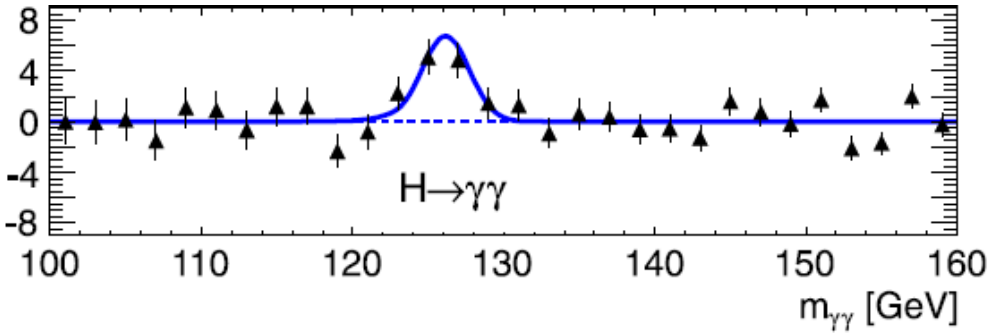
Introduction to Statistical Data Analysis

Chapter 1

Introduction &

Basic Probability



Oct 2023

Introduction

The aim of these notes is to present the most important concepts and methods of statistical data analysis.

- **Statistics** is the study of the collection, organization, analysis, interpretation, and presentation of data.
  - Statistics and physics are similar in that each starts from sets of basic principles.
  - We will study some basic calculations that are commonly performed on sampled data.

- **Data Analysis** is a very broad subject covering many techniques and types of data.
  - In Particle Physics, research methodology is based on 'Statistical Data Analysis'

Introduction

We will deal with a central concept known as **Uncertainty**.

In Particle Physics there are various elements of uncertainty:

- Theory is not deterministic
  (QM)
- Random measurement errors
  (present even without quantum effects)
- Things we could know in principle but don’t
  (e.g. from limitations of cost, time, ...)

We can quantify the uncertainty using **Probability**.

Relative Frequency Experiments

Suppose we repeat an experiment of tossing a die. Let

s be the number of times a “six” appears

n be the number of tosses

Then the ratio *s*/*n* becomes stable in the long run:

$$f = \frac{s}{n}$$

f approaches a limit as n -> ∞

This stability is the basis of probability theory!

Real Experiment

Tossing 10 coins 100 times!

* Result of first 25 experiment is given right.

P(T) = 128/250 = 0.512

P(H) = 122/250 = 0.488

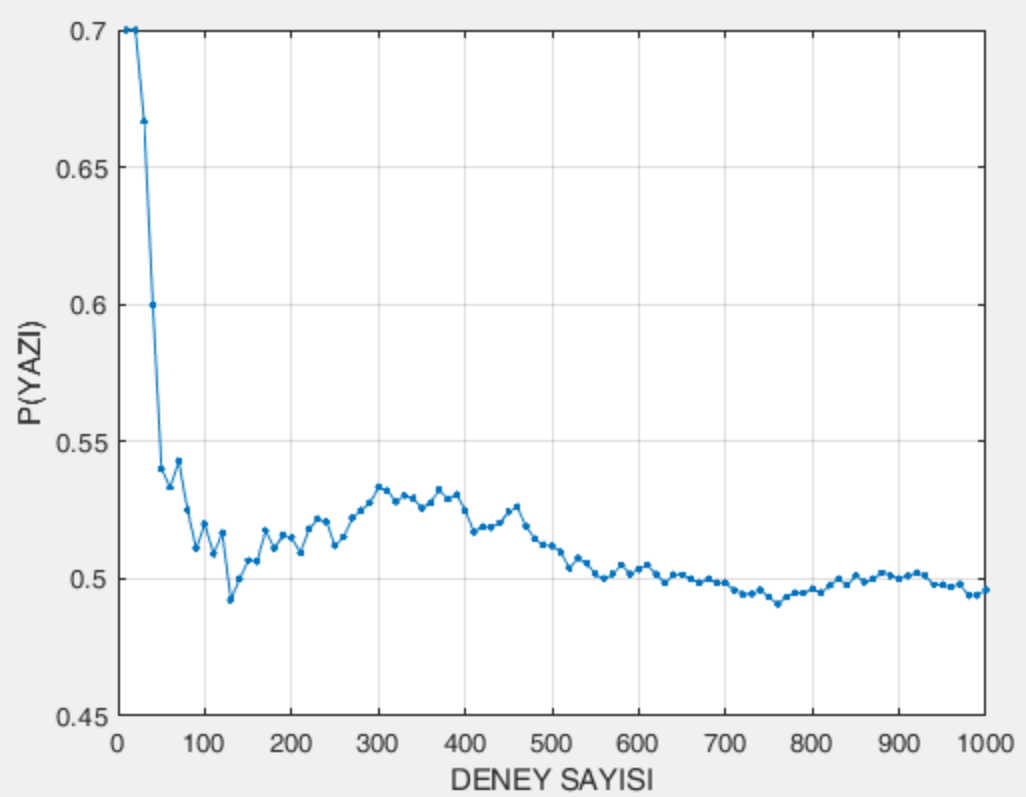
* By using 1000 data, we have:

P(T) = 496 / 1000 = 0.496

P(H) = 504 / 1000 = 0.504

Gözlem#	YAZI	TURA
1	7	3
2	7	3
3	6	4
4	4	6
5	3	7
6	5	5
7	6	4
8	4	6
9	4	6
10	6	4
11	4	6
12	6	4
13	2	8
14	6	4
15	6	4
16	5	5
17	7	3
18	4	6
19	6	4
20	5	5
21	4	6
22	7	3
23	6	4
24	5	5
25	3	7
Toplam	**128**	**122**
Oran	**0.512**	**0.488**

Real Experiment



Simulation

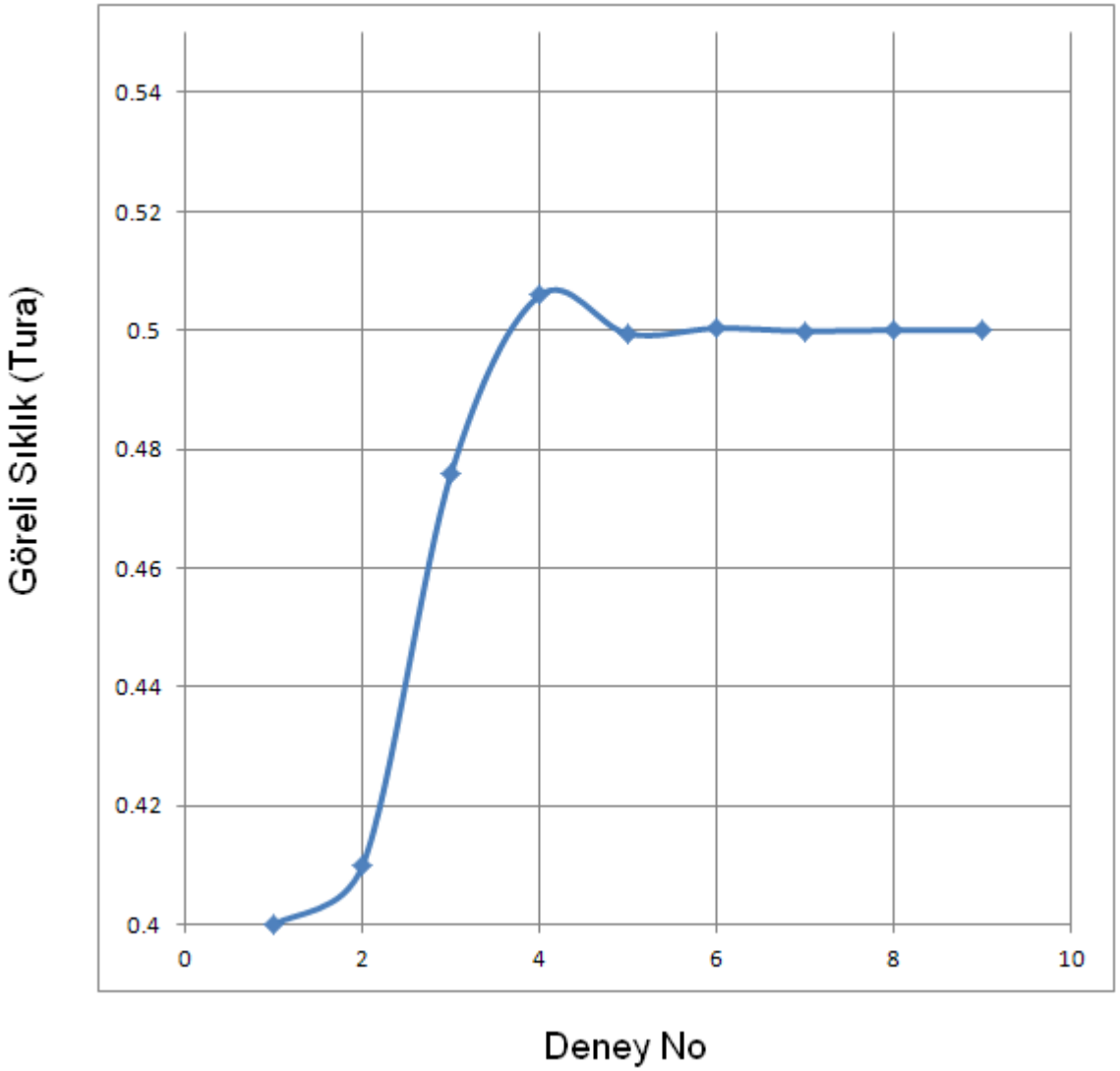
Here is the result obtained from a computer simulation for tossing of a coin and observing frequency of head!

