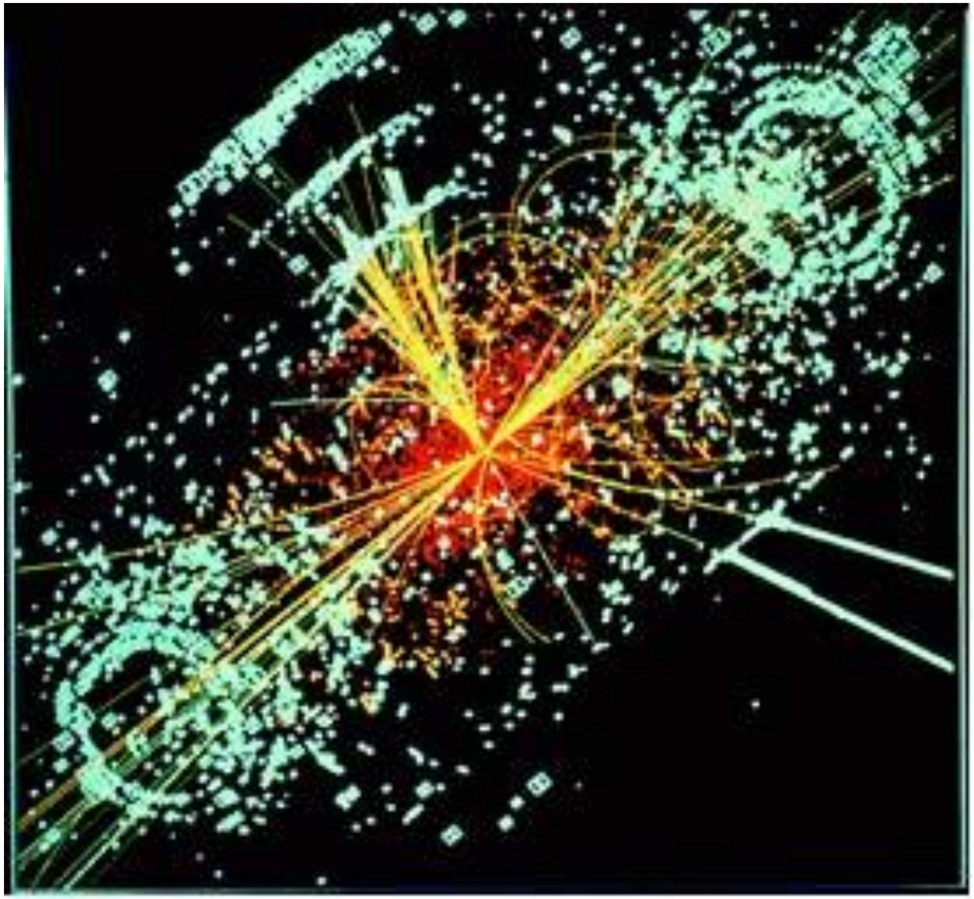


大数据的统计学基础——第9周

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

关注炼数成金企业微信

■提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

两个独立正态总体的均值比较

- 情况一：
- σ_1^2，σ_2^2已知，由于$\bar{X}$，$\bar{Y}$相互独立，且$\bar{X}\sim N(\mu_1,\sigma_1^2/n_1),\bar{Y}\sim N(\mu_2,\sigma_2^2/n_2)$，所以
- $\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)\to\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$
- 故$\mu_1-\mu_2$的1-α的置信区间为
- $(\bar{X}-\bar{Y}\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}})$

例子

某银行对所属两个储蓄所的储户存款情况进行调查，为此从每一家储蓄所抽取 25 个储户组成简单随机样本。甲储蓄所的储户平均存款余额为 7500 元，乙储蓄所的储户平均存款余额为 9000 元。假设两个总体均服从正态分布，标准差分别 $\sigma_1 = 700$ 为 元和 $\sigma_2 = 750$ 元试建立两储蓄所储户平均存款余额之差的 95%的置信区间。

