COMP229: Introduction to Data Science
Lecture 6: Applications of probability and other paradoxes.

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Reminder: Bayes theorem

$P(A \cap B) = P(A|B)P(B)$,
and $P(A \cap B) = P(A)P(B)$ means independence.

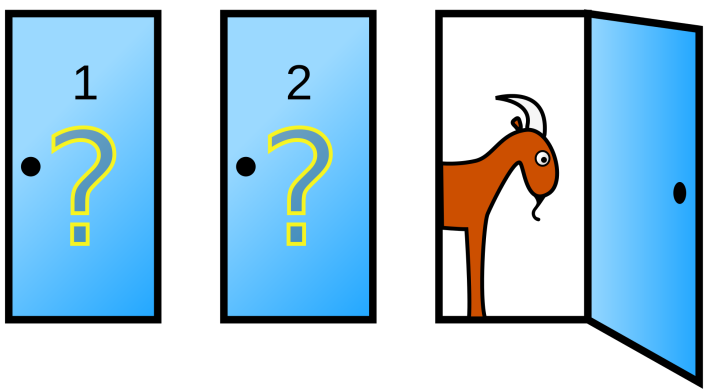
To avoid intersections, we can use

Bayes-Price-Laplace Theorem

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

Monty Hall

"Let's Make a Deal" show, hosted by Monty Hall, allowed to choose one of three doors. Behind one of the doors was a prize, behind the other two was goats. The contestant chooses one of the doors, but before opening it, Monty Hall always opens a door with a goat, and asks:



"Do you want to switch doors or stick to your original choice?"

How will you answer?

Monty Hall from Bayes

C_i = the car is behind door i,

X_i = player's choice of the door i,

G_i = goat is revealed behind door i.

Initially $P(C_i) = \frac{1}{3}$. Suppose the player chooses the door $i = 1, X_1$, and the host opens the door $i = 3, G_3$.

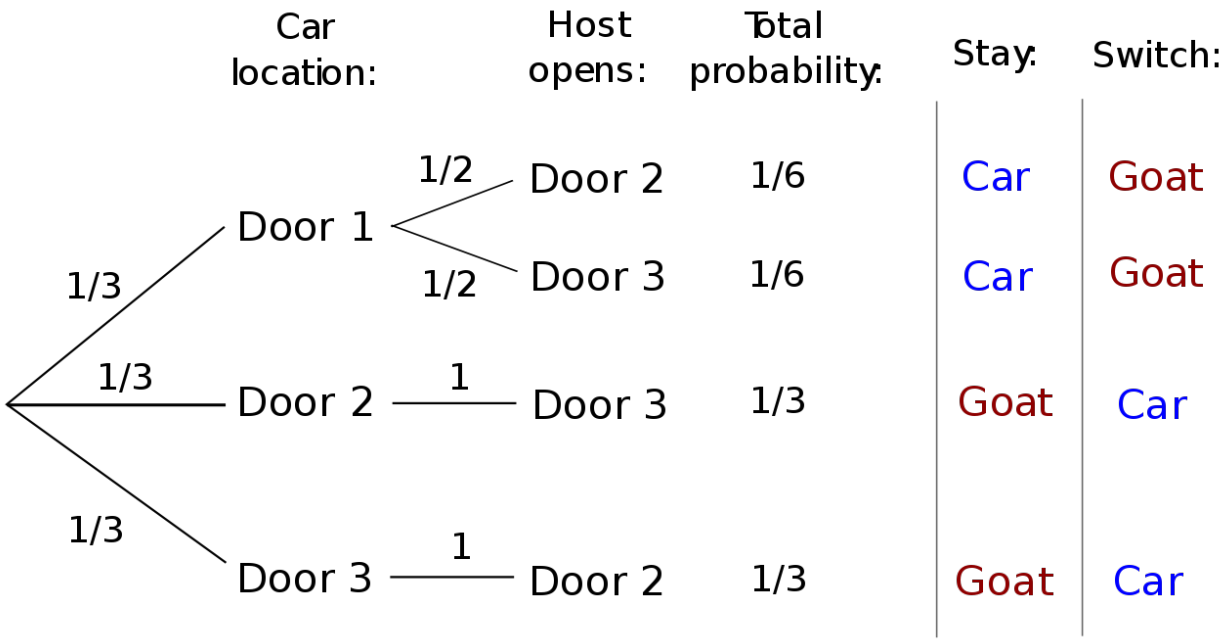
Then clearly $P(C_3|X_1 G_3) = 0$. Let's re-evaluate $P(C_i|X_1 G_3)$:

$$P(C_2|X_1 G_3) = \frac{P(C_2 X_1 G_3)}{P(X_1 G_3)} = = \frac{P(G_3|C_2 X_1)P(C_2 X_1)}{P(G_3 X_1)} = (\text{since}$$

$$P(G_3|C_2 X_1) = 1) = \frac{P(C_2 X_1)}{P(G_3 X_1)} = \frac{P(C_2)P(X_1)}{P(G_3|X_1)P(X_1)} = \frac{\frac{1}{3}\frac{1}{3}}{\frac{1}{2}\frac{1}{3}} = \frac{2}{3}.$$

