

UNIT 1B — Collecting Data

1.10 The Investigative Question Revisited & Data Collection

1.11 Random Sampling

1.12 Potential Problems with Sampling

1.13 Experimental Design

Everything in 1A was about describing data somebody already had. Unit 1B is about where the data comes from — and why that decides what you are allowed to say.



Where We Left Off

In Unit 1A you learned the vocabulary of a statistical study. We pick it back up at **Step 2**.

The Statistical Problem-Solving Process

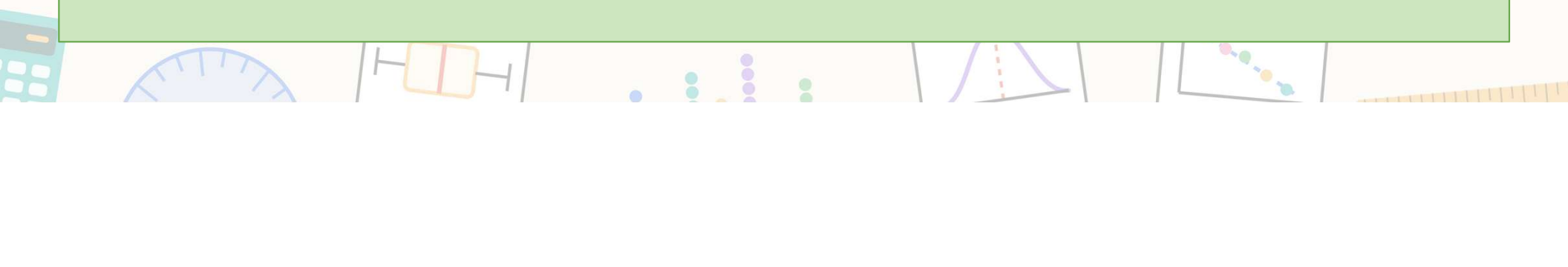
1. **Ask questions** — pose a valid investigative question.
2. **Collect data** — design and carry out a plan for getting data.
3. **Analyze data** — organize, display, summarize.
4. **Interpret results** — answer the question in context.

A **population** is the entire group we want information about.

A **sample** is the subset we actually collect data from.

A **census** records information from **all** items or individuals in the population — usually too expensive, often impossible, sometimes destructive, so we sample instead.

A **parameter** describes a population; a **statistic** describes a sample.



Planning Ahead to Answer a Question

1. What is the question you want to answer?
2. Who or what will you collect data from?
3. What variable(s) will you measure?
4. How will you collect the data?
5. What will you do with the data once you have it?
6. Who can you generalize your results to?

Question 4 has only TWO answers on the AP® exam:

Observational study — you watch and record what is already happening. You change nothing.

Experiment — you deliberately **do something** to the people or things in the study, then see what happens.

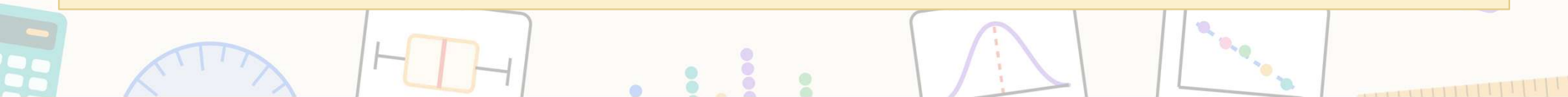
Every study you meet this year is one or the other — we unpack both over the next few slides.

Questions We Could Actually Answer

Every one of these needs data. Which ones could we collect data on this semester?

- How many hours of sleep do Soffer students get on a school night?
- Do students who bring lunch spend less per week than students who buy it?
- Is the average commute to Soffer longer for juniors than for freshmen?
- Does listening to music while studying change quiz scores?
- What percent of Soffer students have a part-time job?

One of these can only be answered with an **experiment**. Can you tell which, and why?



Explanatory and Response Variables

An **explanatory variable** is the variable we think might cause a change.

A **response variable** is the outcome we measure on each unit.

- Either one can be categorical or quantitative.
- Some books say “independent” and “dependent.” We will not — those words already mean something else in statistics.

Back to our five questions: “Does listening to music while studying change quiz scores?”

Explanatory: whether the student studies with music (*categorical*)

Response: quiz score out of 20 (*quantitative*)

You cannot answer this by watching — you have to **decide** who studies with music. That is what makes it the **experiment** of the five.



Experiments: The Vocabulary

An **experiment** is a study in which a researcher **assigns** conditions (treatments) to experimental units.

Experimental unit — the thing a treatment is assigned to. When they are people we call them **subjects** or **participants**.

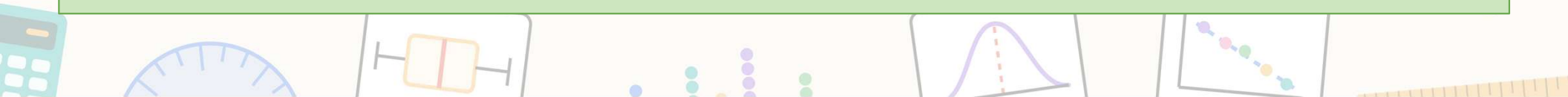
Factor — an explanatory variable whose levels are imposed on the units.

Levels — the different values a factor can take.

Treatment — one level of a factor, or a **combination** of levels when there is more than one factor.

Response variable — what you measure on each unit **after** the treatment has been applied.

Two factors with 3 and 4 levels → $3 \times 4 = 12$ **treatments**.



Practice with the Language of Experiments

✦ **BONUS SLIDE**

Does limiting social media use reduce feelings of loneliness among college students?

Researchers recruited 143 University of Pennsylvania students and randomly placed them into two groups. For three weeks, one group continued using Facebook, Instagram, and Snapchat normally, while the other group was limited to 10 minutes per day on each platform. At the end of each week, every student completed a standardized questionnaire measuring loneliness. The students who limited their social media use showed a greater decrease in loneliness.

Variables • Experimental Units • Factor(s) • Level(s) • Treatments

Practice with the Language of Experiments

★ **BONUS SLIDE**

Does having your phone nearby drain your brainpower even when you never touch it?

Researchers recruited 275 undergraduates for what they were told was a study about memory. Before testing began, each student was randomly told where to put their own phone: face up on the desk, in their pocket or bag, or in a completely different room. Independently, each student was also randomly told to leave the phone powered on or to shut it completely off. Every student then took a working-memory test (holding letters in mind while solving math problems) and a computerized attention task. Nobody was allowed to check their phone at any point. Students whose phones were in another room scored highest; students with phones on the desk scored lowest. The gap was biggest for students who said they depended on their phones the most.

Variables • Experimental Units • Factor(s) • Level(s) • Treatments

Factors and Levels: The Five-Second Rule

Researchers at Rutgers put the five-second rule to the test. They used four foods (watermelon, bread, buttered bread, gummy candy), dropped onto four surfaces (steel, tile, wood, carpet), for four lengths of time (< 1 s, 5 s, 30 s, 300 s), using bacteria prepared two different ways.

Identify the response variable and the experimental units.

List the factors and the number of levels of each. How many treatments?



- Type of food — **4 levels**
- Type of surface — **4 levels**
- Amount of time — **4 levels**
- Bacterial prep — **2 levels**

$$4 \times 4 \times 4 \times 2 = \mathbf{128 \text{ treatments}}$$

Experimental units: the individual pieces of food that were dropped.

Response: amount of bacteria transferred.

Treatments and Experimental Units

Malaria kills hundreds of thousands of people a year, many of them children. Will regularly screening children and treating those who test positive reduce the proportion who develop the disease? Researchers worked with children in 101 schools in Kenya, randomly assigning half the schools to receive regular screening and follow-up treatment and the rest to receive no regular screening. All 101 schools were tested for malaria at the end.

Identify the treatments and the experimental units.



Treatments

- Regular screening + follow-up treatment
- No regular screening

Experimental units

The **101 schools** — not the children. The treatment was assigned to a whole school at a time.

Observational Studies

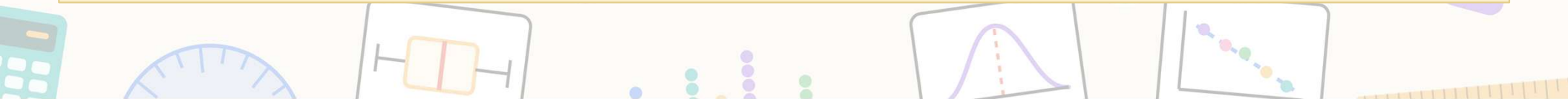
In an **observational study**, treatments are **not** imposed. The researcher simply records the values of the variables of interest.

Retrospective — units are chosen now, and data from the **past** are gathered.

Prospective — units are chosen now, and data are gathered now and **into the future**.

A **survey** is an observational study in which data are collected from **humans** using a standard set of questions.

Every sample survey you have ever seen — Gallup, Pew, a class poll — is an observational study. Nobody was assigned anything.



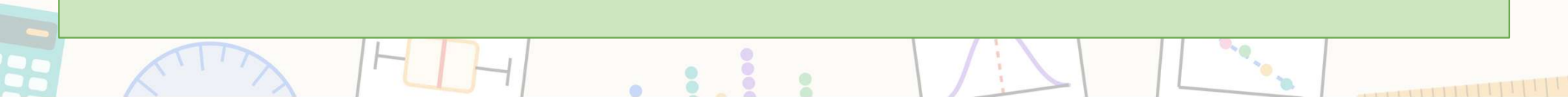
Observational Study or Experiment?

A study finds that mothers who suffer from morning sickness during pregnancy tend to have children with higher IQs. Researchers looked at 121 children aged 3 to 7, many of whose mothers had morning sickness. These children scored higher on IQ, memory, and language tests.

Observational study — nobody assigned mothers to have morning sickness. The researchers just recorded what already happened.

The students who run the Soffer newspaper want to know what students think about lunch. They select a random sample of students in the lunch line and ask how much they spent that day and what they think of the food.

Observational study — and specifically a **survey**. Note the random *selection*: that will matter when we talk about scope.



Observational Study or Experiment?

A study shows that 4th graders who use flashcards learn math facts faster than those who use computer programs. Two hundred 4th graders participated; 100 were **given** flashcards to use for 6 months while the other 100 used computer practice programs. The flashcard group averaged 12 points higher.

Experiment — the researchers decided who used what. A treatment was imposed.

Headline: “Drinking Coffee Gives Jolt to Life Span.” Researchers surveyed 402,260 people about how many cups of coffee they drank per day. Fifteen years later they checked death records. People who drank 2+ cups a day were 10–16% less likely to have died.



Observational study, and **prospective** — chosen then, followed forward. So no, this does not show that coffee *causes* a longer life.

Confounding

An **extraneous variable** is any variable that is known (or believed) to affect the response but is not the explanatory variable being studied.

A **confounding variable** gives an **alternative explanation** for the relationship you observed — so you cannot say the explanatory variable caused the change in the response.

To be confounding, a variable must satisfy BOTH:

- it is associated with the **explanatory** variable, **AND**
- it is associated with the **response** variable.

If you can only argue one of the two, you have named an extraneous variable — not a confounding one. On the AP[®] exam you must explain *both* links.

Extraneous, Not Confounding: Caffeine

✦ **BONUS SLIDE**

Sixty volunteers are **randomly assigned** to drink caffeinated or decaf coffee, then take a reaction-time test.

Extraneous variable: hours of sleep last night. Tired people react more slowly, whatever is in the cup.



Is it confounding? Check BOTH links:

- Linked to the **response**? **Yes** — less sleep, slower reactions.
- Linked to the **explanatory** variable? **No** — random assignment spread the sleepy volunteers evenly across both groups.

One link only → **extraneous, not confounding**. *Flip it:* let volunteers **choose** their drink. If the tired ones grab the caffeine, sleep now has both links — and it **is** confounding.

Extraneous, Not Confounding: Fertilizer

✦ BONUS SLIDE

A gardener **randomly assigns** 10 of her 20 plots to Fertilizer A and 10 to B, then weighs each plot's tomatoes.

