PSM Fall 2016, Exercise Sheet 4

Return on Thursday October 13 in class

Note. You are encouraged to work in teams of two – but no larger. If you work in a team, submit only a single sheet with both names indicated on it. Nicely type-set solutions are highly appreciated.

Problem 1 (25 points) Give a formal description of the following statistical decision problems identifying the data recording procedure, the sample space (data value space), the statistical model $\mathcal{P}$ (including a convenient parametrisation of the statistical model), the decision space D and the loss function L. Make sensible assumptions for it, whenever the given information is not sufficient.

1. The Automobile Emissions Agency of the European Union is introducing a new system to check emissions of new cars. Emissions are recorded at three different speeds for both urban and highway traffic. For all these situations emission limits had been fixed in advance and the cars only get permissions if their emission results are below these limits. Since the commission feels that the loss is roughly like squared error $(d-\theta)^2$ when the true emission score θ is small but is like squared relative error $(\frac{d}{\theta}-1)^2$ when θ is larger, it decided to use loss function $\frac{(d-\theta)^2}{1+\theta^2}$ to reflect this behavior.
   DRP: measuring emissions at six different settings (three different speeds, both for urban and highway traffic)
   Sample space $S = \{x = (x_{u1}, x_{u2}, x_{u3}, x_{h1}, x_{h2}, x_{h3}) : x_{ab} \in \mathbb{R}^+\} \subseteq \mathbb{R}^6_+$.
   Statistical model $\mathcal{P} = \bigotimes P_{X_i}$, with $P_{X_i} \in \{P \text{ with } E[(X_i)] = \vartheta \in \mathbb{R}^+\}$.
   Parametrisation: $\Theta = \mathbb{R}^6_+$.
   Decision space: $D = \{\theta, 0\}$ with θ means emission test has been passed, i.e. $x \leq$ limit.
   Loss function: $L(\theta, d) = \frac{(d-\theta)^2}{1+\theta^2}$
2. An online store wants to learn about the effectiveness of unconscious advertisements. They select a homogeneous customer group and divide it into two parts: one being exposed to multiple short ads on their facebook page, the others not being exposed to them. At defined time points, both groups receive a visual stimulus refering to the product which the users can click on. Based on the number of clicks received by the two groups the effectiveness of the procedure will be assessed.

   We call the customers that are exposed to multiple short ads Group A, the others Group B.

   DRP: for each customer it is measured whether he/she clicks on the product. $X_i : (\Omega, \mathcal{A}, P) \to (\{0,1\}, Pot\{0,1\})$, then we add the responses up for each group. $X_A = \sum_{i=1}^n X_i^A, X_B = \sum_{i=1}^m X_i^B$ (we assume that there are n customers in Group A, and m customers in Group B).

   Sample spaces $S_A = \{0, 1, \ldots, n\}, S_B = \{0, 1, \ldots, m\}$.

   Statistical model $\mathcal{P} = \{(P_A, P_B) : \text{ with } P_A = Bi_{n,p}, P_B = Bi_{m,\tilde{p}}, 0 \leq p, \tilde{p} \leq 1\}$.

   Parametrisation: $\Theta = [0,1]^2$

Decision space: $D = \{-1, 1\}$ indicating whether we decide that p is smaller than $\tilde{p}$ (-1) or not (1).

Loss function: $L(\theta, d) = ((p - \tilde{p}) - d)^2$.

Problem 2 (30 points) Which of the following parametrisations are identifiable, i.e. they hold that from $P_{\theta_1} = P_{\theta_2}$ follows $\theta_1 = \theta_2$? (Prove or disprove!)