Solution in a picture

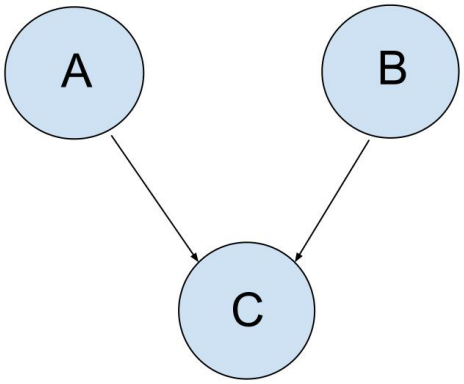
If door 1 is chosen by a player:

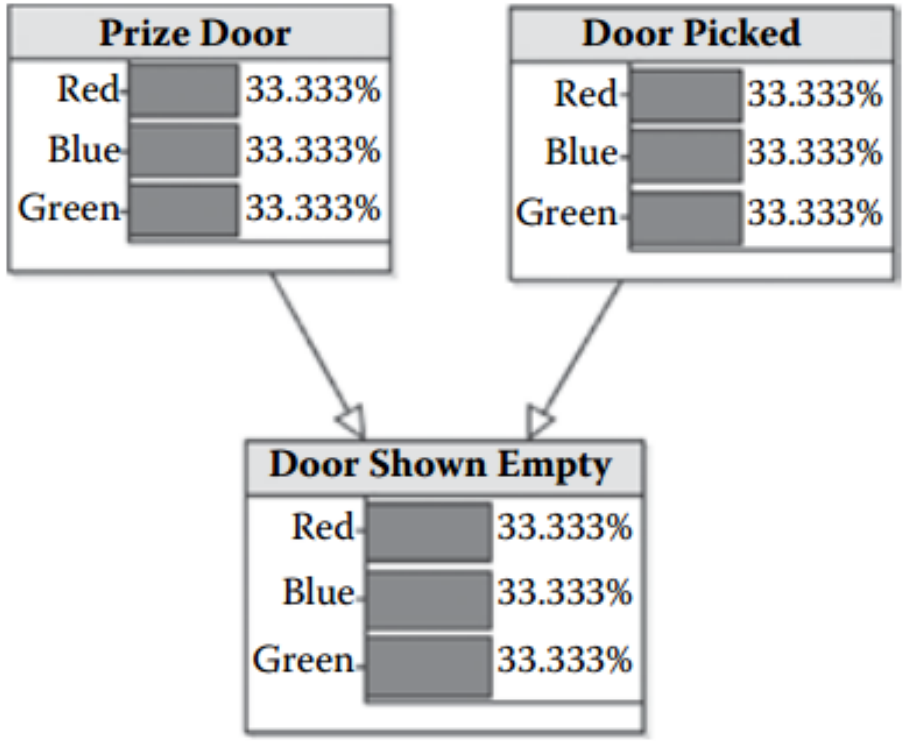


The conditional probability of winning a car by switching is $2/3$.

Bayesian network

A **Bayesian network** (also known as a Bayes network, Bayes net, belief network, or decision network) is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a Directed Acyclic Graph (DAG).





Cancer screening

Problem 6.1. At the age of 40 about 1% of women have a breast cancer. The screening gives a (true) positive result for 80% of women with a breast cancer and (false) positive result for 10% of women without a breast cancer. One woman tested positive. What's the probability of cancer for her?

Solution 6.1. We know: P(cancer)=1%, P(positive $\mid$ cancer)=80%, P(positive $\mid$ no cancer)=10%,
so P(cancer $\mid$ positive) =?? $= P(A|B) = \frac{P(B|A)P(A)}{P(B)} =$

Cancer screening continued

$$\mathrm{P}(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive}|\text{cancer})P(\text{cancer})}{P(\text{positive})}$$

The denominator wasn't given and can be found:

$P(\text{positive}) =$
$\mathrm{P}(\text{positive} \mid \text{cancer})\mathrm{P}(\text{cancer}) + \mathrm{P}(\text{positive} \mid \text{no cancer})\mathrm{P}(\text{no cancer}) =$

$= 0.8 \times 1\% + 0.1 \times 99\% = 10.7\%$. Then

$$\mathrm{P}(\text{cancer} \mid \text{positive}) = \frac{0.8 \times 1\%}{10.7\%} \approx 7.5\%.$$

Cancer screening continued

$$\mathrm{P}(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive}|\text{cancer})P(\text{cancer})}{P(\text{positive})}$$

The denominator wasn't given and can be found:

$P(\text{positive}) =$
$\mathrm{P}(\text{positive} \mid \text{cancer})\mathrm{P}(\text{cancer}) + \mathrm{P}(\text{positive} \mid \text{no cancer})\mathrm{P}(\text{no cancer}) =$

$= 0.8 \times 1\% + 0.1 \times 99\% = 10.7\%$. Then

$$\mathrm{P}(\text{cancer} \mid \text{positive}) = \frac{0.8 \times 1\%}{10.7\%} \approx 7.5\%.$$

It's small, because the cancer-free group is 99 times larger than the cancer group, false positives have an effect.

So-called **Base rate fallacy**: $P(A|B) \neq P(B|A)$.

Changing the parameters

Suppose P(positive | cancer)=70% (instead of 80%), how will P(cancer | positive) change?

$$\text{P(cancer} \mid \text{positive)} = \frac{0.7 \times 1\%}{10.6\%} \approx 6\%.$$

What if P(positive | cancer)=100%?

$$\text{P(cancer} \mid \text{positive)} = \frac{1 \times 1\%}{10.9\%} \approx 9\%.$$

How about the impact of false positives?

