


Exam Study Prep

↗ course	 <u>Data Analysis 1: Exploration</u>
⚙ mastery	none
⚙ progress	not started
📅 date	@October 27, 2023

Chapter 1:

1. What are in the rows and columns of a data table?
 - a. In a data table, each row represents an observation, while each column represents a variable. The rows contain the specific data points or values for each observation, while the columns represent different characteristics or properties of the observations.
2. What are ID variables?
 - a. ID variables are variables that uniquely and unambiguously identify each entity in the data table. They typically contain numeric or text values and are used to identify and link observations across data tables in a dataset. ID variables are essential for maintaining consistency and integrity when working with multiple data tables.
3. What are xsec, tseries, and xt panel data? What's an observation in each? Give an example for each.
 - a. **xsec** refers to cross-sectional data, which includes observations of different cross-sectional units at the same time. Examples of xsec data include people, companies, or countries observed in the same year/month.
 - b. **tseries** refers to time series data, which includes observations of the same cross-sectional unit at different time periods. Examples of tseries data include a single person/company/country observed at different times.
 - c. **xt** panel data refers to cross-section time series data, which includes multiple cross-sectional units observed across multiple time periods. Examples of xt data

include yearly financial data on all companies in a country, weekly sales at various retail stores, or quarterly macroeconomic data on various countries. In xt panel data, each observation is identified using two indices - i for the cross-section and t for the time series.

4. What's the validity and what's the reliability of a variable? Give an example of a variable with high validity and one with low validity.
 - a. Validity refers to the extent to which a variable measures what it is intended to measure. In other words, it assesses whether the variable accurately captures the concept or construct it is supposed to represent. Reliability, on the other hand, refers to the consistency or stability of the measurement. It assesses whether the variable produces consistent and dependable results across multiple measurements or observations.
 - b. An example of a variable with high validity is a thermometer used to measure body temperature. The thermometer accurately measures the body's temperature and provides a valid representation of the concept. It consistently produces accurate readings and is reliable.
 - c. An example of a variable with low validity could be a self-report questionnaire used to measure sleep quality. If individuals inaccurately report their sleep quality, the questionnaire may not validly measure the construct of interest. It may produce inconsistent or unreliable results due to subjective interpretations or biases in self-reporting.
5. What's selection bias? Give an example of data with selection bias and one without.
 - a. Selection bias refers to a systematic error that is introduced into a study or analysis when the selection process of participants or observations is not random or representative of the population being studied. It occurs when certain individuals or groups are more likely to be included in the sample than others, leading to a distortion of the results and rendering them non-generalizable to the entire population.
 - b. An example of data with selection bias would be a study on the effectiveness of a new drug where only patients who voluntarily agree to participate are included in the sample. If those who choose to participate are more likely to have positive experiences with the drug, the results may overstate its effectiveness. This is

because patients who did not have positive experiences or had adverse effects may have chosen not to participate in the study.

- c. On the other hand, an example of data without selection bias would be a study on the prevalence of a certain disease in a population that uses random sampling. In this case, individuals are randomly selected from the population, regardless of their disease status. This ensures that the sample is representative of the entire population and reduces the potential for selection bias.
6. List two common advantages of admin data and two potential disadvantages.
- a. Two common advantages of admin data are high reliability of the variables they measure and high, often complete, coverage. Two potential disadvantages of admin data are that they usually include few variables and may miss many variables that could be useful for analysis, and important variables in admin data may have low validity and may not accurately measure what analysts want to measure.
7. How can we tell if a sample is representative of a population?
- a. We can assess if a sample is representative of a population through two methods: benchmarking and evaluating the sampling process.
 - b. Benchmarking involves comparing the distribution of variables in the sample to the known distribution of those variables in the population. If there are substantial differences between the sample and population distributions, then the sample is not representative. On the other hand, if the distributions are similar, then the sample is representative for those specific variables used in the comparison. However, it may or may not be representative for other variables.
 - c. Evaluating the sampling process involves understanding how the observations were selected and what rules were followed. This allows us to determine whether the selection process was random or non-random. Random sampling, where all observations in the population have an equal chance of being selected, is most likely to produce representative samples. The selection rules for random sampling should be unrelated to the distribution of variables in the data. Random sampling can be done through various methods, such as throwing dice, drawing balls from urns, or using random numbers generated by computers.

- d. Non-random sampling methods, on the other hand, may lead to selection bias and produce non-representative samples. Non-random sampling methods are related to important variables and have a higher or lower likelihood of selecting observations that are different in those variables. As a result, the selected observations tend to be systematically different from the population. Examples of non-random sampling methods include selecting people based on their name or selecting the most recently established firms.#
 - e. In summary, we can assess if a sample is representative by benchmarking its distribution of variables with the known distribution in the population. Additionally, we can evaluate the sampling process to determine if it was random or non-random, with random sampling being the most likely method to produce representative samples.
8. List two sampling rules that likely lead to a representative sample and two sampling rules that don't.
- a. Sampling rules that likely lead to a representative sample:
 - i. Random sampling: This is the best method of producing representative samples. In random sampling, all observations in the population have an equal chance of being selected into the sample. Selection rules are random if they are not related to the distribution of variables in the data. Examples of random sampling include throwing dice or drawing balls from urns, or using random numbers generated by computers.
 - ii. Fixed rules unrelated to the distribution of variables: Another method of random sampling is using fixed rules that are unrelated to the distribution of variables in the data. For example, selecting people with odd-numbered birth dates or people with birthdays on the 15th of every month. These rules may not be "truly random" in a strict sense, but as long as they are not related to the variables used in the analysis, they can lead to a representative sample.
 - b. Sampling rules that don't likely lead to a representative sample:
 - i. Non-random sampling methods related to important variables: Non-random sampling methods are those that are related to important variables and lead

to selection bias. For example, selecting people from the first half of an alphabetic order may lead to selection bias because people with different names may belong to different groups of society. Another example is selecting the most recently established 10% of firms, which is not random and may lead to survivor bias.

- ii. Biased selection due to nonresponse: When conducting surveys, the response rate plays an important role in determining the representativeness of the sample. A low response rate increases the chance of selection bias. For example, if a sample has an 80% response rate, it may be more biased than another sample with a 40% response rate. It is important to report response rates and assess their impact on the representative nature of the sample.

9. List three common features of Big Data. Why does each feature make data analysis difficult?

- a. Three common features of Big Data are volume, complexity, and continuous collection.
 - i. Volume: Big Data refers to datasets that contain a large number of observations and/or variables. This large size can make data analysis difficult because it exceeds the capacity of typical hardware and software to store, manage, and analyze. The sheer volume of data requires advanced technical solutions for collection, structuring, storage, and analysis.
 - ii. Complexity: Big Data often has a complex structure that does not fit into a single data table. It may include various types of data, such as networks with linked observations or multi-dimensional maps with spatial relationships. Additionally, data in the form of text, pictures, or videos may have a complex structure. This complexity makes it challenging to convert the data into a format that can be easily analyzed using traditional methods. Special analytical tools are required to handle the complexity of Big Data.
 - iii. Continuous Collection: Big Data is often automatically and continuously collected and stored. Apps, sensors, social media platforms, and machines collect data as part of their routine operations. This continuous collection and updating of data present challenges for data analysis. Instead of

working with a final data table, analysts need to constantly update and manage the data. The real-time nature of the data requires methods and tools that can handle the continuous flow of information.

- b. Each of these features contributes to the difficulty of analyzing Big Data. The volume of data requires advanced technical solutions to handle the storage and analysis. The complexity of the data necessitates specialized tools and methods that can handle the various forms and structures. The continuous collection and updating of data require ongoing management and analysis processes, making it necessary to develop strategies for handling real-time data. Overall, these features make data analysis with Big Data more complex and challenging compared to traditional data analysis methods.

10. An important principle for research is maintaining confidentiality. How can we achieve that when we collect survey data?

- a. Maintaining confidentiality when collecting survey data can be achieved through several practices.
- b. Firstly, it is crucial to de-identify the data by removing any personally identifiable information such as names and addresses. This ensures that respondents cannot be directly identified from the dataset. However, it is important to note that even without explicit identifiers, certain combinations of variables may still allow for identification. Therefore, additional steps should be taken to ensure that no combination of variables in the dataset can be used to identify individuals or firms.
- c. Secondly, obtaining informed consent from respondents is another way to maintain confidentiality. This means clearly explaining to participants what data will be collected, how it will be used, and any potential linkages to other data sources. By obtaining explicit consent, respondents are aware of and have agreed to the data collection process, which helps to protect their privacy.
- d. Additionally, when data collection is supported by research grants or funding from government or non-governmental foundations, it is often required to adhere to specific ethical and legal principles. These principles may include stricter confidentiality measures and guidelines to ensure the protection of participants' data.

