Quiz Section 7 – Solutions

Review

1) Uniform: $X \sim \text{Uniform}(a, b)$ iff X has the following probability density function:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = \frac{a+b}{2}$ and $\text{Var}(X) = \frac{(b-a)^2}{12}$. This represents each real number from $[a, b]$ to be equally likely.

2) Exponential: $X \sim \text{Exponential}(\lambda)$ iff X has the following probability density function:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$. $F_X(x) = 1 - e^{-\lambda x}$ for $x \geqslant 0$. The exponential random variable is the continuous analog of the geometric random variable: it represents the waiting time to the next event, where $\lambda > 0$ is the average number of events per unit time. Note that the exponential measures how much time passes until the next event (any real number, continuous), whereas the Poisson measures how many events occur in a unit of time (nonnegative integer, discrete). The exponential random variable X is memoryless:

$$\text{for any } s, t \geqslant 0,\ \mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$$

The geometric random variable also has this property.

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

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

$\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$. 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 $\Phi(z) = F_Z(z) = \mathbb{P}(Z \leqslant z)$. Note from symmetry of the probability density function about $z = 0$ that: $\Phi(-z) = 1 - \Phi(z)$.

4) Central Limit Theorem (CLT): Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $X = \sum_{i=1}^n X_i$, which has $\mathbb{E}[X] = n\mu$ and $\text{Var}(X) = n\sigma^2$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, which has $\mathbb{E}[\overline{X}] = \mu$ 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.

Task 1 – Throwing a dart

Consider the closed unit circle of radius r, i.e., $S = \{(x, y) : x^2 + y^2 \leqslant r^2\}$. Suppose we throw a dart onto this circle and are guaranteed to hit it, but the dart is equally likely to land anywhere in S. Concretely this means that the probability that the dart lands in any particular area of size A (that is entirely inside the circle of radius R), is equal to $\frac{\text{A}}{\text{Area of whole circle}}$. The density outside the circle of radius r is 0.

Let X be the distance the dart lands from the center. What is the CDF and pdf of X? What is $\mathbb{E}[X]$ and $\text{Var}(X)$?

Since $F_X(x)$ is the probability that the dart lands inside the circle of radius x, that probability is the area of a circle of radius x divided by the area of the circle of radius r (i.e., $\pi x^2/\pi r^2$). Thus, our CDF looks like

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{r^2} & 0 < x \leqslant r \\ 1 & x > r \end{cases}$$

To find the PDF we just need to take the derivative of the CDF, which give us the following:

$$f_X(x) = \begin{cases} \frac{2x}{r^2} & 0 < x \leqslant r \\ 0 & \text{otherwise} \end{cases}$$

Using the definition of expectation we get

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_0^r x\frac{2x}{r^2}\,\mathrm{d}x = \frac{2}{3r^2}\left(x^3\Big|_0^r\right) = \frac{2}{3}r$$

We know that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x = \int_0^r x^2\frac{2x}{r^2}\,\mathrm{d}x = \frac{2}{4r^2}\left(x^4\Big|_0^r\right) = \frac{1}{2}r^2$$

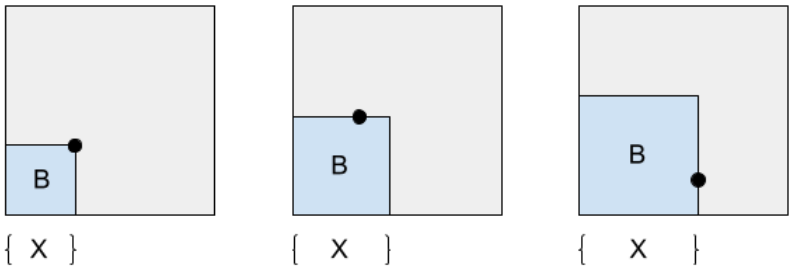
Plugging this into our variance equation gives

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{2}r^2 - \left(\frac{2}{3}r\right)^2 = \frac{1}{18}r^2$$

Task 2 – A square dartboard?

You throw a dart at an $s \times s$ square dartboard. The goal of this game is to get the dart to land as close to the lower left corner of the dartboard as possible. However, your aim is such that the dart is equally likely to land at any point on the dartboard. Let random variable X be the length of the side of the smallest *square* B in the lower left corner of the dartboard that contains the point where the dart lands. That is, the lower left corner of B must be the same point as the lower left corner of the dartboard, and the dart lands somewhere along the upper or right edge of B. For random variable X, find the CDF, PDF, $\mathbb{E}[X]$, and $\mathrm{Var}(X)$.

See the image below for three examples of how X can take on a value.