P(positive | cancer)=80% as before,

set P(positive | no cancer)=20% (instead of 10%), then

$$\text{P(cancer} \mid \text{positive)} = \frac{0.8 \times 1\%}{20.6\%} \approx 4\%$$

Rare diseases

What if we change the P(cancer) = 2% (from 1%)?

If P(positive | cancer)=80%, P(positive | no cancer)=10%, so

$$\text{P(cancer} \mid \text{positive)} = \frac{P(\text{positive}|\text{cancer})P(\text{cancer})}{P(\text{positive})}$$

$= \frac{0.8\times 2\%}{0.8\times 2\%+0.1\times 98\%} = \frac{1.6}{11.4} \approx 14\%$. So the impact of the disease rate change is the largest!

Rates are often used instead of old-fashioned probabilities.

If using "rate" $\frac{P(\text{positive}|\text{cancer})}{P(\text{positive})} = \frac{0.8}{0.8\times 2+0.1\times 98} = \frac{0.8}{11.4} \approx 7\%$,

even smaller than the initial 7.5%.

Accuracy

Problem 6.2. If about 1% of women (at the age of 40) have a breast cancer, suggest an innovative simple way to create a test that is 99% accurate.

Accuracy

Problem 6.2. If about 1% of women (at the age of 40) have a breast cancer, suggest an innovative simple way to create a test that is 99% accurate.

Solution 6.2. Always return a negative result! Then the result will coincide with the ground truth (correct answer) for all non-cancerous cases, i.e. in 99% of cases.

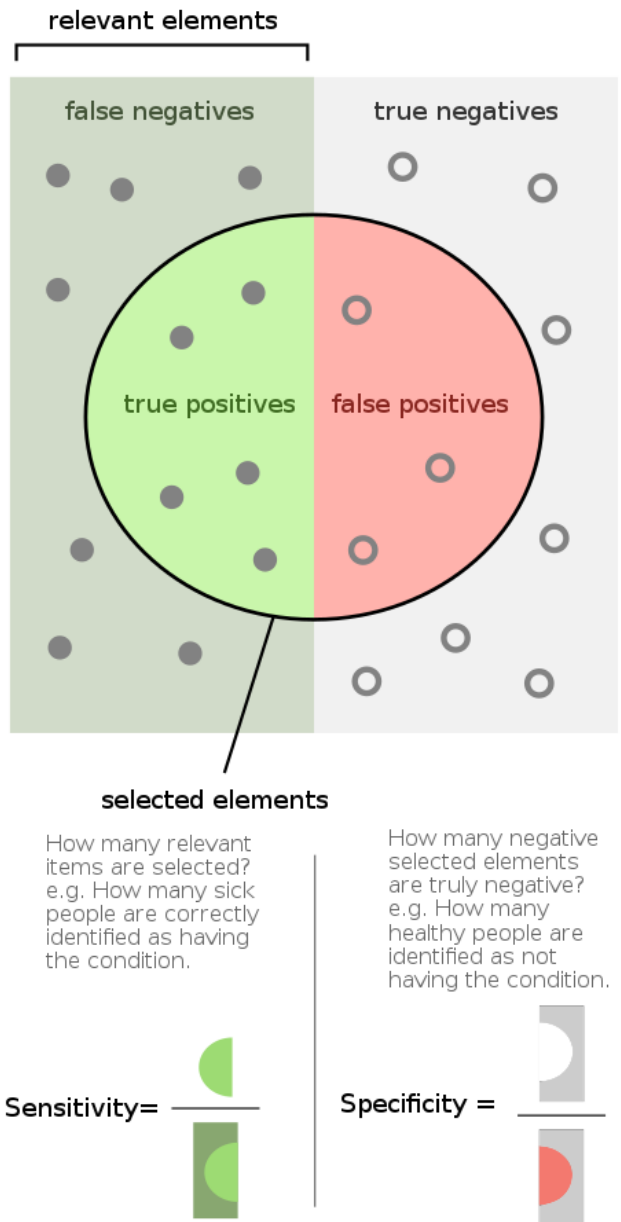
The term "accuracy" is very catchy, but can be very misleading.
Be careful with one-class or unbalanced ground truth in ML!!

Sensitivity and specificity

For any test, at least two rates are measured:

1.**sensitivity**, correct acceptance of a positive result;

2.**specificity**, correct rejection of a negative result.



By

Vaccinations and statistics

Problem 6.3. From the next slide with the data from the Public Health England SARS-CoV-2 variants of concern and variants under investigation in England, Technical briefing 17, compare mortalities (conditional probabilities of death if belonging to the group) between groups of all vaccinated at least ones and all unvaccinated people.

