Sampling Distributions

Applied Statistics

Fall 2025

目录

Outline

1. 抽样与非抽样误差 (Sampling and Non-sampling errors)
2. 样本均值的抽样分布 (Sampling distribution of a sample mean)
3. 其它分布的中心极限定理 (Central limit theorem for other distributions)

1 抽样与非抽样误差 (Sampling and Non-sampling errors)

1.1 估计中的两类误差 (Two types of errors in estimation)

当我们使用样本进行估计时，存在两类误差：**抽样误差 (Sampling error)** 和 **非抽样误差 (Non-sampling error)**。

- **抽样误差 (Sampling error)** 是由于样本与总体之间总存在差异而产生的，无论样本多么具有代表性。
- 抽样误差导致样本均值及其它统计量在相同总体下具有变异性。
- **非抽样误差 (Non-sampling error)** 来源于样本与总体之间系统性的差异。
- 非抽样误差导致估计中的偏倚（偏差）。

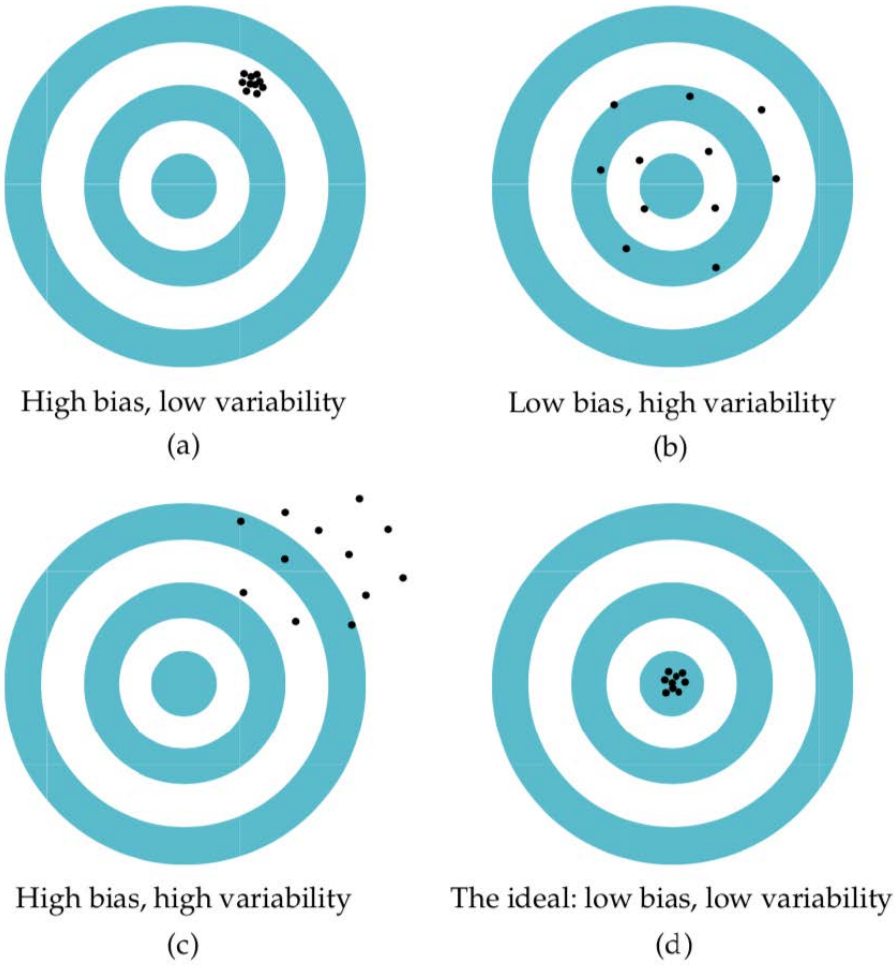

图 1: 偏倚与变异性的四种情况

- (a) 高偏倚，低变异性 (High bias, low variability)
- (b) 低偏倚，高变异性 (Low bias, high variability)
- (c) 高偏倚，高变异性 (High bias, high variability)
- (d) 理想情况：低偏倚，低变异性 (The ideal: low bias, low variability)

1.2 抽样误差 (Sampling Errors)

- 抽样误差是统计分析中 **固有的 (intrinsic)**，只能减少，无法消除。
- 更大的样本量可以减少抽样误差。
- 统计方法用于估计抽样误差的大小：通过设定可能抽样误差大小的界限（边际误差，margin of error）来估计。

1.3 非抽样误差 (Non-Sampling Errors)

非抽样误差是数据收集过程中的缺陷造成的。理论上，如果数据收集过程质量高，非抽样误差是可以避免的。

来源包括：

- 数据获取错误：测量不准确、记录错误、涉及敏感问题的问卷
- 因无回答误差或抽样设计缺陷导致的非代表性样本

注意：增加样本量无助于减少由非抽样误差引起的偏倚。统计方法通常也无能为力。

2 抽样分布 (Sampling Distributions)

- 样本统计量是一个随机变量：其变异来源于抽样的随机性。
- **抽样分布 (Sampling distribution)**：样本统计量的分布
- 抽样分布是衡量抽样误差的一种形式化方法。
- 抽样分布是统计推断的基础。

为什么抽样分布重要

示例：中央银行估计储蓄

某中央银行想要估计该国真实的家庭平均储蓄额（μ）。不可能调查每一个家庭，因此他们随机抽取了 1500 个家庭，计算了样本均值（$\bar{x}$）。
问题：这个单一的 $\bar{x}$ 与真实总体均值有多接近？我们能在多大程度上信任它？
答案在于抽样分布：如果我们了解从该总体中抽取的所有可能样本均值的分布，包括其中心、离散程度和形状，我们就能量化我们的不确定性。
例如，如果我们知道样本均值服从正态分布，我们就能说：在 95% 的情况下，我们的样本估计值落在总体真实平均值加减两个抽样分布标准差的范围内。

2.1 抽样分布理论 (Statistical Theories on Sampling Distributions)

- 统计理论试图了解抽样分布的普遍特征。
- 例如，理论表明样本均值的抽样分布是正态 (normal) 的，因此关于总体均值的推断可以基于此进行。
- 另一个例子是：对于正态总体，样本方差的抽样分布服从卡方分布 (chi-square distribution)，因此在正态总体下关于总体方差的推断基于卡方分布。

