CHAPTER 1

Introduction to Data

What is Statistics?

Statistics is the science of learning from data.

- Examples include, weather forecasting, stock market prediction, opinion polls in politics, market research to estimate demand for a new product

- Statistics involve: Collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical information

- Statisticians are trained in statistical science: collect numerical information in the form of data, evaluate it, and draw conclusions

1.2 Data Basics

Observations, Variables, and Data Matrices

Cases (or observational units) are the objects described by a set of data. Cases may be customers, companies, experimental subjects, or other objects

A variable is a special characteristic of a case.

Data matrix is a convenient and common way to organize data. Each row of the matrix corresponds to a unique case (observational unit), and each column corresponds to a variable.

Example 1:

A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey:

Gender: What is your gender?

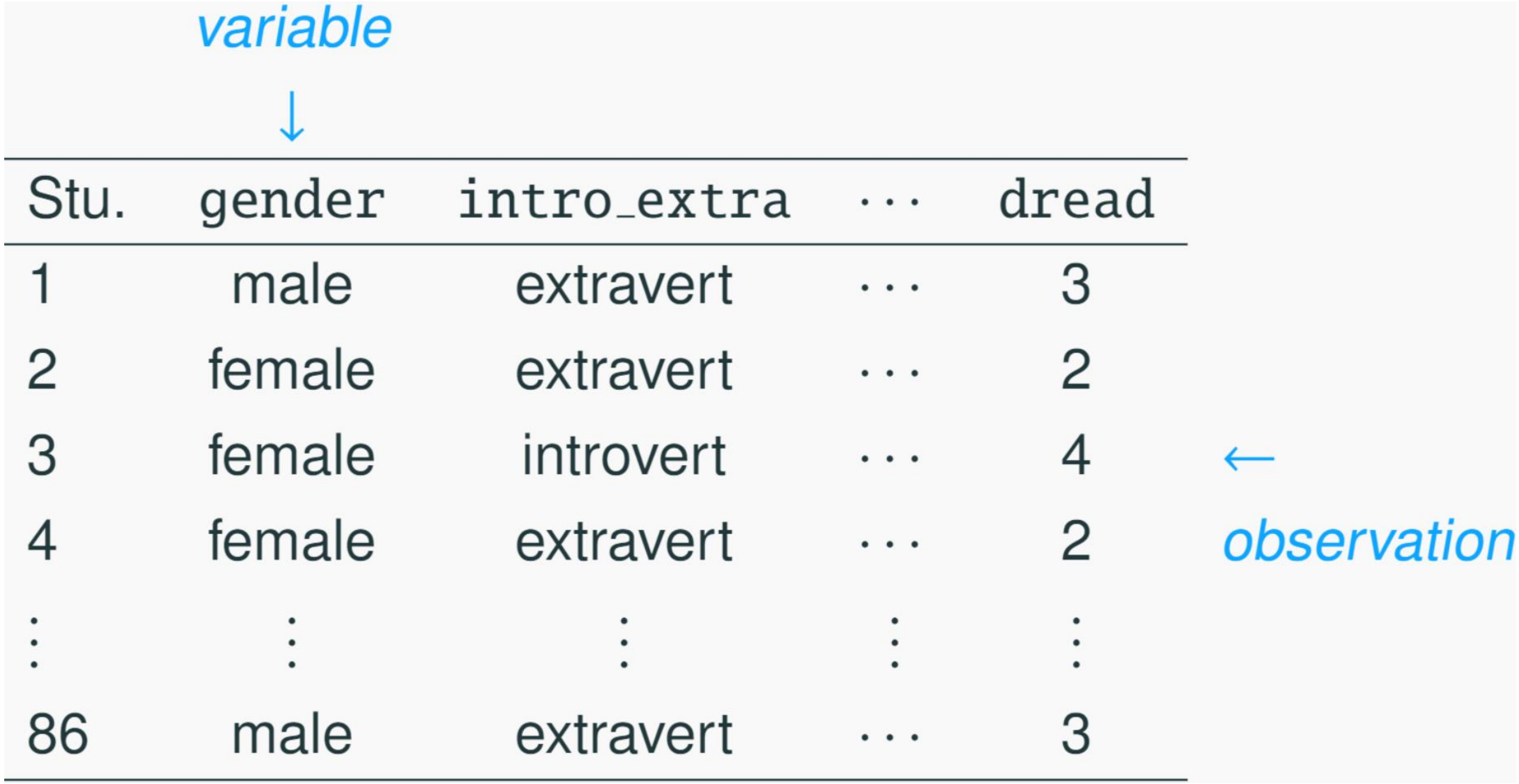
Intro-extra: Do you consider yourself introverted or extraverted?

Sleep: what time do you usually go to bed?

Countries: How many countries have you visited?

Dread: On a scale of 1-5, ho much do you dread being here?

