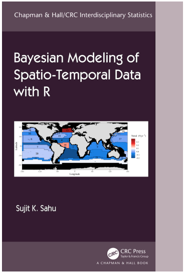
Introduction to Probability & statistics for medical research

Prof Sujit Sahu

https://www.sujitsahu.com

UNIVERSITY OF Southampton



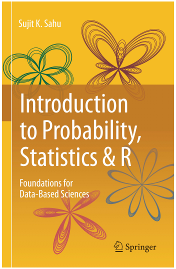
Chapman & Hall/CRC Interdisciplinary Statistics

Bayesian Modeling of Spatio-Temporal Data with R

Sujit K. Sahu

CRC Press Taylor & Francis Group

A CHAPMAN & HALL BOOK



Sujit K. Sahu

Introduction to Probability, Statistics & R

Foundations for Data-Based Sciences

Springer

Royal Hampshire County Hospital, Winchester, 1/5/2024

Biography of Sujit Sahu (https://www.sujitsahu.com)

- Professor of statistics in the University of Southampton (joined 1999).
- Sujit co-authored about 70 research articles published in applied and methodological statistics journals.
- Recently, he also published two textbooks.
- Sujit is also no stranger to medicine and medical science as he gracefully acknowledges the kindness and sincerity of NHS staff (in Winchester, Southampton and Portsmouth hospitals) who saved his life on several occasions by performing major surgeries.
- Sujit is also proud of his daughter who obtained her MBBS degree from the St George's and now a F1 in Charing Cross Hospital, London.

Abstract: This part of the talk will present several examples of statistical ideas in medical research. Examples include (i) success of statistics in exploring the heinous crime of the serial killer British General Practitioner Harold Shipman, (ii) judgement issued in the Sally Clark legal case, (iii) use of the Bayes Theorem in probability in some medical decision making. The talk will also present general ideas of statistics in lighter settings designed to cater for a general audience.
No previous mathematical or statistical background is assumed as the objective here is not to teach difficult (and problem specific) statistical methodologies. Rather, the talk aims to generate interests and awareness so that readers can self-determine the values of statistical contributions they may come across in research papers and conferences.

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 medical research:
   - i. Probability of disease given symptom using the Bayes theorem.
   - ii. Winning the national lottery, playing the Monty Hall game.
   - iii. Sally Clark legal case: Convicted on Statistics?
   - iv. Statistics catches serial killer British GP Harold Shipman.
   - v. Disease mapping of cancer rates in NHS health boards.
3. Interpreting Confidence Intervals and P-Values using practical examples: National lottery and a taste-test.
4. Discussion

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?
- Can the conservatives win the next UK general election?
- Will the hospital waiting list get shorter after the next general election?
- Will the pay and conditions improve for the NHS junior doctors in the near future?

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:

$$\boxed{\text{Uncertain knowledge}} + \boxed{\text{Knowledge of the extent of uncertainty in it}} = \boxed{\text{Usable knowledge}}$$

Expressed differently:

$$\boxed{\text{Noisy data}} + \boxed{\text{Statistical methods}} = \boxed{\text{Sound decision}}$$

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

Importance of calling an uncertainty doctor

- Involving a qualified statistician early, i.e., at the data collection stage, is important.

- On collecting random data: *Man is an orderly animal. He cannot imitate the disorder of nature.*

- *To consult a statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*

- *Use expert opinions in such a way that we stand to gain if they are correct and do not lose if they are wrong.*

- A doctor comforting his patient: *You have a very serious disease. Of ten persons who get this disease only one survives. But do not worry. It is lucky you came to me, for I have recently had nine patients with this disease and they all died of it.*

What does statistics do in medicine?

- Principles of design of experiments are used in screening of drugs and in clinical trials.
- The information supplied by a large number of biochemical and other tests is statistically assessed for diagnosis and prognosis of disease.
- The application of statistical techniques has made medical diagnosis more objective by combining the collective wisdom of the best possible experts with the knowledge on distinctions between diseases indicated by tests.
  - Re: Work of NICE, *National Institute for Health and Care Excellence*

Why should a young person study/care about statistics?

