Midterm practice problems

Solutions

1. (Cumulative distribution function)

   (a) The median is 0.2, because the probability that $\tilde{x} \leq 0.2$ is 0.5.

   (b) The slope of the cdf is clearly larger at 0.2, so the density at 0.2 is higher than at 0.8.

   (c)

$$
\begin{align}
\mathrm{P}(\tilde{x} \leq 0.4 \,|\, \tilde{x} < 0.2) &= \frac{\mathrm{P}(0.2 \leq \tilde{x} \leq 0.4 \,|\, \tilde{x} \geq 0.2)}{\mathrm{P}(\tilde{x} \geq 0.2)} \tag{1} \\
&= \frac{F_{\tilde{x}}(0.4) - F_{\tilde{x}}(0.2)}{1 - F_{\tilde{x}}(0.2)} \tag{2} \\
&= \frac{0.8 - 0.5}{0.5} \tag{3} \\
&= 0.6. \tag{4}
\end{align}
$$

   (d) We apply inverse-transform sampling. Since $F_{\tilde{x}}(0.08) = 0.2$, $F_{\tilde{x}}(0.15) = 0.4$, and $F_{\tilde{x}}(0.25) = 0.55$, the simulated samples are 0.08, 0.15, and 0.25.

2. (Fire alarm)

   (a)

$$
\begin{align}
p_{\tilde{a}}(1) &= \sum_{x \in \{0,1\}} \sum_{b \in \{0,1\}} p_{\tilde{a},\tilde{x},\tilde{b}}(1, x, b) \tag{5} \\
&= p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,0) + p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,1) + p_{\tilde{a},\tilde{x},\tilde{b}}(1,1,0) + p_{\tilde{a},\tilde{x},\tilde{b}}(1,1,1) \tag{6} \\
&= p_{\tilde{x}}(0) p_{\tilde{b}}(0) p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1 \,|\, 0, 0) + p_{\tilde{x}}(0) p_{\tilde{b}}(1) p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1 \,|\, 0, 1) \tag{7} \\
&\quad + p_{\tilde{x}}(1) p_{\tilde{b}}(0) p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1 \,|\, 1, 0) + p_{\tilde{x}}(1) p_{\tilde{b}}(1) p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1 \,|\, 1, 1) \tag{8} \\
&= 0 + 0.8 \cdot 0.5 \cdot 0.2 + 0.2 \cdot 0.5 \cdot 0.8 + 0.2 \cdot 0.5 \cdot 0.4 \tag{9} \\
&= 0.2. \tag{10}
\end{align}
$$

   (b)

$$
\begin{align}
p_{\tilde{x},\tilde{b}\,|\,\tilde{a}}(0, 0 \,|\, 1) &= \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,0)}{p_{\tilde{a}}(1)} \tag{11} \\
&= \frac{p_{\tilde{x}}(0) p_{\tilde{b}}(0) p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1 \,|\, 0, 0)}{p_{\tilde{a}}(1)} \tag{12} \\
&= 0, \tag{13}
\end{align}
$$

$$p_{\tilde{x},\tilde{b}\,|\,\tilde{a}}(0,1\,|\,1) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,1)}{p_{\tilde{a}}(1)} \tag{14}$$
$$= \frac{p_{\tilde{x}}(0)p_{\tilde{b}}(1)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,0,1)}{p_{\tilde{a}}(1)} \tag{15}$$
$$= \frac{0.8 \cdot 0.5 \cdot 0.2}{0.2} \tag{16}$$
$$= 0.4, \tag{17}$$

$$p_{\tilde{x},\tilde{b}\,|\,\tilde{a}}(1,0\,|\,1) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,1,0)}{p_{\tilde{a}}(1)} \tag{18}$$
$$= \frac{p_{\tilde{x}}(1)p_{\tilde{b}}(0)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,1,0)}{p_{\tilde{a}}(1)} \tag{19}$$
$$= \frac{0.2 \cdot 0.5 \cdot 0.8}{0.2} \tag{20}$$
$$= 0.4, \tag{21}$$

$$p_{\tilde{x},\tilde{b}\,|\,\tilde{a}}(1,1\,|\,1) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,1,1)}{p_{\tilde{a}}(1)} \tag{22}$$
$$= \frac{p_{\tilde{x}}(1)p_{\tilde{b}}(1)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,1,1)}{p_{\tilde{a}}(1)} \tag{23}$$
$$= \frac{0.2 \cdot 0.5 \cdot 0.4}{0.2} \tag{24}$$
$$= 0.2. \tag{25}$$

