1.1

The Foundation of Statistical Thinking

STATISTICS

Data Sets

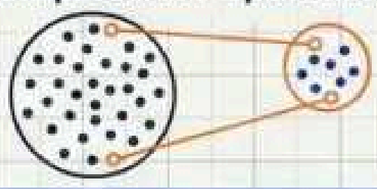
(What data is made of)

Variables



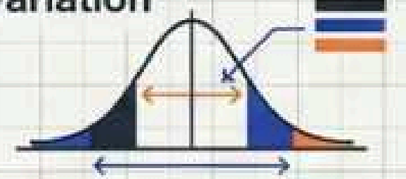
(Separating different kinds of characteristics)

Samples vs. Populations



(How a part relates to the larger group)

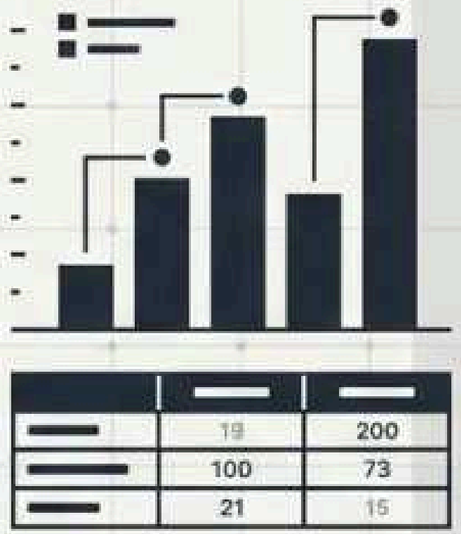
Variation



(The concept that runs through the entire subject)

1.2

The Two Broad Jobs of Statistics



	19	200
	100	73
	21	15

Describe

Describe the data we have.

Uses summaries and graphs to organize the present information.



Draw Conclusions

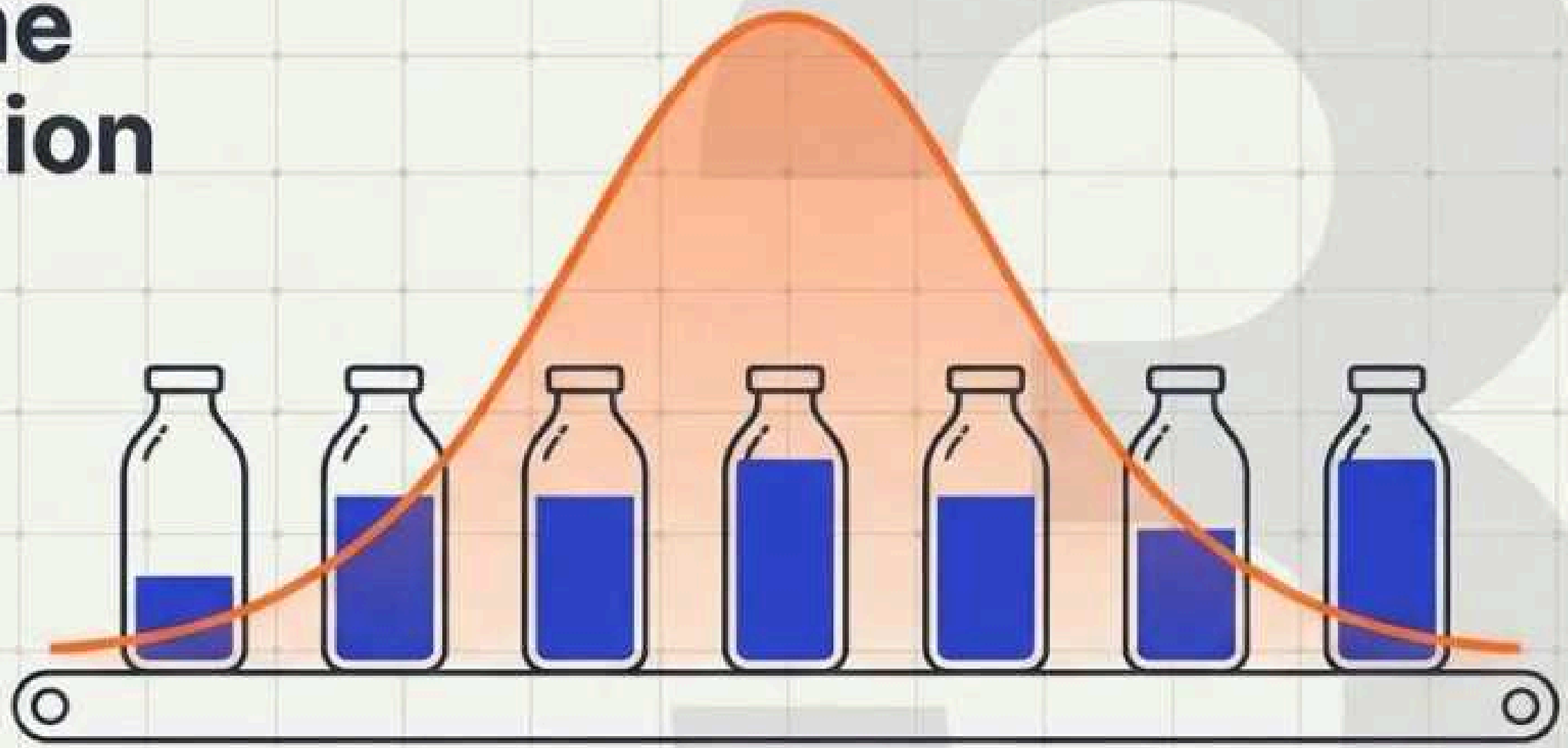
Draw conclusions about unmeasured things.

