CAREERS (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/CAREERS/)

40 Interview Questions On Statistics For Data Scientists

BY ROHIT GARG (HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/F2005636GMAIL-COM/)
02/02/2020

We frequently come out with resources for aspirants and job seekers in data science to help them make a career in this vibrant field. Cracking interviews especially where understating of statistics is needed can be tricky. Here are 40 most commonly asked interview questions for data scientists, broken into basic and advanced.

Here are some other interview questions resources for data scientists.

10 Most Common SQL Questions & Answers You Must Know For Your Next Interview (https://analyticsindiamag.com/10-most-common-sql-questions-answers-you-must-know-for-your-next-interview/)

---

10 Frequently Asked Interview Questions For Machine Learning In 2019 (https://analyticsindiamag.com/10-frequently-asked-interview-questions-for-machine-learning-in-2019/)

5 Mathematical Concepts Every Data Scientist Should Master Before An Interview (https://analyticsindiamag.com/5-mathematical-concepts-every-data-scientist-should-master-before-an-interview/)

10 Important Pandas Interview Questions Every Beginner Must Know (https://analyticsindiamag.com/10-important-pandas-interview-questions-every-beginner-must-know/)

11 Most Commonly Asked NLP Interview Questions For Beginners (https://analyticsindiamag.com/11-most-commonly-asked-nlp-interview-questions-for-beginners/)

12 Most Popular Python Interview Questions You Must Prepare For (https://analyticsindiamag.com/12-most-popular-python-interview-questions-you-must-prepare-for/)

10 Most Frequently Asked Questions In Data Science Interview (https://analyticsindiamag.com/10-most-frequently-asked-questions-in-data-science-interview/)

Top Interview Questions For A Data Engineer Job Profile (https://analyticsindiamag.com/top-interview-questions-for-a-data-engineer-job-profile/)

Part 1 – Basic Statistics and Distributions

20 Question

1. **What is the difference between data analysis and machine learning?**

Data analysis requires strong knowledge of coding and basic knowledge of statistics

Machine learning, on the other hand, requires basic knowledge of coding and strong knowledge of statistics and business.

---

2. **What is big data?**

Big data has 3 major components – volume (size of data), velocity (inflow of data) and variety (types of data)

Big data causes "overloads"

---

3. **What are the four main things we should know before studying data analysis?**

Descriptive statistics

Inferential statistics

Distributions (normal distribution / sampling distribution)

Hypothesis testing

---

4. **What is the difference between inferential statistics and descriptive statistics?**

Descriptive statistics – provides exact and accurate information.

Inferential statistics – provides information of a sample and we need to inferential statistics to reach to a conclusion about the population.

---

5. **What is the difference between population and sample in inferential statistics?**

From the population we take a sample. We cannot work on the population either due to computational costs or due to availability of all data points for the population.

From the sample we calculate the statistics

From the sample statistics we conclude about the population

---

6. **What are descriptive statistics?**

Descriptive statistic is used to describe the data (data properties)

5-number summary is the most commonly used descriptive statistics

---

7. **Most common characteristics used in descriptive statistics?**

- Center – middle of the data. Mean / Median / Mode are the most commonly used as measures.
  - Mean – average of all the numbers
  - Median – the number in the middle
  - Mode – the number that occurs the most. The disadvantage of using Mode is that there may be more than one mode.
- Spread – How the data is dispersed. Range / IQR / Standard Deviation / Variance are the most commonly used as measures.
  - Range = Max – Min
  - Inter Quartile Range (IQR) = Q3 – Q1
  - Standard Deviation (σ) = $\sqrt{(\Sigma(x-\mu)^2 / n)}$
  - Variance = σ^2
- Shape – the shape of the data can be symmetric or skewed
  - Symmetric – the part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median
  - Left skewed – the left tail is longer than the right side
  - Right skewed – the right tail is longer than the left side
- Outlier – An outlier is an abnormal value
  - Keep the outlier based on judgement
  - Remove the outlier based on judgement

---

8. **What is quantitative data and qualitative data?**

Quantitative data is also known as numeric data

Qualitative data is also known as categorical data

---

9. **How to calculate range and interquartile range?**

IQR = Q3 – Q1

Where, Q3 is the third quartile (75 percentile)

Where, Q1 is the first quartile (25 percentile)

---

10. **Why we need 5-number summary?**

Low extreme (minimum)

Lower quartile (Q1)

Median

Upper quartile (Q3)

Upper extreme (maximum)

