CSE 312: Foundations of Computing II

Section 7: Continuous RVs and CLT Solutions

1. Review of Main Concepts

(a) **Closure of the Normal Distribution**: Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. That is, linear transformations of normal random variables are still normal.

(b) **"Reproductive" Property of Normals**: Let $X_1, \ldots, X_n$ be independent normal random variables with $\mathbb{E}[X_i] = \mu_i$ and $Var(X_i) = \sigma_i^2$. Let $a_1, \ldots, a_n \in \mathbb{R}$ and $b \in \mathbb{R}$. Then,

$$X = \sum_{i=1}^{n}(a_i X_i + b) \sim \mathcal{N}\left(\sum_{i=1}^{n}(a_i\mu_i + b), \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

There's nothing special about the parameters – the important result here is that the resulting random variable is still normally distributed.

(c) **Law of Total Probability (Continuous):** A is an event, and X is a continuous random variable with density function $f_X(x)$.

$$\mathbb{P}(A) = \int_{-\infty}^{\infty} \mathbb{P}(A|X = x) f_X(x) dx$$

(d) **Central Limit Theorem (CLT)**: Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[X_i] = \mu$ and $(X_i) = \sigma^2$. Let $X = \sum_{i=1}^{n} X_i$, which has $\mathbb{E}[X] = n\mu$ and $Var(X) = n\sigma^2$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which has $\mathbb{E}\left[\overline{X}\right] = \mu$ and $Var(\overline{X}) = \frac{\sigma^2}{n}$. $\overline{X}$ is called the *sample mean*. Then, as $n \to \infty$, $\overline{X}$ approaches the normal distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Standardizing, this is equivalent to $Y = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ approaching $\mathcal{N}(0, 1)$. Similarly, as $n \to \infty$, X approaches $\mathcal{N}(n\mu, n\sigma^2)$ and $Y' = \frac{X - n\mu}{\sigma\sqrt{n}}$ approaches $\mathcal{N}(0, 1)$.

It is no surprise that $\overline{X}$ has mean μ and variance σ^2/n – this can be done with simple calculations. The importance of the CLT is that, for large n, regardless of what distribution X_i comes from, $\overline{X}$ is *approximately normally distributed with mean μ and variance σ^2/n*. Don't forget the continuity correction, only when $X_1, \ldots, X_n$ are discrete random variables.

2. Bitcoin users

There is a population of n people. The number of Bitcoin users among these n people is i with probability p_i, where, of course, $\sum_{0 \le i \le n} p_i = 1$. We take a random sample of k people from the population (without replacement). Use Bayes Theorem to derive an expression for the probability that there are i Bitcoin users in the population conditioned on the fact that there are j Bitcoin users in the sample. Let B_i be the event that there are i Bitcoin users in the population and let S_j be the event that there are j Bitcoin users in the sample. Your answer should be written in terms of the p_ℓ's, i, j, n and k.

Solution:

$$Pr(B_i|S_j) = \frac{Pr(S_j|B_i)Pr(B_i)}{Pr(S_j)} \qquad \text{by Bayes Theorem}$$

$$= \frac{\frac{\binom{i}{j}\binom{n-i}{k-j}}{\binom{n}{k}} \cdot p_i}{\sum_{\ell=0}^{n} Pr(S_j|B_\ell)Pr(B_\ell)} = \frac{\frac{\binom{i}{j}\binom{n-i}{k-j}}{\binom{n}{k}} \cdot p_i}{\sum_{\ell=0}^{n} \frac{\binom{\ell}{j}\binom{n-\ell}{k-j}}{\binom{n}{k}} \cdot p_\ell} = \frac{\binom{i}{j}\binom{n-i}{k-j} \cdot p_i}{\sum_{\ell=0}^{n} \binom{\ell}{j}\binom{n-\ell}{k-j} \cdot p_\ell}.$$

Above, we used the fact that $Pr(B_\ell) = p_\ell$ and the fact that $Pr(S_j|B_\ell)$ is the probability of choosing a subset of size k, where j of the selected people are from the subset of ℓ Bitcoin users and $k-j$ are from the remaining $n-\ell$ non-Bitcoin users.

