



THE UNIVERSITY
of EDINBURGH



Data Science in Medicine

Lecture 9: Course Outro

Dr Areti Manataki



Usher Institute
The University of Edinburgh

Why this course?

Preparing for the new era of data-intensive medicine

- Growing volume of data in medicine, healthcare and the life sciences
- Useful for your future studies:
 - Intercalation in Year 3
 - Research project in Year 5
- Useful for your future career:
 - Making sense of research findings
 - Doing research, which will most probably involve data
 - Understanding and improving the health of your patients and the way care is provided

Preparing for the new era of data-intensive medicine

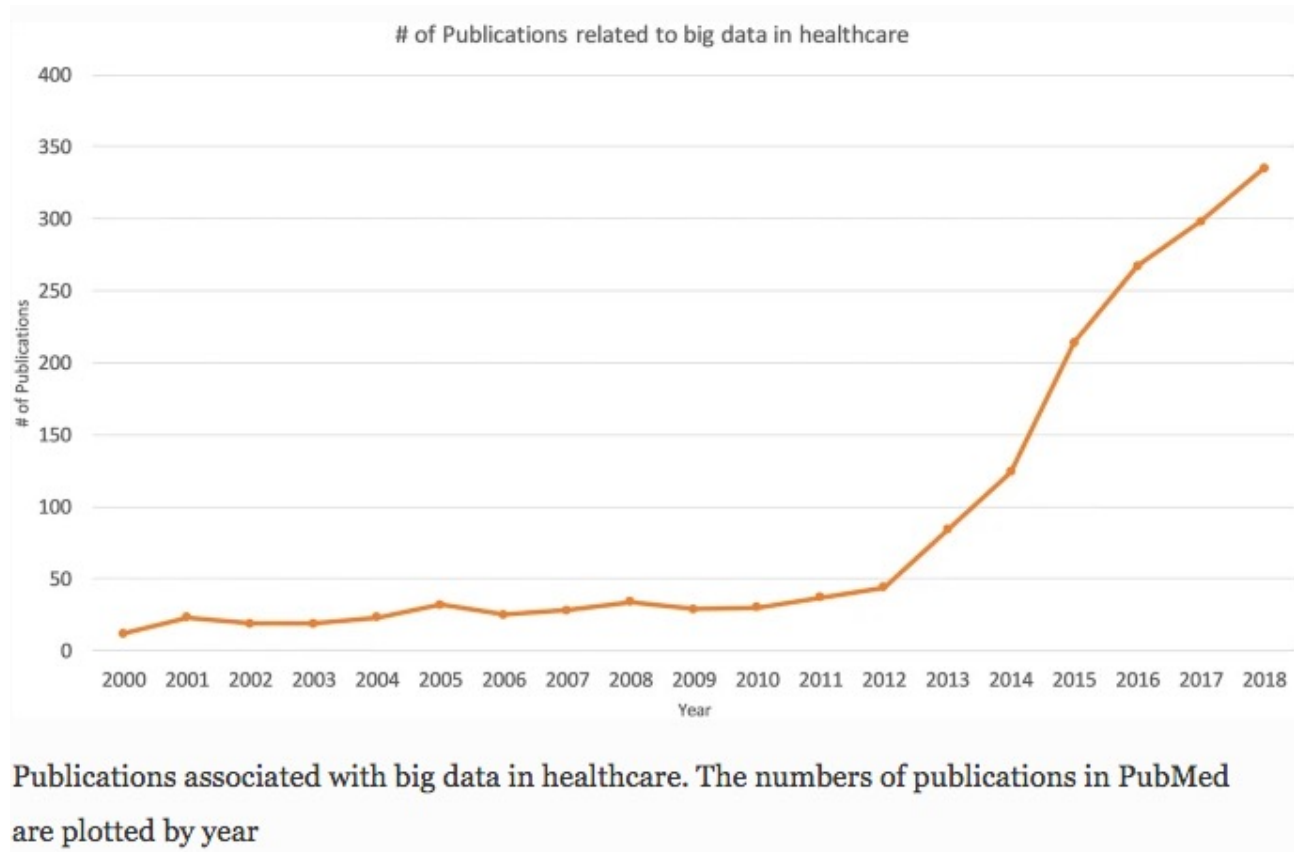


Image from Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis and future prospects. *J Big Data* 6, 54 (2019). <https://doi.org/10.1186/s40537-019-0217-0>