Extraneous variable: soil quality. It varies from plot to plot, and richer soil grows more tomatoes.



Is it confounding? Check BOTH links:

- Linked to the **response**? **Yes** — richer soil, bigger harvest.
- Linked to the **explanatory** variable? **No** — random assignment mixed good and poor soil into both fertilizer groups.

One link only → **extraneous, not confounding**. *Flip it:* put A on the sunny south rows and B on the shady north rows — now sun and soil travel with the treatment, and they **are** confounding.

Confounding in Action: Vitamin D

Hundreds of observational studies show that people with a higher vitamin D concentration in their blood have a lower risk of diabetes.

So should everyone start taking vitamin D?



Variable	Group 1	Group 2
Vitamin D concentration (explanatory)	High vitamin D concentration	Low vitamin D concentration
Quality of diet	Better diet	Worse diet
Amount of exercise	More exercise	Less exercise
⋮	⋮	⋮
Amount of vitamin supplementation	More likely to take other vitamins	Less likely to take other vitamins
Diabetes status (response)	Less likely to have diabetes	More likely to have diabetes

Any of these could be the real cause. Amount of **outdoor exercise** is associated with vitamin D (sun exposure) *and* with diabetes risk — that makes it a **confounding variable**.

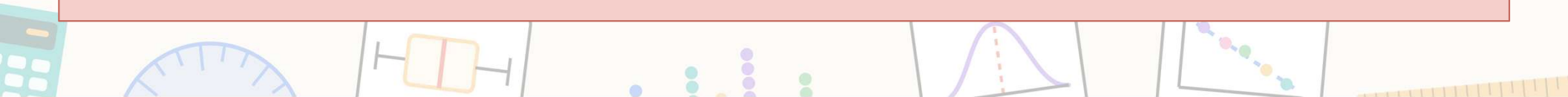
The Fix: Run an Experiment

To find out whether vitamin D actually **causes** a reduction in diabetes risk, researchers in Norway ran an experiment. They randomly assigned 500 people with pre-diabetes to take either a high dose of vitamin D or a placebo. After 5 years, about 40% of **each** group had been diagnosed with diabetes.

Random assignment — a **chance process**, not the researcher, decides which unit gets which treatment. (How to describe one: Topic 1.13.)

A **placebo** is a treatment with no active ingredient that is otherwise identical to the real one.

Same 40% in both groups. The observational studies were picking up confounding, not causation. **This is why the AP[®] exam cares so much about how the data were collected.**



Who Can We Generalize To?

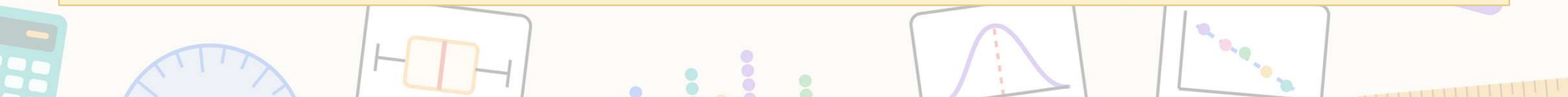
A sample is **random** when every unit in it was selected from the population using a chance process — a random number generator, numbered slips of paper, and so on.

Units were randomly **selected** from a population
→ generalize to **that whole population**.

Units **volunteered** or were hand-picked
→ generalize only to individuals **similar to those in the study**.

Sixty 4-year-olds whose parents **volunteered** were randomly divided into three groups and shown different videos.

Largest population: all children *like these volunteers* — because the children were never randomly selected from any population.



Understanding Check: Scope

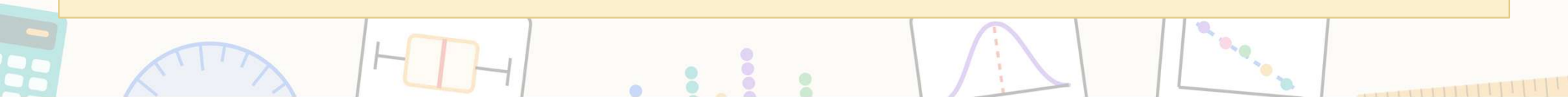
A recent National Health Interview Survey asked a **random sample of U.S. adults** whether they had experienced depression symptoms in the past year. 21.5% of adults with a high school diploma or less had, compared with 12.4% of adults with a college degree.

What is the largest population we can generalize to?

All U.S. adults ages 18 and older, because the sample was randomly selected from that population.

But this is an **observational study** — no treatment was assigned — so we **cannot** say education level *causes* a change in depression risk.

Two separate answers, every time: *who*, and *cause or not*.



Scope of Inference: The Whole Picture

		Were units randomly ASSIGNED to treatments?	
		Yes (experiment)	No (observational)
Were units randomly SELECTED?	Yes	Generalize: YES Cause & effect: YES	Generalize: YES Cause & effect: NO
	No	Generalize: NO Cause & effect: YES	Generalize: NO Cause & effect: NO

Random **selection** buys you the population. Random **assignment** buys you causation. You need both to get both.



The Investigative Question — Revisited

In 1A a *valid* investigative question just needed data that vary — the entry ticket. Now that you know what a study can and cannot conclude, we raise the bar: the question itself has to settle all of it **before** any data are collected.

1A, in one line: Joy Milne's shirts ✓ — some right, some wrong, the data vary. · The 100-metre world record ✗ — one value answers it.

A question can clear that bar and *still* be a bad investigative question. Try this one: **“How tall are basketball players?”**

Heights vary ... so what is wrong with it?

Mean or median height? In what **units**? *Which* players — the NBA, our team, all of them? And are we **estimating**, or **testing a claim**?

Every one of those gaps is one of the three components.

The Three Components of an Investigative Question

A well-written investigative question does three jobs at once — and you name them in this order, every time.

1 VARIABLE(S)

What is measured, on whom, and in what units?

guides DATA COLLECTION

2 TYPE OF INFERENCE

Which parameter? Estimate it, or test a claim (and which way)?

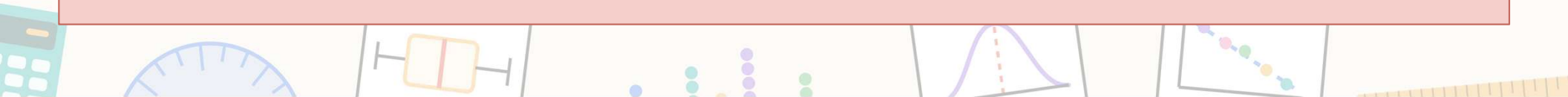
guides the ANALYSIS

3 TYPE OF CONCLUSION

Which population?
Cause and effect allowed?

guides the CONCLUSION

If you cannot point to all three in the question, the question is not finished.



Component 1: The Variable(s) of Interest

It guides the data collection.

The question has to name **what gets measured, on whom, and in what units** — clearly enough that someone else could go collect the data without asking you a single follow-up question.

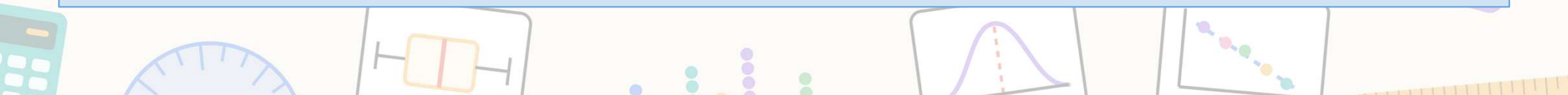
Weak: “Do students sleep enough?”

Measure what? “Enough” is not a variable. No units.

Strong: “What is the mean number of **hours of sleep per school night** for all Soffer students?”

Variable: hours of sleep per school night. Units: hours. Measured on: Soffer students.

In an experiment there are **two** variables to name: the **explanatory** one and the **response** one.



Component 2: The Type of Inference

It guides the choice of analysis.

Name the **parameter**, then say whether we are **estimating** it or **testing a claim** about it.

The only four parameters this course estimates or tests

one **mean** · one **proportion** · a **difference of two means** · a **difference of two proportions**

ESTIMATE a parameter

estimate · plausible values · how much

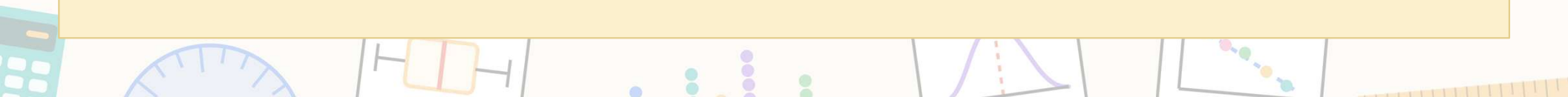
→ **confidence interval**

TEST a claim

is there convincing evidence that ...

→ **significance test**

You do **not** need to know how to build an interval or run a test yet — that is Units 3 and 4. Today you only have to **read the question** and say which one it is asking for.



Component 2, continued: Which Way?

If the question asks for a **test**, the wording must also set the **direction** — that is the alternative hypothesis, written in plain English before you have any data.

- greater than · more than · larger than → one-sided
- less than · fewer than · reduces · lower → one-sided
- differs from · is different from · not equal → two-sided
- is there an association · are they independent → association (you will meet this one in a later unit)

Not directional: “Is there convincing evidence about the mean weight of the chip bags?”

Directional: “Is there convincing evidence that the mean weight of all bags filled at the plant is **less than** 12 ounces?”

Component 3: The Type of Conclusion

It indicates what you are allowed to claim.

The question must make clear **which population** the answer applies to — and whether a **cause-and-effect** claim is even on the table.

Random **selection** from a population
→ name that population.

Random **assignment** of treatments
→ you may use the word **cause**.

Steal this phrasing

Survey, random sample: “... for **all** U.S. adults?”

Experiment, volunteers: “For all patients **like the ones in this experiment**, does X **cause** a reduction in ...?”

The Routine: Naming the Three Components

This is exactly what you will write on the worksheet. Three lines, every time, in this order.

1. **Variable(s).** What is measured, on whom, in what units? **(In an experiment: explanatory and response.)**
2. **Type of inference.** Which parameter — and estimate it, or test a claim about it? **(If a test: which direction?)**
3. **Type of conclusion.** Which population do the results apply to — and is cause and effect allowed?

If you can answer all three from the question **alone**, without being told anything else, the question is well written. If you have to guess at one of them, that is the component that is missing.

Quick Decision Guide

“What is an **estimate** of ...” → **estimate a parameter** → **a confidence interval** (Unit 4)

“Is there **convincing evidence** that ...” → **test a claim** → **a significance test** (Unit 3)

greater than / less than / differs from / association → sets the **direction**

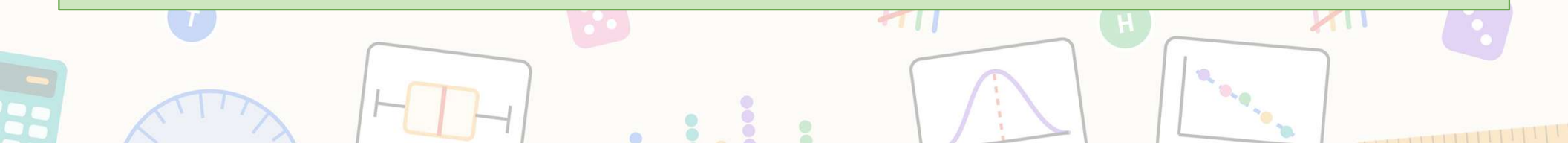
Random **selection** → name the population you generalize to

Random **assignment** → the word **cause** is allowed

Two model questions worth copying

“What is an estimate of the **mean weight, in pounds**, of **all** largemouth bass in Lake Erie?” (**one mean** → **interval**)

“Is there convincing evidence that the **proportion** of **all** men in Ohio who own a vehicle is **larger than** the proportion of **all** women in Ohio who do?” (**difference of two proportions** → **one-sided test**)



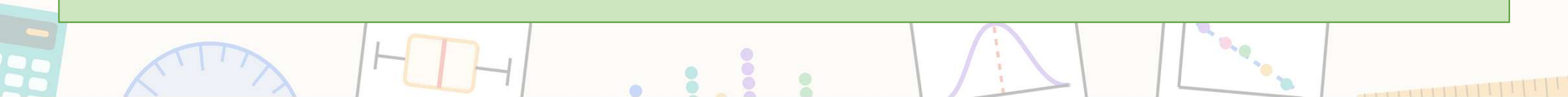
Practice: Name the Three Components

“What is an estimate of the mean number of hours per week that all students at Soffer spend on homework?”