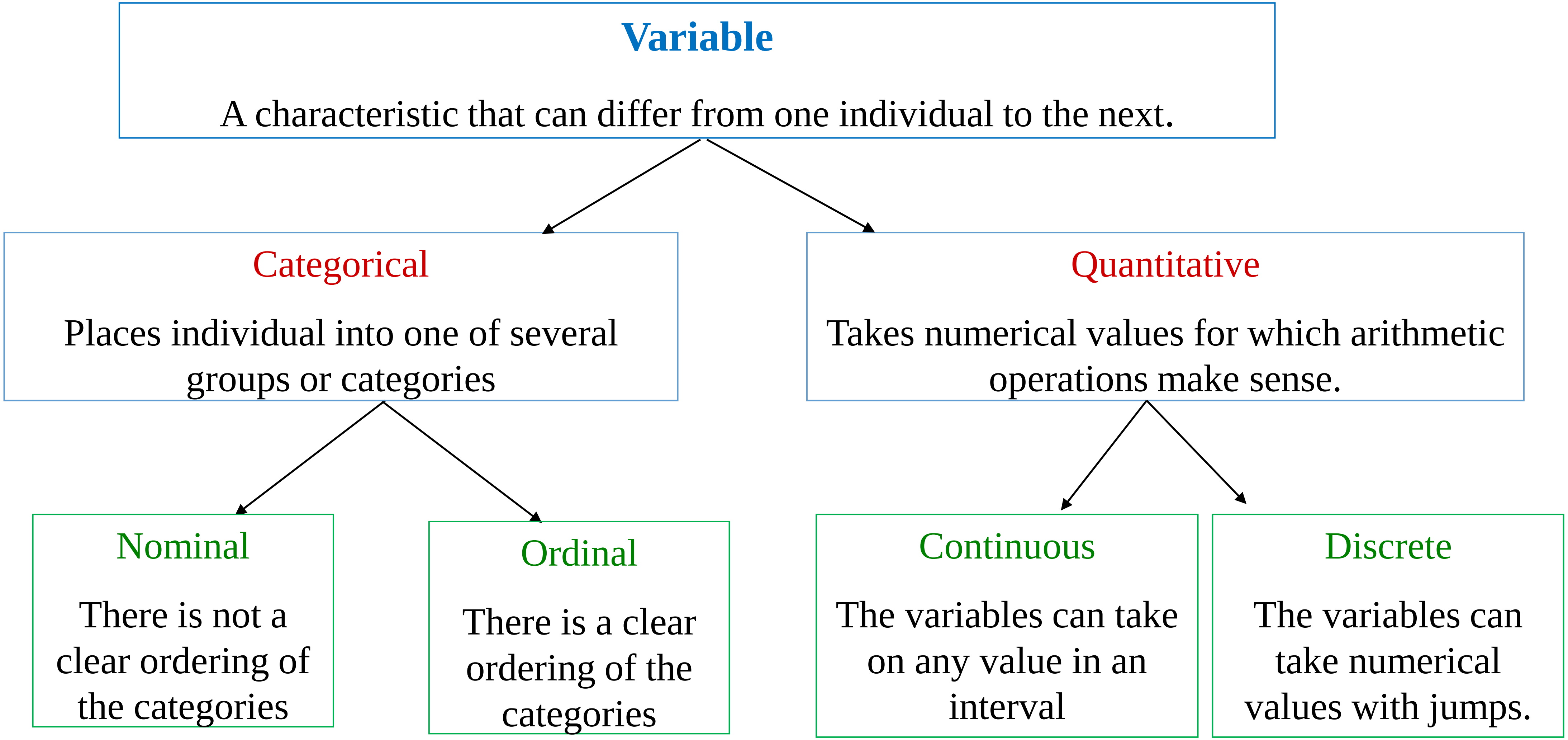
Data collected on students in a statistics class on a variety of variables:



variable
observation

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

Type of Variables



Variable

A characteristic that can differ from one individual to the next.

Categorical

Places individual into one of several groups or categories

Quantitative

Takes numerical values for which arithmetic operations make sense.

