Workshop on GWAS

Introduction to data science and COVID-19 motivating examples

Prof. Lei Sun[1]

Department of Statistical Sciences, FAS

Division of Biostatistics, DLSPH

University of Toronto

21 June, 2021

[1]A big thank you to Boxi Lin who were "on-call" for my Rmarkdown questions.

What is Data Science (Wiki, February 19, 2021)

Data science is an **inter-disciplinary** *field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.Data science is related to data mining, machine learning and big data.*

Data science is a "concept to **unify statistics, data analysis and their related methods**" *in order to "understand and analyze actual phenomena" with data.*

It uses techniques and theories drawn from many fields within the context of **mathematics, statistics, computer science, domain knowledge and information science**.

A great resource from Professor Kerby Shedden

Introduction to Data Science

Materials, e.g.

What is data science *Data science is the study of methods for drawing meaningful insight from data. It is a methodological subject, because it primarily deals with the techniques that we use to learn things, and not the specific domains where the techniques are applied.*

Types of data

Association

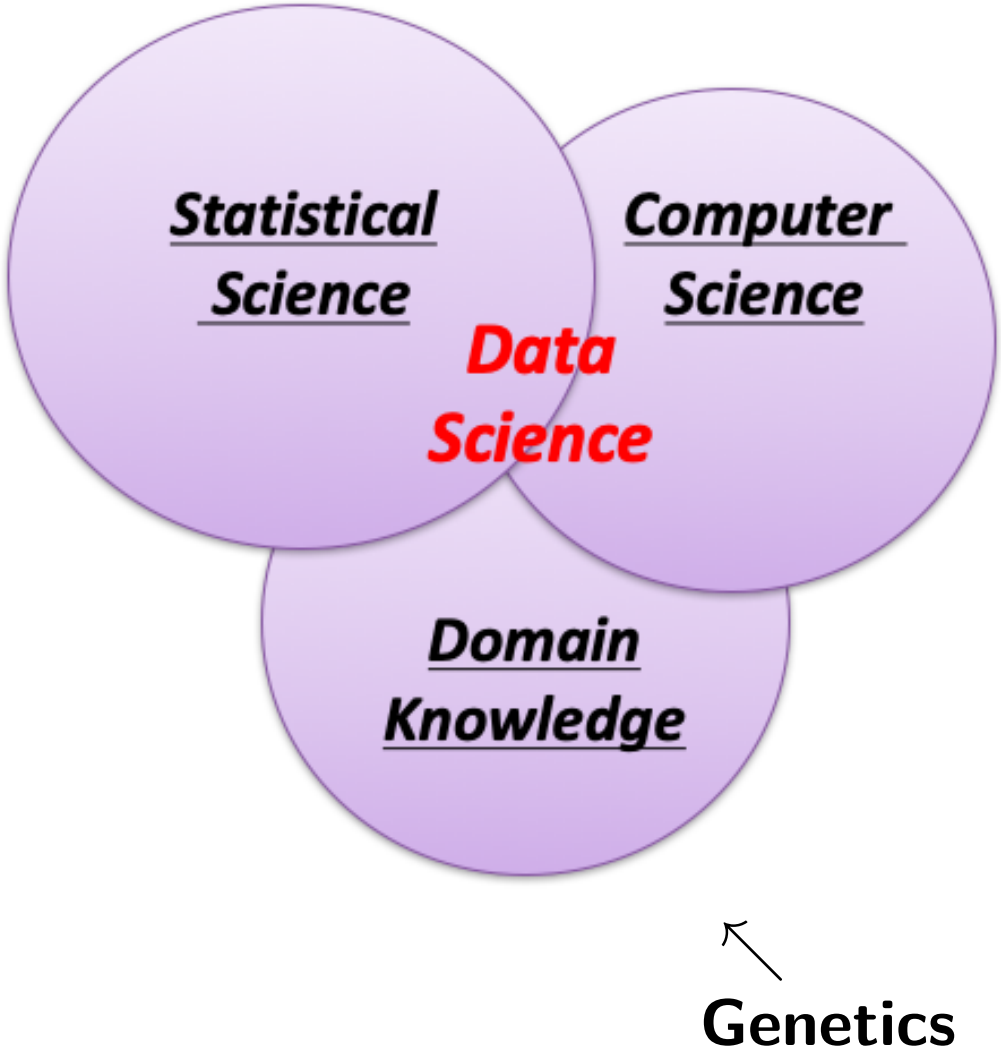
. . . . . .

Hypothesis testing

. . . . . .

A Ebook on Data Science

My take on Data Science



HDSR: A Research + Education Journal

Data Science Education
Reproducibility and Replicability
AI and Responsible Data Science
.....

