Assignment sheet 7

November 1, 2017

The first three assignments are taken from last week.

Assignment 1. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a continuous distribution with continuous density function f. Let $h > 0$ and define a partition $\{I_j\}_{j\in\mathbb{Z}}$ of $\mathbb{R}$, where $I_j = [\kappa + (j-1)h, \kappa + jh)$ for a fixed real number κ.
The *histogram* of $X_1, X_2, \ldots, X_n$ with *bin-width* h and *origin* κ is defined as the function $x \mapsto \mathrm{Hist}_{X_1,X_2,\ldots,X_n}(x)$, where

$$\mathrm{Hist}_{X_1,X_2,\ldots,X_n}(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{1}(X_i \in I_j) \quad \text{if } x \in I_j$$

for each $j \in \mathbb{Z}$.
(a) Show that $\int_{-\infty}^{\infty} \mathrm{Hist}_{X_1,X_2,\ldots,X_n}(x)dx = 1$.
(b) Find the distribution of $nh\mathrm{Hist}_{X_1,X_2,\ldots,X_n}(x)$ for each x. Hence, find its mean and variance.
(c) What happens to $\mathbb{E}[\mathrm{Hist}_{X_1,X_2,\ldots,X_n}(x)]$ when $h \to 0$.
(Note : This limit indicates what $\mathrm{Hist}_{X_1,X_2,\ldots,X_n}(x)$ is estimating for each x for sufficiently small h.)
(d) Using part (b), find $\mathbb{E}\{[\mathrm{Hist}_{X_1,X_2,\ldots,X_n}(x) - f(x)]^2\}$.
(e) What happens to the value of the mean squared error in part (e) when $h \to 0$ and $nh \to \infty$?
(f) Interpret the limits $h \to 0$ and $nh \to \infty$.
(g) Is $\mathrm{Hist}_{X_1,X_2,\ldots,X_n}(x)$ a consistent estimator of $f(x)$? Justify your answer.

Assignment 2. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$ are unknown parameters. In this problem, we will find different types of confidence regions for (μ, σ^2).
(a) Find $100(1-\alpha)\%$ confidence interval for μ (when σ is known) and σ^2 (when μ is unknown) separately using $\overline{X}$ and S^2. Denote these by $R_{1,\alpha}(\mathbf{X})$ and $R_{2,\alpha}(\mathbf{X})$.
(b) Is the region $R_{1,\alpha}(\mathbf{X}) \times R_{2,\alpha}(\mathbf{X})$ a $100(1-\alpha)\%$ confidence region for (μ, σ^2) ? Otherwise, find β (depending on α) such that $R_{1,\beta}(\mathbf{X}) \times R_{2,\beta}(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence region for (μ, σ^2) using the Bonferroni method.
(c) Use the independence of $\overline{X}$ and S^2 to find β such that $R_{1,\beta}(\mathbf{X}) \times R_{2,\beta}(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence region for (μ, σ^2). Denote this region by $R_A(\mathbf{X})$.
(Note : $R_A(\mathbf{X})$ is called the Mood exact region.)
(d) Which one of the above confidence regions obtained in (b) and (c) do you think is preferable ? Justify your answer.
(e) Write down the likelihood ratio statistic for testing the hypothesis $H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$ vs $H_1 : \mu \neq \mu_0, \sigma^2 \neq \sigma_0^2$.
(f) Use Wilks' theorem and the expression of the likelihood ratio statistic to derive an asymptotic $100(1-\alpha)\%$ confidence region for (μ, σ^2). Denote this region by $R_B(\mathbf{X})$.
FACT : It can be proved that $\sqrt{n}\{(\overline{X}, S^2)^\top - (\mu, \sigma^2)^\top\}$ converges in distribution to $(Z_1, Z_2)^\top$, where $Z_1 \sim N(0, \sigma^2)$, $Z^2 \sim N(0, 2\sigma^4)$, and they are independent.
(g) Use the above fact to find the asymptotic distribution of $U_n = n(\overline{X} - \mu)^2/\sigma^2 + n(S^2 - \sigma^2)/(2\sigma^4)$.
(h) Use part (g) to find an asymptotic $100(1-\alpha)\%$ confidence region for (μ, σ^2). Denote this region by $R_C(\mathbf{X})$.
(i) What is the asymptotic distribution of $V_n = n(\overline{X} - \mu)^2/S^2 + n(S^2 - \sigma^2)/(2S^4)$?

(j) Use part (i) to find an asymptotic $100(1-\alpha)\%$ confidence region for (μ, σ^2). Denote this region by $R_D(\mathbf{X})$.
(k) Suppose that $n = 10$, $\overline{x} = 0$, $s^2 = 1$ and $\alpha = 0.05$. Write a code in `R` to understand how the above four 95% confidence regions, namely, $R_A(\mathbf{X}), R_B(\mathbf{X}), R_C(\mathbf{X})$ and $R_D(\mathbf{X})$ look like. (*Hint : Each confidence region will be a set of the form* $\{(\mu, \sigma^2) : H(\mu, \sigma^2) \leq h\}$*, where* H *is a real-valued function and* h *is a real number. You will get the function* H *after you simplify and put the values of* n*,* $\overline{x}$*,* s^2 *and* α*. To draw this set, you can use the following code*)

```
f <- function(a,b) H(a,b)
mu_vals <- seq(from=-1,to=1,length=100)
sig_vals <- seq(from=0.5,to=1.5,length=100)
z <- outer(mu_vals,sig_vals,f)
contour(mu_vals,sig_vals,z,levels=h,drawlabels=FALSE)
abline(h=1,v=0,col="red")
```

What do you observe ? What do you observe if you take $n = 25$ and $n = 100$?
(Note : It can be proved that the likelihood based region $R_B(\mathbf{X})$ asymptotically has the smallest expected area.)