- e. Consulting experts in the legal and ethical aspects of data collection is also considered a good practice. These experts can provide guidance on best practices for maintaining confidentiality and ensuring compliance with relevant laws and regulations.
 - f. Overall, maintaining confidentiality in survey data collection involves de-identifying the data, obtaining informed consent, following ethical and legal principles, and seeking expert guidance. These practices help to protect the privacy and confidentiality of the participants and ensure the ethical and legal use of the collected data.
11. You want to collect data on the learning habits of students in your data analysis class. List two survey methods that you may use and highlight their advantages and disadvantages.
- a. One survey method that can be used to collect data on the learning habits of students in a data analysis class is a self-administered web survey. This method involves asking students to answer questions on their own, typically through an online platform such as Survey Monkey or Google Forms.
 - b. Advantages of a self-administered web survey include:
 - i. Convenience: Students can complete the survey at their own convenience, allowing for flexibility in scheduling and reducing the need for coordination.
 - ii. Anonymity: Web surveys can be conducted anonymously, which may lead to more honest and candid responses from students, especially when addressing sensitive topics
 - c. Disadvantages of a self-administered web survey include:
 - i. Non-response bias: There is a risk of low response rates, as students may choose not to participate or may not take the survey seriously. This can impact the representativeness of the sample and introduce bias.
 - ii. Technology issues: Technical difficulties or limitations in accessing the survey platform may prevent some students from participating, potentially biasing the results

- d. Another survey method that can be used is a personal interview. This involves conducting face-to-face interviews with students in order to gather information on their learning habits
- e. Advantages of a personal interview include:
 - i. Flexibility: Interviewers can adapt the questions and follow-up probes based on the responses of the students, allowing for more in-depth and nuanced understanding of their learning habits.
 - ii. Clarification: In-person interviews provide an opportunity for interviewers to clarify any confusion or provide additional context to ensure that students understand the questions being asked
- f. Disadvantages of a personal interview include:
 - i. Time-consuming: Personal interviews require more time and resources compared to other survey methods, as interviewers need to meet with each student individually.
 - ii. Interviewer bias: Different interviewers may ask questions or record responses differently, potentially introducing bias into the data. Training and standardization are necessary to mitigate this risk.

12. You want to collect data on the friendship network of students in a class. You consider two options: (1) collect their networks of Facebook users using data there (80% of them are on Facebook), or (2) conduct an online survey where they are asked to mark their friends from a list of all students. List arguments for each option, paying attention to representation, costs, and ethical issues.

- a. Option 1: Collecting networks of Facebook users using data from Facebook
- b. Arguments:
 - i. Representation: Collecting friendship networks from Facebook can provide a more accurate representation of the actual friendship connections among the students in the class. Since 80% of the students are on Facebook, it is likely that most of their friends are also on Facebook. This option allows for a comprehensive view of the students' social connections.

- ii. Cost: Collecting data from Facebook may be relatively low-cost compared to other methods. It eliminates the need for manual data entry or survey administration. It also eliminates the need for additional resources such as paper and pencils for the students to mark their friends.
 - iii. Ethical issues: When collecting data from Facebook, it is important to consider ethical issues related to privacy and consent. It is necessary to ensure that the students are aware of the data collection and have given their informed consent to participate in the study. Additionally, steps should be taken to protect the privacy of the participants and their network connections
- c. Option 2: Conducting an online survey where students mark their friends from a list of all student
- d. Arguments:
 - i. Representation: Conducting an online survey allows for equal representation of all students in the class, regardless of whether they are on Facebook or not. It ensures that even those who do not use Facebook or prefer not to share their social connections can participate and provide their input. This option may better capture the diversity of friendships within the class.
 - ii. Cost: Conducting an online survey may require some resources and time to design and administer. However, it eliminates the need for accessing Facebook data and potentially dealing with any associated costs or limitations. It can be a cost-effective alternative, especially if the class size is small.
 - iii. Ethical issues: In an online survey, it is still important to address ethical issues related to informed consent and privacy. Participants should be properly informed about the purpose of the study, how their data will be used, and ensure that their anonymity is protected. It is important to have measures in place to safeguard participant information and ensure data security
- e. Overall, both options have their advantages and drawbacks in terms of representation, cost, and ethical considerations. The choice of the data collection method should be made based on the specific research goals,

resources available, and the ethical guidelines that need to be followed. Consulting experts in the legal and ethical aspects of data collection can provide further guidance in making this decision.

13. You consider surveying a sample of employees at a large firm. List four selection methods and assess whether each would result in a representative sample.

a. Four selection methods that could be used to survey a sample of employees at a large firm are:

- i. **Random Sampling:** This method randomly selects employees from the entire population of the firm. It has the potential to result in a representative sample if every employee has an equal chance of being selected. However, there is still a possibility of chance selecting employees who are not representative of the entire population, especially if the sample size is small.
- ii. **Stratified Sampling:** This method divides the population into different strata or groups based on certain characteristics such as department, job level, or location. Then, a random sample is taken from each stratum in proportion to its size within the entire population. Stratified sampling can result in a representative sample because it ensures that each subgroup is represented in the sample, thereby capturing the diversity within the population.
- iii. **Cluster Sampling:** This method divides the population into clusters or groups based on certain characteristics such as departments or teams. Then, a random sample of clusters is selected, and all employees within the selected clusters are included in the sample. Cluster sampling can be useful when it is not practical to sample individuals directly, such as when departments are geographically dispersed. However, it may not always result in a representative sample if there is a high degree of variability within clusters.
- iv. **Convenience Sampling:** This method selects employees who are readily available or easily accessible, such as those who happen to be present in a certain location at a given time. Convenience sampling is convenient and cost-effective, but it is susceptible to selection bias as it may result in a sample that is not representative of the entire population. This method

should be used with caution and may not provide reliable results if the goal is to obtain a representative sample.

14. You want to examine the growth of manufacturing firms in a country. You have data on all firms that are listed on the stock exchange. Discuss the potential issues of coverage and its consequences. Does it matter which country it is?
- a. The potential issues of coverage in examining the growth of manufacturing firms using data from all firms listed on the stock exchange are as follows:
- i. Exclusion of non-listed firms: By focusing only on firms that are listed on the stock exchange, the analysis excludes non-listed firms. This may introduce a bias in the results because non-listed firms may have different growth patterns compared to listed firms. For example, smaller or newer firms that have not yet met the requirements for listing may have different growth trajectories.
 - ii. Limited sample size: The number of firms listed on the stock exchange is typically smaller compared to the total number of manufacturing firms in a country. Therefore, the analysis may suffer from a limited sample size, which could impact the statistical power and generalizability of the findings.
 - iii. Selection bias: A key issue in using data from listed firms is the possibility of selection bias. Firms that choose to list on the stock exchange may differ systematically from non-listed firms. For example, listed firms may be larger, more established, or have better growth prospects. This selection bias can affect the representativeness of the sample and limit the generalizability of the findings to the entire population of manufacturing firms.
 - iv. Market-driven growth: Focusing on listed firms may bias the analysis towards firms that are more market-driven and growth-oriented. This may exclude firms that prioritize stability or have alternative growth strategies, such as family-owned or founder-owned firms. Consequently, the findings may not capture the full range of growth patterns in the manufacturing sector.

- v. Country-specific effects: The consequences of coverage issues may vary depending on the country under study. Countries differ in terms of their stock market development, the size and nature of the manufacturing sector, and the level of listing requirements. These country-specific factors can influence the representativeness and relevance of using listed firms data for examining growth patterns.

b. Consequences of the coverage issues:

- i. Limited insights into overall sector performance: By excluding non-listed firms, the analysis may provide an incomplete picture of the growth dynamics within the manufacturing sector. It may not capture the diversity of firms and growth patterns, leading to a limited understanding of the industry as a whole.
- ii. Biased conclusions: The analysis based on listed firms' data may produce biased conclusions about the factors influencing growth in the manufacturing sector. The findings may primarily reflect the characteristics and behaviors of listed firms, rather than the broader population of manufacturing firms.
- iii. Potential misrepresentation of policy implications: If the analysis relies solely on data from listed firms, the policy implications may not be applicable to the entire manufacturing sector or may overlook important segments of the industry. This can lead to ineffective or inappropriate policy recommendations.

c. Does it matter which country it is?

