Part II

Principles of Statistics

Year
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
We consider a statistical model \( \{ f(\cdot, \theta) : \theta \in \Theta \} \).

(a) Define the maximum likelihood estimator (MLE) and the Fisher information \( I(\theta) \).

(b) Let \( \Theta = \mathbb{R} \) and assume there exist a continuous one-to-one function \( \mu : \mathbb{R} \to \mathbb{R} \) and a real-valued function \( h \) such that
\[
E_\theta[h(X)] = \mu(\theta) \quad \forall \theta \in \mathbb{R}.
\]

(i) For \( X_1, \ldots, X_n \) i.i.d. from the model for some \( \theta_0 \in \mathbb{R} \), give the limit in almost sure sense of
\[
\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i).
\]
Give a consistent estimator \( \hat{\theta}_n \) of \( \theta_0 \) in terms of \( \hat{\mu}_n \).

(ii) Assume further that \( E_{\theta_0}[h(X)^2] < \infty \) and that \( \mu \) is continuously differentiable and strictly monotone. What is the limit in distribution of \( \sqrt{n}(\hat{\theta}_n - \theta_0) \)? Assume too that the statistical model satisfies the usual regularity assumptions. Do you necessarily expect \( \text{Var}(\hat{\theta}_n) \geq (nI(\theta_0))^{-1} \) for all \( n \)? Why?

(iii) Propose an alternative estimator for \( \theta_0 \) with smaller bias than \( \hat{\theta}_n \) if \( B_n(\theta_0) = E_{\theta_0}[\hat{\theta}_n] - \theta_0 = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) \) for some \( a, b \in \mathbb{R} \) with \( a \neq 0 \).

(iv) Further to all the assumptions in (ii), assume that the MLE for \( \theta_0 \) is of the form
\[
\hat{\theta}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} h(X_i).
\]
What is the link between the Fisher information at \( \theta_0 \) and the variance of \( h(X) \)? What does this mean in terms of the precision of the estimator and why?

[You may use results from the course, provided you state them clearly.]
We consider the exponential model \( \{f(\cdot, \theta) : \theta \in (0, \infty)\} \), where
\[
f(x, \theta) = \theta e^{-\theta x} \quad \text{for } x \geq 0.
\]

We observe an i.i.d. sample \( X_1, \ldots, X_n \) from the model.

(a) Compute the maximum likelihood estimator \( \hat{\theta}_{MLE} \) for \( \theta \). What is the limit in distribution of \( \sqrt{n}(\hat{\theta}_{MLE} - \theta) \)?

(b) Consider the Bayesian setting. Fix \( \alpha, \beta > 0 \) and place a Gamma(\( \alpha, \beta \)) prior for \( \theta \) with density
\[
\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta \theta) \quad \text{for } \theta > 0,
\]
where \( \Gamma \) is the Gamma function satisfying \( \Gamma(a + 1) = a\Gamma(a) \) for all \( a > 0 \). What is the posterior distribution for \( \theta \)? What is the Bayes estimator \( \hat{\theta}_\pi \) for the squared loss?

(c) Show that the Bayes estimator is consistent. What is the limiting distribution of \( \sqrt{n}(\hat{\theta}_\pi - \theta) \)?

[You may use results from the course, provided you state them clearly.]
(a) We consider the model \( \{ \text{Poisson}(\theta) : \theta \in (0, \infty) \} \) and an i.i.d. sample \( X_1, \ldots, X_n \) from it. Compute the expectation and variance of \( X_1 \) and check they are equal. Find the maximum likelihood estimator \( \hat{\theta}_{\text{MLE}} \) for \( \theta \) and, using its form, derive the limit in distribution of \( \sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \).

(b) In practice, Poisson-looking data show overdispersion, i.e., the sample variance is larger than the sample expectation. For \( \pi \in [0, 1] \) and \( \lambda \in (0, \infty) \), let \( p_{\pi, \lambda} : \mathbb{N}_0 \to [0, 1] \),
\[
k \mapsto p_{\pi, \lambda}(k) = \begin{cases} 
\pi e^{-\lambda} \frac{\lambda^k}{k!} & \text{for } k \geq 1 \\
(1 - \pi) + \pi e^{-\lambda} & \text{for } k = 0.
\end{cases}
\]
Show that this defines a distribution. Does it model overdispersion? Justify your answer.

(c) Let \( Y_1, \ldots, Y_n \) be an i.i.d. sample from \( p_{\pi, \lambda} \). Assume \( \lambda \) is known. Find the maximum likelihood estimator \( \hat{\pi}_{\text{MLE}} \) for \( \pi \).

(d) Furthermore, assume that, for any \( \pi \in [0, 1] \), \( \sqrt{n}(\hat{\pi}_{\text{MLE}} - \pi) \) converges in distribution to a random variable \( Z \) as \( n \to \infty \). Suppose we wanted to test the null hypothesis that our data arises from the model in part (a). Before making any further computations, can we necessarily expect \( Z \) to follow a normal distribution under the null hypothesis? Explain. Check your answer by computing the appropriate distribution.

[You may use results from the course, provided you state them clearly. In this question, \( \mathbb{N}_0 \) denotes the non-negative integers.]
In a regression problem, for a given $X \in \mathbb{R}^{n \times p}$ fixed, we observe $Y \in \mathbb{R}^n$ such that

$$Y = X\theta_0 + \varepsilon$$

for an unknown $\theta_0 \in \mathbb{R}^p$ and $\varepsilon$ random such that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ for some known $\sigma^2 > 0$.

(a) When $p \leq n$ and $X$ has rank $p$, compute the maximum likelihood estimator $\hat{\theta}_{MLE}$ for $\theta_0$. When $p > n$, what issue is there with the likelihood maximisation approach and how many maximisers of the likelihood are there (if any)?

(b) For any $\lambda > 0$ fixed, we consider $\hat{\theta}_\lambda$ minimising

$$\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$$

over $\mathbb{R}^p$. Derive an expression for $\hat{\theta}_\lambda$ and show it is well defined, i.e., there is a unique minimiser for every $X,Y$ and $\lambda$.

Assume $p \leq n$ and that $X$ has rank $p$. Let $\Sigma = X^\top X$ and note that $\Sigma = V\Lambda V^\top$ for some orthogonal matrix $V$ and some diagonal matrix $\Lambda$ whose diagonal entries satisfy $\Lambda_{1,1} \geq \Lambda_{2,2} \geq \ldots \geq \Lambda_{p,p}$. Assume that the columns of $X$ have mean zero.

(c) Denote the columns of $U = XV$ by $u_1, \ldots, u_p$. Show that they are sample principal components, i.e., that their pairwise sample correlations are zero and that they have sample variances $\Lambda_{1,1}, \ldots, \Lambda_{p,p}$, respectively. [Hint: the sample covariance between $u_i$ and $u_j$ is $n^{-1}u_i^\top u_j$.]

(d) Let $\hat{Y}_{MLE} = X\hat{\theta}_{MLE}$. Show that

$$\hat{Y}_{MLE} = U\Lambda^{-1}U^\top Y.$$  

Conclude that $\hat{Y}_{MLE}$ is the closest point to $Y$ within the subspace spanned by the normalised sample principal components of part (c).

(e) Let $\hat{Y}_\lambda = X\hat{\theta}_\lambda$. Show that

$$\hat{Y}_\lambda = U(\Lambda + \lambda I_p)^{-1}U^\top Y.$$  

Assume $\Lambda_{1,1}, \Lambda_{2,2}, \ldots, \Lambda_{q,q} \gg \lambda \gg \Lambda_{q+1,q+1}, \ldots, \Lambda_{p,p}$ for some $1 \leq q < p$. Conclude that $\hat{Y}_\lambda$ is approximately the closest point to $Y$ within the subspace spanned by the $q$ normalised sample principal components of part (c) with the greatest variance.
Let \( g : \mathbb{R} \to \mathbb{R} \) be an unknown function, twice continuously differentiable with \( |g''(x)| \leq M \) for all \( x \in \mathbb{R} \). For some \( x_0 \in \mathbb{R} \), we know the value \( g(x_0) \) and we wish to estimate its derivative \( g'(x_0) \). To do so, we have access to a pseudo-random number generator that gives \( U_1^*, \ldots, U_N^* \) i.i.d. uniform over \([0,1]\), and a machine that takes input \( x_1, \ldots, x_N \in \mathbb{R} \) and returns \( g(x_i) + \varepsilon_i \), where the \( \varepsilon_i \) are i.i.d. \( \mathcal{N}(0, \sigma^2) \).

(a) Explain how this setup allows us to generate \( N \) independent \( X_i = x_0 + hZ_i \), where the \( Z_i \) take value 1 or \(-1\) with probability \( 1/2 \), for any \( h > 0 \).

(b) We denote by \( Y_i \) the output \( g(X_i) + \varepsilon_i \). Show that for some independent \( \xi_i \in \mathbb{R} \)

\[
Y_i - g(x_0) = hZ_i g'(x_0) + \frac{h^2}{2} g''(\xi_i) + \varepsilon_i.
\]

(c) Using the intuition given by the least-squares estimator, justify the use of the estimator \( \hat{g}_N \) given by

\[
\hat{g}_N = \frac{1}{N} \sum_{i=1}^{N} Z_i (Y_i - g(x_0)) / h.
\]

(d) Show that

\[
\mathbb{E}[|\hat{g}_N - g'(x_0)|^2] \leq \frac{h^2 M^2}{4} + \frac{\sigma^2}{Nh^2}.
\]

Show that for some choice \( h_N \) of parameter \( h \), this implies

\[
\mathbb{E}[|\hat{g}_N - g'(x_0)|^2] \leq \frac{\sigma M}{\sqrt{N}}.
\]
In the model \( \{N(\theta, I_p), \theta \in \mathbb{R}^p \} \) of a Gaussian distribution in dimension \( p \), with unknown mean \( \theta \) and known identity covariance matrix \( I_p \), we estimate \( \theta \) based on a sample of i.i.d. observations \( X_1, \ldots, X_n \) drawn from \( N(\theta_0, I_p) \).

(a) Define the Fisher information \( I(\theta_0) \), and compute it in this model.

