1. (a) Because $x_i$ are assumed to be independent realizations of a continuous random variable, it is almost surely (a.s.)\(^1\) the case that

$$x'_1 < x'_2 < \cdots < x'_n$$

Thus, the preimage of $x'$ contains the $n!$ possible arrangements of the a.s. distinct values of the (ordered) components of $x'$. As there is no reason to prefer any one arrangement over another, the classical probability assumption of equally likely events\(^2\) yields

$$p_\theta(x \mid T(x) = x') = \frac{1}{n!} \chi \{x \in T^{-1}(x')\} \quad \text{a.s.}$$

which is independent of the unknown parameter vector $\theta$.\(^3\) Thus the vector of order statistics $x' = T(x)$ is sufficient as a consequence of the classical (Fisherian) definition of sufficiency.

(b) Because of the assumption of independence,

$$p_\theta(x) = p_\theta(x_1) \cdots p_\theta(x_n) = p_\theta(x'_1) \cdots p_\theta(x'_n) = g(\theta, T(x)).$$

Thus the vector of order statistics $x' = T(x)$ is sufficient as a consequence of the Neyman–Fisher Factorization Theorem.

2. The solution closely follows the development done in lecture. Invoking the Axiom of Choice, for each coset\(^4\) $[x]$ of the equivalence class (partition) defined in the problem statement we can choose a unique representative point $\xi \in [x]$ to serve as an index for the coset,

$$A_\xi \triangleq [x] \quad \text{iff} \quad \xi \in [x].$$

We then define a statistic $S(\cdot)$ which takes its values in the set of index variables by

$$\xi = S(x) \quad \text{iff} \quad x \in A_\xi.$$

Note this statistic induces the original partition,

$$x \sim x' \quad \text{iff} \quad S(x) = S(x') = \xi \in [x] \quad \text{iff} \quad x, x' \in A_\xi \quad \text{iff} \quad x, x' \in [\xi] \quad \text{iff} \quad x, x' \sim \xi.$$

---

\(^1\)Recall that “almost surely” means “with probability equal to one.”
\(^2\)Sometimes called the “Principle of Insufficient Reason” because there is no “reason” to consider any one outcome as more probable than another.
\(^3\)As usual, $\chi\{A\}$ denotes the characteristic (or indicator) function of the event $A$.
\(^4\)By definition $y \in [x]$ iff $y \sim x$. 

1
Sufficiency of $S$. By definition of the equivalence class (partition) defined in the problem statement, $x \sim \xi = S(x)$ implies that
\[
\frac{p_\theta(x)}{p_\theta(S(x))} = g(x, S(x))
\]
or
\[
p_\theta(x) = p_\theta(S(x)) \cdot g(x, S(x)) \triangleq f(\theta, S(x)) \cdot h(x)
\]
so that $S$ (and hence the partition) is sufficient from the Neyman–Fisher Factorization Theorem.

Necessity of $S$. Let $T(x)$ be any sufficient statistic for $p_\theta(x)$. Then $T(x) = T(x')$ implies that
\[
\frac{p_\theta(x)}{p_\theta(x')} = \frac{p_\theta(x|T(x))}{p_\theta(x'|T(x'))} = \frac{p_\theta(x|T(x))}{p_\theta(x'|T(x'))} = \frac{p(x|T(x))}{p(x'|T(x'))} \triangleq g(x, x')
\]
where $g(x, x')$ is independent of $\theta$. By the definition of the equivalence class induced by the statistic $S$ defined above, this in turn implies that $S(x) = S(x')$. Thus the partition induced by $S$ is coarser than that induced by any sufficient statistic $T$, which is true iff $S$ is a function of any such $T$. Thus (by definition of necessity), $S$ is a necessary statistic for $p_\theta(x)$.

3. If you had difficulty with this problem please come to my office hour.

4. On its domain of positive support a $k$–parameter exponential family distribution has the form
\[
\ln p_\theta(y) = Q(\theta)^T T(y) - b(\theta) + a(y) = \sum_{j=1}^{k} Q_j(\theta) T_j(y) - b(\theta) + a(y). \tag{1}
\]
If the sufficient statistics\textsuperscript{6} are affinely dependent, then at least one of them (say, without loss of generality, $T_k(y)$) can be written as an affine combination of the others,
\[
T_k(y) = \alpha_0 + \alpha_1 T_1(y) + \cdots + \alpha_{k-1} T_{k-1}(y) = \alpha_0 + \sum_{j=1}^{k-1} \alpha_j T_j(y). \tag{2}
\]
Thus
\[
\sum_{j=1}^{k} Q_j(\theta) T_j(y) - b(\theta) + a(y) = \sum_{j=1}^{k-1} [Q_j(\theta) + \alpha_j Q_k(\theta)] T_j(y) - [b(\theta) - \alpha_0 Q_k(\theta)] + a(y)
\]
\[
= \sum_{j=1}^{k-1} Q_j(\theta) T_j(y) - \tilde{b}(\theta) + a(y)
\]
\textsuperscript{5}Recall that $\{x\} = \{x\} \cap \{T(x)\} = \{x\} \cap \{x' : T(x') = T(x)\}$ induces a “cut” $p(x) = p(x, T(x)) = p(x|T(x)) p(T(x))$.
\textsuperscript{6}Why is $T(y)$ always a vector of sufficient statistics?

