Inference Tutorial 1

This week's questions cover background material, and the frequentist and Bayesian approaches to statistical inference.

1. `https://people.maths.bris.ac.uk/~sw15190/TOI/` (also linked from Blackboard) provides some background reading for the course.

   (a) Review Chapter 1 of *Core Statistics*, which revises random variables and probability.

   (b) Review `matrix.pdf` which covers essential matrix algebra needed for the course.

2. This question covers a background result that will be used several times in the course.

   (a) If $\mathbf{Y}$ and $\mathbf{X}$ are random vectors such that $\mathbf{Y} = \mathbf{D}\mathbf{X}$ where $\mathbf{D}$ is a matrix of fixed coefficients, show that if $\mathbf{V}_x$ and $\mathbf{V}_y$ are the covariance matrices for $\mathbf{X}$ and $\mathbf{Y}$ respectively then

   $$\mathbf{V}_y = \mathbf{D}\mathbf{V}_x\mathbf{D}^T.$$

   (Recall that if $\boldsymbol{\mu}_y \equiv E(\mathbf{Y})$, $\mathbf{V}_y = E[(\mathbf{Y} - \boldsymbol{\mu}_y)(\mathbf{Y} - \boldsymbol{\mu}_y)^T]$.)

   (b) Consider a multivariate normal random vector $\mathbf{X} \sim N(\boldsymbol{\mu}_x, \mathbf{V}_x)$, and suppose that the covariance matrix can be decomposed $\mathbf{V}_x = \mathbf{C}\mathbf{C}^T$ (this can always be done for a full rank covariance matrix using e.g. a Choleski decomposition). Show that $\mathbf{V}_x^{-1} = \mathbf{C}^{-T}\mathbf{C}^{-1}$ and that $\mathbf{Y} = \mathbf{C}^{-1}(\mathbf{X} - \boldsymbol{\mu}_x) \sim N(\mathbf{0}, \mathbf{I})$.

   (c) Assuming that $\mathbf{X} \sim N(\boldsymbol{\mu}_x, \mathbf{V}_x)$, show that

   $$(\mathbf{X} - \boldsymbol{\mu}_x)^T\mathbf{V}_x^{-1}(\mathbf{X} - \boldsymbol{\mu}_x) = \mathbf{Y}^T\mathbf{Y} \quad \text{where} \quad \mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$$

   (d) If Z_i are i.i.d. $N(0, 1)$ random variables then

   $$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2.$$

   What is the distribution of

   $$(\mathbf{X} - \boldsymbol{\mu}_x)^T\mathbf{V}_x^{-1}(\mathbf{X} - \boldsymbol{\mu}_x)$$

   if $\mathbf{X} \sim N(\boldsymbol{\mu}_x, \mathbf{V}_x)$?

   Solution

   (a)

   $$\begin{aligned}
   \mathbf{V}_y &= E[(\mathbf{Y} - \boldsymbol{\mu}_y)(\mathbf{Y} - \boldsymbol{\mu}_y)^T] \\
   &= E[(\mathbf{D}\mathbf{X} - \mathbf{D}\boldsymbol{\mu}_x)(\mathbf{D}\mathbf{X} - \mathbf{D}\boldsymbol{\mu}_x)^T] \\
   &= \mathbf{D}E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^T]\mathbf{D}^T \\
   &= \mathbf{D}\mathbf{V}_x\mathbf{D}^T
   \end{aligned}$$

   (b) $\mathbf{Y} = \mathbf{C}^{-1}(\mathbf{X} - \boldsymbol{\mu}_x) \Rightarrow E(\mathbf{Y}) = \mathbf{C}^{-1}[E(\mathbf{X}) - \boldsymbol{\mu}_x] = \mathbf{0}$. By the result from (a)

   $$\mathbf{V}_y = \mathbf{C}^{-1}\mathbf{C}\mathbf{C}^T\mathbf{C}^{-T} = \mathbf{I}.$$

   Each Y_i is a weighted linear sum of normal r.v.s and is hence a normal r.v. so $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$.

