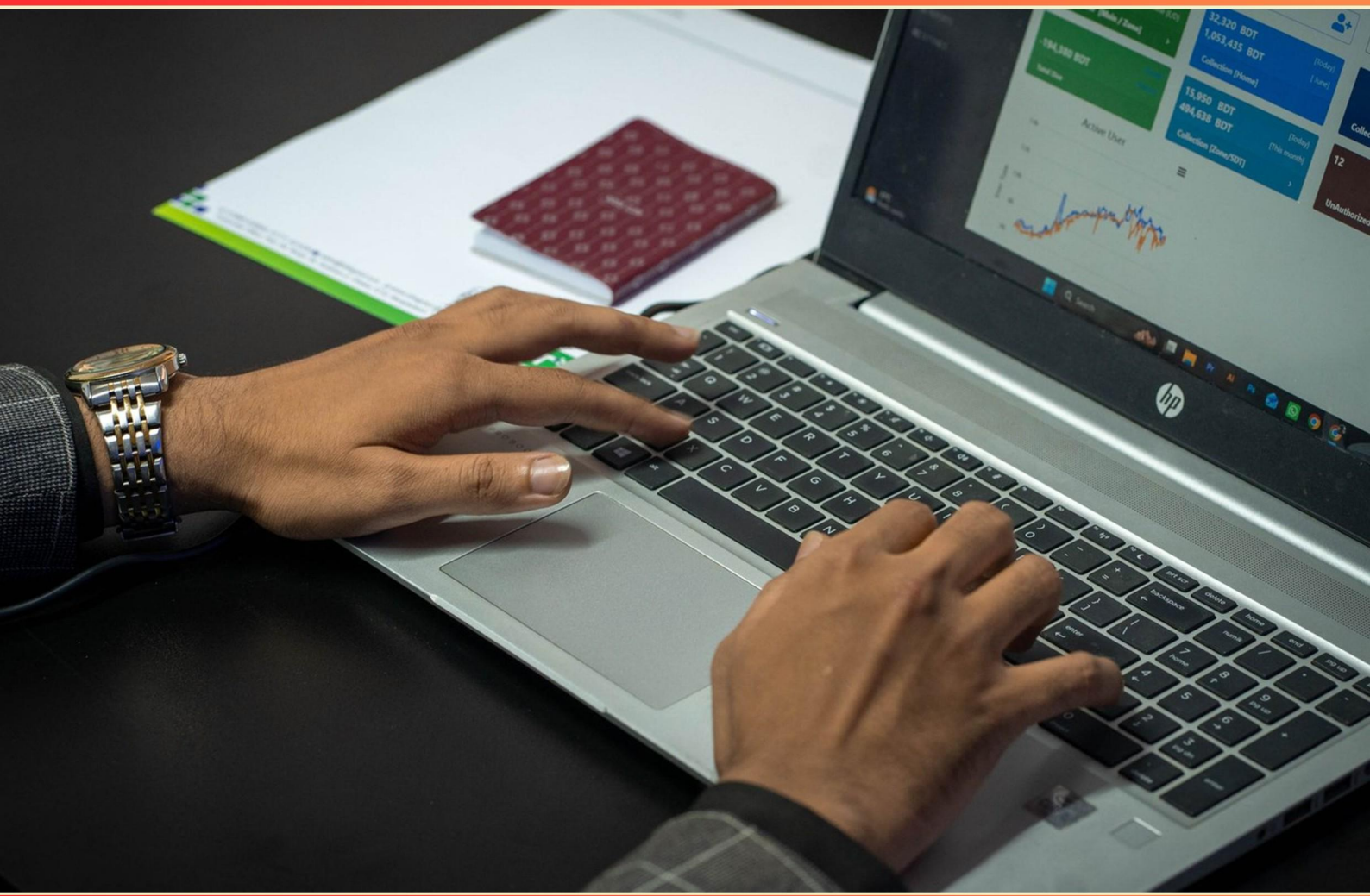
CERTNEXUS CERTIFIED DATA SCIENCE PRACTITIONER (CDSP) PRACTICE EXAM (SAMPLE)

STUDY GUIDE



SAMPLE STUDY GUIDE WITH LIMITED QUESTIONS

VISIT US ONLINE FOR THE FULL VERSION STUDY GUIDE

https://certnexuscdsp.examzify.com

Copyright © 2026 by Examzify - A Kaluba Technologies Inc. product.

