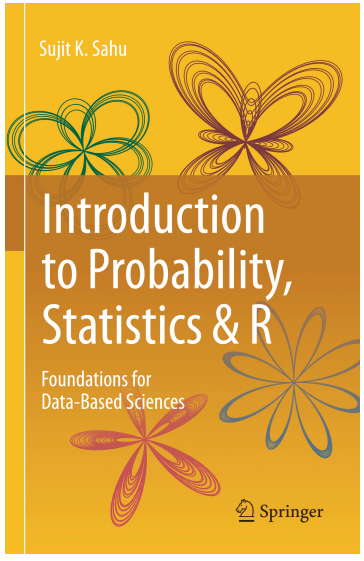
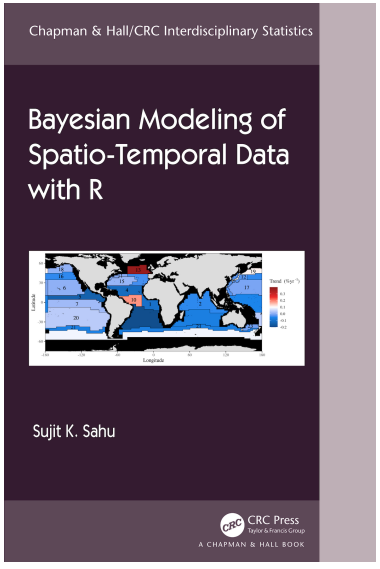
Mathematics Tester Lecture : Probability and Statistics

Prof Sujit Sahu

https://www.sujitsahu.com

UNIVERSITY OF Southampton





Mathematical Sciences, Southampton, 8/9/2024

Plan of the presentation

1. Introduction to the nature of statistics:
   i. Definition of statistics.
   ii. Why should I bother learning statistics and statistical methods?
   iii. Interesting excerpts from the book, *Statistics and Truth* by Prof C R Rao.
2. Examples of statistical ideas in everyday life:
   i. Winning the national lottery, playing the Monty Hall game.
   ii. Probability of disease given symptom using the Bayes theorem.
   iii. Statistics catches serial killer British GP Harold Shipman.
   iv. Disease mapping of cancer rates in NHS health boards.

Definition of statistics

The Oxford English Dictionary defines *statistics* as: "*The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample*."

- Is eating red meat harmful to health - causes cancer?
- Is smoking harmful during pregnancy?
- Is the new therapy/treatment better than the old?
- For a better career prospect should I study Maths or Medicine?

Why isn't it easy to decide one way or the other?

Uncertainty: the main obstacle to decision making

- Uncertainty means: **lack of one-to-one correspondence between cause and effect**.
- For example, having a diet of (well-cooked) red meat is not going to kill me immediately.

> *The only trouble with a sure thing is uncertainty.*
> **Uncertainty is the only certainty there is, ..**

- It is clear that we may never be able to get to the bottom of every case to learn the full truth and so will have to make a decision under uncertainty; thus mistakes cannot be avoided!
- If mistakes cannot be avoided, it is better to know how often we make mistakes (which provides knowledge of the amount of uncertainty) by following a particular rule of decision making: (a statistical method!)
- Such knowledge could be put to use in finding a rule of decision making which does not betray us too often!

Statistics tames uncertainty!

- Everyone (scientists, experts) has their full right to make guesses which can be wild.
- But remember: to guess is cheap, to guess wrongly is expensive!
- Statistical methods allow us to evaluate uncertainty!
- We have the equation:

Uncertain knowledge	+	Knowledge of the extent of uncertainty in it	=	Usable knowledge

Expressed differently:

Noisy data	+	Statistical methods	=	Sound decision

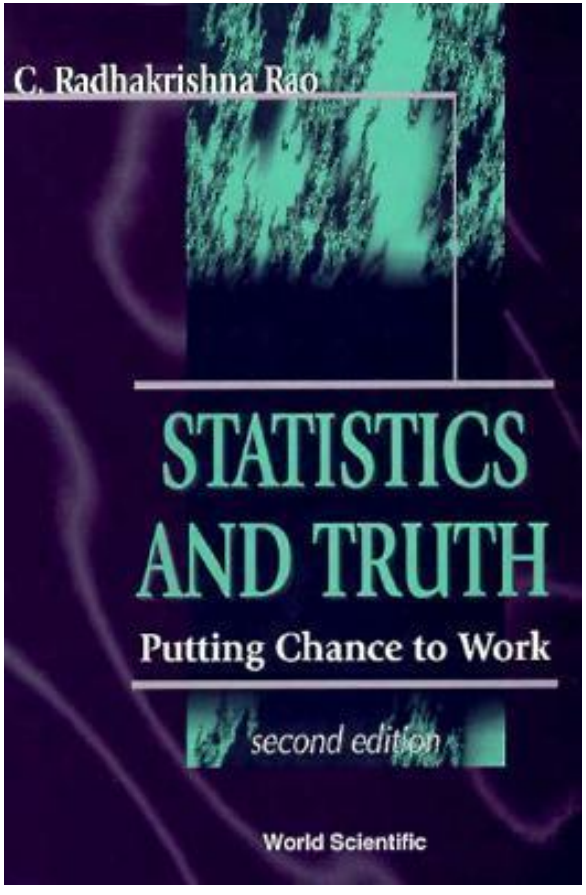
- Whenever we have uncertain & noisy data, we need to call an uncertainty doctor, i.e., a statistician.

