Introduction to Estimation

Applied Statistics

Fall 2025

目录

Outline

1. 点估计与区间估计 (Point and interval estimators)
2. 估计量的性质 (Qualities of estimators)
3. 置信区间 (Confidence intervals)
4. 确定样本量 (Determine the sample size)

1 统计推断 (Statistical Inference)

- **统计推断 (Statistical inference)**：使用样本数据对总体做出结论
- 它以概率的形式表达我们对结论的信心程度
- 有效性需要随机抽样
- 对于数据生产中的基本缺陷（如自愿回应样本）导致的非抽样误差，无法补救

1.1 两种推断类型 (Two Types of Inference)

- **估计 (Estimation)**：使用区间估计总体参数 (population parameter) 的值
  示例：" 估计医院的平均等待时间"
- **假设检验 (Hypothesis Tests)**：评估关于总体 (population) 的某个主张的证据
  示例：" 检验平均等待时间是否超过 20 分钟"
- 在两种技术中，我们都问：" 如果我们多次使用推断方法，会发生什么？"

2 估计 (Estimation)

- **估计 (Estimation)**：基于样本统计量确定总体参数的近似值
- **估计量 vs. 估计值 (Estimator (random variable) vs. Estimate (number))**：
  - " 估计量是一个规则或公式"
  - " 估计值是从样本计算出的数值"
- **两种估计量 (Two types of estimators)**：点估计量和区间估计量

2.1 点估计 (Point Estimators)

- 点估计使用单个值或点来估计参数
- 样本均值 $\bar{X}$ 是总体均值 μ 的点估计
- 样本标准差 s 是 σ 的点估计

2.2 点估计量的性质 (Qualities of Estimators)

在选择点估计量时，我们寻找具有最佳统计性质的估计量：

- **无偏性 (Unbiasedness)**：期望值等于参数的估计量
- **一致性 (Consistency)**：随着样本量增大，估计量与参数之间的差异变小
- **相对有效性 (Relative efficiency)**：方差 (variance) 较小的估计量

2.2.1 无偏估计量 (Unbiased Estimators)

- 总体参数的无偏估计量是其期望值 (expected value) 等于该参数的估计量
- 样本均值 $\bar{X}$ 是总体均值 μ 的无偏估计量：$E(\bar{X}) = \mu$
- 当总体对称时，样本中位数也是总体均值 μ 的无偏估计量
- 样本方差 $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ 是总体方差 σ^2 的无偏估计量
  相比之下，$\frac{\sum(x_i - \bar{x})^2}{n}$ 是有偏 (biased) 估计量

样本方差的无偏性证明 (Unbiasedness of the Sample Variance)
设 $X_1, \ldots, X_n$ 为独立同分布的随机变量，均值为 μ，方差为 σ^2。样本方差定义为：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

关键思路：展开平方偏差的和：

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

两边取期望：

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = n\sigma^2 - n\mathrm{Var}(\bar{X}) = n\sigma^2 - n\left(\frac{\sigma^2}{n}\right) = (n-1)\sigma^2$$

除以 $n-1$：

$$\mathbb{E}[s^2] = \sigma^2$$

2.2.2 一致性 (Consistency)

- 如果随着样本量增大，估计量与参数之间的差异变小，则该估计量是一致的
- $\bar{X}$ 是 μ 的一致估计量，因为随着 n 增大，$\bar{X}$ 的方差 $Var(\bar{X}) = \frac{\sigma^2}{n}$ 变小
- 对于正态总体，样本中位数是 μ 的一致估计量：$Var(\text{样本中位数}) = \frac{2\pi\sigma^2}{4n} \approx \frac{1.57\sigma^2}{n}$

一致性的形式定义 (Formal Definition of Consistency)

一致性正式定义为以概率收敛 (converge in probability) 于参数的真实值，例如：

$$\lim_{n\to\infty} Pr(|\bar{x} - \mu| > \epsilon) = 0 \quad \text{对所有} \epsilon > 0$$