(b) We recall that the observed Fisher information \( i_n(\theta) \) is given by

\[
i_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta} \log f(X_i, \theta) \nabla_{\theta} \log f(X_i, \theta)^\top.
\]

Find the limit of \( \hat{i}_n = i_n(\hat{\theta}_{MLE}) \), where \( \hat{\theta}_{MLE} \) is the maximum likelihood estimator of \( \theta \) in this model.

(c) Define the Wald statistic \( W_n(\theta) \) and compute it. Give the limiting distribution of \( W_n(\theta_0) \) and explain how it can be used to design a confidence interval for \( \theta_0 \).

[You may use results from the course provided that you state them clearly.]

We consider the model \( \{N(\theta, I_p), \theta \in \mathbb{R}^p \} \) of a Gaussian distribution in dimension \( p \geq 3 \), with unknown mean \( \theta \) and known identity covariance matrix \( I_p \). We estimate \( \theta \) based on one observation \( X \sim N(\theta, I_p) \), under the loss function

\[
\ell(\theta, \delta) = \|\theta - \delta\|_2^2.
\]

(a) Define the risk of an estimator \( \hat{\theta} \). Compute the maximum likelihood estimator \( \hat{\theta}_{MLE} \) of \( \theta \) and its risk for any \( \theta \in \mathbb{R}^p \).

(b) Define what an admissible estimator is. Is \( \hat{\theta}_{MLE} \) admissible?

(c) For any \( c > 0 \), let \( \pi_c(\theta) \) be the prior \( N(0, c^2 I_p) \). Find a Bayes optimal estimator \( \hat{\theta}_c \) under this prior with the quadratic loss, and compute its Bayes risk.

(d) Show that \( \hat{\theta}_{MLE} \) is minimax.

[You may use results from the course provided that you state them clearly.]
A scientist wishes to estimate the proportion $\theta \in (0, 1)$ of presence of a gene in a population of flies of size $n$. Every fly receives a chromosome from each of its two parents, each carrying the gene A with probability $(1 - \theta)$ or the gene B with probability $\theta$, independently. The scientist can observe if each fly has two copies of the gene A (denoted by AA), two copies of the gene B (denoted by BB) or one of each (denoted by AB). We let $n_{AA}, n_{BB},$ and $n_{AB}$ denote the number of each observation among the $n$ flies.

(a) Give the probability of each observation as a function of $\theta$, denoted by $f(X, \theta)$, for all three values $X = AA, BB,$ or $AB$.

(b) For a vector $w = (w_{AA}, w_{BB}, w_{AB})$, we let $\hat{\theta}_w$ denote the estimator defined by

$$\hat{\theta}_w = \frac{w_{AA} n_{AA}}{n} + \frac{w_{BB} n_{BB}}{n} + \frac{w_{AB} n_{AB}}{n}.$$ 

Find the unique vector $w^*$ such that $\hat{\theta}_w^*$ is unbiased. Show that $\hat{\theta}_w^*$ is a consistent estimator of $\theta$.

(c) Compute the maximum likelihood estimator of $\theta$ in this model, denoted by $\hat{\theta}_{MLE}$. Find the limiting distribution of $\sqrt{n}(\hat{\theta}_{MLE} - \theta)$. [You may use results from the course, provided that you state them clearly.]
We consider the problem of estimating $\theta$ in the model \{\(f(x, \theta) : \theta \in (0, \infty)\)\}, where
\[
f(x, \theta) = (1 - \alpha)(x - \theta)^{-\alpha}1\{x \in [\theta, \theta + 1]\}.
\]
Here \(1\{A\}\) is the indicator of the set \(A\), and \(\alpha \in (0, 1)\) is known. This estimation is based on a sample of \(n\) i.i.d. \(X_1, \ldots, X_n\), and we denote by \(X_{(1)} < \ldots < X_{(n)}\) the ordered sample.

(a) Compute the mean and the variance of \(X_1\). Construct an unbiased estimator of \(\theta\) taking the form \(\hat{\theta}_n = \bar{X}_n + c(\alpha)\), where \(\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i\), specifying \(c(\alpha)\).

(b) Show that \(\hat{\theta}_n\) is consistent and find the limit in distribution of \(\sqrt{n}(\hat{\theta}_n - \theta)\). Justify your answer, citing theorems that you use.

(c) Find the maximum likelihood estimator \(\hat{\theta}_n\) of \(\theta\). Compute \(P(\hat{\theta}_n - \theta > t)\) for all real \(t\). Is \(\hat{\theta}_n\) unbiased?

(d) For \(t > 0\), show that \(P(n^\beta(\hat{\theta}_n - \theta) > t)\) has a limit in \((0, 1)\) for some \(\beta > 0\). Give explicitly the value of \(\beta\) and the limit. Why should one favour using \(\hat{\theta}_n\) over \(\bar{X}_n\)?
We consider the problem of estimating an unknown $\theta_0$ in a statistical model $\{f(x, \theta), \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}$, based on $n$ i.i.d. observations $X_1, \ldots, X_n$ whose distribution has p.d.f. $f(x, \theta_0)$.

In all the parts below you may assume that the model satisfies necessary regularity conditions.

(a) Define the score function $S_n$ of $\theta$. Prove that $S_n(\theta_0)$ has mean 0.

(b) Define the Fisher Information $I(\theta)$. Show that it can also be expressed as

$$I(\theta) = -\mathbb{E}_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log f(X_1, \theta) \right].$$

(c) Define the maximum likelihood estimator $\hat{\theta}_n$ of $\theta$. Give without proof the limits of $\hat{\theta}_n$ and of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ (in a manner which you should specify). [Be as precise as possible when describing a distribution.]

(d) Let $\psi : \Theta \to \mathbb{R}$ be a continuously differentiable function, and $\tilde{\theta}_n$ another estimator of $\theta_0$ such that $|\tilde{\theta}_n - \theta_0| \leq 1/n$ with probability 1. Give the limits of $\psi(\tilde{\theta}_n)$ and of $\sqrt{n}(\psi(\tilde{\theta}_n) - \psi(\theta_0))$ (in a manner which you should specify).
For the statistical model \( \mathcal{N}_d(\theta, \Sigma), \theta \in \mathbb{R}^d \), where \( \Sigma \) is a known, positive-definite \( d \times d \) matrix, we want to estimate \( \theta \) based on \( n \) i.i.d. observations \( X_1, \ldots, X_n \) with distribution \( \mathcal{N}_d(\theta, \Sigma) \).

(a) Derive the maximum likelihood estimator \( \hat{\theta}_n \) of \( \theta \). What is the distribution of \( \hat{\theta}_n \)?

(b) For \( \alpha \in (0, 1) \), construct a confidence region \( C^\alpha_n \) such that \( P_{\theta}(\theta \in C^\alpha_n) = 1 - \alpha \).

(c) For \( \Sigma = I_d \), compute the maximum likelihood estimator of \( \theta \) for the following parameter spaces:

   (i) \( \Theta = \{ \theta : \|\theta\|_2 = 1 \} \).

   (ii) \( \Theta = \{ \theta : v^\top \theta = 0 \} \) for some unit vector \( v \in \mathbb{R}^d \).

(d) For \( \Sigma = I_d \), we want to test the null hypothesis \( \Theta_0 = \{ 0 \} \) (i.e. \( \theta = 0 \)) against the composite alternative \( \Theta_1 = \mathbb{R}^d \setminus \{ 0 \} \). Compute the likelihood ratio statistic \( \Lambda(\Theta_1, \Theta_0) \) and give its distribution under the null hypothesis. Compare this result with the statement of Wilks’ theorem.
Paper 1, Section II

28K Principles of Statistics

For a positive integer \( n \), we want to estimate the parameter \( p \) in the binomial statistical model \( \{\text{Bin}(n, p), p \in [0, 1]\} \), based on an observation \( X \sim \text{Bin}(n, p) \).

(a) Compute the maximum likelihood estimator for \( p \). Show that the posterior distribution for \( p \) under a uniform prior on \([0, 1]\) is \( \text{Beta}(a, b) \), and specify \( a \) and \( b \).

[The p.d.f. of \( \text{Beta}(a, b) \) is given by

\[
f_{a,b}(p) = \frac{(a + b - 1)!}{(a - 1)!(b - 1)!} p^{a-1}(1 - p)^{b-1}.
\]

(b) (i) For a risk function \( L \), define the risk of an estimator \( \hat{p} \) of \( p \), and the Bayes risk under a prior \( \pi \) for \( p \).

(ii) Under the loss function

\[ L(\hat{p}, p) = \frac{(\hat{p} - p)^2}{p(1 - p)}, \]

find a Bayes optimal estimator for the uniform prior. Give its risk as a function of \( p \).

(iii) Give a minimax optimal estimator for the loss function \( L \) given above. Justify your answer.

Part II, 2017 List of Questions

[TURN OVER
Paper 3, Section II

25J Principles of Statistics

Let $X_1, \ldots, X_n$ be i.i.d. random variables from a $N(\theta, 1)$ distribution, $\theta \in \mathbb{R}$, and consider a Bayesian model $\theta \sim N(0, v^2)$ for the unknown parameter, where $v > 0$ is a fixed constant.

(a) Derive the posterior distribution $\Pi(\theta \mid X_1, \ldots, X_n)$ of $\theta \mid X_1, \ldots, X_n$.

(b) Construct a credible set $C_n \subset \mathbb{R}$ such that

(i) $\Pi(C_n \mid X_1, \ldots, X_n) = 0.95$ for every $n \in \mathbb{N}$, and

(ii) for any $\theta_0 \in \mathbb{R}$,

$$\Pr_{\theta_0}(\theta_0 \in C_n) \to 0.95 \quad \text{as } n \to \infty,$$

where $\Pr_{\theta_0}$ denotes the distribution of the infinite sequence $X_1, X_2, \ldots$ when drawn independently from a fixed $N(\theta, 1)$ distribution.

[You may use the central limit theorem.]

---

Paper 2, Section II

26J Principles of Statistics

(a) State and prove the Cramér–Rao inequality in a parametric model $\{f(\theta) : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}$. [Necessary regularity conditions on the model need not be specified.]

(b) Let $X_1, \ldots, X_n$ be i.i.d. Poisson random variables with unknown parameter $E X_1 = \theta > 0$. For $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ and $S^2 = (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ define

$$T_\alpha = \alpha \bar{X}_n + (1 - \alpha) S^2, \quad 0 \leq \alpha \leq 1.$$

Show that $\text{Var}_\theta(T_\alpha) \geq \text{Var}_\theta(\bar{X}_n)$ for all values of $\alpha, \theta$.