Nominal

There is not a clear ordering of the categories

Ordinal

There is a clear ordering of the categories

Continuous

The variables can take on any value in an interval

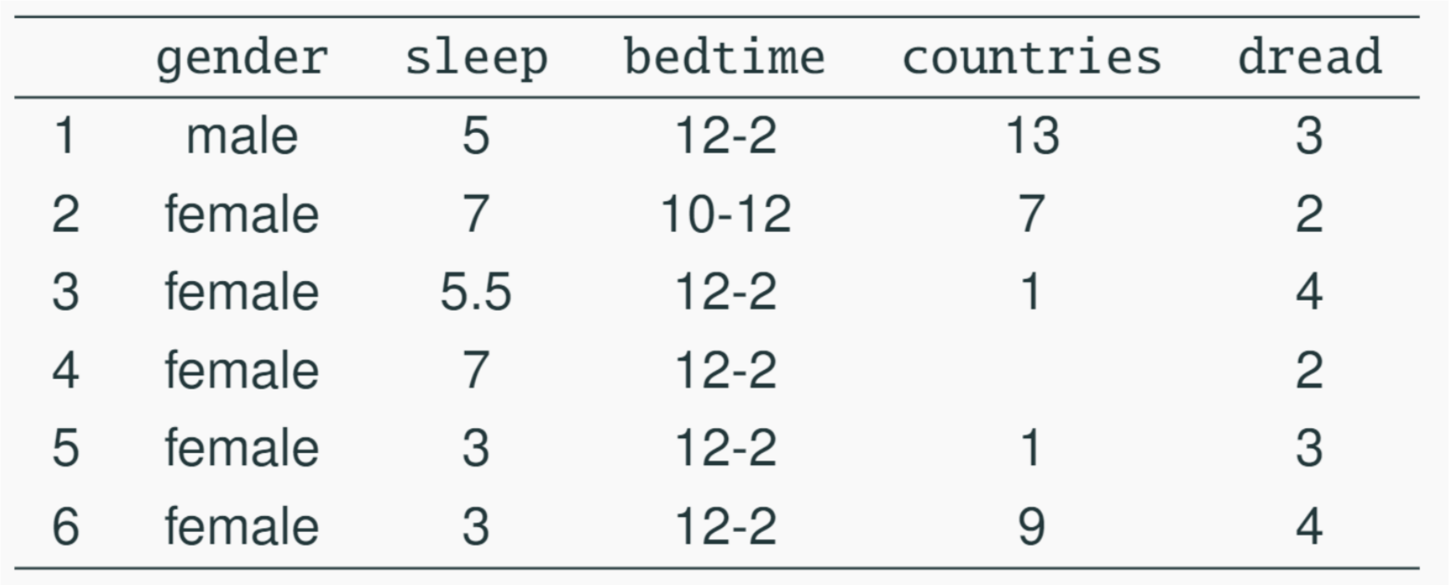
Discrete

The variables can take numerical values with jumps.

Example1 (cont.)

Classify each variable as either numerical or categorical?

- Gender:
- Sleep:
- Bedtime:
- Countries:
- Dread:



	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

Practice Problem

Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa, versicolor and virginica*). There were 50 flowers from each species in the data set.

a) How many cases were included in the data?

b) How many numerical variables are included in the data? Indicate what they are, and they are continuous or discrete.

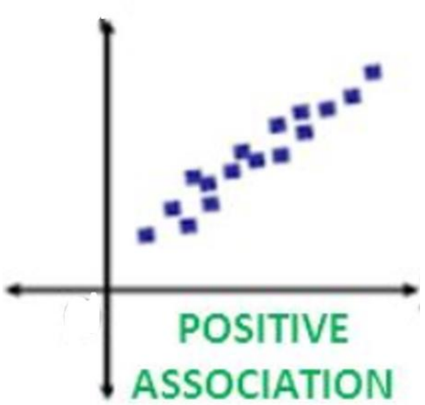
c) How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).

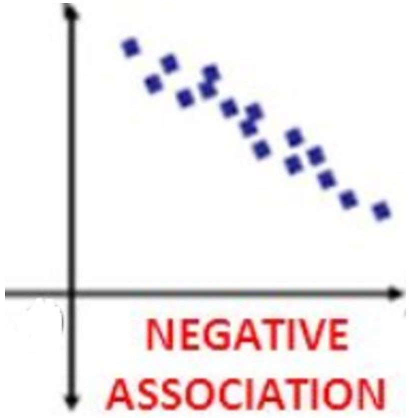
Relationship Between Variables

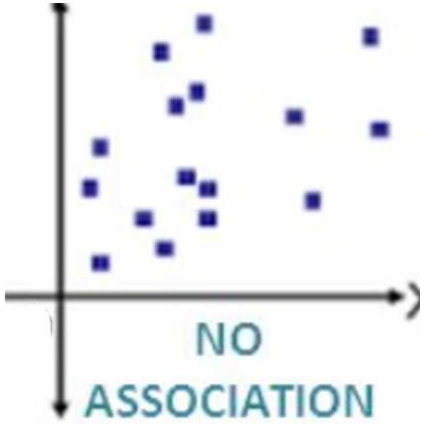
- When two variables show some connection with one another, they are called associated variables
  - Associated variables can also be called dependent variables and vice-versa.
- If two variables are not associated, there is no evident connecting between the two, then they are said to be independent

Scatterplot

Shows the relationship between two quantitative variables measured on the same individuals.