2.1.1 示例：样本均值的抽样分布 (Example: Sampling Distribution of the Sample Mean)

我们使用模拟来模拟从一个已知总体中重复抽样。生成的样本可用于提供抽样分布的证据。

假设总体服从 0 到 100 之间的均匀分布：$U(0, 100)$。总体分布的均值和标准差为：

$$\mu = \frac{a+b}{2} = 50, \quad \sigma = \sqrt{\frac{(b-a)^2}{12}} = 28.87$$

2.2 均值的抽样分布 (Sampling Distribution of the Mean)

假设我们收集一个样本量为 100 的样本来研究样本均值的分布。为了检验抽样分布，我们从总体中抽取 1000 个样本量为 100 的随机样本。然后我们计算 1000 个样本均值；这些样本均值的分布就是抽样分布。

2.2.1 模拟结果 (Simulation Results)

- 样本均值比个体观测值变异性更小。
- 样本均值比个体观测值更接近正态分布。

- 我们计算了 1000 个样本均值的均值和方差：

Variable	Obs	Mean	Std. Dev.	Min	Max
Mean	1,000	49.91832	2.829015	40.30648	57.83766

表 1: 1000 个样本均值的描述统计

- $\mu_{\bar{X}}$ 接近 $\mu = 50$。
- $\sigma_{\bar{X}}$ 接近 $\sigma/\sqrt{100} = 2.887$，其中 100 是样本量。

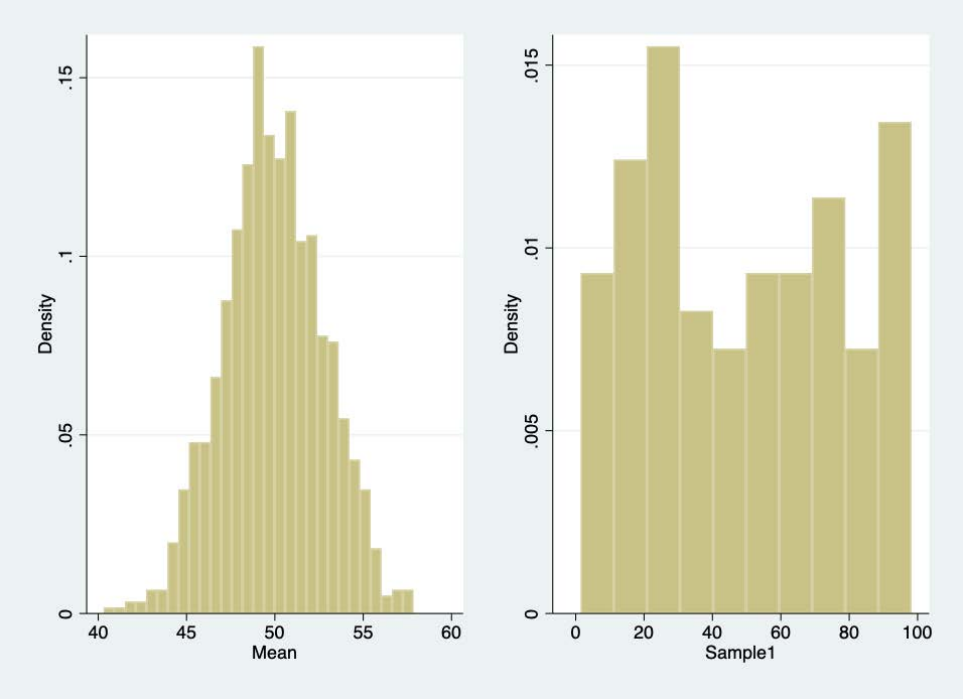

图 2: 左侧：样本均值的分布；右侧：单个样本的分布

2.3 $\bar{x}$ 的均值与标准差 (The Mean and Standard Deviation of $\bar{x}$)

样本均值是随机变量的线性组合：

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

根据随机变量均值和方差的加法规则：

$$\mu_{\bar{x}} = \frac{1}{n}(\mu_{X_1} + \mu_{X_2} + \ldots + \mu_{X_n}) = \mu$$

$$\sigma^2_{\bar{x}} = \left(\frac{1}{n}\right)^2 (\sigma^2_{X_1} + \sigma^2_{X_2} + \ldots + \sigma^2_{X_n}) = \frac{\sigma^2}{n}$$

2.4 正态总体下样本均值的抽样分布 (Sampling distribution of sample mean for normal population)

如果总体服从正态分布：$N(\mu, \sigma)$，那么 n 个独立观测值的样本均值 $\bar{x}$ 服从：

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ 分布}$$

这是基于正态分布在线性组合下的封闭性：两个或多个独立正态随机变量的线性组合也服从正态分布。

这个理论结论对任意样本量 n 都成立。

示例：苏打水瓶

每个“32 盎司”瓶中的苏打水量是一个正态分布的随机变量，均值为 32.2 盎司，标准差为 0.3 盎司。

1. 如果顾客购买一瓶，该瓶苏打水量超过 32 盎司的概率是多少？
2. 如果顾客购买一箱四瓶，四瓶平均苏打水量超过 32 盎司的概率是多少？

图示说明

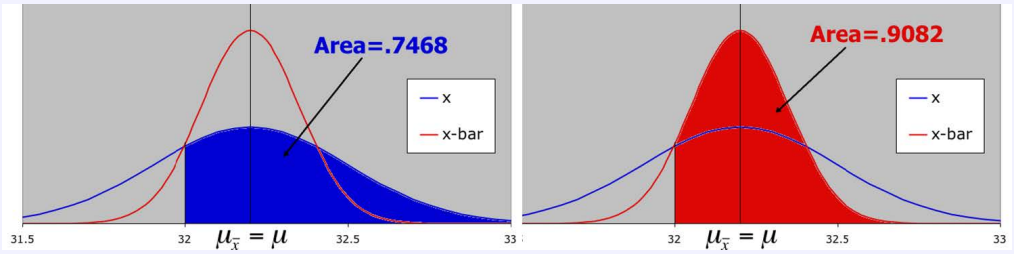

图 3: 单瓶与四瓶平均值的概率比较

练习题：投资基金的日收益率

一个投资基金声称其投资组合的日收益率服从正态分布，均值 (μ) 为 0.1%，标准差 (σ) 为 0.8%。

1. **单日：**任意一天出现亏损（负收益）的概率是多少？
   计算 $P(X < 0)$，其中 $X \sim N(0.1, 0.8)$。
2. **一周平均：**一个 5 天的交易周内，平均日收益率为负的概率是多少？
   这里我们使用抽样分布：$\bar{X} \sim N\left(0.1, \frac{0.8}{\sqrt{5}}\right)$。
   计算 $P(\bar{X} < 0)$。
3. **比较：**周平均出现负收益的概率更低，这体现了平均如何降低变异性。

