Mathematical Statistics - Section 1 - NYU Spring 2019 - Midterm

NAME:

Solutions

Warm-up First, we consider the usual family of exponential distributions. For any $\theta > 0$, we consider the pdf $x \mapsto f(x;\theta)$, defined for $x \in [0, +\infty)$ by $f(x;\theta) := \theta e^{-\theta x}$. We let $\theta_\star$ be the true, unknown value of the parameter, and $X_1, \dots X_n$ be i.i.d. random variables distributed with common pdf $x \mapsto f(x;\theta_\star)$.

1. **Compute** the mean and the variance of a random variable with pdf $x \mapsto f(x;\theta_\star)$.

Mean:

$$\int_0^{+\infty} \theta_* e^{-\theta_* x} x \, dx \overset{(u = \theta_* x)}{=} \int_0^{+\infty} u e^{-u} \frac{1}{\theta_*} du$$

$$= \frac{1}{\theta_*} \int_0^{+\infty} u e^{-u} du \overset{(IBP)}{=} \frac{1}{\theta_*}\left(\int_0^{+\infty} + e^{-u} du + \left[-u e^{-u}\right]_0^{+\infty} \right)$$

$$= \frac{1}{\theta_*}\left[-e^{-u}\right]_0^{+\infty} \boxed{= \frac{1}{\theta_*}}$$

Second moment:

$$\int_0^{+\infty} \theta_* e^{-\theta_* x} x^2 dx \overset{(u = \theta_* x)}{=} \int_0^{+\infty} u^2 e^{-u} \frac{1}{\theta_*^2} du$$

$$= \frac{1}{\theta_*^2} \int_0^{+\infty} u^2 e^{-u} du \overset{(IBP)}{=} \frac{1}{\theta_*^2}\left(2\int_0^{+\infty} u e^{-u} du + \left[-u^2 e^{-u}\right]_0^{+\infty} \right)$$

$$= \frac{2}{\theta_*^2}$$

So Variance $= \dfrac{2}{\theta_*^2} - \left(\dfrac{1}{\theta_*}\right)^2 = \boxed{\dfrac{1}{\theta_*^2}}$

2. **Define** the empirical mean $\hat{m}_n$ of $X_1, \ldots, X_n$.

$$\boxed{\hat{m}_n := \frac{1}{n} \sum_{i=1}^{n} X_i}$$

3. **State** an asymptotic normality result for $\hat{m}_n$. *Make sure that your statement is complete, and includes an explicit expression for the expectation/variance that appear; the mode of convergence etc.*

By the CLT, we know that

$$\sqrt{n} \frac{\hat{m}_n - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}} \xrightarrow[n \to \infty]{} \mathcal{N}(0,1)$$ in distribution. Here we have

$$\boxed{\theta_* \sqrt{n} \left(\hat{m}_n - \frac{1}{\theta_*} \right) \xrightarrow[n \to \infty]{} \mathcal{N}(0,1)}$$

4. **Compute** the max-likelihood estimator for θ_*.

Likelihood function $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} \theta e^{-\theta X_i}$

Log-likelihood $\log \mathcal{L}_n(\theta) = n \log \theta - \theta \sum_{i=1}^{n} X_i$

Derivative $(\log \mathcal{L}_n)'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i$

$= 0$ for $\theta = \frac{1}{\frac{1}{n} \sum X_i}$, unique maximum.

So MLE $\boxed{\hat{\theta}_n = \frac{1}{\frac{1}{n} \sum X_i}}$

Justify, briefly but precisely, why the following convergences hold:

$$\widehat{\mathsf{FM}}_n \longrightarrow_{n\to\infty} \mathsf{FM} \text{ in probability}, \quad \widehat{\mathsf{SM}}_n \longrightarrow_{n\to\infty} \mathsf{SM} \text{ in probability}.$$

This is a direct consequence of the law of large numbers, applied first to the i.i.d. variables $X_1, .., X_n$, for which we obtain

$$\widehat{FM}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[X] = FM$$

and to the iid variables $X_1^2, \dots, X_n^2$, for which we obtain

$$\widehat{SM}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[X^2] = SM .$$

