DATA SCIENCE INTERVIEW

100+ Core Concepts

Interview-এর আগে যে Concepts পরিষ্কার থাকতেই হবে

Statistics, Probability, Machine Learning, Python, SQL ও Production ML—সহজ বাংলায়

লেখক

Md. Aminul Haque Palash

সূচিপত্র

ভূমিকা 2
Chapter 1: Basic Statistics 3
1. Mean, Median এবং Mode 3
2. Variance এবং Standard Deviation 4
3. Covariance এবং Correlation 6
Covariance 6
Correlation 6
4. Population এবং Sample 7
Population 7
Sample 7

ভূমিকা

Data Science interview-এর প্রস্তুতি নিতে গিয়ে সবচেয়ে পরিচিত সমস্যাটি হলো—কোথা থেকে শুরু করব, আর কতটুকু পড়ব?
একদিকে Statistics ও Probability, অন্যদিকে Machine Learning, Python, SQL এবং Production ML। Topic-এর তালিকা যত বড় হয়, preparation তত সহজে এলোমেলো হয়ে যায়। তখন অনেকেই definition ও formula মুখস্থ করেন; কিন্তু interviewer প্রশ্নটি একটু ঘুরিয়ে করলেই উত্তর আটকে যায়।
এই বইটি সেই সমস্যা কমানোর জন্য লেখা। এখানে ১৩২টি গুরুত্বপূর্ণ concept এমনভাবে সাজানো হয়েছে, যাতে আপনি শুধু কী—তা নয়, কেন, কখন এবং কীভাবে—এই তিনটি দিকও বুঝতে পারেন।
প্রতিটি concept শেখার flow হবে:
Concept → সহজ Intuition → Practical Example → প্রয়োজনীয় Formula → ব্যবহার ও সীমাবদ্ধতা → Interview-ready Explanation
সব formula সমান গুরুত্বপূর্ণ নয়, আবার শুধু intuition জানাও যথেষ্ট নয়। Interview-এ ভালো উত্তর দিতে হলে সহজ ভাষায় মূল idea বোঝাতে হবে, একটি relevant example দিতে হবে এবং প্রয়োজনে trade-off বা limitation উল্লেখ করতে হবে।
এই বইটি বিশেষভাবে কাজে লাগবে যদি আপনি:

- Data Science বা Machine Learning interview-এর প্রস্তুতি নিচ্ছেন;
- পরিচিত concepts দ্রুত revise করতে চান;
- formula মুখস্থ না করে intuition তৈরি করতে চান;
- নিজের ভাষায় সংক্ষিপ্ত ও পরিষ্কার answer দিতে শিখতে চান।

একটি কথা মনে রাখুন:

> Interviewer সাধারণত শুধু formula মনে আছে কি না, তা যাচাই করেন না। তিনি দেখতে চান আপনি concept-টি বুঝেছেন কি না এবং বাস্তব problem-এ সেটি প্রয়োগ করতে পারবেন কি না।

Chapter 1: Basic Statistics

1. Mean, Median এবং Mode

ধরুন পাঁচজন মানুষের মাসিক income:

```
20,000
25,000
30,000
35,000
2,00,000
```

এখন আমরা জানতে চাই, এই group-এর একটা "typical income" কত।

এখানেই আসে Mean, Median এবং Mode।

Mean কী?

সব value যোগ করে total observation দিয়ে ভাগ।

$$Mean = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

উপরের example-এ:

$$Mean = \frac{20000 + 25000 + 30000 + 35000 + 200000}{5}$$

$$= 62,000$$

কিন্তু একটা সমস্যা দেখুন।

চারজনের income 20–35 হাজারের মধ্যে, অথচ mean বলছে 62 হাজার।

কারণ 2 লাখের value-টা mean-কে অনেক উপরে টেনে নিয়ে গেছে।

এই extreme value-কে আমরা outlier বলতে পারি।

Median কী?

সব value ছোট থেকে বড় সাজিয়ে মাঝের value।

```
20k, 25k, 30k, 35k, 200k
```

Median:

```
30k
```

এখানে 30k group-টাকে 62k-এর চেয়ে অনেক ভালো represent করছে।

Mode কী?