HEAD

TAIL

n	s	f = s/n
10	4	0.4000000
100	41	0.4100000
1,000	476	0.4760000
10,000	5059	0.5059000
100,000	49942	0.4994200
1,000,000	500351	0.5003510
10,000,000	4998906	0.4998906
100,000,000	50006417	0.5000641
1,000,000,000	500000839	0.5000084

The result approaches a limit as $n \to \infty$


Göreli Sıklık (Tura)
0.54
0.52
0.5
0.48
0.46
0.44
0.42
0.4
0
2
4
6
8
10
Deney No

Real Experiment

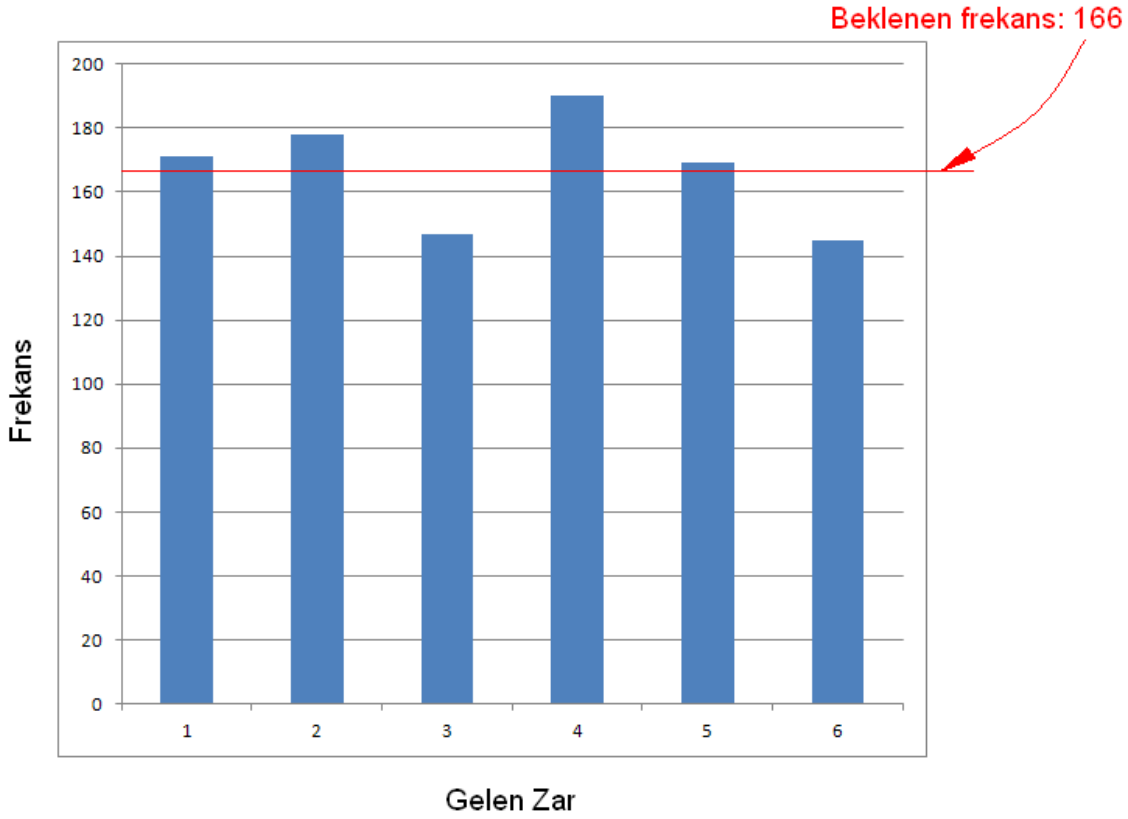
A dice is thrown 1000 times!

P(1) = 171 / 1000 = 0.171

P(2) = 178 / 1000 = 0.178

. . .