Why should a young person study/care about statistics?

- Studying statistics will equip the learner with the basic skills in doing science with data, i.e. data analysis.
- A decent level of statistical knowledge is required no matter what branch of mathematics, engineering, science or medicine a young person will be studying.
- Learning statistical theories gives the maths students the opportunity to practice their deductive mathematical skills on real life problems.
- In this way, Mathematics students will improve at their mathematical methods while studying statistical methods.

What is statistics?

- *Statistics is more a way of thinking or reasoning than a bunch of prescriptions for beating data to elicit answers.*
- Statistics: an inevitable instrument in search of truth.



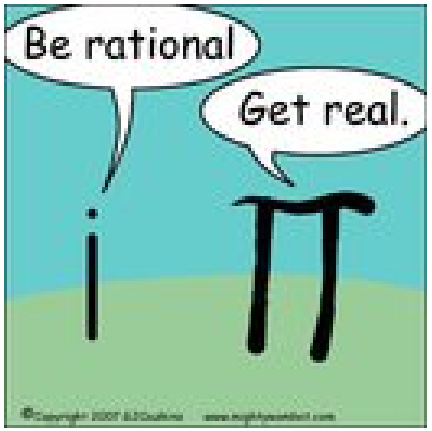
- Statistical methods are results of dialogues between statisticians and practitioners.
- Statistics, here, is not meant to represent just a bunch of numbers or rates!

Nature of statistics

- Statistics is a peculiar subject without any subject matter of its own. It seems to exist and thrive by solving problems in other areas.

- *Statistics is basically parasite: it lives on the work of others. ... Some animals could not digest their food. So it is with many fields of human endeavors, they may not die but they would certainly be a lot weaker without statistics.*
  – Leonard J. Savage.

Lies, Damned Lies and Statistics?



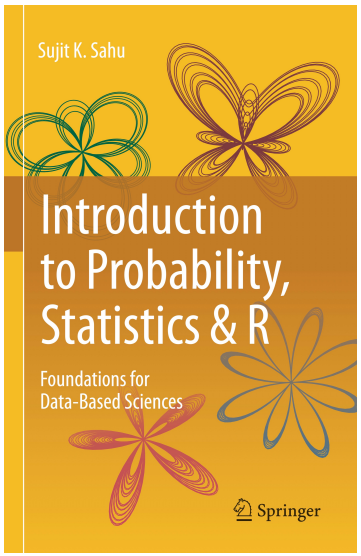
- You can prove anything in statistics!
- *Statistics is like a bikini bathing suit. It reveals the obvious but conceals the vital.*
- Every number is guilty unless proved innocent.
- *I know the answer, give me statistics to substantiate it.*
- Figures won't lie, but liars can figure! – General Charles Grosvenor

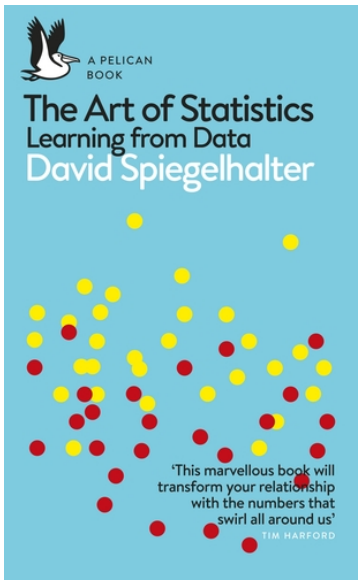
Statistics is an inevitable instrument in search of truth!

- The χ^2-test in statistics is one of the top 20 scientific discoveries!
- Statistical methods prove that cigarette smoking is harmful!
- *According to statistics,* males who remain unmarried die ten years younger.
- *Statistically speaking* tall parents have tall children.
- *A statistical survey* has revealed that a tablet of aspirin every alternate day reduces the risk of a second heart attack.
- *Statistics confirm* that an intake of 500mg of vitamin C everyday prolongs life by six years.

Examples of Statistical ideas

1. A theory book: *Introduction to Probability, Statistics & R* by myself.
2. A wonderful book: *The Art of Statistics: Learning from Data* by Sir David Spiegelhalter.





Probability concepts

- If an experiment has N equally likely possible outcomes then, for any event A,

$$P(A) = \frac{\text{number of outcomes in } A}{\text{total number of possible outcomes of the experiment}}.$$

- ♡ Suppose 4 male and 6 female students are applying for a job, but there are only 3 posts available. How many possible combinations can be formed? How many of those will be female only?

- ♡ The UK National Lottery selects 6 numbers at random from 1 to 49. I bought one ticket - what is the probability that I will win the jackpot?