যে value সবচেয়ে বেশি বার আসে।

```
10, 20, 20, 20, 30, 40
```

Mode:

```
20
```

Categorical data-তেও mode useful।

যেমন:

```
Preferred Payment Method:

Cash
Cash
Card
Cash
bKash
```

Mode:

```
Cash
```

কখন কোনটা ব্যবহার করব?

Symmetric data এবং outlier কম হলে:

Mean

Skewed data বা strong outlier থাকলে:

Median

Most common value জানতে:

Mode

Interviewer জিজ্ঞেস করতে পারে

Why would you use median instead of mean?

আপনি বলতে পারেন:

> যদি data highly skewed হয় অথবা extreme outlier থাকে, mean খুব সহজে affected হয়। Median শুধু ordering-এর উপর depend করে, তাই সেই situation-এ median central tendency-এর better representation হতে পারে।

Salary, house price, delivery time-এর মতো data-তে এই situation অনেক দেখা যায়।

2. Variance এবং Standard Deviation

Mean আমাদের বলে data-এর center কোথায়।

কিন্তু শুধু center জানলে পুরো picture পাওয়া যায় না।

ধরুন:

```
Dataset A:
48, 49, 50, 51, 52
```

এবং:

```
Dataset B:
10, 30, 50, 70, 90
```

দুইটার mean:

```
50
```

কিন্তু clearly দুই dataset একই না।

Dataset A খুব compact।

Dataset B অনেক spread out।

এই spread measure করার জন্য Variance এবং Standard Deviation use করি।

Variance

প্রথমে প্রত্যেক observation mean থেকে কত দূরে সেটা দেখি।

তারপর distance square করি।

$$Variance = \frac{\sum (x_i - \mu)^2}{N}$$

Population-এর জন্য denominator:

$$N$$

Sample-এর ক্ষেত্রে:

$$s^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n - 1}$$

কেন square করি?

ধরুন mean = 50।

একটা value:

```
40 → difference = -10
```

আরেকটা:

```
60 → difference = +10
```

সরাসরি যোগ করলে:

```
-10 + 10 = 0
```

Spread যেন নেই!

Square করলে:

```
100 + 100
```

এ সমস্যা থাকে না।

Standard Deviation

Variance-এর square root।

$$SD = \sqrt{Variance}$$

Standard deviation বেশি intuitive কারণ original data-এর একই unit-এ থাকে।

ধরুন:

```
Delivery time → minutes
```

Variance হবে:

```
minutes²
```

কিন্তু standard deviation:

```
minutes
```

Interview explanation

> Variance এবং standard deviation দুটোই data কতটা spread out সেটা measure করে। Variance squared unit-এ থাকে, আর standard deviation original unit-এ থাকে বলে interpret করা সহজ।

---

3. Covariance এবং Correlation

ধরুন আমরা analyse করছি:

```
Delivery Distance
Delivery Time
```

Normally distance বাড়লে delivery time-ও বাড়বে।
অর্থাৎ দুই variable একসাথে move করছে।
এটাই covariance এবং correlation দিয়ে measure করা যায়।

Covariance

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

সহজভাবে:

Positive covariance

X বাড়লে Y সাধারণত বাড়ে।

```
Distance ↑
Delivery Time ↑
```

Negative covariance

X বাড়লে Y সাধারণত কমে।

```
Product Price ↑
Demand ↓
```

Near-zero covariance

Clear linear movement নেই।

Covariance-এর সমস্যা

এর fixed scale নেই।

ধরুন:

```
Covariance = 25
```

এটা strong নাকি weak relationship?
শুধু এই number দেখে বোঝা কঠিন।
এজন্য correlation বেশি useful।

Correlation

Correlation হচ্ছে normalized covariance।

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

এর range:

$$-1 \leq r \leq 1$$

Interpretation

```
r = +1 → Perfect positive linear relationship
r = -1 → Perfect negative linear relationship
r ≈ 0  → Little/no linear relationship
```