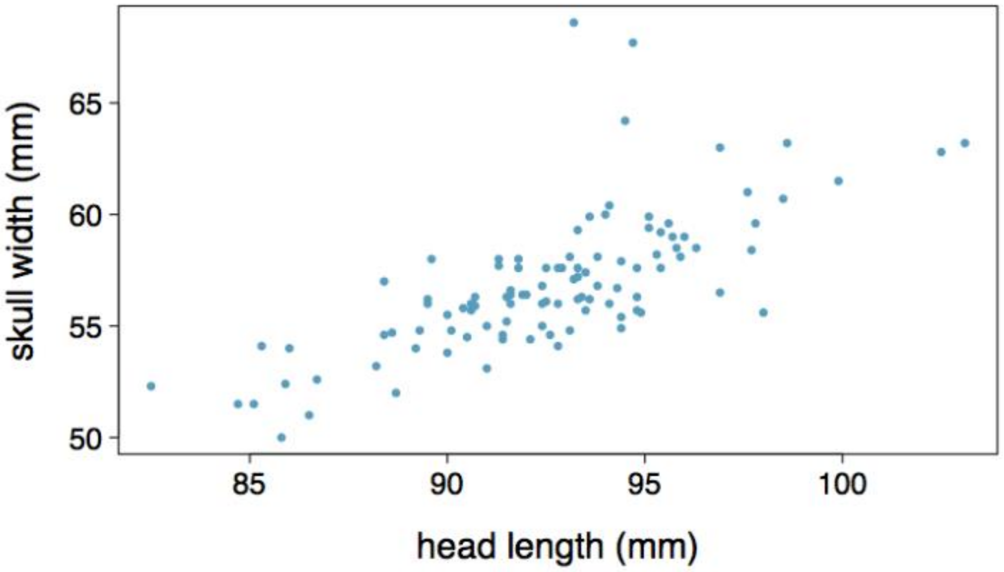






Example

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



a) There is no relationship between head length and skull width (the variables are independent).
b) Head length and skull width are positively associated.
c) Skull width and head length are negatively associated.
d) A longer head causes the skull to be wider.
e) A wider skull causes the head to be longer.

Explanatory and Response Variables

Response variable (y) – measures the outcome of a study

- Also known as the dependent variable

Explanatory variable (x) – Explains or causes a change in the response variable

- Also known as the independent variable

Observation VS. Experiment

Observational study observes individuals and measures variables of interest but does not attempt to influence the response.

- The purpose is to describe some group or situation.
- Example: Surveys, review medical or company records.

Experiment deliberately imposes some treatment on individuals to measure their response.

- The purpose is to study whether the treatment causes a change in the response.

Practice Problem

To assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between *1g* vitamin C, *2g* Vitamin C, or *3g* Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablet has identical appearance and packaging . The nurses who handed the prescribed pills the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four groups, and the placebo group had the shortest duration of symptoms.

1. Was this an experiment or an observational study? Why?

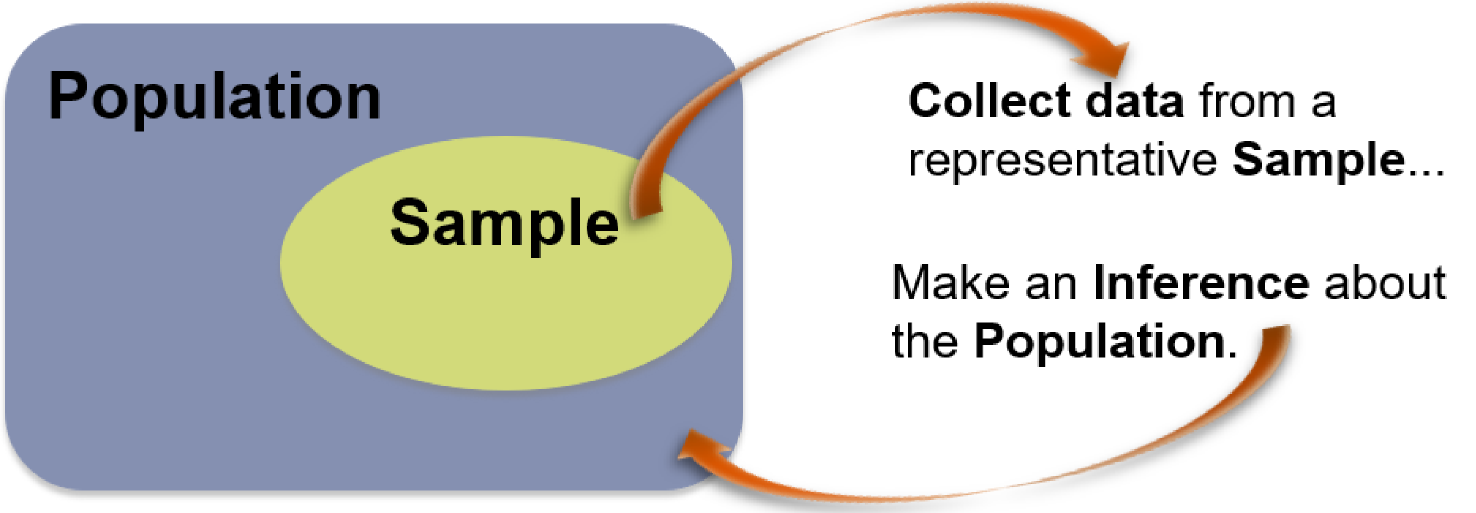
2. What are the explanatory and response variables in this study?

