Informatics 1: Data & Analysis
Session 2013–2014

Exam Feedback

This is a feedback report on the end-of-year examination for the course *Informatics 1: Data & Analysis*, held on Friday 16 May 2014. It reviews the exam itself and gives feedback on the solutions submitted by the students who took the exam. Please note the following:

- This is not a set of "model" answers. It does contain solutions, which can be used to check your own answers; but there are also discussions of different possible answers, key points, possible errors, and comments on the ways people approached each question on the day.
- Not all the questions have a single "right" answer. There can be multiple correct ways to write a database query, explain a concept, or construct an example. This report includes some variants on answers, but still cannot cover every possible correct alternative.
- Studying past papers is one way to learn more about a subject, but it is quite limited and not enough on its own. Even when an exams routinely follow a fixed structure, the questions change and successful performance does essentially depend on a good understanding of the material in the course.

The exam consisted of three questions, each with several subquestions. The rest of this report is a one-page summary and then the full text of each question followed by notes on solution and feedback on the answers given.

Where you find errors in these notes, please send them to me at Ian.Stark@ed.ac.uk

Ian Stark
2014-08-07

Summary

The standard of performance in this exam was generally very high, with many students submitting excellent answers. Even those students who obtained a lower mark overall often did good work one or two individual question. There were many examples of really clear, precise and straightforward solutions to the problems set out.

Some answers and attempts at answers were less confident, with one particular weakness being a lack of precision: writing about something in the area of the question, but not actually answering it; or making a general statement instead of giving an exact answer. Many solutions could have been improved by application of standard exam technique: pay close attention to exactly what the question asks for, and give a precise answer to that.

In the main, though, the distinguishing feature of high-graded submissions was mastery of the material: evidence of detailed knowledge and a good understanding, enough to clearly apply that knowledge to the scenarios provided.

Question 1.

This was the most fully-attempted question of all, with almost everyone making fair progress in their answer. Identifying the components of the entity-relationship diagram was very well done; although many people were then unsure about the operation of a weak entity and its identifying relationship.

A lot of answers were rather too brief — for example, just naming a constraint but not explaining its effect. A key piece of advice here is the mundane but always-relevant "Read The Question": then make sure you give an answer that matches what it asks for. It's also important not only to know the textbook definition of a concept, but also to be able to explain how it works in a specific example.

The final SQL data declaration required an overall understanding of the situation provided in the ER diagram. While many gave a general presentation of the elements required, only a small number of students succeeding in capturing all aspects and details of the scenario.

Question 2.

There were many, many different drawings of the XML tree here, with quite a few getting this correct but also several opportunities for error. There were two rather different levels of challenge in this question: the basic issue of correctly using the syntax of XML, DTD and XPath; and deeper problems in composing regular expressions and path queries to do the job required.

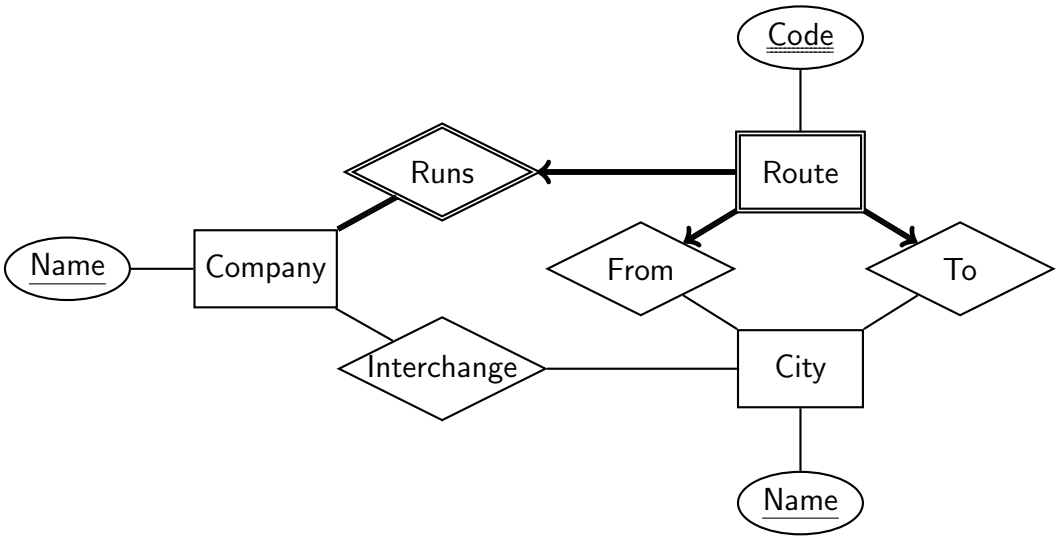
As with Question 1, the bookwork parts in the final two sections require English-language descriptions that are readable but precise, and not just general definitions but specialised to the situation presented here.