Zar çıktısı	s	f = s/n
1	171	0.171
2	178	0.178
3	147	0.147
4	190	0.190
5	169	0.169
6	145	0.145
Toplam	**1000**	**1.000**

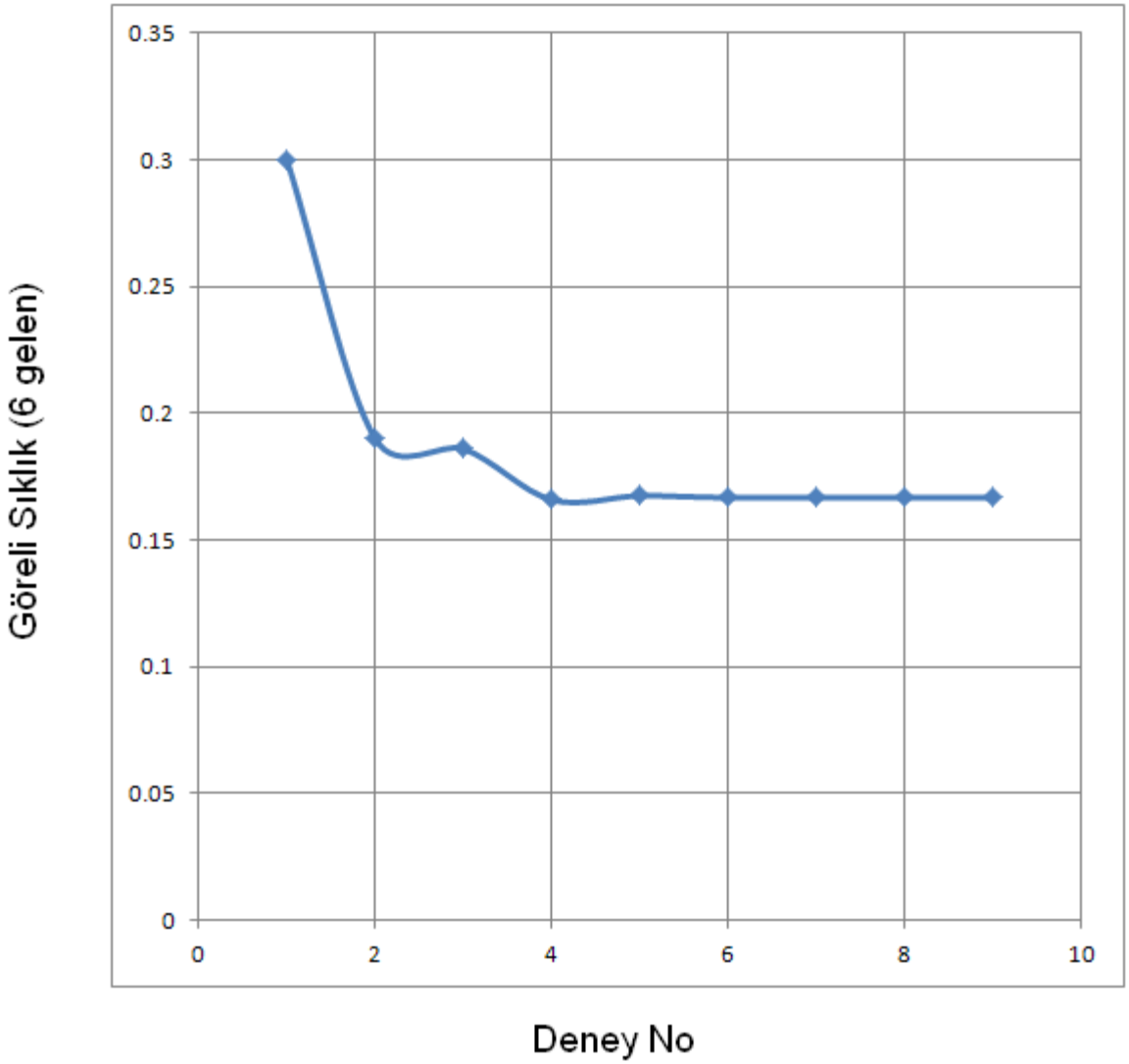


Simulation

Here is the result obtained from a computer simulation for tossing of a die and observing frequency of six!

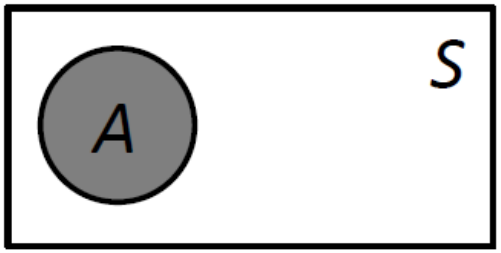
n	s	f = s/n
10	3	0.3000000
100	19	0.1900000
1,000	186	0.1860000
10,000	1659	0.1659000
100,000	16748	0.1674800
1,000,000	166705	0.1667050
10,000,000	1667210	0.1667210
100,000,000	16666290	0.1666629
1,000,000,000	166666653	0.1666666

The result approaches a limit as n -> ∞


Göreli Sıklık (6 gelen)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2
4
6
8
10
Deney No

Probability of an event

An event describes a set of outcomes of interest to which we can assign a probability.



