

# **INSTITUTE AND FACULTY OF ACTUARIES**

## **EXAMINATION**

15 September 2022 (am)

### **Subject CS1 – Actuarial Statistics Core Principles**

#### **Paper A**

Time allowed: Three hours and twenty minutes

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator.

If you encounter any issues during the examination please contact the Assessment Team on T. 0044 (0) 1865 268 873.

- 1** From national statistics, it is known that 7% of all drivers in a country are young drivers. It is also known that 18% of all drivers involved in road accidents are young drivers (less than 25 years old). Define the two events:

$A$ : a randomly chosen driver is involved in a road accident.

$Y$ : a randomly chosen driver is a young driver.

- (i) Determine the conditional probability  $P[A | Y]$  as a function of  $P[A]$ . [2]
- (ii) Comment on the result from part (i). [1]
- [Total 3]

- 2** A warranty is provided for a product worth £10,000 such that the buyer is given £8,000 if it fails in the first year, £6,000 if it fails in the second year, £4,000 if it fails in the third year, £2,000 if it fails in the fourth year and zero after that. The probability of failure in a year is 0.1. Payments are only received at the first failure.

The random variable  $X$  is the number of years before the first failure occurs.

- (i) Determine the distribution of the random variable  $X$ , including the value of the parameter of interest, justifying your answer and stating any assumptions. [2]

The random variable representing the payment under the warranty is denoted by  $Y$ .

- (ii) Calculate the following probabilities:

(a)  $P(Y = 8,000)$

(b)  $P(Y = 6,000)$

(c)  $P(Y = 4,000)$

(d)  $P(Y = 2,000)$

(e)  $P(Y = 0)$ .

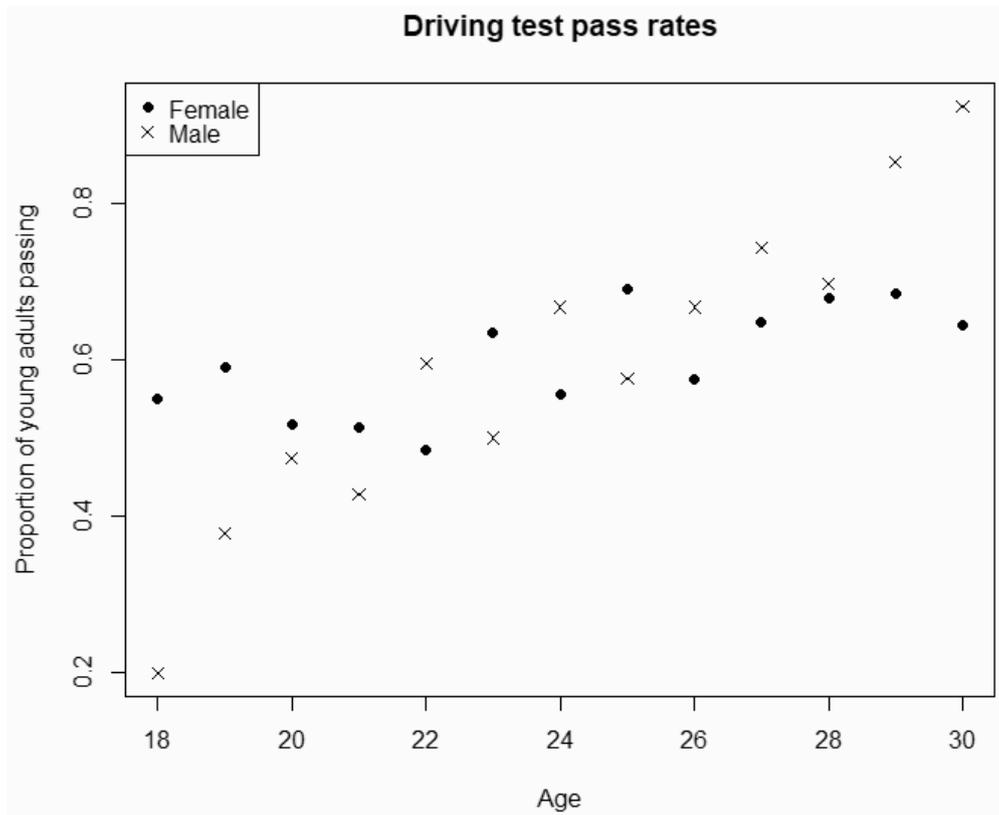
[3]

- (iii) Calculate the expected value of the warranty payment ( $Y$ ) using your answer to part (ii). [2]

[Total 7]

- 3 A study is undertaken in order to devise a model to predict the probabilities of young adults passing a driving test. The data was collected on the basis of results over a 30-day period. An Analyst's observations for any given gender and age group are of the form  $Y/n$ , where  $Y$  is the number passing the test and  $n$  is the number taking the test.