Question 3.

Lots of students correctly completed the contingency tables at the start of this question, successfully extracting relevant numerical data from the textual description. This was great. Many also correctly identified the null hypothesis, specific to this situation. Calculation of χ^2 was often correct, but less often used correctly: making a precise distinction between reliability, correlation and significance turned out to be a challenge.

Incidentally, this is authentic data about real Inf1-DA students, as claimed in the question. However there was an element of data-trawling in its preparation, to pick out which statistics gave a correlation, and that weakens any claim to real scientific discovery about behaviour.

Question 1 [*This question is worth a total of 35 marks.*]



This entity-relationship diagram shows a proposed model for a database recording information about long-distance coach journeys in the UK. It includes details about which companies run which routes between cities, and which cities each company uses as interchanges between its services.

(a) What are the *entities*, *relationships* and *attributes* in this diagram? [*3 marks*]

(b) What is the meaning of the double line around Route? Why is this necessary? What does the double line around Runs indicate? What is the primary key for a Route? [*8 marks*]

(c) How are *total participation* constraints represented in an ER diagram? List all instances of these in the diagram above, and the effect of each. [*5 marks*]

(d) How are *key constraints* represented in an ER diagram? List all the key constraints in the diagram above, and the effect of each. [*4 marks*]

(e) Draw up an SQL data declaration of appropriate tables to implement this entity-relationship arrangement (you should use **not null** where necessary, but need not include **on delete** declarations). [*15 marks*]

Notes on Question 1

(a) The entities are Company, Route and City. The relationships are Runs, Interchange, From and To. The attributes are Name (of Company), Name (of City) and Code (of Route).

Almost everyone correctly read these values off the entity-relationship diagram. There's no trick here: the rectangle, diamond and oval shapes always mean entity, relationship and attribute.

(b) The double line indicates that Route is a *weak entity*. This is necessary because a route code may not uniquely identify the coach route — different companies may reuse the same codes. The double line around Runs indicates that this is the *identifying relationship* for Route, linking each route to its unique operating Company. The primary key for a Route is {Code, Name}, where the Name is that of the operating Company.

Don't call Runs a "weak relationship" — it's a perfectly normal relationship, with the additional feature that it identifies how to distinguish between routes with the same code.

(c) Total participation constraints are indicated by a thick or double line joining an entity to a relationship. There are four in the diagram:

- From Route to Runs, requiring that every route have some operating company.
- From Route to From, requiring that every route have a starting city.
- From Route to To, requiring that every route have a finishing city.
- From Company to Runs, requiring that every company operates some coach routes.

Although this looks quite like part (a), many students lost marks by not fully answering the question. Reading the question carefully, you need to state how total participation constraints are represented, then list all the ones in the diagram, and finally say for each one what is its effect in this case.

This is quite a common feature of questions: give a general "bookwork" definition, and then describe in words how it applies in a particular situation. To answer it successfully, you need to both know the textbook definition of a concept, and also understand what it means so that you can describe its instantiation in the case at hand.

(d) Key constraints are indicated by an arrow on the line joining an entity to a relationship. There are three in the diagram:

- From Route to Runs, requiring that every route have no more than one operating company.
- From Route to From, requiring that every route have no more than one starting city.
- From Route to To, requiring that every route have no more than one finishing city.

