



二、极大似然估计(maximum likelihood estimation)

例：Poisson分布：一个待估参数

$$p(x) = f(x; \lambda) = e^{-\lambda} \lambda^x / x !$$

$$f(x_i; \theta) = f(x_i; \lambda) = e^{-\lambda} \lambda^{x_i} / x_i !$$

$$L(\lambda) = \prod f(x_i; \lambda) = \prod e^{-\lambda} \lambda^{x_i} / x_i !$$

$$\begin{aligned} \ln L(\lambda) &= \ln \prod f(x_i; \lambda) = \ln \prod e^{-\lambda} \lambda^{x_i} / x_i ! \\ &= \sum \ln (e^{-\lambda} \lambda^{x_i} / x_i !) \\ &= \sum (\ln e^{-\lambda} + \ln \lambda^{x_i} - \ln x_i !) \\ &= \sum (-\lambda + x_i \ln \lambda - \ln x_i !) \\ &= -\sum \lambda + \sum x_i \ln \lambda - \sum \ln x_i ! \\ &= -n \lambda + \ln \lambda \sum x_i - \sum \ln x_i ! \end{aligned}$$

$$d \ln L(\lambda) / d\lambda = -n + (1/\lambda)(\sum x_i) = 0$$

$$\lambda = (\sum x_i) / n$$

例：指数分布：一个待估参数

例 2.1.5 设总体服从参数为 λ 的指数分布 $E(\lambda)$, 即 $X \sim p(x; \lambda) = \lambda e^{-\lambda x} (x > 0)$; (X_1, \dots, X_n) 是来自总体的样本, (x_1, \dots, x_n) 是相应的样本观测值, 求 λ 的最大似然估计.

解 因为 $X \sim p(x; \lambda) = \lambda e^{-\lambda x} (x > 0)$, 所以, 样本 (X_1, \dots, X_n) 的联合概率密度, 即 λ 的似然函数为

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i},$$

于是对数似然函数为

$$l(\lambda) = \sum_{i=1}^n \ln \lambda - \sum_{i=1}^n \lambda x_i = n \ln \lambda - \sum_{i=1}^n \lambda x_i,$$

因此

$$\frac{\partial l(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i,$$

令上式等于 0, 立即可以推出 $\hat{\lambda} = n / \sum_{i=1}^n x_i = 1/\bar{x}$.

例：基因连锁分析 ---- 采用多项式分布

的观测个体数为 $n = n_1 + n_2 + n_3 + n_4$ 。

表 10-2 回交群体和 DH 群体中的期望基因型频率

BC ₁	BC ₂	DH 群体	观测次数	理论频率
M ₁ M ₁ M ₂ M ₂	M ₁ m ₁ M ₂ m ₂	M ₁ M ₁ M ₂ M ₂	n_1	$f_1 = \frac{1}{2}(1 - r)$
M ₁ M ₁ M ₂ m ₂	M ₁ m ₁ m ₂ m ₂	M ₁ M ₁ m ₂ m ₂	n_2	$f_2 = \frac{1}{2}r$
M ₁ m ₁ M ₂ M ₂	m ₁ m ₁ M ₂ m ₂	m ₁ m ₁ M ₂ M ₂	n_3	$f_3 = \frac{1}{2}r$
M ₁ m ₁ M ₂ m ₂	m ₁ m ₁ m ₂ m ₂	m ₁ m ₁ m ₂ m ₂	n_4	$f_4 = \frac{1}{2}(1 - r)$

多个独立事件同时发生的概率等于各个独立事件的概率的乘积

然后,采用极大似然法求得重组率的估计公式。以 BC_1 世代的极大似然估计为例,求估计公式的方法和估计值的步骤为:

