Elementi di probabilitá e Statistica

R. Basili

Corso di *Web Mining* e *Retrieval*
a.a. 2008-9

March 10, 2010

Outline

Outline

Elementary Probability Theory

Outline

- Sample Space
- Probability Measures
- Independence
- Conditional Probabilities
- Bayesian Inversion
- Partitions

Sample Space

Sample Space

The sample space is a set of elementary outcomes. An event is a subset of the sample space. Sample spaces are often denoted by Ω and events are often called $A, B, C, \ldots$.

Sample Space

Sample Space

The sample space is a set of elementary outcomes. An event is a subset of the sample space. Sample spaces are often denoted by Ω and events are often called $A, B, C, \ldots$.

Example

Dado. $\Omega = \{'1', '2', \ldots, '6'\}$

- Un tiro del dado in cui si ottiene '1' da' luogo all'evento $\{'1'\}$:
- "Il risultato é meno di 4" consiste nell'evento: $\{'1', '2', '3'\}$
- il numero totale di eventi coincide con il numero totale di sottoinsiemi di Ω.
- Nota: $'1' \neq \{'1'\}$

Probability Measures

Una funzione P a valori reali sullo spazio degli eventi 2^Ω e' una funzione di probabilitá **sse**:

Axioms

1) $0 \leq P(A) \leq 1 \qquad \forall A \in 2^\Omega$

Probability Measures

Una funzione P a valori reali sullo spazio degli eventi 2^{Ω} e' una funzione di probabilitá **sse**:

Axioms

1) $0 \leq P(A) \leq 1 \qquad \forall A \in 2^{\Omega}$

2) $P(\Omega) = 1$

Probability Measures

Una funzione P a valori reali sullo spazio degli eventi 2^{Ω} e' una funzione di probabilitá **sse**:

Axioms

1) $0 \leq P(A) \leq 1 \qquad \forall A \in 2^{\Omega}$
2) $P(\Omega) = 1$
3) $\forall A, B \in 2^{\Omega} \quad (A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B))$

Probability Measures

Una funzione P a valori reali sullo spazio degli eventi 2^{Ω} e' una funzione di probabilitá **sse**:

Axioms

1) $0 \leq P(A) \leq 1 \qquad \forall A \in 2^{\Omega}$
2) $P(\Omega) = 1$
3) $\forall A, B \in 2^{\Omega} \quad (A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B))$

Esempio di Ω: "Il risultato di un tiro di dado e' minore di 7".

Probability Measures

Data la funzione $P : 2^{\Omega} \to [0,1]$

Consequences

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

Probability Measures

Data la funzione $P : 2^{\Omega} \rightarrow [0,1]$

Consequences

- $P(A \setminus B) = P(A) - P(A \cap B)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$

Probability Measures

Data la funzione $P : 2^{\Omega} \to [0, 1]$

Consequences

- $P(A \setminus B) = P(A) - P(A \cap B)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(\bar{A}) = 1 - P(A)$

Probability Measures

Data la funzione $P : 2^{\Omega} \rightarrow [0, 1]$

Consequences

- $P(A \setminus B) = P(A) - P(A \cap B)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(\bar{A}) = 1 - P(A)$
- $P(\emptyset) = 0$

Probability Measures

Data la funzione $P : 2^{\Omega} \rightarrow [0,1]$

Consequences

- $P(A \setminus B) = P(A) - P(A \cap B)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(\bar{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Probability Measures

La situazione in cui due eventi A e B occorrono insieme ammette una probabilita' pari a $P(A \cap B)$.
La conoscenza di un evento B cambia la nostra aspettativa (e quindi la probabilita') di un secondo evento A. Quando questo non avviene allora i due eventi si dicono *indipendenti*.

Independence

A is independent from $B \iff P(A \cap B) = P(A) \cdot P(B)$

Probability Measures

Conditional Probabilities

The probability of A given an event B is written as $P(A|B)$ and it is given by:

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

Probability Measures

Conditional Probabilities

The probability of A given an event B is written as $P(A|B)$ and it is given by:
$P(A|B) = \frac{P(A \cap B)}{P(B)}$

Formula like $P(A)$ are often called *priors* or *a priori* probabilities as nothing is known about A, while $P(A|B)$ are called *posteriors* (or *a posteriori*) probabilities, as B adds information to A.