The Analyst plots the proportion of young adults passing by age for males and females as shown below.



- (i) Comment on the graph. [1]

The Analyst believes that age and gender are variables that influence whether or not a person will pass a driving test. The Analyst fitted a Generalised Linear Model (GLM), with a canonical link function, to investigate such an influence by including the interaction term between the two explanatory variables.

- (ii) Write down a suitable model for the proportion passing the test. [3]

The summary of the fitted model is provided in the form of linear predictors for females (F) and males (M) respectively as:

$$\hat{\eta}_F = -0.968 + (0.056) \times \text{Age} \text{ and } \hat{\eta}_M = -4.584 + (0.209) \times \text{Age}$$

- (iii) Determine the proportion of 22-year-old females predicted by the model to pass the test. [2]

Using the fitted GLM model, the Analyst derives the following expression for the ratio of the probability of passing the test ( $\mu$ ) over the probability of failing ( $1 - \mu$ ) for males:

$$\frac{\hat{\mu}}{1 - \hat{\mu}} = \exp(\hat{\eta}_M) = \exp(-4.584 + 0.209 \times \text{Age})$$

(iv) Comment on this expression with respect to the probability of passing the test.

[1]

[Total 7]

**4** Consider two discrete random variables,  $X$  and  $Y$ , with the joint probability function given by:

	$X=1$	$X=2$	$X=3$
$Y=0$	$P[X=1, Y=0] = 0.3$	$P[X=2, Y=0] = 0.1$	$P[X=3, Y=0] = 0.3$
$Y=1$	$P[X=1, Y=1] = 0.05$	$P[X=2, Y=1] = 0.2$	$P[X=3, Y=1] = 0.05$

(i) Verify that the table above specifies a joint distribution of two discrete random variables. [2]

(ii) Determine the expected value of  $X$ . [3]

(iii) Show that  $X$  and  $Y$  are not independent. [2]

[Total 7]

**5** The claim amounts in an insurance company's car insurance portfolio follow a gamma distribution. The company is modelling the claims it receives and is considering a Generalised Linear Model (GLM), with claim amounts as the response variable and four relevant covariates:

- The age ( $x$ ) of the policyholder
- The experience of the policyholder (a category between 1 and 4, based on the number of years of driving experience)
- The gender of the policyholder (1 = male, 2 = female)
- The car insurance group (a rating between 1 and 20, indicating the level of risk).

(i) State the form of the linear predictor of the GLM when all the covariates are included in the model as main effects. [1]

(ii) Explain all the terms used in the linear predictor in your answer to part (i). [2]

(iii) State how the linear predictor in your answer to part (i) changes if an interaction between the covariates showing policyholder age and car insurance group is also included in the model.

You should explain all the terms used in the new linear predictor. [2]

The company is considering whether to include the interaction term between policyholder age and car insurance group. The scaled deviance of the GLM without the interaction term in part (i) has been calculated as 422.5. For the GLM including the interaction in part (iii), the scaled deviance is equal to 310.3.

(iv) Compare the two models by performing a suitable test for investigating whether the model including the interaction term is a significant improvement over the model without the interaction term. [3]

[Total 8]

**6** Let  $x_1, x_2, \dots, x_n$  independent observations from a Bernoulli distribution with  $P(X_i = 1) = p, i = 1, \dots, n$ . The parameter  $p$  has a beta prior distribution with parameters  $(a, b)$ .