   (c)

   $$(\mathbf{X} - \boldsymbol{\mu}_x)^T\mathbf{V}_x^{-1}(\mathbf{X} - \boldsymbol{\mu}_x) = (\mathbf{X} - \boldsymbol{\mu}_x)^T\mathbf{C}^{-T}\mathbf{C}^{-1}(\mathbf{X} - \boldsymbol{\mu}_x) = \mathbf{Y}^T\mathbf{Y}$$

   where $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$, by (b), above.

(d) The Y_i from (c) are i.i.d. $N(0,1)$ and $\mathbf{Y}^T\mathbf{Y} = \sum_{i=1}^n Y_i^2$ and hence

$$(\mathbf{X} - \boldsymbol{\mu}_x)^T \mathbf{V}_x^{-1} (\mathbf{X} - \boldsymbol{\mu}_x) \sim \chi^2_n$$

where $n = \dim(\mathbf{X})$.

3. Basic matrix algebra revision. You should find these easy, but if not please study the matrix notes from the course web page again.

(a) If $\mathbf{y} = (1, -3)^T$ and

$$\mathbf{B} = \begin{pmatrix} -1 & 2 & -1 \\ 2 & -3 & 0 \end{pmatrix}$$

Find $\mathbf{B}^T\mathbf{y}$.

(b) $\mathbf{A}$ is a full rank 3×3 matrix and $\mathbf{B}$ is a full rank 5×3 matrix. State the dimensions (i.e. number of rows and columns) of the following if they exist. For those that do not exist, explain, in one sentence each, why not.

i. $\mathbf{A}^{-1}\mathbf{B}^T$.
ii. $\mathbf{A}^{-1}\mathbf{B}$.
iii. $\mathbf{B}^{-1}\mathbf{A}$.
iv. $\mathbf{BA}$.
v. $\mathbf{B}^{-1}\mathbf{A}^T$.
vi. $\mathbf{BA}^{-1}$.
vii. $(\mathbf{BA})^{-1}$.
viii. $\mathbf{B}^T\mathbf{A}$.
ix. $\mathbf{B} + \mathbf{A}$.
x. $\mathbf{B} + \mathbf{A}^T$.

Solution

(a) $\mathbf{B}^T\mathbf{y} = (-7, 11, -1)^T$

(b) i. 3×5.
ii. Doesn't exist $\mathbf{A}^{-1}$ has fewer columns than $\mathbf{B}$ has rows.
iii. Doesn't exist. $\mathbf{B}$ is not square, so inverse does not exist.
iv. 5×3.
v. Doesn't exist. $\mathbf{B}$ is not square, so inverse does not exist.
vi. 5×3.
vii. Does not exist. $\mathbf{BA}$ is not square, so inverse does not exist.
viii. Does not exist. $\mathbf{B}^T$ has more columns than $\mathbf{A}$ has rows.
ix. Does not exist. $\mathbf{B}$ and $\mathbf{A}$ have different dimensions.
x. Does not exist, $\mathbf{B}$ and $\mathbf{A}^T$ have different dimensions.

4. The exponential distribution is often a reasonable model of the times between random events. Suppose then, that $x_1, x_2, \ldots x_n$ are observations of times between hardware faults on a computer network, and it is reasonable to treat the faults as independent. To plan for fault tolerance the network managers need a reasonable model for the fault occurrence rate. The p.d.f. of an exponential distribution is

$$f(x) = \lambda e^{-\lambda x} \quad x \geq 0,$$

where λ is a positive parameter. The variance of an exponential random variable is λ^{-2}.

(a) If X is a random variable from an exponential distribution with parameter λ, find $E(X)$.

(b) Hence suggest an estimator, $\hat{\lambda}$, for λ.

(c) What is the variance of $\hat{\lambda}^{-1}$?

(d) Let $\bar{x} = \sum_i x_i/n$. Find a first order Taylor expansion of $\hat{\lambda}$ about $E(\bar{x})$, considering $\hat{\lambda}$ as a function of $\bar{x}$.