ALL RIGHTS RESERVED.

No part of this book may be reproduced or transferred in any form or by any means, graphic, electronic, or mechanical, including photocopying, recording, web distribution, taping, or by any information storage retrieval system, without the written permission of the author.

Notice: Examzify makes every reasonable effort to obtain accurate, complete, and timely information about this product from reliable sources.

SAMPLE

Table of Contents

SAMPLE

Introduction

Preparing for a certification exam can feel overwhelming, but with the right tools, it becomes an opportunity to build confidence, sharpen your skills, and move one step closer to your goals. At Examzify, we believe that effective exam preparation isn't just about memorization, it's about understanding the material, identifying knowledge gaps, and building the test-taking strategies that lead to success.

This guide was designed to help you do exactly that.

Whether you're preparing for a licensing exam, professional certification, or entry-level qualification, this book offers structured practice to reinforce key concepts. You'll find a wide range of multiple-choice questions, each followed by clear explanations to help you understand not just the right answer, but why it's correct.

The content in this guide is based on real-world exam objectives and aligned with the types of questions and topics commonly found on official tests. It's ideal for learners who want to:

- Practice answering questions under realistic conditions,
- Improve accuracy and speed,
- Review explanations to strengthen weak areas, and
- Approach the exam with greater confidence.

We recommend using this book not as a stand-alone study tool, but alongside other resources like flashcards, textbooks, or hands-on training. For best results, we recommend working through each question, reflecting on the explanation provided, and revisiting the topics that challenge you most.

Remember: successful test preparation isn't about getting every question right the first time, it's about learning from your mistakes and improving over time. Stay focused, trust the process, and know that every page you turn brings you closer to success.

Let's begin.

SAMPLE

How to Use This Guide

This guide is designed to help you study more effectively and approach your exam with confidence. Whether you're reviewing for the first time or doing a final refresh, here's how to get the most out of your Examzify study guide:

1. Start with a Diagnostic Review

Skim through the questions to get a sense of what you know and what you need to focus on. Your goal is to identify knowledge gaps early.

2. Study in Short, Focused Sessions

Break your study time into manageable blocks (e.g. 30 – 45 minutes). Review a handful of questions, reflect on the explanations.

3. Learn from the Explanations

After answering a question, always read the explanation, even if you got it right. It reinforces key points, corrects misunderstandings, and teaches subtle distinctions between similar answers.

4. Track Your Progress

Use bookmarks or notes (if reading digitally) to mark difficult questions. Revisit these regularly and track improvements over time.

5. Simulate the Real Exam

Once you're comfortable, try taking a full set of questions without pausing. Set a timer and simulate test-day conditions to build confidence and time management skills.

6. Repeat and Review

Don't just study once, repetition builds retention. Re-attempt questions after a few days and revisit explanations to reinforce learning. Pair this guide with other Examzify tools like flashcards, and digital practice tests to strengthen your preparation across formats.

There's no single right way to study, but consistent, thoughtful effort always wins. Use this guide flexibly, adapt the tips above to fit your pace and learning style. You've got this!

