STAT243 Lecture 9.2 Design of Simulation Studies

1 模拟研究的基本步骤 (Basic Steps of a Simulation Study)

1. **定义单次实验 (Specify an Experiment)**
   明确：
   - 样本量 n
   - 数据分布 (e.g., Normal, t, Poisson)
   - 参数值 (e.g., 均值, 方差, 回归系数)
   - 感兴趣的统计量（估计量、检验统计量等）
   - 数据生成机制 (DGP: Data Generating Process)

Example>

生成 $n = 100$ 个 $X_i \sim N(0,1)$，$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$，
其中 $\varepsilon_i \sim N(0,1)$。

---

2. **确定要变化的输入因子 (Identify Factors to Vary)**
   - 样本量 (n)
   - 参数 (β_1, σ^2)
   - 分布类型 (Normal / t / skewed)
   - 相关结构 (independent vs correlated)
     每种组合形成一个 **scenario (情景)**。

---

3. **编写实验函数 (Write Simulation Code)**
   - 该函数接收上述输入参数；
   - 返回感兴趣的统计量（如估计偏差、方差等）。

Python

```python
def simulate_once(n, beta, dist='normal'):
    X = np.random.normal(size=n)
    if dist == 't':
        eps = np.random.standard_t(df=3, size=n)
    else:
        eps = np.random.normal(size=n)
    Y = beta * X + eps
    est = np.sum(X * Y) / np.sum(X**2)
    return est
```

---

4. **重复实验 (Repeat the Experiment m Times)**
   - 对每个 scenario 重复模拟 m 次。
   - 这可高度并行化（并行于 scenario 与 replicate 两个维度）。

Python

```python
results = [simulate_once(100, 1.0) for _ in range(1000)]
```

---

5. **总结与不确定性量化 (Summarize Results & Quantify Uncertainty)**
   - 计算均值、标准差、偏差、MSE 等；
   - 模拟标准误差：

$$SE_{sim} = \frac{s}{\sqrt{m}}$$

   - 若 SE_{sim} 远小于效应量，可忽略模拟误差。

---

6. **可视化与报告 (Visualization and Reporting)**
   - 表格呈现 bias、RMSE、coverage；
   - 图形化呈现 (箱线图、误差条、曲线对比)。

---

2 各种重要考量 (Design Considerations)

1. **效率与可重现性**
   - 代码应结构化、模块化；
   - 记录随机种子以确保可复现。
2. **与真实数据相似的结构**
   - 模拟分布、参数、依赖性应与目标问题相符；
   - 可引入异常值、随机效应等以测试鲁棒性。
3. **控制随机变异 (Paired Designs)**
   - 在比较两种方法时，**使用相同的模拟数据集**，
     即固定随机数种子，使两方法共享相同噪声。
   - 分析上可使用配对差值以降低方差。
4. **控制随机数的一致性**
   - 例如，在比较 t 分布与 Normal 分布数据时，可以通过相同的累积分布概率匹配：

Python

```python
from scipy.stats import t, norm
import numpy as np, matplotlib.pyplot as plt

devs = np.random.normal(size=100)
tdevs = t.ppf(norm.cdf(devs), df=1)

plt.scatter(devs, tdevs)
plt.xlabel('Normal Deviates')
plt.ylabel('t Deviates (df=1)')
plt.plot([min(devs), max(devs)], [min(devs), max(devs)], color='red')
plt.show()
```

> ⚠ **Remark**
>
> 这样可以使两组数据“成对对应”，减少随机误差的影响。

2.1 重复次数 m 的选择 (Choosing m)

- 理论上可用基本功效 (power) 计算决定 m；
- 实践中常采用**顺序方式 (sequential)**：
  - 逐步增加 m，直到模拟结果的精度达到可接受水平；

- 若 $s/\sqrt{m}$ 已小于差异效应量，可停止。

---

3 实验设计思想在模拟中的应用 (Optional)

3.1 多因素输入的情景 (Multiple Input Variables)

- 通常需评估多个输入变量对输出结果的影响；
- 避免“一次只变一个变量”的低效方式。

3.2 全因子设计 (Full Factorial Design)

- 若输入较少，可离散化为少量水平（levels），进行全因子设计：
  k 个因子，每个有 L 个水平 $\rightarrow L^k$ 个情景。

> **Example**
>
> 示例：3 个输入 × 3 个水平 = $3^3 = 27$ 个组合。

- 适用于输入数量较少的情况。

3.3 分数因子设计 (Fractional Factorial Design)

- 当输入或水平较多时，无法执行全因子设计；
- 分数因子设计通过选择部分组合来保证：
  - 平衡性；
  - 可估计主效应与部分交互项；
  - 高阶交互项与低阶项混叠（aliasing）。

> **Remark**
>
> 目标：在计算资源有限的前提下，最大化可解释性。

3.4 Latin Hypercube Sampling (LHS)

- 适用于输入维度很高的情况；
- 在多维输入空间内均匀采样，保证覆盖性。

步骤：

1. 假设每个输入变量 $X_j \sim U(0,1)$；
2. 将区间 [0,1] 均分为 m 个区间；
3. 在每个区间内随机采样一个点；
4. 对每个变量独立地随机排列区间顺序；
5. 组合形成 m 个输入点。

> **Remark**

这样可在有限样本下均匀探索输入空间。
分析时关注主效应与一、二阶交互作用。

3.5 结果解释建议

- 重点放在**方差分解与效应大小 (effect magnitude)**，而非显著性检验；
- 在模拟研究中，“零假设成立”通常不具意义。