Table 4. Attendance to emergency care and deaths by vaccination status among Delta confirmed cases (sequencing and genotyping) including all confirmed Delta cases in England, 1 February 2021 to 21 June 2021

	Age group (years)	Total	Cases with specimen date in past 28 days	Unlinked	<21 days post dose 1	≥21 days post dose 1	Received 2 doses	Unvaccinated
Delta cases	All cases	92,029	79,336	11,015	6,242	13,715	7,235	53,822
	<50	82,458	71,311	9,892	6,154	9,850	3,689	52,846
	>50	9,571	8,025	1,123	88	3,865	3,546	976
Cases with an emergency care visit§ (excluding cases with the same specimen and attendance dates)‡	All cases	2,406	N/A	33	186	426	190	1,571
	<50	2,013	N/A	25	183	259	68	1,478
	>50	393	N/A	8	3	167	122	93
Cases with an emergency care visit§ (including cases with the same specimen and attendance dates)	All cases	3,460	N/A	51	249	564	348	2,248
	<50	2,728	N/A	40	238	321	94	2,035
	>50	732	N/A	11	11	243	254	213
Cases where presentation to emergency care resulted in overnight inpatient admission§ (excluding cases with the same specimen and attendance dates)‡	All cases	745	N/A	11	55	115	80	484
	<50	564	N/A	8	52	55	17	432
	>50	181	N/A	3	3	60	63	52
Cases where presentation to emergency care resulted in overnight inpatient admission§ (including cases	All cases	1,320	N/A	22	88	189	190	831
	<50	902	N/A	16	79	85	27	695
	>50	418	N/A	6	9	104	163	136

	Age group (years)	Total	Cases with specimen date in past 28 days	Unlinked	<21 days post dose 1	≥21 days post dose 1	Received 2 doses	Unvaccinated
with the same specimen and attendance dates)								
Deaths within 28 days of positive specimen date	Total	117	N/A	3	1	19	50	44
	<50	8	N/A	-	-	2	-	6
	>50	109	N/A	3	1	17	50	38

Data sources: Emergency care attendance and admissions from Emergency Care Dataset (ECDS), deaths from PHE daily death data series (deaths within 28 days)
¥ Cases without specimen dates and unlinked sequences (sequenced samples that could not be matched to individuals) are excluded from this table.
* Cases are assessed for any Emergency Care attendance within 28 days of their positive specimen date. Cases still undergoing within 28-day period may have an emergency care attendance reported at a later date.
§ At least 1 attendance or admission within 28 days of positive specimen date
‡ Cases where specimen date is the same as date of Emergency Care visit are excluded to help remove cases picked up via routine testing in healthcare settings

$$P(\text{death}|\textit{vaccinated}) = \frac{70}{27192} = 0.0026$$

$$P(\text{death}|\textit{unvaccinated}) = \frac{44}{53822} = 0.0008$$

Half truth...

New data for Delta variant from UK.
Deaths of vaxxed = 70/27192 = 0.26% mortality
Deaths of unvaxxed = 44/53822 = 0.08% mortality
Meaning vaxxers have 3.25 higher chance of dying from Indian Coof

Is the statement above truthful? What are the consequences?

Let's recalculate the same probalities, but by 2 different age groups: for under 50 years old and for over 50 years old.

Half truth can be the whole lie

Table 4. Attendance to emergency care and deaths by vaccination status among Delta confirmed cases (sequencing and genotyping) including all confirmed Delta cases in England, 1 February 2021 to 21 June 2021

	Age group (years)	Total	Cases with specimen date in past 28 days	Unlinked	<21 days post dose 1	≥21 days post dose 1	Received 2 doses	Unvaccinated
Delta cases	All cases	92,029	79,336	11,015	6,242	13,715	7,235	53,822
	<50	82,458	71,311	9,892	6,154	9,850	3,689	52,846
	>50	9,571	8,025	1,123	88	3,865	3,546	976
Cases with an emergency care visit§ (excluding cases with the same specimen and attendance dates)‡	All cases	2,406	N/A	33	186	426	190	1,571
	<50	2,013	N/A	25	183	259	68	1,478
	>50	393	N/A	8	3	167	122	93
Cases with an emergency care visit§ (including cases with the same specimen and attendance dates)	All cases	3,460	N/A	51	249	564	348	2,248
	<50	2,728	N/A	40	238	321	94	2,035
	>50	732	N/A	11	11	243	254	213
Cases where presentation to emergency care resulted in overnight inpatient admission§ (excluding cases with the same specimen and attendance dates)‡	All cases	745	N/A	11	55	115	80	484
	<50	564	N/A	8	52	55	17	432
	>50	181	N/A	3	3	60	63	52
Cases where presentation to emergency care resulted in overnight inpatient admission§ (including cases	All cases	1,320	N/A	22	88	189	190	831
	<50	902	N/A	16	79	85	27	695
	>50	418	N/A	6	9	104	163	136

	Age group (years)	Total	Cases with specimen date in past 28 days	Unlinked	<21 days post dose 1	≥21 days post dose 1	Received 2 doses	Unvaccinated
with the same specimen and attendance dates)								
Deaths within 28 days of positive specimen date	Total	117	N/A	3	1	19	50	44
	<50	8	N/A	-	-	2	-	6
	>50	109	N/A	3	1	17	50	38

Data sources: Emergency care attendance and admissions from Emergency Care Dataset (ECDS), deaths from PHE daily death data series (deaths within 28 days)

For < 50:

$$P(\text{death}|\textit{vaccinated}) = \frac{2}{19693} = 0.00010,$$

$$P(\text{death}|\textit{unvaccinated}) = \frac{6}{52846} = 0.00011;$$

For > 50:

$$P(\text{death}|\textit{vaccinated}) = \frac{68}{7499} = 0.009,$$

$$P(\text{death}|\textit{unvaccinated}) = \frac{38}{976} = 0.039$$

Making the right choice

Problem 6.4. A Black box contains 5 red and 6 blue balls, and a White box contains 3 red and 4 blue balls. You are allowed to choose a box and then choose one ball at random from the box. If you choose a red ball, you get a prize. For more chances to win, which box should you choose to draw from?

