



大数据的统计学基础——第8周

【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>

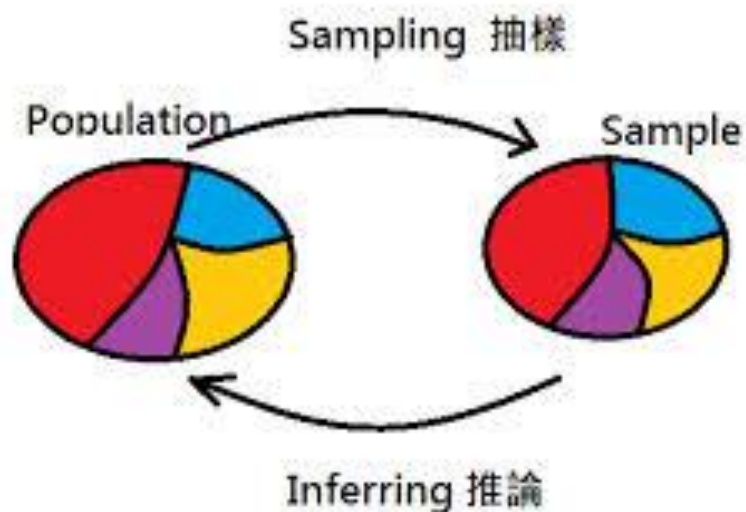
关注炼数成金企业微信



■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！



- ◆ 统计学：描述统计学与推断统计学 ——> 根据样本数据推断总体数据的情况
- ◆ 样本均值→总体均值
- ◆ 样本方差→总体方差
- ◆ 样本比例→总体比例



调查：香港56%富裕投资者未来12个月增持股票

港股要闻 | 财华社 2014-04-24 17:30 | 我要分享 ▾



0

日媒调查：68%民众反对安倍修宪行使
集体自卫权

2014年04月22日09:24 来源：中国新闻网

调查称6%中国人喜欢自身工作 与伊拉克持平

2014-02-02 19:12:00 来源：东方网 有0人参与  分享到 ▾

调查称广州两成大学生无理想 青年最重自身价值

调查显示：半数青少年 读书没计划

调查称84%的人不满明年放假安排 反对除夕不放假

样本比例估计总体比例

- ◆ 要求：
- ◆ 1. 样本要为简单随机样本
- ◆ 2. 二项分布的条件成立
- ◆ 3. 至少有5个成功，5个失败，即 $np \geq 5, nq \geq 5$
- ◆ 样本比例 \hat{p} 是总体比例 p 的最好点估计（Point estimation）——无偏而且最有效

- ◆ 美国的“全国艾滋行为调查”访问了2673位成人同性恋者的随机样本。其中，有170人承认，在前一年曾有超过一个性伴侣，占样本的6.36%。（这个结果可能会存在偏差，因为有人会不愿意把自己的性行为如实告诉别人，但我们在这里假设所有人都说了实话）
- ◆ 根据以上的数据，我们可以推断，美国所有成年同性恋者中有不止一个性伴侣的比例大约是6.36%
- ◆ 但是如果我们再做一次调查，得到的样本比例或许会不一样，假设是6.72%。那么我们应该使用哪个数据区估计总体比例呢？

◆ 刚才的例子中，如果实际上成年异性恋中，有6%的人不止一个性伴侣。则真实的总体比例 $p=0.06$ 。“全国艾滋行为调查”的大小为 $n=2673$ 的样本，如果重复抽取多次的话，得到的样本比例 \hat{p} 的分布会很接近于正态分布（中心极限定理）

◆ 分布的均值：0.06

◆ 分布的标准差： $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.06*0.94}{2673}} \approx 0.0046$

◆ 所有的样本比例 \hat{p} 中，约有95%会落在2个标准差之内，即

◆ $(p-2* \sqrt{\frac{p(1-p)}{n}}, p+2* \sqrt{\frac{p(1-p)}{n}}) = (0.0508, 0.0692)$

◆ 有95%的 \hat{p} 跟 p 的差距的绝对值在 $2* \sqrt{\frac{p(1-p)}{n}}$ 之内。换句话说，95%的 $(\hat{p}-2* \sqrt{\frac{p(1-p)}{n}},$