Preparing for the new era of data-intensive medicine

A screenshot of the BBC News website. The top navigation bar includes the BBC logo, a 'Sign in' button, and links for Home, News, Sport, Weather, iPlayer, and Sounds. Below this is a red banner with the word 'NEWS' in white. Underneath the banner is a secondary navigation bar with links for Home, Coronavirus, US Election, UK, World, Business, Politics, Tech, Science, Health, and Family & Education. The 'Tech' link is highlighted. Below the navigation bar, the word 'Technology' is underlined. The main headline reads 'Excel: Why using Microsoft's tool caused Covid-19 results to be lost'. Below the headline, it says 'By Leo Kelion' and 'Technology desk editor'. At the bottom, it shows a clock icon and the date '5 October'.

What have we learnt?

Data Science in Medicine – in a nutshell

How can we represent and interpret medical data?

Hands-on, practical experience

Topics covered:

- Statistical analysis of biomedical data
- Relational databases for medicine and healthcare
- Medical ontologies and graph data
- Epidemiology

Part 1: Statistical analysis of data

- Data scales
- Summary statistics
- Visualising data
- Hypothesis testing

Summary statistics

- Measures of **central tendency**

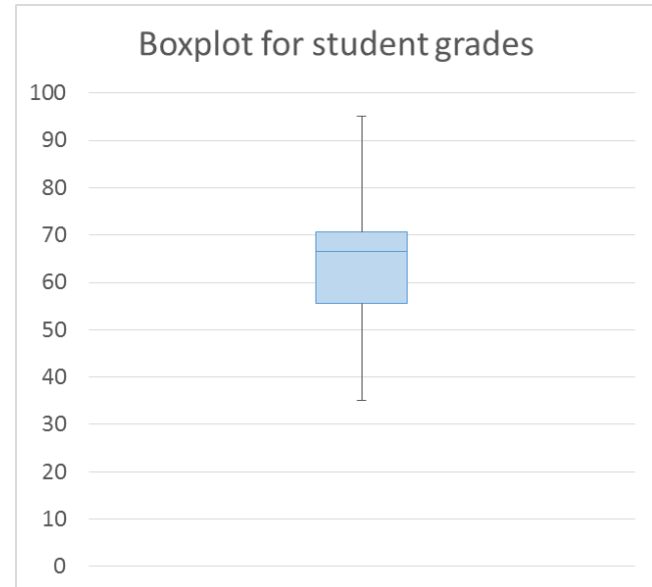
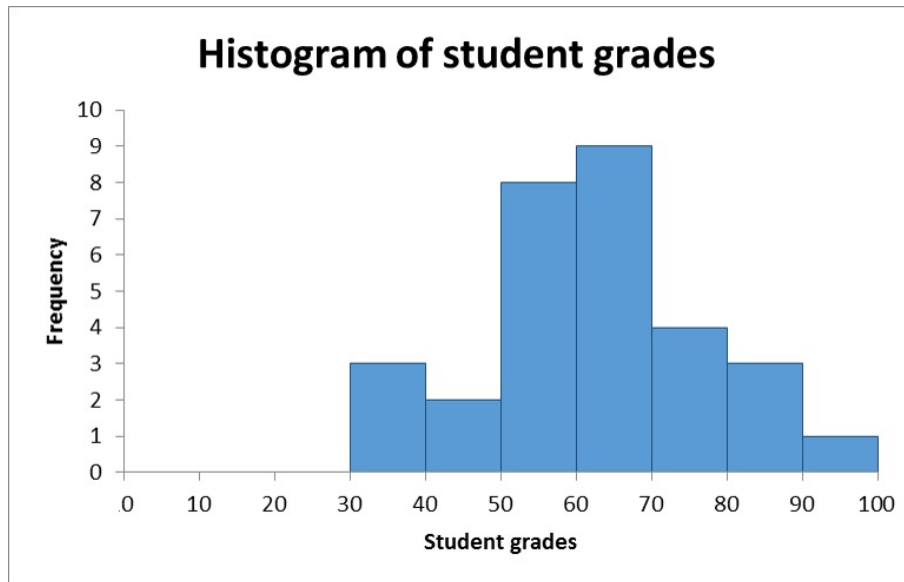
$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{69+70+86+42+54+79+69}{7} = \frac{469}{7} = 67$$