2
where
\[ \tilde{Q}_j(\theta) \triangleq Q_j(\theta) + \alpha_j Q_k(\theta), \quad j = 1, \cdots, k - 1 \]

and
\[ \tilde{b}(\theta) \triangleq b(\theta) - \alpha_0 Q_k(\theta), \]
showing that we can reduce \( p_0(y) \) to a \( k - 1 \)-parameter exponential family distribution.

Similarly, if \( Q_j(\theta), j = 1, \cdots, k \) are an affinely dependent set of functions of \( \theta \), then at least one of them (say \( Q_k(\theta) \)) can be written as an affine combination of the others
\[ Q_k(\theta) = \sum_{j=1}^{k-1} \beta_j Q_j(\theta) \]
and again we can reduce \( p_0(y) \) to a \( k - 1 \)-parameter exponential family distribution.

(With an appropriate possible redefinition of the term \( a(y) \).)

Of course, if there are any remaining affine dependencies, then the order of the distribution can be reduced even further.

5. Note that
\[ y - \langle y \rangle = y - \overline{y} + \overline{y} - \langle y \rangle \]
and therefore
\[
(y - \langle y \rangle)(y - \langle y \rangle)^T = (y - \overline{y})(y - \overline{y})^T + (y - \overline{y})(\overline{y} - \langle y \rangle)^T + (\overline{y} - \langle y \rangle)(\overline{y} - \langle y \rangle)^T.
\]

Now note that
\[
\mathbb{E}\left\{ (y - \overline{y})(\overline{y} - \langle y \rangle)^T \right\} = \mathbb{E}\left\{ \mathbb{E}\left\{ (y - \overline{y})(\overline{y} - \langle y \rangle)^T \mid x \right\} \right\} = \mathbb{E}\left\{ (\overline{y} - \overline{y})(\overline{y} - \langle y \rangle)^T \right\} = 0.
\]

Therefore
\[
\text{Cov}\{y\} = \mathbb{E}\left\{ (y - \overline{y})(y - \overline{y})^T \right\} + \text{Cov}\{\overline{y}\} \geq \text{Cov}\{\overline{y}\} = \text{Cov}\{\mathbb{E}\{y\mid x\}\}.
\] (3)

Note that the left–hand–side of Equation (3) can also be written as
\[
\text{Cov}\{y\} = \mathbb{E}\{\text{Cov}\{y\mid x\}\} + \text{Cov}\{\mathbb{E}\{y\mid x\}\}.
\]

6. This problem is essentially done in Example 5.8 of Kay (once we recognize that an unbiased estimate is provided by twice an unbiased estimate of the mean). The only part of the homework problem which is unsolved in Example 5.8 is proving that the sufficient statistic
\[ T(Y^m) = \max_{1 \leq i \leq m} y_i \]
is complete. As shown in Kay, sufficiency easily follows from
\[ p_\theta(Y^m) = \left( \frac{1}{\theta^m} \chi \left\{ \max_{1 \leq i \leq m} y_i \leq \theta \right\} \right) \cdot \left( \chi \left\{ \min_{1 \leq i \leq m} y_i \geq 0 \right\} \right) = g(T(Y^m), \theta) \cdot h(Y^m) \]
for any \( \theta > 0 \), and the Neyman–Fisher Factorization Theorem.\(^7\) Completeness follows from noting (as per the development on page 115 of Kay) that any measurable function of \( T \), say \( W(t) \), has an expectation given by
\[ E_\theta \{ W(T) \} = \frac{m}{\theta^m} \int_0^\theta W(t) t^{m-1} dt . \]
Let \( W(t) = W^+(t) - W^-(t) \), where \( W^+(t) \geq 0 \) and \( W^-(t) \geq 0 \) are nonnegative functions of \( t \).\(^8\) Then \( E_\theta \{ W(T) \} = 0 \) for all \( \theta > 0 \) if and only if
\[ \int_0^\theta W^+(t) t^{m-1} dt = \int_0^\theta W^-(t) t^{m-1} dt \]
for all \( \theta > 0 \), which, in turn, is true if and only if
\[ \int_{\theta_1}^{\theta_2} W^+(t) t^{m-1} dt = \int_{\theta_1}^{\theta_2} W^-(t) t^{m-1} dt \]
for every \( \theta_1 \) and \( \theta_2 \) such that \( \theta_2 \geq \theta_1 > 0 \). Because the integrands are positive and the equality must hold for any strictly positive \( \theta \), it must therefore be the case that \( W^+(t) = W^-(t) \) for almost all \( t \geq 0 \). Therefore \( W(t) = 0 \) for almost all \( t \) showing that \( T \) is complete.

