Universität Bielefeld
Tanja Pasurek
Ehsan Abedi WiSe 2023/24

Foundations of Statistics

Solutions to Homework 8

Exercise 1 (Empirical distribution functions and histograms).

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

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

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

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

Solution: The height h_k (relative frequency) of the histogram on the class interval (= bean) $(c_k, c_{k+1}]$ is given by the number of samples which lie in this interval divided by a proper normalization i.e. the length of the interval and number of samples n (see also Ch. 2.3):

$$\begin{aligned}
h_k &:= \frac{1}{c_{k+1} - c_k} \cdot \frac{1}{n} \sum\nolimits_{i=1}^{n} \mathbb{I}(x_i \in (c_k, c_{k+1}]) \\
&= \frac{1}{c_{k+1} - c_k} \cdot \frac{1}{n} \left[\sum\nolimits_{i=1}^{n} \mathbb{I}(x_i \leq c_{k+1}) - \sum\nolimits_{i=1}^{n} \mathbb{I}(x_i \leq c_k) \right] \\
&= \frac{1}{c_{k+1} - c_k} \left[\frac{1}{n} \sum\nolimits_{i=1}^{n} \mathbb{I}(x_i \leq c_{k+1}) - \frac{1}{n} \sum\nolimits_{i=1}^{n} \mathbb{I}(x_i \leq c_k) \right] \\
&= \frac{1}{c_{k+1} - c_k} \left[\hat{F}_n(c_{k+1}) - \hat{F}_n(c_k) \right].
\end{aligned}$$

Remark: Here we consider intervals $(c_k, c_{k+1}]$ left-open and right-closed. This is also default for the command `hist` in `R`. More precisely, the default value for the `right` argument in `hist` is `TRUE`, which makes the intervals left-open and right-closed.

One can also define the class intervals as $[c_k, c_{k+1})$, which is less convenient. Observe that in this case, we have

$$\begin{aligned} h_k &:= \frac{1}{c_{k+1}-c_k} \cdot \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_i \in [c_k, c_{k+1})) \\ &= \frac{1}{c_{k+1}-c_k} \cdot \frac{1}{n} \left[\sum_{i=1}^n \mathbb{I}(x_i < c_{k+1}) - \sum_{i=1}^n \mathbb{I}(x_i < c_k) \right] \\ &= \frac{1}{c_{k+1}-c_k} \left[\hat{F}_n(c_{k+1}^-) - \hat{F}_n(c_k^-) \right]. \end{aligned}$$

which differs from the previous formula. Here $F(x^-) := \lim_{\substack{y \to x \\ y<x}} F(y)$ is the so-called *left-limit* of the step function F at point x. We know that CDFs, as defined $\mathbb{P}(X \leq \cdot)$ have left limits but are not necessarily left continuous. They are right continuous. That is, we always have $F(x^+) = F(x)$ but it might happen that $F(x^-) \neq F(x)$.

As we see, it is more convenient to use the beans of the form $(c_k, c_{k+1}]$, which is also default in `R`.

(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.

Solution: We have

$$\alpha := \hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_i \leq x) \text{ for some } x \in \mathbb{R}.$$

Hence $m := n\alpha = \sum_{i=1}^n \mathbb{I}(x_i \leq x)$ will be a nonnegative integer. This tells us that exactly $n\alpha$ observations does not exceed x.

Without loss of generality, we may further restrict to the case $0 < \alpha < 1$. To define the α-quantile for $0 < \alpha < 1$, we first should rearrange the raw sample in the increasing order:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Since $m = n\alpha \in \mathbb{N}$ and exactly $n\alpha$ observations of $x_1, ..., x_n$ are $\leq$ than x, we observe that

$$x_{(1)} \leq \ldots \leq x_{(n\alpha)} \leq x < x_{(n\alpha+1)} \ldots \leq x_{(n)}.$$

Recall (cf. Def (5) in Ch. 2.1) that the α-quantile, x_α, is defined as

$$x_\alpha := \frac{1}{2} \left[x_{(n\alpha)} + x_{(n\alpha+1)} \right]$$

in the particular case of $n\alpha \in \mathbb{N}$. This x_α, as well as our point x, divides the odered dataset into 2 parts, where the bottom group has $n\alpha$ elements and the top group has $(n - n\alpha) = n(1 - \alpha)$ elements. So, there are $\alpha \cdot 100\%$ elements less or equal than x.

(c) From now on, let $x_1, \ldots, x_n$ realizations of i.i.d. continuous random variables $X_1, \ldots X_n$ 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.