- Measures of **dispersion**

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} = \sqrt{\frac{(69-67)^2 + (70-67)^2 + (86-67)^2 + (42-67)^2 + \\ &\quad (54-67)^2 + (79-67)^2 + (69-67)^2}{7}} \\ &= \sqrt{188} = 13.71\end{aligned}$$

Visualising data

- Qualitative data: bar charts, pie charts
- Quantitative data: histograms, box plots
- Bivariate: scatter plots, line graphs

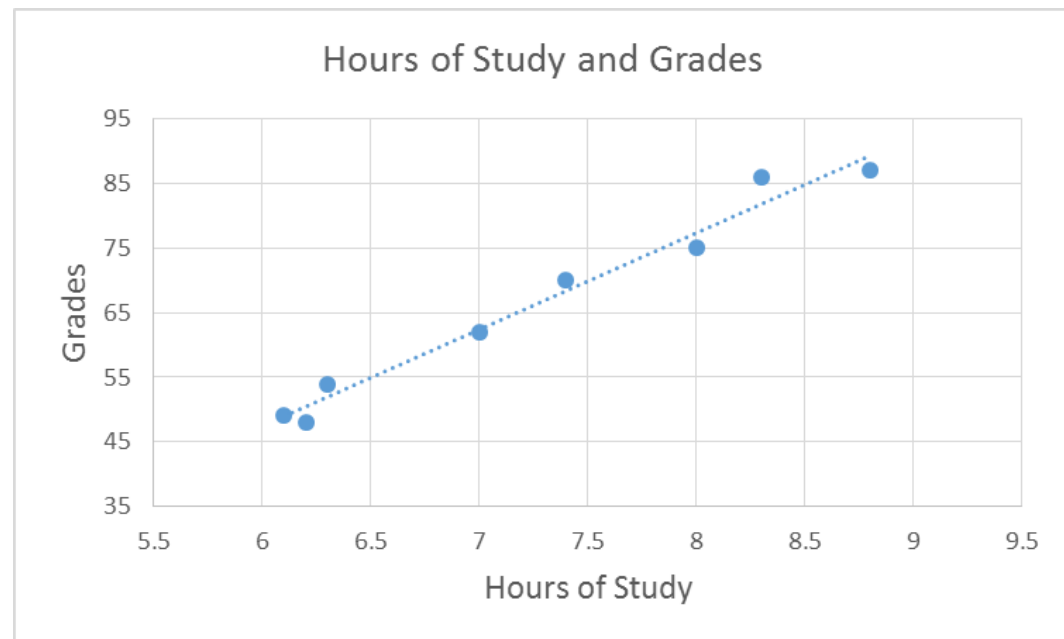


Hypothesis testing

- Correlation between numerical variables
- Association between categorical variables
- Comparing the mean of a sample to a population with a known mean
- Comparing the means of two samples that were independently drawn

Example: correlation between two numerical variables

Weekly hours of study	Grades
8	75
7.4	70
8.3	86
6.2	48
6.3	54
7	62
8.8	87
6.1	49



Example: correlation between two numerical variables

- $\rho_{x,y} \simeq 0.988$
- Hypothesis testing:
 - H0: There is no correlation between weekly hours of study and final exam grades in Statistics.
 - H1: There is a correlation between weekly hours of study and final exam grades in Statistics

ρ	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.001$
$N = 7$	0.669	0.754	0.875	0.951
$N = 8$	0.621	0.707	0.834	0.925
$N = 9$	0.582	0.666	0.798	0.898
$N = 10$	0.549	0.632	0.765	0.872

Analysing data with R

RStudio

Environment History

Global Environment

Object	Size
datasciClass	30 obs. of 4 variables
ED_visit_data	5000 obs. of 27 variables
exdata	12 obs. of 2 variables
yale_data	5000 obs. of 27 variables

