Statistics & Probability

Math for Data Science

Math plays a key role in data science as it forms the foundation for building models, analyzing data and making predictions. Understanding the right math topics helps you apply algorithms effectively in real-world problems.

- Linear Algebra: for working with vectors, matrices and data transformations
- Statistics & Probability: for data analysis, hypothesis testing and predictions
- Calculus: for optimization and understanding how models learn

Linear Algebra

Linear Algebra is the foundation for many machine learning algorithms. It provides the tools to represent and manipulate datasets, features and transformations.

- Scalars, Vectors and Matrices: building blocks of datasets and features.
- Linear Combinations: Key in regression models and Principal Component Analysis (PCA).
- Vector Operations and Dot Product: Used in gradient descent and similarity measures.
- Matrices and Matrix operations: Essential for solving equations and optimizing machine learning models
- Linear Transformation: Operations for reshaping data, often used in PCA and feature scaling.
- Solving systems of linear equations: Essential for finding model parameters, such as in linear regression.
- Eigenvalues and Eigenvectors: for understanding variance and principal components.
- Singular Value Decomposition (SVD): Widely used in dimensionality reduction, data compression and noise reduction.
- Vector norms: for regularization techniques like Lasso and Ridge
- Measures of Distance: Cosine similarity for text similarity, Euclidean and Manhattan distances for clustering.

Probability and Statistics

Probability and Statistics are pillars of Data Science. They help us quantify uncertainty, interpret data and make predictions with confidence.

Probability for data science

- Sample space and Types of events: Foundation for analyzing outcomes.
- Probability Rules: Important for forecasting and evaluating events.
- Conditional Probability: Used in classification, recommender systems and risk modeling.
- Bayes' Theorem: Key for updating predictions with new data, in models like Naive Bayes.
- Random Variables & probability distributions: Basis for modeling uncertainty and hypothesis testing.

Statistics for data science

- Descriptive Statistics: Summarizes dataset characteristics (mean, median, variance), helping understand and visualize data patterns.
- Inferential Statistics: Draws conclusions about a population from a sample, essential for predicting and testing hypotheses in data science.
- Point estimates & confidence intervals: Measuring accuracy of predictions.
- Hypothesis testing: Includes p-value , Type I and II errors
- Statistical Tests: T-test, Paired T-test, F-Test, z-test, Chi-square Test (used for feature selection).
- Correlation: Pearson (linear), Spearman (ranked data), Cosine similarity (vector similarity).
- Sampling techniques: Simple random, stratified, cluster sampling, etc.

Calculus

Calculus is important for optimizing models (the process of adjusting model parameters to minimize error). For a deeper dive into specific areas and their relevance to machine learning, explore the individual articles outlined below:

- Differentiation: Measuring changes in parameters.
- Partial Derivatives: Computing gradients for multivariable functions.
- Gradient Descent: Core optimization algorithm for training ML models.
- Chain Rule: In Backpropagation applying the chain rule in neural networks.
- Jacobian and Hessian Matrices: Providing gradient mapping and second-order optimization.
- Taylor's series: Approximating complex functions for easier computation.
- Higher-Order Derivatives: Capturing curvature for optimization analysis.
- Fourier Transformations: Applied in signal processing and feature extraction.
- Area under the curve: Used in evaluation metrics like AUC-ROC.

Essential Math Topics and Applications to Become a Master in Data Science

# Hi Kagglers,

Mathematics is the bedrock of any contemporary discipline of science. Almost all the techniques of modern data science, including machine learning, have a deep mathematical underpinning.
It goes without saying that you will absolutely need all the other pearls of knowledge—programming ability, some amount of business acumen, and your unique analytical and inquisitive mindset—about the data to function as a top data scientist. But it always pays to know the machinery under the hood, rather than just being the person behind the wheel with no knowledge about the car. Therefore, a solid understanding of the mathematical machinery behind the cool algorithms will give you an edge among your peers.
Here are my suggestions for the topics to study to be at the top of the game in data science.

1. Functions, Variables, Equations, and Graphs
This area of math covers the basics, from the equation of a line to the binomial theorem and everything in between:
1.Logarithm, exponential, polynomial functions, rational numbers
2.Basic geometry and theorems, trigonometric identities
3.Real and complex numbers, basic properties
4.Series, sums, inequalities
5.Graphing and plotting, Cartesian and polar coordinates, conic sections
* Where You Might Use It*
If you want to understand how a search runs faster on a million-item database after you've sorted it, you will come across the concept of "binary search." To understand the dynamics of it, you need to understand logarithms and recurrence equations. Or, if you want to analyze a time series, you may come across concepts like "periodic functions" and "exponential decay."

2. Statistics
The importance of having a solid grasp over essential concepts of statistics and probability cannot be overstated. Many practitioners in the field actually consider classical (non-neural network) machine learning to be nothing but statistical learning. The subject is vast, and focused planning is critical to cover the most essential concepts:
1.Data summaries and descriptive statistics, central tendency, variance, covariance, correlation
2.Basic probability: basic idea, expectation, probability calculus, Bayes' theorem, conditional probability
3.Probability distribution functions: uniform, normal, binomial, chi-square, Student's t-distribution, central limit theorem, Linear regression, regularization
4.Sampling, measurement, error, random number generation
5.Hypothesis testing, A/B testing, confidence intervals, p-values, ANOVA, t-test
Where You Might Use It
In interviews. If you can show you've mastered these concepts, you will impress the other side of the table fast. And you will use them nearly every day as a data scientist.

