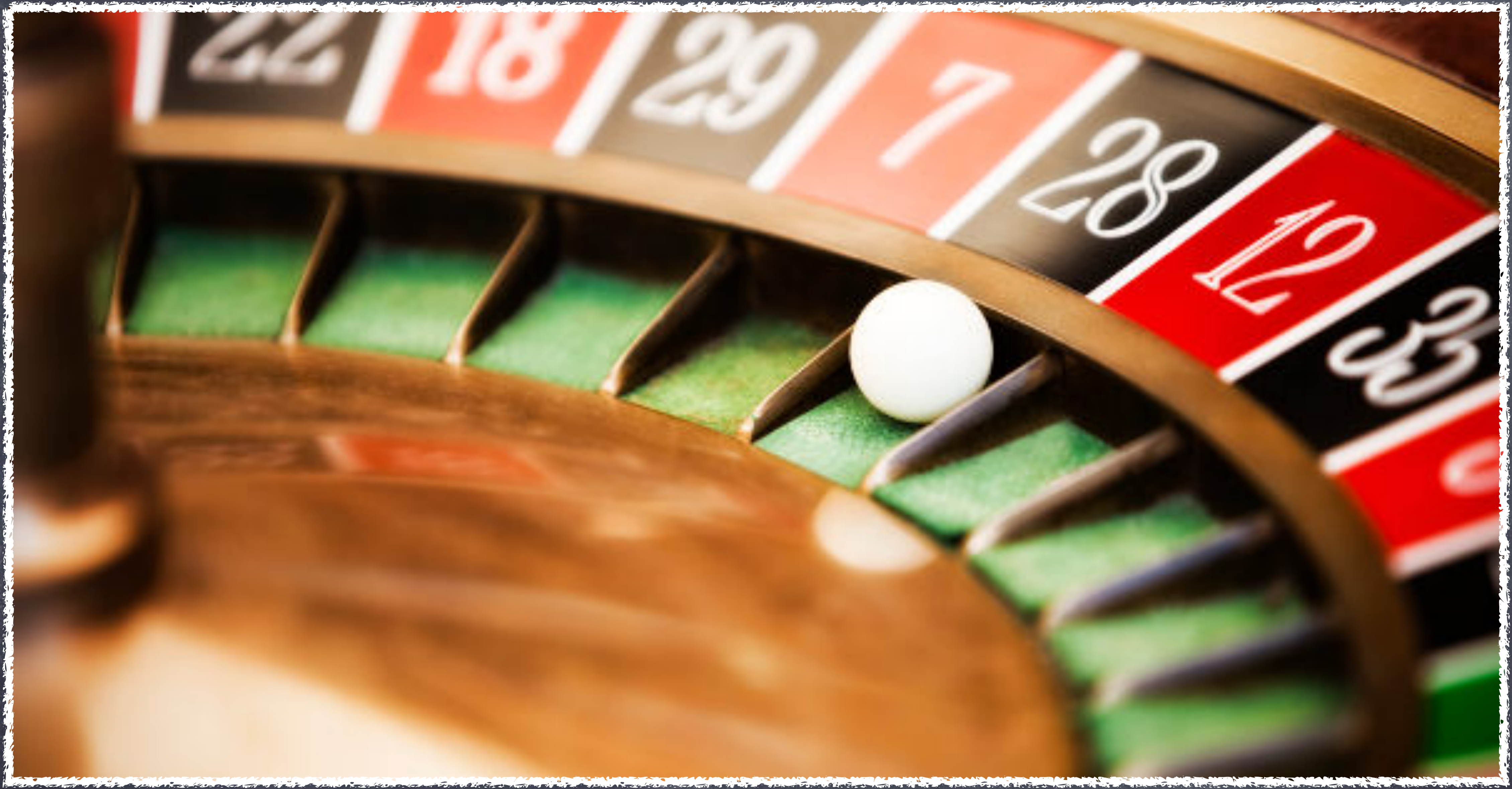
DATA ANALYSIS

Christian Capone
Kazakh-British Technical University
Almaty, Kazakhstan



Probability

chance analysis

Probability

- We need some probability notions (but not too much!) to assess noise in samples and to solve interesting problems in simple fashion
  - Flip a fair coin. What's the probability of a head?
  - Roll a fair 6-sided die. You win the game (on the first roll) if the sum of the uptakes is 7 to 11. What's the probability you win?
  - Roll a pair of fair 6-sided dice. You win the game (on the first roll) if the sum of the uptakes is 7 or 11. What's the probability you win?