Taking balls from an box is similar to flipping a coin whose probabilities of heads/tails are not 50%.

Good outcomes over all outcomes

Solution 6.4. What is the probability to take a red ball from the Black box with 5 red and 6 blue balls?

The black box has 5+6=11 balls, all equally probable, so each ball has probability $\frac{1}{11}$.

We get a prize for each of 5 red balls. These 5 lucky events are mutually exclusive, because we take 1 of 5 balls. By the sum rule, the Black box gives the probability $P_{black} = \frac{5}{11}$ to win.

Comparing two probabilities

If all outcomes are equally likely, the sum rule is the **quotient rule**: $P = \dfrac{\text{number of good outcomes}}{\text{number of all outcomes}}$.

Similarly to the above, the White box with 3 red and 4 blue balls gives the probability $P_{white} = \dfrac{3}{7}$ to win.

Which is larger, $\frac{5}{11}$ or $\frac{3}{7}$?

$\dfrac{5}{11} - \dfrac{3}{7} = \dfrac{5 \times 7 - 3 \times 11}{11 \times 7} = \dfrac{35 - 33}{11 \times 7} > 0$, so we choose the Back box with 5 red and 6 blue balls.

Another box, another choice

Problem 6.5. Consider another game in which a 2nd Black box has 6 red and 3 blue balls, and a 2nd White box has 9 red and 5 blue balls. Which box should you choose if red wins again?

Solution 6.5. The Black box has the probability to win $P_{black} = \frac{6}{6+3} = \frac{2}{3}$. The White box with 9 red and 5 blue balls has $P_{white} = \frac{9}{9+5} = \frac{9}{14}$. Since $\frac{2}{3} - \frac{9}{14} = \frac{2 \times 14 - 9 \times 3}{3 \times 14} = \frac{28 - 27}{3 \times 14} > 0$, we again choose the black box for a larger chance.

Join the lucky boxes

Problem 6.6. The contents of the 2nd Black box are added to the 1st Black box, and the contents of the 2nd White box are added to the 1st White box.

Is the Black box the lucky one again?

The final Black box has $5 + 6 = 11$ red and $6 + 3 = 9$ blue balls. The final White box has $3 + 9 = 12$ red and $4 + 5 = 9$ blue balls.

Luck overturned

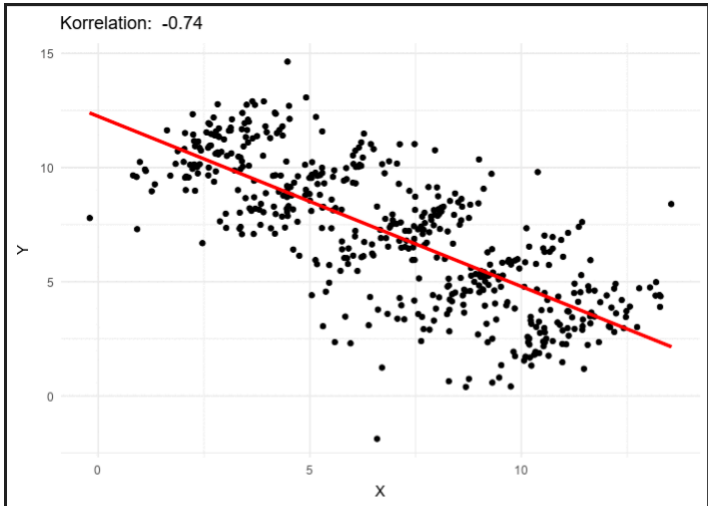
Solution 6.6. The Black box with 11 red and 9 blue balls has the probability to win a prize $P_{black} = \frac{11}{11+9} = \frac{11}{20}$. The White box with 12 red and 9 blue balls has $P_{white} = \frac{12}{12+9} = \frac{12}{21} = \frac{4}{7}$.

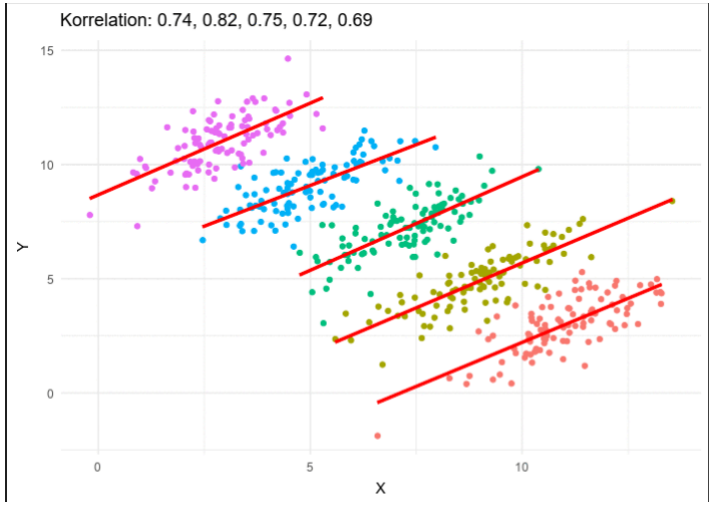
$\frac{11}{20} - \frac{4}{7} = \frac{11 \times 7 - 20 \times 4}{20 \times 7} = \frac{77 - 80}{20 \times 7} < 0$, so we choose the White box for a larger chance to win, not the Black box as in previous two problems.

