Statistical Interview Questions for Machine Learning Engineers

Basic Concepts in Statistics

1. What is the difference between a population and a sample?
2. What is the difference between inferential and descriptive statistics?
3. What are quantitative and qualitative data?
4. What is the difference between long format and wide format data?
5. What role does probability theory play in machine learning?

Descriptive Statistics

6. What is the meaning of standard deviation?
7. Give an example where the median is a better measure than the mean.
8. What is Bessel’s correction?
9. What are left-skewed distribution and right-skewed distribution?
10. What is an Outlier?
11. Mention methods to screen for outliers in a dataset.
12. What is the meaning of an inliner?
13. How would you define Kurtosis?
14. How do mean, median, and mode behave differently in skewed distributions?
15. What are quantiles? How are they used in box plots?
16. Explain the concept of statistical moments (first, second, third, fourth).

Probability Distributions

17. What do you understand by the term Normal Distribution?
18. What is the assumption of normality?
19. How do you convert a normal distribution to standard normal distribution?
20. What are some of the properties of a normal distribution?
21. What is the Binomial Distribution formula?
22. What are the criteria that Binomial distributions must meet?
23. What is a t-distribution, and when would you use it?
24. When would you use a Poisson distribution instead of a Binomial distribution?
25. What are the key properties of the exponential distribution? How does it relate to the Poisson process?
26. Compare the normal distribution and log-normal distribution. Provide real-world examples of each.
27. What is a power law distribution? How does it differ from exponential decay?
28. When is the geometric distribution applicable? Derive its mean.
29. Describe the hypergeometric distribution and its use cases.
30. What are probability distributions? Provide some examples.
31. Explain the Central Limit Theorem and its assumptions.

Sampling and Data Collection

32. How do you calculate the needed sample size?
33. What are the types of sampling in Statistics?
34. What types of biases can you encounter while sampling?
35. What is Resampling and what are the common methods of resampling?
36. Explain bootstrapping and its applications in statistics.
37. What is stratified sampling? When is it more effective than simple random sampling?
38. How do you calculate the minimum sample size for estimating a population proportion?
39. Describe cluster sampling vs. systematic sampling.

Hypothesis Testing

40. What is hypothesis testing?
41. Explain the null hypothesis and the alternative hypothesis.
42. What is the p-value in hypothesis testing? How do you interpret it?
43. When should you use a t-test vs. a z-test?
44. What is the difference between one-tail and two-tail hypothesis testing?
45. What is the difference between type I vs. type II errors? How does statistical power relate to them?
46. Compare the z-test and t-test. When is the t-test necessary?
47. Explain the chi-square test for goodness-of-fit and independence.
48. What is ANOVA? How does it generalize the t-test?
49. When would you use a non-parametric test like the Mann-Whitney U test?
50. What is the difference between parametric and non-parametric hypothesis tests?
51. What is the Bonferroni correction, and why is it used in multiple comparisons?
52. Describe the Wald-Wolfowitz runs test for randomness.
53. How would you test if a new algorithm is statistically better than an existing one?

Confidence Intervals and Estimation

54. What is the difference between Point Estimate and Confidence Interval Estimate?
55. What is a confidence interval, and how is it used?
56. Mention the relationship between standard error and margin of error.
57. What is the proportion of confidence interval that will not contain the population parameter?
58. Derive the formula for a confidence interval for the population mean (σ known vs. unknown).
59. What is the margin of error, and how does sample size affect it?
60. Define unbiasedness, consistency, and efficiency of estimators.

Correlation and Regression

61. What are correlation and covariance in statistics?
62. How do correlation and causation differ?

63. Define covariance and correlation. Why is correlation bounded between -1 and 1?
64. What is Linear Regression?
65. What are the assumptions required for linear regression?
66. What is multicollinearity and how can it be detected?
67. What is heteroscedasticity and how do you address it?
68. How do you interpret interaction effects in regression analysis?

Model Evaluation and Validation

69. Explain the concepts of overfitting and underfitting.
70. Can you describe the bias-variance tradeoff?
71. What is cross-validation and why is it important?
72. What is the purpose of the ROC curve?
73. Define precision, recall, and F1 score.
74. What is a confusion matrix and how do you use it?
75. How do you evaluate the performance of a regression model?
76. What are R-squared and adjusted R-squared statistics?
77. How is hypothesis testing used in validating machine learning models?

