INSTITUTE OF TECHNOLOGY, BLANCHARDSTOWN

Academic Year	2016-7
Year of Programme	2
Semester	Semester 2, First Sitting
Date of Examination	Sample exam
Time of Examination	

Programme Code	Programme Title	Module Code
BN535	Master of Engineering in Internet of Things Technologies	MIOT H6014

Module Title	Statistical Analysis for Engineers

Internal Examiner: Damian Cox

External Examiner(s): Dr Kan Tadd

Instructions to candidates:

1. **To ensure that you take the correct examination, please check that the module and programme that you are following are listed in the table above.**
2. **The information section is at the end of the paper. This includes Statistical Tables.**
3. **Answer all three questions. Question 1 carries 20 marks, Question 2 carries 30 marks and Question 3 carries 50 marks.**

DO NOT TURN OVER THIS PAGE UNTIL YOU ARE TOLD TO DO SO

Question 1 (20 marks)

Answer two of the following three parts of this question. Each part carries 10 marks.

a) For a list of paired numbers (X_i, Y_i), it is required to estimate numbers α and β such that

$$Y_i = \alpha + \beta X_i, \text{ for } i = 1 \text{ to } n.$$

This is treated as a set of n equations for α and β, with the quantities X_i and Y_i as the coefficients. Let Y be the column vector with Y_i as the i-th element and let X be the matrix

$$X = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_1 & X_2 & \ldots & X_n \end{pmatrix}^T.$$

Let a and b be the estimates of α and β found by solving the over-determined system shown by means of the pseudo-inverse:

$$X\begin{pmatrix} a \\ b \end{pmatrix} = Y, \text{ so } \begin{pmatrix} a \\ b \end{pmatrix} = \left(X^T X\right)^{-1} X^T Y.$$

Use these matrix equations to produce individual equations for a and b in terms of the values X_i and Y_i.

b) Answer the following questions.

(i) Give a definition of an event, an experiment and the sample space arising from that experiment.

(ii) Use these ideas to give initial *confidence* and *frequentist* definitions of the probability of an event. You should illustrate your definitions with an example.

(iii) Give an example of an event for which it is difficult to define a frequentist probability and explain why.

c) Let an experiment have a universal set U, also known as the universal event.

Answer the following questions.

(i) Define the sigma algebra for the experiment.

(ii) Define the concept of mutually exclusive events.

(iii) Let P be a function mapping the sigma algebra onto the interval [0, 1]. State the properties the function P must have in order to be a probability measure.

d) Consider the statement made here about two events A and B:

$$P[A \cup B] = P[B|\, A]\, P[A].$$

Answer the following questions:

(i) Explain how this equation may be used to define a conditional probability.

(ii) Expand the denominator of your revised equation to give a statement of Bayes Rule which gives $P[B|\, A]$ in terms of $P[A|\, B]$, $P[A|\, B']$ and $P[B]$.

(iii) Let D be the event of a patient having a given disease; $P[D]$ is known to be 0.01. A testing procedure has been developed for the disease; the probability that the testing procedure gives a positive result, event T, if the patient has the disease, is 0.99. The probability it gives a false positive is 0.02. Calculate the probability the patient has the disease given a positive test result.

e) Consider the statement made here about two events A and B:

$$P[A \cup B] = P[B|A]\,P[A].$$

Answer the following questions:

(i) Use this equation to give a statement of Bayes Rule which gives $P[B|A]$ in terms of $P[A|B]$, $P[A|B']$ and $P[B]$.

(ii) A blood-test on a driver determines whether they are over the legal limit of alcohol for driving. Anonymous polling has suggested that 10% of drivers will still drive after drinking enough to put them over this limit. It is proposed to apply random breathalyser-tests to drivers. It is known that:

- The probability the breathalyser test will give a positive result if the subject is over the limit is 0.95,
- The probability the breathalyser test will give a positive result if the subject is not over the limit is 0.02.

Calculate the probability a driver was not over the limit despite giving a positive result.

Question 2 (30 marks)

Answer two of the following three parts of this question. Each part carries 15 marks.

a) Two players are involved in the following game. For a given players' turn, that player adds a fixed sum into a pot of money and then roll two fair dice; if the numbers that come up are the same, that player looses and the other gets to keep the pot of money. Let variable N be the number of rounds played in the game.

(i) Write down a probability mass function for the variable N.

(ii) Let the function $f(x)$ be given by

$$f(x) = \sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

Use the two forms of the function to calculate the derivative of $f'(x)$. Use this result to calculate the expected value of N.