---

11. **What is the benefit of using box plot?**

Shows the 5-number summary pictorially

Can be used to compare group of histograms

---

12. **What is the meaning of standard deviation?**

It represents how far are the data points from the mean

$(\sigma) = \sqrt{(\Sigma(x-\mu)^2 / n)}$

Variance is the square of standard deviation

---

13. **What is left skewed distribution and right skewed distribution?**

- Left skewed
  - The left tail is longer than the right side
  - Mean < median < mode
- Right skewed
  - The right tail is longer than the right side
  - Mode < median < mean

---

14. **What does symmetric distribution mean?**

The part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

Few examples are – uniform distribution, binomial distribution, normal distribution

---

15. **What is the relationship between mean and median in normal distribution?**

In the normal distribution mean is equal to median

---

16. **What does it mean by bell curve distribution and Gaussian distribution?**

Normal distribution is called bell curve distribution / Gaussian distribution

It is called bell curve because it has the shape of a bell

It is called Gaussian distribution as it is named after Carl Gauss

---

17. How to convert normal distribution to standard normal distribution?

Standardized normal distribution has mean = 0 and standard deviation = 1

To convert normal distribution to standard normal distribution we can use the formula

$X \text{ (standardized)} = (x-\mu) / \sigma$

---

18. **What is an outlier?**

An outlier is an abnormal value (It is at an abnormal distance from rest of the data points).

---

19. **Mention one method to find outliers?**

Shows the 5-number summary can be used to identify the outlier

Widely used – Any data point that lies outside the 1.5 * IQR

Lower bound = Q1 – (1.5 * IQR)

Upper bound = Q3 + (1.5 * IQR)

---

20. **What can I do with outlier?**

- Remove outlier
  - When we know the data-point is wrong (negative age of a person)
  - When we have lots of data
  - We should provide two analyses. One with outliers and another without outliers.
- Keep outlier
  - When there are lot of outliers (skewed data)
  - When results are critical
  - When outliers have meaning (fraud data)

Part 2 – Advance Statistics and Hypothesis Testing

20 Question

21. **What is the difference between population parameters and sample statistics?**

- Population parameters are:
  - Mean = μ
  - Standard deviation = σ
- Sample statistics are:
  - Mean = x (bar)
  - Standard deviation = s

---

22. **Why we need sample statistics?**

Population parameters are usually unknown hence we need sample statistics.

---

23. **How to find the mean length of all fishes in the sea?**

Define the confidence level (most common is 95%)

Take a sample of fishes from the sea (to get better results the number of fishes > 30)

Calculate the mean length and standard deviation of the lengths

Calculate t-statistics

Get the confidence interval in which the mean length of all the fishes should be.

---

24. **What are the effects of the width of confidence interval?**

- Confidence interval is used for decision making
- As the confidence level increases the width of the confidence interval also increases
- As the width of the confidence interval increases, we tend to get useless information also.
  - Useless information – wide CI
  - High risk – narrow CI

25. **Mention the relationship between standard error and margin of error?**

As the standard error increases the margin of error also increases

---

26. **Mention the relationship between confidence interval and margin of error?**

SEE ALSO


(https://analyticsindiamag.com/how-to-work-as-a-freelancer-in-the-field-of-data-science/)

How To Work As A Freelancer In Data Science (https://analyticsindiamag.com/how-to-work-as-a-freelancer-in-the-field-of-data-science/)

As the confidence level increases the margin of error also increases

---

27. **What is the proportion of confidence interval that will not contain the population parameter?**

Alpha is the portion of confidence interval that will not contain the population parameter

$\alpha = 1 - CL$

---

28. **What is the difference between 95% confidence level and 99% confidence level?**

The confidence interval increases as me move from 95% confidence level to 99% confidence level

---

29. **What do you mean by degree of freedom?**

DF is defined as the number of options we have

DF is used with t-distribution and not with Z-distribution

For a series, DF = n-1 (where n is the number of observations in the series)

---

30. **What do you think if DF is more than 30?**

As DF increases the t-distribution reaches closer to the normal distribution

At low DF, we have fat tails

If DF > 30, then t-distribution is as good as normal distribution

---

31. **When to use t distribution and when to use z distribution?**

- The following conditions must be satisfied to use Z-distribution
  - Do we know the population standard deviation?
  - Is the sample size > 30?
  - CI = x (bar) $- Z^*\sigma/\sqrt{n}$ to x (bar) $+ Z^*\sigma/\sqrt{n}$
- Else we should use t-distribution
  - CI = x (bar) $- t^*s/\sqrt{n}$ to x (bar) $+ t^*s/\sqrt{n}$