①建立似然函数。表 10-2 中的观测次数 n_1 、 n_2 、 n_3 、 n_4 服从频率为 f_1 、 f_2 、 f_3 、 f_4 的多项分布,因此概率分布函数或似然函数为

$$L = \frac{n!}{n_1! n_2! n_3! n_4!} \left[\frac{1}{2}(1-r) \right]^{n_1} \left[\frac{1}{2}r \right]^{n_2} \left[\frac{1}{2}r \right]^{n_3} \left[\frac{1}{2}(1-r) \right]^{n_4} \quad [10-1]$$

$$= C(1-r)^{n_1+n_4} (r)^{n_2+n_3}$$

其中 C 为不依赖于重组率 r 的常数。

②建立对数似然函数。对似然函数求对数,有

$$\ln L = \ln C + (n_1 + n_4) \ln(1-r) + (n_2 + n_3) \ln r$$

③求解重组率 r 的极大似然估计。对对数似然函数求导数,并令导数为 0,得

$$\frac{d \ln L}{dr} = -\frac{n_1 + n_4}{1-r} + \frac{n_2 + n_3}{r} = 0$$

上式称为似然方程。解此方程,可以得到重组率的极大似然估计为

$$\hat{r} = \frac{n_2 + n_3}{n_1 + n_2 + n_3 + n_4} = \frac{n_2 + n_3}{n} \quad [10-2]$$

④求信息量 I 。信息量等于对数似然函数二阶导数期望值的负数,信息量的倒数可作为估计量方差的估计。这里

$$I = -E \left(\frac{d^2 \ln L}{dr^2} \right) = -E \left[-\frac{n_1 + n_4}{(1-r)^2} - \frac{n_2 + n_3}{r^2} \right] = \frac{n}{r(1-r)}$$

因此 \hat{r} 的方差的估计值为

$$V_{\hat{r}} = \frac{1}{I} = \frac{\hat{r}(1-\hat{r})}{n} \quad [10-3]$$

⑤应用估计公式求得重组率的估计值和估计值的方差。某回交试验中 P_1 和 P_2 的基因型分别为 AABB 和 aabb,回交 BC_1 世代中 4 种基因型 AABB, AABb, AaBB 和 AaBb 的植株数依次为 162, 40, 41 和 158, 因此重组率的估计值为

$$\hat{r} = \frac{40 + 41}{162 + 40 + 41 + 158} = \frac{81}{401} = 20.20\%$$

估计值的方差为

$$V_{\hat{r}} = \frac{\hat{r}(1-\hat{r})}{n} \approx 4.02 \times 10^{-4}$$

例： 正态分布： 两个待估参数 $N(\mu; \sigma^2)$

例 2.1.6 设总体服从 X 正态分布 $N(\mu, \sigma^2)$, 参数为 $\theta = (\mu, \sigma^2)$, (X_1, \dots, X_n) 是来自总体的样本, (x_1, \dots, x_n) 是相应的样本观测值, 求 (μ, σ^2) 的最大似然估计.

解 因为 X 的概率密度为

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty,$$

所以 (μ, σ^2) 的似然函数为

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

对数似然函数为

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

将 $l(\mu, \sigma^2)$ 分别关于 μ 和 σ^2 求偏导, 并令其为 0, 就得到似然方程

$$\begin{cases} \frac{\partial l(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0, \\ \frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^2} = 0, \end{cases}$$

解此方程可得 μ 和 σ^2 的最大似然估计为

$$\begin{cases} \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = S_n^2. \end{cases}$$

对比例 2.1.1 和例 2.1.4, 例 2.1.2 和例 2.1.5 以及例 2.1.3 和例 2.1.6, 可以看出, 矩估计和最大似然估计可能相同, 也可能不同.

