Probabilistic inference for data science 2

Exercise 1

1. (a) $\mathbb{E}(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\mathbb{E}(\boldsymbol{X}) = 10 - 30 + 100 = 80$

   (b) $\mathbb{E}(\boldsymbol{X}+\boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{Y}) = \begin{bmatrix} 12 \\ 17 \\ 22 \end{bmatrix}$

   (c) $var(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T Var(\boldsymbol{X})\boldsymbol{a} = 500$

   (d)

$$
\begin{aligned}
var(\boldsymbol{X}+\boldsymbol{Y}) &= cov(\boldsymbol{X}+\boldsymbol{Y}, \boldsymbol{X}+\boldsymbol{Y}) \\
&= cov(\boldsymbol{X}, \boldsymbol{X}+\boldsymbol{Y}) + cov(\boldsymbol{Y}, \boldsymbol{X}+\boldsymbol{Y}) \\
&= cov(\boldsymbol{X}, \boldsymbol{X}) + cov(\boldsymbol{X}, \boldsymbol{Y}) + cov(\boldsymbol{Y}, \boldsymbol{X}) + cov(\boldsymbol{Y}, \boldsymbol{Y}) \\
&= \begin{bmatrix} 11 & 5.5 & -4.5 \\ 5.5 & 16 & 0.5 \\ -4.5 & 0.5 & 21 \end{bmatrix}
\end{aligned}
$$

   (e) $\mathbb{E}(\boldsymbol{a}^T(\boldsymbol{X}+\boldsymbol{Y})) = \boldsymbol{a}^T\mathbb{E}(\boldsymbol{X}+\boldsymbol{Y}) = 88$

   (f) $var(\boldsymbol{a}^T(\boldsymbol{X}+\boldsymbol{Y})) = \boldsymbol{a}^T var(\boldsymbol{X}+\boldsymbol{Y})\boldsymbol{a} = 523$

   Evaluation: Each correct item gives 16p. All correct 100p

2. (a)

$$
\begin{aligned}
&\mathbb{P}(X(u_1) > 1 \cap X(u_2) > 1 \cap X(u_3) > 1) \\
&= \mathbb{P}(X(u_1) > 1)\mathbb{P}(X(u_2) > 1)\mathbb{P}(X(u_3) > 1) \\
&= 0.3085375^3 = 0.02937136
\end{aligned}
$$

   (b)

```
> # Distances between locations
> u1.u2 <- sqrt(sum((c(1,0) - c(0,0))^2))
> u1.u3 <- sqrt(sum((c(1,0) - c(0,2))^2))
> u2.u3 <- sqrt(sum((c(0,0) - c(0,2))^2))
>
> # Distance matrix
> sij <- matrix(
+ c(0, u1.u2, u1.u3,
+ u1.u2, 0, u2.u3,
+ u1.u3, u2.u3, 0),
+ nrow=3, byrow=T)
> sij
         [,1] [,2]     [,3]
[1,] 0.000000    1 2.236068
[2,] 1.000000    0 2.000000
[3,] 2.236068    2 0.000000
>
> # Correlation matrix
> rho <- exp( - sij/2)
> rho
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.6065307 0.3269219
[2,] 0.6065307 1.0000000 0.3678794
[3,] 0.3269219 0.3678794 1.0000000
```

```
>
> # Covariance matrix
> sigma <- 2*2*rho
> sigma
         [,1]     [,2]     [,3]
[1,] 4.000000 2.426123 1.307688
[2,] 2.426123 4.000000 1.471518
[3,] 1.307688 1.471518 4.000000
>
> # 3-variate normal density
> library(mvtnorm)
> f <- function(x) dmvnorm(x, mean = c(0,0,0), sigma = sigma)
>
> # Probability P(X(u1)>1, X(u2)>1, X(u3)>1)
> library(cubature)
> hcubature(f, lower = c(1, 1, 1), upper = c(Inf, Inf, Inf))$integral
[1] 0.09364596
```

- item (a) -> 20p
- in item (b) correct correlation matrix -> 20p
- in item (b) correct covariance matrix -> 20p
- in item (b) correctly defined density to be integrated -> 20p
- in item (b) correct end result for the probability -> 20p

3. Assume that we have random variables X_i, $i = 1, 2, \ldots, n$ such that $X_i = 1$ if statistical unit i in our sample is color blind and $X_i = 0$ otherwise. Now X_i's are independent and

$$
\begin{aligned}
0.95 \leq \mathbb{P}(\sum_{i=1}^{n} X_i \geq 1) = 1 - \mathbb{P}\left(\sum_{i=1}^{n} X_i = 0\right) &= 1 - (\mathbb{P}(X_1 = 0))^n \\
\Rightarrow n \cdot \log(\mathbb{P}(X_1 = 0)) &\leq \log(1 - 0.95) \\
\Rightarrow n \geq \frac{\log(0.05)}{\log(\mathbb{P}(X_1 = 0))} = \frac{\log(0.05)}{\log(0.99)} &= 298.0729
\end{aligned}
$$

This implies that n at least 299.