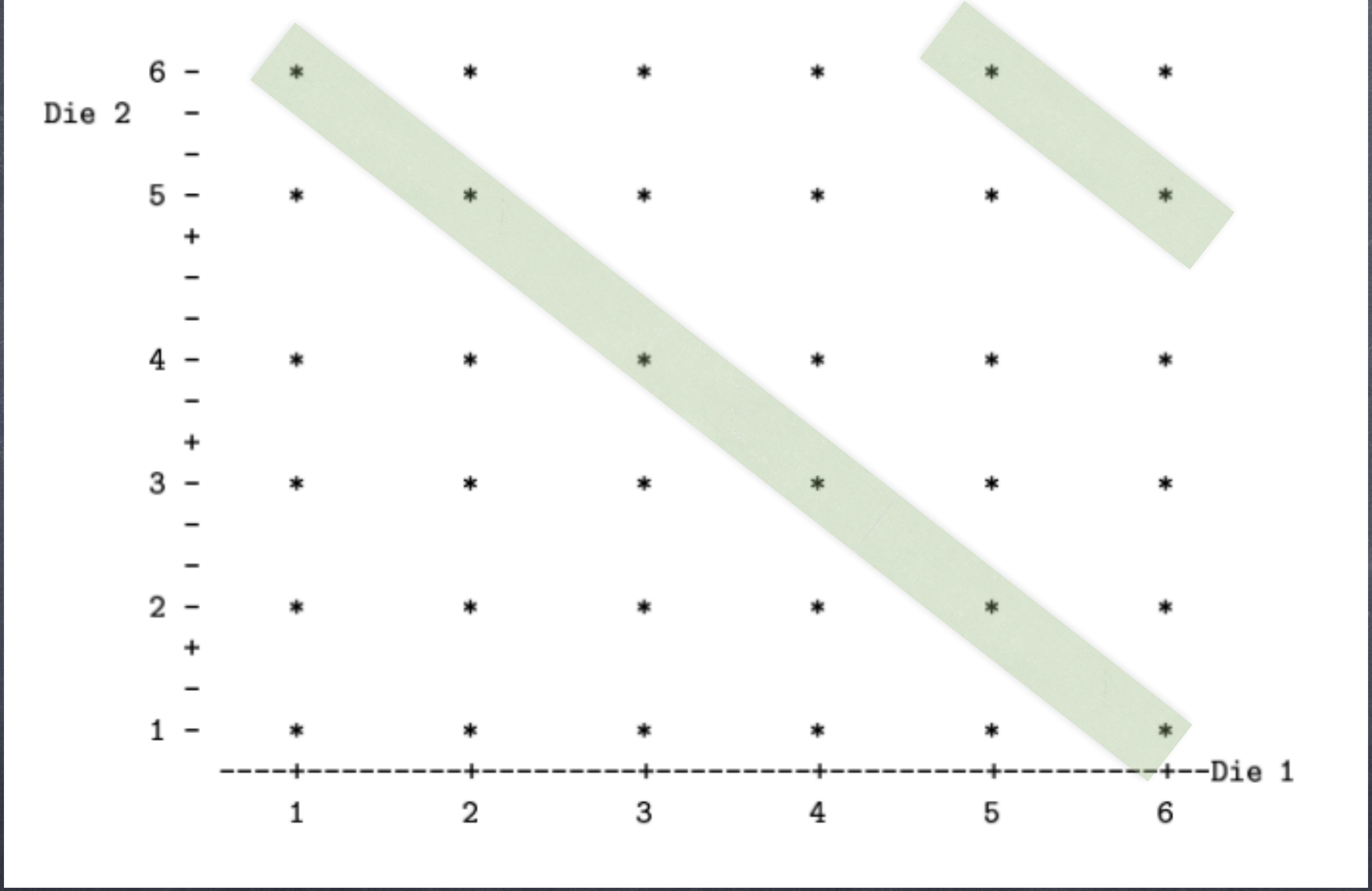
Nomenclature

- An experiment result in an outcome
- The collection of all the outcomes is the sample space. We denote the sample space with the letter "S"
  - Flip a coin: S={H,T}
  - Roll a six sided die: S={1,2,3,4,5,6}
  - Roll a pair of six sided dice: S={(1,1),(1,2),(1,3), ..., (6,6)}
- An event is a subset of S, denoted by A, B, C, ...
  - Flip a coin: A={H}; Comp(A) = {T}
  - Roll a six sided die: B={1,2}
- Exercise: list the sample space, list the event of interest and its complement for this experiment: roll a pair of 6-sided dice; we are interested in the event that both dice have the same outcome