SAMPLE

Questions

SAMPLE

1. What does the p-value represent in hypothesis testing?

A. The probability of obtaining a result given that the alternative hypothesis is true

B. The probability of obtaining a result from the test given that the null hypothesis is true

C. The probability of a Type I error occurring

D. The probability of a Type II error occurring

2. What is the term for variables that change indirectly in an experiment?

A. Independent Variable

B. Dependent Variable

C. Confounding Variable

D. Control Variable

3. Which process converts data from one type to a coded value of a different type?

A. Data Transformation

B. Data Encoding

C. Data Aggregation

D. Data Extraction

4. Which analysis method is primarily focused on predicting future outcomes based on current data?

A. Descriptive Analysis

B. Predictive Analysis

C. Exploratory Analysis

D. Prescriptive Analysis

5. Which clustering algorithm starts with each data example in its own cluster?

A. K-means clustering

B. HAC (hierarchical agglomerative clustering)

C. DBSCAN

D. Mean shift

SAMPLE

6. What is defined as a value that deviates significantly from the main distribution of values?

A. Average

B. Outliers

C. Median

D. Standard deviation

7. What does AUC stand for in the context of model evaluation?

A. Area Under Curve

B. Average Utility Curve

C. Aggregate User Count

D. Algorithmic Understanding Coefficient

8. What best describes the primary goal of data science?

A. Accumulating data

B. Extracting value from data

C. Presenting data in a meaningful way

D. All of the above

9. What term is defined as the average of all numbers in a data set?

A. Median

B. Mode

C. Arithmetic mean

D. Geometric mean

10. What is meant by collinearity in regression analysis?

A. Non-linearity between independent variables

B. Relationship between dependent and independent variables

C. When two features exhibit a linear relationship

D. Multicollinearity issues

SAMPLE

Answers

SAMPLE

1. B
2. B
3. B
4. B
5. B
6. B
7. A
8. D
9. C
10. C

SAMPLE

Explanations

SAMPLE

1. What does the p-value represent in hypothesis testing?

A. The probability of obtaining a result given that the alternative hypothesis is true

B. The probability of obtaining a result from the test given that the null hypothesis is true

C. The probability of a Type I error occurring

D. The probability of a Type II error occurring

The p-value represents the probability of obtaining a result from the test, assuming that the null hypothesis is true. This is a fundamental concept in hypothesis testing. When a p-value is calculated, it helps to determine the strength of the evidence against the null hypothesis. A smaller p-value indicates that the observed data would be very unlikely under the assumption that the null hypothesis holds true, thus suggesting that there is stronger evidence in favor of the alternative hypothesis. By focusing on the probability of the observed data (or something more extreme) under the null hypothesis, researchers can make informed decisions about whether to reject the null hypothesis. If the p-value is less than a predetermined significance level (commonly set at 0.05), this typically leads to the rejection of the null hypothesis, implying that the observed data are statistically significant. In this context, it's essential to understand the other choices. The first option discusses the probability of obtaining a result if the alternative hypothesis is true, which does not accurately describe what the p-value measures. The third option relates the p-value to the probability of a Type I error (rejecting a true null hypothesis), but the p-value itself is not defined as such. Lastly, the fourth option refers to a Type II error, which

2. What is the term for variables that change indirectly in an experiment?

A. Independent Variable

B. Dependent Variable

C. Confounding Variable

D. Control Variable

The term that describes variables that change indirectly in an experiment is indeed associated with the concept of dependent variables. A dependent variable is the one that researchers measure in an experiment to see how it is affected by other variables. This variable is called "dependent" because its value depends on changes made to the independent variable(s). In the context of an experiment, as the independent variables are manipulated, the researcher looks for changes in the dependent variable. For instance, if a study examines how different amounts of fertilizer (independent variable) affect plant growth (dependent variable), the plant growth is what is being measured and is expected to change in response to varying amounts of fertilizer. While independent variables are manipulated directly, and control variables are kept constant to prevent influencing the outcome, confounding variables are factors that could potentially confuse the results. It's the dependent variable that reflects any changes due to the manipulation of the independent variables, making it crucial in understanding the dynamics of the experiment.

SAMPLE

3. Which process converts data from one type to a coded value of a different type?

A. Data Transformation

B. Data Encoding

C. Data Aggregation

D. Data Extraction

The process that converts data from one type to a coded value of a different type is known as Data Encoding. This involves taking raw data and transforming it into a specific format that is suitable for processing or storage, often simplifying or concealing the original data value in the process. Encoding is particularly useful when it involves categorical data that need to be represented numerically so that algorithms can utilize it effectively. For instance, in machine learning, converting categorical variables such as 'Male' and 'Female' into numerical values (e.g., 0 and 1) helps algorithms to interpret the data correctly without ambiguity. This process is essential in preparing data for model building and ensures that the models can function with numeric input efficiently. In contrast, the other processes mentioned are related but serve different purposes. Data Transformation refers to the broader scope of altering the structure or format of data, which can include encoding among other methods. Data Aggregation involves summarizing data, usually to provide insights at a higher level, whereas Data Extraction focuses on retrieving data from various sources. While these processes may intersect, they do not specifically address the conversion of data to coded values as Data Encoding does.