三、估计量的评选标准

无偏性	有效性	一致性
-----	-----	-----

1、无偏性

$$\lim_{n \rightarrow \infty} E(\bar{\theta}) = \theta$$

《中国医学百科全书》 P80

参见P41 表3-6：上次课已讲

无偏估计：在统计上，如果所有可能样本的某一统计量的平均数等于总体的相应参数，则称该统计量为总体相应参数的无偏估计值。

$$E(\bar{x}) = E[\sum x_i / n] = \mu$$

$$E(S^2) = E[\sum (x_i - \bar{x})^2 / (n-1)] = \sigma^2$$

$$E[\sum (x_i - \bar{x})^2 / n] = (1 - 1/n)\sigma^2$$

有偏估计

$$E(S) \neq \sigma,$$

自由度 ---- 那是必需的

上一次课已经讲过：

无偏估计：例 P41

表3—6 $N=3, n=2$ 时所有样本的平均数、方差、标准差

总体： $\{3, 4, 5\}$

$$N=3$$

$$\mu = 4$$

$$\sigma^2 = \sum (x_i - \mu)^2 / N = 0.6667$$

$$\sigma = 0.8165$$

样本： $n=2$

$$\bar{x} = \sum x_i / n$$

$$S^2 = \sum (x_i - \bar{x})^2 / (n-1)$$

所有样本 $N^2=9$

$\{3, 3\}$	$\{3, 4\}$	$\{3, 5\}$
$\{4, 3\}$	$\{4, 4\}$	$\{4, 5\}$
$\{5, 3\}$	$\{5, 4\}$	$\{5, 5\}$

比较结果：

样本平均数 \bar{x} 的平均数： $\mu(\bar{x}) = 36/9 = 4 = \mu$

样本方差 S^2 的平均数： $\mu(S^2) = 6/9 = 0.6667 = \sigma^2$

样本标准差 S 的平均数： $\mu(S) = 5.6568/9 = 0.6285 \neq 0.8165 = \sigma$

所有可能样本的某一统计量的平均数

2、有效性

哪个更有效：方差小的更有效

设 $\bar{\theta}_1$ 和 $\bar{\theta}_2$ 为 θ 的两种无偏估计量，若

$$V(\bar{\theta}_1) < V(\bar{\theta}_2)$$

则称 $\bar{\theta}_1$ 比 $\bar{\theta}_2$ 有效。

例如：样本平均数 \bar{x} 的方差 $V(\bar{x}) = \sigma^2/n$

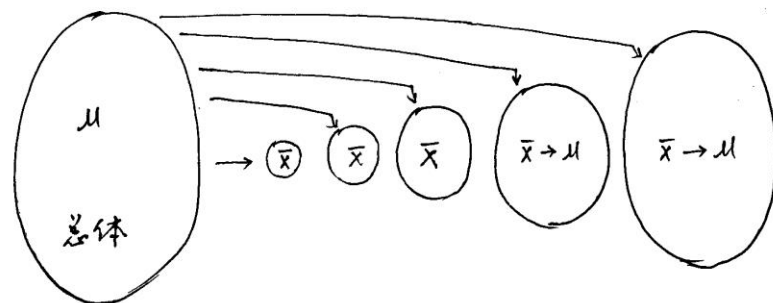
$$\begin{aligned} \text{样本中位数 } m \text{ 的方差 } V(m) &= (\pi/2)\sigma^2/n \\ &= 1.57 \times \sigma^2/n \\ &= 1.57 \times V(\bar{x}) \end{aligned}$$

3、一致性

《百科全书》 P80

$$\lim_{n \rightarrow \infty} E[(\bar{\theta} - \theta)^2] = 0$$

$$\lim_{n \rightarrow \infty} E[(\bar{x} - \mu)^2] = 0$$



设 $\bar{\theta}_n$ 为未知参数 θ 的一个估计量， n 为样本容量，若对任意一个 $\varepsilon > 0$ ，有

$$\lim_{n \rightarrow \infty} P\{|\bar{\theta}_n - \theta| < \varepsilon\} = 1$$

则称 $\bar{\theta}_n$ 为 θ 的一致估计量。

例如：由大数定律知

$$\lim_{n \rightarrow \infty} P\{|\bar{x} - \mu| < \varepsilon\} = 1$$

即，样本平均数 \bar{x} 是总体平均数 μ 的一致估计量。

一致指向何处？
一直指向何处？

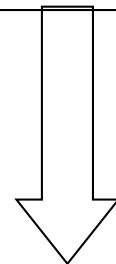
区间

5-2 区间估计

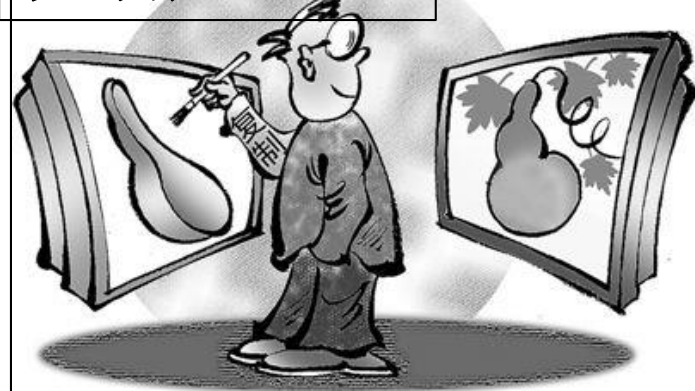
例：男生身高约 **1.7** 米左右
每月花费约 **400** 元左右
左右就是一个区间，但它不同于下述的区间估计