Infers insights beyond the immediate data and explicitly states how sure we can be about those conclusions.

1.3

The Heart of the Subject: Variation

No two people have exactly the same height. The same person weighs a little different each morning. The same machine fills each bottle with slightly different amounts.



“If everything were identical, one measurement would settle every question and statistics would not exist.”

Because things vary, we need methods to find patterns inside the variation.

1.4

The Anatomy of a Data Set

Variable Any characteristic that takes different values from one individual to another. **Each column is a variable.**

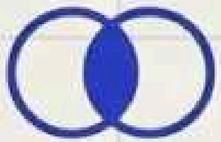
Name	Age	Major	Height
Alice	20	Biology	65
Bob	21	History	70
Charlie	19	Math	68

Individuals
The people, animals, or objects the data describes. Each row is an individual.

A class roster with name, age, major, and height contains four variables.

1.5

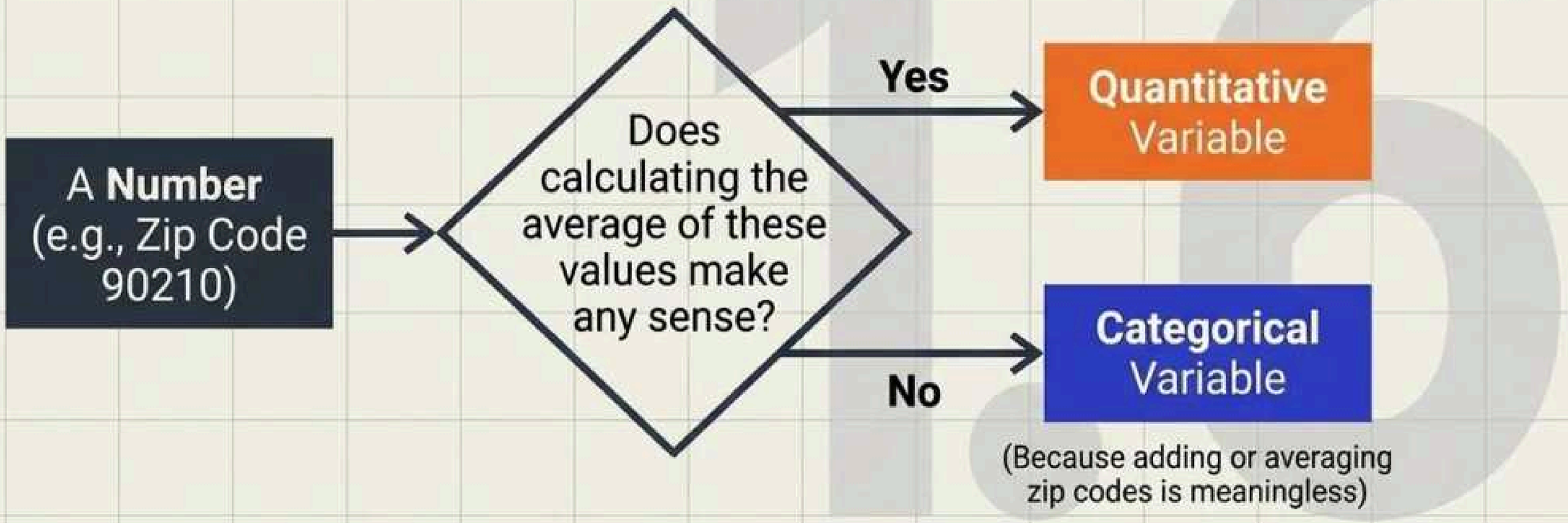
Two Main Types of Variables

Categorical	Quantitative
	
Places each individual into a group or category.	Takes numerical values where arithmetic makes sense.
Eye color, Major, Yes/No.	Height, Price, Age.

Why it matters: The type determines which graphs and summaries are legitimate.

1.6

Beware of Numbers as Labels



1.7

Splitting Variables Further

Quantitative Variables

Discrete



Takes separate counted values.

Example:
Number of siblings.

Continuous



Can take any value in an interval.

Example:
Time to run a mile.

