INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

19th November 2012

Subject CT3 – Probability & Mathematical Statistics

Time allowed: Three Hours (15.00 – 18.00)

Total Marks: 100

INSTRUCTIONS TO THE CANDIDATES

1. **Please read the instructions on the front page of answer booklet and instructions to examinees sent along with hall ticket carefully and follow without exception**
2. **Mark allocations are shown in brackets.**
3. **Attempt all questions, beginning your answer to each question on a separate sheet. However, answers to objective type questions could be written on the same sheet.**
4. **Please check if you have received complete Question Paper and no page is missing. If so, kindly get new set of Question Paper from the Invigilator.**

AT THE END OF THE EXAMINATION

Please return your answer book and this question paper to the supervisor separately.

Q. 1) The probability model for the distribution of claims per policy for an Accidental Disability and Dismemberment Health plan are given in the table below. Independence of claim amounts and occurrences among claim types and claimants is assumed.

Claim Type	**Probability of Claim**	**Distribution of claim amounts given that a claim occurs**	
		Mean	**Standard Deviation**
Disability	0.35	50,000	10,000
Dismemberment	0.65	30,000	7,000

Calculate the mean and the standard deviation of the claim amounts per policy.

[6]

Q. 2) A measure of skewness is defined as:

$$\psi = \frac{mean - mode}{standard\ deviation}.$$

Find the value of ψ for a gamma distribution with parameters $\alpha = 2.5$ and $\lambda = 0.4$.

[4]

Q. 3) A random sample of basic monthly salary (x) of 12 employees of a multinational organization has been selected.

The summary of the data is as follows:

- $\sum x = 213{,}200$;
- $\sum x^2 = 4{,}919{,}860{,}000$;
- Sample Mean = 17,767 ;
- Sample Standard Deviation = 10,144.

However later it was identified that two of these selected employees were temporary employees and needs to be replaced with two permanent employees.

Calculate the sample mean and standard deviation of the revised sample if the salary of the two temporary employees is 11,000 and 48,000 and salary of the two randomly selected permanent employees is 27,500 and 31,500. (*You should show intermediate workings*)

Comment on your answers.

[7]

Q.4) An auto insurance company charges younger drivers a higher premium than it does older drivers because younger drivers as a group tend to have more accidents.

The company analyses its experience data using 3 age groups: Group A includes those less than 25 years old, 22% of all its policyholders. Group B includes those 25-39 years old, 43% of all its policyholders, Group C includes those 40 years old and older.

The probabilities of a claim due to an accident in any one year period for a policyholder belonging to Group A, B or C are 11%, 3% and 2%, respectively.

[i] What percentage of the company's policyholders is expected to make a claim due to an accident during the next 12 months? [2]

[ii] Suppose Mr X has just had a car accident and goes on to register a claim against his insurance policy. If he is one of the company's policyholders, what is the probability that he is under 25? [3]

[5]

Q.5) **[i]** Let the moment generating function (MGF) of a random variable X be$M_X(t)$.
Show that:

(a) $E(X) = M'_X(0)$
(b) $Var(X) = M''_X(0) - \ [M'_X(0)]^2$.

[3]

[ii] Let Z be the sum of two independent random variables X_1 and X_2. Show that the MGF of Z is the product of the MGFs of X_1 and X_2.

A company insures homes in three cities, J, K and L. The losses occurring in these cities are independent. The moment-generating functions for the loss distributions of the cities are:

- $M_J(t) = \ (1-2t)^{-3}$;
- $M_K(t) = \ (1-2t)^{-2.5}$;
- $M_L(t) = \ (1-2t)^{-4.5}$

Let Y represent the combined losses from the three cities.

[2]

[iii] Compute E(Y) and Var(Y).

[5]

[10]

Q.6) Let X_1, X_2 … X_n be a random sample from Uniform distribution over (0, θ), where θ is an unknown parameter (> 0).

[i] Outline why the Cramér-Rao lower bound for the variance of unbiased estimators of θ does not apply in this case.

Consider an estimator of θ: $\hat{\theta}(c) = c\,Y$ for some constant c where $Y = \max_i X_i$

[1]

[ii] Show that the probability density function of Y is given as:

$$g_Y(y) = \frac{n}{\theta^n}.y^{n-1}\ for\ 0 < y < \theta$$

Hence, show that:

$$E[Y^k] = \frac{n\,\theta^k}{n+k}\text{for any non negative real number k}$$

[5]

[iii] Show that the bias and mean square error (MSE) of the estimator$\hat{\theta}$(c) are given as follows:

$$Bias[\hat{\theta}(\mathrm{c})] = \left(\frac{c\,n}{n+1} - 1\right).\theta$$

$$MSE[\hat{\theta}(\mathrm{c})] = c^2.\frac{n\,\theta^2}{n+2} - c.\frac{2n\,\theta^2}{n+1} + \theta^2$$

[4]

[iv] Find the value of c (=c_u) for which $\hat{\theta}(c)$becomes an unbiased estimator of θ.

