Introduction to data science & artificial intelligence (INF7100)

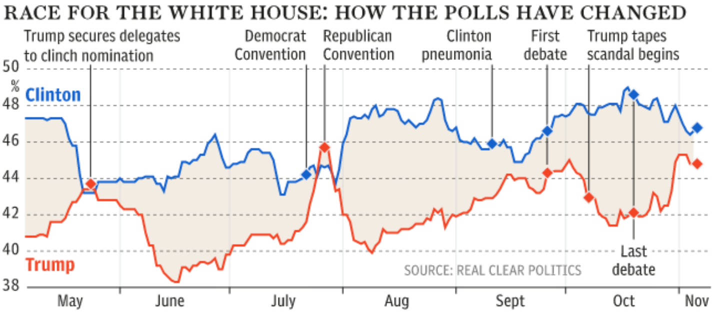
Arthur Charpentier

#131 Uncertainty and Randomness

été 2020

Uncertainty

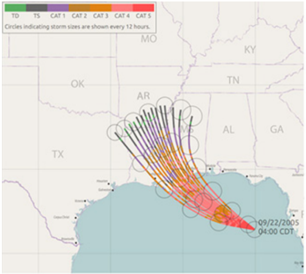
U.S. Elections

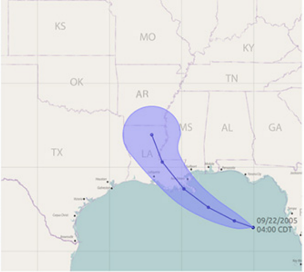


But the key thing to understand is that data science is a tool that is not necessarily going to give you answers, but probabilities (in How Data Failed Us in Calling an Election)

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,

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

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

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)

$$A = \{\text{some pair of them will have the same birthday}\}$$

$$A' = \{\text{no pair of them will have the same birthday}\}$$

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

Poisson approximation, $\lambda = \dfrac{1}{365}\dbinom{23}{2} = \dfrac{253}{365} \simeq -0.6932$ so

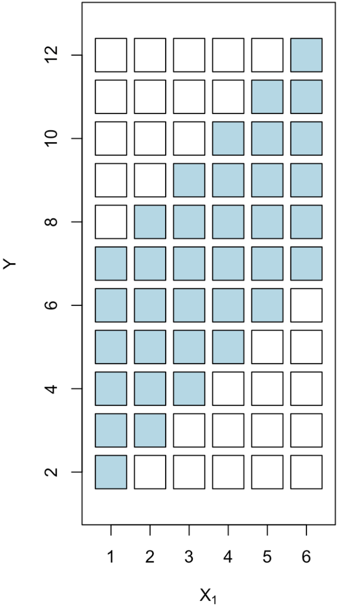
$$\mathbb{P}(X > 0) = 1 - \mathbb{P}(X = 0) \simeq 1 - e^{-0.6932} \simeq 0.500002$$

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 $(\in \{1, 2, \cdots, 6\})$
- y $(\in \{2, 3, \cdots, 12\})$

