Please provide complete and well-written solutions to the following exercises.

Due March 10, 9AM, to be submitted in blackboard, under the Assignments tab.

Homework 4

Exercise 1 (Order Statistics)**.** Let $X\colon\Omega\to\mathbf{R}$ be a random variable. Let $X_1,\ldots,X_n$ be a random sample of size n from X. Define $X_{(1)}:=\min_{1\leq i\leq n}X_i$, and for any $2\leq i\leq n$, inductively define

$$X_i:=\min\Big\{\{X_1,\ldots,X_n\}\smallsetminus\{X_{(1)},\ldots,X_{(i-1)}\}\Big\},$$

so that

$$X_{(1)}\leq X_{(2)}\leq\cdots\leq X_{(n)}=\max_{1\leq i\leq n}X_i.$$

The random variables $X_{(1)},\ldots,X_{(n)}$ are called the **order statistics** of $X_1,\ldots,X_n$.

- Suppose X is a discrete random variable and we can order the values that X takes as $x_1<x_2<\cdots$. For any $i\geq 1$, define $p_i:=\mathbf{P}(X\leq x_i)$. Show that, for any $1\leq i,j\leq n$,

$$\mathbf{P}(X_{(j)}\leq x_i)=\sum_{k=j}^{n}\binom{n}{k}p_i^k(1-p_i)^{n-k}.$$

  (Hint: Let Y be the number of indices $1\leq j\leq n$ such that $X_j\leq x_i$. Then Y is a binomial random variable with parameters n and p_i.)

  You don't have to show it, but if X is a continuous random variable with density f_X and cumulative distribution function F_X, then for any $1\leq j\leq n$, $F_{X_{(j)}}$ has density

$$f_{X_{(j)}}(x):=\frac{n!}{(j-1)!(n-j)!}f_X(x)(F_X(x))^{j-1}(1-F_X(x))^{n-j},\qquad\forall\,x\in\mathbf{R}.$$

  (This follows by differentiating the above identity for the cumulative distribution function.)
- Let X be a random variable uniformly distributed in $[0,1]$. For any $1\leq j\leq n$, show that $X_{(j)}$ is a beta distributed random variable with parameters j and $n-j+1$. Conclude that (as you might anticipate)

$$\mathbf{E}X_{(j)}=\frac{j}{n+1}.$$

- Let $a,b\in\mathbf{R}$ with $a<b$. Let U be the number of indices $1\leq j\leq n$ such that $X_j\leq a$. Let V be the number of indices $1\leq j\leq n$ such that $a<X_j\leq b$. Show that the vector $(U,V,n-U-V)$ is a multinomial random variable, so that for any

nonnegative integers u, v with $u + v \leq n$, we have

$$\begin{aligned}&\mathbf{P}(U = u, V = v, n - U - V = n - u - v)\\&\qquad = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b) - F_X(a))^v (1 - F_X(b))^{n-u-v}.\end{aligned}$$

Consequently, for any $1 \leq i, j \leq n$,

$$\mathbf{P}(X_{(i)} \leq a, X_{(j)} \leq b) = \mathbf{P}(U \geq i, U + V \geq j) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \mathbf{P}(U = k, V = m) + \mathbf{P}(U \geq j).$$

So, it is possible to write an explicit formula for the joint distribution of $X_{(i)}$ and $X_{(j)}$ (but you don't have to write it yourself).

Exercise 2. Using Matlab (or any other mathematical system on a computer), verify that its random number generator agrees with the law of large numbers and central limit theorem. For example, average 10^7 samples from the uniform distribution on $[0, 1]$ and check how close the sample average is to $1/2$. Then, sum up n samples from the uniform distribution on $[0, 1]$, construct this sum n times, make a histogram of the different values of the sum, and check how close the histogram is to a Gaussian (when $n = 10^4$). If you want a challenge, try $n = 10^5$ or $n = 10^6$.

