CS 240: Reasoning Under Uncertainty (Fall 21) — Oct 27

Lecture 14

Instructor: *Profs Peter J. Hass and Jie Xiong* — SI Worksheet: *Juelin Liu*

1 Introduction

In this lecture we discussed about independent continuous random variables, the laws of large numbers, and the Central Limit Theorem.

2 Independent Continuous Random Variables

Two continuous random variables X and Y are **independent** if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. The joint PDF is the product of the marginals.

A more general definition is X and Y are independent if and only if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$. This definition works when both X and Y are continuous random variables they are both discrete random variables.

When X and Y are independent, we have:

- $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for any sets A and B.
- $E(XY) = E(X) \times E(Y)$
- $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for any functions g and h
- $Var[X+Y] = E[(X+Y)^2] - E[X+Y]^2 = (E[X^2] - E[X]^2) + (E[Y^2] - E[Y]^2) + 2(E[XY] - E[X]E[Y]) = Var[X] + Var[Y]$

3 Laws of Large Numbers

Let $X_1, X_2, \dots$ be s sequence of independent and identically distributed (i.i.d.) random variables (discrete or continuous). All these random variables have the same mean μ and variance δ^2. Its sample mean $\overline{X}_n$, which is also a random variable can be defined as

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i; \quad E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

$$var(\overline{X}_n) = var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{\sigma^2}{n}$$

$$std(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$$

3.1 The Weak Law of Large Numbers

The weak law of large numbers states that as n becomes large, the distribution of the sample mean $\overline{X}_n$ is increasingly concentrated around the expected value μ.

$$P(|\overline{X}_n - \mu| \geq \epsilon) \to 0 \; as \; n \to \infty$$

We can use Chebyshev's inequality to prove this:

Proof.

$$P(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{var(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2}$$

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

□

3.2 The Strong Law of Large Numbers

The strong law of large numbers states that as n goes to infinity, the sample mean converges to μ with probability 1.

$$P(\lim_{n\to\infty} \overline{X}_n = \mu) = 1$$

Note that both laws can be applied to either discrete or continuous random variables. Both laws deal with the distribution of the sample mean $\overline{X}_n$ is concentrated around μ when n is large. However, they do not discuss what the distribution of $\overline{X}_n$ looks like.

3.3 Central Limit Theorem

Let

$$Z_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$E(Z_n) = 0; \;\; Var(Z_n) = 1$$

The Central Limit Theorem states that the CDF of Z_n converges to the CDF of a standard normal random variable denoted as Φ. That is:

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x)$$

Z_n is approximately distributed as $N(0,1)$ for large n.
$\overline{X}_n$ is approximately distributed as $N(\mu, \frac{\sigma^2}{n})$.

4 Problems

1. The service times for customers coming through a checkout counter in a retail store are independent random variables with a mean of 1.5 minutes and a variance of 1.0. Approximate the probability that 100 customers can be served in less than 2 hours of total service time.

2. A statistician wants to estimate the mean height h (in meters) of a population, based on n independent samples $X_1, X_2, \ldots, X_n$, chosen uniformly from the entire population. He uses $\overline{X}_n = \frac{(X_1+X_2+\ldots+X_n)}{n}$ the sample mean as the estimate of h, and a rough guess of 1.0m for the standard deviation of the samples X_i.
a) How large should n be so that the standard deviation of $\overline{X}_n$ is at most 1cm?
b) How large should n be so that Chebyshev's inequality guarantees that the estimate is within 5 centimeters from h, with a probability of at least 0.99?

5 Answer

1. From the question we know that $E(X_i) = 1.5$; $Var(X_i) = 1.0$. The size of the sample is $n = 100$ and we can use the Central Limit Theorem to approximate the distribution of the sum of the samples X_n.

$$S_n = X_1 + X_2 + \ldots + X_n$$

$$E(S_n) = 1.5 \times 100 = 150;\ Var(S_n) = 100 \times 1.0 = 100;\ Std(S_n) = 10$$

The Central Limited Theorem states that S_n can be considered as a normal random variable. We can normalize S_n to a standard normal random variable. Let $Z_n = \frac{S_n - 150}{\sqrt{100}}$, and $E(Z_n) = 0$; $Var(Z_n) = 1$. Since S_n has a normal distribution, Z_n also has a normal distribution and it is a standard normal random variable. Instead of computing the integration, we can simply refer to the Z-table to get the answer.

$$P(S_n \le 120) = P(\frac{S_n - 150}{10} \le \frac{120 - 150}{10}) = P(Z_n \le -3) = \Phi(-3) = 0.0013$$

2. a)

$$Var(\overline{X}_n) = Var(\frac{X_1 + X_2 + \ldots + X_n}{n}) = \frac{n * 1.0^2}{n^2} = \frac{1}{n}$$

$$Std(\overline{X}_n) = \sqrt{Var(\overline{X}_n)} = \frac{1}{\sqrt{n}} \le 0.01$$

$$n \ge 10000$$

b) The key insight is the expected value of $\overline{X}_n$ is the mean height of the entire population.

$$P(|\overline{X}_n - h| \le 0.05) \ge 0.99$$

$$P(|\overline{X}_n - E(\overline{X}_n)| \ge 0.05) \le \frac{Var(\overline{X}_n)}{0.05^2} \le 0.01$$

$$Var(\overline{X}_n) = \frac{Var(X_i)}{n} = \frac{1^2}{n} \le 0.01 * 0.05^2$$

$$n \ge 40000$$