So $\frac{5}{5+6} > \frac{3}{3+4}$ and $\frac{6}{6+3} > \frac{9}{9+5}$. After adding numbers the inequality is opposite $\frac{11}{11+9} < \frac{12}{12+9}$

Simpson's paradox

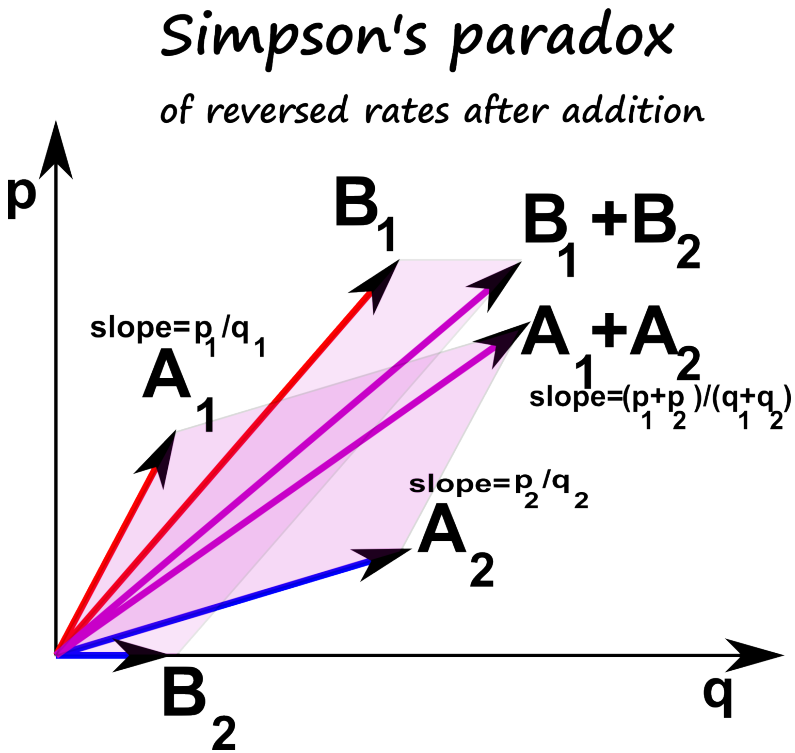
Simpson's (amalgamation) paradox: a trend in different groups can reverse when these groups are combined, see here.





Vector illustration

A probability $P_i = \frac{p_i}{q_i}$ is represented by a vector $A_i = (q_i, p_i)$ with the slope $\frac{p_i}{q_i}$. The vector sum is $A_1 + A_2 = (q_1 + q_2, p_1 + p_2)$ with the slope $\frac{p_1+p_2}{q_1+q_2}$.



Slope of $\bar{B}_1$ is smaller than the slope of $\bar{A}_1$, and slope of $\bar{B}_2$ is smaller than the slope of $\bar{A}_2$, but the sum $\bar{B}_1 + \bar{B}_2$ has a larger slope than $\bar{A}_1 + \bar{A}_2$.

Ecological/population fallacy

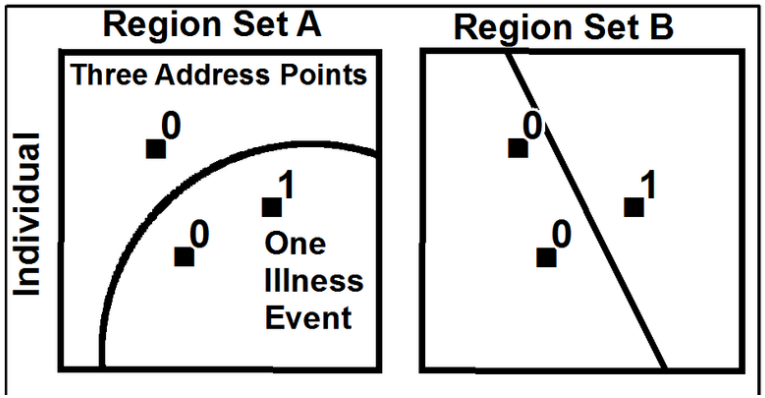
Ecological fallacy refers to wrong interpretations of statistical data that occurs when inferences about individuals are deduced from their group statistics,
https://en.wikipedia.org/wiki/Ecological_fallacy

Publications:

- Wang, Z., Rousseau, R. COVID-19, the Yule-Simpson paradox and research evaluation. Scientometrics 126, 3501–3511 (2021). https://doi.org/10.1007/s11192-020-03830-w
- Charig CR, etc. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed). 1986;292(6524):879-882. doi:10.1136/bmj.292.6524.879