What is a Probability

- Probability is the measure of the likelihood that an event will occur
- The probability of an event is a number between 0 and 1
- The probability of the sample space is 1
- If two events cannot occur in the same time, the probability of one or the other occurs is the sum of the probabilities of the individual events:
  - $P(A+B) = P(A) + P(B)$
- Suppose to throw twice a fair die. Let A be the event A = {1,2}. Calculate P(A)

More on Probability



- What is the probability that the sum of uptakes is 7 or 11 if we have equilikely dice?

Exercise

- Six cards with the numbers 1 through 6 on them are well shuffled and two cards are taken, without replacement. Find the probability that the sum of the numbers on the two cards is 7. Note that the order is not important (i.e. 1,2 = 2,1)
  - Calculate the sample space S
  - Calculate the probability of the event

Relative Frequency

- Suppose we want to determine the probability of some event A
- Suppose we can repeat the experiment over and over again, such that the trials are:
  - Independent of one another
  - Identical, in that conditions do not change from one trial to another
- Let "N" be the number of trials and #(A) the number of times A occurred, then P(A) is approximately #(A)/N and the approximation is better with N larger

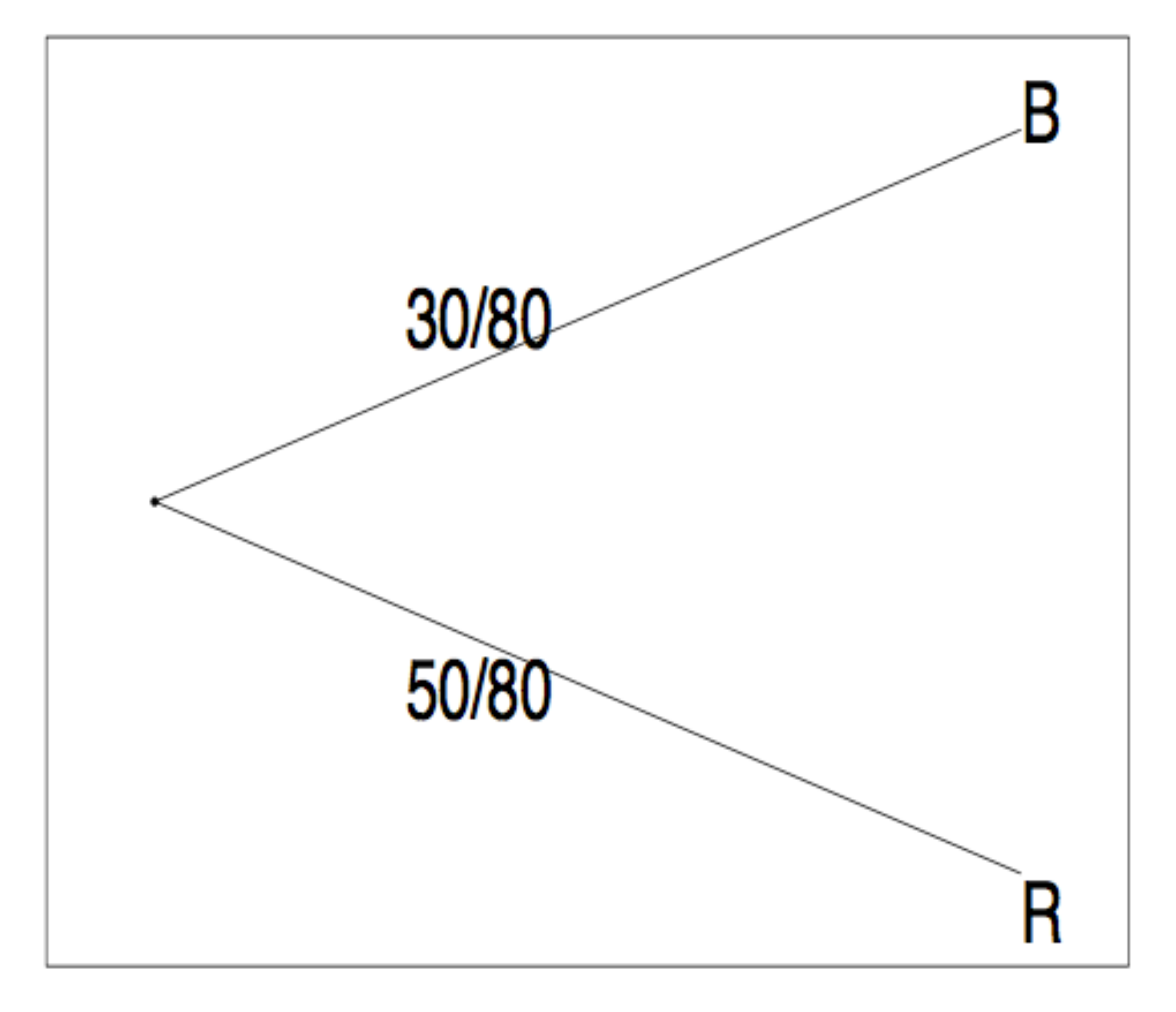
Determination of Probabilities

- ENUMERATION: listing of the sample space
- TREE DIAGRAMS: allow us to see all the possible outcomes of an even and calculate their probability
- RESAMPLING: a method that consists of drawing repeated samples from the original data samples