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

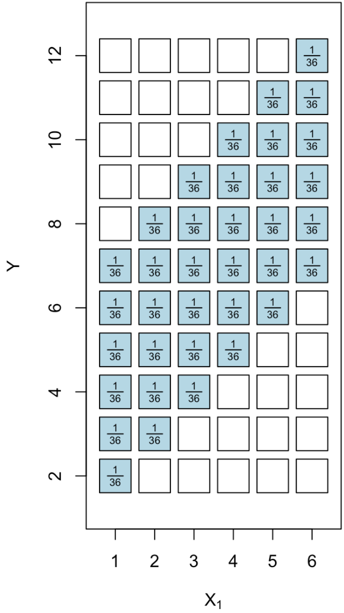


Conditional Events

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

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

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



Conditional Events

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

and similarly

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

We can write

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

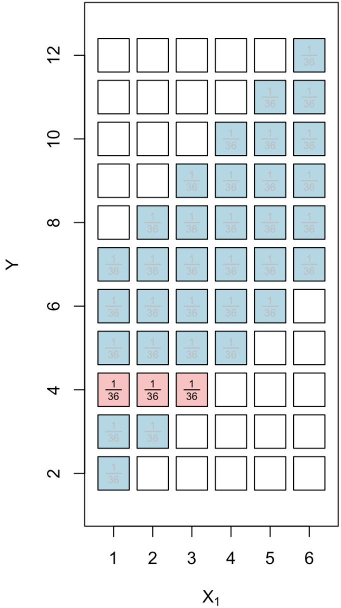
Conditional does not mean causal !

Marginal Events

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

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

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



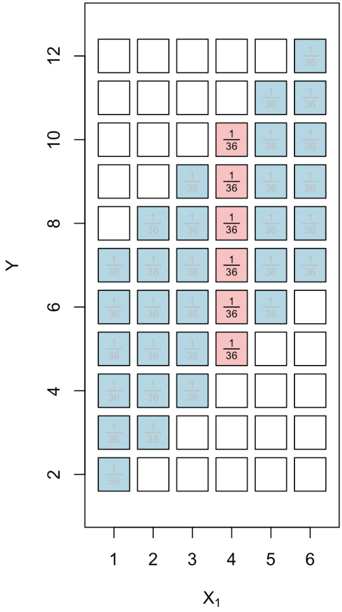
Marginal Events

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

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

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

(no surprise here...)



Trees & Sets

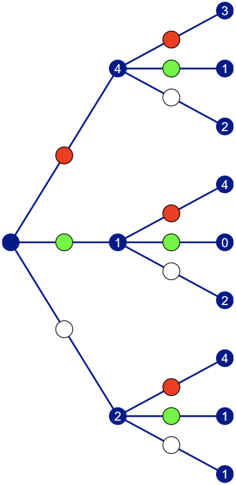
Consider an urn with 10 balls, 5 red ● 2 green ● and 3 white ○.

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
?

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

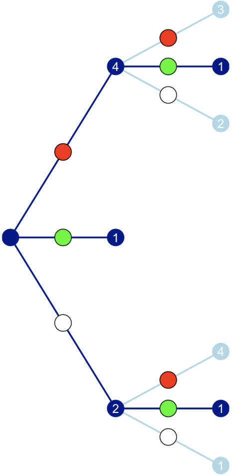
$$p=\mathbb{P}\big[X_1={\color{green}\bullet}\big]+\mathbb{P}\big[X_2={\color{green}\bullet}\text{ and }X_1\neq{\color{green}\bullet}\big]$$

$$\begin{aligned}p=\mathbb{P}\big[X_1={\color{green}\bullet}\big]\quad&+\quad\mathbb{P}\big[X_2={\color{green}\bullet}|X_1={\color{red}\bullet}\big]\cdot\mathbb{P}\big[|X_1={\color{red}\bullet}\big]\\&+\quad\mathbb{P}\big[X_2={\color{green}\bullet}|X_1=\circ\big]\cdot\mathbb{P}\big[|X_1=\circ\big]\end{aligned}$$

