CENTRE DE RECHERCHE PUBLIC SANTÉ

Microarray Center

APPLIED STATISTICS

Lecture 2

Introduction to Probability

Petr Nazarov
petr.nazarov@crp-sante.lu

10-11-2009

- **Experiments, counting rules and assigning probabilities**
  - experiments
  - counting rules, combinations, permutations
  - assigning probabilities
- **Events and their probabilities**
  - event
  - complement event
  - intersection of two events
  - addition law
- **Conditional probability**
  - conditional probability
  - independent events
  - multiplication law
- **Bayes' theorem**
  - the theorem of Bayes and it's application
  - tabular approach

Experiment
A process that generates well-defined outcomes

Experiment	Experimental Outcomes
Toss a coin	Head, tail
Select a part for inspection	Defective, nondefective
Conduct a sales call	Purchase, no purchase
Roll a die	1, 2, 3, 4, 5, 6
Play a football game	Win, lose, tie

$S = \{\text{Head, Tail}\}$

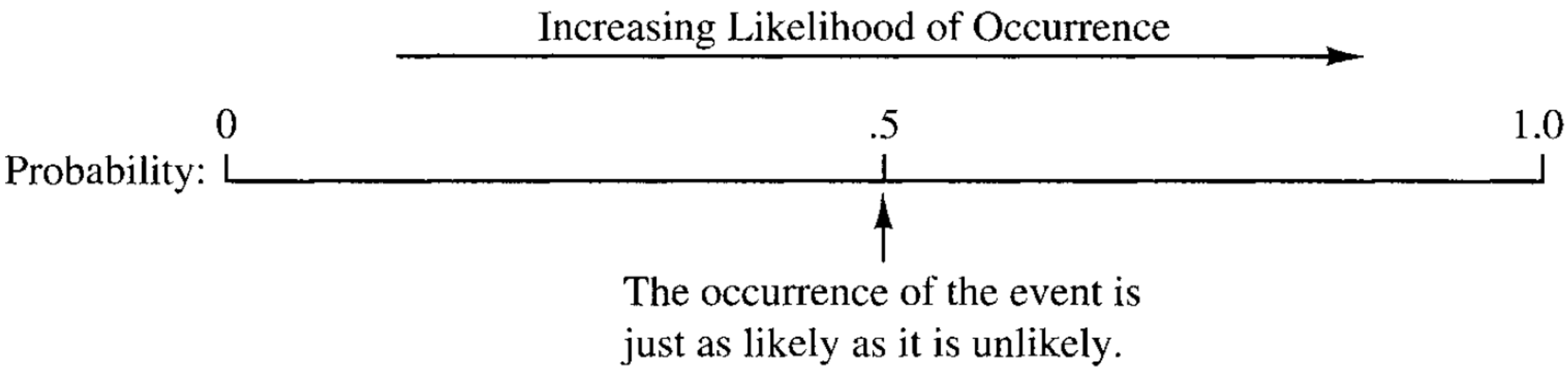
$S = \{\text{Defective, Nondefective}\}$

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

Sample space
The set of all experimental outcomes.

Sample point
An element of the sample space.
A sample point represents an
experimental outcome

Probability
A numerical measure of the likelihood that an event will occur.



Multiple-step experiments

Consider the experiment of tossing two coins. Let the experimental outcomes be defined in terms of the pattern of heads and tails appearing on the upward faces of the two coins. **How many experimental outcomes are possible for this experiment?** The experiment of tossing two coins can be thought of as a two-step experiment in which step 1 is the tossing of the first coin and step 2 is the tossing of the second coin.

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

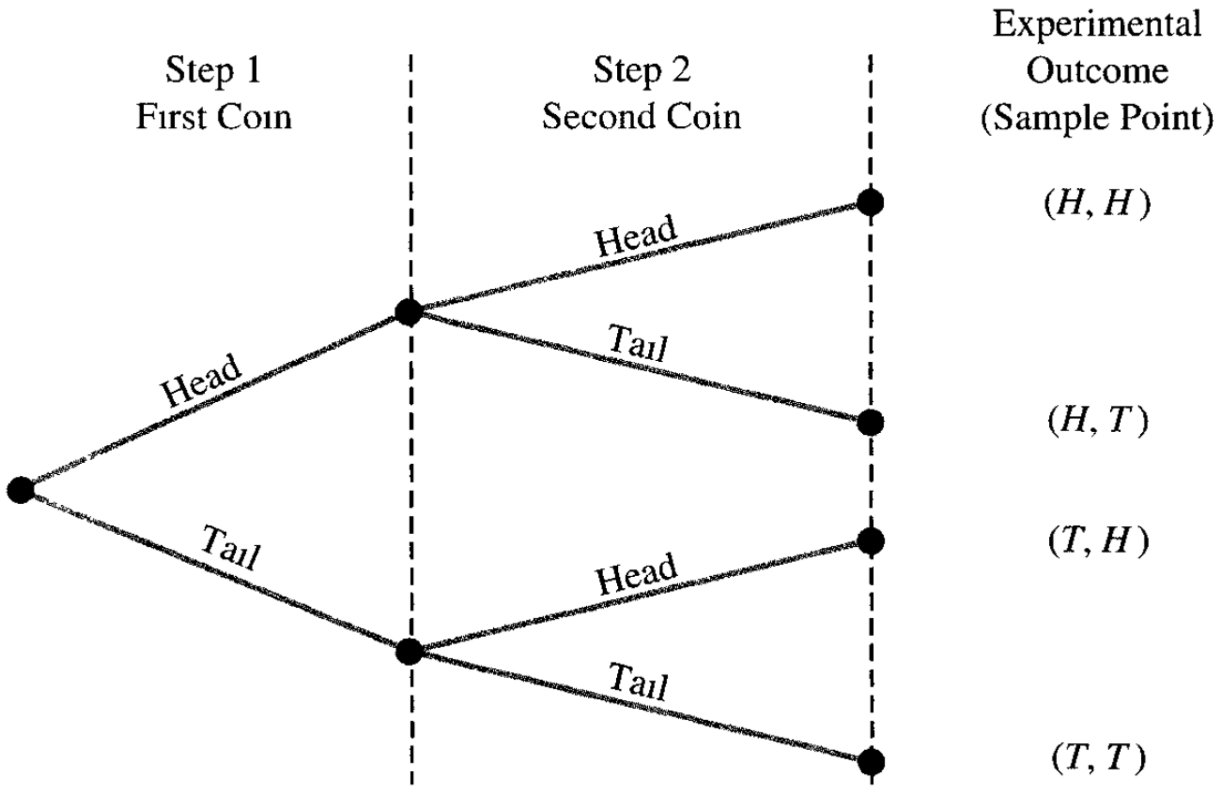
Counting rule for a multi-step experiment