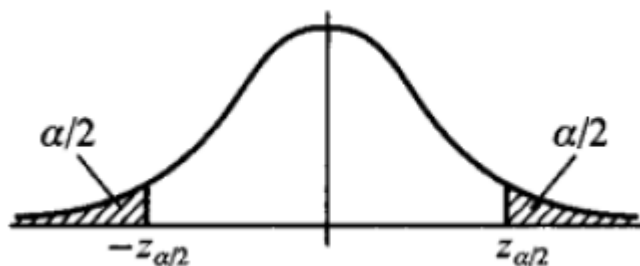
$\hat{p}+2* \sqrt{\frac{p(1-p)}{n}})$ 区间会包含正真的总体比例 p

- ◆ 中心极限定理：样本比例 \hat{p} 近似正态分布 $N(p, p(1-p)/n)$

定理三(棣莫弗—拉普拉斯(De Moivre-Laplace)定理) 设随机变量 η_n ($n=1,2,\dots$) 服从参数为 n, p ($0 < p < 1$) 的二项分布, 则对于任意 x , 有

$$\lim_{n \rightarrow \infty} P\left\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \leq x\right\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x). \quad (2.5)$$

- ◆ 样本比例落在尾部的概率非常小
- ◆ 样本比例落在阴影尾部的总概率为 α
- ◆ 样本比例落在中间部分的概率为 $1-\alpha$



- ◆ 置信区间 (confidence interval) : 用来估计总体参数真实值的一个区间, 通常形式:
估计值 \pm 误差界限
- ◆ 误差界限 (margin of error) : 估计值的最大误差, 使用E表示
- ◆ 置信度 (confidence level) : $1-\alpha$
- ◆ 临界值(critical values): $z_{\alpha/2}$
- ◆ 置信区间边界(confidence interval limits) : 置信上限, 置信下限

◆ $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$, 所以 $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$

◆ $P\left\{\left|\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\right| < \frac{z_{\alpha}}{2}\right\} = 1 - \alpha$, 故 $P\left\{\hat{p} - \frac{z_{\alpha}}{2} * \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + \frac{z_{\alpha}}{2} * \sqrt{\frac{p(1-p)}{n}}\right\} = 1 - \alpha$

◆ 由于p值的真实值不知道，一般采用 $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ 去代替 $\sqrt{\frac{p(1-p)}{n}}$

◆ 所有总体比例p的1- α 置信区间为

$$(\hat{p} - E, \hat{p} + E)$$

◆ 其中, $E = \frac{z_{\alpha}}{2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

- ◆ 要求：
- ◆ 1. 样本要为简单随机样本
- ◆ 2. 二项分布的条件成立
- ◆ 3. 至少有5个成功，5个失败，即 $np \geq 5, nq \geq 5$

- ◆ 之前的例子中，样本比例 $\hat{p}=0.0636$ 。那么所有成年异性恋者中，有不止一个性伴侣的人所占的比例 p 的95%置信区间为：
- ◆
$$\hat{p} \pm z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0636 \pm 1.96 \sqrt{\frac{0.0636*0.9364}{2673}} = 0.0636 \pm 0.0092 = (0.0544, 0.0728)$$
- ◆ 有95%的把握(0.0544,0.0728)会包含真正的总体比例

◆ 更精确的计算方法：

◆ $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$

◆ $1 - \alpha = P\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) = P\{(n + z_{\alpha/2}^2)p^2 - (2n\hat{p} + z_{\alpha/2}^2)p + n\hat{p}^2 < 0\}$

◆ $= P\{p_1 < p < p_2\}$

◆ 令 $(n + z_{\alpha/2}^2) = a$, $-(2n\hat{p} + z_{\alpha/2}^2) = b$, $n\hat{p}^2 = c$, 则

◆ $p_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$, $p_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$

◆ 故p的置信区间为 (p_1, p_2)

样本容量的确定

- ◆ $E = z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow$
- ◆ \hat{p} 已知： $n = \frac{[z_{\alpha/2}]^2 \hat{p}(1-\hat{p})}{E^2}$
- ◆ \hat{p} 未知： $n = \frac{[z_{\alpha/2}]^2 0.25}{E^2}$
- ◆ 取整规则：往上取整。51.1→52
- ◆ 注意适用条件！



