Tripos Questions in Statistics IB (1988–99)

99103

Let $X_1, \ldots, X_6$ be a sample from the uniform distribution on $[0, \theta]$ where $\theta \in [1, 2]$ is an unknown parameter. Find an unbiased estimate for θ of variance less than $1/10$.

99112

Let $X_1, \ldots, X_n$ be a sample from the uniform distribution on $[0, \theta]$, where $\theta \in (0, \infty)$ is an unknown parameter.

(a) Find a one-dimensional sufficient statistic T for θ and construct a 95% confidence interval for θ based on T.

(b) Suppose now that θ is a random variable having prior density

$$\pi(\theta) \propto 1_{\theta \geq a} \theta^{-k}$$

where $a > 0$ and $k > 2$. Compute the posterior density for θ and find the Bayes estimate $\hat{\theta}$ under the quadratic loss function $(\theta - \hat{\theta})^2$.

99203

Write a short account of the standard procedure used by statisticians for hypothesis testing. Your account should explain, in particular, why the null hypothesis is considered differently from the alternative and also say what is meant by a likelihood ratio test.

99212

State and prove the Neyman-Pearson lemma. Explain what is meant by a uniformly most powerful test.

Let $X_1, \ldots, X_n$ be a sample from the normal distribution of mean θ and variance 1, where $\theta \in \mathbb{R}$ is an unknown parameter. Find a uniformly most powerful test of size $1/100$ for

$$H_0 : \theta \leq 0, \quad H_1 : \theta > 0,$$

expressing your answer in terms of an appropriate distribution function. Justify carefully that your test is uniformly most powerful of size $1/100$.

99403

Students of mathematics in a large university are given a percentage mark in their annual examination. In a sample of 9 students the following marks were found:

28 32 34 39 41 42 42 46 56

Students of history also receive a percentage mark. A sample of 5 students reveals the following marks:

53 58 60 61 68

Do these data support the hypothesis that the marks for mathematics are more variable than the marks for history? Quantify your conclusion. Comment on your modelling assumptions.

distribution	$N(0,1)$	$F_{9,5}$	$F_{8,4}$	X^2_{14}	X^2_{13}	X^2_{12}
95% percentile	1.65	4.78	6.04	23.7	22.4	21.0

99412

Consider the linear regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right), \quad i + 1, \ldots, n,$$

where $x_1, \ldots, x_n$ are known, with $\sum_{i=1}^n x_i = 0$, and where $\alpha, \beta \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ are unknown. Find the maximum likelihood estimators $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2$ and write down their distributions.

Consider the following data:

x_i	-3	-2	-1	0	1	2	3
Y_i	-5	0	3	4	3	0	-5

Fit the linear regression model and comment on its appropriateness.

98103

The independent observations X_1, X_2 are distributed as Poisson random variables, with means μ_1, μ_2 respectively, where

$$\begin{aligned} \log \mu_1 &= \alpha, \\ \log \mu_2 &= \alpha + \beta\,, \end{aligned}$$

with α and β unknown parameters. Write down $\ell(\alpha, \beta)$, the log-likelihood function, and hence find the following:

(i) $\dfrac{\partial^2 \ell}{\partial \alpha^2}, \quad \dfrac{\partial^2 \ell}{\partial \alpha \beta}, \quad \dfrac{\partial^2 \ell}{\partial \beta^2},$

(ii) $\hat{\beta}$, the maximum likelihood estimator of β.

98112

The lifetime T of certain electronic components may be assumed to follow the negative exponential density

$$f(t;\theta) = \frac{1}{\theta} \exp\left(-\frac{t}{\theta}\right), \quad \text{for } t \geq 0 \,,$$

where t is the sampled value of T.

Let $t_1, \ldots, t_n$, be a random sample from this density. Quoting carefully the Neyman-Pearson lemma, find the form of the most powerful test of size 0.05 of

$$H_0 : \theta = \theta_0, \quad \text{against} \quad H_1 : \theta = \theta_1$$

where θ_0 andθ_1 are given, $\theta_0 < \theta_1$. Defining the function

$$G_n(u) = \int_0^u e^{-t} \frac{t^{n-1}}{(n-1)!} \, dt \,,$$

show that this test has power $1 - G_n\left(\frac{\theta_0}{\theta_1} \, G_n^{-1}(1-\alpha)\right)$, where $\alpha = 0.05$.

If for $n = 100$, you observed $\sum_i t_i/n = 3.1$, would you accept the hypothesis $H_0 : \theta = 2$? Give reasons for your answer, using the large sample distribution of $(T_1 + \cdots + T_n)/n$.

[*This question can be answered without calculators or statistical tables.*]

98203

Consider the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad \text{for } 1 \leq i \leq n \,,$$

where $x_1, \ldots, x_n$ are given values, with $\sum_i x_i = 0$, and where $\epsilon_1, \ldots, \epsilon_n$ are independent normal errors, each with zero mean and known variance σ^2.

(i) Obtain equations for $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, the maximum likelihood estimates of $(\beta_0, \beta_1, \beta_2)$. Do not attempt to solve these equations.