- i. Yes, the country under study can have implications for the issues of coverage and their consequences. The stock market development, regulatory frameworks, and the composition of the manufacturing sector can vary significantly across countries. For example:
- ii. Stock market size and accessibility: Some countries may have larger and more liquid stock markets, providing a broader representation of the manufacturing sector. In such cases, the coverage issues related to sample size and selection bias may have lesser impact compared to countries with smaller or less accessible stock markets.

- iii. Listing requirements and regulations: Different countries have varying listing requirements and regulations that govern their stock exchanges. These requirements can influence the characteristics and behavior of listed firms, potentially affecting the representativeness of the sample and the generalizability of the findings.
 - iv. Manufacturing sector composition: The composition of the manufacturing sector can differ across countries, with variations in the dominance of specific industries or the prevalence of family-owned or founder-owned firms. These differences can affect the relevance and applicability of using listed firms' data to understand growth dynamics.
 - d. Overall, the country in which the analysis is conducted plays a crucial role in determining the extent to which coverage issues impact the analysis and their consequences. It is important to consider the specific context and characteristics of the country when interpreting the findings and drawing conclusions.
15. 1000 firms are randomly selected from all the SMEs (small and medium enterprises) in a country. What is the population, what is the sample, and is the sample representative?
- a. Population: The population in this case is all small and medium enterprises (SMEs) in the country.
 - b. Sample: The sample is the 1000 firms that were randomly selected from the population of SMEs in the country.
 - c. Representativeness of the sample: To determine if the sample is representative, we need to assess if it has a similar distribution of variables to that of the population. This text does not provide information on the specific variables being considered or any benchmarking results to compare the sample with the population. Therefore, without additional information, we cannot definitively determine if the sample is representative or not.
16. You are doing a survey about the smoking habits of the students of your university and want to reach a 20% sample. Here are some potential sampling rules. Would

each lead to a representative sample? Why or why not?

- (1) Stand at the main entrance and select every fifth entering student.
- (2) Get the students' email list from the administration and select every fifth person in alphabetic order.
- (3) The same, but select the first fifth of the students in alphabetic order.
- (4) The same, but now sort the students according to a random number generated by a computer and select the first fifth.

a. To determine whether each potential sampling rule would lead to a representative sample, we need to consider if the rule is random and unrelated to any variables in the data.

- i. (1) Stand at the main entrance and select every fifth entering student: This sampling rule is systematic rather than random. It selects every fifth student, which means that students with certain characteristics may be overrepresented or underrepresented in the sample. For example, if students tend to arrive in waves or groups, this rule may lead to bias in the sample.
- ii. (2) Get the students' email list from the administration and select every fifth person in alphabetic order: This rule is also systematic and not random. By selecting every fifth person in alphabetic order, students with certain surnames may be overrepresented or underrepresented in the sample. This could lead to a biased sample if, for example, students with last names at the beginning of the alphabet have different smoking habits compared to those with last names at the end of the alphabet.
- iii. (3) The same, but select the first fifth of the students in alphabetic order: This rule is systematic and not random. By selecting only the first fifth of students in alphabetic order, students with certain surnames at the end of the alphabet will be excluded from the sample. This will likely lead to a biased sample as it will not be representative of the entire population of students at the university.
- iv. (4) The same, but now sort the students according to a random number generated by a computer and select the first fifth: This rule is random and unrelated to any variables in the data. By sorting the students according to a random number and selecting the first fifth, each student has an equal chance of being selected into the sample. This random sampling method is

more likely to produce a representative sample compared to the previous rules.

- b. In summary, only option (4) - sorting the students according to a random number and selecting the first fifth - would lead to a representative sample as it is a random and unbiased sampling method. The other options are systematic and could introduce bias into the sample.

Chapter 2:

1. What is the difference between a dataset and a data table, and which do we use for actual analysis?
 - a. A dataset is a broader concept that includes multiple data tables with different kinds of information. On the other hand, a data table is specifically defined as a table consisting of observations (also known as cases) and variables (also called features). In a data table, the rows represent observations, with each row containing information about a specific observation. The columns represent variables, with each column representing a different variable.
 - b. For actual analysis, the text mentions that data is most straightforward to analyze if it forms a single data table. Therefore, it can be inferred that data tables are typically used for actual analysis.
2. What is panel (multi-dimensional) data, and what is xt panel data? Give one example of xt panel data and one of non-xt panel data.
 - a. Panel (multi-dimensional) data refers to a type of data that includes observations across multiple dimensions. It can include data on various cross-sectional units observed across multiple time periods.
 - b. XT panel data, also known as cross-section time series data, is a specific type of panel data where observations are one cross-sectional unit observed at different time periods. This means that each observation represents the same unit (such as a country, retail store, or company) observed multiple times.
 - c. One example of XT panel data could be yearly financial data on all companies in a country. In this case, each observation represents a specific company observed year after year.

- d. On the other hand, a non-XT panel data example could be data on different retail stores observed weekly. In this case, each observation represents a specific retail store, but they are not observed multiple times within different time periods.
3. What is a binary variable? Give two examples.7#
- a. A binary variable is a special case of a qualitative variable that can take on only two values, typically representing a yes/no or true/false answer. It is often used to represent a categorical variable that has a binary distinction. Two examples of binary variables are:
- i. Whether a respondent to a survey is female or not: This binary variable would have a value of 1 for females and 0 for males.
 - ii. Whether a firm is in the manufacturing sector or not: This binary variable would have a value of 1 for firms in the manufacturing sector and 0 for firms in other sectors.
4. What are nominal and ordinal qualitative variables? Give two examples for each.
- a. Nominal qualitative variables are a type of qualitative variable with values that cannot be unambiguously ordered. Two examples of nominal qualitative variables are the chocolate brand names a customer purchased and the headquarter cities for chocolate makers.
- b. Ordinal qualitative variables are also a type of qualitative variable, but they have values that are unambiguously ordered. Two examples of ordinal qualitative variables are subjective health measures (such as whether someone rates their health as poor, fair, good, very good, or excellent) and the strength of an opinion (such as whether one strongly agrees, agrees, disagrees, or strongly disagrees with a statement).
5. What are interval and ratio quantitative variables? Give two examples for each.
- a. Interval quantitative variables are variables that have the property that a difference between values means the same thing regardless of the magnitudes. Examples of interval variables can include temperature differences (e.g. a one degree Celsius difference is the same when comparing 20 to 21 degrees or 30 to 31 degrees) and price differences (e.g. a one dollar price difference of \$3 versus \$4 is the same as \$10 versus \$11).

- b. Ratio quantitative variables, also known as scale variables, are interval variables with the additional property that their ratios mean the same regardless of the magnitudes. These variables also have a meaningful zero in the scale. Examples of ratio variables can include measures of length (e.g. 10 km is twice as long as 5 km), elapsed time, age, value, or size. Another example of a ratio variable could be the cost of a used car (e.g. a used car sold for zero dollars costs nothing unambiguously, while a used car sold for \$8000 is twice as expensive as one sold for \$4000).
6. What are stock and flow variables? Give two examples for each.
- a. Stock variables are quantities that are measured at a specific point in time. They represent the amount of something at a given moment. Examples of stock variables include the amount of water in a reservoir at a specific time, and the amount of government debt at the end of a particular year.
- b. Flow variables, on the other hand, are quantities that are measured over a period of time. They represent the rate of change of something. Examples of flow variables include the amount of water flowing into a reservoir over a specific time period, and the sales of chocolate in a shop during a particular month.
7. What is the difference between long and wide format xt panel data? Which one would you prefer and why?
- a. The difference between long and wide format xt panel data is in the way the data is structured. In long format, each observation is represented by a separate row, with each row containing the variables for a specific time period and cross-sectional unit. This means that for each unit, there can be multiple rows corresponding to different time periods. In wide format, each observation is represented by a separate column, with each column containing the variables for a specific time period and cross-sectional unit. This means that for each unit, there is only one row and the variables for different time periods are arranged horizontally.
- b. In the given text, Table 2.4 represents the xt panel data in long format and Table 2.5 represents the same data in wide format. In Table 2.4, we can see that each country has multiple rows corresponding to different years, while in Table 2.5,

each country has only one row and the variables for different years are arranged horizontally.