Categorical Variables

Ordinal

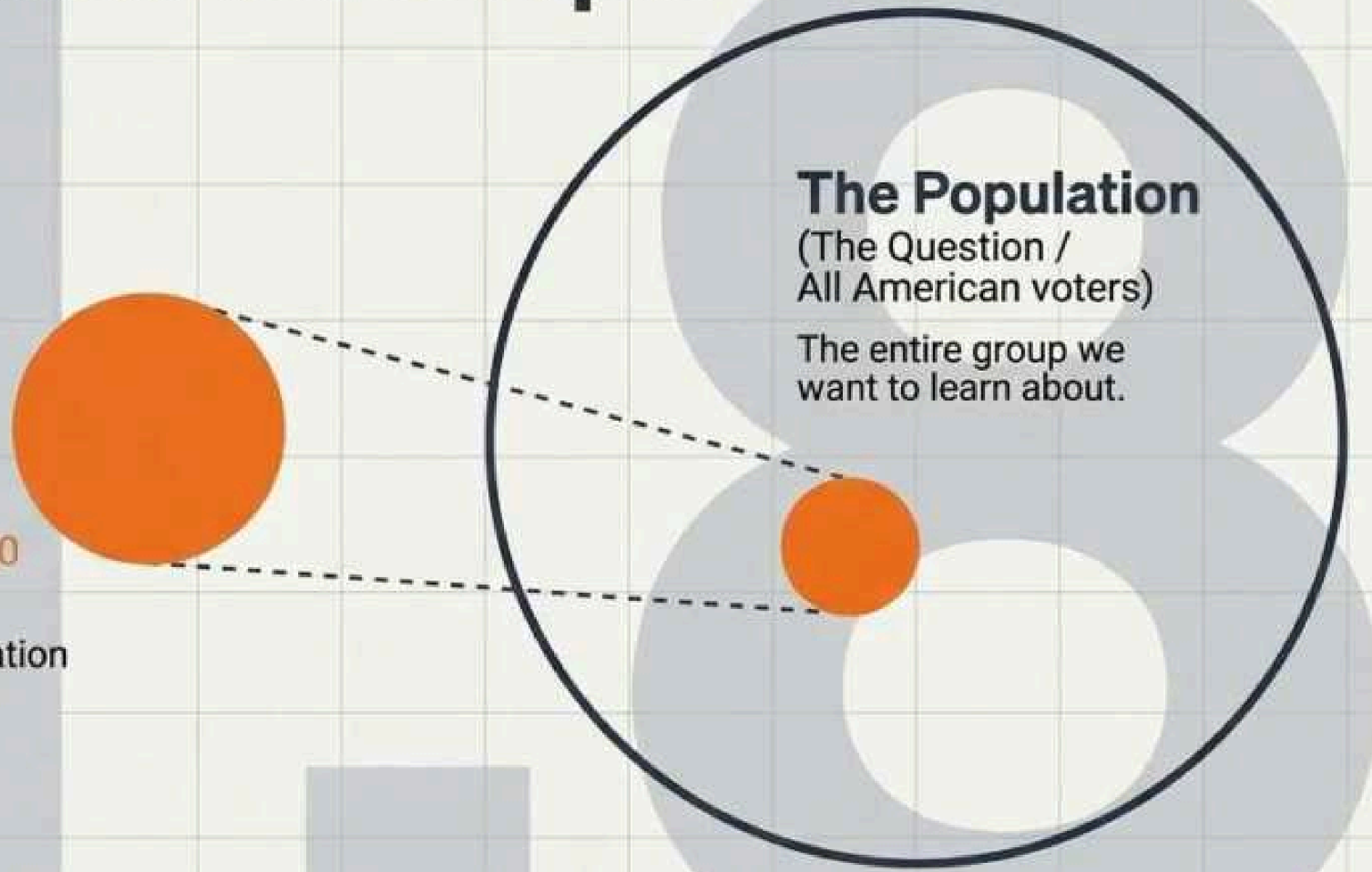
Categories with a natural order.

Example: Ratings from "Poor" to "Excellent".



1.8

The Key Distinction: Population vs. Sample



The Population
(The Question / All American voters)

The entire group we want to learn about.

The Sample
(The Evidence / 1,000 surveyed voters)

The part of the population we actually examine.

1.9

The Vocabulary of Measurement

Population → Parameter

Definition: A number that describes a population.
Example: The average income of all households in a country.

Sample → Statistic

Definition: A number that describes a sample.
Example: The average income of 1,000 sampled households.

1.9

1.10 Why Not Measure Everyone? (The Limits of a Census)



Cost & Time

It simply costs too much or takes too long to measure every individual.

1.10



Destructive Testing

Measuring sometimes destroys the item (e.g., testing how much force breaks a factory part).



