MA930
UNIVERSITY OF WARWICK
CLASS TEST 2014

Data Analysis
Time Allowed: 1.5 Hours

Full marks may be gained by correctly answering three complete questions. Candidates may attempt all questions. Marks will be awarded for the best three answers only.

Please write your name and student number on the answer booklet.

---

1. Summary statistics: Let $X_1, \ldots, X_n$ denote a collection of numbers gathered during an experiment. We will model them as independent, identically distributed random variables with p.d.f. f and c.d.f. F.

   (a) What is a statistic?

   (b) Name and *briefly* describe 6 statistics commonly used to summarize the distribution of a collection of numbers.

   (c) What are the order statistics for $X_1, \ldots, X_n$? What is their joint distribution (i.e. the multi-dimensional pdf)?

   (d) Describe a simple test for checking if the $X_1, \ldots, X_n$ seem to be samples from a given distribution (for example, the $N(0, 1)$ distribution).

   (e) If $X_1, \ldots, X_n$ pass the test from part (d), does that prove that they are independent? Briefly justify your answer.

   (f) Assume that the X_i are independent samples from the Uniform$(0, 1)$ distribution. Show that $X_{(1)}$ converges to zero in probability.

   (g) Assume n is odd (i.e. $n = 2k+1$). State a statistic S that can be used to approximate the median of the distribution of the (X_i). Using the law of large numbers, or otherwise, show that if the $X_i \sim$ Uniform$(0, 1)$ then S converges in probability.

Continued ...

2. Last year, 24 students took a "data analysis" viva, 12 in the morning and 12 in the afternoon. You want to determine if the time of the viva (morning or afternoon) affects the outcome. Let $X_1, \ldots, X_{12}$ denote the morning scores and $Y_1, \ldots, Y_{12}$ denote the afternoon scores. Assume that the students scores can be treated as independent random variables, that the morning scores $X_i \sim N(\mu_1, \sigma^2)$ and the afternoon scores $Y_i \sim N(\mu_2, \sigma^2)$. You want to find out if $\mu_1 < \mu_2$ or $\mu_1 = \mu_2$ or $\mu_1 > \mu_2$.

   (a) What is the distribution of $\bar{X} = \frac{1}{12}\sum_{i=1}^{12} X_i$ and $\bar{Y} = \frac{1}{12}\sum_{i=1}^{12} Y_i$.

   (b) What is the distribution of $Z_1 = \frac{1}{\sigma^2}\sum_{i=1}^{12}(X_i - \bar{X})^2$ and $Z_2 = \frac{1}{\sigma^2}\sum_{i=1}^{12}(Y_i - \bar{Y})^2$?

   (c) Use your answers to part (a) and (b) to form 95% confidence intervals for μ_1 and μ_2.

   (d) If the two confidence intervals from (c) overlap, is there any special statistical significance to that? (Hint: read the rest of the question before answering :-)

   (e) What is the distribution of $(\bar{X} - \bar{Y})/\sigma$ and $Z_1 + Z_2$?

   (f) Use your answer to part (e) to form an (exact) 95% confidence interval for $\mu_1 - \mu_2$. Does the confidence interval containing zero have any special significance?

   (g) Suppose your two confidence intervals from part (c) overlap, but that your confidence interval from part (e) does not include zero. How would you interpret this outcome?

Continued ...

3. An experiment produces independent pairs of observations (X_i, Y_i), $i = 1, \ldots, n$. Assume that $Y_i = A + BX_i + e_i$ where the e_i are random variables with mean zero.

   (a) Define and derive least squares esimators for A and B.

   (b) Let
   $$X := \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix},$$
   $\beta := \begin{pmatrix} A \\ B \end{pmatrix}$, and let I_n denote the $n \times n$ identity matrix. Suppose now that the e_i are independent $N(0, \sigma^2)$ random variables, so that $Y = (Y_i)_{i=1}^n$ has the multivariate normal distribution $N(X\beta, \sigma^2 I_n)$. Stating clearly any results you use, show that $\hat{\beta} = (X^t X)^{-1} X^t y$ is an unbiased estimator for β and give the distribution of $\hat{\beta}$

   (c) You may assume without proof that $\frac{1}{\sigma^2} \sum (Y - X\hat{\beta}) \sim \chi^2_{n-2}$ and is independent of $\hat{\beta}$ You suspect $B = \frac{1}{2}$. How can you test the hypothesis $H_0 : B = \frac{1}{2}$ against the alternative hypothesis $H_1 : B \neq \frac{1}{2}$.

