June 22 Exam Solutions

1 Problems

Problem 1. Assume that observations X_1 and X_2 are jointly normally distributed with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \theta$, $\mathrm{Var}(X_1) = 1$ and $\mathrm{Var}(X_2) = 2$, where θ is the parameter of interest.

(i) Suppose that $\mathrm{Cov}(X_1, X_2) = 0$. Among all unbiased estimators, find one which achieves the minimum variance. Justify your answer.

(ii) Suppose that $\mathrm{Cov}(X_1, X_2) = 1$. Find an unbiased estimator with minimum variance. Justify your answer. Hint: k-variate normal density has the following form:

$$f(x) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-(x-\mu)^T \Sigma^{-1} (x-\mu)/2\right).$$

Solution 1

(i) The Fisher information based on the sample (X_1, X_2) is $3/2$ so that the minimum variance is $2/3$. We consider a linear estimator of the form $aX_1 + bX_2$ so that for it to be unbiased and have variance $2/3$ we must have $a + b = 1$ and $a^2 + 2b^2 = 2/3$. This has solution $a = 2/3$ and $b = 1/3$.

(ii) In this case, we find the Fisher information is 1 so that the minimum variance is 1. The unbiased estimator X_1 achieves this variance.

Problem 2. Let a positive random variable T have a hazard rate function of the form

$$\lambda(t) = \gamma t^{\gamma-1}, \gamma > 0.$$

(i) Find its probability density function f and its median. (Hint: recall that $\lambda(t) = f(t)/(1 - F(t))$ where F is the cumulative distribution function).

(ii) Find its mean for the cases of $\gamma = 1$ and $\gamma = 2$.

Solution 2

(i) Notice that $\gamma t^{\gamma-1} = \frac{\partial}{\partial t} t^\gamma$. On the other hand, we can also observe that $\lambda(t) = \frac{\partial}{\partial t} - \log(1 - F(t))$, where $F(t)$ is the cdf of T. Thus, we have $t^\gamma = -\log(1 - F(t))$ meaning $F(t) = 1 - \exp(-t^\gamma)$ and $f(t) = \exp(-t^\gamma) \cdot \gamma t^{\gamma-1}$. The median is then the solution to $1 - \exp(-t^\gamma) = 1/2$ which is $\log(2)^{1/\gamma}$.

(ii) The mean is the integral $\int_0^\infty \gamma \cdot t^{\gamma-1} \cdot e^{-t^\gamma} \cdot t \, dt$. For $\gamma = 1$, this reduces to the mean of an $\exp(1)$ random variable which is 1. For $\gamma = 2$, the integral becomes $\int_0^\infty 2 \cdot t^2 \cdot e^{-t^2} \, dt = \sqrt{\pi}/2$ using properties of the standard normal pdf.

Problem 3. Let Y be a standard d-dimensional normal random vector and let P be a $d \times d$ symmetric and idempotent matrix, i.e. $P = P^T$ and $P^2 = P$.

(i) Give all the possible values of the eigenvalues of P.

(ii) Show that $Y^T P Y \sim \chi^2(k)$, where $k = \mathrm{rank}(P)$.

Solution 3

(i) See the solution to Problem 20 in Review Doc 1.

(ii) This is the 2018 Summer Practice Exam Problem 10 (see Review Doc 4, Problem 1).

Problem 4. A researcher conducts a completely randomized experiment to test a drug, with n subjects $i = 1, \dots, n$, n_{0+} of which are assigned to a control group, n_{1+} to a treatment group. Let $Y_i = 0$ if i does not improve, $Y_i = 1$ if i improves. The researcher observes n_{00} subjects in the control group and n_{10} subjects in the treatment group with $Y = 0$, n_{01} subjects in the control group and n_{11} subjects in the treatment group with $Y = 1$. You may assume that both n_{0+} and n_{1+} are "large", that subjects in the control group are independent and identically distributed (iid) with $\mathbb{P}(Y = 1|X = 0) = \pi_0$, subjects in the treatment group are iid with $\mathbb{P}(Y = 1|X = 1) = \pi_1$. Let T_0 denote the number of successes in the control group, T_1 the number of successes in the treatment group.