Solution: Fix $x \in \mathbb{R}$. Then there exists a unique k such that

$$x \in (c_k, c_{k+1}] =: I_k.$$

as these intervals form a partition of $\mathbb{R}$. Similar to part (a), we have the following (see also equation (6) Ch. 2.3.)

$$\hat{f}_n(x) = \frac{1}{nb} \sum\nolimits_{i=1}^{n} \mathbb{I}(x_i \in I_k),$$

which simply implies

$$nb\hat{f}_n(x) = \sum\nolimits_{i=1}^{n} \mathbb{I}(x_i \in I_k).$$

The sum above tells us how many samples lie in the interval I_k. For a given random variable X_i, we can consider $X_i \in I_k$ as a *success* and the other case, $X_i \notin I_k$, as a *failure*. Then the sum above simply counts the number of successes. Therefore, we have

$$nb\hat{f}_n(x) \sim \text{Binomial}(n, p_k)$$

with the success probability p_k that depends on the interval I_k

$$p_k = \mathbb{P}(X_i \in I_k)$$

but of course p_k does not depend on i as the samples are i.i.d. Finally recall the mean and variance of the binomial distribution

$$\mathbb{E}[nb\hat{f}_n(x)] = np_k, \qquad \text{Var}(nb\hat{f}_n(x)) = np_k(1 - p_k).$$

(d) Show that

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

Solution: From the result of task (c), we have

$$\mathbb{E}[\hat{f}_n(x)] = \frac{p_k}{b}$$

All we need to do is to take the limit

$$\begin{aligned}\lim_{b\to 0} \mathbb{E}[\hat{f}_n(x)] &= \lim_{b\to 0} \frac{p_k}{b} \\ &= \lim_{b\to 0} \frac{1}{b}\mathbb{P}(X_i \in I_k) \\ &= \lim_{b\to 0} \frac{1}{b}\int_{I_k} f(y)dy = f(x),\end{aligned}$$

where the last step follows from the continuity of f.

(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

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

Solution: From the results of task (c), we have

$$\begin{aligned}\mathbb{E}\big[(\hat{f}_n(x) - f(x))^2\big] &= \mathbb{E}\big[\hat{f}_n(x)^2 + f(x)^2 - 2\hat{f}_n(x)f(x)\big] \\ &= \mathrm{Var}(\hat{f}_n(x)) + \mathbb{E}[\hat{f}_n(x)]^2 + f(x)^2 - 2f(x)\mathbb{E}[\hat{f}_n(x)] \\ &= \mathrm{Var}(\hat{f}_n(x)) + \Big(\mathbb{E}[\hat{f}_n(x)] - f(x)\Big)^2 \\ &= \frac{np_k(1-p_k)}{n^2b^2} + \Big(\frac{p_k}{b} - f(x)\Big)^2 \\ &= \frac{1}{nb}\frac{p_k}{b}(1-p_k) + \Big(\frac{p_k}{b} - f(x)\Big)^2.\end{aligned}$$

As we showed in part (d), $\frac{p_k}{b} \to f(x)$ and $p_k \to 0$ as $b \to 0$. If in addition, we have $nb \to \infty$ i.e. n grows so fast, we obtain

$$\mathbb{E}\big[(\hat{f}_n(x) - f(x))^2\big] \to 0.$$

Note that here both limits are crucial. The first one $b \to 0$ makes the bandwidth parameter smaller and smaller. The second one $nb \to \infty$ ensures that even for small b, we still have enough samples in the interval to be able to accurately estimate $f(x)$.

(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 $f(x)$.

Solution: By Chebyshev's inequality, for any $\epsilon > 0$, we have

$$\mathbb{P}\Big[|\hat{f}_n(x) - f(x)| > \epsilon\Big] \leq \frac{\mathbb{E}\big[(\hat{f}_n(x) - f(x))^2\big]}{\epsilon^2} \to 0$$

as $b \to 0$ and $nb \to \infty$. Therefore, we have $\hat{f}_n(x) \xrightarrow{\mathbb{P}} f(x)$. Note that $\hat{f}_n(x)$ is a random variable (which depends on the random sample), while $f(x)$ is a constant.

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*.

Solution:

(i) (Nonnegative); It is clear that $K(u) \ge 0$ for all $u \in \mathbb{R}$.

(ii) (Symmetry around zero); Observe that for $-1 \le u \le 1$

$$\begin{aligned} K(-u) &= \frac{\pi}{4}\cos\left(\frac{\pi}{2}(-u)\right) \\ &= \frac{\pi}{4}\cos\left(-\frac{\pi}{2}u\right) \\ &= \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) = K(u). \end{aligned}$$

(iii) (Normalization); We have

$$\begin{aligned} \int_{-\infty}^{\infty} K(u)du &= \int_{-1}^{1} K(u)du \\ &= \frac{1}{2}\sin(\frac{\pi}{2}u)\Big|_{-1}^{1} \\ &= \frac{1}{2}\sin(\frac{\pi}{2}) - \frac{1}{2}\sin(-\frac{\pi}{2}) \\ &= \frac{1}{2} + \frac{1}{2} = 1. \end{aligned}$$

(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.