3. Another continuous r.v.

The density function of X is given by

$$f(x) = \begin{cases} a + bx^2 & \text{when } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

If $E(X) = \frac{3}{5}$, find a and b.

Solution:

To find the value of two variables, we need two equations to solve as a system. We know that $\mathbb{E}[X] = \frac{3}{5}$, so we know, by the definition of expected value, that

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) = \frac{3}{5}$$

Since $f(x)$ is defined to be 0 outside of the given range, we can integrate within only that range, plugging in $f(x)$:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) = \int_{-\infty}^{0} xf(x) + \int_{0}^{1} xf(x) + \int_{1}^{\infty} xf(x) = \int_{0}^{1} x(a+bx^2) = \int_{0}^{1} ax+bx^3 = \frac{ax^2}{2} + \frac{bx^4}{4}\bigg|_0^1 = \frac{a}{2} + \frac{b}{4} = \frac{3}{5}$$

We also know that a valid density function integrates to 1 over all possible values. Thus, we can perform the same process to get a second equation:

$$\int_{-\infty}^{\infty} f(x) = \int_{-\infty}^{0} xf(x) + \int_{0}^{1} xf(x) + \int_{1}^{\infty} xf(x) = \int_{0}^{1} (a+bx^2) = ax + \frac{bx^3}{3}\bigg|_0^1 = a + \frac{b}{3} = 1$$

Solving this system of equations we get that $a = \frac{3}{5}, b = \frac{6}{5}$

4. Point on a line

A point is chosen at random on a line segment of length L. Interpret this statement and find the probability that the ratio of the shorter to the longer segment is less than $\frac{1}{4}$.

Solution:

Define RV X to be the distance of your random point from the leftmost side of the stick. Since we're choosing a point at random, this RV has an equal likelihood of any distance from 0 to L, making it a continuous uniform RV with parameters $a = 0, b = L$. For the ratio to be less than $\frac{1}{4}$, the shorter segment has to be less than $\frac{L}{5}$ in length.

This can happen when $X < \frac{L}{5}$ or $X > \frac{4L}{5}$. Thus, using the CDF of a continuous uniform distribution, the probability that the ratio is less than $\frac{1}{4}$ is

$$\mathbb{P}(X \leq \frac{L}{5}) + \mathbb{P}(X > \frac{4L}{5}) = F_X(\frac{L}{5}) + (1 - F_X(\frac{4L}{5})) = \frac{\frac{L}{5} - 0}{L - 0} + (1 - \frac{\frac{4L}{5} - 0}{L - 0}) = \frac{1}{5} + (1 - \frac{4}{5}) = \frac{2}{5}$$

5. Min and max of i.i.d. random variables

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables each with CDF $F_X(x)$ and pdf $f_X(x)$. Let $Y = \min(X_1, \ldots, X_n)$ and let $Z = \max(X_1, \ldots, X_n)$. Show how to write the CDF and pdf of Y and Z in terms of the functions $F_X(\cdot)$ and $f_X(\cdot)$.

Solution:
First we compute the CDFs of Z and Y as follows:

$$\begin{aligned}
F_Z(z) &= P(Z < z) \\
&= P(X_1 < z, ..., X_n < z) && \text{[Definition of max]} \\
&= P(X_1 < z) \cdot ... \cdot P(X_n < z) && \text{[Independence]} \\
&= (F_X(z))^n
\end{aligned}$$

$$\begin{aligned}
F_Y(y) &= P(Y < y) \\
&= 1 - P(Y > y) \\
&= 1 - P(X_1 > y, ..., X_n > y) && \text{[Definition of min]} \\
&= 1 - P(X_1 > y) \cdot ... \cdot P(X_n > y) && \text{[Independence]} \\
&= 1 - (1 - F_X(y))^n
\end{aligned}$$