3 中心极限定理 (Central Limit Theorem)

- 对于非正态总体，我们对样本均值的抽样分布的理解基于中心极限定理。
- **中心极限定理：**对于任何均值为 μ、方差有限为 σ^2 的总体，当样本量足够大时，样本均值的抽样分布近似正态。

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

中心极限定理的应用示例

- **厚尾分布 (Heavy tails)：**股票收益的分布通常是厚尾的（少数极端收益/损失）。然而，包含 50 只股票的投资组合的平均收益将近似服从正态分布。这是现代投资组合理论和风险管理的基础。
- **二元数据/分类数据 (Binary data / categorical data)：**个人的投票行为（例如，投“赞成”票）是一个二元变量。然而，一项对 1000 人进行的民意调查中，“赞成”票的样本比例近似服从正态分布。这使我们能够预测选举结果并为公众意见创建置信区间 (confidence intervals)。
- **高度偏斜的数据 (Highly skewed data)：**公司规模（按收入或员工数）的分布通常是极度右偏的（少数巨头公司，许多小公司）。然而，随机抽取的 100 家公司的平均收入将近似服从正态分布，从而能够对更广泛的经济进行推断。

3.1 中心极限定理与样本量 (Central Limit Theorem and Sample Sizes)

- 样本量越大，$\bar{X}$ 的抽样分布就越接近正态分布。
- “足够大”的定义取决于 X 的非正态程度（例如，严重偏斜、多峰、异常值）。
- 经验法则：$n > 25$。

3.2 不同样本量的抽样分布 (Sampling Distribution of Different Sample sizes)

样本量 10 与样本量 30 的抽样分布比较，总体为均匀分布 $U(0, 100)$。

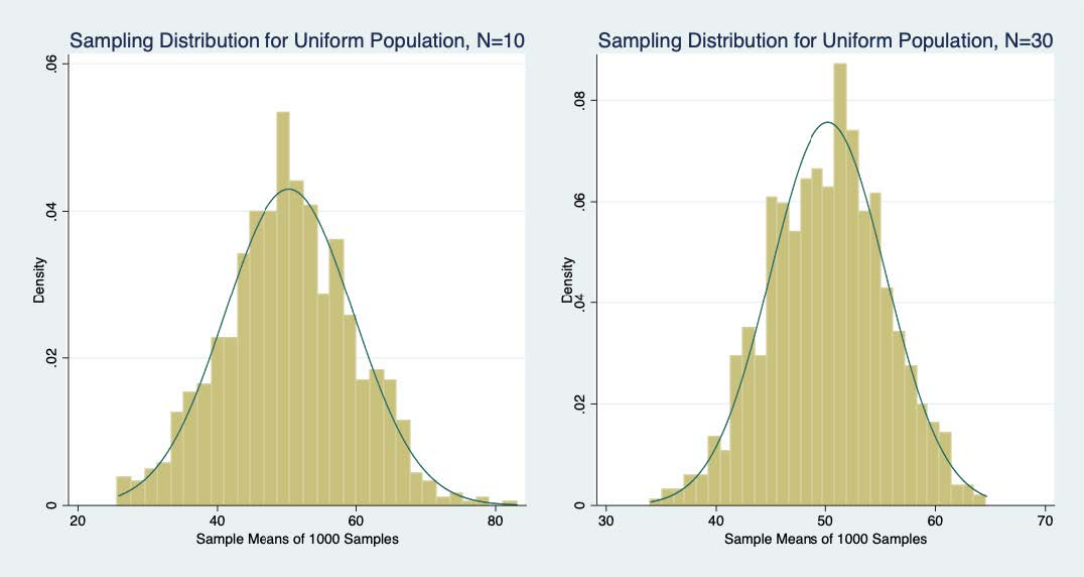

图 4: 样本量 10 与 30 下的抽样分布比较

样本均值的抽样分布何时是正态的？(When is the sampling distribution of the sample mean normal?)

- 如果总体是正态的，那么对于任何样本量，$\bar{x}$ 都服从正态分布。
- 如果总体是非正态的，那么由于中心极限定理，只有当样本量很大时 $\bar{x}$ 才近似正态分布。

示例：商学院毕业生薪资

某大学声称其商学院毕业生毕业一年后的平均周薪为 800 美元，标准差为 100 美元。一名学生调查了 25 名一年前毕业的毕业生，收集了他们的周薪。样本均值为 750 美元。

1. 如果大学的声称是正确的，那么一个 25 名毕业生的样本均值小于等于 750 美元的概率是多少？
2. 95% 的样本均值会落在哪个范围内？

解答：根据中心极限定理，样本均值的抽样分布近似服从正态分布，其参数如下：

- 总体均值：$\mu = 800$ 美元
- 总体标准差：$\sigma = 100$ 美元
- 样本量：$n = 25$
- 样本均值的标准差（标准误）：$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{25}} = \frac{100}{5} = 20$ 美元

因此，样本均值 $\bar{x}$ 的抽样分布为：

$$\bar{x} \sim N\left(\mu = 800, \sigma_{\bar{x}} = 20\right)$$

问题 1：计算 $P(\bar{x} \leq 750)$ 标准化样本均值：

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{750 - 800}{20} = \frac{-50}{20} = -2.5$$