1. Variable: hours per week spent on homework · **2. Inference:** one population **mean**, *estimate* → confidence interval · **3. Conclusion:** all students at Soffer; no cause and effect (nothing was assigned)

“Does the data provide convincing evidence that more than 60% of all registered voters in Miami-Dade County support the proposed school bond?”

1. Variable: whether a voter supports the bond · **2. Inference:** one population **proportion**, test, one-sided (*more than 60%*) · **3. Conclusion:** all registered Miami-Dade voters; no cause and effect



Practice: The Hard One — an Experiment

“For all pneumonia patients **like the ones in this experiment**, do supplemental breathing exercises **cause** a **reduction** in the **proportion** of patients who develop pleurisy, compared with no supplemental breathing exercises?”

Variables: whether the patient did the breathing exercises (**explanatory**) and whether the patient developed pleurisy (**response**) — both categorical.

Type of inference: a **difference of two proportions**; “cause a reduction” asks for a **significance test**, one-sided.

Type of conclusion: cause and effect **is** allowed — it is an experiment with random assignment — but only for patients **like the ones in the study**, because they were not randomly selected.

Read that question again. Every phrase in bold is one of the three components doing its job. Nothing in it is decoration.



Practice: What Is Missing?

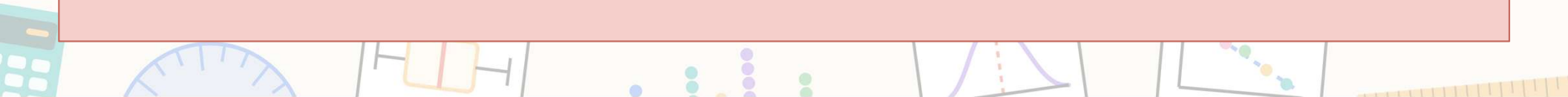
“Are boys and girls different in how much they text?”

No parameter or units (mean texts per day?), no population, and “different” never names what is being compared.

Fix: “Is there convincing evidence of a difference between the mean number of texts sent per day by all U.S. teenage boys and all U.S. teenage girls?”

“Is there convincing evidence that the 30 fish we caught from the pond have a mean length greater than 15 inches?”

This asks about the **sample** we already have — just compute the mean, there is nothing to infer. It has to ask about *all fish in the pond*.



Practice: Write Your Own

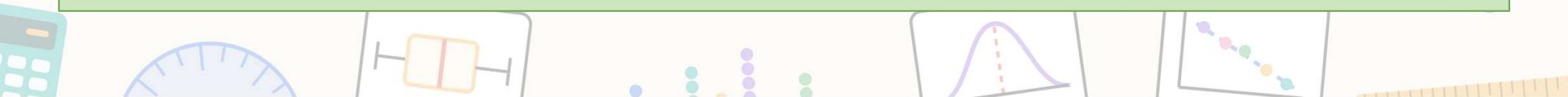
Name the parameter, name the population, and make the analysis clear.

A coffee shop owner wants a good **estimate** of how much a typical customer spends, in dollars.

“What is an estimate of the mean amount spent, in dollars, per customer at the coffee shop?” (**mean** → **interval**)

A sleep researcher **randomly assigns** volunteers to a 65°F or a 75°F room and records minutes to fall asleep. Does the cooler room help?

“Is there convincing evidence that the mean time to fall asleep is lower for people in a 65°F room than for people in a 75°F room?” (**difference of means** → **one-sided test**; **random assignment** → **cause and effect**)



Activity: The Concert

These are the satisfaction scores (0–100) of **50 concertgoers**. Pick **5 scores** that you think are *representative* of the audience and write down the mean of your five.

92	89	90	88	95	100	98	93	95	84
82	86	90	88	86	91	90	89	85	83
80	74	80	67	81	82	76	77	74	65
72	68	74	73	70	69	72	70	68	67
69	67	68	68	64	66	63	63	70	68



What Just Happened?

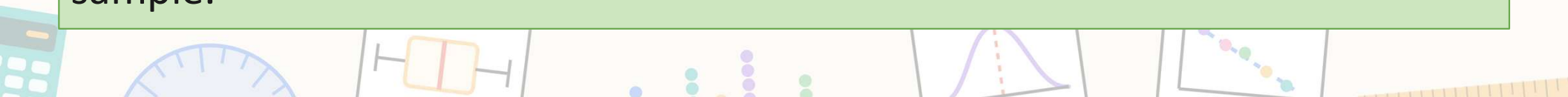
The true mean satisfaction of all 50 attendees is **78.38**.

Where did most of the class means land?

Almost everyone landed **too high**. Your eye was pulled to the big numbers at the top of the table — and you never noticed it happening.

This is **bias**: a systematic error that pushes the estimate the same direction every time. Taking a *bigger* judgment sample would not fix it.

The only fix is to take the choice away from the human. Let **chance** pick the sample.



What Makes a Sample Random?

A sample is **random** when a **chance process** — not a person — decides which members of the population are in it.

A sample chosen by chance rules out both **favoritism by the sampler** and **self-selection by respondents**.

Random sampling does not guarantee a perfect sample. It guarantees that the method is not tilted in any direction — and it lets us *quantify* how far off we are likely to be.



With or Without Replacement?

Sampling without replacement

Once a unit is selected it is **not** returned to the population, so it cannot be chosen again.

This is what we do in this course.

Sampling with replacement

A selected unit **is** returned to the population before the next draw, so it could be chosen again.

Rare in surveys; common in simulation.

Most random number generators sample **with** replacement. So when you use one to build an SRS, you must say that you will **ignore repeats**.



Simple Random Sample (SRS)

In a **simple random sample (SRS)** of size n , every possible sample of size n has the same chance of being selected.

Read that definition again. It is about **samples**, not individuals.

Line up 20 students alphabetically and flip a coin to take either all the odd positions or all the even positions. Every **individual** has a 50% chance of being chosen ... but only **two** of the possible samples can ever happen. That is **not an SRS**.



How to Choose an SRS with Technology

Three steps — say all three for full credit

Label. Give every individual in the population a distinct number from 1 to N .

Randomize. Use a random number generator to get n different integers from 1 to N , ignoring repeats.

Select. The individuals with those labels are your sample.



Example: Choosing an SRS

There are **28 hotels** on the beach strip. A travel magazine wants to inspect a simple random sample of **4** of them.

Describe how to select the sample using technology.



Label the hotels 1 through 28 in alphabetical order.

Randomize: use a random number generator to produce integers from 1 to 28, ignoring any repeat, until you have **4 different** numbers.

Select the four hotels with those labels.

Notice what is *not* here: no coin flips, no “pick the ones that look typical,” no random digit table.

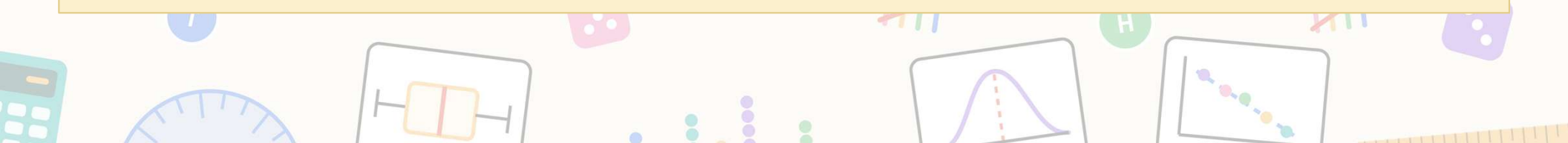
Stratified Random Sampling

Strata are non-overlapping groups of individuals who **share a characteristic** thought to be related to what we are measuring.

In **stratified random sampling** we take an **SRS from every stratum** and combine them into one overall sample.

Works best when individuals **within** each stratum are similar (homogeneous) and there are large differences **between** strata.

When strata are chosen wisely, a stratified random sample gives a **more precise** estimate than an SRS of the same size.



Example: Stratified Random Sampling

Suppose we want the average number of hours of sleep Soffer students get on a school night (Sunday–Thursday), and we can afford a sample of **40** students.

An SRS of 40 might, just by chance, pull mostly seniors — and it is very reasonable to think freshmen and seniors sleep differently.

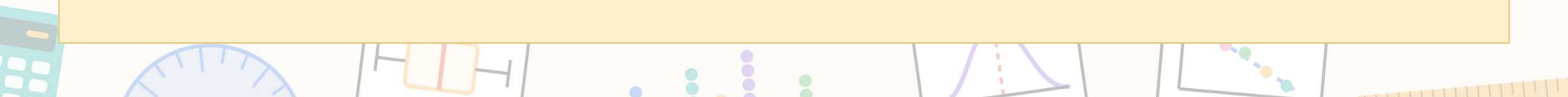
Stratified plan

Split the school into four strata: **9th, 10th, 11th, 12th**.

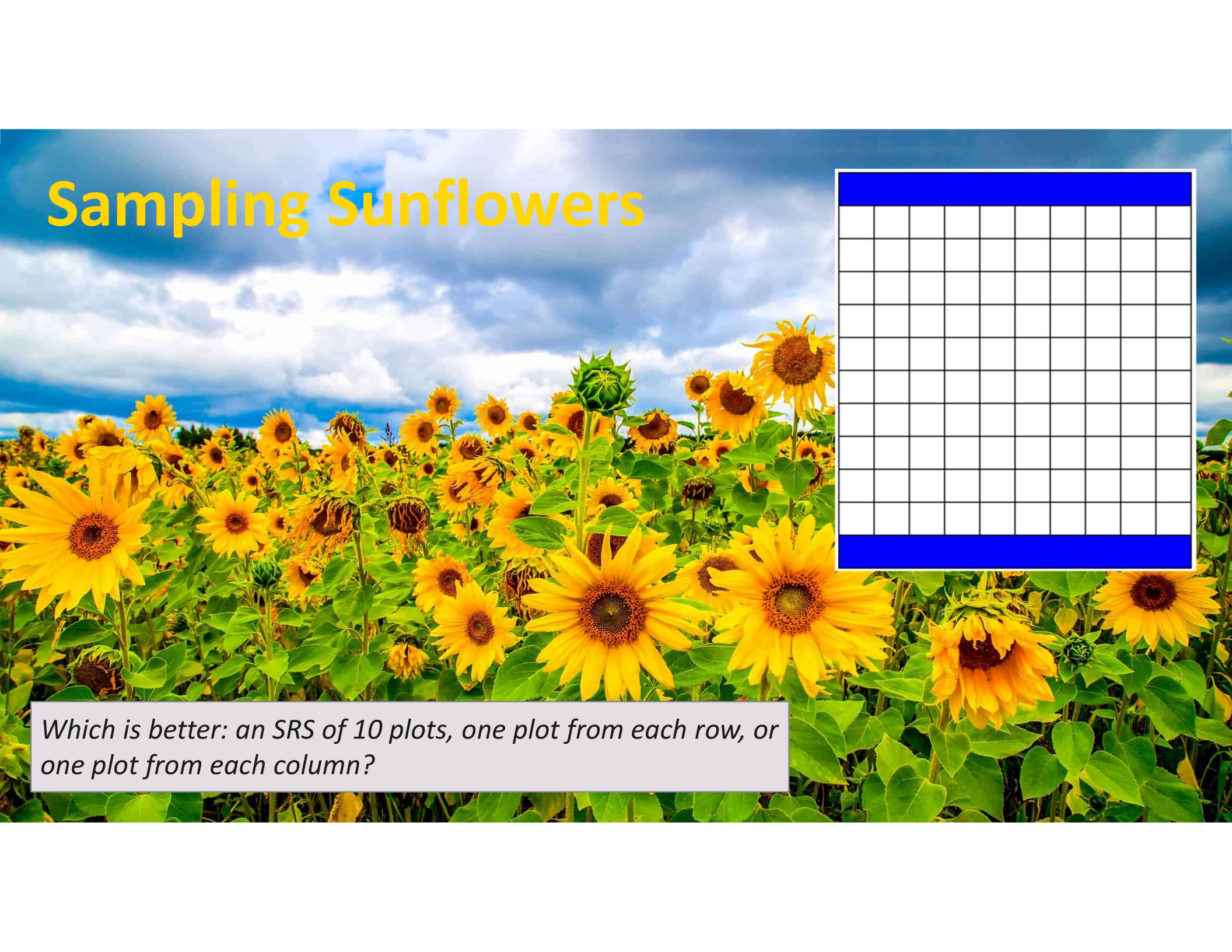
Take an **SRS of 10 students from each grade**.