খুব গুরুত্বপূর্ণ
Correlation does not imply causation.
ধরুন summer-এ:

```
Ice cream sales ↑
Swimming accidents ↑
```

তাহলে কি ice cream swimming accident cause করছে?
না।
দুইটার common কারণ:

```
Temperature / Summer
```

এটাকে confounding variable বলা যায়।

Interview answer

> Covariance direction বলে কিন্তু scale dependent। Correlation covariance-কে standardize করে −1 থেকে +1 range-এ নিয়ে আসে। তবে correlation কোনো causal relationship prove করে না।

---

4. Population এবং Sample

ধরুন FOODI-এর 50 লাখ user আছে।
আপনি জানতে চান:

> Average user প্রতি মাসে কয়টি order করে?

সব 50 লাখ user analyse করাই হচ্ছে population analysis।

Population

যে পুরো group সম্পর্কে আমরা conclusion করতে চাই।
উদাহরণ:

```
Bangladesh-এর সব university student
FOODI-এর সব users
একটা factory-এর সব manufactured products
```

Sample

Population-এর ছোট subset।
যেমন:

```
Randomly selected 20,000 FOODI users
```

এদের behaviour দেখে পুরো population সম্পর্কে estimate করতে পারি।

Population Parameter

Population-এর true value।
যেমন:

$$\mu$$

population mean।

Sample Statistic

Sample থেকে calculated value।
যেমন:

$$\bar{x}$$

sample mean।

আমরা sample statistic দিয়ে population parameter estimate করি।

Sample কেন দরকার?

কারণ পুরো population analyse করা হতে পারে:

- expensive
- slow
- technically difficult
- sometimes impossible

সবচেয়ে গুরুত্বপূর্ণ বিষয়

Sample ideally population-এর representative হতে হবে।

শুধু Dhaka-এর users নিয়ে Bangladesh-wide behaviour predict করলে bias হতে পারে।

ABOUT THE AUTHOR

লেখক পরিচিতি

Md. Aminul Haque Palash

Software Engineer • Machine Learning Practitioner

Md. Aminul Haque Palash একজন Software Engineer এবং Machine Learning practitioner। তিনি Chittagong University of Engineering and Technology (CUET) থেকে Computer Science and Engineering (CSE) বিষয়ে B.Sc. ডিগ্রি অর্জন করেছেন এবং software engineering ও machine learning-এর সমন্বয়ে বাস্তবধর্মী AI solution তৈরিতে কাজ করে আসছেন।

তাঁর কাজের অন্যতম প্রধান আগ্রহ হলো Machine Learning model-কে বাস্তব production environment-এ কার্যকরভাবে ব্যবহারের উপযোগী করে তোলা। তাঁর দৃষ্টিতে একটি ML system-এর সাফল্য শুধু model accuracy দিয়ে নির্ধারিত হয় না; এর সঙ্গে scalability, latency, data quality, reliability, monitoring এবং বাস্তব business requirements-ও সমান গুরুত্বপূর্ণ।

বাংলাভাষী engineers ও শিক্ষার্থীদের জন্য **ML System Design** এবং **Production Engineering** নিয়ে সহজ, structured এবং interview-oriented learning resource তুলনামূলকভাবে সীমিত। এই প্রয়োজন থেকেই বইটি লেখার উদ্যোগ। বইটির লক্ষ্য শুধু বিভিন্ন technical term বা framework মুখস্থ করানো নয়; বরং প্রতিটি concept কেন প্রয়োজন, এর পেছনের engineering trade-off কী এবং একটি real-world production system-এ এগুলো কীভাবে একসঙ্গে কাজ করে—তা সহজ ও practical উপায়ে তুলে ধরা।

এই বইয়ের মাধ্যমে পাঠক Machine Learning system-কে শুধু একটি model হিসেবে নয়, বরং **data থেকে deployment, monitoring এবং continuous improvement পর্যন্ত একটি সম্পূর্ণ engineering system** হিসেবে দেখার সুযোগ পাবেন।