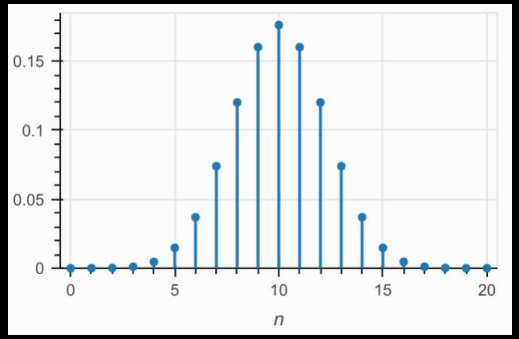
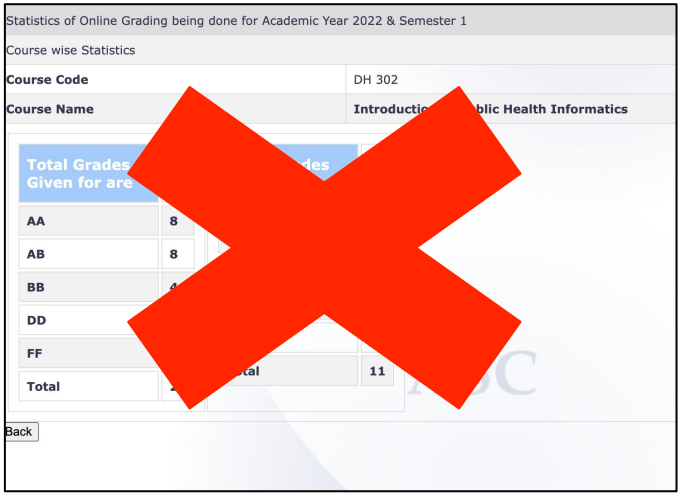
Statistical models for health* data - I

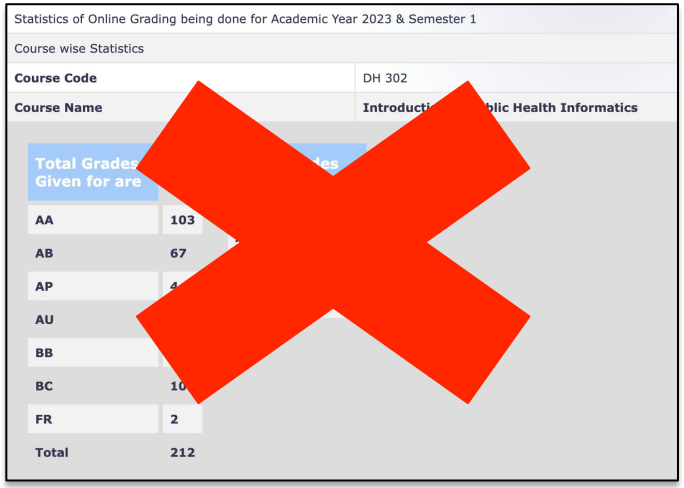


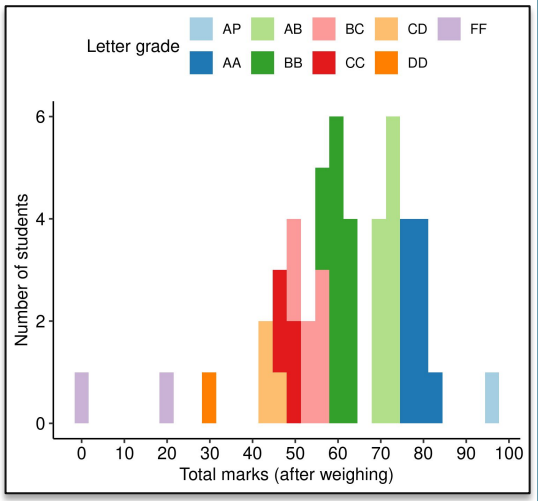
Saket Choudhary
saketc@iitb.ac.in
Introduction to Public Health Informatics
DH 302
Lecture 02 || Friday, 10th January 2025

Logistics review

If you are here because of this, rethink…







Logistics upto Midsems (50%)

- **Assignments: 10% (Best 2/3 out of 3/4)**
  - Due + graded online
  - Best 2 (or 3) will be considered for grading out of 3 (or 4)
  - Late submission policy: 10% penalty per day
  - One submission per student (Attribute if you discussed with someone)
- **Mid-sem: 30%**
  - Closed book and offline (no collaboration)
- **Surprise quizzes: 10%**
  - Based on content taught in the past
  - Participation points if you interact (and not just come to the class)
- **Class participation:** Relative bonus points (Scaled to the maximum total)
- Possibility of absolute bonus points

Final grades: RG (Relative grading)

Logistics - Office hour(s)

- Lecture: Wednesdays and Fridays, 11:05am – 12:30pm || LH 101
- Instructor Office: G-22, KCDH, KReSIT Basement
- Instructor Office Hours: **Wednesdays, 4:00 - 5:00pm** or by appointment
- For appointments outside office hours: https://cal.com/saketkc/
- Contact: saketc@iitb.ac.in | Ext: 3785 (+91 22 2159 3785)

Use email preferably only for personal requests - if you have a question, someone else might also have a similar one.

!!! Collaboration policy and Academic Integrity !!!

- You are expected to work on your own for most part of the course.
- For assignment problems, If you get stuck, you are welcome to discuss it with other students (in-person or online). However, the **solutions must be your work**. If you discussed with someone, **please mention their name and what you received help with** in your submission. If you do not attribute and we find similarities in the final submissions - **this will automatically count as plagiarism**!
- **Mid-semester exam (closed book). No collaboration is allowed.**
- **Write/speak what you understand.** If you write something, it is assumed you understand it - and hence are open to being quizzed by it
- Simply: **DTRT - Do the right thing**

"I declare that I will adhere to all principles of academic honesty and integrity throughout my stay in the Institute. I will not seek or give unauthorized assistance in tests, quizzes, examinations or assignments. I will not misrepresent, fabricate or falsify any idea/data/fact/source in my project submissions. I understand that any violation of the above will be cause for disciplinary action as per the rules and regulations of the Institute."

See Policy

TAs and office hours

Anisha Karmakar
23D1622@iitb.ac.in
Friday, 3-4 pm, BSBE (Lab 605)

Chetan Patil
20b030012@iitb.ac.in
Wednesday, 2-3 PM, KCDH Lab

Devendra Singh
devendrasb@iitb.ac.in
Friday, 5-6 PM KCDH Lab

Kriti A
210100083@iitb.ac.in
Tuesdays, 5-6 PM, ME Department

Shobhit Aggarwal
20d100026@iitb.ac.in
Wednesdays,4-5PM, KCDH Lab

Sunny Gupta
sunnygupta@iitb.ac.in
Friday, 4-5 pm, Medal EE

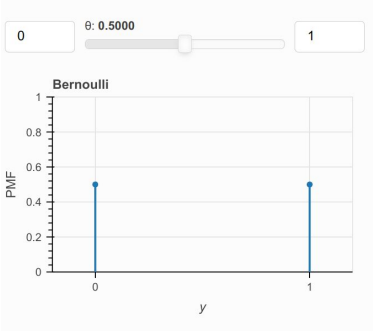
Course vignettes

Probability models for health data

- How should the number of Covid-19 cases be modeled?
- What is the correct statistical model for representing deaths as a function of time?
- What is the distribution of height of males in a village? What about children in village? What about children in a village known to be suffering from stunting?

How to think about distributions? The most important ones..

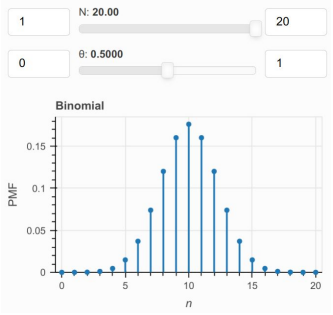
Discrete



$$f(y;\theta) = \begin{cases} 1-\theta & y = 0 \\ \theta & y = 1. \end{cases}$$

Mean: θ

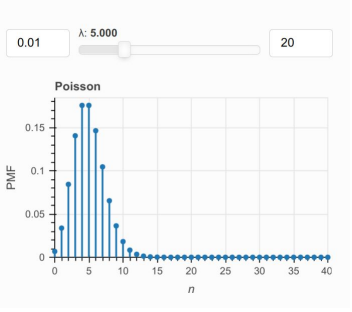
Variance: $\theta(1-\theta)$



$$f(n;N,\theta) = \binom{N}{n}\theta^n(1-\theta)^{N-n}.$$

Mean: $N\theta$

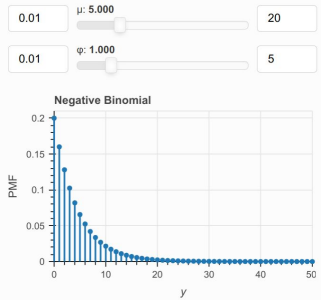
Variance: $N\theta(1-\theta)$



$$f(n;\lambda) = \frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda}.$$

Mean: λ

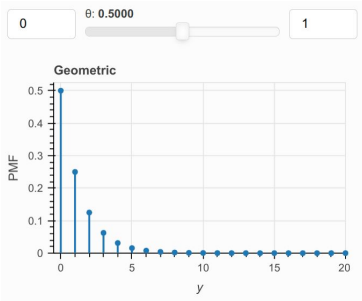
Variance: λ



$$f(y;\mu,\phi) = \frac{\Gamma(y+\phi)}{\Gamma(\phi)\,y!}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}.$$

Mean: μ

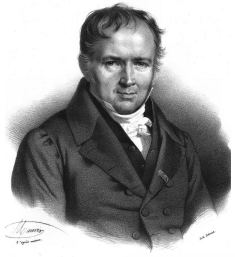
Variance: $\mu\left(1+\frac{\mu}{\phi}\right)$.