(c) We compare $p_{\tilde{x}\,|\,\tilde{a},\tilde{b}}(0\,|\,1,0)$ and $p_{\tilde{x}\,|\,\tilde{a}}(0\,|\,1)$,

$$p_{\tilde{x}\,|\,\tilde{a},\tilde{b}}(0\,|\,1,0) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,0)}{p_{\tilde{a},\tilde{b}}(1,0)} \tag{26}$$
$$= \frac{p_{\tilde{x}}(0)p_{\tilde{b}}(0)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,0,0)}{\sum_{x=0}^{1} p_{\tilde{x}}(x)p_{\tilde{b}}(0)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,x,0)} \tag{27}$$
$$= 0, \tag{28}$$
$$p_{\tilde{x}\,|\,\tilde{a}}(0\,|\,1) = \frac{p_{\tilde{a},\tilde{x}}(1,0)}{p_{\tilde{a}}(1)} \tag{29}$$
$$= \frac{\sum_{b=0}^{1} p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,b)}{p_{\tilde{a}}(1)} \tag{30}$$
$$= \frac{\sum_{b=0}^{1} p_{\tilde{x}}(0)p_{\tilde{b}}(b)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,0,b)}{p_{\tilde{a}}(1)} \tag{31}$$
$$= \frac{0.8 \cdot 0.5 \cdot 0.2}{0.2} = 0.4 \neq 0, \tag{32}$$

which implies that $\tilde{x}$ and $\tilde{b}$ are not conditionally independent given $\tilde{a}$.

Intuitively, if we know that the alarm is ringing, then the state of the battery does provide information about fire. For example, if the battery is fine, then the alarm is not ringing for that reason, so it is more likely that there is a fire.

3. (Rat)

(a) The pdf needs to be nonnegative, which requires $0 \leq \alpha \leq 1$. We also need the pdf to integrate to one, which it does if α is in that range:

$$\int_{a=0}^{1} f_{\tilde{a}}(a)\,\mathrm{d}a = \int_{a=0}^{0.5} 2\alpha\,\mathrm{d}a + \int_{a=0.5}^{1} 2(1-\alpha)\,\mathrm{d}a \tag{33}$$

$$= 2\alpha \cdot 0.5 + 2(1-\alpha)\cdot 0.5 \tag{34}$$

$$= 1. \tag{35}$$

(b) Expressing the pdf as a function of α the likelihood of each data point is equal to 2α if the point is between 0 and 0.5, and to $2(1-\alpha)$ if it is between 0.5 and 1. Let n be the number of data, $n_{[0,0.5]}$ the number of data between 0 and 0.5, and $n_{(0.5,1]}$ the number of points between 0.5 and 1. We have,

$$\log \mathcal{L}(X) = \sum_{i=1}^{n} \log f_\alpha(x_i) \tag{36}$$

$$= n_{[0,0.5]} \log 2\alpha + n_{(0.5,1]} \log(2(1-\alpha)). \tag{37}$$

The first and second derivatives of the log likelihood equal

$$(\log \mathcal{L}(X))' = \frac{n_{[0,0.5]}}{\alpha} - \frac{n_{(0.5,1]}}{1-\alpha}, \tag{38}$$

$$(\log \mathcal{L}(X))'' = -\frac{n_{[0,0.5]}}{\alpha^2} - \frac{n_{(0.5,1]}}{(1-\alpha)^2}. \tag{39}$$