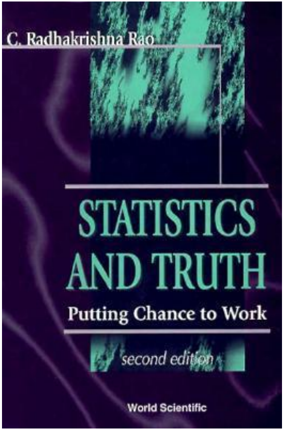
- Studying statistics will equip the learner with the basic skills in data analysis and doing science with data.
- A decent level of statistical knowledge is required no matter what branch of mathematics, engineering, science and *medicine* a young person will be studying.
- Learning statistical theories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

Why should a doctor/physician learn stats?

- For trained doctors (juniors and consultants) it is an essential tool needed for publishing high quality journal articles for the purposes of career progression.
- A strong grasp of statistics (some awareness of the philosophy of statistics and knowledge of the principles of statistical methods) is necessary to hold meaningful and enjoyable conversations with fellow colleagues in scientific conferences.
- In such conversations, one does not necessarily need to know the nitty-gritty details of the statistical techniques another person has applied, but often it helps to have some basic knowledge in statistical methods to (at least) interrogate and raise further questions.
- *“To understand God’s thoughts we must study statistics, for these are the measures of His purpose.”* - Florence Nightingale.

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!

Excerpts from ‘Statistics and Truth’ by CR Rao

- *Scientific laws are not advanced by the principle of authority or justified by faith or medieval philosophy; statistics is the only court of appeal to new knowledge.* – P.C. Mahalanobis.

- *Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.* – H. G. Wells.

- “The more prosperous a country is, the better is its statistics.”

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.

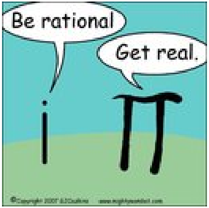
Place of statistics within the universe of knowledge

“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*.
- Diversity in academic disciplines and sciences, medical specialisations etc.
- The motto of the **Indian Statistical Institute** (my alma mater).

STATISTICAL
INDIAN
INSTITUTE
भिन्नेष्वैक्यस्य दर्शनम्
UNITY IN DIVERSITY

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!

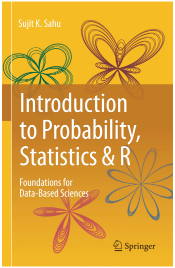
- The χ^2-test in statistics is one of the top 20 scientific discoveries!
- Statistical methods prove that cigarette smoking is harmful!
- *According to statistics,* males who remain unmarried die ten years younger.
- *Statistically speaking* tall parents have tall children.
- *A statistical survey* 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.

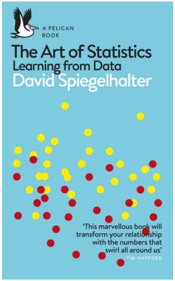
Some comments:

- We apply statistical methods whenever there is uncertainty and complete enumeration is not possible.
- Statistical knowledge is essential for any scientific career in academia, healthcare, industry and government.
- Watch the YouTube video **Joy of Statistics**. https://www.youtube.com/playlist?list=PL4F9E80BCF687CBA6

Examples of Statistical ideas in medical research

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.


Sujit K. Sahu
Introduction to Probability, Statistics & R
Foundations for Data-Based Sciences
Springer


A PELICAN BOOK
The Art of Statistics
Learning from Data
David Spiegelhalter
'This marvellous book will transform your relationship with the numbers that swirl all around us'
TIM HARFORD

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 F2 doctors are applying for training numbers, 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

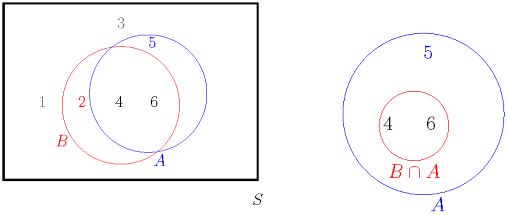
Examples

- 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.