Since $F_X(x)$ is the probability that the dart lands inside the square of side length x, that probability is the area of a square of length x divided by the area of the square of length radius s (i.e., x^2/r^2). Thus, our CDF looks like

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ x^2/s^2, & \text{if } 0 \leqslant x \leqslant s \\ 1, & \text{if } x > s \end{cases}$$

To find the PDF, we just need to take the derivative of the CDF, which gives us the following:

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = \begin{cases} 2x/s^2, & \text{if } 0 \leqslant x \leqslant s \\ 0, & \text{otherwise} \end{cases}$$

Using the definition of expectation and variance we can compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ in the following manner:

$$\mathbb{E}[X] = \int_0^s x f_X(x)\,\mathrm{d}x = \int_0^s \frac{2x^2}{s^2}\,\mathrm{d}x = \frac{2}{s^2}\int_0^s x^2\,\mathrm{d}x = \frac{2}{3s^2}\left[x^3\right]_0^s = \frac{2}{3}s$$

$$\mathbb{E}[X^2] = \int_0^s x^2 f_X(x)\,\mathrm{d}x = \int_0^s \frac{2x^3}{s^2}\,\mathrm{d}x = \frac{2}{s^2}\int_0^s x^3\,\mathrm{d}x = \frac{1}{2s^2}\left[x^4\right]_0^s = \frac{1}{2}s^2$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{1}{2}s^2 - \left(\frac{2}{3}s\right)^2 = \frac{1}{18}s^2$$

Task 3 – Grading on a curve

Each day, the number of customers who come to the CSE 312 probability gift shop is approximately $\mathrm{Poi}(11)$. Approximate the probability that, after the quarter ends ($9 \times 7 = 63$ days), that we had over 700 customers.

The total number of customers that come is $X = X_1 + \cdots + X_{63}$, where each $X_i \sim \mathrm{Poi}(11)$ has $\mathbb{E}[X_i] = \mathrm{Var}(X_i) = \lambda = 11$. By the CLT, $X \approx \mathcal{N}(\mu = 6311, \sigma^2 = 6311)$ (sum of the means and sum of the variances). Hence,

$$\begin{aligned}
\mathbb{P}(X \geqslant 700) &\approx \mathbb{P}(X \geqslant 699.5) && \text{[continuity correction]} \\
&\approx \mathbb{P}(\mathcal{N}(693, 693) \geqslant 699.5) && \text{[CLT]} \\
&= \mathbb{P}\left(Z \geqslant \frac{699.5 - 693}{\sqrt{693}}\right) && \text{[standardize]} \\
&= 1 - \Phi(0.2469) \\
&= 1 - 0.598 \\
&= 0.402.
\end{aligned}$$

Note that you could compute this exactly as well since you know the sum of iid Poissons is Poisson. In fact, $X \sim \mathrm{Poi}(693)$ (the average rate in 63 days is 693 per 63 days), and you could do a sum which would be very annoying.

Task 4 – Flashlights aren't forever

Suppose I have a flashlight which requires one battery to operate, and I have 18 identical batteries. I want to go camping for a week ($24 \times 7 = 168$) hours. If the lifetime of a single battery is $\mathrm{Exp}(0.1)$, what's the probability my flashlight can operate for the entirety of my trip?

The total lifetime of the battery is $X = X_1 + \cdots + X_{18}$ where each $X_i \sim \mathrm{Exp}(0.1)$ has $\mathbb{E}[X_i] = 1/0.1 = 10$ and $\mathrm{Var}(X_i) = 1/0.12 = 100$. Hence, $\mathbb{E}[X] = 180$ and $\mathrm{Var}(X) = 1800$ by linearity of expectation and since variance adds for independent rvs. In fact, $X \sim \mathrm{Gamma}(r = 18, \lambda = 0.1)$, but we don't have a closed-form for its CDF. By the CLT, $X \approx \mathcal{N}(\mu = 180, \sigma^2 = 1800)$, so

$$\begin{aligned}
\mathbb{P}(X \geqslant 168) &\approx \mathbb{P}(\mathcal{N}(180, 1800) \geqslant 168) && \text{[CLT]} \\
&= \mathbb{P}\left(Z \geqslant \frac{168 - 180}{\sqrt{1800}}\right) && \text{[standardize]} \\
&= 1 - \Phi(-0.28284) \\
&= \Phi(0.28284) && \text{[symmetry of Normal]} \\
&= 0.611.
\end{aligned}$$

Note that we don't use the continuity correction here because the RV's we are summing are already continuous RVs.

Task 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 $|X - Y| > 3$?

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 $\text{Unif}(-0.5, 0.5)$. Since $X - Y$ is the sum of 100 i.i.d. random variables with mean $\mu = 0$ and variance $\sigma^2 = \frac{1}{12}$, $X - Y \approx W \sim \mathcal{N}(0, \frac{100}{12})$ by the Central Limit Theorem. For notational convenience let $Z \sim \mathcal{N}(0, 1)$

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

Task 6 – 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!).

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 $\sigma^2 = 1430$, $X \approx N \sim \mathcal{N}(350 \cdot 75, 350 \cdot 1430)$. Thus,

$$\mathbb{P}(26,000 \leqslant X \leqslant 27,000) \approx \mathbb{P}(26,000 \leqslant N \leqslant 27,000)$$

Now, we apply continuity correction:

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

Standardizing this gives the following formula

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

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