Ecole Supérieure
d'Informatique et du Numérique
COLLEGE OF ENGINEERING & ARCHITECTURE

Introduction to Data Analysis

Module leader: **Youness Moukafih** (youness.moukafih@uir.ac.ma)

Lab instructors:

- Youness Moukafih
- Nada Sbihi (nada.sbihi@uir.ac.ma)

20/09/2023

The Role of a Data Analyst

The Role of a Data Analyst

General process of investigation:

The Role of a Data Analyst

General process of investigation:

1. **Identify a question or problem:** Begin by defining a clear and specific question or problem that needs to be addressed.

The Role of a Data Analyst

General process of investigation:

1. **Identify a question or problem:** Begin by defining a clear and specific question or problem that needs to be addressed.

2. **Collect relevant data on the topic:** Gather data that is pertinent to the question or problem at hand.

The Role of a Data Analyst

General process of investigation:

1. **Identify a question or problem:** Begin by defining a clear and specific question or problem that needs to be addressed.

2. **Collect relevant data on the topic:** Gather data that is pertinent to the question or problem at hand.

3. **Analyze the data:** Employ statistical methods and tools to examine and make sense of the data.

The Role of a Data Analyst

General process of investigation:

1. **Identify a question or problem:** Begin by defining a clear and specific question or problem that needs to be addressed.

2. **Collect relevant data on the topic:** Gather data that is pertinent to the question or problem at hand.

3. **Analyze the data:** Employ statistical methods and tools to examine and make sense of the data.

4. **Form a conclusion:** Based on the analysis, draw meaningful conclusions or insights.

The Role of a Data Analyst

Statistics Provides Essential Tools:

The Role of a Data Analyst

Statistics Provides Essential Tools:

- **Statistics provides data analysts with essential tools to:**

The Role of a Data Analyst

Statistics Provides Essential Tools:

- **Statistics provides data analysts with essential tools to:**
  - Effectively collect data.

The Role of a Data Analyst

Statistics Provides Essential Tools:

- **Statistics provides data analysts with essential tools to:**
  - Effectively collect data.
  - Appropriately analyze data.

The Role of a Data Analyst

Statistics Provides Essential Tools:

- **Statistics provides data analysts with essential tools to:**
  - Effectively collect data.
  - Appropriately analyze data.
  - Draw meaningful inferences from the analysis.

Aim of this Module

Aim of this module:

- **Understand Raw Data**: Learn what the data is telling you.
- **Visualize Data**: Make graphs or pictures to see data patterns.
- **Clean Data**: Fix errors or missing parts in the data.
- **Formulate Questions**: Turn real-life questions into math or statistics problems.
- **Choose Analysis Techniques**: Pick the best method to get answers from data.
- **Assess Answers**: Use math to make sure your answers are correct.

Module structure

1. Introduction to Data
2. Summarizing Data
3. Probability
4. Distributions of Random Variables
5. Foundations for Inference
6. Inference for Categorical Data
7. Inference for Numerical Data

Textbook: (https://www.openintro.org/)

OpenIntro

Assessment

- Lab assignments (10%)
- Mini-project (20%)
- Midterm exam (20%)
- Final exam (50%)

So, let’s make sense of the data

Chapter 1: Introduction to data

Chapter 1: Introduction to data

Case study: The Effectiveness of Studying Methods on Exam Performance

- **Objective:** Evaluate the effectiveness of group study vs. individual study for exam preparation.
- **Potential Participants:** 142 students from a large introductory course.
- **Study Participants:** 60 out of the 142 students participated in the study. Some didn't meet the study criteria (like consistent attendance), some had other commitments, and some just didn't want to be part of the study.
- **Study Method Assignment:** Students were randomly divided into two groups, 30 students in each group:
- **Group Study Participants:** students studied collaboratively, discussing and teaching each other.
- **Solo Study Participants:** students studied alone without discussing with peers.

Chapter 1: Introduction to data

Results after the Exam:

The table below shows the distribution of students who scored above 85%. Note that 6 students did not sit for the exam: 3 from the group study and 3 from the solo study.

