Collecting Data

AP Statistics ▪ Unit 3

AP Statistics

What we will cover

1. Trusting the data
2. Observational vs experiment
3. Random sampling
4. When sampling goes wrong
5. Designing experiments
6. Scope of inference

1 Trust

Trusting the data

The *method* of collection decides what we may claim. We study a **sample** 样本 to learn about a **population** 总体.

- **bias** 偏差 tilts a sample systematically off the truth

> Good data comes first – a clever calculation cannot rescue a badly collected sample.

Clinical trial lab

2 Observational

Observational vs experiment

- an **observational study** 观察性研究 watches – shows only **association** 关联
- an **experiment** 实验 applies a **treatment** 处理 – can prove **cause** 因果关系

> A **confounding variable** 混杂变量 tied to both explanatory and response is why observation can't claim cause.

Random dice

3 Sampling

Random sampling

Let chance, not the researcher, choose – four random designs:

- **SRS** 简单随机抽样: every group of n equally likely
- **Stratified** 分层: SRS within similar **strata** 层
- **Cluster** 整群: pick whole **clusters** 群
- **Systematic** 系统: every k-th individual

Sampling bias crowd

> Random selection lets us **generalize** 推广 the sample result to the whole population.

4 Wrong

When sampling goes wrong

Even a "random" plan can bias:

- **undercoverage** 覆盖不足: groups left out
- **voluntary response** 自愿回应: only strong feelings
- **nonresponse** 无回应, **response bias** 回应偏差, wording

> A bigger sample does not fix bias – it just makes a bigger biased sample. Method beats size.

SRS of 1000 beats it

5 Experiments

Designing experiments

Three principles:

- **control** 控制: hold other variables equal
- **randomization** 随机化: balances unknowns
- **replication** 重复: many units per treatment

> A **placebo** 安慰剂 + **double-blind** 双盲 design separate the real effect from belief.

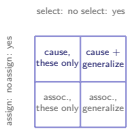
6 Scope

Scope of inference

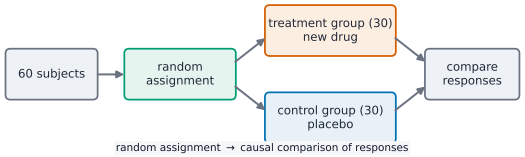
Two uses of randomness set what you may claim:

- random **assignment** 随机分配
  ⇒ cause
- random **selection** 随机选择
  ⇒ generalize

> **Blocking** 区组化 / **matched pairs** 配对设计 remove a known source of variability.



Designing an Experiment



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

Designing an experiment

confounding 混杂	another variable is tied to the treatment, so their effects cannot be separated
random assignment	the cure – it balances *all* other variables, known and unknown, across the groups
a control group	gives something to compare against; a placebo controls for expectation
single-blind 单盲	the subject does not know which treatment they received
double-blind	neither the subject nor the person measuring knows
replication	enough subjects per group that chance differences are small

> Random sampling lets you generalise to the population; random assignment lets you claim causation. The two do different jobs, and the exam asks which you have.

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.

> State how you would carry out the randomization.

Random Sampling, continued

SRS: equal chance · stratified: sample within groups · cluster: take whole clusters · systematic: every kth

> Four random sampling designs: who gets selected, and how

The sampling methods, and the biases

a simple random sample 简单随机样本	every group of n individuals is equally likely – the standard against which the others are judged
stratified	split into groups, then sample within each – more precise when the groups differ
cluster	sample whole groups – cheaper, and less precise
a convenience sample 方便样本	whoever is easy to reach – biased, and never acceptable
response bias 回应偏差	people answer inaccurately – bad wording, or a sensitive topic
non-response bias	the people who do not reply differ from those who do

An **observational study** 观察性研究 measures without intervening, so it can show association only. Only a randomised experiment supports a causal claim.

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.

7 Recap

Worked example – spotting the bias

A school emails a survey to all students; 80 of 1000 reply, and 70 say they want a later start. Identify the flaw.

- Everyone could reply, so this is a voluntary response sample, not a random one.
- Only students who care strongly bother – and those wanting a later start care most.
- So the 87.5% is biased upward; you cannot generalise to all 1000.
- Fix: take a simple random sample and follow up non-responders $\Rightarrow$ a representative estimate.

Name the bias and say which direction it pushes the estimate – that is where the marks are. A big sample does not fix bias; only good selection does.

Unit 3 – key takeaways

1 Study type
observation shows association; experiment proves cause

2 Sampling
SRS, stratified, cluster, systematic; bias $\neq$ fixed by size

3 Experiments
control, randomize, replicate; placebo + blinding

4 Scope
assignment $\rightarrow$ cause; selection $\rightarrow$ generalize

Check yourself

1. Which study type can prove cause and effect?
2. What sampling method takes an SRS within each stratum?
3. Does a larger sample remove bias?
4. What must be random to generalize to a population?

cause	both
assoc.	gen.

Answers **1.** a randomized experiment **2.** stratified **3.** no **4.** random selection