The function is concave, so we can set the first derivative to zero to find the ML estimate, it equals

$$\alpha_{ML} = \frac{n_{[0,0.5]}}{n} \tag{40}$$

$$= \frac{2}{5}. \tag{41}$$

(c) The probability of the rat being in the first half is

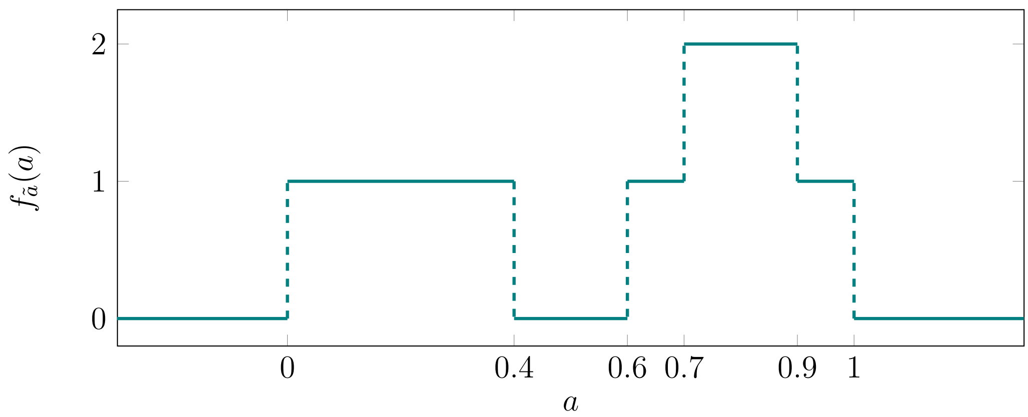
$$\mathrm{P}(0 \leq \tilde{a} \leq 0.5) = \int_{a=0}^{0.5} 2\alpha\,\mathrm{d}a \tag{42}$$

$$= \alpha. \tag{43}$$

Estimating this probability using empirical probabilities yields exactly the same estimate for α as the ML estimate,

$$\alpha_{\hat{ML}} = \frac{n_{[0,0.5]}}{n}. \tag{44}$$

(d) The estimated pdf is:



What is problematic is that the density is zero between 0.4 and 0.6, which seems an artifact of the limited number of data. This can be alleviated by increasing the width of the rectangular kernel.

4. (Election)

(a)

$$\begin{aligned} p_{\tilde{o}}(1) &= \mathrm{P}(\text{A wins all states}) + \sum_{i=1}^{3} \mathrm{P}(\text{A wins all states except } i) && (45)\\ &= 0.6^3 + 3 \cdot 0.4 \cdot 0.6^2 && (46)\\ &= 0.648. && (47) \end{aligned}$$

(b)

$$\begin{aligned} p_{\tilde{s}_2 \mid \tilde{o}}(0 \mid 0) &= \frac{p_{\tilde{s}_2,\tilde{o}}(0,0)}{p_{\tilde{o}}(0)} && (48)\\ &= \frac{\sum_{s_1=0}^{1} \sum_{s_3=0}^{1} p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3,\tilde{o}}(s_1,0,s_3,0)}{p_{\tilde{o}}(0)} && (49)\\ &= \frac{p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3,\tilde{o}}(0,0,0,0) + p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3,\tilde{o}}(1,0,0,0) + p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3,\tilde{o}}(0,0,1,0)}{p_{\tilde{o}}(0)} && (50)\\ &= \frac{p_{\tilde{s}_1}(0)p_{\tilde{s}_2}(0)p_{\tilde{s}_3}(0) + p_{\tilde{s}_1}(1)p_{\tilde{s}_2}(0)p_{\tilde{s}_3}(0) + p_{\tilde{s}_1}(0)p_{\tilde{s}_2}(0)p_{\tilde{s}_3}(1)}{p_{\tilde{o}}(0)} && (51)\\ &= \frac{0.4^3 + 2 \cdot 0.4^2 \cdot 0.6}{1 - 0.648} && (52)\\ &= 0.727. && (53) \end{aligned}$$