Spatial ecologilcal fallacy

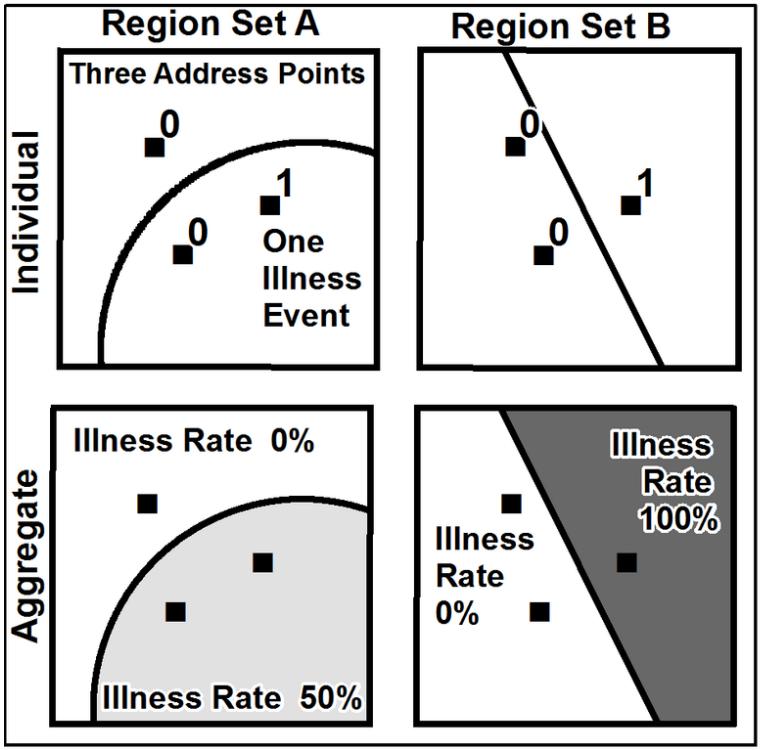
If we combine pitfalls in geometric intuition with the rates, we get a well-known in demography Modifiable Area Unit Problem (MAUP):



What are the disease rates if we draw those boundaries in the same region?

Spatial ecologilcal fallacy

If we combine pitfalls in geometric intuition with the rates, we get a well-known in demography Modifiable Area Unit Problem (MAUP):



Boundaries are important!

Averages do not help: rates are 25% and 50%, so quite differrent.

For a real life example, see Schuurman, etc. (2007). Deprivation indices, population health and geography: an evaluation of the spatial effectiveness of indices at multiple scales.

Mean people

Problem 6.7. Find the probability that a randomly chosen person has an above average number of legs.

Mean people

Problem 6.7. Find the probability that a randomly chosen person has an above average number of legs.

Solution 6.7. Arithmetic mean of legs in the population is close to 1.97. The probability to choose a person with 2 legs is close to 1, so the required probability is close to 0.

Can we locate a person with the average number of legs? Or, can we find a family with an average number of children?

Hint: average number of children per family in the UK is 1.7.

Journal Impact Factor

$$IF(\text{year } n) = \frac{Citations_n}{Publications_{n-1} + Publications_{n-2}}$$

The article "A short history of SHELX" in Acta Crystallographica Section A (2008) included the sentence "This paper could serve as a general literature citation...", it received more than 6,600 citations.
The impact factor of the journal rose from 2.051 to 49.926.

To compare: Nature had IF 31.434 and Science had IF of 28.103.

The 2nd-most cited article only had 28 citations.

More mean people

Group A: 80% of people got 40 points and 20% of them got 95 points. The mean score is 51 points.

Group B: 50% of people got 45 points and 50% got 55 points. The mean score is 50 points.

The comparison of means $51 > 50$ hints that
a random person from group A is likely to have
a higher score (win over) a random person from B, right?

Problem 6.8. Pick two people at random from groups A and B. Who is more likely to win?

Mean and likelihood are different

Solution 6.8. If taking two people at random, there are the following four cases

A: 40, B: 45 (B wins, 40% probability $= 0.8 \times 0.5$)

A: 40, B: 55 (B wins, 40% probability $= 0.8 \times 0.5$)

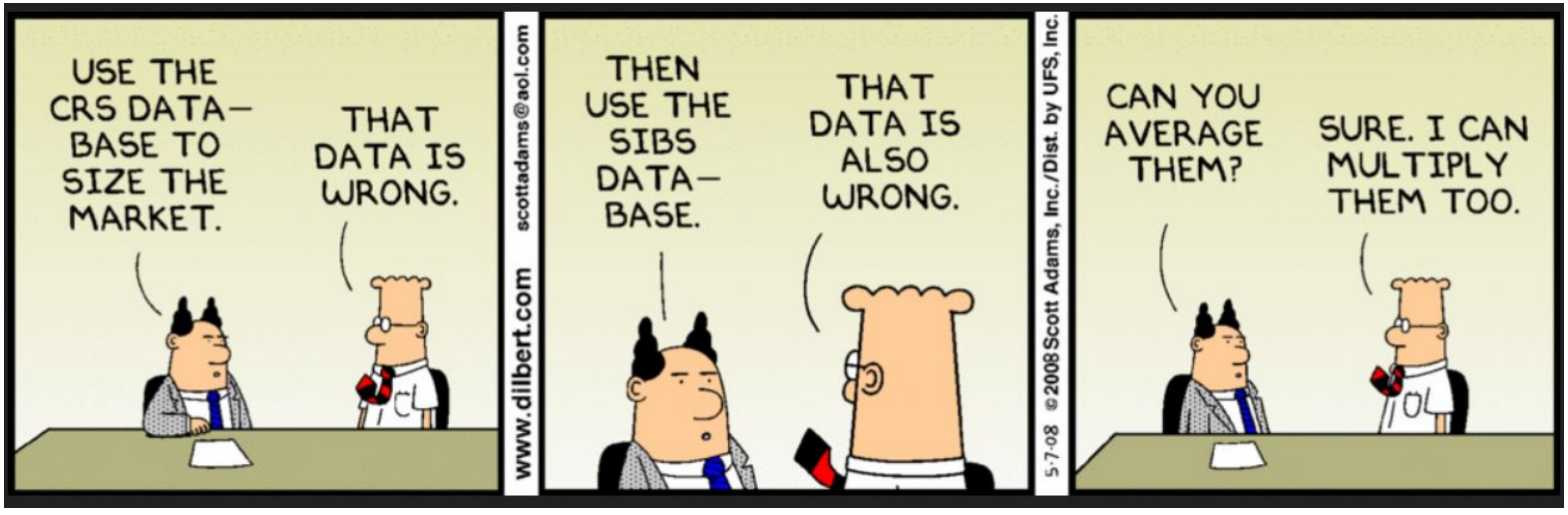
A: 95, B: 45 (A wins, 10% probability $= 0.2 \times 0.5$)