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

- 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 figure 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 $2/3$.

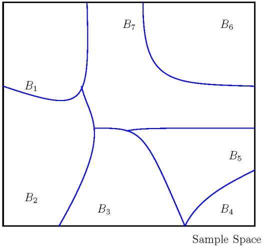
Practical problem

- ♡ 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** After finding the probability, suppose that my phone 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.



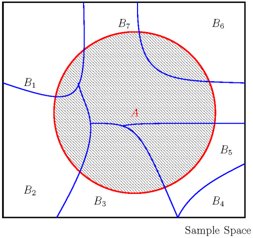


Figure: The left figure shows the mutually exclusive and exhaustive events $B_1, \ldots, B_7$ (they form a partition of the sample space); the right figure shows a possible new event A.

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 the Axiom **A3** 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 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 $Pr(\text{disease} \mid \text{symptom}) =$

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

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

Definition of Independent Events

- Intuitively, events A and B are independent if the occurrence of one event does not affect the probability that the other event occurs.
- Assume $P(A) > 0$ and $P(B) > 0$. The above is equivalent to:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = P(B) \text{ and } P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A).$$

- These give the following formal definition.

$$\boxed{A \text{ and } B \text{ are independent events if } P(A \cap B) = P(A)P(B).}$$

- ♡ Two fair dice when shaken together are assumed to behave independently. Hence the probability of two sixes is $1/6 \times 1/6 = 1/36$.

Dangerous to assume independence wrongly

- When the events are independent then the simpler product formula for joint probability is then used instead of the formula involving more complicated conditional probabilities.

- Independence is often assumed on physical grounds, although sometimes incorrectly.

- There are serious consequences for wrongly assuming independence, e.g. the financial crisis in 2008.

- ♡ Assessing risk in legal cases: In recent years there have been some disastrous miscarriages of justice as a result of incorrect assumption of independence.

- Will discuss the Sally Clark Case.

Sally Clark case

- **December 1996** Sally Clark's son Christopher, aged 11 weeks, is found dead while her husband is out.
- **January 1998** Her second son, Harry, dies, aged eight weeks.
- **February 1998** Mrs Clark is arrested.
- **October 1999** Mrs Clark's trial begins at Chester crown court.
- Professor Roy Meadow appears as a witness, telling the jury there is a "one in 73 million" chance of two children dying from cot deaths in an affluent family.
- **November 1999** Mrs Clark is found guilty and given two life sentences.

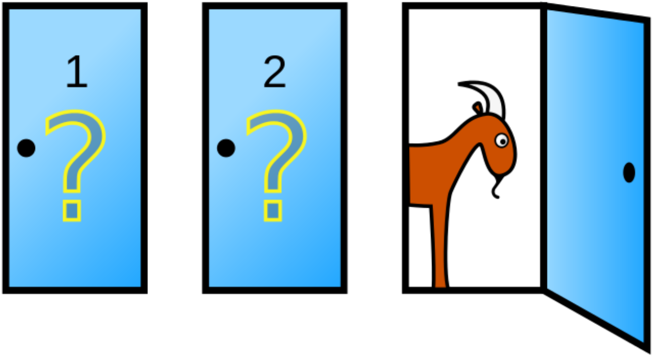
Sally Clark case continues..

- Independence of two child death in a singly family was wrongly assumed by Prof Meadow.
- His calculation: 1 in 8543 $\times$ 1 in 8543 $\approx$ 1 in 73 million.
- Also, “A Prosecutor’s Fallacy”. This consists of showing that the “innocent” explanation for certain facts is highly improbable – and then deducing that the “guilty” explanation is therefore the correct one.
- Moreover, Professor Meadow was not a statistician; yet his ‘expert’ evidence was admitted in the court.

