Section 7 – Solutions

Review

1) 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 $\mathrm{Var}(X) = \sigma^2$. 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: $\Phi(-z) = 1 - \Phi(z)$.

Here is the **Standard normal table.**

2) 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 $Var(Y) = 1$.

3) 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.

4) '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.

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

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

6) 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 $\mathbb{E}[\overline{X}] = \mu$ 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.

7) Multivariate: Discrete to Continuous: To be discussed next week....

	Discrete	Continuous
Joint PMF/PDF	$p_{X,Y}(x,y) = \mathbb{P}(X=x, Y=y)$	$f_{X,Y}(x,y) \neq \mathbb{P}(X=x, Y=y)$
Joint range/support $\Omega_{X,Y}$	$\{(x,y) \in \Omega_X \times \Omega_Y : p_{X,Y}(x,y) > 0\}$	$\{(x,y) \in \Omega_X \times \Omega_Y : f_{X,Y}(x,y) > 0\}$
Joint CDF	$F_{X,Y}(x,y) = \sum_{t \leq x, s \leq y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(t,s)\,dsdt$
Normalization	$\sum_{x,y} p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dxdy = 1$
Marginal PMF/PDF	$p_X(x) = \sum_y p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$
Expectation	$\mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y)p_{X,Y}(x,y)$	$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy$
Independence must have	$\forall x,y, p_{X,Y}(x,y) = p_X(x)p_Y(y)$ $\Omega_{X,Y} = \Omega_X \times \Omega_Y$	$\forall x,y, f_{X,Y}(x,y) = f_X(x)f_Y(y)$ $\Omega_{X,Y} = \Omega_X \times \Omega_Y$

Task 1 – Normal questions at the table (repeat)

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

Let $\frac{X-10}{6} = Z$. By the scale and shift properties of normal random variables $Z \sim \mathcal{N}(0,1)$.

$$\mathbb{P}(4 < X < 16) = \mathbb{P}\left(\frac{4-10}{6} < \frac{X-10}{6} < \frac{16-10}{6}\right) = \mathbb{P}(-1 < Z < 1)$$

$$= \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 0.68268$$

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

Let $\sigma^2 = \mathrm{Var}(X)$. Then,

$$\mathbb{P}(X > 9) = \mathbb{P}\left(\frac{X-5}{\sigma} > \frac{9-5}{\sigma}\right) = 1 - \Phi\left(\frac{4}{\sigma}\right) = 0.2$$

So, $\Phi\left(\frac{4}{\sigma}\right) = 0.8$. Looking up the phi values in reverse lets us undo the Φ function, and gives us $\frac{4}{\sigma} = 0.845$. Solving for σ we get $\sigma \approx 4.73$, which means that the variance is about 22.4.

c) Let X be a normal random variable with mean 12 and variance 4.
Find the value of c such that $\mathbb{P}(X > c) = 0.10$.

$$\mathbb{P}(X > c) = \mathbb{P}\left(\frac{X-12}{2} > \frac{c-12}{2}\right) = 1 - \Phi\left(\frac{c-12}{2}\right) = 0.1$$

So, $\Phi\left(\frac{c-12}{2}\right) = 0.9$. Looking up the phi values in reverse lets us undo the Φ function, and gives us $\frac{c-12}{2} = 1.29$. Solving for c we get $c \approx 14.58$.

Task 2 – 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 3 – 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.

Task 4 – Confidence interval

Suppose that $X_1, \ldots, X_n$ are i.i.d. samples from a normal distribution with unknown mean μ and variance 36. How big does n need to be so that μ is in

$$[\overline{X} - 0.11, \overline{X} + 0.11]$$

with probability at least 0.97?

Recall that

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

You may use the fact that $\Phi^{-1}(0.985) = 2.17$.

Our goal is to find n such that μ lies within 0.11 of $\bar{X}$ 97% of the time. This is equivalent to finding n such that the probability that μ lies outside the range is less than 3%.

$$\mathbb{P}(|\bar{X} - \mu| > 0.11) \leqslant 0.03$$

Let us define $Z = \frac{\bar{X}-\mu}{\sigma}$. We can solve for σ by using the Properties of Variance. Since

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

we can say that

$$\text{Var}(\bar{X}) = \text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i)$$

Using the Properties of Variance and the fact that X_i's are i.i.d., $\text{Var}(\bar{X}) = \frac{1}{n^2} \cdot n \cdot 36 = \frac{36}{n}$, so $\sigma = \frac{6}{\sqrt{n}}$.

$$\begin{aligned}
\mathbb{P}(|\bar{X} - \mu| > 0.11) &\leqslant 0.03 \\
\mathbb{P}(|Z| \cdot \sigma > 0.11) &\leqslant 0.03 && \text{[Definition of } Z\text{]} \\
\mathbb{P}\left(|Z| > \frac{0.11}{6}\sqrt{n}\right) &\leqslant 0.03 \\
\mathbb{P}\left(Z < -\frac{0.11}{6}\sqrt{n}\right) &\leqslant 0.015 && \text{[Symmetry of Normal Dist.]} \\
\Phi\left(-\frac{0.11}{6}\sqrt{n}\right) &\leqslant 0.015 && \text{[CDF of Standard Norm.]} \\
-\frac{0.11}{6}\sqrt{n} &\leqslant -\Phi^{-1}(0.985) \\
\sqrt{n} &\geqslant \frac{6 \cdot \Phi^{-1}(0.985)}{0.11} \\
n &\geqslant \left(\frac{6 \cdot \Phi^{-1}(0.985)}{0.11}\right)^2 \\
&\approx 14009.95
\end{aligned}$$

Then n must be at least 14010.