$$f(y;\theta) = (1-\theta)^y\,\theta.$$

Mean: $\frac{1-\theta}{\theta}$

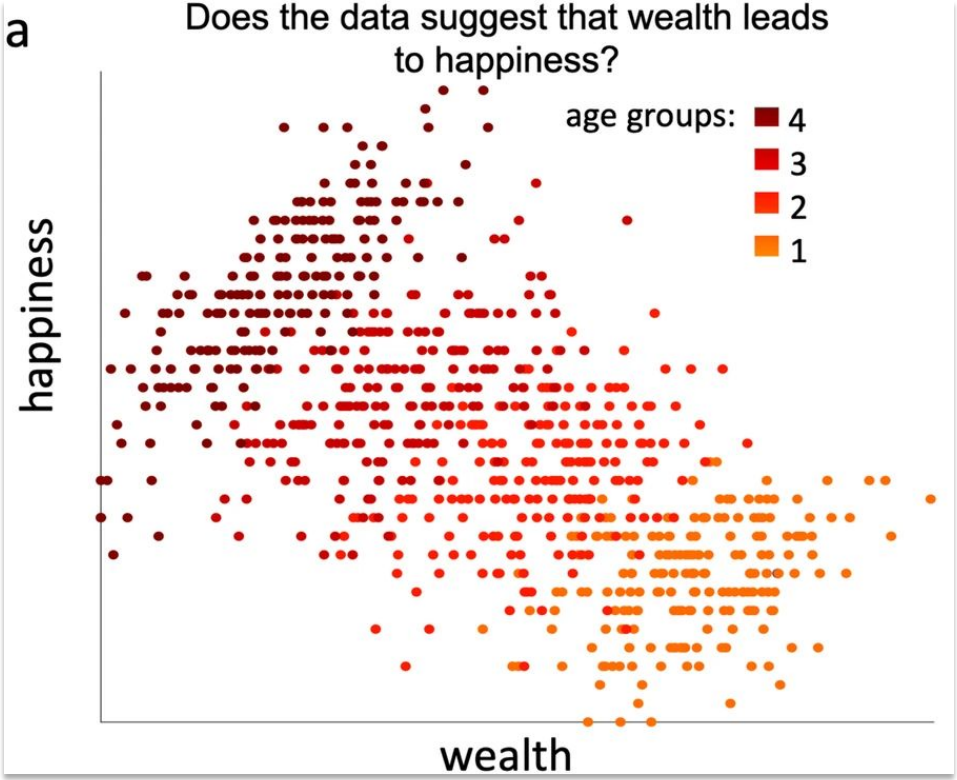
Variance: $\frac{1-\theta}{\theta^2}$



Simeon Poisson,

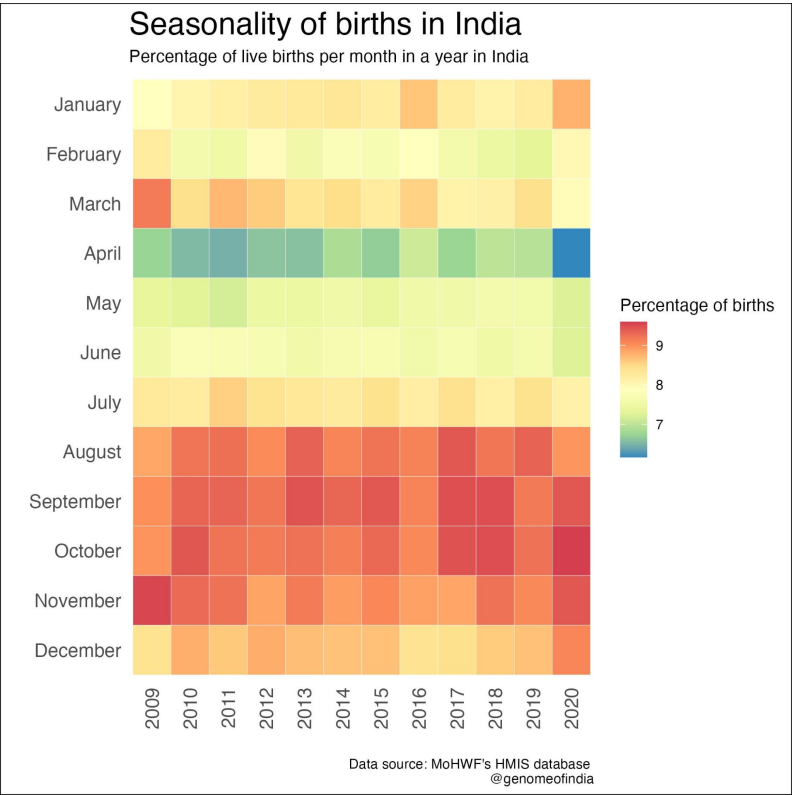
https://distribution-explorer.github.io/

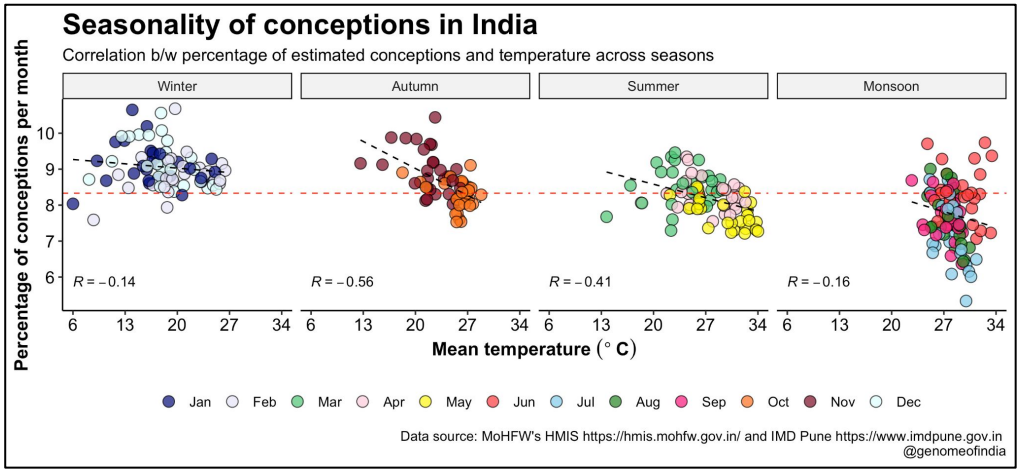
Exploratory data analysis



Exploratory data analysis

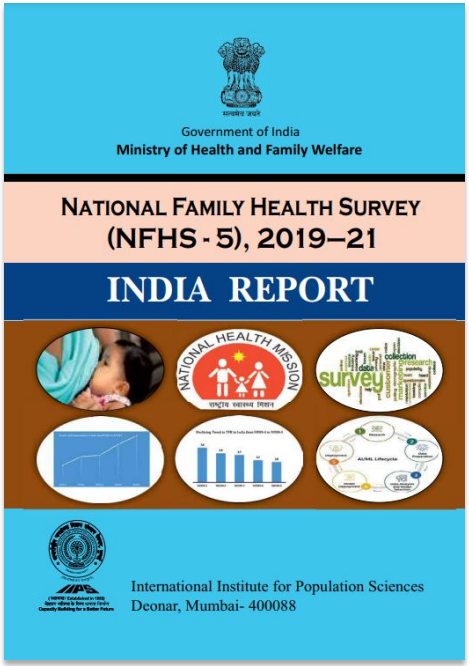
When is your birthday? 1st half or 2nd half?

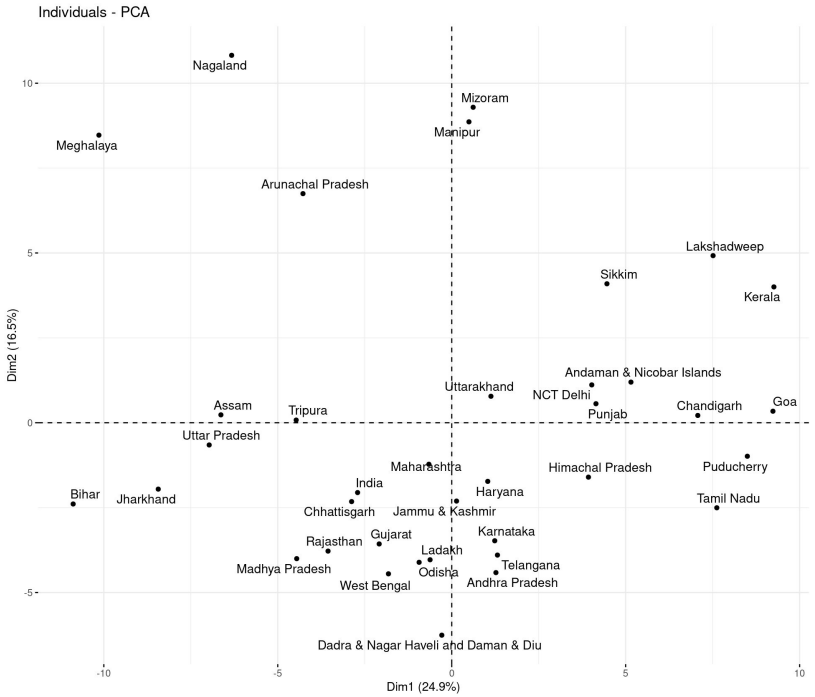




https://genomeofindia.substack.com/p/wake-me-up-when-august-ends

Dimensionality reduction for health care data



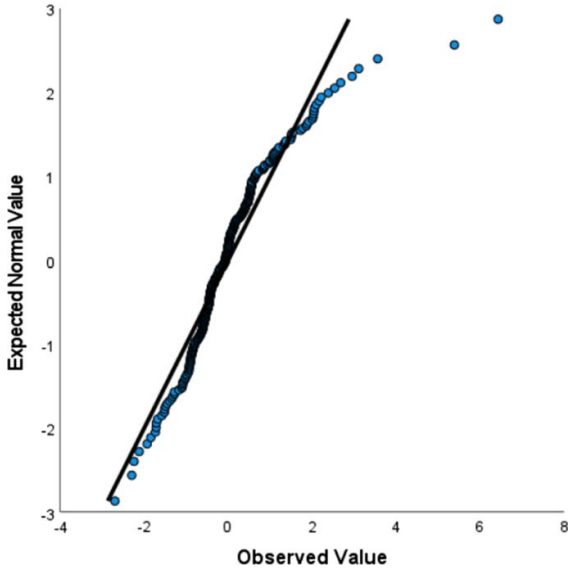


Problem: Are different states performing similarly on multiple health metrics?

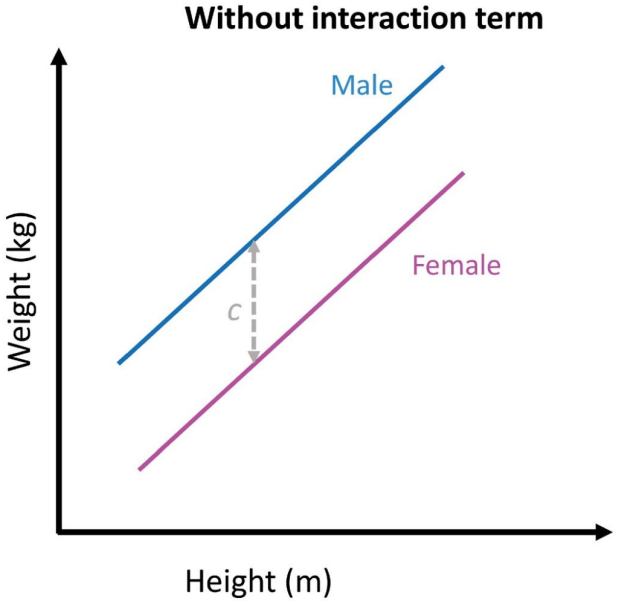
Solution:Dimensionality reduction, clustering

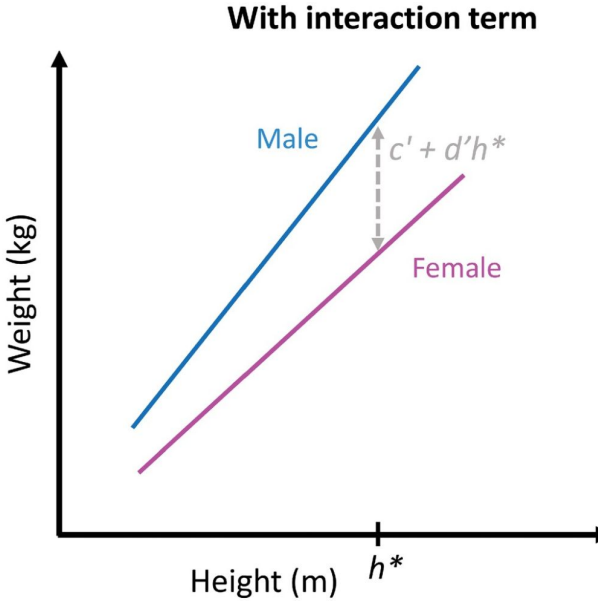
Multivariate regression for health-related data

Problem: Is the relationship between height and weight similar across males and females? Is it linear?



