Universität Bielefeld
Tanja Pasurek
Ehsan Abedi WiSe 2024/25

Foundations of Statistics

Homework 7

Exercise 1 (Empirical distribution functions and histograms).

(a) For a given sample $x_1, \dots, x_n$, consider the empirical distribution function

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq x),$$

and the histogram over the intervals $(c_k, c_{k+1}]$. Show that the height of the histogram on each bin $(c_k, c_{k+1}]$ is given by

$$\frac{\hat{F}_n(c_{k+1}) - \hat{F}_n(c_k)}{c_{k+1} - c_k}.$$

(b) Given $\alpha \in [0, 1]$, let $x \in \mathbb{R}$ so that $\alpha = \hat{F}_n(x)$. Explain that this x can be considered as the α-quantile of the sample.

(c) From now on, let $x_1, \dots, x_n$ realizations of i.i.d. continuous random variables $X_1, \dots X_n$ with continuous density function f. Consider an equidistant histogram $\hat{f}_n(x)$ with bandwidth parameter $b := c_{k+1} - c_k$ for all k. Find the distribution of $nb\hat{f}_n(x)$ for each x and hence, find its mean and variance. Note that in this exercise n is fixed.

(d) Show that

$$\lim_{b \to 0} \mathbb{E}[\hat{f}_n(x)] = f(x).$$

(e) Use task (c) to find $\mathbb{E}\left[(\hat{f}_n(x) - f(x))^2\right]$, which can be regard as a mean squared error. Now take the limit $b \to 0$ and $nb \to \infty$ to prove

$$\lim_{\substack{b \to 0 \\ nb \to \infty}} \mathbb{E}\left[(\hat{f}_n(x) - f(x))^2\right] = 0.$$

(f) Finally, use Chebyshev's inequality to show that for any $\epsilon > 0$,

$$\mathbb{P}\left[|\hat{f}_n(x) - f(x)| > \epsilon\right] \to 0$$

as $b \to 0$ and $nb \to \infty$. This demonstrates that $\hat{f}_n(x)$ is a *consistent* estimator for $f(x)$.

Exercise 2 (Order statistic, part II). In HW 6, Exercise 2, we studied distribution of smallest and largest order statistics. Here we would like to study distribution of *j-th order statistics*.
Let $n \in \mathbb{N}$ be fixed. Given i.i.d. random variables with CDF $x \mapsto F(x)$,

$$X_1, X_2, \cdots, X_n,$$

we sort them in increasing order

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

We have proved that CDF of minimum $X_{(1)}$ and maxium $X_{(n)}$ are given by

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - [1 - F(x)]^n && \forall x \in \mathbb{R}, \\ F_{X_{(n)}}(x) &= [F(x)]^n && \forall x \in \mathbb{R}, \end{aligned}$$

respectively.

(a) Prove that CDF of *j-th order statistics* $X_{(j)}, j \in \{1, 2, \cdots, n\}$, is given by

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} \qquad x \in \mathbb{R}.$$

Hint: Fix x and let Y be the random variable that counts the number of $X_1, X_2, \cdots, X_n$ that are less than or equal to x. Then use distribution of Y to find the distribution of $X_{(j)}$.

(b) If, in particular, F has density function f, find the density function $f_{X_{(j)}}$ of the random variable $X_{(j)}$.

Exercise 3 (Order statistic, part II, examples). Let's consider the same assumptions as in the preceding exercise. The goal of this exercise is to study the distribution of order statistics in `R` when n becomes large.

(a) Consider the standard normal distribution $\mathcal{N}(0,1)$. In `R`, generate $k = 10^5$ random samples of j-th order statistics $X^n_{(j)}$, where $j := \lceil \alpha n \rceil$ for $n \in \{1, 10, 100, 1000\}$ and $\alpha = 0.6$. For each n, use the `density()` function to fit a density to the samples. To compare these densities, overlay them on a single plot with `xlim = c(-0.5, 1.5)`, `ylim = c(0, 10)`, and label them accordingly. What is the theoretical value that $X^n_{(j)}$ is expected to approach as n increases? Add a vertical line to the plot at this theoretical value and comment on your observations.

(b) Consider i.i.d. random variables $X_i \sim \text{Unif}(0,1)$ for $i = 1, \cdots, n$. Find the PDF of j-th order statistics $X_{(j)}$ for a fixed j and n.

(c) Using the PDF you have found in (b), compute the mean and variance of $X_{(j)}$. Now let $j = \lceil \alpha n \rceil$ for a fixed number $\alpha \in [0,1]$ and find the mean and variance of $X_{(j)}$ as $n \to \infty$. What is your conclusion?

Exercise 4.

(a) Assume that $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are independent random variables, with X having a continuous distribution. Assume Y to be either discrete or continuous. Show that

$$\mathbb{P}\Big(\{\omega \in \Omega : X(\omega) \neq Y(\omega)\}\Big) = 1.$$

(b) Let X_n, $n \geq 1$, be a sequence of independent continuous random variables. Show that

$$\mathbb{P}\Big(\{\omega \in \Omega : X_i(\omega) = X_j(\omega) \text{ for some distinct indexes } i, j \geq 1\}\Big) = 0.$$

Exercise 5. The dataset `pi2000` in the package `UsingR` contains the first two thousand digits of π.

(a) Fit a density estimate to the dataset (use the command `density()`). Compare with the appropriate histogram. Why might you want to add an argument like `breaks=0:10-0.5` to `hist`?

(b) Determine the absolute frequencies $n_0, ..., n_9$ of the digits for the π and plot the empirical CDF.

(c) What kind of distribution do you suspect? (If you are interested to know more, read the Wikipedia article about normal numbers!)