(ii) Obtain an expression for β_1^* the maximum likelihood estimate of β_1 in the reduced model

$$H_0 : y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad 1 \leq i \leq n \,,$$

with $\sum_i x_i = 0$ and $\epsilon_1, \ldots, \epsilon_n$ distributed as above.

98212

Let $(x_1, \ldots, x_n)$ be a random sample from the normal density with mean μ and variance σ^2.

(i) Write down the log-likelihood function $\ell(\mu, \sigma^2)$.

(ii) Find a pair of sufficient statistics, for the unknown parameters (μ, σ^2), carefully quoting the relevant theorem.

(iii) Find $(\hat{\mu}, \hat{\sigma}^2)$, the maximum likelihood estimators of (μ, σ^2). Quoting carefully any standard distributional results required, show how to construct a 95% confidence interval for μ.

98403

Suppose that, given the real parameter θ, the observation X is normally distributed with mean θ and variance v, where v is known. If the prior density for θ is

$$\pi(\theta) \propto \exp\left(-(\theta-\mu_0)^2/2v_0\right),$$

where μ_0 and v_0 are given, show that the posterior density for θ is $\pi(\theta|x)$, where

$$\pi(\theta|x) \propto \exp\left(-(\theta-\mu_1)^2/2v_1\right),$$

and μ_1 and v_1 are given by

$$\mu_1 = \frac{\left(\frac{\mu_0}{v_0}+\frac{x}{v}\right)}{\left(\frac{1}{v_0}+\frac{1}{v}\right)}, \quad \frac{1}{v_1} = \frac{1}{v_0}+\frac{1}{v}.$$

Sketch typical curves $\pi(\theta)$ and $\pi(\theta|x)$, with μ_0 and x marked on the θ-axis.

98412

Let (n_{ij}) be the observed frequencies for an $r \times c$ contingency table, let $n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$ and let

$$\mathbb{E}(n_{ij}) = np_{ij}, \quad 1 \le i \le r,\ 1 \le j \le c,$$

thus $\sum_i \sum_j p_{ij} = 1$.

Under the usual assumption that (n_{ij}) is a multinormal sample, show that likelihood ratio statistic for testing

$$H_0 : p_{ij} = \alpha_i \beta_j$$

for all (i, j) and for some α, β, is

$$D = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log(n_{ij}/e_{ij}) \, ,$$

where you should define (e_{ij}). Show further that for $|n_{ij} - e_{ij}|$ small, the statistic D may be approximated by

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (n_{ij} - e_{ij})^2 / e_{ij} \, .$$

In 1843 William Guy collected the following data on 1659 outpatients at a particular hospital, showing their physical exertion at work and whether they had pulmonary consumption or some other disease

	Disease type	
Level of exertion at work	pulmonary consumption	other disease
Little	125	385
Varied	41	136
More	142	630
Great	33	167

For these data, X^2 was found to be 9.84. What do you conclude?

[*Note that this question can be answered without calculators or statistical tables.*]

97103

In a large group of young couples, the standard deviation of the husbands' ages is four years, and that of the wives' ages is three years. Let D denote the age difference within a couple.

Under what circumstances might you expect to find the standard deviation of the Ds in the group to be about 5 years?

Instead you find it to be two years. One possibility is that the discrepancy is the result of random variability. Give another possible explanation.

97112

Suppose that $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ form two independent samples, the first from an exponential distribution with parameter λ, and the second from an exponential distribution with parameter μ.

(i) Construct the likelihood ratio test of $H_0 : \lambda = \mu$ versus $H_i : \lambda \neq \mu$.

(ii) Show that the test in part (i) can be based on the statistic

$$T = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i}.$$

(iii) Describe how the percentiles of the distribution of T under H_0 may be determined from the percentiles of an F-distribution.

97203

Explain what is meant by a *sufficient statistic*.

Consider the independent random variables $X_1, X_2, \ldots, X_n$, where $X_i \sim N(\alpha + \beta c_i, \theta)$ for given constants c_i, $i = 1, 2, \ldots, n$, and unknown parameters α, β and θ. Find three sample quantities that together constitute a sufficient statistic.

97212

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\theta, \sigma^2)$ distribution, and suppose the prior distribution for θ is the $N(\mu, \tau^2)$ distribution, where σ^2, μ, and τ^2 are known. Determine the posterior distribution for θ, given $X_1, X_2, \ldots, X_n$, and the best point estimate of θ under (i) quadratic loss, and (ii) absolute error loss.

97403

$X_1, X_2, \ldots, X_n$ form a random sample from a uniform distribution on the interval $(-\theta, 2\theta)$, where the value of the parameter θ is unknown. Determine the maximum likelihood estimate of the parameter θ.

97412

The χ^2 statistic is often used as a measure of discrepancy between observed frequencies and the expected frequencies under a null hypothesis. Describe the χ^2 statistic, and the χ^2 test for goodness of fit.