On R = statistical computing and R Markdown = reproducible data science

R is a free software environment for statistical computing and graphics. *It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.*

Not going to study or do research in statistics? e.g. Computing for Social Science and Why R?

R Markdown provides an authoring framework for data science.
You can use a single R Markdown file to both
- *save and execute code*
- *generate high quality reports that can be shared with an audience*

R Markdown documents are fully reproducible *and support dozens of static and dynamic output formats.*

On responsible data science, e.g. genetic diversity and ancestry

Akinyemi Oni-Orisan, Yusuph Mavura, Yambazi Banda, Timothy A Thornton, Ronnie Sebro (2021). *New England Journal of Medicine*
Embracing Genetic Diversity to Improve Black Health

A COVID-19 motivating example for learning statistics

An Illustration of the Base Rate Fallacy or Bayes' Rule to understand COVID-19 antibody testing result

If you took an antibody test that was 90 percent accurate, and it determined that you had coronavirus antibodies, how confident should you be that you actually have those antibodies? Most people say about 90 percent, with the average answer being above 50 percent. This makes sense. After all, 90 percent accuracy is pretty high. NYT, May 13th 2020

So, how do you explain that, **if you took the test and received a positive test result, the probability that you actually have COVID antibody is $<$50%**!

An intuitive illustration[2] for the counter-intuitive result

- Assume a total of 110 people (this number is not that important)
- 9.1% truly have the COVID antibody (this proportion is critical). So,
- 10 people (all O's) truly have the antibody and the rest 100 people as X's.

10 Os: people with covid antibodies, of which
9 Os: tested positive (90% accurate)

O O O O O O O O O O

100 Xs: people without covid antibodies, of which
10 Xs: tested positive (90% accurate)

X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X

- Among the 19 people who were tested positive: 9 red O's and 10 red X's,
- **the proportion of red O's is** $\frac{9}{19} = 47.3\% < 50\%$**!**

[2] A challenge: create educational graphics or videos for the general public.

If you knew statistics, use the Bayes' rule formally

- $K = \Pr(\text{COVID})$: **population prevalence** (also known as critical base rate), e.g. about 9.1% (=10/110) of the population truly have COVID antibodies.
- $TPR = \Pr(\text{positive test result} \mid \text{COVID})$: **true positive rate**, e.g. the antibody test is 90% accurate if you have COVID antibody.
- $TNR = \Pr(\text{negative test result} \mid \text{no COVID})$: **true negative rate**, e.g. the antibody test is also 90% accurate if you do not have COVID antibodies.

  $FPR = 1 - TNR = \Pr(\text{positive test result} \mid \text{no COVID})$: **false positve rate**

$$\Pr(\text{COVID} \mid \text{tested positive}) = \frac{\Pr(\text{COVID and tested positive})}{\Pr(\text{tested positive})}$$

$$= \frac{\Pr(\text{COVID and tested positive})}{\Pr(\text{COVID and tested positive}) + \Pr(\text{no COVID and tested positive})}$$

$$= \frac{\Pr(\text{tested positive} \mid \text{COVID}) \cdot \Pr(\text{COVID})}{\Pr(\text{tested positive} \mid \text{COVID}) \cdot \Pr(\text{COVID}) + \Pr(\text{tested positive} \mid \text{no COVID}) \cdot \Pr(\text{no COVID})}$$

$$= \frac{TPR \cdot K}{TPR \cdot K + (1 - TNR) \cdot (1 - K)} = \frac{TPR \cdot K}{TPR \cdot K + FPR \cdot (1 - K)}$$

Bayes' rule in action

1. **The link with the previous graphic representation**,

$$\Pr(\text{COVID} \mid \text{tested positive}) = \frac{TPR \cdot K}{TPR \cdot K + (1 - TNR) \cdot (1 - K)}$$

$$= \frac{0.9 \times 0.091}{0.9 \times 0.091 + (1 - 0.9) \times (1 - 0.091)} = \frac{0.0819}{0.0819 + 0.0909} = \frac{0.0819}{0.1728}$$

$$= \frac{0.0819}{0.0819 + 0.0909} \times \frac{110}{110} \approx \frac{9}{9 + 10} \approx 47\%.$$

2. **If the population prevalence is lower, even the test accuracy increases, the final probability can be substantially smaller than 50%**,

e.g. $K = 2\%$, $TPR = 99\%$ and $TNR = 95\%$ (i.e. $FPR = 5\%$),

$$= \frac{0.99 \times 0.02}{0.9 \times 0.02 + (1 - 0.95) \times (1 - 0.02)} = \frac{0.0198}{0.0198 + 0.049} = \frac{0.0198}{0.0688} \approx 28.8\%.$$

3. **If the prevalence rate K goes up**, then we will be more concerned about the false negative rate, Pr(tested negative | COVID).