Combine all four subsamples into one sample of 40.

Now every grade is guaranteed to be represented — and the grade-to-grade differences no longer add noise to our estimate.

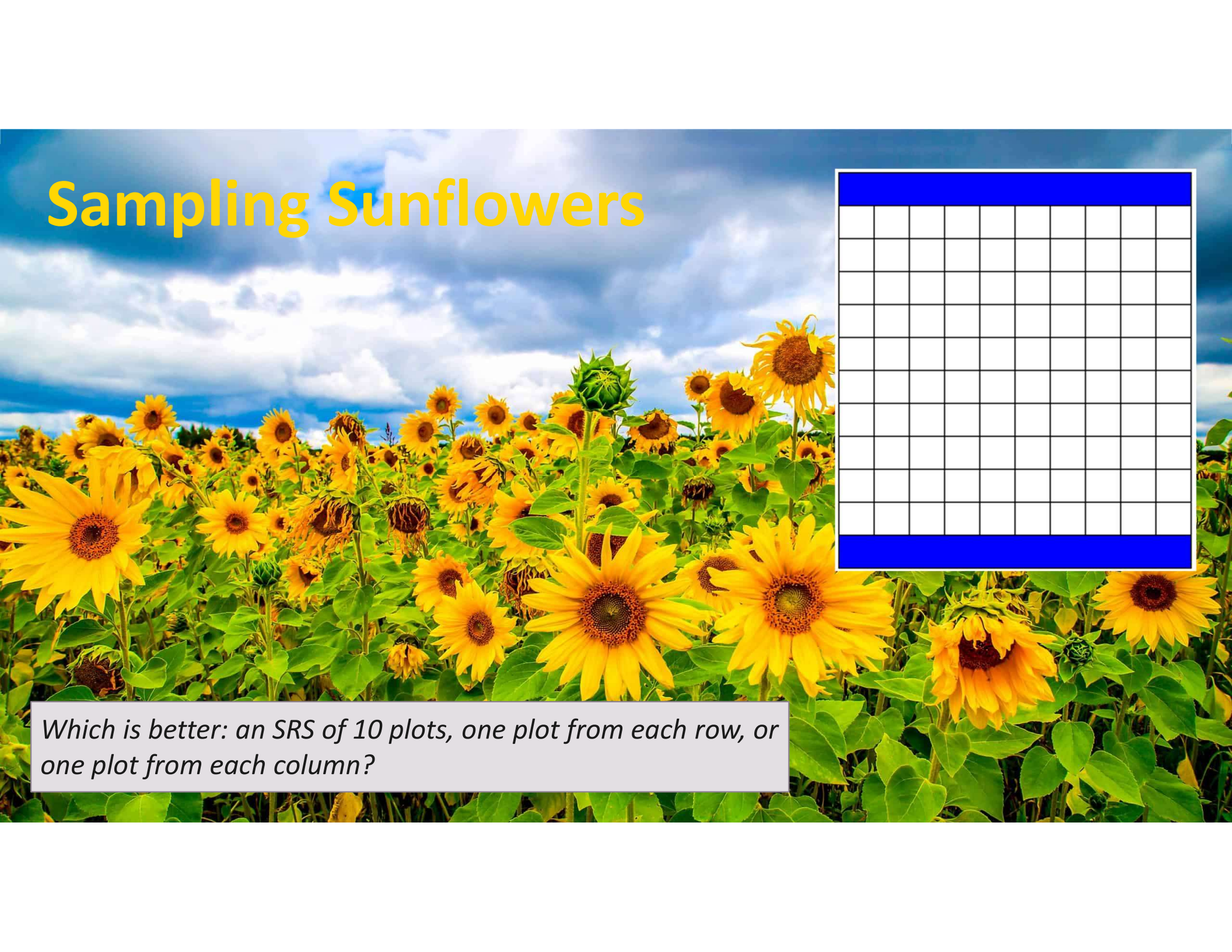


Sampling Sunflowers



Which is better: an SRS of 10 plots, one plot from each row, or one plot from each column?

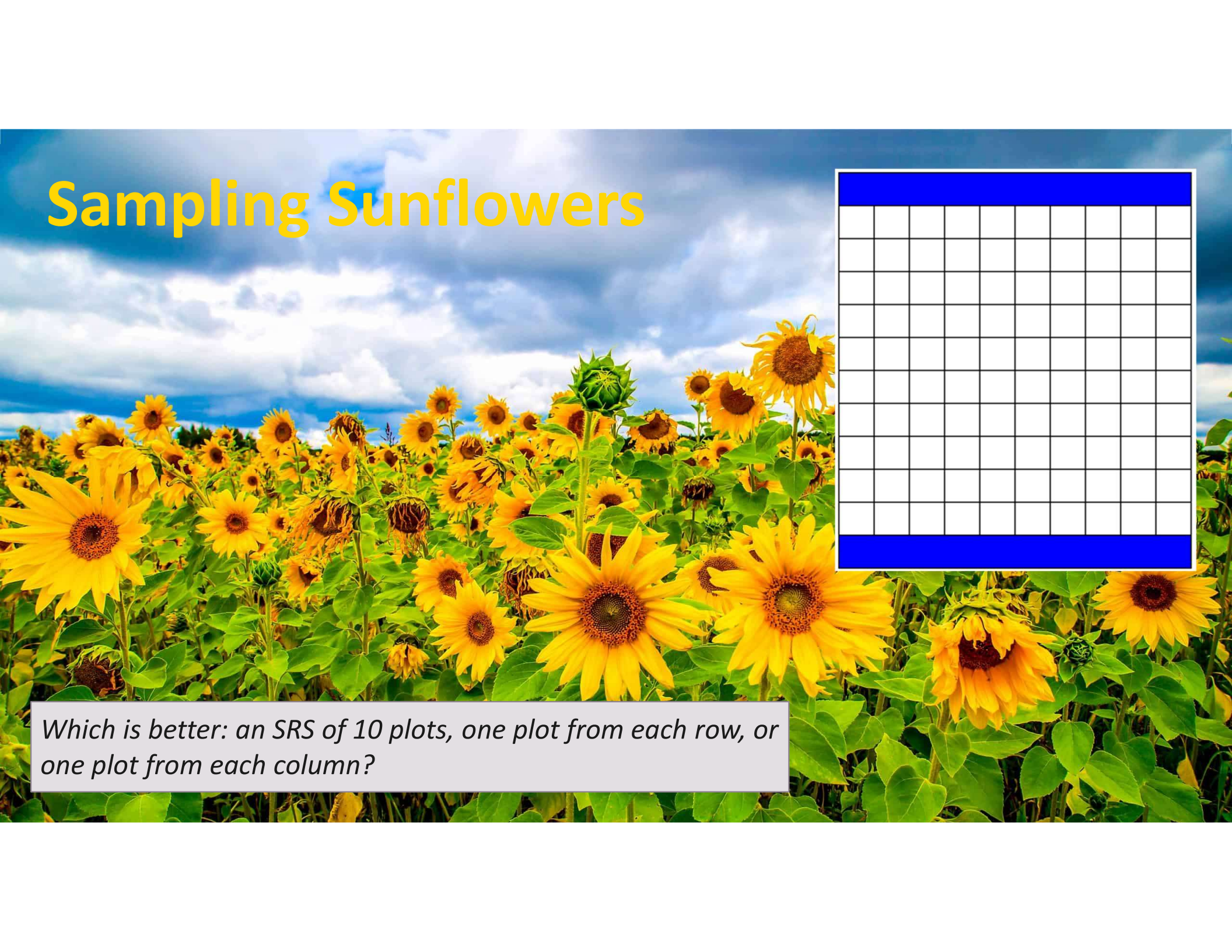
Sampling Sunflowers



A photograph of a sunflower field under a blue sky with white clouds. A 10x10 grid is overlaid on the right side of the image, with blue bars at the top and bottom. The grid is used for sampling the sunflowers.

Which is better: an SRS of 10 plots, one plot from each row, or one plot from each column?

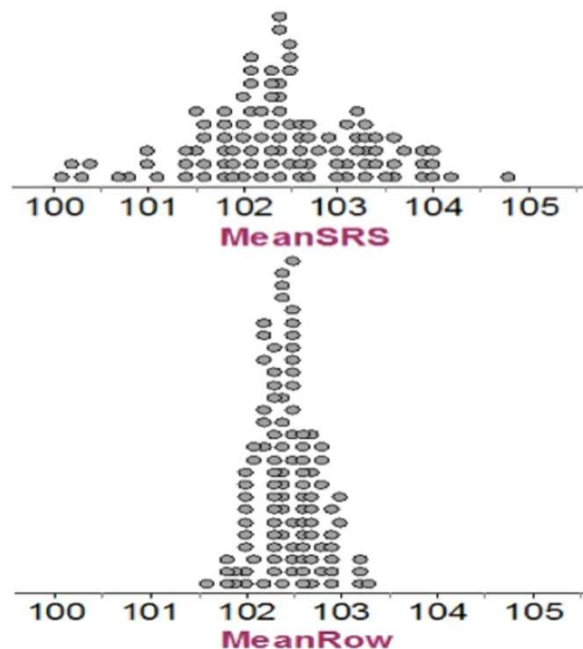
Sampling Sunflowers



A photograph of a sunflower field under a blue sky with white clouds. A 10x10 grid is overlaid on the right side of the image, with blue bars at the top and bottom. The grid is used for sampling the sunflowers.

Which is better: an SRS of 10 plots, one plot from each row, or one plot from each column?

Sampling Sunflowers: The Results



All three dotplots are centered in about the same place — no method is **biased**.

But the row-stratified estimates are much **less variable**: they cluster tightly around the truth.

Water comes from the ditches at the top and bottom, so a plot's row is strongly related to plant health. Rows make good **strata**.

Same sample size, better estimate. That is the whole payoff of stratifying on the right variable.

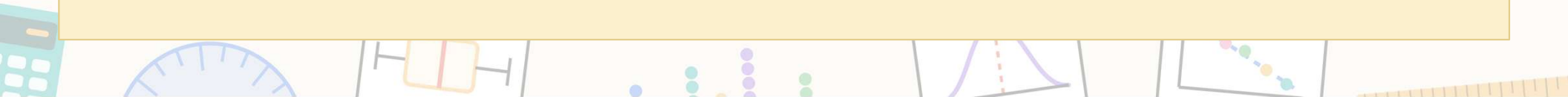
Cluster Random Sampling

A **cluster** is a group of individuals located near one another. In **cluster random sampling** we take an **SRS of whole clusters** and collect data from **every** individual in the chosen clusters.

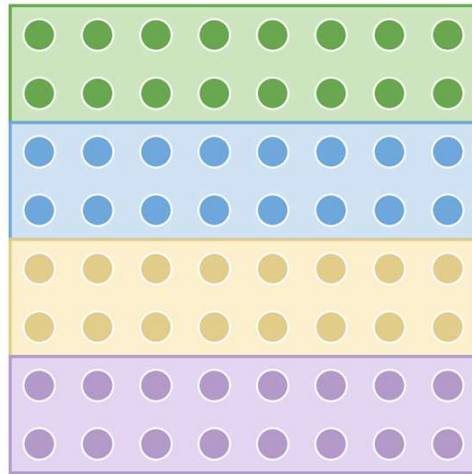
Works best when each cluster is **heterogeneous** — a mini-version of the whole population.