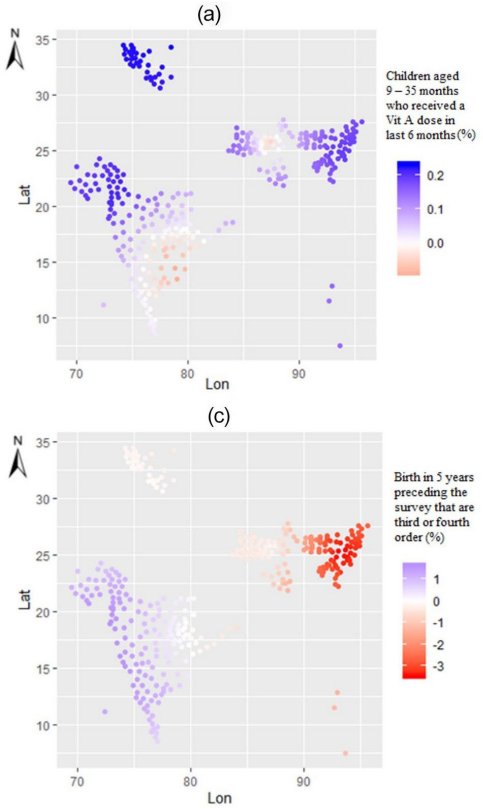
Source



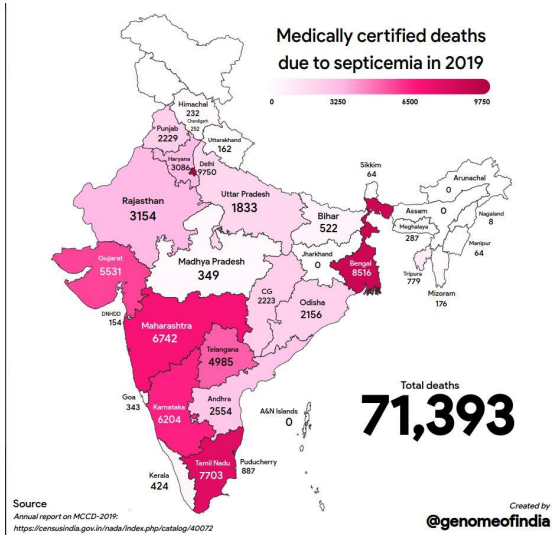


Source

Multivariate regression for health-related data



Source

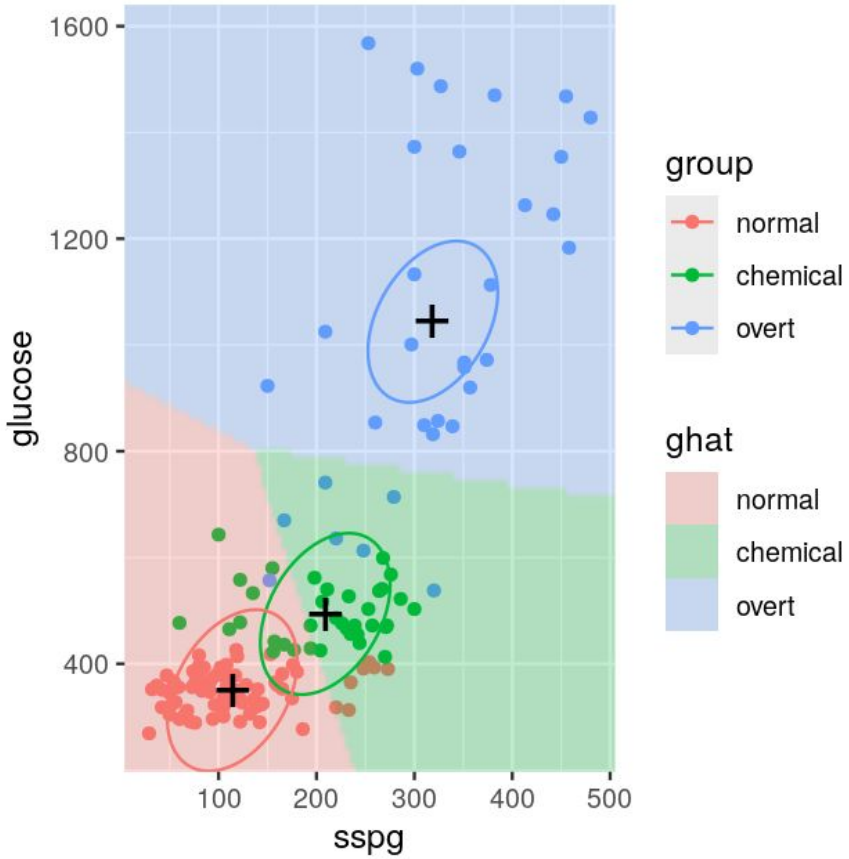


Problem: What are factors associated with disease X in India?

Solution: Geographically aware regression to identify factors and spatial clusters

Source

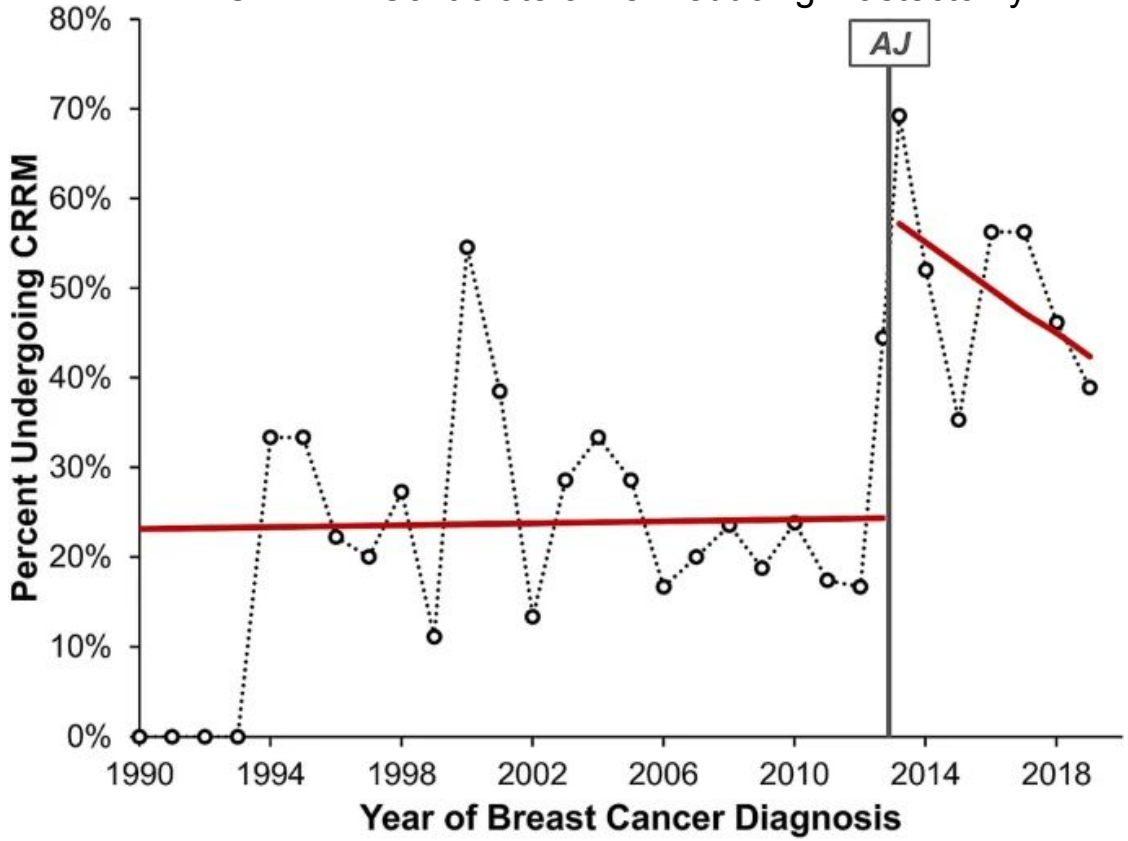
Classification problems in healthcare



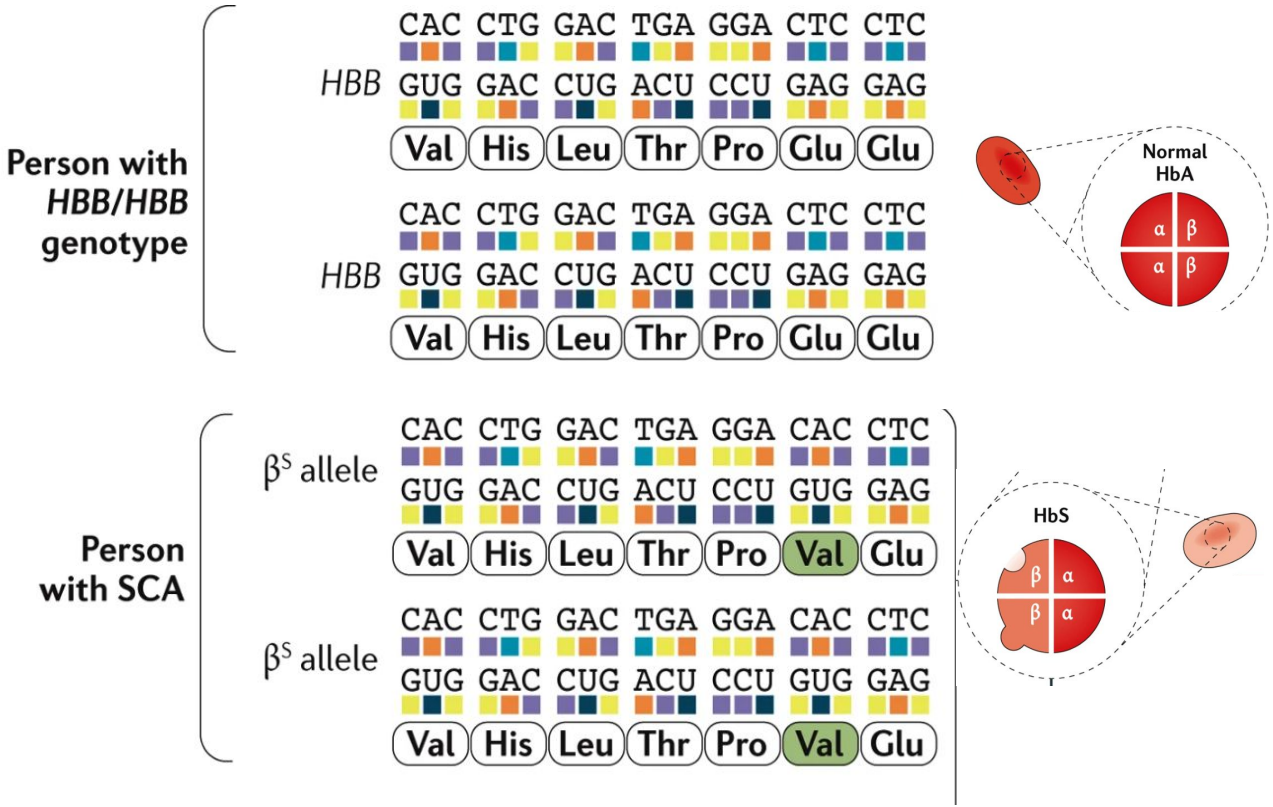
Problem: Predicting the severity of disease (or no-disease)

