Statistics Further Examples Sheet

This examples sheet has some extra questions which you may like to do for further practice or revision. A copy of this sheet can be found at: `http://www.statslab.cam.ac.uk/~rrw1/stats/`

1. (Sufficiency) Let $X_1, \ldots, X_n$ be independent Poisson random variables with X_j having parameter $j\lambda$, where $\lambda > 0$ is an unknown parameter. Find a sufficient statistic for λ. What is its distribution?

2. (Sufficiency, MLE, confidence intervals) Suppose that $X_1, \ldots, X_n$ are independent random variables, $X_j \sim N(\theta, \sigma_j^2)$, where θ is an unknown parameter, but the σ_j^2, $j = 1, \ldots, n$ are known. Find a sufficient statistic for θ. What is its distribution? Find the MLE of θ, and find a 95% confidence interval for θ.

3. (Sufficiency, MLE, confidence intervals) Suppose that $n > 2$, and

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

Find a sufficient statistic $T_n = T_n(X_1, \ldots, X_n)$ for λ, and write down its density. Obtain the maximum-likelihood estimator for λ and show that it is biased, but that some multiple of it is not. Find the distribution of $U_n = 2\lambda \sum_{i=1}^n X_i$ and hence obtain a 95% confidence interval for λ.

4. (Rao-Blackwellization) The random variables $X_1, \ldots, X_n$ are IID Poisson with mean μ. Write down a simple unbiased estimator of $\theta = P(X_1 = 0)$. Is it a function of the sufficient statistic for μ? If not, find a better estimate by conditioning on the sufficient statistic.

5. (Bayes estimation) Show that the posterior mean can never be an unbiased estimator of θ.

Hint: Show that $\mathbb{E}_\theta[E_X(X-\theta)^2]$ *and* $\mathbb{E}_X[E_\theta(X-\theta)^2]$ *are respectively* $\mathbb{E}\theta^2 - \mathbb{E}X^2$ *and* $\mathbb{E}X^2 - \mathbb{E}\theta^2$ *and hence that* $\mathbb{E}(X-\theta)^2 = 0$.

6. (Lecture 9. Chi-squared tests of categorical data) Show that in a 2×2 contingency table, with observed frequencies Y_{ij} $(i, j = 1, 2)$, the usual expression for the statistic X^2 may be put in the form

$$Y_{++}(Y_{11}Y_{22} - Y_{12}Y_{21})^2 / (Y_{1+}Y_{2+}Y_{+1}Y_{+2})$$

where $+$ denotes summation over both values of the corresponding subscript.

7. (regression) Suppose that for given $x_1, \ldots, x_n$,

$$Y_i \sim N(\beta x_i, x_i \sigma^2), \quad i = 1, \ldots, n,$$

independently. Show that this is equivalent to the model

$$\frac{Y_i}{\sqrt{x_i}} = \beta\sqrt{x_i} + \epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_n$ are IID $N(0, \sigma^2)$. Show that both the MLE and LSE of β is $\hat\beta = \bar{Y}/\bar{x}$.

A nationwide chain of 500 retail shops has just started its annual sale. Management would like to estimate, by 9pm on the first day of the sale, the total sales there have been that first day. It is difficult to collect data from all shops by 9pm, but this can be done for 50 of the shops. The shops differ in size: shop i is of size x_i; first day sales in shop i are y_i. The total size of shops in the chain, $\sum_{i=1}^{500} x_i$, is known, as are the pairs (x_i, y_i), for the sample $i = 1, \ldots, 50$.

Explain, using the ideas in the regression model at the start of the question, how you could estimate the total sales. What assumption are you making about the relationship between sales and shop size? Is this reasonable? What assumption are you making about the sample of 50 shops? Supposing that all these assumptions are valid, show that your estimator of $\sum_{i=1}^{500} y_i$ is unbiased.

8. (regression) Suppose that n points are to be chosen in the interval $[-1, 1]$ for estimating a and β in the regression model

$$y_i = a + \beta x_i + \epsilon_i,$$

where the $\epsilon_1, \ldots, \epsilon_n$ are IID $N(0, \sigma^2)$. What values should be given to $x_1, \ldots, x_n$ so as to minimize $\mathrm{var}(\hat\beta)$?

