Tecnológico de Monterrey

Statistics and Stochastic processes

ROBERTO ALEJANDRO CÁRDENAS OVANDO

Outline

- Data
- Probability
- Statistics
- Inference
- Modeling
- Stochastic models
- Stochastic processes
- Finite mixture models
- Markov models

Data

- Variables are random in some way
  - It represents an incompletely, measured variable
  - Sample drawn using random mechanisms
- Data into knowledge:
  - Probability
    - The study of random variables
  - Statistics
    - The discipline of using data samples to support claims about populations.
    - Based on probability
  - Computation
    - A tool well suited to quantitative analyses

Reproducible Research

- Replication
  - Validate findings
  - Some studies cannot be replicated (money/condition)
- Data -> Analytic data -> Reproducible research
- Existing database can be merged into new “mega databases”
- For every field there is a computational field of it

Types of Data Analysis Questions

COMPLEXITY

- Descriptive: First kind of approach, describe a set of data
- Exploratory: Find relationships you didn't know about. No generalizing
- Inferential: Small sample of data to say something about a bigger population
- Predictive: Use data from one object to predict another. No causality
- Causal: To find what happens to one variable when you change another
- Mechanistic: Understand the variables that lead to exact changes for an individual observation

Sources of data

- Census
  - Interested in people
  - Descriptive
- Convenience
  - Depends in how data are sampled
  - Descriptive, Inference and Prediction
  - Highly biased
  - Anecdotal
    - Small number of observations
    - Inaccurate

Sources of data

- Observational
  - Measure a group without replacement
  - Inference
- Randomized trial
  - Find a variable that changes other variables
  - Many subgroups without replacement
  - Each group has different conditions
  - Causal analysis
- Prediction study
  - Two data sets: training and test
  - Predictive

Sources of data - Study over time

- **Longitudinal**
  - It follows along time
  - Inferential and predictive
- **Retrospective**
  - First and last observation
  - Inferential
  - E.g. Outcome and exposure
- **Cross-sectional**
  - Taking samples from different types
  - Inferential
  - E.g. Wildtype vs condition

Probability

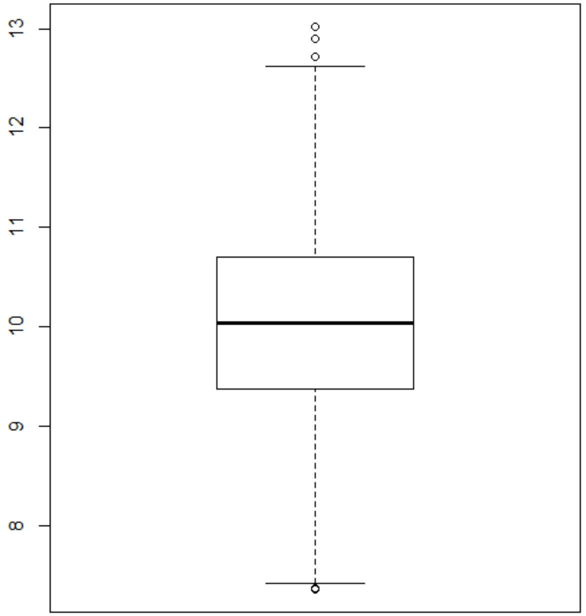
- All the important results are called Events (E)
- In a success or failure trial:
  - $P(E)$ is the probability of success
  - $P(\neg E)$ is the probability of failure
- Two approaches:
  - Frequentist – Depends on observations amount
  - Bayesian – Depends on degree of knowledge

Descriptive statistics

- A small set of parameters can summarize a large amount of data
- Three summary statistics
  - Median
  - Mean
  - Variance

Median

- The value at the center of a sorted dataset
- Value such that the set of values less than itself has a probability of 0.5



Sample mean

- Good description of a set of values

mean≠average

- Average: statistics to describe typical values
- Arithmetic mean is one type of average

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- At least 1 DOF to compute

Sample variance

- It describes the spread of data
- It is the squared deviation from the mean
  - Biased estimator

$$s_X^2 = \frac{1}{N}\sum_{i=i}^{N}(X_i - \mu)^2$$

  - Unbiased estimator

$$s_X^2 = \frac{1}{N-1}\sum_{i=i}^{N}(X_i - \mu)^2$$

- At least 2 DOF to compute

Probability density function (pdf)