♡ Winning the National Lottery

- In Lotto, a winning ticket has six numbers from 1 to 49 matching those on the balls drawn on a Wednesday or Saturday evening.
- The 'experiment' consists of drawing the balls from a box containing 49 balls.
- The 'randomness', the equal chance of any set of six numbers being drawn, is ensured by the spinning machine, which rotates the balls during the selection process.
- What is the probability of winning the jackpot?
- Total number of possible selections of six balls/numbers is given by ${}^{49}C_6$.
- There is only 1 selection for winning the jackpot. Hence

$$P(\text{jackpot}) = \frac{1}{{}^{49}C_6} = 7.15 \times 10^{-8}.$$

which is roughly 1 in 13.98 ($\approx$ 14) million.

♡ Winning the National Lottery

- Other prizes are given for fewer matches.

$$P(5 \text{ matches}) = \frac{{}^{6}C_5 \; {}^{43}C_1}{{}^{49}C_6} = 1.84 \times 10^{-5}.$$

$$P(4 \text{ matches}) = \frac{{}^{6}C_4 \; {}^{43}C_2}{{}^{49}C_6} = 0.0009686197$$

$$P(3 \text{ matches}) = \frac{{}^{6}C_3 \; {}^{43}C_3}{{}^{49}C_6} = 0.0176504$$

- Matching 5 of 6 balls & matching the bonus ball.

$$P(5 \text{ matches } + \text{ bonus}) = \frac{6}{{}^{49}C_6} = 4.29 \times 10^{-7}.$$

- Adding all these probabilities of winning some kind of prize

$$P(\text{winning any prize}) \approx 0.0186 \approx 1/53.7.$$

- So a player buying one ticket each week would expect to win a prize, (most likely a £10 prize for matching three numbers) about once a year

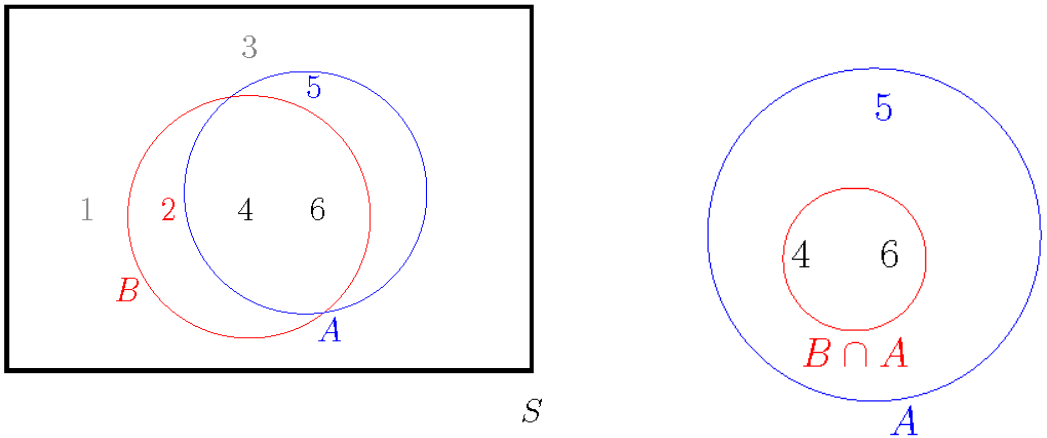
Examples

- Finding probabilities of events conditional on additional information, i.e. something else has happened.
- Find probability of having a disease given that I have the symptom.
- Another example: Die throw.

$$A = (\text{a number greater than } 3) = (4, 5, 6),$$
$$B = (\text{an even number}) = (2, 4, 6).$$

- It is clear that $P(B) = 3/6 = 1/2$. This is the unconditional probability of the event B.
- It is sometimes called the *prior* probability of B.
- We want to find probability of B given that A has already occurred.

Example: Conditional probability



- The left panel shows the two events A and B in the sample space S.
- The right panel shows the event A, which contains three possibilities, two of which are also in the event B.
- To calculate $P(B|A)$, we consider the right panel, where the three events in A $\{4, 5, 6\}$ constitute the whole sample space since we have to assume that A has alreday occurred - the other events 1, 2 and 3 are no longer possible.

Example: Conditional probability...

- Given the partial knowledge that event A has occurred, only the $n_A = 3$ outcomes in $A = (4, 5, 6)$ could have occurred.
- However, only some of the outcomes in B among these n_A outcomes in A will make event B occur; the number of such outcomes is given by the number of outcomes $n_{A \cap B}$ in both A and B, i.e., $A \cap B$, and equal to 2.
- Hence the probability of B, given the partial knowledge that event A has occurred, is equal to

$$\frac{2}{3} = \frac{n_{A \cap B}}{n_A} = \frac{n_{A \cap B}/n}{n_A/n} = \frac{P(A \cap B)}{P(A)}.$$

- Hence we say that $P(B|A) = \frac{2}{3}$, which is often interpreted as the *posterior* probability of B given A.
- The additional knowledge that A has already occurred has helped us to revise the prior probability of $1/2$ to $2/3$.