Sally Clark case continues..

- October 2000 First appeal fails.
- January 2003 Mrs Clark's conviction quashed by the court of appeal.
- March 2007 Sally Clark dies.
- Read the article `https://understandinguncertainty.org/node/545`.

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.

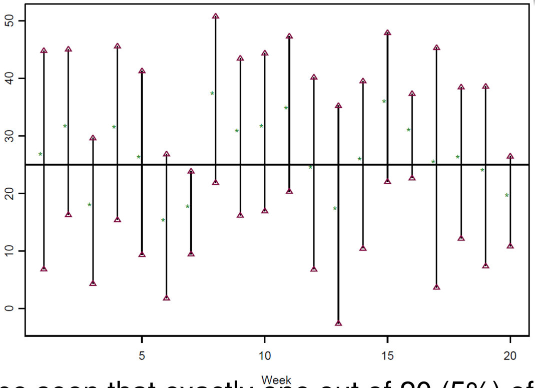
Strategy	$P(W\|C)$	$P(W\|G)$	P(W)
Stay	1	0	$1 \times \frac{1}{3} + 0 \times \frac{2}{3} = \frac{1}{3}$
Swap	0	1	$0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \mathbf{\frac{2}{3}}$
Random	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{2}{3} = \frac{1}{2}$

4. The ‘Swap’ strategy maximises the chance of winning.

Interpretation of confidence intervals.

- Confidence intervals are frequently used, but also frequently misinterpreted.
- A 100($1-\alpha$)% confidence interval for θ is a single observation of a random interval which, under repeated sampling, would include θ 100($1-\alpha$)% of the time.
- The following example from the National Lottery in the UK clarifies the interpretation.
- We collected 6 chosen lottery numbers (sampled at random from 1 to 49) for 20 weeks and then constructed 95% confidence intervals for the population mean $\mu = 25$ and plotted the intervals along with the observed sample means in the following figure.

Some important remarks about confidence intervals.



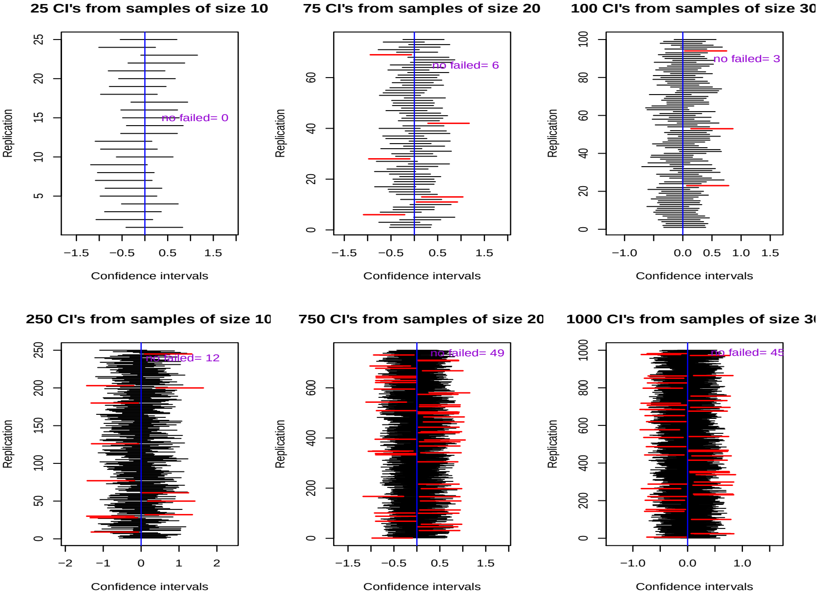
- It can be seen that exactly one out of 20 (5%) of the intervals does not contain the true population mean 25.
- Although this is a coincidence, it explains the main point that if we construct the random intervals with $100(1-\alpha)\%$ confidence levels again and again for hypothetical repetition of the data, on average $100(1-\alpha)\%$ of them will contain the true parameter.

