Informatics 1: Data & Analysis

Lecture 20: Course Review

Ian Stark

School of Informatics
The University of Edinburgh

Tuesday 1 April 2014
Semester 2 Week 11

Plan

This is week 11, the last teaching week of Semester 2.

Your final tutorial for Inf1-DA is this week, in which you should receive back your work on the coursework assignment.

This is the last lecture of Inf1-DA.

- Exam arrangements and format
- Summary of course topics
- Review: Statistics and Hypothesis Testing
- Review: Tuple-Relational Calculus

Time and Place !

Informatics 1: Data & Analysis will be assessed by a single two-hour written examination.

Date: Friday 16 May 2014
Time: 0930–1130
Place: St Leonard's Land Games Hall

This information for course code INFR08015 is current at 2014-04-01; please check the link on the Inf1-DA web page nearer to the date to check this and to confirm all of your exams.

Exam Format

As in previous years, the exam will have three compulsory questions.

- Read all questions before beginning the paper
- You don't need to do the questions in order
- Don't assume a question is only using one part of the course
- If you get stuck on one question: don't waste too much time on it; do go on to the next question; and don't give up!

Calculators are permitted, and will be provided at the exam hall. These are a standard scientific model: you can try one out at the ITO if you wish.

Past Exam Papers

Many of the example questions and solutions on tutorial exercises are taken from past exam papers.

The University Library keeps a full set of past papers online, and you can access them through links from the course web page.

These are a good source of revision material, and I strongly recommend you attempt as many of these questions as you can. However:

- While the overall format of questions remains similar, the exact topics covered do change from year to year.
- Where online "sample solutions" are provided, they may not always be correct and may not provide information you require. (Most were written as guides for external examiners reviewing the paper, not as model answers for students.)
- If you are puzzled by a past question, ask on NB, or email me.
- There have been changes to the course content over the years, so not all past exam questions are relevant.

Questions about Past Exam Questions

(This slide kept mostly blank to provide space for NB queries)

Which Past Exam Questions?

Past Papers

In each year there are exams from the main and resit diet. The following questions are relevant to the current course syllabus.

Informatics 1B 2005:	Questions 1 and 2
Informatics 1B - D&A 2006, 2007:	Questions 4 and 5
2008:	Everything except 2(c) on XQuery
2009, 2010, 2011, 2012, 2013:	All questions

Examinable Material

Unless otherwise specified, all of the following material is examinable:

- Topics covered in lectures
- Directed reading distributed in lectures
- Topics covered in the weekly exercise sheets

Topic Summary

The entity-relationship model, ER diagrams. The relational model, SQL DDL. Translating an ER model into a relational one. Relational algebra, tuple-relational calculus, SQL queries; translating between all three.

Semistructured data models and the XPath data tree. XML documents. Schema languages and DTDs. Relational data converted into XML. XPath as a query language.

Corpora: what they are and how they are made; examples. Annotations and tagging. Concordances, frequencies, n-grams, collocations. Methods for machine translation.

Information retrieval: what it is, evaluating and comparing performance of IR systems; the vector space model and cosine similarity measure.

Data scales, summary statistics, population vs. sample; hypothesis testing and significance; correlation coefficient, χ^2 test.

Some Specific Items

Corpora

In general it is the principles of corpora that are examinable, rather then the precise details of individual corpora. Similarly, you should be familiar with the principles underlying POS-tagging and syntactic annotation, but you do not need to know detailed linguistics or specific tag sets.

You should however, be able to give examples of a corpus or a POS tag.

The CQP tool was used in a tutorial, so is examinable — although again for general principles and use, not every detail of syntax.

Statistics

You are not expected to memorize critical value tables; however, you should be able to use one if provided.

You are expected to know the formulas for the various statistics used, and to be able to calculate with them.

Data Scales

Categorical	Qualitative, fixed set of categories, no order, no possible arithmetic.	Postcodes
Ordinal	Qualitative, fixed set of categories, can be ordered, still no arithmetic.	Exam grades
Interval	Quantitative, values all relative; can take averages, subtract one value from another; no addition or multiplication.	Dates
Ratio	Quantitative, absolute values, can take averages, subtract, add, and take scalar multiples of values.	Mass, energy

Summary Statistics

Mode: All data scales, most common value

Median: Ordinal and quantitative scales, middle value

Mean: $$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance: $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

Standard deviation: $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2}$$

Estimates from Samples

Sample size n from a population of size N, where $n \ll N$

To estimate the mean of the population, use the mean of the sample:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad E(m) = \mu$$

To estimate the variance of the population, use this:

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - m)^2 \qquad\qquad E(s^2) = \sigma^2$$

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - m)^2}$$

The term $(n-1)$ is *Bessel's correction*.