The number of directory enquiry calls arriving each day at a centre are counted over K weeks. It may be assumed that the number of such calls on any given day has a Poisson distribution, that the numbers of calls on different days are independent, and that the expected number of calls depends only on the day of the week. Let n_i, $i = 1, 2, \ldots, 7$ denote, respectively, the total number of calls received on a Monday, Tuesday, ... , Sunday.

Derive an approximate test of the hypothesis that calls are received at the same rate on all days of the week except Sundays.

Find also a test of a second hypothesis, that the expected number of calls received are equal for the three days from Tuesday to Thursday, and that the expected number of calls received are equal on Monday and Friday.

96103

(a) Aerial observations x_1, x_2, x_3, x_4 are made of the interior angles $\theta_1, \theta_2, \theta_3, \theta_4$ of a quadrilateral on the ground. If these observations are subject to small independent errors with zero means and common variance σ^2, determine the least-squares estimates of $\theta_1, \theta_2, \theta_3, \theta_4$.

(b) Obtain an unbiased estimate of σ^2 in the situation described in part (a).

Suppose now that the quadrilateral is known to be a parallelogram with $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$. What now are the least-squares estimate of its angles? Obtain an unbiased estimator of σ^2 in this case.

96203

(a) $X_1, X_2, \ldots, X_n$ form a random sample from a distribution whose probability density function is

$$f(x \mid \theta) = \begin{cases} 2x/\theta^2 & 0 \leq x \leq \theta \\ 0 & \text{otherwise,} \end{cases}$$

where the value of the positive parameter θ is unknown. Determine the maximum likelihood estimate of the median of the distribution.

(b) There is widespread agreement amongst the managers of the Reliable Motor Company that the number x of faulty cars produced in a month has a binomial distribution

$$P(x = s) = \binom{n}{s} p^s (1-p)^{n-s} \quad (s = 0, 1, \ldots, n;\ 0 \leq p \leq 1).$$

There is, however, some dispute about the parameter p. The general manager has a prior distribution for p which is uniform, while the more pessimistic production manager has a prior distribution with density $2p$, both on the interval $[0, 1]$.

In a particular month, s faulty cars are produced. Show that if the general manager's loss function is $(\hat{p} - p)^2$, where $\hat{p}$ is her estimate and p is the true value, then her best estimate of p is

$$\hat{p} = \frac{s+1}{n+2}.$$

The production manager has responsibilities different from those of the general manager, and a different loss function given by $(1-p)(\hat{p} - p)^2$. Find his best estimate of p and show that it is greater than that of the general manager unless $s \geq \frac{1}{2}n$.

[*You may assume that, for non-negative integers α, β,*

$$\int_0^1 p^\alpha (1-p)^\beta \, dp = \frac{\alpha! \beta!}{(\alpha + \beta + 1)!} \ . \Bigg]$$

96403

(a) What is a *simple hypothesis*? Define the terms *size* and *power* for a test of one simple hypothesis against another.

State and prove the Neyman-Pearson lemma.

(b) There is a single observation of a random variable X which has a probability density function $f(x)$. Construct a best test of size 0.05 for the null hypothesis

$$H_0 : f(x) = \tfrac{1}{2} \quad (-1 \leq x \leq 1)$$

against the alternative hypothesis

$$H_1 : f(x) = \tfrac{3}{4}(1 - x^2) \quad (-1 \leq x \leq 1)\,.$$

Calculate the power of your test.

95103

(a) Let $X_1, \ldots, X_m$ be a random sample from the $N(\mu_1, \sigma^2)$-distribution and let $Y_1, \ldots, Y_n$ be an independent sample from the $N(\mu_1, \sigma^2)$-distribution. Here the parameters μ_1, μ_2 and σ^2 are all unknown. Explain carefully how you would test the hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.

(b) Let $X_1, \ldots, X_n$ be a random sample from the distribution with the probability density function

$$f(x \mid \theta) = e^{-(x-\theta)}, \quad \text{for } \theta < x < \infty,$$

where θ has a prior distribution the standard normal $N(0, 1)$. Determine the posterior distribution of θ.

Suppose that θ is to be estimated when the loss function is the absolute error loss, $L(a, \theta) = |a - \theta|$. Determine the optimal point estimate and express it in terms of the function $c_n(x)$ defined by

$$2\Phi(c_n(x) - n) = \Phi(x - n), \quad \text{for } -\infty < x < \infty,$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$ is the standard normal distribution function.

95203

(a) Let $X_1, \ldots, X_n$ be a random sample from the distribution with the probability density function

$$f(x \mid \theta) = \frac{2x}{\theta^2}, \quad \text{for } 0 \leq x \leq \theta.$$

Determine the maximum-likelihood estimate M of θ and show that $\left(M, M/(1-\gamma)^{\frac{1}{2n}}\right)$ is a $100\gamma\%$ confidence interval for θ, where $0 < \gamma < 1$.