4. The method of moments suggests to consider the following estimators for $p_\star, q_\star$:

$$\hat{p}_n := 1 + \frac{\widehat{\mathsf{SM}}_n - 3\widehat{\mathsf{FM}}_n}{2}, \quad \hat{q}_n := 2\widehat{\mathsf{FM}}_n - \widehat{\mathsf{SM}}_n.$$

Prove that $\hat{p}_n$ is an unbiased, consistent estimator of $p_\star$ and that $\hat{q}_n$ is an unbiased, consistent estimator of $q_\star$.
Hint: unbiasedness can be proven directly. For consistency, you may want to use the result of the previous question.

- About $\hat{p}_n$

$$\mathbb{E}[\hat{p}_n] = 1 + \frac{1}{2}\left(\mathbb{E}[\widehat{SM}_n] - 3\mathbb{E}[\widehat{FM}_n]\right)$$

We have
$$\begin{cases} \mathbb{E}[\widehat{SM}_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2] = SM \\ \mathbb{E}[\widehat{FM}_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = FM \end{cases}$$

so $\mathbb{E}[\hat{p}_n] = 1 + \frac{1}{2}(SM - 3FM) \boxed{= p_*}$ (according to Q. 2)

$\rightarrow$ $\hat{p}_n$ is an unbiased estimator of p_*.

By Q.3 we know $\widehat{SM}_n \xrightarrow[n\to\infty]{\mathbb{P}} SM$ and $\widehat{FM}_n \xrightarrow[n\to\infty]{\mathbb{P}} FM$

$$\text{So } \hat{p}_n = 1 + \frac{\widehat{SM}_n - 3\widehat{FM}_n}{2} \xrightarrow[n\to\infty]{\mathbb{P}} 1 + \frac{SM - 3FM}{2} \boxed{= p_*}$$

$\longrightarrow$ $\hat{p}_n$ is a consistent estimator of p_*.

The "multinoulli" distribution For any choice of two parameters p, q in the interval $(0,1)$ such that $0 < p+q < 1$, we consider a random variable X distributed according to the "multinoulli distribution" $\mathbb{P}_{p,q}$. It has three possible outcomes: $0, 1$, or 2 and we let

$$\mathbb{P}_{p,q}(X=0) = p, \quad \mathbb{P}_{p,q}(X=1) = q, \quad \mathbb{P}_{p,q}(X=2) = 1-(p+q).$$

Let $p_\star, q_\star$ be fixed, unknown parameters (the "real ones"), that we now want to estimate.

1. **Compute** the first moment $\mathsf{FM} := \mathbb{E}_{p_\star,q_\star}[X]$ and the second moment $\mathsf{SM} := \mathbb{E}_{p_\star,q_\star}[X^2]$.

$$FM = 0 \cdot p_* + 1 \cdot q_* + 2 \cdot (1-(p_*+q_*)) = q_* + 2 - 2p_* - 2q_*$$

$$\boxed{FM = 2 - 2p_* - q_*}$$

$$SM = 0^2 \cdot p_* + 1^2 \cdot q_* + 2^2 \cdot (1-(p_*+q_*)) = q_* + 4 - 4p_* - 4q_*$$

$$\boxed{SM = 4 - 4p_* - 3q_*}$$

2. **Show** that we have the relations

$$p_\star = 1 + \frac{\mathsf{SM} - 3\mathsf{FM}}{2}, \quad q_\star = 2\mathsf{FM} - \mathsf{SM}.$$

We can solve the linear system

$$\begin{cases} FM = 2 - 2p_* - q_* \\ SM = 4 - 4p_* - 3q_* \end{cases} \qquad 2FM = 4 - 4p_* - 2p_*$$

so $\boxed{2FM - SM = q_*}$

and also $3FM = 6 - 6p_* - 3q_*$

so $SM - 3FM = -2 + 2p_*$

and thus $\boxed{p_* = 1 + \frac{SM - 3FM}{2}}$

3. Let $X_1, \ldots, X_n$ be i.i.d. random variables distributed according to $\mathbb{P}_{p_\star,q_\star}$. We let

$$\widehat{\mathsf{FM}}_n := \frac{1}{n}\sum_{i=1}^n X_i, \quad \widehat{\mathsf{SM}}_n := \frac{1}{n}\sum_{i=1}^n X_i^2.$$

