Probability Theory (revisited)

Summary

- Probability v.s. plausibility
- Random variables
- Simulation of Random Experiments

Challenge

The alarm of a shop rang .

Soon afterwards, a man was seen running in the street, persecuted by a policeman.

Is the man a thief ?

Can a machine reason like us ?

Jacob Bernoulli (1654-1705)

Bernoulli definition

Definition of probability attributed to Jacob Bernoulli (1689):

$$P = \frac{m}{n}$$

m – number of favorable cases
n – total number of cases

This definition establishes a link between probabilities and the output of experiments.

Question: *how to manipulate probabilities in a consistent way ?*

Kolmogorov (1903-1987)

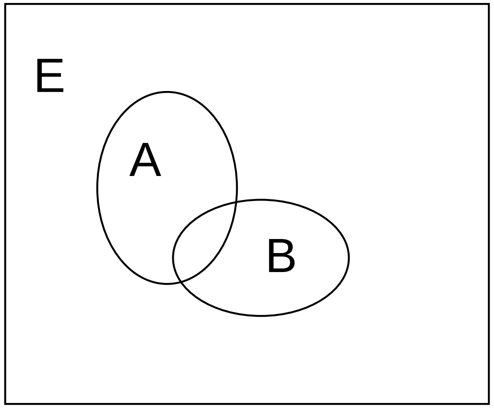
Kolmogorov Axioms

Kolmogorov defined a set of axioms for Probability Calculus based on set theory and measure theory:

He defines:

• sigma algebra of sets, closed with respect to complement and union of a countable number of sets.

• a probability measure for the sets belonging to the sigma field, denoted as events.



Strong point: all the operations with probabilities can be defined from the axioms in a consistent way.

Question: what is the relationship between probabilities and experimental data ?

Probability Space

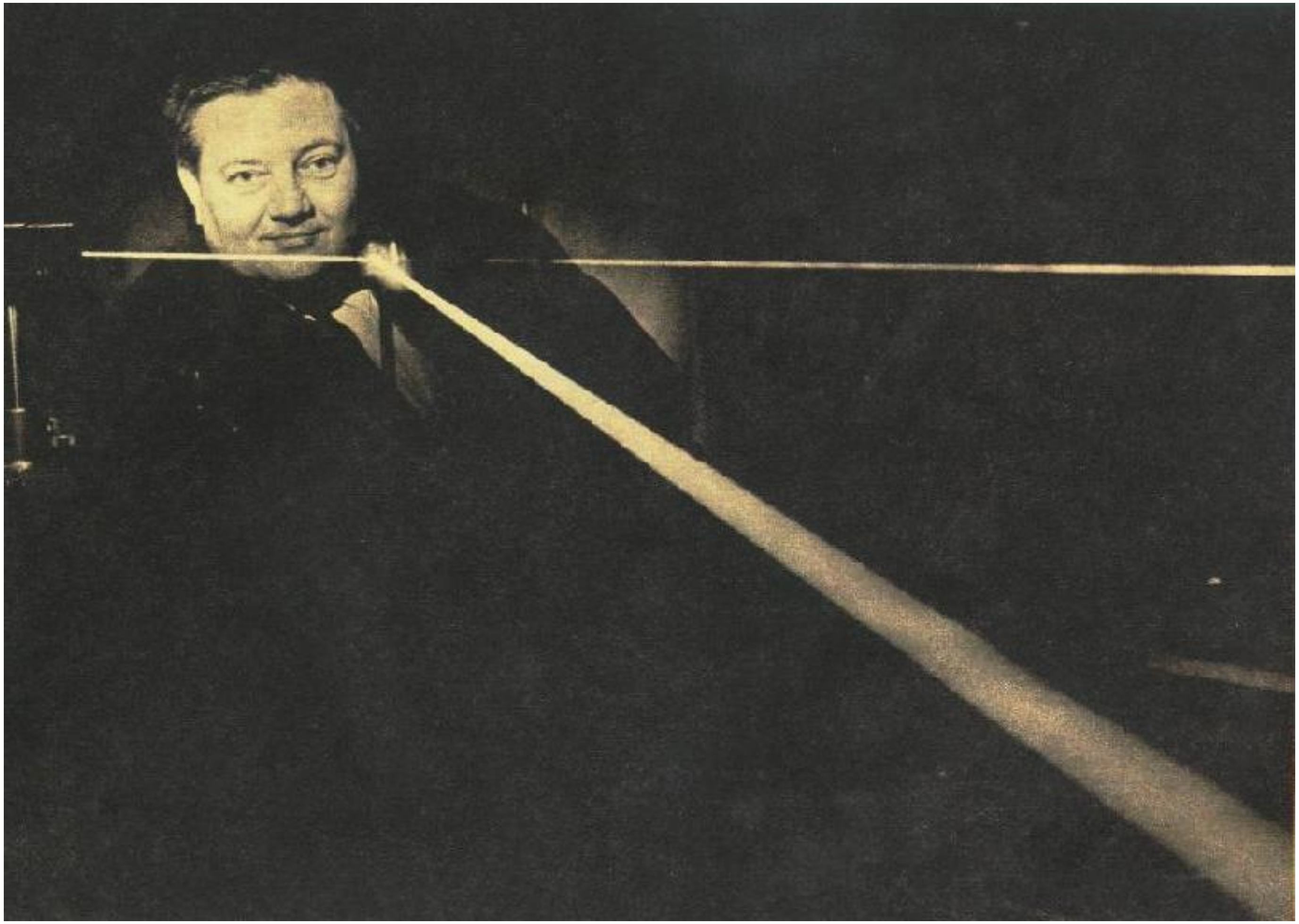
A probability space consists of: a sample space E, an event space F and a **probability measure** P.