Solution: We have

$$C(K)^{5/4} = (k_2(K))^{1/2}\, j_2(K)$$

where

$$j_2(K) = \int_{-\infty}^{\infty} K^2(y)\mathrm{d}y, \qquad k_2(K) = \int_{-\infty}^{\infty} y^2 K(y)\mathrm{d}y.$$

- For the Epanechnikov kernel, we obtain

$$j_2(K_E) = \frac{3}{5\sqrt{5}}, \qquad k_2(K_E) = 1$$

and thus

$$C(K_E)^{5/4} \approx 0.268328 \implies \frac{C(K_E)^{5/4}}{C(K_E)^{5/4}} = 1.$$

- For the Gaussian kernel, we obtain

$$j_2(K_G) = \frac{1}{2\sqrt{\pi}}, \qquad k_2(K_G) = 1$$

and thus

$$C(K_G)^{5/4} \approx 0.282095 \implies \frac{C(K_E)^{5/4}}{C(K_G)^{5/4}} \approx 0.951197.$$

- For the Cosine kernel, we obtain

$$j_2(K_E) = \frac{\pi^2}{16}, \qquad k_2(K_E) = 1 - \frac{8}{\pi^2}$$

and thus

$$C(K_C)^{5/4} \approx 0.268476 \implies \frac{C(K_E)^{5/4}}{C(K_C)^{5/4}} \approx 0.999449.$$

(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?

Solution: The optimal bandwidth, which is obtained by minimizing asymptotic mean integrated squared error $AMISE(\hat{f})$ (cf. Ch.2.4 eq. (24)), is given by

$$b_{opt} = \left(\frac{j_2(K)}{n[k_2(K)]^2 \int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x} \right)^{1/5},$$

which also depends on the density f.
Here, b_{opt} can be written as

$$\begin{aligned} b_{opt} &= \left(\frac{1}{\int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x} \right)^{1/5} \left(\frac{j_2(K)}{n[k_2(K)]^2} \right)^{1/5} \\ &= \left(\frac{1}{n} \right)^{1/5} \sigma \cdot \left(\frac{8}{3}\sqrt{\pi} \right)^{1/5} \left(\frac{j_2(K)}{[k_2(K)]^2} \right)^{1/5}. \end{aligned}$$

n is the sample size and σ can be approximately set as the sample standard deviation. For the last term, we can use the computation from part (a).

Solution:

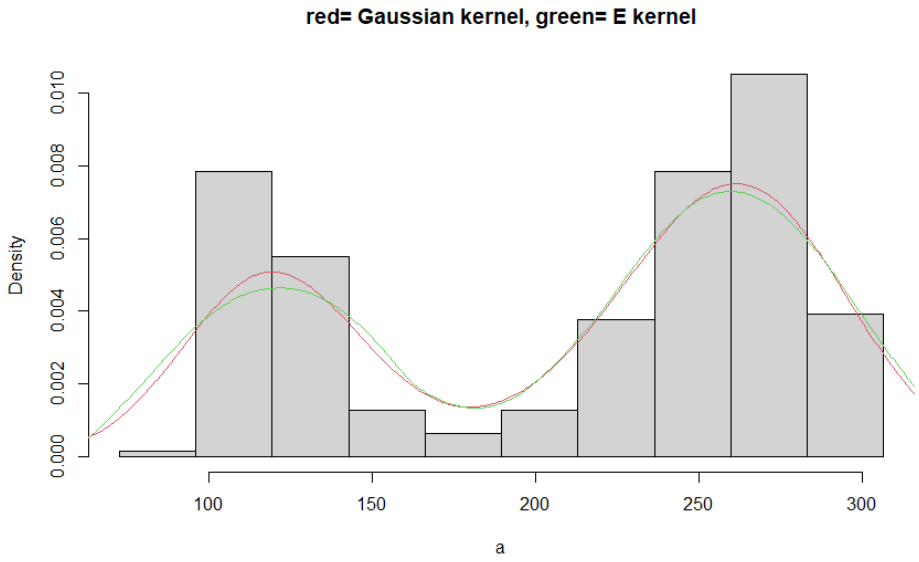
```
a=faithful$eruptions*60
n=length(a)
s=sd(a)

jkG=(2*sqrt(pi))^(-1/5)
jkE=(3/(5*sqrt(5)))^(1/5)

bwG=n^(-1/5)*s*((8/3)*sqrt(pi))^(1/5)*jkG
bwE=n^(-1/5)*s*((8/3)*sqrt(pi))^(1/5)*jkE
bwE
[1] 23.40488
bwG
[1] 23.64025
bw.nrd0(a) #compare with R's Bandwidth selector for Gaussian
kernels
[1] 20.08662

hist(a,prob=T,breaks=seq(min(a)-bwE,max(a)+bwE,by=bwE), main="red=
Gaussian kernel, green= E kernel")
kern=c("gaussian", "epanechnikov")
```

```
lines(density(a,kernel = kern[1],bw=bwG),col=2)
lines(density(a,kernel = kern[2],bw=bwE),col=3)
```