Definition:

Probability is the measure of the *likelihood* that an event will occur.

Given a sample space *S* and event *A*, and if every sample point has an equal likelihood of occuring, we can write:

$$P(A) = \frac{n(A)}{n(S)}$$

We can see that the limits of probability are:

$P(S) = 1$, and $P(\phi) = 0$; and so: $0 \leq P(A) \leq 1$

The Bertrand's Paradox

Consider an equilateral triangle inscribed in a circle. Extracted uniformly one of the possible chords of the circle. What is the probability that the length of the extracted chord is larger than the side of the inscribed triangle?

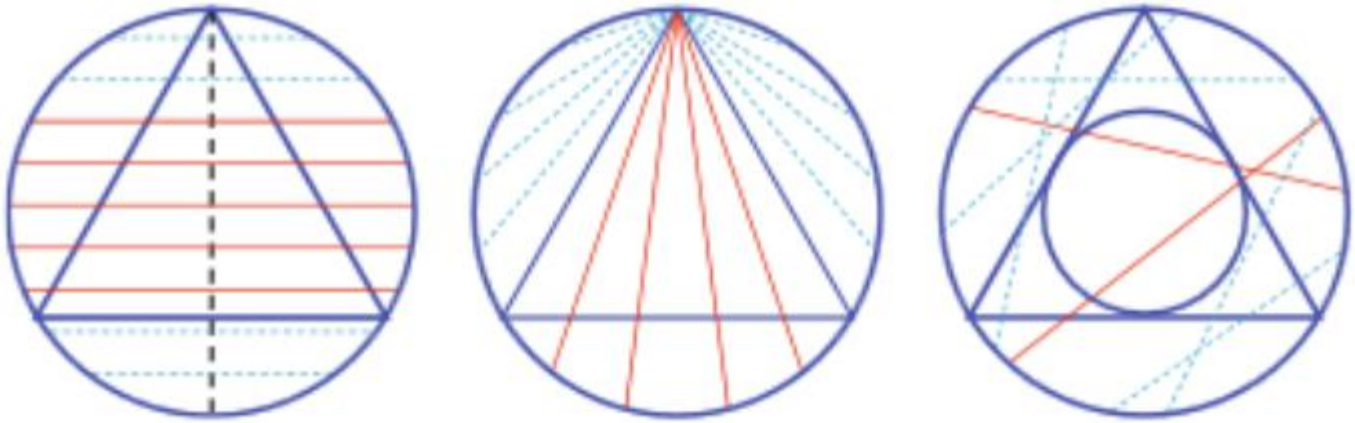
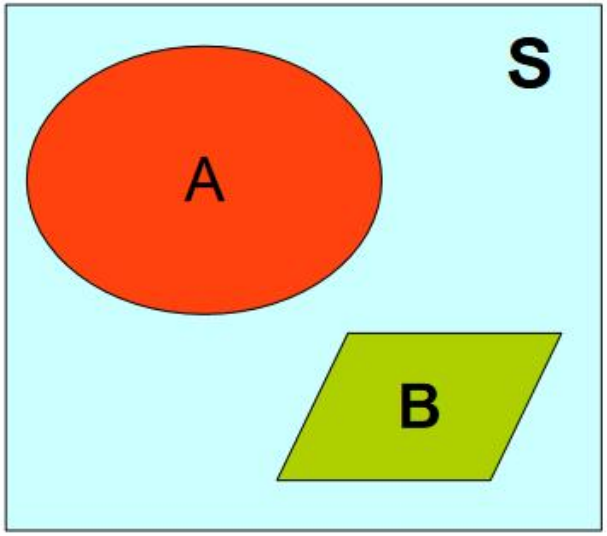


Fig. 1.2 Illustration of Bertrand's paradox: three different choices of *random extraction* of a chord in a circle lead apparently to probabilities that the cord is longer than the inscribed triangle's side of $1/2$ (*left*), $1/3$ (*center*) and $1/4$ (*right*), respectively. *Red solid lines* and *blue dashed lines* represent chords longer and shorter of the triangle's side, respectively

Axioms of Probability



Kolmogorov's axioms (1933)

1. For all $A \subset S$, $P(A) \geq 0$ — Positive definite
2. $P(S) = 1$ — Normalized
3. If $A \cap B = 0$, then $P(A \cup B) = P(A) + P(B)$ — Additive

   (A and B are *disjoint*)