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

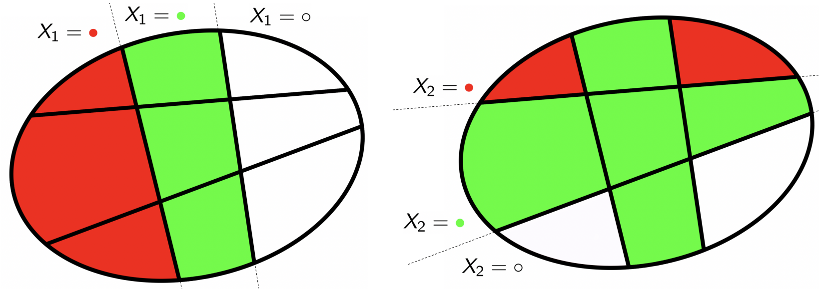
or more simple

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



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

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

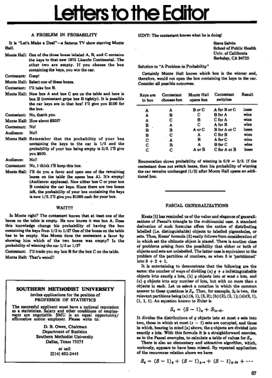
$$\mathbb{P}\big(X_1 = x_1 \big| Y = y\big) = \frac{\mathbb{P}\big(X_1 = x_1\big) \cdot \mathbb{P}\big(Y = y \big| X_1 = x_1\big)}{\sum_x \mathbb{P}\big(X_1 = x\big) \cdot \mathbb{P}\big(Y = y \big| X_1 = x\big)}$$

Monty Hall

Three doors: one has a treasure chest behind it and the other two have goats. You pick a door and indicate it to Monty. He opens one of the other two doors to reveal a goat. Now, should you stick to your initial choice, or switch to the other unopened door?

see wikipedia

X



Letters to the Editor

A PROBLEM IN PROBABILITY

It is "Let's Make a Deal"—a famous TV show starring Monte Hall.

Monte Hall: One of the three boxes labeled A, B, and C contains the keys to that new 1975 Lincoln Continental. The other two are empty. If you choose the box containing the keys, you win the car.

Contestant: Gasp!

Monte Hall: Select one of these boxes.

Contestant: I'll take box B.

Monte Hall: Now box A and box C are on the table and here is box B (contestant grips box B tightly). It is possible the car keys are in that box! I'll give you $100 for the box.

Contestant: No, thank you.

Monte Hall: How about $200?

Contestant: No!

Audience: No!!

Monte Hall: Remember that the probability of your box containing the keys to the car is 1/3 and the probability of your box being empty is 2/3. I'll give you $500.

Audience: No!!

Contestant: No, I think I'll keep this box.

Monte Hall: I'll do you a favor and open one of the remaining boxes on the table (he opens box A). It's empty! (Audience: applause). Now either box C or your box B contains the car keys. Since there are two boxes left, the probability of your box containing the keys is now 1/2. I'll give you $1000 cash for your box.

WAIT!!!!

Is Monte right? The contestant knows that at least one of the boxes on the table is empty. He now knows it was box A. Does this knowledge change his probability of having the box containing the keys from 1/3 to 1/2? One of the boxes on the table has to be empty. Has Monte done the contestant a favor by showing him which of the two boxes was empty? Is the probability of winning the car 1/2 or 1/3?

Contestant: I'll trade you my box B for the box C on the table.

Monte Hall: That's weird!!

HINT: The contestant knows what he is doing!

Steve Selvin
School of Public Health
Univ. of California
Berkeley, CA 94720

Solution to "A Problem in Probability"

Certainly Monte Hall knows which box is the winner and, therefore, would not open the box containing the keys to the car. Consider all possible outcomes:

Keys are in box	Contestant chooses box	Monte Hall opens box	Contestant switches	Result
A	A	B or C	A for B or C	loses
A	B	C	B for A	wins
A	C	B	C for A	wins
B	A	C	A for B	wins
B	B	A or C	B for A or C	loses
B	C	A	C for B	wins
C	A	B	A for C	wins
C	B	A	B for C	wins
C	C	A or B	C for A or B	loses

Enumeration shows probability of winning is 6/9 = 2/3. If the contestant does not switch boxes, then his probability of winning the car remains unchanged (1/3) after Monte Hall opens an additional box.

PASCAL GENERALIZATIONS

Keats [1] has reminded us of the value and elegance of generalizations of Pascal's triangle to the multinomial case. A standard derivation of such formulae offers the notion of distributing labelled (i.e. distinguishable) objects to labelled pigeonholes, or sets. Thus, Keats' formula (3) easily follows from consideration of in which set the ultimate object is placed. There is another class of problems arising from the possibility that either or both of objects and sets are unlabelled. The latter case is equivalent to the problem of the partition of numbers, as when 5 is 'partitioned' into 2 + 2 + 1.