如果四阶矩 (fourth moment) 有限，样本方差 $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ 是总体方差 σ^2 的一致估计量 (consistent estimator)。

在相同条件下，$\frac{\sum(x_i - \bar{x})^2}{n}$ 也是总体方差 σ^2 的一致估计量。

2.2.3 相对有效性 (Relative Efficiency)

如果有两个参数的无偏估计量，方差较小的那个被称为相对有效的。例如，对于正态总体，样本中位数和样本均值都是总体均值的无偏估计量。样本中位数的方差大于样本均值的方差：

$$Var(\text{样本中位数}) = 1.57 \times Var(\bar{X})$$

因此 $\bar{X}$ 相对更有效。

2.3 区间估计 (Interval Estimation)

- 区间估计给出总体参数的一个合理值范围
- 区间估计围绕点估计构建：中心 ± 边际误差

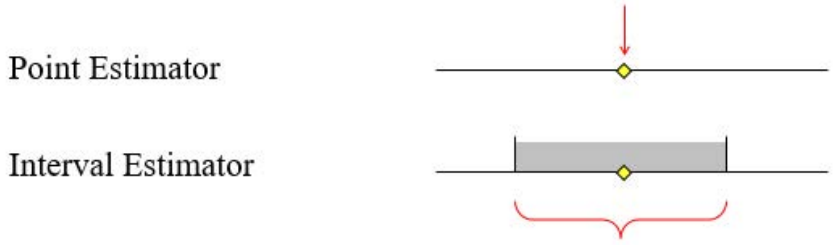

图 1: 点估计量与区间估计量的比较

2.3.1 区间估计的基本思想

我们想要确定总体均值 μ 的合理值区间。

假设有一个正态总体，均值 μ，标准差为 20。我们抽取一个样本量为 16 的样本，样本均值为 240.79。

我们可以猜测 μ " 大概" 在 240.79 附近。那么，μ 可能有多接近 240.79 呢？

为了回答这个问题，我们必须问另一个问题：如果我们从总体中抽取许多大小为 16 的简单随机样本，样本均值 $\bar{x}$ 会如何变化？

在重复抽样中，样本均值的值将服从均值为 μ、标准差为 5 的正态分布。

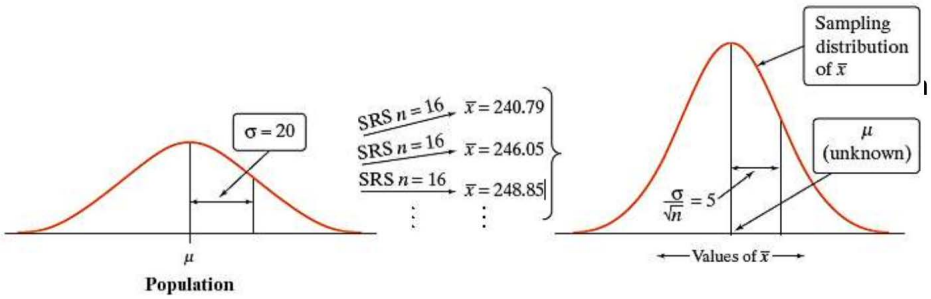

