Statistics - HW3 - Solution

1) We have, by definition, for $i = 1, \dots, n$.

$$\mathbb{1}_{\{X_i \le t\}} = \begin{cases} 1 \text{ if } X_i \le t & \text{proba } \mathbb{P}(X_i \le t) = F(t) \\ 0 \text{ if } X_i > t & \text{proba } 1 - F(t) \end{cases}$$

so they are Bernoulli r.v. with parameter $F(t)$.
They are independant because the X_i's are independant.

2) Apply Hoeffding's inequality, we get, for every $\varepsilon > 0$

$$\mathbb{P}\Big(\Big| \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}}}_{= \hat{F}_n(t)} - F(t) \Big| \geq \varepsilon \Big) \leq 2 \exp(-2\varepsilon^2 n)$$

Since the right-hand side does not depend on t, we may write

$$\sup_{t \in \mathbb{R}} \mathbb{P}\big(|\hat{F}_n(t) - F(t)| \geq \varepsilon \big) \leq \sup_{t \in \mathbb{R}} 2 \exp(-2\varepsilon^2 n)$$
$$= 2 \exp(-2\varepsilon^2 n).$$

3) We always have

$$|\hat{F}_n(t) - F(t)| \leq \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)|,$$

So

$$\mathbb{P}\big(|\hat{F}_n(t) - F(t)| \geq \varepsilon \big) \leq \mathbb{P}\big(\sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| \geq \varepsilon \big).$$

We can take the supremum on t in the left-hand side, the right-hand side does not depend on t! (I know, it's surprising, but observe that for any quantity A, $\sup_t A(t)$ can also be written $\sup_s A(s)$ or with whatever letter you want)

So $\quad \sup_{t \in \mathbb{R}} \mathbb{P}\big(|\hat{F}_n(t) - F(t)| \geq \varepsilon \big) \leq \mathbb{P}\big(\sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| \geq \varepsilon \big),$

and thus inequality (3) implies inequality (2).

4) No and no. Proof:

- We can compute

$$\mathbb{E}[\overline{p}_n(t)] = \frac{1}{\varepsilon}\left(\mathbb{E}[\hat{F}_n(t+\varepsilon)] - \mathbb{E}[\hat{F}_n(t)]\right)$$

$$= \frac{1}{\varepsilon}\left(F(t+\varepsilon) - F(t)\right)$$ because $\hat{F}_n(a)$ is an unbiased estimator of $F(a)$ for all a.

this is "close" to $f'(t)$ for ε small, but not equal to it (in general)

So $\lim_{n\to\infty} \mathbb{E}[\overline{p}_n(t)] \neq F'(t) = p(t)$. $\overline{p}_n(t)$ is Asymptotically biased.

- We know $\hat{F}_n(a)$ is a consistent estimator of $F(a)$, so

$$\hat{F}_n(t+\varepsilon) \xrightarrow[n\to\infty]{\mathbb{P}} F(t+\varepsilon), \quad \hat{F}_n(t) \xrightarrow[n\to\infty]{\mathbb{P}} F(t),$$

and thus $$\overline{p}_n(t) \xrightarrow[n\to\infty]{\mathbb{P}} \frac{1}{\varepsilon}\left(F(t+\varepsilon) - F(t)\right)$$

For the same reason this is not $= F'(t)$.

[So $\overline{p}_n(t)$ is not consistent]

5) We can still write, for any n, t

$$\mathbb{E}[\hat{F}_n(t+\varepsilon_n)] = F(t+\varepsilon_n)$$ because $\hat{F}_n$ is unbiased.

we also have $\mathbb{E}[\hat{F}_n(t)] = F(t)$.