	Scored $>$ 85%	Scored $\leq$ 85%	Total
Group Study	20	7	27
Solo Study	7	20	27

Chapter 1: Introduction to data

Results after the Exam:

The table below shows the distribution of students who scored above 85%. Note that 6 students did not sit for the exam: 3 from the group study and 3 from the solo study.

	Scored > 85%	Scored ≤ 85%	Total
Group Study	20	7	27
Solo Study	7	20	27

Analysis:

- **Proportion scoring >85% among Group Study Participants:** 20/27 ≈ 0.74 → 74%
- **Proportion scoring >85% among Solo Study Participants:** 7/27 ≈ 0.26 → 26%

Chapter 1: Introduction to data

Discussion Question:

Chapter 1: Introduction to data

Discussion Question:

1. Do the data show a significant difference in the effectiveness of group study versus solo study for exam preparation?

Chapter 1: Introduction to data

Understanding the results:

- Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't see exactly 50 heads. This kind of difference happens often when we collect data.

Chapter 1: Introduction to data

Understanding the results:

- Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't see exactly 50 heads. This kind of difference happens often when we collect data.
- The big difference between the Group Study and Solo Study students (74% - 26% = 48%) might be because group study is really better, or it could just be by chance.
- Because the difference is big, it seems more likely that group studying helps students score better.
- To be sure about this difference, we need special tools from statistics.

Chapter 1: Introduction to data

Discussion Question:

2. Are the results of this study generalizable to all students in all courses?

Chapter 1: Introduction to data

Discussion Question:

2. Are the results of this study generalizable to all students in all courses?

These students are from a specific course and chose to join this study. So, they might not represent all students. While we can't immediately say these results apply to all students, this study is a good start. Group study worked for these students, and that gives hope it might help others too.

Chapter 1: Introduction to data

Data basics:

A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey, and the corresponding variables the data from the responses were stored in:

- **gender**: What is your gender?
- **sleep**: How many hours do you sleep at night, on average?
- **bedtime**: What time do you usually go to bed?
- **countries**: How many countries have you visited?

Chapter 1: Introduction to data

Data basics:

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

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Data basics:

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

Variable

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Observation

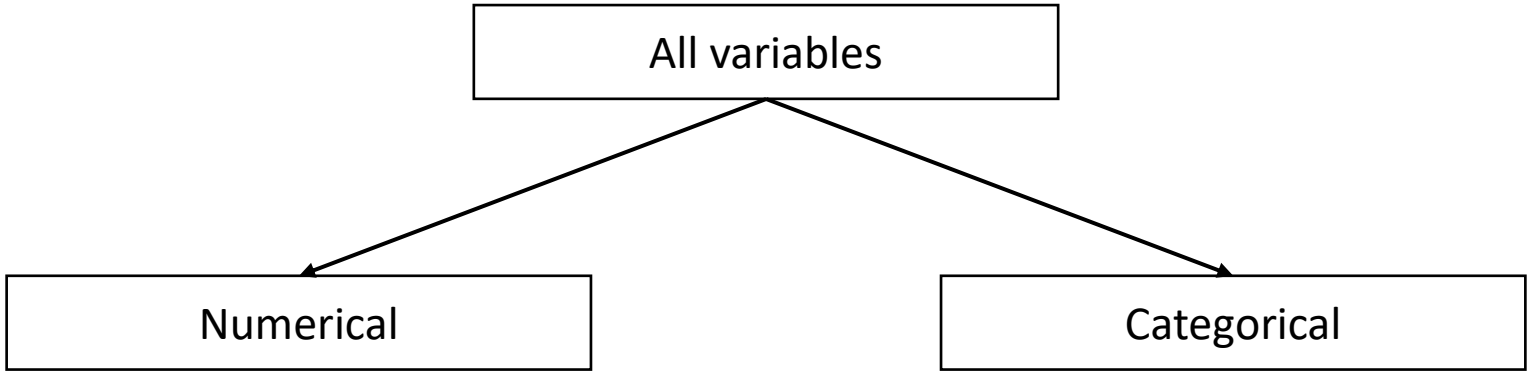
Chapter 1: Introduction to data

Types of variables

All variables

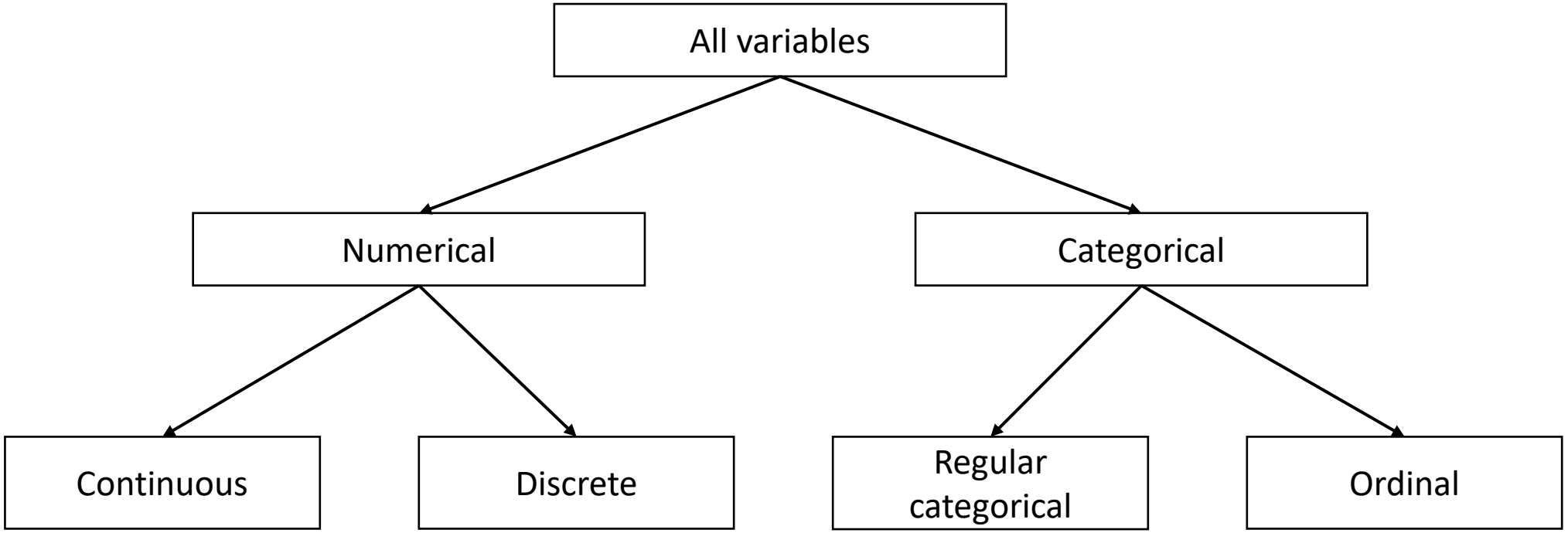
Chapter 1: Introduction to data

Types of variables



Chapter 1: Introduction to data

Types of variables



Chapter 1: Introduction to data

Types of variables

Gender:

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Types of variables

Gender: Categorical

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Types of variables

Gender: Categorical

Sleep:

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Types of variables

Gender: Categorical

Sleep: Numerical, continuous

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Types of variables

Gender: Categorical

Sleep: Numerical, continuous

Bedtime:

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Types of variables

Gender: Categorical

Sleep: Numerical, continuous

Bedtime: Categorical, ordinal

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Types of variables

Gender: Categorical

Sleep: Numerical, continuous

Bedtime: Categorical, ordinal

Countries:

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Types of variables

Gender: Categorical

Sleep: Numerical, continuous

Bedtime: Categorical, ordinal

Countries: numerical, discrete

Student ID	Gender	Sleep (hrs)	Bedtime	Countries
1	Male	8	11:00 PM	3
2	Female	7	10:30 PM	5
3	Male	9	12:00 AM	2
⋮	⋮	⋮	⋮	⋮
84	Female	6	11:00 PM	2
85	Male	7	10:45 PM	4
86	Female	8	11:30 PM	1

Chapter 1: Introduction to data

Practice

What type of variable is a telephone area code?

a. Numerical, continuous
b. Numerical, discrete
c. Categorical
d. Categorical, ordinal

Chapter 1: Introduction to data

Practice

What type of variable is a telephone area code?