If an experiment can be described as a sequence of k steps with n_1 possible outcomes on the first step, n_2 possible outcomes on the second step, and so on, then the total number of experimental outcomes is given by $n_1 \cdot n_2 \cdot \ldots \cdot n_k$.

? How many outcomes can appear after **6 tossing** of a coin?

Tree diagram

A graphical representation that helps in visualizing a multiple-step experiment.



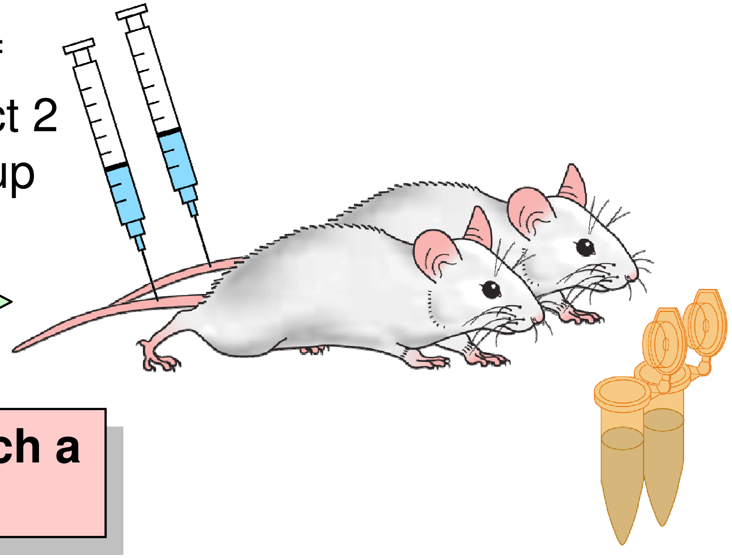
Without the tree diagram, one might think only three experimental outcomes are possible for two tosses of a coin 0 heads, 1 head, and 2 heads

EXPERIMENTS, COUNTING RULES AND PROBABILITIES

Counting Rule 2

Mouse population: 5 mice

To test the state of population, let's select 2 for a routine check-up



How many variants of such a selection exits?

Counting rule for combinations

The number of combinations of *n* objects taken from *N* at a time is:

$$\binom{N}{n} = C_n^N = C_N^n = {}_N C_n = C(N, n) = \frac{N!}{n!(N-n)!}$$

So for mouse selection:

$$C_2^5 = \frac{5!}{2!(5-2)!} = \frac{2 \cdot 3 \cdot 4 \cdot 5}{2 \cdot 2 \cdot 3} = 10$$

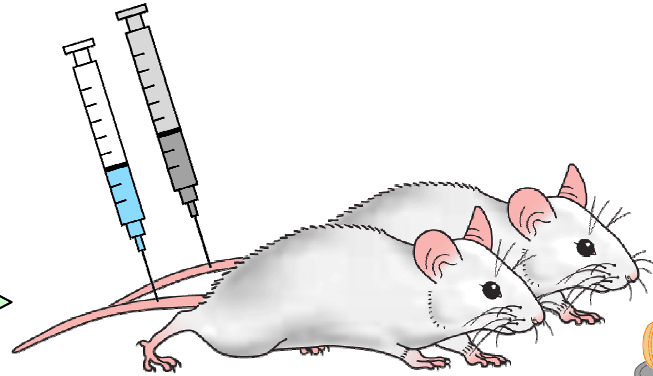
Let's consider the lottery "6 from 47"

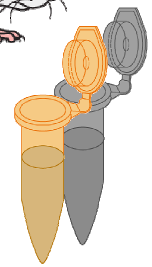
$$C_6^{47} = \frac{47!}{6!(47-6)!} = 10\ 737\ 573$$

Counting Rule 3

Mouse population: 5 mice

To test the state of population, let's select 2 for ***2 different*** check-up





How many variants of such a selection exits?