查标准正态分布表或使用统计软件可得：

$$P(\bar{x} \leq 750) = P(Z \leq -2.5) = 0.0062$$

因此，如果大学的声称是正确的，那么一个 25 名毕业生的样本均值小于等于 750 美元的概率约为 **0.62%**（或 **0.0062**）。

问题 2：95% 的样本均值所在范围对于正态分布，95% 的数据位于均值左右各 1.96 个标准差的范围内。因此：

$$\text{下限} = \mu - 1.96 \times \sigma_{\bar{x}} = 800 - 1.96 \times 20 = 800 - 39.2 = 760.8 \text{ 美元}$$
$$\text{上限} = \mu + 1.96 \times \sigma_{\bar{x}} = 800 + 1.96 \times 20 = 800 + 39.2 = 839.2 \text{ 美元}$$

因此，95% 的样本均值会落在 **(760.8 美元, 839.2 美元)** 范围内

分析与可能误差 (Analysis and Potential Errors)

学生的调查发现样本均值为 750 美元，比大学的声称（$\mu = 800$ 美元）低了 50 美元。计算表明，如果大学的声称正确，这是一个非常不可能的事件。

可能的原因：

- 大学可能夸大了毕业生的薪资。
- 样本收集可能受到非抽样误差的影响。你认为哪些非抽样误差可能导致这种差异？
  - **覆盖不足 (Undercoverage)**：样本未能代表总体中的所有群体。
  - **无回答误差 (Non-response)**：部分被调查者没有回答，导致样本有偏。

增加样本量不仅能使样本均值的抽样分布更接近正态分布，还会减小样本均值的标准差（标准误）。

例如：对于来自 $N(50, 25)$ 的样本，下图显示了样本均值 $\bar{X}$ 的标准误随样本量 $n = 4, 9, 16, 25, 36, 49, 64, 81, 100$ 的变化。

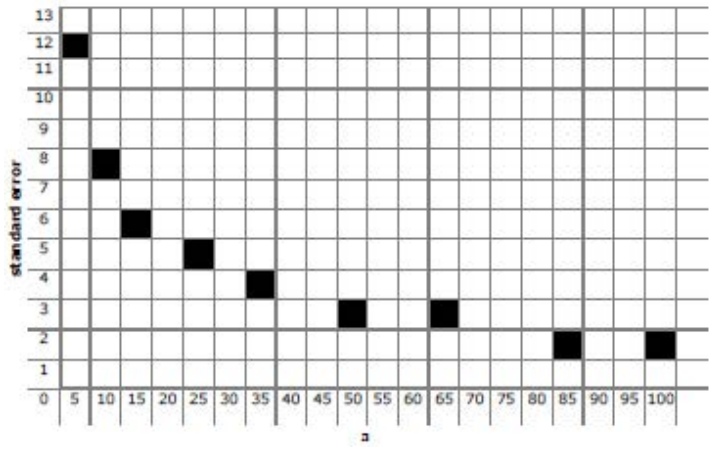

图 5: 样本量对标准误的影响

练习题：中心极限定理与样本量

从一个均值为 200、标准差为 10 的总体中随机抽取一个样本量为 100 的样本。根据中心极限定理，样本均值的近似抽样分布是什么？

1. 使用经验法则描述该样本均值的变异性。95% 的样本均值预计落在什么范围内？
2. 假设我们将样本量增加到 400。使用经验法则描述该样本均值的变异性。比较两个范围。

解答： 根据中心极限定理，对于样本量足够大的随机样本（通常 $n > 30$），样本均值的抽样分布近似服从正态分布，其参数为：

- 总体均值：$\mu = 200$
- 总体标准差：$\sigma = 10$
- 样本量：$n_1 = 100$，$n_2 = 400$

对于样本量 $n_1 = 100$： 样本均值的标准误：

$$\sigma_{\bar{x}_1} = \frac{\sigma}{\sqrt{n_1}} = \frac{10}{\sqrt{100}} = \frac{10}{10} = 1$$

样本均值的抽样分布近似为：

$$\bar{x}_1 \sim N(200, 1)$$

根据经验法则（68-95-99.7 规则）：

- 68% 的样本均值落在 $\mu \pm 1\sigma_{\bar{x}} = 200 \pm 1 = (199, 201)$

- 95% 的样本均值落在 $\mu \pm 1.96\sigma_{\bar{x}} \approx 200 \pm 1.96 = (198.04, 201.96)$
- 99.7% 的样本均值落在 $\mu \pm 3\sigma_{\bar{x}} = 200 \pm 3 = (197, 203)$

因此，95% 的样本均值预计落在 **(198.04, 201.96)** 范围内。

对于样本量 $n_2 = 400$： 样本均值的标准误：

$$\sigma_{\bar{x}_2} = \frac{\sigma}{\sqrt{n_2}} = \frac{10}{\sqrt{400}} = \frac{10}{20} = 0.5$$

样本均值的抽样分布近似为：

$$\bar{x}_2 \sim N(200, 0.5)$$

95% 的样本均值落在：

$$\mu \pm 1.96\sigma_{\bar{x}_2} = 200 \pm 1.96 \times 0.5 = 200 \pm 0.98 = (199.02, 200.98)$$

比较两个范围：

- 样本量 100 时：范围宽度 $= 201.96 - 198.04 = 3.92$
- 样本量 400 时：范围宽度 $= 200.98 - 199.02 = 1.96$

结论：当样本量从 100 增加到 400 时，样本均值的标准误从 1 减少到 0.5，减半了。相应地，95% 的样本均值范围宽度也从 3.92 减少到 1.96，同样减半。这说明增加样本量可以减少样本均值的变异性。

练习题：样本量计算 (Practice: Sample Size Calculation)