- ◆ 日常生活中，我们可以使用不同的工具与他人联系，像电子邮件，qq，微信，电话等等。某位社会学家想知道现今中国内，使用电子邮件的人所占的比例。如果他想要构建一个95%的置信区间，而且要把误差控制在4%以内，那么要调查多少人？
- ◆ （1）如果知道，根据以前的某个调查，在2000年，全国有16.9%的人正使用电子邮件；
- ◆ （2）我们没有任何关于 \hat{p} 的信息。
- ◆ （1）根据过往的调查， $\hat{p}=0.169, 1-\hat{p}=0.831$ 。对于95%的置信区间， $\alpha=0.05$ ， $z_{\alpha/2} = 1.96$ ，误差界限 $E=0.04$ 。
- ◆ 根据公式：
$$n = \frac{[z_{\alpha/2}]^2 \hat{p}(1-\hat{p})}{E^2} = \frac{1.96^2 * 0.169 * 0.831}{0.04^2} = 337.194 = 338$$
- ◆ (2) \hat{p} 未知：
$$n = \frac{[z_{\alpha/2}]^2 * 0.25}{E^2} = \frac{1.96^2 * 0.25}{0.04^2} = 600.25 = 601$$

◆ 假设总体比例的95%的置信区间为(a,b) , 则

◆ $\hat{p} = \frac{a+b}{2}$

◆ $E = \frac{a-b}{2}$

◆ 有限总体校正因子 (Finite Population Correction Factor)

◆ 当总体有限时 , 每次抽样是不放回抽样时

◆ $E = z_{\alpha/2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} * \sqrt{1 - \frac{n}{N}}$ 所以此时 : $n = \frac{N\hat{p}(1-\hat{p})[z_{\alpha/2}]^2}{\hat{p}(1-\hat{p})[z_{\alpha/2}]^2 + (N-1)E^2}$

- ◆ 2012年5月14日，人民银行与西南财经大学共同发布《中国家庭金融调查报告》：中国自有住房拥有率高达89.68%，远超世界60%左右的水平，而城市第一套房平均收益率在300%以上。
- ◆ 2012年5月14日，中国家庭金融调查与研究中心出台《中国家庭金融调查报告》。报告指出：中国城市家庭平均资产为247.6万元，高出美国21%。总体上反映了中国城市家庭拥有较大财富。
- ◆ 2012年8月5日，北京大学发布由北大中国社会科学调查中心完成的《中国民生发展报告2012》。调查称：中国家庭的平均住房面积为116.4平方米，人均住房面积为36.0平方米。2011年中国家庭的平均总支出为3.8万元，比2010年增长了5710元。
- ◆ 2013年2月19日《人民日报海外版》宣布：中国已成为世界第二大经济体，人均GDP已超过5000美元，公共假期已有115天，达到了中等发达国家水平。并指出：解决了“有钱”、“有闲”的问题之后，我国旅游业开始全面增长，进入大众化发展的新阶段。
- ◆ 2013年7月14日新华网刊出“我国人民币存款突破百万亿”、“中国人均存款77623元”。曾任证监会主席的郭树清认为：“中国的储蓄率高达52%，这在世界上是罕见的，而且就大国经济而言历史上不曾有过先例。

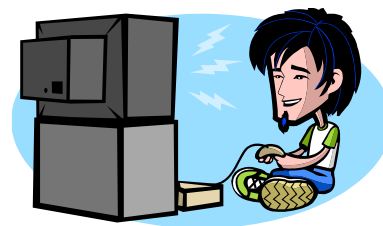
总体均值的估计—— σ 已知

- ◆ 适用条件：
- ◆ 1. 简单随机样本
- ◆ 2. σ 已知
- ◆ 3. 总体为(近似)正态分布或 $n > 30$
- ◆ 点估计：样本均值 \bar{x} 是总体均值 μ 的最好点估计——无偏而且比其他统计量更有效

总体均值的估计—— σ 已知

- ◆ 区间估计
- ◆ 由 $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$, 得 $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- ◆ $1 - \alpha = P\left\{\left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right\} = P\left\{\bar{x} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}\right\} = P\{\bar{x} - E < \mu < \bar{x} + E\}$, 此时 $E = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$
- ◆ 故在 σ 已知的情况下 , 总体均值的置信区间为
- ◆ $(\bar{x} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}})$

- ◆ 某家游戏公司针对某个游戏进行玩家调查，共收集有效问卷125份。问卷中有一个问题是问玩家的年龄。125份问卷中得到的平均年龄为14.75岁。根据之前的调查结果，玩家年龄的标准差为2.45。请根据上述资料构建玩家年龄的95%置信区间。
- ◆ 解： $n=125>30, \bar{x} = 14.75, \sigma = 2.45, z_{\alpha/2}=1.96$
- ◆ 根据公式： $E = \frac{\sigma}{\sqrt{n}} = \frac{2.45}{\sqrt{125}} = 0.219$
- ◆ 故所求置信区间为 $(\bar{x} - z_{\alpha/2} * E, \bar{x} + z_{\alpha/2} * E) = (14.75 - 1.96 * 0.219, 14.75 + 1.96 * 0.219)$