- c. The advantage of the long format is its transparency and ease of management. It is straightforward to add new observations to long format tables, whether they are new cross-sectional units or new time periods. Additionally, it is easier to transform and clean variables in long format. On the other hand, the advantage of the wide format is that it is easier to analyze, especially if there are only a few time periods. However, it is generally considered good practice to store data in long format.
 - d. In terms of preference, it ultimately depends on the specific analysis and the nature of the data. If there are only a few time periods and the analysis is easier to perform in wide format, then wide format may be preferred. However, in general, the tidy approach recommends storing multi-dimensional data in long format and transforming it for analysis when necessary. Long format allows for greater flexibility in analyzing and managing the data, and it is easier to add new observations or variables. Therefore, for most cases, the long format would be preferred.
8. What are missing values and how can we discover them? What options do we have to work with variables that have missing data, and which should we choose when?
- a. Missing values refer to the absence of a value for a variable in some observations. They present a problem because they can be mistaken for valid values and because they reduce the number of observations with valid information.
 - b. Discovering missing values involves analyzing the data variable by variable and identifying the percent of missing values for each variable. This can be done by examining the dataset and determining if there are any observations with blank spaces, dots, or specific characters such as "NA" that represent missing values. Additionally, missing values can also be identified when number values are used to record missing data outside of the usual range.
 - c. To work with variables that have missing data, there are several options available. The first option is to keep observations with missing values and replace the missing values with something else. For qualitative variables, a new

category can be created to represent missing values, and a corresponding binary variable can be generated. For quantitative variables, the missing values can be replaced with an imputed value, such as the overall average of the variable. In both cases, it is recommended to create a binary flag variable indicating that the original value was missing.

- d. The second option is to drop the variable with missing data from the analysis. This is typically the best choice when a large fraction of the observations have missing values for a variable. By dropping the variable, only the predictor variable is lost, and it makes sense to do so if the variable is expected to have many missing values in the live data as well.
 - e. The third option is to drop the observations with missing values. This option is only recommended when there are very few missing values for a few variables. If this option is chosen, it is important to consider the potential loss of a large number of observations from the dataset.
 - f. When deciding which option to choose, it is essential to consider the magnitude of the missing values problem. The fraction of observations affected and the number of variables affected determine the size of the problem. It is also important to check the source of why values are missing and consider the implications of imputing values or removing variables.
 - g. Ultimately, the choice of which option to use depends on the specific situation and the goals of the analysis. Domain knowledge and understanding of the data are important in making these decisions. Additionally, the chosen approach should be documented to provide transparency and to address any potential consequences on the analysis results.
9. What is entity resolution? Give an example.
- a. Entity resolution is the process of resolving issues related to entities in a data table. This includes dealing with duplicate observations, ambiguous identification, and non-entity rows. The goal of entity resolution is to ensure that each entity has a unique identifier and that different data tables have consistent identification of entities.
 - b. For example, in the context of the football manager case study mentioned in the text, entity resolution involves resolving the issue of different names for the same football teams. The data table includes different versions of names for the

teams Manchester City and Manchester United. The process of entity resolution would involve defining unique IDs for each team and determining which names belong to the same team. This ensures that the teams are properly identified and can be linked to other data tables accurately.

10. List four topics that data cleaning documentation should address.

a. Four topics that data cleaning documentation should address are:

- i. Birth of data: This includes information about when, how, and for what purpose the data was collected. It should also mention who collected the data and how it is available.
- ii. Observations: This includes the type of observations (e.g., cross-sectional, time series), how observations are identified, and the number of observations.
- iii. Variables: This includes a list of variables to be used, their type, their content, and the range of values they can take. Additionally, it should mention the number or percentage of missing values and if any generated variables are used, how they were created.
- iv. Data cleaning steps: This includes a detailed description of the steps taken during the data cleaning process. It should outline the specific actions taken to resolve issues such as duplicates, ambiguous entities, non-entity rows, and missing values.

11. What are the benefits of writing code for cleaning data?

a. The benefits of writing code for cleaning data include:

- i. Easy modification and re-doing of data cleaning procedures: Writing code allows for easy modification of specific parts of the data cleaning process and the ability to re-do the entire procedure from start to finish. This is particularly useful when new issues emerge or when the raw data changes.
- ii. Automation and repeatability: Code can automate the data cleaning process, allowing it to be repeated as needed. This is helpful when there is a slight change in the underlying raw data. By having an automated process, it saves time and effort in manually cleaning the data again.

- iii. Reproducibility: Writing code makes it easy for anyone else to reproduce the data cleaning procedure, increasing the credibility of the subsequent analysis. It allows other analysts to check the work, replicate the analysis, or build on it in the future.
- iv. Documentation: Code becomes the skeleton for documentation as it shows the steps in the data cleaning procedure itself. It provides a clear and systematic record of the data cleaning steps taken, making it easier to understand and communicate the process.
- v. Collaboration: Many existing platforms, such as GitHub, assist and promote collaboration. By writing code, it becomes easier to collaborate with others in the data cleaning process, allowing for effective teamwork and sharing of knowledge and insights
- vi. It is important to note that there are trade-offs between these benefits and the work needed to write code, especially for short tasks, small datasets, and novice data analysts. However, the advantages of code tend to outweigh its costs, leading to more efficient and robust data cleaning processes.

12. What does merging two data tables mean, and how do we do it? Give an example.

- a. Merging two data tables means combining the information from two separate tables into a single table based on common variables or identifiers. It allows us to link related information from different tables and create a comprehensive dataset for analysis.
- b. To merge two data tables, we typically start with one data table as the base table and merge observations from the other data table into it. The merge is done based on the common variables between the two tables.
 - i. There are different types of merging based on the relationship between the tables. The most straightforward type is one-to-one (1:1) matching, where each observation in the base table is matched with one observation in the other table. For example, we can merge a data table with customer information (including age and income variables) with another data table on customer ratings of a retail store visit.

- ii. Another type is many-to-one (m:1) matching, where multiple observations in the base table are matched with one observation in the other table. For example, we can merge a data table with average family income in zip code locations (with observations on zip codes) with a data table on customer information (with zip code of residence as a variable).
 - iii. One-to-many (1:m) matching is similar to many-to-one matching but with the order of the tables reversed. We start with the data table with more aggregated observations (such as zip codes or countries) and merge it with the observations from the other dataset (such as customers or country-year observations).
 - iv. The most complex type is many-to-many (m:m) matching, where one observation in the first table may be matched with many observations in the second table, and vice versa. In such cases, a separate table is often needed to connect the IDs. For example, merging information on companies and managers may require a separate table that links the IDs of companies and managers.
- c. An example of merging two data tables could be merging a data table on customers with a data table on purchases. The common variable between the two tables could be the customer ID. By merging the two tables, we can create a new table that combines customer information (such as age, gender, and location) with purchase information (such as item, quantity, and price), allowing us to analyze customer behavior and preferences.
13. Decide what types of variables the following are. Are they qualitative or quantitative? Are they binary variables? Also think about whether they are measured on a nominal, ordinal, interval, or ratio scale.
- (a) IQ
 - (b) Country of origin
 - (c) Number of years spent in higher education
 - (d) The answer to the question in a survey that says: "Indicate on the scale below how much you agree with the following statement: Everyone has to learn data analysis for at least two semesters." with options "5 – Fully agree" "4 – Somewhat agree" "3 – Indifferent" "2 – Somewhat disagree" "1 – Fully disagree"

- (e) A variable that is 1 if an individual bought a car in a given month
- (f) Eye color

(a) IQ is a quantitative variable. It is a ratio variable because it has a meaningful zero point (an IQ of zero is the absence of intelligence) and the ratios between scores are meaningful (an IQ of 100 is twice as intelligent as an IQ of 50).

(b) Country of origin is a qualitative variable. It is a nominal variable because the values (countries) cannot be ordered in a meaningful way.

(c) Number of years spent in higher education is a quantitative variable. It is an interval variable because the differences between values (number of years) are meaningful regardless of the magnitudes. However, it is not a ratio variable because there is no meaningful zero point (having zero years of education does not mean the absence of education).

