Mathematical Statistics - Section 1 - NYU Spring 2019 - Homework 3

Questions preceded by a (*) are either more technical or require more initiative.

Hoeffding and DKW We recall Hoeffding's inequality in the case of Bernoulli random variables. If $H_1, \dots, H_n$ are i.i.d. Bernoulli random variables with parameter $\tau \in (0,1)$ (i.e. $\mathbb{P}(H = 1) = \tau$), we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} H_i - \tau\right| \geq \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2 n\right).$$

Let p be an unknown pdf, F be its cdf, let $X_1, \dots, X_n$ be iid random variables with common pdf p. We recall that the empirical cdf $\hat{F}_n$ is defined as

$$\hat{F}_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq t}.$$

1. Recall why, for a fixed t, the variables $\mathbf{1}_{X_i \leq t}$ are i.i.d. Bernoulli random variables. What is their parameter?

2. Apply Hoeffding's inequality to these variables, and deduce a quantitative bound

   $$\mathbb{P}\left(\left|\hat{F}_n(t) - F(t)\right| \geq \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2 n\right). \tag{1}$$

   Observe that the right-hand side does not depend t, and deduce

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

3. In fact, there is a (much more difficult to prove) inequality named the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Theorem 7.5 in the textbook), that reads

   $$\mathbb{P}\left(\sup_{t \in \mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \geq \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2 n\right). \tag{3}$$

   I want you to convince yourself that DKW's inequality (3) is *stronger* than (2) (i.e. inequality (3) implies inequality (2)).

   Explain why, if A_1, A_2 are two positive quantities, it is better to know that

   $$\mathbb{P}(\max(A_1, A_2) \geq 100) \leq \frac{1}{10}$$

   rather than knowing

   $$\mathbb{P}(A_1 \geq 100) \leq \frac{1}{10} \textbf{ and } \mathbb{P}(A_2 \geq 100) \leq \frac{1}{10}.$$

"How to choose the bin sizes when plotting histograms?" We have learned in class how to estimate the cdf F. We recall that

- For **any** t, and for any n, $\hat{F}_n(t)$ is an unbiased estimator of $F(t)$ (what does that mean?).
- For **any** t, $\hat{F}_n(t)$ is a consistent estimator of $F(t)$ (what does that mean?).

This will be useful to keep in mind for the questions below.

We now turn to the problem of estimating the pdf p itself. We recall that, in general, $F' = p$ (the derivative of the cdf is the pdf), but that is only useful in the continuous setting. When working with empirical quantities based on data sets, we will instead use the fact that

$$F'(t) = \lim_{\varepsilon \to 0} \frac{F(t+\varepsilon) - F(t)}{\varepsilon},$$

and thus intuitively, for ε "small", we have approximately-but-not-exactly

$$F'(t) \approx \frac{F(t+\varepsilon) - F(t)}{\varepsilon}.$$

4. If we fix some $\varepsilon > 0$ and define the empirical pdf $\bar{p}_n$ as

$$\bar{p}_n(t) := \frac{\hat{F}_n(t+\varepsilon) - \hat{F}_n(t)}{\varepsilon},$$

do we get a "good" estimator of $p(t)$? More precisely: is it asymptotically unbiased? is it consistent?

Instead, we decide to let ε go to 0 as $n \to \infty$. We work with a sequence $\{\varepsilon_n\}_n$ of positive real numbers, such that $\lim_{n\to\infty} \varepsilon_n = 0$.

5. If we define now the empirical pdf $\hat{p}_n$ as

$$\hat{p}_n(t) := \frac{\hat{F}_n(t+\varepsilon_n) - \hat{F}_n(t)}{\varepsilon_n}, \tag{4}$$

show that it is asymptotically unbiased.

6. However, the consistency is a delicate matter. As an example, discuss what happens if we take ε_n "way too small" (e.g. $\varepsilon_n = n^{-100}$). *Hint: for a fixed, large n, imagine what the function $t \mapsto \hat{F}_n(t)$ typically looks like, and then what $t \mapsto \hat{F}_n(t+n^{-100}) - \hat{F}_n(t)$ looks like. To fix ideas, you may imagine that you are sampling e.g. from the uniform distribution on* $[0, 1]$.

7. (*) If $\varepsilon_n = n^{-1/4}$, show that for any fixed t, the "mobile window" quantity $\hat{p}_n(t)$ as defined in (4) is a consistent estimator of $p(t)$.

   Hint: you will need to prove the convergence in probability "by hand", using the inequality (1) *obtained in question 2 to control the difference between* $\hat{F}_n$ *and* F) *at certain points).*

In fact, using DKW inequality (3), we can prove that $\sup_{t\in\mathbb{R}} |\hat{p}_n(t) - p(t)|$ converges to 0 in probability (perhaps with an extra assumption on p).