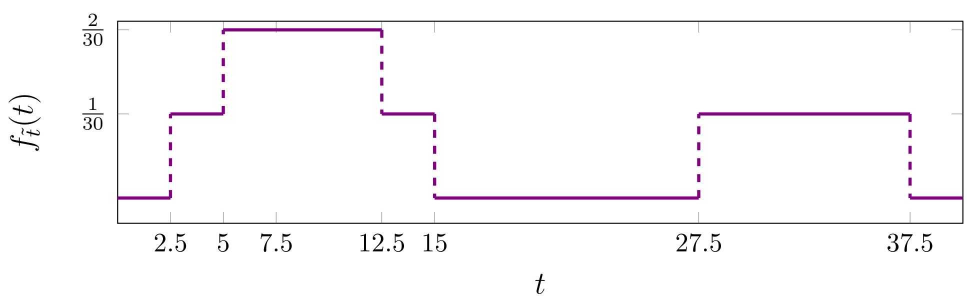
(c) We have

$$
\begin{aligned}
& p_{\tilde{s}_1 \,|\, \tilde{o}}(1 \,|\, 1) && (54)\\
&= \frac{p_{\tilde{s}_1,\tilde{o}}(1,1)}{p_{\tilde{o}}(1)} && (55)\\
&= \frac{\mathrm{P}(\text{A wins all states}) + \mathrm{P}(\text{A wins 1 and 2 but not 3}) + \mathrm{P}(\text{A wins 1 and 3 but not 2})}{p_{\tilde{o}}(1)}\\
&= \frac{0.6^3 + 2 \cdot 0.4 \cdot 0.6^2}{0.648} && (56)\\
&= 0.778. && (57)
\end{aligned}
$$

However, $p_{\tilde{s}_1 \,|\, \tilde{o},\tilde{s}_2}(1 \,|\, 1, 0) = 1$ because if A wins the election then they cannot lose state 1 and state 2. Intuitively, even if the state results are independent, they both determine the result of the election, so revealing who wins the election *connects* them. For example, if candidate A has won the election but lost state 2, then this completely determines the result of state 1.

5. (Earthquake)

(a) The KDE estimate looks like this:



The probability is equal to

$$
\begin{aligned}
\mathrm{P}(\tilde{t} > 10) &= \frac{2 \cdot 2.5}{30} + \frac{2.5}{30} + \frac{10}{30} && (58)\\
&= \frac{17.5}{30} && (59)\\
&= 0.583, && (60)
\end{aligned}
$$

which is the area under the pdf to the right of 10.

(b) From the notes, the ML estimate for the parameter of the exponential is

$$
\begin{aligned}
\lambda_{\mathrm{ML}} &= \frac{3}{7.5 + 10 + 32.5} && (61)\\
&= \frac{3}{50}. && (62)
\end{aligned}
$$

The probability equals

$$
\begin{aligned}
\mathrm{P}(\tilde{t} > 10) &= \int_{10}^{\infty} f_{\tilde{t}}(t)\,\mathrm{d}t && (63)\\
&= \int_{10}^{\infty} \frac{3\exp(-3t/50)}{50}\,\mathrm{d}t && (64)\\
&= -\exp(-3t/50)]_{10}^{\infty} && (65)\\
&= 0.549. && (66)
\end{aligned}
$$

(c) The parametric method requires less data but makes a stronger assumption about the distribution. The nonparametric method is more flexible but requires more data.

6. (Scaling random variables)

(a) False. Recall that the cdf of $\tilde{a}$ equals $F_{\tilde{a}}(a) = 1 - \exp(-\lambda a)$ for $a \geq 0$ and 0 for $a < 0$. The cdf of $\tilde{b}$ equals

$$
\begin{aligned}
F_{\tilde{b}}(b) &:= \mathrm{P}(\tilde{b} \leq b) && (67)\\
&= \mathrm{P}(\alpha\tilde{a} \leq b) && (68)\\
&= \mathrm{P}\left(\tilde{a} \leq \frac{b}{\alpha}\right) && (69)\\
&= F_{\tilde{a}}\left(\frac{b}{\alpha}\right) && (70)\\
&= 1 - \exp\left(-\frac{\lambda}{\alpha} b\right), && (71)
\end{aligned}
$$

so $\tilde{b}$ is actually an exponential random variable with parameter λ/α.

(b) False. If $\alpha = 2$ then $\tilde{b}$ only takes even values, and therefore is not a geometric random variable.

(c) True. If the median of a $\tilde{a}$ is m, then

$$
\mathrm{P}(\tilde{a} \leq m) = \frac{1}{2}, \qquad (72)
$$

which implies

$$
\begin{aligned}
\mathrm{P}(\tilde{b} \leq \alpha m + \beta) &= \mathrm{P}(\alpha\tilde{a} + \beta \leq \alpha m + \beta) && (73)\\
&= \mathrm{P}(\tilde{a} \leq m). && (74)
\end{aligned}
$$

7. (Baby)

(a) We define Bernoulli random variables $\tilde{s}$, $\tilde{b}$ and $\tilde{m}$ to represent the sex of the baby, what Bob thinks, and what Margaret thinks respectively (1=girl, 0=boy).

