Part 1: Theory

(a) Let X be a continuous random variable with pdf $f(x)$ and let $Y = t(X)$, where t is a strictly monotone differentiable function. Let $g(y)$ be the pdf of Y. Show that $f(x) = g(t(x))t'(x)$. [Hint: Write the distribution function of X in terms of that of Y.]

(b) Consider the following estimator of a pdf $f(x)$:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right)$$

where the kernel, K, satisfies the condition $\int_{-\infty}^{\infty} K(z)dz = 1$. Show that

$$E(\hat{f}(x)) = \int_{-\infty}^{\infty}\frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy = \int_{-\infty}^{\infty}K(t)f(x-ht)dt\,.$$

Is $\hat{f}(x)$ an unbiased estimator of $f(x)$?

Part 2: Practical

To carry out the exercises that follow you need to use the functions `hist`, `set.seed`, `sample`, `par(mfrow=c( ))` and `density`. If you are not familiar with these then use the `help` command to learn how to use them.

(a) Import the "car expenditure" data and set up random samples from the population.

(i) Use `pop<-scan("D:/kursdaten_IntActDat/ps105.dat")` to read the data and then divide these population values by 1000 to convert the units from DM to 1000 DM.

(ii) Draw a random sample of size 100 from the population and call this `samp100`. (Use `set.seed(321)` and then `sample` to obtain the same sample of values as those that I will use in the solutions.)

(iii) Store the first 15 entries of `samp100` in a vector called `samp015`, the first 20 in a vector called `samp020` and the first 50 in a vector called `samp050`.

(b) The aim here is to compare the histogram and kernel density estimates for the above four samples:

(i) Use `par(mfrow=c(2,1))` to open a 2×1 graphics window and compare the histogram estimate of the pdf (using `hist()`) with the kernel density estimate (using `density()`) based on the sample `samp015`.

(ii) Repeat the above for the samples `samp020`, `samp050` and `samp100`.

(c) Investigate the effect of the **bin width**, and then of the **sample size**, on the behaviour of histogram estimates:

(i) Use `par(mfrow=c(2,2))` to open a 2×2 graphics window and then draw the histograms for the sample `samp015` using the bin widths 10, 5, 2 and 1.

(ii) Compare the the histograms for the samples `samp015`, `samp020`, `samp050`, `samp100`.

(d) Investigate the effect of each component of a kernel density estimator. Apply the function `density` to the sample `samp100` using

(i) different **bandwidths** (e.g. 0.1, 0.2, 0.5 and 1.0),

(ii) different **kernels** (read the helpfile for `density`),

(iii) different **number of arguments** at which $f(x)$ is estimated (e.g. 4, 8, 16 and 1024).

In each case describe the effect of changing the component.