(b) Let $X_1, \dots, X_n$ be a random sample from the uniform distribution on $[0, \theta_1]$ and let $Y_1, \dots, Y_n$ be an independent random sample from the uniform distribution on $[0, \theta_2]$. Derive the form of the likelihood-ratio test of the hypothesis $H_0 : \theta_1 = \theta_2$ against $H_1 : \theta_1 \neq \theta_2$ and express this test in terms of the statistic

$$T = \frac{\max(M_X, M_Y)}{\min(M_X, M_Y)},$$

where $M_X = \max_{1 \le i \le n} X_i$ and $M_Y = \max_{1 \le i \le n} Y_i$.

By observing that under the hypothesis H_0 the distribution of T is independent of $\theta = \theta_1 = \theta_2$, or otherwise, determine exactly the critical region for the test of size α.

95403

(a) State and prove the Neyman-Pearson lemma.

(b) Let $X_1, \dots, X_n$ be a random sample from the $N(\mu, \sigma^2)$-distribution. Prove that the sample mean $\bar{X}$ and the sample variance $\sum_{i=1}^{n}(X_i - \bar{X})^2$ are independent random variables and determine their distributions.

Suppose that

$$\begin{array}{ccc} X_{1,1} & \cdots & X_{1,n} \\ X_{2,1} & \cdots & X_{2,n} \\ \vdots & & \vdots \\ X_{m,1} & \cdots & X_{m,n} \end{array}$$

are independent random variables and that $X_{i,j}$ has the $N(\mu_i, \sigma^2)$-distribution for $1 \le j \le n$, where $\mu_1, \dots, \mu_m, \sigma^2$ are unknown constants. With reference to the previous result, explain carefully how you would test the hypothesis $H_0 : \mu_1 = \cdots = \mu_m$.

94103

(a) At a particular time three High Street restaurants are observed to have 43, 41 and 30 customers respectively. Detailing carefully the underlying assumptions that you are making, explain how you would test the hypothesis that all three restaurants are equally popular against the alternative that they are not.

(b) Explain the following terms in the context of hypothesis testing:

(i) simple hypothesis;

(ii) composite hypothesis;

(iii) Type I and Type II error probabilities;

(iv) size of a test; and

(v) power of a test.

Let $X_1, \dots, X_n$ be a sample from the $N(\mu, 1)$-distribution. Construct the most powerful size-α test of the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, where $\mu_1 > \mu_0$.

Find the test that minimizes the larger of the two error probabilities. Justify your answer carefully.

94203

(a) Let $X_1, \dots, X_n$ be a sample from the $N(\mu, \sigma_1)$-distribution and let $Y_1, \dots, Y_n$ be an independent sample from the $N(\mu, \sigma_2)$-distribution. Here the parameters μ_1, μ_2, σ_1^2 and σ_2^2 are all unknown. Explain carefully how you would test the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$.

(b) Let $Y_1, \dots, Y_n$ be independent random variables where Y_i has the $N(\beta x_i, \sigma^2)$-distribution, $i = 1, \dots, n$. Here $x_1, \dots, x_n$ are known but β and σ^2 are unknown.

(i) Determine the maximum-likelihood estimates $(\hat{\beta}, \hat{\sigma}^2)$ of (β, σ^2).

(ii) Find the distribution of $\hat{\beta}$.

(iii) By showing that $Y_i - \hat{\beta} x_i$ and $\hat{\beta}$ are independent, or otherwise, determine the joint distribution of $\hat{\beta}$ and $\hat{\sigma}^2$.

(iv) Explain carefully how you would test the hypothesis $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$.

94403

(a) Let X be a random variable with the probability density function

$$f(x|\theta) = e^{-(x-\theta)}, \quad \theta < x < \infty,$$

where θ has as prior distribution the exponential distribution with mean 1. Determine the posterior distribution of θ.

Find the optimal point estimate of θ based on X under quadratic loss.

(b) Let $X_1, \dots, X_n$ be a sample from the probability density function

$$f(x|\lambda, \mu) = \begin{cases} \frac{1}{\lambda+\mu} e^{-x/\lambda}, & x \geq 0, \\ \frac{1}{\lambda+\mu} e^{x/\mu}, & x < 0, \end{cases}$$

where $\lambda > 0$ and$\mu > 0$ are unknown parameters. Find (simple) sufficient statistics for (λ, μ), and determine the maximum-likelihood estimates $(\hat{\lambda}, \hat{\mu})$ of (λ, μ).

Now suppose that $n = 1$. Is $\hat{\lambda}$ an unbiased estimate of λ? Justify your answer

93103

(a) A sample $x_1, \dots, x_n$ is taken from a normal distribution with an unknown mean μ and a known variance σ^2. Show how to construct a most powerful test of a given size $\alpha \in (0, 1)$ for a null hypothesis $H_0 : \mu = \mu_0$ against an alternative $H_1 : \mu = \mu_1$ ($\mu_0 \neq \mu_1$).

What is the value of α for which the power of this test is 1/2?

(b) State and prove the Neyman-Pearson Lemma. For the case of simple null and alternative hypotheses, what sort of test would you propose for minimizing the sum of the probabilities of type I and type II errors? Justify your answer.