解 假设用随机变量 X_1 ， X_2 分别表示甲、乙两个储蓄所的储户存款余额，则由已知条件有 $\overline{X}_1 = 7500$ 元， $\sigma_1 = 700$ 元， $\overline{X}_2 = 9000$ 元， $\sigma_2 = 750$ 元， $n_1 = n_2 = 25$ ，与置信度 95%相对应的 $\alpha = 0.05$ ，查标准正态分布表得 $Z_{0.05/2} = 1.96$ 。由公式，得 $\mu_1 - \mu_2$ 的置信度为 95%的置信区间为

$$
\begin{aligned}
&(\overline{X}_1 - \overline{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\
&\qquad = (7500 - 9000) \pm 1.96 \times \sqrt{\frac{700^2}{25} + \frac{750^2}{25}} \\
&\qquad = -1500 \pm 1.96 \times 205.18 = -1500 \pm 402.15 \doteq (-1902\ ,\ -1098)
\end{aligned}
$$

于是，我们有 95%的把握认为，甲储蓄所储户平均存款余额比乙储蓄所储户平均存款余额大约少 1098 到 1902 元之间。

两个独立正态总体的均值比较

- 情况二：
- $\sigma_1^2 = \sigma_2^2 = \sigma^2$，但$\sigma^2$的值未知
- 当$\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2$时，有

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

其中，$S_w{}^2 = \frac{(n_1-1)S_1{}^2+(n_2-1)S_2{}^2}{n_1+n_2-2}, S_w = \sqrt{S_w{}^2}$

- 所以$\mu_1 - \mu_2$的1-α的置信区间为
- $(\bar{X} - \bar{Y} \pm t_{\alpha/2, n_1+n_n-2} S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$

例子

- 某公司为了解男女推销员的推销能力是否有差别，随机抽取16名男推销员和25名女推销员进行测试。男推销员的平均销售额为30250元，标准差为18400元，女推销员的平均销售额为33750元，标准差为13500元。假设男女推销员的销售额服从正态分布，且方差相等。试建立男女推销员销售额之差的95%的置信区间。

解 假设用随机变量 X_1 ， X_2 分别表示男女推销员的销售额，则由已知条件有 $\bar{X}_1 = 30250$ 元， $S_1 = 18400$ 元， $\bar{X}_2 = 33750$ 元， $S_2 = 13500$ 元， $n_1 = 16$ ， $n_2 = 25$ 。又因两总体方差相等，可以估计出它们的共同方差：

$$S_{合}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(16-1)\times 18400^2 + (25-1)\times 13500^2}{16+25-2} \doteq 15568^2$$

与置信度 95%相对应的 $\alpha = 0.05$ ，查 t 分布表，得到 $t_{0.05/2,16+25-2} = 2.02$ ，由公式，得男女推销员销售额之差的置信度为 95%的置信区间为

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,n_1+n_2-2} S_{合}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= (30250 - 33750) \pm 2.02 \times 15568\sqrt{\frac{1}{16}+\frac{1}{25}}$$

$$= -3500 \pm 2.02 \times 4984.2 = -3500 \pm 10068 = (-13568\text{，}6568)$$

于是，我们有 95%的把握认为：男推销员的销售额既有可能比女推销员多 6568 元，也有可能比女推销员少 13568 元，所以男女推销员的推销能力没有显著差别。

两个独立正态总体的均值比较

- 情况三：
- σ_1^2，σ_2^2未知且，$\sigma_1^2 \neq \sigma_2^2$时，$\bar{X}$，$\bar{Y}$相互独立，且$\bar{X} \sim N(\mu_1, \sigma_1^2/n_1), \bar{Y} \sim N(\mu_2, \sigma_2^2/n_2)$
- 统计量$t = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$近似服从自由度为v的t分布。其中，$v = \frac{(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2})^2}{\frac{(S_1^2/n_1)^2}{n_1}+\frac{(S_2^2/n_2)^2}{n_2}}$
- 故$\mu_1 - \mu_2$的1-α的置信区间为
- $(\bar{X} - \bar{Y} \pm t_{\alpha/2,v}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}})$