Now suppose $\tilde{\theta} = \tilde{\theta}(X_1, \ldots, X_n)$ is an estimator of $\theta$ with possibly nonzero bias $B(\theta) = E_\theta \tilde{\theta} - \theta$. Suppose the function $B$ is monotone increasing on $(0, \infty)$. Prove that the mean-squared errors satisfy

$$E_\theta((\tilde{\theta}_n - \theta)^2) \geq E_\theta(\bar{X}_n - \theta)^2 \quad \text{for all } \theta \in \Theta.$$
Consider a decision problem with parameter space $\Theta$. Define the concepts of a Bayes decision rule $\delta_\pi$ and of a least favourable prior.

Suppose $\pi$ is a prior distribution on $\Theta$ such that the Bayes risk of the Bayes rule equals $\sup_{\theta \in \Theta} R(\delta_\pi, \theta)$, where $R(\delta, \theta)$ is the risk function associated to the decision problem. Prove that $\delta_\pi$ is least favourable.

Now consider a random variable $X$ arising from the binomial distribution $Bin(n, \theta)$, where $\theta \in \Theta = [0, 1]$. Construct a least favourable prior for the squared risk $R(\delta, \theta) = E_\theta(\delta(X) - \theta)^2$. [You may use without proof the fact that the Bayes rule for quadratic risk is given by the posterior mean.]

Derive the maximum likelihood estimator $\hat{\theta}_n$ based on independent observations $X_1, \ldots, X_n$ that are identically distributed as $N(\theta, 1)$, where the unknown parameter $\theta$ lies in the parameter space $\Theta = \mathbb{R}$. Find the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ as $n \to \infty$.

Now define
\[ \tilde{\theta}_n = \begin{cases} \hat{\theta}_n & \text{whenever } |\hat{\theta}_n| > n^{-1/4}, \\ 0 & \text{otherwise}, \end{cases} \]
and find the limiting distribution of $\sqrt{n}(\tilde{\theta}_n - \theta)$ as $n \to \infty$.

Calculate
\[ \lim_{n \to \infty} \sup_{\theta \in \Theta} nE_\theta(T_n - \theta)^2 \]
for the choices $T_n = \hat{\theta}_n$ and $T_n = \tilde{\theta}_n$. Based on the above findings, which estimator $T_n$ of $\theta$ would you prefer? Explain your answer.

[Throughout, you may use standard facts of stochastic convergence, such as the central limit theorem, provided they are clearly stated.]
Given independent and identically distributed observations $X_1, \ldots, X_n$ with finite mean $E(X_1) = \mu$ and variance $\text{Var}(X_1) = \sigma^2$, explain the notion of a bootstrap sample $X^b_1, \ldots, X^b_n$, and discuss how you can use it to construct a confidence interval $C_n$ for $\mu$.

Suppose you can operate a random number generator that can simulate independent uniform random variables $U_1, \ldots, U_n$ on $[0, 1]$. How can you use such a random number generator to simulate a bootstrap sample?

Suppose that $(F_n : n \in \mathbb{N})$ and $F$ are cumulative probability distribution functions defined on the real line, that $F_n(t) \to F(t)$ as $n \to \infty$ for every $t \in \mathbb{R}$, and that $F$ is continuous on $\mathbb{R}$. Show that, as $n \to \infty$,

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \to 0.$$

State (without proof) the theorem about the consistency of the bootstrap of the mean, and use it to give an asymptotic justification of the confidence interval $C_n$. That is, prove that as $n \to \infty$, $P^N(\mu \in C_n) \to 1 - \alpha$ where $P^N$ is the joint distribution of $X_1, X_2, \ldots$.

[You may use standard facts of stochastic convergence and the Central Limit Theorem without proof.]
Define what it means for an estimator $\hat{\theta}$ of an unknown parameter $\theta$ to be consistent.

Let $S_n$ be a sequence of random real-valued continuous functions defined on $\mathbb{R}$ such that, as $n \to \infty$, $S_n(\theta)$ converges to $S(\theta)$ in probability for every $\theta \in \mathbb{R}$, where $S : \mathbb{R} \to \mathbb{R}$ is non-random. Suppose that for some $\theta_0 \in \mathbb{R}$ and every $\varepsilon > 0$ we have

$$S(\theta_0 - \varepsilon) < 0 < S(\theta_0 + \varepsilon),$$

and that $S_n$ has exactly one zero $\hat{\theta}_n$ for every $n \in \mathbb{N}$. Show that $\hat{\theta}_n \to P \theta_0$ as $n \to \infty$, and deduce from this that the maximum likelihood estimator (MLE) based on observations $X_1, \ldots, X_n$ from a $N(\theta, 1)$, $\theta \in \mathbb{R}$ model is consistent.

Now consider independent observations $X_1, \ldots, X_n$ of bivariate normal random vectors

$$X_i = (X_{1i}, X_{2i})^T \sim N_2 \left( (\mu_i, \mu_i)^T, \sigma^2 I_2 \right), \quad i = 1, \ldots, n,$$

where $\mu_i \in \mathbb{R}$, $\sigma > 0$ and $I_2$ is the $2 \times 2$ identity matrix. Find the MLE $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^T$ of $\mu = (\mu_1, \ldots, \mu_n)^T$ and show that the MLE of $\sigma^2$ equals

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n s_i^2, \quad s_i^2 = \frac{1}{2}[(X_{1i} - \hat{\mu}_i)^2 + (X_{2i} - \hat{\mu}_i)^2].$$

Show that $\hat{\sigma}^2$ is not consistent for estimating $\sigma^2$. Explain briefly why the MLE fails in this model.

[You may use the Law of Large Numbers without proof.]
Consider a random variable $X$ arising from the binomial distribution $\text{Bin}(n, \theta)$, $\theta \in \Theta = [0, 1]$. Find the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ and the Fisher information $I(\theta)$ for $\theta \in \Theta$.

Now consider the following priors on $\Theta$:

(i) a uniform $U([0, 1])$ prior on $[0, 1]$,

(ii) a prior with density $\pi(\theta)$ proportional to $\sqrt{I(\theta)}$,

(iii) a Beta($\sqrt{n}/2, \sqrt{n}/2$) prior.

Find the means $E[\theta|X]$ and modes $m_{\theta}|X$ of the posterior distributions corresponding to the prior distributions (i)–(iii). Which of these posterior decision rules coincide with $\hat{\theta}_{\text{MLE}}$? Which one is minimax for quadratic risk? Justify your answers.

[You may use the following properties of the Beta($a, b$) ($a > 0, b > 0$) distribution. Its density $f(x; a, b)$, $x \in [0, 1]$, is proportional to $x^{a-1}(1-x)^{b-1}$, its mean is equal to $a/(a+b)$, and its mode is equal to $\max(a - 1, 0) \max(a, 1) \max(b, 1) - 2$ provided either $a > 1$ or $b > 1$.

You may further use the fact that a unique Bayes rule of constant risk is a unique minimax rule for that risk.]
Consider a normally distributed random vector $X \in \mathbb{R}^p$ modelled as $X \sim N(\theta, I_p)$ where $\theta \in \mathbb{R}^p$, $I_p$ is the $p \times p$ identity matrix, and where $p \geq 3$. Define the Stein estimator $\hat{\theta}_{\text{STEIN}}$ of $\theta$.

Prove that $\hat{\theta}_{\text{STEIN}}$ dominates the estimator $\tilde{\theta} = X$ for the risk function induced by quadratic loss

$$\ell(a, \theta) = \sum_{i=1}^{p} (a_i - \theta_i)^2, \quad a \in \mathbb{R}^p.$$ 

Show however that the worst case risks coincide, that is, show that

$$\sup_{\theta \in \mathbb{R}^p} E_{\theta} \ell(X, \theta) = \sup_{\hat{\theta} \in \mathbb{R}^p} E_{\theta} \ell(\hat{\theta}_{\text{STEIN}}, \theta).$$

[You may use Stein’s lemma without proof, provided it is clearly stated.]
Paper 4, Section II

27J Principles of Statistics

Suppose you have at hand a pseudo-random number generator that can simulate an i.i.d. sequence of uniform \( U[0, 1] \) distributed random variables \( U_1^*, \ldots, U_N^* \) for any \( N \in \mathbb{N} \). Construct an algorithm to simulate an i.i.d. sequence \( X_1^*, \ldots, X_N^* \) of standard normal \( N(0, 1) \) random variables. [Should your algorithm depend on the inverse of any cumulative probability distribution function, you are required to provide an explicit expression for this inverse function.]

Suppose as a matter of urgency you need to approximately evaluate the integral

\[
I = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{1}{(\pi + |x|)^{1/4}} e^{-x^2/2} dx.
\]

Find an approximation \( I_N \) of this integral that requires \( N \) simulation steps from your pseudo-random number generator, and which has stochastic accuracy

\[
\Pr(|I_N - I| > N^{-1/4}) \leq N^{-1/2},
\]

where \( \Pr \) denotes the joint law of the simulated random variables. Justify your answer.

Paper 3, Section II

27J Principles of Statistics

State and prove Wilks’ theorem about testing the simple hypothesis \( H_0 : \theta = \theta_0 \), against the alternative \( H_1 : \theta \in \Theta \backslash \{\theta_0\} \), in a one-dimensional regular parametric model \( \{f(\cdot, \theta) : \theta \in \Theta\}, \Theta \subseteq \mathbb{R} \). [You may use without proof the results from lectures on the consistency and asymptotic distribution of maximum likelihood estimators, as well as on uniform laws of large numbers. Necessary regularity conditions can be assumed without statement.]

Find the maximum likelihood estimator \( \hat{\theta}_n \) based on i.i.d. observations \( X_1, \ldots, X_n \) in a \( N(0, \theta) \)-model, \( \theta \in \Theta = (0, \infty) \). Deduce the limit distribution as \( n \to \infty \) of the sequence of statistics

\[
-n \left( \log(\overline{X}^2) - (\overline{X}^2 - 1) \right),
\]

where \( \overline{X}^2 = (1/n) \sum_{i=1}^{n} X_i^2 \) and \( X_1, \ldots, X_n \) are i.i.d. \( N(0, 1) \).
Paper 2, Section II
28J Principles of Statistics

In a general decision problem, define the concepts of a Bayes rule and of admissibility. Show that a unique Bayes rule is admissible.