1. $X_1, \ldots, X_p$ are independent with $X_i \sim N(\alpha_i + \nu, \sigma^2)$.

   $$\theta = (\alpha_1, \alpha_2, \ldots, \alpha_p, \nu, \sigma^2)$$

   and P_θ is the distribution of $X = (X_1, \ldots, X_p)$.
   Let $\theta_1 = (\alpha_1, \alpha_2, \ldots, \alpha_p, \nu, \sigma^2) \neq \theta_2 = (\alpha_1 + \nu, \alpha_2 + \nu, \ldots, \alpha_p + \nu, 0, \sigma^2)$ with $\nu \neq 0$. Then $P_{\theta_1} = N(\alpha_i + \nu, \sigma^2) = P_{\theta_2}$. Hence this parametrisation is not identifiable.
2. Same as **??** with $\alpha = (\alpha_1, \ldots, \alpha_p)$ restricted to

   $$\{(\alpha_1, \ldots, \alpha_p) : \sum_{i=1}^{p} \alpha_i = 0\}.$$

   Assume there are two parametrisations $\theta_1 = (\alpha_1, \alpha_2, \ldots, \alpha_p, \nu, \sigma^2) \neq \theta_2 = (\beta_1, \beta_2, \ldots, beta_p, \tilde{\nu}, \tilde{\sigma}^2)$ such that $\sum_{i=1}^{p} \alpha_i = 0$ and $\sum_{i=1}^{p} \beta_i = 0$ and $P_{\theta_1} = P_{\theta_2}$, i.e we have for $i = 1, \ldots, p$: $N(\alpha_i + \nu, \sigma^2) = N(\beta_i + \tilde{\nu}, \tilde{\sigma}^2)$. Since the distributions can only be the same, once their variances are the same, it follows that $\tilde{\sigma}^2 = \sigma^2$. Moreover, the means of the distributions must be identical, i.e. $\alpha_i + \nu = \beta_i + \tilde{\nu}$. Summing both sides of the equation over all $i = 1 \ldots, p$ yields

   $$\begin{aligned} \sum_{i=1}^{p} (\alpha_i + \nu) &= \sum_{i=1}^{p} (\beta_i + \tilde{\nu}) \\ \Rightarrow 0 + p\nu &= 0 + p * \tilde{\nu} \\ \Rightarrow \nu &= \tilde{\nu}. \end{aligned}$$

   Hence, in order for the two distributions to be the same we need $\alpha_i + \nu = \beta_i + \nu$ for all $i = 1, \ldots, p$ which results in $\alpha_i = \beta_i$ for all $i = 1, \ldots, p$.
   Hence the two parameters θ_1 and θ_2 are actually identical and consequently, this parametrisation is identifiable.
3. X and Y are independent $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, $\theta = (\mu_1, \mu_2)$ and we observe $Y - X$.
   Let $\theta_1 = (\mu_1 + a, \mu_2 + a), \theta_2 = (\mu_1, \mu_2)$ with $a \neq 0$. Then $\theta_1 \neq \theta_2$ but $P_{\theta_1}(Y - X) = N(\mu_2 + a - (\mu_1 + a), 2\sigma^2) = N(\mu_2 - \mu_1, 2\sigma^2) = P_{\theta_2}(Y - X)$. Hence, this parametrisation is not identifiable.
4. $X_{ij}, i = 1, \ldots, p; j = 1, \ldots, b$ are independent with $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ where $\mu_{ij} = \nu + \alpha_i + \lambda_j$, $\theta = (\alpha_1, \ldots, \alpha_p, \lambda_1, \ldots, \lambda_b, \nu, \sigma^2)$ and P_θ is the joint distribution of $X_{11}, \ldots, X_{pb}$.
5. Same as **??** with $\alpha = (\alpha_1, \ldots, \alpha_p)$ and $\lambda = (\lambda_1, \ldots, \lambda_b)$ restricted to

   $$\{(\alpha_1, \ldots, \alpha_p) : \sum_{i=1}^{p} \alpha_i = 0\} \qquad \{(\lambda_1, \ldots, \lambda_p) : \sum_{i=1}^{p} \lambda_i = 0\}.$$