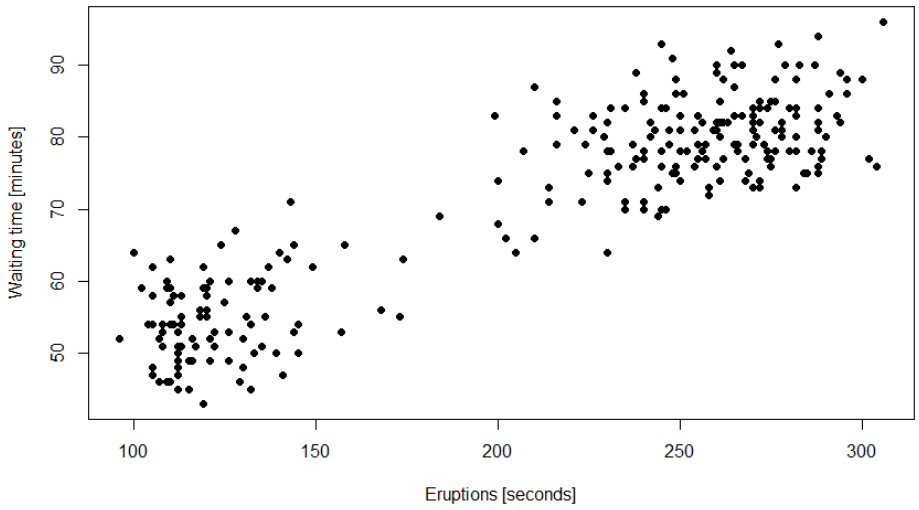


Remark: Note that we see now that density is not Gaussian at all but we used this assumption at the beginning in order to approximate $\int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x$.

(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?

Solution:

```
plot(faithful$eruptions*60, faithful$waiting, pch = 16, xlab
= "Eruptions [seconds]", ylab = "Waiting time [minutes]")
```



We observe that if duration of an eruption is longer, then the time to the next eruption is also longer.

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

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

Solution: By Def. (1) in Ch. 2.4,

$$\hat{f}(x) = \frac{1}{nb}\sum\nolimits_{i=1}^{n} K\left(\frac{x - x_i}{b}\right), \quad x \in \mathbb{R}.$$

Then

$$\begin{aligned}\int_{-\infty}^{+\infty} \hat{f}(x)\,\mathrm{d}x &= \int_{-\infty}^{+\infty} \left[\frac{1}{nb}\sum\nolimits_{i=1}^{n} K\left(\frac{x - x_i}{b}\right)\right] \mathrm{d}x \\ &= \frac{1}{nb}\sum\nolimits_{i=1}^{n} \int_{-\infty}^{+\infty} K\left(\frac{x - x_i}{b}\right)\,\mathrm{d}x.\end{aligned}$$

Introducing new variables

$$y_i = \frac{x - x_i}{b} \quad \text{with} \quad \mathrm{d}y_i = \frac{1}{b}\mathrm{d}x \quad (\text{so that} \quad x = by_i + x_i), \tag{1}$$

we may continue as

$$\begin{aligned}\int_{-\infty}^{+\infty} \hat{f}(x)\,\mathrm{d}x &= \frac{1}{nb}\sum\nolimits_{i=1}^{n} \int_{-\infty}^{+\infty} K\left(y_i\right)\, b\,\mathrm{d}y_i \\ &= \frac{1}{n}\sum\nolimits_{i=1}^{n} \underbrace{\int_{-\infty}^{+\infty} K\left(y\right)\,\mathrm{d}y}_{=1} = \frac{1}{n}\cdot n = 1.\end{aligned}$$

Here we have used the normalization property of the kernel K:

$$\int_{-\infty}^{+\infty} K\left(y\right)\,\mathrm{d}y = 1.$$

(b) Show that

$$m_1(\hat{f}) := \int_{-\infty}^{+\infty} x\hat{f}(x)\,\mathrm{d}x = \bar{x}_n. \tag{2}$$

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.