Consider i.i.d. observations $X_1, \ldots, X_n$ from a Poisson($\theta$), $\theta \in \Theta = (0, \infty)$, model. Can the maximum likelihood estimator $\hat{\theta}_{MLE}$ of $\theta$ be a Bayes rule for estimating $\theta$ in quadratic risk for any prior distribution on $\theta$ that has a continuous probability density on $(0, \infty)$? Justify your answer.

Now model the $X_i$ as i.i.d. copies of $X|\theta \sim \text{Poisson}(\theta)$, where $\theta$ is drawn from a prior that is a Gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$ (given below). Show that the posterior distribution of $\theta | X_1, \ldots, X_n$ is a Gamma distribution and find its parameters. Find the Bayes rule $\hat{\theta}_{BAYES}$ for estimating $\theta$ in quadratic risk for this prior. [The Gamma probability density function with parameters $\alpha > 0$, $\lambda > 0$ is given by

$$f(\theta) = \frac{\lambda^\alpha \theta^{\alpha-1} e^{-\lambda \theta}}{\Gamma(\alpha)}, \quad \theta > 0,$$

where $\Gamma(\alpha)$ is the usual Gamma function.]

Finally assume that the $X_i$ have actually been generated from a fixed Poisson($\theta_0$) distribution, where $\theta_0 > 0$. Show that $\sqrt{n}(\hat{\theta}_{BAYES} - \hat{\theta}_{MLE})$ converges to zero in probability and deduce the asymptotic distribution of $\sqrt{n}(\hat{\theta}_{BAYES} - \theta_0)$ under the joint law $P^{\mathbb{N}}_{\theta_0}$ of the random variables $X_1, X_2, \ldots$. [You may use standard results from lectures without proof provided they are clearly stated.]

---

Paper 1, Section II
28J Principles of Statistics

State without proof the inequality known as the Cramér–Rao lower bound in a parametric model $\{f(\cdot, \theta) : \theta \in \Theta \subseteq \mathbb{R}\}$. Give an example of a maximum likelihood estimator that attains this lower bound, and justify your answer.

Give an example of a parametric model where the maximum likelihood estimator based on observations $X_1, \ldots, X_n$ is biased. State without proof an analogue of the Cramér–Rao inequality for biased estimators.

Define the concept of a minimax decision rule, and show that the maximum likelihood estimator $\hat{\theta}_{MLE}$ based on $X_1, \ldots, X_n$ in a $N(\theta, 1)$ model is minimax for estimating $\theta \in \Theta = \mathbb{R}$ in quadratic risk.
Paper 4, Section II

27K Principles of Statistics

Assuming only the existence and properties of the univariate normal distribution, define \( N_p(\mu, \Sigma) \), the multivariate normal distribution with mean (row-)vector \( \mu \) and dispersion matrix \( \Sigma \); and \( W_p(\nu; \Sigma) \), the Wishart distribution on integer \( \nu > 1 \) degrees of freedom and with scale parameter \( \Sigma \). Show that, if \( X \sim N_p(\mu, \Sigma) \), \( S \sim W_p(\nu; \Sigma) \), and \( b (1 \times q) \), \( A (p \times q) \) are fixed, then \( b + XA \sim N_q(b + \mu A, \Phi) \), \( A^TSA \sim W_p(\nu; \Phi) \), where \( \Phi = A^T \Sigma A \).

The random \((n \times p)\) matrix \( X \) has rows that are independently distributed as \( N_p(\mathbb{M}, \Sigma) \), where both parameters \( \mathbb{M} \) and \( \Sigma \) are unknown. Let \( \bar{X} := n^{-1} 1^T X \), where \( 1 \) is the \((n \times 1)\) vector of 1s; and \( S^c := X^T \Pi X \), with \( \Pi := I_n - n^{-1} 1 1^T \). State the joint distribution of \( \bar{X} \) and \( S^c \) given the parameters.

Now suppose \( n > p \) and \( \Sigma \) is positive definite. Hotelling’s \( T^2 \) is defined as
\[
T^2 := n(\bar{X} - \bar{M})(S^c)^{-1}(\bar{X} - \bar{M})^T
\]
where \( S^c := S^c/\nu \) with \( \nu := (n - 1) \). Show that, for any values of \( \bar{M} \) and \( \Sigma \),
\[
\left( \frac{\nu - p + 1}{\nu p} \right) T^2 \sim F_p^{\nu - p + 1},
\]
the \( F \) distribution on \( p \) and \( \nu - p + 1 \) degrees of freedom.

[You may assume that:

1. If \( S \sim W_p(\nu; \Sigma) \) and \( a \) is a fixed \((p \times 1)\) vector, then
\[
\frac{a^T \Sigma^{-1} a}{a^T S^{-1} a} \sim \chi^2_{\nu - p + 1}.
\]

2. If \( V \sim \chi^2_p \), \( W \sim \chi^2_\lambda \) are independent, then
\[
\frac{V/p}{W/\lambda} \sim F_p^\lambda.
\]
Paper 3, Section II

27K Principles of Statistics

What is meant by a convex decision problem? State and prove a theorem to the effect that, in a convex decision problem, there is no point in randomising. [You may use standard terms without defining them.]

The sample space, parameter space and action space are each the two-point set \{1, 2\}. The observable \(X\) takes value 1 with probability \(2/3\) when the parameter \(\Theta = 1\), and with probability \(3/4\) when \(\Theta = 2\). The loss function \(L(\theta, a)\) is 0 if \(a = \theta\), otherwise 1.

Describe all the non-randomised decision rules, compute their risk functions, and plot these as points in the unit square. Identify an inadmissible non-randomised decision rule, and a decision rule that dominates it.

Show that the minimax rule has risk function \((8/17, 8/17)\), and is Bayes against a prior distribution that you should specify. What is its Bayes risk? Would a Bayesian with this prior distribution be bound to use the minimax rule?

---

Paper 1, Section II

28K Principles of Statistics

When the real parameter \(\Theta\) takes value \(\theta\), variables \(X_1, X_2, \ldots\) arise independently from a distribution \(P_\theta\) having density function \(p_\theta(x)\) with respect to an underlying measure \(\mu\). Define the score variable \(U_n(\theta)\) and the information function \(I_n(\theta)\) for estimation of \(\Theta\) based on \(X^n := (X_1, \ldots, X_n)\), and relate \(I_n(\theta)\) to \(i(\theta) := I_1(\theta)\).

State and prove the Cramér–Rao inequality for the variance of an unbiased estimator of \(\Theta\). Under what conditions does this inequality become an equality? What is the form of the estimator in this case? [You may assume \(E_{\theta}\{U_n(\theta)\} = 0\), \(\text{var}_{\theta}\{U_n(\theta)\} = I_n(\theta)\), and any further required regularity conditions, without comment.]

Let \(\hat{\Theta}_n\) be the maximum likelihood estimator of \(\Theta\) based on \(X^n\). What is the asymptotic distribution of \(n^{1/2}(\hat{\Theta}_n - \Theta)\) when \(\Theta = \theta\)?

Suppose that, for each \(n\), \(\hat{\Theta}_n\) is unbiased for \(\Theta\), and the variance of \(n^{1/2}(\hat{\Theta}_n - \Theta)\) is exactly equal to its asymptotic variance. By considering the estimator \(\alpha\hat{\Theta}_k + (1 - \alpha)\hat{\Theta}_n\), or otherwise, show that, for \(k < n\), \(\text{cov}_{\theta}(\hat{\Theta}_k, \hat{\Theta}_n) = \text{var}_{\theta}(\hat{\Theta}_n)\).
Describe the Weak Sufficiency Principle (WSP) and the Strong Sufficiency Principle (SSP). Show that Bayesian inference with a fixed prior distribution respects WSP.

A parameter $\Phi$ has a prior distribution which is normal with mean 0 and precision (inverse variance) $h_\Phi$. Given $\Phi = \phi$, further parameters $\Theta := (\Theta_i : i = 1, \ldots, I)$ have independent normal distributions with mean $\phi$ and precision $h_\Theta$. Finally, given both $\Phi = \phi$ and $\Theta = \theta := (\theta_1, \ldots, \theta_I)$, observables $X := (X_{ij} : i = 1, \ldots, I; j = 1, \ldots, J)$ are independent, $X_{ij}$ being normal with mean $\theta_i$, and precision $h_X$. The precision parameters $(h_\Phi, h_\Theta, h_X)$ are all fixed and known. Let $\overline{X} := (\overline{X}_1, \ldots, \overline{X}_I)$, where $\overline{X}_i := \sum_{j=1}^J X_{ij}/J$. Show, directly from the definition of sufficiency, that $\overline{X}$ is sufficient for $(\Phi, \Theta)$. [You may assume without proof that, if $Y_1, \ldots, Y_n$ have independent normal distributions with the same variance, and $\overline{Y} := n^{-1} \sum_{i=1}^n Y_i$, then the vector $(Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$ is independent of $\overline{Y}$.]

For data-values $x := (x_{ij} : i = 1, \ldots, I; j = 1, \ldots, J)$, determine the joint distribution, $\Pi_\phi$, say, of $\Theta$, given $X = x$ and $\Phi = \phi$. What is the distribution of $\Phi$, given $\Theta = \theta$ and $X = x$?

Using these results, describe clearly how Gibbs sampling combined with Rao–Blackwellisation could be applied to estimate the posterior joint distribution of $\Theta$, given $X = x$. 

Part II, 2013 List of Questions
For $i = 1, \ldots, n$, the pairs $(X_i, Y_i)$ have independent bivariate normal distributions, with $E(X_i) = \mu_X$, $E(Y_i) = \mu_Y$, $\text{var}(X_i) = \text{var}(Y_i) = \phi$, and $\text{corr}(X_i, Y_i) = \rho$. The means $\mu_X, \mu_Y$ are known; the parameters $\phi > 0$ and $\rho \in (-1, 1)$ are unknown.

Show that the joint distribution of all the variables belongs to an exponential family, and identify the natural sufficient statistic, natural parameter, and mean-value parameter. Hence or otherwise, find the maximum likelihood estimator $\hat{\rho}$ of $\rho$.

