Methods for Causal Inference
Lecture 2

Ava Khamseh
School of Informatics

2021-2022

Causal theory and data

Requires 4 steps:

1. Definition of Causation
2. Clearly formulating causal **assumptions** and creating the **causal model**
3. Link the structure of casual model to features of data
4. **Estimation** given the causal model and data

Defining causation:
A variable X is a cause of a variable Y if Y in any way relies on X for its value.
(Intuitively: X is a cause of Y if Y listens to X and decides its value in response to what it hears)

Pre-requisites: Elementary concepts from probability theory, statistics, graph theory

Causal Inference in Statistics, Pearl (2016)

Basics of Probability

Most causal statements are uncertain: “drinking causes liver disease”, does not mean every person who consumes alcohol is certain to have liver disease

→ Need language and laws of probability.

Variables: Any property or descriptor that can take multiple values, e.g., age (x=40), sex (x’=F), family history of disease (x”=0), … .

Events: An event is any assignment of a **value or set of values** to a variable or set of variables.

Discrete (binary/categorical): Are being treated or not, have a disease or not, …

Continuous (can take infinite set of values): age, weight, …

Drug (yes/no) vs dose of drug (categorical). Sun intake (time is continuous),

Basics of Probability

For probabilistic modelling (of a random experiment) we need to:

- Describe possible outcomes: **sample space**
- **Event:** A subset of sample space
- Describe beliefs about likelihood of these events: **probability law**

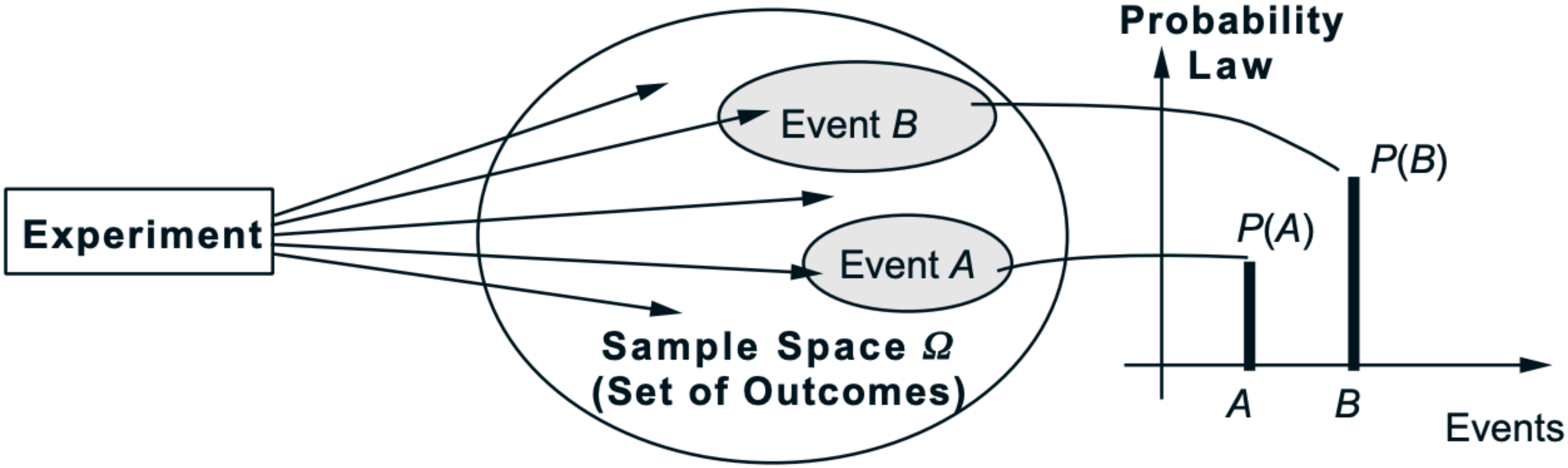
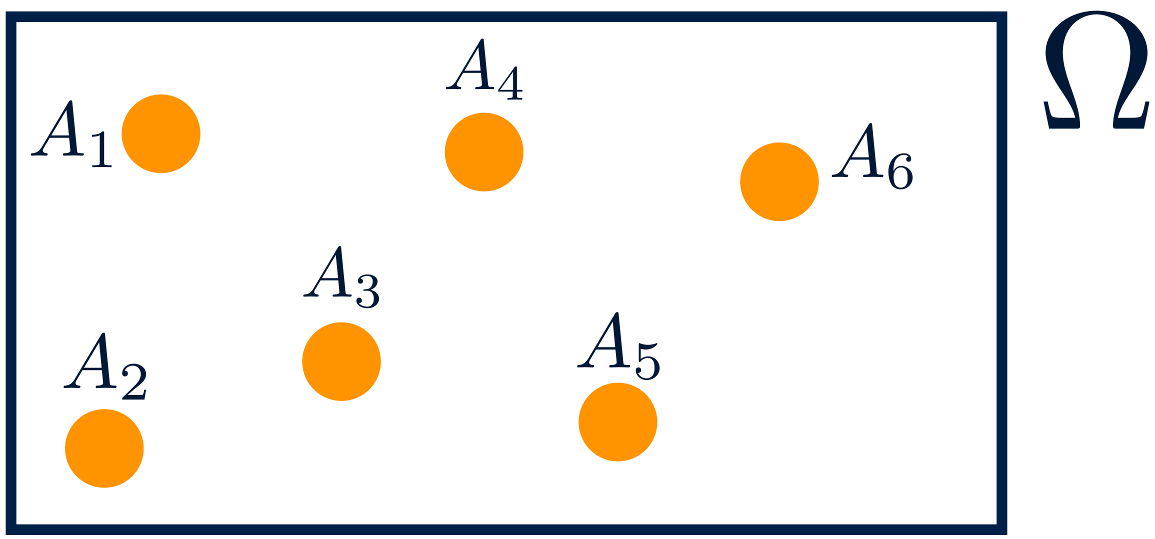


Figure from Tsitsiklis MIT Lecture Notes: Introduction to probability

Sample Space

The sample space is the set of all possible outcomes of the experiment:

e.g. Rolling a dice



Outcomes must be:

- **Mutually Exclusive**: If I tell you, after the experiment, that A_1 happened, then it should not be possible for that A_6 also happened.
- **Collectively Exhaustive**: Collectively, all the outcomes in Ω exhaust all possibilities

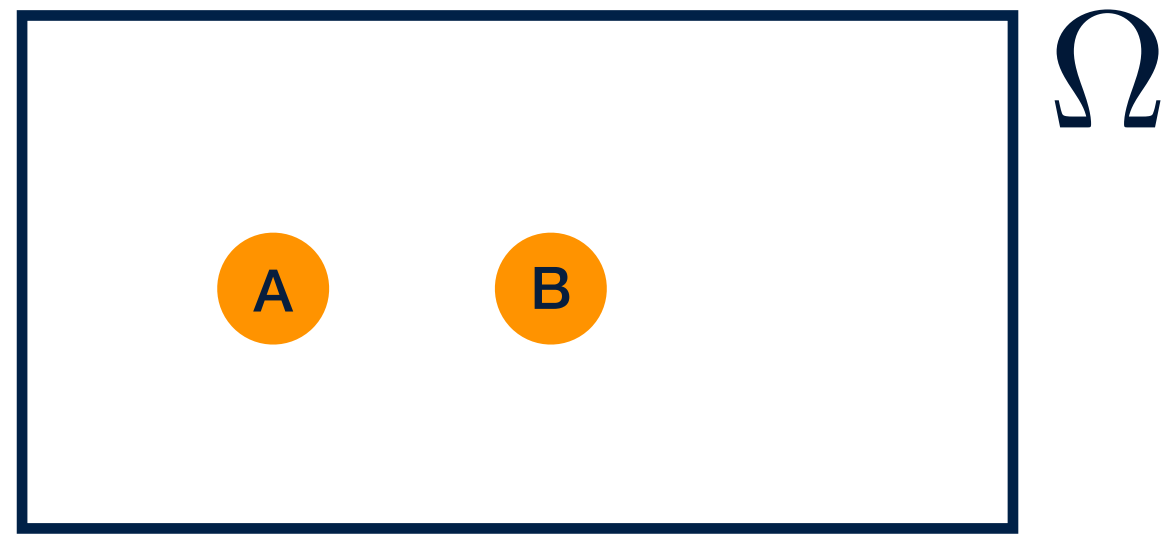
Tsitsiklis: Introduction to probability

Probability Axioms

Non-negativity: $P(A) > 0$

Normalisation: $P(\Omega) = 1$

For any two mutually exclusive events
(e.g. A and B cannot co-occur) we have:



$P(A \text{ or } B) = P(A) + P(B)$,

which implies, $P(A) = P(A, B) + P(A, \text{'not B'})$

A and B are mutually exclusive. If A is true, then either “A and B” or “A and not B” must be true. Generalise for exhaustive, mutually exclusive partitions of B:

Generalise: $P(A) = P(A, B1) + P(A, B2) + \ldots + P(A,Bn)$

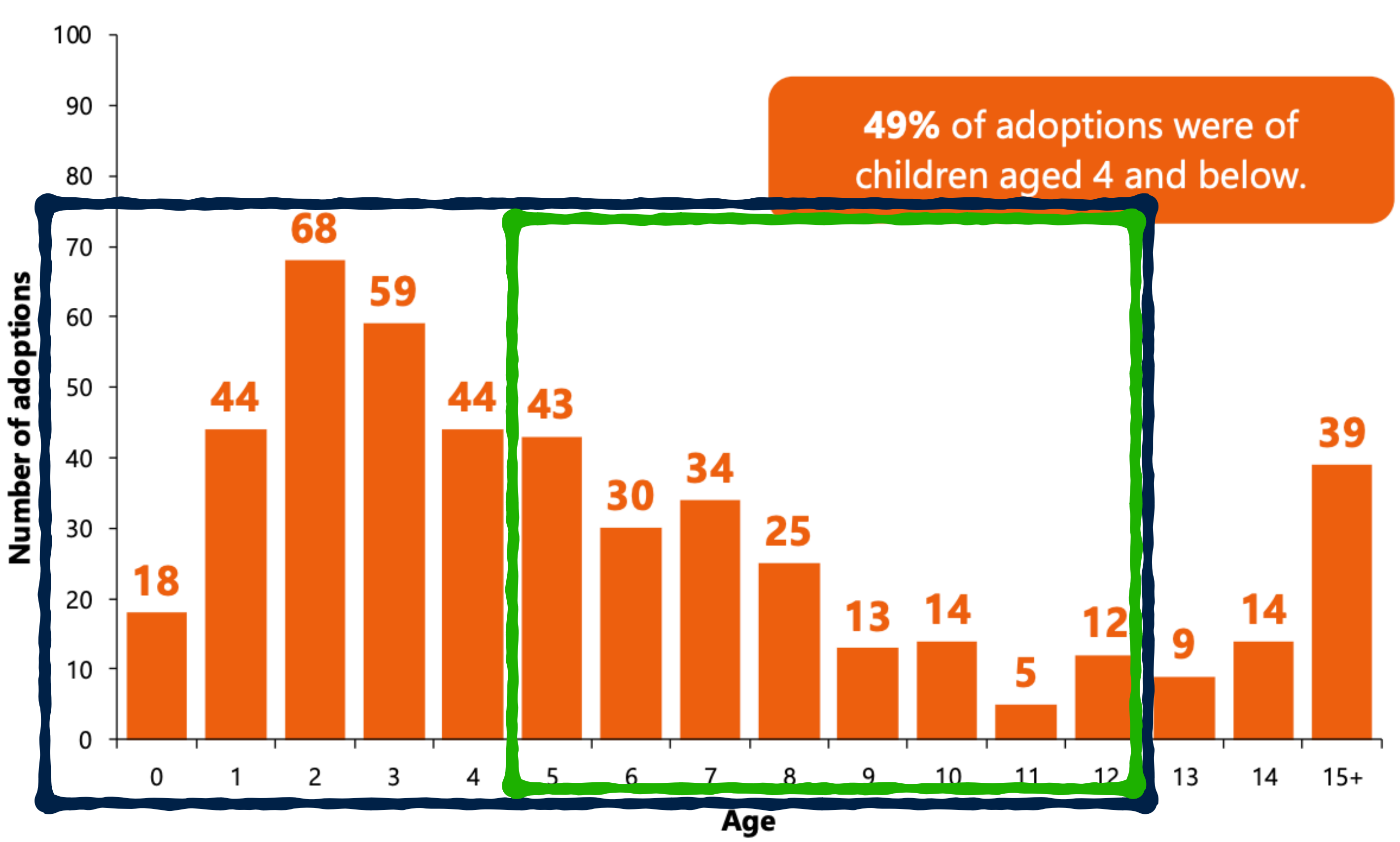
Causal Inference in Statistics, Pearl (2016)