- 一、单个总体平均数的区间估计
- 二、两个总体平均差数的区间估计
- 三、单个二项总体频率的区间估计
- 四、两个二项总体频率差数的区间估计
- 五、正态总体方差的区间估计
- 六、估计置信区间所需的样本量

以一推三、举一反三



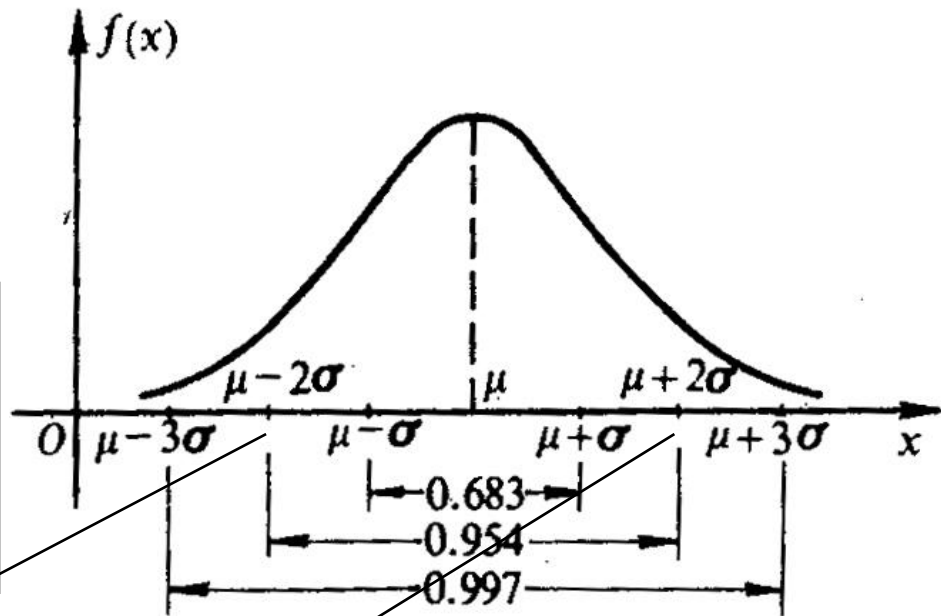
依葫芦画瓢



差数的区间估计，用的不多
差数的假设检验，用的很多

一、单个总体平均数的区间估计

1、 σ^2 已知



前面讲分布有啥用：由分布计算概率

理论分布：

抽样分布：统计量的分布

$$x \sim N(\mu; \sigma^2)$$

教材 P39 (4) :

$$P(\mu - 1.96\sigma < x < \mu + 1.96\sigma) = 0.95$$

$$P(\mu - 2.58\sigma < x < \mu + 2.58\sigma) = 0.99$$

直接写出
抽样分布

$$\bar{x} \sim N(\mu; \sigma^2 / n)$$

$$P(\mu - 1.96\sigma / \sqrt{n} < \bar{x} < \mu + 1.96\sigma / \sqrt{n}) = 0.95$$

$$P(\mu - 2.58\sigma / \sqrt{n} < \bar{x} < \mu + 2.58\sigma / \sqrt{n}) = 0.99$$

$$x \sim N(\mu; \sigma^2) \quad \text{教材 P39 (4):}$$

$$P(\mu - 1.96\sigma < x < \mu + 1.96\sigma) = 0.95$$

$$\bar{x} \sim N(\mu; \sigma^2 / n)$$

$$P(\mu - 1.96\sigma / \sqrt{n} < \bar{x} < \mu + 1.96\sigma / \sqrt{n}) = 0.95$$