Exercise 3. Let $X : \Omega \to \mathbf{R}$ be a random variable on a sample space Ω equipped with a probability law $\mathbf{P}$. For any $t \in \mathbf{R}$ let $F(t) := \mathbf{P}(X \leq t)$. For any $s \in (0, 1)$ define

$$Y(s) := \sup\{t \in \mathbf{R} \colon F(t) < s\}.$$

Then Y is a random variable on $(0, 1)$ with respect to the uniform probability law on $(0, 1)$. (That is, we can consider Y as a random variable $Y(S)$ where S is uniform on $(0, 1)$.) Show that X and Y are equal in distribution. That is, $\mathbf{P}(Y \leq t) = F(t)$ for all $t \in \mathbf{R}$.

Exercise 4 (Box-Muller Algorithm)**.** Let U_1, U_2 be independent random variables uniformly distributed in $(0, 1)$. Define

$$R := \sqrt{-2 \log U_1}, \qquad \Psi := 2\pi U_2.$$
$$X := R \cos \Psi, \qquad Y := R \sin \Psi.$$

Show that X, Y are independent standard Gaussian random variables. So, we can simulate any number of independent standard Gaussian random variables with this procedure.

Now, let $\{a_{ij}\}_{1 \leq i,j \leq n}$ be an $n \times n$ symmetric positive semidefinite matrix. That is, for any $v \in \mathbf{R}^n$, we have

$$v^T a v = \sum_{i,j=1}^{n} v_i v_j a_{ij} \geq 0.$$

We can simulate a Gaussian random vector with any such covariance matrix $\{a_{ij}\}_{1 \leq i,j \leq n}$ using the following procedure.

- Let $X = (X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian random variables (which can be sampled using the Box-Muller algorithm above).
- Write the matrix a in its Cholesky decomposition $a = rr^*$, where r is an $n \times n$ real matrix. (This decomposition can be computed efficiently with about n^3 arithmetic operations.)

- Let $e^{(1)},\ldots,e^{(n)}$ be the rows of r. For any $1\leq i\leq n$, define
$$Z_i := \langle X, e^{(i)}\rangle.$$

Show that $Z:=(Z_1,\ldots,Z_n)$ is a mean zero Gaussian random vector whose covariance matrix is $\{a_{ij}\}_{1\leq i,j\leq n}$, so that
$$\mathbf{E}(Z_iZ_j)=a_{ij},\qquad \forall\, 1\leq i,j\leq n.$$

Exercise 5 (Optional)**.** In the notes we showed that the Delta Method works only assuming that $f'(\theta)$ exists. In fact, the method works even when $f'(\theta)$ does not exist. In this exercise, we assume that
$$f'(\theta^+) := \lim_{y\to\theta^+}\frac{f(y)-f(\theta)}{y-\theta},\qquad f'(\theta^-) := \lim_{y\to\theta^-}\frac{f(y)-f(\theta)}{y-\theta},$$
exist. For example, consider
$$f(y) := \max(y,0),\qquad \forall\, y\in\mathbf{R}.$$
Then $f'(0^+)=1$ while $f'(0^-)=0$, so $f'(0)$ does not exist.

For simplicity, we assume that $\theta=0$ and $f(\theta)=0$.

Let $Y_1,Y_2,\ldots$ be random variables such that $\sqrt{n}(Y_n-\theta)$ converges in distribution to a mean zero Gaussian random variable with variance $\sigma^2>0$.

- Argue as in the notes, and show that for all $y\in\mathbf{R}$, there exists a function h with $\lim_{z\to 0}h(z)/z=0$, and
$$f(y)=f'(0^+)y1_{y>0}+f'(0^-)y1_{y<0}+h(y).$$
- Conclude that
$$\sqrt{n}f(Y_n)=\sqrt{n}\Big(f'(0^+)Y_n1_{Y_n>0}+f'(0^-)Y_n1_{Y_n<0}+h(Y_n)\Big).$$
- Deduce that, as $n\to\infty$, $\sqrt{n}f(Y_n)$ converges in distribution to
$$\Big(\sigma f'(\theta^+)1_{Z>0}+\sigma f'(\theta^-)1_{Z<0}\Big)Z.$$
(Note that $f'(0^+)Y_n1_{Y_n>0}$ and $f'(0^-)Y_n1_{Y_n<0}$ have disjoint supports; this could be useful to prove convergence in distribution as $n\to\infty$.)