(e) Hence find an approximation for the variance of $\hat{\lambda}$, in terms of n and $\bar{x}$. This use of Taylor expansions to compute approximate variances via linearization is known as the *delta method* in statistics (but goes by other names in physics and engineering, for example).

Solution

(a) This is just an integration by parts

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \left[-x\lambda \frac{e^{-\lambda x}}{\lambda} + \int \lambda \frac{e^{-\lambda x}}{\lambda} dx \right]_0^\infty = \left[xe^{-\lambda x} - \frac{e^{-\lambda x}}{\lambda} \right]_0^\infty = \frac{1}{\lambda}$$

(b) So the obvious estimator is $\hat{\lambda} = \bar{x}^{-1}$.

(c) $\hat{\lambda}^{-1} = \bar{x}$ which has variance $\lambda^{-2}n^{-1}$.

(d)

$$\hat{\lambda}(\bar{x}) \simeq E(\bar{x})^{-1} + \frac{\mathrm{d}\hat{\lambda}}{\mathrm{d}\bar{x}}(\bar{x} - E(\bar{x})) = E(\bar{x})^{-1} - \frac{1}{E(\bar{x})^2}(\bar{x} - E(\bar{x})).$$

(e) Hence

$$\mathrm{var}(\hat{\lambda}) \simeq \frac{1}{E(\bar{x})^4} E(\bar{x} - E(\bar{x}))^2 = \lambda^{-2}n^{-1}E(\bar{x})^{-4} \simeq \bar{x}^{-2}n^{-1}.$$

5. Consider again the setup from the previous question, but now taking a Bayesian approach. This means that we need to augment our model with a prior distribution for the parameter: $\lambda \sim \mathrm{gamma}(\alpha, \theta)$. So the prior p.d.f. of λ is

$$f(\lambda) = \frac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\theta^\alpha \Gamma(\alpha)}$$

which has expectation $\alpha\theta$ and variance $\alpha\theta^2$.

(a) Write down the p.d.f. for the joint distribution of the data $x_1, x_2, \ldots$ given λ.

(b) By considering the joint distribution of the λ and $\mathbf{x}$ (the vector of x_i 's), identify the posterior distribution of λ given $\mathbf{x}$.

(c) What are the posterior expectation and variance of λ?

(d) Consider the situation in which $n \to \infty$. What happens to the Bayesian and frequentist inferences about λ in this case?

Solution

(a) By independence of the observations the joint p.d.f. is

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_i x_i}.$$

(b) $f(\lambda|\mathbf{x}) \propto f(\mathbf{x}, \lambda) = f(\mathbf{x}|\lambda)f(\lambda) \propto \lambda^n e^{-\lambda \sum_i x_i} \lambda^{\alpha-1} e^{-\lambda/\theta} = \lambda^{n+\alpha-1} e^{-\lambda(\sum_i x_i + 1/\theta)}$.
i.e. $\lambda|\mathbf{x} \sim \mathrm{gamma}(n + \alpha, (\sum_i x_i + 1/\theta)^{-1})$.

(c) From the given mean and variance of a gamma distribution

$$E(\lambda|\mathbf{x}) = \frac{n+\alpha}{\sum_i x_i + 1/\theta} \text{ and } \mathrm{var}(\lambda|\mathbf{x}) = \frac{n+\alpha}{(\sum_i x_i + 1/\theta)^2}.$$

(d) As $n \to \infty$, $E(\lambda|\mathbf{x}) \to \bar{x}^{-1}$ and $\mathrm{var}(\lambda|\mathbf{x}) \to \bar{x}^{-2}n^{-1}$. i.e. the Bayesian posterior expectation for λ tends to the frequentist $\hat{\lambda}$, and the Bayesian posterior variance tends to the frequentist estimator variance. Of course our frequentist variance was an approximation, but actually it is one for which the error tends to zero as $n \to \infty$.