9. (Lecture 4. confidence intervals) Suppose X_1, X_2 two IID samples from a uniform distribution on $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. Show that $(\min_i x_i, \max_i x_i)$ is a 50% confidence interval for θ. Suppose you observe the data $x = (4.0, 3.4)$. How certain are you that θ lies in the interval $(3.4, 4.0)$?

Hint: $\mathbb{P}(\min_i x_i \leq \theta \leq \max_i x_i) = \mathbb{P}(X_1 \leq \theta \leq X_2) + \mathbb{P}(X_2 \leq \theta \leq X_1)$. *Why?*

10. (contingency table)

11. (theory of multivariate normal) Define the *multivariate normal* distribution as follows. If $\mathbf{y}$ is a p-vector with p-dimensional pdf:

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp -\frac{1}{2}\{(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\}, \mathbf{y} \in \mathbb{R}^p,$$

where μ is a p-vector and Σ a symmetric, positive definite $p \times p$ matrix, then $\mathbf{Y}$ is said to have a multivariate normal distribution, and we write

$$\mathbf{Y} \sim \mathrm{N}_p(\mu, \Sigma).$$

Prove the following:

(i) $Y_1, \ldots, Y_p$ are independent if and only if Σ is diagonal. Also show that when $\Sigma = kI$, $\mu = \mu\mathbf{1}$ then $Y_1, \ldots, Y_p$ are IID.

(ii) If $\mathbf{Y} \sim \mathrm{N}_p(\mu, \Sigma)$, then $\mathbf{Z} = A\mathbf{Y} \sim \mathrm{N}_p(A\mu, A\Sigma A^T)$. (Note the special case: Lemma 3.4.1 from notes.)

(iii) Show that $\Sigma^{-\frac{1}{2}}(\mathbf{Y}-\mu) \sim \mathrm{N}_p(0, I)$.

(iv) From (iii), deduce that $\mathbb{E}[\mathbf{Y}] = \mu$, and that the ij^{th} entry of Σ, σ_{ij}, is $\mathrm{cov}(Y_i, Y_j)$.

12. (theory of bivariate normal) Suppose (X_i, Y_i) $1 \le i \le n$ are IID from a bivariate normal population with mean $\mu = (\mu_X, \mu_Y)^T$, and $\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$.

(a) Calculate the conditional distribution of Y_i given X_i.

(b) Suppose we attempt to fit the linear model

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i.$$

From (a), argue that the hypotheses $H_0 : \beta = 0$ and $H_0 : \rho = 0$ are equivalent.

(c) Letting $S^2 = (n-2)^{-1}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x}))^2$, and $S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2$, show that

$$\frac{\sqrt{S_{XX}}\hat{\beta}}{S} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}.$$

where r is the sample covariance, $\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})/(\Sigma(y_i - \bar{y})^2\Sigma(y_i - \bar{x})^2)^{\frac{1}{2}}$. Therefore construct a test of $H_0 : \beta = 0$ depending only on r^2 and n.

13. (Strong Likelihood Principle) On the Planet Altair A, all nuclear power stations were installed exactly 20 years ago by ANFL. It is known that cataclysmic accidents occur at a Poisson process of constant (but unknown) rate λ per year. The first cataclysmic accident has just happened. ANFL argue that it is not fair to apply statistics just after an accident has occurred. However GreenAlt, the local conservation movement, argues as follows:

1°. The likelihood of getting precisely 1 accident (some time) in a 20-year period is

$$(20\lambda)e^{-20\lambda}$$

by the formula for the Poisson distribution.

2°. The likelihood that the first accident occurs at time 20 is

$$\lambda e^{-20\lambda}$$

by the formula for the exponential distribution.

3°. Since these likelihoods are proportional and since the constant of proportionality, 20, will drop out of every statistical consideration, any inferences about λ made on the basis that the first accident has just occurred must be the same as those based on the fact that precisely 1 accident has occurred in all in 20 years.

Most of the people-in-the-street on Altair A believe that there is substance in ANFL's argument and that GreenAlt are trying to overturn common sense with silly mathematics. What do you think? The Problem is not unknown on Planet Earth; and GreenAlt's argument, if valid, remains valid if 1 accident is replaced by 3 or 4.

R Weber, January 22, 2008