93203

(a) Explain what is meant by constructing a confidence interval for an unknown parameter θ from a given sample $x_1, \ldots, x_n$. Let a family of probability density functions $f(x;\theta)$, $-\infty < \theta < \infty$, be given by

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta, \\ 0, & x < \theta. \end{cases}$$

Suppose that $n = 4$ and $x_1 = -1.0$, $x_2 = 1.5$, $x_3 = 0.5$, $x_4 = 1.0$. Construct a 95% confidence interval for θ.

(b) Let $f(x;\mu,\sigma^2)$ be a family of normal probability density functions with an unknown mean μ and an unknown variance $\sigma^2 > 0$. Explain how to construct a 95% confidence interval for μ from a sample $x_1, \ldots, x_n$. Justify the claims about the distributions you use in your construction.

93403

(a) State and prove the factorization criterion for sufficient statistics, in the case of discrete random variables.

(b) A linear function $y = Ax + B$ with unknown coefficients A and B is repeatedly measured at distinct points $x_1, \ldots, x_k$: first n_1 times at x_1, then n_2 times at x_2, and so on; and finally n_k times at x_k. The result of the ith measurement series is a sample $y_{i1}, \ldots, y_{in_i}$, $i = 1, \ldots, k$. The errors of all measurements are independent normal variables, with mean zero and variance one. You are asked to estimate A and B from the whole sample y_{ij}, $1 \leq j \leq n_i$, $1 \leq i \leq k$. Prove that the maximum likelihood and the least squares estimators of (A, B) coincide and find these.

Denote by $\hat{A}$ the maximum likelihood estimator of A and by $\hat{B}$ the maximum likelihood estimator of B. Find the distribution of $(\hat{A}, \hat{B})$.

92103 sample

(a) Let $x_1, \ldots, x_n$ be a random sample from the probability density function $f(x;\theta)$. What is meant by saying that $t(x_1, \ldots, x_n)$ is sufficient for θ?

Let

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)}, & x > \theta, \\ 0, & x \leq \theta, \end{cases}$$

and suppose $n = 3$. Let $y_1 < y_2 < y_3$ be the ordered values of x_1, x_2, x_3. Show that y_1 is sufficient for θ.

(b) Show that the distribution of $Y_1 - \theta$ is exponential of parameter 3. Your client suggest the following possibilities as estimates of θ:

$$\bar{\theta}_1 = x_3 - 1$$
$$\bar{\theta}_2 = y_1$$
$$\bar{\theta}_3 = \tfrac{1}{3}(x_1 + x_2 + x_3) - 1.$$

Is he being sensible; how would you advise him?

[**Hint**: *any general theorems used should be clearly stated, but need not be proved.*]

92203 sample

(a) Derive the form of the maximum likelihood estimators of α, β and σ^2 in the linear model

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

$1 \le i \le n$, where $\epsilon \sim N(0, \sigma^2)$ and $\sum_{i=1}^n x_i = 0$.

(b) What is the joint distribution of the maximum likelihood estimators $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma^2}$? Construct 95% confidence intervals for

(i) σ^2,

(ii) $\alpha + \beta$.

92403 sample

(a) describe briefly a procedure for obtaining a Bayesian point estimate from a statistical experiment. Include in your description-n definitions of the terms:

(i) prior; (ii) posterior.

(b) Let $X_1, \dots, X_n$ be independent identically distributed random variables, each having a Gamma (k, λ) distribution. Suppose k is known, and *a priori*, λ is exponential of parameter μ. Suppose a penalty of $(a - \lambda)^2$ is incurred on estimating λ by a. Calculate the posterior for λ and find an optimal point estimate for λ.

92106

Let $X_1, X_2, \dots, X_n$ be an independent sample from a normal distribution with unknown mean μ and variance σ^2. Show that the pair $(\bar{X}, \bar{S}^2)$ where

$$\bar{X} = \frac{1}{n} \sum i = 1^n X_i, \quad \bar{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

is a sufficient statistic for (μ, σ^2).

Given $\lambda > 0$, consider $\lambda \bar{A}^2$ as an estimator of σ^2. For what values of λ is $\bar{S}^2$

(a) maximum likelihood,

(b) unbiased?

Which value of λ minimizes the mean square error

$$E(\lambda \bar{S}^2 - \sigma^2)^2 \ ?$$

92206

Suppose you are given a collection of np independent random variables organized in n samples, each of length p:

$$\begin{array}{l} X^{(1)} = (X_{11}, \dots, X_{1p}) \\ X^{(2)} = (X_{21}, \dots, X_{2p}) \\ \cdots \\ X^{(n)} = (X_{n1}, \dots, X_{np}). \end{array}$$

The random variable X_{ij} has a Poisson distribution with an unknown parameter λ_j, $1 \le j \le p$. You are required to test the hypothesis that $\lambda_1 = \cdots = \lambda_p$ against the alternative that at least two of the λ_j's are distinct. Derive the form of the Likelihood Ratio Test Statistic. Show that it may be approximated by

$$\frac{n}{\bar{X}} \sum_{j=1}^{p} (\bar{X}_j - \bar{X})^2$$

with

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}, \quad \bar{X} = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} X_{ij}.$$

Explain how you would test the hypothesis for large n.