从一个 $\mu = 125$、$\sigma = 50$ 的总体中抽取 n 个观测值的随机样本。

1. 假设你希望 $\sigma_{\bar{X}} = 2$。样本量需要多大？
2. 假设你希望将样本均值的标准差减少 50%。样本量需要多大？

解答： 样本均值的标准差（标准误）公式为：

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

其中 $\sigma = 50$ 是总体标准差，n 是样本量。

问题 1：希望 $\sigma_{\bar{x}} = 2$，求 n

$$2 = \frac{50}{\sqrt{n}}$$

解方程：

$$\sqrt{n} = \frac{50}{2} = 25$$

$$n = 25^2 = 625$$

因此，要使样本均值的标准差为 2，需要样本量 $n = \mathbf{625}$。

问题 2：希望将样本均值的标准差减少 50%，求新的样本量 设原来的样本量为 n_0，原来的标准误为 $\sigma_{\bar{x}_0} = \frac{50}{\sqrt{n_0}}$。
我们希望新的标准误是原来的 50%，即：

$$\sigma_{\bar{x}_{\text{new}}} = 0.5 \times \sigma_{\bar{x}_0}$$

代入公式：

$$\frac{50}{\sqrt{n_{\text{new}}}} = 0.5 \times \frac{50}{\sqrt{n_0}}$$

两边同时除以 50：

$$\frac{1}{\sqrt{n_{\text{new}}}} = 0.5 \times \frac{1}{\sqrt{n_0}}$$

取倒数：

$$\sqrt{n_{\text{new}}} = 2 \times \sqrt{n_0}$$

两边平方：

$$n_{\text{new}} = (2\sqrt{n_0})^2 = 4n_0$$

结论：要将样本均值的标准差减少 50%，样本量需要增加到原来的 **4 倍**。
例如：

- 如果原来的样本量 $n_0 = 100$，则需要增加到 $n_{\text{new}} = 4 \times 100 = 400$
- 如果原来的样本量 $n_0 = 400$，则需要增加到 $n_{\text{new}} = 4 \times 400 = 1600$

一般原理：样本均值的标准差与样本量的平方根成反比。要将标准差减少到原来的 k 倍，样本量需要增加到原来的 $\frac{1}{k^2}$ 倍。

- 中心极限定理的更一般形式表明：许多小随机量的和或分布的分布接近于正态。
- 即使这些量不是独立的（只要它们不是高度相关）并且即使它们具有不同的分布（只要没有单个随机量大到主导其他量），这也是成立的。
- 中心极限定理解释了为什么正态分布是观测数据的常见模型。任何由许多小影响

之和构成的变量都将具有近似正态分布。

4 二项分布 (Binomial Distribution)

二项分布来源于独立重复的伯努利试验 (Bernoulli trials)。

- 伯努利试验有两个结果：成功或失败。π 是每次试验中“成功”的概率。
- 令 $X_i = 1$ 表示第 i 次试验成功，$X_i = 0$ 表示失败。
- 成功次数 S_n 是 n 个伯努利随机变量之和，服从二项分布：

$$S_n = X_1 + X_2 + \ldots + X_n$$

- 期望和方差：

$$E(S_n) = n\pi, \quad V(S_n) = n\pi(1-\pi)$$

示例：二项分布应用

1. **Pat 通过猜测回答 10 道五选一的选择题。Pat 一道题也没答对的概率是多少？**

   Pat 回答每道题时，猜对的概率为 $\pi = \frac{1}{5} = 0.2$，猜错的概率为 $1 - \pi = 0.8$。题目数量 $n = 10$。

   设 X 表示 Pat 答对的题目数量，则 $X \sim B(n = 10, \pi = 0.2)$。

   Pat 一道题也没答对的概率即 $P(X = 0)$。

   根据二项分布的概率质量函数：

$$P(X = k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

   代入 $k = 0$：

$$P(X = 0) = \binom{10}{0} (0.2)^0 (0.8)^{10} = 1 \times 1 \times (0.8)^{10}$$

   计算：

$$(0.8)^{10} = 0.1073741824$$

   因此，Pat 一道题也没答对的概率约为 **0.1074**（约 **10.74%**）。

2. **只有 20% 的急诊患者有健康保险。在 20 次急诊就诊中，超过 4 名患者有健康保险的概率是多少？**

设 X 表示 20 次急诊就诊中有健康保险的患者数量。已知 $\pi = 0.2$，$n = 20$，因此 $X \sim B(n = 20, \pi = 0.2)$。

需要计算 $P(X > 4)$，即超过 4 名患者有健康保险的概率:
$P(X > 4) = 1 - P(X \leq 4) = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)]$

使用二项分布概率公式计算各项：

$$P(X = 0) = \binom{20}{0}(0.2)^0(0.8)^{20} = 1 \times 1 \times (0.8)^{20} \approx 0.0115$$

$$P(X = 1) = \binom{20}{1}(0.2)^1(0.8)^{19} = 20 \times 0.2 \times (0.8)^{19} \approx 0.0576$$

$$P(X = 2) = \binom{20}{2}(0.2)^2(0.8)^{18} = 190 \times 0.04 \times (0.8)^{18} \approx 0.1369$$

$$P(X = 3) = \binom{20}{3}(0.2)^3(0.8)^{17} = 1140 \times 0.008 \times (0.8)^{17} \approx 0.2054$$

$$P(X = 4) = \binom{20}{4}(0.2)^4(0.8)^{16} = 4845 \times 0.0016 \times (0.8)^{16} \approx 0.2182$$

求和：

$$P(X \leq 4) \approx 0.0115 + 0.0576 + 0.1369 + 0.2054 + 0.2182 = 0.6296$$

因此：

$$P(X > 4) = 1 - 0.6296 = 0.3704$$

所以，在 20 次急诊就诊中，超过 4 名患者有健康保险的概率约为 **0.3704**（约 **37.04%**）。

注：也可使用统计软件或二项分布表直接计算累积概率。

3. **如果 $\pi = 0.3$，二项分布的形状如何？如果 $\pi = 0.9$ 呢？**

二项分布的形状取决于成功概率 π 和试验次数 n。

对于固定的 n：

- 当 $\pi < 0.5$ 时，二项分布右偏（正偏）
- 当 $\pi = 0.5$ 时，二项分布对称
- 当 $\pi > 0.5$ 时，二项分布左偏（负偏）

情况 1：$\pi = 0.3$（小于 0.5）

- 分布形状：右偏（正偏）

- 峰值出现在 $k \approx n\pi = 0.3n$ 附近
- 尾部向右（较大 k 值方向）延伸

情况 2：$\pi = 0.9$（大于 0.5）

- 分布形状：左偏（负偏）
- 峰值出现在 $k \approx n\pi = 0.9n$ 附近
- 尾部向左（较小 k 值方向）延伸

示例说明：假设 $n = 10$

- 当 $\pi = 0.3$ 时，最可能成功次数约为 3 次，概率分布向右倾斜
- 当 $\pi = 0.9$ 时，最可能成功次数约为 9 次，概率分布向左倾斜