E is the set of results of the random experiment, F is a family of subsets of E such that

$$E \in F$$
$$A \in F \Rightarrow \overline{A} \in F$$
$$A_i \in F, i \in \text{countable set} \Rightarrow \bigcup_i A_i \in F$$

P is a function from F into [0,1] such that:

$$P(A) \geq 0 \quad , \forall A \in F$$
$$P(E) = 1$$
$$P(A \cup B) = P(A) + P(B), \quad \forall A, B \in F \text{ disjoints}$$
$$\text{If } A_1 \supset A_2 \supset \ldots \text{tends to } \varnothing, \mathrm{P(A_i)} \to 0$$



E. Jaynes (1922-1998)

Thinking Robot

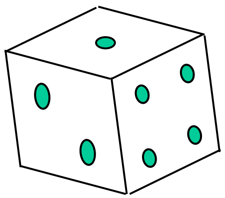
How to build a thinking robot ?

Based on the works of Polya e Cox, Jaynes assigns a **plausibility** to each proposition and demonstrates that a consistent inductive logic must obey the rules of **Probability Calculus**.

In this context, probabilities are assigned not to sets of a sigma field but to propositions.

Random Variables

A random variable assigns a numeric value to each experiment.



discrete

random variables

continuous

Discrete Random Variables

How to characterize a discrete random variable ?

Probability function

$$P(k) = \Pr\{x = k\}$$

Properties

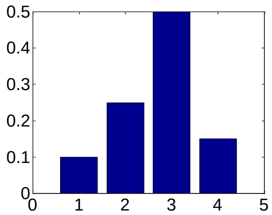
$$P(k) \geq 0, \ \forall k$$

$$\sum_{k=1}^{N} P(k) = 1$$

Ex:

