点估计和区间估计

司继春

[1]上海对外经贸大学

2024年12月

概览

参数估计

- 在参数模型中，我们假设总体P属于某一个参数族$\{P_\theta, \theta \in \Theta\}$，从而推断总体等价于找到一个参数$\theta_0$，使得$P_{\theta_0} = P$。
- 我们一般把θ_0称为**真值（true value）**。
- 而由于总体是不可观测的，我们只能通过样本对总体进行推断，因而我们不可能得到θ_0的精确值，只能对其进行估计，即**参数估计（Estimation）**。

参数估计

- 参数估计包含两部分，即:
  - **点估计（point estimation）**：找到一个统计量$\hat{\theta}(\boldsymbol{x})$，对总体参数$\theta_0$进行推断，而统计量$\hat{\theta}$我们一般称为**估计量（estimator）**。
  - **区间估计（interval estimation）**：找到一组统计量$L(\boldsymbol{x}), U(\boldsymbol{x})$，使得由其组成的区间包含总体参数$\theta_0$的概率为已知的，即

$$P(L(\boldsymbol{x}) \leq \theta_0 \leq U(\boldsymbol{x})) = p$$

    其中统计量$L(\boldsymbol{x}), U(\boldsymbol{x})$被称为**区间估计量（interval estimator）**。
- 估计量即样本的一个函数，即用于估计参数的统计量，而**估计（estimate）**是对于某一个样本，估计量的实现。

评价估计量的标准

对于同一个参数，经常我们有不同的估计量，比如：

方差的估计

- 对于正态总体$x_i \sim \mathrm{N}\left(\mu, \sigma^2\right)$ $i.i.d$，自然地，σ^2的一个估计量为：

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

而类似的，我们也可以使用：

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

作为σ^2的一个估计。

- 以上两个估计量的差别在于分母的不同，很显然，以上两个估计量具有不同的抽样分布。

均方误差

- 那么，在有很多统计量可供选择时，该如何评价这些统计量呢?
- 一个常用的标准是**均方误差**（**mean squared error**, **MSE**），即对于一个参数θ和它的估计量$\hat{\theta}$，其误差平方的期望$\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2$为估计量$\hat{\theta}$的均方误差。
- 注意由于:

$$\begin{aligned}\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2 &= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\ &= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)^2+\mathbb{E}\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2+2\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)\left(\mathbb{E}\hat{\theta}-\theta_0\right) \\ &= \mathbb{V}\left(\hat{\theta}\right)+\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2+2\left(\mathbb{E}\hat{\theta}-\theta_0\right)\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right) \\ &= \mathbb{V}\left(\hat{\theta}\right)+\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\ &= \mathbb{V}\left(\hat{\theta}\right)+\left[\text{Bias}\left(\hat{\theta}\right)\right]^2\end{aligned}$$

- 其中定义**偏差**（**bias**）：$\text{Bias}\left(\hat{\theta}\right)=\mathbb{E}\left(\hat{\theta}-\theta_0\right)$

均方误差

- 从而$\mathrm{MSE}\left(\hat{\theta}\right) = \mathbb{V}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2$，即均方误差等于估计量的方差与偏差平方的和。
  - 因而，降低均方误差有两种途径：降低估计的方差以及降低偏差。
- 此外，根据均方收敛的定义，只要

$$\mathbb{E}\left(\hat{\theta} - \theta_0\right)^2 = \mathbb{V}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2 \to 0$$

那么$\hat{\theta} \xrightarrow{L^2} \theta_0$，从而$\hat{\theta} \xrightarrow{p} \theta_0$。尽管小样本情况下估计量的偏差不为0，但是我们希望当样本量趋向于无穷时，估计量收敛到真值，也是可以接受的。

评价估计量的三个标准

- 这就引出了评价点估计量的三条标准:
  - **无偏性（unbiasedness）**
  - **有效性（efficiency）**
  - **一致性（consistency）**。