Solution: Similarly, we have

$$\begin{aligned} m_1(\hat{f}) &:= \int_{-\infty}^{+\infty} x\hat{f}(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x\left[\frac{1}{nb}\sum\nolimits_{i=1}^n K\left(\frac{x-x_i}{b}\right)\right]\,\mathrm{d}x \\ &= \frac{1}{nb}\sum\nolimits_{i=1}^n \int_{-\infty}^{+\infty} xK\left(\frac{x-x_i}{b}\right)\,\mathrm{d}x. \end{aligned}$$

Making the same change of variables as in (1), we get

$$\begin{aligned} m_1(\hat{f}) &= \frac{1}{nb}\sum\nolimits_{i=1}^n \int_{-\infty}^{+\infty} (by_i + x_i)K\left(y_i\right)\, b\,\mathrm{d}y_i \\ &= \frac{1}{n}\sum\nolimits_{i=1}^n \int_{-\infty}^{+\infty} (by_i + x_i)K\left(y_i\right)\,\mathrm{d}y_i \\ &= \frac{b}{n}\sum\nolimits_{i=1}^n \underbrace{\int_{-\infty}^{+\infty} yK\left(y\right)\,\mathrm{d}y}_{=0} \;+\; \frac{1}{n}\left(\sum\nolimits_{i=1}^n x_i\right)\underbrace{\int_{-\infty}^{+\infty} K\left(y\right)\,\mathrm{d}y}_{=1}. \end{aligned}$$

By the normalization and symmetry proprty of the probability kernel K, we note that

$$\int_{-\infty}^{+\infty} K\left(y\right)\,\mathrm{d}y = 1, \quad \int_{-\infty}^{+\infty} yK\left(y\right)\,\mathrm{d}y = 0. \tag{3}$$

Hence, we immediately conclude from that

$$m_1(\hat{f}) = \frac{1}{n}\left(\sum\nolimits_{i=1}^n x_i\right) = \bar{x}_n.$$

(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.

Solution: We proceed analogously to part (b)

$$\begin{aligned} m_2(\hat{f}) \quad &: \quad = \int_{-\infty}^{+\infty} x^2 \hat{f}(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x^2 \left[\frac{1}{nb}\sum\nolimits_{i=1}^{n} K\left(\frac{x-x_i}{b}\right)\right] \mathrm{d}x \\ &= \frac{1}{nb}\sum\nolimits_{i=1}^{n} \int_{-\infty}^{+\infty} x^2 K\left(\frac{x-x_i}{b}\right) \mathrm{d}x. \end{aligned}$$

Introducing new variables $y_i = \frac{x-x_i}{b}$ as in (1), we continue as

$$\begin{aligned} m_2(\hat{f}) \quad &= \frac{1}{nb}\sum\nolimits_{i=1}^{n} \int_{-\infty}^{+\infty} (by_i + x_i)^2 K(y_i)\; b\,\mathrm{d}y_i \\ &= \frac{1}{n}\sum\nolimits_{i=1}^{n} \int_{-\infty}^{+\infty} (b^2 y_i^2 + 2by_i x_i + x_i^2) K(y_i)\; \mathrm{d}y_i \\ &= \frac{1}{n} b^2 n \int_{-\infty}^{+\infty} y^2 K(y)\; \mathrm{d}y \;+\; \frac{1}{n} 2b \left(\sum\nolimits_{i=1}^{n} x_i\right) \underbrace{\int_{-\infty}^{+\infty} yK(y)\; \mathrm{d}y}_{=0} \\ &\quad + \frac{1}{n}\left(\sum\nolimits_{i=1}^{n} x_i^2\right) \underbrace{\int_{-\infty}^{+\infty} K(y)\; \mathrm{d}y}_{=1}. \end{aligned} \tag{4}$$

Applying (3) to the last 2 integrals in (4), we conclude that

$$\begin{aligned} m_2(\hat{f}) \quad &= b^2 \int_{-\infty}^{+\infty} y^2 K(y)\; \mathrm{d}y \;+\; \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i^2 \\ &= b^2 m_2(K) + \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i^2. \end{aligned}$$

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

$$\mathrm{var}(\hat{f}) := m_2(\hat{f}) - [m_1(\hat{f})]^2 = b^2 m_2(K) + \sum_{1 \le i < j \le n} \left(\frac{x_i - x_j}{n}\right)^2. \tag{5}$$

Solution: From the previous parts, we already know that

$$\begin{aligned} \mathrm{var}(\hat{f}) \quad &: \quad = m_2(\hat{f}) - [m_1(\hat{f})]^2 \\ &= b^2 m_2(K) + \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum\nolimits_{i=1}^{n} x_i\right)^2. \end{aligned} \tag{6}$$

Now we use the following general formula (which is true for any collection of numbers $x_1, ..., x_n$):

$$n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 = \sum_{1 \le i < j \le n} (x_i - x_j)^2 . \tag{7}$$

To prove this we start from the right-hand side and expand it

$$\begin{aligned}
\sum_{1 \le i < j \le n} (x_i - x_j)^2 &= \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(x_i^2 + x_j^2 - 2 x_i x_j \right) \\
&= \sum_{i=1}^{n} \sum_{j=i+1}^{n} x_i^2 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} x_j^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2 x_i x_j \\
&= \sum_{i=1}^{n} (n-i) x_i^2 + \sum_{i=1}^{n} (i-1) x_i^2 - 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} x_i x_j \\
&= n \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i^2 - 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} x_i x_j \\
&= n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 .
\end{aligned}$$

Substituting (7) into (6), we immediately get the result:

$$\operatorname{var}(\hat{f}) = b^2 m_2(K) + \frac{1}{n^2} \sum_{1 \le i < j \le n} (x_i - x_j)^2 .$$

(e) Compare the results of (b) and (d) and comment on your observation.