$S = \{1,2,3,4\}, \quad P(1) = .1, P(2) = .3, P(3) = .4, P(4) = .2$

Realizations: 2 3 2 3 3 3 3 4 2 2 4 1 3 2 3 2 1 3 4 3

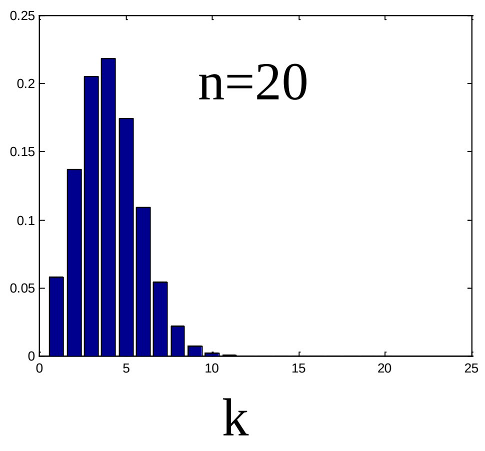


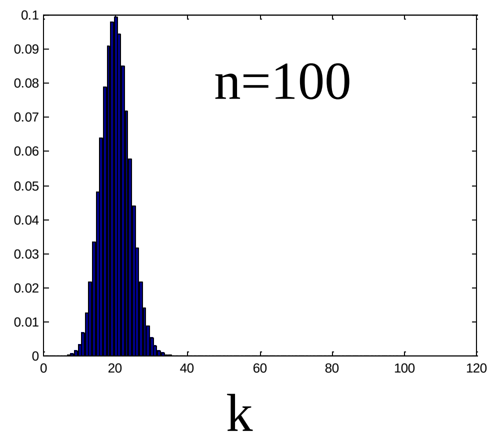
Note: the same symbol will be used to denote the random variable a realization. Different symbols (e.g., capital and lower letters) could be used to make this difference more clear.

Binomial Distribution

It answers the following problem: what is the probability of an event A being observed k times in n random experiments ?

$$P(k) = \binom{n}{k} \alpha^k (1-\alpha)^{n-k} \qquad \alpha = P(A)$$





Continuous Random Variables

How to characterize a continuous random variable x ?

Probability density function, p

$$P\{x \le x_o\} = \int_{-\infty}^{x_0} p(x)\, dx$$

Properties

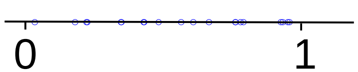
$$p(x) \ge 0, \ \forall x \in R^n$$

$$\int p(x) dx = 1$$

Ex:

$$S = R, \qquad p(x) = 1, x \in [0,1[, p(x) = 0, \text{otherwise}$$

0.8381 0.0196 0.6813 0.3795 0.8318 0.5028 0.7095 0.4289 0.3046
0.1897 0.1934 0.6822 0.3028 0.5417 0.1509 0.6979 0.3784 0.8600
0.8537 0.5936



Normal Distribution

Probability density function

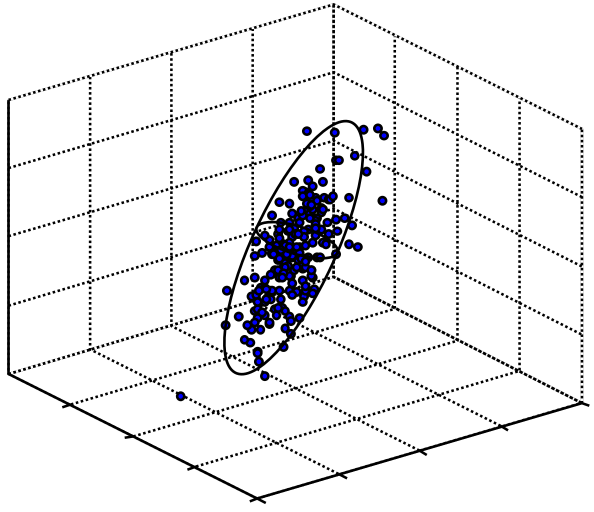
$$N(x;\mu,R) = \frac{1}{(2\pi)^{n/2}|R|^{1/2}} e^{-\frac{1}{2}(x-\mu)' R^{-1} (x-\mu)}$$

μ mean vector

R covariance matrix

The level surfaces in R^n are ellipsoids centered in μ and with axis

$\alpha\sqrt{\lambda_i} v_i$ where λ_i, v_i are eigen values and eigen vectors of R ($\| v_i \| = 1$)