Let $U_i := X_i + Y_i$, $V_i := X_i - Y_i$. What is the joint distribution of $(U_i, V_i)$?

Show that the distribution of

$$\frac{(1 + \hat{\rho})/(1 - \hat{\rho})}{(1 + \rho)/(1 - \rho)}$$

is $F^\mu_n$. Hence describe a $(1 - \alpha)$-level confidence interval for $\rho$. Briefly explain what would change if $\mu_X$ and $\mu_Y$ were also unknown.

[Recall that the distribution $F^\nu_{\nu_1}$ is that of $(W_1/\nu_1)/(W_2/\nu_2)$, where, independently for $j = 1$ and $j = 2$, $W_j$ has the chi-squared distribution with $\nu_j$ degrees of freedom.]
The parameter vector is $\Theta \equiv (\Theta_1, \Theta_2, \Theta_3)$, with $\Theta_i > 0$, $\Theta_1 + \Theta_2 + \Theta_3 = 1$. Given $\Theta = \theta \equiv (\theta_1, \theta_2, \theta_3)$, the integer random vector $X = (X_1, X_2, X_3)$ has a trinomial distribution, with probability mass function
\[
p(x \mid \theta) = \frac{n!}{x_1! x_2! x_3!} \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3}, \quad \left( x_i \geq 0, \sum_{i=1}^{3} x_i = n \right). \tag{1}
\]
Compute the score vector for the parameter $\Theta^* := (\Theta_1, \Theta_2)$, and, quoting any relevant general result, use this to determine $E(X_i)$ ($i = 1, 2, 3$).

Considering (1) as an exponential family with mean-value parameter $\Theta^*$, what is the corresponding natural parameter $\Phi \equiv (\Phi_1, \Phi_2)$?

Compute the information matrix $I$ for $\Theta^*$, which has $(i, j)$-entry
\[
I_{ij} = -E \left( \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right), \quad (i, j = 1, 2),
\]
where $l$ denotes the log-likelihood function, based on $X$, expressed in terms of $(\theta_1, \theta_2)$.

Show that the variance of $\log(X_1/X_3)$ is asymptotic to $n^{-1}(\theta_1^{-1} + \theta_3^{-1})$ as $n \to \infty$. [Hint. The information matrix $I_\Phi$ for $\Phi$ is $I^{-1}$ and the dispersion matrix of the maximum likelihood estimator $\hat{\Phi}$ behaves, asymptotically (for $n \to \infty$) as $I_\Phi^{-1}$.]
Carefully defining all italicised terms, show that, if a sufficiently general method of inference respects both the **Weak Sufficiency Principle** and the **Conditionality Principle**, then it respects the **Likelihood Principle**.

The position $X_t$ of a particle at time $t > 0$ has the Normal distribution $N(0, \phi t)$, where $\phi$ is the value of an unknown parameter $\Phi$; and the time, $T_x$, at which the particle first reaches position $x \neq 0$ has probability density function

$$p_x(t) = \frac{|x|}{\sqrt{2\pi\phi t^3}} \exp\left(-\frac{x^2}{2\phi t}\right) \quad (t > 0).$$

Experimenter $E_1$ observes $X_\tau$, and experimenter $E_2$ observes $T_\xi$, where $\tau > 0$, $\xi \neq 0$ are fixed in advance. It turns out that $T_\xi = \tau$. What does the Likelihood Principle say about the inferences about $\Phi$ to be made by the two experimenters?

$E_1$ bases his inference about $\Phi$ on the distribution and observed value of $X_\tau^2/\tau$, while $E_2$ bases her inference on the distribution and observed value of $\xi^2/T_\xi$. Show that these choices respect the Likelihood Principle.
Prove that, if $T$ is complete sufficient for $\Theta$, and $S$ is a function of $T$, then $S$ is the minimum variance unbiased estimator of $E(S \mid \Theta)$.

When the parameter $\Theta$ takes a value $\theta > 0$, observables $(X_1, \ldots, X_n)$ arise independently from the exponential distribution $E(\theta)$, having probability density function

$$p(x \mid \theta) = \theta e^{-\theta x} \quad (x > 0).$$

Show that the family of distributions

$$\Theta \sim \text{Gamma} (\alpha, \beta) \quad (\alpha > 0, \beta > 0),$$

with probability density function

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta \theta} \quad (\theta > 0),$$

is a conjugate family for Bayesian inference about $\Theta$ (where $\Gamma(\alpha)$ is the Gamma function).

Show that the expectation of $\Lambda := \log \Theta$, under prior distribution (1), is $\psi(\alpha) - \log \beta$, where $\psi(\alpha) := (d/d\alpha) \log \Gamma(\alpha)$. What is the prior variance of $\Lambda$? Deduce the posterior expectation and variance of $\Lambda$, given $(X_1, \ldots, X_n)$.

Let $\tilde{\Lambda}$ denote the limiting form of the posterior expectation of $\Lambda$ as $\alpha, \beta \downarrow 0$. Show that $\tilde{\Lambda}$ is the minimum variance unbiased estimator of $\Lambda$. What is its variance?
**Paper 1, Section II**

28K Principles of Statistics

Define admissible, Bayes, minimax decision rules.

A random vector \( X = (X_1, X_2, X_3)^T \) has independent components, where \( X_i \) has the normal distribution \( N(\theta_i, 1) \) when the parameter vector \( \Theta \) takes the value \( \theta = (\theta_1, \theta_2, \theta_3)^T \). It is required to estimate \( \Theta \) by a point \( a \in \mathbb{R}^3 \), with loss function \( L(\theta, a) = \|a - \theta\|^2 \).

What is the risk function of the maximum-likelihood estimator \( \hat{\Theta} := X \)? Show that \( \hat{\Theta} \) is dominated by the estimator \( \tilde{\Theta} := (1 - \|X\|^2) X \).

**Paper 2, Section II**

28K Principles of Statistics

Random variables \( X_1, \ldots, X_n \) are independent and identically distributed from the normal distribution with unknown mean \( M \) and unknown precision (inverse variance) \( H \).

Show that the likelihood function, for data \( X_1 = x_1, \ldots, X_n = x_n \), is

\[
L_n(\mu, h) \propto h^{n/2} \exp \left( -\frac{1}{2} h \left( n \overline{x} - \mu \right)^2 + S \right),
\]

where \( \overline{x} := n^{-1} \sum_i x_i \) and \( S := \sum_i (x_i - \overline{x})^2 \).

A bivariate prior distribution for \( (M, H) \) is specified, in terms of hyperparameters \( (\alpha_0, \beta_0, m_0, \lambda_0) \), as follows. The marginal distribution of \( H \) is \( \Gamma(\alpha_0, \beta_0) \), with density

\[
\pi(h) \propto h^{\alpha_0 - 1} e^{-\beta_0 h} \quad (h > 0),
\]

and the conditional distribution of \( M \), given \( H = h \), is normal with mean \( m_0 \) and precision \( \lambda_0 h \).

Show that the conditional prior distribution of \( H \), given \( M = \mu \), is

\[
H \mid M = \mu \sim \Gamma \left( \alpha_0 + \frac{1}{2}, \beta_0 + \frac{1}{2} \lambda_0 (\mu - m_0)^2 \right).
\]

Show that the posterior joint distribution of \( (M, H) \), given \( X_1 = x_1, \ldots, X_n = x_n \), has the same form as the prior, with updated hyperparameters \( (\alpha_n, \beta_n, m_n, \lambda_n) \) which you should express in terms of the prior hyperparameters and the data.

[You may use the identity

\[
p(t - a)^2 + q(t - b)^2 = (t - \delta)^2 + pq(a - b)^2,
\]

where \( p + q = 1 \) and \( \delta = pa + qb \).]

Explain how you could implement Gibbs sampling to generate a random sample from the posterior joint distribution.
Random variables $X_1, X_2, \ldots$ are independent and identically distributed from the exponential distribution $E(\theta)$, with density function
\[ p_X(x \mid \theta) = \theta e^{-\theta x} \quad (x > 0), \]
when the parameter $\Theta$ takes value $\theta > 0$. The following experiment is performed. First $X_1$ is observed. Thereafter, if $X_1 = x_1, \ldots, X_i = x_i$ have been observed ($i \geq 1$), a coin having probability $\alpha(x_i)$ of landing heads is tossed, where $\alpha : \mathbb{R} \to (0, 1)$ is a known function and the coin toss is independent of the $X$'s and previous tosses. If it lands heads, no further observations are made; if tails, $X_{i+1}$ is observed.

Let $N$ be the total number of $X$'s observed, and $\mathbf{X} := (X_1, \ldots, X_N)$. Write down the likelihood function for $\Theta$ based on data $\mathbf{X} = (x_1, \ldots, x_n)$, and identify a minimal sufficient statistic. What does the likelihood principle have to say about inference from this experiment?

Now consider the experiment that only records $Y := X_N$. Show that the density function of $Y$ has the form
\[ p_Y(y \mid \theta) = \exp\{a(y) - k(\theta) - \theta y\}. \]
Assuming the function $a(\cdot)$ is twice differentiable and that both $p_Y(y \mid \theta)$ and $\partial p_Y(y \mid \theta)/\partial y$ vanish at 0 and $\infty$, show that $a'(Y)$ is an unbiased estimator of $\Theta$, and find its variance.

Stating clearly any general results you use, deduce that
\[ -k''(\theta) E_{\theta}\{a''(Y)\} \geq 1. \]
What does it mean to say that a \((1 \times p)\) random vector \(\xi\) has a multivariate normal distribution?

Suppose \(\xi = (X, Y)\) has the bivariate normal distribution with mean vector \(\mu = (\mu_X, \mu_Y)\), and dispersion matrix

\[
\Sigma = \begin{pmatrix}
\sigma_{XX} & \sigma_{XY} \\
\sigma_{XY} & \sigma_{YY}
\end{pmatrix}.
\]

Show that, with \(\beta := \sigma_{XY}/\sigma_{XX},\ Y - \beta X\) is independent of \(X\), and thus that the conditional distribution of \(Y\) given \(X\) is normal with mean \(\mu_Y + \beta(X - \mu_X)\) and variance \(\sigma_{YY;X} := \sigma_{YY} - \sigma_{XY}^2/\sigma_{XX}\).

