经济统计 – 6

刁莉男

diaoln@jlu.edu.cn

吉林大学商学院

April 18, 2012

复习

- 抽样误差
- 样本均值的抽样分布
- 中心极限定理(CLT)
- 均值的点估计与置信区间（σ已知）

提纲

9. 均值点估计与置信区间

- σ未知
- 比例的置信区间
- 有限总体修正因子
- 选择合适的样本容量

1. 均值的点估计与区间估计，σ未知情况。

均值的点估计与区间估计，σ未知

当总体标准差未知时，用样本标准差s代替。构造统计量

- $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$，服从$t(n-1)$分布，但是基于X服从正态分布假设；
- $t = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}}}$
- 置信区间：$\bar{X} \pm t\frac{s}{\sqrt{n}}$；
- t值与置信水平相关；

t分布性质

- 与正态分布相似，t分布也是连续分布；
- 对称、钟形分布；
- 均值为0，方差为$\frac{n}{n-2}$;
- 与正态分布相比更平坦，随着n增大，逐渐趋近于正态分布。

均值的点估计与区间估计，σ未知

例1：一轮胎生产厂商调查其轮胎花纹的使用寿命。随机抽取了10个行程5万英里以上的轮胎，样本均值为0.32 inch，标准差0.09 inch。构造总体均值的95%置信区间；生产厂商得到5万英里之后轮胎花纹的平均厚度（总体均值）为0.30 inch这一结论是否合理?

- 置信水平为95%，$df = 10 - 1 = 9$时，t=2.262。
- $\bar{X} \pm t\frac{s}{\sqrt{n}} = 0.32 \pm 2.262\frac{0.09}{\sqrt{10}} = [0.256, 0.384]$；

均值的点估计与区间估计，σ未知

例2：一商场管理者想估计每位顾客的平均消费金额，下表给出了20个顾客的消费金额：

48.16	42.22	46.82	51.45	23.78	41.86	54.86
37.92	52.64	48.59	50.82	46.94	61.83	61.69
49.17	61.46	51.35	52.68	58.84	43.88	

- 总体均值的最好估计量是多少？
- 确定总体均值的95%置信区间，
- 解释结果，
- 总体均值为50（60）这一结论是否合理？

2. 比例的置信区间

(A Confidence Interval for a Proportion)

比例的置信区间 (A Confidence Interval for a Proportion)

- 定比尺度数据：重量、容量、收入、距离、年龄等。
- 定类尺度数据：
  - 南理工学院报告中指出，80%学生毕业后从事与本专业相关工作。
  - Burger King 45%的销售额都来自于汽车专用道窗口。
  - 调查表明芝加哥地区85%的建筑都有中央空调。
  - 调查表明，63%的35-50岁之间的已婚男士都认为双方都应该有工作。

比例的置信区间

比例：样本或总体中，比例是指具有某一特性的个体占全部样本或总体的比重(fraction)、比率(ratio)或百分比(percent)。

- 总体比例用π表示。
- 样本比例用p来表示，$p = \frac{X}{n}$。

比例的置信区间

构造比例的置信区间需要满足的假设

- 二项分布（或者说伯努利分布Bernoulli）的假设必须满足；
  - 样本数据为计数（成功的次数）。
  - 只有两种可能：成功、失败。
  - 独立同分布（iid）。
- $n\pi > 5$ 并且$n(1-\pi) > 5$，使我们能够使用CLT并且应用标准正态分布以及z统计量。

总体比例的置信区间：$p \pm z\sqrt{\frac{p(1-p)}{n}}$

比例的置信区间

例1：BBA（Bottle Blower America）工会在考虑与Teamsters工会合并。根据工会规章，3/4以上的工会成员通过才能进行合并。BBA工会一个2000人的随机样本中，1600人表示同意合并。

- 总体比例的点估计是多少？
- 总体比例的置信区间是多少？ $[.80 \pm 0.018]$
- 基于此样本信息，我们能够得出结论：BBA工会的成员中有足够多的比例希望合并？为什么？ [0.782, 0.818]

对置信区间的解释

例2：Cliff Obermeyer竞选区议员，为了赢得选举Obermeyer必须得到50%以上的支持率 。随机联系了500个选民，其中275名表示支持他。因此，对总体支持率的点估计为55%。问题是：

- 我们是不是从一个支持率少于50%的总体中抽到了支持率为55%的样本？【50%落在置信区间内】
- 或者，总体中支持率本身就高于50%（为55%）。也就是说，5%的抽样误差完全由于随机抽样产生。【50%在置信区间外（右边）】