(d) The answer to the question in a survey is a qualitative variable. It is an ordinal variable because the options (5 – Fully agree, 4 – Somewhat agree, 3 – Indifferent, 2 – Somewhat disagree, 1 – Fully disagree) have an unambiguous order.

(e) A variable that is 1 if an individual bought a car in a given month is a binary variable. It is a qualitative variable with only two values. It is best to code it as 0 for not buying a car and 1 for buying a car.

(f) Eye color is a qualitative variable. It is a nominal variable because the values (eye colors) cannot be ordered in a meaningful way.

14. Consider the following data tables. Data table 1 includes countries in its rows; its columns are the name of the country, its area, and whether it has access to the sea. Data table 2 includes countries in its rows; its columns are GDP and population for various years, each year in a different column. Is this a tidy dataset? If yes, why? If not, why not, and how can you transform it into a tidy dataset?

a. No, the given data table is not a tidy dataset.

b. The reasons are as follows:

- The columns in data table 2 represent different variables (GDP and population) for various years. According to the principles of tidy data, each

variable should have its own column. In this data table, the multiple columns represent multiple variables, violating the principle

- ii. To transform the data table into a tidy dataset, we need to reshape it so that each variable has its own column. This can be done using the `pivot_longer` function in R or the `melt` function in Python.
 - iii. The transformed tidy dataset will have the following columns: country, year, variable (which can be GDP or population), and value (the corresponding value for the variable and year). Each row in the transformed dataset will represent a unique observation.
 - iv. By transforming the data table in this way, we adhere to the principles of tidy data, where each observation is a row and each variable is a column. This makes the dataset easier to work with and analyze.
15. Consider the following data tables. Data table 1 includes single-plant manufacturing companies in its rows; its columns are the name of the company, the zip code of its plant, its industry, and the average wage the company pays in 2019. Data table 2 includes zip codes in its rows; its columns are total population, population density, and average house prices in 2019 in the zip code area. You want to analyze how the average wage paid by companies is related to the average house prices. How would you create a work file from these data tables for that analysis?
- a. To create a workfile from these data tables for the analysis of how the average wage paid by companies is related to the average house prices, we would follow these steps:
 - i. Open a spreadsheet software (such as Excel or Google Sheets) to create the workfile.
 - ii. In a new sheet or tab, copy the relevant columns from Data table 1 and Data table 2.
 - iii. Copy the columns "name of the company", "zip code of its plant", "industry", and "average wage" from Data table 1 into the corresponding columns in the workfile.
 - iv. Copy the columns "zip codes", "total population", "population density", and "average house prices" from Data table 2 into the corresponding columns in the workfile.

- v. Make sure that the columns in the workfile are properly labeled and organized.
- vi. Check for any missing or incomplete data in the workfile and, if necessary, fill in or correct the data.
- vii. If needed, calculate any additional variables or transformations that may be required for the analysis. For example, if there is a need to calculate the wage-to-house price ratio, this could be done by dividing the average wage by the average house price for each row.
- viii. Save the workfile in a suitable format, such as a CSV or Excel file, for further analysis
- ix. By following these steps, we have created a workfile that combines the relevant data from Data table 1 and Data table 2, allowing us to analyze the relationship between the average wage paid by companies and the average house prices. This workfile can now be used for various statistical analyses, such as regression or correlation analysis, to explore the relationship between these variables.

Chapter 3

1. The distribution of quantitative variables may be visualized by a histogram or a density plot (kernel density estimate). What's the difference between the two and which one would you use? List at least one advantage for each. How about qualitative variables with a few values?
 - a. The difference between a histogram and a density plot is that a histogram represents the distribution of quantitative variables by dividing the data into bins and counting the number of data points in each bin, whereas a density plot represents the distribution by estimating the probability density function (PDF) of the variable.
 - b. Advantages of a histogram:
 - i. Provides a visual representation of the frequency distribution of the data.
 - ii. Allows for easy identification of the shape, central tendency, and spread of the data.

- c. Advantages of a density plot:
 - i. Provides a smooth curve that can reveal patterns and trends in the data.
 - ii. Removes the dependence on the number of bins, allowing for better comparison of distributions with different sample sizes.
 - d. For qualitative variables with a few values, a bar plot or a pie chart is commonly used to visualize the distribution. These plots show the frequency or proportion of each value of the variable.
2. The mean, median, and mode are statistics of central tendency. Explain what they are precisely.
 - a. The mean is the average value of a set of data. It is calculated by summing all the values of the data and dividing the sum by the number of data points. The median is the middle value of a set of ordered data. It divides the data into two equal halves. The mode is the value that appears most frequently in the data.
 3. The standard deviation, variance, and inter-quartile range are statistics of spread. Explain what they are and give the formula for each.
 - a. The standard deviation measures the average deviation of each data point from the mean. It is calculated by taking the square root of the variance. The variance measures the spread of the data by calculating the average squared difference between each data point and the mean. The inter-quartile range (IQR) is the range between the first quartile (25th percentile) and the third quartile (75th percentile). It represents the spread of the middle half of the data.
 - b. The formulas for each are as follows:
 - i. Variance: $\text{Var} = \sum (x_i - \bar{x})^2 / n$
 - ii. Standard deviation: $\text{Std} = \sqrt{\sum (x_i - \bar{x})^2 / n}$
 - iii. Inter-quartile range: $\text{IQR} = Q3 - Q1$
 4. What are percentiles, quartiles, and quintiles? Is the median equal to a percentile?
 - a. Percentiles, quartiles, and quintiles are measures of position in a distribution. Percentiles divide the data into 100 equal parts, quartiles divide the data into four equal parts, and quintiles divide the data into five equal parts. The median is equal to the 50th percentile.

5. Why do we take the sum of squared deviations from the mean as a measure of spread, not the sum of the deviations themselves?
 - a. The sum of squared deviations from the mean is used as a measure of spread because it takes into account both positive and negative deviations from the mean. Squaring the deviations ensures that all deviations contribute equally to the measure, and the sum of squared deviations is always non-negative.
6. A distribution with a mean higher than the median is skewed. In what direction? Why? Give an intuitive explanation.
 - a. A distribution with a mean higher than the median is skewed to the right. This means that the tail of the distribution is elongated towards the right side. It occurs when there are a few extreme values on the right side of the distribution that pull the mean in that direction. An intuitive explanation is that the mean is more sensitive to extreme values than the median, so if there are a few very high values, they will have a larger impact on the mean, pulling it towards the right.
7. Extreme values are a challenge to data analysis if they are relevant for the question of the analysis. List two reasons why.
 - a. They can have a disproportionate influence on summary statistics such as the mean, causing them to be misleading.
 - b. They can affect the distribution itself, making it non-normal or altering its shape, and subsequently affecting the appropriateness of statistical models and assumptions.
8. What kind of real-life variables are likely to be well approximated by the normal distribution?
 - a. Real-life variables that are likely to be well approximated by the normal distribution include heights, weights, IQ scores, and error terms in regression models.
9. What are well approximated by the lognormal distribution? Give an example for each.
 - a. Real-life variables that are well approximated by the lognormal distribution include incomes, stock prices, and radioactive decay.

10. What is a box plot, what is a violin plot, and what are they used for?
- A box plot is a graphical representation of the distribution of a quantitative variable that displays the minimum, first quartile, median, third quartile, and maximum values. A violin plot is similar to a box plot but also includes a rotated kernel density plot on each side, providing a different way to visualize the distribution.
11. Based on what you have learnt about measurement scales and descriptive statistics, decide if it is possible to calculate the mean, mode, and median of the following variables that tell us information about the employees at a company:
- number of years spent in higher education
 - the level of education (high school, undergraduate, graduate, doctoral school)
 - field of education (e.g., IT, engineering, business administration)
 - binary variable that shows whether someone has a university degree.
12. Take Figure 3.9 in Case Study 3.C1. Describe its usage, its main geometric object and how it encodes information, and scaffolding. Would you want to add annotation to it? What and why?
13. Take Table 3.6 in Case Study 3.B1, Comparing hotel prices in Europe: Vienna vs. London. Describe its usage, encoding, and scaffolding. Would you do some things differently? What and why?
14. What kind of real-life variables are likely to be well approximated by the Bernoulli distribution? Give two examples.
- Real-life variables that are well approximated by the Bernoulli distribution include binary variables such as success/failure outcomes, heads/tails in a coin toss, and yes/no responses.
15. What kind of real-life variables are likely to be well approximated by the binomial distribution? Give two examples.
- Real-life variables that are well approximated by the binomial distribution include the number of successes in a fixed number of independent Bernoulli trials, such as the number of correct answers on a multiple-choice exam or the number of defective items in a production sample.
16. What kind of real-life variables are likely to be well approximated by the power-law distribution? Give two examples.