4. Which analysis method is primarily focused on predicting future outcomes based on current data?

A. Descriptive Analysis

B. Predictive Analysis

C. Exploratory Analysis

D. Prescriptive Analysis

Predictive analysis is focused specifically on forecasting future events or outcomes by utilizing current and historical data. It employs various statistical techniques and machine learning algorithms to identify patterns and relationships within the data, which can then be extrapolated to make informed predictions about future trends. This method is particularly valuable in fields such as finance, marketing, and healthcare, where understanding potential future scenarios can guide decision-making. By analyzing patterns in existing data, predictive analysis allows organizations to assess risks, allocate resources more effectively, and improve strategic planning. In contrast, descriptive analysis is primarily concerned with summarizing historical data to provide insight into what has happened, while exploratory analysis is used to understand the underlying structure of the data and discover patterns without a specific focus on prediction. Prescriptive analysis goes a step further by suggesting actions based on predictions, but it does not primarily focus on forecasting itself. Thus, predictive analysis is the most appropriate choice for the question about predicting future outcomes based on current data.

SAMPLE

5. Which clustering algorithm starts with each data example in its own cluster?

A. K-means clustering

B. HAC (hierarchical agglomerative clustering)

C. DBSCAN

D. Mean shift

The correct answer is hierarchical agglomerative clustering (HAC). This clustering algorithm begins by treating each individual data point as a separate cluster. As the algorithm progresses, it merges these clusters based on a specified linkage criterion, gradually reducing the total number of clusters until the desired number is achieved or until all points are merged into a single cluster. This approach is particularly useful when the structure of the data is hierarchical in nature, allowing for a detailed exploration of how clusters are formed at various levels of granularity. The initial state of having each data point in its own cluster enables HAC to build a dendrogram, which visually represents the merging of clusters and can help to reveal the data's intrinsic structure. K-means clustering, by contrast, starts with predefined cluster centroids and assigns data points to the nearest centroid, which is a fundamentally different approach. DBSCAN forms clusters based on density and does not require a predetermined number of clusters, while mean shift finds modes in the data distribution rather than starting with individual points as clusters. Each of these algorithms has distinct methodologies and objectives that set them apart from hierarchical agglomerative clustering.

6. What is defined as a value that deviates significantly from the main distribution of values?

A. Average

B. Outliers

C. Median

D. Standard deviation

The term that describes a value that deviates significantly from the main distribution of values is outliers. Outliers are observations that fall far away from the overall pattern of data, often located in the tails of the distribution. They can arise from variability in the data, measurement errors, or other external factors that do not fit within the general trend of the dataset. Identifying outliers is crucial in data analysis because they can significantly influence statistical measures, such as the mean and standard deviation, leading to potential misinterpretation of the data. In many data analyses, outliers may be investigated further to determine their cause and whether they should be included or excluded from analysis, as they provide insight into data quality and underlying processes. The other options, such as average, median, and standard deviation, relate to central tendency and dispersion measures rather than representing deviations from the norm. While the average and median describe typical values around which data might cluster, and standard deviation measures the spread or dispersion of data points, they do not specifically refer to extreme values or deviations like outliers do.

SAMPLE

7. What does AUC stand for in the context of model evaluation?

A. Area Under Curve

B. Average Utility Curve

C. Aggregate User Count

D. Algorithmic Understanding Coefficient

AUC stands for Area Under Curve, which is a critical metric used in evaluating the performance of classification models, particularly in the context of binary classification problems. The AUC measures the ability of a model to distinguish between the positive and negative classes. Specifically, it refers to the area under the ROC (Receiver Operating Characteristic) curve, which is a graphical representation plotting the true positive rate against the false positive rate at various threshold settings. An AUC value of 1 indicates a perfect model that can perfectly distinguish between the two classes, while a value of 0.5 suggests that the model performs no better than chance. Values below 0.5 indicate a model that is performing poorly or worse than random guessing. AUC is particularly valuable because it provides a single metric that summarizes the overall ability of a model to discriminate between classes across all possible thresholds, making it widely applicable in practice for model comparison and selection. In contrast to the other options, which do not accurately capture the relevant meaning in the context of model evaluation, AUC distinctly represents a well-established standard in assessing classification model effectiveness.