- Also known as probability distribution
- It describes how often a value appears [Frequency]

$$P(\, a < X \leq b) = \int_{a}^{b} f\,(x)dx$$

- Histogram
  - Frequency of each value
- Probability mass function (pmf)
  - It describes a discrete random variable

$$P(X = a)$$

Cumulative distribution function

- The CDF is the function that maps values to their percentile rank in a distribution

$$P(X \leq x)$$

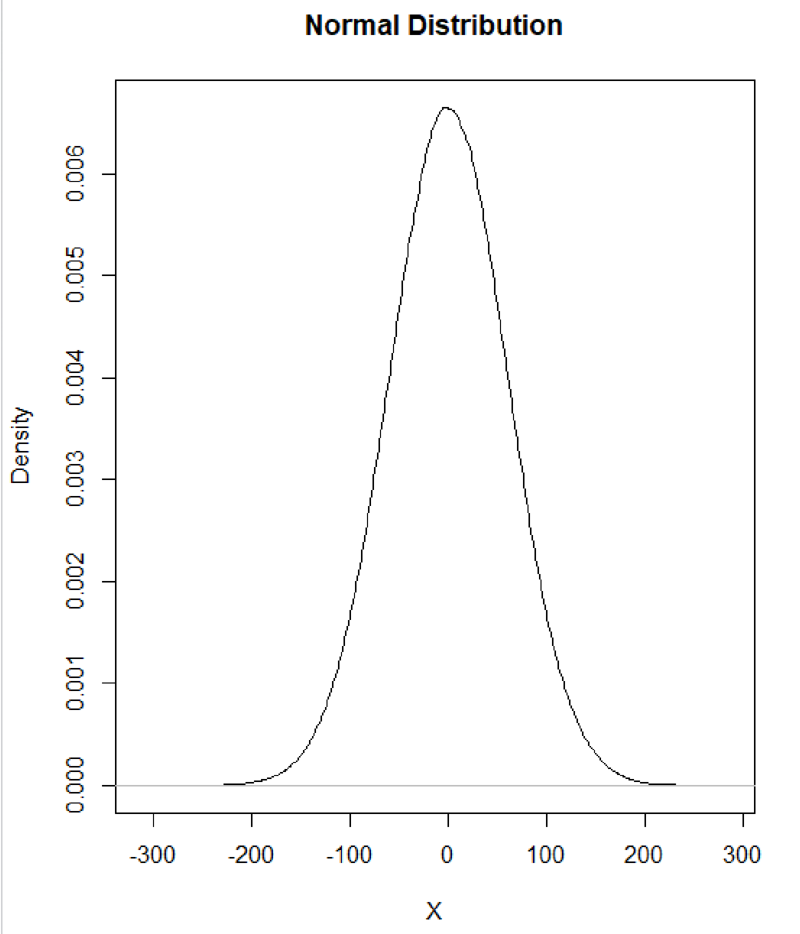
- The CDF is a function of X, where X is any value that might appear in the distribution