注意：当 n 很大时，无论 π 是多少，二项分布都近似对称（中心极限定理），且可以用正态分布近似。

4.1 二项分布的形态 (Shape of Binomial Distribution)

- 当 $\pi < 0.5$ 时，二项分布右偏；当 $\pi > 0.5$ 时，二项分布左偏。
- 然而，当 n 变大时，二项分布变得更对称。

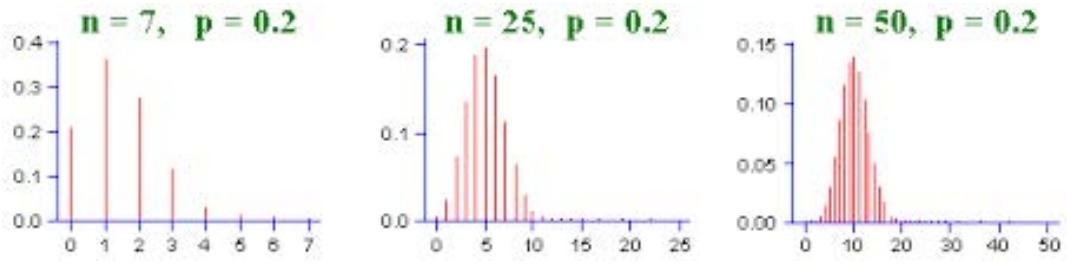

图 6: 不同 n 下二项分布的形态（$p = 0.2$）

4.2 二项分布的正态近似 (Normal Approximation of Binomial Distributions)

- S_n 是 n 个相互独立的随机变量之和。
- 中心极限定理表明，S_n 的分布函数可以用正态密度函数很好地近似。
- 当 n 足够大时，S_n 的分布近似服从 $N(n\pi, \sqrt{n\pi(1-\pi)})$。

5 样本比例 (Sample Proportions)

样本比例 p 与样本中“成功”次数 S_n 之间存在重要联系：

$$p = \frac{\text{样本中成功次数}}{\text{样本量}} = \frac{S_n}{n}$$

样本比例的均值和方差：

$$\mu_p = \frac{\mu_{S_n}}{n} = \frac{n\pi}{n} = \pi$$

$$\sigma_p^2 = \frac{\sigma_{S_n}^2}{n^2} = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$$

示例：大学生对在线购物体验的满意度

随机抽取 2500 名大学生，询问他们是否同意” 我对我的在线购物体验非常满意” 这一说法。假设所有大学生中有 60% 同意这一说法。样本中同意比例至少为 58% 的概率是多少？

解答过程：

已知：总体比例 $\pi = 0.60$，样本量 $n = 2500$，样本比例 p。
根据中心极限定理，当样本量足够大时，样本比例 p 的抽样分布近似正态分布。

1. **检查正态近似条件：**

$$n\pi = 2500 \times 0.60 = 1500 \geq 10$$

$$n(1-\pi) = 2500 \times 0.40 = 1000 \geq 10$$

两个条件都满足，因此可以使用正态近似。

2. **计算样本比例的抽样分布参数：**样本比例的均值：$\mu_p = \pi = 0.60$ 样本比例的标准差：$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.60\times0.40}{2500}} = \sqrt{\frac{0.24}{2500}} = \sqrt{0.000096} = 0.0098$
因此，$p \sim N(0.60, 0.0098)$

3. **计算** $P(p \geq 0.58)$：首先进行标准化：

$$z = \frac{p - \mu_p}{\sigma_p} = \frac{0.58 - 0.60}{0.0098} = \frac{-0.02}{0.0098} \approx -2.04$$

所以：

$$P(p \geq 0.58) = P(Z \geq -2.04)$$

由于正态分布的对称性：

$$P(Z \geq -2.04) = 1 - P(Z < -2.04) = 1 - 0.0207 = 0.9793$$

4. **结果：**样本中同意比例至少为 58% 的概率约为 **0.9793**（约 97.93%）。

这个高概率表明，从同意率为 60% 的总体中抽取 2500 名大学生的样本，其同意比例至少为 58% 的可能性非常大。如果实际观测到的样本比例显著低于 58%，可能需要重新考虑总体比例的假设。

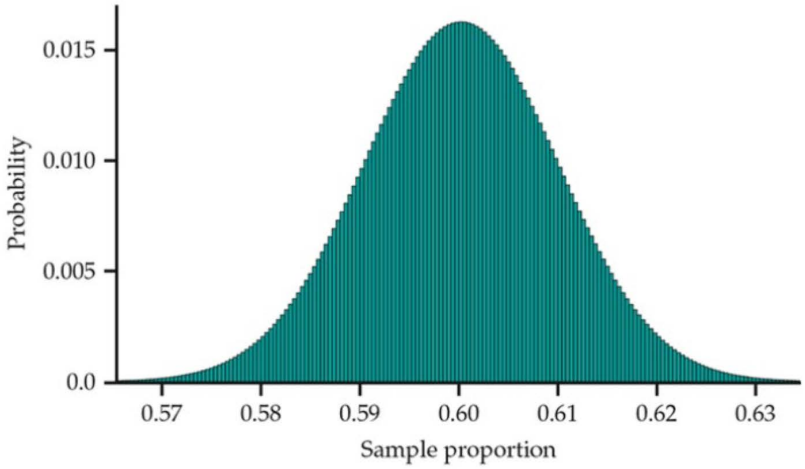

图 7: 基于二项分布 $B(2500, 0.6)$ 的样本比例 p 的概率分布

6 样本比例的抽样分布 (Sampling Distribution for Sample Proportions)

当 n 很大时，样本比例的抽样分布近似正态：

$$p \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

正态性来源于当 n 很大时二项随机变量 X 的正态近似。作为经验法则，当 n 足够大，使得

$$n\pi \geq 10 \quad \text{和} \quad n(1-\pi) \geq 10$$

时，我们使用正态近似。

示例计算

在购物体验示例中，我们想计算 $P(p > 0.58)$。

1. **检查 $n\pi$ 和 $n(1-\pi)$ 是否都大于 10。若是，则 p 近似正态分布。**

检查正态近似条件：

$$n\pi = 2500 \times 0.60 = 1500 \geq 10$$
$$n(1-\pi) = 2500 \times 0.40 = 1000 \geq 10$$

