Informatics 1: Data & Analysis

Lecture 22: More Practice Exam Questions

Ian Stark

School of Informatics
The University of Edinburgh

Friday 6 April 2018
Semester 2 Week 11

Student Course Feedback !

Please complete the online survey for Inf1-DA. It's anonymous and I read every submission.

MyEd $\longrightarrow$ Studies $\longrightarrow$ Course Enhancement Questionnaire

You can also reach the questionnaires in the following ways.

- Direct link **http://edin.ac/CEQ**
- Find survey email titled "Course Enhancement Questionnaires"

There are feedback forms for every course where you are enrolled. For all Informatics courses there are questions where can give feedback to staff and also offer your own advice to future students. We publish these comments online so you can find out yourself about courses you plan to take: http://www.inf.ed.ac.uk/teaching/student-feedback

Optional Exam Practice Tutorial !

Ten of the Inf1-DA tutors are coordinating to give additional exam practice. It's entirely optional: if you do sign up then please do try to follow through with completing the work and attending the tutorial.

- Locate and download the written coursework assignment for Inf1-DA 2016/17. This is a mock exam paper which you will work through and submit.
- Choose a time and tutor at http://is.gd/da18extra.
- Complete the mock exam paper, writing out your answers in full with pen and paper.
- Submit your work by 4pm on Wednesday 18 April. Check with the tutor how they want you to do this: in the box by the ITO, or scan/photograph it and send to them by email.
- Tutor will mark your work with written feedback.
- Attend tutorial. Get your work back, discuss feedback with tutor.

Homework From Tuesday

Follow These Instructions

- Locate and download the main and resit Inf1-DA exam papers 2017. For this you will need to find the library website with past papers.
- Read Question 3 from each paper.
- Work through and write out answers to both questions.

  When doing this it's fine to look at past lectures, read your notes, look things up, and ask for help from others. This is about practising to write exam answers that are as good as you can make them.

- Bring your solutions along to the lecture on Friday.

In the lecture I shall (again) explain some sample solutions and guide you through marking your own.

May 2017 Question 3 Section (a)

A database consulting company is evaluating the performance of a remote server. They test this by sending a number of simple queries, one after another, and measuring the time to receive a response for each one. This *round-trip time* is a key performance measure for user experience. Here are the first seven measurements.

	n	1	2	3	4	5	6	7
Time(ms)	t_n	140	500	203	233	303	140	175

Times are in milliseconds, and the test has a 500-millisecond timeout: if there is no answer for query n within that time then it gives up and records t_n as 500.

(a) For this set of values calculate their *mean*, their *median*, their *mode* and their *standard deviation*. Show your working and explain your calculation, giving formulas if appropriate. [8 marks]

May 2017 Question 3 Section (a)

$$\text{Mean } \mu = \frac{1}{7}\sum_{n=1}^{7} t_n = \frac{140 + 500 + 203 + 233 + 303 + 140 + 175}{7} = 242$$

$$\text{Median} = \text{Middle value when sorted arithmetically} = 203$$

$$\text{Mode} = \text{Most common value} = 140$$

$$\text{Standard deviation } \sigma = \sqrt{\frac{1}{7}\sum_{n=1}^{7} (t_n - \mu)^2} = 117.8$$

The question specifically asks for the standard deviation of the values given, so the factor $\frac{1}{7}$ here is appropriate: using σ_{n-1} with $\frac{1}{6}$ would be wrong.

May 2017 Question 3 Section (b)

(b) Mean, median and mode are all kinds of *average*. For each of these three:

- Is it a sensible measure to use to evaluate system performance?
- If not, then explain why not.
- If it is a sensible measure, then briefly describe how it is affected by the use of a timeout.

[9 marks]

May 2017 Question 3 Section (b)

The *mean* is a sensible measure to evaluate system performance. The use of a 500ms timeout will tend to *reduce* the measure: waiting until the response eventually does come in, possibly much later, would give an even larger value.

(The use of a timeout here does mitigate one of the disadvantages of mean, that it can be skewed by extreme values, which in this case can only be very high as there is a lower bound of zero.)

The *median* is a sensible measure to evaluate system performance. The 500ms timeout does not affect the measure, unless half or more of the tests take this long.

(This is an advantage of the median, that it is not strongly affected by individual eccentric results.)

May 2017 Question 3 Section (b)

The *mode* is not a sensible measure to evaluate system performance. Although there are in this case two tests with the same measured time, this is entirely by chance and in many cases every value would be different.

(Indeed, the value itself depends on the precision to which we record the time interval — for example, if we recorded to the nearest 1/10 of a second then the mode would be 0.2s.)