The Angelina Jolie Effect

CRRM = Contralateral risk-reducing mastectomy



Genomics for health

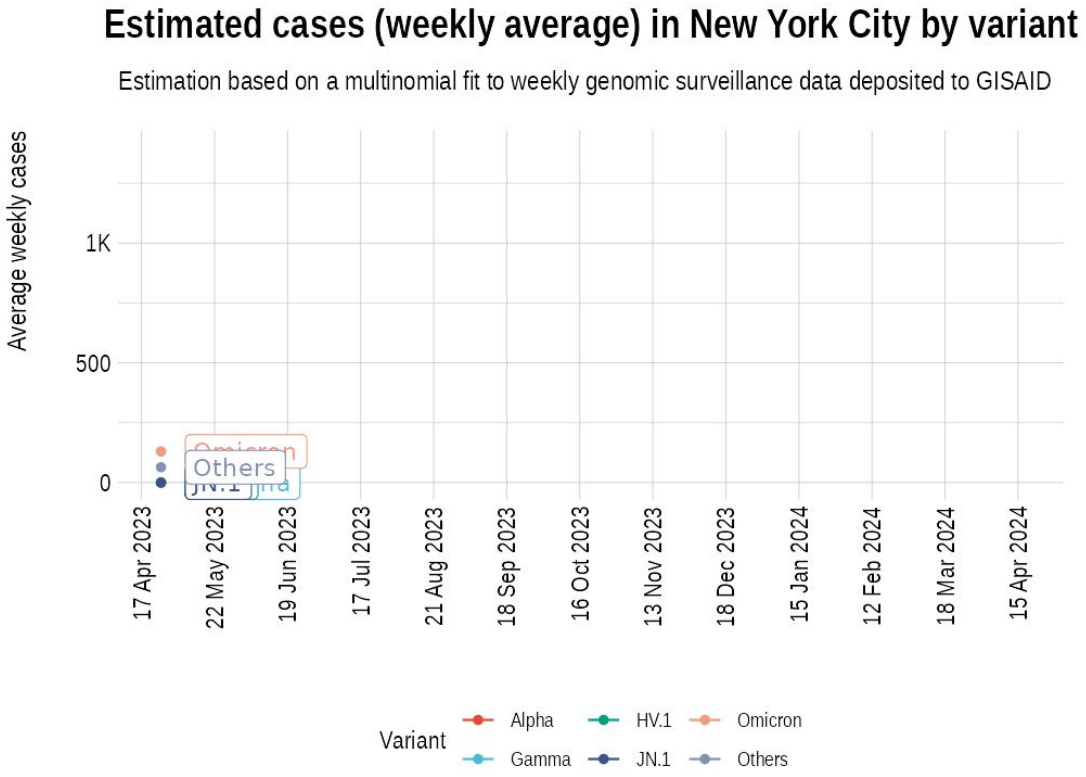


Problem: How are DNA mutations identified?

NHGRI

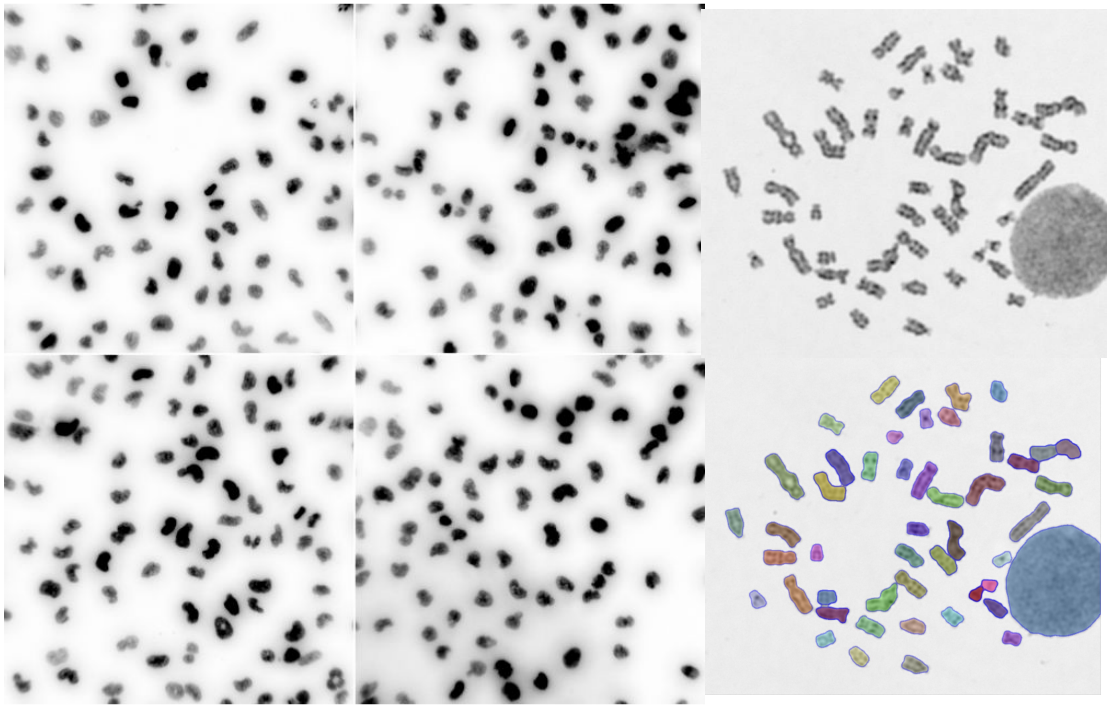
Genomics for health

Problem: Which variants are (were) driving the uptick in Covid-19 cases?



https://saket-choudhary.me/covmuller

Analysis of image data for healthcare



Problem: Extracting features from medical images; Classifying images into grades of diseases

Short(est) introduction to public health informatics

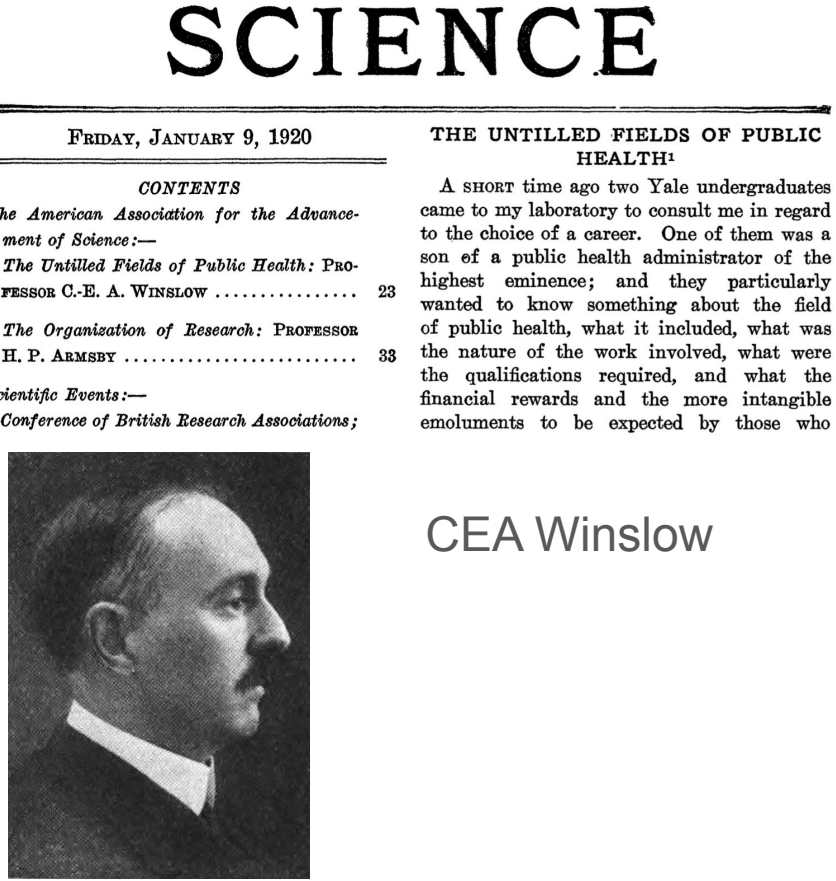
What is health?

Health is a state of **complete** physical, mental and social well-being and not merely the absence of disease or infirmity – WHO

Complete = Holistic and not “perfect” health

Schramme 2023

What is public health?