In combination with the total participation constraints, these ensure that every route has exactly one operating company, start and end point.

As with part (c), a complete answer means not just finding all three arrows in the diagram, but also explaining in words what is their specific effect in this model.

(e) The following is an appropriate set of tables.

```
create table Company (
    name varchar(20),
    primary key (name)
)

create table Route (
    code           varchar(10),
    companyName varchar(20),
    origin         varchar(20) not null,
    destination    varchar(20) not null,
    foreign key (origin)          references City(name),
    foreign key (destination)     references City(name),
    foreign key (companyName) references Company(name),
    primary key (code,companyName)
)

create table City (
    name varchar(20),
    primary key (name)
)

create table Interchange (
    cityName       varchar(20),
    companyName varchar(20),
    foreign key (companyName) references Company(name),
    foreign key (cityName)       references City(name),
    primary key (companyName,cityName)
)
```

Here I have used a string, a varchar(10), for each bus route code. Using an integer would also be acceptable.

In all the **references** clauses I have given a full description of the target key, with **foreign key** (origin) **references** City(name) and similar. This isn't essential — writing **foreign key** (origin) **references** City is also acceptable, and is unambiguous because City has a single-field primary key — but I think the more explicit form is clearer, and it may also be more robust against any future changes in schema.

Some students recorded the origin and destination of bus routes using additional tables From and To. This isn't quite as good a solution: it doesn't capture the constraint that every route must have a starting city and a final destination. By putting origin and destination with **not null** declarations in the Route table itself we ensure that every route is listed with exactly one starting city and one finishing city.

Question 2 [*This question is worth a total of 35 marks.*]

The following XML document records some books held by village school libraries.

