Math 6070-1, Spring 2014; Partial Solutions to Assignment #6

2. *Let $X_1, \ldots, X_n$ be a random sample from a population with a continuous strictly increasing CDF F, and define $\hat{F}_n$ to be the empirical CDF. Suppose F is continuously differentiable with derivative f, which is of course the PDF of the X_j's. Define*

$$\bar{D}_n := \int_{-\infty}^{\infty} \left[\hat{F}_n(x) - F(x)\right]^2 f(x)\,\mathrm{d}x$$

   (a) *Prove that $\bar{D}_n$ is finite, as well as distribution free.*
   Solution. Since $\hat{F}_n(x) + F(x) \le 2$, it follows that

$$\bar{D}_n \le 4 \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 4.$$

   This shows that $\bar{D}_n$is finite. To show that $\bar{D}_n$ is distribution free we change variables with $y = F(x)$. Since $\mathrm{d}y = f(x)\,\mathrm{d}x$,

$$\begin{aligned}\bar{D}_n &= \int_0^1 \left[\hat{F}_n\left(F^{-1}(y)\right) - y\right]^2 \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \left[\hat{F}_n\left(F^{-1}(y)\right) - F_{\text{unif}}(y)\right]^2 f_{\text{unif}}(y)\,\mathrm{d}y,\end{aligned}$$

   where F_{unif} denotes the Unif$(0\,,1)$ cdf and $f_{\text{unif}} := F'_{\text{unif}}$ the corresponding pdf. Finally, recall that $\hat{F}_n(F^{-1}(y))$ is the empirical cdf of $F(X_1), \ldots, F(X_n)$, which are i.i.d. Unif$(0\,,1)$'s. This proves that $\bar{D}_n$ is distribution free.

   (b) *In the case that the X_i's are Unif$(0\,,1)$, express $\bar{D}_n$ explicitly in terms of the order statistics $X_{1:n}, \ldots, X_{n:n}$.*
   Solution. Let $X_{0:n} := 0$ and $X_{n+1:n} := 1$ in order to see that, in

this particular case,

$$\begin{aligned}
\bar{D}_n &= \int_0^1 \left[\hat{F}_n(x) - x\right]^2 \mathrm{d}x \\
&= \sum_{j=1}^{n+1} \int_{X_{j-1:n}}^{X_{j:n}} \left[\frac{j-1}{n} - x\right]^2 \mathrm{d}x \\
&= \sum_{j=1}^{n+1} \int_{X_{j-1:n}-(j-1)/n}^{X_{j:n}-(j-1)/n} z^2 \, \mathrm{d}z \\
&= \tfrac{1}{3} \sum_{j=1}^{n+1} \left\{ \left(X_{j:n} - \frac{j-1}{n}\right)^3 - \left(X_{j-1:n} - \frac{j-1}{n}\right)^3 \right\}.
\end{aligned}$$

(c) *Prove that $\bar{D}_n \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$ in two different ways:*

i. *Do this by appealing to the Glivenko–Cantelli theorem.*
Solution. Clearly, $\bar{D}_n \le D_n^2 \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = D_n^2$, which goes to zero, as $n \to \infty$, in probability thanks to the Glivenko–Cantelli theorem.

ii. *Do this by first computing the mean of $\bar{D}_n$.*
Solution. By the distribution-free property, we need to only consider the Unif$(0\,,1)$ case.
In that case, $\mathrm{E}\hat{F}_n(x) = x$ and $\mathrm{Var}(\hat{F}_n(x)) = x(1-x)/n$ for $0 \le x \le 1$. It is easy to see from calculus that $g(x) = x(1-x)$ attains its maximum at $x = 1/2$ and $g(1/2) = 1/4$. Therefore,

$$\mathrm{E}(\bar{D}_n) = \int_0^1 \mathrm{Var}(\hat{F}_n(x))\,\mathrm{d}x \le \frac{1}{4n} \to 0 \quad \text{as } n \to \infty.$$

Since $\bar{D}_n \ge 0$, this and the Markov inequality together show that $\bar{D}_n \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$.

3. *Let $X_1, \ldots, X_n$ be a random sample from a CDF F and $Y_1, \ldots, Y_n$ an independent random sample from a CDF G. We wish to test $H_0 : F = G$ versus the two-sided alternative, $H_1 : F \ne G$. Let $\hat{F}_n$ and $\hat{G}_n$ denote the respective empirical CDFs of the X_i's and the Y_j's.*

(a) *Describe a condition on F and/or G under which*

$$\Delta_n := \max_x |\hat{F}_n(x) - \hat{G}_n(x)|$$

is distribution free; you need to justify your assertions.
Solution. Let us assume that F^{-1} exists. Under H_0, so does G^{-1} [$= F^{-1}$!]. Now

$$\begin{aligned}
\Delta_n &= \max_{0<y<1} \left|\hat{F}_n\left(F^{-1}(y)\right) - \hat{G}_n\left(F^{-1}(y)\right)\right| \\
&= \max_{0<y<1} \left|\hat{F}_n\left(F^{-1}(y)\right) - \hat{G}_n\left(G^{-1}(y)\right)\right|,
\end{aligned}$$

the last line holding under H_0. We always have $\hat{F}_n(F^{-1}(y)$ = empirical cdf of i.i.d. Unif$(0\,,1)$ random variables, $F(X_1),\ldots,F(X_n)$; and $\hat{G}_n(G^{-1}(x)$ = empirical cdf of i.i.d. Unif$(0\,,1)$ random variables, $G(Y_1),\ldots,G(Y_n)$. Therefore, *under* H_0, the distribution of Δ_n is the same as the distribution of $\max_{0<y<1}|\hat{U}_n(y)-\hat{V}_n(y)|$, where $\hat{U}_n$ and $\hat{V}_n$ are the empirical cdfs of two independent samples from Unif$(0\,,1)$. This shows that Δ_n is distribution free under H_0.