比例的置信区间

练习：一项市场调查是关于家庭主妇们是否能够根据瓶子的形状以及颜色识别出清洁剂的品牌。1400个样本中，420名家庭主妇可以识别。

- 估计总体比例；
- 建立99%的置信区间；
- 请对结果做出解释。

3. 有限总体修正因子

(Finite-Population Correction Factor)

有限总体修正因子 Finite-Population Correction Factor

- 区间估计中，我们考虑的都是无限总体不重复抽样，相当于有限总体重复抽样，抽样可能性N^n。
- 现实中，多为有限总体不重复抽样C_N^n，需要使用有限总体修正因子（FPC）对均值以及比例的方差进行修正。
- $FPC = \sqrt{\frac{N-n}{N-1}}$。

有限总体修正因子 (FPC)

- 方差已知时，均值区间估计：
$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$
- 方差未知时，均值区间估计：
$$\bar{X} \pm t\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$
- 总体比例区间估计：
$$p \pm z\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

why FPC

- 假设总体为1000，样本为100。
  $FPC = \sqrt{\frac{900}{999}} = 0.9492$，乘以FPC后使得标准差减少5%=(1-0.9492)/1，缩小了置信区间。
- 假设总体为1000，样本为200。
  $FPC = \sqrt{\frac{800}{999}} = 0.8949$，乘以FPC后使得标准差减少10.5%=(1-0.8949)/1。

Why FPC

当总体容量为1000时，不同样本容量的FPC：

样本容量	占总体比重	FPC
10	0.01	0.9955
25	0.025	0.9879
50	0.05	0.9752
100	0.10	0.9492
200	0.20	0.8949
500	0.50	0.7075

当样本占总体比重不超过5%时，FPC影响比较小，可以忽略不计。

有限总体修正因子(FPC)

例1：Pennsylvania，Scandia地区有250户居民，随机抽取40户作为样本，得到平均每户每年捐给教堂的金额为$450，标准差为$75。总体均值可能为$445或$425吗？步骤为：

- 总体均值为多少？总体均值的最好估计是什么？
- 有必要用到FPC吗？为什么？
- 请为总体均值构造90%置信区间。

  [431.65， 468.35]
- 对置信区间做出解释。

有限总体修正因子(FPC)

例2：同上例，如果随机抽取的40户居民中，有15户定期去教堂。请构造定期去教堂居民比例的95%置信区间。是否应该用到FPC？为什么？

4. 选择合适的样本容量

选择合适的样本容量

受三方面影响：

- 受置信水平影响。置信水平越高，需要选取的样本数量越多。
- 允许的误差程度。小的误差程度则需要较多样本；反之，大误差程度需要较少样本。
- 总体的分散程度。总体越分散（异质），需要的样本越多；反之，总体越集中（同质），需要样本越少。

估计总体标准差

- 使用具有可比性的其它样本标准差。
- 使用全距估计。(Max-Min)/6。
- 先做一个试验研究。从小样本数据代替大样本标准差，进而估计合适样本容量。

选择合适的样本容量

$$E = z\frac{\sigma}{\sqrt{n}}, \quad n = \left(\frac{z\sigma}{E}\right)^2$$

- n为样本容量；
- z为标准正态分布下，置信水平所决定的z统计量；
- σ为总体方差；
- E为最大允许的误差程度。

选择合适的样本容量

例1：一学生想调查大城市市政厅工作人员的平均工资。最大允许误差为$100，置信水平为95%，劳工部一项调查表明工资标准差为$1,000。则需要的样本数量为多少？

$$n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{1.96*1000}{100}\right)^2 = 384.16 = 385$$

选择合适的样本容量

例2：如果这学生想提高置信水平到99%，则需要样本数量为多少？

$$n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{2.58*1000}{100}\right)^2 = 665.64 = 666$$

选择合适的样本容量

同样，该方法也适用于决定比例的样本容量。

$$n = p(1-p)\left(\frac{z}{E}\right)^2$$

- p可以从试验研究或其他具有可比性的数据中得到；
- 否则，可以使用0.5。

选择合适的样本容量

例3：现要估计有私营垃圾收集车的城市比例。最大允许误差比例为0.10，置信水平为90%，无法获得关于p的估计，需要的样本容量为多少？

$n = p(1-p)(\frac{z}{E})^2 = 0.5 * (1-0.5)(\frac{1.65}{0.1})^2 = 68.0625 = 69.$

选择合适的样本容量

例4：学校的注册主任想统计过去10年所有毕业班同学的GPA，GPA在2.0和4.0之间，估计的GPA均值要求在总体均值正负0.05范围内，标准差为0.279，在99%置信水平下，需要的样本容量为多少？

$n = (\frac{2.58*0.279}{0.05})^2 = 207.26 = 208.$