92306

Let $X_1, X_2, \dots, X_n$ be an independent sample from a normal distribution with a known mean μ and an unknown variance σ^2 taking one of two values σ_1^2 and σ_2^2. Explain carefully how to construct a most powerful test of size α of the hypothesis $\sigma = \sigma_1$ against the alternative $\sigma = \sigma_2$. Does there exist a most powerful test of size α with power strictly less than α? Justify your answer.

92406

Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ be independent random variables each with the $N(0,1)$ distribution, and $x_1, x_2, \ldots, x_n$ be fixed real numbers. Let the random variables $Y_1, Y_2, \ldots, Y_n$ be given by

$$Y_i = \alpha + \beta x_i + \sigma \epsilon_i, \quad 1 \leq i \leq n,$$

where $\alpha, \beta \in \mathbb{R}$ and $\sigma \in (0, \infty)$ are unknown parameters. Derive the form of the Least Squares Estimator for the pair (α, β) and establish the form of its distribution. Explain how to test the hypothesis $\beta = 0$ against $\beta \neq 0$ and how to construct a 95% confidence interval for β.

[*General results used should be stated carefully, but need not be proved.*]

91106

A sample $X_1, \ldots, X_n$ of independent observations comes from a normal distribution with mean θ and variance 1. Find a test of the hypothesis $\theta = \theta_0$ against the alternative $\theta = \theta_1$ (where θ_0 and θ_1 are given values with $\theta_0 < \theta_1$) which has the property that no other test of the same size has larger power.

Find an expression for the smallest value of n for which it is possible for this test to have size ≤ 0.025 and power ≥ 0.95.

[*If Φ is the standard normal distribution function, then*

$$\Phi(1.282) = 0.90, \quad \Phi(1.645) = 0.95, \quad \Phi(1.960) = 0.975.]$$

91206

A treatment is suggested for a particular illness, and the results of treating a number of patients chosen at random from those in a hospital suffering from the illness are shown in the following table, in which the entries a, b, c, d are numbers of patients.

	Recovery	Non-recovery
Untreated	a	b
Treated	c	d

Describe the use of Pearson's χ^2 statistic in testing whether the treatment affects recovery, and outline a justification for its use. Show that

$$\chi^2 = \frac{(ad - bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}.$$

[*You may find it helpful to observe that*

$$(a(a+b+c+d) - (a+b)(a+c))^2 = (ad - bc)^2.]$$

Comment on the use of this statistical technique when

$$a = 50, \quad b = 10, \quad c = 15, \quad d = 5.$$

91306

Let $x_1, x_2, \ldots, x_n$ be real numbers such that

$$\sum_{i=1}^{n} x_i = 0, \quad \sum_{i=1}^{n} x_i^2 = 1.$$

Suppose that $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent random variables each with the $N(0,1)$ distribution and that

$$Y_i = \alpha x_i + \sigma \epsilon_i \quad (1 \leq i \leq n),$$

where a in $\mathbb{R}$ and σ in $(0, \infty)$ are unknown parameters. Let (A, V) denote the Maximum-Likelihood Estimator for the pair (a, σ^2). Prove that $(A - a)/\sigma$ has the $N(0,1)$ distribution and that

$$nV + (A-a)^2 = \sigma^2 \epsilon^\top \epsilon, \quad (\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top).$$

[*You may find it helpful to prove that*

$$nV = \eta^\top \eta, \text{ where } \eta = (I - xx^\top)Y.]$$

Assuming that nV is independent of $A - a$, derive the distribution of nV and show how to test

$$H_0 : a = 1, \sigma > 0 \text{ against } H_1 : a \neq 1, \sigma > 0.$$

91406

Let $X_1, X_2, \ldots, X_n$ be a random sample from the density $f(x_1, x_2, \ldots, x_n \mid \theta)$. What is meant by saying that $T(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for the unknown parameter θ? State, without proof, the factorization theorem for a sufficient statistic.

An engineer interested in failures in a large computer system observes that in the first 7 months of 1990 there were x_1 failures, and in the first 4 months of 1991 there were x_2 failures. The engineer proposes to estimate the rate per month by using the estimator $S = \frac{1}{2}(\frac{x_1}{7} + \frac{x_2}{4})$. Assuming the failures occur as a Poisson process of rate θ per month, show that S is unbiased for θ and find its variance. Find a sufficient statistic for θ, and hence construct an unbiased estimator of θ with a smaller variance than S.

90106

(a) State the Neyman-Pearson lemma for testing $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda = \lambda_1$, given $(x_1, \ldots, x_n)$ a random sample of size n from a known density $f(x \mid \lambda)$.

(b) Find the form of the corresponding test for the case where

$$f(x \mid \lambda) = \frac{1}{\lambda} e^{-x/\lambda} \text{ for } \lambda_0 = 1 \text{ and } \lambda_1 \text{ given, } \lambda_1 > 1.$$

Show that for $n = 1$, and α fixed, β is a decreasing function of λ_1, where α, β are the Type I and Type II error probabilities.