- a. Real-life variables that are well approximated by the power-law distribution include the distribution of city populations, the distribution of file sizes on a computer, and the distribution of website link connections.

Chapter 4:

1. Give an example with two independent events. Can independent events happen at the same time?
 - a. Two independent events are events that have no influence on each other's outcome. For example, flipping a coin and rolling a dice are two independent events. The outcome of flipping a coin does not affect the outcome of rolling a dice. Independent events can happen at the same time, as the outcome of one event does not impact the outcome of the other event.
2. Give an example of two mutually exclusive events. Can mutually exclusive events happen at the same time?
 - a. Two mutually exclusive events are events that cannot occur at the same time. If one event happens, the other event cannot happen. For example, getting heads and tails when flipping a coin are two mutually exclusive events. If the coin lands on heads, it cannot land on tails at the same time. Mutually exclusive events cannot happen at the same time.
3. What's the conditional probability of an event? Give an example.
 - a. The conditional probability of an event is the probability of that event happening given that another event has already occurred. For example, the conditional probability of getting heads on a coin flip given that the coin is fair is 0.5. It represents the likelihood of an event occurring given certain conditions.
4. What's the conditional mean of a variable? Give an example.
 - a. The conditional mean of a variable is the average value of that variable given certain conditions or criteria. For example, the conditional mean of the height of individuals given their age group would calculate the average height of individuals within specific age ranges. It provides insight into how a variable varies based on different conditions.

5. How is the correlation coefficient related to the covariance? What is the sign of each when two variables are negatively associated, positively associated, or independent?
 - a. The correlation coefficient measures the strength and direction of the linear relationship between two variables. It is related to the covariance as the covariance divided by the product of the standard deviations of the two variables. The sign of the correlation coefficient indicates the direction of the association: positive for a positive association, negative for a negative association, and zero for no association.
6. Describe in words what it means that hotel prices and distance to the city center are negatively correlated.
 - a. Hotel prices and distance to the city center being negatively correlated means that as the distance from the city center increases, the hotel prices tend to decrease. It suggests that hotels located closer to the city center generally have higher prices, while those further away have lower prices.
7. When we want to compare the mean of one variable for values of another variable, we need variation in the conditioning variable. Explain this.
 - a. When we want to compare the mean of one variable for values of another variable, we need variation in the conditioning variable. This is because if there is no variation in the conditioning variable, there will be no difference in the mean values of the other variable. Having variation allows us to observe how the mean of one variable changes across different values of the conditioning variable.
8. What's the difference between the sources of variation in x in experimental data and observational data?
 - a. The sources of variation in x in experimental data come from intentional manipulation by the researcher. The researcher controls the values of x to observe its effect on the outcome variable. In observational data, the sources of variation in x arise naturally and are not deliberately controlled by the researcher. Observational data reflects the existing variation in x in the population.
9. What's the joint distribution of two variables, and how can we visualize it?

- a. The joint distribution of two variables represents the distribution of their values occurring together. It provides information about the relationship between the two variables. It can be visualized using a scatterplot, which plots the values of one variable on the x-axis and the values of the other variable on the y-axis.
10. What's a scatterplot? What does it look like for two quantitative variables, each of which can be positive only, if the two variables are positively correlated?
- a. A scatterplot is a graphical representation of the relationship between two quantitative variables. For two positively correlated variables, the scatterplot would show a positive sloping pattern, where higher values of one variable correspond to higher values of the other variable. It would exhibit an upward trend from left to right.
11. What's a bin scatter, and what is it used for?
- a. A bin scatter is a visualization tool used to compare the means or other statistics of a variable across different bins or categories of another variable. It plots the means of the variable on the y-axis for each bin of the other variable on the x-axis. It allows for a comparison of means between different categories or bins.
12. What's a latent variable, and how can we use latent variables in data analysis? Give an example.
- a. A latent variable is a variable that cannot be directly observed or measured. It is inferred from other observable variables. In data analysis, latent variables can be used to represent underlying constructs or dimensions that cannot be measured directly. For example, intelligence or happiness can be latent variables.
13. List two ways to combine multiple measures of the same latent variable in your data for further analysis, and list an advantage and a disadvantage of each way.
- a. Two ways to combine multiple measures of the same latent variable in data analysis are factor analysis and summation of scores.
 - b. Factor analysis: It involves identifying a smaller number of underlying factors that explain the covariation among the multiple measures. An advantage of factor analysis is that it reduces the complexity of the data and provides a clearer understanding of the latent variable. However, a disadvantage is that the

interpretation of the factors may be subjective, and the results may not always be easily interpretable.

- c. Summation of scores: It involves adding up the individual scores on each measure to create a composite score. An advantage of this method is its simplicity and ease of interpretation. However, a disadvantage is that it assumes equal weighting of each measure, which may not always be appropriate or accurate.
14. You want to know if working on case studies in groups or working on them independently is a better way to learn coding in R. What would be your y and x variables here and how would you measure them?
- a. In the context of learning coding in R, the y variable could be the coding proficiency or mastery level, and the x variable could be the method of learning (working in groups or working independently). The y variable could be measured using a coding assessment or a self-assessment questionnaire. The x variable could be measured using a survey asking participants about their preferred method of learning. The measurement of y and x would depend on the specific criteria or metrics used to assess coding proficiency and the method of learning.
15. Can you tell from the shape of a bin scatter if y and x are positively correlated? Can you tell from it how strong their correlation is?
- a. From the shape of a bin scatter, we can determine if y and x are positively correlated. If the scatterplot shows an upward sloping pattern, it suggests a positive correlation. The strength of their correlation cannot be determined solely from the shape of the bin scatter plot. The correlation coefficient or other statistical measures would be needed to quantify the strength of their correlation.

Chapter 5:

1. When do we call a sample representative and is it connected to random sampling?

- a. A sample is called representative when it accurately reflects the characteristics and diversity of the population it is taken from. It is connected to random sampling in the sense that random sampling is a method used to ensure that a sample is representative. Random sampling involves selecting individuals from a population in a way that every individual has an equal chance of being selected. This helps to minimize bias and ensure that the sample is representative of the population.
2. What are the two parts of the inference process? List them, explain them, and give an example with the two parts.
 - a. The two parts of the inference process are:
 - i. Computing the confidence interval: This involves using statistical methods to calculate a range of values that likely contains the true value of the statistic in the population. The confidence interval provides a measure of uncertainty and represents the range of values that can be reasonably assumed to contain the true value. For example, a 95% confidence interval for the mean height of a population might be [160 cm, 170 cm], indicating that we can be 95% confident that the true mean height falls within this range.
 - ii. Assessing external validity: This step involves evaluating how well the sample represents the population of interest. It involves considering factors such as the similarity of the population and sample in terms of characteristics and dynamics. For example, if a study is conducted on a specific demographic group, the external validity would be high if the sample accurately reflects the demographic characteristics and experiences of the entire population.
 3. When do we say that results from analyzing our data have high external validity? Low external validity? Give an example for each.
 - a. Results from analyzing data have high external validity when the population or general pattern that the data represents is very similar to the population or general pattern that we care about. This means that the findings from the analysis can be confidently generalized to the larger population.
 - b. On the other hand, results have low external validity when there are significant differences between the population or general pattern represented by the data

and the population or general pattern we care about. This means that the findings may not accurately reflect the larger population and should be interpreted with caution.