Some important remarks about confidence intervals.

3. A confidence interval is not a probability interval.

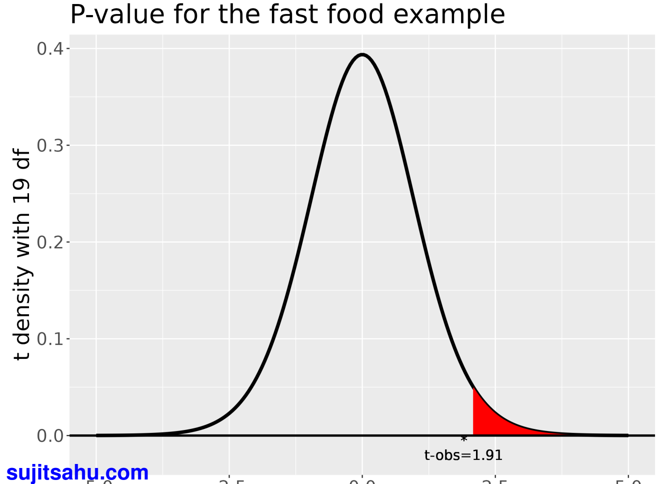
- You should avoid making statements like $P(1.3 < \theta < 2.2) = 0.95$, since
- $P(1.3 < \theta < 2.2) =$ either 0 or 1.
- In the classical approach to statistics you can only make probability statements about random variables, and θ here is assumed to be a constant.
- But such a statement makes perfect sense if you are a Bayesian.

Interpreting Confidence Intervals



Interpreting p-value (tail area)

- The p-value corresponding to an observed test statistic $T(\mathbf{x}) = t_{\text{obs}}$ is defined to be the probability of $T(\mathbf{X})$ lying at and beyond t_{obs} in the “direction of the more extreme values” of the alternative, computed under the null distribution.



Example: t-test

- Consider the t-test for $H_0 : \mu = 60$ against $H_1 : \mu > 60$. Suppose that $n = 20$ and $t_{\text{obs}} = 1.91$.

- Then

$$\text{p-value} = P(T_{n-1} > t_{\text{obs}}) = P\left(T_{19} > 1.91\right) = 0.036.$$

P-values

- Small p-values ($< \alpha$) mean that t_{obs} does lie in C.
- Hence, for small p-values the data provides a strong evidence against the null hypothesis.
- P-values are often incorrectly interpreted as the *probability* that H_0 is true.
  No such interpretation can be given!
  In fact, p-values are problematic since they depend on not only the observations but also their distributions, *i.e.* other values which may have occurred, but did not occur.
- Moreover, these depend on the sampling design not only on the likelihood, as example below shows.

Example ...

- In an experiment to determine whether an individual possesses discriminating powers, she has to identify correctly which of the two brands she is provided with, over a series of trials.
- Let θ denote the probability of her choosing the correct brand in any trial and X_i be the Bernoulli r.v. taking the value 1 for correct guess in the ith trial.
- Suppose that in first 6 trials the results are $1, 1, 1, 1, 1, 0$.
- We wish to test that the tester does not have any discriminatory power against the alternative that she does:

$$H_0 : \theta = \tfrac{1}{2} \text{ versus } H_1 : \theta > \tfrac{1}{2}.$$

- It is a simple versus composite case with $\Theta^{(0)} = \{\frac{1}{2}\}$ and $\Theta_1 = \left(\frac{1}{2}, 1\right]$.

Example (continued)

- Here two cases arise depending on the sampling design.
- Case 1: The number of trials, n, is fixed in advance, *i.e.* binomial distribution.
- If X is the number of correct guesses in the $n = 6$ trials, then

$$\begin{aligned} \text{p-value} &= P\left(X = 5 \text{ or something more extreme} | \theta = \tfrac{1}{2}\right) \\ &= P\left(X = 5 \text{ or } X = 6 | \theta = \tfrac{1}{2}\right) \\ &= (6+1) \times \left(\tfrac{1}{2}\right)^6 = 0.109. \end{aligned}$$

Example (continued)

- Case 2: Continue the trials until the first zero appears, *i.e.* geometric distribution.