Counting rule for permutations

The number of permutation of *n* objects taken from *N* at a time is:

$$P_n^N = \frac{N!}{(N-n)!}$$

So for mouse selection:

$$P_2^5 = \frac{5!}{(5-2)!} = 20$$

How many triplets we can build from 4 nucleotides without repetition?

$$P_3^4 = \frac{4!}{(4-3)!} = 24$$

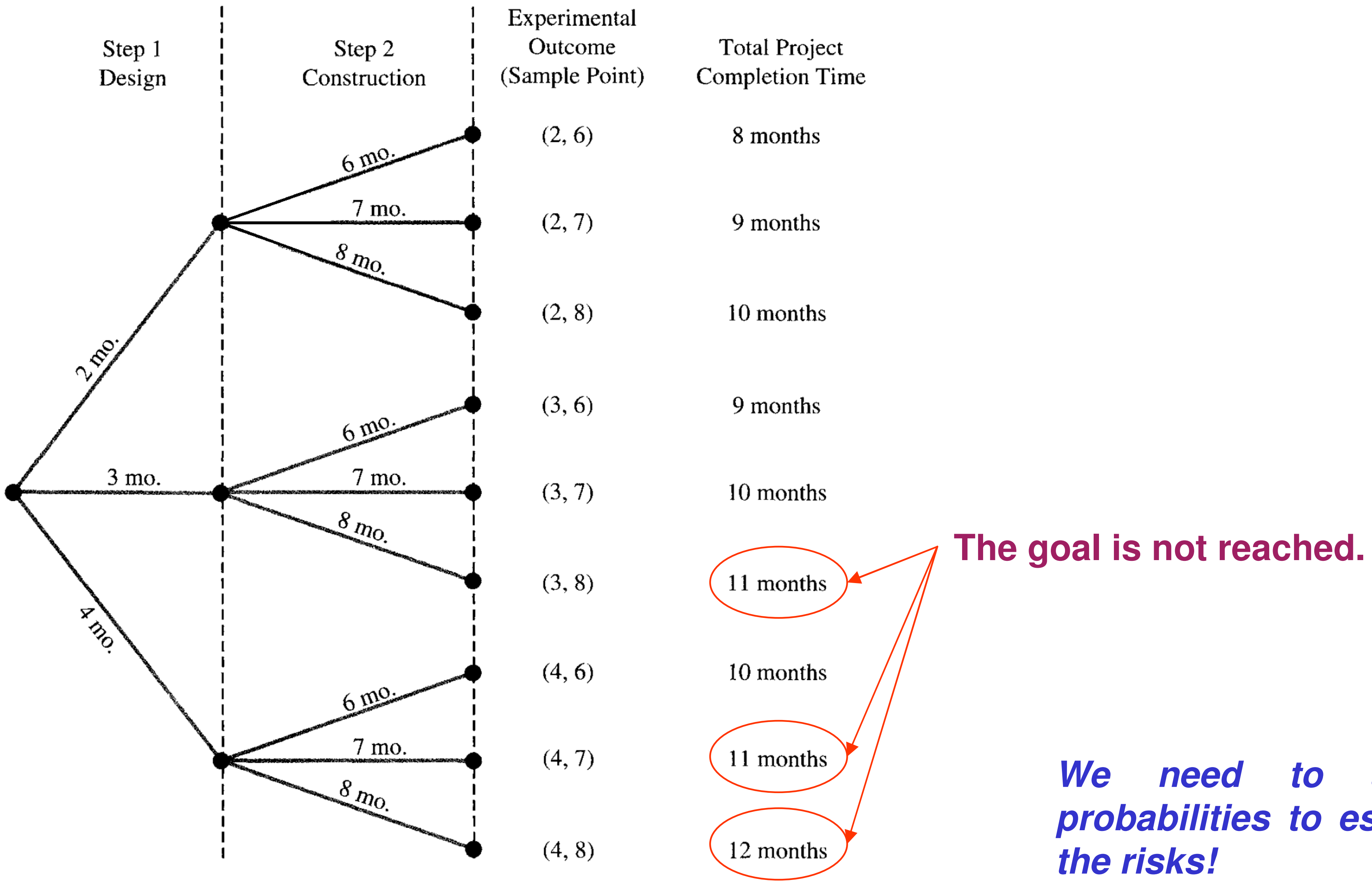
Example: KP&L

Kentucky Power & Light Company (KP&L) is starting a project designed to increase the generating capacity of one of its plants in northern Kentucky. The project is divided into **two sequential stages** or steps: **stage 1 (design)** and **stage 2 (construction).** Even though each stage will be scheduled and controlled as closely as possible, management cannot predict beforehand the exact time required to complete each stage of the project. An analysis of similar construction projects revealed possible completion times for the **design stage of 2, 3, or 4** months and possible completion times for the **construction stage of 6, 7, or 8 months**. In addition, because of the critical need for additional electrical power, management set a **goal of 10 months** for the completion of the entire project.

Completion Time (months)			
Stage 1 Design	Stage 2 Construction	Notation for Experimental Outcome	Total Project Completion Time (months)
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

Anderson et al Statistics for Business and Economics

Example: KP&L



We need to assign probabilities to estimate the risks!