For \(i = 1, \ldots, n\), \(\xi_i = (X_i, Y_i)\) are independent and identically distributed with the above distribution, where all elements of \(\mu\) and \(\Sigma\) are unknown. Let

\[
S = \begin{pmatrix}
S_{XX} & S_{XY} \\
S_{XY} & S_{YY}
\end{pmatrix} := \sum_{i=1}^n (\xi_i - \bar{\xi})^T (\xi_i - \bar{\xi}),
\]

where \(\bar{\xi} := n^{-1} \sum_{i=1}^n \xi_i\).

The sample correlation coefficient is \(r := S_{XY}/\sqrt{S_{XX}S_{YY}}\). Show that the distribution of \(r\) depends only on the population correlation coefficient \(\rho := \sigma_{XY}/\sqrt{\sigma_{XX}\sigma_{YY}}\).

Student’s t-statistic (on \(n - 2\) degrees of freedom) for testing the null hypothesis \(H_0 : \beta = 0\) is

\[
t := \frac{\hat{\beta}}{\sqrt{S_{YY;X}/(n - 2)S_{XX}}},
\]

where \(\hat{\beta} := S_{XY}/S_{XX}\) and \(S_{YY;X} := S_{YY} - S_{XY}^2/S_{XX}\). Its density when \(H_0\) is true is

\[
p(t) = C \left(1 + \frac{t^2}{n - 2}\right)^{-\frac{1}{2}(n-1)},
\]

where \(C\) is a constant that need not be specified.

Express \(t\) in terms of \(r\), and hence derive the density of \(r\) when \(\rho = 0\).

How could you use the sample correlation \(r\) to test the hypothesis \(\rho = 0\)?
Paper 1, Section II

28J Principles of Statistics

The distribution of a random variable $X$ is obtained from the binomial distribution $B(n; \Pi)$ by conditioning on $X > 0$; here $\Pi \in (0, 1)$ is an unknown probability parameter and $n$ is known. Show that the distributions of $X$ form an exponential family and identify the natural sufficient statistic $T$, natural parameter $\Phi$, and cumulant function $k(\phi)$. Using general properties of the cumulant function, compute the mean and variance of $X$ when $\Pi = \pi$. Write down an equation for the maximum likelihood estimate $\hat{\Pi}$ of $\Pi$ and explain why, when $\Pi = \pi$, the distribution of $\hat{\Pi}$ is approximately normal $N(\pi, \pi(1 - \pi)/n)$ for large $n$.

Suppose we observe $X = 1$. It is suggested that, since the condition $X > 0$ is then automatically satisfied, general principles of inference require that the inference to be drawn should be the same as if the distribution of $X$ had been $B(n; \Pi)$ and we had observed $X = 1$. Comment briefly on this suggestion.

Paper 2, Section II

28J Principles of Statistics

Define the Kolmogorov–Smirnov statistic for testing the null hypothesis that real random variables $X_1, \ldots, X_n$ are independently and identically distributed with specified continuous, strictly increasing distribution function $F$, and show that its null distribution does not depend on $F$.

A composite hypothesis $H_0$ specifies that, when the unknown positive parameter $\Theta$ takes value $\theta$, the random variables $X_1, \ldots, X_n$ arise independently from the uniform distribution $U[0, \theta]$. Letting $J := \arg \max_{1 \leq i \leq n} X_i$, show that, under $H_0$, the statistic $(J, X_J)$ is sufficient for $\Theta$. Show further that, given $\{J = j, X_j = \xi\}$, the random variables $(X_i : i \neq j)$ are independent and have the $U[0, \xi]$ distribution. How might you apply the Kolmogorov–Smirnov test to test the hypothesis $H_0$?
Define the normal and extensive form solutions of a Bayesian statistical decision problem involving parameter $\Theta$, random variable $X$, and loss function $L(\theta, a)$. How are they related? Let $R_0 = R_0(\Pi)$ be the Bayes loss of the optimal act when $\Theta \sim \Pi$ and no data can be observed. Express the Bayes risk $R_1$ of the optimal statistical decision rule in terms of $R_0$ and the joint distribution of $(\Theta, X)$.

The real parameter $\Theta$ has distribution $\Pi$, having probability density function $\pi(\cdot)$. Consider the problem of specifying a set $S \subseteq \mathbb{R}$ such that the loss when $\Theta = \theta$ is $L(\theta, S) = c|S| - 1_S(\theta)$, where $1_S$ is the indicator function of $S$, where $c > 0$, and where $|S| = \int_S dx$. Show that the “highest density” region $S^* := \{\theta : \pi(\theta) \geq c\}$ supplies a Bayes act for this decision problem, and explain why $R_0(\Pi) \leq 0$.

For the case $\Theta \sim \mathcal{N}(\mu, \sigma^2)$, find an expression for $R_0$ in terms of the standard normal distribution function $\Phi$.

Suppose now that $c = 0.5$, that $\Theta \sim \mathcal{N}(0, 1)$ and that $X|\Theta \sim \mathcal{N}(\Theta, 1/9)$. Show that $R_1 < R_0$.

Paper 4, Section II

**27J Principles of Statistics**

Define completeness and bounded completeness of a statistic $T$ in a statistical experiment.

Random variables $X_1$, $X_2$, $X_3$ are generated as $X_i = \Theta^{1/2} Z + (1 - \Theta)^{1/2} Y_i$, where $Z, Y_1, Y_2, Y_3$ are independently standard normal $\mathcal{N}(0, 1)$, and the parameter $\Theta$ takes values in $(0, 1)$. What is the joint distribution of $(X_1, X_2, X_3)$ when $\Theta = \theta$? Write down its density function, and show that a minimal sufficient statistic for $\Theta$ based on $(X_1, X_2, X_3)$ is $T = (T_1, T_2) := (\sum_{i=1}^3 X_i^2, (\sum_{i=1}^3 X_i)^2)$.

[Hint: You may use that if $I$ is the $n \times n$ identity matrix and $J$ is the $n \times n$ matrix all of whose entries are 1, then $aI + bJ$ has determinant $a^{n-1}(a + nb)$, and inverse $cI + dJ$ with $c = 1/a, d = -b/(a(a + nb))$.]

What is $\mathbb{E}_\theta(T_1)$? Is $T$ complete for $\Theta$?

Let $S := \text{Prob}(X_1^2 \leq 1 \mid T)$. Show that $\mathbb{E}_\theta(S)$ is a positive constant $c$ which does not depend on $\theta$, but that $S$ is not identically equal to $c$. Is $T$ boundedly complete for $\Theta$?
Paper 1, Section II
28I Principles of Statistics

(i) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, having the exponential distribution $E(\lambda)$ with density $p(x|\lambda) = \lambda \exp(-\lambda x)$ for $x, \lambda > 0$. Show that $T_n = \sum_{i=1}^{n} X_i$ is minimal sufficient and complete for $\lambda$.

[You may assume uniqueness of Laplace transforms.]

(ii) For given $x > 0$, it is desired to estimate the quantity $\phi = \text{Prob}(X_1 > x|\lambda)$. Compute the Fisher information for $\phi$.

(iii) State the Lehmann–Scheffé theorem. Show that the estimator $\tilde{\phi}_n$ of $\phi$ defined by

$$\tilde{\phi}_n = \begin{cases} 0, & \text{if } T_n < x, \\ (1 - \frac{x}{T_n})^{n-1}, & \text{if } T_n \geq x \end{cases}$$

is the minimum variance unbiased estimator of $\phi$ based on $(X_1, \ldots, X_n)$. Without doing any computations, state whether or not the variance of $\tilde{\phi}_n$ achieves the Cramér–Rao lower bound, justifying your answer briefly.

Let $k \leq n$. Show that $E(\tilde{\phi}_k | T_n, \lambda) = \tilde{\phi}_n$.

Paper 2, Section II
28I Principles of Statistics

Suppose that the random vector $X = (X_1, \ldots, X_n)$ has a distribution over $\mathbb{R}^n$ depending on a real parameter $\theta$, with everywhere positive density function $p(x | \theta)$. Define the maximum likelihood estimator $\hat{\theta}$, the score variable $U$, the observed information $\hat{j}$ and the expected (Fisher) information $I$ for the problem of estimating $\theta$ from $X$.

For the case where the $(X_i)$ are independent and identically distributed, show that, as $n \to \infty$, $I^{-1/2} U \overset{d}{\to} N(0, 1)$. [You may assume sufficient conditions to allow interchange of integration over the sample space and differentiation with respect to the parameter.] State the asymptotic distribution of $\hat{\theta}$.

The random vector $X = (X_1, \ldots, X_n)$ is generated according to the rule

$$X_{i+1} = \theta X_i + E_i,$$

where $X_0 = 1$ and the $(E_i)$ are independent and identically distributed from the standard normal distribution $N(0, 1)$. Write down the likelihood function for $\theta$ based on data $x = (x_1, \ldots, x_n)$, find $\hat{\theta}$ and $\hat{j}$ and show that the pair $(\hat{\theta}, \hat{j})$ forms a minimal sufficient statistic.

A Bayesian uses the improper prior density $\pi(\theta) \propto 1$. Show that, in the posterior, $S(\theta - \hat{\theta})$ (where $S$ is a statistic that you should identify) has the same distribution as $E_1$. 

Part II, 2009 List of Questions
Paper 3, Section II

27I Principles of Statistics

What is meant by an equaliser decision rule? What is meant by an extended Bayes rule? Show that a decision rule that is both an equaliser rule and extended Bayes is minimax.

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with the normal distribution $N(\theta, h^{-1})$, and let $k > 0$. It is desired to estimate $\theta$ with loss function $L(\theta, a) = 1 - \exp\left\{-\frac{1}{2}k(a - \theta)^2\right\}$.

Suppose the prior distribution is $\theta \sim N(m_0, h_0^{-1})$. Find the Bayes act and the Bayes loss posterior to observing $X_1 = x_1, \ldots, X_n = x_n$. What is the Bayes risk of the Bayes rule with respect to this prior distribution?

Show that the rule that estimates $\theta$ by $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ is minimax.

---

Paper 4, Section II

27I Principles of Statistics

Consider the double dichotomy, where the loss is 0 for a correct decision and 1 for an incorrect decision. Describe the form of a Bayes decision rule. Assuming the equivalence of normal and extensive form analyses, deduce the Neyman-Pearson lemma.