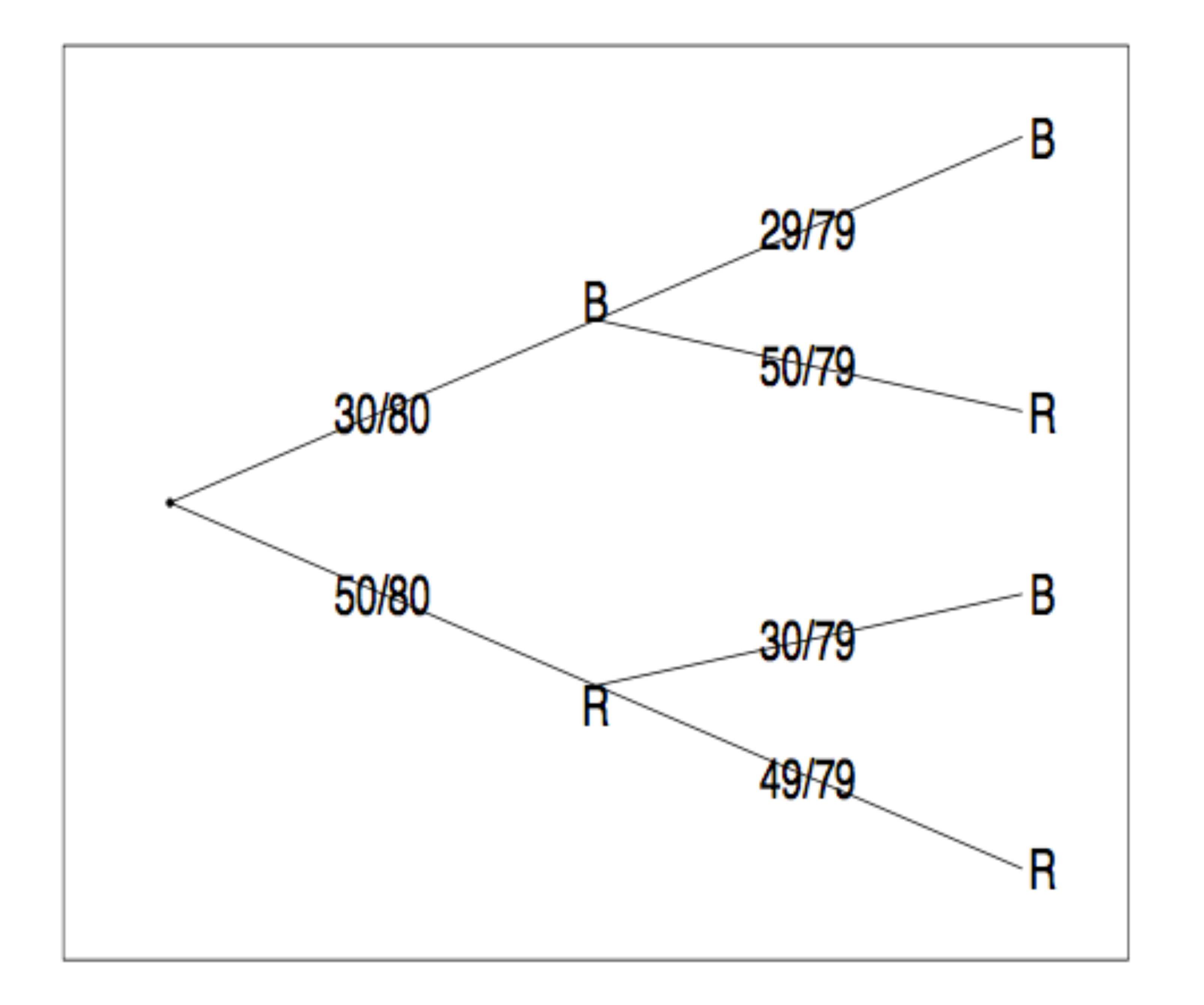
Tree Diagrams

- URN PROBLEM: suppose we have an urn with 30 blue balls and 50 red balls in it and that these balls are identical except for the color. Suppose further the balls are well mixed and that we draw 3 balls, without replacement. Determine the probability that the balls are all of the same color
- Even for this simple problem there are 82160 elements in the sample space
- Also note that this problem is sequential
- In such case a simple tree diagram can help to solve it