(i) Write out the probability distribution $\mathbb{P}(T_0 = n_{01}, T_1 = n_{11})$

(ii) The researcher is interested in testing $H_0 : \pi_0 = \pi_1$ vs. $H_A : \pi_0 \neq \pi_1$. As a test statistic, she uses $\frac{p_1 - p_0}{\left(\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_0(1-p_0)}{n_{0+}}\right)^{1/2}}$, where $p_0 = n_{01}/n_{0+}$, $p_1 = n_{11}/n_{1+}$ and compares this to the normal distribution with mean 0 and variance 1. Justify the use of this statistic.

(iii) Construct the likelihood ratio test for H_0 vs. H_A above.

(iv) The researcher is also interested in estimating the relative risk RR $= \mathbb{P}(Y = 1|X = 1)/\mathbb{P}(Y = 1|X = 0)$. What is the maximum likelihood estimator of RR?

Solution 4

1. $\mathbb{P}(T_0 = n_{01}, T_1 = n_{11}) = \binom{n_{0+}}{n_{01}} \pi_0^{n_{01}} \cdot (1 - \pi_0)^{n_{00}} \cdot \binom{n_{1+}}{n_{11}} \pi_1^{n_{11}} \cdot (1 - \pi_1)^{n_{10}}$.

2. We claim the test statistic converges to $\mathcal{N}(0, 1)$ in distribution under the null hypothesis. This means the test is pretty good since n_{0+}, n_{1+} are both "large". Let $\pi = \pi_0 = \pi_1$. Note that we can first rewrite the statistic as

$$\frac{\sqrt{n_{1+}}(p_1 - \pi)}{(p_1(1 - p_1) + p_0(1 - p_0) \cdot (n_{1+}/n_{0+}))^{1/2}} + \frac{\sqrt{n_{0+}}(p_0 - \pi)}{(p_0(1 - p_0) + p_1(1 - p_1) \cdot (n_{0+}/n_{1+}))^{1/2}}.$$

To make life easier, let's first assume $n_{1+}/n_{0+} \overset{n_{1+}, n_{0+} \to \infty}{\longrightarrow} c \in \mathbb{R}_{>0}$. Then, by the continuous mapping theorem, the denominator of the first term goes to $\sqrt{(\pi(1 - \pi) \cdot (1 + c)}$ so that by CLT and Slutsky, the first term in the sum above goes to $\mathcal{N}(0, (1 + c)^{-1})$. Similarly, the second term goes to $\mathcal{N}(0, c(1 + c)^{-1})$. Thus, the sum goes to $\mathcal{N}(0, 1)$ in distribution.

In fact, essentially the same argument would have worked if $\lim_{n_{1+}, n_{0+} \to \infty} n_{1+}/n_{0+} = 0$ or $+\infty$ since then the first term in the sum above goes to $\mathcal{N}(0, 1)$ while the second term goes to 0 (or vice-versa).

In fact, this will be enough to show the desired convergence result for general sequences $\{(n_{0+}, n_{1+})\}$. In particular, a subsequence argument will allow us to remove the assumption that n_{1+}/n_{0+} converges or diverges. To do so, we need to consider an alternative characterization of convergence in distribution in terms of subsequences.

Claim: $X_n \xrightarrow{d} X$ if every subsequence $\{X_{n_k}\}_k \subseteq \{X_n\}_n$ has a further subsequence $\{X_{n_m}\}_m \subseteq \{X_{n_k}\}_k$ such that $X_{n_m} \xrightarrow{d} X$ as $m \to \infty$. This claim can be proven using the Portmanteau theorem.

Next, recall from real analysis that any sequence of real numbers has a monotone subsequence. Thus, for any sequence of values n_{1+}/n_{0+}, we can find a monotonic subsequence in which case $n_{1+}/n_{0+} \to c \geq 0$ or $n_{1+}/n_2 \to +\infty$ since the sequence is positive. Combining this fact with our claim aboves gives us the desired convergence result for all sequences $\{(n_{1+}, n_{0+})\}$.

3. The MLE under H_0 for $\pi = \pi_0 = \pi_1$ is $\frac{n_{01}+n_{11}}{n_{0+}+n_{1+}}$ since all the data is iid. The unconstrained MLE for (π_0, π_1) is (p_0, p_1) where p_0, p_1 are defined as in (ii). Plugging these into the likelihood from (i) and taking the ratio gives us the likelihood ratio test statistic.

4. By functional invariance of the MLE, the MLE for RR $= \pi_1/\pi_0$ must be p_1/p_0.

Problem 5. Suppose that observations Y_i, $i = 1, 2, 3$, follow independent Poisson distributions with parameters λ_i.