Unreachable/Unfinished

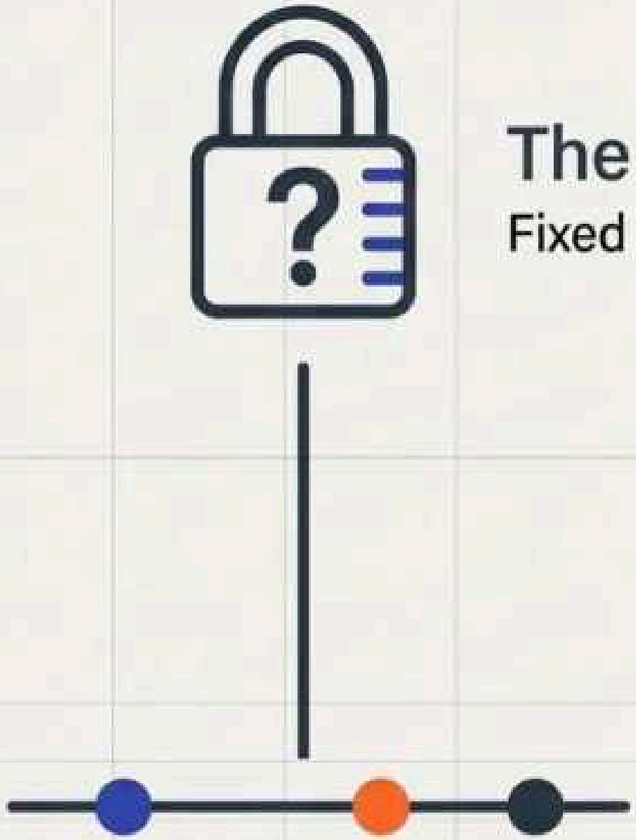
The population might not be accessible, or hasn't even been fully created yet.



Inevitable Misses

Even a census, which attempts to count absolutely everyone, always misses some people.

1.11 The Central Difficulty: Sampling Variability



The Parameter

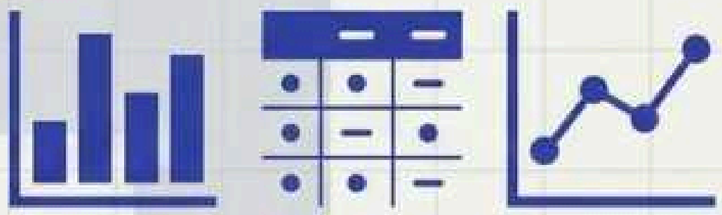
Fixed but usually unknown.

The Statistic

Known, but changes from sample to sample.

Two different samples from the same population will give two different averages. This variability is not a mistake—it is a fact of life.

1.12 The Two Branches of Statistics



Descriptive Statistics

Summarizes the data currently in hand using graphs, tables, and numbers (like averages).

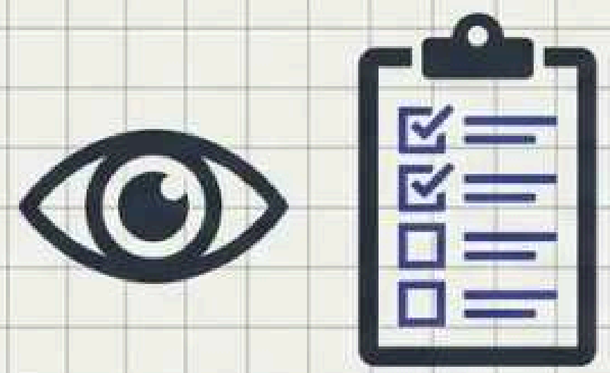


Inferential Statistics

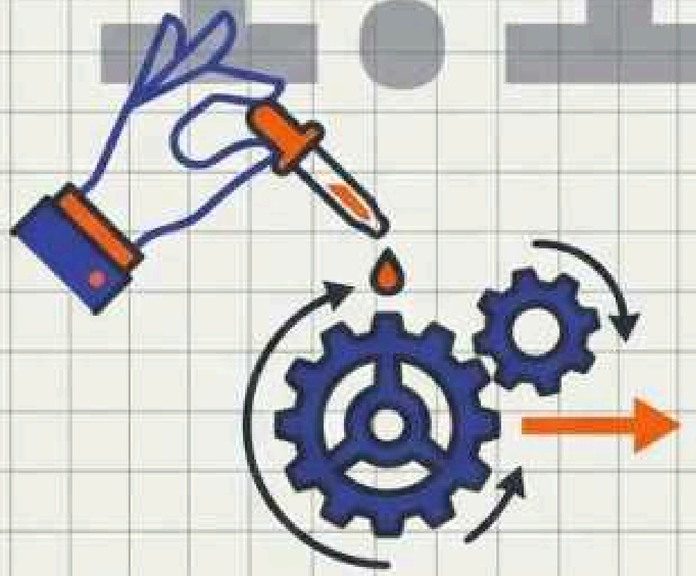
Uses a sample to draw conclusions about a wider population, crucially attaching a measure of uncertainty.