无偏性

- 无偏性要求估计量的偏差为0。
- 当估计量的偏差为0，即$\mathbb{E}\left(\hat{\theta}\right)=\theta_0$时，我们称估计量$\hat{\theta}$为**无偏的（unbiased）**。
- 无偏性意味着，尽管对于每个样本，$\hat{\theta}$对θ_0的估计不可能完全准确而是有误差的，但是估计量$\hat{\theta}$总是围绕在真值θ_0的周围，平均而言（其期望）误差为0。

无偏性

均值的偏差

如果$\mathbb{E}x_i = \mu_0$，那么样本均值的期望:

$$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \mu_0$$

因而$\bar{x}$是总体均值μ_0的无偏估计。

无偏性

方差的偏差

根据之前的计算，我们知道对于$x_i \sim \left(\mu_0, \sigma_0^2\right)\ i.i.d$:

$$\mathbb{E}\left(s^2\right) = \sigma_0^2$$

因而$\hat{\sigma}^2$的期望为:

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \mathbb{E}\left(\frac{N-1}{N}s^2\right) = \frac{N-1}{N}\sigma_0^2$$

因而$\text{Bias}\left(s^2\right) = 0, \text{Bias}\left(\hat{\sigma}^2\right) = \frac{1}{N}\sigma_0^2$，只有$s^2$是$\sigma_0^2$的无偏估计量。

有效性

- 为了降低MSE，除了降低偏差以外，降低估计量的方差$\mathbb{V}\left(\hat{\theta}\right)$也是非常重要的手段。
- 一般而言，如果两个估计量$\hat{\theta}_1$和$\hat{\theta}_2$，如果

$$\mathbb{V}\left(\hat{\theta}_1\right) < \mathbb{V}\left(\hat{\theta}_2\right)$$

那么我们称$\hat{\theta}_1$相对于$\hat{\theta}_2$是**有效（efficient）**的。

有效性

样本方差的有效性

在样本方差的例子中，因为$\hat{\sigma}^2 = \frac{N-1}{N} s^2$，从而

$$\mathbb{V}\left(\hat{\sigma}^2\right) = \left(\frac{N-1}{N}\right)^2 \mathbb{V}\left(s^2\right) < \mathbb{V}\left(s^2\right)$$

因而$\hat{\sigma}^2$是相对于s^2更有效的估计量。

偏差-方差权衡

- 我们注意到，尽管s^2比$\hat{\sigma}^2$的偏差更小，但是s^2比$\hat{\sigma}^2$的方差更大。
- 在很多应用问题，比如非参数回归或者监督学习（supervised learning）中，我们都会碰到类似的**偏差-方差权衡**（**bias-variance tradeoff**），即很多时候，同时降低偏差和方差是不可能的。

一致性

- 很多时候，尽管在有限样本下，一个估计量的偏差不为零，但是如果样本量足够大时，估计量与真值之间的误差充分的小，那么在样本量比较大时，我们也可以得到一个足够好的估计量。
- 如果一个估计量$\hat{\theta}$依概率收敛到真值θ_0，即$\hat{\theta} \xrightarrow{p} \theta_0$，那么我们称估计量$\hat{\theta}$为一**致（consistent）**估计量。
- 如果一个估计量是不一致的，也就是说即便我们拥有无限多的样本，我们也不能获得真值θ_0的估计，因而一致性是对一个估计量的最低要求。

一致性

样本均值的一致性

在样本均值的例子中，如果对样本$\{x_i\}$做额外的假设（比如x_i独立同分布且可积），那么根据大数定律，有

$$\bar{x} \xrightarrow{p} \mathbb{E}(x_i) = \mu_0$$

因而样本均值是总体均值的一致估计量。

一致性

样本方差的一致性

在样本方差的例子中，由于

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2$$

其中$\bar{x} \xrightarrow{p} \mu_0$，而

$$\frac{1}{N}\sum_{i=1}^{N}x_i^2 \xrightarrow{p} \mathbb{E}\left(x^2\right) = \mu_0^2 + \sigma_0^2$$