- c. For example, if a study on the effects of a new drug is conducted on a small sample of healthy young adults, the results may have low external validity if the population of interest is elderly individuals with chronic health conditions.
4. What's the population, or general pattern, represented by the monthly time series of unemployment rates in Chile between 2000 and 2018?
 - a. The population or general pattern represented by the monthly time series of unemployment rates in Chile between 2000 and 2018 is the employment situation in Chile during that period. It represents the fluctuations in unemployment rates and the overall labor market conditions in the country over those 18 years.
5. What's the population, or general pattern, that represents the time series of GDP in Vietnam between 1990 and 2020?
 - a. The population or general pattern that represents the time series of GDP in Vietnam between 1990 and 2020 is the economic performance and growth of Vietnam during that period. It captures the changes in the country's gross domestic product over the 30-year timeframe.
6. Does it make sense to create a confidence interval for a statistic that you calculated using cross-country data with all countries in the world? If not, why not? If yes, what's the interpretation of that CI?
 - a. It may not make sense to create a confidence interval for a statistic calculated using cross-country data with all countries in the world. The reason is that the data from different countries may have different characteristics and dynamics, making it difficult to generalize the results to the entire world population.
 - b. If a confidence interval is created for such a statistic, the interpretation would involve acknowledging the limitation of generalizing the results to the entire world population. The confidence interval would provide a range of values that represents the uncertainty in estimating the statistic based on the available cross-country data.

7. What does the confidence interval show? What are the typical likelihoods used for them?
- The confidence interval shows a range of values within which the true value of the statistic is likely to fall. It represents the uncertainty associated with estimating the statistic from a sample.
 - Typical likelihoods used for confidence intervals are 90% and 95%. A confidence interval with a 90% likelihood indicates that if the sampling and estimation process were repeated many times, approximately 90% of the resulting confidence intervals would contain the true value of the statistic. Similarly, a confidence interval with a 95% likelihood indicates that approximately 95% of the resulting confidence intervals would contain the true value.
8. The proportion of daily losses of more than 2% on an investment portfolio is 5% in the data. Its confidence interval is [4,6] percent. Interpret these numbers.
- The confidence interval [4,6] percent for the proportion of daily losses of more than 2% on an investment portfolio (which is 5% in the data) indicates that we can be 95% confident that the true proportion of such losses in the population (from which the data is a sample) falls within the range of 4% to 6%. This means that there is a 95% likelihood that the actual proportion of losses is between 4% and 6%.
9. In the data that is a random sample of the population of your country from last year, 30-year-old market analysts earn 30% more than 25-year-old market analysts, on average. The 95% CI of this difference is [25,35] percent. Interpret these numbers.
- The 95% confidence interval [25,35] percent for the difference in earnings between 30-year-old and 25-year-old market analysts indicates that we can be 95% confident that the true average difference in earnings between these two groups in the population falls within the range of 25% to 35%. This means that there is a 95% likelihood that the actual average difference is between 25% and 35%.
10. In the example above, what can you conclude about the expected wage difference between 30 and 25 year-old market analysts in your country five years from now? In a different country five years from now?

- a. Based on the example above, it is difficult to conclude anything about the expected wage difference between 30 and 25-year-old market analysts in a specific country five years from now. The data used for the estimation came from a random sample of the population of a past year, and the confidence interval represents the uncertainty around the estimated difference in earnings. It does not provide information about future changes in wages or differences between countries.
11. How would you estimate the bootstrap standard error of the average of a variable from cross-sectional data?
- a. The bootstrap standard error of the average of a variable from cross-sectional data can be estimated by repeatedly resampling the data with replacement and calculating the average for each resampled dataset. The standard deviation of the averages from the resampled datasets gives an estimate of the bootstrap standard error.
12. What's the standard error formula for an average, and what does it imply about what makes the SE larger or smaller? Under what assumption does this formula work?
- a. The standard error formula for an average is given by the standard deviation of the variable divided by the square root of the sample size. The standard error measures the precision of the sample mean estimate and represents the variation in sample means that would be expected if multiple samples were taken from the same population.
- b. The formula tells us that the standard error gets smaller as the sample size increases. This means that larger sample sizes provide more precise estimates of the true population mean. The formula works under the assumption of random sampling and independence between observations.
13. How do you create the 95% CI of a statistic if you know its SE? How do you create the 90% CI?
- a. To create a 95% confidence interval of a statistic if its standard error (SE) is known, we can use the formula:
- b. $CI = \text{statistic} \pm (1.96 * SE)$
- c. To create a 90% confidence interval, we can use the formula:
- d. $CI = \text{statistic} \pm (1.645 * SE)$

- e. The confidence interval represents the range of values within which the true value of the statistic is likely to fall, with a given level of confidence.
14. Name two kinds of stability that are important challenges to external validity. Give an example for each when they are likely to result in low external validity.
- a. Two kinds of stability that are important challenges to external validity are temporal stability and contextual stability.
 - i. Temporal stability refers to the stability of the relationship or pattern over time. If the relationship between two variables changes significantly over time, it can result in low external validity. For example, if a study finds a strong positive correlation between education level and income in the 1990s, but this relationship weakens or becomes negative in the present day, the external validity of the study's findings may be low.
 - ii. Contextual stability refers to the stability of the relationship or pattern across different contexts or settings. If the relationship between variables differs depending on the context, it can result in low external validity. For example, if a study finds a strong positive correlation between exercise and weight loss in a controlled laboratory setting, but this relationship does not hold in real-world environments, the external validity of the study's findings may be low.
15. You downloaded data from the World Development Indicators database on GDP per capita and CO2 emission per capita, and find that their correlation coefficient is 0.7, with SE = 0.05. Create a 95% CI and interpret it.
- a. Given a correlation coefficient of 0.7 and a standard error of 0.05, we can create a 95% confidence interval by multiplying the standard error by 1.96 (the critical value for a 95% confidence level) and adding/subtracting the result from the correlation coefficient:
 - i. $CI = 0.7 \pm (1.96 * 0.05) = 0.7 \pm 0.098$
 - b. The resulting 95% confidence interval is [0.602, 0.798]. This means that we can be 95% confident that the true correlation coefficient between GDP per capita and CO2 emission per capita in the population falls within the range of 0.602 to 0.798.

Chapter 6:

1. Write down an example for a null and a two-sided alternative hypothesis.
 - a. Null hypothesis: The mean height of male students is equal to the mean height of female students.
Alternative hypothesis: The mean height of male students is not equal to the mean height of female students.
2. Write down an example for a null and a one-sided alternative hypothesis.
 - a. Null hypothesis: The new medication has no effect on reducing blood pressure.
Alternative hypothesis: The new medication leads to a decrease in blood pressure.
3. What is false positive? What is false negative? Give an example of a test, and explain what a false positive and a false negative would look like.
 - a. False positive refers to rejecting the null hypothesis when it is actually true. This means that we conclude there is a significant effect or relationship when there is none. An example of false positive is rejecting the null hypothesis that a new drug has no side effects, when in reality it does not have any side effects.
 - b. False negative refers to failing to reject the null hypothesis when it is actually false. This means that we fail to identify a significant effect or relationship that exists. An example of false negative is failing to reject the null hypothesis that a new treatment is ineffective, when in reality it is effective.
4. What is the level of significance or size of a test, and what is the power of a test? Give an example of a test, and explain the concept of size, significance, and power in the context of this example.
 - a. The level of significance, also known as the size of a test, determines how willing we are to make a Type I error (false positive). It represents the probability of incorrectly rejecting the null hypothesis when it is actually true. Typically, a level of significance is set at 0.05 or 0.01.
 - b. The power of a test is the probability of correctly rejecting the null hypothesis when it is false (avoiding a false negative). It represents the ability of a test to correctly detect a true effect or relationship.

- c. Example: Let's say we want to test if a new drug reduces the average cholesterol level in a population. The level of significance is set at 0.05, meaning we are willing to accept a 5% chance of making a Type I error. The power of the test depends on factors such as sample size, effect size, and variability. A larger sample size increases the power of the test, making it more likely to detect a true difference if it exists.
5. Tests on larger datasets have more power in general. Why?
- a. Tests on larger datasets have more power in general because they provide more information and reduce sampling error. With a larger sample size, the estimates of the parameters become more precise and the variability decreases. This allows for a better detection of true effects or relationships, resulting in higher power.
6. What is the p-value? The p-value of a test is 0.01. What does that mean and what can you do with that information?
- a. The p-value is a measure of the evidence against the null hypothesis. It represents the probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming the null hypothesis is true. A p-value of 0.01 means that if the null hypothesis is true, there is a 1% chance of observing a test statistic as extreme or more extreme than the one obtained.
 - b. With a p-value of 0.01, we can conclude that the observed data provide strong evidence against the null hypothesis. It suggests that the results are unlikely to occur by chance alone. Based on this information, we can reject the null hypothesis in favor of the alternative hypothesis.
7. Why is testing multiple hypotheses problematic?
- a. Testing multiple hypotheses is problematic because it increases the likelihood of making a Type I error (false positive). When multiple tests are performed, each with a certain level of significance, the overall probability of at least one test producing a significant result by chance alone increases. This is known as the problem of multiple comparisons.
8. What is p-hacking, and how can you minimize its perils when presenting the results of your analysis? Give an example.