- Formulation through random variables -> 40p
- Correct lower bound for sample size n -> 40p
- Final result that $n \geq 299$ -> 20p

4.
```
rounds <- 10000
# Create vectors for order statistic
os24 <- os210 <- os1010 <- rep(NA, n)
for(i in 1:rounds){
u4 <- runif(4, 0, 1)
os24[i] <- sort(u4)[2]
u10 <- runif(10, 0, 1)
os210[i] <- sort(u10)[2]
os1010[i] <- sort(u10)[10]
}
```

```
x <- seq(0,1, by=.01)
hist(os24, main = expression("Uniform "*X[2:4]), freq=FALSE)
j <- 2
n <- 4
lines(x, dbeta(x, j, n-j+1))


x <- seq(0,1, by=.01)
hist(os210, main = expression("Uniform "*X[2:10]), freq=FALSE)
j <- 2
n <- 10
lines(x, dbeta(x, j, n-j+1))

x <- seq(0,1, by=.01)
hist(os1010, main = expression("Uniform "*X[10:10]), freq=FALSE)
j <- 10
n <- 10
lines(x, dbeta(x, j, n-j+1))
```

- Simulation of observed values for each order statistic -> 15p
- Each histogram and each density function -> 9p
- Completely correct -> 100p

5.

$$
\begin{aligned}
F(x) &= \mathbb{P}(X_{k:n} \leq x) \\
&= \binom{n}{k} \mathbb{P}(X_1 \leq x \cap X_2 \leq x \cap \ldots \cap X_k \leq x \cap X_{k+1} > x \cap \ldots \cap X_n > x) \\
&+ \binom{n}{k+1} \mathbb{P}(X_1 \leq x \cap X_2 \leq x \cap \ldots \cap X_{k+1} \leq x \cap X_{k+2} > x \cap \ldots \cap X_n > x) + \cdots \\
&+ \binom{n}{n} \mathbb{P}(X_1 \leq x \cap X_2 \leq x \cap \ldots \cap X_n \leq x) \\
&= \binom{n}{k} x^k (1-x)^{n-k} + \binom{n}{k+1} x^{k+1} (1-x)^{n-(k+1)} + \cdots + \binom{n}{n} x^n \\
&= 1 - F_{binom}(k-1) \\
&= 1 - (n-(k-1)) \binom{n}{k-1} \int_0^{1-x} t^{n-(k-1)-1} (1-t)^{k-1} dt,
\end{aligned}
$$

where F_{binom} is the cdf of $Bin(n, x)$-distribution.

$$
\begin{aligned}
\frac{d}{dx} F(x) &= (n-(k-1)) \binom{n}{k-1} (1-x)^{n-(k-1)-1} (1-(1-x))^{k-1} \\
&= (n-(k-1)) \binom{n}{k-1} (1-x)^{(n-1)-(k-1)} x^{k-1} \\
&= n \binom{n-1}{k-1} x^{k-1} (1-x)^{(n-1)-(k-1)}
\end{aligned}
$$

$$= \frac{\Gamma(k+(n-k+1)}{\Gamma(k)\Gamma(n-k+1)}x^{k-1}(1-x)^{(n-1)-(k-1)}$$

This implies that

$$X_{k:n} \sim Beta(k, n-k+1)$$

$$\begin{aligned}
&(n-(k-1))\frac{n!}{(k-1)!(n-(k-1))!} \\
&= \frac{n!}{(k-1)!(n-(k-1)-1)!} \\
&= n\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} \\
&= n\binom{n-1}{k-1} = n\frac{(n-1)!}{(k-1)!(n-1-(k-1))!} \\
&= \frac{n!}{(k-1)!((n-k+1)-1)!} = \frac{(k+(n-k+1)-1)!}{(k-1)!((n-k+1)-1)!} \\
&= \frac{\Gamma(k+(n-k+1))}{\Gamma(k)\Gamma(n-k+1)}
\end{aligned}$$

- Finding correct formula for the cumulative distribution function -> 50p
- Correct formula for pdf deduced from cdf -> 50p

6. If $Y \sim Unif(0,1)$ then $10Y \sim unif(0,10)$. Transformation $g(y) = 10y$. $g^{-1}(x) = \frac{x}{10}$ and $\frac{\partial}{\partial x}g^{-1}(x) = \frac{1}{10}$ gives $f_X(x) = f_Y(g^{-1}(x))\frac{1}{10} = f_Y(\frac{1}{10}x)\frac{1}{10}$.

   Now plug-in the density of Beta-distribution:

$$\begin{aligned}
f_{X_{k:n}}(x) = f_{Y_{k:n}}(g^{-1}(x))\frac{1}{10} &= f_{Y_{k:n}}(\frac{1}{10}x)\frac{1}{10} \\
&= \frac{\Gamma(k+(n-k+1)}{\Gamma(k)\Gamma(n-k+1)}\left(\frac{1}{10}x\right)^{k-1}\left(1-\frac{1}{10}x\right)^{(n-1)-(k-1)}\frac{1}{10} \\
&= \frac{\Gamma(k+(n-k+1)}{\Gamma(k)\Gamma(n-k+1)}\frac{1}{10^k}x^{k-1}\left(1-\frac{1}{10}x\right)^{(n-k)}, x \in (0,10).
\end{aligned}$$

$$f_{X_{k:n}}(x) = 0, x \notin (0,10)$$

- Finding correct transformation -> 33p
- Finding correct formula for f_Y using the transformation -> 33p
- Finding correct pdf using the transformation and the pdf of Beta-distribution -> 34p

7. X_i has the same distribution as $F^{-1}(U_i)$, where $U_i \sim unif(0,1)$.

   The function F^{-1} does not change order $\Rightarrow F^{-1}(U_{k:n})$ has the same distribution as $X_{k:n}$.

   For transformation $g = F^{-1}$ we can use the transformation formula from Probabilistic inference for data science 1

$$f_{X_{k:n}}(x) = f_{U_{k:n}}(F(x))f(x), x \in (-\infty, \infty)$$

   where $f(x) = F'(x)$.

   Correct formula for the pdf of the order statistic -> 100p