It is entertaining to demonstrate that the following are the same: the number of ways of dividing (a) $q + s$ indistinguishable objects into *exactly* s lots, (b) q objects into *at most* s lots, and (c) q objects into any number of lots, but with no more than s objects in each. Let us select a notation in which the common answer to these questions is S_q. Then, for example, 2_3 is two, the relevant partitions being (a):(4, 1), (3, 2); (b):(3), (2, 1); (c):(2, 1), (1, 1, 1). An equation known to Euler is

$$S_q = (S-1)_q + S_{(q-s)}.$$

It divides the distributions of q objects into at most s sets into two, those in which at most $(s-1)$ sets are occupied, and those in which, bearing in mind (a) above, the q objects are divided into exactly s lots. With this formula it is a straightforward exercise, as in the Pascal examples, to calculate a table of values for S_q.

There is also an elementary and attractive algorithm, which, curiously, appears to have been missed. By repeated application of the recurrence relation above we have

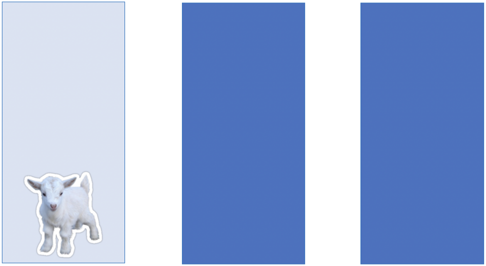
$$S_q = (S-1)_q + (S-1)_{q-s} + (S-1)_{q-2s} + \cdots$$

67

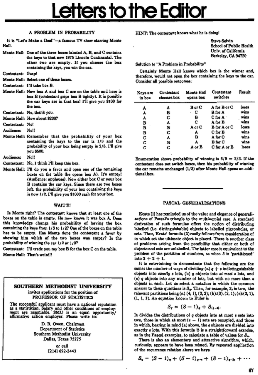
Monty Hall

Three doors: one has a treasure chest behind it and the other two have goats. You pick a door and indicate it to Monty. He opens one of the other two doors to reveal a goat. Now, should you stick to your initial choice, or switch to the other unopened door?

see wikipedia



X



Letters to the Editor

A PROBLEM IN PROBABILITY

It is "Let's Make a Deal"—a famous TV show starring Monte Hall.

Monte Hall: One of the three boxes labeled A, B, and C contains the keys to that new 1975 Lincoln Continental. The other two are empty. If you choose the box containing the keys, you win the car.

Contestant: Gasp!

Monte Hall: Select one of these boxes.

Contestant: I'll take box B.

Monte Hall: Now box A and box C are on the table and here is box B (contestant grips box B tightly). It is possible the car keys are in that box! I'll give you $100 for the box.

Contestant: No, thank you.

Monte Hall: How about $200?

Contestant: No!

Audience: No!!

Monte Hall: Remember that the probability of your box containing the keys to the car is 1/3 and the probability of your box being empty is 2/3. I'll give you $500.

Audience: No!!

Contestant: No, I think I'll keep this box.

Monte Hall: I'll do you a favor and open one of the remaining boxes on the table (he opens box A). It's empty! (Audience: applause). Now either box C or your box B contains the car keys. Since there are two boxes left, the probability of your box containing the keys is now 1/2. I'll give you $1000 cash for your box.

WAIT!!!!

Is Monte right? The contestant knows that at least one of the boxes on the table is empty. He now knows it was box A. Does this knowledge change his probability of having the box containing the keys from 1/3 to 1/2? One of the boxes on the table has to be empty. Has Monte done the contestant a favor by showing him which of the two boxes was empty? Is the probability of winning the car 1/2 or 1/3?

Contestant: I'll trade you my box B for the box C on the table.

Monte Hall: That's weird!!

HINT: The contestant knows what he is doing!