Careful: in 1A a “cluster” was a **clump of data values** in a distribution. Here it is a **group of individuals you sample all of**. Same word, unrelated idea.

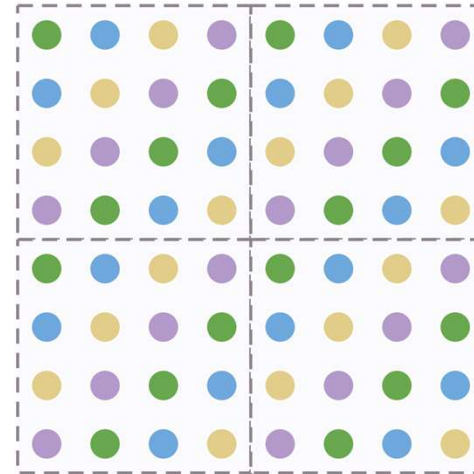
Usually chosen for **practical** reasons: it saves time and money, and you do not need a list of every individual in the population — only a list of clusters.



Strata vs. Clusters — Do Not Mix These Up



STRATA
similar **WITHIN** each group,
different **BETWEEN** groups



CLUSTERS
mixed **WITHIN** each group,
each group mirrors the population

Strata: sample from **every** group.
everyone inside.

Clusters: sample **whole groups**, then take

Systematic Random Sampling

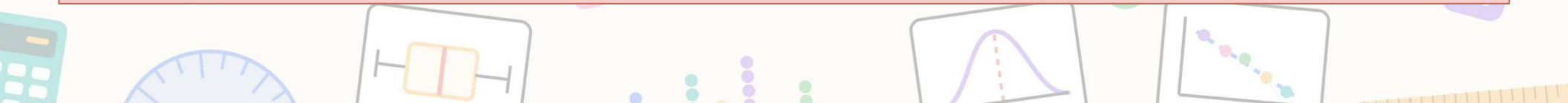
In **systematic random sampling** we take an ordered list, randomly pick one of the first k individuals, and then take **every k -th individual** after that.

Useful when you cannot list the population in advance — **exit polling** on Election Day is the classic case. You have no idea who will show up, but you can survey every 20th voter who walks out.



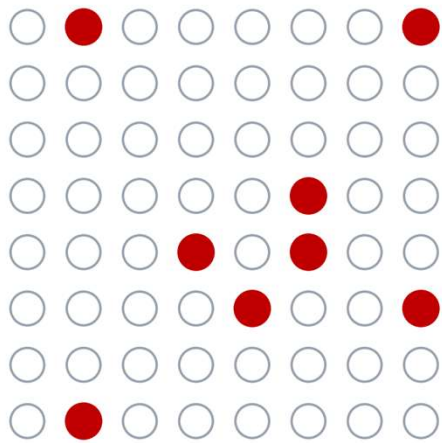
CAUTION

If there is a **pattern** in the way the population is ordered that lines up with your interval k , the sample will not be representative. (Every 10th house on a block is a corner lot.)



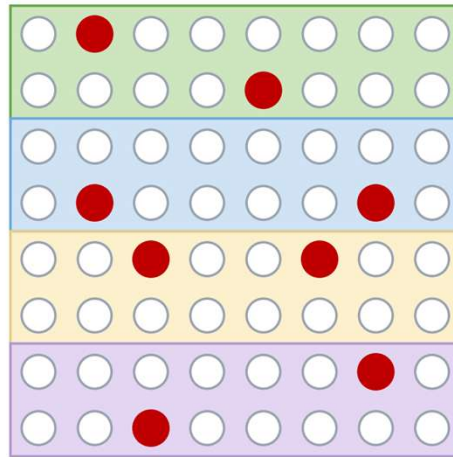
The Four Random Sampling Methods

Simple Random Sample



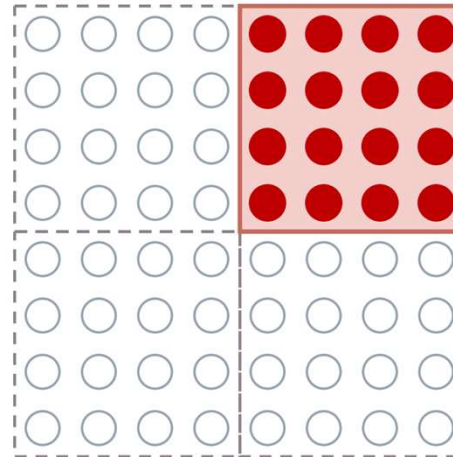
Every possible sample of size n has the same chance.

Stratified Random Sample



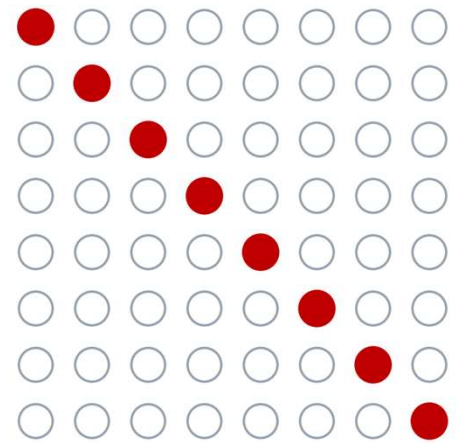
Split into strata (similar within).
Take an SRS from EVERY stratum.

Cluster Random Sample



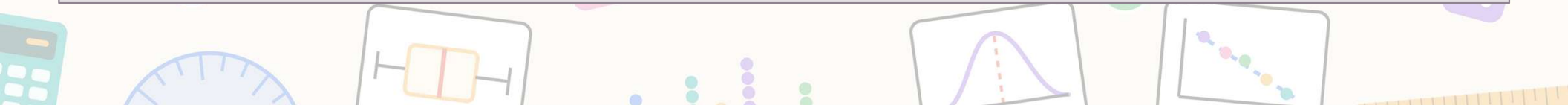
Split into clusters (each mirrors the population). Pick whole clusters.

Systematic Random Sample



Random start, then every k -th individual after that.

A **census** is the fifth way to get data from a population — but it is not a sampling method, because you take **everybody**.



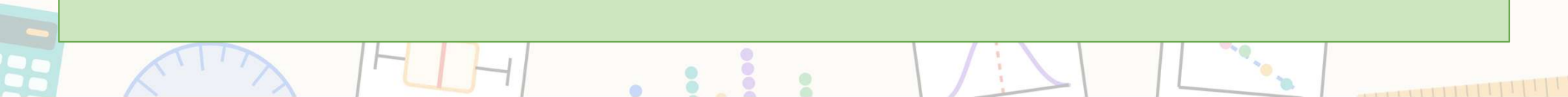
Identify the Sampling Method

The Educational Testing Service needed a sample of colleges. ETS first divided all colleges into groups of similar type (small public, small private, large public, ...). Then they randomly selected 3 colleges from **each** group.

Stratified random sampling — groups of similar colleges are the strata, and an SRS was taken from every one of them.

A county commissioner randomly selects **blocks** in her district and then surveys **everyone** who lives on those blocks.

Cluster random sampling — whole blocks were chosen, and every resident of a chosen block is in the sample.



Identify the Sampling Method

A restaurant manager randomly picks a number between 1 and 10 each night. He gives a survey to that customer and to **every 10th customer after them**.

Systematic random sampling — random start, then a fixed interval.

A teacher puts all 30 students' names in a hat and, without looking, draws 5.

Simple random sample — every group of 5 students had the same chance of being drawn.

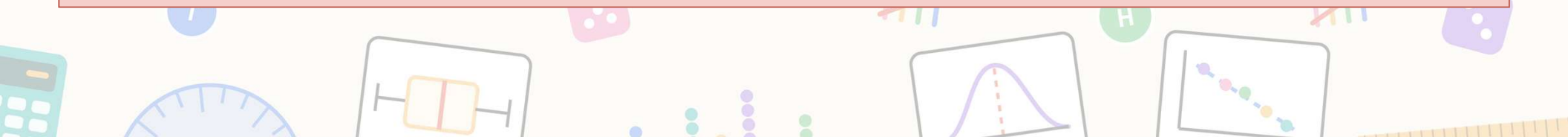


Which Method Should I Use?

Match the method to the situation

- You have a **full list** and no strong subgroups → **SRS**. Simple and unbiased.
- The population splits into groups that **differ a lot** and you want precision → **stratified**.
- The population is **spread out**, or you have no list of individuals, and you need to save time and money → **cluster**.
- You cannot list the population in advance but they come by in order → **systematic**.

On the AP[®] exam, “justify a sampling method” means naming the feature of *this* population that makes *this* method a good fit. “Because it is random” earns nothing.



Understanding Check

A warehouse has **100 boxes**, each containing **50 identical toys**. You want to test **200 toys** for quality.

Using the framework on the last slide: describe how each method would work and choose the one you think would work best. Be ready to say why!

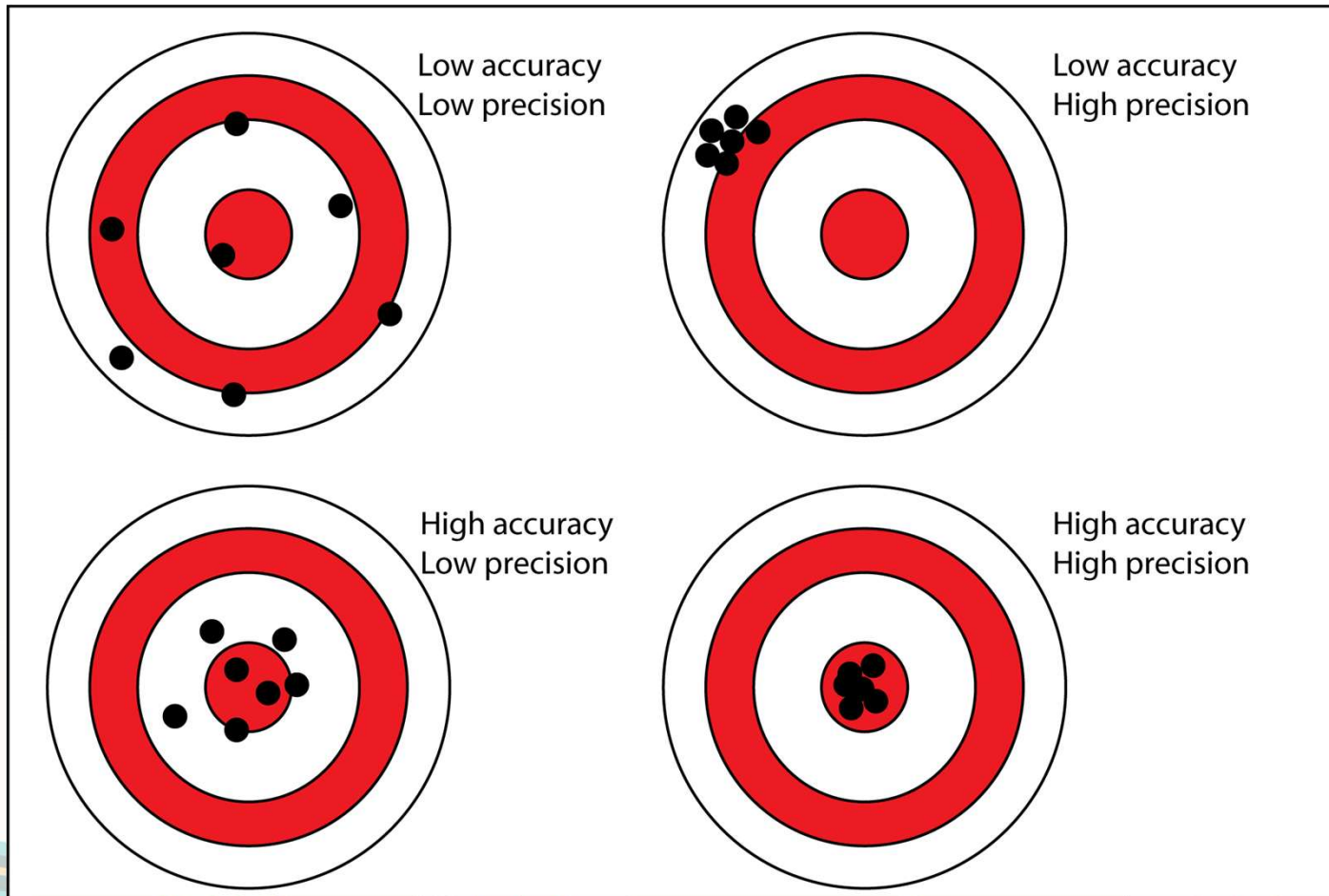
- Simple Random Sample (SRS)
- Stratified Random Sample
- Cluster Random Sample
- Systematic Random Sample
- Census

A **cluster random sample** is the practical choice: randomly select **4 boxes** and test all 50 toys in each.

