













































# 置信区间

- 注意由于未知参数 $\theta_0$ 是一个未知的常数，而统计量 $L(\mathbf{x})$ 和 $U(\mathbf{x})$ 是随着抽样的变化而变化的，因此我们不能说「 $\theta_0$ 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 $P(\theta_0 \in [L(\mathbf{x}), U(\mathbf{x})])$ 称为覆盖概率 (coverage probability)。
- 注意由于总体参数 $\theta_0$ 未知，因而概率 $P(\mu_0 \in [L(\mathbf{x}), U(\mathbf{x})])$ 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 $\inf_{\theta} P_{\theta}(\theta_0 \in [L(\mathbf{x}), U(\mathbf{x})])$ 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 $[L(\mathbf{x}), U(\mathbf{x})]$ 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。
- 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  $x_i \sim N(\mu_0, \sigma_0^2)$  *i.i.d.*,  $i = 1, \dots, N$ , 那么:

- ① 统计量  $\bar{x} \sim N\left(\mu_0, \frac{\sigma_0^2}{N}\right)$ , 其分布依赖于两个未知参数;
- ② 统计量  $\bar{x} - \mu_0 \sim N\left(0, \frac{\sigma_0^2}{N}\right)$ , 其分布仍然依赖于未知参数  $\mu_0$ ;
- ③ 统计量  $\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma_0^2}{N}}} \sim N(0, 1)$  分布不依赖于任何未知参数, 因而是基准统计量;
- ④ 统计量  $\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \sim t(N - 1)$  分布不依赖于任何未知参数, 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  $x_i \sim N(\mu_0, \sigma_0^2)$  *i.i.d.*,  $i = 1, \dots, N$ , 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，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  $x_i \sim N(\mu_0, \sigma_0^2)$  *i.i.d.*,  $i = 1, \dots, N$ , 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  $\mu_0$  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

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

满足以上条件。

# 置信区间的构造

## 正态分布均值的置信区间

如果令  $t_{\alpha/2}^{(N-1)} = F_{t(N-1)}^{-1} \left( 1 - \frac{\alpha}{2} \right)$ , 我们有:

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

因而我们可以得到:

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

# 置信区间的构造

## 正态分布均值的置信区间

- 从而  $\left[ \bar{x} - t_{\alpha/2}^{(N-1)} \sqrt{\frac{s^2}{N}}, \bar{x} + t_{\alpha/2}^{(N-1)} \sqrt{\frac{s^2}{N}} \right]$  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 (d.f.=29) 得到  $t_{\alpha/2} = 2.0452$ ，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

## 正态分布样本方差的置信区间

如果样本  $x_i \sim N(\mu_0, \sigma_0^2)$  *i.i.d.*,  $i = 1, \dots, N$ , 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  $\sigma_0^2$  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

$$\chi_{\alpha/2}^{2,(N-1)} = F_{\chi^2}^{-1} \left( 1 - \frac{\alpha}{2} \right), \quad \chi_{1-\alpha/2}^{2,(N-1)} = F_{\chi^2}^{-1} \left( \frac{\alpha}{2} \right)$$

我们有:

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

因而:

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

$\sigma_0^2$  的 95% 的置信区间为:  $\left[ \frac{(N-1)s^2}{\chi_{\alpha/2}^{2,(N-1)}}, \frac{(N-1)s^2}{\chi_{1-\alpha/2}^{2,(N-1)}} \right]$ 。

# 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  $F^{-1}(\frac{\alpha}{2})$  以及  $F^{-1}(1 - \frac{\alpha}{2})$ ；
- 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 $N(5, 100)$ 的正态总体中进行抽样，接下来使用公式

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

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：

$$\sqrt{N}(\bar{x} - \mu_0) \stackrel{a}{\sim} N(0, \mathbb{V}(x))$$

- 根据上式，我们有：

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

然而以上公式中， $\mathbb{V}(x_i)$ 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：

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

因而我们可以使用上式进行区间估计。

- 使用类似的技巧，有：

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

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

# 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:

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

从而置信区间也可以写为

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

- 正态总体小样本可以用  $t_{\alpha/2}(N-1)$  代替  $z_{\alpha/2}$ , 即

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

# 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，但是在大量样本条件下，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  $x_i \sim Ber(p_0)$ ，那么  $\overline{x^2} = \bar{x}^2$ ，因而  $\overline{x^2} = \bar{x}^2$ ，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:

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

- 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 $N(-3, 1)$ 和 $N(3, 1)$ 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

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

由于独立性，方差为

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

# 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：

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

- 同样，由于  $\sigma_1^2$  和  $\sigma_2^2$  不可观测，我们分别用其样本方差代替，即：

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

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

- 从而置信区间

为  $[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \text{s.e.}(\bar{x}_1 - \bar{x}_2), \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \text{s.e.}(\bar{x}_1 - \bar{x}_2)]$ 。

# 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 $\sqrt{N}(\bar{x} - \mu_0) \stackrel{a}{\sim} N(0, \mathbb{V}(x))$

# 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： $\left[ \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{N}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{N}} \right]$  区间长度应为： $\frac{2z_{\alpha/2}\sigma}{\sqrt{N}}$ 。
- 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 $N = \left[ \frac{2z_{\alpha/2}\sigma}{l} \right]^2$ ，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：

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

即需要8667户样本。