Intervals

P(age > 4) = 1 - P(age <= 4) = 1- 0.49 = 0.51

P(4 < age < 12) =
(43+30+34+25+13+14+5+12) / 471 = 0.37

Figure 7.2: Age at adoption, Scotland, 2018



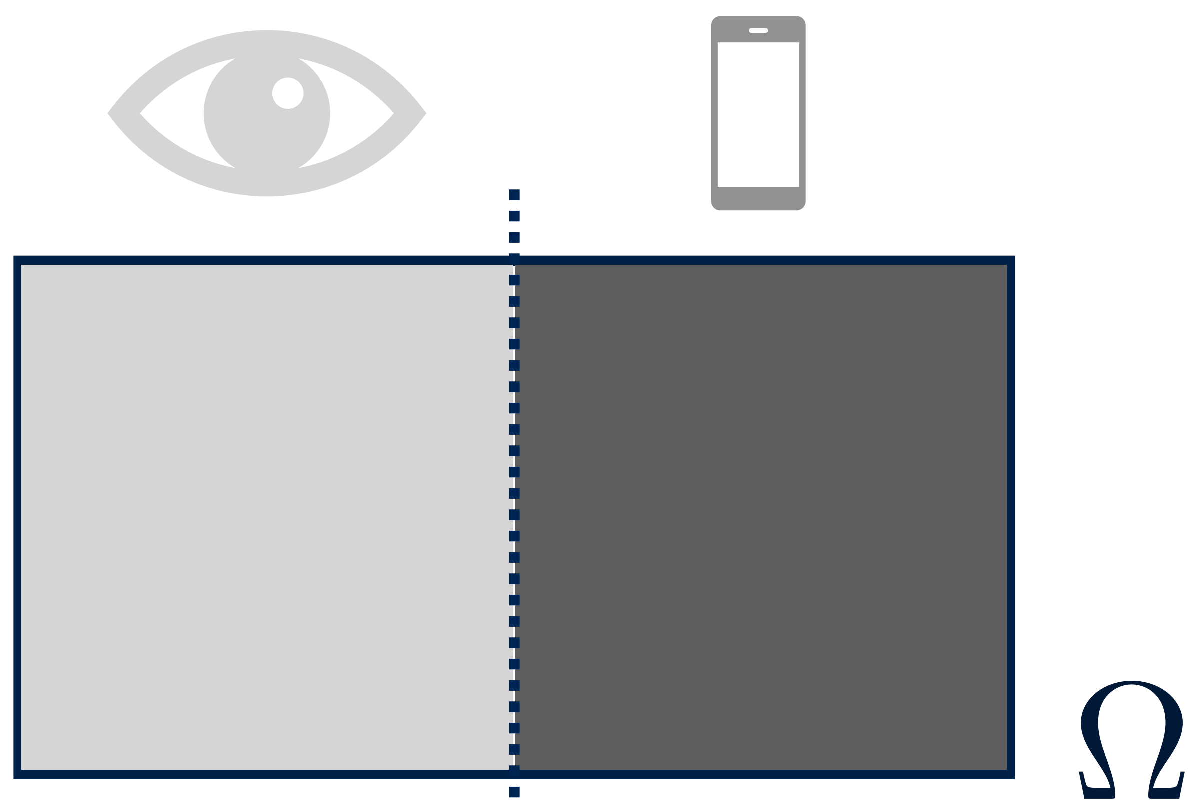
Total = 471

National Records of Scotland 2018

Law of Total probability: Example

Assuming ‘no multi-tasking’, the event:
“Passing the causality exam AND not being on your phone during the lectures”
is **mutually exclusive** from
“Passing the causality exam and being entirely on your phone during the lectures”

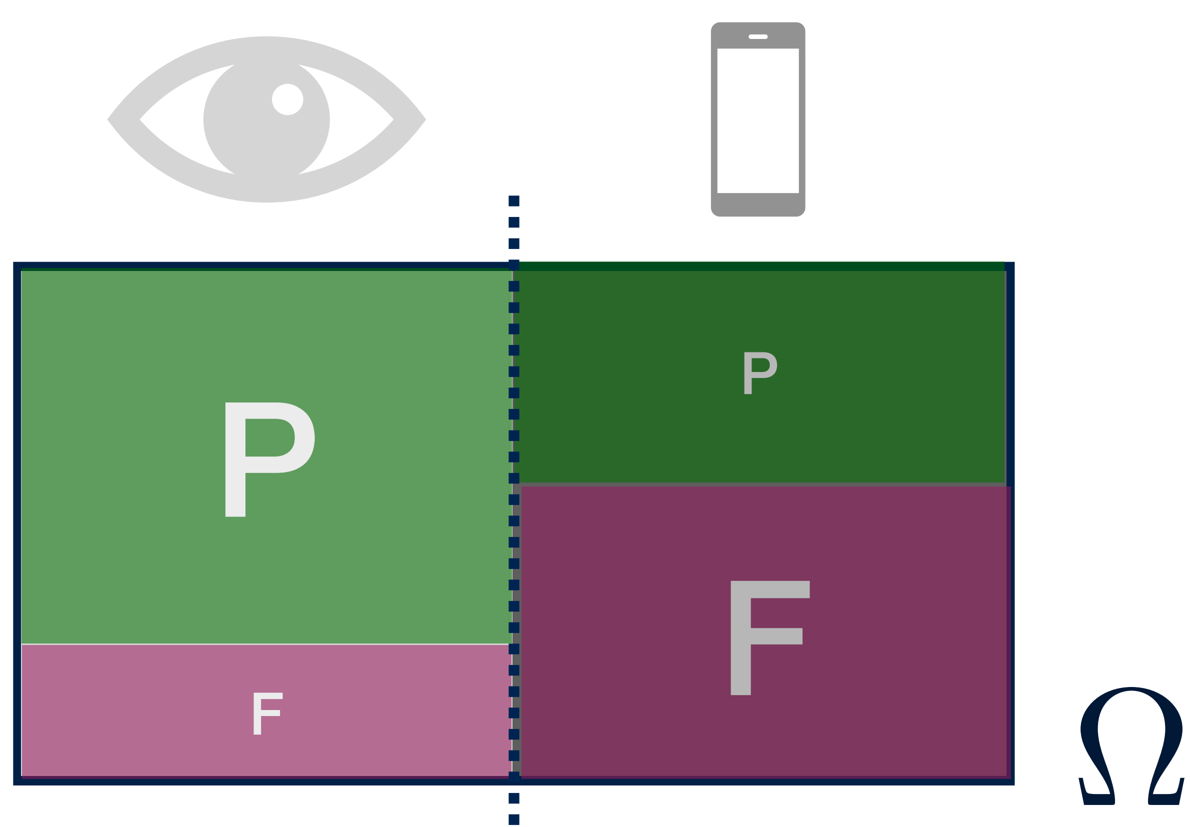
P(passing the causality exam) =
P(passing the exam, being entirely on your phone during the lecture) +
P(passing the exam, fully paying attention during the lecture)



Law of Total probability: Example

Assuming ‘no multi-tasking’, the event:
“Passing the causality exam AND not being on your phone during the lectures”
is **mutually exclusive** from
“Passing the causality exam and being entirely on your phone during the lectures”

P(passing the causality exam) =
P(passing the exam, being entirely on your phone during the lecture) +
P(passing the exam, fully paying attention during the lecture)



Conditional Probability

The probability that event A occurs, given that we know some other event B has occurred. (Think of filtering the data based on the value of some variable)

P(X=x) vs P(X=x|Y=y): The probability of X=x can drastically change depending on the knowledge Y=y

