**Mathematical Tripos: Part IB** **DJS/Lent 2015**

# Statistics: Example Sheet 1 (of 3)

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

1. Ask your supervisor to test you on the sheet of common distributions handed out in lectures.

2. **(Probability review)** If $X \sim \text{Exponential}(\lambda)$ and $Y \sim \text{Exponential}(\mu)$ are independent, derive the distribution of $\min(X, Y)$. If $X \sim \text{Gamma}(\alpha, \lambda)$ and $Y \sim \text{Gamma}(\beta, \lambda)$ are independent, derive the distributions of $X + Y$ and $X/(X + Y)$.

3. In a genetics experiment, a sample of $n$ individuals was found to include $a, b, c$ of the three possible genotypes $GG$, $Gg$, $gg$ respectively. The population frequency of a gene of type $G$ is $\theta/(\theta + 1)$, where $\theta$ is unknown, and it is assumed that the individuals are unrelated and that two genes in a single individual are independent. Show that the likelihood of $\theta$ is proportional to $\theta^{2a+b}/(1 + \theta)^{2a+2b+2c}$ and that the maximum likelihood estimate of $\theta$ is $(2a + b)/(b + 2c)$.

4. (a) Let $X_1, \ldots, X_n$ be independent Poisson random variables, with $X_i$ having mean $i\theta$, for some $\theta > 0$. Show that $T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for $\theta$ and write down the distribution of $T$. Show that the maximum likelihood estimator $\hat{\theta}$ of $\theta$ is a function of $T$, and show that it is unbiased.

   (b) For some $n > 2$, let $X_1, \ldots, X_n$ be iid with $X_i \sim \text{Exponential}(\theta)$. Find a minimal sufficient statistic $T$ and write down its distribution. Show that the maximum likelihood estimator $\hat{\theta}$ of $\theta$ is a function of $T$, and show that it is biased, but asymptotically unbiased. Find an injective function $h$ on $(0, \infty)$ such that, writing $\psi = h(\theta)$, the maximum likelihood estimator $\hat{\psi}$ of the new parameter $\psi$ is unbiased.

5. Suppose $X_1, \ldots, X_n$ are independent random variables with distribution $\text{Bin}(1, p)$.

   (a) Show that a sufficient statistic for $\theta = (1 - p)^2$ is $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ and that the MLE for $\theta$ is $\left(1 - \frac{1}{n}T\right)^2$.
   *Hint: use the chain rule, $df/d\theta = (df/dp)(dp/d\theta)$.*

   (b) Show that the MLE is a biased estimator for $\theta$. Let $\tilde{\theta} = 1(X_1 + X_2 = 0)$. Show that $\tilde{\theta}$ is unbiased for $\theta$. Use the Rao–Blackwell theorem to find a function of $T$ which is an unbiased estimator for $\theta$.

6. For some $n \geq 2$, suppose that $X_1, \ldots, X_n$ are iid random variables uniformly distributed on $[\theta, 2\theta]$ for some $\theta > 0$. Show that $\tilde{\theta} = \frac{2}{3}X_1$ is an unbiased estimator of $\theta$. Show that $T(\mathbf{X}) = (\min_i X_i, \max_i X_i)$ is a minimal sufficient statistic for $\theta$. Use the Rao–Blackwell theorem to find an unbiased estimator $\hat{\theta}$ of $\theta$ which is a function of $T$ and whose variance is strictly smaller than the variance of $\tilde{\theta}$ for all $\theta > 0$.

7. (a) Let $X_1, \ldots X_n$ be iid with $X_i \sim U[0, \theta]$. Find the maximum likelihood estimator $\hat{\theta}$ of $\theta$. Show that the distribution of $R(\mathbf{X}, \theta) = \hat{\theta}/\theta$ does not depend on $\theta$, and use $R(\mathbf{X}, \theta)$ to find a $100(1-\alpha)\%$ confidence interval for $\theta$ for $0 < \alpha < 1$.

   (b) The lengths (in minutes) of calls to a call centre may be modelled as iid exponentially distributed random variables, and $n$ such call lengths are observed. The original sample is lost, but the data manager has noted down $n$ and $t$ where $t$ is the total length of the $n$ calls in minutes. Derive a 95% confidence interval for the probability that a call is longer than 2 minutes if $n = 50$ and $t = 105.3$.

8. Suppose that $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$ independently, where $\theta_1$ and $\theta_2$ are unknown. Show that $(\theta_1 - X_1)^2 + (\theta_2 - X_2)^2$ has a $\chi^2_2$ distribution and that this is the same as Exponential($\frac{1}{2}$), i.e., the exponential distribution with mean 2.