1.3 Sampling Principles and Strategies

Population and Sample

The population in a statistical study is the entire group of individuals about which we want information.

A sample is the part of the population from which we collect information.



Collect data from a representative **Sample**...

Make an **Inference** about the **Population**.

Examples:

Identify the population and sample:

1. What is the average mercury content in swordfish in the Atlantic Ocean?

2. Over the last 5 years, what is the average time to complete a degree for Duke undergrads?

Parameters and Statistics

As we begin to use sample data to draw conclusions about a wider population, we must be clear about whether a number describes a sample or a population.

- A *parameter* is a number that describes some characteristic of the population. In statistical practice, the value of a parameter is not known because we cannot examine the entire population.
- A *statistic* is a number that describes some characteristic of a sample. The value of a statistic can be computed directly from the sample data. We often use a statistic to estimate an unknown parameter.

Remember **s** and **p**: **s**tatistics come from **s**amples and

parameters come from **p**opulations.

Anecdotal Evidence

- Be careful of data collected in a haphazard fashion. Such evidence maybe true and verifiable, but it may only represent extraordinary cases.
- Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

Examples:

- A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
- I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.

Sampling from a Population

In general, we always seek to randomly select sample from a population. We pick sample randomly to reduce the chance we introduce biases.

- The design of a sample is biased if it systematically favors certain outcomes.

The most basic random sample is called simple random sample, consist of *n* cases from the population chosen in such a way that every case in the population has an equal chance of being included.

Sampling Bias

Non-response: if only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.

Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.

Convenience sample: Individuals who are easily accessible are more likely to be included in the sample.

Observational Studies

- Researchers collect data in a way that does not directly interfere with how the data arise.
- Results of an observational study can generally be used to establish an association between the explanatory and response variables.

Prospective vs. Retrospective Studies

A prospective study identifies individuals and collects information as events unfold.

Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.

Retrospective studies collect data after events have taken place.

Example: Researchers reviewing past events in medical records.

Four Sampling Methods

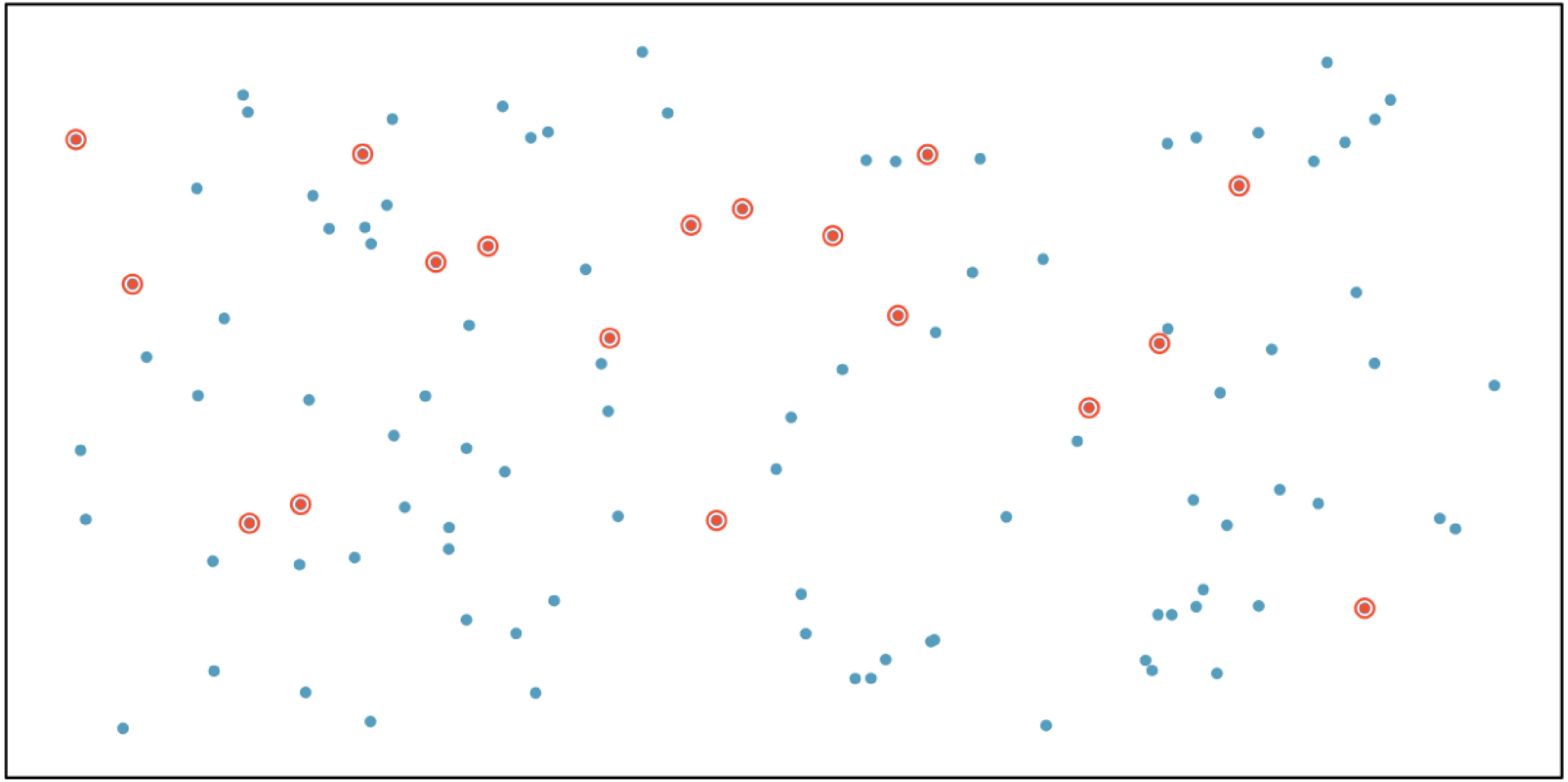
- Almost all statistical methods are based on the notion of implied randomness.