Tests of Significance / Hypothesis Testing

To test for a statistical result, start with a specified *null hypothesis*, that there is nothing out of the ordinary in the data.

Compute some statistic R from the data.

Consult a table of *critical values* to see what is the chance p of getting a statistic as extreme as R if the null hypothesis holds.

If p is small — getting a value like R is very unlikely — then the result is *significant* and we reject the null hypothesis.

For example: if $p < 0.05$, the result is significant "at the 95% level".

Example test statistics:

- Correlation coefficient $\rho_{x,y}$ for paired quantitative data;
- χ^2 statistic for summary tables counting categorical data.

Example

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Sold
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

The company plan to use a χ^2 test to investigate whether colour affects sales.

Example

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Sold
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

The company plan to use a χ^2 test to investigate whether colour affects sales.

Null hypothesis: Colour makes no difference to sales

Example

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Obs.
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

Colour	Expected
Red	250
Yellow	250
Green	250
Blue	250
Total	1000

The company plan to use a χ^2 test to investigate whether colour affects sales.

Null hypothesis: Colour makes no difference to sales

Example

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Obs.
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

Colour	Expected
Red	250
Yellow	250
Green	250
Blue	250
Total	1000

$$\begin{aligned}\chi^2 &= \textstyle\sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} \\ &= \frac{15^2}{250} + \frac{25^2}{250} + \frac{15^2}{250} + \frac{25^2}{250} \\ &= 6.8\end{aligned}$$

The company plan to use a χ^2 test to investigate whether colour affects sales.

Null hypothesis: Colour makes no difference to sales

Example

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Obs.
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

Colour	Expected
Red	250
Yellow	250
Green	250
Blue	250
Total	1000

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$$= \frac{15^2}{250} + \frac{25^2}{250} + \frac{15^2}{250} + \frac{25^2}{250}$$

$$= 6.8$$

Critical values for χ^2 test with three degrees of freedom

p	0.1	0.05	0.01
χ^2	6.25	7.81	11.35

Example

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Obs.
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

Colour	Expected
Red	250
Yellow	250
Green	250
Blue	250
Total	1000

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$$= \frac{15^2}{250} + \frac{25^2}{250} + \frac{15^2}{250} + \frac{25^2}{250}$$

$$= 6.8$$

Critical values for χ^2 test with three degrees of freedom

p	0.1	0.05	0.01
χ^2	6.25	7.81	11.35

Example

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Obs.
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

Colour	Expected
Red	250
Yellow	250
Green	250
Blue	250
Total	1000

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$$= \frac{15^2}{250} + \frac{25^2}{250} + \frac{15^2}{250} + \frac{25^2}{250}$$

$$= 6.8$$

Critical values for χ^2 test with three degrees of freedom

p	0.1	0.05	0.01
χ^2	6.25	7.81	11.35

$6.8 > 6.25$ so significant at 90% level.

Example

A company making consumer-grade widgets wants to know whether they can sell more by careful choice of the colour of box the widget is sold in. Their initial test is to supply widget boxes in four different colours and see how many they sell of each colour. The following table shows the box colours of the first thousand widgets sold.

Colour	Obs.
Red	235
Yellow	275
Green	225
Blue	265
Total	1000

Colour	Expected
Red	250
Yellow	250
Green	250
Blue	250
Total	1000

$$\begin{aligned}\chi^2 &= \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} \\ &= \frac{15^2}{250} + \frac{25^2}{250} + \frac{15^2}{250} + \frac{25^2}{250} \\ &= 6.8\end{aligned}$$

Critical values for χ^2 test with three degrees of freedom

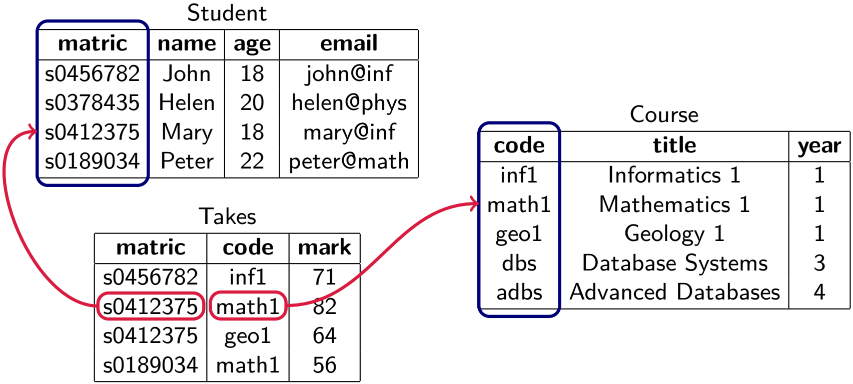
p	0.1	0.05	0.01
χ^2	6.25	7.81	11.35

$6.8 > 6.25$ so significant at 90% level.
Reject null hypothesis. Colour affects sales.