3. Linear Algebra
This is an essential branch of mathematics for understanding how machine-learning algorithms work on a stream of data to create insight. Everything from friend suggestions on Facebook, to song recommendations on Spotify, to transferring your selfie to a Salvador Dali-style portrait using deep transfer learning involves matrices and matrix algebra.
Here are the essential topics to learn:
1.Basic properties of matrix and vectors: scalar multiplication, linear transformation, transpose, conjugate, rank, determinant
2.Inner and outer products, matrix multiplication rule and various algorithms, matrix inverse
3.Special matrices: square matrix, identity matrix, triangular matrix, idea about sparse and dense matrix, unit vectors, symmetric matrix, Hermitian, skew-Hermitian and unitary matrices
4.Matrix factorization concept/LU decomposition, Gaussian/Gauss-Jordan elimination, solving Ax=b linear system of equation
5.Vector space, basis, span, orthogonality, orthonormality, linear least square

Eigenvalues, eigenvectors, diagonalization, singular value decomposition
Where You Might Use It
If you have used the dimensionality reduction technique principal component analysis, then you have likely used the singular value decomposition to achieve a compact dimension representation of your data set with fewer parameters. All neural network algorithms use linear algebra techniques to represent and process network structures and learning operations.

4. Calculus
Whether you loved or hated it in college, calculus pops up in numerous places in data science and machine learning. It lurks behind the simple-looking analytical solution of an ordinary least squares problem in linear regression or embedded in every back-propagation your neural network makes to learn a new pattern. It is an extremely valuable skill to add to your repertoire.
Here are the topics to learn:
1.Functions of a single variable, limit, continuity, differentiability
2.Mean value theorems, indeterminate forms, L'Hospital's rule
3.Maxima and minima
4.Product and chain rule
5.Taylor's series, infinite series summation/integration concepts
Fundamental and mean value-theorems of integral calculus, evaluation of definite and improper integrals
6.Beta and gamma functions
7.Functions of multiple variables, limit, continuity, partial derivatives
8.Basics of ordinary and partial differential equations
Where You Might Use It
Ever wondered how exactly a logistic regression algorithm is implemented? There is a high chance it uses a method called "gradient descent" to find the minimum loss function. To understand how this works, you need to use concepts from calculus: gradient, derivatives, limits, and chain rule.

Discrete Mathematics:
This area is not discussed as often in data science, but all modern data science is done with the help of computational systems, and discrete math is at the heart of such systems. A refresher in discrete math will include concepts critical to daily use of algorithms and data structures in analytics project:
1.Sets, subsets, power sets
2.Counting functions, combinatorics, countability
3.Basic proof techniques: induction, proof by contradiction
4.Basics of inductive, deductive, and propositional logic
5.Basic data structures: stacks, queues, graphs, arrays, hash tables, trees
6.Graph properties: connected components, degree, maximum flow/minimum cut concepts, graph coloring
7.Recurrence relations and equations
8.Growth of functions and O(n) notation concept

Where You Might Use It

In any social network analysis, you need to know the properties of a graph and fast algorithm to search and traverse the network. In any choice of algorithm, you need to understand the time and space complexity—i.e., how the running time and space requirement grows with input data size, by using O(n) (Big-Oh) notation.

Optimization and Operation Research Topics

These topics are most relevant in specialized fields like theoretical computer science, control theory, or operation research. But a basic understanding of these powerful techniques can also be fruitful in the practice of machine learning. Virtually every machine-learning algorithm aims to minimize some kind of estimation error subject to various constraints—which is an optimization problem.

Here are the topics to learn:

1. Basics of optimization, how to formulate the problem
2. Maxima, minima, convex function, global solution
3. Linear programming, simplex algorithm
4. Integer programming
5. Constraint programming, knapsack problem
6. Randomized optimization techniques: hill climbing, simulated annealing, genetic algorithms

Where You Might Use It

Simple linear regression problems using least-square loss function often have an exact analytical solution, but logistic regression problems don’t. To understand the reason, you need to be familiar with the concept of “convexity” in optimization. This line of investigation will also illuminate why we must remain satisfied with “approximate” solutions in most machine-learning problems.

Though there are a lot of things to learn, there are excellent resources online. After a refresher on these topics (which you probably studied as an undergrad) and learning new concepts, you will be empowered to hear the hidden music in your daily data analysis and machine-learning projects. And that’s a big leap toward becoming an amazing data scientist.

Essential Math for Data Science

The key topics to master to become a better data scientist

Mathematics is the bedrock of any contemporary discipline of science. Almost all the techniques of modern data science, including machine learning, have a deep mathematical underpinning.

It goes without saying that you will absolutely need all the other pearls of knowledge—programming ability, some amount of business acumen, and your unique analytical and inquisitive mindset—about the data to function as a top data scientist. But it always pays to know the machinery under the hood, rather than just being the person behind the wheel with no knowledge about the car. Therefore, a solid understanding of the mathematical machinery behind the cool algorithms will give you an edge among your peers.

The knowledge of this essential math is particularly important for newcomers arriving at data science from other professions: hardware engineering, retail, the chemical process industry, medicine and health care, business management, etc. Although such fields may require experience with spreadsheets, numerical calculations, and projections, the math skills required in data science can be significantly different.

Consider a web developer or business analyst. They may be dealing with a lot of data and information on a daily basis, but there may not be an emphasis on rigorous modeling of that data. Often, the emphasis is on using the data for an immediate need and moving on, rather than on deep scientific exploration. Data science, on the other hand, should always be about the science (not the data). Following that thread, certain tools and techniques become indispensable. Most are the hallmarks of the sound scientific process:

- Modeling a process (physical or informational) by probing the underlying dynamics
- Constructing hypotheses
- Rigorously estimating the quality of the data source
- Quantifying the uncertainty around the data and predictions

- Identifying the hidden pattern from the stream of information
- Understanding the limitation of a model
- Understanding mathematical proof and the abstract logic behind it

Data science, by its very nature, is not tied to a particular subject area and may deal with phenomena as diverse as cancer diagnoses and social behavior analysis. This produces the possibility of a dizzying array of n-dimensional mathematical objects, statistical distributions, optimization objective functions, etc.