(iii) Calculate the second derivative $f''(x)$ and use this result to calculate the variance of the variable N.

(iv) Calculate the probability that the same player who starts the game loses. [Hint: recall that for any number k, $2k + 1$ is an odd number.]

b) Let X be a discrete random variable with the binomial distribution

$$P[X = r] = {}^nC_r p^r (1-p)^{n-r}.$$

Prove that when the probability p is small, the distribution becomes the Poisson distribution as n becomes large, with mean $\mu = np$. You should use Stirling's Approximation and the definition of the quantity ${}^nC_r$, both given here:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \quad {}^nC_r = \frac{n!}{r!(n-r)!}.$$

c) Let X be a discrete random variable with the binomial distribution with parameters n and p, so that the probability mass function is

$$P[X = r] = {}^{n}C_{r}\, p^{r}(1-p)^{n-r}, \text{ for } 0 \le r \le n,\ P[X = r] = 0 \text{ otherwise.}$$

Answer the following questions.

(i) Derive the equation for the probability mass function from the laws of probability.

(ii) When the probability p is not close to 0 or 1, explain why the distribution becomes the Normal distribution as n becomes large. Identify the mean and variance.

d) Answer all three parts of this question.

(i) Let X_i, for $i = 1$ to n, be n normally distributed random variables all with mean μ and standard deviation σ. State the Central limit theorem for these variables.

(ii) Define the variable S as

$$S^2 = \frac{\sum_i X_i^2 - n\overline{X}^2}{n-1}.$$

State the Central limit theorem for the list of normally distributed random variables X_i, for $i = 1$ to n, when n is very large.

(iii) State the central limit theorem for a list of n variables X_i, for $i = 1$ to n, all of which have a mean μ and where n is very large and S is defined as above.

(iv) Let Z be the standard normal variable. Define the number z_a by the equation

$$P[Z < z_a] = \alpha, \text{ for probability } \alpha.$$

Define the concept of a confidence interval for a probability of $1 - \alpha$, that is, the $100(1 - \alpha)\%$ confidence interval. Explain what the confidence interval will mean using a frequentist approach to a probability if actual values of the interval boundaries are calculated from data.

Question 3 (50 marks)

In the relevant question parts where it arises, the multiple regression model is

$$Y = X\beta + \varepsilon, \text{ with } \sigma^2 = \text{var}[\varepsilon_i] \text{ for all } i.$$

where Y is an n by 1 vector of observable 'response' random variables, X is the n by p 'design' matrix of random variables with full rank p, β is a p by 1 vector of unknown parameters and ε is an n by 1 vector of random variables, the 'random error'. Let b be the Ordinary Least Squares estimator for β, given by

$$b = (X^T X)^{-1} X^T Y.$$

It is assumed that the elements of ε are independent and identically distributed with mean 0 and variance σ^2 and that if X is random, ε and X are independent.

Answer five of the following six parts of this question. Each part carries 10 marks.

a) Let p be the probability that one particular side of a coin ('heads') is left facing up when the coin is spun up and then lands on a flat surface. It is to be determined whether a coin is fair by means of a Null Hypothesis based statistical test with '$p = 0.5$' as the Null hypothesis.

Answer the following questions.

(i) It has been decided that the test will be carried out with just ten throws of the coin. Identify an appropriate statistic to carry out a two-tailed test on the Null Hypothesis. Assuming the value of the level of significance α is to be 0.05 or less, identify the critical values of the relevant statistic.

(ii) If $\alpha < 0.05$, identify the minimal number of throws needed for a meaningful test.

(iii) The number of throws is now to be a large number above 1,000. Identify a suitable test statistic and write an equation for it.

b) Let Z be a standard normal variable and let Z_i, for $i = 1$ to k, be k random variables with the standard normal distribution. Let T_k be the variable with the Student t-distribution with k degrees of freedom, which is defined by the equation

$$T = \frac{Z}{\sqrt{\sum_i Z_i^2 \Big/ k}}.$$

The probability density function $f_k(x)$ for this variable is given by the equation

$$f_k(x) = C_k \left(1 + \frac{x^2}{k}\right)^{-\frac{1}{2}(k+1)} \text{ where } C_k = \frac{\Gamma(\frac{1}{2}(k+1))}{\sqrt{\pi k}\,\Gamma(\frac{1}{2}k)}.$$

Answer the following questions.