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE libraries SYSTEM "libraries.dtd">
<libraries>
    <library village="Aberfoyle">
        <section name="Fiction">
            <book author="Scott" title="Rob Roy"></book>
        </section>
        <section name="Nonfiction">
            <book author="Ray" title="Learning XML"/>
        </section>
    </library>
    <library village="Kippen">
        <book author="Hill">Where's Spot?</book>
    </library>
</libraries>
```

It is planned to extend this with information from other libraries. However, as you can see, the document structure is already rather complex: some libraries are divided into "Fiction" and "Nonfiction" sections, others are not; and while all books have an author listed in the element attributes, for some the title is held as an attribute and in others as content text.

(a) Draw the XPath data model tree for this XML document. [*10 marks*]

(b) Write a suitable DTD to be held in the file libraries.dtd. If you make any assumptions about the data structure, beyond those listed above, then state what they are. [*10 marks*]

(c) Write XPath expressions to list the following from such a document:

(i) All of the villages included.

(ii) All authors of books listed as Nonfiction.

(iii) Names of villages whose school library holds a book by Dickens. [*7 marks*]

Once the record of library holdings is more complete, it will be used as the basis of an information retrieval system. Given a suitable search topic, this should return a list of relevant books.

(d) The performance of an information retrieval system like this can be evaluated in terms of its *precision*, P, and *recall*, R. Give an English-language definition of these two terms. [*2 marks*]

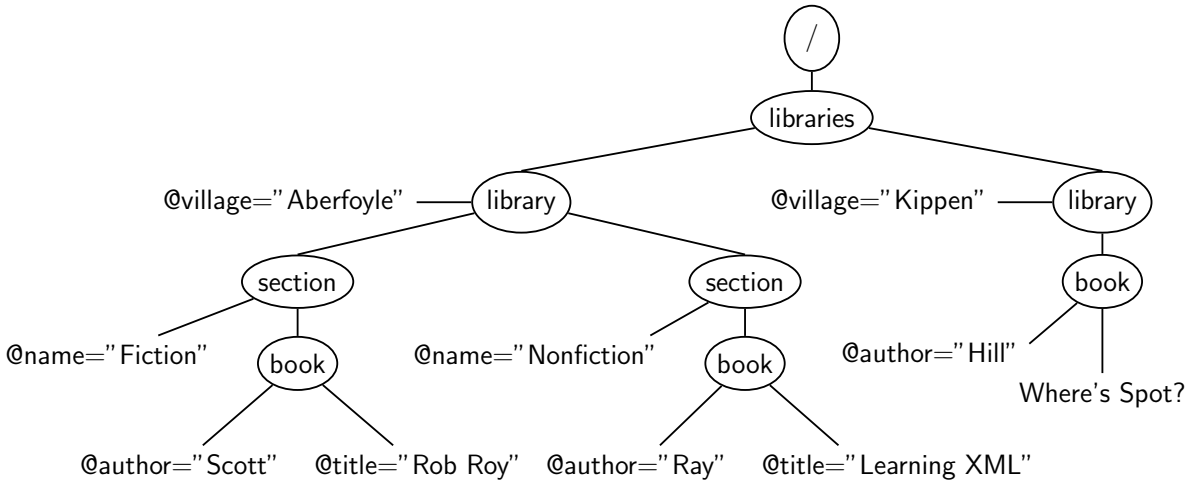
(e) Precision and recall are computed as follows:

$$P = \frac{TP}{TP + FP} \qquad\qquad R = \frac{TP}{TP + FN}$$

Name and define the three values TP, FP, FN appearing here. [*6 marks*]

Notes on Question 2

(a) Here is the corresponding tree.



Many students had a structure something like this, but there were also lots of details missed out. Notice, for example: the root node at the top; attribute nodes like @village="Aberfoyle" with both name (village) and value (Aberfoyle), properly attached to the relevant element node; some element nodes with multiple attribute nodes; and element nodes on the fringe of the tree both with and without subsidiary text nodes.

(b) Here is the contents of libraries.dtd

<!**ELEMENT** libraries (library)+>
<!**ELEMENT** library (section+|book+)>
<!**ELEMENT** section (book)*>
<!**ELEMENT** book (#**PCDATA**)>
<!**ATTLIST** library village **CDATA** #**REQUIRED**>
<!**ATTLIST** section name (Fiction|Nonfiction) #**REQUIRED**>
<!**ATTLIST** book author **CDATA** #**REQUIRED** title **CDATA** #**IMPLIED**>

Some variation is possible here: for example, the use of "+" indicates that there must be at least one library, and each library must contain at least one section or one book; but the "*" means that a particular section might be empty. Changing between "*" and "+" would give different restrictions.

Notice that this DTD does not include a <!**DOCTYPE**...> statement. That's because we already have one in the XML file, which points to this DTD.

The final declaration line could be split into two for an alternative answer:

<!**ATTLIST** book author **CDATA** #**REQUIRED**>
<!**ATTLIST** book title **CDATA** #**IMPLIED**>

The order of declarations in a DTD doesn't matter.

However, there are also many possible wrong answers, with regular expressions that don't fit the data or description given in the question. For example:

- The library element is correctly specified as (section+|book+). This means that a library either has sections, or it has just books. The following incorrect expressions fail to capture this properly:

(section|book)∗	This states that library can be a mixture of both sections (which contain books) and individual books.
(section+,book+)	This says that the XML record must contain some sections, followed by some books.

Many students used expressions with comma "," (which forces an ordering of XML elements) rather than "|" (which is the symbol for alternatives). Remember that order is significant in XML trees.

- To divide libraries into "Fiction" and "Nonfiction" sections means having precisely two possible values for the name attribute of a section. This can be captured with the "(Fiction|Nonfiction)" attribute declaration in the DTD. Using the more general CDATA would weaken the constraint.
- The element declaration (#PCDATA) for book includes the possibility that the element is empty. There is no need to try (#PCDATA|EMPTY), which is anyway not legal XML. The EMPTY keyword is reserved for the case where an element *must* be empty.