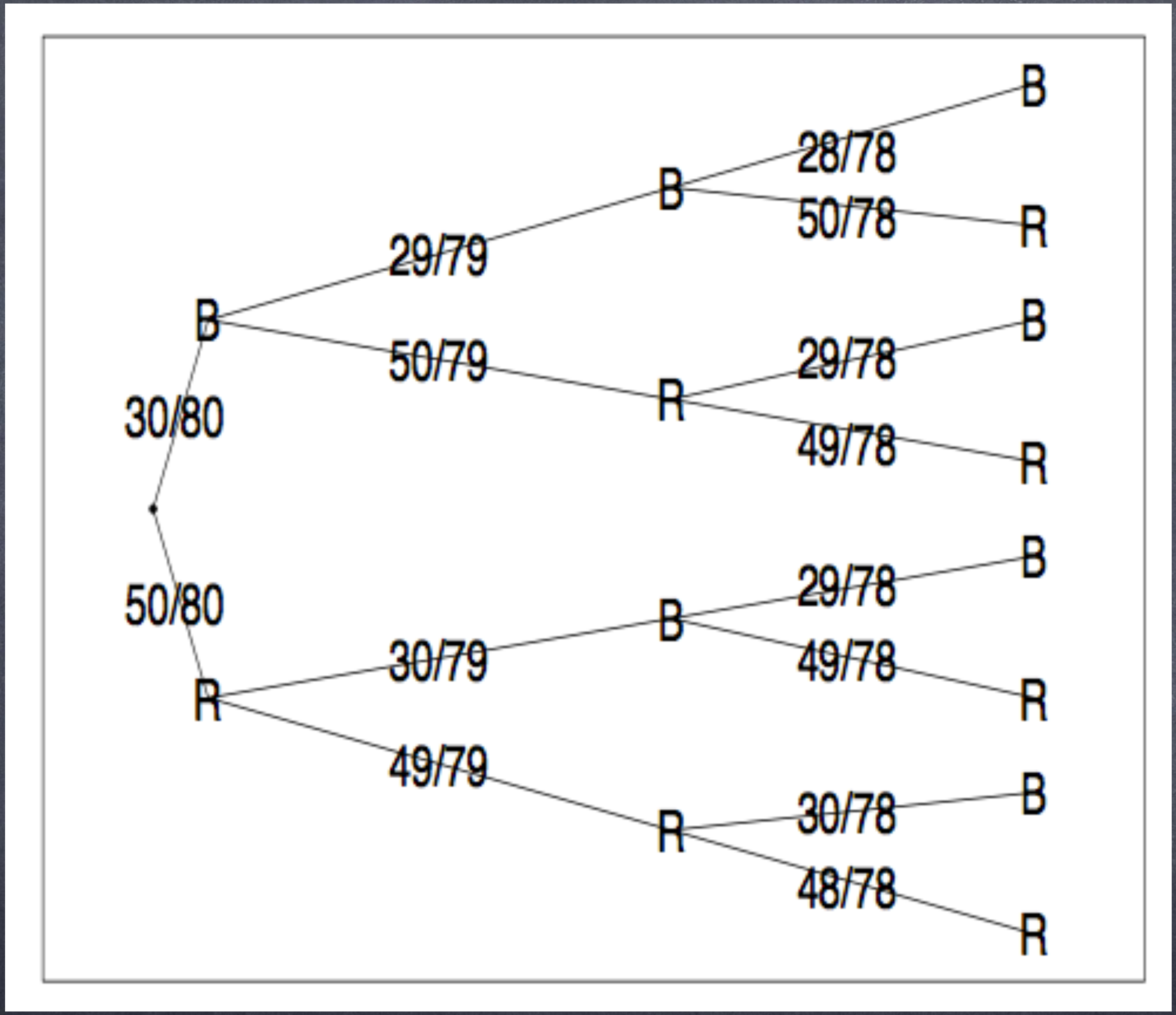
Tree Diagram: first draw



Tree Diagram: second draw



Tree Diagram: third draw



Solution: 28.8%

Exercises

- In the urn example, find the probability of getting
  - 2 red balls and one blue ball
  - the first two balls red
  - 4 balls of the same color if we draw one more time
- Six cards with the number 1 through 6 on them are well chuffed and two cards are taken without replacement. Use a tree diagram to determine the probability of that the sum of the numbers on the two cards is 7