Joint Distribution

The joint distribution of $x_1,\ldots, x_N$, is defined on the set of values of the sequence, being characterized by

Probability function	$P(x_1,\ldots, x_N)$	(discrete variables)
Probability density function	$p(x_1,\ldots, x_N)$	(continuous variables)

Marginalization

$$p(x_1) = \sum_{x_2} P(x_1, x_2)$$

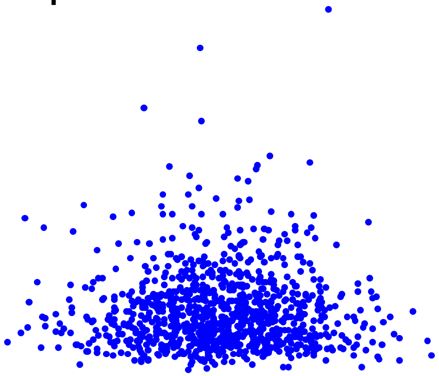
$$p(x_2) = \sum_{x_1} P(x_1, x_2)$$

Independence

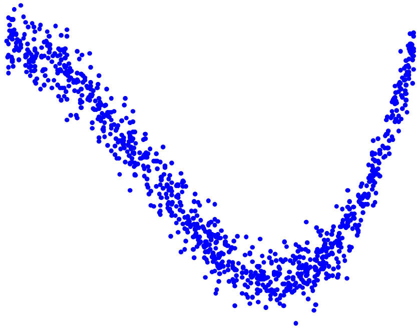
Def: the r.v. $x_1,..., x_N$ are independent if and only if

$$p(x_1,...,x_N) = \prod_i p(x_i)$$

independent variables



dependent variables



Covariances: $R = \begin{bmatrix} 1 & 0 \\ 0 & .37 \end{bmatrix}$ $R = \begin{bmatrix} 8.3 & -1.6 \\ -1.6 & 1.1 \end{bmatrix}$

Note: independent r.v. are converted into dependent ones by applying a non diagonal linear transformation

Correlation

Def: $x_1,\ldots, x_N$ are correlated r.v. if their covariance matrix is non diagonal.

Notes:

- independent r.v. are always uncorrelated. The converse is not true.
- given n r.vs. it is possible to decorrelate them by applying a suitable linear transformation (e.g. KLT or PCA)

Conditional Probabilities

Definition (conditional probability):

$$P(x \mid y) = P(x,y)/P(y) \quad \text{if } P(y) \neq 0$$

P(x|y) is interpreted as the probability of occurring x knowing that y occurred.

Note: if x,y are continuous random variables the conditional probability density function p(x|y) is defined in an analogous way .

Expectation

Definition (Expectation): Let $f: S \rightarrow R^n$

$$E\{f(x)\} = \sum_x f(x)P(x) \qquad \text{(x - discrete r.v.)}$$

$$E\{f(x)\} = \int f(x)p(x)dx \qquad \text{(x - continuous r.v.)}$$

Relationship with the aritmetic mean:

$$E\{f(x)\} = \lim_N \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

$x_1, x_2, \ldots.$ are realization of x

Definition (*mean and covariance matrix):* Let x be a random variable

mean $\mu = E\{x\}$ covariance matrix: $R = E\{(x-\mu)(x-\mu)'\}$

Properties of the Covariance Matrix

A is a covariance matrix if and only if it is a square matrix, symmetric and semi definite positive.