Here are my suggestions for the topics to study to be at the top of the game in data science.

Functions, Variables, Equations, and Graphs

This area of math covers the basics, from the equation of a line to the binomial theorem and everything in between:

- Logarithm, exponential, polynomial functions, rational numbers
- Basic geometry and theorems, trigonometric identities
- Real and complex numbers, basic properties
- Series, sums, inequalities
- Graphing and plotting, Cartesian and polar coordinates, conic sections

Where You Might Use It

If you want to understand how a search runs faster on a million-item database after you've sorted it, you will come across the concept of "binary search." To understand the dynamics of it, you need to understand logarithms and recurrence equations. Or, if you want to analyze a time series, you may come across concepts like "periodic functions" and "exponential decay."

Statistics

The importance of having a solid grasp over essential concepts of statistics and probability cannot be overstated. Many practitioners in the field actually consider classical (non-neural network) machine learning to be nothing but statistical learning. The subject is vast, and focused planning is critical to cover the most essential concepts:

- Data summaries and descriptive statistics, central tendency, variance, covariance, correlation
- Basic probability: basic idea, expectation, probability calculus, Bayes' theorem, conditional probability
- Probability distribution functions: uniform, normal, binomial, chi-square, Student's t-distribution, central limit theorem
- Sampling, measurement, error, random number generation
- Hypothesis testing, A/B testing, confidence intervals, p-values
- ANOVA, t-test
- Linear regression, regularization

Where You Might Use It

In interviews. If you can show you've mastered these concepts, you will impress the other side of the table fast. And you will use them nearly every day as a data scientist.

Linear Algebra

This is an essential branch of mathematics for understanding how machine-learning algorithms work on a stream of data to create insight. Everything from friend suggestions on Facebook, to song recommendations on Spotify, to transferring your selfie to a Salvador Dali-style portrait using deep transfer learning involves matrices and matrix algebra. Here are the essential topics to learn:

- Basic properties of matrix and vectors: scalar multiplication, linear transformation, transpose, conjugate, rank, determinant
- Inner and outer products, matrix multiplication rule and various algorithms, matrix inverse

- Special matrices: square matrix, identity matrix, triangular matrix, idea about sparse and dense matrix, unit vectors, symmetric matrix, Hermitian, skew-Hermitian and unitary matrices
- Matrix factorization concept/LU decomposition, Gaussian/Gauss-Jordan elimination, solving Ax=b linear system of equation
- Vector space, basis, span, orthogonality, orthonormality, linear least square
- Eigenvalues, eigenvectors, diagonalization, singular value decomposition

Where You Might Use It

If you have used the dimensionality reduction technique principal component analysis, then you have likely used the singular value decomposition to achieve a compact dimension representation of your data set with fewer parameters. All neural network algorithms use linear algebra techniques to represent and process network structures and learning operations.

Calculus

Whether you loved or hated it in college, calculus pops up in numerous places in data science and machine learning. It lurks behind the simple-looking analytical solution of an ordinary least squares problem in linear regression or embedded in every back-propagation your neural network makes to learn a new pattern. It is an extremely valuable skill to add to your repertoire. Here are the topics to learn:

- Functions of a single variable, limit, continuity, differentiability
- Mean value theorems, indeterminate forms, L'Hospital's rule
- Maxima and minima
- Product and chain rule
- Taylor's series, infinite series summation/integration concepts
- Fundamental and mean value-theorems of integral calculus, evaluation of definite and improper integrals
- Beta and gamma functions
- Functions of multiple variables, limit, continuity, partial derivatives
- Basics of ordinary and partial differential equations

Where You Might Use It

Ever wondered how exactly a logistic regression algorithm is implemented? There is a high chance it uses a method called “gradient descent” to find the minimum loss function. To understand how this works, you need to use concepts from calculus: gradient, derivatives, limits, and chain rule.

Discrete Math

This area is not discussed as often in data science, but all modern data science is done with the help of computational systems, and discrete math is at the heart of such systems. A refresher in discrete math will include concepts critical to daily use of algorithms and data structures in analytics project:

- Sets, subsets, power sets
- Counting functions, combinatorics, countability
- Basic proof techniques: induction, proof by contradiction
- Basics of inductive, deductive, and propositional logic
- Basic data structures: stacks, queues, graphs, arrays, hash tables, trees
- Graph properties: connected components, degree, maximum flow/minimum cut concepts, graph coloring
- Recurrence relations and equations
- Growth of functions and O(n) notation concept

Where You Might Use It

In any social network analysis, you need to know the properties of a graph and fast algorithm to search and traverse the network. In any choice of algorithm, you need to understand the time and space complexity—i.e., how the running time and space requirement grows with input data size, by using O(n) (Big-Oh) notation.

Optimization and Operation Research Topics

These topics are most relevant in specialized fields like theoretical computer science, control theory, or operation research. But a basic understanding of these powerful techniques can also be fruitful in the practice of machine learning. Virtually every machine-learning algorithm aims to minimize some kind of estimation error subject to various constraints—which is an optimization problem. Here are the topics to learn:

- Basics of optimization, how to formulate the problem
- Maxima, minima, convex function, global solution
- Linear programming, simplex algorithm
- Integer programming
- Constraint programming, knapsack problem
- Randomized optimization techniques: hill climbing, simulated annealing, genetic algorithms

Where You Might Use It

Simple linear regression problems using least-square loss function often have an exact analytical solution, but logistic regression problems don't. To understand the reason, you need to be familiar with the concept of "convexity" in optimization. This line of investigation will also illuminate why we must remain satisfied with "approximate" solutions in most machine-learning problems.