   Show that both the square $S$ and circle $C$ in $\mathbb{R}^2$, given by

$$S = \{(\theta_1, \theta_2) : |\theta_1 - X_1| \leq 2.236;\ |\theta_2 - X_2| \leq 2.236\}$$
$$C = \left\{(\theta_1, \theta_2) : (\theta_1 - X_1)^2 + (\theta_2 - X_2)^2 \leq 5.991\right\}$$

   are 95% confidence regions for $(\theta_1, \theta_2)$.

   *Hint:* $\Phi(2.236) = (1 + \sqrt{.95})/2$, *where* $\Phi$ *is the distribution function of* $N(0, 1)$. What might be a sensible criterion for choosing between $S$ and $C$?

9. Suppose that the number of defects on a roll of magnetic recording tape is modelled with a Poisson distribution for which the mean $\lambda$ is known to be either 1 or 1.5. Suppose the prior mass function for $\lambda$ is

$$\pi_\lambda(1) = 0.4, \quad \pi_\lambda(1.5) = 0.6\,.$$

   A random sample of five rolls of tape has $\mathbf{x} = (3, 1, 4, 6, 2)$ defects respectively. Show that the posterior distribution for $\lambda$ given $\mathbf{x}$ is

$$\pi_{\lambda|\mathbf{X}}(1 \mid \mathbf{x}) = 0.012, \quad \pi_{\lambda|\mathbf{X}}(1.5 \mid \mathbf{x}) = 0.988\,.$$

10. Suppose $X_1, \ldots, X_n$ are iid with (conditional) probability density function $f(x \mid \theta) = \theta x^{\theta-1}$ for $0 < x < 1$ (and is zero otherwise), for some $\theta > 0$. Suppose that the prior for $\theta$ is Gamma$(\alpha, \beta)$, $\alpha > 0, \beta > 0$. Find the posterior distribution of $\theta$ given $\mathbf{X} = (X_1, \ldots, X_n)$ and the Bayesian estimator of $\theta$ under quadratic loss.

+11 For some $n \geq 3$, let $\epsilon_1, \ldots, \epsilon_n$ be iid with $\epsilon_i \sim N(0, 1)$. Set $X_1 = \epsilon_1$ and $X_i = \theta X_{i-1} + (1 - \theta^2)^{1/2}\epsilon_i$ for $i = 2, \ldots, n$ and some $\theta \in (-1, 1)$. Find a sufficient statistic for $\theta$ that takes values in a subset of $\mathbb{R}^3$.

**Mathematical Tripos: Part IB** **DJS/Lent 2015**

# Statistics: Example Sheet 2 (of 3)

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

1. Let $X$ have density function $f(x;\theta) = \frac{\theta}{(x+\theta)^2}$, $x > 0$, where $\theta \in (0,\infty)$ is an unknown parameter. Find the likelihood ratio test of size 0.05 of $H_0 : \theta = 1$ against $H_1 : \theta = 2$, and show that the probability of Type II error is 19/21.

2. Let $X_1, X_2, \ldots, X_n$ be iid random variables, each with a Poisson distribution with parameter $\theta$ (and therefore with mean $\theta$ and variance $\theta$). Find the form of the likelihood ratio test of $H_0 : \theta = 1$ against $H_1 : \theta = 1.21$. By using the Central Limit Theorem to approximate the distribution of $\sum_i X_i$, show that the smallest value of $n$ required to make $\alpha = 0.05$ and $\beta \leq 0.1$ (where $\alpha$ and $\beta$ are the Type I and Type II error probabilities) is somewhere near 212.

3. Let $f_0$ and $f_1$ be probability mass functions for $\mathbf{X} = (X_1 \ldots, X_n)$ on a countable set $\mathcal{X}^n$. State and prove a version of the Neyman–Pearson lemma for a size $\alpha$ test of $H_0 : f = f_0$ against $H_1 : f = f_1$, assuming that $\alpha$ is such that there exists a likelihood ratio test of exact size $\alpha$.

4. Let $X \sim \text{Bin}(2,\theta)$ and consider testing $H_0 : \theta = \frac{1}{2}$ against $H_1 : \theta = \frac{3}{4}$. Find the possible values of $\alpha$ for which there exists a likelihood ratio test with size exactly $\alpha$.

5. Let $X_1, \ldots, X_n$ be iid random variables each with a $N(\mu_0, \sigma^2)$ distribution, where $\mu_0$ is known and $\sigma^2$ is unknown. Find the best (most powerful) test of size at most $\alpha$ for testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 = \sigma_1^2$ for known $\sigma_0^2$ and $\sigma_1^2$ $(> \sigma_0^2)$. Show that this test is a size $\alpha$ uniformly most powerful test for testing $H_0' : \sigma^2 \leq \sigma_0^2$ against $H_1' : \sigma^2 > \sigma_0^2$.

6. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exponential}(\theta)$. Find the likelihood ratio test of size $\alpha$ of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ $(> \theta_0)$ and derive an expression for the power function. Is the test uniformly most powerful for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$? Is it uniformly most powerful for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$?

7. Let $X_1, \ldots X_n, Y_1, \ldots, Y_n$ be independent, with $X_1, \ldots, X_n \sim \text{Exponential}(\theta_1)$ and $Y_1, \ldots, Y_n \sim \text{Exponential}(\theta_2)$. Recalling the forms of the relevant MLEs from Sheet 1, show that the likelihood ratio of $H_0 : \theta_1 = \theta_2$ and $H_1 : \theta_1 \neq \theta_2$ is a monotone function of $|t - 1/2|$, where $t$ is the observed value of the statistic $T$ given by

$$T = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i + \sum_{i=1}^n Y_i}.$$

By writing down the distribution of $T$ under $H_0$, express the likelihood ratio test of size $\alpha$ in terms of $|T - 1/2|$ and the percentage points of a beta distribution.
*Hint: use Question 2 on Example Sheet 1.*

8. A machine produces plastic articles (many of which are defective) in bunches of three articles at a time. Under the null hypothesis that each article has a constant (but unknown) probability $\theta$ of being defective, write down the probabilities $p_i(\theta)$ of a bunch having $i$ defective articles, for $i = 0, 1, 2, 3$. In an trial run in which 512 bunches were produced, the numbers of bunches with $i$ defective articles were 213 ($i = 0$), 228 ($i = 1$), 57 ($i = 2$) and 14 ($i = 3$). Carry out Pearson's chi-squared test at the 5% level of the null hypothesis, explaining carefully why the test statistic should be referred to the $\chi^2_2$ distribution.

9. A random sample of 59 people from the planet Krypton yielded the results below.

| | | Eye-colour | |
|---|---|---|---|
| | | 1 (Blue) | 2 (Brown) |
| Sex | 1 (Male) | 19 | 10 |
| | 2 (Female) | 9 | 21 |

Carry out Pearson's chi-squared test at the 5% level of the null hypothesis that sex and eye-colour are independent factors on Krypton. Now carry out the corresponding test at the 5% level of the null hypothesis that each of the cell probabilities is equal to 1/4. Comment on your results.

10. Write down from lectures the model and hypotheses for a test of homogeneity in a two-way contingency table. By first deriving the MLEs under each hypothesis, show that the likelihood ratio and Pearson's chi-squared tests are identical to those for the independence test. Apply the homogeneity test to the data below from a clinical trial for a drug, obtained by randomly allocating 150 patients to three equal groups (so the row totals are fixed).

| | Improved | No difference | Worse |
|---|---|---|---|
| Placebo | 18 | 17 | 15 |
| Half dose | 20 | 10 | 20 |
| Full dose | 25 | 13 | 12 |

$^+$11 In Question 3, does there exist a version of the Neyman–Pearson lemma when a likelihood ratio test of exact size $\alpha$ does not exist?

**Mathematical Tripos: Part IB** **DJS/Lent 2015**

# Statistics: Example Sheet 3 (of 3)

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

1. (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 $\mathbf{X_1} \sim N_{n_1}(\boldsymbol{\mu_1}, \Sigma_{11})$.

2. 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.

3. 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.]

4. 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 $\sigma_X^2/\sigma_Y^2$.

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 $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$ 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 13.2 (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 one-way analysis of variance (ANOVA) model $Y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, where $(\varepsilon_{ij}) \overset{iid}{\sim} N(0, \sigma^2)$. Derive from first principles explicit expressions for the MLEs $\hat{\mu}_1, \ldots, \hat{\mu}_I$ and $\hat{\sigma}^2$. Show that we can obtain the same expressions if we regard the ANOVA model as a special case of the general linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and specialise the formulae $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$ and $\hat{\sigma}^2 = n^{-1}\|\mathbf{Y} - X\hat{\boldsymbol{\beta}}\|^2$.

9. 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 $\mathrm{cov}(\mathbf{R}, \hat{\mathbf{Y}}) = 0$.

10. 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.]

+11 Consider the one-way ANOVA model of Question 8. Letting $\bar{Y}_{i+} = n_i^{-1}\sum_{j=1}^{n_i} Y_{ij}$ and $\bar{Y}_{++} = n^{-1}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}$ with $n = n_1 + \ldots + n_I$, show from first principles that the size $\alpha$ likelihood ratio test of equality of means rejects $H_0$ if

$$F \equiv \frac{\frac{1}{I-1}\sum_{i=1}^{I} n_i(\bar{Y}_{i+} - \bar{Y}_{++})^2}{\frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i+})^2} > F_{I-1,n-I}(\alpha),$$

i.e. if 'the ratio of the between groups sum of squares to the within groups sum of squares is large'.