Other properties:

- the eigen values of a covariance matrix are non negative.
- $$R = \sum_{i=1}^{m} \lambda_i v_i v_i' \qquad |R| = \prod_{i=1}^{m} \lambda_i$$

  $\lambda_i, v_i \quad (\| v_i \| = 1)$ eigen values and eigen vectors

Properties of Normal Distribution

- If x_1, ..., x_n, are r.v. with normal distribution, any subset of variables x_{p1}, ..., x_{pm} are also r.v. with normal distribution.
- Given a r.v. $x \sim N(\bar{x}, R)$ the distribution of y=Ax+b is $N(\bar{y}, R_{yy})$

$$\bar{y} = A\bar{x} + b, \quad R_{yy} = ARA'$$

- Given 2 variables $x \sim N(\bar{x}, R_{xx}), y \sim N(\bar{y}, R_{yy})$ then

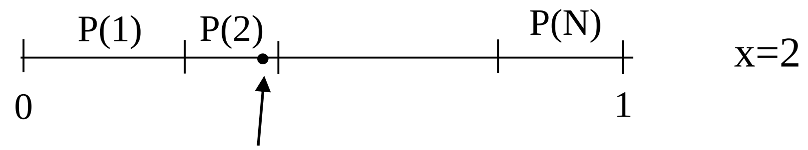
$$x + y \sim N(\bar{x} + \bar{y}, P) \qquad P = R_{xx} + R_{yy} \text{ if x, y are independen t}$$

Generation of Random Values

Discrete variables:

•Split [0,1[interval into subintervals of length P(i).

•generate a random value with uniform distribution in [0,1[. The value of x is the index of the subinterval which was selected.



Continuous variables:

•specific algorithms for some distributions

•Metropolis algorithm

•importance sampling

•Gibbs sampler

Metropolis Algorithm

How to generate random values with a given distribution ?

Metropolis Algorithm:

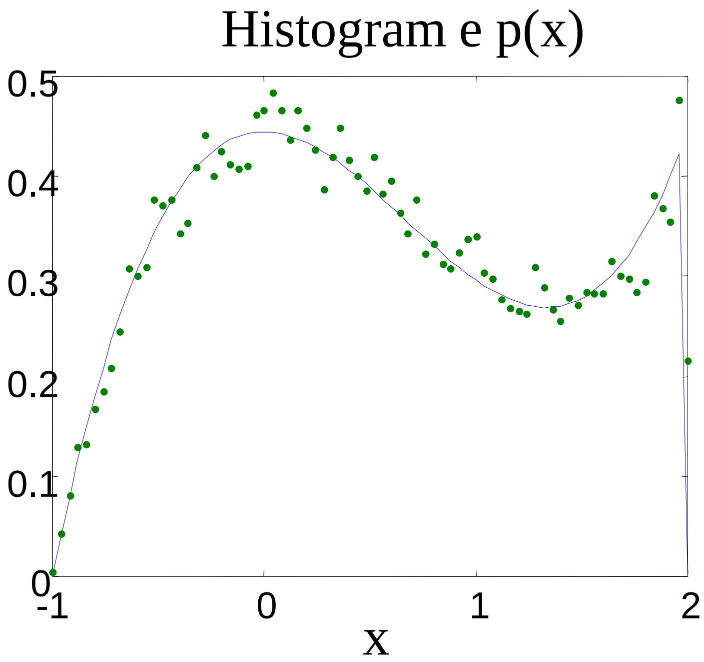
- move x randomly
- accept the new value *x'* with probability

$$P = \min(1, p(x') / p(x))$$

otherwise make *x'*=*x*

Example

$p(x) = (2x^3 - 4x^2 + 6)/13.5$



Importance Sampling

It is used to compute expected values when it is difficult to generate random values with the true distribution p(x) but it is possible to generate samples with an auxiliary distribution q(x).

Algorithm:

- generate n independent realizations $x_i \sim q(x)$
- assign a weight to each realization (importance) $w_i = p(x_i) / q(x_i)$.

Expectation: $$E\{f(x)\} \approx \frac{1}{n} \sum_{i=1}^{n} w_i f(x_i)$$

Note:

- $q(x) > 0, \ \forall x : p(x) > 0$
- poor performance in high dimension spaces

Example

We wish to estimate 2 moments of a distribution N(0,1) using importance sampling.