- If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.

- Most commonly used random sampling techniques are *simple, stratified*, *cluster* and multistage sampling.

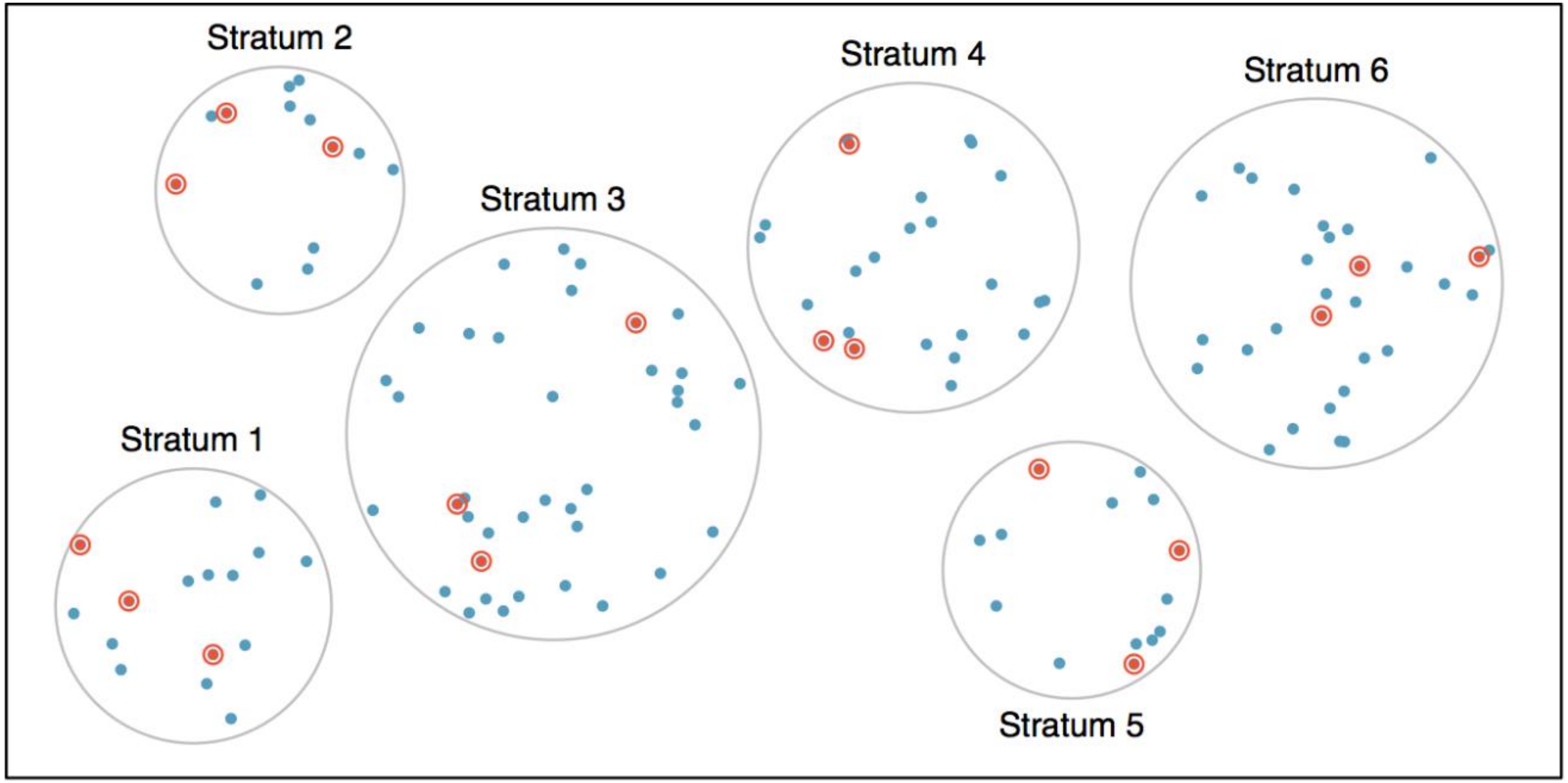
Simple Random Sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