Practical (phone) problem

- ♡ 3 manufacturing companies: B_1 Pale, B_2 Sung and B_3 Windows.
- Their market shares are respectively 30, 40 and 30 percent.
- Suppose also that respectively 5, 8, and 10 percent of their phones become faulty within one year.
- **Q1** If I buy a phone randomly (ignoring the manufacturer), what is the probability that my phone will develop a fault within one year?
- **Q2** Suppose that the phone I bought developed a fault in the first year - what is the probability that it was made by B_1 Pale?Summary table of information:

Company	Market share	Percent defective
B_1 Pale	30%	5%
B_2 Sung	40%	8%
B_3 Windows	30%	10%

A partition of the sample space, S

- Let $B_1, \ldots, B_k$ denote a set of mutually exclusive (cannot occur together) and exhaustive (fill the whole space) events.
- That is, they are disjoint events and together make up the whole sample space, S.

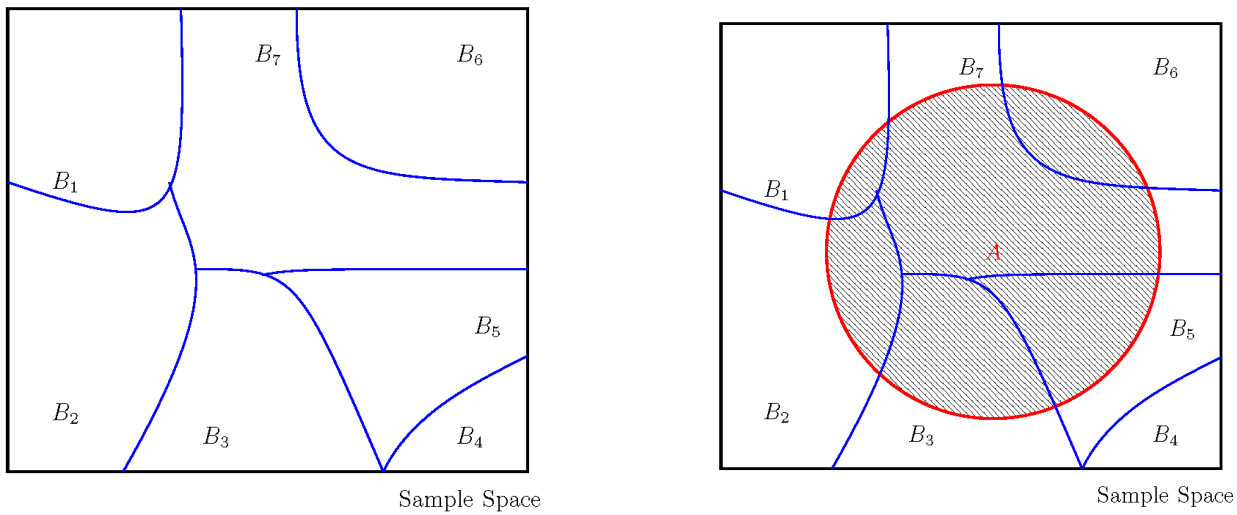


Figure: The left figure shows the mutually exclusive and exhaustive events $B_1, \ldots, B_7$ (they form a partition of the sample space); the right figure shows a possible new event A.

Total probability formula

- Let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive, i.e. $B_i \cap B_j = \emptyset \;\; \text{for all } 1 \leq i \neq j \leq k$
- and exhaustive events, i.e.: $B_1 \cup B_2 \cup \ldots \cup B_k = S$.
- Now any event A can be represented by

$$A = A \cap S = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_k)$$

  where $(A \cap B_1), (A \cap B_2), \ldots, (A \cap B_k)$ are mutually exclusive events.
- Hence one of the three Axioms (Universal truth) of probability gives the **the total probability formula** for $P(A)$.

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_k) \\ &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_k)P(A|B_k). \end{aligned}$$

♡ Return to the phone problem

- We can now find the probability of the event, say A, that a randomly selected phone develops a fault within one year.
- Let B_1, B_2, B_3 be the events that the phone is manufactured respectively by companies B_1 Pale, B_2 Sung and B_3 Windows.
- Then we have:

$$\begin{aligned} P(A) &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3) \\ &= 0.30 \times 0.05 + 0.40 \times 0.08 + 0.30 \times 0.10 \\ &= 0.077. \end{aligned}$$

- Now suppose that my phone has developed a fault within one year.
- What is the probability that it was manufactured by B_1 Pale?
- To answer this we need to introduce the Bayes Theorem.

The Bayes theorem

- Let A be an event, and let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive and exhaustive events.
- Then, for any $i = 1, \ldots, k$,

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_k)P(A|B_k)},$$

- The Bayes theorem is used to update the probability of a possible cause, B_i, given that an effect, A, has already taken place.
- For example, a murder has already taken place and then trying to investigate "who dunnit!"

The Bayes theorem...