图 2: 区间估计示例

2.3.2 经验法则

经验法则 (empirical rule) 告诉我们，在所有大小为 16 的样本中，95% 的样本均值将在 μ 的 10 个单位内（两个标准差）。

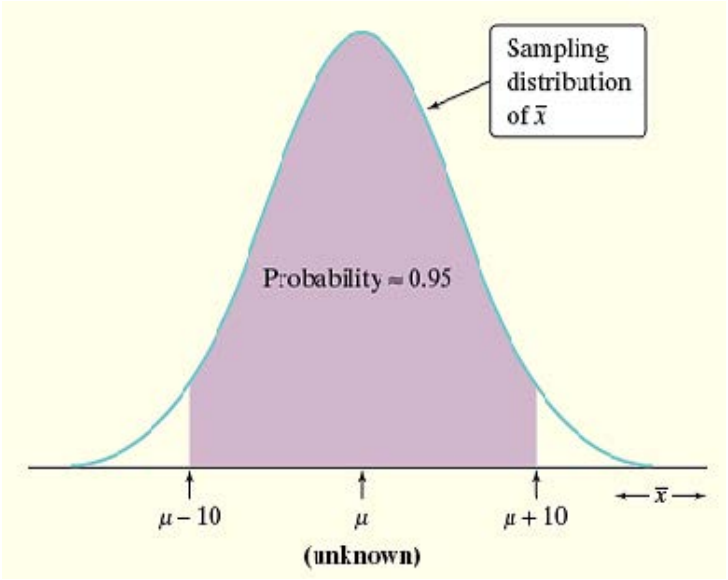

图 3: 样本均值的抽样分布

如果样本均值在 μ 的 10 个单位内，那么 μ 就在样本均值的 10 个单位内：

$$\mu - 10 \leq \bar{x} \leq \mu + 10 \quad \leftrightarrow \quad \bar{x} - 10 \leq \mu \leq \bar{x} + 10$$

从样本均值下方 10 个点到上方 10 个点的区间将在约 95% 的所有大小为 16 的样本中” 捕获” μ。

如果我们估计 μ 在 230.79 到 250.79 之间的某个区间内，我们将使用一种方法，在该大小的所有可能样本中约 95% 的情况下捕获真实的 μ。

2.3.3 使用模拟解释置信区间

考虑估计均匀分布 $U(0, 100)$ 数据的样本均值，其中 $\mu = 50$，$\sigma = \frac{100}{\sqrt{12}} = 28.9$。

我们抽取大小为 $n = 100$ 的样本并计算样本均值。95% 置信区间估计量为 $\bar{X} \pm 5.8$。

如果我们重复抽样多次，95% 的 $\bar{X}$ 值将使 μ 位于 $\bar{X} \pm 5.8$ 区间内，5% 将产生不包括 μ 的区间。

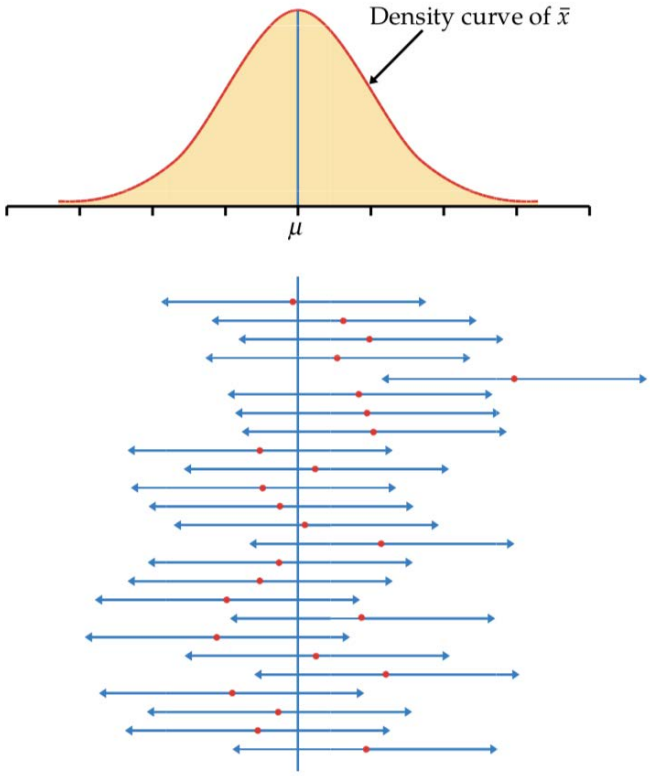

图 4: 置信区间的模拟情况

3 置信区间 (Confidence Interval)