从而$\hat{\sigma}^2 \xrightarrow{p} \sigma_0^2$，即$\hat{\sigma}^2$是$\sigma_0^2$的一致估计量。而

$$s^2 = \frac{N}{N-1}\hat{\sigma}^2 \xrightarrow{p} \sigma_0^2$$

因而s^2也是σ_0^2的一致估计量。

无偏性与一致性

- 需要注意的是：
  - 无偏性关注的是估计量的期望
  - 而一致性则是当样本足够大时估计量的性质
- 两者并没有任何必然联系，无偏性和一致性并不是彼此的充分或者必要条件。

区间估计

- 在上一节中，无偏性、一致性都是使用单个估计量（如样本均值、样本方差）对未知参数进行估计，这种估计被称为点估计（point estimation）。
- 尽管我们可以使用点估计方法对参数值进行推断，然而我们知道，参数的点估计值$\hat{\theta}$与真值θ_0相等的概率一般为0，即$P\left(\hat{\theta}=\theta_0\right)=0$。
- 因而更进一步的，我们很多时候希望得到一个区间，使得这个区间能够以正的概率包含真值θ_0。这就诞生了**区间估计（interval estimation）**的概念。
- 区间估计即对于样本$\boldsymbol{x}=(x_1,...,x_N)$，通过一对统计量$L(\boldsymbol{x})$和$U(\boldsymbol{x})$，满足$L(\boldsymbol{x})\leq U(\boldsymbol{x})$，我们可以使用区间$[L(\boldsymbol{x}),U(\boldsymbol{x})]$对未知参数$\theta_0$进行推断。

区间估计

区间包含真值的概率

如果样本$x_i \sim \mathrm{N}(\mu_0, 1)\ i.i.d, i = 1, ..., N$，那么区间$[\bar{x} - 0.5, \bar{x} + 0.5]$包含真值$\mu_0$的概率为:

$$\begin{aligned} P(\mu_0 \in [\bar{x} - 0.5, \bar{x} + 0.5]) &= P(\mu_0 - 0.5 \leq \bar{x} \leq \mu_0 + 0.5) \\ &= P\left(-\frac{0.5}{\sqrt{\frac{1}{N}}} \leq \frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N}}} \leq \frac{0.5}{\sqrt{\frac{1}{N}}}\right) \end{aligned}$$

由于$\bar{x} \sim \mathrm{N}\left(\mu_0, \frac{1}{N}\right)$，因而

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N}}} \sim \mathrm{N}(0, 1)$$

区间估计

区间包含真值的概率

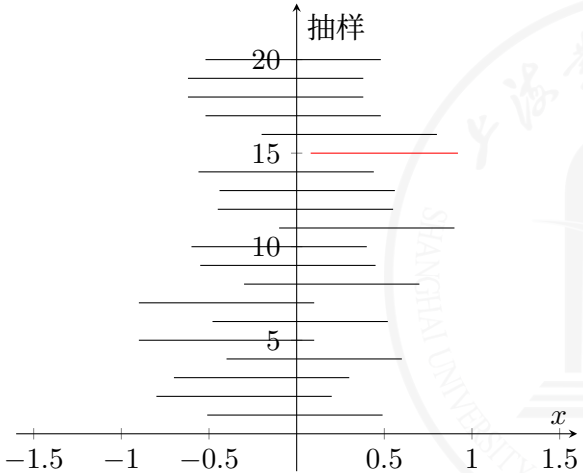
因而

$$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = \Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - \Phi\left(-\frac{0.5}{\sqrt{\frac{1}{N}}}\right)$$

$$= 2\Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - 1$$

例如，当$N = 16$时，查表可得，$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = 2\Phi(2) - 1 \approx 2 \times 0.9772 - 1 = 0.9544$，即区间$[\bar{x}-0.5, \bar{x}+0.5]$包含真值$\mu_0$的概率为95.44%。

