

Some Key Statistical Concepts for Interview Prep

Rafiq Islam

2024-09-05

Table of contents

Basic Statistical Terminologies	1
The mean	1
Variance	2
Covariance	3
Correlation	3
P-Values and Hypothesis Testing	4
Maximum Likelihood Estimation (MLE)	4
Maximum A Posteriori (MAP)	4

In the world of data analysis and machine learning, statistics plays a vital role in making sense of the data. Whether you're estimating parameters, testing hypotheses, or understanding relationships between variables, statistical concepts guide how we interpret data. In this post, I want to summarise and collect some fundamental statistical ideas that are quite common and asked in many data science, machine learning, and quant interviews

Basic Statistical Terminologies

The mean

The **mean** is one of the most basic statistical concepts and represents the average value of a dataset. It's calculated by summing all the values in a dataset and then dividing by the number of observations.

Mathematically, for a set of discrete observations x_1, x_2, \dots, x_n , the mean μ or *Expected Value* is defined as:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\Rightarrow \mathbb{E}[X] = \sum_{i=1}^n x_i \mathbb{P}(X = x_i)$$

For a continuous random variable X , the mean

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

where, $\mathbb{P}(X = x)$ is the *probability mass function (pmf)* and $f_X(x)$ is the *probability density function (pdf)* of the random variable X , depending on whether it is discrete or continuous type. The mean helps describe the central tendency of data, but it can be sensitive to outliers.

Variance

Variance measures the spread or dispersion of a dataset relative to its mean. It tells us how far the individual data points are from the mean. A small variance indicates that data points are clustered closely around the mean, while a large variance means they are spread out.

The formula for variance σ^2 is:

$$\begin{aligned} \sigma^2 = \text{Var}(X) &= \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned}$$

However, the population and sample variance formula are slightly different. For discrete observations, the sample variance is given as

$$s = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2$$

Instead of dividing by n we divide by $n-1$ to have the sample variance unbiased and bigger than the population variance so that it contains the true population variance.

Examples

1. Normal Distribution with mean μ and variance σ^2 has the pdf $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x-\mu)^2}{2\sigma^2} \right]$
2. Standard Normal Distribution with mean 0 and variance 1 has the pdf $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{x^2}{2} \right]$
3. Now if $\log X \sim \mathcal{N}(0, 1)$ then what is the distribution of X ?

Covariance

Covariance measures how two variables move together. If the covariance is positive, the two variables tend to increase or decrease together. If negative, one variable tends to increase when the other decreases.

The formula for covariance between two variables X and Y is:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y)$$

However, covariance doesn't indicate the strength of the relationship, which brings us to **correlation**.

Correlation

Correlation is a standardized measure of the relationship between two variables. It ranges from -1 to 1 , where 1 indicates a perfect positive relationship, -1 a perfect negative relationship, and 0 no relationship.

The most common correlation metric is **Pearson correlation**, defined as:

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

Unlike covariance, correlation gives a clearer picture of the strength and direction of a linear relationship between variables.

P-Values and Hypothesis Testing

P-values and **hypothesis testing** form the backbone of inferential statistics. Hypothesis testing is used to determine if a given assumption (the **null hypothesis** H_0) about a population parameter is true or not.

- The **null hypothesis** H_0 typically suggests no effect or no difference.
- The **alternative hypothesis** H_1 is the claim you want to test.

The **p-value** is the probability of observing a result as extreme as, or more extreme than, the one obtained, assuming the null hypothesis is true. A small p-value (usually less than 0.05) indicates that the null hypothesis is unlikely, and we may reject it in favor of the alternative hypothesis.

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is a method for estimating the parameters of a statistical model. The idea behind MLE is to find the parameter values that maximize the likelihood function, which represents the probability of observing the given data under a particular model.

Given a parameter θ and observed data X , the likelihood function is:

$$L(\theta|X) = P(X|\theta)$$

MLE finds the parameter $\hat{\theta}$ that maximizes this likelihood:

$$\hat{\theta} = \arg \max_{\theta} L(\theta|X)$$

MLE is widely used in statistical modeling, from simple linear regression to complex machine learning algorithms.

Maximum A Posteriori (MAP)

While MLE focuses on maximizing the likelihood, **Maximum A Posteriori (MAP)** estimation incorporates prior information about the parameters. MAP is rooted in Bayesian statistics, where the goal is to find the parameter that maximizes the **posterior** distribution.

The posterior is given by Bayes' Theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

MAP finds the parameter $\hat{\theta}_{\text{MAP}}$ that maximizes the posterior probability:

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} P(\theta|X)$$

Unlike MLE, MAP estimation incorporates the prior distribution $P(\theta)$, making it more robust when prior knowledge is available

Share on



You may also like