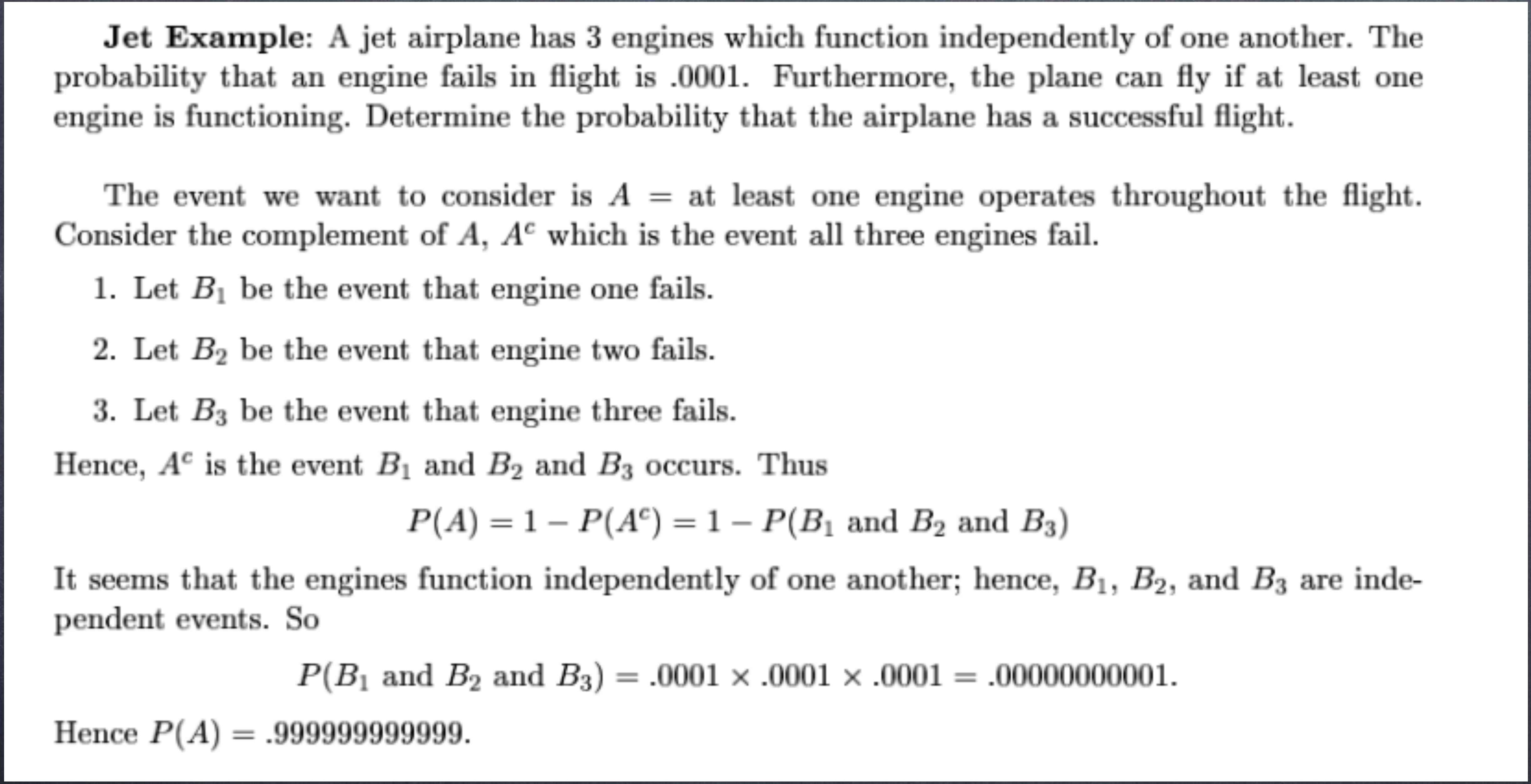
Independence

- If we recall the urn problem and
  - let C_3 denote the event that the third ball is blue and
  - let B_2 denote the event that the second ball is blue and
  - let A_1 denote the event that the first ball is blue
- $P(B_2)$ depends on A_1 and we represent this by $P(B_2|A_1) = 29/79$ and we read it as $P(B_2)$ given A_1
- Given two events A and B, if $P(B|A)=P(B)$ then we say that A and B are independent events
- How much is $P(B_2)$ then? And $P(C_3)$? Are B_2 and A_1 independent events?
- What happens if we do sampling with replacement?

Conditional Probability

- Let A and B be arbitrary events. We want to determine P(B|A).
- Suppose that the tree diagram is too complicated and therefore we have to use relative frequency
- We repeat the experiment many time, say "N"
- Since we want the probability given A, among these "N", we select only the experiments in which A occurred and those are #(A)
- Among these, let's count the time B occurred. But this is #(A and B)
- So P(B|A) is approximately #(A and B) / #(A) or we can say
- P(B|A) = P(A and B) / P(A)
- P(A and B) = P(B|A)P(A) (moltiplicative law)
- P(A and B) = P(B)P(A) if events A and B are independent

Solved Exercise



Jet Example: A jet airplane has 3 engines which function independently of one another. The probability that an engine fails in flight is .0001. Furthermore, the plane can fly if at least one engine is functioning. Determine the probability that the airplane has a successful flight.

The event we want to consider is A = at least one engine operates throughout the flight. Consider the complement of A, A^c which is the event all three engines fail.

1. Let B_1 be the event that engine one fails.
2. Let B_2 be the event that engine two fails.
3. Let B_3 be the event that engine three fails.

Hence, A^c is the event B_1 and B_2 and B_3 occurs. Thus

$$P(A) = 1 - P(A^c) = 1 - P(B_1 \text{ and } B_2 \text{ and } B_3)$$

