36-705 Intermediate Statistics Fall 2016: Practice Test II.

Date: 4th November 2020.

This is my attempt at the practice test questions.

Correction status: pending.

1) Let $X_1, X_2, \ldots, X_n \sim \text{Uniform}(-\theta, \theta)$ where $\theta > 0$.

a) Find the maximum likelihood estimator $\widehat{\theta}_n$.

b) Find a minimal sufficient statistic.

c) Show that $\widehat{\theta}_n \xrightarrow{p} \theta$.

d) Find the limiting distribution of $n(\theta - \widehat{\theta}_n)$

1a)

The PDF of $X_i \sim \text{Uniform}(-\theta, \theta)$ is:

$$f_{X_i}(x_i;\theta) = \begin{cases} \frac{1}{2\theta} & \text{if} -\theta \leq x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

As the X_i are IID, the likelihood function is the product of individual PDFs. This is evaluated for a specific realisation of the data, i.e. fixing $X_1 = x_1, \ldots, X_n = x_n$, and varies as a function of θ over a parameter space Θ:

$$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta)$$

The likelihood function $L(\theta)$ will be 0 when any of the PDFs $f_{X_i}(x_i;\theta)$ are 0. This occurs when either $-\theta > \min\{X_1, \ldots, X_n\}$ or when $\theta < \max\{X_1, \ldots, X_n\}$.

At these values, there will be data points $X_k = x_k$ where $x_k < -\theta$ or $x_k > \theta$ which have a zero-probability of occurring, according to that particular parametrisation. The likelihood will then be 0 because given the observed values $X_1 = x_1, \ldots, X_n = x_n$, it will be almost impossible for that parametrisation to have generated the data.

Hence the likelihood function is:

$$L(\theta) = \begin{cases} 0 & \text{if} \quad -\theta > X_{(1)} \\ 0 & \text{if} \quad \theta < X_{(n)} \\ \left(\frac{1}{2\theta}\right)^n & \text{if} \quad \theta > X_{(n)}, -\theta > X_{(1)} \end{cases}$$

Which can be written as:

$$L(\theta) = \left(\frac{1}{2\theta}\right)^n \mathbb{I}(-\theta \leq X_{(1)})\mathbb{I}(\theta \geq X_{(n)})$$

In order to maximise this, we need to set the parameter θ so that the constraints imposed by the indicator functions are met, otherwise the likelihood function becomes 0. The form of the likelihood function also suggests that we estmate θ such that $\theta = X_{(n)}$ or $\theta = -X_{(1)}$, depending on whichever of $-X_{(1)}$ or $X_{(n)}$ is larger, thereby 'trapping' the entirety of the data, summarised by the 1st order statistic $X_{(1)}$ and nth order statistic $X_{(n)}$. This can be condensed into the more concise condition that we select θ to be the data point that is the largest in magnitude i.e. absolute value.

Furthermore the factor $(1/2\theta)^n$ is decreasing in θ provided the conditions within the indicators are met, meaning that any further increases in θ will decreas $L(\theta)$. With that in mind, we have that the maximum likelihood estimator is

$$\hat{\theta}_n = \max\{-X_{(1)}, X_{(n)}\} = \max_i\{|X_i|\}$$

1b)

A statistic T is minimal sufficient if it is the case that for two observed data sets x^n and y^n, where the notation x^n refers to $X_1 = x_1, \ldots, X_n = x_n$, we have that

$$R = \frac{p(x^n;\theta)}{p(y^n;\theta)} = c \iff T(x^n) = T(y^n)$$

Where c is a constant that does not depend on the parameter θ.

The condition that $T(x^n) = T(y^n)$ implies that R does not depend on θ means that T is sufficient. If it is the case that the reverse implication that R does not depend on θ implies $T(x^n) = T(y^n)$ additionally holds, then T is additionally minimal sufficient.

Notice that we can express the likelihood function in the following way:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \left(\frac{1}{2\theta}\right) \mathbb{I}(-\theta < x_i < \theta) \\ &= \left(\frac{1}{2\theta}\right)^n \prod_{i=1}^n \mathbb{I}(0 < |x_i| < \theta) \\ &= \left(\frac{1}{2\theta}\right)^n \cdot \mathbb{I}\left(\max_i\{|x_i|\} < \theta\right) \end{aligned}$$

For sufficiency, using the notation X^n to refer to $X_1, \ldots, X_n$, notice that if we define $T(X^n) = \max_i\{|X_i|\}$ then $T(x^n) = T(y^n)$ implies that

$$R = \frac{\left(\frac{1}{2\theta}\right)^n \mathbb{I}\left(\max_i |x_i| < \theta\right)}{\left(\frac{1}{2\theta}\right)^n \mathbb{I}\left(\max_i |x_i| < \theta\right)} = 1$$

and R does not depend on θ, meaning that T is sufficient. As the reverse implication also holds true by inspection, we have that

$$T(X^n) = \max_i \{|X_i|\}$$

is minimal sufficient.

In this context, the maximum likelihood estimator $\hat{\theta}_n$ is also the minimal sufficient statistic T.

1c) In order to explicitly show that the maximum likelihood estimator $\hat{\theta}_n$ converges in probability to the 'true' parameter θ of the Uniform$(-\theta, \theta)$ distribution, i.e. that $\hat{\theta}_n$ is a consistent estimator, we opt for directly showing that as $n \to \infty$, then

$$P(|\hat{\theta}_n - \theta| > \epsilon) \longrightarrow 0$$

for all $\epsilon > 0$.

Notice that we can simplify the left hand side of the above to get

$$\begin{aligned} P(|\hat{\theta}_n - \theta| > \epsilon) &= P\left(\{\hat{\theta}_n - \theta > \epsilon\} \cup \{-(\hat{\theta}_n - \theta) > \epsilon\}\right) \\ &= P(\hat{\theta}_n > \theta + \epsilon) + P(\hat{\theta}_n < \theta - \epsilon) \\ &= P(\hat{\theta}_n < \theta - \epsilon) \end{aligned}$$

because

$$X_i \sim \text{Uniform}(-\theta, \theta) \implies \max_i \{|X_i|\} \leq \theta \implies P(\hat{\theta}_n > \theta + \epsilon) = 0$$

To simplify this further, notice that

$$\begin{aligned} P(|\hat{\theta}_n - \theta| > \epsilon) &= P(\hat{\theta}_n < \theta - \epsilon) \\ &= P(\max_i \{|X_i|\} < \theta - \epsilon) \\ &= P\left(\bigcap_{i=1}^n \{|X_i| < \theta - \epsilon\}\right) \\ &= \prod_{i=1}^n P(|X_i| < \theta - \epsilon) \end{aligned}$$

Using the symmetry of the Uniform$(-\theta, \theta)$ about 0, the probability mass contained within the interval $[-(\theta-\epsilon), \theta-\epsilon]$ consists of a rectangle of height $1/2\theta$ and length $2(\theta-\epsilon)$, meaning that

$$\begin{aligned} P(|\hat{\theta}_n - \theta| > \epsilon) &= \prod_{i=1}^{n} P(|X_i| < \theta - \epsilon) \\ &= \left(\frac{\theta-\epsilon}{\theta}\right)^n \longrightarrow 0 \end{aligned}$$

as $n \to \infty$ for all $0 < \epsilon < \theta$, which is a consequence of the fact that

$$\frac{\theta-\epsilon}{\theta} < 1$$

In the case that $\epsilon \geq \theta$, then $P(|X_i| < \theta - \epsilon) = 0$ because the absolute value function is, by definition, non-negative. Hence as $n \to \infty$, we have that $P(|\hat{\theta}_n - \theta| > \epsilon) \longrightarrow 0$ for all $\epsilon > 0$, and hence $\hat{\theta}_n \xrightarrow{p} \theta$.