Steve Selvin
School of Public Health
Univ. of California
Berkeley, CA 94720

Solution to "A Problem in Probability"

Certainly Monte Hall knows which box is the winner and, therefore, would not open the box containing the keys to the car. Consider all possible outcomes:

Keys are in box	Contestant chooses box	Monte Hall opens box	Contestant switches	Result
A	A	B or C	A for B or C	loses
A	B	C	B for A	wins
A	C	B	C for A	wins
B	A	C	A for B	wins
B	B	A or C	B for A or C	loses
B	C	A	C for B	wins
C	A	B	A for C	wins
C	B	A	B for C	wins
C	C	A or B	C for A or B	loses

Enumeration shows probability of winning is 6/9 = 2/3. If the contestant does not switch boxes, then his probability of winning the car remains unchanged (1/3) after Monte Hall opens an additional box.

SOUTHERN METHODIST UNIVERSITY
invites applications for the position of
PROFESSOR OF STATISTICS

The successful applicant must have a national reputation as a statistician. Salary and other conditions of employment are negotiable. SMU is an equal opportunity/affirmative action employer. Please write to:

D. B. Owen, Chairman
Department of Statistics
Southern Methodist University
Dallas, Texas 75275
or call
[214] 692-2443

PASCAL GENERALIZATIONS

Keats [1] has reminded us of the value and elegance of generalizations of Pascal's triangle to the multinomial case. A standard derivation of such formulae offers the notion of distributing labelled (i.e. distinguishable) objects to labelled pigeonholes, or sets. Thus, Keats' formula (3) easily follows from consideration of in which set the ultimate object is placed. There is another class of problems arising from the possibility that either or both of objects and sets are unlabelled. The latter case is equivalent to the problem of the partition of numbers, as when 5 is 'partitioned' into 2 + 2 + 1.

It is entertaining to demonstrate that the following are the same: the number of ways of dividing (a) $q + s$ indistinguishable objects into *exactly* s lots, (b) q objects into *at most* s lots, and (c) q objects into any number of lots, but with no more than s objects in each. Let us select a notation in which the common answer to these questions is S_q. Then, for example, 2_3 is two, the relevant partitions being (a):(4, 1), (3, 2); (b):(3), (2, 1); (c):(2, 1), (1, 1, 1). An equation known to Euler is

$$S_q = (S-1)_q + S_{(q-s)}.$$

It divides the distributions of q objects into at most s sets into two, those in which at most $(s-1)$ sets are occupied, and those in which, bearing in mind (a) above, the q objects are divided into exactly s lots. With this formula it is a straightforward exercise, as in the Pascal examples, to calculate a table of values for S_q.

There is also an elementary and attractive algorithm, which, curiously, appears to have been missed. By repeated application of the recurrence relation above we have

$$S_q = (S-1)_q + (S-1)_{q-s} + (S-1)_{q-2s} + \cdots$$

67

Monty Hall

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

so yes, we should switch...

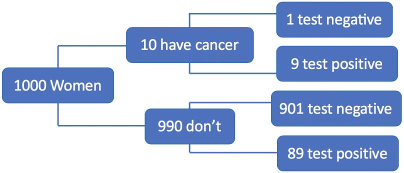
Exercise

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: $A = \{$ first coin is heads $\}$ and $B = \{$ 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, $A = \{$ first dice is 6 $\}$ and $B = \{$ sum > 6 $\}$

Quizz: with two cards (deck of 52),
$A = \{$ first is heart $\}$ and $B = \{$ second is club $\}$

Quizz: with two cards (deck of 52),
$A = \{$ first is heart $\}$ and $B = \{$ second is 10 $\}$

Quizz: $A = \{$ first child boy$\}$ and $B = \{$ 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: $A = \{$ first coin is heads $\}$ and $B = \{$ 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, $A = \{$ first dice is 6 $\}$ and $B = \{$ sum > 6 $\}$

Not independent ($A \subset B$)

Quizz: with two cards (deck of 52),

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

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

Quizz: with two cards (deck of 52),

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

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

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