Values

Variable	Range	Summary
age_of_admitted	int [1:1488]	53 25 85 69 59 59 31 72 58 37 ...
age_of_discharged	int [1:3512]	37 29 37 88 24 54 61 32 69 32 ...
birth_weights	num [1:1000]	2.73 3.27 3.99 2.72 2.74 ...
dbp_admitted	num [1:1488]	NA NA NA 75 NA 83 NA NA NA ...
EDvars	chr [1:27]	"disposition" "age" "gender" "ethnicity" "rac...
freq	'xtabs' int [1:2, 1:2]	17 15 17 23
freq_insure	'table' int [1:5(1d)]	1686 1778 1062 464 10

```
156
157 # boxplot by group
158 boxplot(datasciClass$Grades~datasciClass$Degree,
159         main = "Student grades in the Data Science class by Degree",
160         xlab = "Degree",
161         ylab = "Grade",
162         col="lightblue",
163         ylim = c(30, 100)
164       )
165
166
167 # ==> Exercise 3.2: Get the boxplot of Hours.of.sleep. The colour should be "darkseagreen".
168
169
170 # ==> Exercise 3.3: Get the boxplot of Hours.of.sleep by Degree.
171
```

Console

```
~/Documents/tutoring/MedicalInformatics/UGMedical_Informatics/2019-2020/labs/Lab1/ >
+   main = "Histogram of student grades in the Data Science class",
+   xlab = "Student grades",
+   ylab = "Frequency",
+   col = "grey",
+   xlim = c(0, 100), # change the scale of the x-axis
+   ylim = c(0, 10), # change the scale of the y-axis
+   labels = TRUE # add frequency labels to each bar
+ )
> boxplot(datasciClass$Grades~datasciClass$Degree,
+         main = "Student grades in the Data Science class by Degree",
+         xlab = "Degree",
+         ylab = "Grade",
+         col="lightblue",
+         ylim = c(30, 100)
+       )
>
```

Student grades in the Data Science class by Degree

Degree	Min	Q1	Median	Q3	Max
Informatics	55	60	70	85	90
Mathematics	40	50	65	75	85
Medicine	40	55	60	70	80
Psychology	35	45	55	65	75

Part 2: Relational databases

Employee

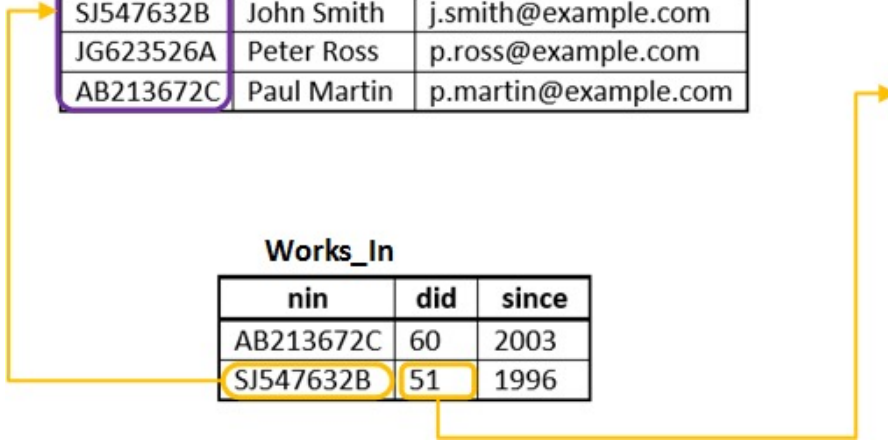
nin	name	email
SK728468L	Kate Taylor	k.taylor@example.com
SJ547632B	John Smith	j.smith@example.com
JG623526A	Peter Ross	p.ross@example.com
AB213672C	Paul Martin	p.martin@example.com

Department

did	dname	budget
51	Information Technology	80,000
56	Human Resources	50,000
60	Accounting	40,000

Works_In

nin	did	since
AB213672C	60	2003
SJ547632B	51	1996



Part 2: Relational databases

- How to build a database: Relational model
- How to query a database: SQL

Relational model

Student

mn	name	email	age
s0785212	Andrew	andrew@maths	19
s1253477	Jenny	jenny@inf	23
s1456381	Rhona	rhona@inf	18
s1489673	Stuart	stuart@law	34
s1473612	Alan	alan@law	20

Course

cid	title	credits
db5	Database Systems	20
inf1	Informatics 1	10
sls	Scottish Legal System	10
lalg	Linear Algebra	10

Takes

mn	cid
s0785212	lalg
s1253477	db5
s1253477	inf1
s1489673	sls

```
CREATE TABLE Takes (  
  mn CHAR(8),  
  cid CHAR(20),  
  PRIMARY KEY (mn, cid),  
  FOREIGN KEY (mn) REFERENCES Student,  
  FOREIGN KEY (cid) REFERENCES Course  
)
```