例子

- 在上一道题中，假设两个正态分布方差不等，试建立男女推销员销售额之差的95%的置信区间。

首先根据公式 计算自由度 v，

$$v=\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1}+\frac{\left(S_2^2/n_2\right)^2}{n_2}}=\frac{\left(\frac{18400^2}{16}+\frac{13500^2}{25}\right)^2}{\frac{\left(18400^2/16\right)^2}{16}+\frac{\left(13500^2/25\right)^2}{25}}\doteq 27$$

查 t 分布表，得到 $t_{0.05/2,27}=2.05$ ，由公式，得男女推销员销售额之差的置信度为 95%的置信区间为

$$(\bar{X}_1-\bar{X}_2)\pm t_{\alpha/2,v}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}=(30250-33750)\pm 2.05\sqrt{\frac{18400^2}{16}+\frac{13500^2}{25}}$$
$$=-3500\pm 2.05\times 5334=-3500\pm 10934.4=(-14434,7434)$$

于是，我们有 95%的把握认为：男推销员的销售额既有可能比女推销员多 7434 元，也有可能比女推销员少 14434 元，所以男女推销员的推销能力没有显著差别。

两个非正态分布总体的均值比较

- 对于两个来自非正态总体的样本，只有样本量足够大（一般要求$n_1, n_2 \geq 30$)，就可以根据中心极限定理，有$\overline{X_1} - \overline{X_2}$服从正态分布$N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$
- 当两个总体的方差已知时：

$\mu_1 - \mu_2$的1-α置信区间为$((\overline{X_1} - \overline{X_2}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$

- 当两个总体的方差未知时，用样本方差代替总体方差：

$\mu_1 - \mu_2$的1-α置信区间为$((\overline{X_1} - \overline{X_2}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}})$

例子

- 某连锁店准备在两个不同地点选择一个地方开一家新店，为此需调查这两个地点居民收入的差别。在甲地点调查了100户居民，年平均收入为19000元，标准差为70元，在乙地点调查了80户居民，年平均收入为17000元，标准差为75元。试建立两个地点居民年平均收入差别的99%的置信区间。

解 假设用随机变量 X_1，X_2 分别表示甲、乙两地居民的年收入，虽然已知条件没有给出 X_1，X_2 服从正态分布的假设，但由于 $n_1 = 100$，$n_2 = 80$，均为大样本，所以可以应用公式(6.13)。根据题意，$\overline{X}_1 = 19000$ 元，$\overline{X}_2 = 17000$ 元，$S_1 = 70$ 元，$S_2 = 75$ 元，与置信度 99%相对应的 $\alpha = 0.01$，查标准正态分布表，得到 $Z_{0.01/2} = 2.58$。由公式(6.13)，得两总体均值之差 $\mu_1 - \mu_2$ 的置信度为 99%的置信区间为

$$(\overline{X}_1 - \overline{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$= (19000 - 17000) \pm 2.58 \times \sqrt{\frac{70^2}{100} + \frac{75^2}{80}}$$

$$= 2000 \pm 2.58 \times 10.92 = 2000 \pm 28.17 \doteq (1972,\ 2028)$$

于是，我们有 99%的把握认为，甲地居民年平均收入比乙地居民年平均收入大约高 1972 元到 2028 元之间。

两个总体比例比较

- 假设两个总体的比例分别为p_1和p_2，从中分别抽取容量为n_1和n_2的两个相互独立的简单随机样本，样本比例分别为$\hat{p}_1$和$\hat{p}_2$。由中心极限定理可知，在大样本条件下，两个样本比例之差$\hat{p}_1-\hat{p}_2$的抽样分布近似服从正态分布，其数学期望为p_1-p_2，方差为$\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}$
- 故$\hat{p}_1-\hat{p}_2 \sim N(p_1-p_2, \frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2})$
- 从而p_1-p_2的1-α置信区间为$(\hat{p}_1-\hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}})$
- 要求：1.n_1，$n_2 \geq 30$