1d)

To find the limiting distribution of $n(\theta - \hat{\theta}_n)$, we begin by considering

$$\begin{aligned} P(n(\theta - \hat{\theta}_n) \leq t) &= P\left(\hat{\theta}_n \geq \theta - \frac{t}{n}\right) \\ &= P\left(\max_i\{|X_i|\} \geq \theta - \frac{t}{n}\right) \\ &= 1 - P\left(\max_i\{|X_i|\} \leq \theta - \frac{t}{n}\right) \\ &= 1 - P\left(\bigcap_{i=1}^{n}\left\{|X_i| \leq \theta - \frac{t}{n}\right\}\right) \\ &= 1 - \prod_{i=1}^{n} P\left(|X_i| \leq \theta - \frac{t}{n}\right) \end{aligned}$$

In order to evaluate the term within the product operator, we compute the CDF of a transformed $Y = r(X) = |X|$ where $X \sim$ Uniform$(-\theta, \theta)$. To find the CDF, we need to find the set $A_y = \{x : |x| \leq y\}$ for all y, which will yield

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\{x : |x| \leq y\}) \\ &= \int_{A_y} f_X(x)dx \end{aligned}$$

Noting that the transformed $y \in [0, \theta]$, the set A_y constitutes a line segment of length y, so that $A_y = y$. Meaning that

$$\int_{A_y} f_X(x)dx = \int_{A_y} \frac{1}{\theta} dx = \int_0^y \frac{1}{\theta} dx = \frac{y}{\theta}$$

The CDF of the transformed $Y = |X|$ is therefore

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \frac{y}{\theta} & 0 \leq y \leq \theta \\ 1 & y > \theta \end{cases}$$

which is the CDF for a $Y \sim \text{Uniform}(0, \theta)$ distribution. Returning to our previous expression, we have that

$$\begin{aligned} P(n(\theta - \hat{\theta}_n) \leq t) &= 1 - \prod_{i=1}^{n} P\left(|X_i| \leq \theta - \frac{t}{n}\right) \\ &= 1 - \prod_{i=1}^{n} F_Y\left(\frac{\theta - t/n}{\theta}\right) \\ &= 1 - \left(1 - \frac{t}{n\theta}\right)^n \end{aligned}$$

Using the power series representation of the exponential function together with the Binomial theorem, the exponential function has the following limit representation:

$$\exp(t) = \lim_{n \to \infty} \left(1 + \frac{t}{n}\right)^n$$

Consequently, we have that

$$\exp\left(-\frac{1}{\theta} t\right) = \lim_{n \to \infty} \left(1 - \frac{t}{n\theta}\right)^n$$

Considering the limit as $n \to \infty$ in the previous expression, we have that

$$\lim_{n \to \infty} P(n(\theta - \hat{\theta}_n) \leq t) = 1 - \lim_{n \to \infty} \left(1 - \frac{t}{n\theta}\right)^n = 1 - \exp\left(-\frac{1}{\theta} t\right)$$

Hence we have that

$$n(\theta - \hat{\theta}_n) \xrightarrow{d} \text{Exponential}\left(\frac{1}{\theta}\right)$$

i.e. the limiting distribution of $n(\theta - \hat{\theta}_n)$ is an exponential distribution with parameter $1/\theta$.

2) Let $X \sim \text{Bernoulli}(\theta)$ be a single coin flip.

Suppose that $\theta \in \Theta = \{1/3, 2/3\}$. Hence θ can only take two possible values.

a) Find the maximum likelihood estimator.

b) Let the loss function be:

$$L(\theta, \widehat{\theta}) = \begin{cases} 1 & \text{if} \quad \theta \neq \widehat{\theta} \\ 0 & \text{if} \quad \theta = \widehat{\theta} \end{cases}$$

Find the risk function of the maximum likelihood estimator. Since θ only takes the values $1/3$ and $2/3$, you only need to find $R(1/3, \widehat{\theta})$ and $R(2/3, \widehat{\theta})$.

c) Show that the maximum likelihood estimator is minimax.

2a)

The maximum likelihood estimator in this instance will be $\widehat{\theta} = \widehat{\theta}(X) = X$, that is, our observation of the single coin flip.

The log likelihood function is:

$$l(\theta) = x \log \theta - (1 - x) \log(1 - \theta)$$

Maximising this by computing derivatives, setting to 0 and solving for θ, we have that:

$$l'(\theta) = \frac{x}{\theta} - \frac{1 - x}{(1 - \theta)} = 0 \implies (1 - \theta)x - \theta(1 - x) = 0 \implies \widehat{\theta} = x$$

2b)

The loss function specified is known as zero-one loss. The risk function $R(\theta, \widehat{\theta})$ is the average loss incurred by the estimator $\widehat{\theta}$ over all possible values that the data X can take:

$$\begin{aligned} R(\theta, \widehat{\theta}(X)) &= \mathbb{E}_\theta[L(\theta, \widehat{\theta}(X)] \\ &= \sum_x L(\theta, \widehat{\theta}(x)) f_X(x; \theta) \\ &= L(\theta, \widehat{\theta}(0)) \cdot P(X = 0) + L(\theta, \widehat{\theta}(1)) \cdot P(X = 1) \\ &= (1 - \theta) \cdot \mathbb{I}(\theta \neq 0) + \theta \cdot \mathbb{I}(\theta \neq 1) \end{aligned}$$

Which is the risk function of the maximum likelihood estimator. We now have that:

$$R(1/3, \widehat{\theta}) = \frac{2}{3}\mathbb{I}(1/3 \neq 0) + \frac{1}{3}\mathbb{I}(1/3 \neq 1) = 1$$

And that:

$$R(2/3, \widehat{\theta}) = \frac{1}{3}\mathbb{I}(2/3 \neq 0) + \frac{2}{3}\mathbb{I}(2/3 \neq 1) = 1$$

Hence over the restricted parameter space $\Theta = \{1/3, 2/3\}$, the risk function $R(\cdot, \widehat{\theta})$ for the maximum likelihood estimator $\widehat{\theta}$ is constant.

2c)

To show that the maximum likelihood estimator $\widehat{\theta}$ is minimax, we use the result that Bayes estimators with respect to a prior π that have a constant risk function are minimax estimators. We now assume that the parameter of interest θ is a random variable.

Under zero-one loss, the Bayes estimator is the posterior mode.

Noting that the likelihood function is a Bernoulli distribution, we use its conjugate prior, the Beta distribution, $\pi(\theta) = \text{Beta}(\alpha, \beta)$ for hyperparameters α, β. Omitting the normalisation constants, we have that:

$$\begin{aligned} p(\theta|X) &\propto p(X|\theta)\pi(\theta) \\ &\propto \theta^X(1-\theta)^{1-X}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{X+\alpha-1}(1-\theta)^{\beta-X} \end{aligned}$$

Hence the posterior distribution is $\text{Beta}(X+\alpha, \beta - X + 1)$. As the mode of a $\text{Beta}(\alpha, \beta)$ distribution is $(\alpha - 1)/(\alpha + \beta + 2)$, we have that the posterior mode, and hence Bayes estimator under a general $\text{Beta}(\alpha, \beta)$ prior is:

$$\frac{X + \alpha - 1}{(X + \alpha) + (\beta - X + 1) - 2} = \frac{X + \alpha - 1}{\alpha + \beta - 1}$$

Hence the maximum likelihood estimator $\widehat{\theta}(X) = X$ is the posterior mode under a $\text{Beta}(1, 1)$ prior, which is also a uniform prior $\pi(\theta) = 1$.

And so for zero-one loss, $\widehat{\theta}(X) = X$ is the Bayes estimator for a uniform prior $\pi(\theta) = 1$. As the risk function $R(\cdot, \widehat{\theta})$ is constant, $\widehat{\theta}$ is minimax.

3) Let $X_1, \ldots, X_n$ be IID with distribution $\text{Binomial}(k, \theta)$.

a) Find a minimal sufficient statistic S for θ.

b) For each of the following say whether it is sufficeint, minimal sufficient, or not sufficient:

$$T = X_1, \quad T = \sum_i X_i, \quad T = \left(X_1, \sum_i X_i\right), \quad T = \left(X_1, \sum_{i=1}^n X_i\right)$$

c) Let $\tau(\theta) = P(X = 1) = k\theta(1-\theta)^{k-1}$. Define $U = 1$ if $X_1 = 1$ and 0 otherwise. Show that U is an unbiased estimator of τ.

d) Find the maximum likelihood estimator $\widehat{\tau}$ of τ.

e) Show that $\widehat{\tau} \xrightarrow{p} \tau$.

f) Find the limiting distribution of $\widehat{\tau}$.

3a)

The parameter of interest is θ, so assume that the number of trials k is fixed and known.