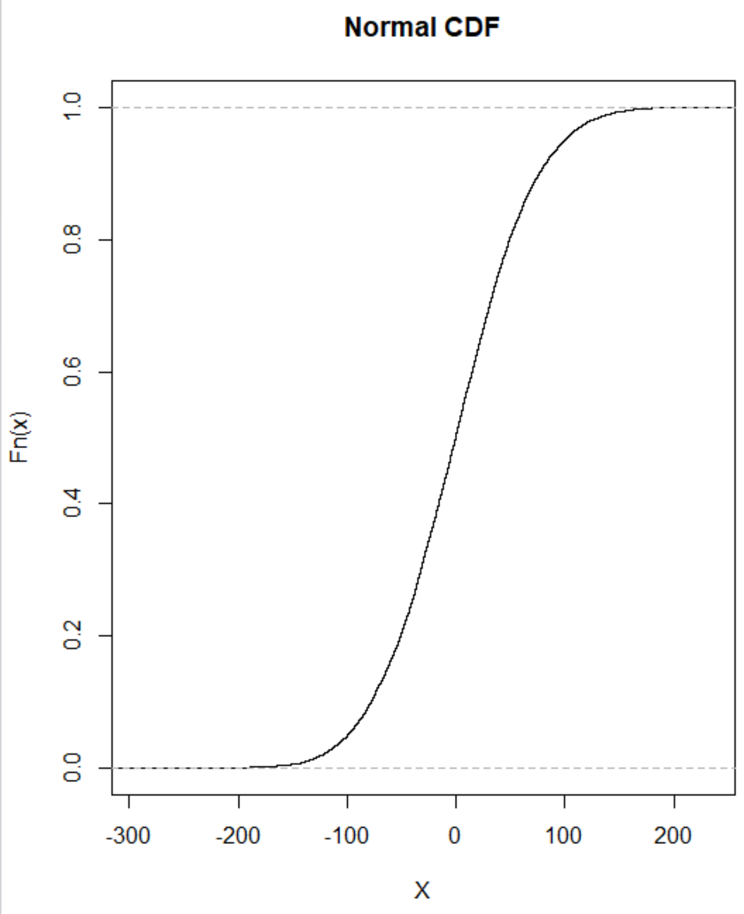
$$\lim_{X \to -\infty} cdf(X) = 0$$

$$\lim_{X \to \infty} cdf(X) = 1$$

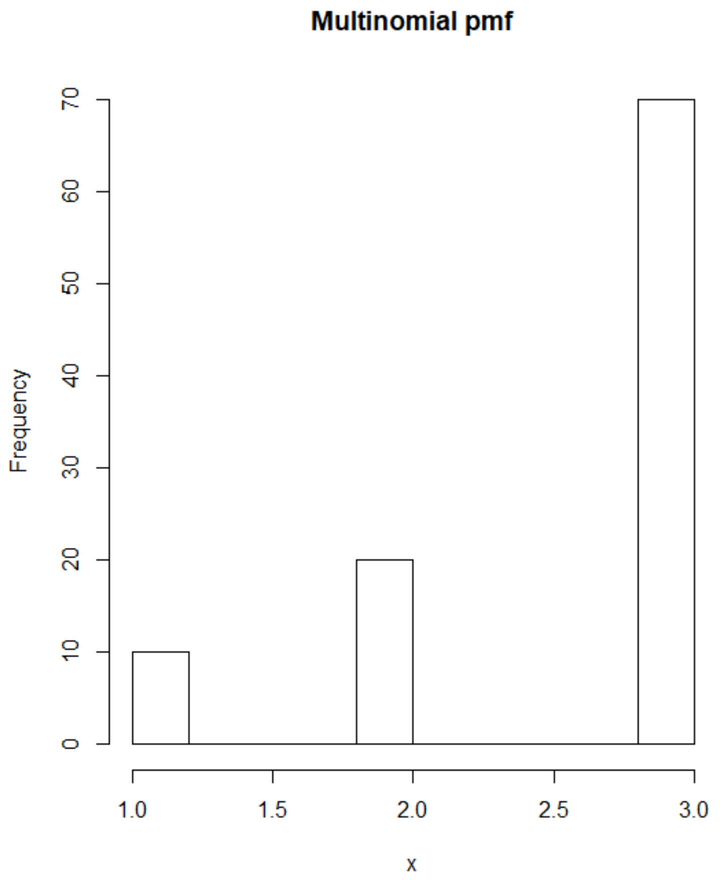
- Cumulative mass function (cmf)
  - It describes a discrete random variable