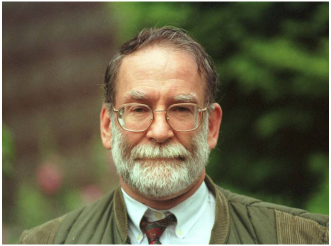
- If X is the number of guesses up to and including first incorrect guess, then

$$\begin{aligned}\text{p-value} &= P\left(X = 6 \text{ or something more extreme} | \theta = \tfrac{1}{2}\right)\\ &= P\left(X = 6, 7, \ldots | \theta = \tfrac{1}{2}\right)\\ &= \left(\tfrac{1}{2}\right)^6 + \left(\tfrac{1}{2}\right)^7 + \cdots = \frac{0.5^6}{1 - 0.5} = 0.031.\end{aligned}$$

- Despite exactly the same sequence of events being observed, different inferences are made.

Harold Shipman : Serial murder example

- Shipman was a British family doctor 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 ...

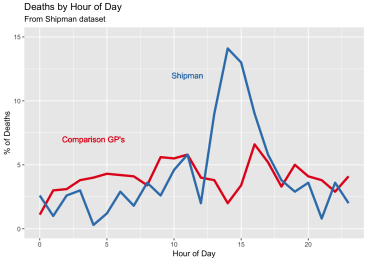


Figure: The time at which Shipman's patients died, compared to the times at which patients of other local family doctors died.

- This is Figure 0.2 in the book "The Art of Statistics" by Sir David Spiegelhalter.
- 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.

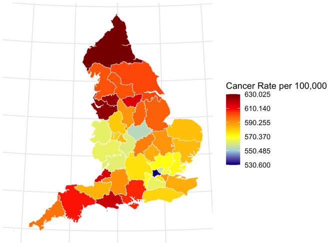
Disease mapping of cancer rates in the NHS regions

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

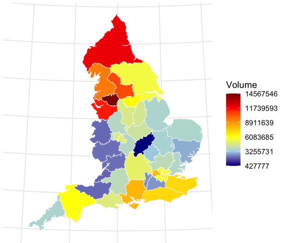
- Iona is interested in assessing spatio-temporal trends in cancer incidence in English health regions during the period 2001-2018.
- One in two people in the UK will develop some form of cancer in their lifetime.
- Cancer rates are associated with characteristics such as hereditary, lifestyle and health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

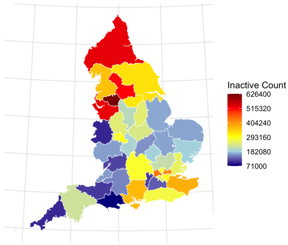
Cancer Rates by NHS Region



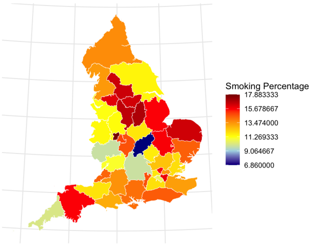
Volume of Alcohol Sales



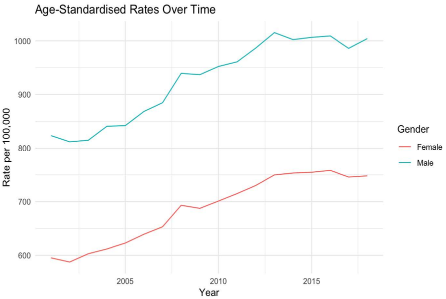
Number of People Inactive

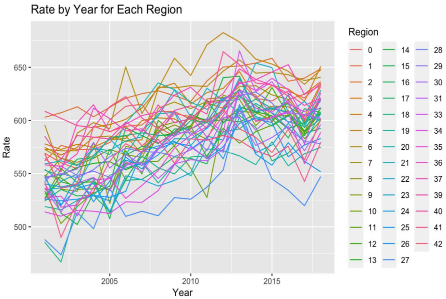


Smoking percentage across regions



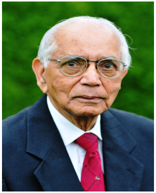
Trend in cancer rates: Graphs from Iona's dissertation