SCIENCE

FRIDAY, JANUARY 9, 1920

CONTENTS

The American Association for the Advancement of Science:—

The Untilled Fields of Public Health: PROFESSOR C.-E. A. WINSLOW 23

The Organization of Research: PROFESSOR H. P. ARMSBY 33

Scientific Events:—

Conference of British Research Associations;

THE UNTILLED FIELDS OF PUBLIC HEALTH[1]

A SHORT time ago two Yale undergraduates came to my laboratory to consult me in regard to the choice of a career. One of them was a son of a public health administrator of the highest eminence; and they particularly wanted to know something about the field of public health, what it included, what was the nature of the work involved, what were the qualifications required, and what the financial rewards and the more intangible emoluments to be expected by those who

CEA Winslow

"*Public health is the science and* ***the art of preventing disease, prolonging life, and promoting physical health and efficiency*** *through organized community efforts for the sanitation of the environment, the control of community infections, the education of the individual in principles of personal hygiene, the organization of medical and nursing service for the early diagnosis and preventive treatment of disease, and the development of the social machinery which will ensure to every individual in the community a standard of modest duties which their task entails. living adequate for the maintenance of health*"

– CEA Winslow, Science (1920)

The ten essential public health services

Assessment

1. Monitor health status to identify community health problems
2. Diagnose and investigate health problems and health hazards in the community

Policy Development

3. Inform, educate, and empower people about health issues
4. Mobilize community partnerships to identify and solve health problems
5. Develop policies and plans that support individual and community health efforts

Assurance

6. Enforce laws and regulations that protect health and ensure safety
7. Link people to needed personal health services and assure the provision of health care when otherwise unavailable
8. Assure a competent public health and personal healthcare workforce
9. Evaluate effectiveness, accessibility, and quality of personal and population-based health services

Serving All Functions

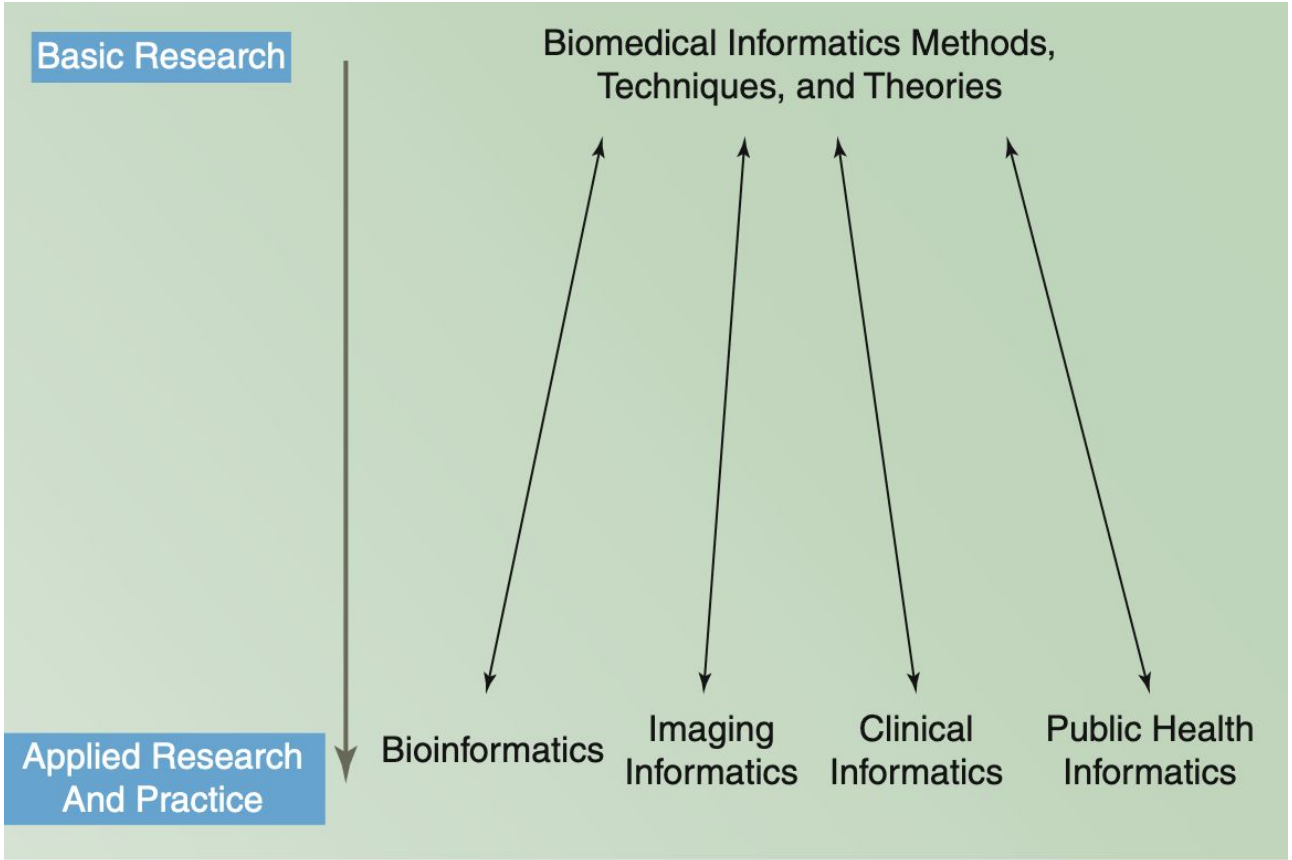
10. Research for new insights and innovative solutions to health problems

Source

What is health informatics?

Public health informatics = Systematic application of data, technology, and information systems to public health practice and research

Goal: Use data to improve efficiency, accuracy and diagnosis of health care needs

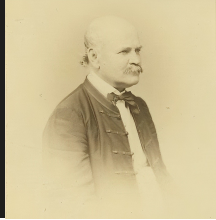


Source

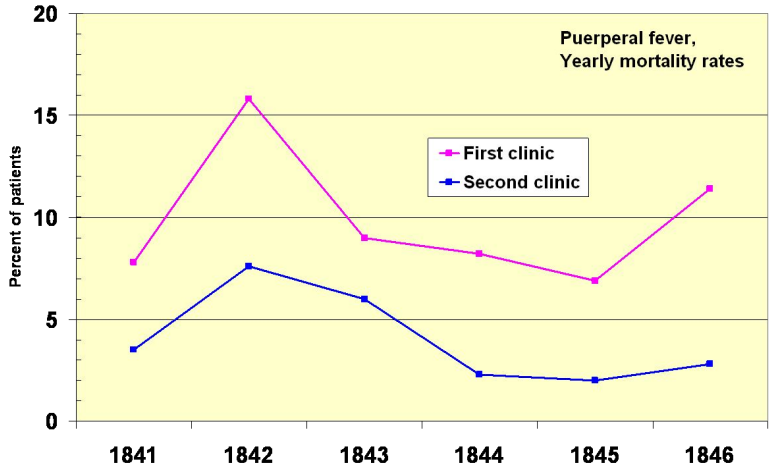
Story time

Blast from the past

Ignaz Semmelweis: Saviour of Mothers



- In 19th century, postpartum infection also called puerperal fever was common cause of death with its reason largely remaining unknown (germ theory had not been proven yet!)
- Iganz was appointed as an assistant physician at vienna general hospital → came across pregnant women “begging” to not be admitted to the “First” clinic
- First clinic mortality rate: 10%, Second clinic (just down the road, more crowded): 4%
- Mortality rate of women giving birth on road (to avoid the “First”): << 10%



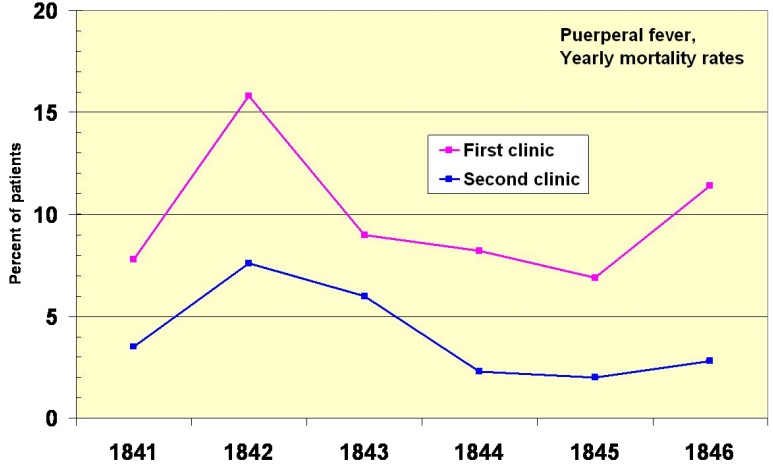
Why is the mortality different across the two clinics?

First clinic had doctors training who would conduct everything; Second clinic had mostly midwives assisting in birth

Source
Source

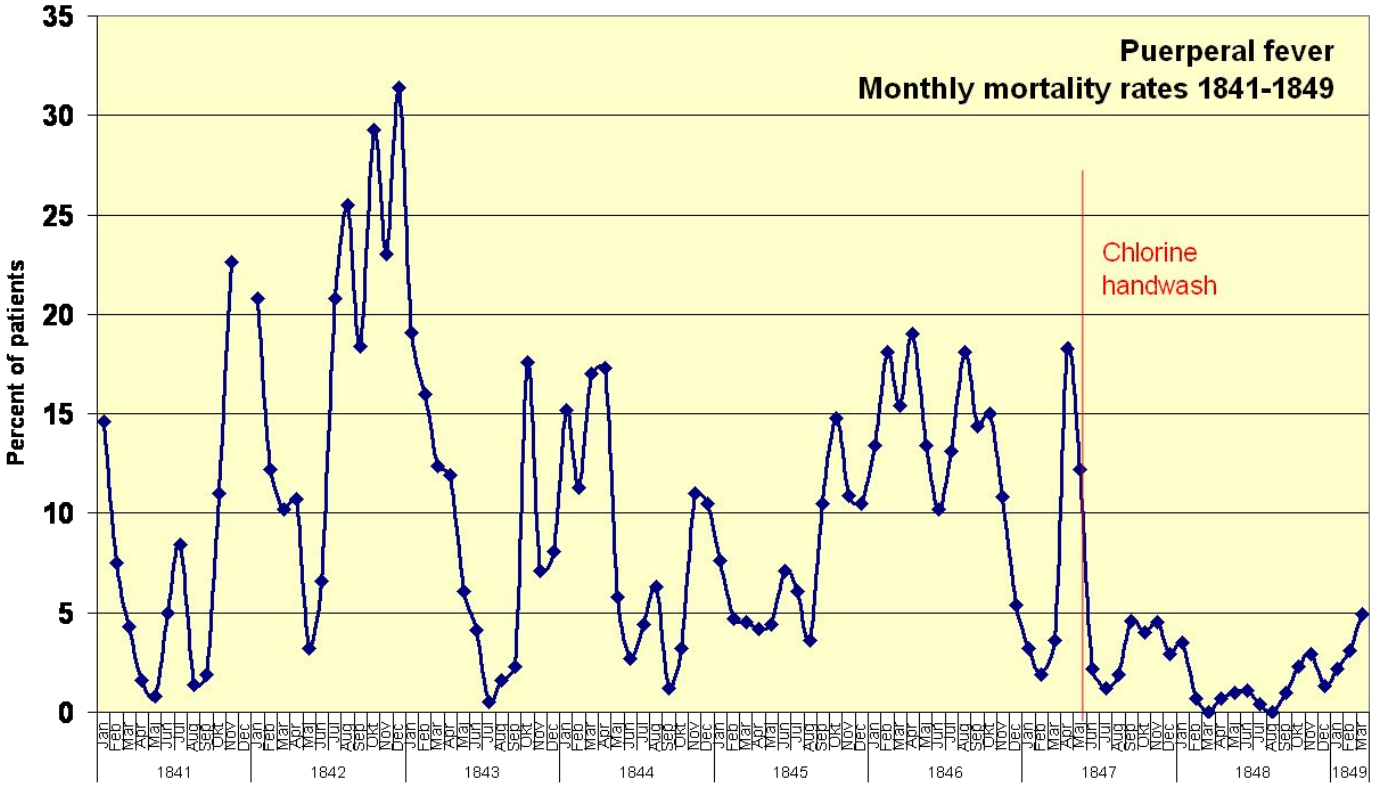
Ignaz Semmelweis: Saviour of Mothers

- Semmelweis' own friend died in 1847 after being accidentally poked with a scalpel while conducting an autopsy → his autopsy had similar pathophysiology of that of women dyeing in "First" clinic
- Doctors in the first clinic would conduct autopsies in the morning and then go and help in deliveries during the afternoon