The Binomial PMF is:

$$f_{X_i}(x_i;\theta) = P(X_i = x_i) = \binom{k}{x_i}\theta^{x_i}(1-\theta)^{k-x_i}$$

Given two datasets $x^n = X_1, ...X_n$ and $y^n = Y_1, ..., Y_n$, we compute:

$$R(x^n, y^n; \theta) = \frac{P(x_n;\theta)}{P(y^n;\theta)}$$

Which involves computing the following joint PMF, or probability as the X_i are discrete:

$$\begin{aligned} f_{X^n}(x^n;\theta) &= \prod_{i=1}^{n}\binom{k}{x_i}\theta^{x_i}(1-\theta)^{k-x_i} \\ &= \frac{(k!)^n}{\prod_{i=1}^{n} x_i!(k-x_i)!}\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{nk-\sum_{i=1}^{n} x_i} \end{aligned}$$

Hence we have that:

$$R(x^n, y^n; \theta) = \frac{\frac{(k!)^n}{\prod_{i=1}^{n} x_i!(k-x_i)!}\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{nk-\sum_{i=1}^{n} x_i}}{\frac{(k!)^n}{\prod_{i=1}^{n} y_i!(k-y_i)!}\theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{nk-\sum_{i=1}^{n} y_i}}$$

Now if R does not depend on θ, then it must be the case that $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$. The reverse implication also holds true, and R will be some constant c that does not depend on θ.

Hence $T(X^n) = \sum_{i=1}^{n} X_i$ is a minimal sufficient statistic.

3b)

Given that we have found a minimal sufficient statistic (MSS) to be $T(X^n) = \sum_{i=1}^{n} X_i$, we can use the definition that the MSS must be expressible as a function of another sufficient statistic U, that is it must be the case that $T = g(U)$ for some function $g(\cdot)$.

We can use this to test whether another statistic U is sufficient. If the MSS cannot be expressed as a function of U, then U is not sufficient. If the MSS can be expressed as a function of U then U is sufficient. If it is the case that the MSS can be expressed as a function of U, where the function $g(\cdot)$ is 1-to-1, then T and U are equivalent, and U is also an MSS.

Hence:

$T = X_1$ is not sufficient because no function $g(\cdot)$ can be applied to it to get $S = \sum_{i=1}^{n} X_i$.

$T = \sum_i X_i$ is minimal sufficient because it is exactly equivalent to the MSS.

$\mathbf{T} = (X_1, \sum_i X_i)$ is sufficient because the MSS can be expressed as $S = \sum_{i=1}^{n} X_i = g(\mathbf{T})$ where:

$$g(T) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} X_1 \\ \sum_i X_i \end{pmatrix}$$

However, it is not minimal sufficient, because we can replace X_1 in the 1st entry of $\mathbf{T}$ with any other X_i to still yield the same MSS, $S = \sum_{i=1}^{n}$. Therefore the function $g(\cdot)$ is not one-to-one, and therefore $\mathbf{T}$ is not minimal sufficient, as it contains redundant information.

$\mathbf{T} = (X_1, \sum_{i=2}^{n}) X_i$ is minimal sufficient, as we note that the MSS can be expressed as $S = \sum_{i=1}^{n} X_i = h(\mathbf{T})$, where:

$$h(\mathbf{T}) = \left(X_1, \sum_{i=2}^{n} X_i\right)^T \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

which is one-to-one.

3c)

An estimator $\widetilde{\theta}$ of a parameter θ is unbiased if we have that:

$$\mathbb{E}_\theta[\widetilde{\theta}] = \theta$$

Where the expectation is with respect to the ***

Noting that U is a Bernoulli random variable with parameter $\tau(\theta)$, we have that

$$\mathbb{E}_{\tau(\theta)}[U] = \mathbb{E}_{\tau(\theta)}[\mathbb{I}(X_1 = 1)] = 1 \cdot \tau(\theta) + 0 \cdot (1 - \tau(\theta)) = \tau(\theta)$$

Hence U is an unbiased estimator of τ.

3d)

We first compute the maximum likelihood estimator $\widehat{\theta}$ of θ. The log-likelihood is:

$$l(\theta) = \log f_{X^n}(x^n; \theta) = \sum_{i=1}^{n} \log \binom{k}{x_i} + \left(\sum_{i=1}^{n} x_i\right) \log \theta + \left(nk - \sum_{i=1}^{n} x_i\right) \log(1 - \theta)$$

Taking derivatives, setting to 0, and solving for the parameter to maximise the log-likelihood, we have

$$
\begin{aligned}
l'(\theta) &= \frac{\sum_{i=1}^n x_i}{\theta} - \frac{nk - \sum_{i=1}^n x_i}{1-\theta} \\
&= \frac{(1-\theta)\sum_{i=1}^n x_i - \theta\left(nk - \sum_{i=1}^n x_i\right)}{\theta(1-\theta)} \\
&= \sum_{i=1}^n x_i - \theta\sum_{i=1}^n x_i - \theta nk + \theta\sum_{i=1}^n x_i = 0 \\
\implies \widehat{\theta} &= \frac{\sum_{i=1}^n x_i}{nk}
\end{aligned}
$$

Hence the maximum likelihood estimator is $\widehat{\theta} = k^{-1}\bar{X}_n$ which is the proportion of success events out of the maximum number of successes observable in n experiments, where each experiment consists of k trials.

Noting that τ is a function of the parameter θ, we can use the equivariance of the MLE to get the maximum likelihood estimator $\widehat{\tau}$ of τ, using our estimate $\widehat{\theta}$:

$$
\widehat{\tau} = \tau(\widehat{\theta}) = k\widehat{\theta}(1-\widehat{\theta})^{k-1} = k\left(\frac{\sum_{i=1}^n X_i}{nk}\right)\left(1 - \frac{\sum_{i=1}^n X_i}{nk}\right)^{k-1} = \bar{X}_n\left(1 - \frac{1}{k}\bar{X}_n\right)^{k-1}
$$

3e)

The weak law of large numbers (WLLN) states that a sample mean $\bar{X}_n$ of IID random variables $X_1, \ldots, X_n$ converges in probability to their mean $\mathbb{E}[X_i] = \mu$, $\bar{X}_n \xrightarrow{p} \mu$. Hence we have that:

$$
\bar{X}_n \xrightarrow{p} k\theta
$$

because the X_i are Binomial distributed with mean $\mathbb{E}[X_i] = k\theta$. Noting that $\widehat{\tau} = h(\bar{X}_n) = \bar{X}_n\left(1 - \frac{1}{k}\bar{X}_n\right)^{k-1}$, that $h(\cdot)$ is a polynomial of degree k and continuous, we have via the continuous mapping theorem:

$$
h(\bar{X}_n) \xrightarrow{p} h(k\theta) = k\theta(1-\theta)^{k-1} = \tau(\theta)
$$

Hence $\widehat{\tau} \xrightarrow{p} \tau$.

3f)

Under appropriate regularity conditions and assuming this for the entirety of the answer, the maximum likelihood estimator $\widehat{\theta}$ is asymptotically Normal, that is:

$$
\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)
$$

where $I(\theta)$ is the Fisher information for a single data point X_i.

We can use the delta method to yield the asymptotic distribution of smooth functions of estimators with known limiting distribution, and it states:

$$\frac{\widehat{\tau}-\tau}{\text{se}(\tau)} \xrightarrow{d} N(0,1) \implies \tau(\widehat{\theta}) \xrightarrow{d} N\left(\tau(\theta), \frac{\tau'(\theta)^2}{I_n(\theta)}\right)$$

where $\text{se}(\cdot)$ is the (asymptotic) standard error, and $I_n(\theta)$ is the Fisher information for n data points $X_1, ..., X_n$.