a. Numerical, continuous
b. Numerical, discrete
c. Categorical
d. Categorical, ordinal

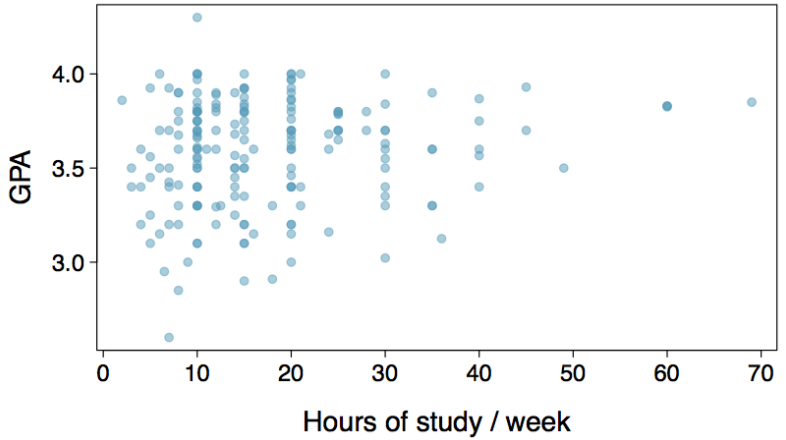
Chapter 1: Introduction to data

Relationships among variables

Chapter 1: Introduction to data

Relationships among variables

Does there appear to be a relationship between Grade Point Average (GPA) and number of hours students study per week?

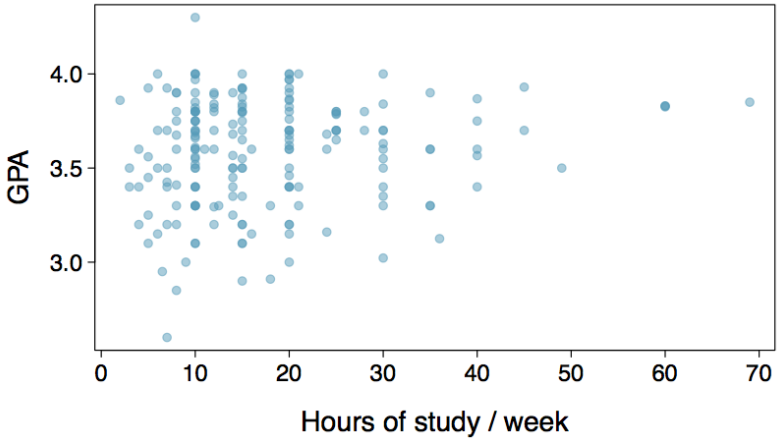


Chapter 1: Introduction to data

Relationships among variables

Does there appear to be a relationship between Grade Point Average (GPA) and number of hours students study per week?

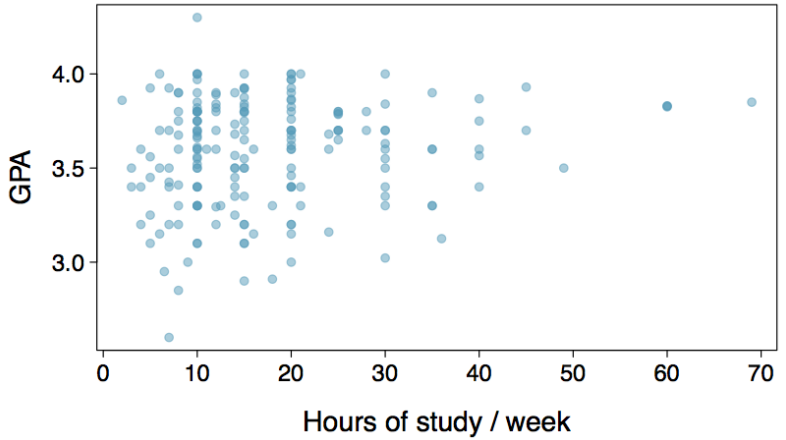
Can you spot anything unusual about any of the data points?



Chapter 1: Introduction to data

Relationships among variables

Does there appear to be a relationship between Grade Point Average (GPA) and number of hours students study per week?