Theorems of Probability

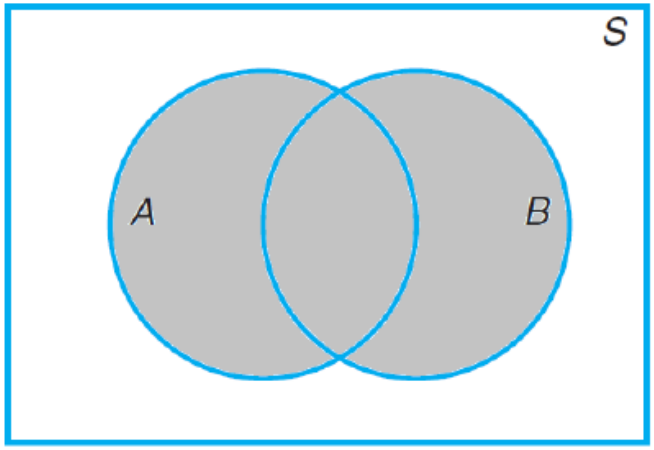
From Kolmogorov's axioms we can derive further properties:

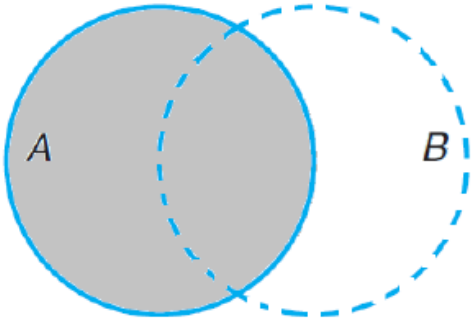
- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup \bar{A}) = 1$
- $P(\emptyset) = 0$
- If $A \subset B$, then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Subsets A and B are said independent if $P(A \cap B) = P(A)P(B)$

Do not confuse with disjoint subsets i.e. $A \cap B = 0$

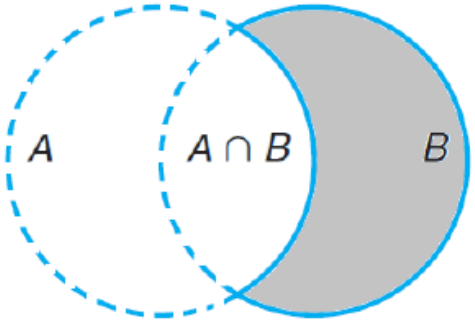
Theorem: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$





$P(A)$

$+$

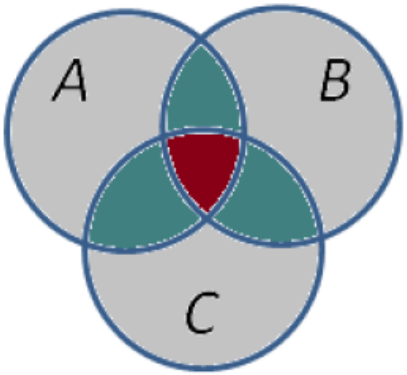


$P(B) - P(A \cap B)$

Theorem

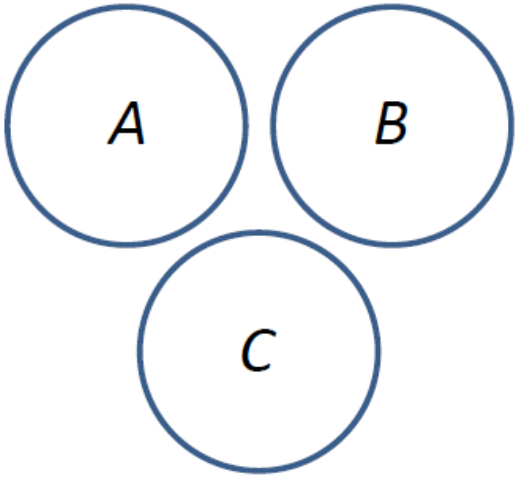
For three events A, B, and C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$



And for disjoint events we obtain:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - \cancel{P(A \cap B)} - \cancel{P(A \cap C)} - \cancel{P(B \cap C)} + \cancel{P(A \cap B \cap C)}$$



And if $C = \emptyset$, we obtain:

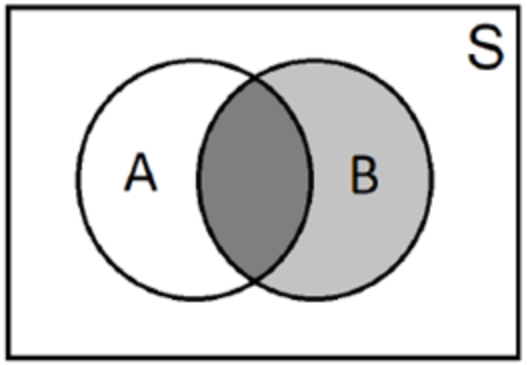
$$P(A \cup B \cup \cancel{C}) = P(A) + P(B) + \cancel{P(C)} - P(A \cap B) - \cancel{P(A \cap C)} - \cancel{P(B \cap C)} + \cancel{P(A \cap B \cap C)}$$