(c) How is your answer to (b) affected if now $\lambda_1 < 1$?

90206

(a) Show that if Z has a χ^2_ν distribution, then Z has mean ν, variance 2ν.

(b) For a fixed sample size n, the cell frequencies $(n_0, n_1, \dots, n_k)$ have the multinomial distribution with frequency function

$$f(n_0, n_1, \dots, n_k \mid p) = \frac{1}{n!} \prod_{j=0}^{k} \frac{p_j^{n_j}}{n_j!}$$

for $n_0, \dots, n_k \geq 0$, $n_0 + \cdots + n_k = n$, where $p_0 + \cdots + p_k = 1$ and $(p_0, \dots, p_k)$ unknown.

Given the frequencies $(n_0, n_1, \dots, n_k)$, explain how to test the null hypothesis H_0 of a binomial distribution, i.e.,

$$H_0 : p_j = \left(\begin{array}{c} k \\ j \end{array} \right) \theta^j (1-\theta)^{k-j}, \quad 0 \leq j \leq k,$$

where θ is unknown, $0 \leq \theta \leq 0$, and n is large enough for an approximate method to be used. Illustrate your test in the case $k = 3$, $n = 80$, $n_0 = 40$, $n_1 = 4$, $n_2 = 6$, $n_3 = 30$.

[*Any standard theorems used should be carefully quoted. Note that standard tables are not needed for this question.*]

90306

Let $x_1, \dots, x_n$ be a random sample from the probability distribution $f(x \mid \theta)$. What is meant by saying that $t(x)$ is a *sufficient statistic* for the unknown parameter θ? State, without proof, the factorisation theorem for a sufficient statistic.

Suppose now that $f(x \mid \theta) = \dfrac{1}{\sqrt{2\pi\theta}} \exp\left(\dfrac{-x^2}{2\theta}\right), \quad -\infty < x < \infty.$

Find $t(x)$, a sufficient statistic for θ, and find $\hat{\theta}$, the maximum likelihood estimate of θ. What are the mean and variance of $\hat{\theta}$?

[*Note that* $\int_{-\infty}^{\infty} x^2 e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = 3$.]

90406

Observations $y_1, \dots, y_n$ are distributed according to the following model

$$y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i,$$

where $\epsilon_1, \dots, \epsilon_n$ are independent normal errors with mean 0, variance σ^2, and α, β, σ^2 are unknown, $x_1, \dots, x_n$ are known and fixed.

(a) Find $(\hat{\alpha}, \hat{\beta})$, the least squares estimators of (α, β).

(b) Show that $\hat{\alpha}, \hat{\beta}$ are independent, and that $\hat{\beta}$ is normal, mean β, variance σ^2/S_{xx}, where $S_{xx} = \sum(x_i - \bar{x})^2$.

(c) Let $R = \sum[y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2$. For a given data set $\hat{\beta} = 3.7$, $\dfrac{R}{S_{xx}(n-2)} = 0.8$, $n = 12$. Would you accept the hypothesis that $\beta = 0$?

[*The upper 99% point of* t_{10} *is 2.76.*]

[*Any standard results may be quoted without proof.*]

89106

A sample $X_1, \ldots, X_n$ is believed to arise from a normal distribution with mean θ and variance 1. We wish to decide which of two given values $\theta_0 < \theta_1$ is taken by θ. The *most powerful* test of a given *size* uses a *critical region* of the form $\{x \in \mathbb{R}^n : \sum_{i=1}^n x_i \geq k\}$. Explain the italicised terms, and prove from first principles the assertion of the preceding sentence.

Suppose $\theta_0 = 0$ and $n = 9$, and the following observations are recorded: 0.3, 0.3, 0.6, 0.9, 0.9, 1.2, 1.5, 1.5. At what level of significance could you reject the hypothesis $\theta = \theta_0$?

If you had no prior knowledge of the variance of the sample, would you have been more or less sure that $\theta \neq \theta_0$?

$P(N(0,1) \geq z)$	0.25	0.10	0.05	0.025
z	0.67	1.28	1.64	1.96

$P(t_8 \geq z)$	0.25	0.10	0.05	0.025
z	0.71	1.40	1.82	2.31

89206

A sample of size n is drawn from a normal distribution of unknown mean, and unknown variance θ. Derive the form of generalized likelihood ratio tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Prove that a multiple of the sample variance has χ^2_{n-1} distribution.

The following data are assumed to arise from a normal distribution of unknown mean and variance: 8, 9, 9, 10, 12, 12, 13, 15. Are they consistent with the contention that the variance is less than 5?

$P(\chi^2_7 \geq z)$	0.05	0.025	0.01
z	14.1	16.0	18.5

89306

State and prove the factorization criterion for sufficient statistics of discrete random variables. Why can the search for a maximum likelihood estimator of any parameter be confined to sufficient statistics?