7. Kay 5.13. Note that like the previous problem the density is not regular (in particular the area of positive support again depends on the unknown parameter \( \theta \)) so that we cannot compute a Cramér–Rao lower bound. Note that we can write the sample data pdf as
\[ p_\theta(X^N) = \left( e^{N\theta} \chi \left\{ \min_{1 \leq n \leq N} x[n] \geq \theta \right\} \right) \cdot \left( e^{-\sum_{n=1}^N x[n]} \right) = g(T(X^N), \theta) \cdot h(X^N) . \]
Therefore, from the Neyman–Fisher Factorization Theorem a sufficient statistic is determined to be
\[ T(X^N) = \min_{1 \leq n \leq N} x[n] . \]
\(^7\)Here, \( \chi \{ \} \) denotes the so-called characteristic (or indicator) function.
\(^8\)This can be done for any real function \( W(t) \).
We now proceed to find the distribution function of $T$,

$$P_\theta(T \leq t) = 1 - P_\theta(T > t) = 1 - P_\theta(\min_{1 \leq n \leq N} x[n] > t)$$

$$= 1 - P_\theta(x[1] > t, \ldots, x[N] > t)$$

$$= 1 - \left( \int_{\max(t, \theta)}^\infty e^{\theta-x} \, dx \right)^N$$

$$= 1 - e^{N(\theta - \max(t, \theta))}.$$

Differentiating the distribution function with respect to $t$ we obtain the pdf,

$$p_\theta(t) = \begin{cases} Ne^{N(\theta-t)} & t \geq \theta \\ 0 & t < \theta \end{cases}.$$

The expected value of $T$ can now be computed,

$$E_{\theta} \{T\} = \theta + \frac{1}{N}.$$

An unbiased estimator is then obviously given by

$$\hat{\theta} = T - \frac{1}{N} = \min_{1 \leq n \leq N} x[n] - \frac{1}{N}.$$

It can be shown that $T$ is a complete sufficient statistic. Therefore, from the RBLTS Theorem, we have found the UMVUE of the unknown parameter $\theta$. Note that from the pdf of $T$ we can compute the (uniformly optimal, parameter dependent) error variance if we so desire.

8. Moon 10.5.8. This is a generalization of the previous problem. The general class of such non–regular exponential families (i.e., exponential family–like, but with parameter–dependent support) is discussed in the text by Ferguson cited in Footnote 9.

(a) For $\sigma$ known, $\mu$ unknown we have

$$p_\theta(\mathcal{X}^n) = \left( e^{N\frac{\mu}{\sigma}} \chi \left\{ \min_{1 \leq k \leq n} x_k \geq \mu \right\} \right) \cdot \left( \frac{e^{-1/\sigma \sum_{k=1}^n x_k}}{\sigma^n} \right) = g(T(\mathcal{X}^n), \theta) \cdot h(\mathcal{X}^n).$$

From the N–F Factorization Theorem, $T(\mathcal{X}^n) = \min_{1 \leq k \leq n} x_k$ is a sufficient statistic for $\mu$. It is also complete.

(b) For $\sigma$ unknown and $\mu$ known we have

$$p_\theta(X^n) = \left(\frac{e^{\frac{n\mu}{2}} - \frac{1}{2} \sum_{k=1}^{n} x_k}{\sigma^n}\right) \cdot \left(\chi \left\{ \min_{1\leq k\leq n} x_k \geq \mu \right\} \right) = g(T(X^n), \theta) \cdot h(X^n).$$

Thus $T(X^n) = \sum_{k=1}^{n} x_k$ is a sufficient statistic for $\sigma$.