(i) What is the distribution of $Y_1 + Y_2 + Y_3$? Justify.

(ii) The researcher believes $\lambda_i = i \times \lambda$ for all $i = 1, 2, 3$. How might you test $H_0 : \lambda_i = i \times \lambda$ for all i vs. $H_A : \lambda_i \neq i \times \lambda$ for some i?

(iii) Presuming the researcher's belief is correct, obtain the maximum likelihood estimator of λ.

(iv) Again, presuming the researcher's belief is correct, obtain the likelihood ratio test for $H_0 : \lambda = \lambda_0$ vs. $H_A : \lambda \neq \lambda_0$.

Solution 5

(i) The mgf of $Y_1 + Y_2 + Y_3$, by independence, is the product of the mgf's of Y_1, Y_2, Y_3. This is

$$\exp\left((\lambda_1 + \lambda_2 + \lambda_3)(e^t - 1)\right),$$

which is the mgf of a Poisson distribution with parameter $\lambda_1 + \lambda_2 + \lambda_3$.

(ii) We can use the likelihood ratio test. The joint likelihood under H_0 is

$$L(\lambda|Y_1, Y_2, Y_3) = \frac{\lambda^{Y_1+Y_2+Y_3} e^{-6\lambda} 2^{Y_2} 3^{Y_3}}{Y_1! Y_2! Y_3!}.$$

So that taking $\log$ and derivatives gives us the MLE for λ under H_0 is $\frac{Y_1+Y_2+Y_3}{6}$. Meanwhile, the unconstrained MLE for $(\lambda_1, \lambda_2, \lambda_3)$ is just (Y_1, Y_2, Y_3). Plugging these into the likelihood above give us the likelihood ratio test statistic.

(iii) See (ii).

(iv) Here, the unconstrained MLE for λ is $\lambda_{\text{MLE}} = \frac{Y_1+Y_2+Y_3}{6}$, so that $L(\lambda_0|Y_1, Y_2, Y_3)/L(\lambda_{\text{MLE}}|Y_1, Y_2, Y_3)$ is the likelihood ratio test statistic.

Problem 6. Suppose that $\hat{\theta}$ has the p-dimensional normal distribution with expectation the vector all of whose components are zero and positive definite variance-covariance matrix Σ. Let $c(\hat{\theta})$ be the value of c that maximizes

$$t = \frac{c^T \hat{\theta}}{\sqrt{c^T \Sigma c}}.$$

Find the distribution of

$$t_{\max} = \frac{c(\hat{\theta})^T \hat{\theta}}{\sqrt{c(\hat{\theta})^T \Sigma c(\hat{\theta})}}.$$

Solution 6

First, we note that objective function t, is invariant under positive scalar multiplication of c. Thus, we can assume WLOG that $c^T \Sigma c = 1$. Then, $c(\hat{\theta})$ just maximizes $c^T \hat{\theta}$ which means c is the vector on the ellipse $\{c : c^T \Sigma c = 1\}$ parallel to $\hat{\theta}$. Let $c = \alpha \cdot \hat{\theta}$. Then, since $c^T \Sigma c = \alpha^2 \hat{\theta}^T \Sigma \hat{\theta} = 1$, we must have $\alpha = \sqrt{\frac{1}{\hat{\theta}^T \Sigma \hat{\theta}}}$. Thus,

$$t_{\max} = \alpha \hat{\theta}^T \hat{\theta} = \frac{\hat{\theta}^T \hat{\theta}}{\sqrt{\hat{\theta}^T \Sigma \hat{\theta}}}.$$

Letting $\hat{\theta} = \Sigma^{1/2} \mathbf{z}$ where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_p, \mathrm{Id})$, we have

$$t_{\max} = \frac{\mathbf{z}^T \Sigma \mathbf{z}}{\sqrt{\mathbf{z}^T \Sigma^2 \mathbf{z}}}.$$

Problem 7. Suppose that Y_{i1}, Y_{i2}, independent pairs of Bernoulli observations with expectations, respectively,

$$\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \text{ and } \frac{e^{\alpha_i}}{1+e^{\alpha_i}},$$

i from 1 to n, where α_i and β are unspecified parameters. Let $\hat{\beta}$ be the maximum likelihood estimate of β. What is the limiting expectation of $\hat{\beta}$ when the true value of β is 0.5 and the true values of α_i are all equal to 0? Hint: for each i, think about the four possibilities for (Y_{i1}, Y_{i2}) separately.