(i) Determine the posterior distribution of parameter  $p$ . [6]

(ii) Determine the Bayesian estimate of parameter  $p$  under quadratic loss. [1]

(iii) Determine the Bayesian estimate of parameter  $p$  under quadratic loss as a credibility estimate, stating the credibility factor. [2]

[Total 9]

- 7 The prevalence of an infectious disease at a particular time during an epidemic, in a certain large population, is assumed to be  $r$ . This means that the probability that a randomly selected person from the population has the disease is  $r$ . A test for detecting the presence of the disease is available. The following events are defined:

$T$ : the test returns a **positive** result, i.e. it indicates that an individual has the disease.

$N$ : the test returns a **negative** result, i.e. it indicates that an individual does not have the disease.

$D$ : an individual in the population has the disease.

$H$ : an individual in the population does not have the disease.

The test is 100% accurate when used on people who do not have the disease, i.e.

$P(T|H) = 0$ . However, the test is imperfect when used on people who have the disease, with  $p = P(T|D)$ , where  $0 < p < 1$ .

- (i) Determine the probability  $P(N)$ . [4]

An individual from this population was tested for the disease using this test. The test gave a **negative result**.

- (ii) Determine the conditional probability  $P(D|N)$  that this particular individual has the disease, given that the test gave a **negative result**, in terms of  $p$  and  $r$ . [2]

Consider now that another individual was tested  $k$  times using this test, and all  $k$  tests gave **negative results**. We denote this event, for a randomly selected individual, by  $N^*$ . (You can assume that the outcome of each test, conditional on disease status, is independent of the outcome of all other tests.)

- (iii) (a) Determine the conditional probability  $P(D|N^*)$  that this individual has the disease, given that all  $k$  tests gave negative results, in terms of  $p$  and  $r$ . [5]

- (b) Comment on the impact of an increasing number of negative test results on the probability that an individual has the disease based on your answer to part (iii)(a). [1]

- (iv) State one other assumption that needs to be made to determine the conditional probability in part (iii)(a), with a brief comment on its validity. [2]

[Total 14]

- 8** Following a recent Climate Action Plan (CAP) conference, a particular model was agreed for measuring global earth tremors. A series of  $n$  positive measurements are to be taken, which are assumed to be independent observations of a random variable that are Uniformly distributed on  $(0, \nu)$ , where  $\nu > 0$ .

A Climate Risk Actuary adopts the model agreed by the CAP conference. The Actuary knows only that the number,  $R$ , of the measurements that are less than 1 is  $r$ , with the remaining  $n - r$  being greater than 1.

- (i) (a) Show that the probability for a single measurement to be less than 1 is  $\frac{1}{\nu}$ . [2]

- (b) Show that the maximum likelihood estimate of  $\nu$  is  $\hat{\nu} = \frac{n}{r}$ . [5]

- (c) Identify which **one** of the following expressions gives the Cramer–Rao lower bound for estimating  $\nu$ .

A  $\frac{\nu(1-\nu)^2}{n}$

B  $\frac{\nu^2(1-\nu)^2}{n}$

C  $\frac{(\nu-1)}{n}$

D  $\frac{\nu^2(\nu-1)}{n}$

[3]

- (d) Write down the asymptotic distribution of  $\hat{\nu}$ , using your answer from part (i)(c). [2]

In a random sample of 500 measurements, exactly 75 measurements are less than 1.

- (ii) (a) Calculate an estimate for the standard error of  $\hat{\nu}$ . [2]

- (b) Determine an approximate two-sided 99% confidence interval for  $\nu$ . [2]

- (c) Perform a test for the following hypotheses, using your asymptotic distribution of  $\hat{\nu}$  from part (i)(d):

$$H_0: \nu = 10 \text{ vs } H_A: \nu < 10$$

[5]

[Total 21]



- (vi) Determine the 95% confidence interval for the expected price of a property with 1,930 square feet of living space. [3]
- (vii) Determine the 95% prediction interval for the price of a property with 1,930 square feet of living space. [3]
- (viii) Comment on your answer to parts (vi) and (vii). [2]

The Banking Analyst fitted another least squares regression line for the price of the properties, depending on the square feet of living space and also the year the property was built. The coefficient of determination for this regression line is  $R^2 = 60\%$ .

- (ix) Comment on the result from this second regression line and your answer to part (iii). [2]
- [Total 24]

**END OF PAPER**