Machine Learning Concepts

78. What is regularization and what are L1 and L2 regularization?
79. What is the difference between parametric and non-parametric models?
80. Explain the differences between supervised and unsupervised learning.
81. How do you perform feature selection in machine learning?
82. What is dimensionality reduction, and why is it used?
83. What is Principal Component Analysis (PCA) and how does it work?
84. How does PCA differ from Linear Discriminant Analysis (LDA)?
85. How do you handle missing data in a dataset?
86. How do you handle imbalanced datasets in machine learning?
87. What is statistical significance and how does it impact feature importance?

Advanced Statistical Methods

88. What is the goal of A/B testing?
89. Explain the concept and process of A/B testing.
90. What do you understand by sensitivity and specificity?
91. What are the differences between frequentist and Bayesian statistics?
92. Explain Bayes' theorem and its application in machine learning.
93. What is a Markov Chain and how is it used in modeling?
94. What is maximum likelihood estimation (MLE)? Provide an example.
95. Explain the difference between confidence intervals and credible intervals (Bayesian vs. Frequentist).
96. What is Simpson's paradox? Provide an example.
97. Explain the concept of p-hacking and why it is a problem.

98. How does the Jensen-Shannon divergence differ from the Kullback-Leibler divergence?
99. Explain the Bias-Variance Tradeoff in the context of statistical estimation.

Problem-Solving and Case Studies

100. A coin is flipped 10 times, resulting in 8 heads. Test if the coin is fair.
101. Two groups have means of 50 and 55 with standard deviations of 5. Is the difference statistically significant (assume n=30)?
102. Calculate the probability of drawing two aces consecutively from a deck without replacement.
103. Interpret a 95% confidence interval of [25, 30] for the mean height of a population.
104. A company runs an A/B test on two website designs. Design A has a conversion rate of 12% with 500 users, and Design B has a conversion rate of 15% with 600 users. Determine if the difference is statistically significant at a 5% significance level.
105. You're analyzing a dataset of customer purchase amounts with a sample mean of $50 and a standard deviation of $10 (n=100). Construct a 95% confidence interval and explain what it implies about the population mean.
106. A machine learning model predicts customer churn with 80% accuracy on a test set of 1,000 customers, where 20% actually churned. Calculate the precision and recall, assuming 160 true positives, and discuss if the model is balanced for this imbalanced dataset.
107. In a clinical trial, a drug reduces symptom duration from 10 days (control group, n=50, std=2) to 8 days (treatment group, n=50, std=3). Test if the reduction is statistically significant, and discuss the implications for model validation.
108. A factory's machine produces parts with lengths following a normal distribution (mean=10 cm, std=0.5 cm). If 5% of parts are rejected for being too short, calculate the rejection threshold and estimate the proportion rejected if the mean shifts to 9.8 cm.
109.

Updated Probability Distributions Category (25 Questions)

17. **What do you understand by the term Normal Distribution?**
18. **What is the assumption of normality?**
19. **How do you convert a normal distribution to standard normal distribution?**
20. **What are some of the properties of a normal distribution?**
21. **What is the Binomial Distribution formula?**
22. **What are the criteria that Binomial distributions must meet?**
23. **What is a t-distribution, and when would you use it?**
24. **When would you use a Poisson distribution instead of a Binomial distribution?**
25. **What are the key properties of the exponential distribution? How does it relate to the Poisson process?**
26. **Compare the normal distribution and log-normal distribution. Provide real-world examples of each.**
27. **What is a power law distribution? How does it differ from exponential decay?**
28. **When is the geometric distribution applicable? Derive its mean.**

29. **Describe the hypergeometric distribution and its use cases.**
30. **What are probability distributions? Provide some examples.**
31. **Explain the Central Limit Theorem and its assumptions.**
32. **What is the Uniform Distribution, and how is it used in machine learning?**
33. **Explain the Beta Distribution and its significance in Bayesian inference.**
34. **What is the Exponential Distribution, and how does it relate to time-to-event modeling in machine learning?**
35. **Describe the Gamma Distribution and its applications in machine learning.**
36. **What is the Bernoulli Distribution, and how does it differ from the Binomial Distribution?**
37. **Explain the Multinomial Distribution and its use in natural language processing.**
38. **What is the Chi-Square Distribution, and how is it applied in hypothesis testing?**
39. **Describe the Weibull Distribution and its role in reliability analysis or failure prediction.**
40. **What is the Cauchy Distribution, and why is it considered a ‘pathological’ distribution in statistics?**
41. **Explain the Dirichlet Distribution and its importance in topic modeling (e.g., LDA).**