区间估计



置信区间

- 注意由于未知参数θ_0是一个未知的常数，而统计量$L(\boldsymbol{x})$和$U(\boldsymbol{x})$是随着抽样的变化而变化的，因此我们不能说「θ_0落入区间$[\bar{x}-0.5, \bar{x}+0.5]$的概率是多少」，而只能说「区间$[\bar{x}-0.5, \bar{x}+0.5]$包含$\theta_0$的概率是多少」。
- 我们把概率$P(\theta_0 \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$称为**覆盖概率（coverage probability）**。
- 注意由于总体参数θ_0未知，因而概率$P(\mu_0 \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$可能依赖于未知的参数$\theta_0$，因而我们通常将而覆盖概率的下界，即$\inf_\theta P_\theta(\theta_0 \in [L(\boldsymbol{x}), U(\boldsymbol{x})])$称为**置信度（confidence coefficient）**或者**置信水平**，通常用$1-\alpha$表示。
- 在某一置信度下，区间$[L(\boldsymbol{x}), U(\boldsymbol{x})]$又被称为**置信区间（confidence interval）**。
  - 因而上例中，我们可以说在95.44%的置信水平下，置信区间为$[\bar{x}-0.5, \bar{x}+0.5]$。

基准统计量

- 此外还需要注意的是，在上例中，为了求得置信区间和覆盖概率，我们首先将统计量$\bar{x}$做了标准化处理，即使用$(\bar{x}-\mu_0)/\sqrt{\frac{1}{N}}$推算概率，而不是直接使用$\bar{x}$。
- 使用$(\bar{x}-\mu_0)/\sqrt{\frac{1}{N}}$的好处是，此统计量的抽样分布不依赖于任何未知参数，因而其分布不会随着未知参数的变化而变化，即服从一个“标准的”分布，这样一来，我们得到的覆盖概率不依赖于任何未知参数。
- $(\bar{x}-\mu_0)/\sqrt{\frac{1}{N}}$依赖未知参数$\mu_0$，严格意义上不是一个统计量，然而$\mu_0$是我们要构建区间估计的目标变量，我们在此允许$\mu_0$的存在。
- 一般的，我们把分布不依赖于未知参数的统计量成为**基准统计量（pivotal statistic）**。

基准统计量

样本均值与基准统计量

如果样本$x_i \sim \mathrm{N}\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i = 1, ..., N$，那么：

1. 统计量$\bar{x} \sim \mathrm{N}\left(\mu_0, \frac{\sigma_0^2}{N}\right)$，其分布依赖于两个未知参数；
2. 统计量$\bar{x} - \mu_0 \sim \mathrm{N}\left(0, \frac{\sigma_0^2}{N}\right)$，其分布仍然依赖于未知参数$\sigma_0^2$；
3. 统计量$\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma_0^2}{N}}} \sim \mathrm{N}(0,1)$分布不依赖于任何未知参数，然而计算过程中$\sigma_0^2$未知；
4. 统计量$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \sim \mathrm{t}(N-1)$分布不依赖于任何未知参数，且计算过程除需要构建置信区间的μ_0之外没有未知参数，因而是基准统计量。

基准统计量

样本方差与基准统计量

如果样本$x_i \sim \mathrm{N}\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i = 1, ..., N$，那么

$$(N-1)\frac{s^2}{\sigma_0^2} \sim \chi^2(N-1)$$

其分布不依赖于任何未知参数，且计算过程除需要构建置信区间的σ_0^2之外没有未知参数，因而是基准统计量。

置信区间的构造

- 之前我们首先给出了区间，进而计算了该区间的置信度。
- 然而现实中，我们经常希望得到在一定置信水平下的置信区间，即一般的区间估计过程。
  - $1-\alpha$常取90%、95%、99%三个数值，其中95%最为常用。
- 由于基准统计量的分布不依赖任何未知参数，因而一般情况下，使用基准统计量可以很方便地构造置信区间。

置信区间的构造

正态分布均值的置信区间

如果样本$x_i \sim \mathrm{N}\left(\mu_0, \sigma_0^2\right)\ i.i.d, i = 1, ..., N$，为了得到$\mu_0$的95%的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有μ_0是未知的，其他都是已知的（包括已知常数以及已知统计量）。在上例中，只有统计量

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \sim \mathrm{t}\left(N - 1\right)$$