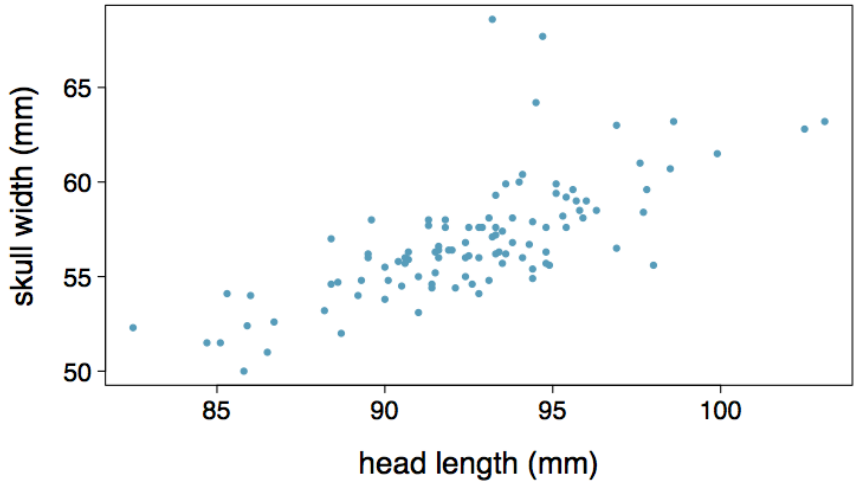
Can you spot anything unusual about any of the data points?

There is one student with GPA > 4.0, this is likely a data error.

Chapter 1: Introduction to data

Practice:

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, i.e. the variables are independent.

b. Head length and skull width are positively associated.

c. Skull width and head length are negatively associated.

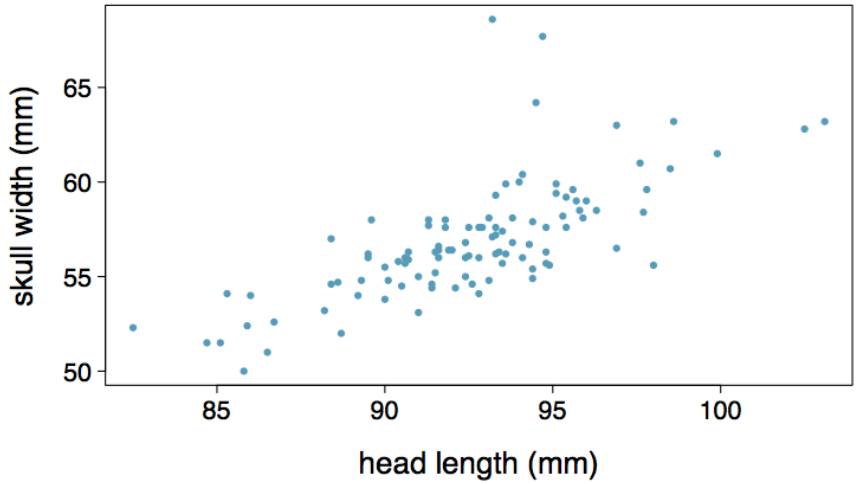
d. A longer head causes the skull to be wider.

e. A wider skull causes the head to be longer.

Chapter 1: Introduction to data

Practice:

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, i.e. 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.

Chapter 1: Introduction to data

Associated vs. independent

- 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, i.e. there is no evident connection between the two, then they are said to be *independent*.

Chapter 1: Introduction to data

Sampling Principles and Strategies

- Understanding the Core Concepts

Chapter 1: Introduction to data

Sampling Principles and Strategies

- Understanding the Core Concepts

Definition: Sampling is the process of selecting a subset of individuals from a population to estimate population parameters**.**

Chapter 1: Introduction to data

Sampling Principles and Strategies

- **Importance of Sampling:**
  - It's impractical to study an entire population.
  - Samples provide a feasible way to gather data.
  - Proper sampling can give accurate representations of the whole.

Chapter 1: Introduction to data

Sampling Principles and Strategies

- **Key Principles:**

    1. **Randomness:** Every member should have an equal chance of selection.
    2. **Representation:** The sample should mirror the population's diversity.
    3. **Size:** Larger samples are usually better, but not always by much.

