PhD Qualifying Examination–Part I

Department of Statistics
University of South Carolina
August 10, 2020 - 9:00AM–1:00PM

READ FIRST THESE INSTRUCTIONS

1. DO NOT write your name on any of your answer sheets. Instead, write your pre-assigned codename.
2. There are four (4) problems on this examination.
3. Formulas relating to distributions potentially relevant to the problems are provide in the last page.
4. You are **not allowed** to use search engines during the examination. Please adhere to the HONOR CODE in this instance. Any violation of the HONOR CODE (such as using search engines) will lead to a zero for the exam.
5. You have four hours for this examination. All four problems will be graded and are of equal weight.

The Problems

1. In medical diagnosis, the *sensitivity* of a test, for a particular virus for example, is the probability of an individual testing positive given that he/she carries the virus; the *specificity* is the probability of an individual testing negative given that he/she does not carry the virus.

   Consider the population of South Carolina (SC), and a particular test for COVID-19 that has a specificity of p_0. Denote by $d \in (0, 1)$ the COVID-19 prevalence of SC. A COVID-19 carrier is either symptomatic (i.e., showing some disease symptoms) or asymptomatic (i.e., showing no symptoms). The accuracy of this test depends on whether or not the testing subject is symptomatic. In particular, the test returns a positive result with probability p_1 when the testing subject is a symptomatic carrier; whereas the test returns a positive result with probability p_2 when the testing subject is an asymptomatic carrier. Among the COVID-19 carriers in SC, $a \times 100\%$ of them are asymptomatic, where $a \in (0, 1)$.

   (a) Derive the sensitivity of the test.

   (b) Three randomly selected individuals in SC take the test. Assume that these three individuals are mutually independent in regard to the disease status (including being symptomatic or not) and also in terms of the test result. If all three tests return negative, what is the probability that at least one of the three individuals is a carrier?

   (c) Three randomly selected individuals in SC take the test. Assume that these three individuals are mutually independent in regard to the disease status (including being symptomatic or not) and also in terms of the test result. Given that at least one of the three tests return positive, what is the probability that at least one of the three individuals is an asymptomatic carrier?

2. Consider a simple linear regression model in (1), where the response variable Y and the scalar predictor X are both centered so that they each has a mean of zero,

$$Y = \beta X + \epsilon;\ \epsilon \sim N(0, \sigma^2), \tag{1}$$

where β is the regression coefficient and σ^2 is the error variance, both unknown. A random sample of size n is collected, $\{(x_i, y_i),\ \text{for } i = 1, \ldots, n\}$. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{x} = (x_1, \ldots, x_n)'$.

(a) Find the least squares estimator for β, denoted by $\hat{\beta}$. Construct a $100(1-\alpha)\%$ confidence interval for β based on the least squares estimator.

(b) To infer the unknown parameters in (β, σ^2) under the regression setting, one may view $\mathbf{y}$ as **data** and $\mathbf{x}$ as fixed constants, e.g., fixed design points. Derive the likelihood function for inferring (β, σ^2). Show that

$$\sum_{i=1}^{n}(y_i - \beta x_i)^2 = \sum_{i=1}^{n} e_i^2 + (\beta - \hat{\beta})^2 \sum_{i=1}^{n} x_i^2,$$

where $e_i = y_i - \hat{\beta} x_i$.

(c) Instead of using the least squares method, we now want to incorporate prior information when inferring β and σ^2. More specifically, assume that the prior distribution of β is $N(\beta_0,\, \sigma_0^2)$, and the prior distribution of σ^2 is an inverse gamma (IG), IG$(a,\, \lambda)$, of which the probability density function (pdf) is given by

$$f(\sigma^2; a,\, \lambda) = \frac{\lambda^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left(-\frac{\lambda}{\sigma^2}\right),$$

in which $a > 0$, $\lambda > 0$, and $\Gamma(a) = \int_0^{+\infty} t^{a-1} e^{-t}\, dt$ is the gamma function.

i. Derive the *conditional* posterior distribution of β, $p(\beta|\sigma^2, \mathbf{y})$.

ii. Derive the *conditional* posterior distribution of σ^2, $p(\sigma^2|\beta, \mathbf{y})$.

(d) When prior information are unavailable, noninformative priors are used to carry out Bayesian inference. A noninformative prior may not be a valid distribution, even though one can still draw sensible Bayesian inference when such a prior is used. Suppose that the noninformative priors for β and σ^2 are specified by the following pdfs,

$$p(\beta) = 1, \quad \beta \in (-\infty, +\infty),$$
$$p(\sigma^2) = \frac{1}{\sigma^2}, \quad \sigma^2 \in (0, +\infty).$$

i. Derive the *conditional* posterior distribution of β, $p(\beta|\sigma^2, \mathbf{y})$.

ii. Derive the *conditional* posterior distribution of σ^2, $p(\sigma^2|\beta, \mathbf{y})$.

(e) Under the noninformative prior specification in part (d), derive the *marginal* posterior distribution of β, $p(\beta|\mathbf{y})$. Based on this marginal posterior distribution, propose a point estimator for β; also suggest a way to construct an interval estimator for β.

3. Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) according to $N(\mu_1, 1)$ and $Y_1, \ldots, Y_n$ be iid according to $N(\mu_2, 1)$. The two samples are mutually independent. Define $\Theta_0 = \{(\mu_1, \mu_2) : \max(\mu_1, \mu_2) \leq 0\}$ and let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$. We want to test $H_0 : (\mu_1, \mu_2) \in \Theta_0$ versus $H_1 :$ not H_0 at significance level α where $0 < \alpha < 0.25$.

   (a) Suppose we use $T_n = \sqrt{n} \min\{\bar{X}_n I(\bar{X}_n \geq 0), \bar{Y}_n I(\bar{Y}_n \geq 0)\}$ as a test statistic and reject H_0 if $T_n > t_\alpha$ at significance level α. Find the value t_α such that resulting test is of size α; i.e.,

$$\sup_{(\mu_1, \mu_2) \in \Theta_0} P(T_n > t_\alpha) \leq \alpha.$$

   (b) We now develop a likelihood ratio test for this purpose.

      (i) Let the likelihood ratio test statistic be λ_n. Show that

$$-2 \log \lambda_n = n\bar{X}_n^2 I(\bar{X}_n \geq 0) + n\bar{Y}_n^2 I(\bar{Y}_n \geq 0).$$

      (ii) For $\alpha < 0.25$, let q_α be the $1 - \alpha$th quantile of the distribution of $Z_1^2 I(Z_1 \geq 0) + Z_2^2 I(Z_2 \geq 0)$ where Z_1 and Z_2 are independent standard normal random variables; i.e.,

$$P\{Z_1^2 I(Z_1 \geq 0) + Z_2^2 I(Z_2 \geq 0) > q_\alpha\} = \alpha.$$

      Prove that

$$\sup_{(\mu_1, \mu_2) \in \Theta_0} P(-2 \log \lambda_n > q_\alpha) \leq \alpha;$$

      that is, we reject H_0 at significance level α if $-2 \log \lambda_n > q_\alpha$.

4. The diameter of a certain particle of interest follows an exponential distribution with a mean of 1 unit. The particles pass through a sieve system, and the diameters of n particles coming out of the sieve system are measured. When a particle is so big such that its diameter exceeds θ units, it will not pass through the system. Consequently, the observed data, denoted by $(X_1, \ldots, X_n)$, are a random sample from a distribution supported on $(0, \theta)$.

   (a) Show that the observed data form a random sample from a distribution whose probability density function (pdf) is given by

   $$f(x) = \frac{e^{-x}}{1 - e^{-\theta}}, \text{ for } 0 < x < \theta. \tag{2}$$

   (b) Derive the maximum likelihood estimator (MLE) of θ based on the observed data $(X_1, \ldots, X_n)$. Show that this MLE underestimates θ.

   (c) Provide the uniformly most powerful (UMP) level-α test for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is a positive constant.

   (d) Derive the Θ'-UMA (uniformly most accurate) $(1 - \alpha)$ confidence set of θ, where $\Theta' = \{\theta' > 0 : \theta' < \theta\}$.

   (e) The distribution specified by the pdf in (2) is known as a *truncated exponential* distribution, not to confused with a *shifted exponential* distribution. Now return to the regular exponential distributions (by setting $\theta = +\infty$ in (2) for instance). Suppose $X \sim$ exponential(1). Consider a sequence of independent random variables, $Y_1, Y_2, \ldots$, that are also independent of X, where $Y_n \sim$ exponential$(1 + n^{-1})$, for $n = 1, 2, \ldots$. As $n \to \infty$, does Y_n converge to X in probability? Does it converge to X in distribution? Does it converge to X almost surely? Explain.

Formulas relating to some distributions

- Exponential(β)

  pdf: $f(x|\beta) = \beta^{-1}e^{-x/\beta}$, $0 \le x < \infty$, $\beta > 0$

  mgf: $M(t) = (1-\beta t)^{-1}$, for $t < 1/\beta$

moments: $E(X) = \beta$, $\mathrm{Var}(X) = \beta^2$

- Gamma(α, β)

  pdf: $f(x|\alpha,\beta) = \dfrac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$, $0 \le x < \infty$, $\alpha, \beta > 0$

  mgf: $M(t) = (1-\beta t)^{-\alpha}$, for $t < 1/\beta$

moments: $E(X) = \alpha\beta$, $\mathrm{Var}(X) = \alpha\beta^2$

notes: Gamma(1, β) is exponential(β). Gamma($p/2$, 2) is χ^2_p

- Normal(μ, σ^2)

  pdf: $f(x|\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$, $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$

  mgf: $M(t) = e^{\mu t + \sigma^2 t^2/2}$

- t distribution with ν degrees of freedom.

  pdf: $f(x|\nu) = \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\dfrac{1}{\sqrt{\nu\pi}}\left(1+x^2/\nu\right)^{-(\nu+1)/2}$, $-\infty < x < \infty$, $\nu = 1, 2, \ldots$

moments: $E(X) = 0$, for $\nu > 1$; $\mathrm{Var}(X) = \nu/(\nu-2)$, for $\nu > 2$