Tutorial 10：实验方法与验证

题目

1. [单选题] 为什么单次最好结果通常不足以形成强证据？A. 因为系统实验只看图表颜色 B. 因为它可能只是偶然命中有利条件 C. 因为平均值永远没意义 D. 因为 benchmark 都不能用

---

2. [判断题] 多次稳定趋势通常比单次峰值更具说服力。

---

3. [多选题] 下列哪些信息属于实验结论成立的前提？A. baseline B. workload 说明 C. 变量控制方式 D. 标题字号

---

4. [简答题] 为什么 warmup 不能被随意省略？

5. [计算题] 某实验共运行 5 次，吞吐结果分别为 98、100、101、99、102。请计算平均吞吐，并判断这一组结果是否体现出较稳定的趋势。

---

6. [判断题] 只要结果表格足够整齐，即使日志与解释链对不上，实验仍然可信。

---

7. [简答题] 为什么实验开始前就应写清楚失败条件？

8. [伪代码题] 请写出一个简短伪代码：若结果未超过 baseline 且成本上升，则返回 `reject_hypothesis`；否则返回 `keep_testing`。

9. [多选题] 下列哪些因素会直接影响实验可信度？A. 随机种子 B. 批次设置 C. 输入分布 D. 目录颜色主题

---

10. [简答题] 什么样的 ablation 才真正具有解释力？

参考答案

1. B。单次最好结果可能只是偶然命中有利条件。
2. 对。稳定趋势更能说明结果具有可复现性。
3. A、B、C。这些都是实验结论成立的前提。
4. 因为 warmup 会显著影响稳态阶段的观测结果。
5. 平均吞吐 = (98 + 100 + 101 + 99 + 102) / 5 = 100。该组结果波动较小，可视为较稳定。
6. 错。日志、变量控制和解释链必须能够互相支撑。
7. 因为可否定性是实验可信的前提。
8. 示例伪代码：

```
function evaluate(result, baseline, cost_up):
     if result <= baseline and cost_up == true:
          return "reject_hypothesis"
     return "keep_testing"
```

9. A、B、C。这些因素都会直接改变结果口径。
10. 只有在能够隔离关键因素时，ablation 才真正具有解释力。