Chapter 1: Introduction to data

Sampling Principles and Strategies

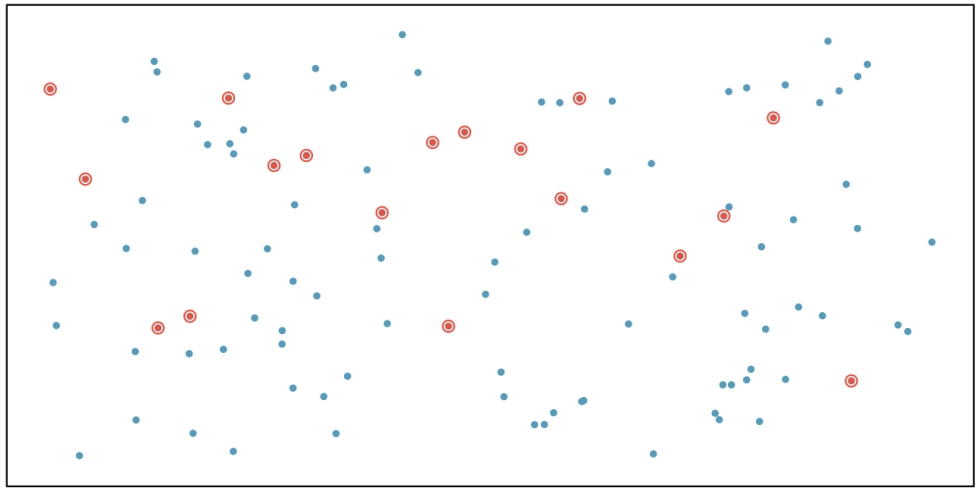
- **Types of Sampling:**

1. Random Sampling
2. Stratified Sampling
3. Cluster Sampling

Chapter 1: Introduction to data

Sampling Principles and Strategies

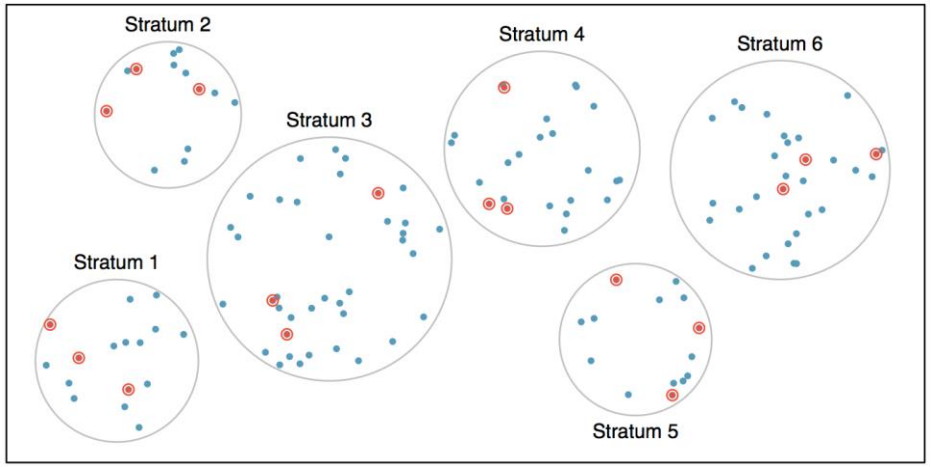
- **Random Sampling:**
  - Randomly select cases from the population, where there is no implied connection between the points that are selected.