Problem 3 (20 points) The cumulative distribution function for the random variable X is given by

$$F(x) = \begin{cases} 0, & \text{for } x \leq 0, \\ \frac{x}{4}, & \text{for } 0 \leq x \leq 2, \\ \frac{x^2}{8}, & \text{for } 2 \leq x \leq \sqrt{8}, \\ 1, & \text{for } 4 \leq x. \end{cases}$$

1. Find and graph $f(x)$.
   The PDF for X is:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{1}{4} & \text{if } 0 \leq x \leq 2 \\ \frac{x}{4} & \text{if } 2 \leq x \leq \sqrt{8} \\ 0 & \text{if } \sqrt{8} \leq x \end{cases}$$

```
f <- function(x){
  firstInds<-intersect(which(x >= 0), which(x < 2))
  secInds<-intersect(which(x >= 2), which(x < sqrt(8)))
  y <- x
  y[firstInds] <- 1/4
  y[secInds] <- x[secInds]/4
  y[-c(firstInds, secInds)]<-0
  y
}
x<-seq(-2,4,by=0.001)
pdf<-f(x)
data<-data.frame(cbind(x,pdf))
plot(data,cex=0.1, main="PDF of X", xlim=c(-1.5,4))
segments(0,0,0,1/4)
segments(2,1/4,2,1/2)
segments(sqrt(8),0,sqrt(8),sqrt(8)/4)
```

2. Find $E[X]$ and $STD[X]$.
   The expected value is

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x) \mathrm{d}x \\ &= \int_{0}^{4} x f(x) \mathrm{d}x \\ &= \int_{0}^{2} x \frac{1}{4} \mathrm{d}x + \int_{2}^{\sqrt{8}} x \frac{x}{4} \mathrm{d}x \\ &= \frac{x^2}{8}\Big|_0^2 + \frac{x^3}{12}\Big|_2^{\sqrt{8}} \\ &= \frac{1}{2} + \frac{2}{3}\sqrt{8} - \frac{2}{3} \\ &\approx 1.719. \end{aligned}$$