$$
\begin{aligned}
p_{\tilde{m},\tilde{b}}(1,1) &= p_{\tilde{s},\tilde{m},\tilde{b}}(0,1,1) + p_{\tilde{s},\tilde{m},\tilde{b}}(1,1,1) && (75)\\
&= 0.5 \cdot 0.2 \cdot 0.9 + 0.5 \cdot 0.8 \cdot 0.9 && (76)\\
&= 0.45. && (77)
\end{aligned}
$$

(b)

$$p_{\tilde{s}\,|\,\tilde{b}}(1\,|\,1) = \frac{p_{\tilde{s},\tilde{b}}(1,1)}{p_{\tilde{b}}(1)} \tag{78}$$

$$= \frac{p_{\tilde{s},\tilde{m},\tilde{b}}(1,0,1) + p_{\tilde{s},\tilde{m},\tilde{b}}(1,1,1)}{p_{\tilde{s},\tilde{m},\tilde{b}}(0,0,1) + p_{\tilde{s},\tilde{m},\tilde{b}}(0,1,1) + p_{\tilde{s},\tilde{m},\tilde{b}}(1,0,1) + p_{\tilde{s},\tilde{m},\tilde{b}}(1,1,1)} \tag{79}$$

$$= \frac{0.5 \cdot 0.2 \cdot 0.1 + 0.5 \cdot 0.8 \cdot 0.9}{0.5 \cdot 0.8 \cdot 0.1 + 0.5 \cdot 0.2 \cdot 0.9 + 0.5 \cdot 0.2 \cdot 0.1 + 0.5 \cdot 0.8 \cdot 0.9} \tag{80}$$

$$= 0.74. \tag{81}$$

(c) Bob's beliefs only depends on what Margaret tells him, so they are conditionally independent from the sex of the baby given what she says, for $m \in \{0,1\}$

$$p_{\tilde{b}\,|\,\tilde{m}}(b\,|\,m) = \begin{cases} 0.9 & \text{if } b = m, \\ 0.1 & \text{if } b \neq m, \end{cases} \tag{82}$$

and for $s \in \{0,1\}$

$$p_{\tilde{b}\,|\,\tilde{m},\tilde{s}}(b\,|\,m,s) = \begin{cases} 0.9 & \text{if } b = m, \\ 0.1 & \text{if } b \neq m. \end{cases} \tag{83}$$

(d) The probability that the baby is a girl is 0.5, which is different from $p_{\tilde{s}\,|\,\tilde{b}}(1\,|\,1)$, so what Bob thinks is not independent from the sex.

8. (Missing data)

(a) No, otherwise there wouldn't be such a large discrepancy between the number of missing data for men and women.

(b) 70 out of 140 observed cases have side effects, so the probability is 0.5.

(c) Let $\tilde{\text{se}}$, $\tilde{o}$ and $\tilde{s}$ be Bernoulli random variables representing the side effects, whether the data are observed ($\tilde{o} = 1$) or not ($\tilde{o} = 1$) and the sex respectively. We have

$$p_{\tilde{\text{se}}}(1) = \sum_{s \in \{\text{man},\text{woman}\}} p_{\tilde{s}}(s) p_{\tilde{\text{se}}\,|\,\tilde{s}}(1\,|\,s). \tag{84}$$

If the side effects are conditionally independent from the data being observed or missing given the patient sex, then

$$p_{\tilde{\text{se}}\,|\,\tilde{s}}(1\,|\,\text{man}) = p_{\tilde{\text{se}}\,|\,\tilde{o},\tilde{s}}(1\,|\,1,\text{man}) \tag{85}$$

$$= \frac{40}{50} = 0.8, \tag{86}$$

$$p_{\tilde{\text{se}}\,|\,\tilde{s}}(1\,|\,\text{woman}) = p_{\tilde{\text{se}}\,|\,\tilde{o},\tilde{s}}(1\,|\,1,\text{woman}) \tag{87}$$

$$= \frac{30}{90} = 0.33. \tag{88}$$

Consequently, since $p_{\tilde{s}}(\text{man}) = p_{\tilde{s}}(\text{woman}) = \frac{1}{2}$,

$$p_{\tilde{\text{se}}}(1) = \frac{0.8 + 0.33}{2} = 0.57. \tag{89}$$