- Many students made small errors in writing out the DTD: such as missing out the libraries element, or writing ATTRIBUTE instead of ATTLIST.
- There are certainly other possible XML structures to represent this sort of information — for example making author and title element nodes — but here it's essential to give a DTD that matches the XML document and design choices presented in the question.

(c) (i) Here are three possible variations on an answer:

- //@village
- /libraries/library/@village
- /descendant::library/attribute::village

(ii) Three ways to do it:

- //section[@name="Nonfiction"]/book/@author
- //book[../../section/@name="Nonfiction"]/@author
- //∗[@name="Nonfiction"]//@author

(iii) Another three alternatives:

- //library[.//@author="Dickens"]/@village
- /libraries/library[descendant::book/attribute::author='Dickens']/attribute::village
- //book[@author='Dickens']/ancestor::library/@village

In each case I would probably use the first choice, but all of these are correct.

One incorrect proposal for the final part of this question was quite common:

- //∗[book/@author='Dickens']/../@village ← Incorrect.

This fails because it doesn't work for a library that is divided into Fiction and Nonfiction sections. It's necessary to use the descendant or ancestor axes correctly to handle both cases, of libraries with and without section divisions.

Notice that in all cases the attribute values are addressed directly: there is no need to use text(), as none of these are extracting the content of a text node. In fact, the only text node in the XML tree given is the book title "Where's Spot?".

(d) The *precision* P is the proportion of books returned which are relevant to the search topic.

The *recall* R is the proportion of the relevant books in the collection which are successfully retrieved.

The question asks for an "English-language definition". That means using words, not equations, but the answer should still be precise. Also, notice that these answers don't give just a general textbook definition of precision and recall, but specialise them to the situation of searching for books in a library catalogue.

(e) The names and definitions are as follows:

- TP True positives: the number of relevant books returned by the search.
- FP False positives: the number of books returned by the search which are in fact not relevant.
- FN False negatives: the number of books which are relevant but the search does not return.

Again, these answers are the standard definitions made specific for this scenario of a book search. Most people correctly described true and false positives; quite a few gave an incorrect definition for false negatives.

Precision is important here, and answering the question as asked: some students gave the name of the value ("True positive"), but not the definition; or the definition without the name. The question asked for both.

Question 3 [*This question is worth a total of 30 marks.*]

An anonymous survey at an Inf1-DA lecture in 2011 collected data on students' home countries, the hours they spent on physical exercise each week, and how long they had slept the night before. Here is some information from that survey.

- There were 64 responses, of which 50 reported a home country in the European Union and the remainder were from outside the EU.
- Overall 45 students reported that they took at least 2.5 hours of physical exercise each week: 32 of these were students from the EU.
- Half of the students estimated they had spent more than 7 hours asleep on the previous night. Of these, 10 were from outside the EU and 22 were from EU countries.

You are asked to analyse whether this data provides evidence of any correlation between home country and either exercise or sleeping patterns.

(a) What is the *null hypothesis* for this investigation? [*2 marks*]

(b) Draw up two contingency tables of frequencies from this data: one showing home country against exercise and the other home country against sleeping hours. [*6 marks*]

(c) The following formula is used to calculate the χ^2 statistic for tables of data.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

What are the values O_{ij} and E_{ij} in this equation? How are they calculated? [*6 marks*]

(d) Compute the χ^2 statistic for each of your contingency tables. [*4 marks*]

(e) Here each table of data has *one degree of freedom*. Explain what that means. [*2 marks*]

(f) The appropriate critical values for this χ^2 test are as follows.

p	0.10	0.05	0.01	0.001
χ^2	2.71	3.84	6.64	10.83

Answer for each of your contingency tables:

- Is the χ^2 test reliable for this table?
- If it is: does it show a correlation?
- If if does: is it significant?

[*6 marks*]

(g) Summarise in words the evidence that this data provides regarding correlation of a student's home country with exercise and sleeping patterns. [*4 marks*]

Notes on Question 3