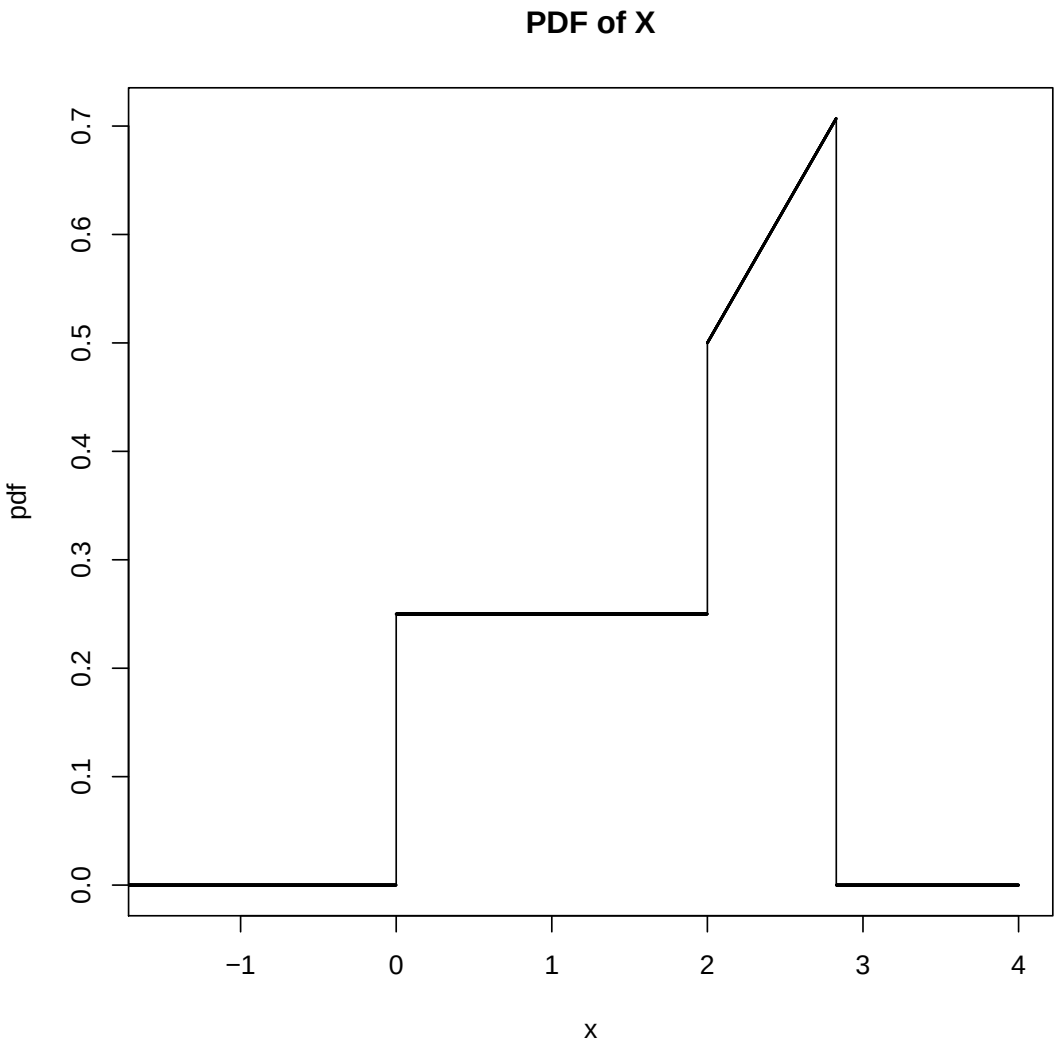

Figure 1: Graph of pdf.

To compute $STD[X]$ we first compute the variance via $VAR[X] = E[X^2] - (E[X])^2$.

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x \\ &= \int_0^4 x^2 f(x)\mathrm{d}x \\ &= \int_0^2 x^2 \frac{1}{4}\mathrm{d}x + \int_2^{\sqrt{8}} x^2 \frac{x}{4}\mathrm{d}x \\ &= \frac{x^3}{12}\Big|_0^2 + \frac{x^4}{16}\Big|_2^{\sqrt{8}} \\ &= \frac{2}{3} + 4 - 1 = \frac{11}{3} \end{aligned}$$

$$\begin{aligned} VAR[X] &= E[X^2] - (E[X])^2 \\ &= \frac{11}{3} - \left(\frac{2}{3}\sqrt{8} - \frac{1}{6}\right)^2 \\ &= \frac{11}{3} - \left(\frac{4}{9}\cdot 8 - 2\frac{2}{3}\frac{1}{6}\sqrt{8} + \frac{1}{36}\right) \\ &= \frac{11}{3} - \left(\frac{43}{12} - \frac{2}{9}\sqrt{8}\right) \\ &= \frac{1}{12} + \frac{2}{9}\sqrt{8} \\ &\approx 0.7119. \end{aligned}$$

Then we have

$$\begin{aligned} STD[X] &= \sqrt{VAR[X]} = \sqrt{\frac{4}{9}\sqrt{8}} \\ &\approx 0.8437. \end{aligned}$$

