

Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#131 Uncertainty and Randomness

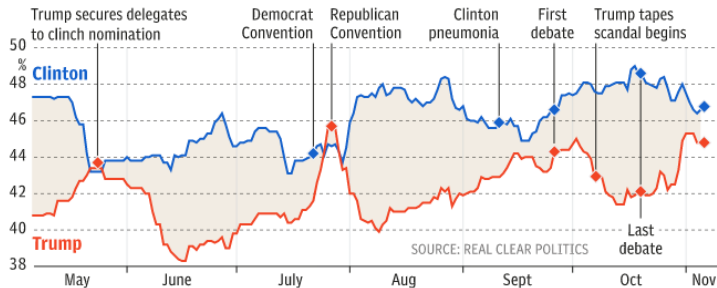
été 2020

Uncertainty



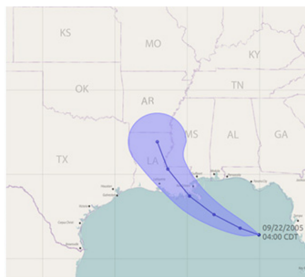
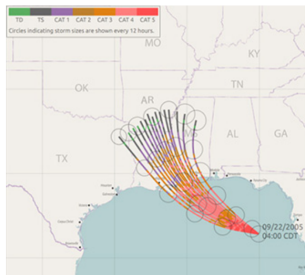
U.S. Elections

RACE FOR THE WHITE HOUSE: HOW THE POLLS HAVE CHANGED



Uncertainty

Data from an experiment may appear rock solid. Upon further examination, the data may morph into something much less firm. A knee-jerk reaction to this conundrum may be to try and hide uncertain scientific results, which are unloved fellow travelers of science. After all, words can afford ambiguity, but with visuals, “we are damned to be concrete,” says Bang Wong, who is the creative director of the Broad Institute of MIT and Harvard. The alternative is to face the ambiguity head-on through visual means.,
via **Data visualization: ambiguity as a fellow traveler** (see also **How data visualizations can clarify and confound uncertainty**)



Probabilities

See [visualizing uncertainty](#)

Consider a (standard) dice, taking values $\{1, 2, 3, 4, 5, 6\}$

Here, lower case denotes specific values x , e.g. $x = 3$

while upper case denotes a random variable X .

To describe X use can give its [distribution](#),

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = \dots = \mathbb{P}(X = 5) = \mathbb{P}(X = 6) = \frac{1}{6}$$

For any subset \mathcal{X} of $\{1, 2, 3, 4, 5, 6\}$ we can compute $\mathbb{P}(X \in \mathcal{X})$

Birthday Paradox

Consider a set of n randomly chosen people. If $n \geq 23$, there is more than 50% chances that some pair of them will have the same birthday.

(assuming that each day of the year is equally probable for a birthday)

"Class 1.pdf" p15

$$\mathbb{P}(A') = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \cdots \times \frac{343}{365} \simeq 49.2703\%$$

$$\text{Poisson approximation, } \lambda = \frac{1}{365} \binom{23}{2} = \frac{253}{365} \simeq 0.6932 \text{ so}$$

$$\mathbb{P}(X > 0) = 1 - \mathbb{P}(X = 0) \simeq 1 - e^{-0.6932} \simeq 0.500002$$

Joint Distribution

Consider two dices.

X_1 denotes the value of the first one,

X_2 denotes the value of the second one.

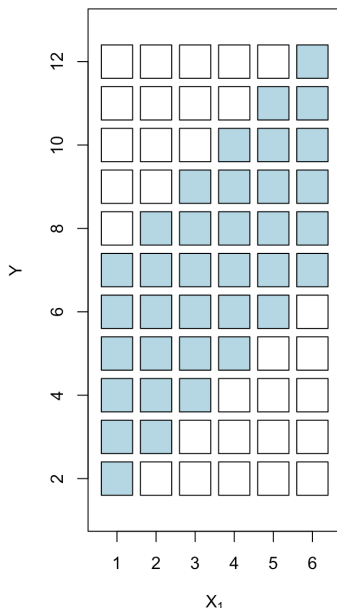
Let $Y = X_1 + X_2$.