  2.两个样本成功和失败次数都不少于5

  3. p_1, p_2不要太接近1或0

例子

- 对某个电视广告的收视率进行调查。在甲地区调查了200人，有128人收看过该广告，在乙地区调查了225人，有90人收看过该广告。试以90%的可靠性对该广告在两地收视率的差别作出区间估计。

假设用p_1, p_2分别表示某电视广告在甲、乙两地区的收视率，则由已知条件可知，$\hat{p}_1 = \frac{128}{200} = 0.64$，$\hat{p}_2 = \frac{90}{225} = 0.40$，$\alpha = 0.10$，$z_{0.10/2} = 1.645$，得到该广告在两地收视率之差的 90 的置信区间为

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= (0.64 - 0.40) \pm 1.645$$

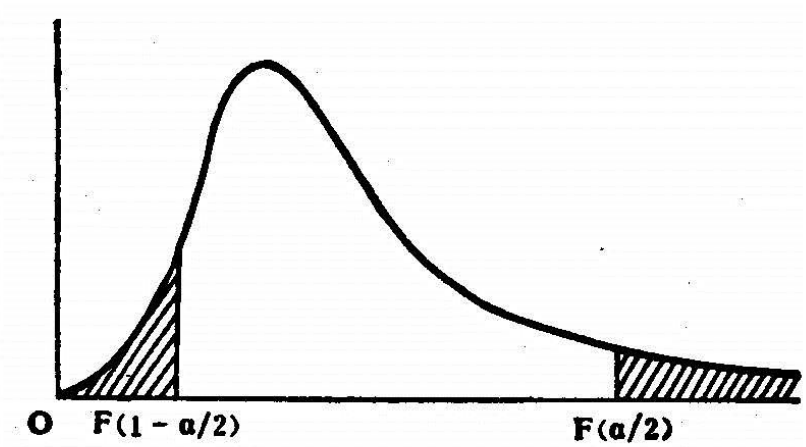
$$\times \sqrt{0.64 * \frac{1-0.64}{200} + 0.40 * \frac{1-0.40}{225}} = (0.1625, 0.3175)$$

两个独立正态总体的方差比较

- 设X1，X2，……Xn1与Y1,Y2,……，Yn2分别来自正态总体$N(\mu_1,{\sigma_1}^2)$，$N(\mu_2,{\sigma_2}^2)$的样本，且这两个样本相互独立。其样本均值分别为$\bar{X}$，$\bar{Y}$，其样本方差分别为${S_1}^2$, ${S_2}^2$,则有

- $\frac{{S_1}^2/{S_2}^2}{{\sigma_1}^2/{\sigma_2}^2} \sim F(n_1-1,n_2-1)$

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



例子

- 在便士铸造的例子中，假设从原来的机器随机抽取10个硬币，测量其重量，得到其标准差为0.0158g。试比较两个机器的铸币重量的标准差。
- $s_1^2 = 0.0158 * 0.0158 = 0.00024964; s_2^2 = 0.0125 * 0.0125 = 0.00015625$；
- 取$\alpha = 0.05$，$F_{1-\alpha/2,n_1-1,n_2-1} = \frac{1}{F_{\alpha/2,n_2-1,n_1-1}} = \frac{1}{4.03}$；$F_{\alpha/2,n_1-1,n_2-1} = 4.03$
- 故$\frac{{\sigma_1}^2}{{\sigma_2}^2}$的95%置信区间为
- $\left(\frac{{S_1}^2/{S_2}^2}{F_{\alpha/2,n_1-1,n_2-1}}, \frac{{S_1}^2/{S_2}^2}{F_{1-\alpha/2,n_1-1,n_2-1}}\right)$=(0.39645,6.43871)
- 所以$\frac{\sigma_1}{\sigma_2}$的95%置信区间为
- （0.6294,2.5375）