So for any fixed t, for any n, we have

$$\mathbb{E}[\hat{p}_n(t)] = \frac{1}{\varepsilon_n}\left(F(t+\varepsilon_n) - F(t)\right).$$

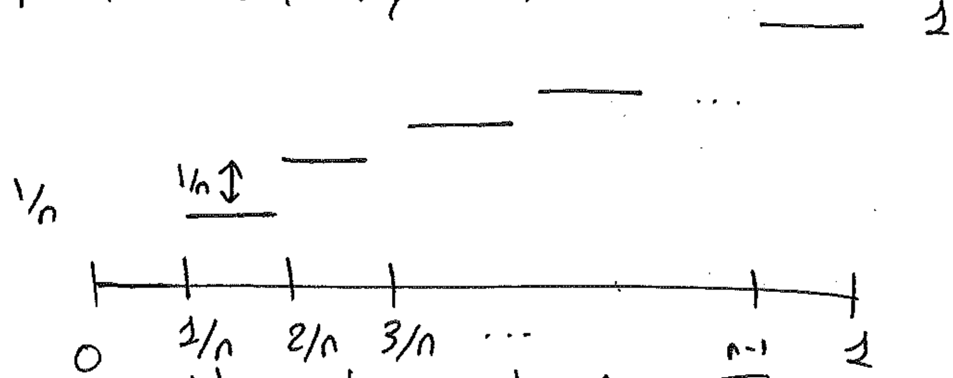
Then, sending $n \to +\infty$, since $\varepsilon_n \xrightarrow[n\to\infty]{} 0$, we have

$$\lim_{n\to\infty} \mathbb{E}[\hat{p}_n(t)] = \lim_{n\to\infty}\left(\frac{1}{\varepsilon_n}\left(F(t+\varepsilon_n) - F(t)\right)\right) = F'(t) = p(t),$$

so $\hat{p}_n(t)$ is asymptotically unbiased.

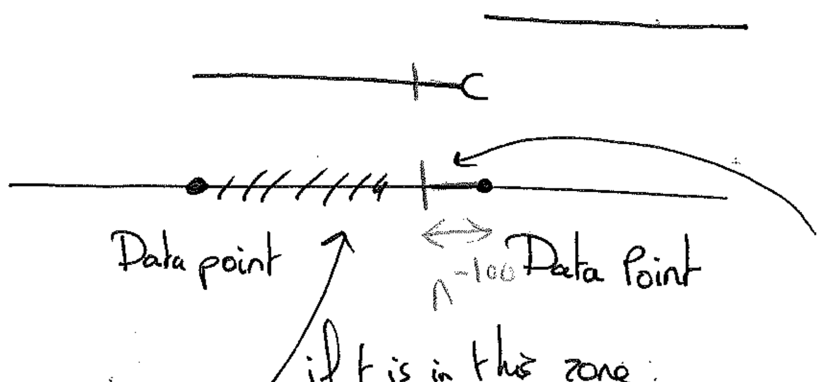
6] Suppose we are sampling the uniform distribution on $[0,1]$ (2)

Typically we will get n data points that are (approximately) uniformly spread on $[0,1]$, so $\hat{F}_n(t)$ looks like



data points $\approx \frac{k}{n}$ for $k = 1, \dots, n$.

For t arbitrary in $[0,1]$, $\hat{F}_n(t + n^{-100})$ and $\hat{F}_n(t)$ are equal, except if t is near a data point (more precisely, except if t is at distance $\leq n^{-100}$ on the left of a data point).



$\hat{F}_n$ jumps when we reach a data point

if t is in this zone, $\hat{F}_n(t + n^{-100}) - \hat{F}_n(t) = \text{jump} = 1/n$

$$\hat{F}_n(t + n^{-100}) - \hat{F}_n(t) = 0.$$

So "most of the time", $\dfrac{\hat{F}_n(t+n^{-100}) - \hat{F}_n(t)}{n^{-100}} = 0$.

At some "exceptional" times, $\dfrac{\hat{F}_n(t+n^{-100}) - \hat{F}_n(t)}{n^{-100}} = \dfrac{\frac{1}{n}}{n^{-100}} = \underline{n^{99}}$ Very large.