Computing $\tau'(\theta)^2$:

$$\begin{aligned}
\tau'(\theta)^2 &= \left\{\frac{\partial}{\partial\theta} k\theta(1-\theta)^{k-1}\right\}^2 \\
&= \left\{-k\theta(k-1)(1-\theta)^{k-2} + k(1-\theta)^{k-1}\right\}^2 \\
&= \left\{k(1-\theta)^{k-2}(1-k\theta)\right\}^2
\end{aligned}$$

Now the Fisher information $I_n(\theta)$ is attained by computing negative expectations of the second derivative of the log-likelihood, that is $I_n(\theta) = -\mathbb{E}_\theta[l''(\theta)]$. Using our results from 3d), and setting $S = \sum_{i=1}^n x_i$ we first compute $l''(\theta)$:

$$\begin{aligned}
l''(\theta) &= \frac{\partial}{\partial\theta} l'(\theta) \\
&= \frac{\partial}{\partial\theta} \frac{(1-\theta)S - \theta(nk-S)}{\theta(1-\theta)} \\
&= \frac{\partial}{\partial\theta} \frac{S-\theta nk}{\theta(1-\theta)} \\
&= \frac{-nk\theta(1-\theta) - (S-\theta nk)(1-2\theta)}{\theta(1-\theta)} \\
\implies l''(\theta) &= \frac{-\theta^2 nk + 2\theta S - S}{[\theta(1-\theta)]^2}
\end{aligned}$$

Taking negative expectations, we have that the Fisher information is:

$$I_n(\theta) = -\mathbb{E}[l''(\theta)] = -\frac{1}{[\theta(1-\theta)]^2}\mathbb{E}[-\theta^2 nk + 2\theta S - S] = \frac{\theta^2 nk - 2\theta^2 nk + \theta nk}{[\theta(1-\theta)]^2} = \frac{nk}{\theta(1-\theta)}$$

Where we have used the fact that $\mathbb{E}[S] = \mathbb{E}[n\bar{X}_n] = \theta nk$.

Hence the asymptotic variance of $\tau(\widehat{\theta})$ is:

$$\frac{\tau'(\theta)^2}{I_n(\theta)} = \frac{\left[k(1-\theta)^{k-2}(1-k\theta)\right]^2 \theta(1-\theta)}{nk} = \frac{k\theta(1-\theta)^{2k-3}(1-k\theta)^2}{n}$$

And $\widehat{\tau}$ has the limiting distribution:

$$\widehat{\tau} \xrightarrow{d} N\left(k\theta(1-\theta)^{k-1}, \frac{k\theta(1-\theta)^{2k-3}(1-k\theta)^2}{n}\right)$$

We can also estimate the asymptotic variance by evaluating it at $\widehat{\theta} = k^{-1}\bar{X}_n$, yielding the following limiting distribution for $\widehat{\tau}$:

$$\widehat{\tau} \xrightarrow{d} N\left(k\theta(1-\theta)^{k-1}, \frac{k\bar{X}_n(1-k^{-1}\bar{X}_n)^{2k-3}(1-\bar{X}_n)^2}{n}\right)$$

4) Construct an example where $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, but $X_n + Y_n$ does not converge in distribution to $X + Y$.

5) Let $X_i \sim N(\theta_i, 1)$ for $i = 1, 2, \ldots, n$. Let $\gamma = \sum_{i=1}^n \theta_i^2$.

Find the maximum likelihood estimator $\widehat{\gamma}$.

Let $L(\gamma, \widehat{\gamma}) = (\gamma - \widehat{\gamma})^2$. Find the risk of the maximum likelihood estimator.

We assume that the X_i are independent; but it is clear that they are not identically distributed.

Denoting $\boldsymbol{\theta} = (\theta_1, ..., \theta_n) \in \mathbb{R}^n$, we first compute the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, ..., \widehat{\theta}_n)$. The log-likelihood function is:

$$l(\boldsymbol{\theta}) = \log \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \theta_i)^2\right) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n (x_i - \theta_i)^2$$

Taking derivatives with respect to each θ_i, we have that:

$$\frac{\partial l}{\partial \theta_i} = \frac{1}{2}2(x_i - \theta_i) = x_i - \theta_i = 0 \implies \widehat{\theta}_i = x_i$$

Hence we have that the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (x_1, ..., x_n)$, which in this case is just the data that has been observed.

Defining the scalar valued function $g : \mathbb{R}^n \to \mathbb{R}$, with $g(\mathbf{u}) = \mathbf{u}^T\mathbf{u}$, we have that $\gamma = g(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\boldsymbol{\theta} = \sum_{i=1}^n \theta_i^2$. Hence denoting the data as $\mathbf{x} = (x_1, ..., x_n)$, via equivariance of the maximum likelihood estimator, we have that:

$$\widehat{\gamma} = g(\widehat{\boldsymbol{\theta}}) = \mathbf{x}^T\mathbf{x} = \sum_{i=1}^n x_i^2$$

The risk function $R(\gamma, \widehat{\gamma})$ is the expected value of the loss function $L(\gamma, \widehat{\gamma})$:

$$R(\gamma, \widehat{\gamma}) = \mathbb{E}_\gamma[L(\gamma, \widehat{\gamma})]$$

Under L_2, or squared-error loss, $L(\gamma, \widehat{\gamma}) = (\gamma - \widehat{\gamma})^2$, the risk function is just the mean squared error (MSE), $\mathbb{E}_\gamma[(\gamma - \widehat{\gamma})^2]$. As the mean-squared error is the sum of the variance V and the squared bias B^2, we have:

$$\text{MSE} = B^2 + V = (\mathbb{E}_\gamma[\widehat{\gamma}] - \gamma)^2 - \text{Var}_\gamma[\widehat{\gamma}]$$

Computing the expectations of the estimator $\widehat{\gamma}$:

$$\mathbb{E}_\gamma[\widehat{\gamma}] = \mathbb{E}\left[\sum_{i=1}^{n} x_i^2\right] = \sum_{i=1}^{n} \mathbb{E}[x_i^2] = \sum_{i=1}^{n} \text{Var}(x_i) + \mathbb{E}[x_i]^2 = \sum_{i=1}^{n} (1 + \theta_i)^2 = n + \sum_{i=1}^{n} \theta_i^2$$

And so the squared bias is:

$$B^2 = \left(n + \sum_{i=1}^{n} \theta_i^2 - \sum_{i=1}^{n} \theta_i^2\right)^2 = n^2$$

Computing the variance of the estimator relies on the observation that it is the sum of squares of n independent Normally distributed random variables, each with individual mean θ_i and unit variance. Hence the estimator will have a non-central Chi-squared distribution with n degrees of freedom and non-centrality parameter $\lambda = \sum_{i=1}^{n} \theta_i^2$.

Stating this in a multivariate form, we have that for a multivariate Normal random vector $\mathbf{X} = (X_1, ...X_n)$ with mean $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ and isotropic covariance $\Sigma = \sigma^2 I$, it is the case that:

$$\frac{\|\mathbf{X}\|_2^2}{\sigma^2} = \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} \sim \chi_n^2\left(\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{\sigma^2}\right)$$

Setting $\sigma^2 = 1$ yields the non-centrality parameter $\lambda = \boldsymbol{\theta}^T\boldsymbol{\theta} = \sum_{i=1}^{n} \theta_i^2$. The variance of this distribution is $2(n + 2\lambda)$, and so we have that the risk function is:

$$R(\gamma, \widehat{\gamma}) = n^2 + 2\left(n + 2\sum_{i=1}^{n} \theta_i^2\right)$$

6) Let $X_1, \ldots, X_n \sim \text{Uniform}(0, 2)$. Let

$$W_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

a) Show that there is a number μ such that W_n converges in quadratic mean to μ.

b) Show that W_n converges in probability to μ.

c) What is the limiting distribution of $\sqrt{n}(W_n^2 - \mu^2)$?

6a)

The PDF of $X_i \sim \text{Uniform}(0, 2)$ is:

$$f_{X_i}(x_i) = \begin{cases} \frac{1}{2} & 0 \leq x_i \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

To show that W_n converges in quadratic mean to μ, that is, $W_n \overset{\text{q.m.}}{\to} \mu$, it is necessary to show that as $n \to \infty$:

$$\mathbb{E}[(W_n - \mu)^2] \to 0$$

Defining the transformation $Y = r(X) = X^2$, notice that we can express W_n as a sample mean $\bar{Y}_n$ of the transformed IID random variables $Y_1, ..., Y_n$:

$$W_n = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}_n$$

We now need to ascertain the properties of Y by computing its mean and variance. We evaluate the CDF and PDF in order to do so for practice, but a quicker method would be to use the law of the unconscious statistician:

The CDF of Y can be found by considering:

$$F_Y(y) = P(r(X) \leq y) = P(\{x : x^2 \leq y\}) = \int_{A_y} f_X(x)dx$$

Where the integral is over the set $A_y = \{x : r(x) \leq y)\}$. The support of X is $0 \leq x \leq 2$, so we consider $0 \leq y \leq 4$. Hence we have that:

$$\begin{aligned} F_Y(y) &= P(\{x : x^2 \leq y\}) \\ &= P(\{x : \sqrt{y} \leq x \leq \sqrt{y}\}) \\ &= \int_{\{x : \sqrt{y} \leq x \leq \sqrt{y}\}} f_X(x)dx \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= \frac{1}{2}\sqrt{y} - \left(-\frac{1}{2}\sqrt{y}\right) \\ &= \sqrt{y} \end{aligned}$$

And so the CDF of Y is:

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \sqrt{y} & 0 \leq y \leq 4 \\ 1 & y > 4 \end{cases}$$

And taking derivatives, we have the following PDF of Y:

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} & 0 < y < 4 \\ 0 & \text{otherwise} \end{cases}$$

We have that the mean and variance are given by:

$$\mathbb{E}[$$

Defining the transformation $Y = r(X) = X^2$, notice that we can express W_n as a sample mean $\bar{Y}_n$ of the transformed IID random variables $Y_1, ..., Y_n$:

$$W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n$$

We now want to compute the mean and variance of Y. The Uniform(a, b) distribution has mean $(a+b)/2$ and variance $(b-a)^2/12$, and as the mean of Y is also the 2nd moment of X, that is, $\mathbb{E}[Y] = \mathbb{E}[X^2]$, we have that $\mathbb{E}[Y] = \text{Var}[X] + \mathbb{E}[X]^2 = 4/12 + 1^2 = 4/3 = \mu_Y$.