满足以上条件。

置信区间的构造

正态分布均值的置信区间

记$F_{\mathrm{t}(N-1)}(x)$为自由度为$N-1$的t分布的分布函数，令$\mathrm{t}_{\alpha/2}^{(N-1)} = F_{\mathrm{t}(N-1)}^{-1}\left(1-\frac{\alpha}{2}\right)$，我们有：

$$
\begin{aligned}
P\left(-\mathrm{t}_{\alpha/2}^{(N-1)} \leq \frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \leq \mathrm{t}_{\alpha/2}^{(N-1)}\right) &= F_{\mathrm{t}(N-1)}\left(\mathrm{t}_{\alpha/2}^{(N-1)}\right) - F_{\mathrm{t}(N-1)}\left(-\mathrm{t}_{\alpha/2}^{(N-1)}\right) \\
&= 1 - 2F_{\mathrm{t}(N-1)}\left(\mathrm{t}_{\alpha/2}^{(N-1)}\right) \\
&= 1 - 2F_{\mathrm{t}(N-1)}\left(F_{\mathrm{t}(N-1)}^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1-\alpha
\end{aligned}
$$

因而我们可以得到：

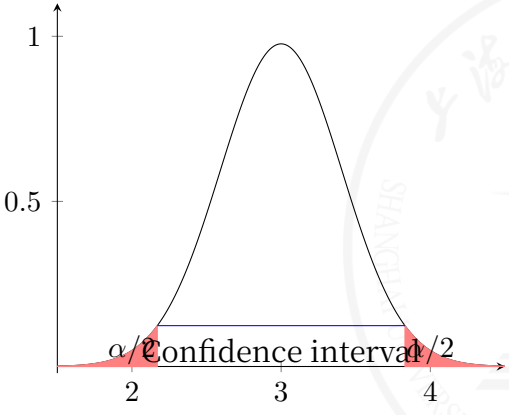
$$
P\left(\bar{x} - \mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}} \leq \mu_0 \leq \bar{x} + \mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}}\right) = 1-\alpha
$$

置信区间的构造

正态分布均值的置信区间

- 从而$\left[\bar{x}-\mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}},\bar{x}+\mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}}\right]$就是我们想要的置信区间。
- 例如，对于一个$N=30$的正态样本，$\bar{x}=3$，$s^2=5$，如果我们想要得到95%置信水平下的置信区间，查表得到$\mathrm{t}_{\alpha/2}^{29}=2.0452$，因而置信下界为$3-2.0452\times\sqrt{5/30}\approx 2.17$，置信上界为$3+2.0452\times\sqrt{5/30}\approx 3.83$。
- 如下图所示，其中红色区域为左右两个概率为$\alpha/2$的区域，中间的一块即为所要求的置信区间。

置信区间的构造



置信区间的构造

正态分布样本方差的置信区间

如果样本$x_i \sim \mathrm{N}\left(\mu_0, \sigma_0^2\right)\ i.i.d, i = 1, ..., N$，为了得到$\sigma_0^2$的95%的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有σ_0^2是未知的，其他都是已知的。在上例中，统计量

$$(N-1)\frac{s^2}{\sigma_0^2} \sim \chi^2(N-1)$$

满足以上条件。

置信区间的构造

正态分布样本方差的置信区间

如果令

$$\chi^{2,(N-1)}_{\alpha/2} = F^{-1}_{\chi^2}\left(1-\frac{\alpha}{2}\right),\ \chi^{2,(N-1)}_{1-\alpha/2} = F^{-1}_{\chi^2}\left(\frac{\alpha}{2}\right)$$

我们有:

$$P\left(\chi^{2,(N-1)}_{1-\alpha/2} \le (N-1)\frac{s^2}{\sigma_0^2} \le \chi^{2,(N-1)}_{\alpha/2}\right) = 1-\alpha$$

因而:

$$P\left(\frac{(N-1)s^2}{\chi^{2,(N-1)}_{\alpha/2}} \le \sigma_0^2 \le \frac{(N-1)s^2}{\chi^{2,(N-1)}_{1-\alpha/2}}\right) = 1-\alpha$$