Probability Measures

Conditional Probabilities

The probability of A given an event B is written as $P(A|B)$ and it is given by:
$P(A|B) = \frac{P(A \cap B)}{P(B)}$

Formula like $P(A)$ are often called *priors* or *a priori* probabilities as nothing is known about A, while $P(A|B)$ are called *posteriors* (or *a posteriori*) probabilities, as B adds information to A.

Note that:

- $P(A|A) = 1$, $P(A|\bar{A}) = 0$
- If A and B are independent: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$

Bayesian Inversion

The probability $P(A|B)$ can be more difficult to estimate than $P(B|A)$. A way to invert the conditional probability $P(A|B)$ is known as **Bayes rule**:

Bayesian Inversion

The probability $P(A|B)$ can be more difficult to estimate than $P(B|A)$. A way to invert the conditional probability $P(A|B)$ is known as **Bayes rule**:

Bayesian Inversion

As $P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$
then it follows that:
$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$

Bayesian Inversion

The probability $P(A|B)$ can be more difficult to estimate than $P(B|A)$. A way to invert the conditional probability $P(A|B)$ is known as **Bayes rule**:

Bayesian Inversion

As $P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$
then it follows that:
$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$
In the Bayes formula

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

the posteriors $P(B|A)$ are used instead of $P(A|B)$.

Partitions

When a partion in n events $A_i \quad (i = 1, \ldots, n)$ is available for Ω, i.e.

$$\begin{cases} \quad \Omega & = & \bigcup_{i=1}^{n} A_i \\ \forall i \neq j \quad A_i \cap A_j & = & \emptyset \end{cases}$$

Partitions

When a partion in n events $A_i \quad (i = 1, \ldots, n)$ is available for Ω, i.e.

$$\begin{cases} & \Omega = \bigcup_{i=1}^{n} A_i \\ \forall i \neq j & A_i \cap A_j = \emptyset \end{cases}$$

then:

$$P(B) = P(B \cap \Omega) = P(B \cap (\bigcup_{1}^{n} A_i)) = P(\bigcup_{1}^{n} (B \cap A_i)) =$$

Partitions

When a partion in n events $A_i \quad (i = 1, \ldots, n)$ is available for Ω, i.e.

$$\begin{cases} & \Omega = \bigcup_{i=1}^{n} A_i \\ \forall i \neq j & A_i \cap A_j = \emptyset \end{cases}$$

then:

$$P(B) = P(B \cap \Omega) = P(B \cap (\bigcup_{1}^{n} A_i)) = P(\bigcup_{1}^{n} (B \cap A_i)) =$$

$$= \sum_{i}^{n} P(B \cap A_i) = \sum_{i}^{n} P(B|A_i) P(A_i)$$

Stochastic Variables

Stochastic Variables

- Distribution Functions
- Probability Measures
- Discrete and Continuous Stochastic Variables
- Frequency Function
- Expectation Value
- Variance
- Two dimensional Stochastic Variables

Stochastic Variable

Sample space of a stochastic variable

A stochastic or random variable ξ is a function from a sample space Ω to the set of real numbers R.
Thus if $u \in \Omega$ then $\xi(u) \in R$.

Stochastic Variable

Sample space of a stochastic variable

A stochastic or random variable ξ is a function from a sample space Ω to the set of real numbers R.
Thus if $u \in \Omega$ then $\xi(u) \in R$.

In the "Dado" example, the image of ξ is $\{1,...,6\}$, and $\xi('1') = 1,...,\xi('6') = 6$.