Example: P(lung cancer | smoker) vs

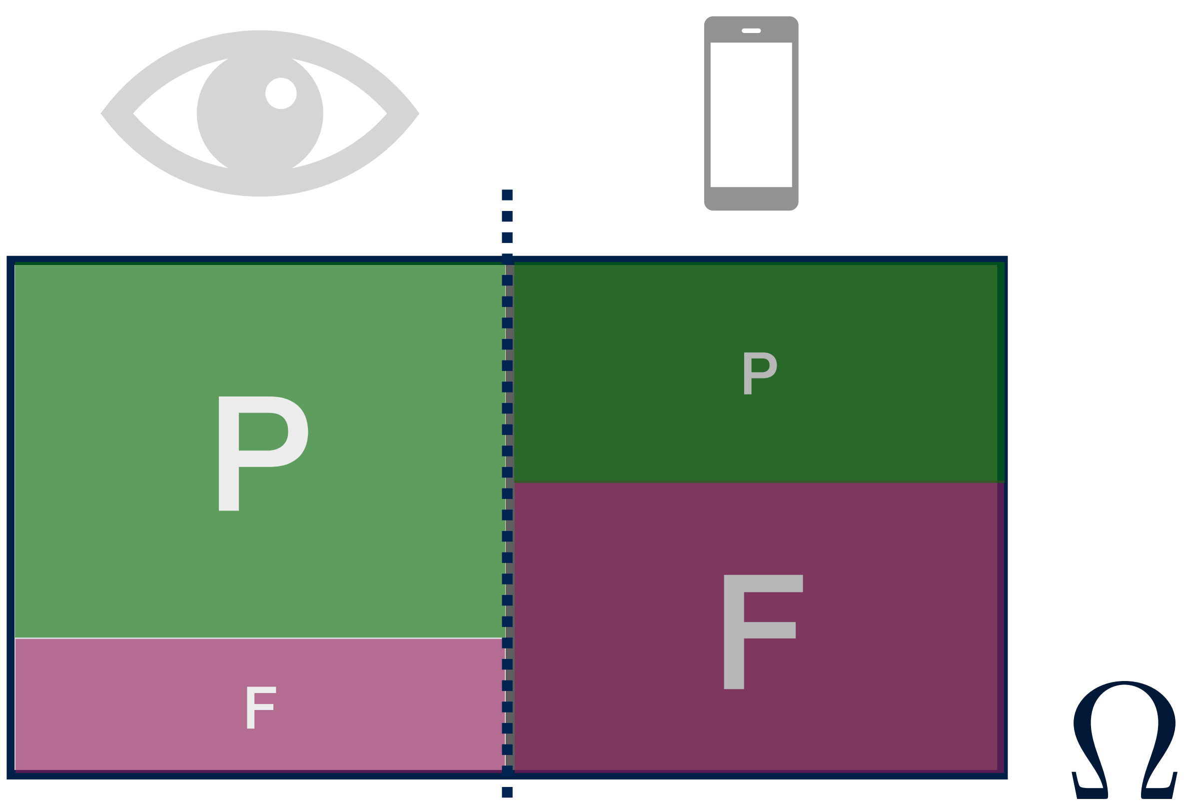
$$P(\underbrace{\text{lung cancer | smoker}}_{x}\ ,\ \underbrace{\text{socio-economic status}}_{Y})$$

Given that the patient is a smoker, does knowing their socio-economic status add further information to the probability of lung cancer?

$$P(X,Y) = P(X|Y)P(Y)$$

Conditional Probabilities

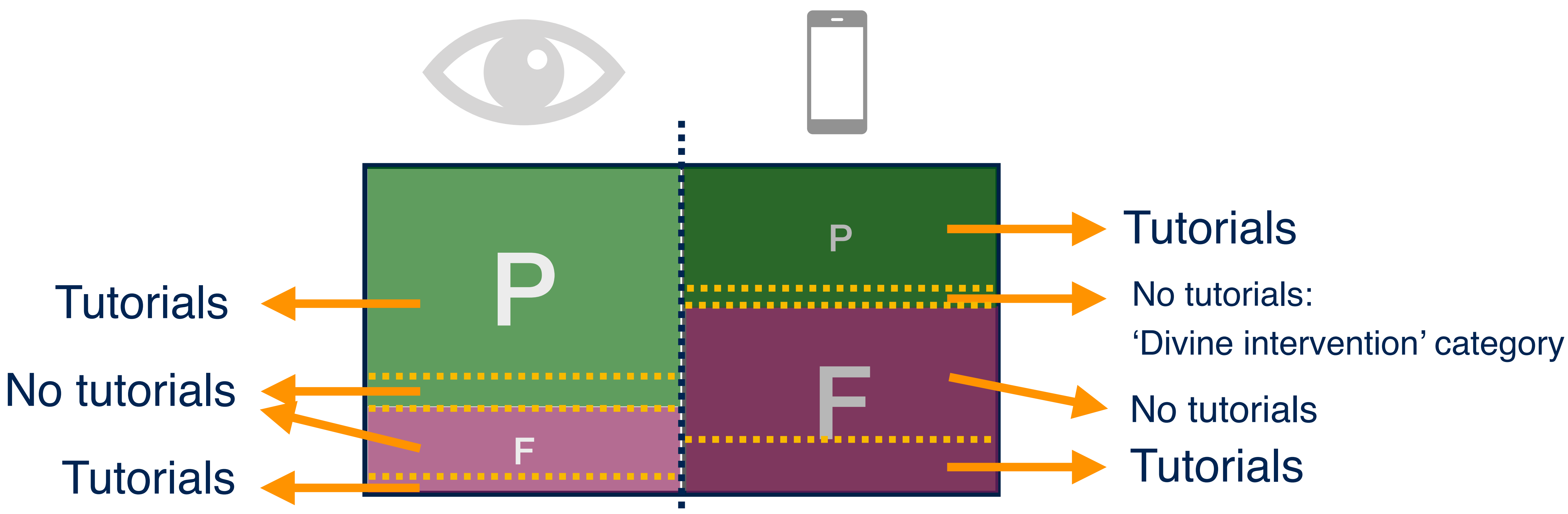
P(passing the causality exam | paying attention) >
P(passing the causality exam | being on your phone)



Conditional Law of Total probability: Example

P(passing the causality exam | fully paying attention during the lecture) =
P(passing the exam , attending tutorials | attention in lecture) +
P($\underbrace{\text{passing the exam}}_{X}$, $\underbrace{\text{not attending tutorials | attention in lecture}}_{Y}$)

P(passing the causality exam | being on one's phone during the lectures) =
P(passing the exam , attending tutorials | being on phone during lecture) +
P(passing the exam, not attending tutorials | being on phone lecture)



Bayes' Rule

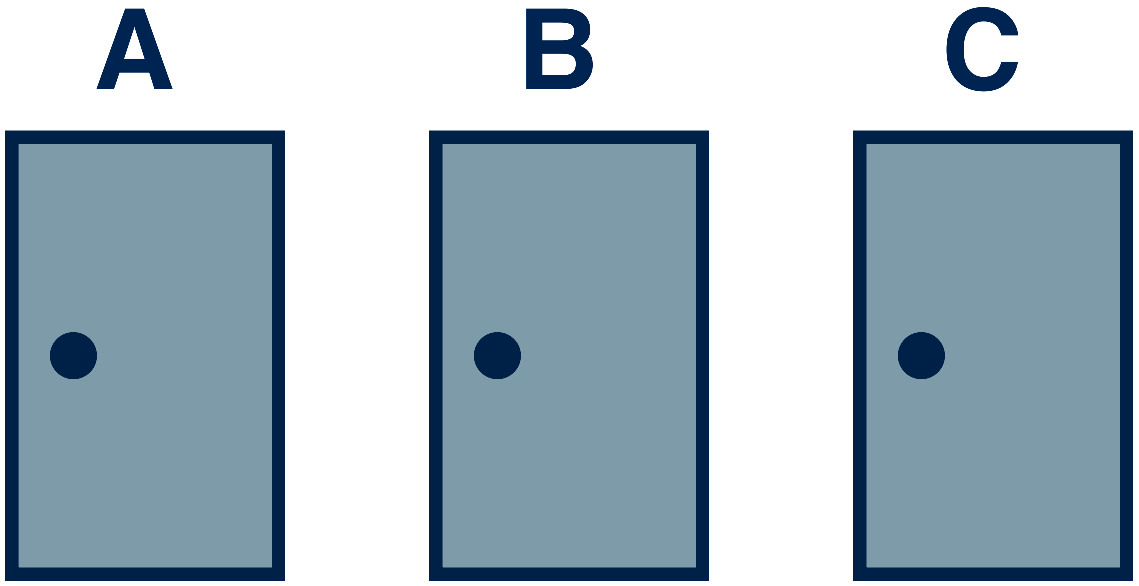
$X_1, X_2, \ldots, X_n$ are disjoint events forming a partition of the sample space

and $P(X_i) > 0$, $\forall X_i$. Then for any event Y , $P(Y) > 0$, Bayes' rule states:

$$P(X_i|Y) = \frac{P(X_i)P(Y|X_i)}{P(Y)}$$

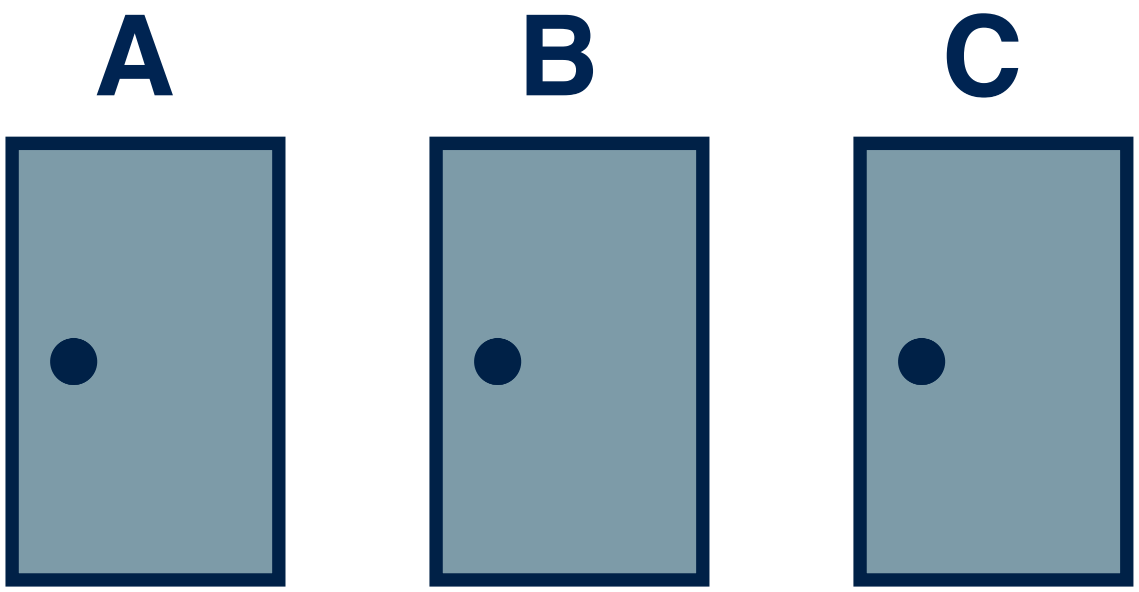
$$= \frac{P(X_i)P(Y|X_i)}{\underbrace{P(X_1)P(Y|X_1) + \cdots + P(X_n)P(Y|X_n)}_{\text{this is just normalised notation}}} = P(Y)$$

Monty Hall Problem & Application of Bayes' Rule



Car or Goat?