Basic requirements for assigning probabilities

Two requirements that restrict the manner in which probability assignments can be made:

1. For each experimental outcome E_i , probability must be $\mathbf{0 \leq P(E) \leq 1}$
2. Considering all experimental outcomes, we must have
   $\mathbf{P(E_1) + P(E_2) + \ldots + P(E_n) = 1}$

But how to assign it?

3 basic methods of probability assigning:

Classical method

Frequency method

Subjective method

Theoretical method

Classical method
A method of assigning probabilities that is appropriate when all the experimental outcomes are equally likely.

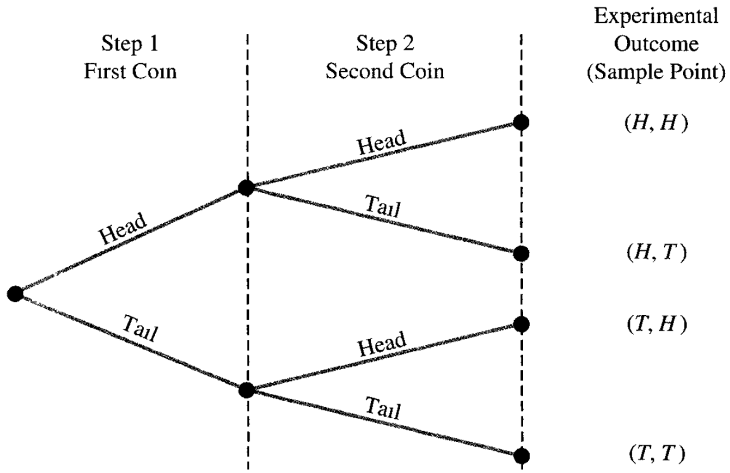
If ***n*** experimental outcomes are possible, a probability of **1/*n*** is assigned to each experimental outcome. When using this approach, the two basic requirements for assigning probabilities are automatically satisfied.

P(1)=1/6
P(2)=1/6
P(3)=1/6
P(4)=1/6
P(5)=1/6
P(6)=1/6

P(H)=1/2
P(T)=1/2

P(2xH)=
P(2xT)=
P(T+H)=

?



1/4
1/4
1/4
1/4

P(2xH)=1/4
P(2xT)=1/4
P(T+H)=1/2

Relative frequency method

A method of assigning probabilities that is appropriate when data are available to estimate the proportion of the time the experimental outcome occurs if the experiment is repeated a large number of times.

Example: A clerk recorded the number of patients waiting for service in the X-ray department for a local hospital at 9:00 a.m. on 20 successive days, and obtained the following results.

Number waiting	Number of days outcome occurred
0	2
1	5
2	6
3	4
4	3
Total	**20**

→

Number waiting	Relative frequency
0	0.1
1	0.25
2	0.3
3	0.2
4	0.15
Total	1

Subjective method
A method of assigning probabilities on the basis of judgment

? What is a probability to meet a dinosaur, going out from this building?

Classical method: ***50%***

Frequency method: ***no data***

Subjective method: ***no way!***

Theoretical method
A method of assigning probabilities on the basis knowledge of probability distribution… *(to be discussed later)*

Management decided to conduct a study of the completion times for similar projects undertaken by KP&L over the past three years. The results of a study of **40 similar projects** are summarized in Table. After reviewing the results of the study, management decided to employ the relative frequency method of assigning probabilities. Management could have provided subjective probability estimates, but felt that the current project was quite similar to the 40 previous projects. Thus, the relative frequency method was judged best.

Sample Point	Project Completion Time	Probability of Sample Point
(2, 6)	8 months	$P(2, 6) = 6/40 = .15$
(2, 7)	9 months	$P(2, 7) = 6/40 = .15$
(2, 8)	10 months	$P(2, 8) = 2/40 = .05$
(3, 6)	9 months	$P(3, 6) = 4/40 = .10$
(3, 7)	10 months	$P(3, 7) = 8/40 = .20$
(3, 8)	11 months	$P(3, 8) = 2/40 = .05$
(4, 6)	10 months	$P(4, 6) = 2/40 = .05$
(4, 7)	11 months	$P(4, 7) = 4/40 = .10$
(4, 8)	12 months	$P(4, 8) = 6/40 = .15$
		Total 1.00

Event

Event
A collection of sample points.

Sample Point	Project Completion Time	Probability of Sample Point
(2, 6)	8 months	$P(2, 6) = 6/40 = .15$
(2, 7)	9 months	$P(2, 7) = 6/40 = .15$
(2, 8)	10 months	$P(2, 8) = 2/40 = .05$
(3, 6)	9 months	$P(3, 6) = 4/40 = .10$
(3, 7)	10 months	$P(3, 7) = 8/40 = .20$
(3, 8)	11 months	$P(3, 8) = 2/40 = .05$
(4, 6)	10 months	$P(4, 6) = 2/40 = .05$
(4, 7)	11 months	$P(4, 7) = 4/40 = .10$
(4, 8)	12 months	$P(4, 8) = 6/40 = .15$
		Total 1.00