1.13

Data Gathering: Observation vs. Experiment



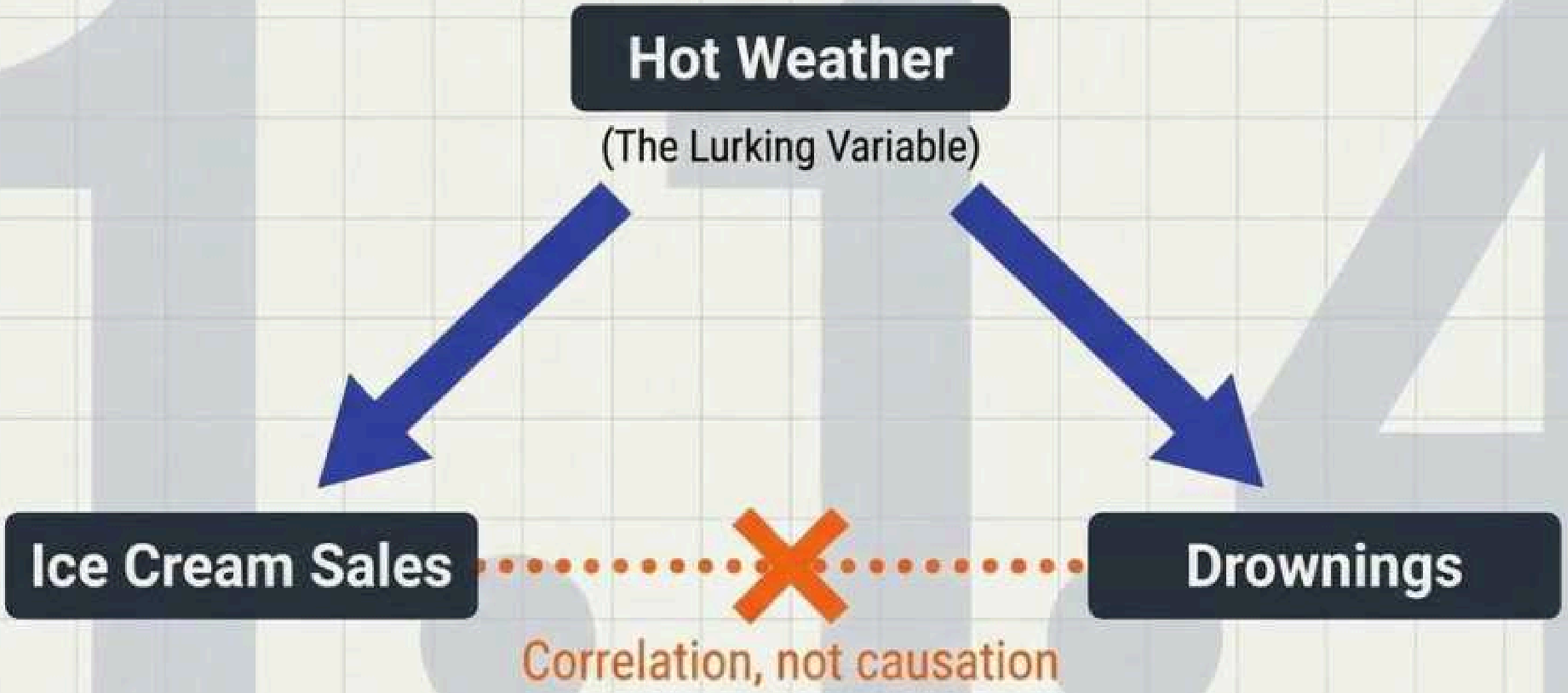
Observational Study
Watch and record what happens without interfering.



Experiment
Deliberately impose a treatment on individuals to observe the response.

The Decisive Difference: Only a well-designed experiment can show that one thing causes another.