σ_0^2的95%的置信区间为: $\left[\frac{(N-1)s^2}{\chi^{2,(N-1)}_{\alpha/2}}, \frac{(N-1)s^2}{\chi^{2,(N-1)}_{1-\alpha/2}}\right]$。

置信区间的构造步骤

总结上述两个置信区间的计算，一般而言我们得到置信区间的步骤如下：

1. 给定置信度$1-\alpha$；
2. 找到一个基准统计量，其中只有所要求的参数是未知的，其他都是已知的；
3. 找到这个基准统计量的分布函数$F(\cdot)$；
4. 查表或使用计算机计算$F^{-1}\left(\frac{\alpha}{2}\right)$以及$F^{-1}\left(1-\frac{\alpha}{2}\right)$；
5. 通过不等式变换得到置信区间。

置信区间的模拟

- 我们可以使用程序验证以上步骤得到的置信区间包含真值的概率刚好为置信度$1-\alpha$。
- 代码CI_small_sample.do给出了正态均值的区间估计中小样本均值的区间估计的模拟。
- 在以上程序中，我们首先定义了一个抽样过程，即从$\mathrm{N}(5,100)$的正态总体中进行抽样，接下来使用公式

$$\left[\bar{x}-\mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}},\bar{x}+\mathrm{t}_{\alpha/2}^{(N-1)}\sqrt{\frac{s^2}{N}}\right]$$

计算置信区间，并验证置信区间是否包含了真值（5）。重复抽样10000次，我们就得到了10000个置信区间，最终计算出置信区间包含真值的比率为94.93%，结果显示与与95%相差无几。

大样本下的置信区间

- 尽管上两例给出了正态总体的均值和方差的置信区间的计算方法，然而很多时候我们的总体并不是一定来自于正态总体，很多时候我们很难计算在非正态总体下样本均值的精确分布。
- 然而根据中心极限定理，独立同分布、二阶矩有限的条件下，有：

$$\sqrt{N}\left(\bar{x}-\mu_0\right) \overset{a}{\sim} \mathrm{N}\left(0, \mathbb{V}\left(x\right)\right)$$

- 根据上式，我们有：

$$\frac{\sqrt{N}\left(\bar{x}-\mu_0\right)}{\sqrt{\mathbb{V}\left(x_i\right)}} \overset{a}{\sim} \mathrm{N}\left(0,1\right)$$

然而以上公式中，$\mathbb{V}\left(x_i\right)$是未知的，因而不能直接用于假设检验。

大样本下的置信区间

- 根据Slutsky定理，我们可以使用样本方差s^2代替$\mathbb{V}(x_i)$，由于$s^2 \xrightarrow{p} \mathbb{V}(x_i)$，因而不改变分子上的渐近分布，即有：

$$\frac{\sqrt{N}(\bar{x}-\mu_0)}{\sqrt{s^2}} \overset{a}{\sim} \mathrm{N}(0,1)$$

因而我们可以使用上式进行区间估计。

- 使用类似的技巧，有：

$$P\left(\left|\frac{\sqrt{N}(\bar{x}-\mu_0)}{s}\right| \leq \mathrm{z}_{\alpha/2}\right) = 1-\alpha$$

其中$\mathrm{z}_{\alpha/2} = \Phi_t^{-1}\left(1-\frac{\alpha}{2}\right)$为标准正态分布的上$\alpha/2$分位数，从而置信区间为$\left[\bar{x}-\mathrm{z}_{\alpha/2}\sqrt{\frac{s^2}{N}}, \bar{x}+\mathrm{z}_{\alpha/2}\sqrt{\frac{s^2}{N}}\right]$。

大样本下的置信区间

- 注意$\sqrt{\frac{s^2}{N}}$实际上是$\bar{x}$的标准误$\sqrt{\frac{\sigma_0^2}{N}}$的估计：

$$\text{s.e.}\left(\bar{x}\right)=\sqrt{\frac{s^2}{N}}$$

从而置信区间也可以写为

$$\left[\bar{x}-\text{z}_{\alpha/2}\text{s.e.}\left(\bar{x}\right),\bar{x}+\text{z}_{\alpha/2}\text{s.e.}\left(\bar{x}\right)\right]$$