条件满足，可以使用正态近似。

2. **基于正态近似求概率。**

样本比例的抽样分布：$p \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}) = N(0.60, \sqrt{\frac{0.60\times0.40}{2500}})$

计算标准差：

$$\sigma_p = \sqrt{\frac{0.24}{2500}} = \sqrt{0.000096} = 0.0098$$

标准化：

$$z = \frac{0.58 - 0.60}{0.0098} = \frac{-0.02}{0.0098} \approx -2.04$$

查标准正态分布表：

$$P(Z > -2.04) = 1 - P(Z \leq -2.04) = 1 - 0.0207 = 0.9793$$

精确二项分布计算（验证）：使用二项分布 $B(2500, 0.6)$ 直接计算：

$$P(p > 0.58) = P(X > 2500 \times 0.58) = P(X > 1450)$$

其中 X 是同意的人数。

精确计算较为复杂，但正态近似与精确值非常接近。

结果： $P(p > 0.58) \approx 0.9793$

注意：这里计算的是 $P(p > 0.58)$，不包括等于 0.58 的情况。如果题目要求 $P(p \geq 0.58)$，由于连续分布中单点概率为 0，两者相等。

7 泊松分布 (Poisson Distribution)

- 泊松分布是计数数据的主要模型。
- 泊松分布描述在特定时间段或空间区域内事件发生的次数。
- 这些计数可以是 0, 1, 2, 3, ……无限。
- 示例：一小时内到达服务站的汽车数量、一匹布上的瑕疵数量。

7.1 泊松分布的特性 (Properties of Poisson Distribution)

- 发生在两个不重叠的度量单位中的成功次数是独立的。
- 成功发生在一个度量单位中的概率对于所有相同大小的单位是相同的，并且与单位的大小成比例。
- 在一个非常小的区间内，发生多于一个事件的概率可以忽略不计。换句话说，事件一次发生一个。

7.2 泊松分布的定义 (Definition of Poisson Distribution)

泊松分布由一个参数 λ 定义：在定义区间内的平均成功次数。X 的可能取值为整数 0, 1, 2, 3, ……对于任意整数 k，泊松变量取该值的概率为：

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

泊松随机变量的均值和标准差为：

$$E(X) = V(X) = \lambda$$

7.3 泊松分布的正态近似 (Normal Approximation of Poisson Distribution)

泊松分布是右偏分布。当 λ 变大时，分布变得更对称。当 λ 足够大时（经验法则是 $\lambda > 20$），泊松分布可以用正态分布 $N(\lambda, \sqrt{\lambda})$ 近似。

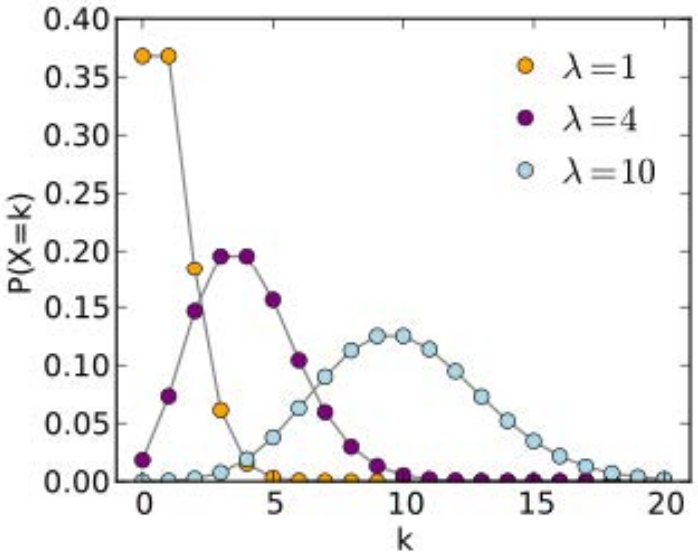

图 8: 不同 λ 下泊松分布的形态及正态近似

示例：泊松分布应用

1. **假设学校 Wi-Fi 中断次数变化但服从泊松分布，平均每天 3.7 次。明天出现不超过两次中断的概率是多少？**

   设 X 表示每天 Wi-Fi 中断次数，则 $X \sim \text{Poisson}(\lambda = 3.7)$。

   需要计算 $P(X \leq 2)$，即：

   $$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

   使用泊松分布概率质量函数：

   $$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

   计算：

   $$P(X = 0) = \frac{e^{-3.7} \times 3.7^0}{0!} = e^{-3.7} \approx 0.0247$$
   $$P(X = 1) = \frac{e^{-3.7} \times 3.7^1}{1!} = 3.7 \times e^{-3.7} \approx 0.0915$$
   $$P(X = 2) = \frac{e^{-3.7} \times 3.7^2}{2!} = \frac{13.69 \times e^{-3.7}}{2} \approx 0.1693$$

   求和：

   $$P(X \leq 2) \approx 0.0247 + 0.0915 + 0.1693 = 0.2855$$

   因此，明天出现不超过两次中断的概率约为 **0.2855**（约 28.55%）。

2. **一家高频交易公司使用泊松分布（均值 $\lambda = 4.2$）模拟每小时股票的大幅波动“价格尖峰”次数。在给定小时内观察到超过 7 次此类尖峰的概率是多少？**