---

32. **What is H0 and H1? What is H0 and H1 for two-tail test?**

- H0 is known as null hypothesis. It is the normal case / default case.
  - For one tail test x <= μ
  - For two-tail test x = μ
- H1 is known as alternate hypothesis. It is the other case.
  - For one tail test x > μ
  - For two-tail test x <> μ

---

33. **What is p-value in hypothesis testing?**

- If the p-value is more than then critical value, then we fail to reject the H0
  - If p-value = 0.015 (critical value = 0.05) – strong evidence
  - If p-value = 0.055 (critical value = 0.05) – weak evidence
- If the p-value is less than the critical value, then we reject the H0
  - If p-value = 0.055 (critical value = 0.05) – weak evidence
  - If p-value = 0.005 (critical value = 0.05) – strong evidence

---

34. **How to calculate p-value using manual method?**

Find H0 and H1

Find n, x(bar) and s

Find DF for t-distribution

Find the type of distribution – t or z distribution

Find t or z value (using the look-up table)

Compute the p-value to critical value

---

35. **How to calculate p-value using EXCEL?**

Go to Data tab

Click on Data Analysis

Select Descriptive Statistics

Choose the column

Select summary statistics and confidence level (0.95)

---

36. **What do we mean by – making decision based on comparing p-value with significance level?**

If the p-value is more than then critical value, then we fail to reject the H0

If the p-value is less than the critical value, then we reject the H0

---

37. What is the difference between one tail and two tail hypothesis testing?

- 2-tail test: Critical region is on both sides of the distribution
  - H0: x = μ
  - H1: x <> μ
- 1-tail test: Critical region is on one side of the distribution
  - H1: x <= μ
  - H1: x > μ