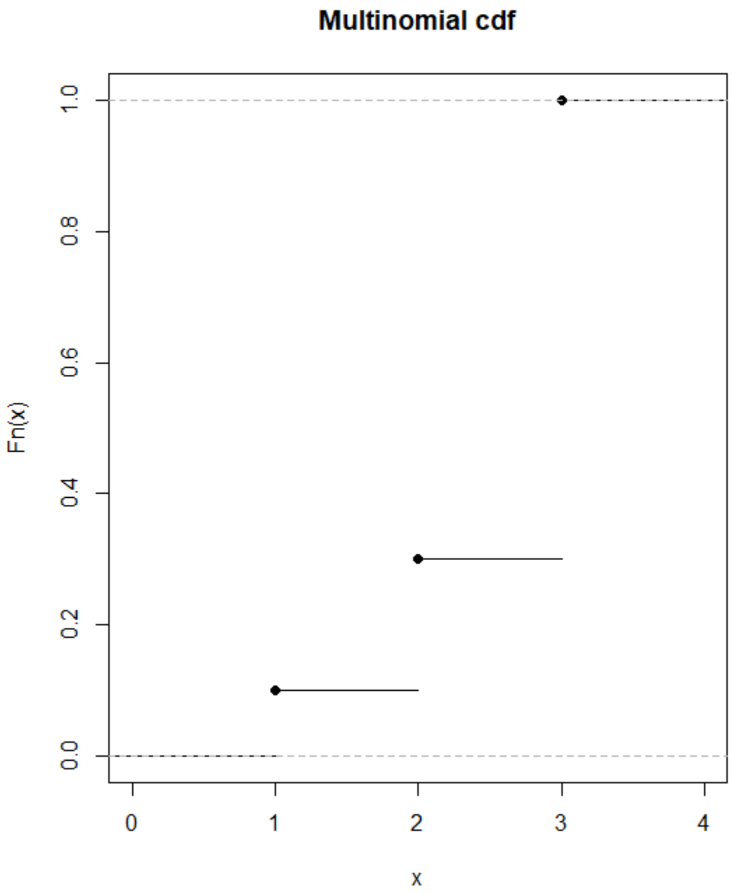
Example - Normal distribution





Example - Multinomial distribution





Law of large numbers

- The law of large numbers describes the result of performing the same experiment a large number of times

- Strong law of large numbers states that the sample average converges almost surely to the expected value

$$Average(X_{1:n}) \to \mu \qquad \text{when } n \to \infty$$

Central Limit Theorem

- This explains the prevalence of normal distribution in the natural world

- The characteristics we measure are the sum of a huge number of small effects
  - Therefore, the distribution tends to be normal

Example

- Binomial distribution
  - Probability of success: 0.3
  - Number of trials: 100
  - Number of observations

  - Binomial mean for one trial: p
  - Binomial variance for one trial: p(1-p)

Hypothesis testing

- The fundamental question we want to address is whether the effects are real or due to randomness
- Two steps:
  - Effect is significant, didn’t happen by chance
  - Interpret the result as an answer to the original question

Statistical significance

- Null hypothesis: Assumption that the apparent effect was actually due to chance (H_o)
- P-value: Probability of the apparent effect under the null hypothesis

$$P(Effect\ |Null\ hypothesis)$$

  - If the p-value is low enough, the null hypothesis unlikely true
- Interpretation: Based on the p-value, we conclude if the effect is real or nor
  - i.e. The effect is false until there is a contradiction. If there is a contradiction, then the effect is true

Example

- Testing a difference in Means
  - Null hypothesis – the distribution for the two groups are the same. Difference are due to chance

$$\begin{cases} H_o & \mu_X = \mu_{null} \\ H_A & \mu_X \neq \mu_{null} \end{cases}$$

Choosing a threshold

- Hypothesis testing error
  - False positive – accept hypothesis when it is false
  - False negatives – reject hypothesis when it is true

		True condition	
	Total population	Condition positive	Condition negative
Prediction	Predicted positive	True positive (Power)	False positive Type I error
	Predicted negative	False negative Type II error	True negative