Conditional probability

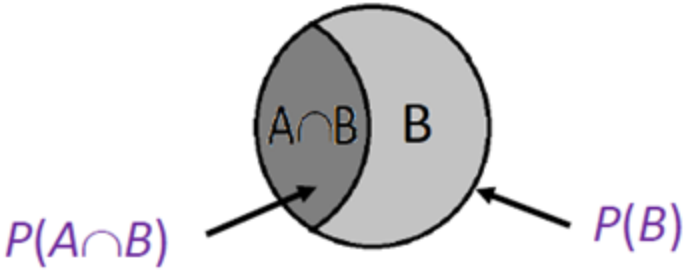
The probability of an event *A* occurring given that is it known that event *B* has occurred is called the *conditional probability* and is written as *P*(*A*|*B*); "the probability of *A* given *B*".

It can be shown that:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Probability is represented by the ***areas*** of events in a *proportional* Venn diagram. Given that event *B* has occured, the sample space is now reduced to *B*.



Example

In rolling a dice we have the sample space:

$$x = \{1, 2, 3, 4, 5, 6\}$$

$$P(x < 3 \mid x \text{ even}) = \frac{P((x < 3) \cap (x \text{ even}))}{P(x \text{ even})} = \frac{1/6}{3/6} = \frac{1}{3}$$

Independent events

Two events A and B are **independent** if

$$P(B|A) = P(B)$$

and

$$P(A|B) = P(A)$$

That is, if the probability of event *B* has no dependence on the occurrence of event *A*, and vice versa, then events *A* and *B* are *independent*.

Given *independence*, and using *conditional probability*:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A) \Rightarrow \boxed{P(A \cap B) = P(A)P(B)}$$

*This is a **test** for the independence of two events.*

We used *independence* in the previous lecture to simplify $P(A \cap B)$ to $P(A)P(B)$, in those cases independence was an *assumption*; such assumptions can be poor and lead to wrong results!

Now, we will use this relation as a tool to ***test*** for independence.

Bayes' Theorem

From the definition of conditional probability we have:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

But $P(A \cap B) = P(B \cap A)$, so:

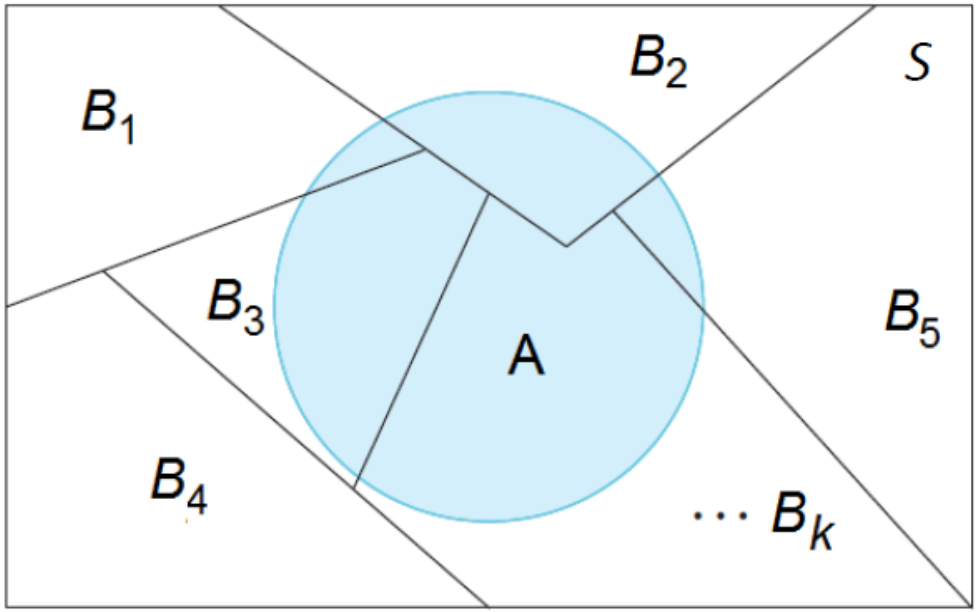
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem

First published (posthumously) by the Reverend Thomas Bayes (1702−1761)

An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Soc. **53c**(1763) 370; reprinted in Biometrika, **45** (1958) 293.