样本容量的确定

- ◆ 在误差界限E和总体标准差 σ 已知的情况下：
- ◆ 由 $E = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$ 得
- ◆ $n = \left[\frac{z_{\alpha/2} \sigma}{E} \right]^2$
- ◆ 当 σ 未知时：
- ◆ （1）使用 极差(range)/4 来近似 σ
- ◆ （2）进行预实验，使用预实验中的样本标准差S来近似 σ
- ◆ （3）使用早期做的其他实验数据

总体均值估计—— σ 未知

- ◆ 适用条件：
 - ◆ 1. 简单随机样本
 - ◆ 2. 总体正态分布或 $n > 30$
- ◆ 点估计：样本均值 \bar{x} 是总体均值 μ 的最好点估计
- ◆ 区间估计： $\frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t(n - 1)$
- ◆ $1 - \alpha = P\left\{\left|\frac{\bar{x} - \mu}{S/\sqrt{n}}\right| < t_{\alpha/2, n-1}\right\} = P\left\{\bar{x} - t_{\alpha/2, n-1} * \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2, n-1} * \frac{S}{\sqrt{n}}\right\} = P\{\bar{x} - E < \mu < \bar{x} + E\}$, 此时, $E = t_{\alpha/2, n-1} * \frac{S}{\sqrt{n}}$

- ◆ 某公司的全部职工中，随机抽取了23名员工，收集了他们的年龄数据，如下：
- ◆ 34,37,37,38,41,42,43,44,44,45,45,45,46,48,49,53,53,54,54,55,56,57,60
- ◆ 求公司平均年龄的点估计与95%置信区间。
- ◆ 解：（1）点估计： $\bar{x} = \frac{34+37+\cdots+57+60}{23} = 47.0$
- ◆ （2）区间估计：
- ◆ $n = 23, n - 1 = 22, \alpha = 0.05, t_{\frac{\alpha}{2}} = 2.074$
- ◆ $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{(34-47)^2 + (37-47)^2 + \cdots + (60-47)^2}{23-1} = 52.1, s = 7.2$
- ◆ $E = t_{\alpha/2} \frac{s}{\sqrt{n}} = 2.074 * \frac{7.2}{\sqrt{23}} = 3.114$
- ◆ 故95%的置信区间为 (43.9, 50.1)

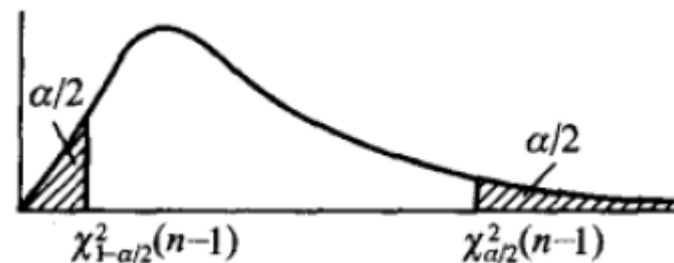
总体方差的估计

- ◆ 要求：
- ◆ 1. 简单随机样本
- ◆ 2. 总体**必须**服从正态分布
- ◆ 点估计：样本方差是总体方差的最好点估计——无偏
- ◆ 一般使用样本标准差估计总体标准差，尽管它是有偏的

- ◆ 区间估计： $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

- ◆ $1 - \alpha = P(\chi_{1-\alpha/2, n-1}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2, n-1}^2)$

- ◆ $= P(\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2})$



- ◆ 英国的硬币——便士现在铸造机器的标准差为0.0165g。现想提高铸造工艺，降低便士的铸造标准差而引进了新的铸币机器。从新的机器铸造的一批硬币中随机抽取了10个硬币，测量其重量。根据测量的数据知道，这10个硬币的重量标准差为0.0125g。
已知便士的重量服从正态分布，求新机器铸造的硬币的标准差的95%置信区间，并由此判断新的机器的铸造技艺是否有所改进。