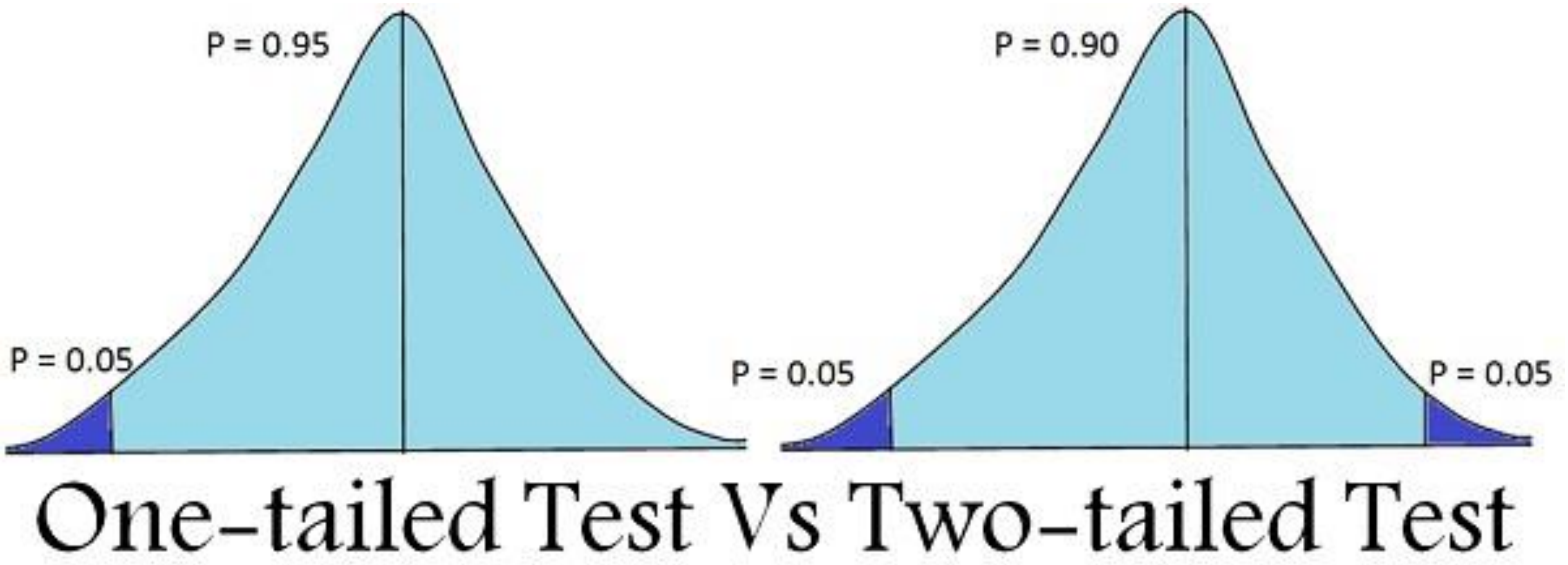
Choosing a threshold

- Statistical Power – It is the probability that the test will be positive if the null hypothesis is false
- False Discovery Rate (FDR) – Rate of false positives and number of true values predicted
- Precision - Rate of true positives and number of true values predicted
- Sensitivity – Rate of true positive and real true values

Choosing a threshold

- Choose an α threshold for p-values and to accept as significant when p-value<α
- Common choice: $\alpha \leq 5\%$
- The probability of a false positive is α
- If lower alpha then it is lower the chance of false positive
  - However, it may reject a valid hypothesis
- Trade-off between false positives and false negatives

Choosing a threshold



Interpreting the result

- Classical
  - If p-value$<\alpha$, then it is statistical significant

- Practical
  - The lower the p-value, the higher the confidence the effect is real

Statistic test/Contrast test

- They are used to verify or reject a hypothesis from data
- They must have:
  - Data
  - Null hypothesis
  - Alternative hypothesis
  - Contrast statistic – p-value
- Type of contrasts:
  - Parametric
  - Non-parametric

T-test (Univariate)

- Parametric test
- It contrasts the mean of a population
- The population follows a Normal distribution
  - But the variance is unknown
- Hypothesis

$$\begin{cases} H_o: \mu_1 = \mu_0 \\ H_A: \mu_1 \neq \mu_0 \end{cases}$$

Mann-Whitney U Test

- Non-Parametric test
  - N < 25
- It contrasts the centrality of a population (median)
- Symmetric distribution
- Hypothesis

$$\begin{cases} H_o: Median(X) = Median_0 \\ H_A: Median(X) \neq Median_0 \end{cases}$$

T-test (2 Samples)