The toys are identical, so any box is a fine mini-version of the warehouse — and you only have to open 4 boxes instead of hunting 200 toys across 100 boxes.



Precision vs. Accuracy



Accuracy — is the method centered on the truth? A biased method is inaccurate no matter how big the sample.

Precision — how tightly do repeated samples cluster? Bigger samples and smart strata improve precision.

Random sampling buys **accuracy**. Sample size and good strata buy **precision**.

Bias

Bias in a sampling method is a **systematic error** in the procedure that makes the statistic consistently **larger** — or consistently **smaller** — than the parameter it is estimating.

- Bias is a property of the **method**, not of one unlucky sample.
- A bigger sample does **not** fix bias. It just gives you a more precise wrong answer.

Remember the concert. Everyone was trying to be fair, everyone landed high, and picking 10 scores instead of 5 would have landed high too.



Convenience Sampling

Choosing individuals who are **easy to reach** produces a **convenience sample**.

A gardener with 20 tomato plants counts the tomatoes on the **5 plants closest to the house** — 53 of them — and estimates $53 \times 4 = 212$ tomatoes in the whole garden.

Which way is this estimate off, and why?



The plants nearest the house are the easiest to water, weed, and fertilize, and they sit on the sunny southern edge. They are **healthier than average**, so 212 is almost certainly an **overestimate**.

Voluntary Response Sampling

Voluntary response sampling lets people **choose themselves** into the sample by responding to a general invitation.

Internet polls, call-in and text-in polls, write-in surveys, “rate us” links, and the review section of any app.

People who bother to respond usually have **strong opinions**, and usually **negative** ones. They are not a cross-section of anybody.

Ask yourself: when was the last time you left a restaurant review because the food was exactly as expected?



Undercoverage Bias

Undercoverage bias happens when the sampling method **leaves out part of the population**, or makes part of it less likely to be selected.

Ideally we sample from a **sampling frame** — a list that includes every member of the population. In practice, that list is usually incomplete.

Undercoverage in the wild

- Sampling Soffer students by pulling names from the **yearbook club roster**.
- Polling voters by **landline** in 2026.
- Surveying “all shoppers” at a mall **on a Tuesday at 10 a.m.**



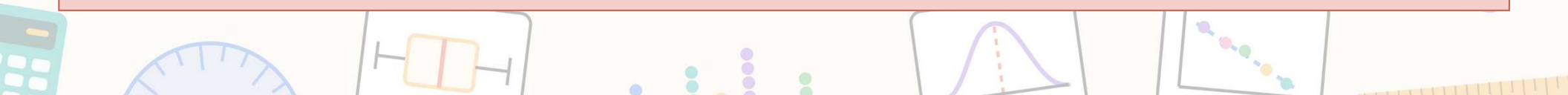
Nonresponse Bias

Nonresponse bias happens when individuals **who were selected** cannot be contacted or **refuse to participate** — and they differ from those who do respond in a way that matters.

- Undercoverage = you never had a chance to be picked.
- Nonresponse = you were picked and did not answer.

CAUTION

Students sometimes call nonresponse “voluntary response” because participation is optional. That is wrong. **Voluntary response** means people invited *themselves* into the sample.



Response Bias

Response bias happens when the answers people give tend to differ from the truth **in one direction**.

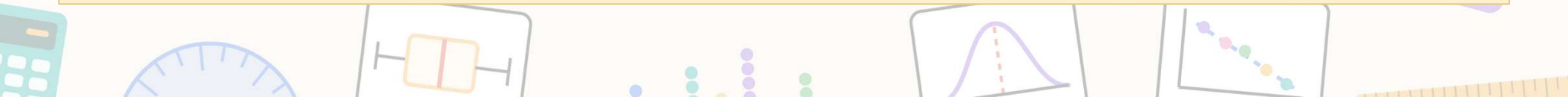
Question wording — confusing or leading questions.

Self-reporting — people round their sleep up and their screen time down.

Social desirability — answering the way that sounds good.

The interviewer — who is asking, and whether the survey is anonymous.

Notice: response bias can show up even in a **perfect random sample**. Randomizing fixes *who* you ask, not *how* you ask.



Question Wording in Action

Residents of a small Pennsylvania town were upset about a Walmart opening. At the town council meeting one resident presented a poll of a **random sample** of township residents: **92% were against** the store.

A council member asked to see the question. It read:

“Are you against a Walmart opening in our township, which would increase *traffic delays, noise, and pollution?*”

Almost nobody is in favour of traffic, noise, and pollution — so the question no longer measures opinion about **Walmart**. The 92% is an **overestimate** of the true percentage opposed.



AP[®] Exam Tip: How to Describe Bias

For **ANY** of the five, if you are asked how a design leads to bias, do **TWO** things:

- Describe **how** the people in the sample would answer differently from the rest of the population.
- Explain whether that leads to an **overestimate** or an **underestimate** of the parameter.

Not enough: “This is a voluntary response sample, so it is biased.”

Full credit: “Students who feel strongly about the dress code are more likely to answer the poll than students who do not care, so the percentage opposed in the sample will be *larger* than the true percentage of all students opposed.”

Source of Bias?

Before the 1936 presidential election (Roosevelt vs. Landon), *Literary Digest* mailed a survey to **2.8 million people** and predicted Landon would win 3-to-2. George Gallup surveyed only **50,000** and predicted Roosevelt.

The Digest built its mailing list from magazine subscribers, car registrations, and telephone directories. Roosevelt won 46 of 48 states.



Undercoverage bias — in 1936, magazine subscribers, car owners, and telephone owners were disproportionately wealthy, and wealthy voters favored Landon. The poor were largely absent from the frame.

Also **nonresponse bias**: only about 24% of the 10 million mailed surveys came back.

2.8 million biased responses lost to 50,000 well-chosen ones. **Sample size never fixes bias.**

Source of Bias?

To estimate the total amount Soffer students spend on lunch each week, you stand outside the cafeteria at lunch on one Wednesday and record what students say they spent.



Convenience sampling — you sampled whoever was easy to reach.

Also **undercoverage** — students who bring lunch from home, or who leave campus, never had a chance to be selected. Since they spend less, this estimate is probably an **overestimate**.

To estimate the average value of a home in Aventura, you average the asking prices of homes currently listed with one realtor.

Undercoverage — homes that are not for sale, and homes listed with other realtors, are left out entirely.

Bias: The Round-Up

Five names, five different failures

Convenience — you sampled whoever was easy to reach.

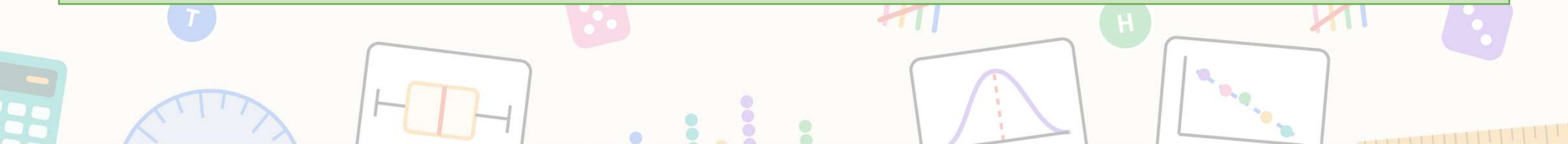
Voluntary response — people put *themselves* in the sample.

Undercoverage — part of the population could never be selected.

Nonresponse — selected people did not answer.

Response bias — the answers themselves lean one way.

Random sampling fixes the **first four**. Nothing but a better questionnaire fixes the fifth.



Let's Start With a Bad Design

Give 20 students a caffeinated cola, wait 10 minutes, and measure their pulse. Their pulse went up. Caffeine raises your pulse — right?



Their pulse went up compared to **what**? Time of day, the walk to class, the sugar in the cola, the excitement of an experiment — every one of these is confounded with caffeine. There is nothing to compare against.

The remedy is a **comparative experiment** with two or more groups.

The Four Elements of a Well-Designed Experiment

1. **Comparison** — compare at least two treatment groups, one of which may be a control group.
2. **Random assignment** — use chance to decide which unit gets which treatment.
3. **Replication** — use enough experimental units in each group.
4. **Direct control** — hold other variables the same for every unit.

Every experiment FRQ you will ever write is graded against these four.



You Are Going to Build One of These

Every August these floors are polished. By December they are scuffed grey. I buy **new tennis balls** for the chair legs — it works, and it is expensive.

Is there something cheaper that works just as well? I do not know. Nobody in this building knows.



Before this topic is over, your period designs a **real experiment** on a real problem here — and volunteers actually run it.

P3 floor protectors · **P4** the school garden · **P6** cafeteria food waste · **P8** recycling signs

So for the rest of 1.13, every element we meet is one you will have to write down for your own study. Take notes accordingly.

1. Comparison and the Control Group

A **control group** is a collection of experimental units created for **comparison** — a baseline. It may get a placebo, a different active treatment, or no treatment at all.

Rwanda wanted to raise childhood literacy. Researchers randomly assigned school districts to **teacher training, teacher training + a community literacy program, or neither**.

Why was the “neither” group necessary?



Without it, you could not tell whether rising literacy was caused by the programs or simply by children **getting older and better at reading** over the same period.

Not every experiment needs a control group — but every experiment needs **comparison**.

Placebos and the Placebo Effect

A **placebo** is a treatment with **no active ingredient** that is otherwise identical to the real treatment.

The **placebo effect** is the difference between the average response to a placebo and the average response to **no treatment at all**.

People genuinely feel better after taking a sugar pill they believe is medicine. The effect is real and it is measurable — which is exactly why a new drug has to beat a **placebo**, not just beat nothing.



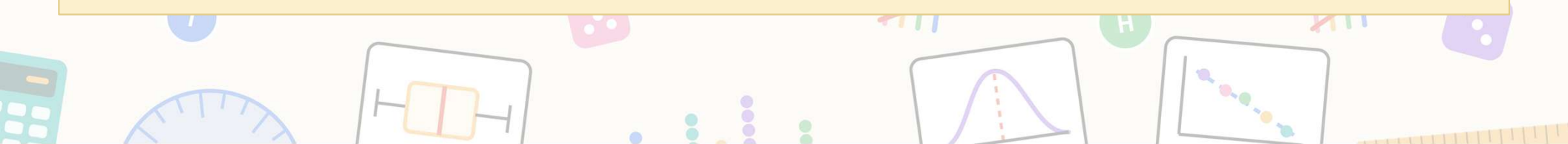
Blinding

In a **single-blind** (single-masked) experiment, the **participants** do not know which treatment they are getting — or the **researchers** who interact with them do not — but not both.

In a **double-blind** (double-masked) experiment, **neither** the participants **nor** the researchers who interact with them know which treatment each participant is getting.

Blinding protects the study twice: participants cannot talk themselves into feeling better, and researchers cannot unconsciously treat or score one group differently.

When the units are seeds, mice, or machines they are blind automatically — so the study is double-blind if the **researchers** do not know either.



