Mathematical Statistics - Section 1 - NYU Spring 2019 - Homework 2

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

We recall that, if p is a probability distribution function (pdf) on $\mathbb{R}$, the associated cumulative distribution function (cdf) F is given by

$$F(t) := \int_{-\infty}^{t} p(x)dx.$$

"Probabilistically speaking", if X a random variable with pdf p, the quantity $F(t)$ corresponds to $F(t) = P(X \leq t)$. The cdf F is (among other things) used to define important statistical quantities named *quantiles*.

We will often assume that p is a continuous, non-negative probability distribution function, e.g. a Gaussian distribution. Then F is strictly increasing on $(-\infty, \infty)$ and it is in fact a bijection from $(-\infty, \infty)$ to $(0, 1)$. In particular, for any $p \in (0, 1)$, there exists a unique real number t such that $F(t) = p$. We define this as the *quantile of order* p and we denote it by $Q(p)$. In short, for $p \in (0, 1)$, we have:

$$Q(p) = t \iff F(t) = p.$$

A specific case is $p = \frac{1}{2}$, in which case $Q\left(\frac{1}{2}\right)$ is called *the median* of the distribution. The median is thus the value t such that $F(t) = \frac{1}{2}$, which can be read as $P(X \leq t) = \frac{1}{2}$, so of course we also have $P(X > t) = \frac{1}{2}$. In this case, the median is the only value such that $P(X \leq t) = P(X > t)$, i.e. it is equally likely for X to be above t or below t.

1. Let p_θ be an exponential distribution of parameter $\theta > 0$. Compute the cumulative distribution function $t \mapsto F_\theta(t)$ of p_θ.

2. Compute the values Q_1, Q_2, Q_3 such that $F(Q_1) = 0.25$, $F(Q_2) = 0.5$, $F(Q_3) = 0.75$. You have obtained the first quartile, the median and the third quartile of the distribution.

3. The inter-quartile range is defined as the quantity $Q_3 - Q_1$. Is this quantity increasing or decreasing as a function of θ?

In the sequel, let p be a general pdf such that F is strictly increasing on $(-\infty, \infty)$. We denote by M its median. Let $X_1, \dots, X_n$ be an observation of p (n random variables, i.i.d, of common distribution p). We form the *empirical cdf*, as defined in class, by:

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

4. Explain why there exists either no t or more than one t such that $\hat{F}_n(t) = \frac{1}{2}$. (Hint: imagine an ”example” of a “data set” with $n = 4$ and $n = 5$).

5. Suggests (at least) two different ways in which you could define an “empirical median”. Which one would you prefer? Why?

Now, we let $\hat{M}_n$ be a statistic defined in such a way that we can always guarantee:

$$\left| \hat{F}_n(\hat{M}_n) - \frac{1}{2} \right| \leq \frac{1}{n}.$$

[All your possible definitions of “empirical medians” in question 5 should satisfy this.] We want to prove that $\hat{M}_n$ is a “good” estimator of M.

6. For a given $\varepsilon > 0$, show that

$$P\left(\hat{M}_n \geq M + \varepsilon\right) \leq P\left(\hat{F}_n(M + \varepsilon) \leq \frac{1}{2} + \frac{1}{n}\right).$$

   (Remember that the empirical cdf $\hat{F}_n$ is always a non-decreasing function).

7. What do we know (from class) about $\hat{F}_n(M + \varepsilon)$ and $F(M + \varepsilon)$?

8. (*) Conclude that $\hat{M}_n$ is a consistent estimator of M.