若不知总体参数 μ ，但却知样本平均数 \bar{x} ，可由样本 \bar{x} 估计总体 μ ：

$$\text{变 } P(-1.96\sigma / \sqrt{n} < \bar{x} - \mu < +1.96\sigma / \sqrt{n}) = 0.95$$

$$\text{变 } P(-\bar{x} - 1.96\sigma / \sqrt{n} < -\mu < -\bar{x} + 1.96\sigma / \sqrt{n}) = 0.95$$

$$\text{变 } P(\bar{x} + 1.96\sigma / \sqrt{n} > \mu > \bar{x} - 1.96\sigma / \sqrt{n}) = 0.95$$

$$\text{变 } P(\bar{x} - 1.96\sigma / \sqrt{n} < \mu < \bar{x} + 1.96\sigma / \sqrt{n}) = 0.95$$

万变不离其宗 ---- 只变其中

这是我们要的

$$\text{记: } L_{1,2} = \bar{x} \pm u_{\alpha} \sigma / \sqrt{n}$$

$$\text{当 } \alpha = 0.05 \text{ 时, } L_{1,2} = \bar{x} \pm 1.96 \sigma / \sqrt{n}$$

$$\text{当 } \alpha = 0.01 \text{ 时, } L_{1,2} = \bar{x} \pm 2.58 \sigma / \sqrt{n}$$

对区间的两种理解：落在、覆盖

$$P(\bar{x} - 1.96 \sigma / \sqrt{n} < \mu < \bar{x} + 1.96 \sigma / \sqrt{n}) = 0.95$$

$$P(\bar{x} - u_{\alpha} \sigma / \sqrt{n} < \mu < \bar{x} + u_{\alpha} \sigma / \sqrt{n}) = 1 - \alpha$$

置信度： $P = 1 - \alpha = 0.95$ 或 0.99

值得相信的程度
值得相信的区间

置信区间：

$(\bar{x} - u_{\alpha} \sigma / \sqrt{n}, \bar{x} + u_{\alpha} \sigma / \sqrt{n})$ 为 μ 的 $1 - \alpha$ 置信区间

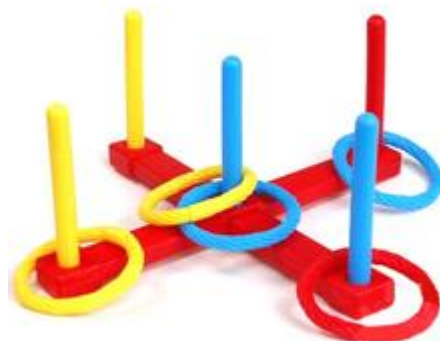
显著性水平： $\alpha = 0.05$ 或 0.01

临界值： u_{α} 为正态分布下置信度 $P = 1 - \alpha$ 时的 u 临界值



投掷炸弹: 区域 (区间) 不动, 炸弹 (点) 动
 飞镖游戏: 镖盘 (区间) 不动, 飞镖 (点) 动

飞圈套物: 物体 (点) 不动, 飞圈 (区间) 动
 降落伞: 落点 (点) 不动, 降落伞 (区间) 动



$$x \sim N(\mu; \sigma^2)$$

$$P(\mu - 1.96 \sigma < x < \mu + 1.96 \sigma) = 0.95 \quad \text{公式 1}$$

$$\bar{x} \sim N(\mu; \sigma^2 / n)$$

$$P(\mu - 1.96 \sigma / \sqrt{n} < \bar{x} < \mu + 1.96 \sigma / \sqrt{n}) = 0.95 \quad \text{公式 2}$$

$$P(\bar{x} - 1.96 \sigma / \sqrt{n} < \mu < \bar{x} + 1.96 \sigma / \sqrt{n}) = 0.95 \quad \text{公式 3}$$

解释含义：

投炸弹；掷飞镖

$(\mu - 1.96 \sigma, \mu + 1.96 \sigma)$ 区域不动，炸弹动；镖盘不动，飞镖动

如一个总体， μ 、 σ 已知，可算出一个区间 $(\mu - 1.96 \sigma, \mu + 1.96 \sigma)$ ，从总体中抽100个个体，有95个落在该区间内。

如一个总体， μ 、 σ 已知，由抽样分布也可算出一个区间 $(\mu - 1.96 \sigma / \sqrt{n}, \mu + 1.96 \sigma / \sqrt{n})$ ，从总体中抽100套样本，算出100个平均数，有95个落在该区间内。