Monty Hall Problem & Application of Bayes' Rule

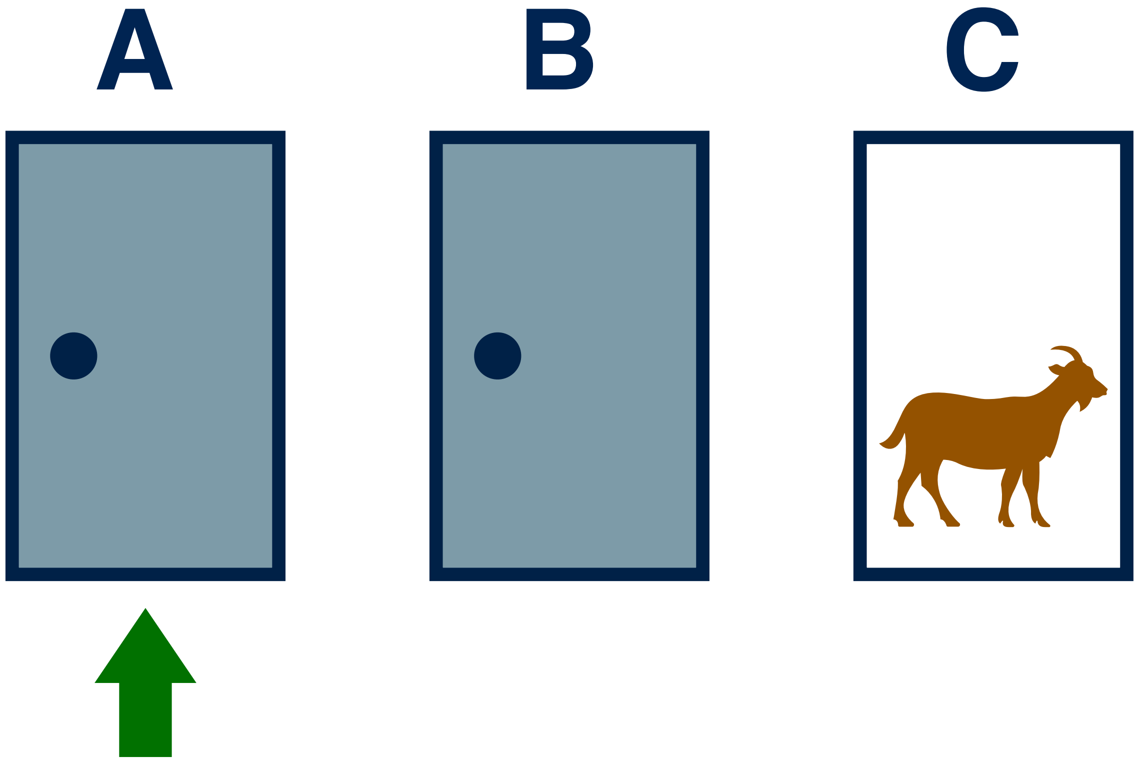


X = Door chosen by player

Y = Door hiding the car

Z = Door opened by host

Monty Hall Problem & Application of Bayes' Rule



X = Door chosen by player

Y = Door hiding the car

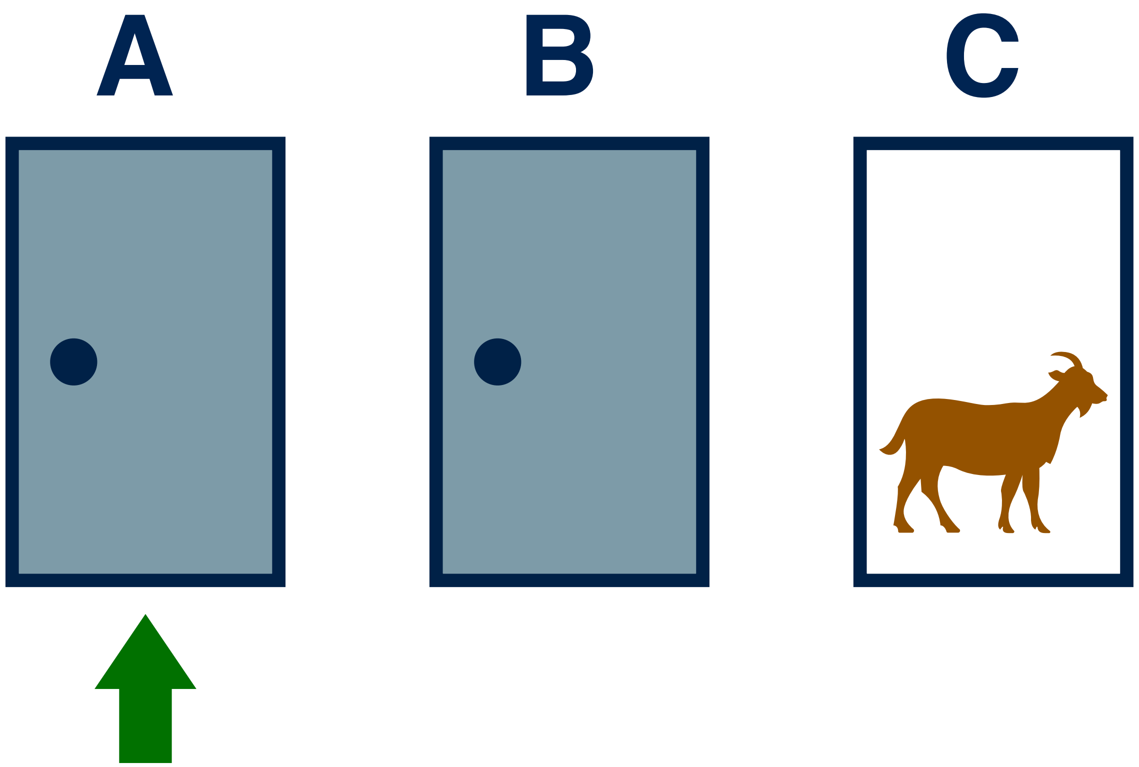
Z = Door opened by host

Prove that switching doors improves our chance of winning the car.

Note the assumptions:

1. The host will not open the door we have chosen
2. **The host will never open a door with a car behind**
3. Given a choice of doors, the host will choose at **random** (whilst 2)
4. Given no info, the car is equally likely to be behind any door

Monty Hall Problem & Application of Bayes' Rule



X = Door chosen by player

Y = Door hiding the car

Z = Door opened by host

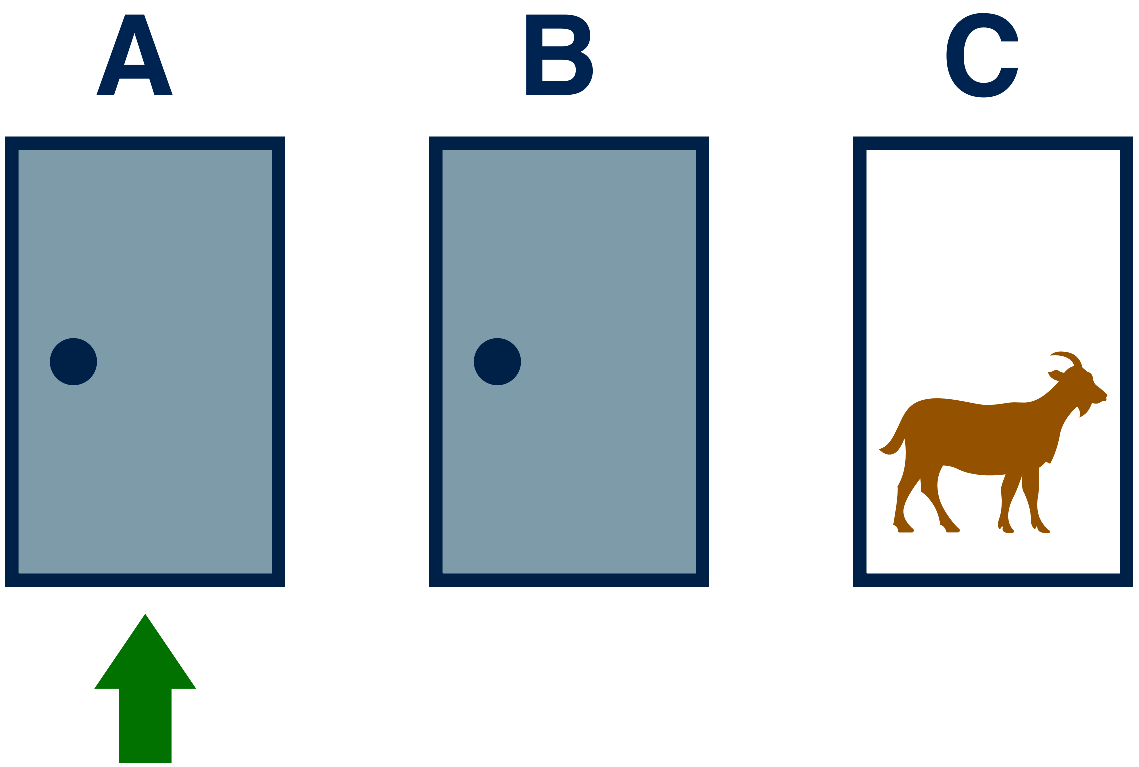
Prove that switching doors improves our chance of winning the car.

Need to show (given the we have selected A and host has shown us C):

$$P(Y = A | X = A, Z = C) < P(Y = B | X = A, Z = C)$$

Is the car more likely to be behind B than A, i.e. switching improves our chance.