Stochastic Variables

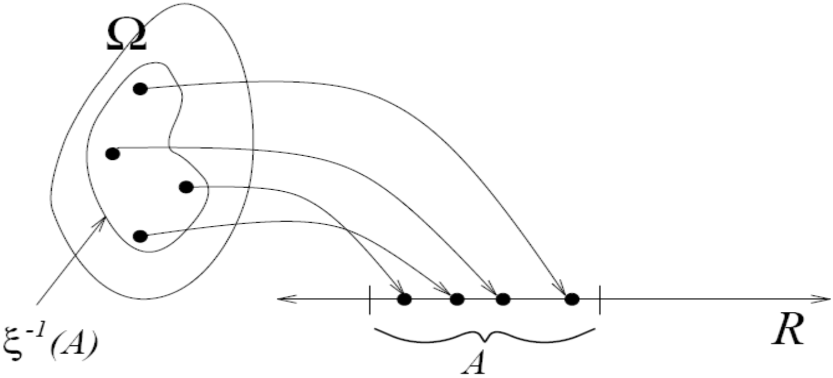


Figure 1.3: $P(\xi \in A)$ is defined through $P(\xi^{-1}(A)) = P(\{u : \xi(u) \in A\})$.

Stochastic Variables

Sample space of a stochastic variable

The image of the sample space Ω in R under the random variable ξ, i.e. the range of ξ, is called *the sample space of the stochastic variable* ξ and is denoted by Ω_ξ.
In short, $\Omega_\xi = \xi(\Omega)$

Distribution Function

Distribution Function

Let A be a subset of R and consider the inverse image of A under ξ, i.e. $\xi^{-1}(A) = \{u \in \Omega : \xi(u) \in A\} \subseteq \Omega$.

Distribution Function

Distribution Function

Let A be a subset of R and consider the inverse image of A under ξ, i.e. $\xi^{-1}(A) = \{u \in \Omega : \xi(u) \in A\} \subseteq \Omega$.
We will let $P(\xi \in A)$ denote the probability of this set, i.e. $P(\xi^{-1}(A)) = P(\{u : \xi(u) \in A\}) = P(\xi \in A)$.

Distribution Function

Distribution Function

Let A be a subset of R and consider the inverse image of A under ξ, i.e. $\xi^{-1}(A) = \{u \in \Omega : \xi(u) \in A\} \subseteq \Omega$.
We will let $P(\xi \in A)$ denote the probability of this set, i.e. $P(\xi^{-1}(A)) = P(\{u : \xi(u) \in A\}) = P(\xi \in A)$.
If A is the interval $(-\infty, x]$ then the real-valued function F denoted by
$F(x) = P(\{u : \xi(u) \leq x\}) = P(\xi \leq x) \qquad \forall x \in R$
is called the *distribution function of the random variable* ξ.
Sometimes F is denoted F_ξ to indicate that it is the distribution function of the particular random variable ξ.

Stochastic Variables

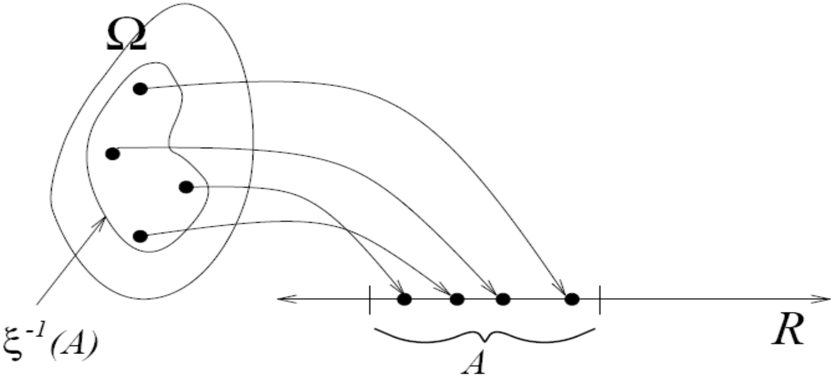


Figure 1.3: $P(\xi \in A)$ is defined through $P(\xi^{-1}(A)) = P(\{u : \xi(u) \in A\})$.

Distribution Function

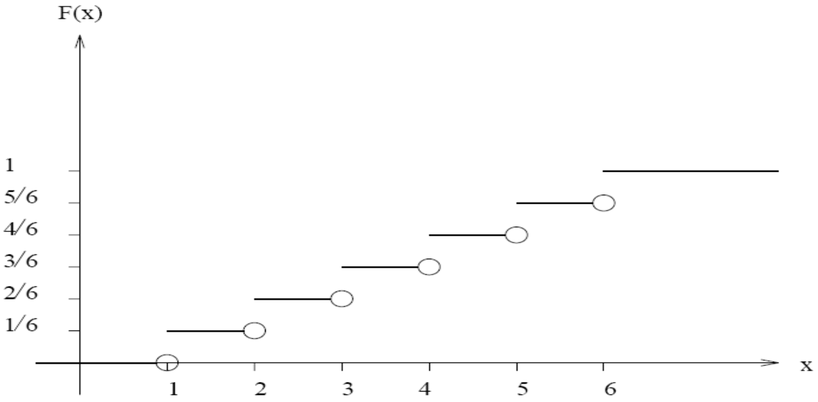


Figure 1.4: Fair die: Graph of the distribution function.