(c) For both $\sigma$ and $\mu$ unknown we have

$$p_\theta(X^n) = \left(\frac{e^{\frac{n\mu}{2}} - \frac{1}{2} \sum_{k=1}^{n} x_k}{\sigma^n}\right) \cdot \chi \left\{ \min_{1\leq k\leq n} x_k \geq \mu \right\} \cdot 1 = g(T(X^n), \theta) \cdot h(X^n).$$

Therefore $T(X^n) = (\sum_{k=1}^{n} x_k, \min_{1\leq k\leq n} x_k)^T$ is a sufficient statistic for $\theta = (\sigma, \mu)^T$.

9. Kay 5.15. If you have trouble with the first two parts of this problem, please come see me at my office hour. Note that the Gaussian, Rayleigh, and Exponential distributions are regular exponential families and that $T$ is a complete (and hence minimal) sufficient statistic. Note that below we can find an UMVUE for each case, but only if we choose an appropriate parameterization. This is because of the strong constraint that the estimator be uniformly unbiased.

(a) Gaussian. Here $\theta = \mu$, $T(x) = \sum_{n=1}^{N} x[n]$ and $E_\theta \{T(x)\} = N\mu$. Therefore the UMVUE for $\mu$ is given by

$$\hat{\mu} = \frac{1}{N} T(x) = \frac{1}{N} \sum_{n=1}^{N} x[n].$$

(b) Rayleigh. Here $\theta = \sigma^2$, $T(x) = \sum_{n=1}^{N} x^2[n]$ and $E_\theta \{T(x)\} = 2\sigma^2 N$. Therefore

$$\hat{\sigma}^2 = \frac{1}{2N} T(x) = \frac{1}{2N} \sum_{n=1}^{N} x^2[n]$$

is the UMVUE for $\sigma^2$.

(c) Exponential. Here the appropriate parameterization is a little trickier. Now we take $\theta = \frac{1}{\lambda}$. We also have $T(x) = \sum_{n=1}^{N} x[n]$ and $E_\theta \{T(x)\} = \frac{N}{\lambda}$. Thus the UMVUE is given by

$$\hat{\theta} = \left(\frac{1}{\lambda}\right) = \frac{1}{N} T(x) = \frac{1}{N} \sum_{n=1}^{N} x[n].$$

10. Under the multivariate gaussian assumption we have that

$$p_\theta(y) = c \exp \left\{-\frac{1}{2} \|y - A\theta\|^2 \right\}$$

where the normalizing constant $c$ is independent of $\theta$ and the full column–rank matrix $A$ is $k \times n$. 

6
(a) With \( T(y) \triangleq A^T R^{-1} y \) we have
\[
\|y - A\theta\|_{R^{-1}}^2 = \|y\|_{R^{-1}}^2 - 2\theta^T T(y) + \|A\theta\|_{R^{-1}}^2.
\]
Thus, as a consequence of the NFFT, \( T \) is sufficient and \( p_\theta(y) \) is seen to be an exponential family distribution. Because of the full column rank assumption on \( A \), the rows of \( A^T \) are linearly independent which means that the components of \( T \) are linearly independent functions of \( \theta \). Because the parameter vector is unconstrained, the parameter space has nonempty interior so that \( T \) is minimal and complete.\(^\text{10}\) Because of the assumption that \( A \) has full column rank, it must be the case that \( n \leq k \). Thus the \( n \)-dimensional minimum sufficient statistic \( T \) is no larger than the dimension \( k \) of the raw data. If it is the case that \( n < k \) then data compression has occurred, which is especially nice in the case when \( n \ll k \).

(b) With \( T \) a complete, minimum sufficient statistic, the RBLS Theorem tells us that the UMVUE (if it exists\(^\text{11}\)) must be a function of \( T \). Noting that
\[
E_\theta \{ T \} = A^T R^{-1} A \theta,
\]
with \( A \) full column–rank, it is evident that the UMVUE is given by
\[
\hat{\theta}(y) = (A^T R^{-1} A)^{-1} T(y) = (A^T R^{-1} A)^{-1} A^T R^{-1} y.
\]

11. The solution to this problem involves a very simple application of the result from the previous section. Define \( T_k \triangleq T(y_k) \) and
\[
\Sigma_k \triangleq E \{ v_k v_k^T \} = \text{diag}(\sigma_1, \ldots, \sigma_k) = \text{diag}(\Sigma_{k-1} \sigma_k).
\]
Then
\[
T_k = A_k^T \Sigma_k^{-1} y_k = A_{k-1}^T \Sigma_{k-1}^{-1} y_{k-1} + \frac{1}{\sigma_k} y[k] = T_{k-1} + b_k y[k]
\]
where \( b_k \triangleq \frac{r_k}{\sigma_k} \). Note that both \( T_k \) and \( b_k \) are \( n \)-dimensional for all times \( k \).