Monty Hall Problem & Application of Bayes' Rule



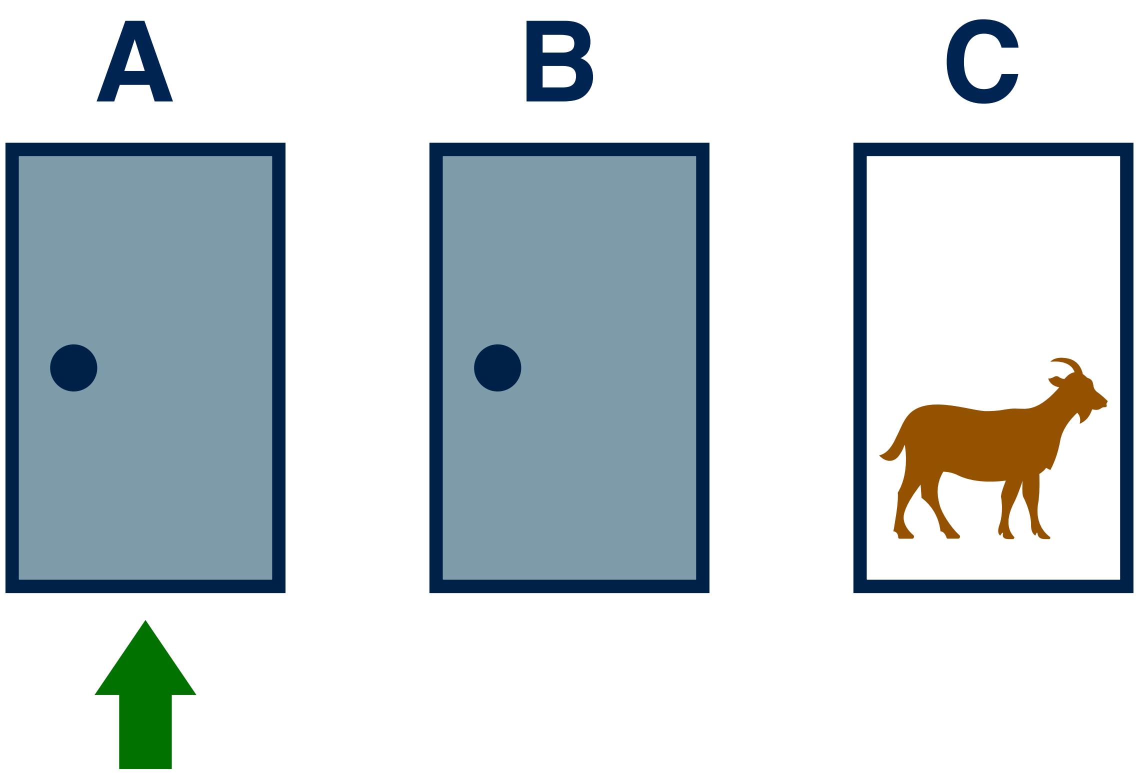
X = Door chosen by player

Y = Door hiding the car

Z = Door opened by host

$$P(Y=A|X=A,Z=C)=\frac{P(Z=C|X=A,Y=A)P(Y=A|X=A)}{P(Z=C|X=A)}$$

Monty Hall Problem & Application of Bayes' Rule



X = Door chosen by player

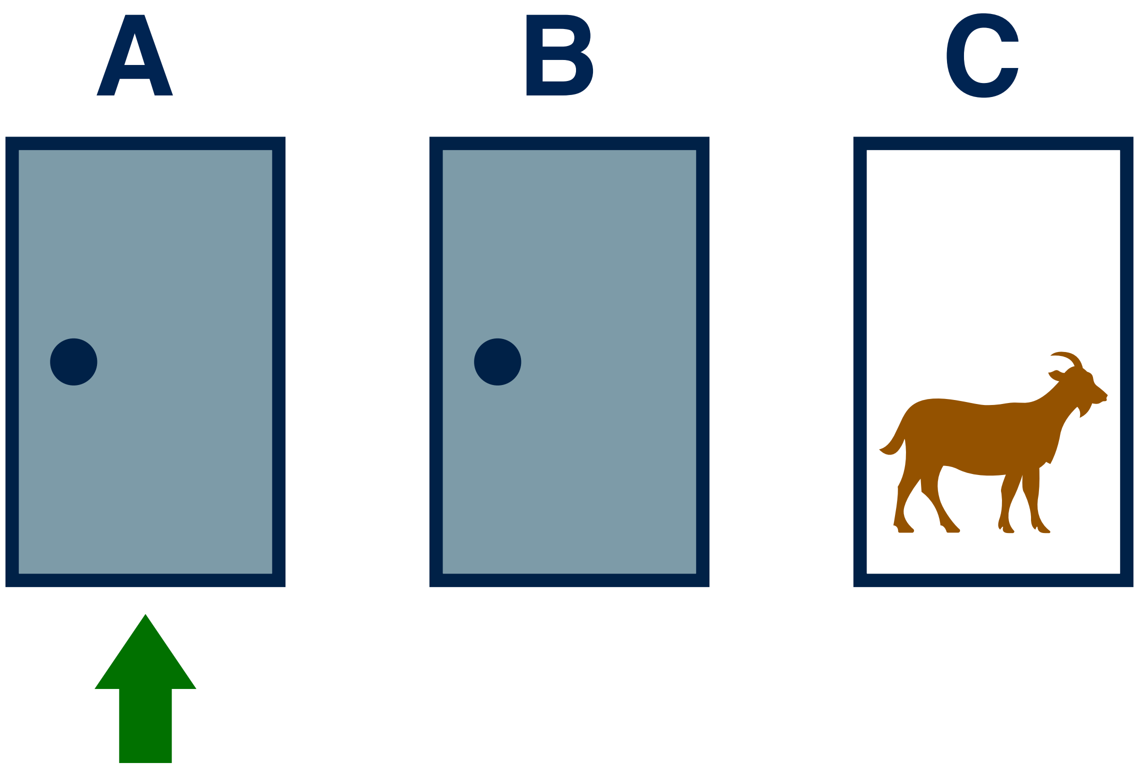
Y = Door hiding the car

Z = Door opened by host

$$P(Y = A|X = A, Z = C) = \frac{\overset{1/2}{\boxed{P(Z = C|X = A, Y = A)}}P(Y = A|X = A)}{P(Z = C|X = A)}$$

Given we choose A (X=A), and the car is in A (Y=A), then the host is allowed to choose either B or C, as neither has the car behind it. Since the host choses randomly (assumption 3), we get 1/2.

Monty Hall Problem & Application of Bayes' Rule



X = Door chosen by player

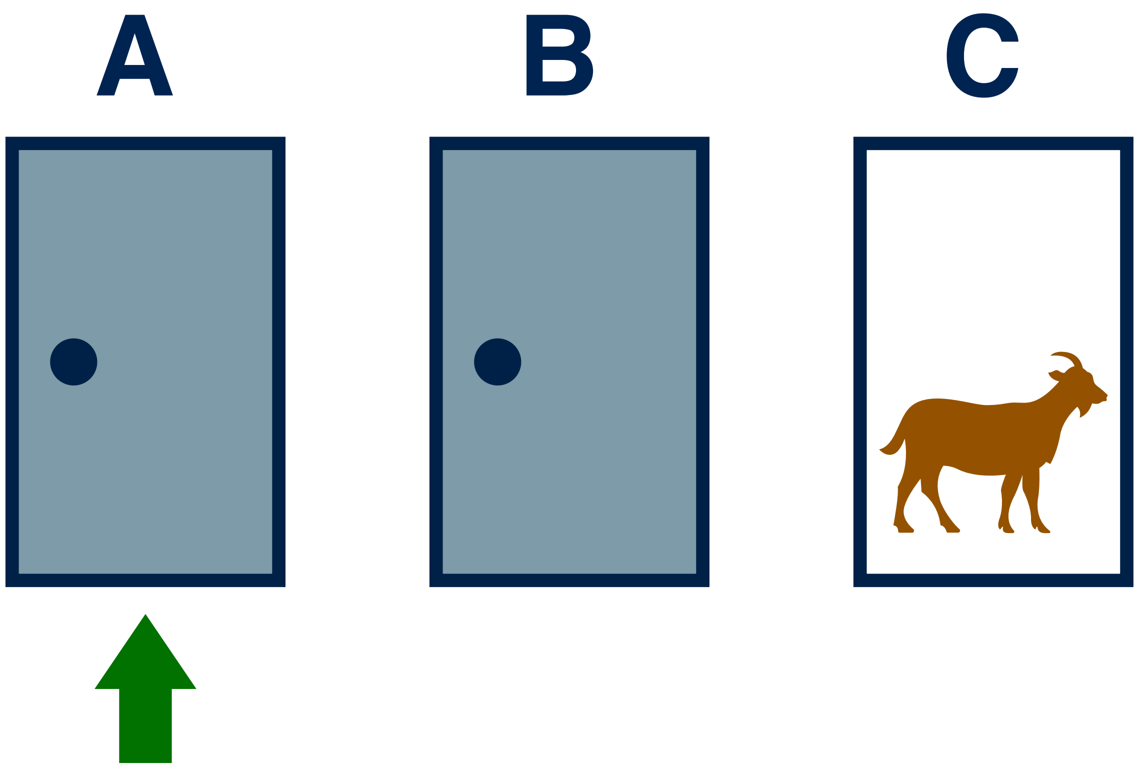
Y = Door hiding the car

Z = Door opened by host

$$P(Y=A|X=A,Z=C) = \frac{P(Z=C|X=A,Y=A)\overset{1/3}{\boxed{P(Y=A|X=A)}}}{P(Z=C|X=A)}$$

Given we choose A (X=A), what is the probability that the car is behind A? With no further information, this is equal to 1/3.

Monty Hall Problem & Application of Bayes' Rule



X = Door chosen by player

Y = Door hiding the car

Z = Door opened by host

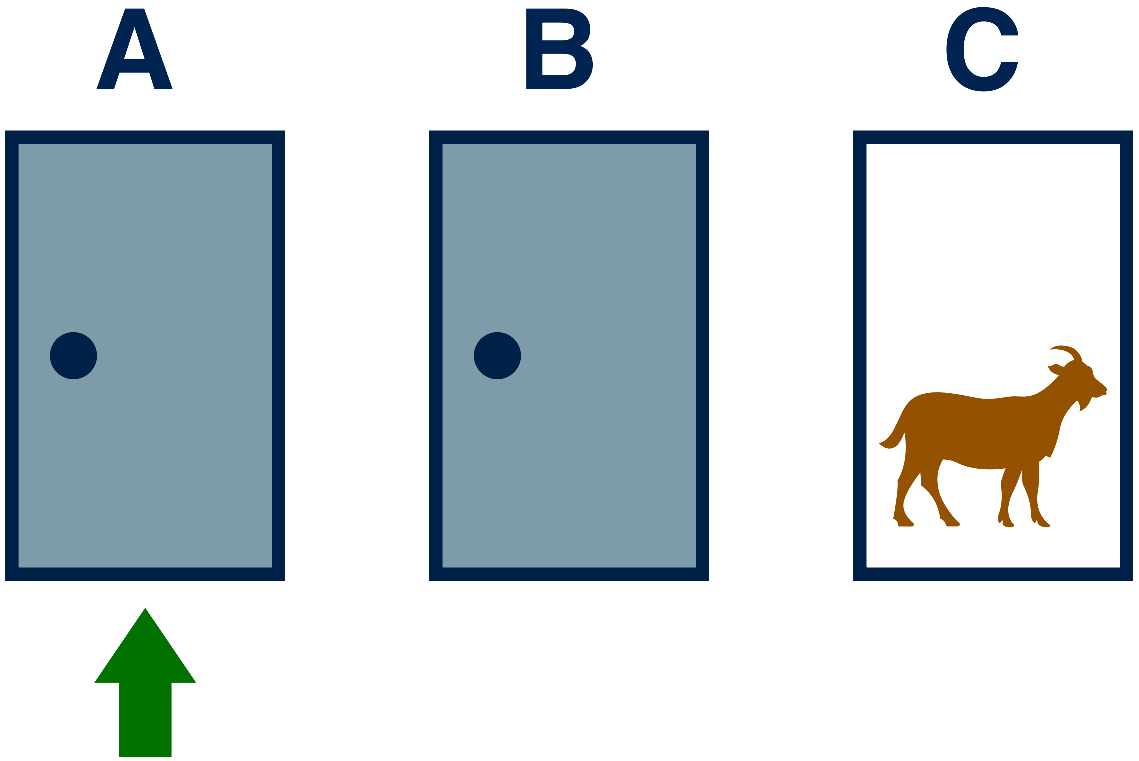
$$P(Y=A|X=A,Z=C)=\frac{P(Z=C|X=A,Y=A)P(Y=A|X=A)}{\boxed{P(Z=C|X=A)}}$$ 1/2