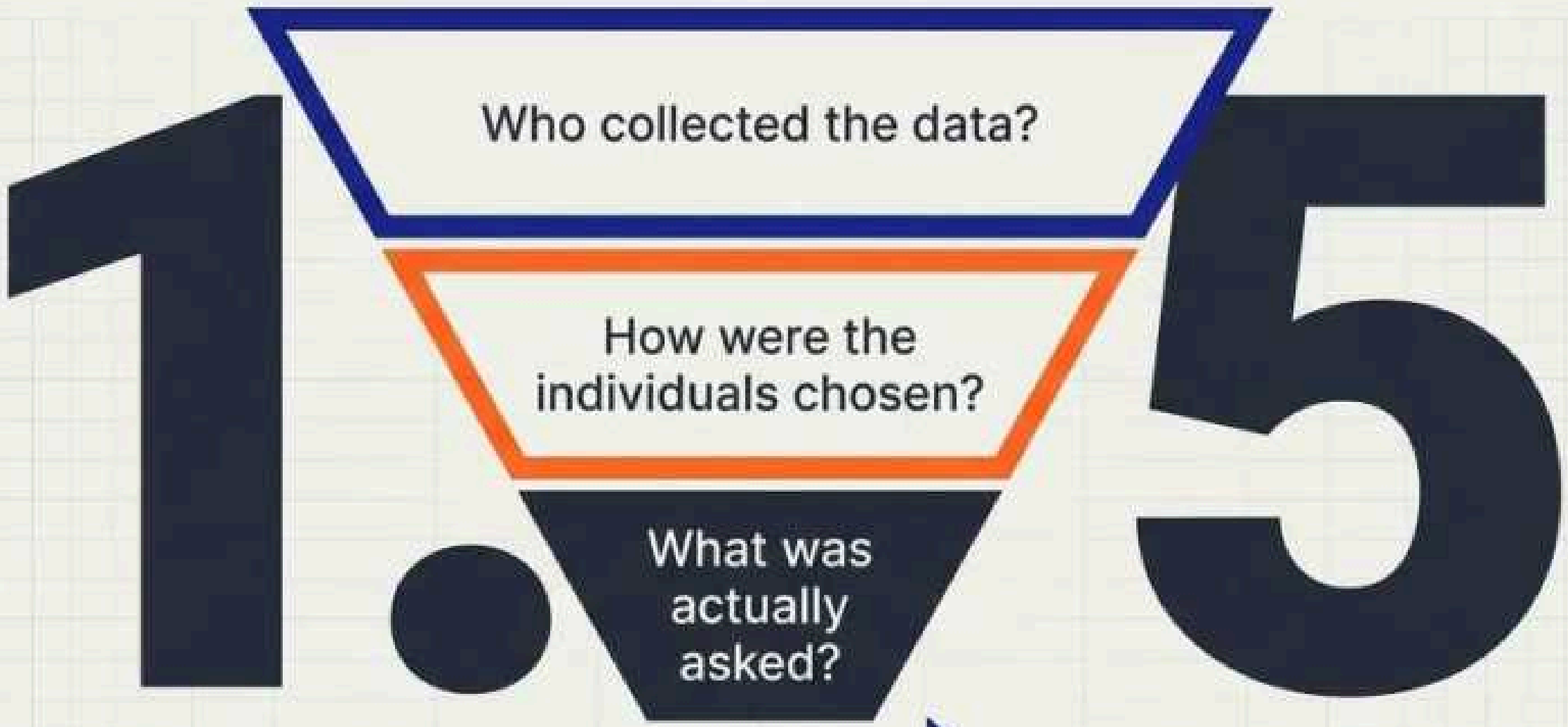
1.14 The Danger of Lurking Variables



Observation alone cannot show cause. Hidden factors can affect results. When a lurking variable cannot be separated from the one we are studying, they are confounded.

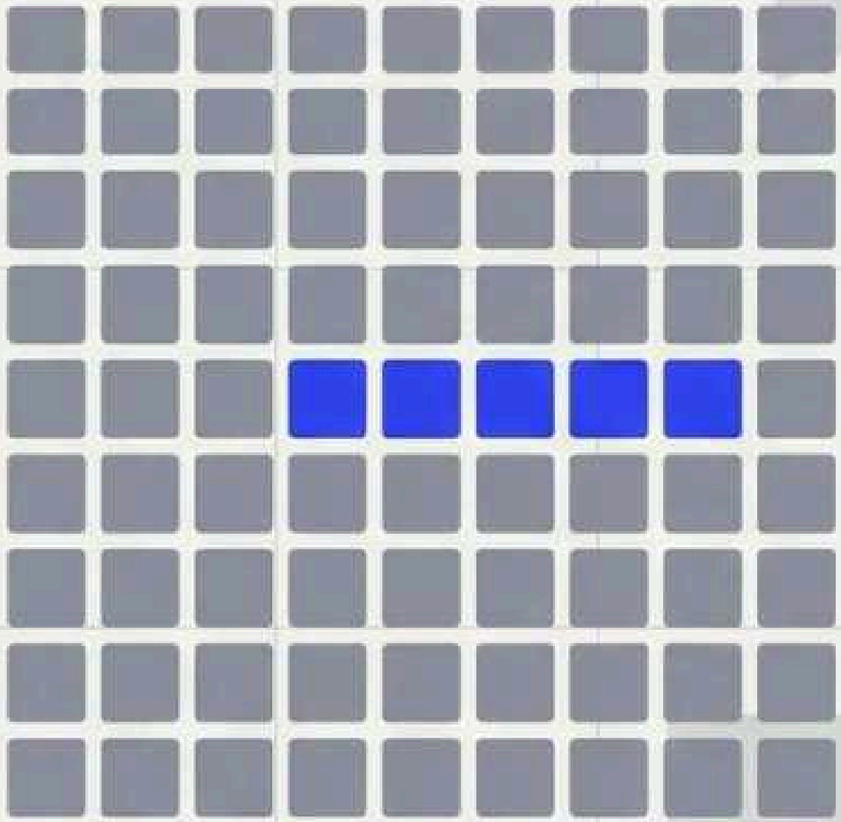
1.15

Data Provenance (Trusting the Numbers)



"A precise answer computed from biased data is a precise answer to the wrong question, and no later calculation can repair it."