Another COVID-19 motivating example[3] for learning statistics

Kügelgen et al. (2021). Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects. *e-prints posted on arXiv are not peer-reviewed by arXiv*.

Code and data for the Simpson's paradox in Covid-19 case fatality rates

Comparing data of 44,672 cases from China with early reports from Italy (9th March), we find that Covid-19 case fatality rates are lower in Italy for every age group, but higher overall.

Confounding and Simpson's paradox Quiz: what is the explanation? (The answer is in the title.) Any other real-life examples?

[3] a big thank you to Michal Malyska who answered my call for covid-related examples, through the departmenal Slack teaching channel, and provided other good teaching examples as well!

But we are here to learn genetics!

Severe Covid-19 GWAS Group (2020). ***The New England Journal of Medicine*****. Genomewide Association Study of Severe Covid-19 with Respiratory Failure.** (Cited by 382 PubMed articles as of June 16, 2021)

We conducted ***a genomewide association study*** *involving 1980 patients with Covid-19 and severe disease (defined as respiratory failure) at seven hospitals in the Italian and Spanish epicenters of the SARS-CoV-2 pandemic in Europe.*

After quality control, ***835 patients and 1255 control participants from Italy*** *and* ***775 patients and 950 control participants from Spain*** *were included in the final analysis. In total, we* ***analyzed 8,582,968 single-nucleotide polymorphisms*** *and conducted a meta-analysis of the two case-control panels.*

Quiz: who should be the controls or individuals do not have the disease? (Hint: infectious disease).

Severe Covid-19 GWAS Group (2020) cont'd

At the end of this workshop, you will have a basic understanding of the following results!

We detected cross-replicating associations with ***rs11385942 at locus 3p21.31*** *and with rs657152 at locus 9q34.2, which were* ***significant at the genomewide level*** *(P [**p-value**]* $< 5 \times 10^{-8}$*).*

The association signal at locus 9q34.2 coincided with the ABO blood group locus; a blood-group-specific analysis showed a protective effect in blood group O as compared with other blood groups.
(Quiz: potential confounders here?)

Another GWAS of COVID-19

Pairo-Castineira et al. (2021). ***Nature*****. Genetic mechanisms of critical illness in COVID-19.**

We report the results of the GenOMICC (Genetics Of Mortality In Critical Care) ***genome-wide association study*** *in 2,244 critically ill patients with COVID-19 from 208 UK intensive care units.*

Ancestry-matched control individuals *were selected from the large population-based cohort of UK Biobank (****five controls were included for each case****). Controls with a known positive COVID-19 test were excluded.*

Similar quiz: How to define controls in the context of infectious disease?

Pairo-Castineira et al. (2021) cont'd

We have identified and replicated the following new genome-wide significant associations:

on chromosome 12q24.13 (rs10735079, $P = 1.65 \times 10^{-8}$) in a gene cluster that encodes antiviral restriction enzyme activators (OAS1, OAS2 and OAS3);

on chromosome 19p13.2 (rs74956615, $P = 2.3 \times 10^{-8}$) near the gene that encodes tyrosine kinase 2 (TYK2);

on chromosome 19p13.3 (rs2109069, $P = 3.98 \times 10^{-12}$) within the gene that encodes dipeptidyl peptidase 9 (DPP9);

and on chromosome 21q22.1 (rs2236757, $P = 4.99 \times 10^{-8}$) in the interferon receptor gene IFNAR2.

A Canadian effort: Covid-19 CGEn Host Genome Sequencing Project of the Canadian COVID Genomics Network.

At the end of this workshop[4]

√ Successfully run a genome-wide association study (GWAS), albeit a black-boxed approach by using the PLINK software. But, you will

√ Learn about variations in the human genome and the structure of the human population from Dr. Andrew Paterson.

√ Have fun with the 1000 Genomes project data and obtain >1 million p-values, using the manual created by Anton Sugolov, Eric Emmenegger. GWAS hands-on materials. And we will peek into the black-box, and

√ Have a basic understanding of regression, association test and one single p-value.

√ Have a basic understanding of multiple hypothesis testing and dealing with >1 million p-values simultaneously.

√ Have a conceptual understanding of unsupervised vs. supervised learning, principle component analysis (dimension reduction), overfitting, and confounding and Simpson's paradox!

---

[4] Monday/June 21-Thursday/June 24 2021: 9:30-noon lectures (open to all UTS S5 and S6 students; Zoom ID (code) = 826 6918 2798 (741780)); 1-3:30pm hands-on sessions (only open to the students who have registered by May 25th 2021); Friday/June 25: 10-noon Q&A and meet with Prof. Radu Craiu, Chair of the Department of Statistical Sciences.