- a. P-hacking refers to the practice of selectively analyzing data or conducting multiple tests until a desired result is obtained. It is a form of data manipulation that can lead to false positive results and invalid conclusions. To minimize its perils, it is important to pre-specify the analysis plan and hypotheses before conducting the tests. Researchers should also report all analyses performed, even those that did not yield significant results.
 - b. Example: A researcher conducts a study to investigate the effect of a new diet on weight loss. The study measures weight at multiple time points and performs various statistical tests. However, the researcher selectively reports only the tests that show a significant effect, while neglecting to mention the other tests that were not significant. This is an example of p-hacking.
9. What is the effect of p-hacking on published results that examine the same question? What does that imply for the usefulness of reading through the literature of published results? Give an example.
- a. P-hacking can have a significant effect on published results that examine the same question. If researchers selectively report or manipulate data until a desired result is obtained, it can lead to a distorted body of literature. This can result in biased conclusions and can undermine the reliability and replicability of the research findings. It is important to critically evaluate and consider the potential influence of p-hacking when reading through the literature of published results.
 - b. Example: A meta-analysis examines the effect of a certain medication on reducing blood pressure. However, many of the studies included in the meta-analysis have selectively reported significant results, while disregarding the non-significant results. This can lead to an overestimation of the true effect of the medication, creating a biased representation of the overall evidence.
10. You examine the wages of recent college graduates, and you want to test whether the starting wage of women is the same, on average, as the starting wage of men. Define the statistic you want to test. Define the population for which you can carry out the test if your data is a random sample of college graduates from your country surveyed in 2015. Write down the appropriate null and alternative hypotheses, and describe how you would carry out the test. What would be a false negative in this case? What would be a false positive?

- a. Statistic to test: The mean starting wage of women and men.
Population: College graduates from the country surveyed in 2015.
 - b. Null hypothesis: The mean starting wage of women is equal to the mean starting wage of men.
Alternative hypothesis: The mean starting wage of women is not equal to the mean starting wage of men.
 - c. To carry out the test, the researcher would collect a random sample of college graduates from the country surveyed in 2015. The sample would include data on the starting wages of both women and men. The researcher would then calculate the mean starting wage for each group and perform a statistical test, such as a t-test, to compare the means.
 - d. A false negative in this case would occur if the statistical test fails to reject the null hypothesis, suggesting that there is no significant difference in the mean starting wages of women and men, when in reality there is a difference.
 - e. A false positive would occur if the statistical test rejects the null hypothesis, suggesting that there is a significant difference in the mean starting wages of women and men, when in reality there is no difference.
11. You are testing whether an online advertising campaign had an effect on sales by comparing the average spending by customers who were exposed to the campaign with the average spending of customers who were not exposed. Define the statistic you want to test. What is the null and the alternative if your question is whether the campaign had a positive effect? What would be a false negative in this case? What would be a false positive? Which of the two do you think would have more severe business consequences?
- a. Statistic to test: Average spending by customers exposed to the advertising campaign compared to average spending by customers not exposed.
Null hypothesis: The average spending by customers exposed to the campaign is equal to the average spending by customers not exposed.
Alternative hypothesis: The average spending by customers exposed to the campaign is greater than the average spending by customers not exposed.
 - b. A false negative in this case would occur if the statistical test fails to reject the null hypothesis, indicating that there is no significant difference in average

spending between the two groups, when in reality there is a positive effect of the campaign.

- c. A false positive would occur if the statistical test rejects the null hypothesis, indicating that there is a significant difference in average spending between the two groups, suggesting a positive effect of the campaign, when in reality there is no effect.
 - d. False positive would have more severe business consequences as it may lead to investing in an advertising campaign that does not actually result in increased sales, leading to financial losses.
12. Consider the null hypothesis that there is no difference between the likelihood of bankruptcy of firms that were established more than three years ago and firms that were established less than three years ago. You carry out a test using data on all firms from a country in 2015, and the test produces a p-value of 0.001. What is your conclusion? What if the p-value was 0.20?
- a. Based on the given information, if the test produced a p-value of 0.001, the conclusion would be to reject the null hypothesis. This suggests that there is a significant difference in the likelihood of bankruptcy between firms established more than three years ago and firms established less than three years ago.
 - b. If the p-value was 0.20, the conclusion would be to fail to reject the null hypothesis. This suggests that there is not enough evidence to support a significant difference in the likelihood of bankruptcy between the two types of firms.
13. A randomly selected half of the employees of a customer service firm participated in a training program, while the other half didn't. How would you test whether the training had an effect on the satisfaction of the customers they serve? Describe all steps of the test procedure.
- a. To test whether the training program had an effect on customer satisfaction, the researcher would follow these steps:
 - i. Randomly select half of the employees to participate in the training program, while the other half does not participate.

- ii. Collect data on the satisfaction ratings of the customers served by each group of employees.
 - iii. Calculate the mean satisfaction rating for the group that participated in the training program and the group that did not participate.
 - iv. Perform a statistical test, such as a t-test, to compare the means of the two groups.
 - v. Determine the p-value, which represents the probability of observing a difference in satisfaction ratings as extreme as the one obtained, assuming the null hypothesis (the training program has no effect).
 - vi. If the p-value is below the pre-set level of significance (e.g., 0.05), reject the null hypothesis and conclude that the training program had a significant effect on customer satisfaction. Otherwise, fail to reject the null hypothesis.
14. Consider our example of online ads in Section 6.10, in which two versions of the same ad were shown to 1 million people each, and 100 followed up from the group that was shown the old version, while 130 followed through from the other group. Write down the test statistic, the null, and the alternative (go for two-sided alternative). Carry out the test using the t-statistic with the help of the following statistics: the follow-up rate for the new version is 0.000 13; for the old version 0.000 10; their difference is 0.000 03; the SE of that difference is 0.000 021. Interpret your decision.
- a. Test statistic: Difference in follow-up rates between the two versions of the ad.
Null hypothesis: The follow-up rates for the old and new versions of the ad are equal.
Alternative hypothesis: The follow-up rates for the old and new versions of the ad are not equal.
 - b. Using the given statistics (follow-up rate for the new version is 0.00013, for the old version is 0.00010, their difference is 0.00003, and the standard error of the difference is 0.000021), we can calculate the t-statistic.
 - c. $t\text{-statistic} = (\text{difference in follow-up rates} - \text{hypothesized difference}) / \text{standard error of the difference}$
 $t\text{-statistic} = (0.00003 - 0) / 0.000021$

- d. Interpreting the decision would involve comparing the calculated t-statistic to the critical value (e.g., using a t-distribution table or statistical software) for a given level of significance (e.g., 0.05 or 0.01). If the calculated t-statistic is greater than the critical value, we would reject the null hypothesis and conclude that there is a significant difference in the follow-up rates between the two versions of the ad. Otherwise, if the calculated t-statistic is less than the critical value, we would fail to reject the null hypothesis.
15. Consider the same online ad example as in the previous question, but now go for a one-sided alternative. Write down the test statistic, the null, and the alternative, and argue why you did it that way. Carry out the test using the information that the p-value of the two-sided test would be 0.07.
- a. For a one-sided alternative, we would focus on the deviation in one direction. Let's assume we are interested in testing if the new version of the ad has a higher follow-up rate than the old version.
- b. Test statistic: Difference in follow-up rates between the two versions of the ad.
Null hypothesis: The follow-up rate for the new version is equal to or lower than the follow-up rate for the old version.
Alternative hypothesis: The follow-up rate for the new version is higher than the follow-up rate for the old version.
- c. Using the given information that the p-value of the two-sided test would be 0.07, we would perform a one-sided test by dividing the p-value by 2 (since we are only interested in one side of the distribution). So the adjusted p-value would be $0.07 / 2 = 0.035$.
- d. To interpret the decision, we would compare the adjusted p-value to the pre-set level of significance (e.g., 0.05). If the adjusted p-value is less than the level of significance, we would reject the null hypothesis and conclude that the new version of the ad has a significantly higher follow-up rate than the old version. If the adjusted p-value is greater than or equal to the level of significance, we would fail to reject the null hypothesis.