[1]

[v] Find the value of c(=c_m) for which the mean square error of $\hat{\theta}(c)$is minimised.

[2]

[vi] Which of the two estimators $\hat{\theta}(c_u)$ or$\hat{\theta}(c_m)$ will you prefer for estimating θ? Give reasons. What happens when n is large? [2]

[15]

Q.7) In a pre-match program, 280 callers (random callers) were asked which team they think will win the upcoming cricket match between Australia and England.

If 40% of the population to which these callers belong thinks that team England will be the winner, calculate (approximately) the probability that at least 106 callers think that England will win. **[3]**

Q.8) The number of breakages in a damaged chromosome X follows a discrete distribution with an unknown parameter λ (> 0). The probability mass function of X is given by:

$$P(X = k) = \frac{e^{-\lambda}}{1 - e^{-\lambda}} \cdot \frac{\lambda^k}{k!}, \quad k = 1, 2, \ldots$$

A random sample of 33 damaged chromosomes was studied and the number of breakages in each chromosome was recorded in the form of a frequency table as follows:

Number of Breakages	Number of Chromosomes
1	11
2	6
3	4
4	5
5	0
6	1
7	0
8	2
9	1
10	0
11	1
12	1
13	1
14+	0

[i] Find an equation satisfied by $\hat{\lambda}$, the maximum likelihood estimator of λ. [5]

[ii] It is found using the observed data, the maximum likelihood estimate of λ is$\hat{\lambda} = 3.6$. Using this value of $\hat{\lambda}$, test the null hypothesis that the number of breakages in a damaged chromosome follows the given probability distribution. [7]

[12]

Q.9) The following data (x) are the number of germinations per square foot observed in an experiment where a particular type of plant seed was applied at four different rates.

	Rate of Application			
	1	**2**	**3**	**4**
	29	180	332	910
	13	90	444	880
	21	120	190	460
Mean	21	130	322	750
Variance	64	2,100	16,204	63,300

[i] State the assumptions required for a one-way analysis of variance (ANOVA) and whether these data appear to violate any of them. Give reasons.

[2]

[ii] An agricultural scientist working on this data applied the following three transformations to the data:

$\sqrt{x}$, $\log_e(x)$ and $(1/x)$.

The following table contains values for the means and variances of the transformed data although he forgot to fill in a couple of them and thus is missing (the ones marked as ****):

Rate	**Transformation**					
	$\sqrt{x}$		$\log_e(x)$		$1/x$	
	mean	**variance**	**mean**	**variance**	**mean**	**variance**
1	****	0.79	2.992	0.1630	0.05301	0.0004721
2	11.29	3.94	4.827	****	0.00833	0.0000077
3	17.69	13.49	5.716	0.1861	0.00351	0.0000025
4	27.09	23.96	6.575	0.1479	0.00147	0.0000004

Complete the table for him by determining the missing values.

[3]

[iii] The scientist insisted (in order to perform a one-way ANOVA) we must consider the $\log_e$ transformation of the data only. Argue heuristically why this is the case?

[2]

[iv] The scientist now decides to go ahead and performs the one-way ANOVA.

(a) State the ANOVA model for the transformed data along with the null hypothesis he is testing.
(b) Perform the ANOVA for the transformation chosen.
(c) What conclusions can be drawn from the ANOVA?

[8]

[v] Calculate the 95% confidence intervals for the mean values of each of the rates on the chosen transformed scale. Back-transform these intervals onto the original scale.

[7]

[22]

Q.10) For a study into the density of population around a large city, a random sample of 10 residential areas was selected. For each area the distance (x) in kilometres from the city centre and the population density in hundreds per square kilometre (y) were recorded.

The following table shows the data and also the $\log_e$ of each measurement:

	x	**y**	**$\log_e$(x)**	**$\log_e$(y)**
	0.4	149	-0.916	5.004
	1.0	141	0.000	4.949
	3.1	102	1.131	4.625
	4.5	46	1.504	3.829
	4.7	72	1.548	4.277
	6.5	40	1.872	3.689
	7.3	23	1.988	3.135
	8.2	15	2.104	2.708
	9.7	7	2.272	1.946
	11.7	5	2.460	1.609
Mean	5.710	60.000	1.396	3.577
Variance	13.425	2912.667	1.153	1.451

Three different regression models were considered for analysing this data.

- $Model\ 1\text{:}\ \mathbb{E}[\,Y\mid x\,] = \alpha + \beta x$
- $Model\ 2\text{:}\ \mathbb{E}[\,Y\mid log_e\, x] = \alpha + \beta\, log_e\, x$
- $Model\ 3\text{:}\ \mathbb{E}[log_e\, Y\mid x\,] = \alpha + \beta x$

Here $\mathbb{E}[.]$ stands for the Expectation of the response variable.