For a problem with random variable $X$ and real parameter $\theta$, define monotone likelihood ratio (MLR) and monotone test.

Suppose the problem has MLR in a real statistic $T = t(X)$. Let $\phi$ be a monotone test, with power function $\gamma(\cdot)$, and let $\phi'$ be any other test, with power function $\gamma'(\cdot)$. Show that if $\theta_1 > \theta_0$ and $\gamma(\theta_0) > \gamma'(\theta_0)$, then $\gamma(\theta_1) > \gamma'(\theta_1)$. Deduce that there exists $\theta^* \in [-\infty, \infty]$ such that $\gamma(\theta) \leq \gamma'(\theta)$ for $\theta < \theta^*$, and $\gamma(\theta) \geq \gamma'(\theta)$ for $\theta > \theta^*$.

For an arbitrary prior distribution $\Pi$ with density $\pi(\cdot)$, and an arbitrary value $\theta^*$, show that the posterior odds

$$\frac{\Pi(\theta > \theta^* | X = x)}{\Pi(\theta \leq \theta^* | X = x)}$$

is a non-decreasing function of $t(x)$. 
An angler starts fishing at time 0. Fish bite in a Poisson Process of rate $\Lambda$ per hour, so that, if $\Lambda = \lambda$, the number $N_t$ of fish he catches in the first $t$ hours has the Poisson distribution $\mathcal{P}(\lambda t)$, while $T_n$, the time in hours until his $n$th bite, has the Gamma distribution $\Gamma(n, \lambda)$, with density function

$$p(t \mid \lambda) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} \quad (t > 0).$$

Bystander $B_1$ plans to watch for 3 hours, and to record the number $N_3$ of fish caught. Bystander $B_2$ plans to observe until the 10th bite, and to record $T_{10}$, the number of hours until this occurs.

For $B_1$, show that $\tilde{\Lambda}_1 := N_3/3$ is an unbiased estimator of $\Lambda$ whose variance function achieves the Cramér–Rao lower bound.

Find an unbiased estimator of $\Lambda$ for $B_2$, of the form $\tilde{\Lambda}_2 = k/T_{10}$. Does it achieve the Cramér–Rao lower bound? Is it minimum-variance-unbiased? Justify your answers.

In fact, the 10th fish bites after exactly 3 hours. For each of $B_1$ and $B_2$, write down the likelihood function for $\Lambda$ based their observations. What does the Likelihood Principle have to say about the inferences to be drawn by $B_1$ and $B_2$, and why? Compute the estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ produced by applying $\tilde{\Lambda}_1$ and $\tilde{\Lambda}_2$ to the observed data. Does the method of minimum-variance-unbiased estimation respect the Likelihood Principle?
**Principles of Statistics**

Under hypothesis $H_i (i = 0, 1)$, a real-valued observable $X$, taking values in $\mathcal{X}$, has density function $p_i(\cdot)$. Define the *Type I error* $\alpha$ and the *Type II error* $\beta$ of a test $\phi : \mathcal{X} \to [0, 1]$ of the null hypothesis $H_0$ against the alternative hypothesis $H_1$. What are the size and power of the test in terms of $\alpha$ and $\beta$?

Show that, for $0 < c < \infty$, $\phi$ minimises $c\alpha + \beta$ among all possible tests if and only if it satisfies

$$
p_1(x) > c p_0(x) \Rightarrow \phi(x) = 1,
$$
$$
p_1(x) < c p_0(x) \Rightarrow \phi(x) = 0.
$$

What does this imply about the admissibility of such a test?

Given the value $\theta$ of a parameter variable $\Theta \in [0, 1)$, the observable $X$ has density function

$$
p(x | \theta) = \frac{2(x - \theta)}{(1 - \theta)^2} \quad (\theta \leq x \leq 1).
$$

For fixed $\theta \in (0, 1)$, describe all the likelihood ratio tests of $H_0 : \Theta = 0$ against $H_\theta : \Theta = \theta$.

For fixed $k \in (0, 1)$, let $\phi_k$ be the test that rejects $H_0$ if and only if $X \geq k$. Is $\phi_k$ admissible as a test of $H_0$ against $H_\theta$ for every $\theta \in (0, 1)$? Is it uniformly most powerful for its size for testing $H_0$ against the composite hypothesis $H_1 : \Theta \in (0, 1)$? Is it admissible as a test of $H_0$ against $H_1$?

**Principles of Statistics**

Define the notion of exponential family (EF), and show that, for data arising as a random sample of size $n$ from an exponential family, there exists a sufficient statistic whose dimension stays bounded as $n \to \infty$.

The log-density of a normal distribution $\mathcal{N}(\mu, \sigma)$ can be expressed in the form

$$
\log p(x | \phi) = \phi_1 x + \phi_2 x^2 - k(\phi)
$$

where $\phi = (\phi_1, \phi_2)$ is the value of an unknown parameter $\Phi = (\Phi_1, \Phi_2)$. Determine the function $k$, and the natural parameter-space $\Phi$. What is the mean-value parameter $H = (H_1, H_2)$ in terms of $\Phi$?

Determine the maximum likelihood estimator $\hat{\Phi}_1$ of $\Phi_1$ based on a random sample $(X_1, \ldots, X_n)$, and give its asymptotic distribution for $n \to \infty$.

How would these answers be affected if the variance of $X$ were known to have value $v_0$?
Principles of Statistics

Define *sufficient statistic*, and state the factorisation criterion for determining whether a statistic is sufficient. Show that a Bayesian posterior distribution depends on the data only through the value of a sufficient statistic.

Given the value $\mu$ of an unknown parameter $\mu$, observables $X_1, \ldots, X_n$ are independent and identically distributed with distribution $\mathcal{N}(\mu, 1)$. Show that the statistic $\overline{X} := n^{-1}\sum_{i=1}^{n} X_i$ is sufficient for $\mu$.

If the prior distribution is $\mu \sim \mathcal{N}(0, \tau^2)$, determine the posterior distribution of $\mu$ and the predictive distribution of $\overline{X}$.

In fact, there are two hypotheses as to the value of $\mu$. Under hypothesis $H_0$, $\mu$ takes the known value 0; under $H_1$, $\mu$ is unknown, with prior distribution $\mathcal{N}(0, \tau^2)$. Explain why the *Bayes factor* for choosing between $H_0$ and $H_1$ depends only on $\overline{X}$, and determine its value for data $X_1 = x_1, \ldots, X_n = x_n$.

The frequentist 5%-level test of $H_0$ against $H_1$ rejects $H_0$ when $|\overline{X}| \geq 1.96/\sqrt{n}$. What is the Bayes factor for the critical case $|\overline{X}| = 1.96/\sqrt{n}$? How does this behave as $n \to \infty$? Comment on the similarities or differences in behaviour between the frequentist and Bayesian tests.
Suppose that $X$ has density $f(\cdot|\theta)$ where $\theta \in \Theta$. What does it mean to say that statistic $T = T(X)$ is sufficient for $\theta$?

Suppose that $\theta = (\psi, \lambda)$, where $\psi$ is the parameter of interest, and $\lambda$ is a nuisance parameter, and that the sufficient statistic $T$ has the form $T = (C, S)$. What does it mean to say that the statistic $S$ is ancillary? If it is, how (according to the conditionality principle) do we test hypotheses on $\psi$? Assuming that the set of possible values for $X$ is discrete, show that $S$ is ancillary if and only if the density (probability mass function) $f(x|\psi, \lambda)$ factorises as

$$f(x|\psi, \lambda) = \varphi_0(x) \varphi_C(C(x), S(x), \psi) \varphi_S(S(x), \lambda)$$  \hspace{1cm} (*)

for some functions $\varphi_0$, $\varphi_C$, and $\varphi_S$ with the properties

$$\sum_{x \in C^{-1}(c) \cap S^{-1}(s)} \varphi_0(x) = 1 = \sum_s \varphi_S(s, \lambda) = \sum_c \sum_s \varphi_C(c, s, \psi)$$

for all $c$, $s$, $\psi$, and $\lambda$.

Suppose now that $X_1, \ldots, X_n$ are independent observations from a $\Gamma(a, b)$ distribution, with density

$$f(x|a, b) = (bx)^{a-1}e^{-bx}I_{\{x>0\}}/\Gamma(a).$$

Assuming that the criterion (*) holds also for observations which are not discrete, show that it is not possible to find $(C(X), S(X))$ sufficient for $(a, b)$ such that $S$ is ancillary when $b$ is regarded as a nuisance parameter, and $a$ is the parameter of interest.
2/II/27I  Principles of Statistics

(i) State Wilks' likelihood ratio test of the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$, where $\Theta_0 \subset \Theta_1$. Explain when this test may be used.

(ii) Independent identically-distributed observations $X_1, \ldots, X_n$ take values in the set $S = \{1, \ldots, K\}$, with common distribution which under the null hypothesis is of the form

$$P(X_1 = k|\theta) = f(k|\theta) \quad (k \in S)$$

for some $\theta \in \Theta_0$, where $\Theta_0$ is an open subset of some Euclidean space $\mathbb{R}^d$, $d < K - 1$. Under the alternative hypothesis, the probability mass function of the $X_i$ is unrestricted.

Assuming sufficient regularity conditions on $f$ to guarantee the existence and uniqueness of a maximum-likelihood estimator $\hat{\theta}_n(X_1, \ldots, X_n)$ of $\theta$ for each $n$, show that for large $n$ the Wilks' likelihood ratio test statistic is approximately of the form

$$\sum_{j=1}^{K} (N_j - n\hat{\pi}_j)^2 / N_j,$$

where $N_j = \sum_{i=1}^{n} I\{X_i = j\}$, and $\hat{\pi}_j = f(j|\hat{\theta}_n)$. What is the asymptotic distribution of this statistic?

---

3/II/26I  Principles of Statistics

(i) In the context of decision theory, what is a Bayes rule with respect to a given loss function and prior? What is an extended Bayes rule?

Characterise the Bayes rule with respect to a given prior in terms of the posterior distribution for the parameter given the observation. When $\Theta = \mathcal{A} = \mathbb{R}^d$ for some $d$, and the loss function is $L(\theta, a) = \|\theta - a\|^2$, what is the Bayes rule?