估计量与估计值

DATAGURU 炼数成金

- 对于已知类型的分布，估计分布函数参数是关键。
- 设总体X的分布函数F(x；θ)的形式为已知， θ是待估参数。X1，X2，……Xn是X的一个样本，x1,x2,……,xn是相应的一个样本值。点估计就是要构造一个适当的统计量$\hat{\theta}(X_1, X_2, \dots\dots, X_n)$,其相应的观察值$\hat{\theta}(x_1, x_2, \dots\dots, x_n)$作为未知参数θ的一个近似值。
- 我们称$\hat{\theta}(X_1, X_2, \dots\dots, X_n)$为θ的估计量， $\hat{\theta}(x_1, x_2, \dots\dots, x_n)$为θ的估计值。
- 估计量不唯一，使用不同方法得到的估计量可能不一样
- 最常用的的构造估计量的方法：矩估计法和最大似然估计法

无偏性

- 设 X_1 ，X_2 ，……，X_n 是总体X的一个样本，θ∈Θ是包含在总体X的分布中的待估函数，Θ表示θ的取值范围。
- 若θ的估计量$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots\ldots, X_n)$的数学期望$E(\hat{\theta})$存在，且对于任意θ∈Θ都有

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

- 则称$\hat{\theta}$是θ的无偏估计量。
- 样本均值是总体均值的无偏估计量：
- $E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$
- 样本方差是总体方差的无偏估计量：
- $E(S^2) = E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2\right)\right] = \frac{1}{n-1}\sum_{i=1}^{n} E({X_i}^2) - \frac{n}{n-1}E(\bar{X}^2) = \frac{1}{n-1}\sum_{i=1}^{n}\{D(X_i) +$

有效性

- 设$\widehat{\theta_1} = \widehat{\theta_1}(X_1, X_2, \ldots \ldots, X_n)$与$\widehat{\theta_2} = \widehat{\theta_2}(X_1, X_2, \ldots \ldots, X_n)$都是θ的无偏估计量，若对于任意θ ∈ Θ，有
- $D(\widehat{\theta_1}) \le D(\widehat{\theta_2})$
- 且至少对于某一个θ不等号成立，则称$\widehat{\theta_1}$比$\widehat{\theta_2}$有效

- 对于一个样本中的任一个数据X_i,由于其数学期望也是为总体均值μ，所以也是一个无偏估计量。证明样本均值$\bar{X}$比X_i有效。
- 证明：$D(X_i) = \sigma^2, D(\bar{X}) = \sigma^2/n$.当n>1时，$\frac{\sigma^2}{n} < \sigma^2$恒成立，所以样本均值$\bar{X}$比$X_i$有效。

相合性

- 设$\hat{\theta}(X_1, X_2, \ldots \ldots, X_n)$为参数θ的估计量，若对于任意θ ∈ Θ，当n → ∞时，$\hat{\theta}(X_1, X_2, \ldots \ldots, X_n)$依概率收敛于θ，则称$\hat{\theta}$为θ的相合估计量。
- 即，若对于任意θ ∈ Θ都有：对于任意ε<0,有$\lim_{n\to\infty} P\{|\hat{\theta}-\theta|< \varepsilon\} = 1$
- 相合性也称为一致性，是对一个估计量的基本要求。如果一个估计量不具有相合性，那么无论我们将样本容量取得多大，都不能将θ估计得足够准确。
- 可以证明，样本均值 $\bar{X}$ 是总体均值μ的一致估计量；样本比例 $\hat{p}$ 是总体比例 p 的一致估计量；样本方差 S^2 是总体方差 σ^2 的一致估计量；样本标准差 S 是总体标准差σ的一致估计量。