Total law of prob

Product rule

$$P(Z=C|X=A)=\sum_{d=A,B,C}P(Z=C,Y=d|X=A)=\sum_{d=A,B,C}P(Z=C|X=A,Y=d)P(Y=d)$$

Monty Hall Problem & Application of Bayes' Rule



X = Door chosen by player

Y = Door hiding the car

Z = Door opened by host

$$P(Y=A|X=A,Z=C) = \frac{P(Z=C|X=A,Y=A)P(Y=A|X=A)}{\boxed{P(Z=C|X=A)}}$$ 1/2

Total law of prob — Product rule

$$P(Z=C|X=A) = \sum_{d=A,B,C} P(Z=C,Y=d|X=A) = \sum_{d=A,B,C} P(Z=C|X=A,Y=d)P(Y=d)$$

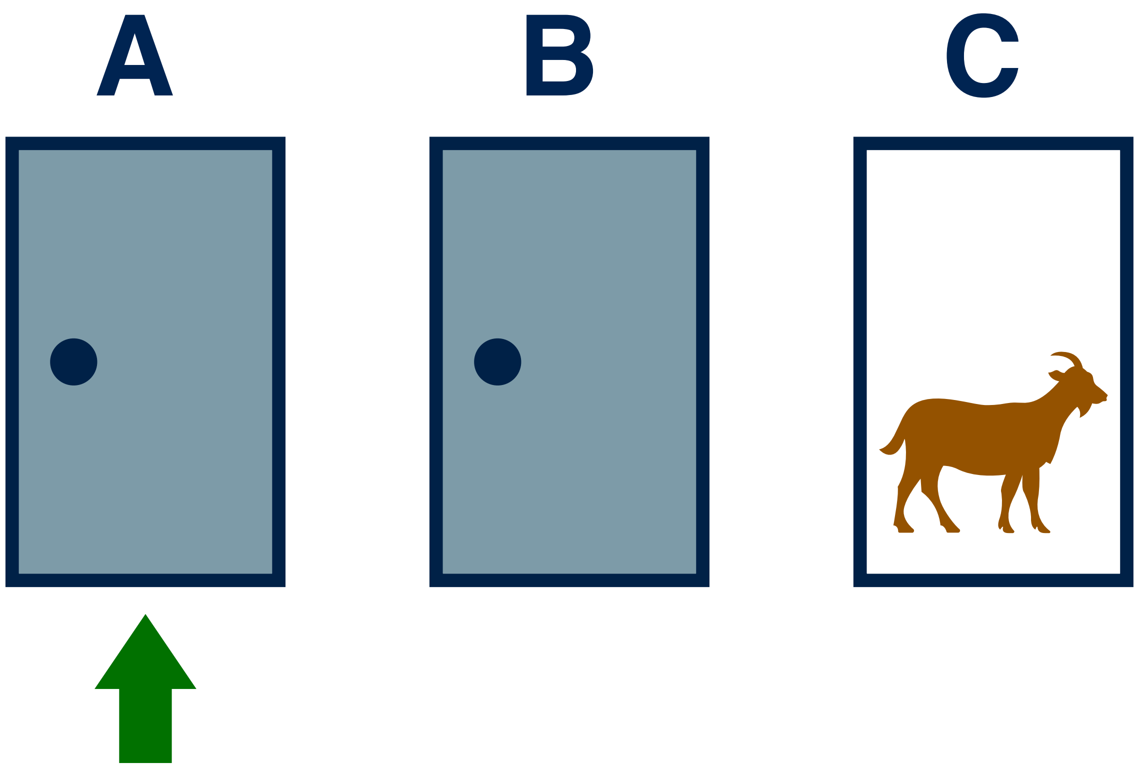
$$= \frac{1}{3}\Big(P(Z=C|X=A,Y=A) + P(Z=C|X=A,Y=B) + P(Z=C|X=A,Y=C)\Big) = 1/2$$

1/2 as above

1: Given we chose A and car is behind B, host is **forced** to choose C (Assumption 2)

0: Given we chose A and car is behind C, the host cannot choose C (Assumption 2)

Monty Hall Problem & Application of Bayes' Rule



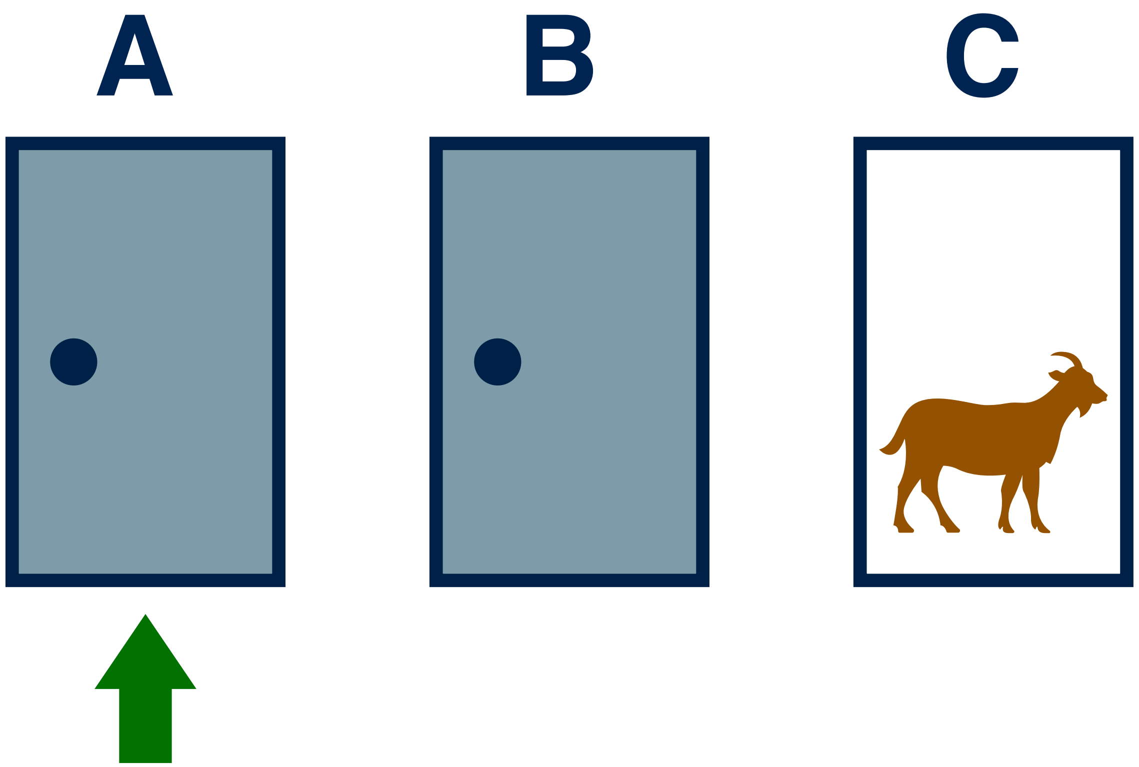
X = Door chosen by player

Y = Door hiding the car

Z = Door opened by host

$$P(Y = A|X = A, Z = C) = \frac{\overset{1/2}{P(Z = C|X = A, Y = A)}\,\overset{1/3}{P(Y = A|X = A)}}{P(Z = C|X = A)\quad 1/2} = 1/3$$

Monty Hall Problem & Application of Bayes' Rule



X = Door chosen by player

Y = Door hiding the car

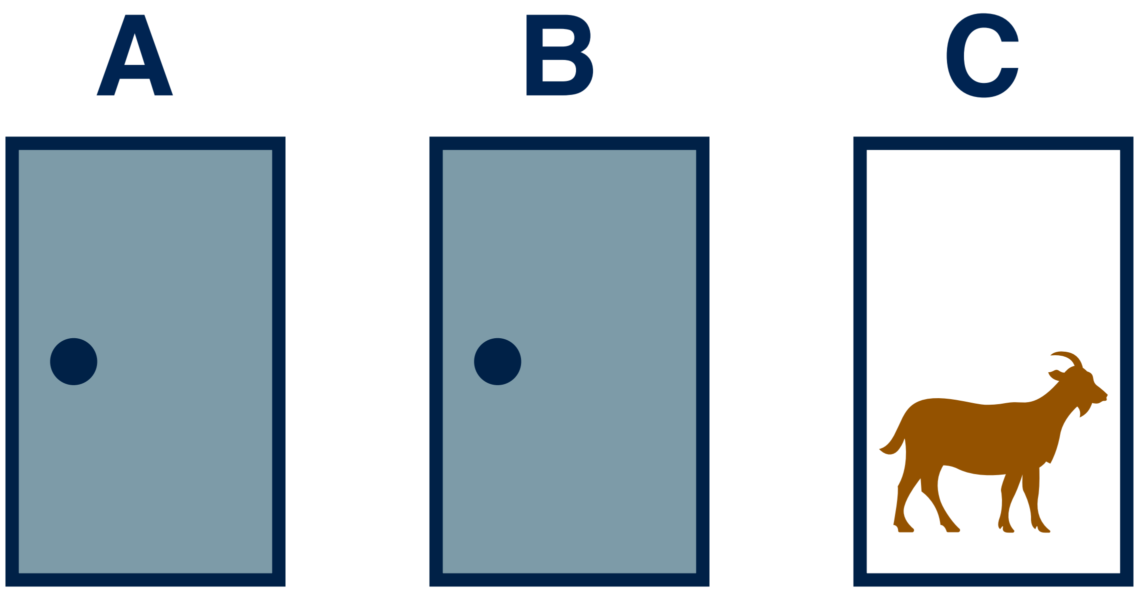
Z = Door opened by host

$$P(Y=A|X=A,Z=C) = \frac{\overset{1/2}{P(Z=C|X=A,Y=A)}\,\overset{1/3}{P(Y=A|X=A)}}{P(Z=C|X=A)\;\;1/2} = 1/3$$

$$\begin{aligned} P(Y=B|X=A,Z=C) =& 1 - P(Y=A|X=A,Z=C) - P(Y=C|X=A,Z=C) \\ =& 1 - \frac{1}{3} - 0 = 2/3 \end{aligned}$$

Mamma Mia!

Monty Hall Problem & Application of Bayes' Rule



X = Door chosen by player

Y = Door hiding the car

Z = Door opened by host

Importance: Incorporating knowledge about the process that generated the data. The first step towards **causal inference**.