Exercise 6. Let A,B,Ω be sets. Let $u\colon\Omega\to A$ and let $t\colon\Omega\to B$. Assume that, for every $x,y\in\Omega$, if $u(x)=u(y)$, then $t(x)=t(y)$. Show that there exists a function $s\colon A\to B$ such that
$$t=s(u).$$

Exercise 7. Let $\{f_\theta\colon\theta\in\Theta\}$ be a k-parameter exponential family $\{f_\theta\colon\theta\in\Theta,\, a(w(\theta))<\infty\}$ of probability density functions or probability mass functions, where
$$f_\theta(x) := h(x)\exp\Big(\sum_{i=1}^k w_i(\theta)t_i(x)-a(w(\theta))\Big),\qquad \forall\, x\in\mathbf{R}.$$
For any $\theta\in\Theta$, let $w(\theta):=(w_1(\theta),\ldots,w_k(\theta))$. Assume that the following subset of $\mathbf{R}^k$ is k-dimensional:
$$\{w(\theta)-w(\theta')\in\mathbf{R}^k\colon \theta,\theta'\in\Theta\}.$$

That is, if $x \in \mathbf{R}^k$ satisfies $\langle x, y\rangle = 0$ for all y in this set, then $x = 0$. (Note that the assumption of the exercise is always satisfied for an exponential family in canonical form.)

Let $X = (X_1, \ldots, X_n)$ be a random sample of size n from f_θ. Define $t\colon \mathbf{R}^n \to \mathbf{R}^n$ by

$$t(X) := \sum_{j=1}^{n} (t_1(X_j), \ldots, t_k(X_j)).$$

Show that $t(X)$ is minimal sufficient for θ. (Hint: if you get stuck, look at Example 3.12 in Keener.)

Conclude that if we sample from a Gaussian with unknown mean μ and variance $\sigma^2 > 0$, then $\overline{X}$ is minimal sufficient for μ and $(\overline{X}, S)$ is minimal sufficient for (μ, σ^2).

Warning: the f_θ exponential family mentioned here is a function of one variable. If you use the Theorem from class about checking the ratio of $f_\theta(x)/f_\theta(y)$, the functions there are *joint* density functions (i.e. the product of n copies of the same function).

Optional: If the f_θ functions are always positive, you should be able to change the assumption to the following. For any $\theta \in \Theta$, let $w(\theta) := (w_1(\theta), \ldots, w_k(\theta))$. Assume that the following subset of $\mathbf{R}^k$ is k-dimensional:

$$\{w(\theta) \in \mathbf{R}^k \colon \theta, \theta' \in \Theta\}.$$

Exercise 8. Let $\mathbf{P}_1, \mathbf{P}_2$ be two probability laws on the sample space $\Omega = \mathbf{R}$. Suppose these laws have densities $f_1, f_2\colon \mathbf{R} \to [0, \infty)$ so that

$$\mathbf{P}_i(A) = \int_A f_i(x)dx, \qquad \forall\, i = 1, 2, \qquad \forall\, A \subseteq \mathbf{R}.$$

Show that

$$\sup_{A \subseteq \mathbf{R}} |\mathbf{P}_1(A) - \mathbf{P}_2(A)| = \frac{1}{2}\int_{\mathbf{R}} |f_1(x) - f_2(x)|\, dx.$$

(Hint: consider $A := \{x \in \mathbf{R} \colon f_1(x) > f_2(x)\}$.)