Solution: We note that $m_1(\hat{f})$ does not depend on the choice of kernel and it is always equal to the sample mean $\bar{x}_n$. However, $m_2(\hat{f})$ and thus $\operatorname{var}(\hat{f})$ do depend on the choice of kernel.

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`.

Solution:

```
data <- WWWusage
summary(data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    83.0    99.0   138.5   137.1   167.5   228.0
```

```
# The Interquartile Range
quantile(data, p =0.75) - quantile(data, p =0.25)
```

```
##  75%
## 68.5
```

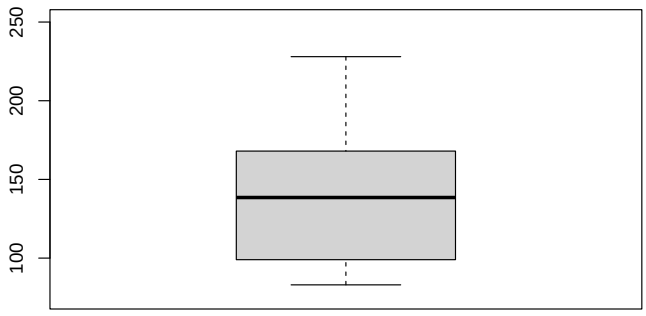
```
IQR(data)
```

```
## [1] 68.5
```

```
# mode is the value that appears most often in the dataset
# write a function that returns the mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(data)
```

```
## [1] 85
```

```
boxplot(data, ylim = c ( min(data)*0.9, max(data)*1.1))
```



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

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

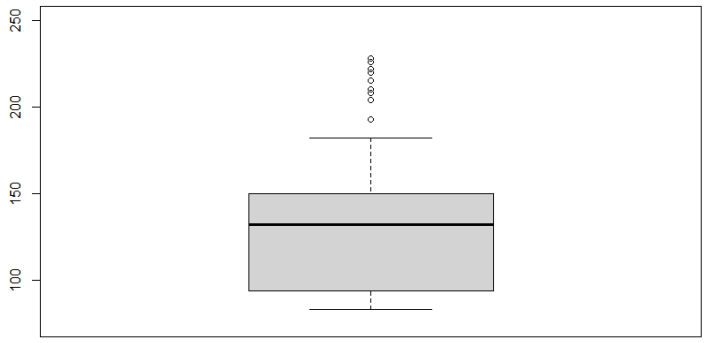
Here by x_α we mean the α-quantiles. Given this definition, are there outliers among this dataset? Now, draw a boxplot where 10% of the biggest data are plotted as outliers.

Solution: Note that for this dataset we have

$$x_{0.25} - 1.5 \times \text{IQR} = -3.75,$$
$$x_{0.75} + 1.5 \times \text{IQR} = 270.25,$$

while, Min. = 83, Max. = 228. As a result, with this definition, there are *no* outliers among the dataset. Alternatively, we can consider 10% of the biggest data as outliers and draw a boxplot, whose `R` code is as follows:

```
data = WWWusage
box_data = data[data < quantile(data, 0.9)]
outliers_data = data[data >= quantile(data, 0.9)]
outlier_n = length(outliers_data)
boxplot(box_data, ylim = c(min(data) * 0.9, max(data) * 1.1),
at = 1, outline= FALSE)
points(rep(1,outliers_n), outliers_data)
```

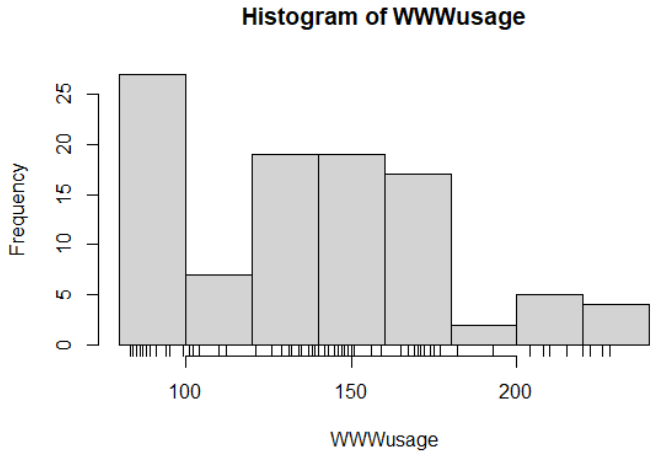


(c) With the command `histo <- hist(WWWusage)` draw a default histogram (with automatically chosen bins and absolute frequencies). Add the dataset using the command `rug(WWWusage)`.
Get more numerical information about this histogram by using the command `str(histo)`.