Relational Database Modelling

Relational models

- Relations: Tables matching schemas
- Schema: A set of field names and their domains
- Table: A set of tuples of values for these fields



Student

matric	name	age	email
s0456782	John	18	john@inf
s0378435	Helen	20	helen@phys
s0412375	Mary	18	mary@inf
s0189034	Peter	22	peter@math

Takes

matric	code	mark
s0456782	inf1	71
s0412375	math1	82
s0412375	geo1	64
s0189034	math1	56

Course

code	title	year
inf1	Informatics 1	1
math1	Mathematics 1	1
geo1	Geology 1	1
dbs	Database Systems	3
adbs	Advanced Databases	4

Tuple-Relational Calculus

The tuple-relational calculus (TRC) is a mathematical language for expressing queries over a relational database.

Standard TRC Idioms

To pick out some tuples from an existing table:

$$\{\, R \in \mathrm{Table} \;\mid \exists S \in \mathrm{OtherTable}, T \in \mathrm{FurtherTable} \,.\, \langle \mathit{Test} \rangle \,\}$$

With $\langle \mathit{Test} \rangle$ a Boolean expression using R, S, T,...

To obtain tuples not directly appearing in any other table:

$$\{\, R \;\mid \exists S \in \mathrm{OtherTable}, T \in \mathrm{FurtherTable} \,.\, \langle \mathit{Test} \rangle \,\}$$

With $\langle \mathit{Test} \rangle$ including $(R.\mathrm{field} = S.\mathrm{otherfield}) \land \ldots$

Example: Records from Existing Table

All records for students more than 19 years old

$$\{\, S \in \mathsf{Student} \;\mid\; S.\mathsf{age} > 19 \,\}$$

The set of all tuples S in the table "Student" with field "age" greater than 19.

All records for students taking math1

$$\{\, S \in \mathsf{Student} \;\mid\; \exists \mathsf{T} \in \mathsf{Takes} \,.\, S.\mathsf{mn} = \mathsf{T.mn} \wedge \mathsf{T.code} = \mathsf{math1} \,\}$$

The set of all tuples S in the table "Student" for which there is a tuple T in "Takes" linking the matriculation number of S to course "math1".

Example: Projecting and Building New Records

Results for a single course

$$\{\, \mathrm{R} \mid \exists \mathrm{T} \in \text{Takes} \,.\, \mathrm{T.code} = \text{"math1"} \wedge \mathrm{R.mn} = \mathrm{T.mn} \wedge \mathrm{R.mark} = \mathrm{T.mark} \,\}$$

The set of tuples R where there is a "Takes" tuple T with code "math1" such that R and T have matching matriculation number and mark.

Students on courses

$$\begin{aligned}\{\, \mathrm{R} \mid \exists S \in \text{Student}, \mathrm{T} \in \text{Takes}, \mathrm{C} \in \text{Course} \,.\, & \\ & S.\mathrm{mn} = \mathrm{T.mn} \;\wedge\; \mathrm{T.code} = \mathrm{C.code} \\ & \wedge\; \mathrm{R.name} = S.\mathrm{name} \;\wedge\; \mathrm{R.title} = \mathrm{C.title} \,\}\end{aligned}$$

The set of all tuples R where there is a "Student" tuple S, a "Takes" tuple T, and a "Course" tuple C with matching matriculation numbers and course codes, and where R takes student name from tuple S and course title from tuple C.

Bonus Reading

Tim Harford.
Big Data: Are We Making a Big Mistake?
FT Magazine, 28 March 2014.
http://is.gd/ftbigdata

March 28, 2014 11:38 am

Big data: are we making a big mistake?

By Tim Harford

Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media

Finis

Anything Else?

If you have further questions about the course material, past lectures, exercises, the exam, or anything else, please:

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

Course Survey

Please complete the online survey for this course. This is anonymous, and I read every submission. Ideally, do it immediately after this lecture.

Go to **http://www.inf.ed.ac.uk/teaching** and follow the link to the Semester 2 Course Surveys.