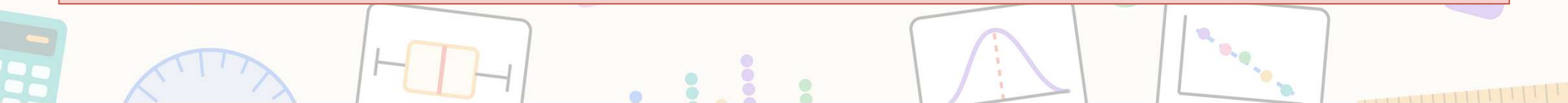
2. Random Assignment

Random assignment means experimental units are assigned to treatments using a **chance process**.

Its purpose is to create treatment groups that are as **similar as possible** with respect to every extraneous variable — the ones you thought of and the ones you did not.

If the groups start out equivalent, the **only** systematic difference between them is the treatment. That is what licenses a **cause-and-effect** conclusion.

Random **selection** → who the results apply to. Random **assignment** → whether you may say “caused.”



Describing a Random Assignment

Twenty students volunteer for an experiment comparing **caffeinated cola** and **caffeine-free cola** on pulse rate. Describe how to randomly assign 10 students to each treatment.

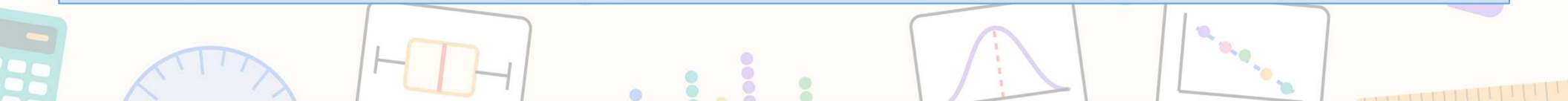


With slips of paper

Write “A” on 10 slips of paper and “B” on 10 more. Shuffle them and hand one to each volunteer. The “A” students drink the caffeinated cola; the “B” students drink the caffeine-free cola.

With technology

Label the students 1 to 20. Use a random number generator to produce 10 different integers from 1 to 20, ignoring repeats. Those students drink the caffeinated cola; the remaining 10 drink the caffeine-free cola.



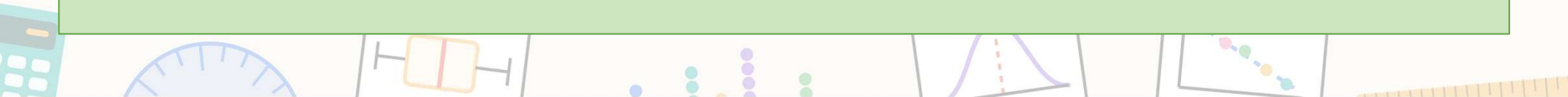
AP[®] Exam Tip: Stay Away From Coin Flips

Suppose each of the 20 students flips a coin: heads → caffeine, tails → caffeine-free. As long as **all 20** flip, that really is a completely randomized design — the groups just will not be exactly 10 and 10.

The trouble starts when you try to **force** equal groups: “keep flipping until one group has 10, then dump everyone left into the other group.”

Now the last two students in line are almost certain to land in the **same** group — but in a genuine completely randomized design they should have only a 50% chance of that. The assignment is no longer random.

Just describe the slips-of-paper method or the random number generator. They always work.



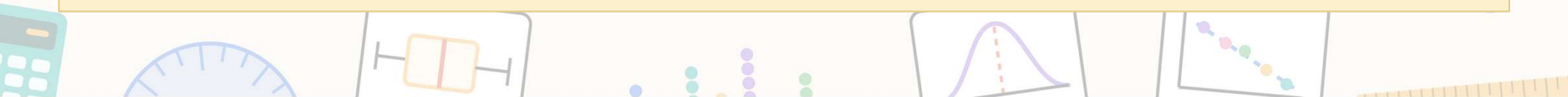
3. Replication

Replication means using **enough experimental units in each treatment group** to tell a real difference between treatments apart from ordinary chance variation in the random assignment.

CAUTION

In statistics, replication means “**use enough units.**” In other sciences it means “repeat the whole study somewhere else.” On the AP[®] exam, use the statistics meaning.

Two students per group can produce a big difference purely by luck. Two hundred per group cannot.



4. Direct Control

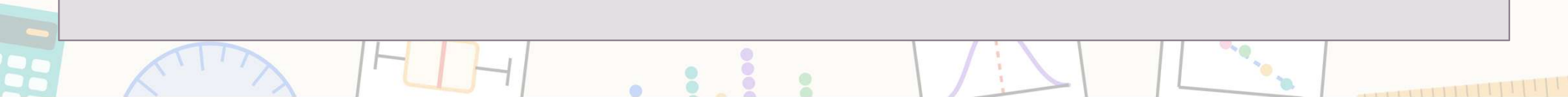
An **extraneous variable** is a variable that affects the response but is not the explanatory variable you are studying.

Direct control means keeping those variables **the same for every experimental unit**.

Direct control does two jobs

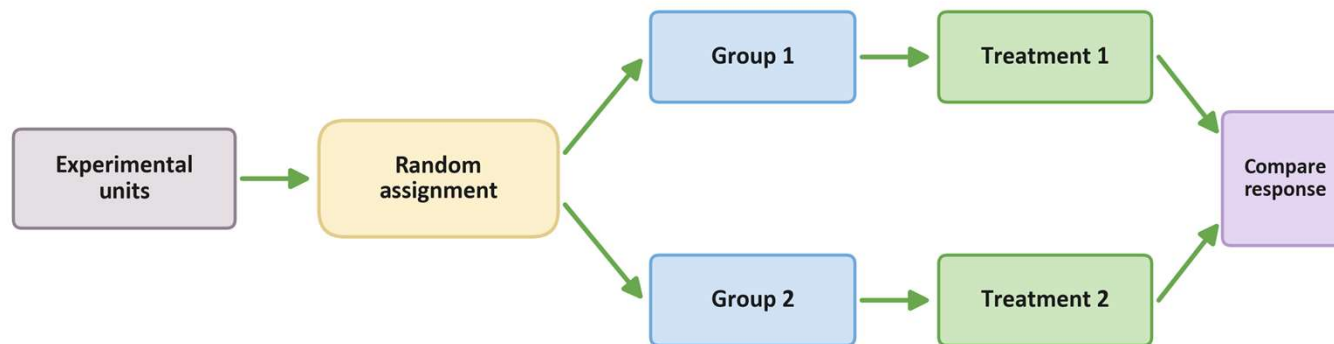
- **Prevents confounding** — a variable held constant cannot explain away your result.
- **Reduces variability** in the response — which makes a real treatment effect easier to see.

In the Trikafta cystic-fibrosis trial, every patient took the same **dose**, at the same **times of day**, with **food**, for the same **24 weeks**. None of those can be confounded with the drug, because none of them varied.



Completely Randomized Design

In a **completely randomized design**, treatments are assigned to experimental units **completely at random**, with no restrictions.

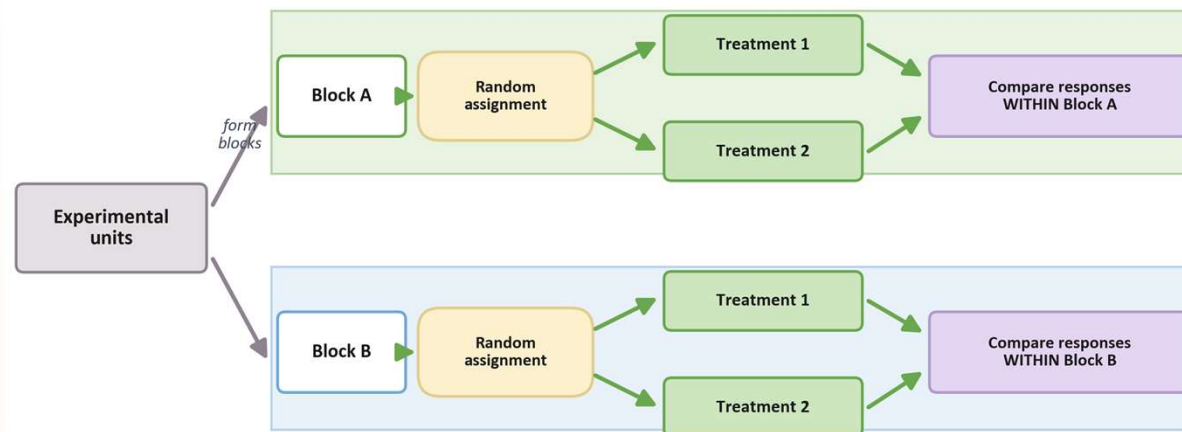


Groups do not have to be the same size — but there are good statistical reasons to make them close.

Randomized Block Design

A **blocking variable** is a source of extraneous variation in the response.

In a **randomized block design** we first group units into **blocks** that are similar with respect to that variable, then randomly assign **all** the treatments **separately within each block**.



Why Block?

Blocking **separates** the variation caused by the blocking variable from the rest of the noise, so treatments can be compared **within** blocks — where that variable is not changing.

Two students want the fastest route to school. They will test three routes over 15 days. The **morning** commute is rush hour; the **afternoon** commute is not.

What should they block on, and why?

Block on **time of day**. Randomly assign 5 mornings to each route, and separately assign 5 afternoons to each route.

Otherwise a route could look slow simply because it happened to be tested in the morning more often.

Blocking is like stratifying — “similar within, different between” — but for **experiments** instead of samples.

So: Control It, Block It, or Randomize It

You know about it
and you **can** hold it
constant

→ **direct control**.

You know about it
but **cannot** hold it
constant

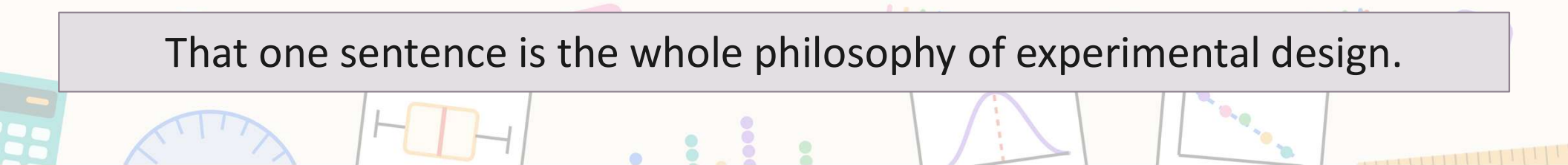
→ **block** on it.

You do not know
about it at all

→ **random
assignment** balances it
out.

**Control what you can, block on what you cannot control, and
randomize to balance out the rest.**

That one sentence is the whole philosophy of experimental design.



AP[®] Exam Tip: Do Not Mix the Two Vocabularies

SAMPLING words

Population, sample, strata, clusters, random **selection**, bias, generalize.

EXPERIMENT words

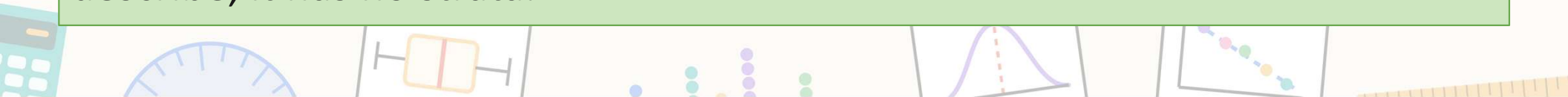
Experimental units, treatments, blocks, random **assignment**, confounding, cause and effect.

You will lose credit for sentences like these:

“Use a randomized block design to select the sample for this survey.”

“Randomly assign the treatments using three different strata.”