置信区间：**估计值** ± **边际误差**，具有一定置信度 (degree of interval)

- 置信度表示区间捕获真实参数的可能性
- 常见置信度 $1 - \alpha$：99%，95%，90%
- ”95% 的置信区间并不意味着真实均值在此特定区间内的概率是 95%”
- 相反，该方法产生的区间在重复抽样中 95% 的情况下包含真实均值

3.1 构建总体均值的置信区间

回顾样本均值的抽样分布：

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

标准化给出 Z 统计量：

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

要构建 $1-\alpha$ 置信区间，我们从以下概率陈述开始：

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

这是关于 $\bar{X}$ 的概率陈述：

$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

将其解释为 μ 的区间估计量，我们得到 $(1-\alpha)$ 置信区间：

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

边际误差：$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

3.2 常见临界 Z 值 (Common Critical z Values)

我们可以使用 z/t 值表来查找 $Z_{\alpha/2}$，也可以使用软件查找临界 z 值。

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
置信水平 C	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

表 1: 常见置信水平对应的临界 z 值

计算机需求示例

为了改进库存规划，公司需要估计平均需求。经理注意到交货期内的需求服从正态分布。假设经理知道标准差为 75 台计算机。经理希望得到平均需求的 95% 置信区间估计。他记录了 25 个交货期的需求。从数据中，我们计算出平均需求为 370.16 台。

解答： 已知：$\bar{X} = 370.16$，$\sigma = 75$，$n = 25$，置信水平 $1 - \alpha = 0.95$

临界值：$z_{\alpha/2} = z_{0.025} = 1.96$

标准误：$\frac{\sigma}{\sqrt{n}} = \frac{75}{\sqrt{25}} = \frac{75}{5} = 15$

边际误差：$z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = 1.96 \times 15 = 29.4$

置信区间：$370.16 \pm 29.4 = (340.76, 399.56)$

解释： 平均需求的 95% 置信区间在 340.76 到 399.56 之间。

置信区间解释

不要解释为" 总体均值在 340.76 和 399.56 之间的概率为 95%"，因为这意味着总体均值是一个我们可以做出概率陈述的变量。
置信区间估计量是关于样本均值的概率陈述。
正确解释：置信区间在 95% 的情况下可能捕获总体均值。

练习题：滑雪度假村客人年龄

滑雪度假村客人的年龄通常呈右偏分布。假设年龄的标准差为 14.5 岁。

1. 随机样本中 40 位客人平均年龄 $\bar{X}$ 的分布形状是什么？
2. 从 40 位客人的随机样本中，平均年龄为 36.4 岁。求 μ（滑雪度假村客人的真实平均年龄）的 95% 置信区间。
3. 解释你的结果。

解答：

1. 根据中心极限定理，对于样本量 $n = 40$（大于 30），样本均值的分布近似正态，尽管原始总体是右偏的。
2. 已知：$\bar{X} = 36.4$，$\sigma = 14.5$，$n = 40$，$z_{\alpha/2} = 1.96$
   标准误：$\frac{\sigma}{\sqrt{n}} = \frac{14.5}{\sqrt{40}} \approx 2.29$
   边际误差：$1.96 \times 2.29 \approx 4.49$
   置信区间：$36.4 \pm 4.49 = (31.91, 40.89)$
3. 我们有 95% 的信心认为，滑雪度假村客人的真实平均年龄在 31.91 岁到 40.89 岁之间。

3.3 区间宽度 (Interval Width)

- 宽区间提供的信息很少
- 比较两个估计：
  - 以 95% 置信度，会计师的平均起薪在 \$15,000 到 \$100,000 之间
  - 以 95% 置信度，起薪的置信区间估计在 \$42,000 到 \$45,000 之间
- 当你试图决定是否获得 CPA 证书时，哪一个更有帮助？
- 宽度取决于置信度 $1 - \alpha$、离散程度 σ 和样本量 n

