中心极限定理(Author:Black)

前情提要：

当我们计算一个总体的均值和方差时，经常会遇到总体数目过大无法计算的情况，这是我们可以通过采样来进行估计，这里我们使用的估计方法的依据就是中心极限定理

中心极限定理性质：

对总体进行多次随机取样并且每次取样的**维数相同(假设为n)**，这些样本的均值服从正太分布，该正太分布的均值$\mu_{\overline{x}}$近似于总体均值μ, 正态分布的方差$\sigma_{\overline{x}}^2$等于总体方差除以样本维数n：$\sigma_{\overline{x}}^2 \approx \frac{\sigma^2}{n}$。当**n越大**时估计值于实际值越接近

置信区间

前情提要：

我们通过中心极限定理估计出总体均值，但这个值的准确度我们是无法衡量的，我们只知道维数越大准确度越高，但是具体多少维度对应多少的准确度我们还是无法确定的。这样的话我们还是无法衡量我们估计出来的数值是否可信，但是我们可以定义一个数值区间，由于样本均值的分布我们已知，那么总体均值在这个区间的概率就是可求的，这个区间就是置信区间，置信区间要比估计值更直观。

求解置信区间得方法：

1.选择总体统计量

也就是说，我们希望为那个统计量构建置信区间。常见的如均值和比例。比如身高平均值、药效持续时长、治愈率等。选择好统计量，则可以开始进行下一步。

2.求其抽样分布

为了求出统计量的抽样分布，需要知道其期望、方差以及分布。以均值为例（我们构建总体均值的置信区间），我们知道对于均值抽样分布(推导过程，详见前文链接)：$E(\bar{X}) = \mu \quad Var(\bar{X}) = \frac{\sigma^2}{n}$

知道了期望和方差，下面就需要知道抽样分布了。我们知道，根据中心极限定理，当样本很大的时候，均值抽样分布符合正太分布。那如果样本比较小的时候呢？答案是：当样本比较小的时候，均值抽样分布符合t分布。用数学方法表示就是：

- 样本很大的时候，$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$。这里，尽管我们不知道总体的方差，但可以用总体方差的点估计量来估计。因此，改写为：$\bar{X} \sim N(\mu, \frac{s^2}{n})$
- 样本比较小的时候，$\bar{X} \sim t(v)$。这里，v是表示自由度，且$v = n - 1$，其中n为样本大小。（这里不对t分布做更多的讨论）

3.决定执行水平

置信水平表明，我们有多大的信心使得总体统计量位于区间(a, b)内。常用的置信水平是95%，需要注意的是：**置信水平越高，区间越宽，置信区间包含总体统计量的几率也就越大。但是如果置信区间过大，就会失去其意义。**举例来说，“某个地区男性的平均身高介于100cm和200cm之间”，这个概率几乎可以说是100%，但是这样的论断，完全没有实际的意义。现在确定了置信区间，最后就剩下求解置信上下限了。

4.求出置信上下限

均值抽样分布符合正太分布，且置信水平为95%时：我们已知$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$，将其标准化后得到：

$Z = \frac{\bar{X}-\mu}{\frac{\sigma^2}{n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \quad Z \sim N(0,1)$ 查表可得: 当$P(Z < Z_a) = 0.025$时，$Z_a = -1.96$；当$P(Z < Z_b) = 0.975$时，$Z_b = 1.96$。因此，我们需要求解下面的不等式，其中$\bar{X}$用均值点估计量替换，σ用方差点估计量替换：

$-1.96 < \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} < 1.96 \quad \bar{X} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{n}}$ 到此为止，就求出了置信水平为95%下的置信区间为：$(\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}})$

均值抽样分布符合t分布，且置信水平为95%时：我们已知$\bar{X} \sim t(v)$，将其标准化后得到：$T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$

求解时，我们将$\bar{X}$和s分别用均值和方法的点估计量代入即可。类似的，变换不等式则可以求出置信区间为：$(\bar{X} - t\frac{s}{\sqrt{n}}, \bar{X} + t\frac{s}{\sqrt{n}})$，其中t通过查表得出。