Using the fact that $f_X(x) = \frac{d}{dx}F_X(x)$ and the CDFs that we found we can compute the pdfs of Z and Y as follows:

$$\begin{aligned}
f_Z(z) &= \frac{d}{dz}F_Z(z) \\
&= \frac{d}{dz}(F_X(z))^n \\
&= n \cdot F_X(z)^{n-1} \cdot \left(\frac{d}{dz}F_X(z)\right) \\
&= n \cdot F_X(z)^{n-1} \cdot f_X(z)
\end{aligned}$$

$$\begin{aligned}
f_Y(y) &= \frac{d}{dy}F_Y(y) \\
&= \frac{d}{dy}\left(1 - (1 - F_X(y))^n\right) \\
&= -n \cdot (1 - F_X(y))^{n-1} \cdot \frac{d}{dy}(1 - F_X(y)) \\
&= n \cdot (1 - F_X(y))^{n-1} \cdot f_X(y)
\end{aligned}$$

6. Round off error

Let X be the sum of 100 real numbers, and let Y be the same sum, but with each number rounded to the nearest integer before summing. If the roundoff errors are independent and uniformly distributed between -0.5 and 0.5, what is the approximate probability that $|X - Y| > 3$?

Solution:
Let $X = \sum_{i=1}^{100} X_i$, and $Y = \sum_{i=1}^{100} r(X_i)$, where $r(X_i)$ is X_i rounded to the nearest integer. Then, we have

$$X - Y = \sum_{i=1}^{100} X_i - r(X_i)$$

Note that each $X_i - r(X_i)$ is simply the round off error, which is distributed as $Unif(-0.5, 0.5)$. Since $X - Y$ is the sum of 100 i.i.d. random variables with mean $\mu = 0$ and variance $\sigma^2 = \frac{1}{12}$, $X - Y \approx W \sim \mathbb{N}(0, \frac{100}{12})$ by the Central Limit Theorem. For notational convenience let $Z \sim \mathbb{N}(0,1)$

$$
\begin{aligned}
\mathbb{P}(|X-Y|>3) &\approx \mathbb{P}(|W|>3) && \text{[CLT]}\\
&= \mathbb{P}(W>3)+\mathbb{P}(W<-3) && \text{[No overlap between } W>3 \text{ and } W<-3]\\
&= 2\,\mathbb{P}(W>3) && \text{[Symmetry of normal]}\\
&= 2\,\mathbb{P}\left(\frac{W}{\sqrt{100/12}} > \frac{3}{\sqrt{100/12}}\right)\\
&\approx 2\,\mathbb{P}(Z>1.039) && \text{[Standardize } W]\\
&= 2\,(1-\Phi(1.039)) \approx 0.29834
\end{aligned}
$$

7. Tweets

A prolific twitter user tweets approximately 350 tweets per week. Let's assume for simplicity that the tweets are independent, and each consists of a uniformly random number of characters between 10 and 140. (Note that this is a discrete uniform distribution.) Thus, the central limit theorem (CLT) implies that the number of characters tweeted by this user is approximately normal with an appropriate mean and variance. Assuming this normal approximation is correct, estimate the probability that this user tweets between 26,000 and 27,000 characters in a particular week. (This is a case where continuity correction will make virtually no difference in the answer, but you should still use it to get into the practice!).

Solution:

Let X be the total number of characters tweeted by a twitter user in a week. Let $X_i \sim Unif(10,140)$ be the number of characters in the ith tweet (since the start of the week). Since X is the sum of 350 i.i.d. rvs with mean $\mu = 75$ and variance $\sigma^2 = 1430$, $X \approx N \sim \mathbb{N}(350 \cdot 75, 350 \cdot 1430)$. Thus,

$$\mathbb{P}(26,000 \le X \le 27,000) \approx \mathbb{P}(25,999.5 \le N \le 27,000.5)$$

Standardizing this gives the following formula

$$
\begin{aligned}
\mathbb{P}(25,999.5 \le N \le 27,000.5) &\approx \mathbb{P}\left(-0.3541 \le \frac{N-350\cdot 75}{\sqrt{350\cdot 1430}} \le 1.0608\right)\\
&= \mathbb{P}\left(-0.3541 \le Z \le 1.0608\right)\\
&= \Phi(1.0608) - \Phi(-0.3541)\\
&\approx 0.4923
\end{aligned}
$$

So the probability that this user tweets between 26,000 and 27,000 characters in a particular week is approximately 0.4923.