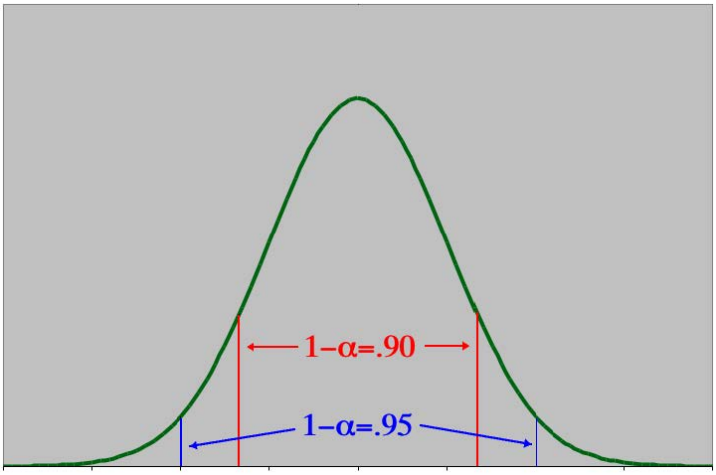

图 5: 不同置信水平对区间宽度的影响

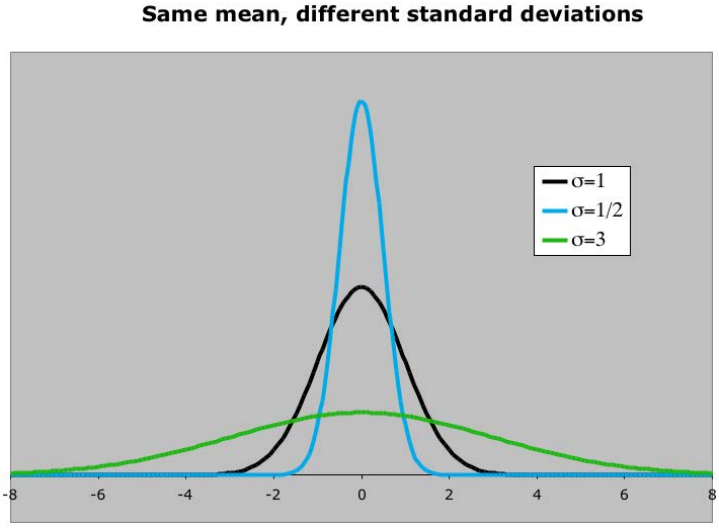

图 6: 不同 σ 对区间宽度的影响

3.3.1 影响区间宽度的因素

- 较大的置信水平产生较宽的置信区间
- 较大的 σ 值产生较宽的置信区间
- 增加样本量减小置信区间的宽度，而置信水平可以保持不变
- 注意：这也增加了获取额外数据的成本

4 选择样本量 (Selecting the Sample Size)

一旦我们决定了期望的估计精度，即置信区间的宽度和置信水平，我们就可以计算达到该精度所需的样本量。

表示期望的宽度为：$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq B$

估计均值的样本量：

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2$$

计算机需求样本量计算

在收集数据之前，经理已经决定他需要在样本平均值周围 16 个单位内估计平均需求。因此边际误差为 16。

我们还有 $1-\alpha = 0.95$ 和 $\sigma = 75$。我们计算：

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2 = \left(\frac{1.96 \times 75}{16}\right)^2 = \left(\frac{147}{16}\right)^2 = (9.1875)^2 = 84.41$$