1.16 Human Intuition vs. Randomness



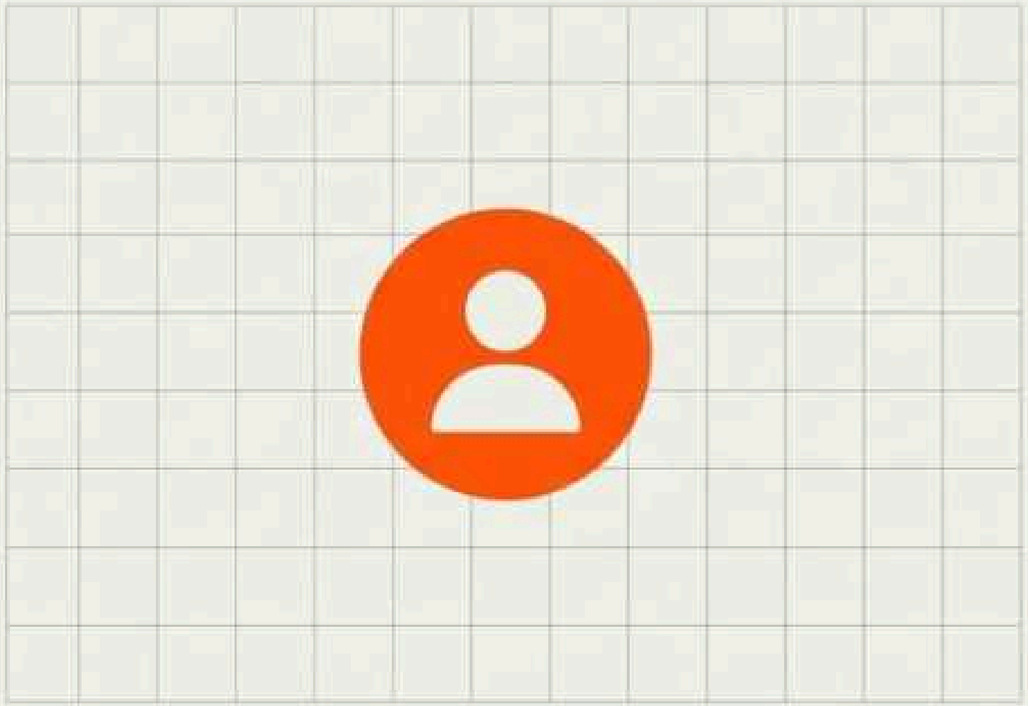
The Intuition:
People avoid long runs when writing fake coin flips because they look suspicious.

The Reality:
In 100 flips of a fair coin, a run of 6 or more heads or tails happens about 81% of the time.

Takeaway:
Patterns appear by chance far more often than intuition expects.

1.17 Everyday Thinking vs. Statistical Thinking

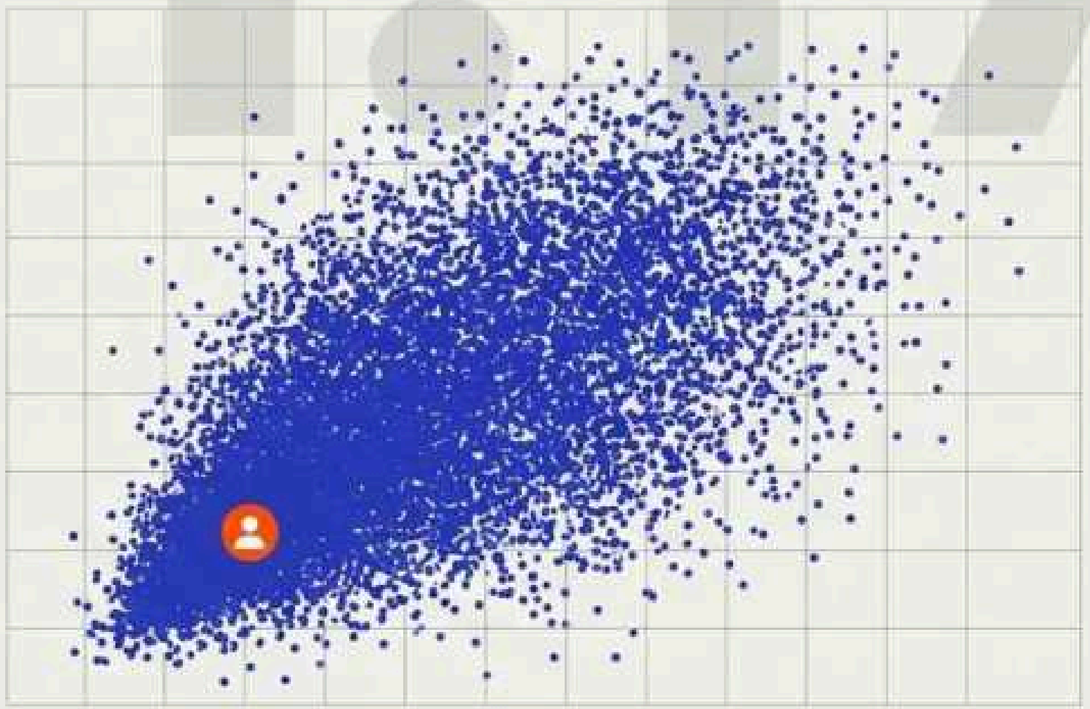
Everyday Thinking



Leans on vivid, single stories.