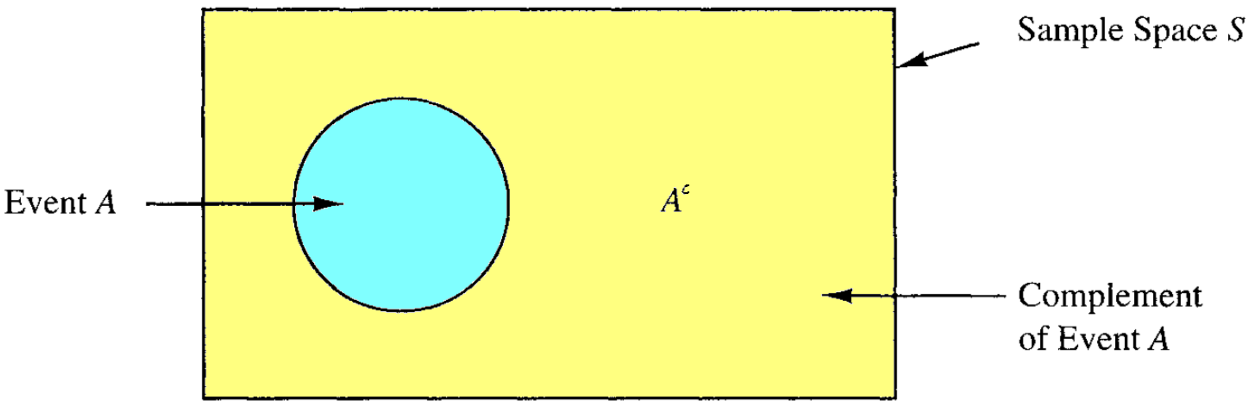
Let us denote the event of successful completion as C, so
C={(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)}

Probability of an event
The probability of any event is equal to the sum of the probabilities of the sample points in the event.

$P(C) = 0.7$

Complement of event A
The event consisting of all sample points that are not in A.

$$P(A) + P(A^C) = 1$$

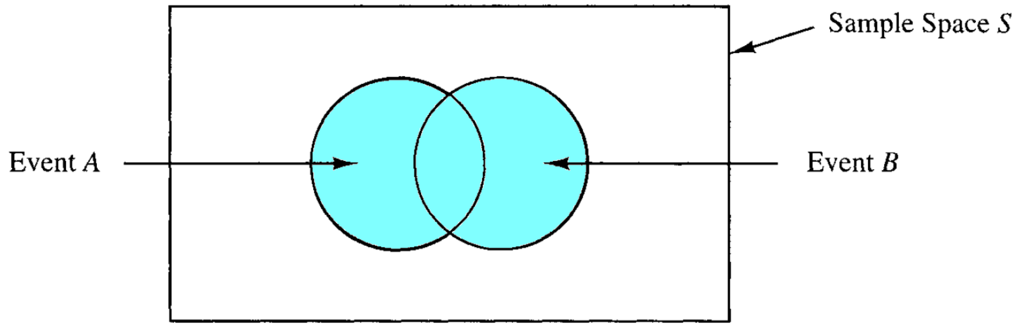


Venn diagram
A graphical representation for showing symbolically the sample space and operations involving events. Usually the sample space is represented by a rectangle and events are represented as circles within the sample space.

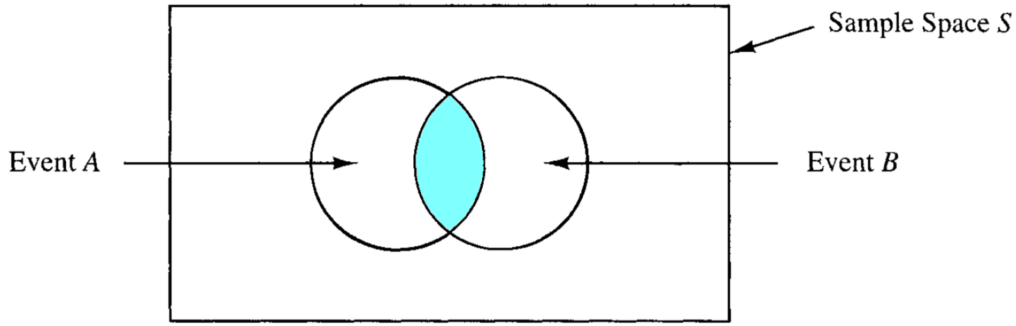
if $P(C) = 0.7$ $\quad$ $P(C^C) = 0.3$

Union, Intersection and Addition Law for Two Events

Union of A and B
The event containing all sample points belonging to *A* or *B* or both. The union is denoted $A \cup B$.



Intersection of A and B
The event containing the sample points belonging to both *A* and *B*. The intersection is denoted $A \cap B$.



Addition Law
The event containing the sample points belonging to both *A* and *B*. The intersection is denoted $A \cap B$.

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

Practical Example

Consider a study conducted by the personnel manager of a big computer software company. The study showed that **30%** of the employees who left the firm within two years did so primarily because they were **dissatisfied with their salary**, **20%** left because they were dissatisfied with their work **assignments**, and **12%** of the former employees indicated **dissatisfaction with both their salary and their work assignments**. What is the probability that an employee who leaves within two years does so because of dissatisfaction with salary, dissatisfaction with the work assignment, or both?