$\mathbb{P}(X_1 = x_1, Y = y)$ is the (joint) probability

- ▶ $x_1 \in \{1, 2, \dots, 6\}$

- ▶ $y \in \{2, 3, \dots, 12\}$

E.g. $\mathbb{P}(X_1 = 4, Y = 3) = 0$

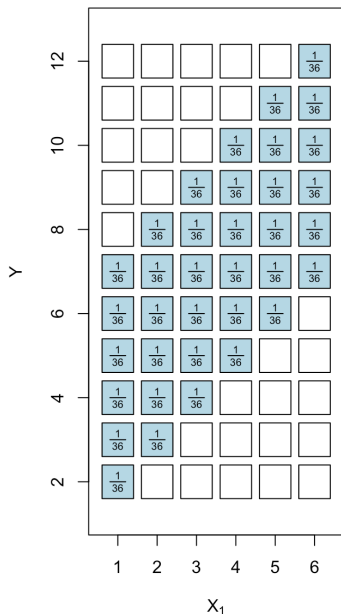


Conditional Events

$$\mathbb{P}(Y = y | X_1 = x_1) \stackrel{\text{def}}{=} \frac{\mathbb{P}(X_1 = x_1, Y = y)}{\mathbb{P}(X_1 = x_1)}$$

or $\mathbb{P}(X_1 = x_1, Y = y)$ is equal to

$$\underbrace{\mathbb{P}(Y = y | X_1 = x_1)}_{=\mathbb{P}(X_2 = y - x_1)} \cdot \mathbb{P}(X_1 = x_1)$$



Conditional Events

$$\mathbb{P}(Y = y | X_1 = x_1) \stackrel{\text{def}}{=} \frac{\mathbb{P}(X_1 = x_1, Y = y)}{\mathbb{P}(X_1 = x_1)}$$

and similarly

$$\mathbb{P}(X_1 = x_1 | Y = y) = \frac{\mathbb{P}(X_1 = x_1, Y = y)}{\mathbb{P}(Y = y)}$$

We can write

$$\mathbb{P}(Y = y | X_1 = x_1) = \frac{\mathbb{P}(Y = y)}{\mathbb{P}(X_1 = x_1)} \cdot \mathbb{P}(X_1 = x_1 | Y = y)$$

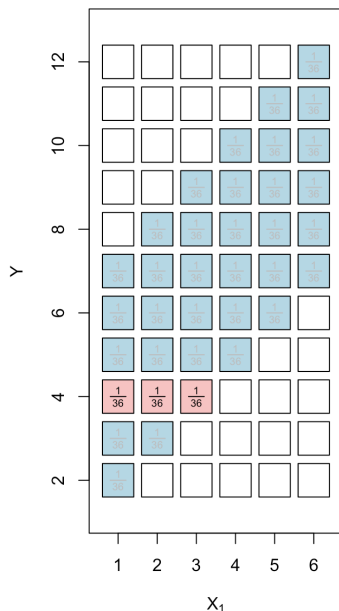
Conditional does not mean causal !

Marginal Events

$$\mathbb{P}(Y = y) = \sum_{x_1} \mathbb{P}(X_1 = x_1, Y = y)$$

$$= \sum_{x_1} \mathbb{P}(Y = y | X_1 = x_1) \cdot \mathbb{P}(X_1 = x_1)$$

$$\text{e.g. } \mathbb{P}(Y = 4) = \frac{3}{36} = \frac{1}{12}$$



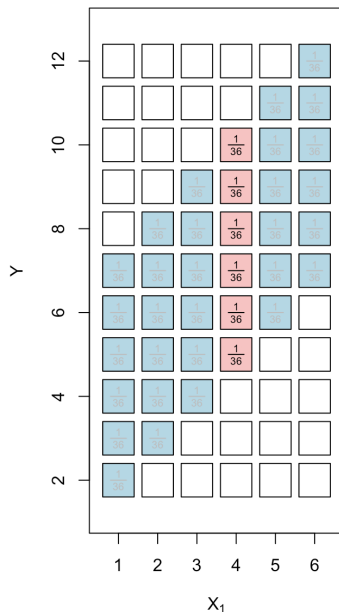
Marginal Events

$$\mathbb{P}(X_1 = x_1) = \sum_y \mathbb{P}(X_1 = x_1, Y = y)$$

$$= \sum_y \mathbb{P}(X_1 = x_1 | Y = y) \cdot \mathbb{P}(Y = y)$$

$$\text{e.g. } \mathbb{P}(X_1 = 4) = \frac{6}{36} = \frac{1}{6}$$

(no surprise here...)