(ii) Suppose that $\mathcal{A} = \Theta = \mathbb{R}$, with loss function $L(\theta, a) = (\theta - a)^2$ and suppose further that under $P_\theta$, $X \sim N(\theta, 1)$.

Supposing that a $N(0, \tau^{-1})$ prior is taken over $\theta$, compute the Bayes risk of the decision rule $d_\lambda(X) = \lambda X$. Find the posterior distribution of $\theta$ given $X$, and confirm that its mean is of the form $d_\lambda(X)$ for some value of $\lambda$ which you should identify. Hence show that the decision rule $d_1$ is an extended Bayes rule.
Assuming sufficient regularity conditions on the likelihood \( f(x|\theta) \) for a univariate parameter \( \theta \in \Theta \), establish the Cramér–Rao lower bound for the variance of an unbiased estimator of \( \theta \).

If \( \hat{\theta}(X) \) is an unbiased estimator of \( \theta \) whose variance attains the Cramér–Rao lower bound for every value of \( \theta \in \Theta \), show that the likelihood function is an exponential family.
1/II/27J  Principles of Statistics

(a) What is a loss function? What is a decision rule? What is the risk function of a decision rule? What is the Bayes risk of a decision rule with respect to a prior \( \pi \)?

(b) Let \( \theta \mapsto R(\theta, d) \) denote the risk function of decision rule \( d \), and let \( r(\pi, d) \) denote the Bayes risk of decision rule \( d \) with respect to prior \( \pi \). Suppose that \( d^* \) is a decision rule and \( \pi_0 \) is a prior over the parameter space \( \Theta \) with the two properties

(i) \( r(\pi_0, d^*) = \min_d r(\pi_0, d) \)

(ii) \( \sup_{\theta} R(\theta, d^*) = r(\pi_0, d^*) \).

Prove that \( d^* \) is minimax.

(c) Suppose now that \( \Theta = \mathcal{A} = \mathbb{R} \), where \( \mathcal{A} \) is the space of possible actions, and that the loss function is

\[
L(\theta, a) = \exp(-\lambda a \theta),
\]

where \( \lambda \) is a positive constant. If the law of the observation \( X \) given parameter \( \theta \) is \( N(\theta, \sigma^2) \), where \( \sigma > 0 \) is known, show (using (b) or otherwise) that the rule

\[
d^*(x) = x/\sigma^2 \lambda
\]

is minimax.

2/II/27J  Principles of Statistics

Let \( \{f(\cdot|\theta) : \theta \in \Theta\} \) be a parametric family of densities for observation \( X \). What does it mean to say that the statistic \( T \equiv T(X) \) is sufficient for \( \theta \)? What does it mean to say that \( T \) is minimal sufficient?

State the Rao–Blackwell theorem. State the Cramér–Rao lower bound for the variance of an unbiased estimator of a (scalar) parameter, taking care to specify any assumptions needed.

Let \( X_1, \ldots, X_n \) be a sample from a \( U(0, \theta) \) distribution, where the positive parameter \( \theta \) is unknown. Find a minimal sufficient statistic \( T \) for \( \theta \). If \( h(T) \) is an unbiased estimator for \( \theta \), find the form of \( h \), and deduce that this estimator is minimum-variance unbiased. Would it be possible to reach this conclusion using the Cramér–Rao lower bound?
Principles of Statistics

Write an essay on the rôle of the Metropolis–Hastings algorithm in computational Bayesian inference on a parametric model. You may for simplicity assume that the parameter space is finite. Your essay should:

(a) explain what problem in Bayesian inference the Metropolis–Hastings algorithm is used to tackle;
(b) fully justify that the algorithm does indeed deliver the required information about the model;
(c) discuss any implementational issues that need care.

Principles of Statistics

(a) State the strong law of large numbers. State the central limit theorem.

(b) Assuming whatever regularity conditions you require, show that if

\[ \hat{\theta}_n \equiv \hat{\theta}_n(X_1, \ldots, X_n) \]

is the maximum-likelihood estimator of the unknown parameter \( \theta \) based on an independent identically distributed sample of size \( n \), then under \( P_\theta \)

\[ \sqrt{n}(\hat{\theta}_n - \theta) \to N(0, J(\theta)^{-1}) \]

in distribution

as \( n \to \infty \), where \( J(\theta) \) is a matrix which you should identify. A rigorous derivation is not required.

(c) Suppose that \( X_1, X_2, \ldots \) are independent binomial \( \text{Bin}(1, \theta) \) random variables. It is required to test \( H_0 : \theta = \frac{1}{2} \) against the alternative \( H_1 : \theta \in (0, 1) \). Show that the construction of a likelihood-ratio test leads us to the statistic

\[ T_n = 2n\{ \hat{\theta}_n \log \hat{\theta}_n + (1 - \hat{\theta}_n) \log(1 - \hat{\theta}_n) + \log 2 \}, \]

where \( \hat{\theta}_n \equiv n^{-1} \sum_{k=1}^{n} X_k \). Stating clearly any result to which you appeal, for large \( n \), what approximately is the distribution of \( T_n \) under \( H_0 \)? Writing \( \hat{\theta}_n = \frac{1}{2} + Z_n \), and assuming that \( Z_n \) is small, show that

\[ T_n \simeq 4nZ_n^2. \]

Using this and the central limit theorem, briefly justify the approximate distribution of \( T_n \) given by asymptotic maximum-likelihood theory. What could you say if the assumption that \( Z_n \) is small failed?
1/II/27I Principles of Statistics

State Wilks’ Theorem on the asymptotic distribution of likelihood-ratio test statistics.

Suppose that \(X_1, \ldots, X_n\) are independent with common \(N(\mu, \sigma^2)\) distribution, where the parameters \(\mu\) and \(\sigma\) are both unknown. Find the likelihood-ratio test statistic for testing \(H_0 : \mu = 0\) against \(H_1 : \mu \) unrestricted, and state its (approximate) distribution.

What is the form of the \(t\)-test of \(H_0 \) against \(H_1 \)? Explain why for large \(n\) the likelihood-ratio test and the \(t\)-test are nearly the same.

2/II/27I Principles of Statistics

(i) Suppose that \(X\) is a multivariate normal vector with mean \(\mu \in \mathbb{R}^d\) and covariance matrix \(\sigma^2 I\), where \(\mu\) and \(\sigma^2\) are both unknown, and \(I\) denotes the \(d \times d\) identity matrix. Suppose that \(\Theta_0 \subset \Theta_1\) are linear subspaces of \(\mathbb{R}^d\) of dimensions \(d_0\) and \(d_1\), where \(d_0 < d_1 < d\). Let \(P_i\) denote orthogonal projection onto \(\Theta_i\) \((i = 0, 1)\). Carefully derive the joint distribution of \(|X - P_1X|^2, |P_1X - P_0X|^2\) under the hypothesis \(H_0 : \mu \in \Theta_0\). How could you use this to make a test of \(H_0 \) against \(H_1 : \mu \in \Theta_1\)?

(ii) Suppose that \(I\) students take \(J\) exams, and that the mark \(X_{ij}\) of student \(i\) in exam \(j\) is modelled as

\[X_{ij} = m + \alpha_i + \beta_j + \varepsilon_{ij}\]

where \(\sum_i \alpha_i = 0 = \sum_j \beta_j\), the \(\varepsilon_{ij}\) are independent \(N(0, \sigma^2)\), and the parameters \(m, \alpha, \beta\) and \(\sigma\) are unknown. Construct a test of \(H_0 : \beta_j = 0\) for all \(j\) against \(H_1 : \sum_j \beta_j = 0\).

3/II/26I Principles of Statistics

In the context of decision theory, explain the meaning of the following italicized terms: loss function, decision rule, the risk of a decision rule, a Bayes rule with respect to prior \(\pi\), and an admissible rule. Explain how a Bayes rule with respect to a prior \(\pi\) can be constructed.

Suppose that \(X_1, \ldots, X_n\) are independent with common \(N(0, v)\) distribution, where \(v > 0\) is supposed to have a prior density \(f_0\). In a decision-theoretic approach to estimating \(v\), we take a quadratic loss: \(L(v,a) = (v-a)^2\). Write \(X = (X_1, \ldots, X_n)\) and \(|X| = (X_1^2 + \ldots + X_n^2)^{1/2}\).

By considering decision rules (estimators) of the form \(\hat{v}(X) = \alpha|X|^2\), prove that if \(\alpha \neq 1/(n+2)\) then the estimator \(\hat{v}(X) = \alpha|X|^2\) is not Bayes, for any choice of prior \(f_0\).

By considering decision rules of the form \(\hat{v}(X) = \alpha|X|^2 + \beta\), prove that if \(\alpha \neq 1/n\) then the estimator \(\hat{v}(X) = \alpha|X|^2\) is not Bayes, for any choice of prior \(f_0\).

[You may use without proof the fact that, if \(Z\) has a \(N(0,1)\) distribution, then \(EZ^4 = 3\).]
A group of \( n \) hospitals is to be ‘appraised’; the ‘performance’ \( \theta_i \) of hospital \( i \) has a \( N(0,1/\tau) \) prior distribution, different hospitals being independent. The ‘performance’ cannot be measured directly, so an expensive firm of management consultants has been hired to arrive at each hospital’s Standardised Index of Quality [SIQ], this being a number \( X_i \) for hospital \( i \) related to \( \theta_i \) by the commercially-sensitive formula

\[
X_i = \theta_i + \epsilon_i,
\]

where the \( \epsilon_i \) are independent with common \( N(0,1/\tau_\epsilon) \) distribution.

(i) Assume that \( \tau \) and \( \tau_\epsilon \) are known. What is the posterior distribution of \( \theta \) given \( X \)? Suppose that hospital \( j \) was the hospital with the lowest SIQ, with a value \( X_j = x \); conditional on \( X \), what is the distribution of \( \theta_j \)?

(ii) Now, instead of assuming \( \tau \) and \( \tau_\epsilon \) known, suppose that \( \tau \) has a Gamma prior with parameters \( (\alpha, \beta) \), density

\[
f(t) = (\beta t)^{\alpha-1} e^{-\beta t} / \Gamma(\alpha)
\]

for known \( \alpha \) and \( \beta \), and that \( \tau_\epsilon = \kappa \tau \), where \( \kappa \) is a known constant. Find the posterior distribution of \( (\theta, \tau) \) given \( X \). Comment briefly on the form of the distribution.