Let

S = the event that the employee leaves because of salary

W = the event that the employee leaves because of work assignment

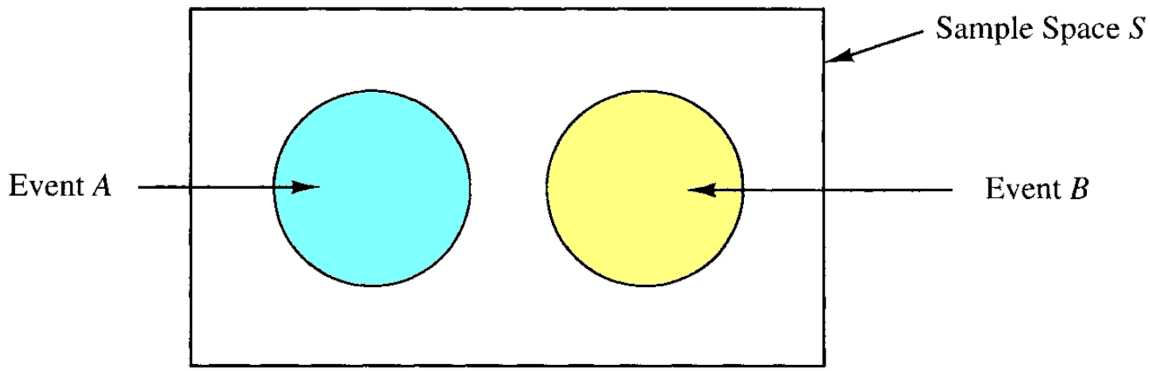
We have $P(S) = 0.30$, $P(W) = 0.20$, and $P(S \cap W) = 0.12$. Using equation for the addition law, we have

$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.30 + 0.20 - 0.12 = \mathbf{0.38}$

Anderson et al Statistics for Business and Economics

Mutual Exclusive Events

Mutually exclusive events
Events that have no sample points in common; that is, $A \cap B$ is empty and $P(A \cap B) = 0$.



Addition Law for mutually exclusive events

$$P(A \cup B) = P(A) + P(B)$$

What is the probability that a die shows 5 or 6 ?

Conditional probability
The probability of an event given that another event already oc-curred. The conditional probability of *A* given *B* is denoted *P*(*A* | *B*)

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

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

Example:
Let's consider the situation of the promotion status of male and female officers of a major metropolitan police force in the eastern United States. The police force consists of 1200 officers, 960 men and 240 women. Over the past two years, 324 officers on the police force received promotions. The specific breakdown of promotions for male and female officers is shown in Table.

	Men	Women	Totals
Promoted	288	36	324
Not Promoted	672	204	876
Totals	960	240	1200

Female association raised a discrimination case !

Example: Police

	Men	Women	Totals
Promoted	288	36	324
Not Promoted	672	204	876
Totals	960	240	1200

M = event an officer is a man

W = event an officer is a woman

A = event an officer is promoted

A^c = event an officer is not promoted

Joint probability table

Joint probabilities appear in the body of the table

	Men (M)	Women (W)	Totals
Promoted (A)	.24	03	.27
Not Promoted (A^c)	.56	.17	73
Totals	.80	.20	1.00

Marginal probabilities appear in the margins of the table

CRP SANTÉ
CENTRE DE RECHERCHE PUBLIC

CONDITIONAL PROBABILITY

Example: Police

	Men	Women	Totals
Promoted	288	36	324
Not Promoted	672	204	876
Totals	960	240	1200

Joint probabilities appear in the body of the table

	Men (M)	Women (W)	Totals
Promoted (A)	.24	03	.27
Not Promoted (A^c)	.56	.17	73
Totals	.80	.20	1.00

Marginal probabilities appear in the margins of the table

$$P(A \mid M) = \frac{288}{960} = \frac{288/1200}{960/1200} = \frac{.24}{.80} = .30$$

$$P(A \mid M) = \frac{P(A \cap M)}{P(M)} = \frac{.24}{.80} = .30 \qquad P(A \mid W) = \frac{P(A \cap W)}{P(W)} = \frac{.03}{.20} = .15$$

? What conclusion can you make, seeing this result?

Only that a randomly taken man will be promoted with x2 higher probability than a randomly selected woman.

NO CONCLUSION CAN BE MADE CONCERNING THE DISCRIMINATION !!!

Multiplication law
A probability law used to compute the probability of the intersection of two events.