STATISTICS & INFERENCE

Key Concepts

P-Value

Confidence Interval

Statistical Significance

Practical Significance

Type I Error (False Positive) + Type II Error (False Negative)

Testing Methods

T-test: Compare means between two groups

Z-test: Large sample comparison of proportions

Chi-square test: Test independence/association for categorical data

ANOVA: Compare means across 3+ groups

Bootstrap: Resampling method to estimate sampling distribution without parametric assumptions

Common Pitfalls

Peeking/Early stopping: Checking results repeatedly increases false positive rate

Multiple testing: Testing many hypotheses without correction inflates Type I error

Solution: Bonferroni correction, False Discovery Rate (FDR), sequential testing procedures

Novelty effects: Users behave differently initially then revert to baseline

Primacy effects: Experienced users resist change initially

Selection bias: Non-random assignment or differential attrition between groups

Network effects: Treatment on one user affects others (violation of SUTVA)

Simpson's Paradox: Trend reverses when data is segmented

Causal Inference

Methods

Difference-in-Differences (DiD): Compare trend changes between treatment and control groups pre/post intervention

Regression Discontinuity: Exploit threshold-based treatment assignment (e.g., users above certain score get feature)

Synthetic Control: Create synthetic control group from weighted combination of untreated units

Propensity Score Matching: Match treated/untreated units with similar probability of treatment

Instrumental Variables: Use variable affecting treatment but not outcome directly to isolate causal effect

Interrupted Time Series: Analyze metric before/after intervention accounting for trends

Usage

Geographic rollouts (can't randomize at user level)

Time-based rollouts (everyone gets feature eventually)

Retrospective analysis of past initiatives

Policy changes affecting entire platform

Marketing campaigns with spillover effects

CORE TOPICS

Probability & Distributions

Descriptive Statistics

Inferential Statistics

Regression Analysis

Probability & Distributions (PDS)

BROAD TOPIC

SUB-TOPICS

DETAILED TOPICS

Probability

Fundamentals

- Sample spaces and events

- Probability rules (addition, multiplication)

- Conditional probability

- Bayes’ Theorem

Probability Distributions

- Discrete distributions (Binomial, Poisson)

- Continuous distributions (Normal, Exponential)

- Expected value and variance

- Central Limit Theorem

Descriptive Statistics

Descriptive Statistics

Measures of Central Tendency

- Mean, median, mode
- Weighted averages

Measures of Dispersion

- Variance, standard deviation
- Range, interquartile range
- Skewness and kurtosis

Inferential Statistics

Inferential Statistics

Hypothesis Testing

- Null vs. alternative hypothesis
- P-values and significance levels
- Type I and Type II errors
- T-tests, ANOVA

Confidence Intervals

- Estimation and interpretation
- Margin of error
- Confidence levels (e.g., 95%)

Regression Analysis

Regression Analysis

Linear Regression

- Simple and multiple linear regression
- Least squares method
- R-squared and adjusted R-squared

Logistic Regression

- Binary and multinomial logistic regression
- Odds ratios
- Model evaluation (ROC, AUC)

Regression Modeling w/ Cases

1. Linear Regression

Use: To measure the impact of one or more independent variables (e.g., ad spend, campaign impressions) on a dependent variable (e.g., sales or revenue).

Adobe Example:

Situation: Adobe’s Creative Cloud sales team needed to understand the relationship between marketing spend across channels and overall subscription revenue.

Task: Develop a model to identify which channels were driving the most impact on sales.

Action: Used linear regression to analyze historical spend data across channels, including TV, social media, and search. The model identified key contributors to sales and diminishing returns for overspending on certain channels.

Result: Reallocated 15% of the budget to underperforming but high-potential channels, resulting in a 20% increase in ROI within one quarter.

2. Logistic Regression

Use: To predict binary outcomes (e.g., purchase vs. no purchase, churn vs. retention) based on independent variables.

Credit Karma Example:

Situation: Credit Karma aimed to reduce churn among users who frequently accessed their credit reports but didn't engage with other financial tools.

Task: Predict the likelihood of churn based on user behaviors and engagement metrics.

Action: Built a logistic regression model using features like frequency of logins, product usage patterns, and email engagement rates. Identified high-risk users and implemented targeted outreach campaigns.

Result: Reduced churn by 18% in the identified high-risk cohort, contributing to a 12% improvement in overall retention rates.

3. Ridge and Lasso Regression

Use: To handle multicollinearity in data by regularizing coefficients (Ridge) or performing feature selection (Lasso).

Adobe Example:

Situation: Adobe marketing managers faced multicollinearity in campaign data across overlapping channels (e.g., digital ads and social media).

Task: Build a robust model to identify key performance drivers without overfitting.

Action: Applied Ridge regression to reduce the impact of multicollinearity and Lasso regression to select the most impactful variables.

Result: Improved model accuracy by 25%, enabling the team to confidently reallocate budgets and focus on high-performing channels, driving a 15% sales uplift.

4. Time Series Regression

Use: To forecast outcomes based on temporal data, accounting for trends, seasonality, and cycles.

Credit Karma Example:

Situation: Credit Karma needed to predict daily site traffic during the tax season to optimize marketing efforts.

Task: Build a time series model to forecast traffic and allocate campaign budgets accordingly.

Action: Used an ARIMA model to forecast traffic, accounting for seasonal spikes in tax-related queries. Recommended scaling ad spend during high-traffic days.

Result: Increased site traffic by 25% during peak days while maintaining a cost-per-click 10% below the industry benchmark.

5. Logistic Regression with Interaction Terms

Use: To understand how combinations of variables influence a binary outcome.

Adobe Example:

Situation: Adobe wanted to understand how email engagement and webinar attendance together impacted trial-to-paid conversion rates.

Task: Build a model to identify the combined effects of user engagement behaviors.

Action: Built a logistic regression model with interaction terms for email opens and webinar participation. Insights revealed that users who attended webinars and opened emails were 3x more likely to convert.

Result: Focused resources on email follow-ups for webinar attendees, boosting trial-to-paid conversions by 30%.

