SOLUTIONS MAS472: 2017-18

1. (i) (Bookwork)

$$F_X(x) = \mathbb{P}(X \le x)\checkmark$$

(ii) (Bookwork)

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \le x}\checkmark\checkmark$$

$$\begin{aligned}
\mathbb{E}(\hat{F}_X(x)) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \le x}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\mathbb{I}_{X_i \le x}\right) \qquad \text{(as } \mathbb{E} \text{ is a linear operator)}\checkmark \\
&= \frac{n}{n}\mathbb{P}\left(X_i \le x\right)\checkmark \\
&= \mathbb{P}\left(X_1 \le x\right) = F_X(x)\checkmark.
\end{aligned}$$

(iii) (Bookwork) As $n \to \infty$, $\hat{F}_X(x) \to F_X(x)\checkmark$ with probability one (mark just for saying it converges), and by the CLT, it has an approximate normal distribution $\checkmark$.

$$\widehat{F}_X(x) \sim N\left(F_X(x), \frac{F_X(x)(1-F_X(x))}{n}\right)\checkmark$$

(iv) (Unseen) By the plug-in principle we need to find m s.t.

$$\begin{aligned}
\frac{1}{2} &\le \int_{-\infty}^{m} \mathrm{d}\hat{F}(x)\checkmark \\
&= \int_{-\infty}^{m} \frac{1}{n}\sum \delta(X_i - x)\mathrm{d}x\checkmark \\
&= \frac{1}{n}\sum \mathbb{I}_{X_i \le m}\checkmark
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{2} &\le \int_{m}^{\infty} \mathrm{d}\hat{F}(x) \\
&= \frac{1}{n}\sum \mathbb{I}_{X_i \ge m}
\end{aligned}$$

So $\hat{m} = X_{\left(\frac{n+1}{2}\right)}$, i.e., the midpoint/median of the dataset$\checkmark$.

(v) (Unseen) For $i = 1, \ldots, B\checkmark$

- Sample $X_1^{(i)}, \ldots, X_n^{(i)}$ with replacement from $\{X_1, \ldots, X_n\}$.$\checkmark$

- Set $\hat{m}^{(i)} = X^{(i)}_{(\frac{n+1}{2})}$.

Calculate the standard error as

$$se(\hat{m}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{m}^{(i)} - \bar{\hat{m}})^2$$

where $\bar{\hat{m}} = \frac{1}{B}\sum_{i=1}^{B}\hat{m}^{(i)}$.

(vi) (Routine) A 95% CI can be found by calculating the 2.5th and 97.5th percentiles of $\hat{m}^{(1)}, \ldots, \hat{m}^{(B)}$.
Or, as $\bar{\hat{m}} \pm 1.96 se(\hat{m})$, but only if the $\hat{m}^{(i)}$ are approximately normally distributed.

2. (i) (Routine) Estimates are $\hat{M} = 10419.7$ and $\hat{P}_1 = 65/1000 = 0.065$.
Confidence intervals are

$$\hat{M} \pm 1.96\sqrt{\frac{141763122}{1000}} : (9682, 11158)$$

$$\hat{P} \pm 1.96\sqrt{\frac{0.065 \times 0.935}{1000}} : (0.050, 0.080)$$

The width of the CI for M is $2 \times 1.96 \times \sqrt{\frac{141763122}{1000}}$. So to make the width less than 10 we would need

$$n = \left(2 \times 1.96 \times \frac{\sqrt{141763122}}{10}\right)^2 = 21783888.$$

(ii) (Unseen) Inversion sampling has been used to generate x, with antithetic sampling used to generate negatively correlated pairs. This reduces the variance of the sample mean.

$$\begin{aligned} Var(\bar{c}) &= Var\left\{\frac{1}{1000}\sum_{i=1}^{1000} c_i\right\} \\ &= \frac{1}{1000^2}\{1000 \times Var(c_i) + 2 \times 500 \times Cov(c_i, c_{i+500})\} \\ &= \frac{1}{1000}\{153930901 \times (1 - 0.505)\} \\ &= 76083. \end{aligned}$$

So 95% confidence interval is $10794 \pm 1.96\sqrt{76083}$, i.e. (10253, 11334).