We compute the variance of the Y using the law of the unconscious statistician:

$$\begin{aligned}
\text{Var}[Y] &= \mathbb{E}[(Y - \mu_Y)^2] \\
&= \mathbb{E}[Y^2 - 2\mu_Y Y + \mu_Y^2] \\
&= \mathbb{E}[X^4 - 2\mu_Y X^2 + \mu_Y^2] \\
&= \int_0^2 (x^4 - 2\mu_Y x^2 + \mu_Y^2) f_X(x) dx \\
&= \int_0^2 \frac{1}{2}x^4 - \mu_Y x^2 + \frac{1}{2}\mu_Y^2 dx \\
&= \left[\frac{1}{10}x^5 - \frac{\mu_Y}{3}x^3 - \frac{1}{2}\mu_Y^2 x\right]_0^2 \\
&= \frac{1}{10}(2)^5 - \frac{4}{9}(2)^3 - \frac{1}{2}\frac{16}{9}(2) = \frac{64}{45} = \sigma_Y^2
\end{aligned}$$

Now we have that:

$$\mathbb{E}[(W_n - \mu)^2] = \mathbb{E}[(\bar{Y}_n - \mu_Y)^2] = \mathbb{E}[\bar{Y}_n^2 - 2\mu_Y \bar{Y}_n + \mu_Y^2] = \mathbb{E}[\bar{Y}_n^2] - 2\mu_Y \mathbb{E}[\bar{Y}_n] + \mu_Y^2$$

As $\bar{Y}_n$ is a sample mean, its mean, variance and second moment can be specified in terms of the underlying random variables Y_i, and so we have that:

$$\mathbb{E}[\bar{Y}_n] = \mu_Y, \quad \text{Var}[\bar{Y}_n] = \frac{\sigma_Y^2}{n}, \quad \mathbb{E}[\bar{Y}_n^2] = \frac{\sigma_Y^2}{n} + \mu_Y^2$$

Hence we have that:

$$\mathbb{E}[(W_n - \mu)^2] = \mathbb{E}[(\bar{Y}_n - \mu_Y)^2] = \left(\frac{\sigma_Y^2}{n} + \mu_Y^2\right) - 2\mu_Y^2 + \mu_Y^2 = \frac{64}{45n} \to 0$$

as $n \to \infty$, which proves the required result, $W_n \overset{\text{q.m.}}{\to} \mu$ and where the number $\mu = \mu_Y = 4/3$.

6b)

Because of the way we have set up the problem, the convergence in probability result, $W_n \overset{p}{\to} \mu$ is a consequence of the weak law of large numbers (WLLN) applied to the sample mean $\bar{Y}_n$, so that we have:

$$\bar{Y}_n \overset{p}{\to} \mu_Y = \frac{4}{3}$$

Alternatively, can use Hoeffding's inequality, which states that for bounded random variables $a \leq Y_i \leq b$ with mean μ_Y:

$$P(|\bar{Y}_n - \mu_Y| > \epsilon) \leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

Hence we have that

$$P(|W_n - \mu| > \epsilon) = P(|\bar{Y}_n - \mu_Y| > \epsilon) \leq 2\exp\left(\frac{-n\epsilon^2}{8}\right) \longrightarrow 0$$

as $n \to \infty$ and so $W_n \overset{p}{\to} \mu$ at a rate of ***.

6c)

As the sample mean Y_n is asymptotically Normal with mean μ_Y and variance σ_Y^2/n, we define the function $g(u) := u^2$, which is smooth, and using the delta method, which states:

$$\sqrt{n}(g(\bar{Y}_n) - g(\mu_Y) \overset{d}{\to} N\left(0, \sigma_Y^2 g'(\mu_Y)^2\right)$$

We have that:

$$g'(\mu_Y)^2 = (2\mu_Y)^2 \implies \sqrt{n}(W_n^2 - \mu^2) \overset{d}{\to} N(0, (2\mu_Y\sigma_Y)^2)$$

And hence the $\sqrt{n}(W_n^2 - \mu)$ is asymptotically Normal with mean 0 and variance $(2\mu_Y\sigma_Y)^2$, where $\mu_Y = \mu = 4/3$ and $\sigma_Y^2 = 64/45$.

7) Let $X_1, \ldots X_n \sim$ Bernoulli(p).

a) Let $T = X_3$. Show that T is not sufficient.

b) Show that $U = \sum_{i=1}^n X_i^2$ is minimal sufficient.

7a) and 7b)

We first show that $U = \sum_{i=1}^n X_i^2$ is minimal sufficient before showing that $T = X_3$ is not sufficient.

For $X_i \sim$ Bernoulli(p), we have that $P(X_i = 1) = p$ and $P(X_i = 0) = (1-p)$ with PMF $f_{X_i}(x_i) = p^{x_i}(1-p)^{1-x_i}$.

Using similar arguments to 3a), computing $R(x^n, y^n; p)$ yields:

$$R(x^n, y^n; p) = \frac{\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}}{\prod_{i=1}^n p^{y_i}(1-p)^{1-y_i}} = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}} = \frac{p^{\sum_{i=1}^n x_i^2}(1-p)^{n-\sum_{i=1}^n x_i^2}}{p^{\sum_{i=1}^n y_i^2}(1-p)^{n-\sum_{i=1}^n y_i^2}}$$

Where we have used the fact that because X_i is a Bernoulli random variable, $X_i^2 = X_i$ in going from the second to third equality.

If we set $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$, then $R(x^n, y^n; p) = 1$ which does not depend on the parameter p. Furthermore, the only way in which $R(x^n, y^n; p)$ does not depend on p is when it is equal to 1, and this occurs when $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$.

Hence $U(X^n) = \sum_{i=1}^n X_i^2$ is a minimal sufficient statistic.

Using similar arguments to 3b), we have that $T = X_3$ is not sufficient because we cannot express a known minimal sufficient statistic $U = \sum_{i=1}^n X_i^2$ as a function of T. That is, there exists no function g such that $U = g(T)$.

8) Let:

$$X_1, \ldots, X_n \sim \frac{1}{2}N(0,1) + \frac{1}{2}N(\theta, 1)$$

In other words, with probability $1/2$, X_i is drawn from $N(0,1)$ and with probabiltiy $1/2$, X_i is drawn from $N(\theta, 1)$.

a) Find the method of moments estimator $\widehat{\theta}$ of θ.

b) Find the mean squared error of θ.

8a)

It is necessary to note that each X_i is a mixture of Normal distributions. We denote the mixture component Normal probability densities as $f_{X_{i1}}(x_i) = N(0,1)$ and $f_{X_{i2}}(x_i) = N(\theta, 1)$, and the mixture weights are $(1/2, 1/2)$.

Hence the PDF of X_i is:

$$f_{X_i}(x_i) = \frac{1}{2}f_{X_{i1}}(x_i) + \frac{1}{2}f_{X_{i2}}(x_i) = \frac{1}{2}\left(\frac{1}{\sqrt{2\pi}}e^{-x_i^2/2}\right) + \frac{1}{2}\left(\frac{1}{\sqrt{2\pi}}e^{-(x_i-\theta_i)^2/2}\right)$$

To compute the method of moments estimator $\widehat{\theta}$ of θ, we equate the kth theoretical moment $\mathbb{E}[X^k]$ with the kth sample moment $m_k = \frac{1}{n}\sum_{i=1}^n X_i^k$. As there is only one parameter θ for which we are looking to compute an estimator, this means we only require the 1st moment i.e. the expectation and sample mean.