(i) Show that as k becomes very large, the variable T becomes the standard normal variable.

(ii) Let X_i, for $i = 1$ to n, be n normally distributed random variables all with mean μ and standard deviation σ. The quantities $\overline{X}$ and S are defined by the equations

$$\overline{X} = \frac{\sum_i X_i}{n} \text{ and } S^2 = \frac{\sum_i X_i^2 - n\overline{X}^2}{n-1}.$$

Show that the random variable T defined in the equation below follows the Student t-distribution with $n - 1$ degrees of freedom:

$$T = \frac{\overline{X} - \mu}{S \Big/ \sqrt{n}}.$$

Explain how this result leads to a statistical test on the mean.

c) The Null-Hypothesis-based statistical test on 'goodness-of-fit' to a distribution uses the test statistic Q, defined in the equation below:

$$Q = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

Answer the following questions.

(i) Explain the structure of the test, its assumptions and the origin of the distribution for the statistic Q. A formal mathematical proof is not required.

(ii) Explain how this test is used to set up the standard 'Test for independence' for two categorical variables, with data represented in a contingency table.

d) It is to be determined whether a dice is fair, that is to say, whether rolling the dice gives any one of the six sides with equal probability.

Answer the following questions.

(i) Set up the Null and alternative Hypothesis and other elements of the test.

(ii) The number of throws is to be a large number above 100. Identify a suitable test statistic Q for the goodness-of-fit test and write an equation for it. Simplify this equation under the Null hypothesis to show that the statistic is

$$Q = \frac{\sum_i O_i^2}{\overline{O}} - \overline{O},$$

where the O_i are the observed values.

(iii) It has been decided that the test will be done with just 50 rolls of the dice. Identify an appropriate statistic to carry out a two-tailed test on the Null Hypothesis.

(iv) Assuming the value of the level of significance α is 0.05 or less, identify the critical values of the relevant statistic.

e) Answer the following questions.

(i) For a Null Hypothesis based statistical test (NHBST), define the terms Type I error, Type II error and power.

(ii) Let a NHBST 'test on a mean' be carried out with Null Hypothesis $\mu = \mu_0$. Derive an expression for the power of the test in terms of a true mean μ_1. State what this tells about the sample size.

f) Let X and Y be two random variables with means μ_X, μ_Y and standard deviations σ_X, σ_Y respectively. Define the covariance of X and Y as

$$\text{cov}(X, Y) = E[\ (X - \mu_X)(Y - \mu_Y)\ ] = E[\ XY\ ] - \mu_X\mu_Y.$$

The correlation coefficient is then defined by

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Answer the following two questions.

(i) It is required to link the variable Y to X by a linear equation $Y = \alpha + \beta X$. Estimates a and b of the coefficients α and β respectively are given by the equations

$$\mu_Y = a + b\mu_X \text{ and } b = \frac{\rho_{XY}\sigma_Y}{\sigma_X}.$$

Show that these equations arise from minimising the quantity Q, defined by

$$Q = E[(Y - a - bX)^2],$$

with respect to a and b. You may use the simplified form of Q:

$$Q = E[Y^2] - 2a\mu_Y - 2bE[XY] - 2ab\mu_X + b^2E[X^2] + a^2.$$

(ii) Show further that the value of Q at this minimum point is the value

$$E[(Y - a - bX)^2] = \sigma_Y^2(1 - \rho_{XY}^2).$$

g) For the multiple regression model is as stated above, prove the following statements.

(i) $X^T e = \underline{0}$, where $e = Y - Xb$ and $\underline{0}$ represents the p by 1 vector of zeros.

(ii) The vector of residuals e is perpendicular to Xb and so $|Y|^2 = |Xb| + |e|^2$.

(iii) If γ is any p by 1 vector, the function $Q(\gamma) = |Y - X\gamma|^2$ is minimised when $\gamma = b$. [Hint: write $Y - X\gamma$ as $Y - Xb + X(b - \gamma)$.]

(iv) $E(b|X) = \beta$. Explain the importance of this result.

h) The covariance matrix of a vector U of random variables is defined as

$$\text{cov}(U) = E[UU^T] - E[U]E[U^T].$$

For the multiple regression model, answer the following two questions.