A: 95, B: 55 (A wins, 10% probability $= 0.2 \times 0.5$)

Group A has a higher mean, but in 80% cases a random person of A scores lower than a random person of B.

In ML: ensembles of poorly performing models can outperform better models.

Averages are not always the solution!



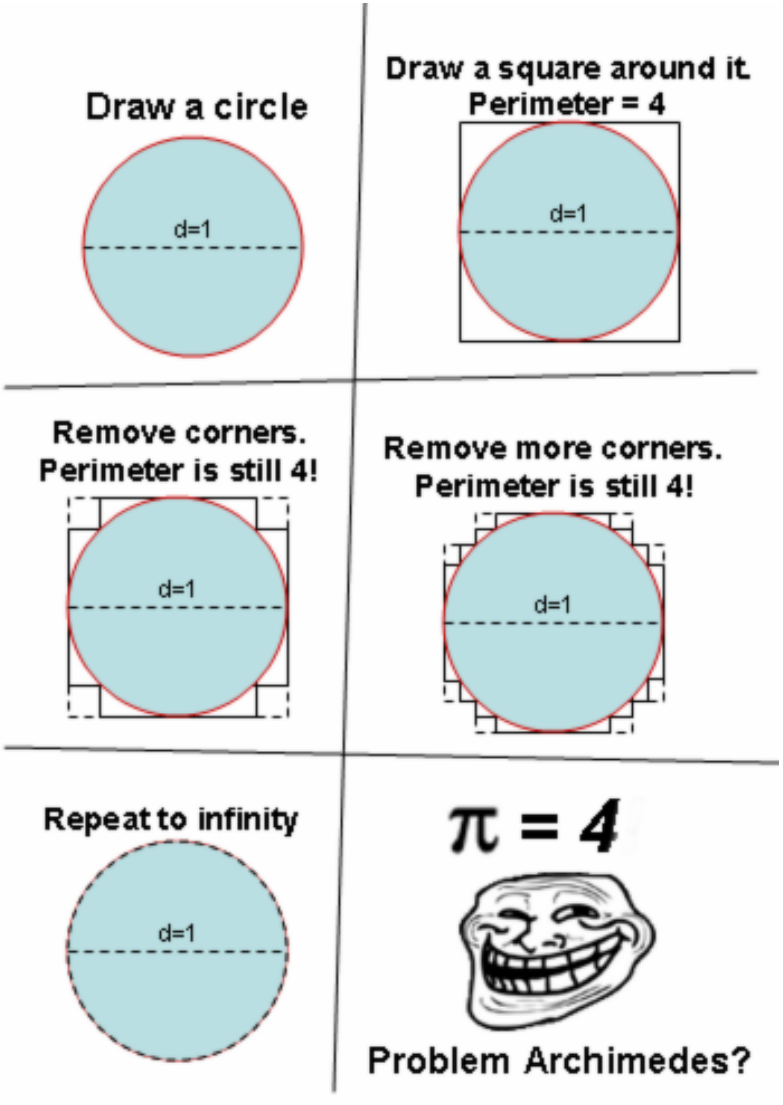
From discrete to continuos world

$$\int_a^b \int_c^d f(x,y)dx\,dy \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f(x_i,y_j)\Delta x \Delta y$$

Everything can be approximated by discrete cases, right?

For example, any smooth curve can be well approximated by polynomial lines, right?

New formula for π



The reason why the Archimedean approximation works and the rectilinear approximation doesn't is because for a 1-dimensional measurement in 2-dimensional space we need two measurements: not just the position of the curve, but also its direction.

Summary

- "Accuracy" can be misleading.
- Ecological fallacies: conclusions about individuals by using group statistics are not always justified.
- Beware mean subjects!
- Discrete approximation is not always suitable.
- Choose your tools wisely, depending on your problem and deep understanding!

Useful literature

- John A. Rice. Mathematical Statistics and Data Analysis.
- Martin Gardner. Aha! Gotcha: Paradoxes to Puzzle and Delight.
- Links from these slides.
- Applications of statistical tools, too many to name.