   设 Y 表示每小时价格尖峰次数，则 $Y \sim \text{Poisson}(\lambda = 4.2)$。

   需要计算 $P(Y > 7)$，即：

   $$P(Y > 7) = 1 - P(Y \leq 7)$$

   首先计算 $P(Y \leq 7)$，即计算 $P(Y = 0)$ 到 $P(Y = 7)$ 的和。

使用泊松分布公式：

$$P(Y=0)=\frac{e^{-4.2}\times 4.2^0}{0!}=e^{-4.2}\approx 0.0150$$
$$P(Y=1)=\frac{e^{-4.2}\times 4.2^1}{1!}=4.2e^{-4.2}\approx 0.0630$$
$$P(Y=2)=\frac{e^{-4.2}\times 4.2^2}{2!}=\frac{17.64\times e^{-4.2}}{2}\approx 0.1323$$
$$P(Y=3)=\frac{e^{-4.2}\times 4.2^3}{3!}=\frac{74.088\times e^{-4.2}}{6}\approx 0.1852$$
$$P(Y=4)=\frac{e^{-4.2}\times 4.2^4}{4!}=\frac{311.1696\times e^{-4.2}}{24}\approx 0.1945$$
$$P(Y=5)=\frac{e^{-4.2}\times 4.2^5}{5!}=\frac{1306.9123\times e^{-4.2}}{120}\approx 0.1634$$
$$P(Y=6)=\frac{e^{-4.2}\times 4.2^6}{6!}=\frac{5489.0317\times e^{-4.2}}{720}\approx 0.1144$$
$$P(Y=7)=\frac{e^{-4.2}\times 4.2^7}{7!}=\frac{23053.9331\times e^{-4.2}}{5040}\approx 0.0686$$

求和：

$$P(Y\le 7)\approx 0.9364$$

因此：

$$P(Y>7)=1-0.9364=0.0636$$

所以，在给定小时内观察到超过 7 次价格尖峰的概率约为 **0.0636**（约 6.36%）。

可选的正态近似：

由于 $\lambda=4.2<20$，正态近似可能不够精确，但可以作为验证：正态近似：$Y\approx N(\lambda,\sqrt{\lambda})=N(4.2,\sqrt{4.2})\approx N(4.2,2.049)$。

标准化：

$$z=\frac{7.5-4.2}{2.049}\approx 1.61\quad\text{(使用连续性校正)}$$
$$P(Y>7)\approx P(Z>1.61)=1-0.9463=0.0537$$

正态近似结果为 0.0537，与精确值 0.0636 有一定差距，说明对于较小的 λ，正态近似不够精确。

8 历史注记 (Historical Remarks)

二项分布的正态近似是中心极限定理的最早版本。它首先由亚伯拉罕·棣莫弗 (Abraham de Moivre) 证明，并出现在他的著作《机会的学说》(*The Doctrine of Chances*)

（1718 年首次出版）中。棣莫弗因新教背景在法国被监禁了 18 至 21 岁，释放后他离开法国前往英格兰，在那里他担任贵族子弟的家庭教师，并接触到了牛顿 (Newton) 的《自然哲学的数学原理》(*Principia Mathematica*)，开始学习现代数学。他通过为赌徒计算赔率开始了他的概率工作。完整的证明和关于棣莫弗生平的有趣讨论可以在 F. N. David 的《Games, Gods and Gambling》一书中找到。

Summary

- **抽样误差与非抽样误差 (Sampling and Non-sampling Errors)：**
  - **抽样误差 (Sampling error)**：由于样本与总体之间的随机差异导致，无法完全消除但可通过增加样本量减少，是统计分析中的固有变异性
  - **非抽样误差 (Non-sampling error)**：由数据收集过程中的缺陷导致，如测量错误、无回答偏差等，理论上可以避免但难以通过增加样本量消除
- **偏倚与变异性 (Bias and Variability)：**
  - **偏倚 (Bias)**：系统性地偏离真实值，由非抽样误差引起
  - **变异性 (Variability)**：样本统计量的波动程度，由抽样误差引起
  - **理想估计量 (Ideal estimator)**：低偏倚、低变异性
- **抽样分布 (Sampling Distributions)：**
  - 样本统计量的概率分布，反映抽样误差的分布特征
  - 是统计推断的理论基础，连接样本统计量与总体参数
  - 提供量化抽样不确定性的方法
- **样本均值的抽样分布 (Sampling Distribution of the Sample Mean)：**
  - 对于任意总体，样本均值的期望等于总体均值：$E(\bar{x}) = \mu$
  - 样本均值的标准差（标准误）：$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$，随样本量增加而减小
  - **正态总体**：无论样本量大小，$\bar{x}$ 精确服从正态分布
  - **非正态总体**：当样本量足够大时，$\bar{x}$ 近似服从正态分布（中心极限定理）
- **中心极限定理 (Central Limit Theorem, CLT)：**
  - 对于任何均值为 μ、方差有限为 σ^2 的总体，当样本量 n 足够大时，样本均值的抽样分布近似正态分布：$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
  - **应用条件**：经验上 $n > 25$ 或 $n > 30$ 通常足够，但取决于总体的非正态程度
  - **重要意义**：解释正态分布的普遍性，是参数统计推断的理论基础
- **二项分布与样本比例 (Binomial Distribution and Sample Proportions)：**
  - **二项分布 (Binomial distribution)**：描述 n 次独立伯努利试验中成功次数的分布，参数为 n 和 π

  - **形态特征**：当 $\pi = 0.5$ 时对称，$\pi < 0.5$ 时右偏，$\pi > 0.5$ 时左偏；n 增大时趋于对称
  - **样本比例 (Sample proportion)**：$p = S_n/n$，其中 S_n 为成功次数
  - **样本比例的抽样分布**：当 n 足够大（$n\pi \geq 10$ 且 $n(1-\pi) \geq 10$）时，$p \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

- **泊松分布 (Poisson Distribution)**：
  - 描述单位时间或空间内随机事件发生次数的分布，参数为 λ（平均发生率）
  - **特征**：期望和方差均为 λ：$E(X) = V(X) = \lambda$
  - **正态近似**：当 $\lambda > 20$ 时，泊松分布可用正态分布 $N(\lambda, \sqrt{\lambda})$ 近似

- **关键公式与计算 (Key Formulas and Calculations)**：
  - 样本均值标准误：$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
  - 样本比例标准误：$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$
  - 样本量计算：$n = \left(\frac{\sigma}{\sigma_{\bar{x}}}\right)^2$
  - 95% 样本均值范围：$\mu \pm 1.96\sigma_{\bar{x}}$

- **实际应用与解释 (Practical Applications and Interpretation)**：
  - 使用抽样分布评估样本结果的合理性和异常性
  - 根据已知的总体参数计算特定样本结果的概率
  - 解释中心极限定理在实际数据分析中的重要性
  - 区分统计显著性与实际重要性