Computing the mean, we have:

$$
\begin{aligned}
\mathbb{E}[X_i] &= \int_{\infty}^{\infty} x_i f_{X_i}(x_i) dx_i \\
&= \int_{\infty}^{\infty} x_i \left(\frac{1}{2} f_{X_{i1}}(x_i) + \frac{1}{2} f_{X_{i2}}(x_i) \right) dx_i \\
&= \frac{1}{2} \int_{\infty}^{\infty} x_i f_{X_{i1}}(x_i) dx_i + \frac{1}{2} \int_{\infty}^{\infty} x_i f_{X_{i2}}(x_i) dx_i \\
&= \frac{1}{2} \mathbb{E}[X_{i1}] + \frac{1}{2} \mathbb{E}[X_{i2}] \\
&= \frac{1}{2}(0) + \frac{1}{2}(\theta) \\
&= \frac{1}{2}\theta
\end{aligned}
$$

Equating $\mathbb{E}[X_i]$ with the sample mean $m_1 = \bar{X}_n$:

$$
\bar{X}_n = \frac{1}{2}\theta
$$

Hence the method of moments estimator is:

$$
\widehat{\theta} = 2\bar{X}_n
$$

8b)

The mean squared error (MSE) of an estimator is the sum of its variance and its squared bias:

$$
\text{MSE} = \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] = B^2 + V = \left(\mathbb{E}_\theta[\widehat{\theta}] - \theta)^2 + \text{Var}_\theta[\widehat{\theta}] \right)
$$

Suppressing θ from the expectation, the squared bias is:

$$
B^2 = \left(\mathbb{E}[2\bar{X}_n] - \theta \right)^2 = \left(2\mathbb{E}[\bar{X}_n] - \theta \right)^2 = \left(2 \cdot \frac{1}{2}\theta - \theta \right)^2 = 0 \implies B = 0
$$

Hence the method of moments estimator for this particular mixture of Normal distributions is unbiased.

In order to compute the variance of $\widehat{\theta}$, we compute the 2nd theoretical moment of X_i, and in addition the variance of the latter. Following similar calculations in 8a) we have that:

$$
\begin{aligned}
\mathbb{E}[X_i^2] &= \frac{1}{2}\mathbb{E}[X_{i1}^2] + \frac{1}{2}\mathbb{E}[X_{i2}^2] \\
&= \frac{1}{2}\left(\text{Var}[X_{i1}] + \mathbb{E}[X_{i1}]^2 \right) + \frac{1}{2}\left(\text{Var}[X_{i2}] + \mathbb{E}[X_{i2}]^2 \right) \\
&= \frac{1}{2}\left(1 + 0^2 \right) + \frac{1}{2}\left(1 + \theta^2 \right) \\
&= 1 + \frac{1}{2}\theta^2
\end{aligned}
$$

Hence the variance of X_i is:

$$\text{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = \left(1 + \frac{1}{2}\theta^2\right) - \left(\frac{\theta}{2}\right)^2 = 1 + \frac{1}{4}\theta^2 = \sigma^2$$

As $\text{Var}[\bar{X}_n] = \sigma^2/n$, we have the variance of the estimator $\widehat{\theta}$:

$$\text{Var}[2\bar{X}_n] = 4\text{Var}[\bar{X}_n] = \frac{4\sigma^2}{n} = \frac{4\left(1 + \frac{\theta^2}{4}\right)}{n} = \frac{4 + \theta^2}{n} = \text{MSE}$$

And because the estimator is unbiased, $B = 0$, the mean squared error is equal to the variance of the estimator.

9) Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Let $\tau = e^\theta + 1$.

a) Consider some loss function $L(\tau, \widehat{\tau})$. Define what it means for an estimator to be the minimax estimator for τ.

b) Let π be some prior for θ. Find the Bayes estimator for τ under the loss $L(\tau, \widehat{\tau}) = (\widehat{\tau} - \tau)^2/\tau$.

9a)

The minimax estimator $\widetilde{\tau}$ minimises the maximum risk. That is, out of all possible estimators $\widehat{\tau}$, each with their own maximum risk $\bar{R}(\widehat{\tau})$, where the maximum risk is with respect to the parameter τ, the maximum risk of the minimax estimator $\bar{R}(\widetilde{\tau})$ is the lowest out of all possible estimators $\widehat{\tau}$.

Formally $\widetilde{\tau}$ is the minimax estimator if:

$$\bar{R}(\widetilde{\tau}) = \sup_\tau R(\tau, \widetilde{\tau}) = \inf_{\widehat{\tau}} \sup_\tau R(\tau, \widehat{\tau})$$

And that is with respect to a particular specification of the loss function $L(\tau, \widehat{\tau})$. It can be viewed as selecting an estimator on the basis of its worst case behaviour.

There is some machinery to unpack here, and we take this opportunity to do so whilst clarifying the role of the loss function, risk function, maximum risk, and minimax risk.

A loss function $L(\tau, \widehat{\tau})$ is a measure of the "quality" of an estimator, and we can endow it with particular functional forms to capture how we want to "penalise" an estimator $\widehat{\tau}$ deviating from the parameter τ, or how we wish to assign numerical values to discrepancies between $\widehat{\tau}$ and τ.

The loss function is a function of both the unknown parameter τ, and the estimator $\widehat{\tau}(X_1, \ldots, X_n)$. The loss function is a random quantity because it is dependent on the data $X_1, ..., X_n$ through the estimator. Formally the loss function is mapping from $T \times T \to \mathbb{R}$. In the context of statistical inference, where we have observed the data as the outcome of some experiment, that is $X_1 = x_1, ..., X_n = x_n$, our estimator becomes an estimate $\widehat{\tau}(x_1, \ldots, x_n)$. However, as we do not know the value of the loss function due to the presence of the unknown parameter τ, we cannot compute it.

Having specified a loss function, the risk function $R(\tau, \widehat{\tau})$ is the average loss incurred by the estimator $\widehat{\tau}$, that is, $\mathbb{E}_\tau[L(\tau, \widehat{\tau})]$. We are interested in this quantity because we are interested in the performance of an estimator not for one particular dataset $X^n = x^n$, but all possible X^n. By taking expectations with respect to the joint distribution that generated the data, i.e. $p(X_1, ..., X_n; \theta)$, we

average out the data, and for a particular estimator $\widehat{\tau}$, the risk function $R(\cdot, \widehat{\tau})$ is a function of the parameter, mapping from $T \to \mathbb{R}^+$.

Given a particular estimator $\widehat{\tau}$, we cannot compute the value of the risk function, as the true parameter τ is unknown, and also fixed. However, as it is unknown we now imagine that the the parameter τ can vary within the range defined by the parameter space T, and specify the risk function in terms of τ.

Hence for a number of different estimators $\widehat{\tau}_1, \cdots, \widehat{\tau}_n$, each will yield an individual risk function $R_i(\cdot, \widehat{\tau}_i)$ whose output varies as a function of the parameter τ over the parameter space T. In order to compare risk functions of different estimators, we rely on one-number summaries such as the maximum risk $\bar{R}(\widehat{\tau}) = \sup_\tau R_i(\tau, \widehat{\tau}_i)$, which is a maximum with respect to the parameter τ, for a particular estimator $\widehat{\tau}_i$.

The minimax estimator $\widetilde{\tau}$ is that estimator whose maximum risk $\bar{R}(\widetilde{\tau})$ is the lowest amongst all possible estimators $\widehat{\tau}$, and so $\bar{R}(\widetilde{\tau}) = \inf_{\widehat{\tau}} \sup_\tau R(\tau, \widehat{\tau})$.

