MATH 205: Statistical methods

Lecture 15: Continuous random variables

Random variables and expectations

Google data scientist interview questions - Statistics and machine learning

General Statistics

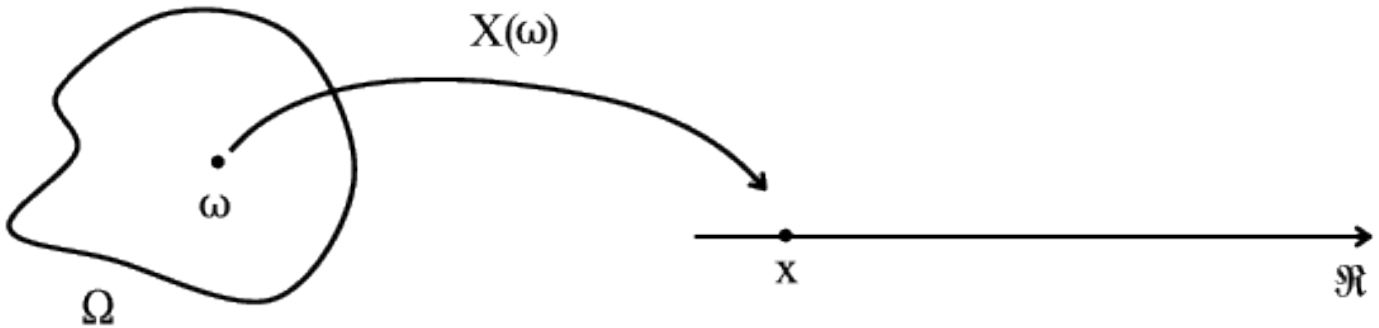
- In what situation would you consider mean over median?
- For sample size n, the margin of error is 3. How many more samples do we need to make the margin of error 0.3?
- What is the assumption of error in linear regression? (Solution)
- Given data from two product campaigns, how could you do an A/B test if we see a 3% increase for one product?

Statistical Probability

- I have a deck and take one card at random. What is the probability you guess it right?
- Explain a probability distribution that is not normal and how to apply that.
- Given uniform distributions X and Y and the mean 0 and standard deviation 1 for both, what's the probability of 2X > Y? (Solution)
- There are four people in an elevator and four floors in a building. What's the probability that each person gets off on a different floor?
- Make an unfair coin fair. (Solution)

16. There were essentially typical probability questions about confidence intervals, sample size, and hypothesis testing, MC simulations.
17. Write a code to generate random normal distribution and plot it.
18. Write code to generate iid draws from distribution X when we only have access to a random number generator.
19. For sample size n, the margin of error is 3. How many more samples do we need to make the margin of error to 0.3? 3. X is draw from the normal distribution of mu with standard deviation of 1. Decides whether mu is not zero.
20. Write a function to generate N sample from normal distribution and plot the histogram.
21. Explain basic maximum likelihood estimator (MLE), Confidential Interval and Hypothesis testing.

Random variables



Notations:

- random variables are denoted by uppercase letters (e.g., X);
- the observed values of the random variables are denoted by lowercase letters (e.g., x)

Discrete random variable

A random variables X is discrete if the set of all possible values of X

- is finite
- is countably infinite

A random variables is characterized by its probability mass function, usually represented as a table

x	1	2	3	4	5	6	7
$p(x)$	.01	.03	.13	.25	.39	.17	.02

Continuous random variables

Continuous random variables

Definition

Let X be a random variable. Suppose that there exists a nonnegative real-valued function $f : \mathbb{R} \to [0, \infty)$ such that for any subset of real numbers A, we have

$$P(X \in A) = \int_A f(x)dx$$

Then X is called **absolutely continuous** or, for simplicity, **continuous**. The function f is called the **probability density function**, or simply the **density function** of X.

Whenever we say that X is continuous, we mean that it is absolutely continuous and hence satisfies the equation above.

Properties

Let X be a continuous r.v. with density function f, then

- $f(x) \geq 0$ for all $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)\, dx = 1$
- For any fixed constant a, b,

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

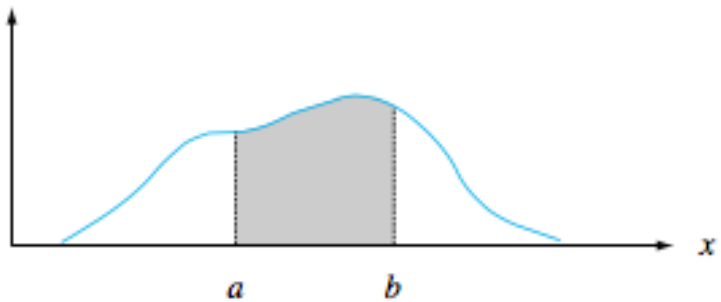


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

Example

Problem

Let X be a continuous r.v. with density function

$$f(x) = \begin{cases} 2x & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Compute $P(X \in [0.25, 0.75])$

Distribution function

Definition

If X is a random variable, then the function F defined on $(-\infty, \infty)$ by

$$F(t) = P(X \leq t)$$

is called the distribution function of X.

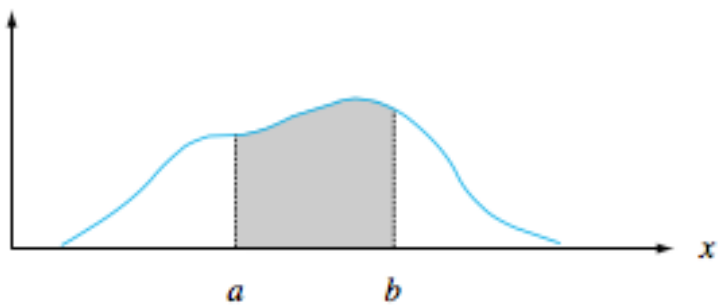


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

Distribution function

For continuous random variable:

$$F(t) = P(X \leq t) = \int_{(-\infty,t]} f(x)dx$$
$$= \int_{-\infty}^{t} f(x)dx$$

Distribution function

For continuous random variable:

$$P(a \leq X \leq b) = \int_a^b f(x)\ dx = F(b) - F(a)$$

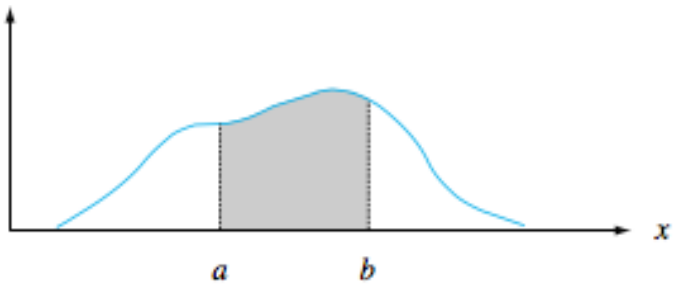


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

Moreover:

$$f(x) = F'(x)$$