6. Multivariate Regression

Use: To assess the impact of multiple dependent variables simultaneously.

Credit Karma Example:

Situation: Credit Karma wanted to evaluate the simultaneous effects of marketing campaigns on user engagement and revenue.

Task: Develop a model to analyze multiple outcomes.

Action: Built a multivariate regression model to measure how variations in campaign spending impacted revenue and engagement metrics like time on site and click-through rates.

Result: Revealed optimal budget thresholds for maximizing both engagement and revenue, increasing average session duration by 20% and revenue by 10%.

7. Stepwise Regression

Use: To identify the most statistically significant predictors in large datasets.

Adobe Example:

Situation: Adobe needed to identify the most critical drivers of customer satisfaction from a survey dataset with over 50 variables.

Task: Develop a model to highlight significant predictors while reducing noise.

Action: Used stepwise regression to iteratively add and remove predictors based on statistical significance. The analysis identified three key drivers of satisfaction: ease of use, feature breadth, and customer support.

Result: Informed product and support improvements, increasing customer satisfaction scores by 15% in six months.

Introduction to Statistics

Statistics For Data Science

11 min read

Descriptive Statistic

5 min read

What is Inferential Statistics?

7 min read

Bayes' Theorem

13 min read

Probability Data Distributions in Data Science

8 min read

Parametric Methods in Statistics

6 min read

Hypothesis Testing

9 min read

ANOVA for Data Science and Data Analytics

9 min read

Bayesian Statistics & Probability

6 min read

/

02Week 2: Maths for Data Analytics

Descriptive Statistics: Mean, Median, Mode, Variance, Standard Deviation

Range, Quartiles, Percentiles

Probability: Basic Concepts & Distributions (Normal, Binomial, Poisson, etc.)

Covariance and Correlation

Hypothesis Testing: CLT, Z-test, T-test, ANOVA, MANOVA

Non-parametric Tests: Mann-Whitney, Kruskal-Wallis

Data Skewness Detection and Handling

Probability and Statistics Concepts to Review for the Data Science Interview

Because probability & statistics are foundational to the field of Data Science, before the interview you should review:

Central Limit Theorem

Probability Distributions

Regression Analysis

Hypothesis Testing

If are unfamiliar with these concepts I recommend reading some of the books from the 13 Best Books for Data Scientists list.

Central Limit Theorem

Understanding the Central Limit Theorem is crucial. It states that the distribution of the sample mean of a large enough sample from any population will be approximately normally distributed, regardless of the population's underlying distribution. This theorem is fundamental when dealing with inferential statistics and hypothesis testing.

Probability Distributions

Hypothesis testing involves formulating null and alternative hypotheses, collecting data, and using statistical methods to determine whether there is enough evidence to reject the null hypothesis. You should be proficient in different types of hypothesis tests (e.g., t-tests, chi-squared tests) and their applications.

Regression Analysis

Familiarity with common probability distributions like the normal distribution, binomial distribution, and Poisson distribution is essential. You should understand their probability density functions, cumulative distribution functions, and how to use them in real-world scenarios.

Hypothesis Testing

Regression analysis is a fundamental statistical technique used for modeling relationships between variables. You should know about linear regression, multiple regression, logistic regression (for classification), and how to interpret regression coefficients, p-values, and R-squared values. Understanding regression allows you to make predictions and draw insights from data.

Beginner Probability and Statistic Questions and Answers asked by FAANG

1. What is the probability of rolling a 6 on a fair six-sided die?

The probability of rolling a 6 on a fair six-sided die is 1/6.

2. Calculate the expected value of a fair coin flip.

The expected value of a fair coin flip is 0.5 (or 1/2).

3. Explain the concept of simple random sampling in statistics.

Simple random sampling is a method where every member of the population has an equal chance of being selected in the sample.

4. Define the Central Limit Theorem and its significance in statistics.

The Central Limit Theorem states that the distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the shape of the original population distribution.

5. What is a p-value, and how is it used in hypothesis testing?

A p-value is a probability measure used in hypothesis testing that quantifies the evidence against a null hypothesis. A smaller p-value suggests stronger evidence against the null hypothesis.

More Questions

Not enough? Try these questions for FREE:

Two Consecutive Sixes

Coin Fairness Test

Medium Probability and Statistic Questions and Answers asked by FAANG

6. Given two events A and B, how do you calculate P(A|B) (the conditional probability of A given B)?

Conditional probability P(A|B) is calculated as the probability of both events A and B occurring (P(A ∩ B)) divided by the probability of event B occurring (P(B)).

7. Explain the Bayesian probability theory and its application in data science.

Bayesian probability is a framework that incorporates prior beliefs and updates them with new evidence using Bayes' theorem, allowing for probabilistic reasoning and decision-making.

8. What is the confidence interval, and how do you interpret a 95% confidence interval?

A 95% confidence interval means that if we were to take many random samples and construct confidence intervals from them, we would expect approximately 95% of those intervals to contain the true population parameter.

9. Describe the sampling distribution of the sample mean.

The sampling distribution of the sample mean is a normal distribution with the same mean as the population and a standard deviation equal to the population standard deviation divided by the square root of the sample size.

10. Calculate the z-score for a data point in a standard normal distribution.

The z-score for a data point in a standard normal distribution is calculated as (X - μ) / σ, where X is the data point, μ is the mean, and σ is the standard deviation.

More Questions

Not enough? Try these questions for FREE:

Consecutive Fives

Biased Coin?

Hard Probability and Statistic Questions and Answers asked by FAANG

11. Compare and contrast the Poisson and Binomial distributions.

The Poisson distribution models the number of events occurring in a fixed interval of time or space, while the Binomial distribution models the number of successes in a fixed number of independent trials.

12. What is the difference between Type I and Type II errors in hypothesis testing?

Type I error occurs when we reject a true null hypothesis, while Type II error occurs when we fail to reject a false null hypothesis.