We considered n=100; 100 estimation experiments were performed

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} w_i x_i \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} w_i x_i^2$$

	$\overline{m}_1$	σ_{m_1}	$\overline{m}_2$	σ_{m_2}	
q=N(0,.25)	-0.0194	0.2801	0.6549	0.4575	tail mismatch
q=N(0,1)	0.0034	0.0951	0.9977	0.1540	
q=N(0,4)	0.0170	0.0927	0.9959	0.0611	

Gibbs Sampler

Problem: generate random values with a known distribution

$$P(x_1,\ldots,x_N)$$

Algorithm:

begin

- generate x_1 with distribution $P(x_1 / x_2, x_3, \ldots, x_N)$
- generate x_2 with distribution $P(x_2 / x_1, x_3, \ldots, x_N)$

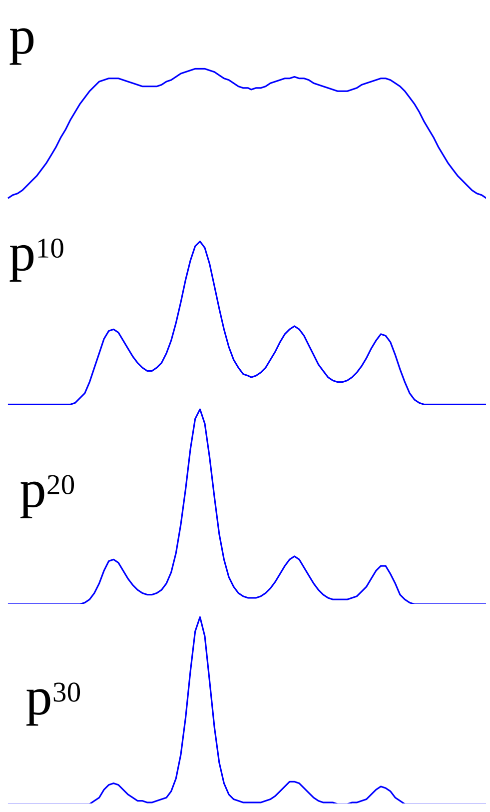
- generate x_N with distribution $P(x_N / x_1, x_2, \ldots, x_{N-1})$

repeat

This algorithm generates a Markov with asymptotic distribution

$$P(x_1,\ldots,x_N)$$

Optimization with the Gibbs Sampler



Generate realizations of a r.v. x with distribution $p(x)^a$

Change a until a dominant mode is observed

In the limit the algorithm will only generate values which maximize p.

Difficulty: there are no optimal criteria for the evolution of a

Problems

x	y	P
1	1	.1
1	2	.2
1	3	.1
2	1	.3
2	2	.1
2	3	.2

1. Given a distribution P(x,y) defined by:
   i) P(x) ii) P(y) iii) P(x|y) iv) E{x} v) E{y} vi) E{x+y} vii) E{xy}

2. The meaning of variables x,y,z is the following : x-the is gas in the tank; y – battery is OK; z- motor starts at first attempt. Define a probability distribution for these variables.

3. A random variable x~N(0,R) has an uncertainty ellipsoid with semi axis [3 1], [-.2 .6]. Compute the covariance matrix R knowing that $E\{x_1^2\}=1$.

4. We known that a bridge falls with probability .8 if the main structure elements break and this happens with probability .001. Which is the break probability knowing that the bridge has fallen? Discuss if this problem can be solved.

5. Three prisoners A, B, C are in separate cells. One is going to be released and the other two will be condemned to die. Prisoner A asks the jailer to deliver a farewell letter to one of the other prisoners which will be condemned. The next day the jailer tells him that he delivered the letter to prisoner B. What is the probability of A being set free before and after the jailer answer?

Work

Let x be a random variable with distribution N(0,1). Determine in an exact or approximate way:

$E\{x^2\}$, $E\{x^4\}$, $E\{\cos(x)\}$, $E\{\tan(x)\}$, $E\{\tan^{-1}(x)\}$

Bibliography

E. T. Jaynes, Probability Theory: the Logic of Science, 1995.

J. Marques, Reconhecimento de Padrões. Métodos Estatísticos e Neuronais, IST Press, 1999.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6:721-741, 1984.