Universität Bielefeld
Tanja Pasurek
Ehsan Abedi WiSe 2023/24

Foundations of Statistics

Homework 8

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) Let $\alpha := \hat{F}_n(x)$ for some $x \in \mathbb{R}$. Show 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) Using (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 that

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

(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.

(a) Show that

$$K(u) = \begin{cases} \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) & -1 \le u \le 1, \\ 0 & \text{elsewhere,} \end{cases}$$

is a kernel density function. It is called the *Cosine kernel*.

(b) In the lectures (Ch. 2.4) the relative efficiency of kernels are defined by the ratio of their values of $C(K)^{5/4}$. Calculate those ratios for the Epanechnikov, Gaussian and Cosine kernels.

(c) From the built-in `faithful` dataset define `A=faithful$eruptions*60` to be the eruption time in seconds. We assume the density f to be normal distributed with mean μ and standard deviation σ. Then we get

$$\left(\frac{1}{\int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x}\right)^{1/5} = \sigma \cdot \left(\frac{8}{3}\sqrt{\pi}\right)^{1/5}.$$

Calculate the optimal bandwidth for the Gaussian and Epanechnikov kernels for the eruption data. Plot the histogram of the eruption data with the command `hist(A,prob=T)` and in the same graphic the kernel density of the Epanechnikov kernel with the optimal bandwidth of the Epanechnikov kernel. Why does the result does not contradict the optimality of the Epanechnikov kernel?

(d) Draw a scatterplot of the duration and the time to the next eruption in seconds. Does the scatterplot give reason to believe that the duration of an eruption influences the time to the next eruption?

Exercise 3 (Moments of kernel density estimators).

(a) Show that the kernel density estimator $\hat{f}(x)$ as defined in Ch. 2.4 for a sample $x_1, ..., x_n$ is a probability density, that is

$$\int_{-\infty}^{+\infty} \hat{f}(x)\,\mathrm{d}x = 1.$$

(b) Show that

$$m_1(\hat{f}) := \int_{-\infty}^{+\infty} x\hat{f}(x)\,\mathrm{d}x = \bar{x}_n. \tag{1}$$

This means that the 1st moment of $\hat{f}$ (= mean value of the corresponding probability distribution on $\mathbb{R}$ having the density $\hat{f}$) is the arithmetic mean of the sample $\bar{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$.
Hint: use the normalization and symmetry property of the kernel.

(c) *(*optional*) Let b denote the bandwidth. Show that the 2nd moment of $\hat{f}$ reads

$$m_2(\hat{f}) = \int_{-\infty}^{+\infty} x^2\hat{f}(x)\,\mathrm{d}x = b^2 m_2(K) + \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i^2,$$

where

$$m_2(K) := \int_{-\infty}^{+\infty} x^2 K(x)\,\mathrm{d}x$$

is the 2nd moment of the corresponding kernel K.

(d) *(*optional*) Finally, show that the variance of $\hat{f}$ is given by

$$\mathrm{var}(\hat{f}) := m_2(\hat{f}) - [m_1(\hat{f})]^2 = b^2 m_2(K) + \sum_{1\le i<j\le n}\left(\frac{x_i - x_j}{n}\right)^2. \tag{2}$$

(e) Compare (1) and (2) and comment on your observation.

Exercise 4. The built-in-dataset `WWWusage` in the package `stats` contains a time series of the numbers of users connected to the Internet through a server every minute.

(a) Calculate the quartiles, maximum, minimum, mean, median, IQR and mode with `R`.

(b) A value x of the dataset is called an outlier if

$$x < x_{0.25} - 1.5\times\mathrm{IQR} \quad \text{or} \quad x > x_{0.75} + 1.5\times\mathrm{IQR}.$$

Here by x_α we mean the α-quantiles. Given this definition, are there outliers among this dataset? Now, draw a boxplot where 10% of the biggest data are plotted as outliers.

(c) With the command `histo <- hist(WWWusage)` draw a default histogram (with automatically chosen bins and absolute frequencies). Add the dataset using the command `rug(WWWusage)`.
Get more numerical information about this histogram by using the command `str(histo)`.

(d) Draw a relative frequency histogram with bins of length 15 and add the dataset. Calculate the probability that a value x of the sample lies in the bin $(95, 110]$.

(e) With the commands `density` and `lines` add a kernel density plot to your histogram. Try Gaussian, Epanechnikov, rectangular, and triangular kernels and vary the bandwidth. Describe the results.

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!)