13. Explain the concept of MLE and provide an example of its application.

Maximum Likelihood Estimation is a method used to estimate the parameters of a statistical model by maximizing the likelihood function. For example, in the case of a normal distribution, MLE estimates the mean and standard deviation.

14. What is covariance, and how does it differ from correlation?

Covariance measures the degree to which two variables change together, while correlation measures the strength and direction of the linear relationship between two variables.

15. Describe stratified sampling and its advantages over simple random sampling.

Stratified sampling involves dividing the population into subgroups or strata and then taking random samples from each stratum. It is advantageous when there is significant variation within strata.

More Questions

Not enough? Try these questions for FREE:

Product vs. Square

Minimum of Two Uniform Variables

Expert Probability and Statistic Questions and Answers asked by FAANG

Questions:

16. How does Monte Carlo simulation work, and what are its applications in data science?

Monte Carlo simulation is a computational technique that uses random sampling to solve complex problems or estimate numerical results. It has applications in finance, engineering, and optimization problems.

17. Define bootstrapping and discuss its use in estimating population parameters.

Bootstrapping is a resampling technique where samples are drawn with replacement from the observed data to estimate population parameters. It is useful when parametric assumptions are uncertain.

18. Explain the principles of Bayesian networks and their role in probabilistic graphical models.

Bayesian networks are graphical models that represent probabilistic relationships among a set of variables. They are used for probabilistic reasoning, decision-making, and risk analysis.

19. What are autoregressive (AR) and moving average (MA) models in time series analysis?

Autoregressive (AR) models describe a time series using its own past values while moving average (MA) models describe a time series using past forecast errors.

20. Discuss the concept of familywise error rate and methods to control it in multiple hypothesis testing scenarios.

Familywise error rate is the probability of making at least one Type I error when conducting multiple hypothesis tests. Methods to control it include Bonferroni correction and false discovery rate control.

S&P Topology

Module 1: Probability Basics

📄 Review of Probability Theory

▶️ Expected Number of Tosses to Get Consecutive Heads (9:31)

▶️ Expectation and Variance of Number of Ads (5:54)

▶️ Combinatorics (7:37)

📄 Permutation and Combination

Module 2: Probability Distribution

📄 Review of Probability Distributions

▶️ Displaying Ads (Binomial Distribution) (9:06)

▶️ Normal Distribution and Normality Tests (10:29)

▶️ The Central Limit Theorem (7:58)

▶️ Describing Distributions of Real Data (9:30)

[Advanced] Probability Distribution

▶️ [Advanced] Coupon Collector Problem (Geometric Distribution) (7:17)

Module 3: Conditional Probability

Interview Q&A

▶️ Bayes' Theorem (8:52)

▶️ Fraudster (6:42)

▶️ Simpson's Paradox (9:41)

▶️ Monty Hall Problem (7:18)

Module 4: Hypothesis Testing Basics

▶️ Hypothesis Testing Terminology (11:17)

▶️ The Hypothesis Testing Process (11:31)

▶️ Types of Errors and Statistical Power (6:55)

▶️ Confidence Intervals (8:42)

▶️ The p-value (5:48)

▶️ [Advanced] Tradeoff between Type I and Type II Error Rates (4:34)

Module 5: Parametric Tests

▶️ Z-test for Means (12:26) [Preview]

▶️ Z-test for Proportions (14:17) [Preview]

▶️ One-sample T-test (11:22) [Preview]

▶️ Two-sample T-test (13:13) [Preview]

▶️ Test if a Coin is Fair Pt 1 (5:52)

▶️ Test if a Coin is Fair Pt 2 (4:53)

▶️ Introduction to Sample Size Estimation (7:19)

📄 A/B Tests: Sample Size Estimation and Power Analysis

[Advanced] Parametric Tests

▶️ [Advanced] Introduction to ANOVA (8:11)

▶️ [Advanced] One-Way and Two-Way ANOVA (6:57)

▶️ [Advanced] Multiple Testing (Multiple Comparisons) (6:58)

▶️ [Advanced] Multiple Testing Correction (8:44)

▶️ [Advanced] Power Analysis (Derivation) (9:34)

Module 6: Product-Specific Questions

▶️ Covariance (5:52)

▶️ Correlation Coefficient (5:17)

▶️ Assumptions of Linear Regression (9:05)

📄 Linear Regression

Module 7: Sampling and Estimation

▶️ Introduction to Sampling (6:26)

▶️ Sampling Without Replacement (7:59)

▶️ Sampling With Replacement (5:46)

▶️ Maximum Likelihood Estimation (MLE) (8:25)

[Advanced] Sampling and Estimation

▶️ [Advanced] Sampling Methods Pt 1 (4:19)

▶️ [Advanced] Sampling Methods Pt 2 (7:46)

▶️ [Advanced] Introduction to Resampling (8:54)

▶️ [Advanced] Bootstrapping Means and Medians (7:08)

Module 8: Nonparametric Tests

📄 Introduction to Nonparametric Tests

▶️ Chi-Squared Tests (8:58)

▶️ How to Select the Right Test Pt 1 (9:15)

▶️ How to Select the Right Test Pt 2 (6:06)

[Advanced] Nonparametric Tests

📄 [Advanced] Goodness-of-fit Tests

▶️ [Advanced] Permutation Tests (9:13)

Introduction to Statistics and A/B Testing

In this lesson, we're going to go over problems you might face in interviews focused on A/B testing and statistics.

0 of 2 Completed

Hypothesis Testing

Hypothesis testing covers the fundamental theory and background behind A/B Testing. In this course we'll cover Z and T test, multiple hypothesis testing, and the different type errors.

0 of 11 Completed

A/B Testing & Experiment Design

Let's start with a general framework for A/B testing. In practice, an A/B testing and experimentation all follow a step by step process of setting metrics and designing experiments.