It seems that the engines function independently of one another; hence, B_1, B_2, and B_3 are independent events. So

$$P(B_1 \text{ and } B_2 \text{ and } B_3) = .0001 \times .0001 \times .0001 = .00000000001.$$

Hence $P(A) = .999999999999$.

Exercises

- Suppose to flip a fair coin 4 times. What is the probability of 4 heads?
- Suppose in the Jet airplane example, that one engine is broken before takeoff, but the plane takes off anyway. Determine the probability that the plane arrives safely

Probability in Complex Events

- Enumerating and tree diagrams are useful to identify probability for small problems
- Which is the probability of opening with a pair in 5 card poker? It is not possible to calculate with these methodologies
- We can still use math, but this require a deep knowledge of theory of probability e pretty good math skills
- So what can we do?

Resampling

- Using resampling we can estimate the probability of an event and..

- … we can increase the accuracy of the estimation by simply increasing the number of resamples

- Another advantage of resampling is that you need to build up a correct model to accomplish it and you can do that only if understand clearly the problem

The Four Steps of Resampling

- Let A be the event of interest

    1. Choose a model and define a trial: identify the sample space and the event
    2. Define the event of interest in term of step 1
    3. Obtain N trials of the experiment. Count the occurrence of the event A. Denote this count as #(A). Note that is extremely important that the trials are independent of one another and that the trial are performed under identical conditions
    4. Estimate P(A) as #(A)/N

- The error of this estimation is $2\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$

Exercise

- Try to solve the previous simple probability calculations
- Now we can calculate which is the probability of opening with a pair in 5 card poker…

Q&A