- The probability, $P(B_i|A)$ is called the posterior probability of B_i and $P(B_i)$ is called the prior probability.
- The Bayes theorem is the rule that converts the prior probability into the posterior probability by using the additional information that some other event, A above, has already occurred.
- ♡ The probability that my faulty phone was manufactured by B_1 Pale is

$$\begin{aligned} P(B_1|A) &= \frac{P(B_1)P(A|B_1)}{P(A)} \\ &= \frac{0.30 \times 0.05}{0.077} = 0.1948. \end{aligned}$$

- Similarly, the probability that the faulty phone was manufactured by B_2 Sung is 0.4156,
- and the probability that it was manufactured by B_3 Windows is 0.3896.
- Note that $P(B_1|A) + P(B_2|A) + P(B_3|A) = 1$. Why?

♡ Example: Disease given symptom

- In medical statistics, the Bayes theorem determines the probability that someone has the disease given that they have the symptom.
- ♡ Consider a disease that is thought to occur in 1% of the population.
- Using a particular blood test a physician observes that out of the patients with disease 98% possess a particular symptom.
- Also assume that 0.1% of the population without the disease have the same symptom.
- A randomly chosen person from the population is blood tested and is shown to have the symptom.
- What is the conditional probability that the person has the disease?
- Example taken from youtube: search Bayesian trap.

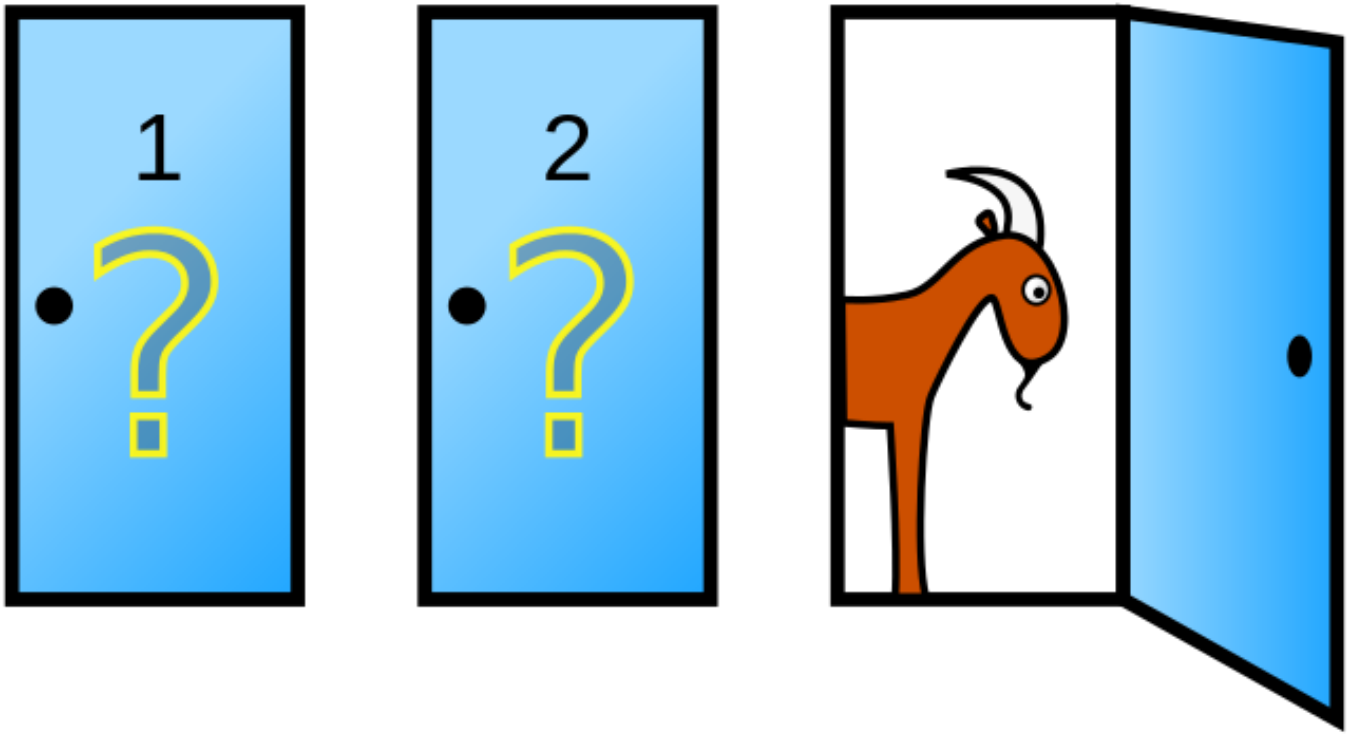
♡ Example: Disease given symptom

- Here $k = 2$ and let B_1 be the event that a randomly chosen person has the disease and B_2 is the complement of B_1.
- Let A be the event that a randomly chosen person has the symptom. The problem is to determine $Pr(B_1|A)$.
- We have $Pr(B_1) = 0.01$ since 1% of the population has the disease, and $Pr(A|B_1) = 0.98$.
- Also $Pr(B_2) = 0.99$ and $Pr(A|B_2) = 0.001$, this is the probability of having the symptom without having the disease. Now $Pr(\text{disease} \mid \text{symptom}) =$

$$\begin{aligned} Pr(\text{D} \mid \text{S}) = Pr(B_1|A) &= \frac{Pr(A|B_1)\,Pr(B_1)}{Pr(A|B_1)\,Pr(B_1)+Pr(A|B_2)\,Pr(B_2)} \\ &= \frac{0.98\times 0.01}{0.98\times 0.01+0.001\times 0.99} \\ &= 0.9082. \end{aligned}$$

- Thus, $Pr(\text{disease}) = Pr(B_1) = 0.01$ gets revised to: $Pr(\text{disease} \mid \text{symptom}) = Pr(B_1|A) = 0.9082$.