参数估计的一般方法

可选内容

矩估计法

- 设X是连续型随机变量，其概率密度为$f(x;\theta_1,\theta_2,\dots\dots,\theta_k)$，或X为离散型随机变量，其分布律为$P\{X=x\}=p(x;\theta_1,\theta_2,\dots\dots,\theta_k)$，其中$\theta_1,\theta_2,\dots\dots,\theta_k$为待估计参数，$X_1$，$X_2$，……，$X_n$是来自X的样本。假设总体X的前k阶矩为
- $\mu_l=E(X^l)=\int_{-\infty}^{\infty}x^l f(x;\theta_1,\theta_2,\dots\dots,\theta_k)dx$　(X为连续型)
- $\mu_l=E(X^l)=\sum_{x\epsilon R_X}x^l p(x;\theta_1,\theta_2,\dots\dots,\theta_k)$　(X为离散型）l=1,2,……,k
- 样本矩：$A_l=\frac{1}{n}\sum_{i=1}^{n}{X_i}^l$——估计总体矩$\mu_l$

矩估计法

- 联立方程 $\begin{cases} \mu_1 = \mu_1(\theta_1, \theta_2, \ldots\ldots, \theta_k) \\ \mu_2 = \mu_2(\theta_1, \theta_2, \ldots\ldots, \theta_k) \\ \ldots\ldots \\ \mu_k = \mu_k(\theta_1, \theta_2, \ldots\ldots, \theta_k) \end{cases}$
- 得 $\begin{cases} \theta_1 = \theta_1(\mu_1, \mu_2, \ldots\ldots, \mu_k) \\ \theta_2 = \theta_2(\mu_1, \mu_2, \ldots\ldots, \mu_k) \\ \ldots\ldots \\ \theta_k = \theta_k(\mu_1, \mu_2, \ldots\ldots, \mu_k) \end{cases}$
- 将$A_l = \frac{1}{n}\sum_{i=1}^{n} {X_i}^l$代替上式中的$\mu_l$，就可以得到各个参数的估计量：
- $\widehat{\theta_i} = \theta_i(A_1, A_2, \ldots\ldots, A_k)$，$i = 1,2,\ldots\ldots,k$

均匀分布的矩估计量

例 2　设总体 X 在 $[a,b]$ 上服从均匀分布，a,b 未知．$X_1,X_2,\cdots,X_n$ 是来自 X 的样本，试求 a,b 的矩估计量．

解
$$\mu_1 = E(X) = (a+b)/2,$$
$$\mu_2 = E(X^2) = D(X) + [E(X)]^2 = (b-a)^2/12 + (a+b)^2/4.$$

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

解这一方程组得
$$a = \mu_1 - \sqrt{3(\mu_2-\mu_1^2)}, \quad b = \mu_1 + \sqrt{3(\mu_2-\mu_1^2)}.$$

分别以 A_1,A_2 代替 μ_1,μ_2，得到 a,b 的矩估计量分别为（注意到 $\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$）：

$$\hat{a} = A_1 - \sqrt{3(A_2-A_1^2)} = \overline{X} - \sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2},$$
$$\hat{b} = A_1 + \sqrt{3(A_2-A_1^2)} = \overline{X} + \sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}. \qquad \square$$

正态分布的矩估计量

例 3 设总体 X 的均值 μ 及方差 σ^2 都存在，且有 $\sigma^2>0$. 但 μ,σ^2 均为未知. 又设 $X_1,X_2,\cdots,X_n$ 是来自 X 的样本. 试求 μ,σ^2 的矩估计量.

解
$$\begin{cases}\mu_1=E(X)=\mu,\\ \mu_2=E(X^2)=D(X)+[E(X)]^2=\sigma^2+\mu^2.\end{cases}$$

解得
$$\begin{cases}\mu=\mu_1,\\ \sigma^2=\mu_2-\mu_1^2.\end{cases}$$

分别以 A_1,A_2 代替 μ_1,μ_2，得 μ 和 σ^2 的矩估计量分别为

$$\hat{\mu}=A_1=\overline{X},$$

$$\hat{\sigma}^2=A_2-A_1^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

所得结果表明，总体均值与方差的矩估计量的表达式不因不同的总体分布而异.