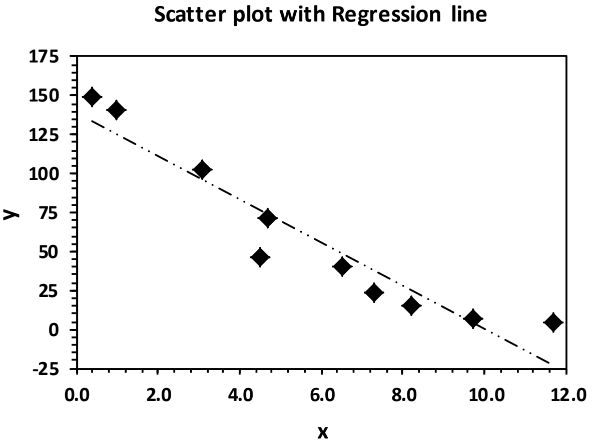
The following is the extract from the computer output on the analysis done for Model 1:

The regression equation: $y = *** - 14.0\,x$

```
Parameter   Estimate      s.e
    α          ***       11.120
    β       -13.958        ***

Parameter      t        p-value
    α        12.56       0.000
    β         ***        0.000
```

$\hat{\sigma} = 18.2834$ $R^2 =$ ***



[“s.e”: standard error; “t”: test statistic value for testing the significance of the coefficient]

[i] In the above output some of the entries have gone missing (the ones marked as ***). Making any necessary assumptions, use the available information to compute these missing entries and perform a statistical test for significance of β at the 5% level.

The following are similar extracts from the computer output on the analysis done for Models 2 and 3:

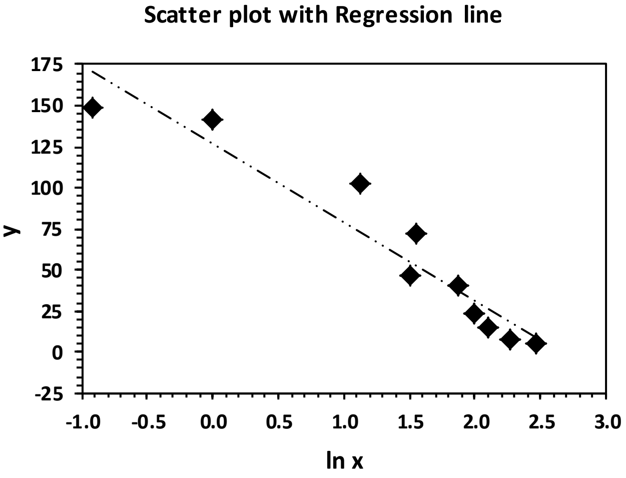
Regression Analysis: Model 2

The regression equation: $y = 127.0 - 48.0\,log_e x$

```
Parameter   Estimate      s.e
    α        126.990     9.147
    β        -47.980     5.293

Parameter      t        p-value
    α        13.88       0.000
    β        -9.07       0.000
```

$\hat{\sigma} = 17.0492$ $R^2 = 91.1\ \%$



[6]

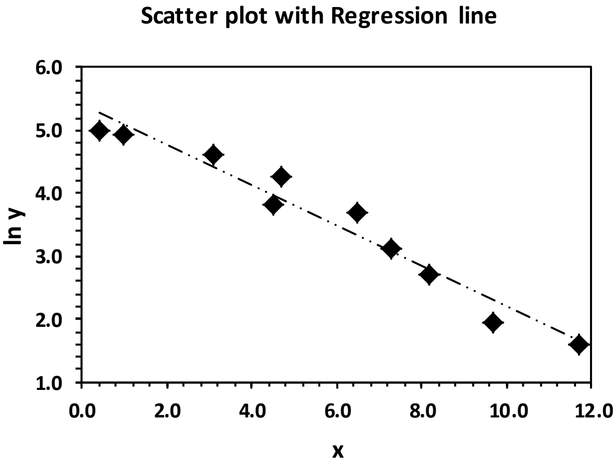
Regression Analysis: Model 3

The regression equation: $\boldsymbol{log_e y = 5.4 - 0.3\,x}$

Parameter	Estimate	s.e
α	5.413	0.162
β	-0.322	0.024

Parameter	t	p-value
α	33.40	0.000
β	-13.26	0.000

$\hat{\sigma} = 0.266544$ $R^2 = 95.6\ \%$



[ii] On the basis of the regression results given above, which of the three regression models would you consider to be the best? Justify your answer by reference to the diagnostic criteria given in the output and relating those to the corresponding plots.

[5]

[iii] For the model you consider to be the '**best**' in part ii, write down an expression for y in terms of x.

[1]

[iv] Using the chosen model, estimate the density of the population at a distance of 5 km from the city centre.

[2]

[v] State any reservations you have about using the model to predict population density.

[2]

[16]

****