Frequency Function

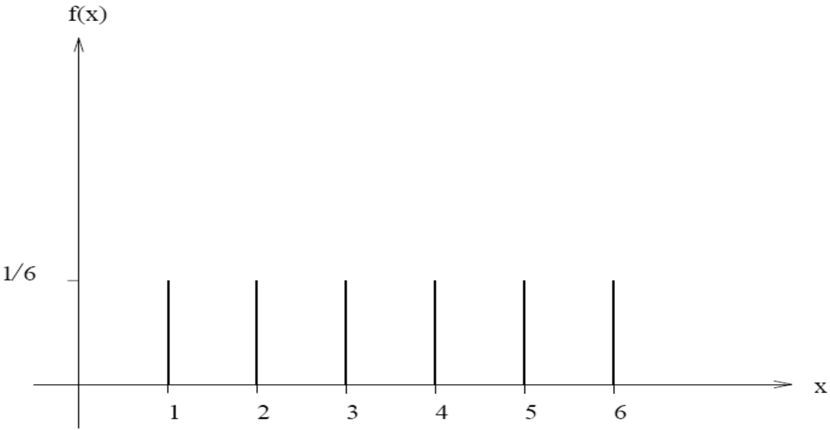
Frequency Function

Another way of seeing the distribution of a random variable is through its *frequency function*, f, given by:

- Discrete Case: $f(x) = P(\xi = x)$
- Continuous Case: $f(x) = F'(x) = \frac{dF(x)}{dx}$

In order to explicit the reference to the random varable ξ f is often denoted as f_ξ.

Frequency Function



Frequency Function and Probabilities

Frequency and Distribution Function

The probability distribution of a random variable ξ can be computed from its frequency function f_ξ as follows:

- Discrete Case: $P(\xi \in A) = \sum_{x \in A} f_\xi(x)$
- Continuous Case: $P(\xi \in A) = \int_A f_\xi(x)dx$

Frequency Function and Probabilities

Consequences

- Discrete Case:

$$P(\Omega_\xi) = \sum_{x \in \Omega_\xi} f_\xi(x) = 1$$

- Continuous Case:

$$P(\Omega_\xi) = \int_{-\infty}^{+\infty} f_\xi(x)dx = 1$$

Expectation

Expectation or Mean value

A way to summaize the distribution of a random variable is through its *expectation value*, or statistical mean, $E[\xi]$, given by:

- Discrete Case:

$$E[\xi] = \sum_{x \in \Omega_\xi} x \cdot f_\xi(x) = \sum_i x_i \cdot f_\xi(x_i)$$

- Continuous Case:

$$E[\xi] = \int_{-\infty}^{+\infty} x \cdot f_\xi(x) dx$$

In both cases $E[\xi]$ is often denoted by μ.

Variance

Variance

A second aspect is to express how much is the mean value of a random variable is representative of the entire distribution. This is given by the notion of standard deviation or, more commonly, the *variance* $Var[\xi]$:

Variance

Variance

A second aspect is to express how much is the mean value of a random variable is representative of the entire distribution. This is given by the notion of standard deviation or, more commonly, the *variance* $Var[\xi]$:

- Discrete Case:

$$Var[\xi] = \sum_{x \in \Omega_\xi} (x-\mu)^2 \cdot f_\xi(x) = \sum_i (x_i - \mu)^2 \cdot f_\xi(x_i)$$

Variance

Variance

A second aspect is to express how much is the mean value of a random variable is representative of the entire distribution. This is given by the notion of standard deviation or, more commonly, the *variance* $Var[\xi]$:

Variance

Variance

A second aspect is to express how much is the mean value of a random variable is representative of the entire distribution. This is given by the notion of standard deviation or, more commonly, the *variance* $Var[\xi]$:

- Continuous Case:

$$Var[\xi] = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f_{\xi}(x)dx$$

Variance

Variance

It is clearly true that $Var[\xi] = E[(\xi - \mu)^2]$.
The variance of a variable ξ is often denoted by σ^2, whereas σ denotes the *standard deviation*.

