Mathematics, Computing and Technology
M140 Introducing statistics

M140

Written EMA 2013J

Covers the whole module. **Cut-off date: 29 May 2014**

The end-of-module assessment (EMA) consists of **two parts**: this written EMA and the final iCMA (called 'iCMA 44').

Before you begin work on this written EMA, please read the instructions on the M140 website for preparing and submitting the EMA. Note in particular:

- Your solutions to this part of the EMA should *not* be sent to your tutor, but to the address given in a letter from the Project Portfolios and Dissertation Office, at the Open University.
- Your tutor is not permitted to grant an extension for your EMA.

This part of the EMA is marked out of 70. The marks allocated to each part of each question are indicated in brackets in the margin.

The Minitab files you require for this part of the EMA should be downloaded from the M140 website.

Unless otherwise specified in the question, it is up to you whether you use Minitab to arrive at your answer.

Remember that this is just one part of the EMA and you must submit **both parts** of the EMA by the respective cut-off dates.

WEB 03472 8
1.1

This assignment covers *the whole module.*

In this written EMA you are going to analyse some data from an orienteering competition ('event') which was held in 2013.

Orienteering is a sport where participants aim to navigate their way round a series of points in the shortest time possible. The location of the points are given on a map, so doing well depends on map-reading and compass skills as well as running ability.

The data you are going to analyse are given in the file **orienteering.mtw**. In this file there are the following variables.

- **id**: a serial number which identifies each individual participant ('orienteer')
- **gender**: '1' – male, '2' – female'
- **age**: '1' – 21 to 44 years old, '2' – 45 to 64 years old, '3' – 65 or more years old
- **actual**: the actual time (in minutes) the orienteer took to complete the course
- **model**: the time the orienteer was predicted to complete the course (in minutes) based on a statistical model.
- **pace**: the pace at which the orienteer was estimated to have been travelling around the course, in minutes per kilometre.

Question 1 – 3 marks

The fastest orienteer is the one whose pace was smaller than everyone else's. By looking at the data file **orienteering.mtw** answer the following.

(a) What is the serial number of the fastest orienteer? [*1*]

(b) Was the fastest orienteer a man or a woman? Which age group were they in? [*2*]

Question 2 – 13 marks

A contingency table of the age and gender of the orienteers is given below.

	Gender		
	Male	Female	Total
21 to 44 years old	23	14	37
45 to 64 years old	74	37	111
65 or more years old	29	10	39
Total	126	61	187

Note that the file **orienteering demographics.mtw** gives this table.

(a) Explain carefully why this table is correctly described as a contingency table. [*3*]

(b) What is the probability that a randomly selected orienteer at this event was male? Give your answer to three decimal places. [*1*]

(c) What is the probability that a randomly selected female orienteer at this event was aged between 21 and 44 years old? Give your answer to three decimal places. [*1*]

(d) Suppose that a researcher is interested in the following question:

Is the distribution of age different for male and female orienteers?

(i) Write down suitable null and alternative hypotheses. [*2*]

(ii) Using the χ^2 test for contingency tables, test the hypotheses that you wrote down in part (i). Make sure that you include the following in your answer:

- a table of the expected values (with the expected values given to two decimal places)
- the value of the test statistic
- the degrees of freedom
- the *p*-value, or values of CV5 and CV1
- your conclusion from the test. [*6*]

Question 3 – 5 marks

(a) Using Minitab, produce a stemplot of the orienteers' paces. Include a copy of this stemplot in your answer. (When producing the stemplot, you should leave the **Increment** field blank and the **Trim outliers** option unselected, and make sure that the stemplot includes all the orienteers.) [*1*]

(b) Use your stemplot to describe the shape of the distribution of the paces. Justify your answers. [*4*]

Question 4 – 19 marks

(a) Using Minitab, or otherwise, produce a diagram containing boxplots of the actual times the male participants took and the actual times the female participants took. (These times are given in separate columns in the file **times.mtw**.)

You should ensure that the boxplots are horizontal and drawn on the same scale. You should prepare these boxplots ready for inclusion in a report by ensuring that the title and horizontal-axis label are clear and informative. Include the finished boxplots in your answer. *[5]*

(b) Complete the following table of summary statistics. (The number of orienteers should be exact. All the other values should be rounded to two decimal places.) *[2]*

	Male orienteers	Female orienteers
Number		
Mean		
Median		
Standard deviation		
Interquartile range		
Range		

(c) Using your answers to parts (a) and (b), informally compare the distribution of times that the male and female orienteers took. *[3]*

(d) Based on previous experience with similar events, an orienteer claims that male and female orienteers have the same mean time. Write down suitable null and alternative hypotheses for use in a statistical test of this claim, stating clearly the meaning of any symbols you use. *[2]*

(e) By hand, use the two-sample z-test to test the hypotheses you wrote down in part (d). Show your workings. *[7]*

Question 5 – 11 marks

A researcher is interested in comparing the actual times taken by orienteers in the sample with the times predicted by the statistical model, as recorded in the **model** column of the file **orienteering.mtw**.

(a) Using the data in **orienteering.mtw**, obtain a scatterplot of the actual time the orienteer took (on the vertical axis) against the time predicted by a model (on the horizontal axis). Include this scatterplot in your answer. Interpret this scatterplot, giving reasons for your interpretation. *[9]*

(b) The correlation coefficient for these two variables is 0.883. With reference to the scatterplot you produced in part (a), state whether this correlation coefficient provides a good representation of the strength of the relationship between the actual times and those predicted by the model. Justify your answer. *[2]*

Question 6 – 4 marks

Occasionally orienteers injure their ankle (not necessarily whilst orienteering). Suppose that a physiotherapist has come up with a new way of treating a severely sprained ankle and wishes to investigate whether this new treatment is more successful than a standard treatment.

(a) The physiotherapist is most interested in the following outcome to a trial: time (in days) before being able to compete again fully. Explain why such data arising out of this trial could be considered to be interval scale data. *[1]*

(b) Give a reason why a crossover design is not suitable for this trial. *[1]*

(c) Suppose a matched-pairs design is used for the trial. Suggest a reasonable hypothesis test that could be used to analyse the data from such a trial. Give a reason why this test might not turn out to be suitable. *[2]*

Question 7 – 6 marks

Suppose that the physiotherapist in Question 6 is also interested in the current physical fitness of orienteers. He intends on investigating this by assessing the fitness of a sample of a few orienteers.

(a) Suppose that the physiotherapist decides that he can afford to assess approximately one eighteenth of the orienteers who were at the event.

Select by hand a suitable sample for this physiotherapist, using *systematic random sampling.* Any random numbers you require should be obtained from the random number table given in the appendix to Unit 4, starting at row **61**. Show your workings. *[4]*

(b) At the event, each orienteer completed one of seven different courses. Before the **id** numbers for the orienteers were allocated sequentially, the orienteers were grouped by which course they did.

Explain why the sample you obtained in part (a) is similar to a stratified sample. In what way is it different? *[2]*

Question 8 – 9 marks

For the data in the **pace** column of **orienteering.mtw**, the sample mean is 12.432 min/km and the sample standard deviation is 4.775 min/km.

(a) Calculate by hand the 95% confidence interval for the population mean pace. Show your workings, and round each end of the interval to two decimal places. *[5]*

(b) Interpret the interval you calculated in part (a). *[2]*

(c) For the calculation in part (a), why does it not matter whether the distribution of pace is or isn't normal? *[2]*

Remember that this is just one part of the EMA and you must submit both parts of the EMA by the respective cut-off dates.