**Mathematical Tripos: Part IB** **DJS/Lent 2016**

# Statistics: Example Sheet 3 (of 3)

*Comments and corrections to david@statslab.cam.ac.uk*

1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where $\sigma^2$ is unknown, and suppose we are interested in testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Letting $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$, show that the likelihood ratio can be expressed as

$$\Lambda_{\mathbf{X}}(H_0, H_1) = \left(1 + \frac{T^2}{n-1}\right)^{n/2},$$

where $T = \frac{n^{1/2}(\bar{X}-\mu_0)}{\{S_{XX}/(n-1)\}^{1/2}}$. Determine the distribution of $T$ under $H_0$, and hence determine the size $\alpha$ likelihood ratio test.

2. Statisticians A and B obtain independent samples $X_1, \ldots, X_{10}$ and $Y_1, \ldots, Y_{17}$ respectively, both from a $N(\mu, \sigma^2)$ distribution with both $\mu$ and $\sigma^2$ unknown. They estimate $(\mu, \sigma^2)$ by $(\bar{X}, S_{XX}/9)$ and $(\bar{Y}, S_{YY}/16)$ respectively, where, for example, $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$ and $S_{XX} = \sum_{i=1}^{10}(X_i - \bar{X})^2$. Given that $\bar{X} = 5.5$ and $\bar{Y} = 5.8$, which statistician's estimate of $\sigma^2$ is more probable to have exceeded the true value by more than 50%? Find this probability (approximately) in each case. [Hint: This is something of a 'trick' question. Why? You may find $\chi^2$ tables helpful.]

3. Suppose that $X_1, \ldots, X_m$ are iid $N(\mu_X, \sigma_X^2)$, and, independently, $Y_1, \ldots, Y_n$ are iid $N(\mu_Y, \sigma_Y^2)$, with $\mu_X$, $\mu_Y$, $\sigma_X^2$ and $\sigma_Y^2$ unknown. Write down the distributions of $S_{XX}/\sigma_X^2$ and $S_{YY}/\sigma_Y^2$. Derive a $100(1-\alpha)\%$ confidence interval for $\sigma_X^2/\sigma_Y^2$.

4. (a) Let $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, and let $A$ be an arbitrary $m \times n$ matrix. Prove directly from the definition that $A\mathbf{X}$ has an $m$-variate normal distribution. Show that $\mathrm{cov}(A\mathbf{X}) = A\Sigma A^T$, and that $A\mathbf{X} \sim N_m(A\boldsymbol{\mu}, A\Sigma A^T)$. Give an alternative proof that $A\mathbf{X} \sim N_m(A\boldsymbol{\mu}, A\Sigma A^T)$ using moment generating functions.

   (b) Let $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, and let $\mathbf{X_1}$ denote the first $n_1$ components of $\mathbf{X}$. Let $\boldsymbol{\mu_1}$ denote the first $n_1$ components of $\boldsymbol{\mu}$, and let $\Sigma_{11}$ denote the upper left $n_1 \times n_1$ block of $\Sigma$. Show that $\mathbf{X_1} \sim N_{n_1}(\boldsymbol{\mu_1}, \Sigma_{11})$.

5. Consider the simple linear regression model $Y_i = a + bx_i + \varepsilon_i$, $i = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ and $\sum_{i=1}^n x_i = 0$. Derive from first principles explicit expressions for the MLEs $\hat{a}$, $\hat{b}$ and $\hat{\sigma}^2$. Show that we can obtain the same expressions if we regard the simple linear regression model as a special case of the general linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and specialise the formulae $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$ and $\hat{\sigma}^2 = n^{-1}\|\mathbf{Y} - X\hat{\boldsymbol{\beta}}\|^2$.

6. Consider the model $Y_i = bx_i + \varepsilon_i$, $i = 1, \ldots, n$, where the $\varepsilon_i$ are independent with mean zero and variance $\sigma^2$ (regression through the origin). Write this in the form $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and find the least squares estimator of $b$.