3. Simulate 5000 values for X and find the sample average and sample standard deviation. Repeat this procedure 10 times. Compare the results from the simulation with the theoretical values obtained in part b). For the simulation use the following algorithm (*cumulative distribution function method*):
   - Obtain a random number u with $0 \leq u \leq 1$.
   - Set $u = F(x)$ for the given CDF F of X.
   - Solve for x
   - Repeat until you have the required number of simulations.

```
F.inv <- function(y){
  firstInds<-intersect(which(y >= 0), which(y < 1/2))
  secInds<-intersect(which(y >= 1/2), which(y <= 1))
  x <- y
  x[firstInds] <- 4*y[firstInds]
```

```
  x[secInds] <- sqrt(8*y[secInds])
  x[-c(firstInds, secInds)]<-0
  x
}
sample.means <- 0
sample.stds <- 0
for (i in 1:10){
z<-runif(5000)
sample5000 <- F.inv(z)
sample.means[i] <- mean(sample5000)
sample.stds[i] <- sd(sample5000)
}
cbind(sample.means,sample.stds)
```

```
##       sample.means sample.stds
##  [1,]     1.708627   0.8446512
##  [2,]     1.716735   0.8446312
##  [3,]     1.723311   0.8495704
##  [4,]     1.708402   0.8474581
##  [5,]     1.704981   0.8496538
##  [6,]     1.723827   0.8405949
##  [7,]     1.717352   0.8410855
##  [8,]     1.718497   0.8417824
##  [9,]     1.729247   0.8479120
## [10,]     1.707888   0.8444259
```

```
mean(sample.means)
```

```
## [1] 1.715887
```

```
mean(sample.stds)
```

```
## [1] 0.8451765
```

Problem 4 (15 points) Let X_1 and X_2 have the joint pdf

$$f_{X_1,X_2}(x_1,x_2) = 2e^{-x_1-x_2} I_{(0,x_2)}(x1) I_{(0,\infty)}(x2),$$

where $I_{(a,b)}(x) = \begin{cases} 1 & \text{for } a < x < b \\ 0 & \text{elsewhere} \end{cases}$.

1. Find the marginal pdfs for X_1 and X_2.
2. Are X_1 and X_2 independent?
3. Suppose that $Y_1 = 2X_1$ and $Y_2 = X_2 - X_1$. Show that Y_1 and Y_2 are independent.

Problem 5 (10 points)

The temperature X (measured in degrees Fahrenheit) at a randomly selected point in a commercial refrigerator is a random variable with PDF

$$f(x) = \begin{cases} \frac{(x-32)^2}{1944}, & \text{for } 32 \le x \le 50 \\ 0, & \text{otherwise} \end{cases}$$

1. Find $E[X]$ and $STD[X]$.

```
f <- function(x){x*(x-32)^2/1944}
f2 <- function(x){x^2*(x-32)^2/1944}
expv<-integrate(f,32,50)$value
expx2 <- integrate(f2,32,50)$value
std<-sqrt(expx2-expv^2)
```

The expected value is

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x) \mathrm{d}x \\ &= \int_{32}^{50} x f(x) \mathrm{d}x \\ &= \int_{32}^{50} x \frac{(x-32)^2}{1944} \mathrm{d}x \\ &\approx 45.5. \end{aligned}$$

To compute $STD[X]$ we first compute the variance via $VAR[X] = E[X^2] - (E[X])^2$.

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x \\ &= \int_{32}^{50} x^2 f(x) \mathrm{d}x \\ &= \int_{32}^{50} x^2 \frac{(x-32)^2}{1944} \mathrm{d}x \\ &\approx 2082.4 \end{aligned}$$

$$\begin{aligned} VAR[X] &= E[X^2] - (E[X])^2 \\ &= 2082.4 - 45.4^2 \\ &\approx 12.15. \end{aligned}$$

Then we have

$$\begin{aligned} STD[X] &= \sqrt{VAR[X]} \\ &\approx 3.4857. \end{aligned}$$

2. Let Y be the temperature in degrees Celsius. Find $E[Y]$ and $STD[Y]$.
By $Y = \frac{5}{9}(X - 32)$ we have

$$\begin{aligned} E[Y] &= \frac{5}{9}(E[X] - 32) \\ &\approx 7.5 \end{aligned}$$

and

$$\begin{aligned} STD[Y] &= \frac{5}{9} STD[X] \\ &\approx 1.94 \end{aligned}$$