- 正态总体小样本可以用$\text{t}_{\alpha/2}\left(N-1\right)$代替$\text{z}_{\alpha/2}$，即

$$\left[\bar{x}-\text{t}_{\alpha/2}^{(N-1)}\text{s.e.}\left(\bar{x}\right),\bar{x}+\text{t}_{\alpha/2}^{(N-1)}\text{s.e.}\left(\bar{x}\right)\right]$$

大样本下的置信区间

收入的置信区间

根据2009年中国城镇住户调查，在37480户家庭中，已知家庭年收入均值为54157.63 元，标准差为38533.96元，那么全国家庭家庭平均收入的95%置信区间是多少?

- 一般而言，收入不服从正态分布，但是在大样本条件下，我们知道样本均值近似服从正态分布
- 如果取$1-\alpha=95\%$，查表知$z_{2.5\%}=1.96$，因而:
  - 置信下界为：$54157.63-1.96\times\sqrt{\frac{38533.96^2}{37480}}\approx 53767.50$
  - 置信上界约为54547.75因而全国家庭家庭平均收入的95%置信区间为$[53767.50, 54547.75]$。

大样本下的置信区间

比率的置信区间

根据2013年中国家庭金融调查，样本7711户家庭中，有6%的家庭有信用卡，请问全国持有信用卡的家庭比例的95%置信区间是多少?

- 同样的，比例一般不服从正态分布，但是如果把每个家庭是否持有信用卡假设为独立同分布的伯努利分布，即$x_i \sim \text{Ber}(p_0)$，那么$x_i^2 = x_i$，因而$\overline{x^2} = \bar{x}$，从而

$$\frac{s^2}{N} = \frac{N}{N-1}\frac{\overline{x^2} - \bar{x}^2}{N} \approx \frac{\overline{x^2} - \bar{x}^2}{N} = \frac{\bar{x} - \bar{x}^2}{N} = \frac{\bar{x}(1-\bar{x})}{N}$$

其中比例$\hat{p} = \bar{x}$，从而

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \approx \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}}$$

大样本下的置信区间

比率的置信区间

- 由于$\hat{p} \xrightarrow{p} p_0$，从而$\hat{p}(1-\hat{p}) \xrightarrow{p} p_0(1-p_0) = \mathbb{V}(x_i)$，从而根据Slutsky定理:

$$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \overset{a}{\sim} \mathrm{N}(0,1)$$

- 那么
  - 置信下界为

$$\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = 0.06 - 1.96 \times \sqrt{\frac{0.06 \times (1-0.06)}{7711}} \approx 5.47\%$$

  - 同理置信上界约为6.53%
  - 因而全国家庭持有信用卡比例的95%置信区间为$[5.47\%, 6.53\%]$。

大样本下的置信区间的模拟

- 我们同样可以使用程序验证以上大样本条件下均值的区间估计。
- CI_large_sample.do给出了一个模拟
- 以上程序与之前小样本的区间估计模拟类似，区别在于我们放弃了正态性的假设，转而生成了一个$\mathrm{N}(-3,1)$和$\mathrm{N}(3,1)$的混合正态，并在不同的样本量下观察其覆盖率。
- 结果发现，当样本量仅为10时，覆盖率为91.37%，的确不能保证置信区间95%的覆盖了真值，但是当样本量逐渐增大时，覆盖率会接近95%。

两个样本均值比较

- 此外，很多时候我们还对两个样本的差值感兴趣。如果假设两个独立的样本x_1和x_2，其均值分别为$\bar{x}_1$和$\bar{x}_2$，
- 样本均值差值的期望为

$$\mathbb{E}\left(\bar{x}_1 - \bar{x}_2\right) = \mu_1 - \mu_2$$

由于独立性，方差为

$$\mathbb{V}\left(\bar{x}_1 - \bar{x}_2\right) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

两个样本均值比较

- 由于$\bar{x}_k, k = 1, 2$在大样本条件下分别渐近服从正态分布，从而其差值经过标准化后：

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \overset{a}{\sim} \mathrm{N}(0, 1)$$