Variance

Variance

It is clearly true that $Var[\xi] = E[(\xi - \mu)^2]$.
The variance of a variable ξ is often denoted by σ^2, whereas σ denotes the *standard deviation*.

In the "Dado" example obviously follows:

- $E[\xi] = \sum_{i=1}^{6} \frac{1}{6} \cdot i = \frac{6 \cdot (6+1)}{2} \cdot \frac{1}{6} = \frac{7}{2}$
- $Var[\xi] = \sum_{i=1}^{6} (i - \frac{7}{2})^2 \cdot \frac{1}{6} = \frac{35}{12}$

Frequency Function

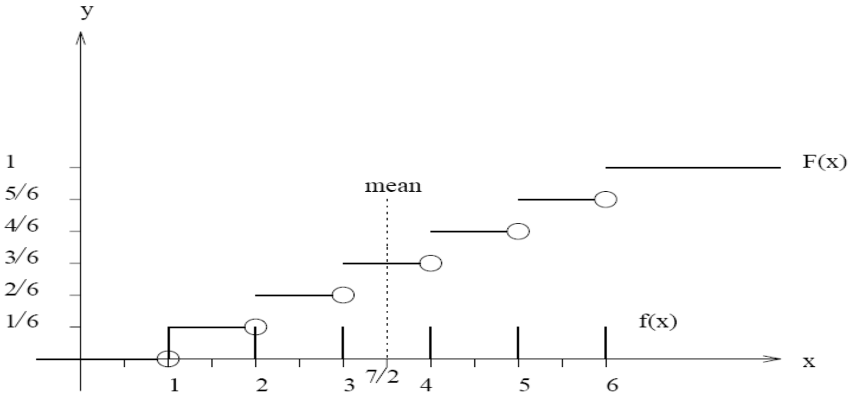


Figure 1.6: Fair die: Expectation value (mean)

Multiple Random Variables

Multiple Variable

Let ξ and η be two random variables defined on the same sample space Ω.

Multiple Random Variables

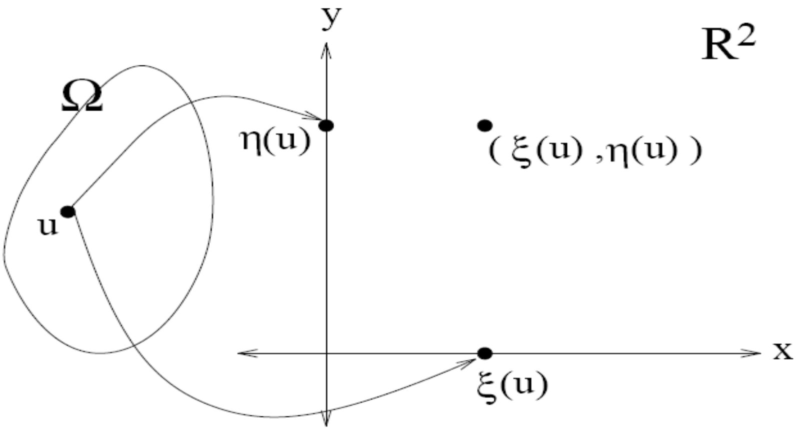
Multiple Variable

Let ξ and η be two random variables defined on the same sample space Ω.

Then (ξ,η) is a two-dimensional random variable from Ω to $\Omega_{(\xi,\eta)} = \{(\xi(u),\eta(u)) : u \in \Omega\} \subseteq R^2$.

Here $R^2 = R \times R$ is the Cartesian product of the set of real numbers R with itself.

Multiple Random Variables



Multiple Random Variables

Generalizations: Discrete Case

A two-dimensional random variable (ξ, η) is discrete **iff** $\Omega_{(\xi,\eta)}$ is finite or countable.