Trees & Sets

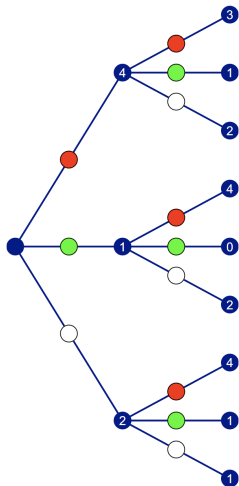
Consider an urn with 10 balls, 5 red ● 2 green ● and 3 white ○.

The **conditional probability** of event A occurring given that event B occurred is defined as

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \text{ and } B]}{\mathbb{P}[B]}$$

or $\mathbb{P}[A \text{ and } B] = \mathbb{P}[A|B] \cdot \mathbb{P}[B]$
(so called **chain rule**)

We draw two balls, without replacement
what is the probability to have (at least) one green ?



Trees & Sets

We draw two balls, without replacement
what is the probability to have (at least) one green
?

$$p = \mathbb{P}[X_1 = \bullet \text{ or } X_2 = \bullet]$$

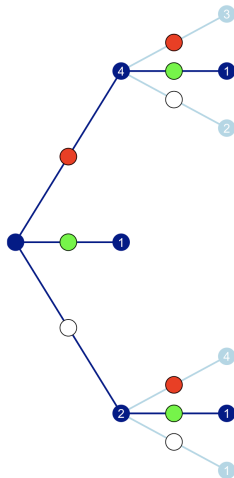
$$p = \mathbb{P}[X_1 = \bullet] + \mathbb{P}[X_2 = \bullet \text{ and } X_1 \neq \bullet]$$

$$\begin{aligned} \rho &= \mathbb{P}[X_1 = \bullet] &+& \mathbb{P}[X_2 = \bullet | X_1 = \bullet] \cdot \mathbb{P}[X_1 = \bullet] \\ &&+& \mathbb{P}[X_2 = \bullet | X_1 = \circ] \cdot \mathbb{P}[X_1 = \circ] \end{aligned}$$

$$p = \frac{1}{10} + \frac{1}{9} \cdot \frac{5}{10} + \frac{1}{9} \cdot \frac{3}{10} = \dots = \frac{17}{45}$$

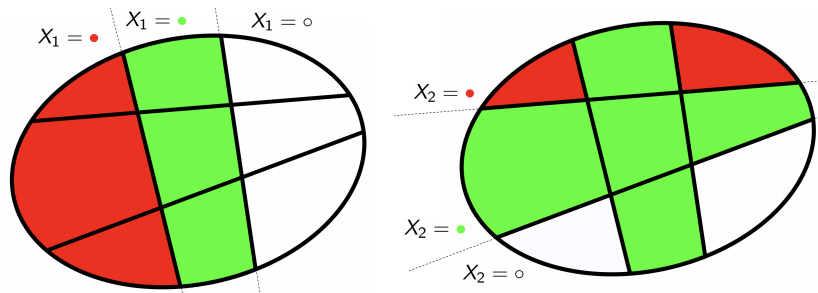
or more simple

$$p = 1 - \mathbb{P}[X_1 \in \{\bullet, \circ\} \text{ and } X_2 \in \{\bullet, \circ\}] = 1 - \frac{8}{10} \cdot \frac{7}{9} = \frac{17}{45}$$



Trees & Sets

One can also consider sets,



$$\mathbb{P}[A \text{ and } B] = \mathbb{P}[A \cap B] \text{ and } \mathbb{P}[A \text{ or } B] = \mathbb{P}[A \cup B]$$