- 同样，由于σ_1^2和σ_2^2不可观测，我们分别用其样本方差代替，即：

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\mathrm{s.e.}(\bar{x}_1 - \bar{x}_2)} \overset{a}{\sim} \mathrm{N}(0, 1)$$

其中$\mathrm{s.e.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$

- 从而置信区间为$\left[\bar{x}_1 - \bar{x}_2 - \mathrm{z}_{\alpha/2}\mathrm{s.e.}(\bar{x}_1 - \bar{x}_2), \bar{x}_1 - \bar{x}_2 + \mathrm{z}_{\alpha/2}\mathrm{s.e.}(\bar{x}_1 - \bar{x}_2)\right]$。

两个样本均值比较

不同性别收入比较

在2009年中国城镇住户调查中，共有23440位20-50岁的男性，以及21184位20-50岁的女性。已知男性年平均收入为28367.96元，标准差为21811.88元；女性年平均收入为20145.77元，标准差为16541.08元。如果假设男女收入独立，请问男女收入差异的95%置信区间是多少?

- 同上，尽管收入不服从正态分布，但是大样本情况下可以使用正态分布近似。其中$\bar{x}_1 - \bar{x}_2 = 28367.96 - 20145.77 = 8222.19$,

$$\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} = \sqrt{\frac{21811.88^2}{23440} + \frac{16541.08^2}{21184}} = 182.24$$

因而其差值的置信区间的下界为$8222.19 - 1.96 \times 182.24 = 7864.99$，同理得到置信上界，最终置信区间为：$[7864.99, 8579.38]$。

样本量的确定

- 区间估计中区间的长度实际上度量了精度：如果区间估计的长度越短，那么意味着在一定的置信水平下，我们可以把真值的范围缩的更小。
- 而反过来，根据区间估计，我们还能确定为了达到某一精度所需要样本量的大小。
- 比如，为了确定给定精度下样本均值对总体均值的样本量，根据中心极限定理，在一定条件下有$\sqrt{N}\left(\bar{x}-\mu_0\right) \overset{a}{\sim} \mathrm{N}\left(0, \mathbb{V}\left(x\right)\right)$

样本量的确定

- 因而大样本条件下，均值在$1-\alpha$置信水平下的置信区间为：$\left[\bar{x}-\mathrm{z}_{\alpha/2}\frac{\sigma}{\sqrt{N}},\bar{x}+\mathrm{z}_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right]$区间长度应为：$\frac{2\mathrm{z}_{\alpha/2}\sigma}{\sqrt{N}}$。
- 可以看到，区间大小随着样本量的增加而减小。
- 如果我们要求在$1-\alpha$置信水平下的置信区间的长度为l，那么样本量应为$N=\left[\frac{2\mathrm{z}_{\alpha/2}\sigma}{l}\right]^2$，即样本量与精度成二次方关系。

样本量的确定

比率问题中样本量的确定

在信用卡比例的例子中，为了使95%置信区间长度不超过1%，需要的样本量为:

$$N = \left[\frac{2\mathrm{z}_{\alpha/2}\sigma}{l}\right]^2 = \left[\frac{2 \times 1.96 \times \sqrt{0.06 \times (1-0.06)}}{0.01}\right]^2 \approx 8667$$

即需要8667户样本。

样本量的确定

比率问题中样本量的确定

然而实际情况中，我们不太可能知道σ，所以有时预调研是非常重要的。不过在这个例子中，我们可以得到一个样本量的上界：

$$N = \left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2 = \left[\frac{2z_{\alpha/2}\sqrt{p(1-p)}}{l}\right]^2$$
$$\leq \left[\frac{2z_{\alpha/2}\sqrt{0.5(1-0.5)}}{l}\right]^2 = \left[\frac{z_{\alpha/2}}{l}\right]^2$$

即在这个例子中，如果我们不知道持有信用卡的家庭约为6%，那么需要的样本数量上界为38416户家庭。

作业

- 9.1
- 9.3
- 9.5