8. What best describes the primary goal of data science?

A. Accumulating data

B. Extracting value from data

C. Presenting data in a meaningful way

D. All of the above

The primary goal of data science is comprehensively encapsulated by the idea of extracting value from data. However, this process inherently includes various steps such as accumulating data and presenting it in a meaningful way. Accumulating data is essential because it forms the foundational layer upon which all analyses are built. Without sufficient and relevant data, the potential for meaningful insights is severely limited. Once the data is collected, the next step involves extracting value, which is critical as it focuses on deriving insights, patterns, and actionable information that can influence decision-making and drive strategies within an organization. This is often the ultimate aim of a data science endeavor. Finally, presenting data in a meaningful way ensures that the insights gleaned from analysis are communicated effectively to stakeholders. Visualization and presentation are crucial for making complex data more understandable and relatable to audiences who may not have technical expertise. By intertwining accumulating data, extracting value, and presenting findings, the comprehensive answer that encompasses all these factors reflects the full scope and goal of data science efforts. Therefore, acknowledging all these aspects together correctly illustrates the multifaceted nature of data science.

SAMPLE

9. What term is defined as the average of all numbers in a data set?

A. Median

B. Mode

C. Arithmetic mean

D. Geometric mean

The term that accurately describes the average of all numbers in a data set is known as the arithmetic mean. This statistical measure is calculated by summing all the values in the data set and then dividing that total by the number of values present. The arithmetic mean is widely used because it provides a representative value of the data set, especially when the data points are evenly distributed. In contrast, the median refers to the middle value in a data set when the numbers are arranged in order, and it is sensitive to the data distribution, particularly in skewed distributions. The mode indicates the most frequently occurring value in a data set; it highlights the value that appears the most often but does not represent an average. The geometric mean is another type of average, calculated by multiplying all the numbers in a data set together and then taking the nth root (where n is the count of values). This measure is particularly useful in situations with multiplicative relationships or for data that spans several orders of magnitude. Understanding these distinctions helps in selecting the appropriate averaging method based on the characteristics of the data being analyzed.

10. What is meant by collinearity in regression analysis?

A. Non-linearity between independent variables

B. Relationship between dependent and independent variables

C. When two features exhibit a linear relationship

D. Multicollinearity issues

Collinearity in regression analysis refers to the scenario where two or more features (independent variables) exhibit a strong linear relationship with each other. This means that one independent variable can be predicted from the other(s) with a high degree of accuracy. In practice, this can create challenges in regression, as it complicates the estimation of the individual effects of each feature on the dependent variable. When collinearity is present, it becomes difficult to assess the contribution of each independent variable to the prediction because their effects are intertwined. Understanding collinearity is crucial for data scientists and statisticians as it can lead to inflated standard errors of the coefficients, making hypothesis tests unreliable. While this term is often mentioned in relation to multicollinearity, which refers specifically to the situation where multiple independent variables are correlated, collinearity broadly addresses the linear relationship between two features. In contrast, non-linearity between independent variables does not directly relate to collinearity, as collinearity specifically involves linear relationships. The relationship between dependent and independent variables pertains more to the overall regression model rather than the interrelationships of the independent variables themselves. Multicollinearity issues are a specific scenario under the umbrella of collinearity but pertain more to situations where three

SAMPLE

Next Steps

Congratulations on reaching the final section of this guide. You've taken a meaningful step toward passing your certification exam and advancing your career.

As you continue preparing, remember that consistent practice, review, and self-reflection are key to success. Make time to revisit difficult topics, simulate exam conditions, and track your progress along the way.

If you need help, have suggestions, or want to share feedback, we'd love to hear from you. Reach out to our team at hello@examzify.com.

Or visit your dedicated course page for more study tools and resources:

https://certnexuscdsp.examzify.com

We wish you the very best on your exam journey. You've got this!

SAMPLE