'Host could have opened', 'he was forced to open', 'randomly opened', 'about to open', …

Independence

X and Y are independent events: $P(X,Y) = P(X)P(Y)$

Equivalently: $P(X|Y) = P(X)$ (where $P(Y)$ is non-zero, otherwise $P(X|Y)$ not defined)

Conditional independence: $P(X,Y|Z) = P(X|Z)P(Y|Z)$

Equivalently: $P(X|Y,Z) = P(X|Z)$ (again, for $P(Y,Z)$ non-zero)

Independence of several events: $$P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P(X_i)$$

Remark: Pairwise independence does not imply independence

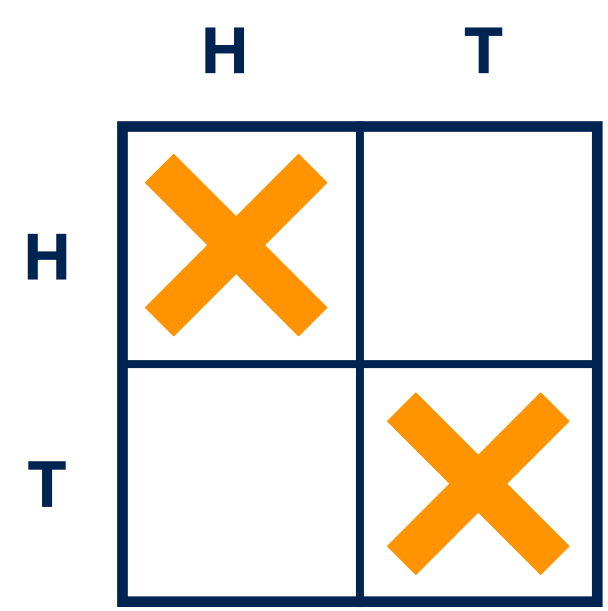
Example: 2 independent fair coin tosses (p1, p2 = 0.5)

Consider 3 events:

H1 = first coin is a head

H2 = second coin is a head

J = the two tosses have the same results



Independence

X and Y are independent events: P(X,Y) = P(X)P(Y)

Equivalently: P(X|Y) = P(X) (where P(Y) is non-zero, otherwise P(X|Y) not defined)

Conditional independence: P(X,Y|Z) = P(X|Z)P(Y|Z)

Equivalently: P(X|Y,Z) = P(X|Z) (again, for P(Y,Z) non-zero)

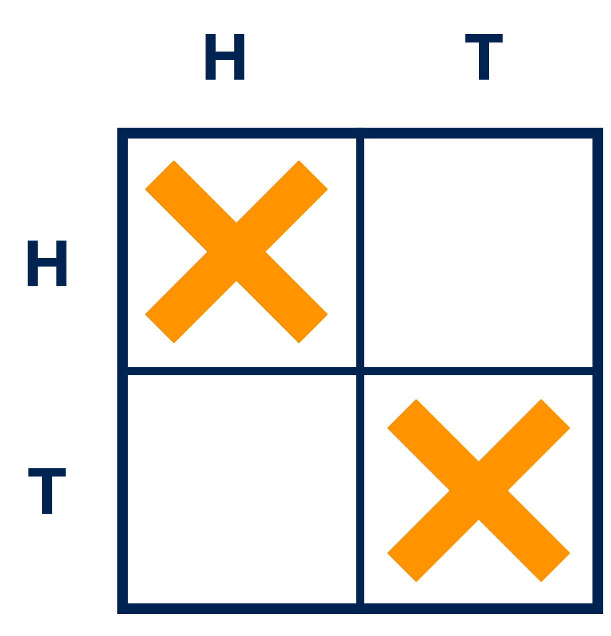
Independence of several events: $P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P(X_i)$

Remark: Pairwise independence does not imply independence

Example: 2 independent fair coin tosses (p1, p2 = 0.5)

H1 & H2: independent coin tosses

P(H1,H2) = P(H1|H2)P(H2) = 0.5x0.5 = P(H1)P(H2)



Independence

X and Y are independent events: P(X,Y) = P(X)P(Y)

Equivalently: P(XlY) = P(X) (where P(Y) is non-zero, otherwise P(XlY) not defined)

Conditional independence: P(X,YlZ) = P(XlZ)P(YlZ)

Equivalently: P(XlY,Z) = P(XlZ) (again, for P(Y,Z) non-zero)

Independence of several events: $P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P(X_i)$

Remark: Pairwise independence does not imply independence

Example: 2 independent fair coin tosses (p1, p2 = 0.5)

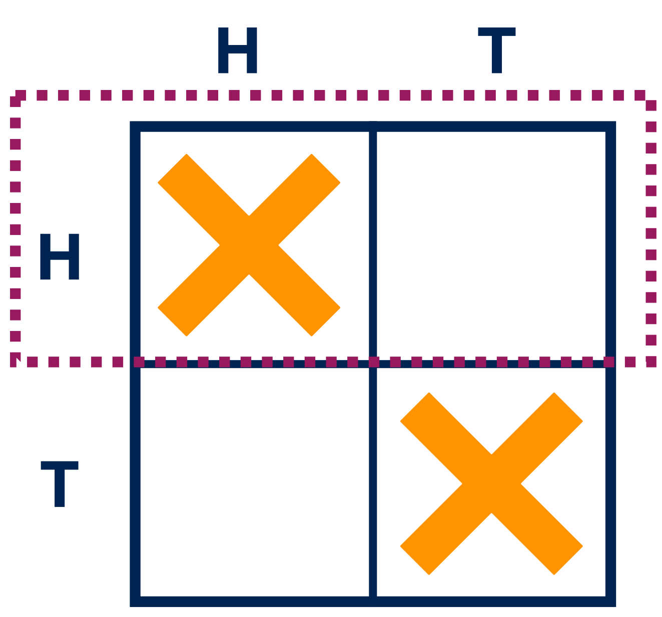
H1 & H2: independent coin tosses

P(H1,J) = P(J l H1)P(H1) =

Given H1, what is the probability of J

(i.e second toss also being a head)

So: P(J l H1) = 0.5



Independence

X and Y are independent events: P(X,Y) = P(X)P(Y)

Equivalently: P(XIY) = P(X) (where P(Y) is non-zero, otherwise P(XIY) not defined)

Conditional independence: P(X,YIZ) = P(XIZ)P(YIZ)

Equivalently: P(XIY,Z) = P(XIZ) (again, for P(Y,Z) non-zero)

Independence of several events: $P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P(X_i)$

Remark: Pairwise independence does not imply independence

Example: 2 independent fair coin tosses (p1, p2 = 0.5)

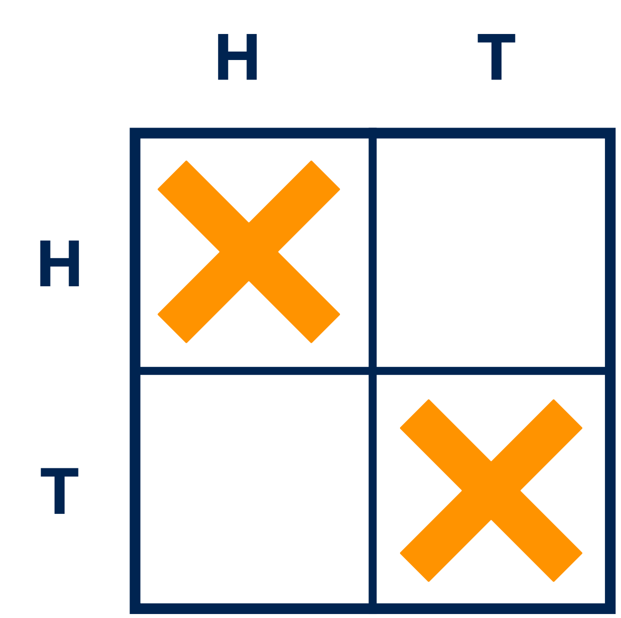
H1 & H2: independent coin tosses

P(H1,J) = P(J I H1)P(H1) = 0.5 x 0.5 = P(J)P(H1)

Given H1, what is the probability of J

(i.e second toss also being a head)

So: P(J I H1) = 0.5



Independence

X and Y are independent events: P(X,Y) = P(X)P(Y)

Equivalently: P(X|Y) = P(X) (where P(Y) is non-zero, otherwise P(X|Y) not defined)

Conditional independence: P(X,Y|Z) = P(X|Z)P(Y|Z)

Equivalently: P(X|Y,Z) = P(X|Z) (again, for P(Y,Z) non-zero)

Independence of several events: $P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P(X_i)$

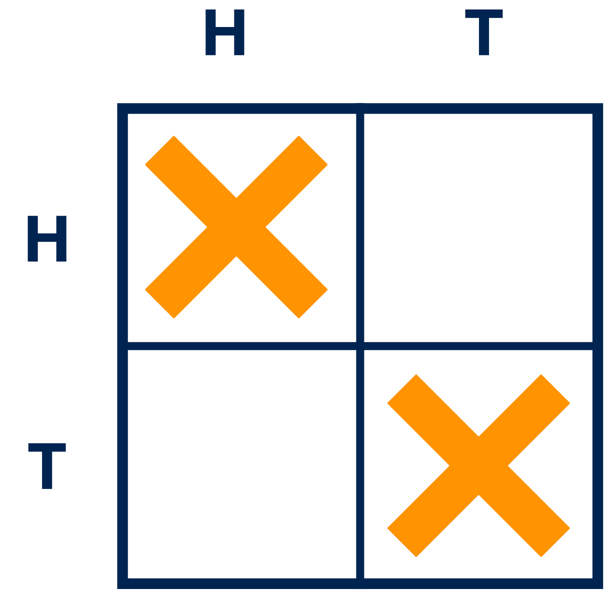
Remark: Pairwise independence does not imply independence

Example: 2 independent fair coin tosses (p1, p2 = 0.5)

H1 & H2: independent coin tosses

P(H2,J) = P(J | H2)P(H2) = 0.5 x 0.5 = P(J)P(H2)

So pair-wise independent. BUT …



Independence

X and Y are independent events: P(X,Y) = P(X)P(Y)

Equivalently: P(X|Y) = P(X) (where P(Y) is non-zero, otherwise P(X|Y) not defined)

Conditional independence: P(X,Y|Z) = P(X|Z)P(Y|Z)

Equivalently: P(X|Y,Z) = P(X|Z) (again, for P(Y,Z) non-zero)

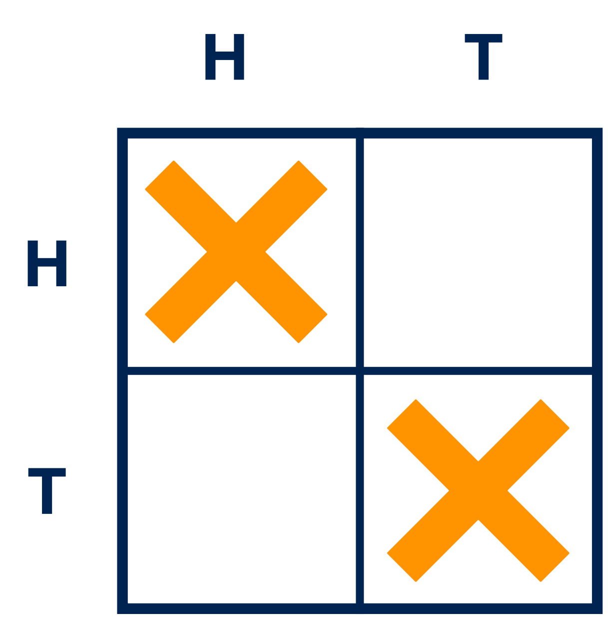
Independence of several events: $P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P(X_i)$

Remark: Pairwise independence does not imply independence

Example: 2 independent fair coin tosses (p1, p2 = 0.5)

H1 & H2: independent coin tosses

P(H1,H2,J) = P(H1 | H2,J) P(H2,J) = 1 x 0.25 = 0.25



Independence

X and Y are independent events: P(X,Y) = P(X)P(Y)