Solution 7

The log-likelihood is (after simplifying some terms):

$$L(\beta, \{\alpha_i\}_i | \{Y_{i1}, Y_{i2}\}_i) = \sum_{i=1}^{n} Y_{i1} \cdot (\alpha_i + \beta) + Y_{i2} \cdot \alpha_i - \log(1 + e^{\alpha_i+\beta}) - \log(1 + e^{\alpha_i}). \tag{1}$$

We first fix a value of β and consider maximizing with respect to α_i for a fixed i. There are four cases depending on the values of Y_{i1}, Y_{i2}.

(a) First, suppose that $Y_{i1} = Y_{i2} = 0$. Then, the summand in (1) simplifies to:

$$\log\left(\frac{1}{(1+e^{\alpha_i+\beta})(1+e^{\alpha_i})}\right),$$

which, for any β, is increasing with decreasing α_i. Thus, the maximizer of the above for any fixed β, is $\alpha_i = -\infty$. We note the maximum value of the i-th summand in (1) in this case is $\log(1) = 0$.

(b) Next, suppose $Y_{i1} = Y_{i1} = 1$. Then, the part of the log-likelihood depending on α_i is

$$2\alpha_i + \beta - \log(1 + e^{\alpha_i+\beta}) - \log(1 + e^{\alpha_i}). \tag{2}$$

First, we note that the derivative of the above with respect to α_i is

$$2 - \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} - \frac{e^{\alpha_i}}{1+e^{\alpha_i}},$$

which is positive for all $\alpha_i, \beta \in \mathbb{R}$. Thus, the log-likelihood is increasing in α_i in this case which means the maximizer with respect to α_i is $\alpha_i = +\infty$. Next, we compute the limit of (2) as $\alpha_i \to +\infty$. We rewrite (2) as

$$\log\left(\frac{e^{2\alpha_i+\beta}}{(1+e^{\alpha_i})(1+e^{\alpha_i+\beta})}\right),$$

so that taking $\alpha_i \to +\infty$ the ratio above goes to 1 which means the log-likelihood goes to 0.

(c) Suppose $Y_{i1} = 1$ and $Y_{i2} = 0$. In this case, our likelihood is

$$\alpha_i + \beta - \log(1 + e^{\alpha_i+\beta}) - \log(1 + e^{\alpha_i}).$$

For this case, we can proceed as usual and take the derivative with respect to α_i and set it equal to 0. Omitting the details of this, we will find the maximizer of the above is at $\alpha_i = -\beta/2$.

(d) Consider the case where $Y_{i1} = 0$ and $Y_{i2} = 1$, for which the likelihood is

$$\alpha_i - \log(1 + e^{\alpha_i+\beta}) - \log(1 + e^{\alpha_i}).$$

This has the same derivative with respect to α_i as in the last case, so again our maximizer is $\alpha_i = -\beta/2$.

Putting all the cases together, letting $\hat{\alpha}_i$ be the maximizer with respect to α_i (whose value we allow to possibly be $\pm\infty$), we have:

$$L(\beta, \{\hat{\alpha}_i\}_i) = \sum_{i=1}^{n} \mathbf{1}\{Y_{i1} = 1, Y_{i2} = 0\} \log\left(\frac{e^{\beta/2}}{(1+e^{\beta/2})(1+e^{-\beta/2})}\right) + \mathbf{1}\{Y_{i1} = 0, Y_{i2} = 1\} \log\left(\frac{e^{-\beta/2}}{(1+e^{\beta/2})(1+e^{-\beta/2})}\right).$$

Let $N_1 := \sum_{i=1}^n \mathbf{1}\{Y_{i1} = 1, Y_{i2} = 0\}$ and $N_2 := \sum_{i=1}^n \mathbf{1}\{Y_{i1} = 0, Y_{i2} = 1\}$. Then, we have the above is

$$\frac{\beta}{2}(N_1 - N_2) - n\left(\log(1 + e^{\beta/2}) + \log(1 + e^{-\beta/2})\right).$$

Taking derivative of the above with respect to β and setting equal to 0, a quadratic with respect to $e^{\beta/2}$ will appear which we can solve for β getting (detailed omitted):

$$\beta_{\textsf{MLE}} = 2\log\left(\frac{n}{N_2 - N_1 + n}\right).$$

Since $N_1/n \xrightarrow{\mathsf{P}} \mathbb{P}(Y_{i1} = 1, Y_{i2} = 0) = \frac{e^{1/2}}{2(1+e^{1/2})}$ and $N_2/2 \xrightarrow{\mathsf{P}} \mathbb{P}(Y_{i1} = 0, Y_{i2} = 1) = \frac{1}{2(1+e^{1/2})}$ by LLN, $\beta_{\textsf{MLE}}$ goes to a constant in distribution (whose value can be determined by plugging these limits into the above).