Monty Hall: TV Game



- Three doors: only one has a brand new car behind it.
- The other two have goats.
- You are asked to choose one door to win the prize behind.
- The host then reveals one of the other two and reveals a goat behind it.
- Now do you stick with your choice, swap or choose again randomly between the two?

Probability calculation for the Monty Hall problem.

1. Note that the player can win under two mutually exclusive ways if initially they choose:
   (i) the car (C) which has probability $1/3$ and
   (ii) a goat(G) which has probability $2/3$.
2. Hence probability of eventual win (W):

$$\begin{aligned} P(W) &= P(W \cap C) + P(W \cap G) \text{ since } C \cup G = S \\ &= P(W|C)P(C) + P(W|G)P(G) \\ &= P(W|C)\,\tfrac{1}{3} + P(W|G)\,\tfrac{2}{3} \end{aligned}$$

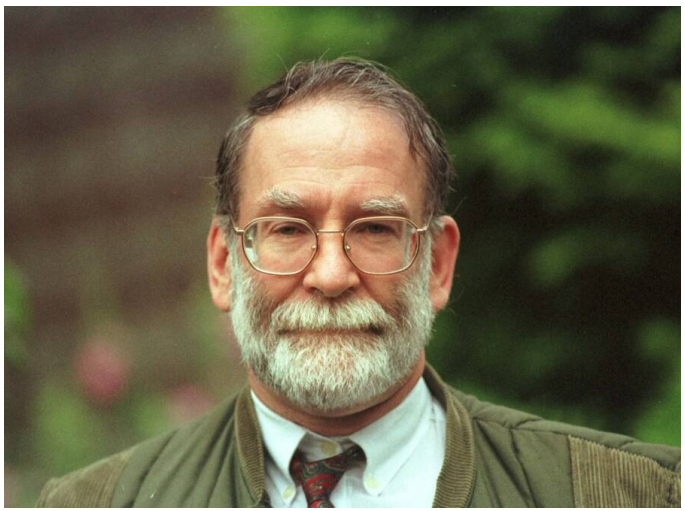
3. The two probabilities $P(W|C)$ and $P(W|G)$ will depend on their strategy after the initial selection. Three possible strategies: Stay, Swap and Randomly choose again.

Strategy	$P(W\|C)$	$P(W\|G)$	P(W)
Stay	1	0	$1 \times \frac{1}{3} + 0 \times \frac{2}{3} = \frac{1}{3}$
Swap	0	1	$0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \mathbf{\frac{2}{3}}$
Random	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{2}{3} = \frac{1}{2}$

4. The 'Swap' strategy maximises the chance of winning.

Harold Shipman : Serial murder example

- Shipman was a British family doctor (GP) and serial killer who killed at least 215 of his most elderly patients by injecting opium between 1975 and 1998.
- His killing spree went undetected until an enquiry was launched during 1998-1999 and convicted in 2000.



Harold Shipman : Serial murder example ...

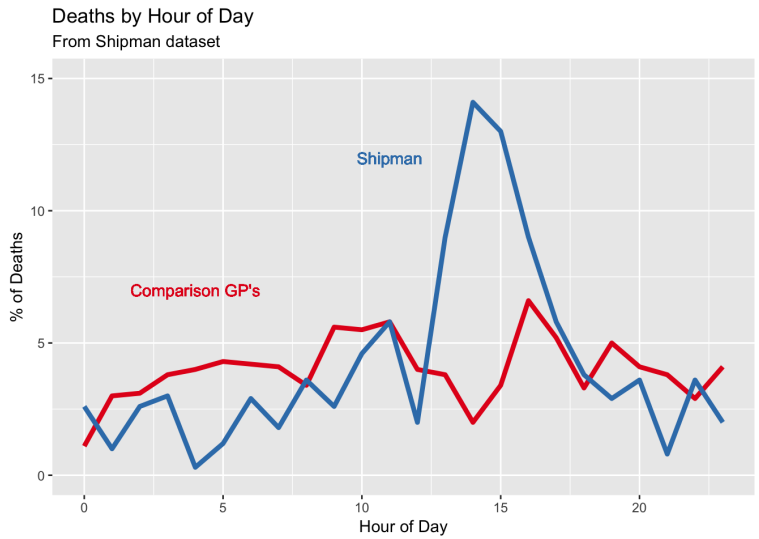


Figure: The time at which Shipman's patients died, compared to the times at which patients of other local family doctors died.

