MATH-UA.0234 Mathematical Statistics Spring 2022
Dr. M. Nitzschner

Problem Set 4

Submission:
Thursday, 03/03/2022, until 1 PM, to be uploaded on the NYU Brightspace course homepage.

---

1. **A brief reminder on conditional distributions** [4 Points]

   In this problem we recall the notion of ***conditional distributions*** *for two random variables X and Y. This is in particular relevant for the notion of sufficient statistics.*

   - *If X and Y are discrete random variables with values in Ω_X and Ω_Y respectively, then for any $y \in \Omega_Y$ with $\mathbf{P}[Y = y] > 0$, the* conditional distribution of X given $Y = y$ *is characterized by the conditional probability mass function*
   $$p_{X|Y=y}(x) = \frac{\mathbf{P}[X = x, Y = y]}{\mathbf{P}[Y = y]}, \qquad x \in \Omega_X.$$
   This is the distribution of X under the probability measure $\mathbf{P}[\,\cdot\,|Y = y]$.
   - *If X and Y are continuous real random variables, then for any $y \in \mathbb{R}$ with $f_Y(y) > 0$, the* conditional distribution of X given $Y = y$ *is characterized by the conditional probability density function*
   $$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \qquad x \in \mathbb{R},$$
   where $f_{X,Y}$ is the joint probability density function of (X, Y) and f_Y is the probability density function of Y.

   (a) A fair coin is tossed 4 times. Let X denote the number of times that heads comes up and $Y = 1$ if heads comes up on the first toss and $Y = 0$ otherwise. Determine the conditional distribution of X given $Y = 1$ and the conditional distribution of Y given $X = 3$.

   (b) Consider jointly continuous random variables X, Y with density
   $$f_{X,Y}(x, y) = 4ye^{-2y(x+1)}\mathbb{1}_{\{x,y>0\}}.$$
   Determine the conditional probability density function $f_{X|Y=y}$. What is the conditional distribution of X given $Y = y$?

2. **Sufficient statistics** [4 Points]

   (a) Let $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} Pois(\lambda)$ with unknown parameter $\lambda > 0$. Show that $T(\mathbf{X}) = \sum_{j=1}^n X_j$ is a sufficient statistic for λ

   (i) ...directly, using the definition of sufficiency.

   (ii) ...using the Neyman-characterization of sufficiency.

   Hint: For (i), use the result of Problem 2 (a) on Problem set 1. You need to calculate $\mathbf{P}[X_1 = x_1, ..., X_n = x_n | T(\mathbf{X}) = t]$ for all possible values of $x_1, ..., x_n, t \in \mathbb{N}_0$.

   (b) Let $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} Pareto(\lambda, a)$ with *known* $a > 0$ and unknown $\lambda > 0$, where we say that $X \sim Pareto(\lambda, a)$ if
   $$f_X(x) = \frac{\lambda a^\lambda}{x^{\lambda+1}}\mathbb{1}_{(a,\infty)}(x).$$
   Find a sufficient statistic $T(\mathbf{X}) \in \mathbb{R}$ for λ, using the Neyman-characterization.

3. **Method of moments** [4 Points]

(a) Consider $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,\theta])$ with unknown $\theta > 0$. Calculate an estimator $\widehat{\theta}_n$ for θ based on the method of moments. Check this estimator for consistency and unbiasedness.

(b) Consider $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} Bin(N,p)$, where $\theta = (N,p) \in \mathbb{N} \times (0,1)$ is unknown (meaning both N and p are unknown). Determine an estimator $(\widehat{N}_n, \widehat{p}_n)$ for (N,p) based on the method of moments.

Hint: You need both $\mathbf{E}_{(N,p)}[X_1]$ and $\mathbf{E}_{(N,p)}[X_1^2]$.

4. **(R exercise) The empirical distribution** [4 Points]

Note: Please provide your source code and images obtained with your solution.

Suppose that $X_1, ..., X_n$ are i.i.d. real random variables such that X_1 has cumulative distribution function F. Given a realization of these random variables, we can consider the *empirical cumulative distribution function*

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

(a) Explain why $\widehat{F}_n(x)$ fulfills the following:

(i) For every $x \in \mathbb{R}$, one has $\mathbf{E}[\widehat{F}_n(x)] = F(x)$ and $\mathrm{Var}[\widehat{F}_n(x)] = \frac{1}{n} F(x)(1 - F(x))$.

(ii) For every $x \in \mathbb{R}$, one has $\widehat{F}_n(x) \xrightarrow[n\to\infty]{\mathbf{P}} F(x)$.

Hint: For any event A in some probability space, one has $\mathbf{E}[\mathbb{1}_A] = \mathbf{P}[A]$.

(b) Use R to generate $n \in \{5, 20, 500\}$ samples of i.i.d. random variables $X_1, ..., X_n$ following a $\mathcal{U}([0,1])$-distribution or a $\mathcal{E}(3)$-distribution. Plot the empiricial cumulative distribution function together with the graph of the cumulative distribution function F_{X_1}. What do you observe?

Hint: The R-command `ecdf()` calculates the empirical distribution function of a vector and `plot(ecdf())` plots the respective graph.

(c) Data on the magnitudes of earthquakes near Fiji are available from R, using the command `quakes`. [1] For help on this dataset type `?quakes`. Plot a histogram and the empirical cumulative distribution function for the *magnitudes*. Calculate the average $\overline{X}_n$ and sample variance S_n^2 for the magnitude.

Hint: The data set `quakes` is a data frame containing information on 5 observations (i.e. a table with 5 columns). To obtain a vector *only* containing the data in column 1, use `quakes[,4]`.

(d) Suppose it is suggested that the data for the magnitudes $X_1, ..., X_n$ can be modelled by a $\Gamma(\alpha,\beta)$ distribution. Find consistent estimators $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ for α and β, and calculate the estimates using the data from `quakes`. Plot the cumulative distribution function of $\Gamma(\widehat{\alpha}_n, \widehat{\beta}_n)$ together with the empirical distribution function of the data. What do you observe?

Hint: Recall that for $X \sim \Gamma(\alpha,\beta)$, we have $\mathbf{E}[X] = \frac{\alpha}{\beta}$ and $\mathrm{Var}[X] = \frac{\alpha}{\beta^2}$.

[1] alternatively, you can find this data on *https://www.stat.cmu.edu/ larry/all-of-statistics/=data/fijiquakes.dat*