Multiple Random Variables

Generalizations: Discrete Case

A two-dimensional random variable (ξ, η) is discrete **iff** $\Omega_{(\xi,\eta)}$ is finite or countable.

The frequency function f of (ξ, η) is then defined by:

$$f(x,y) = P(\xi = x, \eta = y) = P((\xi, \eta) = (x,y)) \qquad \forall (x,y) \in R^2$$

Multiple Random Variables

Generalizations: Discrete Case

A two-dimensional random variable (ξ, η) is discrete **iff** $\Omega_{(\xi,\eta)}$ is finite or countable.

The frequency function f of (ξ, η) is then defined by:

$f(x,y) = P(\xi = x, \eta = y) = P((\xi, \eta) = (x,y)) \qquad \forall (x,y) \in R^2$

Furthermore:

$\forall A \subseteq \Omega_{(\xi,\eta)}$

$$P(A) = P((\xi, \eta) \in A) = \sum_{(x,y) \in A} f(x,y)$$

Multiple Random Variables

Marginal distributions

We can recover the frequency functions of either of the individual variables by summing or integrating over the other.

Multiple Random Variables

Marginal distributions

We can recover the frequency functions of either of the individual variables by summing or integrating over the other. If (ξ, η) is discrete:

$$f_\xi(x) = \sum_{y \in \Omega_\eta} f(x, y)$$

$$f_\eta(y) = \sum_{x \in \Omega_\xi} f(x, y)$$

Multiple Random Variables

Marginal distributions

We can recover the frequency functions of either of the individual variables by summing or integrating over the other.
If (ξ, η) is discrete:

$$f_\xi(x) = \sum_{y \in \Omega_\eta} f(x, y)$$

$$f_\eta(y) = \sum_{x \in \Omega_\xi} f(x, y)$$

In this context f_ξ and f_η are often referred to as the marginal distributions of ξ and η respectively.

Functions over Multiple Random Variables

Special functions $\Psi(u)$ of two random variables (i.e. $\Psi(u) = g(\xi(u), \eta(u))$ can be easily derived from the single variable case.

Mean

The *expectation value* of $g(\xi, \eta)$ when (ξ, η) is discrete, is given by:

$$E[g(\xi, \eta)] = \sum_{(x,y) \in \Omega_{(\xi,\eta)}} g(x,y) \cdot f_{(\xi,\eta)}(x,y).$$

Expectation (Continuous Case)

$E[g(\xi, \eta)] = \int_{-\infty}^{+\infty} g(x,y) \cdot f_{(\xi,\eta)}(x,y) dx\, dy$ if (ξ, η) is continuous.

Stochastic or Random Processes

Random Processes

A *stochastic* or *random process* is a sequence $\xi_1, \xi_2, \ldots \xi_n$ of random variables based on the same sample space Ω.

Stochastic or Random Processes

Random Processes

A *stochastic* or *random process* is a sequence $\xi_1, \xi_2, \ldots \xi_n$ of random variables based on the same sample space Ω.
The possible outcomes of the random variables are called the set of *possible states* of the process. The process will be said to be in state ξ_t at time t.

Stochastic or Random Processes

Random Processes

A *stochastic* or *random process* is a sequence $\xi_1, \xi_2, \ldots \xi_n$ of random variables based on the same sample space Ω.
The possible outcomes of the random variables are called the set of *possible states* of the process. The process will be said to be in state ξ_t at time t.

Independence

Note that the random variables are in general not independent (i.e. $P(\xi_{t+1}|\xi_t) \neq P(\xi_{t+1})$ in general). In fact, the interesting thing about stochastic processes is the dependence between the random variables ξ_{t+1} and ξ_t, for the different t.

Selected Probability Distributions

Useful Distribution

- Binomial Distribution
- Normal Distribution
- Other Distributions
- Distribution Tables
- Probability Measures

See them in (Krenn & Samuelsson, 1997)

References

Introduction to Probability

- (Krenn & Samuelsson, 1997), Brigitte Krenn, Christer Samuelsson, *The Linguist's Guide to Statistics Don't Panic*, Univ. of Saarlandes, 1997.
  URL:
  `http://nlp.stanford.edu/fsnlp/dontpanic.pdf`