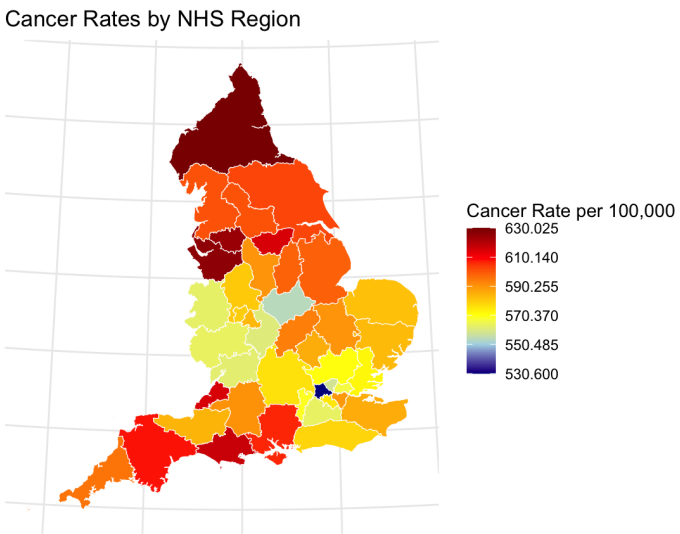
- This is Figure 1.1 in my introduction to probability and statistics book.
- 2PM is the very unusual peak time of death for Shipman's patients.
- No sophisticated statistical analysis is required to detect the obvious pattern in the data.

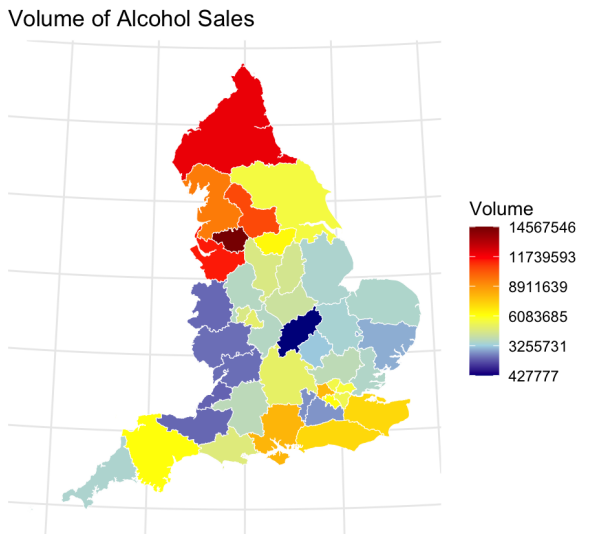
Disease mapping of cancer rates in the NHS regions

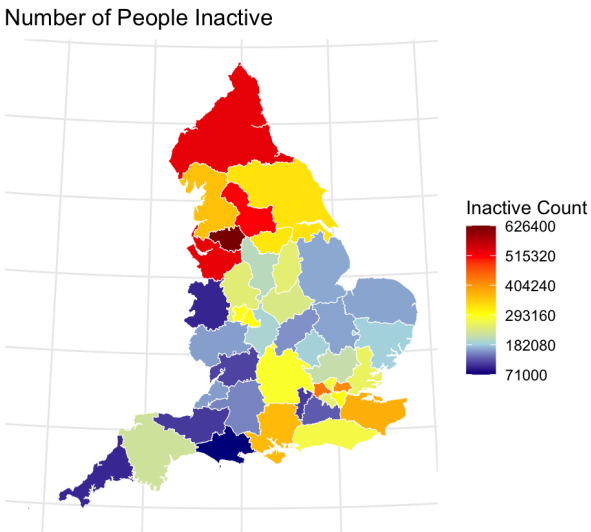
The slides are based on the work of Ms Iona Baxter, a BSc third year project student.

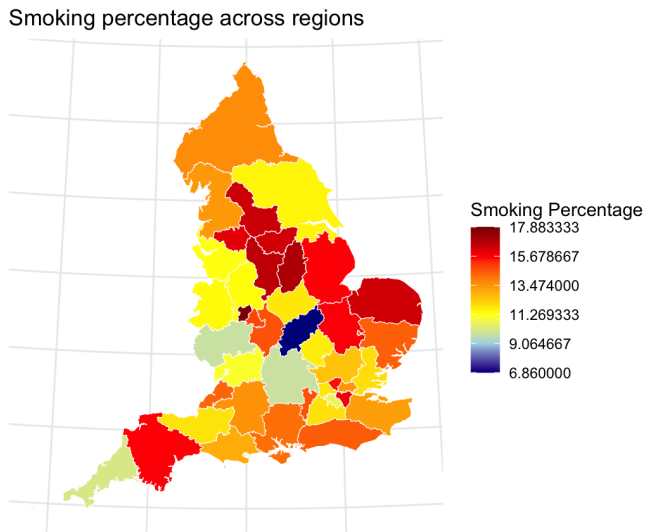
- Iona is interested in assessing spatio-temporal trends in cancer incidence in English health regions during the period 2001-2018.
- One in two people in the UK will develop some form of cancer in their lifetime.
- Cancer rates are associated with characteristics such as hereditary, lifestyle and health related factors.
- Iona's project studies the effect of such factors on geographical and temporal variations in the incidence rates.