- Hand-washing did NOT exist!
- Semelweiss suggested washing hands with chlorine after any operation/autospy
- Backlash; driven to inssanity → asylum → beaten → died

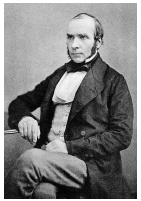
Ignaz Semmelweis: Saviour of Mothers

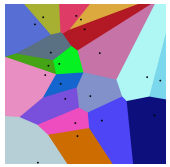


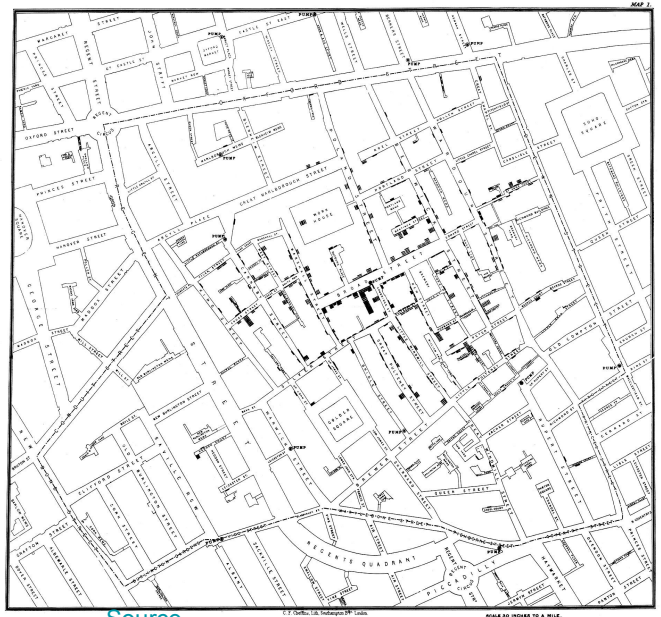
The Cholera outbreak of 1854 on Broad street, London

- 616 people died because of cholera in 1854
- Attended by Florence Nightingale
- John Snow identified the source of outbreak to be a handpump on Broad street
- Chemical and microscopic identification did not prove anything
- But a dot map was convincing → the handle of the pump was disabled → cases declined (but possibly were already declining before his argument)
- The pump was dug next to a cesspit!

John Snow used a dot map to illustrate how cases occurred around this pump → one of the earliest uses of voronoi diagram







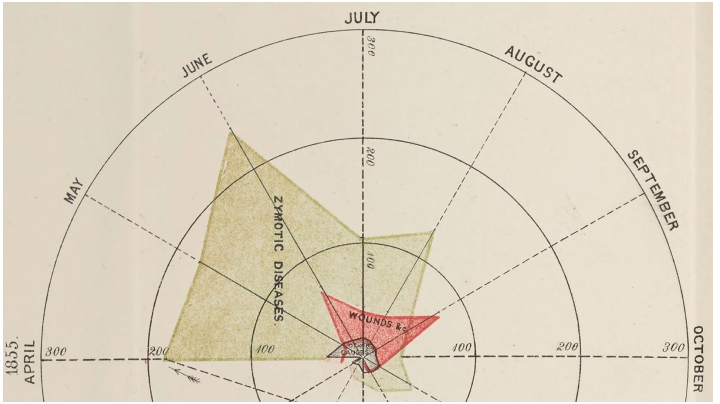
Source

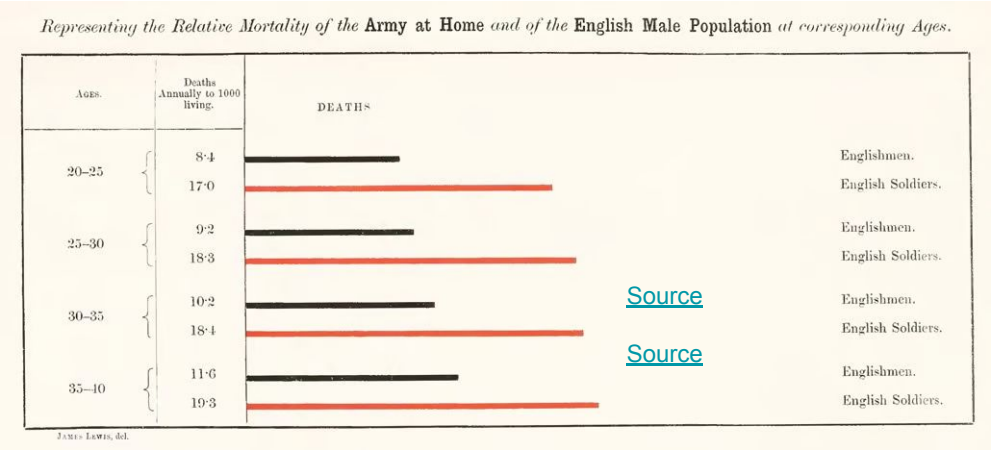
The pioneer of data visualization in health

Florence Nightingale

Lady with the lamp.
Source

Mortality in British East Asia army



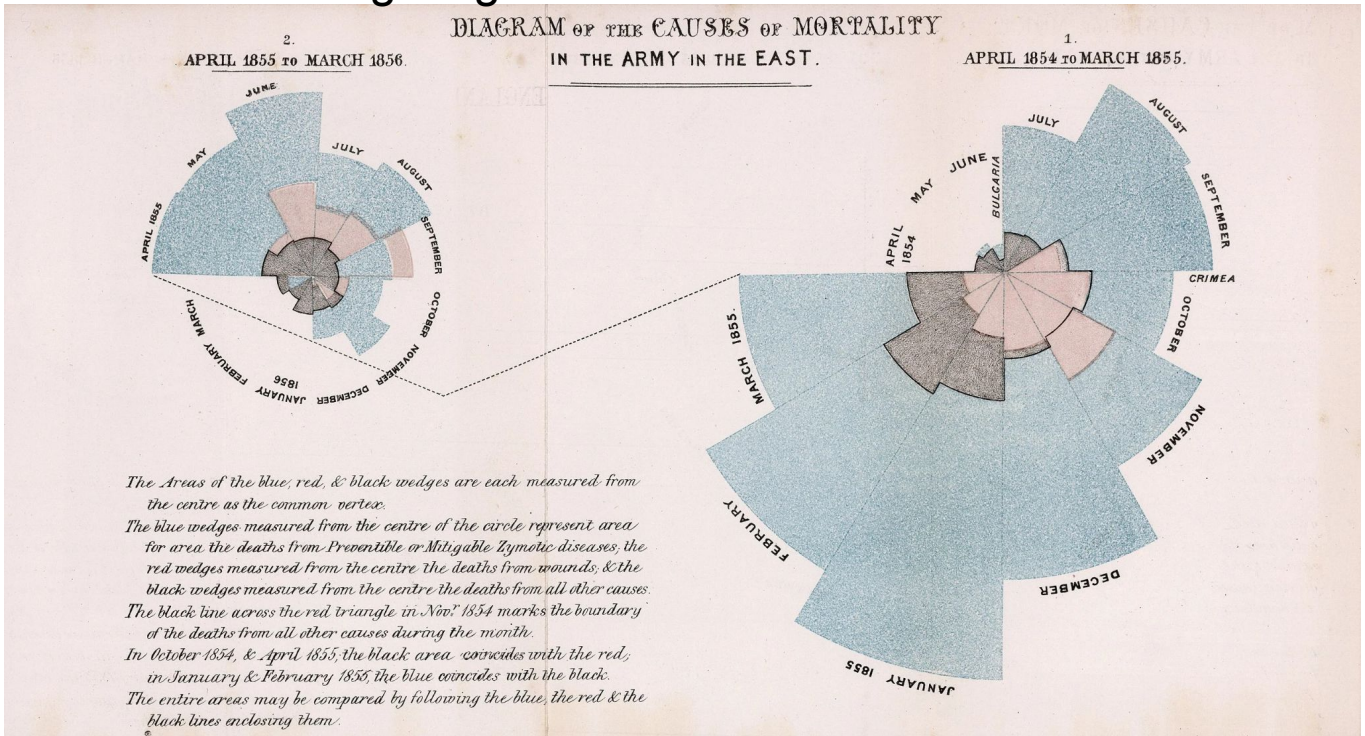


Source

Source

The pioneer of data visualization in health

“Printed tables and all-in double columns, I do not think anyone will read. None but scientific men ever look in the Appendix of a Report. And this is for the vulgar* public.” – Florence Nightingale

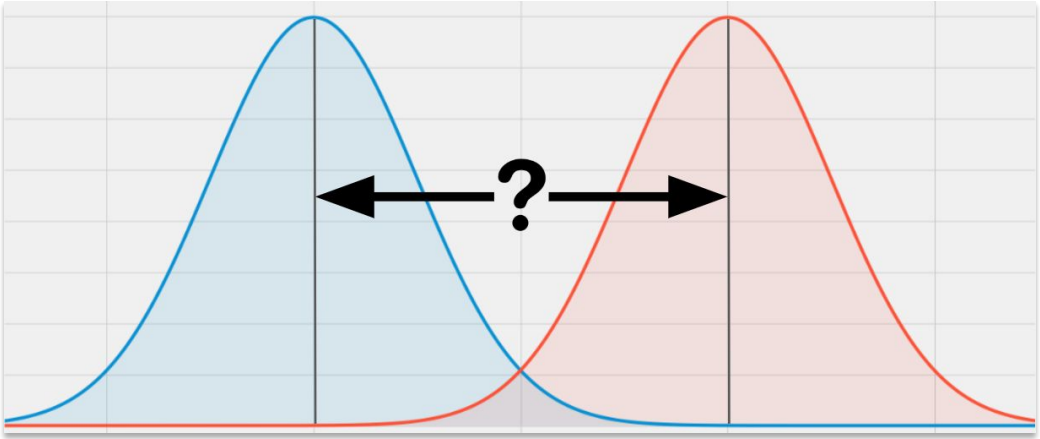


Source
Source

Statistical Models

Why care about modelling (health) data?

Problem: You obtain a set of “readings” from a ‘control’ population and from a suspected ‘case’ population. Are these readings statistically different? “Control” population can also be based on prior “expectation”



Goal: Develop a model that best explains the observations. Quantify the “chance” of observing something ‘at least as different’.

Some digression…

Working with infinite data is elementary, a good statistician works with limited data.