12. Kay 7.7. In particular, we will show that \( \hat{\theta}_{ML} \) is the solution to the equation
\[
E_\theta \{ B(x) \} = B(x)
\]
where \( x \) denotes the vector of the \( N \) iid samples. Recall that we have defined \( f'(\theta) \) as
\[
f'(\theta) = \frac{\partial}{\partial \theta} f(\theta).
\]
\(^\text{10}\)This is a fundamental property of regular, full–rank, nonempty parameter–set exponential family distributions.
\(^\text{11}\)Remember that the set of uniformly unbiased estimators might be empty for an unfelicitous choice of parameterization.
Because the samples are iid, we have that
\[ p_\theta(x) = \exp \left\{ A(\theta)B(x) + C(x) + ND(\theta) \right\} \]
where
\[ B(x) = \sum_{k=1}^{N} B(x_i) \]  
\[ C(x) = \sum_{k=1}^{N} C(x_i) . \]

The MLE is found as a solution of the likelihood equation,
\[ S_\theta(x) = 0 , \]  
where \( S_\theta(x) \) is the score function,
\[ S_\theta(x) = \frac{\partial}{\partial \theta} \ln p_\theta(x) = A'(\theta)B(x) + ND'(\theta) . \]
I.e., the MLE is a solution to
\[ A'(\theta)B(x) + ND'(\theta) = 0 . \]

It is evident, then, that the MLE can be equivalently found as a solution to the equation
\[ B(x) = -N \frac{D'(\theta)}{A'(\theta)} \]  

Now because we are working with a regular exponential family, we can interchange the order of differentiation and integration in expressions involving \( p_\theta(x) \). In particular, differentiating the expression
\[ \int p_\theta(x)dx = 1 \]
with respect to \( \theta \) yields the fact that the score has zero mean uniformly in \( \theta \)
\[ E_\theta \{ S_\theta(x) \} = 0 \]  
an expression derived and discussed in class last quarter. The likelihood equation (5) and the zero-mean condition (9) are the key equations needed to solve this homework problem.
From (6), equation (9) is seen to be equivalent to
\[ E_\theta \{ A'(\theta)B(x) + ND'(\theta) \} = 0 \]
which can be rearranged as
\[ E_\theta \{ B(x) \} = -N \frac{D'(\theta)}{A'(\theta)} \tag{10} \]
If we take \( \theta \) itself to be the natural parameter
\[ \theta = A \]
then this becomes
\[ E_A \{ B(x) \} = -N D'(A) \tag{11} \]
where \( D \) is here taken to be a function of the natural parameter \( A \) and \( D'(A) \) denotes the derivative of \( D \) with respect to \( A \).

Finally, comparison of equations (8) and (10) shows that the MLE can be found as a solution to the equation
\[ E_\theta \{ B(x) \} = B(x) \]
as claimed.

(a) Unit variance, unknown mean Gaussian case. We have,
\[ p(x; \mu) = \phi(x; \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (x - \mu)^2 \right\} = \exp \left\{ \mu x - \frac{1}{2} x^2 - \frac{1}{2} (\mu^2 + \ln 2\pi) \right\} . \]
Thus we have,
\[ A(\mu) = \mu, \quad B(x) = x, \quad C(x) = -\frac{x^2}{2}, \quad D(\mu) = -\frac{1}{2} (\mu^2 + \ln 2\pi), \]
and therefore,
\[ A'(\mu) = 1, \quad D'(\mu) = -\mu, \quad B(x) = \sum_{n=1}^{N} B(x[n]) = \sum_{n=1}^{N} x[n] . \]
Substitution into equation (7) above and rearranging yields the answer,
\[ \hat{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} x[n] . \]

(b) Exponential Distribution. Here we have,
\[ p(x; \lambda) = \lambda \exp \{ -\lambda x \} = \exp \{ -\lambda x + \ln \lambda \} . \]
Thus,

\[ A(\lambda) = -\lambda, \quad B(x) = x, \quad C(x) = 0, \quad D(\lambda) = \ln \lambda, \]

so that

\[ A'(\lambda) = -1, \quad D'(\lambda) = \frac{1}{\lambda}, \quad B(x) = \sum_{n=1}^{N} B(x[n]) = \sum_{n=1}^{N} x[n]. \]

Substituting into equation (7) above and rearranging yields

\[ \hat{\lambda}_{ML} = \left( \frac{1}{N} \sum_{n=1}^{N} x[n] \right)^{-1}. \]