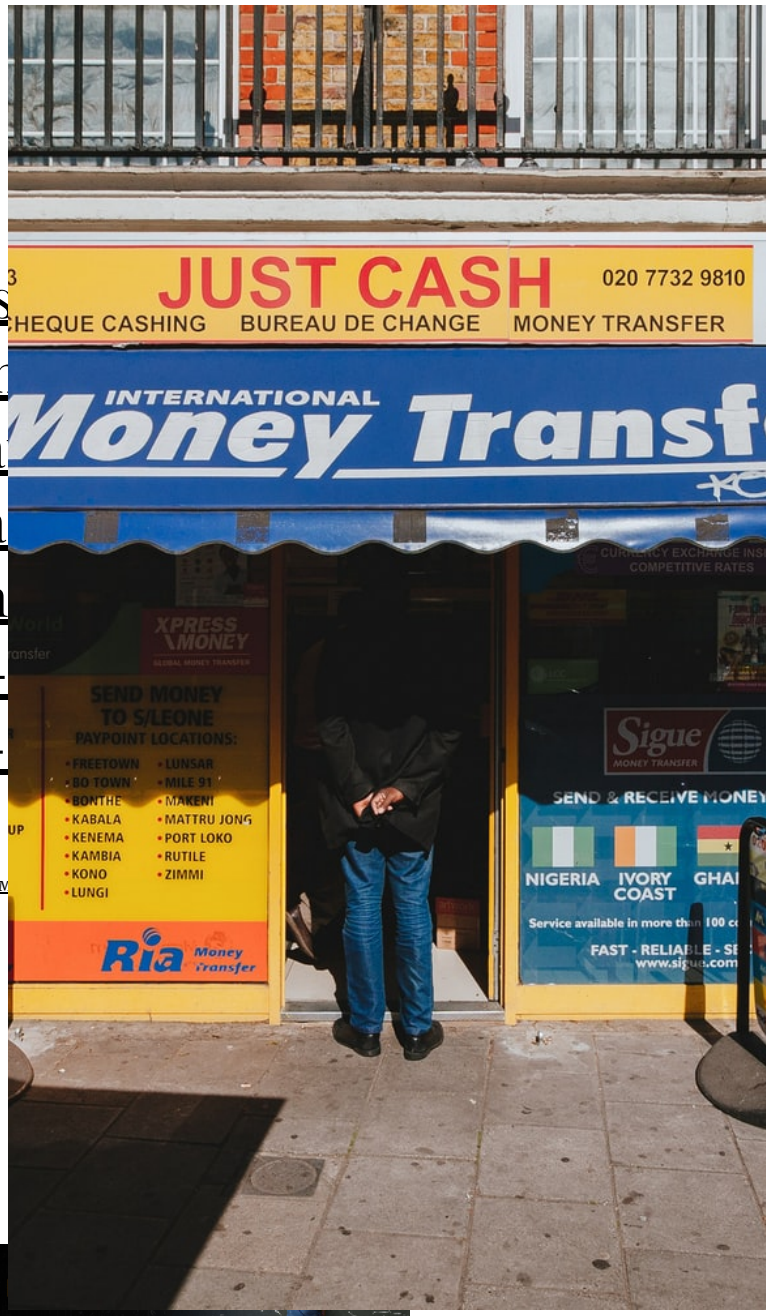
---

38. What do you think of the tail (one tail or two tail) if H0 is equal to one value only?

It is a two-tail test

OPINIONS (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/ARTICLES/)

Case Study: How This Forex ncumbent Improved Their Cros order Payment Process With Th lp Of An AI & ML Data Aggrega tps://analyticsindiamag.com/ca udy-how-this-forex-incumben improved-their-cross-border- yment-process-with-the-help- an-ai-ml-data-aggregator/)



BY SEJUTI DAS (HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/SEJUTI-DASANALYTICSINDIAMAG-COM
02/02/2020

Money transfer companies make a considerable amount of their revenue from the fees they levy on each transaction, and therefore factors like transfer rate, fees associated, and transfer speed plays a vital role for a company to thrive in this competitive world of money transfer. With banks offering attractive prices, money transfer companies started to struggle to stay on top of their game. In a bid to stay relevant, one such veteran forex company switched to artificial intelligence-based solutions for critical data in order to tackle the changing forex landscape.

The Challenge

With the rise in advanced technologies in other parts of our lives, tech-savvy customers expect BFSI companies to deliver seamless financial experience. And that's why BFSI companies are equipping themselves with tools that are required to stay ahead of competitors.

A Colorado-based cross border, cross-currency, money movement payment company (wishes to remain anonymous) that allows transactions, online as well as offline, have been struggling a long time in streamlining their business process in order to deliver enhanced customer service.

39. **What is the critical value in one tail or two-tail test?**

Critical value in 1-tail = alpha

Critical value in 2-tail = alpha / 2

---

40. **Why is the t-value same for 90% two tail and 95% one tail test?**

P-value of 1-tail = P-value of 2-tail / 2

It is because in two tail there are 2 critical regions

If you loved this story, do join our Telegram Community (https://t.me/joinchat/NJLxnhZB7GkX3CPvjs9QGQ).

Also, you can write for us and be one of the 500+ experts who have contributed stories at AIM. Share your nominations here (https://analyticsindiamag.com/write-for-us/).

ROHIT GARG (HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/F2005636GMAIL-COM/)

Rohit Garg has close to 7 years of work experience in field of data analytics and machine learning. He has worked extensively in the areas of predictive modeling, time series analysis and segmentation techniques. Rohit holds BE from BITS Pilani and PGDM from IIM Raipur.

Considering the massive scale of operation, with five lakh agent locations in over 200 countries and territories, and a plethora of online payment options, the customer had to deal with numerous challenges, such as volatile market rates, competitors trying to undercut them, and a changing marketplace. An incumbent in the money transfer space, the company had completed more than 800 million transactions in 2018 and has moved over $300 billion in principal amount.

To tackle these problems efficiently, the customer required a seamless framework that can provide detailed information on the competitors' positioning and their pricing strategies, along with their transfer fees, transfer speed, and the exchange rate of the competitors per transaction.

The Solution

The answer to the company's problem was data. And, therefore, the customer demanded an AI-based framework that could obtain the high quality, hard to get, competitor intelligence, which Bridged agreed on to provide.

Bridged is a company that provides hard to obtain human-powered data for artificial intelligence (AI) models at a scale that can create a competitive advantage and grow revenue for its customers. The company used a combination of artificial intelligence technologies and a 13,000 strong, highly skilled workforce to develop unique and vast data at a scale, significantly improving the quality of data models. Whether it be training data for machine learning models or competitive intelligence, they drew a "bridge" between the problem and solution for the customer.

The scalable solution included retail intelligence, content generation, content categorisation, and video and image tagging for companies across the world. The massive workforce captured over a million data points from top-tier competitors daily, which they converted into statistical information and leveraged it to provide insights into new market opportunities, competitor behaviour, and price optimisation, amongst many others.