Age-Standardised Rates Over Time


Rate by Year for Each Region

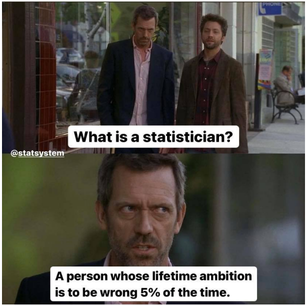
Discussion

- Some familiarity with statistics is a must for career progression.
- Do not hesitate to seek help if needed.
- It is good to build a vocabulary: e.g., by reading
  1. *The Art of Statistics* by Sir David Spiegelhalter.
  2. Sir David's blog: understandinguncertainty.org
  3. *Statistics and Truth* by Prof CR Rao.
  4. *Intro to Probability, Statistics and R* by myself.



Discussion: Final slide

- I have not discussed any serious statistical methods which you can use immediately.
- For example, power and sample size calculation, sensitivity-specificity, survival analysis, Kaplan-Meyer, disease mapping.
- I will be delighted to give a presentation on any such topics to any interested group.
- Please reach out either through Pankaj or my web site.


What is a statistician?
@statsystem
A person whose lifetime ambition is to be wrong 5% of the time.

Some excerpts from Statistics and Truth

- A report submitted to the Tea Board by a consulting statistician contained a table with the caption: Estimated number of people taking tea with standard error.
  - Soon a letter was sent to the statistician asking what standard error is, which peopke take with tea.
- The figure of 2.2 children per adult female is in some respects absurd. It is suggested that the middle classes be paid money to increase the average to a rounded and more convenient number.
- A health minister was intrigued by the statement in the report submitted by a statistician that 3.2 persons out of 1000 suffering from a disease died during last year. He asked his secretary, how 3.2 persons can die. The secretary replied,
- Sir, when a statistician says 3.2 persons died he means that 3 persons actually died and 2 are at the point of death.

Some interesting quotes:

- Weather forecasting: A reliable forecaster is one whose microphone is close enough to the window so that he can decide whether to use official forecast or make up one of his own.

- Public opinion polls: *Once I make up my mind, I am full of indecision.* – Oscar Levant

- Superstition: When asked why he does not believe in astrology, the logician Raymond Smullyan responds that he is a Gemini, and Gemini never believe in astrology.

- A Christian friend of Prof Rao donated his first month’s salary in his first job to the church. When asked whether he believed in God, he replied, “I do not know whether God exists or not, but it would be on the safe side to believe that God exists and act accordingly.”

Some important remarks about confidence intervals.

4. If a confidence interval is interpreted as a probability interval, this may lead to problems.

- For example, suppose that X_1 and X_2 are i.i.d. $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ random variables.
- Then $P[\min(X_1, X_2) < \theta < \max(X_1, X_2)] = \frac{1}{2}$.
- This is because, there are four equally likely outcomes for x_1 and x_2 with regards to the positioning of θ:
- (i) $x_1 < \theta < x_2$, (ii) $x_2 < \theta < x_1$, (iii) $x_1 < x_2 < \theta$ (iv) $\theta < x_1 < x_2$.
- So $[\min(x_1, x_2), \max(x_1, x_2)]$ is a 50% confidence interval for θ, where x_1 and x_2 are the observed values of X_1 and X_2.
- Now suppose that $x_1 = 0.3$ and $x_2 = 0.9$.
- What is $P(0.3 < \theta < 0.9)$? 0 or 1?