All questions may be attempted but only marks obtained on the best **four** *solutions will count.*
The use of an electronic calculator **is** *permitted in this examination.*

New Cambridge Statistical Tables are provided.
Candidates are reminded of the following series expansions:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} .$$

$$\log(1-x) = -\sum_{k=1}^{\infty} x^k/k \text{ for } |x| < 1.$$

1. (a) Give the definition of the probability generating function, $\Pi(z)$ say, of a random variable X. State, clearly and completely, any restrictions on the types of random variables for which probability generating functions are defined. Show that $\Pi'(1) = \mathrm{E}[X]$ and that $\Pi''(1) = \mathrm{E}[X(X-1)]$, where $\Pi'(\cdot)$ and $\Pi''(\cdot)$ denote the first and second derivatives of the probability generating function.

   (b) A random variable X has probability generating function $\Pi(z) = [(3z+1)/4]^n$, for some positive integer n. Find the probability mass function of X, along with its expectation and variance. Name the distribution of X, giving the values of its parameters.

2. (a) Suppose that X is a continuous random variable with probability density function $f_X(\cdot)$, and that $Y = g(X)$ where $g(\cdot)$ is a differentiable strictly monotonic (increasing or decreasing) function. Show that, for y in the range of $g(\cdot)$, the probability density function of Y can be expressed as

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dx}{dy}\right| ,$$

   where dx/dy is the gradient of the function $x = g^{-1}(y)$ at the point y.
   Explain carefully why $g(\cdot)$ must be strictly monotonic for this result to hold.

   (b) A random variable U is uniformly distributed on $(0,1)$. Find the probability density functions of the following random variables. Where the resulting densities correspond to one of the standard distributions covered during the course, name the distribution and give the values of its parameters:

   (i) $-3\log(1-U)$.

   (ii) $\Phi^{-1}(U) - 7$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

   (iii) U^2.

3. (a) A random variable X has moment generating function $M_X(t)$. If a and b are constants and $Y = aX + b$, show that the moment generating function of Y can be written as $e^{bt}M_X(at)$.

   (b) Show that the moment generating functions of the $N(0,1)$ and $\Gamma(\alpha,\lambda)$ distributions are $M(t) = e^{t^2/2}$ and $M(t) = \lambda^\alpha(\lambda - t)^{-\alpha}$ $(t < \lambda)$, respectively.

   (c) Suppose that $X_1, \ldots, X_n$ are independent, identically distributed random variables, each distributed as $\Gamma(\alpha,\lambda)$. Let $Z_n = (n\alpha)^{-1/2}\sum_{i=1}^n (\lambda X_i - \alpha)$ and let $M_n(\cdot)$ denote the moment generating function of Z_n. Use the results from part (a) to find an exact expression for $M_n(t)$. By considering the series expansion of $\log M_n(t)$, or otherwise, show that the distribution of Z_n tends to $N(0,1)$ as $n \to \infty$.

4. (a) X is a random variable taking values 0 and 1 with probabilities $(1-p)$ and p respectively. Y is another random variable taking values 0 and 1. If $X = 0$ then Y takes the value 1 with probability q_0; if $X = 1$, Y takes the value 1 with probability q_1.

   (i) Produce a table showing the joint probability mass function of X and Y.

   (ii) Name the marginal distribution of Y, giving the value(s) of its parameter(s).

   (b) Give a clear statement of the Iterated Expectation Law, taking care to define or explain any notation that you use. Verify that this law holds for $E(Y)$ in the joint distribution from part (a).

   (c) In a certain part of the world, if it rains on a particular day (i.e. it is 'wet') then the following day is also wet with probability 2/3, and dry otherwise. Conversely, if a particular day is dry then the following day is wet with probability 1/4 and dry otherwise. Consider a sequence of days labelled $1, 2, \ldots$, and let X_t be a random variable taking the values 0 and 1 if day t is respectively dry and wet.

   (i) If day 0 is dry, write down the joint distribution of X_1 and X_2 and give their marginal distributions.

   (ii) Again if day 0 is dry, find the marginal distribution of X_3.

   (iii) Suppose that day 1 is wet with probability p. What value of p ensures that X_1 and X_2 have the same marginal distributions?

5. (a) Suppose that $X_1, \ldots X_n$ are independent, identically distributed (IID) random variables from some distribution with an unknown parameter θ, and consider a statistic $T_n = T(X_1, \ldots, X_n)$ which is an estimator of θ.

   (i) Denote the bias, variance and mean squared error of T_n by $b(T_n)$, $\mathrm{Var}(T_n)$ and $\mathrm{MSE}(T_n)$ respectively. Show that

   $$\mathrm{MSE}(T_n) = \mathrm{Var}(T_n) + b^2(T_n) \ .$$

   (ii) What is meant by saying that T_n is a consistent estimator of θ?

   (b) $X_1, \ldots, X_n$ are IID random variables from some distribution with mean μ and variance σ^2. $Y_1, \ldots, Y_n$ are IID random variables from another distribution with mean 2μ and variance $2\sigma^2$; the Xs and Ys are independent of each other. Let $\overline{X}_n$ and $\overline{Y}_n$ denote the sample means from the two samples.

   (i) Write down expressions for the expected value and variance of $\overline{X}_n$ and $\overline{Y}_n$.

   (ii) It is proposed to combine the information in the two samples to obtain an estimator of μ, of the form

   $$T_n = a\overline{X}_n + b\overline{Y}_n \ ,$$

   where a and b are constants to be determined. Show that for T_n to be an unbiased estimator of μ, we must have $a = 1 - 2b$. In this case, find the value of b for which the variance of T_n is minimised.

   (iii) Explain carefully why it is of interest to minimise the variance of T_n in part (ii) above.

   (iv) An alternative way to estimate μ from the same number of observations is to take a single sample $X_1, \ldots, X_{2n}$ of $2n$ observations from the first distribution above, and to estimate μ as the sample mean $\overline{X}_{2n}$. Would this be better or worse than the unbiased estimator with minimum variance from part (ii)?

6. (a) Consider the theoretical derivation of the two-sample t-test. What *specific* aspect of this derivation breaks down if the data are paired? Typically, how would you expect the test results to be affected in practice if a two-sample test is carried out when the data are clearly paired? Explain your reasoning clearly.

   (b) It is sometimes claimed that boys and girls have different levels of academic attainment at school. To test this theory, the high school maths exam results of 25 boys and 16 girls were analysed. Suppose that the boys' marks (as percentage scores) are mutually independent, and all drawn from a normal distribution with mean μ_b and variance σ_b^2. Similarly, suppose that the girls' marks are drawn independently from a normal distribution with mean μ_g and variance σ_g^2. The summary statistics for the observed values of the marks are as follows:

   Boys: sample mean $\bar{x}_b = 72$; sample standard deviation $s_b = 8$

   Girls: sample mean $\bar{x}_g = 68$; sample standard deviation $s_g = 7$

   (i) Show that, at the 5% level, these data are consistent with the hypothesis that $\sigma_b^2 = \sigma_g^2$.

   (ii) Test, at the 5% level, the hypothesis that in fact there is *no difference* between boys and girls in terms of their expected maths exam performance. State your conclusions clearly.