Probability Overview

-Sathiya Ramesh

What is probability?

Fundamentally related to the frequencies of repeated events - Frequentist

Fundamentally related to our own certainty and uncertainty of events - Bayesians.

<https://courses.edx.org/asset-v1:UCSanDiegoX+DSE210x+3T2017+type@asset+block@HistoryOfProbAndStat.pdf>

Event

An event is a set of outcomes of an experiment (a subset of the sample space) to which a probability is assigned.

Sample Space

The sample space , denoted *S*, is the collection of all possible outcomes of a random study.

Counting

(CLASSICAL DEFINITION OF PROBABILITY)

$$P(\text{Event}) = \frac{\#\ \text{Favorable Outcomes}}{\#\ \text{Possible Outcomes}}$$

What are our favorable outcomes?
What are the possible outcomes?

Motivation Problem

I need to choose a password for a computer account. The rule is that the password must consist of two lowercase letters (a to z) followed by one capital letter (A to Z) followed by four digits (0,1,⋯,9). For example, the following is a valid password

$$ejT3018$$

- Find the total number of possible passwords, N.
- A hacker has been able to write a program that randomly and independently generates 10^8 passwords according to the above rule. Note that the same password could be generated more than once. If one of the randomly chosen passwords matches my password, then he can access my account information. What is the probability that he is successful in accessing my account information?

Counting Techniques

- Ordered sampling with replacement
- Ordered sampling without replacement
- Unordered sampling without replacement
- Unordered sampling with replacement

Problem

A 4 digit PIN is selected. What is the probability that there are no repeated digits?

Problem

Ten passengers get on an airport shuttle at the airport. The shuttle has a route that includes 5 hotels, and each passenger gets off the shuttle at his/her hotel. The driver records how many passengers leave the shuttle at each hotel. How many different possibilities exist?

Summary

ordered sampling with replacement	n^k
ordered sampling without replacement	$P_k^n = \frac{n!}{(n-k)!}$
unordered sampling without replacement	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$
unordered sampling with replacement	$\binom{n+k-1}{k}$

Partitions

Definition. The **number of distinguishable permutations** of n objects, of which:

- n_1 are of one type
- n_2 are of a second type
- ... and ...
- n_k are of the last type

and $n = n_1 + n_2 + ... + n_k$ is given by:

$$\binom{n}{n_1 n_2 n_3 \ldots n_k} = \frac{n!}{n_1!n_2!n_3!\ldots n_k!}$$

Problem

Fifteen dogs are available to use in a study to compare three different diets. Each of the diets (let's say, A, B, C) is to be used on five randomly selected dogs.. In how many ways can the diets be assigned to the dogs?

Independent Events

The probability one event occurs in no way affects the probability of the other event occurring. Eg: Owning a dog and growing your own herb garden.