Similarly, if $\mathbf{P}_1, \mathbf{P}_2$ are probability laws on $\Omega = \mathbf{Z}$, show that

$$\sup_{A \subseteq \mathbf{Z}} |\mathbf{P}_1(A) - \mathbf{P}_2(A)| = \frac{1}{2}\sum_{z \in \mathbf{Z}} |\mathbf{P}_1(z) - \mathbf{P}_2(z)|.$$

Exercise 9. Give an example of a statistic Y that is complete and nonconstant, but such that Y is not sufficient.

Exercise 10. This exercise shows that a complete sufficient statistic might not exist.

Let $X_1, \ldots, X_n$ be a random sample of size n from the uniform distribution on the three points $\{\theta, \theta + 1, \theta + 2\}$, where $\theta \in \mathbf{Z}$.

- Show that the vector $Y := (X_{(1)}, X_{(n)})$ is minimal sufficient for θ.
- Show that Y is not complete by considering $X_{(n)} - X_{(1)}$.
- Using minimal sufficiency, conclude that any sufficient statistic for θ is not complete.

Warning: An earlier version of this exercise considered all $\theta \in \mathbf{R}$, whereas now we only consider $\theta \in \mathbf{Z}$. The case $\theta \in \mathbf{R}$ was unintentionally difficult.

Exercise 11 ((Optional) This exercise requires some measure theory so it is optional.)**.** Let $\{f_\theta : \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta : \theta \in \Theta, a(w(\theta)) < \infty\}$ of **joint** probability density functions or probability mass functions in canonical form, where

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x) - a(w)\Big), \qquad \forall\, x \in \mathbf{R}^n, \qquad \forall\, w \in \{w \in \mathbf{R}^k : a(w) < \infty\}.$$

Assume that the following subset of $\mathbf{R}^k$ contains an open set in $\mathbf{R}^k$:

$$\{w \in \mathbf{R}^k : a(w) < \infty\}.$$

Assume also that there is no redundancy in the functions $t_1, \ldots, t_k$, i.e. assume: if $\exists$ $\alpha_1, \ldots, \alpha_k \in \mathbf{R}$ such that $\sum_{i=1}^k \alpha_i t_i(x) = 0$ for all $x \in \mathbf{R}^n$, then $\alpha_1 = \cdots = \alpha_k = 0$.

Let X be a random sample **of size** 1 from f_θ (so $X = (X_1, \ldots, X_n)$, and $X_1, \ldots, X_n$ are all real valued). Define $t : \mathbf{R}^n \to \mathbf{R}^n$ by

$$t(X) := (t_1(X), \ldots, t_k(X)).$$

Show that $t(X)$ is complete for θ.

Hint: if you get stuck, look at Theorem 4.3.1 in Lehmann-Romano. An early step in the proof uses the change of variables formula for the pushforward measure.

Once we know the above statement, we can deduce the following about repeated random samples from a single variable exponential family.

Let $\{f_\theta : \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta : \theta \in \Theta, a(w(\theta)) < \infty\}$ of probability density functions or probability mass functions in canonical form, where

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x) - a(w)\Big), \qquad \forall\, x \in \mathbf{R}, \qquad \forall\, w \in \{w \in \mathbf{R}^k : a(w) < \infty\}.$$

Assume that the following subset of $\mathbf{R}^k$ contains an open set in $\mathbf{R}^k$:

$$\{w \in \mathbf{R}^k : a(w) < \infty\}.$$

Assume also that there is no redundancy in the functions $t_1, \ldots, t_k$, i.e. assume: if $\exists$ $\alpha_1, \ldots, \alpha_k \in \mathbf{R}$ such that $\sum_{i=1}^k \alpha_i t_i(x) = 0$ for all $x \in \mathbf{R}$, then $\alpha_1 = \cdots = \alpha_k = 0$.

Let $X_1, \ldots, X_n$ be a random sample **of size** n from f_θ. Define $t : \mathbf{R}^n \to \mathbf{R}^n$ by

$$t(X) := \sum_{j=1}^n (t_1(X_j), \ldots, t_k(X_j)).$$

Show that $t(X)$ is complete for θ.