(a) The null hypothesis is that there is no correlation between home country and either amount of physical exercise or hours slept.

A number of people wrote a null hypothesis about whether home country affected exercise or sleep. This isn't quite right: the null hypothesis does not say anything about possible causality or mechanism, just about whether things are correlated.

(b) Filling out the values provided in the question gives us this:

Exercise	EU	Overseas	
> 2.5 hours	32		45
≤ 2.5 hours			
	50		64

Sleep	EU	Overseas	
> 7 hours	22	10	32
≤ 7 hours			
	50		64

This data is then enough to calculate everything else in the contingency tables:

Exercise	EU	Overseas	
> 2.5 hours	32	13	45
≤ 2.5 hours	18	1	19
	50	14	64

Sleep	EU	Overseas	
> 7 hours	22	10	32
≤ 7 hours	28	4	32
	50	14	64

(c) The O_{ij} are the observed values in each cell of the contingency table. They are calculated as the table is built.

The E_{ij} are the expected values for each cell of the contingency table, assuming the null hypothesis. The are calculated in proportion from the marginals, the totals of each row and column.

(d) These are the tables of expected values assuming the null hypothesis.

Exercise	EU	Overseas	
> 2.5 hours	35.16	9.84	45
≤ 2.5 hours	14.84	4.16	19
	50	14	64

Sleep	EU	Overseas	
> 7 hours	25	7	32
≤ 7 hours	25	7	32
	50	14	64

These should not be rounded to the nearest integer, even though of course the original tables only contain integer values. Keeping fractional values is necessary for the correct χ^2 statistic.

The χ^2 value for weekly exercise hours against home country is:

$$\begin{aligned}\chi^2 &= \frac{3.16^2}{35.16} + \frac{3.16^2}{9.84} + \frac{3.16^2}{14.84} + \frac{3.16^2}{4.16} \\ &= 4.36\end{aligned}$$

The χ^2 for sleep against home country is:

$$\begin{aligned}\chi^2 &= \frac{3^2}{25} + \frac{3^2}{25} + \frac{3^2}{7} + \frac{3^2}{7} \\ &= 3.29\end{aligned}$$

(e) One degree of freedom means that given a fixed value for any single cell in the contingency table, the value of all the others is forced by the constraints on row and column totals.

(f) For the exercise table, one of the expected cell values is below 5 which means that the χ^2 is not reliable. (It does show a correlation, which would be significant if reliable.)

For the sleep table, all expected cell values are above 5, so that the χ^2 test is reliable (even though one of the observed values is much lower). It does show a correlation, and the evidence is significant at the 90% level.

An acceptable alternative answer would be to say that the test shows a correlation, but this evidence is not significant at the 95% level.

Several people said that 90% or 95% significance was good, but wouldn't be enough for a scientific experiment. In fact that level of significance is strong evidence for a correlation, and certainly enough to be part of a scientific result. However, an important part of science is that it be repeatable: having seen significant evidence on this occasion, do we see it again and again in later tests?

Some students described χ^2 as indicating a positive correlation. The χ^2 value is always a positive number: it doesn't say anything about whether correlation is positive or negative, just whether the values in the table vary from those predicted by the null hypothesis. Positive or negative correlation applies to quantitative data and Pearson's correlation coefficient — here we have qualitative (categorical) data.

(g) This data provides no evidence of any correlation between students' home country and exercise patterns; it does provide evidence of correlation, sufficient to reject the null hypothesis, suggesting that students from outside the EU get a longer night's sleep. (Or, if you were look for confidence at the 95% level, then the data does provide evidence of a correlation but not enough to reject the null hypothesis.)

Notice that because the χ^2 statistic for the exercise table is not reliable, owing to small sample size, it doesn't provide *any evidence at all*, either for or against the null hypothesis. We just can't squeeze any information out of it.

Many people said that the 90% critical value meant there was a 90% chance of a correlation. This is incorrect: there either is or is not an underlying correlation in the system, and what we are doing is experiments to assess this. High values of χ^2 increase our confidence that there is a correlation. The 90% measures something quite different: whether the result we see would happen 10% of the time anyway, even if there was no actual correlation.

Quite a few students made general statements about sample size or apparent correlation that were either intuition about the numbers, or based on proposed reasons why things might be linked. That's not really enough here: the point of statistical tests is that we can make precise these ideas, and give firm statements about correlation and significance based on the data.