Some digression…

NEET UG 2024 : Centre Asks CBI To Investigate Alleged Irregularities In NEET UG Exam

LIVELAW NEWS NETWORK

23 June 2024 1:58 AM

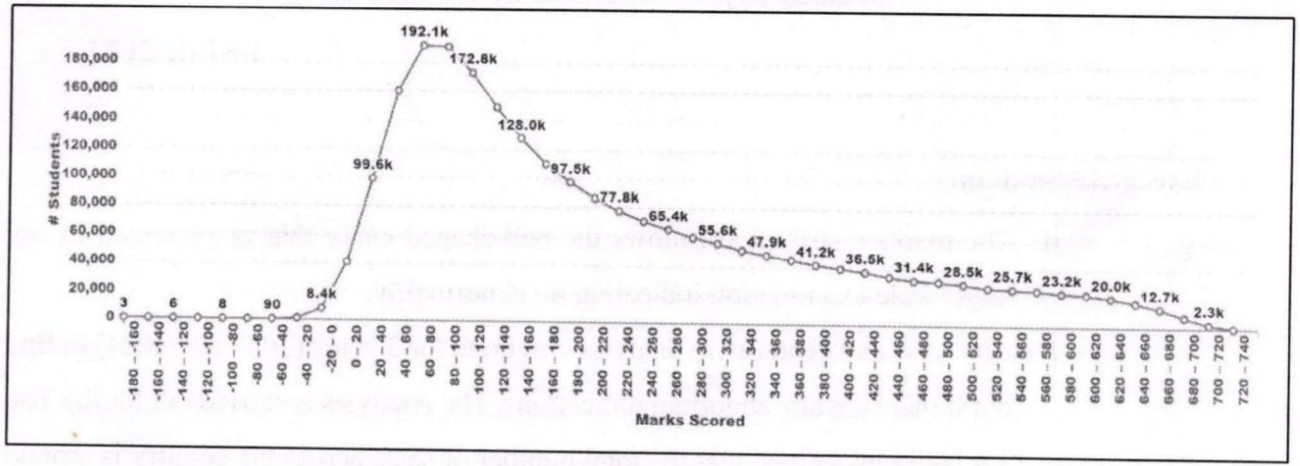
Source

Source

Some digression…

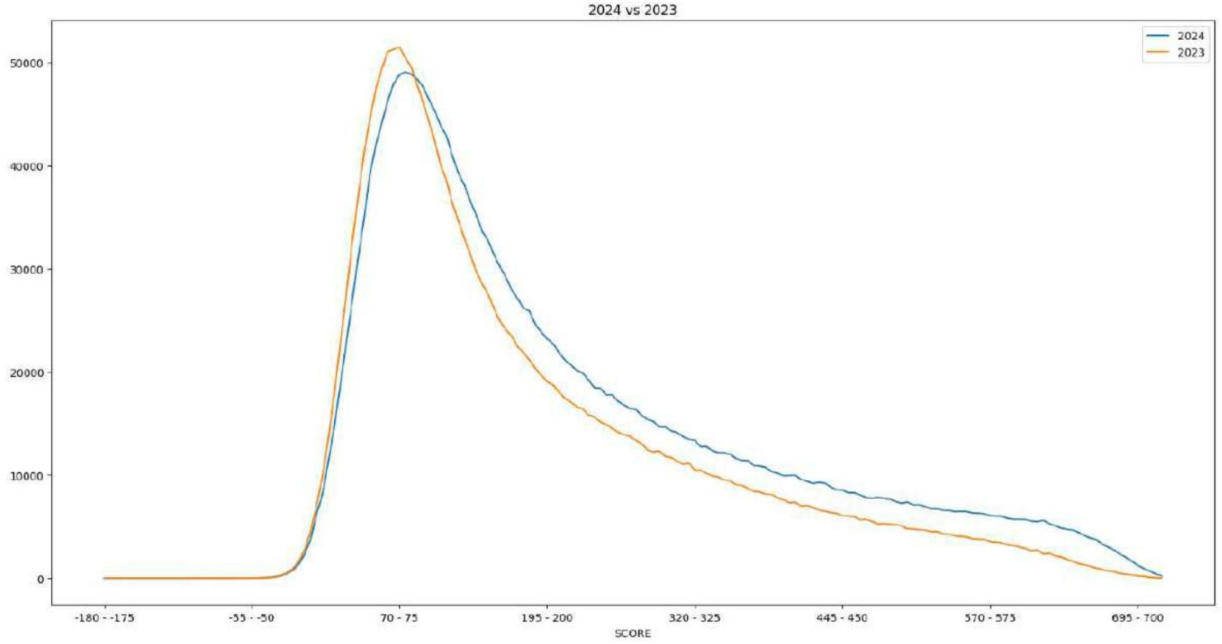
Detailed analysis

1. **The marks distribution follows the bell-shaped curve that is witnessed in any large-scale examination indicating no abnormality.**



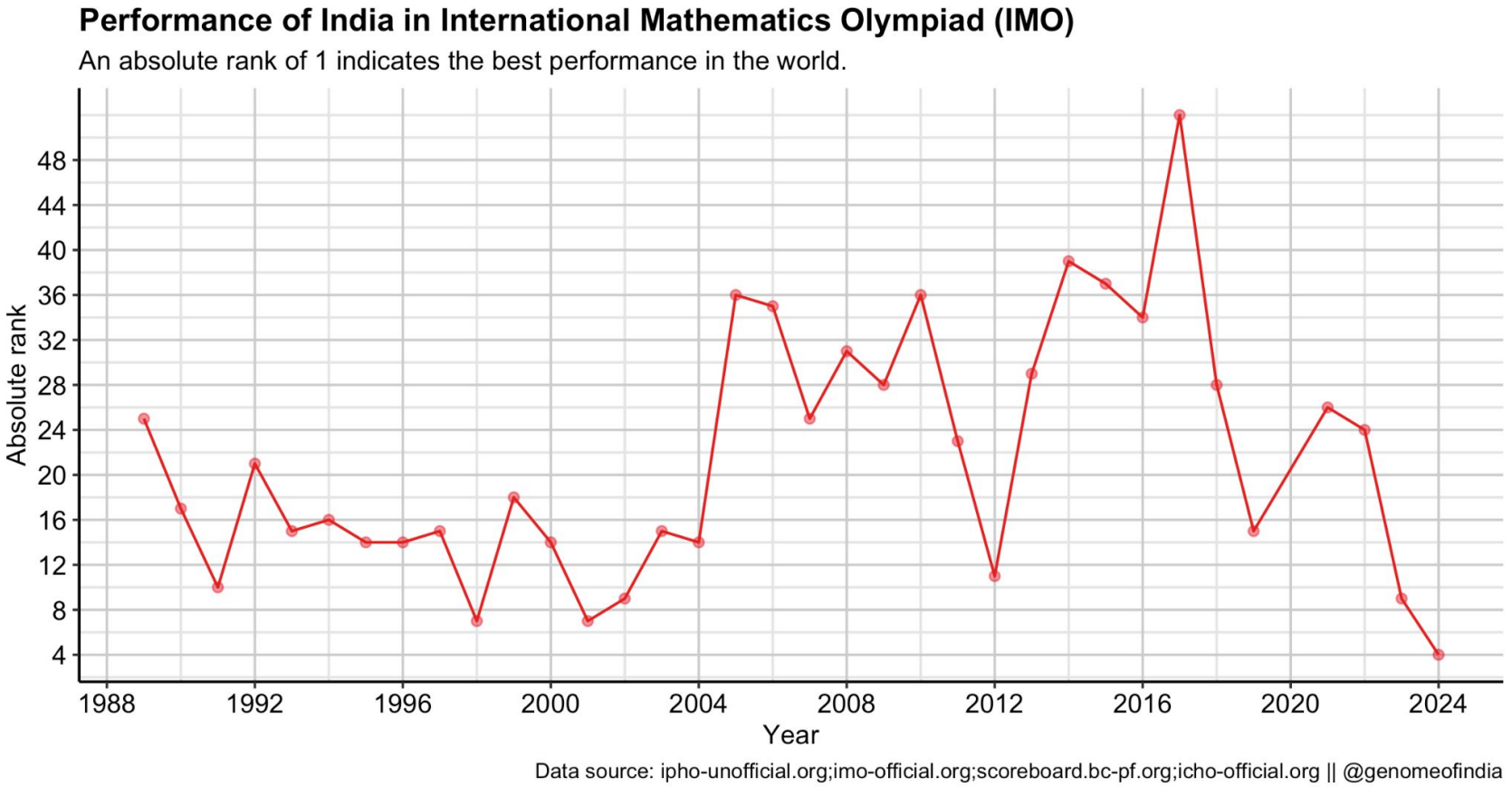
The above graph is the plot of "The marks obtained Vs Number of candidates" for all the candidates who took the examination. Similar analysis was carried out by NTA for every city and the results and the results followed the same distribution. This was verified by IIT Madras team. This further strengthens our view that there are no major abnormalities.

Some digression…



Comparison of the frequency distribution of marks obtained by candidates in 2023 and 2024. “Moreover, NTA has also carried out an analysis of the distribution of marks of candidates in NEET (UG) 2024 at the National, State and City levels and also the Centre level. This analysis indicates that the distribution of marks is quite normal and there seems to be no extraneous factor, which would influence the distribution of marks. The distribution of the pattern of marks at the National level, State Level, City Level & Centre Level has been carried out.”

Some more digression…



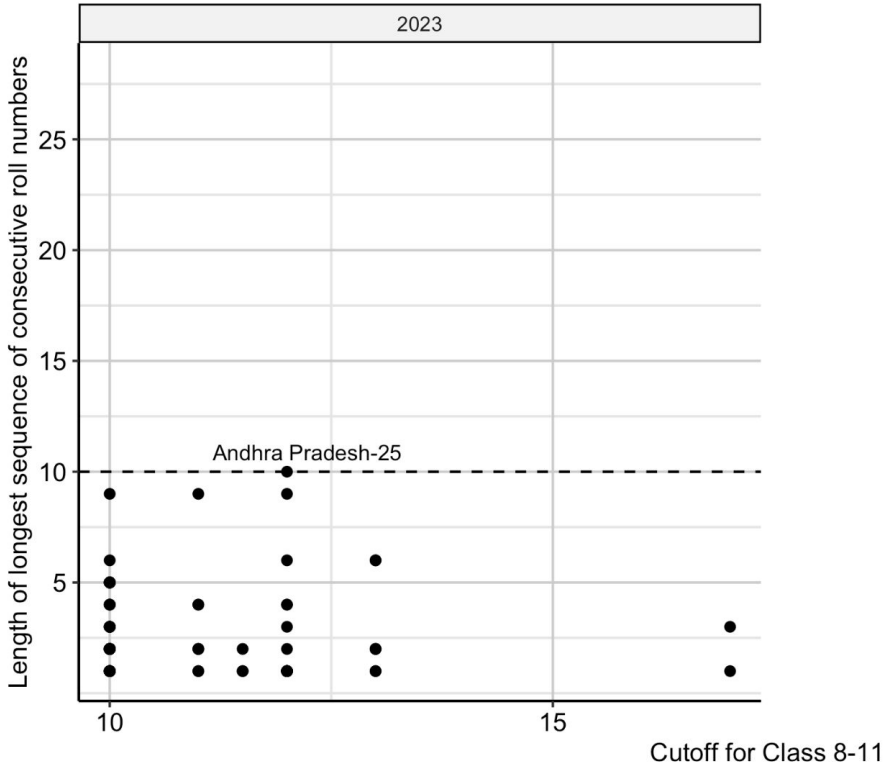
Source

Some more digression…

IOQM = Indian Olympiad Qualifier in Mathematics

Number of students with consecutive roll numbers who qualified IOQM in 2023 and 2024

Number of qualified students who had consecutive roll numbers (difference = 1) are shown on the Y-axis and the cutoff for corresponding state-region (e.g Bihar-14 is South Bihar) is shown on X-axis.
Indicated labels are referred as 'integrity checkpoints'



Source

Random variable

Random variable: A random variable, is a function from a sample space S into the set of real numbers. In its mathematical definition, a random variable neither has randomness nor variability.