(iii) (Unseen) We want to calculate

$$\begin{aligned} M &= \mathbb{E}c(X, Y) \\ &= \int c(x, y)\pi_X(x)\pi_Y(y)\mathrm{d}x\mathrm{d}y \\ &= \int c(x, y)\pi_X(x)\frac{\pi_Y(y)}{g(y)}g(y)\mathrm{d}x\mathrm{d}y \\ &\approx \frac{1}{n}\sum c(x_i, y_i)h(y_i) \end{aligned}$$

where $g(y)$ is the $N(10,4)$ pdf, and

$$\begin{aligned} h(y) &= \frac{\pi_Y(y)}{g(y)} \\ &= \frac{\frac{1}{4}ye^{-y/2}}{\frac{1}{\sqrt{8\pi}}e^{-(y-10)^2/8}}\mathbb{I}_{y>0} \\ &= \frac{\sqrt{8\pi}ye^{-y/2+(y-10)^2/8}}{4}\mathbb{I}_{y>0} \checkmark \end{aligned}$$

Thus, an estimate of M is

$$\hat{M} = \frac{1}{1000}\sum_{i=1}^{1000} c_i \frac{y_i\exp(-0.5y_i + (y_i-10)^2/8)\sqrt{8\pi}}{4}\mathbb{I}_{y_i>0} \checkmark$$

3. (i) (a) (routine) Method I is a two sample randomisation test.
Method II is a Monte Carlo hypothesis test.
Null hypothesis is that group means are equal $\mu_x = \mu_y$.
Alternative is that $\mu_x \neq \mu_y$.

(b) (bookwork) Assumption is that subjects have been allocated to the two groups randomly.

(c) (routine) There are ${}^{12}C_6 = 924$ possible allocations of patients into group. Smallest p-value obtained when, for the observed data, every measurement in one group is greater than every measurement in the other. So p-value in this case would be $2/924 = 0.0022$ for two-sided alternative.

(d) (routine) p-value for the randomisation test is 0.041, and for the Monte Carlo test it is 0.031. So in both cases we would reject H_0.

(ii) (a) (routine) We require

$$\int_{-k}^{k} f(x)\mathrm{d}x = 1.$$

But we know

$$\int_{-k}^{k} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\mathrm{d}x = \Phi(k) - \Phi(-k)$$

where $\Phi(\cdot)$ is the CDF of a standard normal random variable. Thus

$$r = \frac{1}{\Phi(k) - \Phi(-k)}.$$

.

(b) (unseen) If we use a uniform proposal, then $g(x) = \frac{1}{2k}$ and thus the max of $f(x)/g(x)$ occurs at $x = 0$ and we find

$$M = \sup \frac{f(x)}{g(x)} = \frac{2kr}{\sqrt{2\pi}}$$

So the rejection algorithm in this case is:
1. Simulate $Y \sim U[-k, k]$ and $U \sim U[0, 1]$
2. If $U \leq e^{-X^2/2}$ set $X = Y$. Otherwise return to step 1.

The acceptance rate of this algorithm will be

$$\frac{1}{M} = \frac{\sqrt{2\pi}}{2kr}.$$

(c) (unseen) If we use a truncated normal as a proposal, then the acceptance rate of the rejection algorithm is simply

$$\frac{1}{r} = \Phi(k) - \Phi(-k).$$

Thus the uniform proposal has a higher acceptance rate if

$$\frac{\sqrt{2\pi}}{2kr} > \frac{1}{r}$$

which happens if and only if

$$k < \sqrt{\frac{\pi}{2}}.$$