Assignment 3. In this exercise we will explore how, in testing many hypotheses simultaneously, compiling a list on tests based on a small p-values cut-off might result in many false positives with high probability.

Mars Reconnaissance Orbiter is a NASA orbiter that aims to prove that water persisted on the surface on Mars for long period of times. Assume Europe sent its own orbiter to check for presence of liquid water on Mars. The orbiter will snap a picture at 100 pre-sampled location. Under the null, each sampled location is assumed to have a probability of 5% to host water. To have more certainty, the orbiter will gravitate around the planet until it has taken 200 pictures of each location.

(i). Run the following code and comment it.

```
set.seed(25102017)
positions <- 100;
trials <- 200;
true.p <- 0.05;
alpha <- 0.01;
nrep <- 1000


p <- matrix(replicate(positions*nrep,prop.test(rbinom(1,trials,true.p),
              trials,true.p)$p.value),nrep)
dim(p)
mean(apply(p,1,min)<alpha)
```

(ii). What happens if $\alpha = 0.05$? Comment the result.

(iii). Use the function `p.adjust` to adjust the p-values using Bonferroni, Holms and Hochberg's correction.

(iv). * What happens when you change the numbers ? For example, if trials "small" ? How could you overcome this ?

Assignment 4. This assignment pertains to confidence bands and histograms. Let $f : [A, B] \to [0, \infty)$ be a continuous density function, and $X_1, \ldots, X_n$ a sample from f. Let I be a histogram bin of length h, and recall that the histogram estimates (the restriction to I of) f by $(nh)^{-1} \sum_{j=1}^{n} \mathbf{1}\{X_j \in I\}$; this is the average number of sample points in I, normalised by the length of the interval (h).

(i). Invert a Wald test to construct an approximate $(1 - \alpha)$-confidence interval for $p_I = \mathbb{P}(X \in I)$ at level $\alpha \in (0, 1)$. *Hint : something very similar was done in Assignment 5(b).*

(ii). Suppose that h is small. Let $x = \inf I$ be the leftmost point of I. What is the (approximate) relation between $f(x)$ and p_I ?

(iii). Using the previous two points, construct an approximate confidence interval for $f(x)$. What happens if we chose another point $x \in I$?

(iv). What is the length of that interval ?

(v). We would now like to construct a simultaneous confidence band for the entire density f on $[A, B]$ using the histrogram. We need to correct for the multiple testing, so the first step is to understand : how many bins does the histogram contain ?

(vi). Let $I_1, I_2, \ldots, I_m$ denote the intervals corresponding to the histogram bins, and let $x = \inf I_1$. Using a Bonferroni correction, construct an approximate confidence region (product of intervals) that contains simulateneously all the density values $f(x), f(x + h), f(x + 2h), \ldots, f(x + mh)$.

Remark. Since the number of points in the bins are correlated, we cannot use the independence correction.

(vii). What is the length of each of these intervals ? Compare with the length in part (d). How does the length behave as $n \to \infty$ and $h \to 0$? *Hint : use the following fact (which you don't have to prove, but it is not hard). The $1 - q$ quantile of a χ^2_1 distribution behaves like $-2 \log q$ as $q \to 0$.*

Assignment 5. Chosing the bandwidth is a crucial step in KDE. In class you saw that it regulates the variance-bias tradeoff. In this exercise we are going to see this through a practical example.

The dataset `faithful` collects the duration of the eruptions and the waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

(i). Search for and download the dataset. Save the waiting times in a vector `x`.

(ii). Use the functions `plot` and `density` to plot an estimated density for `x`. Which is the default kernel used by `density` ?

(iii). Plot an histogram of `x` and overline the curve plotted by the `density` function.

(iv). Repeat the previous point for different kernels.

(v). Run the following command and comment

```
par(mfrow=c(1,1))
kernels <- eval(formals(density.default)$kernel)
plot(density(precip),main = "Different kernels, bw not selected")
for(i in 2:length(kernels))
  lines(density(precip,kern=kernels[i]),col=i)
legend("topright",legend=kernels,
       col=seq(kernels),lty=1).
```

(vi). By default, bandwidth selection is done with the normal reference rule, but can also be done manually. Select manually the bandwidth within `density`. For example, plot several estimated densities over the histograms with bandwidth varying from 1 to 10 and chose the most suitable one by eye.

(vii). Plot the chosen bandwidth agains the default one and the one picked by cross validation.

Assignment 6. In this exercise we are going to investigate how KDE might encour in problems at the border.

(i). Run the following code. Try to discuss what is happening.

```
curve(dexp(x), -0.5, 1, ylim=c(0,1.5))
n <- 10^5;
for (i in 1:100) lines(density(rexp(n)), col="red")
```

(ii). Run again the previous code for different values of n. Say

```
n <- c(10^3, 10^4, 10^5, 10^6)
```

.

(iii). For each value of n, use the function `density` to compute the value of the KDE of the exponential function at 0. Take several estimates (i.e. 100 or more). Write down function to compute the bias, the MSE and the variance. What is happening ?