??th December 2016
??.?? – ??.?? ?.m.

**EXAMINATION FOR THE DEGREES OF M.Sc., M.Sci.
AND Ph.D. Integrated Study**

Probability Level M (STATS 5024)

"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

This paper consists of five pages and contains four questions. Candidates should attempt THREE out of FOUR questions. If more than three questions are attempted, please indicate which questions should be marked; otherwise, only the first three questions will be marked.

Question 1	20 marks
Question 2	20 marks
Question 3	20 marks
Question 4	20 marks

The following material is made available to you:

INSERT PICTURE OF FORMULA SHEET HERE!

CONTINUED OVERLEAF/

The standard normal distribution function

When Z is a $\mathrm{N}(0,1)$ random variable, this table gives $\Phi(z) = P(Z \leq z)$ for values of z from 0.00 to 3.67 in steps of 0.01. When $z < 0$, $\Phi(z)$ can be found from this table using the relationship $\Phi(-z) = 1 - \Phi(z)$.

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9999		

CONTINUED OVERLEAF/

1. The continuous random vector $\boldsymbol{X} = (X_1, X_2)$ has probability density function

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = kx_1^2x_2^2, \qquad \boldsymbol{x} = (x_1, x_2) : 0 < x_1 < 1; 0 < x_2 < 1; 0 < x_1 + x_2 < 1.$$

(a) Show that k is equal to $3/\mathbb{B}(3,4)$, where $\mathbb{B}$ is the Beta function. Evaluate k. **[3,1 MARKS]**

(b) Calculate the marginal probability density function of X_1 (being sure to indicate its range space). **[2,1 MARKS]**

(c) Hence derive the conditional density function of X_2 given X_1 (and the corresponding range space). **[2,1 MARKS]**

(d) Compute the conditional expectation of X_2 given X_1 and hence derive $\mathrm{E}(X_2)$. **[3,3 MARKS]**

(e) Write down a double integral expression for $P(X_1 + X_2 \leq y)$ $(0 < y < 1)$ but **do not** evaluate the integral. Instead suppose it evaluates to give the answer y^6. What is the probability density function of the random variable $Y = X_1 + X_2$? **[2,2 MARKS]**

2. (a) Suppose $X \sim \mathrm{Poi}(\lambda)$ for $\lambda > 0$.

i. Show that the moment generating function (m.g.f.) of X is

$$M_X(t) = \mathrm{e}^{\lambda(\mathrm{e}^t - 1)}.$$

[4 MARKS]

ii. For what values of t does the m.g.f. exist? **[1 MARK]**

iii. Use the m.g.f. to derive $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ **[2,2 MARKS]**

(b) Suppose $X_1, \ldots, X_k$ are a collection of independent random variables such that $X_i \sim \mathrm{Poi}(\lambda_i)$ for $i = 1, \ldots, k$, with $\lambda_i > 0$. Let $Y = X_1 + \cdots + X_k$.

i. Derive the moment generating function of Y. Hence identify the distribution of Y. **[3,1 MARKS]**

ii. State the Central Limit Theorem. Use it to approximate the distribution of Y when k is large. **[3,1 MARKS]**

iii. The number of mutations that have occurred at each position in the human genome in the course of human evolution is modelled as a $\mathrm{Poi}(0.1)$ distribution. If we look at 1000 positions, and can assume that they experience mutation

CONTINUED OVERLEAF/

independently, estimate the approximate probability that more than 120 mutations in total have affected these positions.

[3 MARKS]

3. The Olympic triathlon is a sporting event consisting of three elements: swimming, cycling and running. The finishing times for the last two events of the triathlon are to be modelled. Suppose that X_1 is a random variable describing the time it takes for a random athlete to complete the cycling segment and X_2 is a random variable describing the time to complete the run. Let us suppose that X_1 and X_2 are independent, X_1 has a $\mathrm{Ga}(\alpha_1, \theta)$ distribution and X_2 has a $\mathrm{Ga}(\alpha_2, \theta)$ distribution.

   (a) Write down the probability density function of $\boldsymbol{X} = (X_1, X_2)$, and its range space. **[3 MARKS]**

   (b) Suppose we are interested in the joint distribution of the total time spent cycling and running, $Y_1 = X_1 + X_2$ and the fraction of that time that is spent cycling $Y_2 = X_1/(X_1 + X_2)$. Show that the joint probability density of $\boldsymbol{Y} = (Y_1, Y_2)$ is

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) = & \frac{\theta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)} y_1^{\alpha_1+\alpha_2-1} \mathrm{e}^{-\theta y_1} \\ & \times \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y_2^{\alpha_1-1}(1-y_2)^{\alpha_2-1}. \end{aligned}$$

   [10 MARKS]

   (c) What is the range space of $\boldsymbol{Y}$? **[1 MARK]**

   (d) Are Y_1 and Y_2 independent or not? Explain. **[2 MARKS]**

   (e) Identify the marginal distributions of both Y_1 and Y_2. **[2 MARKS]**

   (f) Identify one feature of our probability model for $\boldsymbol{X}$ that seems questionable. **[2 MARKS]**

4. (a) Consider the random vector $\boldsymbol{X} = (X_1, X_2)^{\mathrm{T}}$ with the following distribution:

$$\boldsymbol{X} \sim \mathrm{N}_2\left(\begin{bmatrix}2\\1\end{bmatrix}, \begin{bmatrix}4 & 1\\1 & 1\end{bmatrix}\right).$$

   i. What is the marginal distribution of X_1? **[1 MARK]**

   ii. What is the correlation between X_1 and X_2? **[2 MARKS]**

CONTINUED OVERLEAF/

iii. Suppose $Y_1 = X_1 + X_2$ and $Y_2 = 2X_1 - X_2$. If $\boldsymbol{Y} = (Y_1, Y_2)^{\mathrm{T}}$, what are $\mathrm{E}(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{Y})$? Are Y_1 and Y_2 independent? Justify your answer.

[5,2 MARKS]

(b) Suppose that events $E_1, \ldots, E_k$ partition a sample space S. Let F be any event in S. Prove the Law of Total Probability, which states that:

$$P(F) = \sum_{i=1}^{k} P(F|E_i)P(E_i).$$

[4 MARKS]

(c) A diagnostic test for a genetic disease that appears late in life gives a positive result on those who will actually develop the disease 80% of the time, but also gives a positive result (a false positive) on 20% of those who will not develop the disease. The disease occurs in 1% of the population. A doctor performs the test on me and it is positive. He tells me that since the positive result is four times more likely to occur in those who will develop the disease compared to those who will not that I am four times more likely to develop the disease than not. By calculating the probability of developing the disease given that a positive test has occurred, using Bayes' theorem, or otherwise, discuss the validity of the doctor's argument. **[6 MARKS]**

END OF QUESTION PAPER.