Section 7: Continuous Random Variables and the Central Limit Theorem

Review of Main Concepts

- **Normal (Gaussian, "bell curve")**: $X \sim \mathcal{N}(\mu,\ \sigma^2)$ iff X has the following probability density function:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},\ \ x \in \mathbb{R}$$

  $\mathbb{E}[X] = \mu$ and $Var(X) = \sigma^2$. The "standard normal" random variable is typically denoted Z and has mean 0 and variance 1: if $X \sim \mathcal{N}(\mu,\ \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. The CDF has no closed form, but we denote the CDF of the standard normal as $\Phi(z) = F_Z(z) = \mathbb{P}(Z \leq z)$. Note from symmetry of the probability density function about $z = 0$ that: $\Phi(-z) = 1 - \Phi(z)$.

- **Standardizing**: Let X be any random variable (discrete or continuous, not necessarily normal), with $\mathbb{E}[X] = \mu$ and $Var(X) = \sigma^2$. If we let $Y = \frac{X-\mu}{\sigma}$, then $\mathbb{E}[Y] = 0$ and $Var(Y) = 1$.

- **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.

- **"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.

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

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

- **Central Limit Theorem (CLT)**: Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[X_i] = \mu$ and $Var(X_i) = \sigma^2$. 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.

1. Content Review

(a) True or False: For any random variable X, $\mathbb{P}(X = 5) = \mathbb{P}(X - 5 = 0)$.

(b) True or False: For some continuous random variable X, $\mathbb{P}(X \leq 5) \neq \mathbb{P}(X < 5)$.

(c) True or False: Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a, b \in \mathbb{R}$. Then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

(d) Select one: Suppose we have n independent and identically distributed random variables $X_1, X_2, \ldots, X_n$, each with mean μ and variance σ^2. Let $X = \sum_{i=1}^{n} X$. Then as n grows large, the Central Limit Theorem tells us that X behaves similarly to which normal distribution?

- ◯ $X \sim \mathcal{N}(n\mu, n\sigma^2)$
- ◯ $X \sim \mathcal{N}(\mu, n\sigma^2)$
- ◯ $X \sim \mathcal{N}(n\mu, \sigma^2)$
- ◯ $X \sim \mathcal{N}(n\mu, n^2\sigma^2)$

2. Will the battery last?

Suppose that the number of miles that a car can run before its battery wears out is exponentially distributed with expectation 10,000 miles. If the owner wants to take a 5000 mile road trip, what is the probability that she will be able to complete the trip without replacing the battery, given that the car has already been used for 2000 miles on the same trip?

3. Normal questions

(a) Let X be a normal random with parameters $\mu = 10$ and $\sigma^2 = 36$. Compute $\mathbb{P}(4 < X < 16)$.

(b) Let X be a normal random variable with mean 5. If $\mathbb{P}(X > 9) = 0.2$, approximately what is $Var(X)$?

(c) Let X be a normal random variable with mean 12 and variance 4. Find the value of c such that

$$\mathbb{P}(X > c) = 0.10.$$

Central Limit Theorem Problems
The next few problems are CLT focused problems. Here's a general template for that! Sometimes we'll be trying to solve for the probability of something (e.g., $P(X \leq 10)$, and sometimes, we'll be trying to find a value of some parameter that will allow for the probability to be in a certain range (e.g., $P(X \leq 10) \leq 0.2$). Regardless, we still will want to apply CLT on X, and follow the same process (the only difference is that we may be solving for different things).

(a) Setup the problem - write event you are interested in, in terms of sum of random variables. (what do we want to solve for/what is the probability we want to be true?)
   - Write the random variable we're interested in as a sum of i.i.d., random variables
   - Apply CLT to $X = X_1 + X_2 + ... + X_n$ (we can approximate X as a normal random variable $Y \sim N(\mu, \sigma^2)$)
   - Write the probability we're interested in

(b) *If the RVs are discrete*, apply continuity correction.
(c) Normalize RV to have mean 0 and standard deviation 1: $Z = \frac{Y-\mu}{\sigma}$
(d) Replace RV in probability expression with $Z \sim N(0, 1)$
(e) Write in terms of $\Phi(z) = P(Z \leq z)$
(f) Look up in the Phi table (or do a reverse Phi table lookup if we're looking for a value of z that gives us a certain probability)

4. Do it in Reverse

(a) Let X be a normal random variable with parameters $\mu = 8$ and $\sigma^2 = 9$. Find x such that $\mathbb{P}(X \leq x) = 0.6$.

(b) Lots of statistics (like standardized test scores or heights) use *percentiles* to give context to where outcomes fall in a distribution. The nth percentile marks the outcome at which n% of the data points are less than the outcome. Let Y be a normal random variable with parameters $\mu = 15$ and $\sigma^2 = 4$. What value y marks the 85th percentile? What value b marks the 15th percentile?

5. 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 $|X - Y| > 3$?

6. Bad Computer

Each day, the probability your computer crashes is 10%, independent of every other day. Suppose we want to evaluate the computer's performance over the next 100 days.

(a) Let X be the number of crash-free days in the next 100 days. What distribution does X have? Identify $\mathbb{E}[X]$ and $Var(X)$ as well. Write an exact (possibly unsimplified) expression for $\mathbb{P}(X \geq 87)$.

(b) Approximate the probability of at least 87 crash-free days out of the next 100 days using the Central Limit Theorem. Use continuity correction.

Important: continuity correction says that if we are using the normal distribution to approximate

$$\mathbb{P}(a \leq \sum_{i=1}^{n} X_i \leq b)$$

where $a \leq b$ are integers and the X_i's are i.i.d. **discrete** random variables, then, as our approximation, we should use

$$\mathbb{P}(a - 0.5 \leq Y \leq b + 0.5)$$

where Y is the appropriate normal distribution that $\sum_{i=1}^{n} X_i$ converges to by the Central Limit Theorem.[1]

For more details see pages 209-210 in the book.

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!).

[1] The intuition here is that, to avoid a mismatch between discrete distributions (whose range is a set of integers) and continuous distributions, we get a better approximation by imagining that a discrete random variable, say W, is a continuous distribution with density function

$$f_W(x) := p_W(i) \quad \text{when } i - 0.5 \leq x < i + 0.5 \text{ and } i \text{ integer}$$

8. 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 $\mathbb{E}[X] = \frac{3}{5}$, find a and b.

9. 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 $\frac{1}{4}$.

10. 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, $\sum_{0 \leq i \leq n} p_i = 1$. 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.

11. 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 $f_X(\cdot)$.