降落伞；飞圈套物

(μ) 点不动，降落伞（区间）动；物体不动，飞圈动

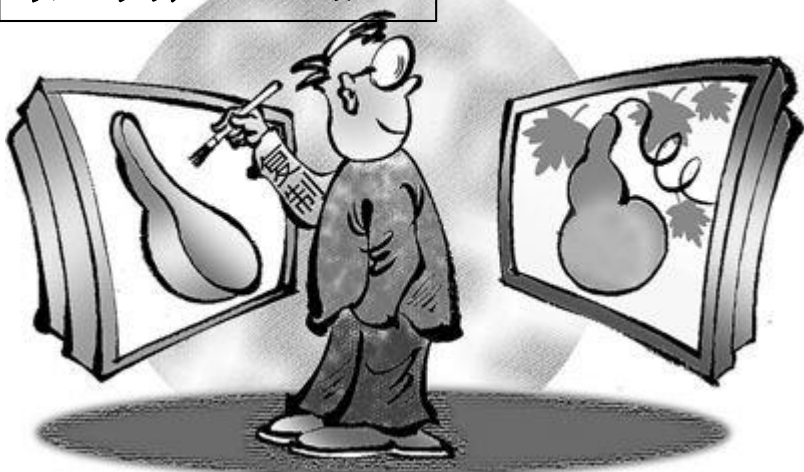
如100套样本，算出100个平均数构成100个区间，有95区间覆盖点 μ 。

$$P(\bar{x} - u_{\alpha}\sigma / \sqrt{n} < \mu < \bar{x} + u_{\alpha}\sigma / \sqrt{n}) = 1 - \alpha$$

$$L_{1,2} = \bar{x} \pm u_{\alpha}\sigma / \sqrt{n}$$

依葫芦画瓢写公式

依葫芦画瓢



一、单个总体平均数的区间估计

1、 σ^2 已知

2、 σ^2 未知

参考前者： $L_{1,2} = \bar{x} \pm u_{\alpha} \sigma / \sqrt{n}$
其中 u_{α} 一般取 1.96、2.58

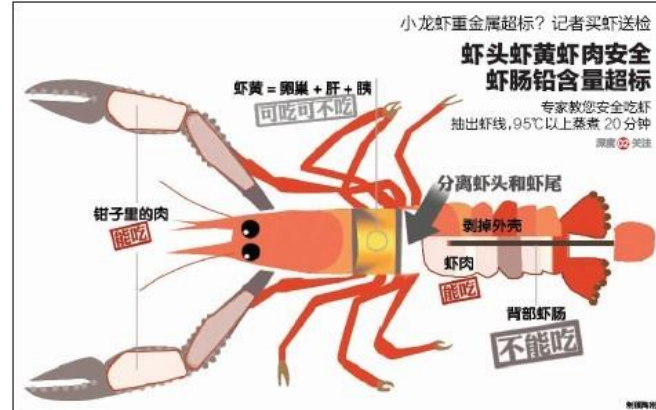
当 σ^2 未知且为小样本时，同样：

$L_{1,2} = \bar{x} \pm t_{\alpha} S / \sqrt{n}$
其中 t_{α} 随自由度 $df = n - 1$ 而变

例：教材P67 随机取20尾对虾，平均体长 $\bar{x} = 120\text{mm}$ 、标准差 $S = 15\text{mm}$ ，求置信度为99%的对虾总体的平均体长。

解：自由度 $df = 20 - 1 = 19$ 时， $t_{0.01(19)} = 2.861$ ，
标准误 $S / \sqrt{n} = 15 / \sqrt{20} = 3.354$

$$L_{1,2} = \bar{x} \pm t_{\alpha} S / \sqrt{n} = 120 \pm 2.861 \times 3.354 \\ = 120 \pm 9.6$$



一、单个总体平均数的区间估计：

$$L_{1,2} = \bar{x} \pm u_{\alpha} \sigma / \sqrt{n}$$

$$L_{1,2} = \bar{x} \pm t_{\alpha} S / \sqrt{n}$$



二、两个总体平均差数的区间估计：

简单扩展

1、 σ_i 已知，或者 σ_i 未知但为大样本时

$$L_{1,2} : (\bar{x}_1 - \bar{x}_2) \pm u_{\alpha} \sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}$$