SQL querying

```
SELECT *  
FROM Student  
WHERE age > 19
```

Student

mn	name	email	age
s0785212	Andrew	andrew@maths	19
s1253477	Jenny	jenny@inf	23
s1456381	Rhona	rhona@inf	18
s1489673	Stuart	stuart@law	34
s1473612	Alan	alan@law	20

Course

cid	title	credits
dbs	Database Systems	20
inf1	Informatics 1	10
sls	Scottish Legal System	10
lalg	Linear Algebra	10

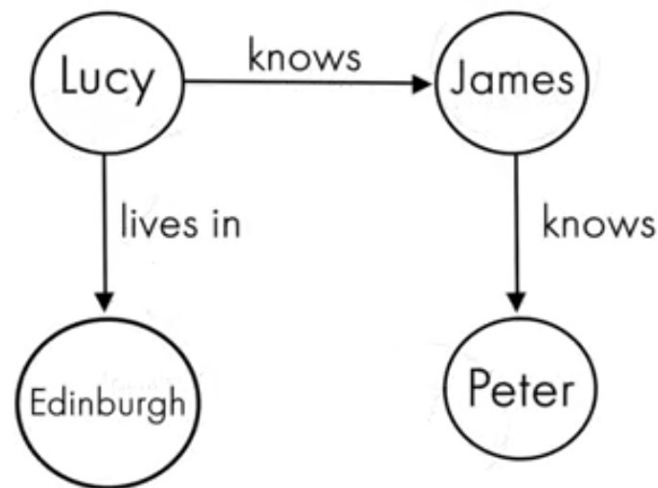
Takes

mn	cid
s0785212	lalg
s1253477	dbs
s1253477	inf1
s1489673	sls

```
SELECT S.email  
FROM Student S, Takes T, Course C  
WHERE S.mn = T.mn  
      AND T.cid = C.cid  
      AND C.title = 'Medical Informatics'
```

Part 3: Medical ontologies and graph data

- Graph databases follow an alternative data representation approach to relational databases.
- The objective here is to easily integrate data.



Part 4: Epidemiology

- Measuring the occurrence of disease
- Evaluating treatment and prognosis
- Assessing risk of disease
- Determining cause and reporting research

Disease prevalence

MEASURING THE OCCURRENCE OF DISEASE IN EPIDEMIOLOGY



1. Prevalence of disease

- Very simple measure
- **Cross-sectional study** results in prevalence estimates
- It can not be smaller than **0%** or greater than **100%**
- In theory, it is calculated as:

$$P = \frac{\text{Number of persons with disease}}{\text{Number of people checked for presence of disease}}$$

- What could possibly be more simple than prevalence?

DIFFERENT WAYS TO EXPRESS THE MEASURED PREVALENCE IN A SAMPLE



1. **Point prevalence:** study can be conducted over a period of few months, but the result is still “point prevalence” – from the perspective of EACH SUBJECT, the information is on one point in time only, and **counts only those with active symptoms of disease at that point in time;**
2. **Period prevalence:** if we asked about presence of active symptoms **within the past e.g. 6 months or 3 years;**
3. **Lifetime prevalence:** if we asked whether there **were ever any symptoms of schizophrenia**, regardless of the status of disease in the present (e.g. medication or remission);
4. **Lifetime morbid risk:** if we followed everyone in the sample until they die, and then **added all further new cases** to the already noted lifetime prevalence

Next steps, if of interest

- Explore real data and practise further
 - WHO COVID-19 data: <https://covid19.who.int/>
 - Public Health Scotland open data:
<https://www.opendata.nhs.scot/>
 - NHS England data: <https://data.england.nhs.uk/>
- Develop further skills in R programming
 - HealthyR book:
https://argoshare.is.ed.ac.uk/healthyr_book/
 - Free online courses on Coursera, edX, Datacamp and other platforms

Next steps, if of interest

- Engage with the Data-Driven Innovation programme

- Innovative training
- World-class data infrastructure



- Free online course: **Data Science in Stratified Healthcare and Precision Medicine**

- 5 weeks, self-paced, free
- 11,900+ learners worldwide
- <https://www.coursera.org/learn/datascimed/>



Thank you!

Time to play!

Enter our kahoot.it quiz to
win a place in the DSM “Hall of Fame”