• About $\hat{q}_n$

using previous computation

$\mathbb{E}[\hat{q}_n] = 2\mathbb{E}[\widehat{FM}_n] - \mathbb{E}[\widehat{SM}_n] = 2FM - SM = q_*$

$\longrightarrow$ $\hat{q}_n$ is an unbiased estimator of q_*.

Also $\hat{q}_n = 2\widehat{FM}_n - \widehat{SM}_n \xrightarrow[n \to \infty]{\mathbb{P}} 2FM - SM$ using Q.3

$= q_*$

So $\underline{\hat{q}_n \text{ is a consistent estimator of } q_*}$.

Do it yourself: a fallible, fair coin. We consider the following situation: a rudimentary computer is programmed as a "random number generator". Each time that we run the program, it answers 0 or 1 with equal probability. However, once in a while, something goes wrong, in which case the output of the program is $\frac{1}{2}$.

We let $R_1, \ldots, R_n$ be the results obtained in the course of n runs, we assume that the results are independent. We let p_* be the (unknown) probability that the program "fails" and answers $\frac{1}{2}$.

Frame the problem as a parametric statistical model. **Design** an estimator for p_* (there is more than one possible answer) and **justify**, concisely but precisely, its properties: (asymptotic) unbiasedness, consistency, asymptotic normality... ?

p_* = probability it fails and answers $\frac{1}{2}$. R = result.

$\mathbb{P}(R=0) = \mathbb{P}(R=1)$; $\mathbb{P}(R=\frac{1}{2}) = p_*$ so

$$\boxed{\mathbb{P}(R=0) = \mathbb{P}(R=1) = \frac{1}{2} - \frac{p_*}{2} \ ; \ \mathbb{P}(R=\tfrac{1}{2}) = p_*}$$

Family indexed by p_*.

Estimator for p_*?

Compute for example $\mathbb{E}[R^2]$ (second moment)

$$\mathbb{E}[R^2] = 0^2 \cdot \left(\frac{1}{2} - \frac{p_*}{2}\right) + 1^2 \cdot \left(\frac{1}{2} - \frac{p_*}{2}\right) + \left(\frac{1}{2}\right)^2 \cdot p_*$$

$$= \frac{1}{2} - \frac{p_*}{2} + \frac{1}{4} p_* = \frac{1}{2} - \frac{p_*}{4}$$

So $\boxed{p_* = 4\mathbb{E}[R^2] - 2}$ we can use method of moments

$$\boxed{\hat{p}_n := 4\frac{1}{n}\sum_{i=1}^{n} R_i^2 - 2 \quad \text{estimator}}.$$

$\mathbb{E}[\hat{p}_n] = 4\mathbb{E}[R^2] - 2 = p_*$ $\underline{\text{Unbiased}}$ ✓

Law of large number

$\frac{1}{n}\sum_{i=1}^{n} R_i^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[R_1^2]$ so

$\hat{p}_n \xrightarrow[n\to\infty]{\mathbb{P}} 4\mathbb{E}[R^2] - 2 = p_*$

$\underline{\text{Consistent}}$ ✓

CLT ensures

$$\frac{\frac{1}{n}\sum_{i=1}^{n} R_i^2 - \mathbb{E}[R^2]}{\sqrt{Var(R^2)}}\sqrt{n} \longrightarrow \mathcal{N}(0,1)$$

So

$$\boxed{\sqrt{n}\,\frac{\hat{P}_n - \mathbb{E}[R^2]}{4\sqrt{Var(R^2)}} \longrightarrow \mathcal{N}(0,1) \quad \text{asymptotically normal}}$$

(can be computed if needed)

Other solutions:

- Using the empirical pdf by simply counting how many times we get 1/2, and dividing by n.
- MLE
- ....