Total Probability and Bayes' theorem



The probability of event *A* can be constructed from the total probability of all intersecting events:

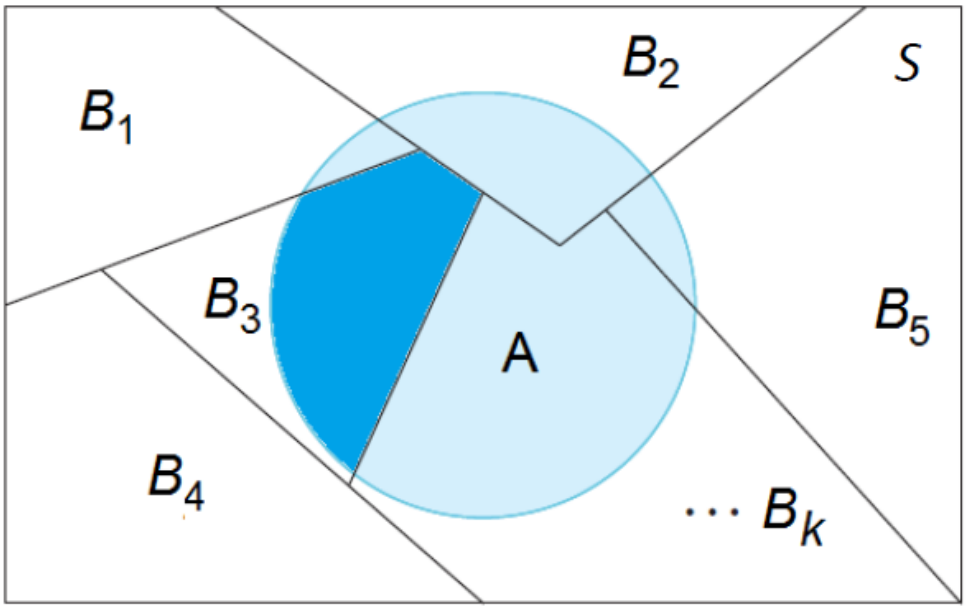
$$P(A) = \sum_{i=1}^{k} P(A \cap B_i)$$

$$P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)}$$

$$\Rightarrow \; P(A \cap B_i) = P(A|B_i)P(B_i) \quad \Rightarrow$$

Total probability of event A

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$



Bayes' theorem:

$$P(B_r|A) = \frac{P(A|B_r)P(B_r)}{P(A)}$$

With:

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Proof:

$$P(B_r|A) = \frac{P(B_r \cap A)}{P(A)} = \frac{P(A \cap B_r)}{P(A)} = \frac{P(A|B_r)\,P(B_r)}{P(A)}$$

Example

Suppose the probability (for anyone) to have the disease A is:

$$P(A) = 0.001$$
$$P(\text{no } A) = 0.999$$

Consider a test for that disease. The result can be 'pos' or 'neg' :

$$P(\text{pos}|A) = 0.98$$
$$P(\text{neg}|A) = 0.02$$

← probabilities to (in)correctly Identify an infected person

$$P(\text{pos}|\text{no } A) = 0.03$$
$$P(\text{neg}|\text{no } A) = 0.97$$

← probabilities to (in)correctly Identify a healthy person

Suppose your result is 'pos'. How worried should you be?

The probability to have the disease A, given a 'pos' result is:

$$P(A|pos) = \frac{P(pos|A)P(A)}{P(pos|A)P(A)+P(pos|no\,A)P(no\,A)}$$

$$= \frac{0.98\times0.001}{0.98\times0.001+0.03\times0.999}$$

$$= 0.032$$

i.e. you're probably OK!

Your viewpoint: my degree of belief that I have disease A is 3.2%

Your doctor's viewpoint: 3.2% of people like this will have disease A

Example

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\text{no } A)[1 - P(A)]}$$

Detector for particle identification

In proton-proton collisions we have: 90% pions, 10% kaons

1. Kaon identification: 95% efficient
2. Pion misidentification: 6%

Question: if the particle identification indicates a kaon, what is the probability that it is a real kaon / a real pion?

Solution: Output can be particle is identified (detected) $= I$
or particle is not identified (not detected) $= N$

$$P(\pi) = 0.90$$
$$P(K) = 0.10$$
$$P(I|K) = 0.95$$
$$P(N|K) = 0.05$$
$$P(I|\pi) = 0.94$$
$$P(N|\pi) = 0.06$$

$$P(K|I) = \frac{P(I|K)P(K)}{P(I|K)P(K) + P(I|\pi)P(\pi)}$$

$$= \frac{(0.95)(0.10)}{(0.95)(0.10) + (0.94)(0.90)} = 0.101$$

$$P(\pi|I) = 0.899$$