   (d) Pearson's famous "fathers and sons " dataset contains the heights of 1078 pairs of fathers and sons. Here is linear model for the heights of the sons in terms of their fathers' heights, fitted in R.

```
# Y = sons.height
# X = fathers.height
# Y = A + BX + e

Estimator   Estimate     Std. Error     t value    Pr(>|t|)
A-hat       33.88660       1.83235       18.49      <2e-16 ***
B-hat        0.51409       0.02705       19.01      <2e-16 ***
```

   The standard errors correspond to

   $$S\sqrt{(X^t X)^{-1}_{11}} \quad \text{and} \quad S\sqrt{(X^t X)^{-1}_{22}} \quad \text{with} \quad S^2 = \frac{1}{1078 - 2} \sum_{i=1}^{1078} (Y_i - \hat{A} - \hat{B} X_i)^2.$$

   Briefly explain how to calculate confidence intervals for A and B?

Continued ...

4. Consider two collections of random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. The X_i all have p.d.f. f and the Y_i all have p.d.f. g .

   (a) What does it mean to say that X_1 and X_2 are independent?

   (b) What does it mean to say that $X_1, X_2, \ldots, X_n$ are independent?

   (c) If $X_1, \ldots, X_n$ are independent, and $Y_1, \ldots, Y_n$ are independent, is it generally the case that $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are independent. Give a brief proof or a counter-example.

   (d) Define $\mathbb{E}[X_i]$ and $\mathrm{Var}(X_i)$ in terms of f .

   (e) Use the properties of expectation to show that for random variables X, Y

   $$\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

   What is this quantity called?

   (f) If $\mathrm{Var}(X_i) < \infty$ and $\mathrm{Var}(Y_i) < \infty$, and X_i, Y_i are independent, is it always true that $\mathrm{Var}(\mathrm{X_i Y_i}) < \infty$? Give a proof or a counterexample. Can you express $\mathrm{Var}(\mathrm{X_i Y_i})$ in terms of the mean and variances of the X and Y distributions.

   (g) Using theorems mentioned during the course, what can you say about the distribution of

   $$C := \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - \bar{X}\bar{Y}, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

   in the case that both the X_i and Y_i have mean zero, variance 1, and with full independence for $X_1, \ldots, X_n, Y_1, \ldots, Y_n$? For large n, for which values of C would you reject a hypothesis that the X_i and Y_i are independent. Your answer does not need to be exact, but you should try to be as accurate as possible in an asymptotic sense.

End

Common distributions

- $X \sim \text{Bin}(n, p)$ if
$$\mathbb{P}(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$
- $X \sim \text{Uniform(a, b)}$ if
$$f(x) = \begin{cases} 1/(b-a), & a < x < b \\ 0 & \text{otherwise} \end{cases}$$
- Exponential: $X \sim \text{Exp}(\lambda)$ if
$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & x > 0 \\ 0 & x < 0 \end{cases}$$
- $X \sim N(\mu, \sigma^2)$ if it has p.d.f. $\frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/(2\sigma^2))$
- if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent then
$$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$
- if $X_1, \ldots X_n \sim N(0, 1)$ then $\sum_{i=1}^n X_i^2 \sim \chi_n^2$.
- Students t-distribution: If $Z \sim N(0, 1)$ and $U \sim \chi_n^2$ then
$$\frac{Z}{\sqrt{U/n}} \sim t_n$$
- Beta pdf mean variance
$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(a, b)}$$
mean $\frac{\alpha}{\alpha+\beta}$ variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Gamma
$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} exp(-\beta x)$$
mean $\frac{\alpha}{\beta}$ variance $\frac{\alpha}{\beta^2}$.
- Multivariate normal $N(\mu, \Sigma)$, $\mu \in \mathbb{R}^n$, Σ an $n \times n$ matrix:
$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (X-\mu)\right).$$
If $X \sim N(\mu, \Sigma)$ and M is an $m \times n$ matrix then then $MX \sim N(M\mu, M\Sigma M^{-1})$.