$$P(A \cap B) = P(B)P(A \mid B)$$

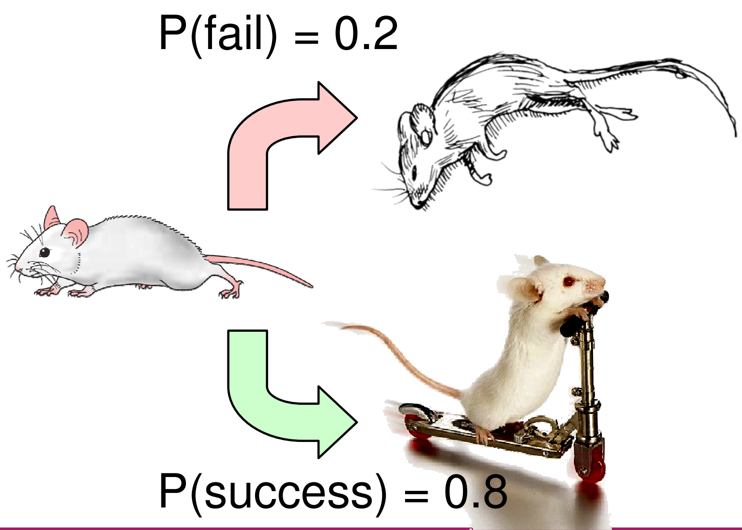
$$P(A \cap B) = P(A)P(B \mid A)$$

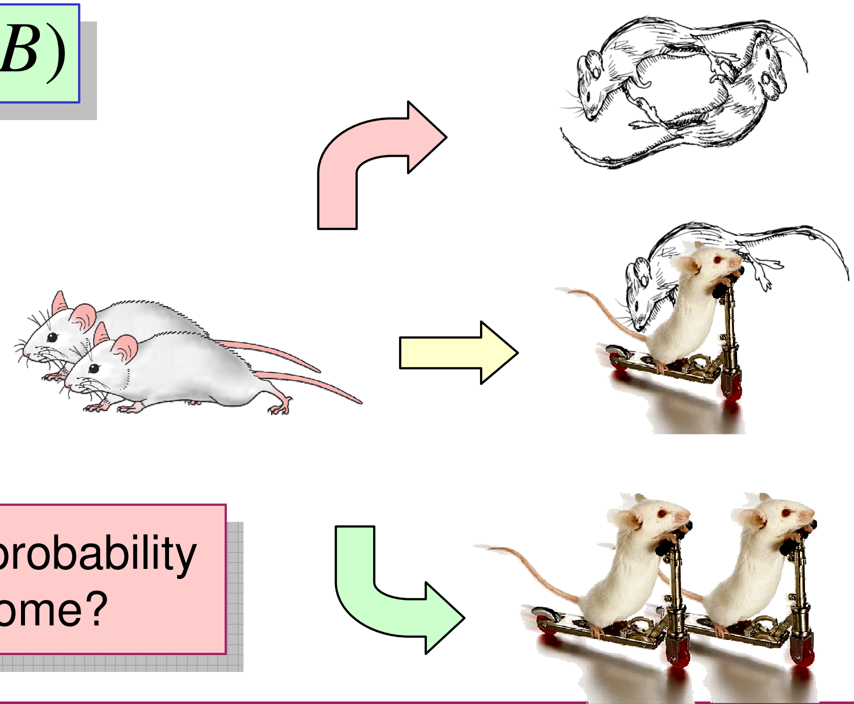
Independent events
Two events A and B that have no influence on each other.

$$P(A \mid B) = P(A)$$

$$P(B \mid A) = P(B)$$

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





? What is the probability of each outcome?

The Theorem of Bayes

Prior probabilities
Initial estimates of the probabilities of events.

Posterior probabilities
Revised probabilities of events based on additional information

Bayes' theorem
A method used to compute posterior probabilities.

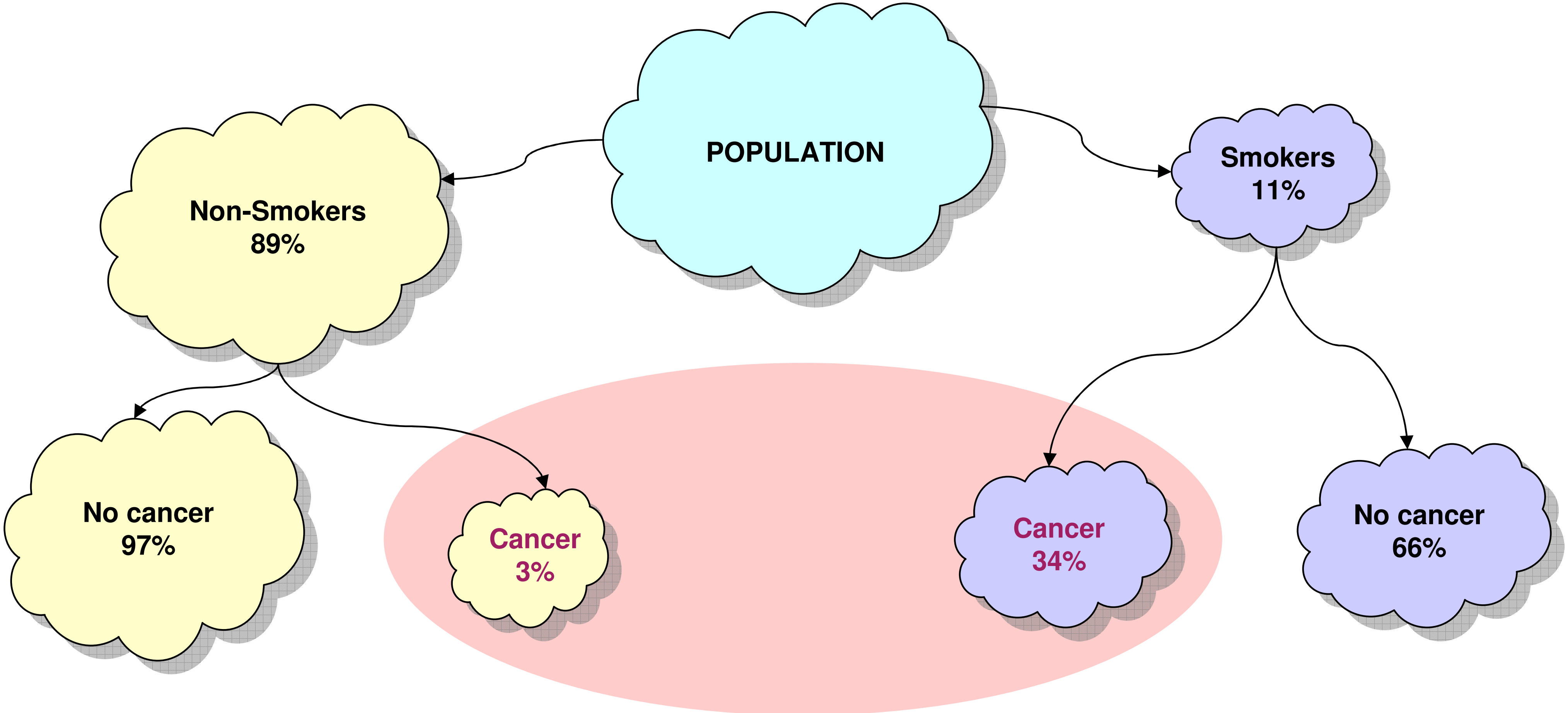
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B)}$$