(Example: One person who smoked all his life and lived to ninety is a memorable story.)

Statistical Thinking



Asks about groups, rates, and comparisons.

(Example: The rate of lung disease among smokers and nonsmokers can be checked against the whole picture.)

1.18

Embracing Disciplined Uncertainty



Statistics cannot give **exact certainty**. It gives a disciplined way to talk about uncertainty using terms like “about”, “probably”, and “give or take”.

A good statement does not say a poll is exactly right; it says the true value is probably within a stated distance. Honest uncertainty is a strength of the method.

1.19

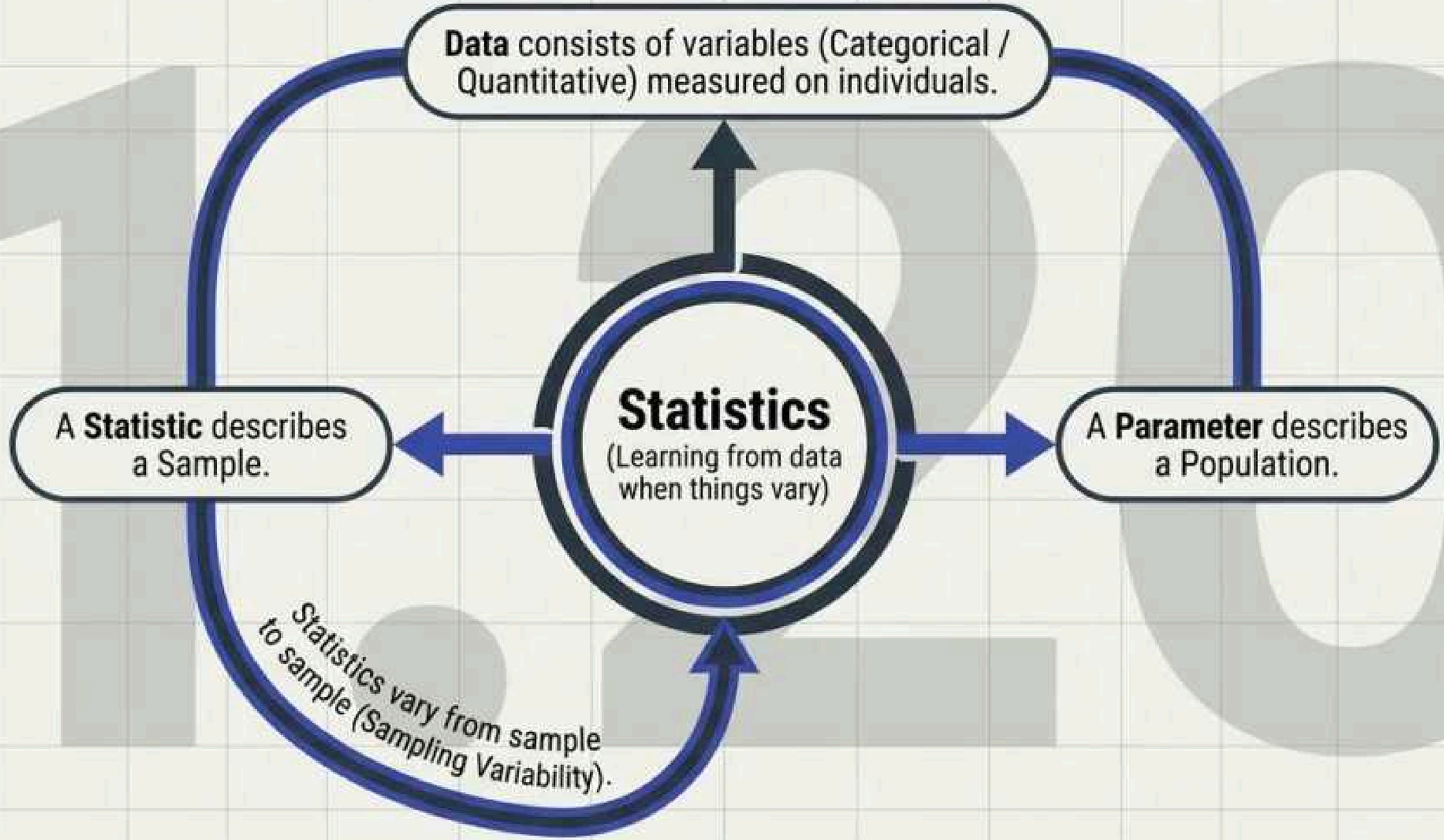
The News Checklist

- ☐ Who was measured, and how were they chosen?
- ☐ What exactly was measured?
- ☐ What is being compared to what?
- ☐ What might be missing? (e.g., a lurking variable, a hidden group, or context).

These four questions catch most misleading claims.

1.20

Lecture Summary: Learning from Variation



Next: How to describe a single variable well.