9b)

Denote the observed values of the data $X_1 = x_1, \ldots, X_n = x_n$ as x^n. Using the property that the Bayes estimator $\hat{\tau}_B$ minimises the posterior risk $r(\hat{\tau}|x^n)$, and the law of the unconscious statistician, we have that

$$\hat{\tau}_B = \min_{\hat{\tau}} r(\hat{\tau}|x^n) = \min_{\hat{\tau}} \mathbb{E}_{p(\theta|x^n)}[L(\tau(\theta), \hat{\tau})|X^n = x^n] = \int L(\tau(\theta), \hat{\tau}) p(\theta|x^n) d\theta$$

With a view to minimisation, we compute the derivative with respect to the estimator $\hat{\tau}$:

$$\begin{aligned}
\frac{d}{d\hat{\tau}} \int \frac{(\hat{\tau} - t)^2}{\tau} p(\theta|x^n) d\theta &= \int \frac{\partial}{\partial \hat{\tau}} \frac{(\hat{\tau} - \tau)}{t} p(\theta|x^n) d\theta \\
&= 2 \int \left(\frac{\hat{\tau} - \tau}{\tau} p(\theta|x^n) \right) d\theta \\
&= 2 \left(\hat{\tau} \int \frac{1}{\tau} p(\theta|x^n) d\theta - \int p(\theta|x^n) d\theta \right)
\end{aligned}$$

Assuming that the posterior distribution $p(\theta|x^n)$ is appropriately normalised, setting the above to 0 and solving for $\hat{\tau}$ yields

$$\hat{\tau}_B = \frac{1}{\int [1/\tau(\theta)] \cdot p(\theta|x^n) d\theta}$$

An alternative solution.

The above solution might be viewed suspiciously. In particular, because it involves differentiating under the integral sign, and because when we compute the derivative with respect to an estimator $\hat{\tau}$, i.e. a *function* of data, we are really computing the derivative of the posterior risk *functional*.

To avoid these complications, we can go about this in a different way.

The loss function can be interpreted as a generalisation of squared error loss, known as *weighted squared error loss*. That is,

$$L(\tau, \hat{\tau}) = \frac{1}{\tau} \cdot (\hat{\tau} - \tau)^2,$$

where $w(\theta) = 1/\tau(\theta)$ is a weight, and $L'(\tau, \hat{\tau}) = (\hat{\tau} - \tau)^2$ is squared error loss.

In this case, the Bayes estimator $\hat{\tau}_B$ minimises posterior risk $r'(\hat{\tau}|x^n)$ under squared error loss $L'(\tau, \hat{\tau})$. Where the weight $1/\tau(\theta)$ has been absorbed into the original posterior $p(\theta|x^n)$ with renormalisation to form a new posterior $p'(\theta|x^n)$:

$$\hat{\tau}_B = \min_{\hat{\tau}} r'(\hat{\tau}|x^n) = \min_{\hat{\tau}} \int (\hat{\tau} - \tau)^2 \cdot p'(\theta|x^n) d\theta.$$

Using the result that the Bayes estimator under squared error loss is mean of the posterior r', and the lazy statistician rule, we have that

$$\hat{\tau}_B = \mathbb{E}_{p'(\theta|x^n)}[\tau(\theta)|X^n = x^n].$$

Rewriting the right hand side in terms of our original posterior p, the solution is:

$$\begin{aligned}
\hat{\tau}_B &= \tau(\theta) \left(\frac{[1/\tau(\theta)] \cdot p(\theta|x^n)}{\int [1/\tau(\theta')] \cdot p(\theta'|x^n) d\theta'} \right) d\theta \\
&= \frac{\int p(\theta|x^n) d\theta}{\int [(1/\tau(\theta')] \cdot p(\theta'|x^n) d\theta'} \\
&= \frac{1}{\int [(1/\tau(\theta')] \cdot p(\theta'|x^n) d\theta'}.
\end{aligned}$$

Where in the 2nd line the normalisation constant has been factored out, and in the 3rd line it has been assumed that $p(\theta|x^n)$ has been appropriately normalised.

The above result is an instance of the more general solution (in this context) that

$$\hat{\tau}_B = \frac{\mathbb{E}_{p(\theta|x^n)}[w(\theta)\tau(\theta)|X^n = x^n]}{\mathbb{E}_{p(\theta|x^n)}[w(\theta)|X^n = x^n]}.$$

A statement of this can be found in Corollary 2.5.2. of The Bayesian Choice by Christian Robert (2003).

10) Let $X_1, \ldots X_n \sim \text{Normal}(\theta, 1)$. Suppose that $\theta \in \{1, -1\}$. In other words, θ can only take two possible values.

a) Find a minimal sufficient statistic.

b) Find the maximum likelihood estimator. Is the maximum likelihood estimator a sufficient statistic?

c) Find the risk function using the loss function

$$L(\theta, \widehat{\theta}) = \begin{cases} 1 & \text{if} \theta \neq \widehat{\theta} \\ 0 & \text{if} \theta = \widehat{\theta} \end{cases}$$

10a)

A reasonable guess for a minimal sufficient statistic would be $T = \bar{X}_n$ - the sample mean.

Following the same arguments as 3a), we compute the ratio of Normal densities with mean θ and unit variance for two datasets $X^n = x^n$ and $Y^n = y^n$:

$$R(x^n, y^n; \theta) = \frac{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)\right)^2}{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \theta)^2\right)} = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\right) \exp\left(-\frac{n}{2}(\bar{x} - \theta)^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \bar{y})^2\right) \exp\left(-\frac{n}{2}(\bar{y} - \theta)^2\right)}$$

Now if we set $\bar{x} = \bar{y}$, we have that:

$$R(x^n, y^n; \theta) = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \bar{x})^2\right)} = c$$

i.e. R is constant, and does not depend on θ. To show the reverse implication holds true, i.e. if R does not depend on θ, then it must be the case that $\bar{x} = \bar{y}$, we compute derivatives of R with respect to θ:

$$\begin{aligned} \frac{\partial}{\partial \theta} R(x^n, y^n; \theta) =& \frac{\partial}{\partial \theta}\left[c \cdot \frac{\exp\left(-\frac{n}{2}(\bar{x} - \theta)^2\right)}{\exp\left(-\frac{n}{2}(\bar{y} - \theta)^2\right)}\right] \\ =& c \cdot \frac{\exp\left(-\frac{n}{2}(\bar{x} - \theta)^2\right) n(\bar{x} - \theta) \exp\left(-\frac{n}{2}(\bar{y} - \theta)^2\right)}{\exp\left(-\frac{n}{2}(\bar{y} - \theta)^2\right)^2} \\ & - \frac{\exp\left(-\frac{n}{2}(\bar{x} - \theta)^2\right) \exp\left(-\frac{n}{2}(\bar{y} - \theta)^2\right) n(\bar{y} - \theta)}{\exp\left(-\frac{n}{2}(\bar{y} - \theta)^2\right)^2} \end{aligned}$$

If R does not depend on θ, then this derivative must be 0 and rearranging we have that:

$$\frac{\partial}{\partial \theta} R(x^n, y^n; \theta) = 0 \implies n(\bar{x} - \theta) = n(\bar{y} - \theta) \implies \bar{x} = \bar{y}$$

Hence $T(X^n) = \bar{X}_n$ is a minimal sufficient statistic.

10b)

We show that the maximum likelihood estimator $\widehat{\theta}$ of the mean θ in this setting is the sample mean, $\widehat{\theta} = \bar{X}_n$.

Computing the log-likelihood by taking the product of Normal densities with mean θ and unit variance, and evaluating for a particular realisation of the data $X^n = x^n$, we have that $l(\theta)$:

$$l(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$$

Taking derivatives with respect to θ, setting to 0 and solving for θ, we have that:

$$l'(\theta) = 0 \implies \sum_{i=1}^{n}(x_i - \theta) = 0 \implies \widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

As we established in the previous section that the sample mean is minimal sufficient in this case, the maximum likelihood estimator $\widehat{\theta} = \bar{X}_n$ is also minimal sufficient.

If we did not have the previous result, and were interested in whether the the maximum likelihood estimator $\widehat{\theta} = \bar{X}_n$ was sufficient, observe the following factorisation of the joint PDF:

$$\begin{aligned} f_{X^n}(x^n;\theta) &= (2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \\ &= \underbrace{(2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)}_{h(x^n)}\underbrace{\exp\left(-\frac{n}{2}(\bar{x}-\theta)^2\right)}_{g(T(x^n),\theta)} \end{aligned}$$

We have factorised the joint PDF into a function $h(x^n)$ that depends only on the data, and a function $g(T(x^n), \theta)$ that depends on the parameter θ and the data x^n, but with the depenendence on the latter only through a statistic $T(x^n)$. So by the Fisher-Neyman factorisation theorem, the maximum likelihood estimator is sufficient.