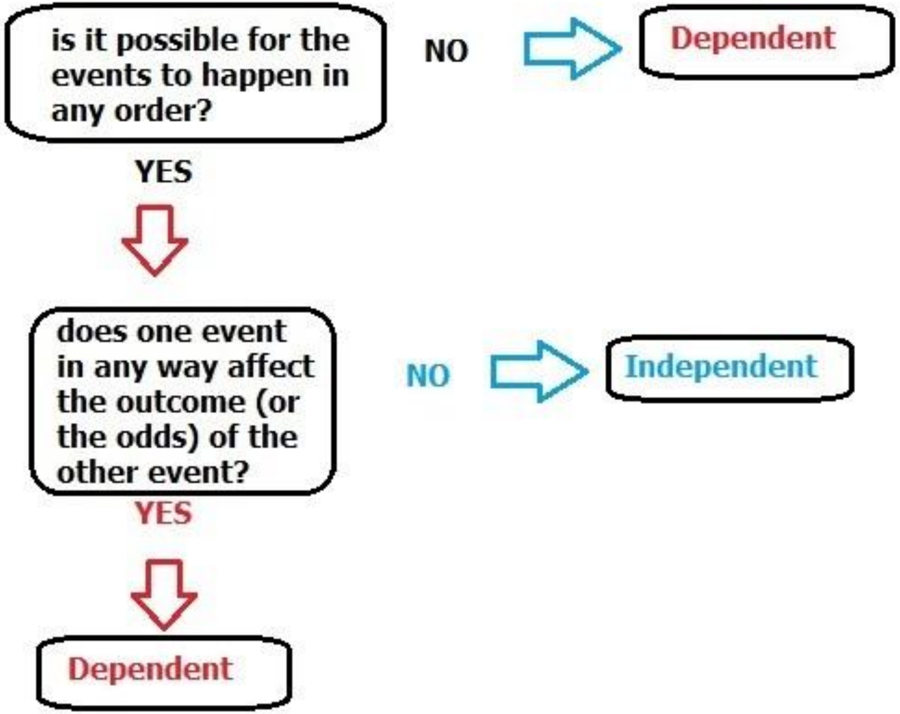
Dependent Events

When two events are said to be dependent, the probability of one event occurring influences the likelihood of the other event. Eg: Robbing a bank and going to jail.

Mutual Exclusivity

Two events are said to be mutually exclusive if they can't occur at the same time. For a given sample space, its either one or the other but not both. Eg: Head or Tail.

Dependent or Independent?



Joint, Marginal & Conditional Probabilities

Joint, Marginal and Conditional

- Joint probabilities for rain and wind:

	no wind	some wind	strong wind	storm
no rain	0.1	0.2	0.05	0.01
light rain	0.05	0.1	0.15	0.04
heavy rain	0.05	0.1	0.1	0.05

- Marginalize to get simple probabilities:
  - $P(\text{no wind}) = 0.1 + 0.05 + 0.05 = 0.2$
  - $P(\text{light rain}) = 0.05 + 0.1 + 0.15 + 0.04 = 0.34$
- Combine to get conditional probabilities:
  - $P(\text{no wind}|\text{light rain}) = \frac{0.05}{0.34} = 0.147$
  - $P(\text{light rain}|\text{no wind}) = \frac{0.05}{0.2} = 0.25$