- ◆ $n=10, \alpha = 0.05$ ，故 $\chi^2_{1-\alpha/2, n-1} = 2.700$ ，

- ◆ $\chi^2_{\frac{\alpha}{2}, n-1} = 19.022$

- ◆ 代入数据，

- ◆ $\frac{(10-1)0.0125^2}{19.022} < \sigma^2 < \frac{(10-1)0.0125^2}{2.700}$

- ◆ 两边开方，得 (0.0086, 0.0228)

- ◆ 不能说明铸造技艺提升

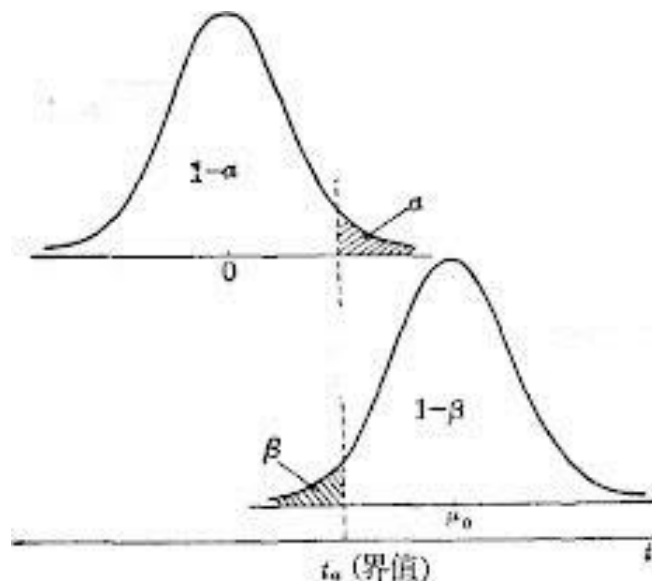
α	0.995	0.99	0.975	0.025
n				
1	0.000	0.000	0.001	5.025
2	0.010	0.020	0.051	7.378
3	0.072	0.115	0.216	9.348
4	0.207	0.297	0.484	11.143
5	0.412	0.554	0.831	12.832
6	0.676	0.872	1.237	14.440
7	0.989	1.239	1.690	16.012
8	1.344	1.646	2.180	17.534
9	1.735	2.088	2.700	19.022
10	2.156	2.558	3.247	20.483

◆ 单侧置信区间：（置信下限， ∞ ）或是（ $-\infty$ ，置信上限）

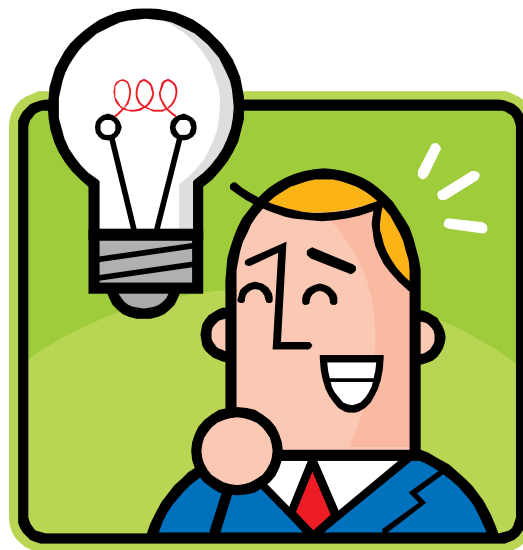
◆ 总体均值的单侧置信区间：

◆ $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1) \rightarrow 1-\alpha = P\left\{\frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha,n-1}\right\} = P\left\{\bar{X} - t_{\alpha,n-1} * \frac{S}{\sqrt{n}} < \mu < \infty\right\}$

◆ $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1) \rightarrow 1-\alpha = P\left\{t_{1-\alpha,n-1} < \frac{\bar{X}-\mu}{S/\sqrt{n}}\right\} = P\left\{-\infty < \mu < \bar{X} + t_{\alpha,n-1} * \frac{S}{\sqrt{n}}\right\}$



- ◆ 从一批灯泡中随机地取5只做寿命试验，测得寿命（以h计）为：
- ◆ 1050, 1100, 1120, 1250, 1280
- ◆ 设灯泡的寿命服从正态分布。求灯泡寿命平均值的95%的单侧置信区间。
- ◆ $\alpha=0.05$, $n=5, t_{\alpha, n-1} = t_{0.05, 4} = 2.1318, \bar{x} = 1160, s^2 = 9950$
- ◆ $\left(\bar{x} - \frac{s}{\sqrt{n}} * t_{\alpha, n-1}, \infty \right) = (1065, \infty)$



- ◆ 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