Solution:

```
histo <- hist(WWWusage)
rug(WWWusage)
str(histo)
```

```
List of 6
$ breaks :  int [1:9] 80 100 120 140 160 180 200 220 240
$ counts :  int [1:8] 27 7 19 19 17 2 5 4
$ density :  num [1:8] 0.0135 0.0035 0.0095 0.0095 0.0085 0.001
0.0025 0.002
$ mids :  num [1:8] 90 110 130 150 170 190 210 230
$ xname :  chr "WWWusage"
$ equidist:  logi TRUE
- attr(*, "class")= chr "histogram"
```



Note that since the length of (equidistance) bins is 20, we have

```
sum(histo$density)
[1] 0.05
sum(histo$density)*20
[1] 1
```

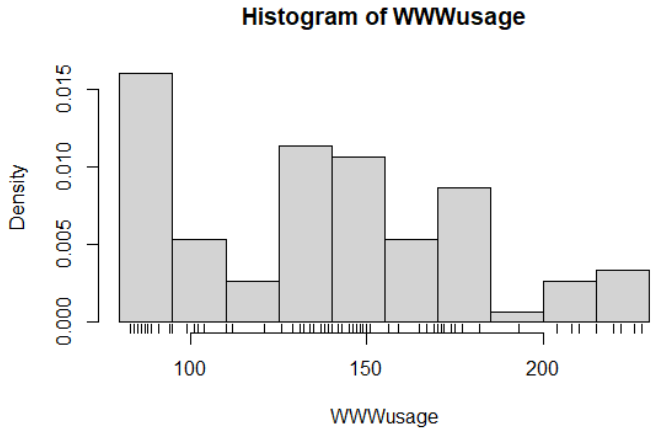
(d) Draw a relative frequency histogram with bins of length 15 and add the dataset. Calculate the probability that a value x of the sample lies in the bin $(95, 110]$.

Solution:

```
histf <- hist(WWWusage,probability = TRUE,breaks = seq(80,
230, by= 15))
rug(WWWusage)
str(histf)
```

```
List of 6
```

```
$ breaks :  num [1:11] 80 95 110 125 140 155 170 185 200 215
...
$ counts :  int [1:10] 24 8 4 17 16 8 13 1 4 5
$ density :  num [1:10] 0.016 0.00533 0.00267 0.01133 0.01067
...
$ mids :  num [1:10] 87.5 102.5 117.5 132.5 147.5 ...
$ xname :  chr "WWWusage"
$ equidist:  logi TRUE
- attr(*, "class")= chr "histogram"
```



Now since the length of (equidistance) bins is 15, we note that

```
sum(histf$density)
[1] 0.06666667
sum(histf$density)*15
[1] 1
```

The probability that a value x of the sample lies in the bin $(95, 110]$ is

```
histf$density[2]*15
[1] 0.08
```

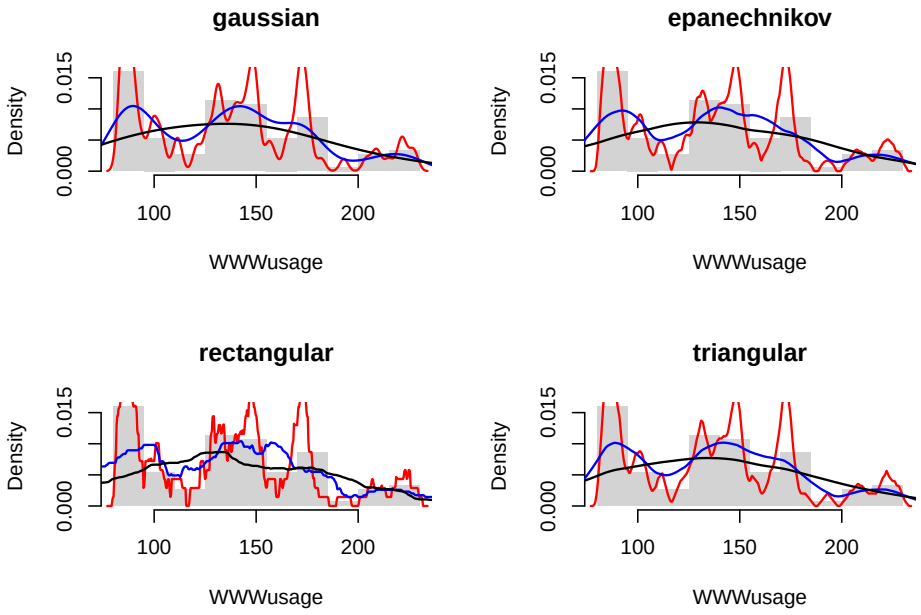
or can be directly checked

```
mean( (95<WWWusage) & (WWWusage <= 110))
[1] 0.08
```

(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.