Why is it called Random? Randomness comes from the underlying sample space.

A clinical outcome experiment

Example : A clinical trial studying the effectiveness of a new drug has two possible outcomes

- Success = patient shows improvement (S)
- Failure = no improvement (F)

For a given set of *N* patients, we can use random variable X to represent the number of successes. However, note that the observation SFFS is itself not a random variable (even though each such realization itself is random). Why?

Because there is no mapping to a set of real numbers

Discrete vs continuous

Discrete random variable: When the sample space S contains countable number of distinct values, X is discrete random variable. Example. In a coin toss experiment, the realizations are X = {H, T} so X is a discrete random variable.

Probability mass function (PMF)

$$p_X(x) = P(X = x)$$

$$p : \mathbb{R} \to [0, 1]$$
$$-\infty < x < \infty$$

Continuous random variable: When the random variable X can take any value in an interval, X is continuous.

Probability density function (PDF)

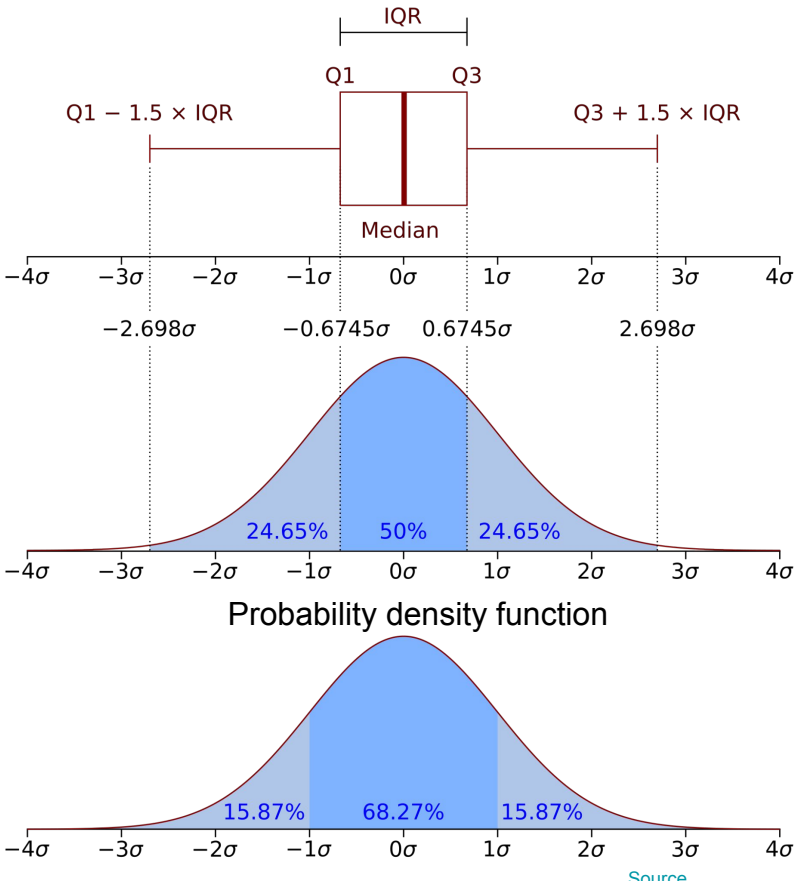
$$\Pr[a \leq X \leq b] = \int_a^b f_X(x)\,dx.$$

Properties of a PMF/PDF

Suppose we have a process that outputs a discrete realization X that takes exactly J values in a set X = {x_1, x_2,..., x_J}. Assume x_j occurs with probability p_j, j = 1, 2, ...J, then p1, p2,..., pJ is called the probability mass (distribution of X is continuous) function as long as it satisfies:

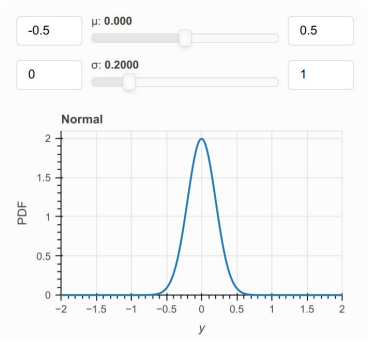
$p_1, p_2, \ldots p_J >= 0$

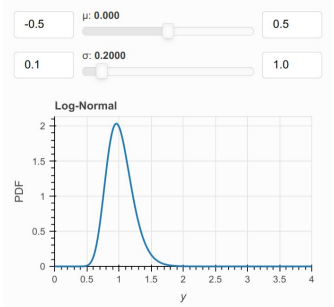
$p_1 + p_2 + \cdots + p_J = 1$ Law of total probability

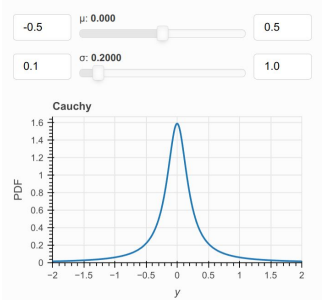


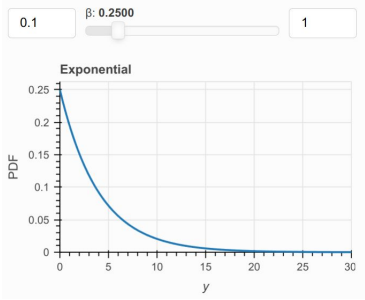
How to think about distributions? The most important ones..

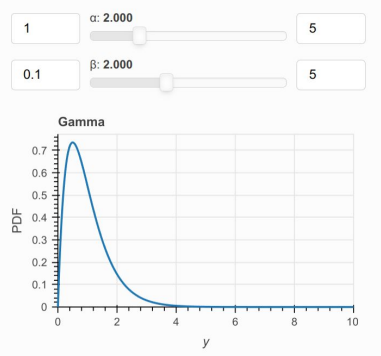
Continuous



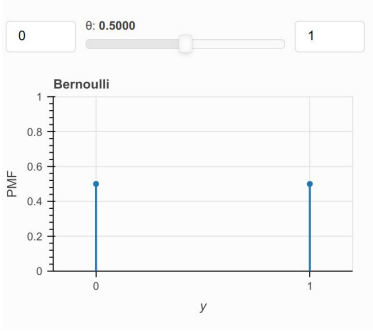


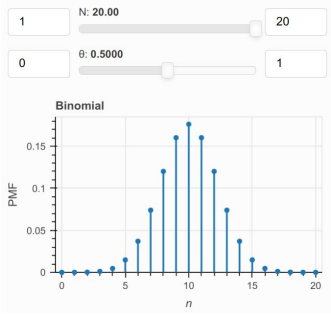


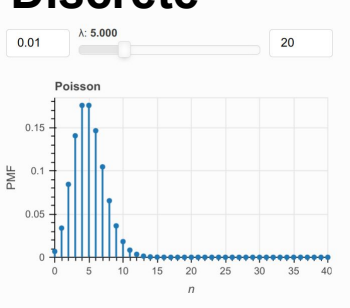


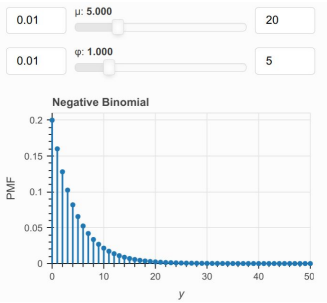


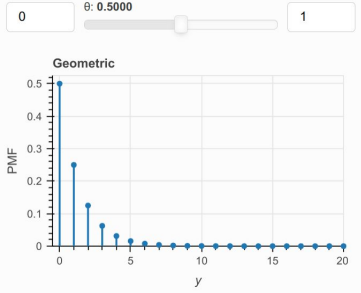
Discrete







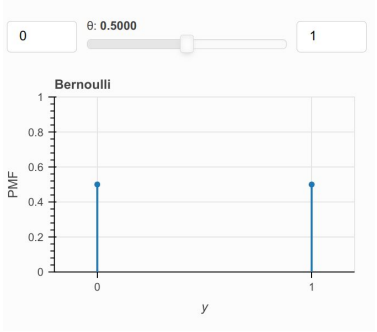




https://distribution-explorer.github.io/

We will focus on discrete case for now…

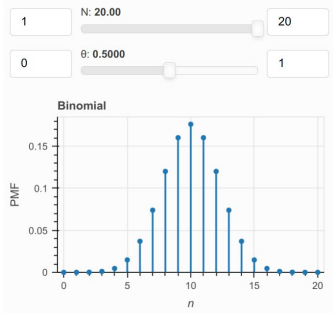
Discrete



$$f(y;\theta) = \begin{cases} 1-\theta & y = 0 \\ \theta & y = 1. \end{cases}$$

Mean: θ

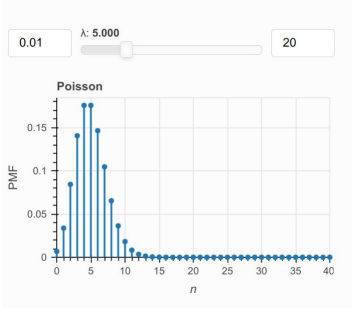
Variance: $\theta(1-\theta)$



$$f(n;N,\theta) = \binom{N}{n}\theta^n(1-\theta)^{N-n}.$$

Mean: $N\theta$

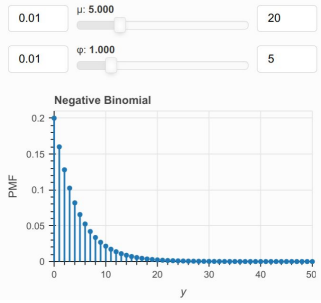
Variance: $N\theta(1-\theta)$



$$f(n;\lambda) = \frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda}.$$

Mean: λ

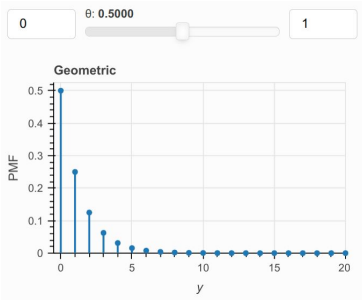
Variance: λ



$$f(y;\mu,\phi) = \frac{\Gamma(y+\phi)}{\Gamma(\phi)\,y!}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}.$$

Mean: μ

Variance: $\mu\left(1+\frac{\mu}{\phi}\right)$.



$$f(y;\theta) = (1-\theta)^y\,\theta.$$

Mean: $\frac{1-\theta}{\theta}$

Variance: $\frac{1-\theta}{\theta^2}$

https://distribution-explorer.github.io/

Questions?