May 2017 Question 3 Sections (c)–(e)

The same consultants are interested to find out whether there are differences in choice of database among potential customers in different European countries. Here is a small sample of the data they have available on system installations among their target market.

	UK	France	Germany	Total
Oracle	11	8	31	50
MySQL	19	12	19	50
Total	30	20	50	100

Each entry in the table counts target companies in the country named who use either *Oracle* or *MySQL* as their primary database engine.

The consultants propose χ^2 testing to see if there are correlations in this data.

(c) What is the *null hypothesis* for this investigation?

(d) Assuming fixed row and column totals, how many *degrees of freedom* has this table?

(e) Calculate the χ^2 value for this table. [10 marks]

May 2017 Question 3 Sections (c)–(d)

(c) The null hypothesis is that there is no correlation between country and choice of database system.

(d) The table has two degrees of freedom.

May 2017 Question 3 Section (e)

(e) Here is the table of expected values assuming the null hypothesis.

	UK	France	Germany	Total
Oracle	15	10	25	50
MySQL	15	10	25	50
Total	30	20	50	100

This gives a χ^2 calculation as follows.

$$\chi^2 = \frac{(11-15)^2}{15} + \frac{(19-15)^2}{15} + \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(31-25)^2}{25} + \frac{(31-25)^2}{25}$$
$$= 5.81$$

Notice that the numerators are not all the same, unlike with 2×2 tables.

May 2017 Question 3 Section (f)

The full table of data includes many more countries and other database engines. The consultants calculate a χ^2 value of 44.2 for this larger table. The appropriate critical values are as follows:

p	0.10	0.05	0.025	0.01
χ^2	33.2	36.4	39.4	43.0

(f) Summarise briefly what evidence this data from the full table provides about any possible correlation between country and database use among potential customers for the consultancy. [3 marks]

Note (not in the exam): These are the appropriate critical values for a table of the top 5 database engines across 7 countries, giving 24 degrees of freedom.

May 2017 Question 3 Section (f)

(f) The χ^2 value of 44.2 provides evidence to *reject* the null hypothesis at the 99% level, suggesting quite strongly that among the target customers of the consultancy there is a correlation between their country of operation and choice of database system.

What it doesn't give is any indication of the details of the correlation — which country favours which system, or whether it only applies to certain countries — or any possible cause.

August 2017 Question 3 Sections (a)–(b)

(a) Give brief explanations of the following four kinds of data scale, making clear how each differs from the others: *categorical*, *ordinal*, *interval* and *ratio*. For each one, give an example of data that is measured using that kind of scale. [8 marks]

(b) A supermarket quality tester selects at random for inspection five packets of sweets from a supply crate. All are sold as "225g" and in fact are found to weigh the following amounts: 230g, 225g, 232g, 234g, 229g.

Use this set of sample data to calculate estimates for the mean and standard deviation of the weights of all the packets of sweets in the crate. Use a calculator and write down all your working. [6 marks]

August 2017 Question 3 Section (a)

A *categorical* scale divides data into different named categories. There is no ordering or numerical content. For example, classifying words as different parts of speech.

An *ordinal* scale gives a recognised ordering between data items, but there is no arithmetic content. Numbers may still be used to record the ordering, but there is no way to calculate on them. For example, dividing degrees as 1st, 2.1, 2.2 and 3rd class.

An *interval* scale assigns numeric values to data, but where these values are relative to each other. Values can be compared, averaged, and subtracted; but not multiplied or added together. For example, times of day.

A *ratio* scale uses numeric values which have an absolute notion of zero; this means they can sensibly be added, and multiplied by real numbers. For example the mass of an object or the speed of a vehicle.

August 2017 Question 3 Section (b)

We can estimate the mean weight of all packets in the crate simply using the mean weight of the five packets sampled.

$$\mathrm{m} = \frac{230 + 225 + 232 + 234 + 229}{5} = \frac{1150}{5} = 230$$

This gives an estimate of 230g for the mean weight across the whole crate.

August 2017 Question 3 Section (b)

For the standard deviation, we must correct for sampling.

$$s = \sqrt{\frac{(230-230)^2 + (225-230)^2 + (232-230)^2 + (234-230)^2 + (229-230)^2}{(5-1)}}$$

$$= \sqrt{\frac{0^2 + 5^2 + 2^2 + 4^2 + 1^2}{4}} = \sqrt{\frac{46}{4}} = 3.39$$

This gives an estimate of 3.39g for the standard deviation of weights in the crate. Giving more digits 3.39116... doesn't really offer any more accuracy but in the exam it wasn't considered an error as the course did not discuss precision, errors, or significant figures.