例如，$X\sim N(\mu,\sigma^2)$，μ,σ^2 未知，即得 μ,σ^2 的矩估计量为

$$\hat{\mu}=\overline{X},\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

极大似然估计法

- 有1个射手射击3次，命中0次。试问该射手的命中概率最有可能为3个命中概率：1/5、8/15和4/5中的哪一个？回答该问题可以从两方面来看，一方面，该射手的命中率为0，与此最接近的命中概率为1/5，即1/5最有可能；另一方面，分别假定该射手的命中率为1/5、8/15和4/5，根据二项分布原理分别计算出该射手射击3次命中0次的概率分别为：

$$C_3^0\left(\frac{1}{5}\right)^0\left(1-\frac{1}{5}\right)^3=\frac{1728}{3375},\quad C_3^0\left(\frac{8}{15}\right)^0\left(1-\frac{8}{15}\right)^3=\frac{343}{3375},\quad C_3^0\left(\frac{4}{5}\right)^0\left(1-\frac{4}{5}\right)^3=\frac{27}{3375}$$

- 因此，选择使事件发生概率最大的可能命中概率为1/5，从而认为该射手的命中概率最有可能为1/5。这种参数估计方法称为极大似然法。

似然函数

DATAGURU 炼数成金

- 若总体X是离散型，其分布律P{X=x}=p(x, θ), θ ∈ Θ的形式为已知， θ为待估参数，Θ是θ可能取值的范围。设X_1，X_2，……，X_n是来自X的样本，则X_1，X_2，……，X_n的联合分布律为$\prod_{i=1}^{n} p(x_i;\theta)$.
- 设x_1，x_2，……，x_n是相应于样本X_1，X_2，……，X_n的一个观测值。
- 事件$\{X_1 = x_1, X_2 = x_2 \ldots\ldots X_n = x_n\}$发生的概率是

$$L(\theta) = L(x_1, x_2, \ldots\ldots, x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta), \theta \epsilon \Theta$$

- $L(\theta)$是关于θ的一个函数，称为样本的似然函数。

似然函数

- 若总体X是连续型，其概率密度f(x; θ), θ ∈ Θ的形式已知， θ为待估参数， Θ是θ可能取值德范围。设X_1，X_2，……，X_n是来自X的样本，则X_1，X_2，……，X_n的联合密度分布函数为$\prod_{i=1}^{n} f(x_i, \theta)$
- 设x_1，x_2，……，x_n是相应于样本X_1，X_2，……，X_n的一个观测值。
- 则随机点(X_1，X_2，……，X_n)落在点(x_1，x_2，……，x_n)的邻域内的概率近似为

$$L(\theta) = L(x_1, x_2, \ldots\ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

- $L(\theta)$是关于θ的一个函数，称为样本的似然函数

极大似然估计量

- 在θ的取值范围里，取使得L(θ)达到最大值的$\hat{\theta}$作为θ的极大似然估计量。
- 求解方法：
- 1. 对极大似然函数取对数：$\ln L(\theta) = \ln L(x_1, x_2, \cdots, x_n; \theta) = \sum_{i=1}^{n} \ln f(x_i, \theta)$
- 2. 对待测参数求导并令导数为0：

$$\frac{\partial}{\partial \theta_k} \ln L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2, \cdots, \theta_n) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_k} f(x_i; \theta_1, \theta_2, \cdots, \theta n) = 0, \mathrm{k} = 1, 2 \ldots\ldots, \mathrm{n}$$

- 3. 求解上述方程组，得到θ的极大似然估计量$\hat{\theta}$

二项分布的极大似然估计量

例 4 设 $X \sim b(1,p)$. $X_1, X_2, \cdots, X_n$ 是来自 X 的一个样本，试求参数 p 的最大似然估计量.

解 设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的一个样本值. X 的分布律为

$$P\{X=x\}=p^x(1-p)^{1-x}, \quad x=0,1.$$

故似然函数为

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

而

$$\ln L(p)=\left(\sum_{i=1}^{n} x_i\right)\ln p+\left(n-\sum_{i=1}^{n} x_i\right)\ln(1-p),$$

令

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p)=\frac{\sum_{i=1}^{n} x_i}{p}-\frac{n-\sum_{i=1}^{n} x_i}{1-p}=0,$$

解得 p 的最大似然估计值

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

p 的最大似然估计量为

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n} X_i=\bar{X}.$$

我们看到这一估计量与矩估计量是相同的.