向上取整到 85。

这意味着要以 95% 的置信度确保估计误差不大于 16，我们需要随机抽样 85 个交货期区间。

练习题：西班牙洋葱重量

一名农产品经理希望以 95% 的置信度和 ±1 盎司的误差估计供应商交付的西班牙洋葱的平均重量。12 个洋葱的初步样本显示标准差为 3.6 盎司。

1. 我们应该称量多少个洋葱来求平均重量的置信区间？
2. 如果我们希望更准确，误差在 ±0.5 以内，样本量是多少？

解答：

1. 已知：$B = 1$，$\sigma \approx 3.6$，$z_{\alpha/2} = 1.96$
   $n \geq \left(\frac{1.96\times3.6}{1}\right)^2 = (7.056)^2 = 49.79$
   向上取整到 50。需要称量 50 个洋葱。

2. 对于 $B = 0.5$：
   $n \geq \left(\frac{1.96\times3.6}{0.5}\right)^2 = (14.112)^2 = 199.2$
   向上取整到 200。需要称量 200 个洋葱。

5 如何估计 σ?

当总体标准差未知时，我们需要估计它来计算样本量。

5.1 方法 1：抽取初步样本

计算小规模试点样本的样本标准差 s。

5.2 方法 2：使用关于分布形式的先验知识

假设保守的均匀分布 $U(a,b)$，基于上限和下限估计 σ：

$$\sigma \approx \frac{b-a}{\sqrt{12}}$$

示例：如果已知卡车重量范围从 1500 公斤到 3500 公斤，那么：

$$\sigma \approx \frac{3500-1500}{\sqrt{12}} \approx 577 \text{ 公斤}$$

5.3 方法 3：使用经验法则

对于近似正态的数据，值通常落在 $\mu \pm 3\sigma$ 内，因此：

$$\sigma \approx \frac{\text{最大值-最小值}}{6}$$

示例：如果交货时间通常在 20 到 50 分钟之间，那么：

$$\sigma \approx \frac{50-20}{6} = 5 \text{ 分钟}$$

5.4 方法 4：对于计数数据

假设泊松到达并猜测均值，然后使用 $\sigma = \sqrt{\lambda}$。

示例：如果平均每小时有 20 位顾客到达，那么：

$$\sigma = \sqrt{20} \approx 4.47$$

Summary

- **统计推断 (Statistical Inference)**：
  - 使用样本数据对总体做出结论
  - 以概率形式表达置信度
  - 有效性需要随机抽样
- **估计类型 (Types of Estimation)**：
  - **点估计 (Point Estimation)**：使用单个值估计参数
  - **区间估计 (Interval Estimation)**：使用值范围估计参数
- **估计量的性质 (Qualities of Estimators)**：
  - **无偏性 (Unbiasedness)**：$E(\hat{\theta}) = \theta$
  - **一致性 (Consistency)**：随着 $n \to \infty$，$\hat{\theta} \to \theta$
  - **相对有效性 (Relative Efficiency)**：方差较小的无偏估计量更有效
- **置信区间 (Confidence Intervals)**：
  - 形式：估计值 $\pm$ 边际误差
  - 置信水平 $1 - \alpha$：区间捕获真实参数的概率
  - 对于正态总体，均值 μ 的置信区间：$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- **区间宽度的影响因素 (Factors Affecting Interval Width)**：
  - 置信水平：置信水平越高，区间越宽
  - 总体变异性：σ 越大，区间越宽
  - 样本量：n 越大，区间越窄
- **样本量确定 (Sample Size Determination)**：
  - 给定期望边际误差 B，样本量 $n \geq \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2$
  - 需要估计 σ 的方法：初步样本、先验知识、经验法则
- **关键公式 (Key Formulas)**：
  - 置信区间：$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
  - 边际误差：$B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
  - 样本量：$n \geq \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2$

- **常见误解 (Common Misconceptions)：**
  - 置信区间不是” 参数在区间内的概率”
  - 而是” 该方法产生的区间在重复抽样中以一定比例包含参数”
- **实际应用 (Practical Applications)：**
  - 在商业、医学、社会科学等领域估计总体参数
  - 根据精度要求规划研究样本量
  - 正确解释置信区间的结果