Solution:

```
par(mfrow=c(2,2))
kernels = c("gaussian", "epanechnikov",
            "rectangular","triangular")
bandwidths = c(2, 10, 30)
colours = c("red", "blue", "black")
for (i in 1:4){
  hist(WWWusage,
       breaks = seq(80, 230, by = 15),
       prob = TRUE,
       main = kernels[i],
       border = FALSE)
  for (j in 1:3){
    lines(density(WWWusage,
                  kernel = kernels[i],
                  bw = bandwidths[j]),
                  col = colours[j], lwd = 1.5)
  }
}
```



A large bandwidth results in *oversmoothing* of the density estimate and will hide most of the data structure. A small bandwidth will *undersmooth* the density estimate, making it spiky and difficult to interpret.

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

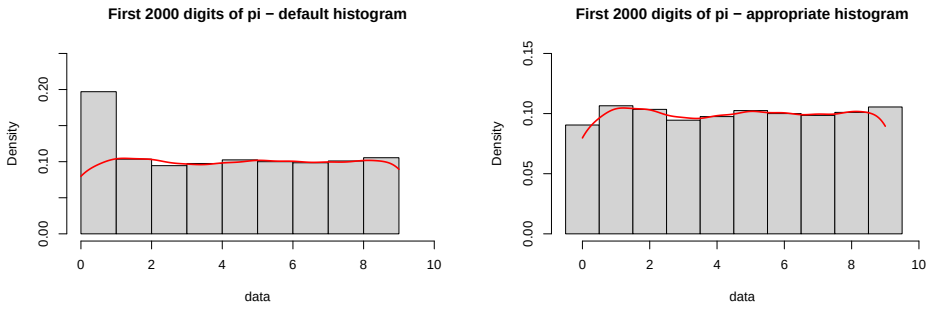
(a) Fit a density estimate to the dataset (use the command `density()`). Compare with the appropriate histogram. Why might you want to add an argument like `breaks=0:10-0.5` to `hist`?

Solution:

```
library(UsingR)
data <- pi2000
head(data)
[1] 3 1 4 1 5 9
```

```
 # default histogram
h1 < - hist(data, probability = TRUE, main = "First 2000 digits
of pi - default histogram", ylim = c(0,0.25), xlim = c(0,10))
lines(density(data,from = 0,to = 9), col = "red", lwd = 2)
```

```
 # appropriate histogram
h2 <- hist(data, probability = TRUE, main = "First 2000 digits
of pi - appropriate histogram", breaks=0:10-0.5, ylim = c(0,0.15),
xlim = c(-0.5,10))
lines(density(data,from = 0,to = 9), col = "red", lwd = 2)
```



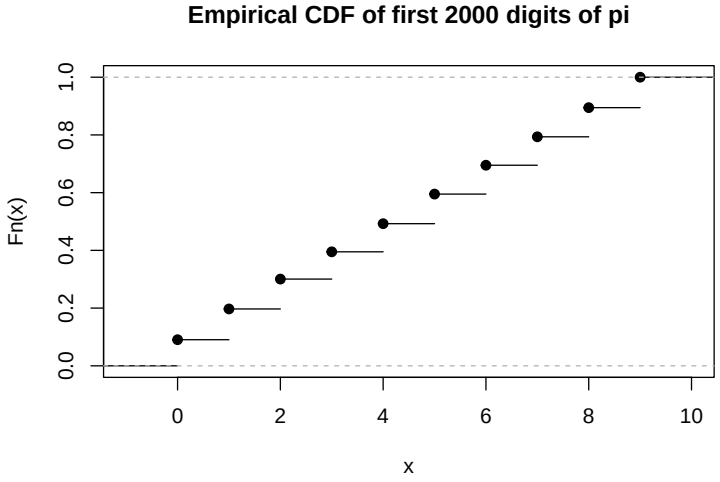

```
 h1$breaks
[1] 0 1 2 3 4 5 6 7 8 9
h2$breaks
[1] -0.5 0.5 1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5 9.5
h1$counts[1]
[1] 394
h2$counts[1]+h2$counts[2]
[1] 394
```

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

Solution:

```
table(data)
0   1   2   3   4   5   6   7   8   9
181 213 207 189 195 205 200 197 202 211
```

```
plot(ecdf(data), main = "Empirical CDF of first 2000 digits
of pi")
```



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

Solution: We expect a uniform distribution. Despite this observation, it is not known that π is a normal number! Determining if numbers are normal by mathematical proof is very difficult.