Note that the sample is small enough that this is clearly different to the standard deviation among the five bags σ_n, which is only 3.03g.

August 2017 Question 3 Sections (c)–(d)

The standard "information retrieval (IR) task" is, given a *query* and a collection of *documents*, to find those documents that are relevant to the query.

(c) One measure of performance of an information retrieval algorithm is its *recall* R, usually computed using the following formula.

$$\mathrm{R} = \frac{TP}{TP + FN}$$

State the matching formula for the other common performance measure, *precision* P.

[2 marks]

(d) The terms TN and FN stand for the numbers of "True Negatives" and "False Negatives", respectively. What do these phrases mean? [2 marks]

August 2017 Question 3 Sections (c)–(d)

(c) Precision is computed using the following formula.

$$\mathrm{P} = \frac{TP}{TP + FP}$$

(d) True Negatives: the number of documents not returned by the system that are not relevant to the query.

False Negatives: the number of documents that are relevant to the query but are not returned by the system.

August 2017 Question 3 Section (e)

(e) You are evaluating information retrieval systems to use with a collection of 12000 documents from an online archive of TV scripts. There are two candidate systems: *Scoop* and *Drill*. Each is tested on the query "MacGyver", for which there are 30 relevant documents in the archive.

Scoop returns 240 documents, with 24 of those being relevant; *Drill* returns only 30, but 21 of them are relevant.

For each system, draw up a table of retrieval against relevance from this data, and use it to calculate precision and recall on the "MacGyver" test. Show your working. [8 marks]

August 2017 Question 3 Section (e)

(e) The following tables give all the necessary figures for the calculation.

Scoop	Relevant	Not relevant	Total
Retrieved	24	216	240
Not retrieved	6	11754	11760
Total	30	11970	12000

Drill	Relevant	Not relevant	Total
Retrieved	21	9	30
Not retrieved	9	11961	11970
Total	30	11970	12000

From these we can evaluate the performance of each system on this single query.

Scoop precision $\mathrm{P} = \dfrac{24}{240} = 1/10 = 0.1$ *Drill* precision $\mathrm{P} = \dfrac{21}{30} = 7/10 = 0.7$

Scoop recall $\mathrm{R} = \dfrac{24}{30} = 8/10 = 0.8$ *Drill* recall $\mathrm{R} = \dfrac{21}{30} = 7/10 = 0.7$

August 2017 Question 3 Section (f)

(f) The F_α performance measure combines precision and recall using the formula

$$\mathrm{F}_\alpha = \frac{1}{\alpha\frac{1}{\mathrm{P}} + (1-\alpha)\frac{1}{\mathrm{R}}}\,.$$

Compute $\mathrm{F}_{0.2}$ for *Scoop* and *Drill*. Which performs more strongly?

The *balanced* F-score uses $\alpha = 0.5$. Does the choice of $\alpha = 0.2$ favour systems that are good at precision, or recall? [4 marks]

August 2017 Question 3 Section (f)

(f) The *Drill* system has the stronger $\mathrm{F}_{0.2}$ score.

$$\mathrm{F}_{0.2}(\textit{Scoop}) = \frac{1}{0.2\frac{1}{0.1} + 0.8\frac{1}{0.8}} = 0.33$$

$$\mathrm{F}_{0.2}(\textit{Drill}) = \frac{1}{0.2\frac{1}{0.7} + 0.8\frac{1}{0.7}} = 0.70$$

Choosing $\alpha = 0.2$ favours recall (but, in this case, not enough to offset *Scoop*'s terrible precision).

Adding it all up

Question 1 was out of 40 marks, Question 2 out of 30 marks. You can add these up and convert them into grades with the following table.

Mark out of						
30	40	70	100	Grade	Description	Degree Class
27	36	63	90	A1	Excellent	1st
24	32	56	80	A2	Excellent	1st
21	28	49	70	A3	Excellent	1st
18	24	42	60	B	Very Good	2:1
15	20	35	50	C	Good	2:2
12	16	28	40	D	Pass	3rd
9	12	21	30	E	Marginal Fail	
6	8	14	20	F	Clear Fail	
3	4	7	10	G	Bad Fail	
0	0	0	0	H	Bad Fail	

You can do this

The Inf1-DA syllabus and exam questions are written to be achievable. Every year large numbers of students pass the exam writing straightforward correct answers about things they understand. You can do this too.

Anything Else?

If you have further questions about the course content, tutorial exercises, the exam, where to buy a disco calculator, or anything else, please:

- Post a question on *Piazza*; *or*
- Ask your course tutor, in person or by email; *or*
- Ask me, in person or by email.

Thank you for your attention

We're done here