Disease mapping of cancer rates in the NHS regions

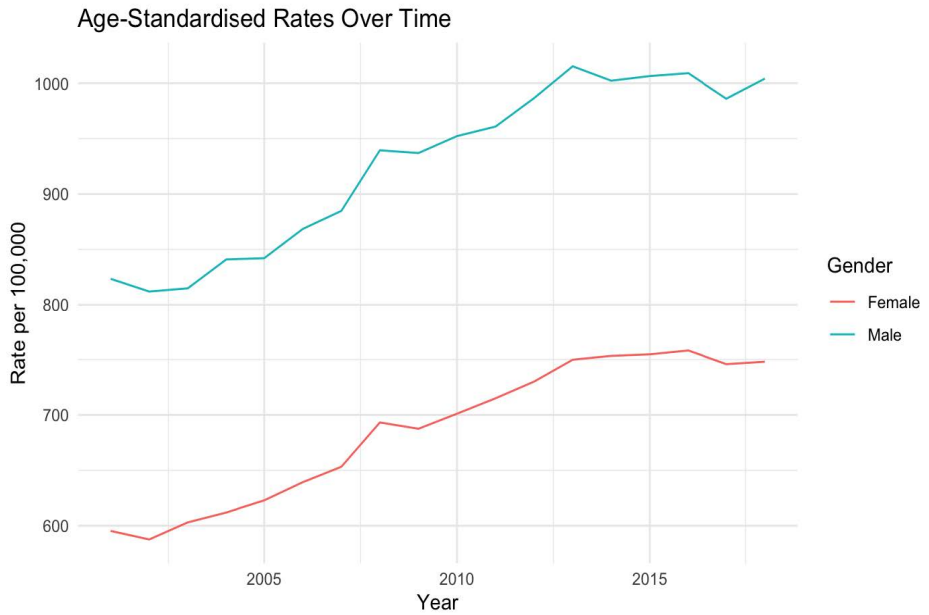


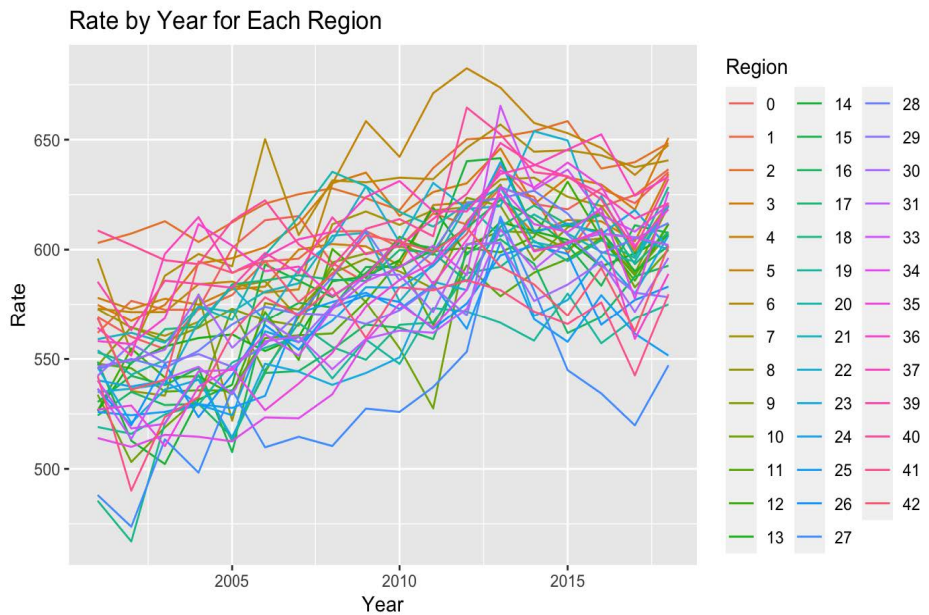






Trend in cancer rates: Graphs from Iona's dissertation





Discussion

- The subject of statistics is vast, powerful and essential in public life.
- A serious study (and devotion) of statistics can both be challenging and rewarding.
- A mathematics department, such as ours, is the best place to study statistics.

"All knowledge is, in final analysis, history.
All sciences are, in the abstract, mathematics.
All judgements are, in their rationale, statistics."

- Statistics and statistical methods bring *Unity in Diversity*.

Take home summary slide

1. Search youtube: Joy of Statistics.
2. Bayes theorem: search youtube : Bayesian Trap
3. Search: For Today's Graduate, Just One Word: Statistics
4. Learn more about the Harold Shipman murder enquiry.
5. Read the book: *The Art of Statistics* by Sir David Spiegelhalter.
6. Find these slides from my website: https://www.sujitsahu.com/bookipsrdbs/resources/

Personalise: Butterfly/Starfish picture

- Start learning R (a free to use computer programme). R allows you to be creative: For example, you can re-draw the following (cover image) of my book (see my website).