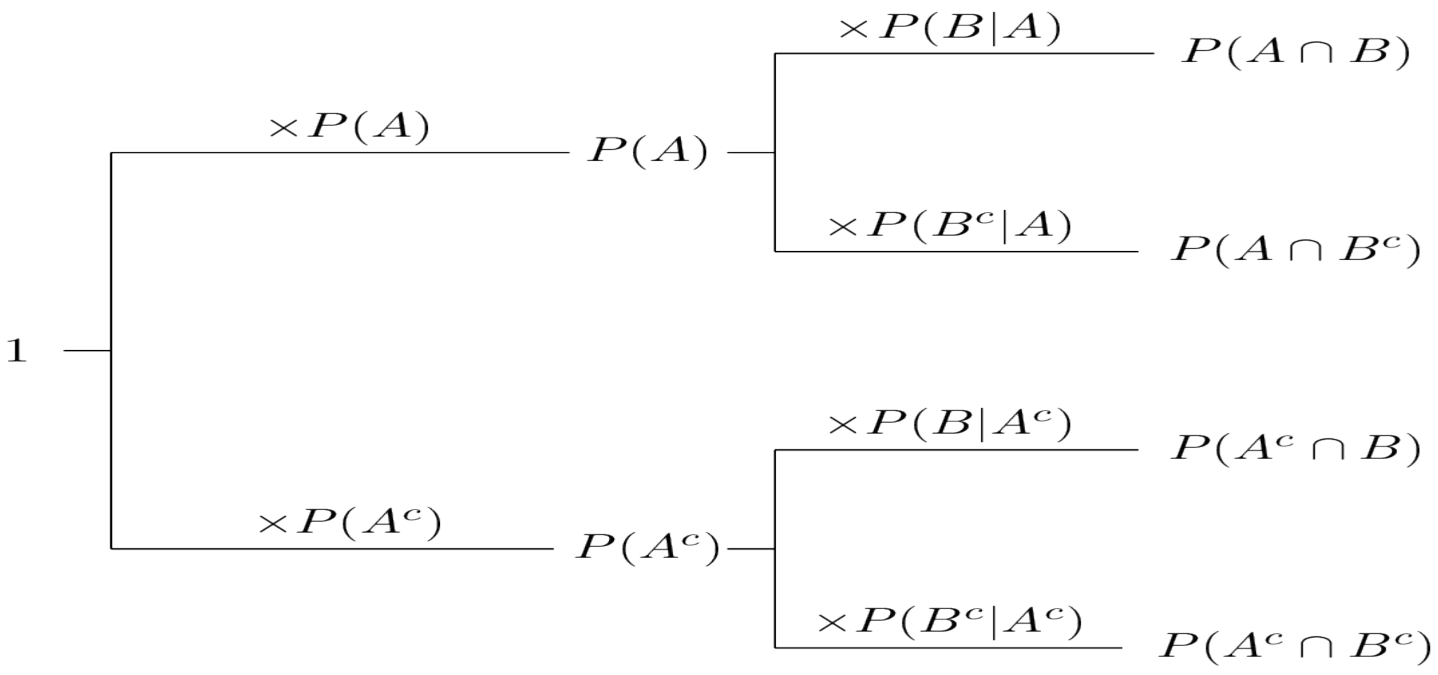
Problem

Consider a family that has two children. We are interested in the children's genders.

Our sample space is $S=\{(G,G),(G,B),(B,G),(B,B)\}$. Also assume that all four possible outcomes are equally likely.

1. What is the probability that both children are girls given that the first child is a girl?
2. We ask the father: "Do you have at least one daughter?" He responds "Yes!" Given this extra information, what is the probability that both children are girls? In other words, what is the probability that both children are girls given that we know at least one of them is a girl?

Chain Rule


1
$\times P(A)$
$P(A)$
$\times P(A^c)$
$P(A^c)$
$\times P(B|A)$
$P(A \cap B)$
$\times P(B^c|A)$
$P(A \cap B^c)$
$\times P(B|A^c)$
$P(A^c \cap B)$
$\times P(B^c|A^c)$
$P(A^c \cap B^c)$

Generalization

Chain rule for conditional probability:

$$P(A1 \cap A2 \cap \cdots \cap An) = P(A1)P(A2|A1)P(A3|A2,A1) \cdots P(An|An-1An-2 \cdots A1)$$

Problem

In a factory there are 100 units of a certain product, 55 of which are defective. We pick three units from the 100 units at random. What is the probability that none of them are defective?

Bayes Theorem

Likelihood
Probability of collecting this data when our hypothesis is true

Bill Howe, UW

Prior
The probability of the hypothesis being true before collecting data

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?

Problem

In a particular pain clinic, 10% of patients are prescribed narcotic pain killers. Overall, five percent of the clinic's patients are addicted to narcotics (including pain killers and illegal substances). Out of all the people prescribed pain pills, 8% are addicts. If a patient is an addict, what is the probability that they will be prescribed pain pills?

Problem

ASSUMPTIONS

- 100 out of 10,000 women aged forty who participate in a routine screening have breast cancer
- 80 of every 100 women with breast cancer will get positive tests
- 950 out of 9,900 women without breast cancer will also get positive tests

PROBLEM

If 10,000 women in this age group undergo a routine screening, about what fraction of women with positive tests will actually have breast cancer?

Prior Probability

$$P(A|B) = \frac{\boxed{P(B|A)\,P(A)}}{\boxed{P(B|A)P(A)} + \boxed{P(B|\neg A)P(\neg A)}}$$

Same formula as the prior probability, but this time for not-A.

Problem

A factory production line is manufacturing bolts using three machines, A, B and C. Of the total output, machine A is responsible for 25%, machine B for 35% and machine C for the rest. It is known from previous experience with the machines that 5% of the output from machine A is defective, 4% from machine B and 2% from machine C. A bolt is chosen at random from the production line and found to be defective. What is the probability that it came from (a) machine A

Extras

What is a random variable.

What is a distribution.

What are mostly used distributions.

What is PDF, PMF, CDF.

What is mean, variance , covariance.

References

- Statistics and Probability in Data Science using Python - EDX
- Introduction to Probability by Bertsekas and Tsitsiklis
- https://lagunita.stanford.edu/courses/course-v1:OLI+ProbStat+Open_Jan2017/about
- https://www.probabilitycourse.com
- https://onlinecourses.science.psu.edu/stat414/node/46
- http://www.personal.soton.ac.uk/jav/soton/HELM/workbooks/workbook_35/35_4_tot_prob_bayes_thm.pdf
- http://www.statisticshowto.com/bayes-theorem-problems/
- http://www.statisticshowto.com/wp-content/uploads/2009/09/dependent-or-independent-event.jpg
- https://www.slideshare.net/marinasantini1/mathematics-for-language-technology
- https://gerardnico.com/data_mining/bayes