$$P(A_i \mid B) = \frac{P(A_i) P(B \mid A_i)}{P(A_1) P(B \mid A_1) + P(A_2) P(B \mid A_2) + \cdots + P(A_n) P(B \mid A_n)}$$

Example: Bayes' Theorem

Medical researchers know that the probability of getting lung cancer if a person smokes is **0.34**. The probability that a non-smoker will get lung cancer is **0.03**. It is also known that **11%** of the population smokes. What is the probability that a person with lung cancer will have been a smoker?



CONDITIONAL PROBABILITY

Example: Bayes' Theorem

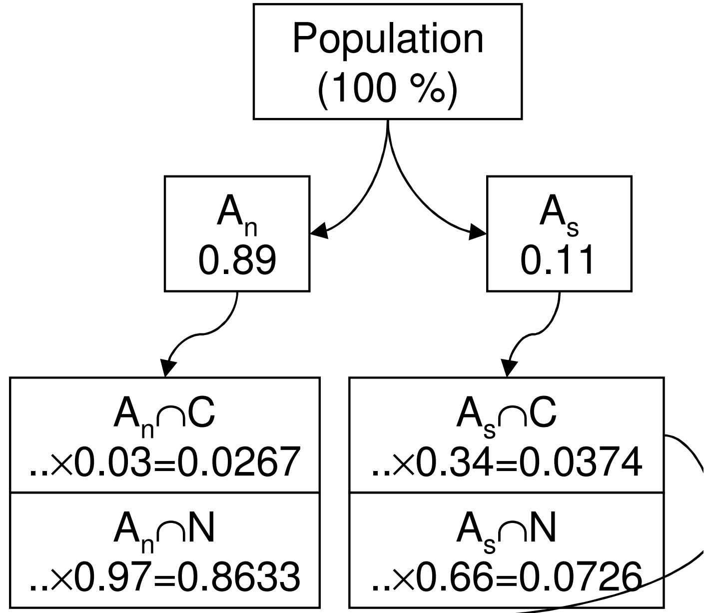
Medical researchers know that the probability of getting lung cancer if a person smokes is **0.34**. The probability that a non-smoker will get lung cancer is **0.03**. It is also known that **11%** of the population smokes. What is the probability that a person with lung cancer will have been a smoker?

Let's define events:

A_s – smoker
A_n – non-smoker
C – get a lung cancer
N – no cancer

$P(A_s) = 0.11$
$P(C \mid A_s) = 0.34$
$P(C \mid A_n) = 0.03$

$P(A_s \mid C)$ – ?



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \rightarrow P(A_s \mid C) = \frac{P(A_s \cap C)}{P(C)}$$

$$P(A \cap B) = P(A)P(B \mid A) \rightarrow P(A_i \cap C) = P(A_i)P(C \mid A_i)$$

$$P(A \cup B) = P(A) + P(B) \rightarrow P(C) = P(A_s \cap C) + P(B_s \cap C)$$

$P(C) = P(A_n \cap C) + P(A_s \cap C) = 0.0641$

$$P(A_s \mid C) = \frac{0.0374}{0.0641} \approx 0.58$$

Example: Bayes' Theorem and Tabular Approach

Medical researchers know that the probability of getting lung cancer if a person smokes is **0.34**. The probability that a non-smoker will get lung cancer is **0.03**. It is also known that **11%** of the population smokes. What is the probability that a person with lung cancer will have been a smoker?

Let's define events:

- A_s – smoker
- A_n – non-smoker
- C – get a lung cancer

$P(A_s) = 0.11$
$P(C|A_s) = 0.34$
$P(C|A_n) = 0.03$

$P(A_s|C)$ – ?

$$P(A \cap B) = P(A)P(B \mid A)$$

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

	prior probability $P(A_i)$	conditional probability $P(C\|A_i)$	joint probability $P(A_i \cap C)$	posterior probability $P(A_i\|C)$
Smokers (A_s)	0.11	0.34	0.0374	0.5834633
Non-smokers (A_n)	0.89	0.03	0.0267	0.4165367
Totals	1		P(C)=0.0641	1

Thank you for your attention

to be continued…