   The relationship between the range in metres, $Y$, of a howitzer with muzzle velocity $v$ metres per second fired at angle of elevation $\alpha$ degrees is assumed to be $Y = \frac{v^2}{g}\sin(2\alpha) + \varepsilon$, where $g = 9.81$ and where $\varepsilon$ has mean zero and variance $\sigma^2$. Estimate $v$ from the following independent observations made on 9 shells.

| $\alpha$ (deg) | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 |
|---|---|---|---|---|---|---|---|---|---|
| $\sin 2\alpha$ | 0.1736 | 0.3420 | 0.5 | 0.6428 | 0.7660 | 0.8660 | 0.9397 | 0.9848 | 1 |
| range (m) | 4860 | 9580 | 14080 | 18100 | 21550 | 24350 | 26400 | 27700 | 28300 |

7. Consider the model $Y_i = \mu + \varepsilon_i$, $i = 1, \ldots, n$, where $\varepsilon_i$ are iid $N(0, \sigma^2)$ random variables. Write this in matrix form $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and find the MLE $\hat{\boldsymbol{\beta}}$. Find the fitted values, the residuals and the residual sum of squares. Show how applying Theorem 13.3 (in lectures) to this case gives the independence of $\bar{Y}$ and $S_{YY}$ for an iid sample from $N(\mu, \sigma^2)$. Write down an unbiased estimate $\tilde{\sigma}^2$ of $\sigma^2$.

8. Consider the linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{Y}$ is an $n \times 1$ vector of observations, $X$ is a known $n \times p$ matrix of rank $p$, $\boldsymbol{\beta}$ is a $p \times 1$ unknown parameter vector and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of uncorrelated random variables with mean zero and variance $\sigma^2$ (i.e. we are *not* assuming that the $\varepsilon_i$ are normally distributed). Let $\hat{\boldsymbol{\beta}}$ denote the least squares estimate of $\boldsymbol{\beta}$, $\hat{\mathbf{Y}}$ denote the vector of fitted values, and let $\mathbf{R}$ be the vector of residuals. Find $\mathbb{E}(\mathbf{R})$ and $\mathrm{cov}(\mathbf{R})$. Show that $\mathrm{cov}(\mathbf{R}, \hat{\boldsymbol{\beta}}) = 0$ and $\mathrm{cov}(\mathbf{R}, \hat{\mathbf{Y}}) = 0$.

9. For the simple linear regresssion model $Y_i = a + bx_i + \varepsilon_i$, $i = 1, \ldots, n$, where $\sum_i x_i = 0$ and where the $\varepsilon_i$ are iid $N(0, \sigma^2)$ random variables, the MLEs $\hat{a}$ and $\hat{b}$ were found in Question 5. Find the distribution of $\hat{\boldsymbol{\beta}} = (\hat{a}, \hat{b})^T$. Find a 95% confidence interval for $b$ and for the mean value of $Y$ when $x = 1$. [Hint: Look at "Applications of the distribution theory" in lectures.]

10. *Question 19H from 2013 Part II paper*

    Consider the general linear model $Y = X\theta + \epsilon$ where $X$ is a known $n \times p$ matrix, $\theta$ is an unknown $p \times 1$ vector of parameters, and $\epsilon$ is an $n \times 1$ vector of independent $N(0, \sigma^2)$ random variables with unknown variance $\sigma^2$. Assume the $p \times p$ matrix $X^T X$ is invertible. Let

$$\begin{aligned} \hat{\theta} &= (X^T X)^{-1} X^T Y \\ \hat{\epsilon} &= Y - X\hat{\theta} \end{aligned}$$

    What are the distributions of $\hat{\theta}$ and $\hat{\epsilon}$? Show that $\hat{\theta}$ and $\hat{\epsilon}$ are uncorrelated.

    Four apple trees stand in a $2 \times 2$ rectangular grid. The annual yield of the tree at coordinate $(i, j)$ conforms to the model

$$y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}, \quad i, j \in \{1, 2\},$$

where $x_{ij}$ is the amount of fertilizer applied to tree $(i, j)$, $\alpha_1, \alpha_2$ may differ because of varying soil across rows, and the $\epsilon_{ij}$ are $N(0, \sigma^2)$ random variables that are independent of one another and from year to year. The following two possible experiments are to be compared:

$$\text{I} : (x_{ij}) = \begin{pmatrix} 0 & 1 \\ 2 & 3 \end{pmatrix} \text{ and } \text{II} : (x_{ij}) = \begin{pmatrix} 0 & 2 \\ 3 & 1 \end{pmatrix}.$$

Represent these as general linear models, with $\theta = (\alpha_1, \alpha_2, \beta)$. Compare the variances of estimates of $\beta$ under I and II.

With II the following yields are observed:

$$(y_{ij}) = \begin{pmatrix} 100 & 300 \\ 600 & 400 \end{pmatrix}.$$

Forecast the total yield that will be obtained next year if no fertilizer is used. What is the 95% predictive interval for this yield?