正态分布的极大似然估计量

设 y_1， $y_2,\cdots$， y_n 是正态总体 $N(\mu，\sigma^2)$ 的随机样本，求正态分布 $N(\mu，\sigma^2)$ 参数的极大似然估计量。

似然函数为：

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

取对数，得：

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

那么似然方程组为：

$$\begin{cases}\dfrac{\partial}{\partial\mu}\ln L(\mu，\sigma^2)=\dfrac{1}{\sigma^2}\sum\limits_{1}^{n}(y_i-\mu)=0\\ \dfrac{\partial}{\partial\sigma^2}\ln L(\mu，\sigma^2)=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\sum\limits_{1}^{n}(y_i-\mu)^2=0\end{cases}$$

解得：

$$\begin{cases}\mu=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i=\bar{y}\\ \sigma^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}(y_i-\mu)^2\end{cases}$$

因此，正态分布总体平均数的极大似然估计量为： $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}y_i=\bar{y}$ 。当总体平均值为未知时， 方 差 估 计 量 为 $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$ ； 当 总 体 平 均 值 为 已 知 时 ， 方 差 估 计 量 为 $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2$ 。

均匀分布的极大似然估计量

例 6 设总体 X 在 $[a,b]$ 上服从均匀分布，a,b 未知，$x_1,x_2,\cdots,x_n$ 是一个样本值. 试求 a,b 的最大似然估计量.

解 记 $x_{(1)}=\min\{x_1,x_2,\cdots,x_n\}$，$x_{(n)}=\max\{x_1,x_2,\cdots,x_n\}$. X 的概率密度是

$$f(x;a,b)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b,\\ 0, & \text{其他}.\end{cases}$$

似然函数为

$$L(a,b)=\begin{cases}\dfrac{1}{(b-a)^n}, & a\leqslant x_1,x_2,\cdots,x_n\leqslant b,\\ 0, & \text{其他}.\end{cases}$$

由于 $a\leqslant x_1,x_2,\cdots,x_n\leqslant b$，等价于 $a\leqslant x_{(1)},x_{(n)}\leqslant b$. 似然函数可写成

$$L(a,b)=\begin{cases}\dfrac{1}{(b-a)^n}, a\leqslant x_{(1)}, & b\geqslant x_{(n)},\\ 0, & \text{其他}.\end{cases}$$

于是对于满足条件 $a\leqslant x_{(1)},b\geqslant x_{(n)}$ 的任意 a,b 有

$$L(a,b)=\frac{1}{(b-a)^n}\leqslant\frac{1}{(x_{(n)}-x_{(1)})^n}.$$

即 $L(a,b)$ 在 $a=x_{(1)},b=x_{(n)}$ 时取到最大值 $(x_{(n)}-x_{(1)})^{-n}$. 故 a,b 的最大似然估计值为

$$\hat{a}=x_{(1)}=\min_{1\leqslant i\leqslant n}x_i,\quad \hat{b}=x_{(n)}=\max_{1\leqslant i\leqslant n}x_i.$$

a,b 的最大似然估计量为

$$\hat{a}=\min_{1\leqslant i\leqslant n}X_i,\quad \hat{b}=\max_{1\leqslant i\leqslant n}X_i.$$ □

极大似然估计的不变性

- 设θ的函数u=u(θ), θ∈Θ具有单值反函数θ=θ(u),u∈$\mathscr{U}$。若$\hat{\theta}$是X的概率分布函数中参数θ的极大似然估计，则$\hat{u}=\hat{u}(\hat{\theta})$是u(θ)的极大似然估计。

- 对于正态总体，总体方差σ^2的极大似然估计量为$\widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$。函数$u=u(\sigma^2)=\sqrt{\sigma^2}$有单值反函数$\sigma^2=u^2$。故标准差σ的极大似然估计量为

$$\hat{\sigma}=\sqrt{\widehat{\sigma^2}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间