(i) Prove that the covariance matrix of the estimator b is given by

$$\text{cov}(b|X) = \sigma^2(X^T X)^{-1}.$$

(ii) Since the variance σ^2 is unknown, identify an unbiased estimator for it and hence a corresponding estimator for $\text{cov}(b|X)$.

i) In the multiple regression model, let q be an integer with $0 < q \leq p$. Under the hypothesis that the last q elements of β are 0, let b_s be the resulting OLS estimate for β. Treating b_s as a p by 1 vector of parameters with the last q elements 0, set F to be the quantity

$$F = \frac{|Xb|^2 - |Xb_s|^2}{q} \Big/ \frac{|e|^2}{n - p}.$$

Answer the following questions.

(i) Define the F-distribution for two degrees of freedom parameters and give a brief explanation of why the quantity shown above follows this distribution.

(ii) This result is typically used for the well-known 'F-test' with $q = p - 1$. Explain the meaning of such a test and how a large value of F should be interpreted.

j) A random sample of 23 components from four different suppliers were tested to destruction, with the time recorded in 10^6 s, shown in the table below. The overall sum of squares is $\Sigma X_i^2 = 139{,}511$. Answer the following questions.

(i) State the Null Hypothesis H_0 and the assumptions that are made for an appropriate test to determine if any supplier is producing longer-lasting components test.

(ii) Explain briefly how this Null hypothesis relates to the standard *F*-test on the regression model. State what conclusions should be drawn if H_0 is rejected.

(iii) Carry out the test using the information given in the table below.

Supplier	Lifetimes:	Total:
A	65, 87, 73, 79, 81, 69	454
B	75, 69, 83, 81, 72, 79, 90	549
C	59, 78, 67, 62, 83, 76	425
D	94, 89, 80, 88	351

k) A trial was conducted to determine the reduction of cholesterol due to four different exercise regimes (Pilates, spinning, circuits or trekking) and three relaxation regimes (yoga, mediation and Tai Chai). Ten subjects were assigned to each combination of exercise and relaxation. All subjects started with approximately equal cholesterol levels, and their differences were recorded at the end of the experiment.

(i) State the Null Hypothesis and the assumptions that are made for an appropriate test to determine whether the difference in cholesterol depends on the four exercise regimes and three relaxation methods.

(ii) The total 'sum of squares' is $\Sigma X_i^2 = 32.653$. Carry out the test using the information given in the table below.

Source	*Sum of squares*

Exercise	2.634
Relaxation	3.539
Interaction effect	6.947

Information

Product rule: if $y = uv$ then $\frac{\mathrm{d}y}{\mathrm{d}x} = u\frac{\mathrm{d}v}{\mathrm{d}x} + v\frac{\mathrm{d}u}{\mathrm{d}x}$.

Quotient rule: if $y = \frac{u}{v}$ then $\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{v\frac{\mathrm{d}u}{\mathrm{d}x} - u\frac{\mathrm{d}v}{\mathrm{d}x}}{v^2}$.

Integration by parts: $\int u\frac{\mathrm{d}v}{\mathrm{d}x}\mathrm{d}x = uv - \int v\frac{\mathrm{d}u}{\mathrm{d}x}\mathrm{d}x$.

The gamma function is defined by $\Gamma(a+1) = \int_0^\infty u^a e^{-u}\,\mathrm{d}u$.

Correlation

For a list of paired values (X_1 , Y_1) to (X_n , Y_n):

The sample correlation coefficient r is: $r = \frac{\sum_i X_i Y_i - n\overline{X}\,\overline{Y}}{\sqrt{\sum_i X_i^{\,2} - n\overline{X}^2}\sqrt{\sum_i Y_i^{\,2} - n\overline{Y}^2}}$.

One-way analysis of Variance

Using the symbol m for the overall mean, and m_j to be the mean for group j:

$$m = \overline{X} = \frac{\sum X_i}{N} \text{ and } m_j = \overline{X}_j = \frac{\sum_{\text{group } j} X_i}{n_j}.$$

The *total sum of squares*, denoted S_1, is: $S_1 = \sum \left(X_i - \overline{X}\right)^2 = \sum (X_i - m)^2 = \sum X_i^{\,2} - Nm^2$.

The sum of squares between the groups, denoted S_T, is $S_T = \sum_j n_j \left(m_j - m\right)^2$.

The sum of squares for the error is $S_E = \sum_{i,j} \left(X_i - m_j\right)^2 = \sum_j \sum_{\text{within } j} X_i - n_j m_j^{\,2}$.

Critical Values for the F-distribution for level of significance 0.05

	Numerator degrees of freedom							
Denominator degrees of freedom.	1	2	3	4	5	6	7	8
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180
45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94