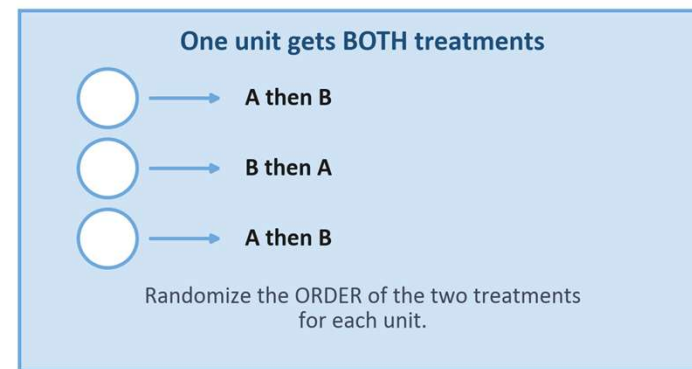
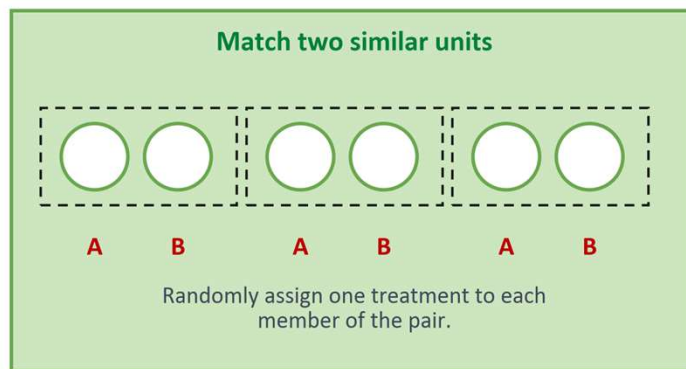
If the study has no treatments, it has no blocks. If it has no population to describe, it has no strata.



Matched Pairs Design

A **matched pairs design** is a randomized block design with exactly **two treatments**, where the blocks have size 2.

Two ways to build the pairs

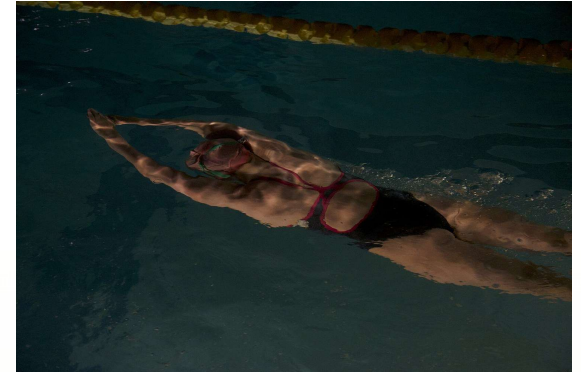


Pairing removes the variation caused by the differences between individuals — the single biggest source of noise in most experiments.

Example: Matched Pairs

Drag suits are swimsuits with extra fabric that increases resistance. Two students want to know whether a drag suit really slows down a 50-metre freestyle compared with a regular suit. There are **24 swimmers** on the team.

Describe a matched pairs experiment.



Have **each swimmer swim with both suits** — each swimmer is a pair. Some swimmers are simply faster than others, and pairing removes that variation.

Number the swimmers 1 to 24 and use a random number generator to pick 12 different integers. Those 12 swim the **drag suit first**, then rest, then the regular suit. The other 12 swim in the **opposite order**.

Record each swimmer's two times and compare them **within** each swimmer.

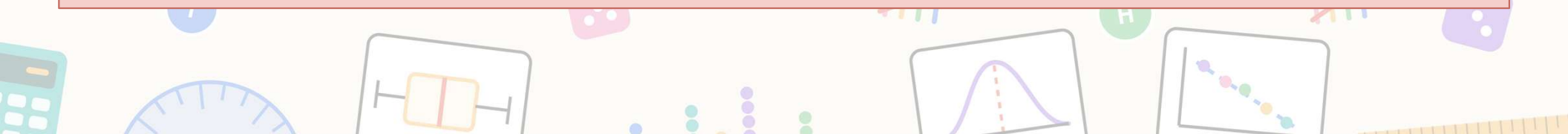
Even when a subject is compared to themselves, you must still describe the **random assignment of order** completely.

Which Design Should I Use?

Match the design to the situation

- Units are fairly similar, or you cannot identify a big source of variation → **completely randomized design**.
- You can identify a variable that strongly affects the response and can group units by it → **randomized block design**.
- There are exactly two treatments and units can be paired (or each unit can take both) → **matched pairs**.

Justifying a design means naming the **specific source of variation** in *this* study that your design handles. “Because blocking is better” earns nothing.



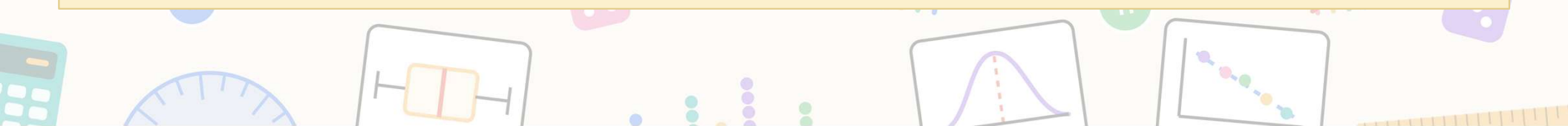
Inference for Experiments

Well-designed experiments let us make **cause-and-effect** conclusions — but only when the results are convincing enough.

When one group does better, there are only two explanations:

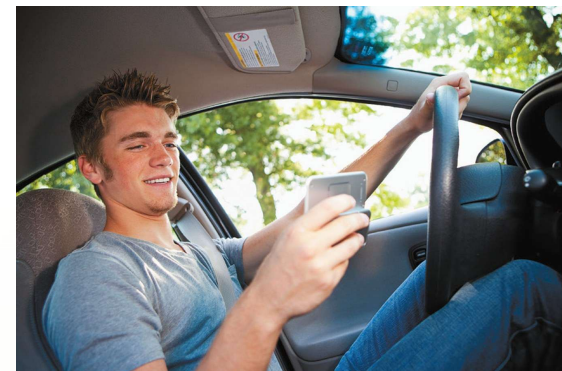
1. The treatment has no effect, and the difference we saw is just **chance variation in the random assignment**.
2. The treatment **causes** a difference in the average response.

When the observed result is **too unusual to be explained by chance alone**, we call it **statistically significant**.



Is a Phone Worse Than a Passenger?

Is talking on a cell phone while driving more distracting than talking to a passenger? Scientists at the University of Utah randomly assigned **48 undergraduates**: half drove a simulator while talking on a phone, half while talking to a passenger. The response was whether the driver stopped at a rest area specified in advance.



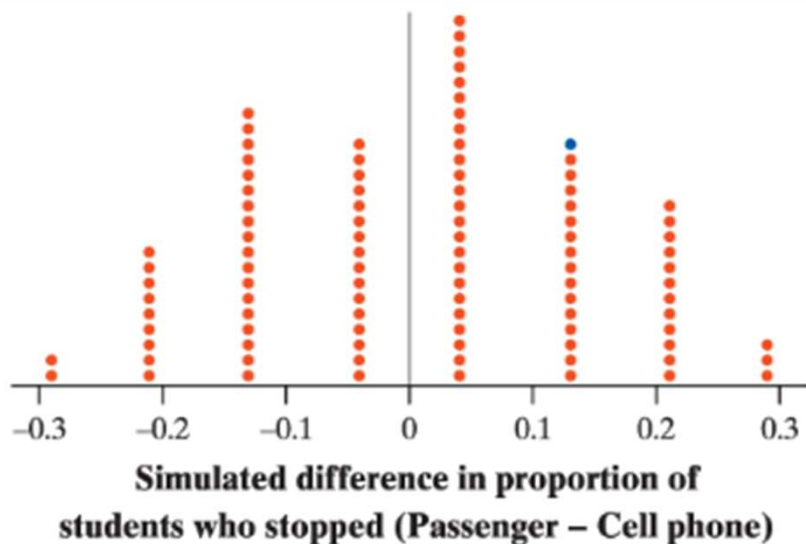
		Treatment		
		Cell phone	Passenger	Total
Response	Stopped at rest area	12	21	33
	Didn't stop	12	3	15
	Total	24	24	48

(a) Calculate the difference (Passenger – Cell phone) in the proportion who stopped.

$$21/24 - 12/24 = 0.875 - 0.500 = \mathbf{0.375}$$

Could Chance Alone Do That?

We simulated **100 random assignments** of the same 33 “stoppers” and 15 “non-stoppers” into two groups of 24, assuming the type of distraction makes **no difference**. Each dot is one simulated difference in proportions.

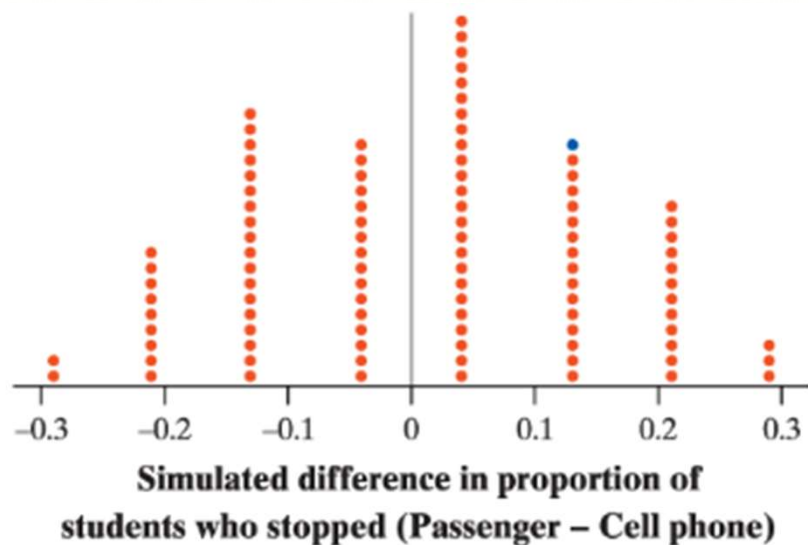


This dotplot has a name: a **randomization distribution** — the distribution of a statistic produced by repeatedly reassigning the responses to treatment groups by chance.

(b) The three dots at 0.29 mean: in three of the 100 simulated random assignments, the difference came out to 0.29 — with no real treatment effect at all.

Statistically Significant?

(c) Use the simulation to decide whether the observed difference of **0.375** is statistically significant.



A difference of 0.375 or larger **never happened** in 100 simulated random assignments.

So the result is **statistically significant**: it is extremely unlikely to get a difference this big from chance variation in the random assignment alone.

Because treatments were **randomly assigned**, we can conclude the phone *caused* more drivers to miss the rest area.

Scope of Inference for Experiments

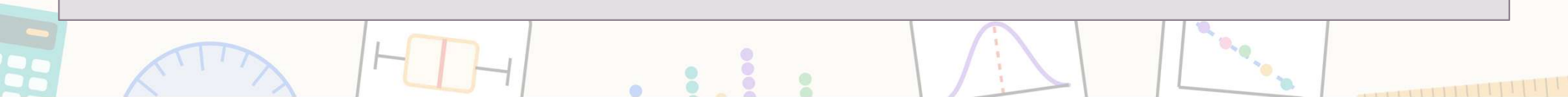
Random **assignment** of treatments allows a **cause-and-effect** conclusion, because it reduces the potential for confounding variables.

But it is often unethical or impossible to randomly **select** experimental units from a population. Most experiments run on **volunteers**.

So the honest conclusion for most experiments is:

“This treatment *caused* the difference **for people like the volunteers in this study.**”

Those 48 Utah undergraduates were not randomly selected from all drivers. The phone caused the difference — for undergraduates like them.



Scope of Inference: Putting It All Together

		Were units randomly ASSIGNED to treatments?	
		Yes (experiment)	No (observational)
Were units randomly SELECTED?	Yes	Generalize: YES Cause & effect: YES	Generalize: YES Cause & effect: NO
	No	Generalize: NO Cause & effect: YES	Generalize: NO Cause & effect: NO

Ask both questions, every time. **Who were the units?** (selection) and **What was done to them?** (assignment).

The One Idea to Take From Unit 1B

Random **SELECTION** decides **WHO** your answer is about.

Random **ASSIGNMENT** decides whether you may say **CAUSED**.

Neither one is optional, neither one substitutes for the other, and no amount of extra data replaces either.

Next unit: what to do once the data are in.