0 of 10 Completed

Confidence Intervals

Confidence intervals help us deal with this imprecision by giving us a way to talk about a range of values with some certainty where the true value of the statistic is contained in.

0 of 6 Completed

A/B Testing Common Scenarios

The next couple of chapters will cover common scenarios and concepts involved in A/B testing. As A/B testing involves statistical concepts, there may be terms that you need refreshing on.

0 of 9 Completed

A/B Testing Tradeoffs

There are scenarios where A/B testing is not necessarily the best course of action. Often, there are technical, infrastructure, or practical concerns that come up while planning an A/B test.

0 of 6 Completed

Statistics

This is a refresher on some important statistical concepts that will help us with A/B testing and beyond. While by no means a comprehensive guide, this chapter will go over some important basics about statistical testing and probability distributions.

0 of 11 Completed

Data Analytics Fundamentals: Causal Inference

In this course we'll go over the core concepts of causality, significance, and analyzing data. This is meant as a quick refresher and a high level overview of causal inference basics to eventually apply them in data analytics problems.

0 of 9 Completed

Generalized Linear Models and Regression

Regression models are used to predict the value of a dependent variable from one or more independent variables.

Basic Probability

Probability Theory is the branch of mathematics that deals with uncertainty, underpinning all of statistics and machine learning.

0 of 10 Completed

Discrete Distributions

All areas of study in math can roughly be divided into two camps: discrete mathematics and continuous mathematics. Perhaps the best way to describe the difference between the two is to talk about what each of the branches means by "number."

0 of 12 Completed

Continuous Distributions

Continuous probability distribution: A probability distribution in which the random variable X can take on any value (is continuous).

0 of 6 Completed

Multivariate Distributions

Multivariate distributions show comparisons between two or more measurements and the relationships among them. For each univariate distribution with one random variable, there is a more general multivariate distribution.

0 of 9 Completed

Sampling Theorems

Thus far in this course, we have considered random variables under an idealized scenario where we know the distribution of the random variable.

0 of 7 Completed

Probability Questions: Hard

Let's tackle harder probability questions you'd expect to see from quantitative finance or trading interviews.

0 of 5 Completed

S&P Applied

40 Probability & Statistics Data Science Interview Questions Asked By FAANG & Wall Street

Software Engineering Career Advice

We can't lie - Data Science Interviews are TOUGH. Especially tricky - probability and statistics questions asked by top tech companies & hedge funds during the Data Science, Data Analyst, and the Quant Trading Interview process.

That's why we put together 40 real probability & statistics data science interview questions asked by companies like Facebook, Amazon, Two Sigma, & Bloomberg. We have solutions to all 40 problems, and to 161 other data interview problems on SQL, Machine Learning, and Product/Business Sense in our book, Ace The Data Science Interview. You can also practice some of these same exact questions on DataLemur's statistics interview questions section.

DataLemur has hundreds of real Statistics and Probability Interview questions, sourced from real Data Science and Data Analyst interviews at companies like Facebook and Google.

So, without further ado, here are:

the probability & stat concepts to review before your DS interview

20 probability questions asked by top tech-companies & Wall Street

20 statistics questions asked by FANG & Hedge Funds

solutions to 5 of the probability questions

solutions to 5 of the statistics questions

links to more data science interview resources

Probability & Statistics Concepts To Review Before Your Data Science Interview

Because probability & statistics is foundational to the field of Data Science, before the interview you should review:

Probability Basics & Random Variables

Probability Distributions

Hypothesis Testing

Regression Analysis

In case these statistical concepts sound alien to you, check out some of our favorite Statistics Books for Data Analysts to get a gentle refresher.

Probability Basics and Random Variables

The beginnings of probability start with thinking about sample spaces, basic counting and combinatorial principles. Although it is not necessary to know all of the ins-and-outs of combinatorics, it is helpful to understand the basics for simplifying problems. One classic example here is the “stars and bars” counting method.

The other core topic to study is random variables. Knowing concepts related to expectation, variance, covariance, along with the basic probability distributions is crucial.

Probability Distributions

For modeling random variables, knowing the basics of various probability distributions is essential. Understanding both discrete and continuous examples, combined with expectations and variances, is crucial. The most common distributions discussed in interviews are the Uniform and Normal but there are plenty of other well-known distributions for particular use cases (Poisson, Binomial, Geometric).

Most of the time knowing the basics and their applications should suffice. For example, which distribution would flipping a coin be under? What about waiting for an event? It never hurts being able to do the derivations for expectation, variance, or other higher moments.

Hypothesis Testing

Hypothesis testing is the backbone behind statistical inference and can be broken down into a couple of topics. The first is the Central Limit Theorem, which plays an important role in studying large samples of data. Other core elements of hypothesis testing: sampling distributions, p-values, confidence intervals, type I and II errors. Lastly, it is worth looking at various tests involving proportions, and other hypothesis tests.

Most of these concepts play a crucial role in A/B testing, which is a commonly asked topic during interviews at consumer-tech companies like Facebook, Amazon, and Uber. It's useful to not only understand the technical details but also conceptually how A/B testing operates, what the assumptions are, possible pitfalls, and applications to real-life products.

Modeling

Modeling relies on a strong understanding of probability distributions and hypothesis testing. Since it is a broad term, we will refer to modeling as the areas which have a strong statistical intersection with Machine Learning. This includes topics such as: linear regression, maximum likelihood estimation, & bayesian statistics. For interviews focused on modeling and machine learning, knowing these topics is essential.

20 Probability Interview Problems Asked By Top-Tech Companies & Wall Street

[Facebook - Easy] [Coin Fairness Test on DataLemur] There is a fair coin (one side heads, one side tails) and an unfair coin (both sides tails). You pick one at random, flip it 5 times, and observe that it comes up as tails all five times. What is the chance that you are flipping the unfair coin?