Bridged has serviced numerous AI/machine learning companies with their data requirements and therefore, has become the perfect choice for the forex exchange company to deal with their challenges. With the help of Bridged's expertise, the customer was able to receive agent store location data of its competitors for 50+ countries; and pricing data such as transfer speed, transfer rate, and the exchange rate of 13 competitors spanning 1300+ countries and 650+ currency pairs in the online and offline space.

With the combination of technology and crowdsourcing Bridged delivered artificial intelligence as a service, providing scalable data solution for the customer. Along with a scaled process to capture real-time data on multiple competitors, Bridged also designed a fully managed process to cover multiple sources, and utilise multiple APIs, to ensure the data shared was accurate and complete. Additionally, the customer used the data provided, giving it the ability to move to a dynamic pricing structure and thus winning business.

Benefits

Being a brand that services artificial intelligence and machine learning companies with their data requirements, Bridged understood the company's requirement for high-quality data. And therefore, it delivered 150k+ data rows per day, thus giving the customer precise insights for the company to

adjust its offerings and prices. With the help of the structured data, it adjusted its offering and pricing parameters, which, in turn, increased their earnings from each transaction. The customer also enjoyed the luxury of obtaining data that they've struggled to find and utilise in the past.

By collaborating with Bridged, the customer was able to gain a strong data-driven understanding of the competitive landscape across geographies. Alongside, they had the privilege of receiving critical data on store locations, for its offline market, that wasn't accessible before. This information also helped the customer to analyse into which geographical locations they can maximise their bottom-line.

In terms of return of investment, working with Bridged, helped the customer to reduce its expenses by 70% and time to access this data by 90%.

Future Prospects

The customer aims to expand its collaboration with Bridged and the scope of its requirements with a focus on capturing the location data for more countries, and the pricing data across multiple channels, such as mobile apps, etc. Besides, Bridged is looking to use its crowd for auditing the retail locations to ensure better customer service.

Presently, they're working on an FX product — Smart Pricing System, that has already captured 25M+ data rows from leading money transfer companies. It captures data points such as transfer fees, transfer speed, and exchange rate for various send amounts and currency corridors. The product has been designed to help money transfer companies to locate new market opportunities, track their competition, increase their revenue, price their transactions strategically, and analyse competitors' reactions to a given change in price.

SEJUTI DAS (HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/SEJUTI-DASANALYTICSINDIAMAG-COM/)

Sejuti currently works as Senior Technology Journalist at Analytics India Magazine (AIM). Reach out at sejuti.das@analyticsindiamag.com

RELATED POSTS

Hands-On Guide To Pandas Visual Analysis – Way To Speed-Up Data Visualization (https://analyticsindiamag.com/hands-on-guide-to-pandas-visual-analysis-way-to-speed-up-data-visualization/)

(https://analytics
on-guide-to-
pandas-visual-
analysis-way-
to-speed-up-
data-
visualization/)

25/09/2020 · 4 MINS READ

Complete Guide To Visualizer: Python Library for Automating Visualization (https://analyticsindiamag.com/complete-guide-to-visualizer-python-library-for-automating-visualization/)

(https://analytics
guide-to-
visualizer-
python-library-
for-automating-
visualization/)

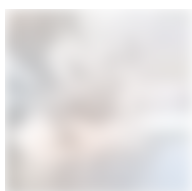
24/09/2020 · 3 MINS READ

50 Latest Data Science And Analytics Jobs From Past Week (https://analyticsindiamag.com/50-latest-data-science-and-analytics-jobs-from-past-week/)

(https://analytics
latest-data-
science-and-
analytics-jobs-
from-past-
week/)

24/09/2020 · 6 MINS READ

Why Are Analytics Interviews Flawed (https://analyticsindiamag.com/why-are-analytics-interviews-flawed/)



(https://analytics
are-analytics-
interviews-

flawed/)

22/09/2020 · 5 MINS READ

Hands-On Tutorial On Lens: Python Tool For Swift Statistical Analysis (https://analyticsindiamag.com/hands-on-tutorial-on-lens-python-tool-for-swift-statistical-analysis/)

(https://analytics
on-tutorial-on-
lens-python-
tool-for-swift-
statistical-
analysis/)

17/09/2020 · 4 MINS READ

Is More Data Always Better For Building Analytics Models? (https://analyticsindiamag.com/is-more-data-always-better-for-building-analytics-models/)

(https://analytics
more-data-
always-better-
for-building-
analytics-
models/)

More than 1,00,000 people are subscribed to our newsletter

Subscribe now to receive in-depth stories on AI & Machine Learning.

ENTER YOUR EMAIL HERE... YES! SUBSCRIBE ME