Collecting Data

AP Statistics

Can We Trust the Data We Collected?

A conclusion is only as good as the data behind it. **How** data are collected decides what you may conclude –whether you can generalize to a **population** 总体, and whether you can claim **cause and effect**. Poorly collected data can be worse than none.

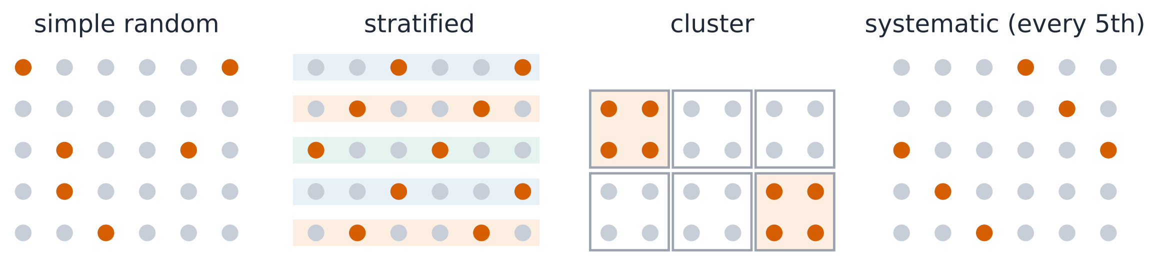
Observational Studies and Experiments

- In an **observational study** 观察性研究 you measure individuals without trying to influence them. It can show **association**, but not causation, because lurking variables may explain the link.
- In an **experiment** 实验 you deliberately impose a **treatment** 处理 and compare responses. A well-designed experiment **can** establish cause and effect.

Random Sampling

To learn about a population you take a **sample** 样本. **Random sampling** 随机抽样 removes selection **bias** 偏差 and lets you generalize. Common designs:

- **Simple random sample (SRS)** 简单随机样本: every group of the chosen size is equally likely.
- **Stratified** 分层: split the population into similar strata, then sample within each.
- **Cluster** 整群: split into clusters, randomly choose whole clusters.
- **Systematic** 系统: pick every kth individual from a random start.



Four random sampling designs: who gets selected, and how

A **convenience sample** 方便样本 or **voluntary response** sample is **not** random and is biased.

Worked example. To survey a school, an administrator lists all students by grade and randomly selects 20 from each grade. This is a **stratified** sample –the grades are the strata –which guarantees every grade is represented, unlike an SRS that might by chance draw few from one grade.

When Sampling Goes Wrong

Bias makes estimates systematically miss the truth:

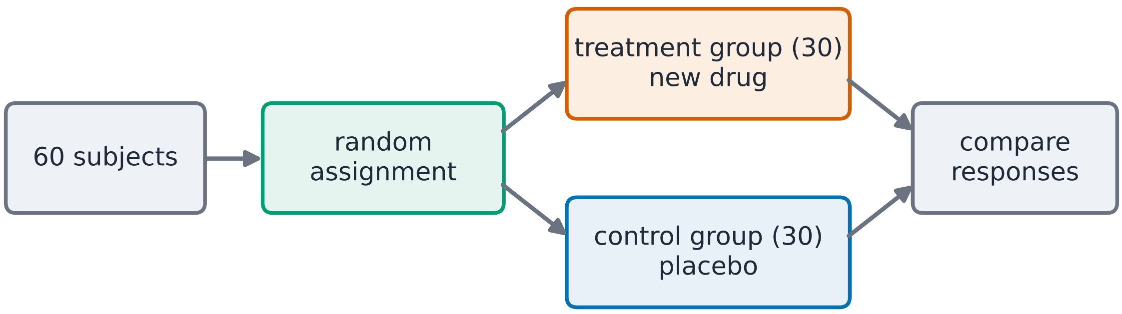
- **Undercoverage** 覆盖不足: some groups are left out of the sampling frame.
- **Nonresponse** 无回应: selected people do not answer.
- **Response bias** 回应偏差: people answer inaccurately (bad wording, sensitive topics).

Bias is about a **consistent** error in one direction –increasing the sample size does not fix it.

Designing an Experiment

Good experiments follow three principles:

- **Comparison** with a **control group** 对照组 (often a **placebo** 安慰剂).
- **Random assignment** 随机分配 of subjects to treatments, to balance out other variables.
- **Replication** 重复: enough subjects per treatment to see a real effect.



A completely randomized experiment compares a treatment group with a control group

Confounding 混杂 occurs when another variable is tied to the treatment so their effects cannot be separated; random assignment guards against it. **Blinding** 盲法 hides who is getting which treatment to prevent expectation effects: in a **single-blind** 单盲 study only one side is kept unaware (usually the subjects, or only the people assessing the result), while in a **double-blind** 双盲 study **neither** the subjects nor the researchers who interact with them know, blocking both the placebo effect and biased assessment. **Blocking** 区组 groups similar subjects and randomizes within each block to reduce variability.

Choosing the Right Design

Match the design to the goal: use a **completely randomized design** for uniform subjects; a **randomized block design** when a known variable (sex, age) affects the response; a **matched-pairs** design when each subject can serve as its own control. State how you would carry out the randomization.

What an Experiment Lets You Conclude

Two questions decide the scope of a conclusion:

- **Random assignment** used? Then a significant difference can be attributed to the treatment (**causation**) –for these subjects.
- **Random sampling** from a population? Then results **generalize** to that population.

Only an experiment with random assignment supports a cause-and-effect claim; only random sampling supports generalization. Say exactly which you have.

Worked example. Researchers randomly assign 100 **volunteers** to a new drug or a placebo, and the drug group improves significantly more. Because of the **random assignment**, the improvement can be attributed to the drug (causation) –but because the subjects were **not randomly sampled**, the conclusion applies only to these volunteers and does not automatically generalize to everyone.

Exam tips

- Distinguish an **observational study** (finds association) from an **experiment** (can show causation).
- Good sampling is **random** (SRS, stratified, cluster) —beware bias (voluntary response, undercoverage, nonresponse).
- Good experiments use **control, randomization, and replication**; blocking handles a known nuisance variable.
- Only a randomized experiment supports a cause-and-effect conclusion.
- Name the population, sample, and any confounding clearly.