Problem 8. Suppose we have a normal random sample $X_1, \ldots, X_n$ such that $\mathbb{E}[X_i] = \mu$, and that we conduct a t-test with level α of the hypothesis $H_0 : \mu = \mu_0$. Show that if the t-test achieves an asymptotic level α when we have n independent $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$ if $0 < m \le \sigma_i \le M < \infty$ for all $i = 1, \ldots, n$, where $m < M$ are positive constants that do not depend on n.

Solution 8

It suffices to show the Student's t-statistic converges in distribution to $\mathcal{N}(0,1)$. WLOG, we may assume $\mu_0 = 0$ since we can recenter each X_i by $X_i - \mu_0$. So, under the null hypothesis, our test statistic is distributed as

$$\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Z_i \cdot \sigma_i\right)}{\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^n Z_i \cdot \sigma_i^2 - \left(\frac{1}{n}\sum_{i=1}^n Z_i \cdot \sigma_i\right)^2 \cdot n\right)}},$$

where $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Note that I wrote the sample variance in the denominator using the formula $S^2 = \sum_i x_i^2 - n \cdot \overline{x}_n^2$, which will serve more convenient later when taking limits.

We first observe that:

$$\frac{1}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \sum_{i=1}^n Z_i \cdot \sigma_i \sim \mathcal{N}(0,1), \tag{3}$$

since $\sum_{i=1}^n Z_i \cdot \sigma_i \sim \mathcal{N}\left(0, \sum_{i=1}^n \sigma_i^2\right)$. Then, it suffices to show the sample variance behaves like $\frac{1}{n}\sum_{i=1}^n \sigma_i^2$ in the limit. In particular, by Slutsky, it suffices to show that:

$$\frac{\sum_{i=1}^n Z_i^2 \cdot \sigma_i^2 - \left(\frac{1}{n}\sum_{i=1}^n Z_i \cdot \sigma_i\right)^2 \cdot n}{\sum_{i=1}^n \sigma_i^2} \xrightarrow{\mathsf{d}} 1.$$

(3) already tells us that the second term in the difference above goes to 0:

$$\frac{\left(\frac{1}{n}\sum_{i=1}^n Z_i \cdot \sigma_i\right)^2 \cdot n}{\sum_{i=1}^n \sigma_i^2} = \left(\frac{1}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \sum_{i=1}^n Z_i \cdot \sigma_i\right)^2 \cdot \frac{1}{n} \xrightarrow{\mathsf{d}} 0,$$

by Slutsky. Thus, it suffices to show

$$\frac{\sum_{i=1}^n Z_i^2 \cdot \sigma_i^2}{\sum_{i=1}^n \sigma_i^2} \xrightarrow{\mathsf{d}} 1.$$

This will essentially follow from the law of large numbers for triangular arrays. I'll write out the key steps: first observe by Chebyshev that for any $\epsilon > 0$:

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n Z_i^2 \sigma_i^2}{\sum_{i=1}^n \sigma_i^2} - 1\right| \ge \epsilon\right) \le \frac{\mathrm{Var}\left(\sum_{i=1}^n Z_i^2 \cdot \sigma_i^2\right)}{\epsilon^2 \cdot \left(\sum_{i=1}^n \sigma_i^2\right)^2} \propto \frac{\sum_{i=1}^n \sigma_i^4}{\epsilon \cdot \left(\sum_{i=1}^n \sigma_i^2\right)^2}.$$

Then, it suffices to show this last RHS goes to 0 as $n \to \infty$. Alternatively, we can show the reciprocal of the RHS above goes to ∞. We have

$$\frac{\left(\sum_{i=1}^n \sigma_i^2\right)^2}{\sum_{i=1}^n \sigma_i^4} = 1 + \frac{\sum_{i \ne j} \sigma_i^2 \sigma_j^2}{\sum_{i=1}^n \sigma_i^2} \le 1 + \frac{n(n-1) \cdot M^4}{n \cdot m^4} \xrightarrow{n \to \infty} \infty.$$