[Lyft - Easy] You and your friend are playing a game. The two of you will continue to toss a coin until the sequence HH or TH shows up. If HH shows up first, you win. If TH shows up first, your friend wins. What is the probability of you winning?

[Google - Easy] What is the probability that a seven-game series goes to 7 games?

[Facebook - Easy] Facebook has a content team that labels pieces of content on the platform as spam or not spam. 90% of them are diligent raters and will label 20% of the content as spam and 80% as non-spam. The remaining 10% are non-diligent raters and will label 0% of the content as spam and 100% as non-spam. Assume the pieces of content are labeled independently from one another, for every rater. Given that a rater has labeled 4 pieces of content as good, what is the probability that they are a diligent rater?

[Bloomberg - Easy] Say you draw a circle and choose two chords at random. What is the probability that those chords will intersect?

[Amazon - Easy] 1/1000 people have a particular disease, and there is a test that is 98% correct if you have the disease. If you don’t have the disease, there is a 1% error rate. If someone tests positive, what are the odds they have the disease?

[Facebook - Easy] There are 50 cards of 5 different colors. Each color has cards numbered between 1 to 10. You pick 2 cards at random. What is the probability that they are not of same color and also not of same number?

[Tesla - Easy] A fair six-sided die is rolled twice. What is the probability of getting 1 on the first roll and not getting 6 on the second roll?

[Facebook - Easy] What is the expected number of rolls needed to see all 6 sides of a fair die?

[Microsoft - Easy] Three friends in Seattle each told you it's rainy, and each person has a 1/3 probability of lying. What is the probability that Seattle is rainy? Assume the probability of rain on any given day in Seattle is 0.25.

[Uber - Easy] Say you roll three dice, one by one. What is the probability that you obtain 3 numbers in a strictly increasing order?

[Bloomberg - Medium] Three ants are sitting at the corners of an equilateral triangle. Each ant randomly picks a direction and starts moving along the edge of the triangle. What is the probability that none of the ants collide? Now, what if it is k ants on all k corners of an equilateral polygon?

[Two Sigma - Medium] What is the expected number of coin flips needed to get two consecutive heads?

[Amazon - Medium] How many cards would you expect to draw from a standard deck before seeing the first ace?

[Robinhood - Medium] A and B are playing a game where A has n+1 coins, B has n coins, and they each flip all of their coins. What is the probability that A will have more heads than B?

[Airbnb - Medium] Say you are given an unfair coin, with an unknown bias towards heads or tails. How can you generate fair odds using this coin?

[Quora - Medium] Say you have N i.i.d. draws of a normal distribution with parameters μ and σ. What is the probability that k of those draws are larger than some value Y?

[Spotify - Hard] A fair die is rolled n times. What is the probability that the largest number rolled is r, for each r in 1..6?

[Snapchat - Hard] There are two groups of n users, A and B, and each user in A is friends with those in B and vice versa. Each user in A will randomly choose a user in B as their best friend and each user in B will randomly choose a user in A as their best friend. If two people have chosen each other, they are mutual best friends. What is the probability that there will be no mutual best friendships?

[Tesla - Hard] Suppose there is a new vehicle launch upcoming. Initial data suggests that any given day there is either a malfunction with some part of the vehicle or possibility of a crash, with probability p which then requires a replacement. Additionally, each vehicle that has been around for n days must be replaced. What is the long-term frequency of vehicle replacements?

20 Statistics Problems Asked By FAANG & Hedge Funds

[Facebook - Easy] How would you explain a confidence interval to a non-technical audience?

[Two Sigma - Easy] Say you are running a multiple linear regression and believe there are several predictors that are correlated. How will the results of the regression be affected if they are indeed correlated? How would you deal with this problem?

[Uber - Easy] Describe p-values in layman's terms.

[Facebook - Easy] How would you build and test a metric to compare two user's ranked lists of movie/tv show preferences?

[Microsoft - Easy] Explain the statistical background behind power.

[Twitter - Easy] Describe A/B testing. What are some common pitfalls?

[Google - Medium] How would you derive a confidence interval from a series of coin tosses?

[Stripe - Medium] Say you model the lifetime for a set of customers using an exponential distribution with parameter λ, and you have the lifetime history (in months) of n customers. What is your best guess for λ?

[Lyft - Medium] Derive the mean and variance of the uniform distribution U(a, b).

[Google - Medium] Say we have X ~ Uniform(0, 1) and Y ~ Uniform(0, 1). What is the expected value of the minimum of X and Y?

[Spotify - Medium] You sample from a uniform distribution [0, d] n times. What is your best estimate of d?

[Quora - Medium] You are drawing from a normally distributed random variable X ~ N(0, 1) once a day. What is the approximate expected number of days until you get a value of more than 2?

[Facebook - Medium] Derive the expectation for a geometric distributed random variable.

[Google - Medium] A coin was flipped 1000 times, and 550 times it showed up heads. Do you think the coin is biased? Why or why not?

[Robinhood - Medium] Say you have n integers 1…n and take a random permutation. For any integers i, j let a swap be defined as when the integer i is in the jth position, and vice versa. What is the expected value of the total number of swaps?

[Uber - Hard] What is the difference between MLE and MAP? Describe it mathematically.

[Google - Hard] Say you have two subsets of a dataset for which you know their means and standard deviations. How do you calculate the blended mean and standard deviation of the total dataset? Can you extend it to K subsets?

[Lyft - Hard] How do you randomly sample a point uniformly from a circle with radius 1?

[Two Sigma - Hard] Say you continually sample from some i.i.d. uniformly distributed (0, 1) random variables until the sum of the variables exceeds 1. How many times do you expect to sample?

[Uber - Hard] Given a random Bernoulli trial generator, how do you return a value sampled from a normal distribution