11) Let $X_i \sim \text{Bernoulli}(p_i)$ for $i = 1, 2, ..., n$. The observations are independent but each observation has a different mean. The unknown parameter $p = (p_1, \ldots, p_n)$.

a) Let $\psi = \sum_{i=1}^{n} p_i$. Find the maximum likelihood estimator of ψ.

b) Find the mean squared error (MSE) of the maximum likelihood estimator of ψ.

c) Suppose we use the following prior distribution:

$$\pi(p_1, ...p_n) = 1$$

Find the Bayes estimator of ψ.

Hint: Recall that the Beta (α, β) density is:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

If $W \sim \text{Beta}(\alpha, \beta)$ then $\mathbb{E}[W] = \alpha/(\alpha+\beta)$ and $\text{Var}[W] = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$.

11a)

We first compute the maximum likelihood estimator $\widehat{p} = (\widehat{p_1}, ..., \widehat{p_n})$

This is a multiparameter problem. As the X_i are independent, the joint PMF can be written as a product of Bernoulli PMFs.

Evaluating the joint PMF for a particular data set $X^n = x^n$, the likelihood function is a function of the p_i:

$$L(p_1, ..., p_n) = \prod_{i=1}^{n} p_i^{x_i}(1 - p_i)^{1-x_i}$$

The log-likelihood is:

$$l(p_1, ..., p_n) = \sum_{i=1}^{n} x_i \log p_i + (1 - x_i) \log(1 - p_i)$$

Computing partial derivatives with respect to p_i, setting this equal to 0, and solving for p_i:

$$\frac{\partial}{\partial p_i} l(p_1, ..., p_n) = \frac{x_i}{p_i} - \frac{1 - x_i}{1 - p_i} = 0 \implies \widehat{p_i} = x_i$$

Hence we have that the maximum likelihood estimator is $\widehat{p} = (X_1, ..., X_n)$, i.e. the particular observation of the data set that is realised.

Now the $\psi = g(p_1, ..., p_n) = \sum_{i=1}^{n} p_i$ is a function of the n parameters $p_1, ..., p_n$. We can compute the maximum likelihood estimator $\widehat{\psi}$ using the equivariance property:

$$\widehat{\psi} = g(\widehat{p_1}, ..., \widehat{p_n}) = \sum_{i=1}^{n} \widehat{p_i} = \sum_{i=1}^{n} X_i$$

Hence the maximum likelihood estimator $\widehat{\psi} = n\bar{X}_n$

11b)

The mean squared error of the maximum likelihood estimator $\widehat{psi}$ is the sum of the variance and squared bias of the estimator:

$$\text{MSE} = B^2 + V$$

Computing the squared bias, we have***:

$$B^2 = (\mathbb{E}_\psi[\widehat{\psi}] - \psi)^2 = (\mathbb{E}_\psi\left[\sum_{i=1}^{n} X_i\right] - \psi)^2 = (\sum_{i=1}^{n} \mathbb{E}[X_i] - \psi)^2 = (\sum_{i=1}^{n} p_i - \psi)^2 = 0$$

Hence this maximum likelihood estimator $\widehat{\psi}$ is unbiased $B = 0$.

As the the estimator is unbiased, the mean squared error will be equal to the variance of the estimator, which is:

$$\text{MSE} = V = \text{Var}_\psi[\widehat{\psi}] = \text{Var}_\psi\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \text{Var}[X_i] + \sum_{i\neq j} \text{Cov}[X_i, X_j] = \sum_{i=1}^n \text{Var}[X_i] = \sum_{i=1}^n p_i(1-p_i)$$

11c)

To compute the Bayes estimator, we assume that the parameters p_i are random variables, and not fixed unknown parameters. In this context, we can now define a prior distribution on the value of the parameters p_i, and also a posterior distribution.

We have that the posterior is:

$$\begin{aligned} p(p_1, ..., p_n | X_1, ..., X_n) &\propto p(X_1, ..., X_n | p_1, ..., p_n)\pi(p_1, ..., p_n) \\ &\propto \prod_{i=1}^n p_i^{X_i}(1-p_i)^{1-X_i} \\ &= \frac{\prod_{i=1}^n p_i^{X_i}(1-p_i)^{1-X_i}}{\int ... \int \prod_{i=1}^n p_i^{X_i}(1-p_i)^{1-X_i} dp_1 ... dp_n} \end{aligned}$$

As the prior is uniform, and the likelihood factorises into n individual likelihoods on X_i, it is also the case that the posterior distribution of the parameters can be factorised into n individual posterior distributions on each of the p_i, which implies independence of the parameters p_i:

$$p(p_1, ..., p_n | X_1, ..., X_n) = p(p_1|X_1) \cdot p(p_2|X_2) \cdot ... \cdot p(p_n|X_n)$$

Where each individual posterior has the form:

$$p(p_i|X_i) = p_i^{X_i}(1-p_i)^{1-X_i}$$

This posterior can be written as a Beta distribution up to a constant of proportionality, that is $p_i \sim \text{Beta}(X_i + 1, -X_i + 2)$, that is:

$$p(p_i|X_i) \propto \frac{\Gamma(3)}{\Gamma(X_i+1)\Gamma(-X_i+2)} p_i^{X_i}(1-p_i)^{1-X_i}$$

And hence p_i has posterior mean:

$$\mathbb{E}[p_i|X_i] = \frac{X_i + 1}{(X_i+1) + (-X_i+2)} = \frac{X_i+1}{3}$$

Using the above results, and recalling that $\psi = g(p_1, ..., p_n)$, we can compute the Bayes estimator $\widehat{\psi}$ in the following manner:

$$
\begin{aligned}
\widehat{\psi} &= \mathbb{E}[\psi|X_1...X_n] \\
&= \mathbb{E}\left[\textstyle\sum_{i=1}^{n} p_i|X_1,...,X_n\right] \\
&= \sum_{i=1}^{n} \mathbb{E}[p_i|X_1,...,X_n] \\
&= \sum_{i=1}^{n} \mathbb{E}[p_i|X_i] \\
&= \sum_{i=1}^{n} \frac{X_i+1}{3} = \frac{n}{3}(\bar{X}_n+1)
\end{aligned}
$$

Where we have used the law of the unconscious statistician, linearity of expectation; and the fact that the p_i are conditionally independent of the remaining X_{-i} in going from the 3rd to the 4th equality. This needs further justification in terms of what distributions expectations are being taken with respect to, so we show the derivation using integrals:

$$
\begin{aligned}
\mathbb{E}[\psi|X_1...X_n] &= \int ... \int \left(\sum_{i=1}^{n} p_i\right) p(p_1,...,p_n|X_1,...,X_n) dp_1...dp_n \\
&= \int ... \int p_1 p(p_1,...,p_n|X_1,...,X_n) dp_1...dp_n + ... \\
&\quad + \int ... \int p_n p(p_1,...,p_n|X_1,...,X_n) dp_1...dp_n \\
&= \sum_{i=1}^{n} \mathbb{E}[p_i|X_1...X_n]
\end{aligned}
$$

Where we have the used the fact that definite integration is distributive. Denoting X_{-i} and p_{-i} to refer to all the X_k except for X_i, and all the p_k except p_i respectively, each summand can be simplified as follows:

$$
\begin{aligned}
\mathbb{E}[p_i|X_1...X_n] &= \int ... \int p_i p(p_1,...,p_N|X_1,...,X_n) dp_1...dp_n \\
&= \int ... \int p(p_{-i}|X_{-i}) \int p_i p(p_i|X_i) dp_i dp_{-i} \\
&= \int ... \int p(p_{-i}|X_{-i}) \int_0^1 p_i p(p_i|X_i) dp_i dp_{-i} \\
&= \int ... \int p(p_{-i}|X_{-i}) \mathbb{E}[p_i|X_i] dp_{-i} \\
&= \mathbb{E}[p_i|X_i] \underbrace{\int ... \int p(p_{-i}|X_{-i}) dp_{-i}}_{=1} \\
&= \mathbb{E}[p_i|X_i]
\end{aligned}
$$

Yielding the required result.

12) Let $X_1, ..., X_n \sim N(\mu, 1)$.

a) Let $T = \max\{X_1, ..., X_n\}$. Show that T is not sufficient.

b) Let use the improper prior $\pi(\mu) \propto 1$. Find the Bayes estimator of $\psi = \mu^2$.