Chapter 1: Introduction to data

Sampling Principles and Strategies

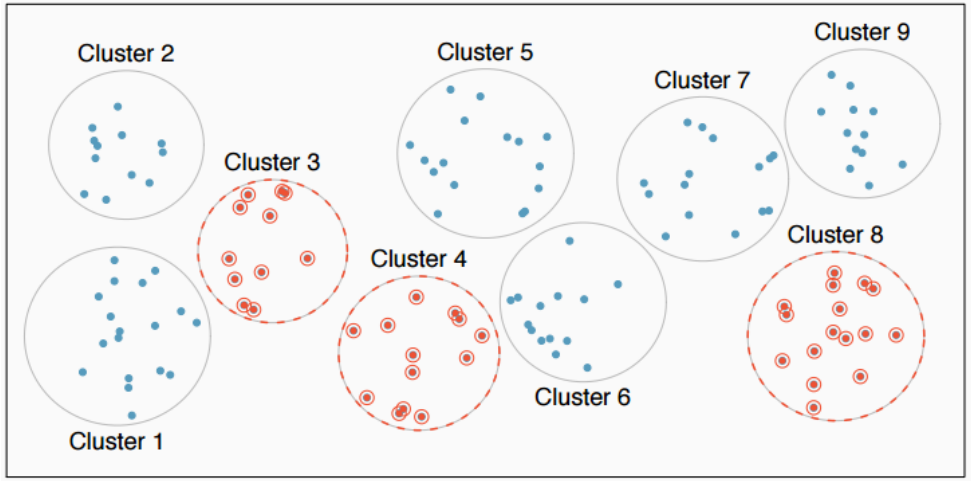
- **Stratified Sampling:**
  - Population divided into subgroups (strata) based on a characteristic. Then samples are taken proportionally from each subgroup.



Chapter 1: Introduction to data

Sampling Principles and Strategies

- **Cluster Sampling:**
  - Population divided into clusters. A random sample of clusters is chosen, and all members in chosen clusters are surveyed.



Chapter 1: Introduction to data

Sampling Principles and Strategies

- **Sampling Errors:**

    1. **Selection Bias**: When some members have a lower or higher probability of being chosen.

    2. **Non-response Bias**: When respondents differ from non-respondents, skewing the sample.

    3. **Undercoverage**: When some groups are left out or underrepresented.

Chapter 1: Introduction to data

Sampling Principles and Strategies

- **Practice:**

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?

  - Simple random sampling
  - Cluster sampling
  - Stratified sampling
  - Blocked sampling

Chapter 1: Introduction to data

Sampling Principles and Strategies

- **Practice:**

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?

  - Simple random sampling
  - Cluster sampling
  - Stratified sampling
  - Blocked sampling

Chapter 1: Introduction to data

Do you have any questions?

Chapter 1: Introduction to data

Do you have any questions?

Chapter 1: Introduction to data

Exercices

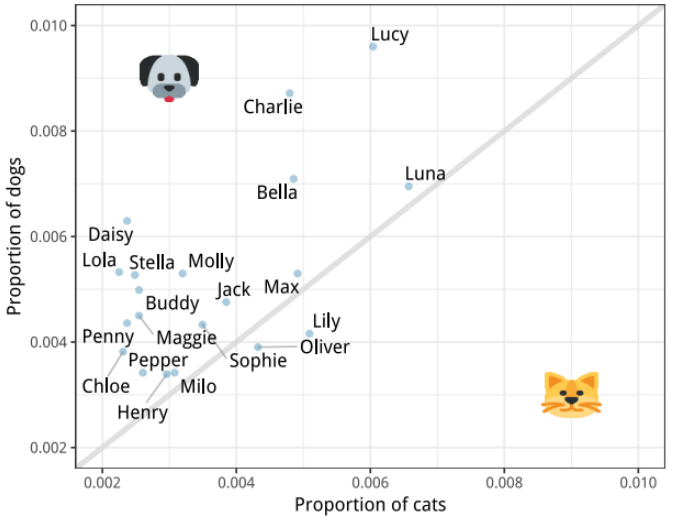
Light and exam performance. A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

(a) What is the response variable?
(b) What is the explanatory variable? What are its levels?
(c) What is the blocking variable? What are its levels?

Chapter 1: Introduction to data

Exercices

Pet names. The city of Seattle, WA has an open data portal that includes pets registered in the city. For each registered pet, we have information on the pet's name and species. The following visualization plots the proportion of dogs with a given name versus the proportion of cats with the same name. The 20 most common cat and dog names are displayed. The diagonal line on the plot is the x = y line; if a name appeared on this line, the name's popularity would be exactly the same for dogs and cats.



(a) Are these data collected as part of an experiment or an observational study?
(b) What is the most common dog name? What is the most common cat name?
(c) What names are more common for cats than dogs?
(d) Is the relationship between the two variables positive or negative? What does this mean in context of the data?