Bayes Rule

$$\mathbb{P}(X_1 = x_1 | Y = y) = \frac{\mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(Y = y | X_1 = x_1)}{\mathbb{P}(Y = y)}$$

$$\mathbb{P}(X_1 = x_1 | Y = y) = \frac{\mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(Y = y | X_1 = x_1)}{\sum_x \mathbb{P}(X_1 = x) \cdot \mathbb{P}(Y = y | X_1 = x)}$$

Monty Hall

$$\begin{aligned} & \mathbb{P}(\text{treasure behind the other door}) \\ &= \mathbb{P}(\text{treasure behind the other door} | \text{X was correct}) \cdot \mathbb{P}(\text{X was correct}) \\ &+ \mathbb{P}(\text{treasure behind the other door} | \text{X was wrong}) \cdot \mathbb{P}(\text{X was wrong}) \\ &= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3} \end{aligned}$$

so yes, we should switch...

Exercise

The probability that a woman has breast cancer is 1%

If a woman has breast cancer, probability to test positive is 90%

If a woman does not have breast cancer, the probability that she nevertheless tests positive is 9%

A 50-year-old woman, no symptoms, participates in routine mammography screening. She tests positive, is alarmed, and wants to know from you whether she has breast cancer for certain or what the chances are. Apart from the screening results, you know nothing else about this woman. How many women who test positive actually have breast cancer? What is the best answer?

- A) nine in 10
- B) eight in 10
- C) one in 10
- D) one in 100

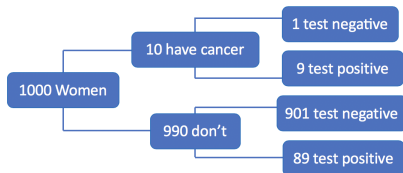


Exercise & Probability Trees

The probability that a woman has breast cancer is 1%

If a woman has breast cancer, probability to test positive is 90%

If a woman does not have breast cancer, the probability that she nevertheless tests positive is 9%



$$\mathbb{P}[\text{have cancer}|\text{test positive}] = \frac{9}{9 + 89} \simeq \frac{1}{10}$$

See [Do doctors understand test results?](#) half the group of 160 gynaecologists responded that the woman's chance of having cancer was nine in 10

Independence

Definition: Two events A and B are independent if the probability of B occurring is the same whether or not A occurs.

Example: $A = \{ \text{first coin is heads} \}$ and $B = \{ \text{second coin is heads} \}$

Formally, $\mathbb{P}[B|A] = \mathbb{P}[B]$ or $\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B]$

Quizz: with two dices, $A = \{ \text{first dice is 6} \}$ and $B = \{ \text{sum} > 6 \}$

Quizz: with two cards (deck of 52),

$A = \{ \text{first is heart} \}$ and $B = \{ \text{second is club} \}$

Quizz: with two cards (deck of 52),

$A = \{ \text{first is heart} \}$ and $B = \{ \text{second is 10} \}$

Quizz: $A = \{ \text{first child boy} \}$ and $B = \{ \text{second child boy} \}$

Independence

Definition: Two events A and B are independent if the probability of B occurring is the same whether or not A occurs.

Example: $A = \{ \text{first coin is heads} \}$ and $B = \{ \text{second coin is heads} \}$

Formally, $\mathbb{P}[B|A] = \mathbb{P}[B]$ or $\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B]$

Quizz: with two dices, $A = \{ \text{first dice is 6} \}$ and $B = \{ \text{sum} > 6 \}$

Not independent ($A \subset B$)

Quizz: with two cards (deck of 52),

$A = \{ \text{first is heart} \}$ and $B = \{ \text{second is club} \}$

Not independent $\mathbb{P}[A \cap B] = \frac{1}{4} \cdot \frac{13}{51} > \frac{1}{4} \cdot \frac{13}{52} = \mathbb{P}[A] \cdot \mathbb{P}[B]$

Quizz: with two cards (deck of 52),

$A = \{ \text{first is heart} \}$ and $B = \{ \text{second is 10} \}$

Independent $\mathbb{P}[A \cap B] = \frac{12}{52} \cdot \frac{4}{51} + \frac{1}{52} \cdot \frac{3}{51} = \frac{1}{52} = \mathbb{P}[A] \cdot \mathbb{P}[B]$

Quizz: $A = \{ \text{first child boy} \}$ and $B = \{ \text{second child boy} \}$