Find the maximum likelihood estimator of (p, q, r) based on a sample $X_1, \ldots, X_n$ from a trinomial distribution

$$P(X_i = (k_1, k_2, k_3)) = \frac{m!}{(k_1)!(k_2)!(k_3)!} p^{k_1} p^{k_2} p^{k_3},$$

where $m \in \mathbb{Z}^+$ is known, $k_1 + k_2 + k_3 = m$, and where $p, q, r \geq 0$, $p + q + r = 1$.

89406

In a survey, thirty members of each of four social classes, A, B, C, D were asked to reveal their personal income: the first table summarizes their answers.

Income (£1000's)	Social Class			
	A	B	C	D
0 to 10	2	0	5	13
10 to 20	16	13	18	13
More than 20	12	17	7	4

Do the data indicate a significant correlation between income and class?

An official of the Inland Revenue points out that if participants were able to reply anonymously a more accurate conclusion might be reached. So, a new survey (with a new set of people) is performed as suggested, the results being tabulated as before.

Income (£1000's)	Social Class			
	A	B	C	D
0 to 10	3	0	1	12
10 to 20	14	7	22	12
More than 20	13	23	7	6

Can you say, with any certainty, whether the new survey technique made a difference (a) to class B, (b) overall?

[*Theoretical justification of general principles is not required here. You will need to refer to the table of percentiles below.*]

	95%	99%	99.5%
χ_1^2	3.84	6.64	7.88
χ_2^2	5.99	9.21	10.60
χ_3^2	7.82	11.34	12.84
χ_4^2	9.49	13.28	14.86
χ_6^2	12.59	16.81	18.55
χ_8^2	15.51	20.09	21.96

88106

Explain what is meant by the term *sufficient statistic*.

A radioactive source emits both α-particles and β-particles, the times between emissions being exponential of rates λ and μ respectively. A Geiger counter A registers the incidence of α particles only; another counter B is sensitive to both sorts of particle. In an experiment using counter A the number of particles registered in the ith unit of time is X_i, $1 \leq i \leq n$. The experiment is later repeated with counter B, with outcomes $Y_, \ldots, Y_n$. Assuming that all emitted particles reach the counters being used, what are the distributions of X_1 and Y_1?

Show that the maximum likelihood estimate of μ when λ is known is given by $\max\{(\bar{Y}-\lambda), 0\}$, where $\bar{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n)$. Suppose now that λ is unknown. Show that $(\bar{X}, \bar{Y})$ is sufficient for (λ, μ) and find the maximum likelihood estimator of μ.

88306

A scientist performs a series of experiments on a fixed mass of gas at constant temperature. In the ith experiment the gas us subjected to a pressure P_i and its volume measured as V_i. The observed values of $\log P_i$ and $\log V_i$ are given below. We believe that $\log V_i$ is measured subject to an error of distribution $N(0, \frac{1}{10})$ and that the true volume satisfies PV^γ =constant. Are the data consistent with the conjecture that $\gamma = 1$?

$\log P_i$	$\log V_i$
-0.70	0.80
-0.50	0.60
-0.40	0.50
-0.30	0.30
-0.10	0.00
0.10	-0.10
0.30	-0.40
0.40	-0.50
0.50	-0.60
0.70	-0.80

$P(\chi_1^2 \geq x)$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
x	2.71	1.64	1.07	0.71	0.46	0.27	0.15	0.06	0.02

88406

For $x = (x_{ij} : i = 1, \ldots, n; j = 1, \ldots, m)$, set $\bar{x} = \frac{1}{mn}\sum_{i,j} x_{ij}$ and $\bar{x}_i = \frac{1}{m}\sum_j x_{ij}$, furthermore let

$$S_1(x) = \sum_{i,j}(x_{ij} - \bar{x}_i)^2, \quad S_2(x) = \sum_i (\bar{x}_i - \bar{x})^2.$$

Show that

$$S_1(x) + S_2(x) = \sum_{i,j}(x_{ij} - \bar{x})^2.$$

Consider n normal distributions with common unknown variance. A sample of size m is drawn from each distribution. It is proposed to test the hypothesis that the means of the distributions are all equal, using a procedure of rejecting the hypothesis if any two of the sample means differ by more than some predetermined value. Criticize briefly this proposal.

Describe a better test based on the same data in which the distribution of the test statistic when the means coincide is independent of the unknown variance. State what distribution the test statistic has when the means coincide.

In an experiment with $m = 9$ and $n = 3$, the following data are recorded:

i	1	2	3
$\bar{X}_i$	1.05	1.02	0.93
Σ_i	0.06	0.09	0.05

Here $X_i = \frac{1}{9}\sum_{j=1}^{9} X_{ij}$ and $\Sigma_i = \sum_{j=1}^{9}(X_{ij} - \bar{X}_i)^2$.

Would you reject the null hypothesis that the means coincide?

Distribution	χ^2_8	χ^2_9	t_8	$F_{2,8}$	$F_{2,24}$
95th percentile	15.5	16.9	1.86	4.46	3.40
99th percentile	20.1	21.7	2.90	8.65	5.61