2、 σ_i 未知且为小样本时，经F 测验，若 $\sigma_1 = \sigma_2$

$$L_{1,2} : (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha} \sqrt{(S^2 / n_1 + S^2 / n_2)},$$

$$df = n_1 + n_2 - 2$$

注意 $S^2 = !$ 教材P 58 公式4.9

3、 σ_i 未知且为小样本时，经F 测验，若 $\sigma_1 \neq \sigma_2$

$L_{1,2}$ ：同上

注意 $df = !$ 教材P 59 公式4.16

后者，用的不多

三、单个二项总体频率的区间估计 P61-62

正态总体: $\bar{x} \sim N(\mu; \sigma^2 / n)$

$$L_{1,2} = \bar{x} \pm u_{\alpha} \sigma / \sqrt{n}$$

二项总体: $\bar{x}_p \sim N(p, pq/n)$

$$L_{1,2} = p \pm u_{\alpha} \sqrt{(pq / n)}$$

四、两个二项总体频率差数的区间估计 P61-62

直接扩展: 明确二者有显著差异这才有意义

$$(\bar{x}_{p1} - \bar{x}_{p2}) \sim N \{ p_1 - p_2, p_1 q_1 / n_1 + p_2 q_2 / n_2 \}$$

$$L_{1,2} = (\bar{x}_{p1} - \bar{x}_{p2}) \pm u_{\alpha} \sqrt{(p_1 q_1 / n_1 + p_2 q_2 / n_2)}$$

$$= (\bar{x}_{p1} - \bar{x}_{p2}) \pm u_{\alpha} S_{p1-p2}$$

后者, 用的不多

→ 例

单个二项总体: $\bar{x}_p \sim N(p, pq/n)$

$$L_{1,2} = p \pm u_{\alpha} \sqrt{(pq/n)}$$

P69 例4.18:

调查100株玉米, 20株受玉米螟危害, 即 $p=0.2$, $np=20$, 求95%的置信区间。

$$\sqrt{(pq/n)} = \sqrt{(0.2*0.8/100)} = 0.04$$

$$\begin{aligned} L_{1,2} &= p \pm u_{\alpha} \sqrt{(pq/n)} \\ &= 0.2 \pm 1.96*0.04 \\ &= 0.2 \pm 0.0784 \end{aligned}$$

两个二项总体频率差数的区间估计

$$\begin{aligned} L_{1,2} &= (\bar{x}_{p1} - \bar{x}_{p2}) \pm u_{\alpha} \sqrt{(p_1 q_1 / n_1 + p_2 q_2 / n_2)} \\ &= (\bar{x}_{p1} - \bar{x}_{p2}) \pm u_{\alpha} S_{p1-p2} \end{aligned}$$

P70 例4.19:

调查两块麦田锈病发病率:

	调查株	锈病株	发病率 p
低洼地	378	342	0.905
高坡地	396	313	0.790

求99%的置信区间。

$$\text{P64: } S_{p1-p2} = [p^- q^- (1/n_1 + 1/n_2)]^{1/2} = 0.026$$

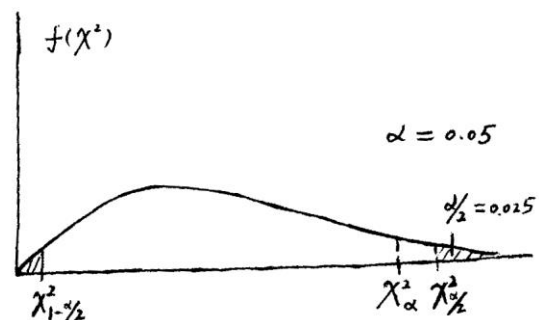
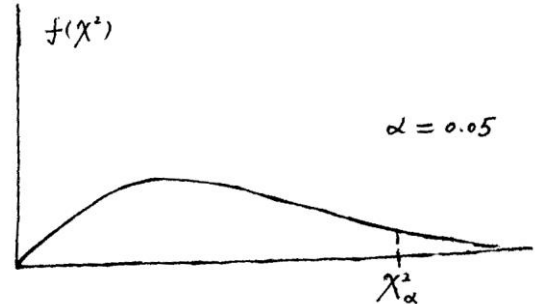
$$\begin{aligned} L_{1,2} &= (\bar{x}_{p1} - \bar{x}_{p2}) \pm u_{\alpha} S_{p1-p2} \\ &= (0.905 - 0.790) \pm 2.58 \times 0.026 \\ &= 0.115 \pm 0.0671 \end{aligned}$$