In expectation, things work, because

Often $\times 0$ + Rarely $\times$ Very large $=$ OK $\rightarrow$ asymptotically unbiased, see Q.5.

However, "in probability", we only see the "often" part, and choosing ε_n too small leads to an estimator $\hat{p}_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. Not consistent!

7] let t be fixed

We want to prove that $\hat{p}_n(t)$ is a consistent estimator of $p(t)$, which means $\hat{p}_n(t) \xrightarrow[n\to\infty]{\mathbb{P}} p(t)$.

So we fix $\delta > 0$, and we want to prove

$$\mathbb{P}\left(\left|\hat{p}_n(t) - p(t)\right| > \delta \right) \xrightarrow[n\to\infty]{} 0 .$$

Let us write the definition of $\hat{p}_n$:

$$\hat{p}_n(t) = \frac{\hat{F}_n(t+n^{-1/4}) - \hat{F}_n(t)}{n^{-1/4}} .$$

We also have, since $p(t) = F'(t)$.

$$p(t) = \frac{F(t+n^{-1/4}) - F(t)}{n^{-1/4}} + h(n), \quad \text{with } h(n) \xrightarrow[n\to\infty]{} 0,$$

by definition of a derivative.

So
$$\left|\hat{p}_n(t) - p(t)\right| \leq \frac{\left|\hat{F}_n(t+n^{-1/4}) - F(t+n^{-1/4})\right|}{n^{-1/4}} + \frac{\left|\hat{F}_n(t) - F(t)\right|}{n^{-1/4}} + |h(n)|$$
by triangular inequality.

By inequality (1), we have, choosing $\varepsilon = \frac{\delta}{10} n^{-1/4}$:

$$\mathbb{P}\left(\left|\hat{F}_n(t) - F(t)\right| \geq \frac{\delta}{10} n^{-1/4} \right) \leq 2\exp\left(-2\frac{\delta^2}{100} n^{-1/2} \cdot n\right) = \overbrace{2\exp\left(\frac{-\delta^2}{50} n^{1/2}\right)}^{\xrightarrow[\text{as } n\to\infty]{} 0}$$

similarly, we have

$$\mathbb{P}\left(\left|\hat{F}_n(t+n^{-1/4}) - F(t+n^{-1/4})\right| \geq \frac{\delta}{10} n^{-1/4} \right) \leq 2\exp\left(\frac{-\delta^2}{50} n^{1/2}\right).$$

So
$$\mathbb{P}\left(\frac{\left|\hat{F}_n(t+n^{-1/4}) - F(t+n^{-1/4})\right|}{n^{-1/4}} + \frac{\left|\hat{F}_n(t) - F(t)\right|}{n^{-1/4}} \geq \frac{\delta}{5} \right) \quad \left(\frac{\delta}{5} = \frac{\delta}{10} + \frac{\delta}{10}\right)$$

Union bound
$$\leq 2 \cdot 2 \cdot \exp\left(-\frac{\delta^2}{50} n^{1/2}\right).$$

Since $h(n) \longrightarrow 0$, it is less than $\frac{\delta}{5}$ for n large enough.

In conclusion,

$$\frac{\left|\hat{F}_n(t+n^{-1/4}) - F(t+n^{-1/4})\right|}{n^{-1/4}} + \frac{\left|\hat{F}_n(t) - F(t)\right|}{n^{-1/4}} + |h(n)|$$

is less than $\delta/5 + \dfrac{\delta}{5} = \dfrac{2\delta}{5} < \delta$, for n large enough,

with probability $\geq 1 - 4\,\underbrace{\exp\left(\frac{-\delta^2}{50} n^{1/2}\right)}_{\xrightarrow[\text{as } n \to \infty]{} 0}$,

and thus $\mathbb{P}\left(|\hat{p}_n(t) - p(t)| \geq \delta\right) \xrightarrow[n \to \infty]{} 0.$