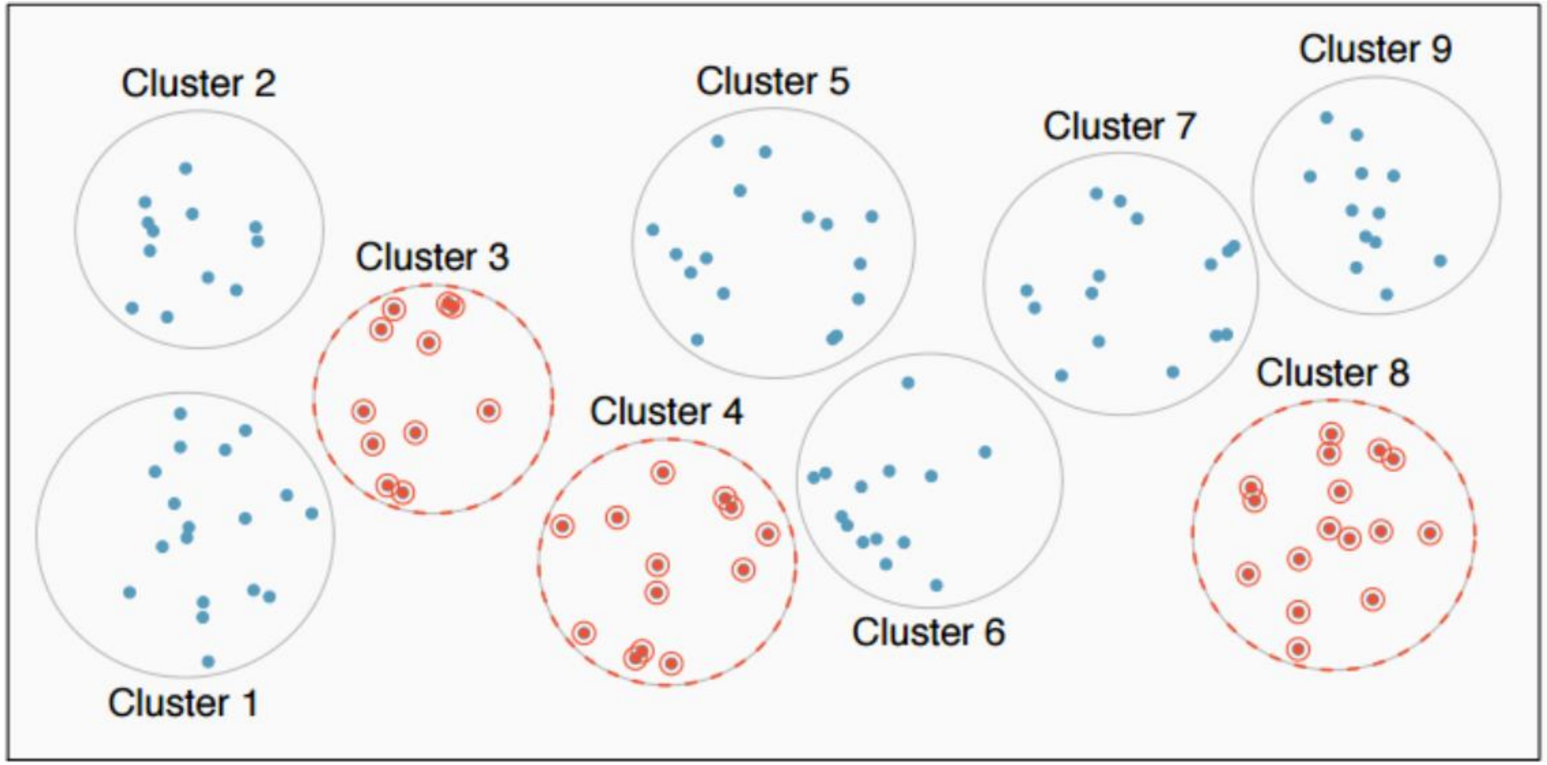
Stratified Sample

Strata are made up of similar observations. We take a simple random sample from each stratum.