五、正态总体方差的区间估计

由：P71的4.67公式

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

可以推导出：



$$P [\chi^2_{(n-1), (1-\alpha/2)} < (n-1)S^2/\sigma^2 < \chi^2_{(n-1), (\alpha/2)}] = 1-\alpha$$

变 $P [1/\chi^2_{(n-1), (1-\alpha/2)} > \sigma^2 / (n-1)S^2 > 1/\chi^2_{(n-1), (\alpha/2)}] = 1-\alpha$

变

$$P [(n-1)S^2 / \chi^2_{(n-1), (1-\alpha/2)} > \sigma^2 > (n-1)S^2 / \chi^2_{(n-1), (\alpha/2)}] = 1-\alpha$$

变

$$P [(n-1)S^2 / \chi^2_{(n-1), (\alpha/2)} < \sigma^2 < (n-1)S^2 / \chi^2_{(n-1), (1-\alpha/2)}] = 1-\alpha$$

六、估计置信区间所需的样本量

- 1、正态总体平均数置信区间的样本量
- 2、二项总体概率 置信区间的样本量

香港学者 Nature 综述：探讨大规模基因组研究

2014年5月7日 来源：生物通

- ❖ 近来常有人说我们进入了大数据时代，确实，**2005年**人类（据不完全统计）创造了**150EB**的数据，而在**2010年**，达到了**1200EB**，在生命科学研究领域，随着新一代基因组测序技术的发展，近十年来大规模基因组测序研究越来越多，由此也积累出了庞大的数据群。
- ❖ 数据爆炸使科学的研究方法都落伍了，比如说要计算什刹海的鱼，原来的统计方式，是先测量每段水域内鱼的数量，然后根据样本估计，其实这种方法在现在来说已经不准确了。现在用一种极端的方式来说，就是抽干什刹海的水，然后一条一条数。统计学盛行不过百年，但是现在已经过时了，最好的统计方法就是穷举，这就是统计学的革命。
- ❖ 香港大学的两位学者发表了“**Statistical power and significance testing in large-scale genetic studies**”的综述，针对基因组研究中常用方法：显著性检验进行了深入探讨，回顾了显著性检验的基础原则和应用方法，以及就最近的罕见突变基因研究为例进行介绍。
- ❖ 显著性检验（**Significance Testing**）是为某个假说总结统计证据的一种客观方法，这种方法在基因研究中被广泛使用，如全基因组关联研究，还有最近的外显子组测序研究。
- ❖ 但是无论是在全基因组研究，还是外显子组范围内的研究中，显著性检验都需要严格意义的阈值，以便进行多方检测，并且这种方法也只适用于已有充足统计意义的研究，而这依赖于表型的特征，以及假定的遗传变异，还有研究设计方案。

1、设：c 为常数，x 和 y 为两个独立随机变量，求：

$$E(c) =$$

$$V(c) =$$

$$E(cx) =$$

$$V(cx) =$$

$$E(x \pm y) =$$

$$V(x \pm y) =$$

$$\text{另外：} \sum c =$$

$$\sum c x_i =$$

$$\sum (x_i \pm y_i) =$$

2、已知二项分布的总体平均数（次数）： $E(X) = \mu_x = n p$

二项分布的总体方差（次数）： $V(X) = \sigma_x^2 = n p q$

写出二项成数（百分数）分布的平均数： $E = \mu_p =$

二项成数（百分数）分布的方差： $V = \sigma_p^2 =$

3、已知 $x \sim N(\mu, \sigma^2)$ ，其平均数 $\bar{x} \sim N(,)$

4、已知 $x \sim N(\mu, \sigma^2)$ ，如何变换出一个新变量服从标准正态分布 $N(0, 1)$