Equivalently: P(X|Y) = P(X) (where P(Y) is non-zero, otherwise P(X|Y) not defined)

Conditional independence: P(X,Y|Z) = P(X|Z)P(Y|Z)

Equivalently: P(X|Y,Z) = P(X|Z) (again, for P(Y,Z) non-zero)

Independence of several events: $P(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P(X_i)$

Remark: Pairwise independence does not imply independence

Example: 2 independent fair coin tosses (p1, p2 = 0.5)

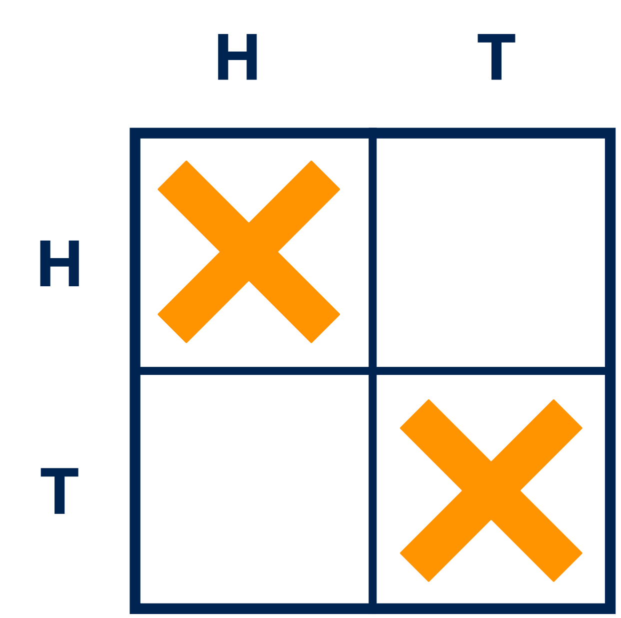
H1 & H2: independent coin tosses

P(H1,H2,J) = P(H1 | H2,J) P(H2,J) = 1 x 0.25 = 0.25

$\neq$

However, P(H1)P(H2)P(J)=0.5x0.5x0.5=0.125

i.e. not jointly independent



Expected Values

The probability distribution of a random variable X provides us with probabilities of all possible values of X.

Summarise information, with some loss of information, represented by:
The **expected value** or **mean**:

$$\mathbb{E}[X] = \sum_{x} x \; P(X = x)$$

For a dice: (1x1/6) + (2x1/6) + (3x1/6) + (4x1/6) + (5x1/6) + (6x1/6) = 3.5

Expected Values

The probability distribution of a random variable X provides us with probabilities of all possible values of X.

Summarise information, with some loss of information, represented by:
The **expected value** or **mean**:

$$\mathbb{E}[X] = \sum_{x} x \; P(X = x)$$

For a dice: (1x1/6) + (2x1/6) + (3x1/6) + (4x1/6) + (5x1/6) + (6x1/6) = 3.5

The expected value of any function of X, e.g. g(x):

$$\mathbb{E}[g(X)] = \sum_{x} g(x) \; P(X = x)$$

Dice: (1x1/6) + (4x1/6) + (9x1/6) + (16x1/6) + (25x1/6) + (36x1/6) = 15.17

Expected Values

The probability distribution of a random variable X provides us with probabilities of all possible values of X.

Summarise information, with some loss of information, represented by:
The **expected value** or **mean**:

$$\mathbb{E}[X] = \int x \ P(x) dx$$

for a continuous variable X.

Variance

The variance of a random variable X, denoted Var(X) or σ_X^2:

$$var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

and can be calculated as

$$var(X) = \sum_x (X - \mathbb{E}[X])^2 p_X(x)$$

(Integral of continuous variables), and measure how "spread out" the values of X in a data set are relative to their mean.

The standard deviation σ_X , (has the same units as X).

For a normal distribution, ~2/3 of the population values of X fall within one σ_X, 95% fall between 2 σ_X , etc.

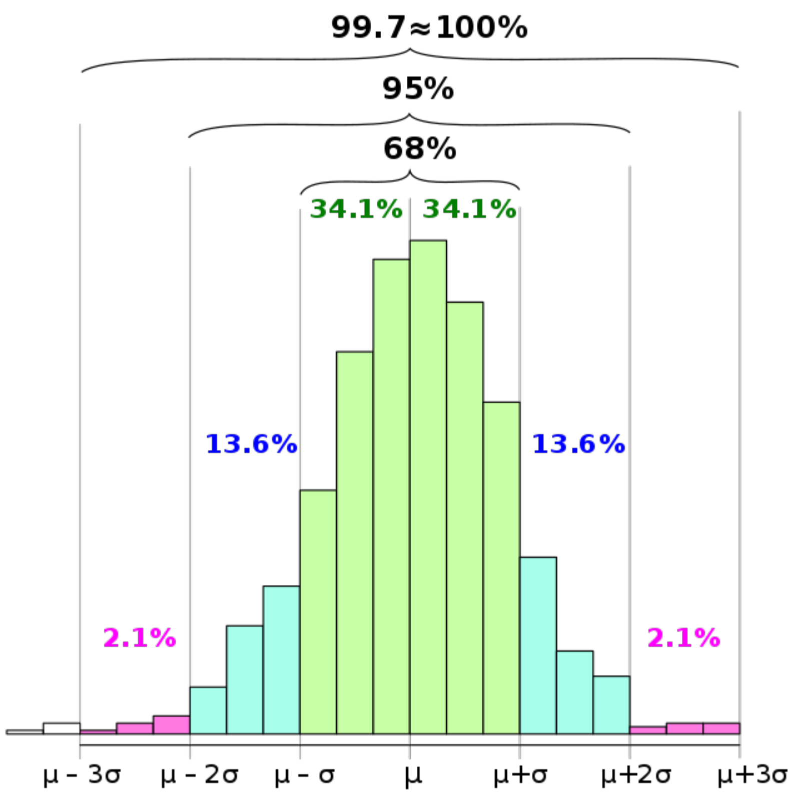


Figure from WikiPedia

Covariance

The degree to which two random variables X and Y covary (degree associated):

$$\sigma_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

and measures a specific way X and Y covary, i.e., **linearly**. When normalised, it yields the correlation coefficient (Pearson correlation):

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

a dimensionless quantity between -1 and 1.

When X and Y are independent, then $\rho_{XY} = 0$.
The reverse is not true!
(e.g. ρ_{XY} may be zero, but not linear-correlation, hence dependence exists. This requires more complex methods of demonstrating if $P(Y|X) = P(Y)$)

Anscombe's Quartet

Group of 4 datasets with nearly identical simple descriptive statistical properties:

- Mean and sample variance of X
- Mean and sample variance of Y
- Correlation between X and Y
- Linear regression line (coefficient the same up to 2 or 3 decimal places)
- R^2 coefficient

A note on R^2: A measure for goodness-of-fit

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)}{\sum_i (y_i - \bar{y})} \ , \ y_i = f(x_i) \ , \ \bar{y} = \frac{1}{n} \sum_i y_i$$

If the fit y=f(x) is a perfect fit, the numerator is zero, $R^2 = 1$, and
$R^2 = 0$ implies the fit f(x) is no better than baseline average $\bar{y}$.
Negative values corresponds to models worse than the baseline average.

Anscombe’s Quartet

Group of 4 datasets with nearly identical simple descriptive statistical properties:

- Mean and sample variance of X
- Mean and sample variance of Y
- Correlation between X and Y
- Linear regression line (coefficient the same up to 2 or 3 decimal places)
- R^2 coefficient

Yet, very different distributions, which can be observed by plotting the graphs

Same Pearson correlation, but,
different dependence structure
(X causes Y, but it different ways)

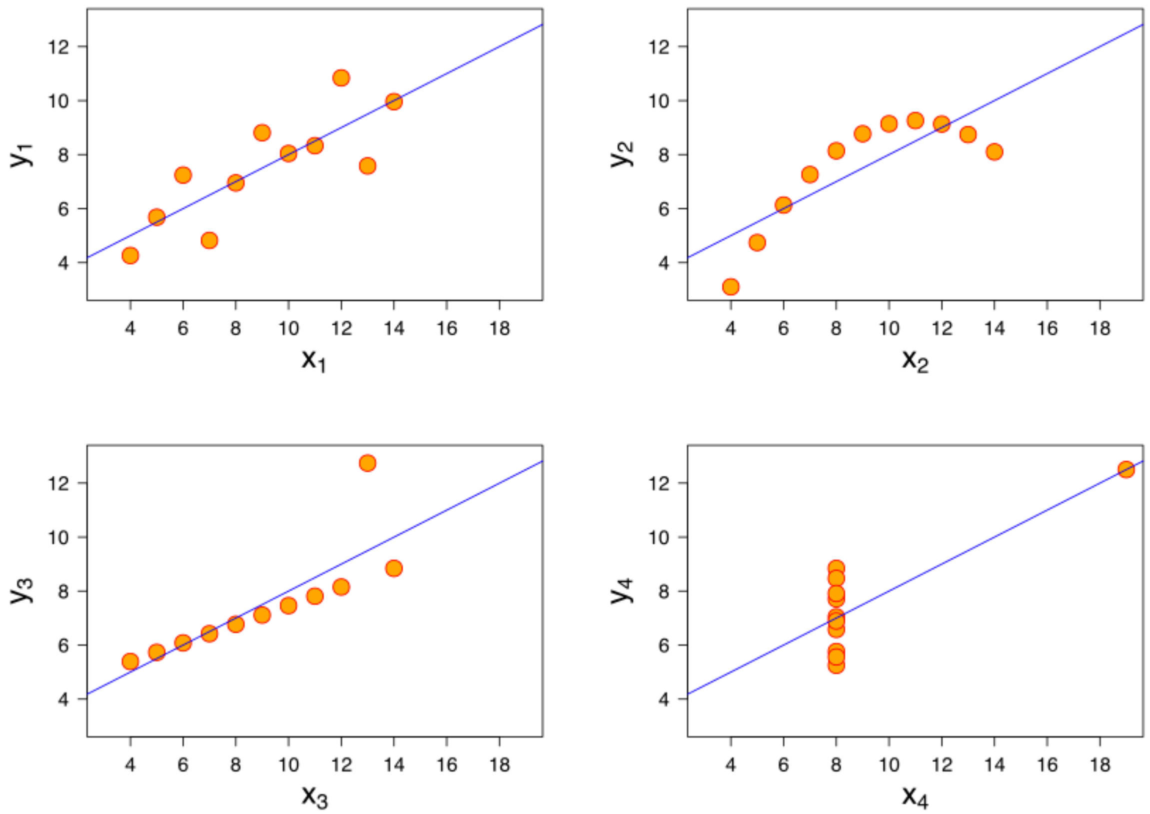


Figure from WikiPedia

Next time: Lecture 3

Regression, graphs, Structural Causal Models

Methods for Causal Inference
Lecture 2

Ava Khamseh

School of Informatics

2021-2022