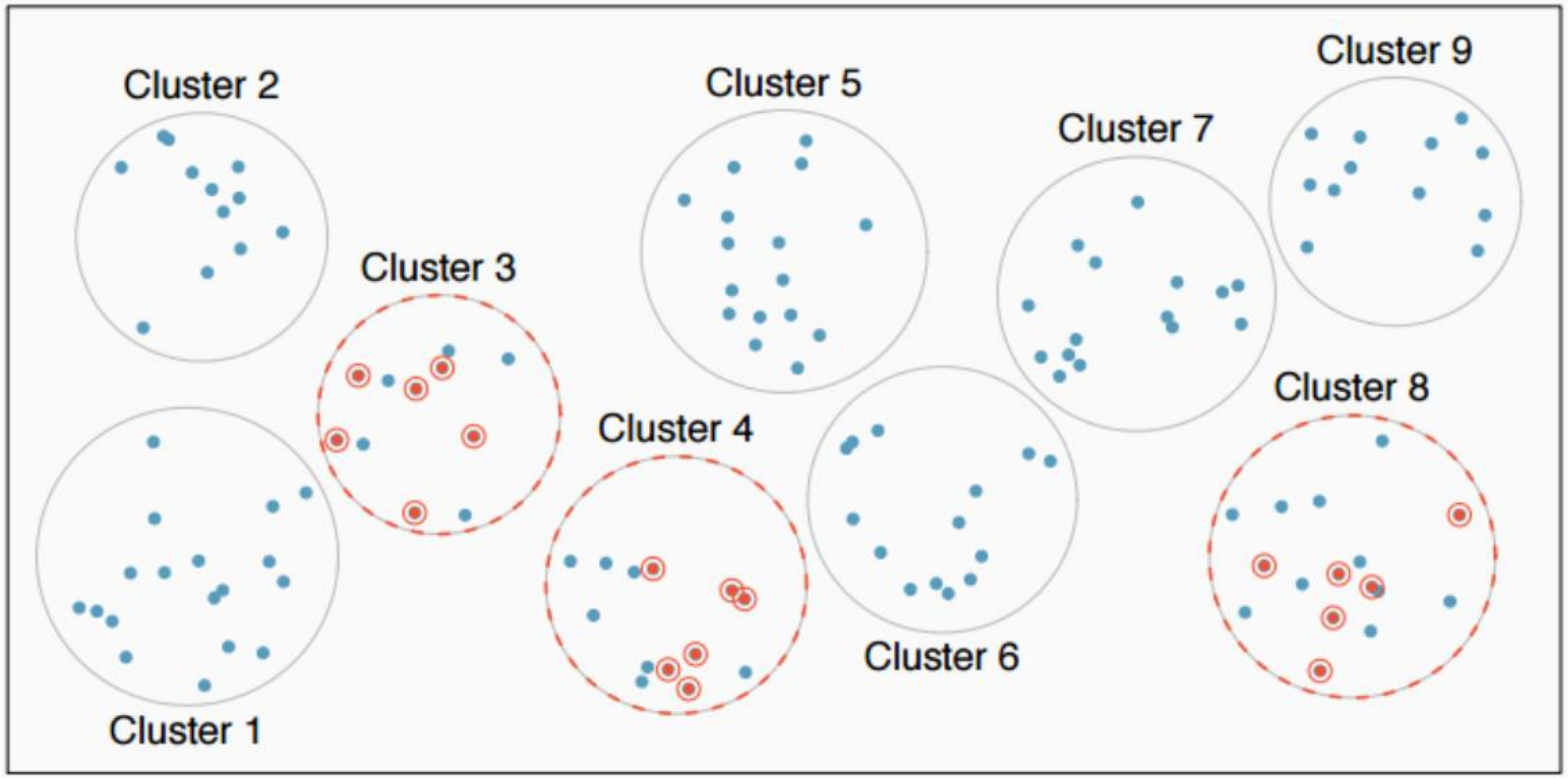
Cluster Sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.



Multistage Sample

Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters



Example

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the *least* effective?

a) Simple random sampling

b) Cluster sampling

c) Stratified sampling

Practice Problem

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. For each part below, identify the sampling method described, and describe the statistical pros and cons of the method in the city's context.

a) Randomly sample 200 households from the city.

b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhoods.

c) Divide he city into 20 neighborhoods, randomly sample 3 neighborhoods, and then sample all households from those 3 neighborhoods.

d) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.

1.4 Experiments

Experiments

- Studies where the researchers assign treatments to cases are called experiments. When this assignments include randomization , it is called randomized experiment.

- Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

Principles of experimental design

1. Control: Compare treatment of interest to a control group.
2. Randomize: Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
3. Replicate: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. Block: If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups.

Example

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

b) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

More Experimental Design Terminology...

Placebo: fake treatment, often used as the control group for medical studies

Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment

Blinding: when experimental units do not know whether they are in the control or treatment group

Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group