- Parametric test
  - N<25
- It contrasts the mean of two populations
  - Independent variables
- Both populations follow a Normal distribution
  - But the variance is unknown in both
- Hypothesis

$$\begin{cases} H_o: \mu_1 = \mu_2 \\ H_A: \mu_1 \neq \mu_2 \end{cases}$$

Wilcoxon Test

- Non-Parametric test
  - Small sample
  - Paired data
- It contrasts the centrality of a population (median)
- Symmetric distribution
- Hypothesis

$$\begin{cases} H_o: Median(X) = Median_0 \\ H_A: Median(X) \neq Median_0 \end{cases}$$

Z-test

- Parametric test
  - N >= 25
- It contrasts the mean of two populations
  - Independent variables
- Both populations follow a Normal distribution
- Hypothesis

$$\begin{cases} H_o: \mu_1 = \mu_2 \\ H_A: \mu_1 \neq \mu_2 \end{cases}$$

Correlation test

- Contrast to test for independence between two variables
- If data follows a normal distribution
- Hypothesis

$$\begin{cases} H_o: \rho = 0 \\ H_A: \rho \neq 0 \end{cases}$$

- If data does not follows a normal distribution a Kendall’s Tau correlation coefficient is used

χ^2-test/ Categoric data test

- Contrast to test for homogeneity and/or independence
- Two-way tables
- For each factor the events are summed and are compared to the expected value
- Hypothesis

$$\begin{cases} H_o: Homogeneous \\ H_A: Non-homogeneous \end{cases}$$

Example

- In the dataset "Popular Kids," students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them.

Original Table

Goals	Grade 4	Grade 5	Grade 6	Total
Grades	49	50	69	168
Popular	24	36	38	98
Sports	19	22	28	69
Total	92	108	135	335

Expected Values Table

Goals	Grade 4	Grade 5	Grade 6
Grades	46.1	54.2	67.7
Popular	26.9	31.6	39.5
Sports	18.9	22.2	27.8

- DOF: 4 and $\chi^2 = 1.51 \therefore p - value = 0.8244$

Example

- Dataset from “Popular kids”, now associated by type of school

	School Area			
Goals	Rural	Suburban	Urban	Total
Grades	57	87	24	168
Popular	50	42	6	98
Sports	42	22	5	69
Total	149	151	35	335

- DOF: 4, $\chi^2 = 18.564 \therefore p - value = 0.001$

Modeling

- Model
  - A system’s representation
  - It incorporates the knowledge of the system
- Constraints:
  - All the system variables are observable – maybe not
  - Are the system variables quantifiable?
- Requirements:
  - Representation
  - Learning
  - Inference

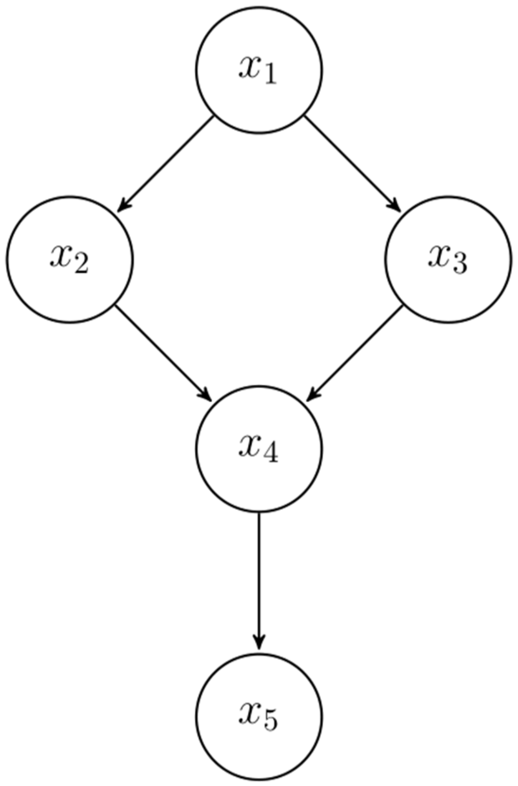
Stochastic models

- Stochastic models are used to model the relationships between random variables
- To model relationships they use independence and probability distributions
- Stochastic modeling is needed when the studied system can be only measured partially

Probabilistic graphic models (PGM)

- PGM are stochastic models that use graphs to represent the system

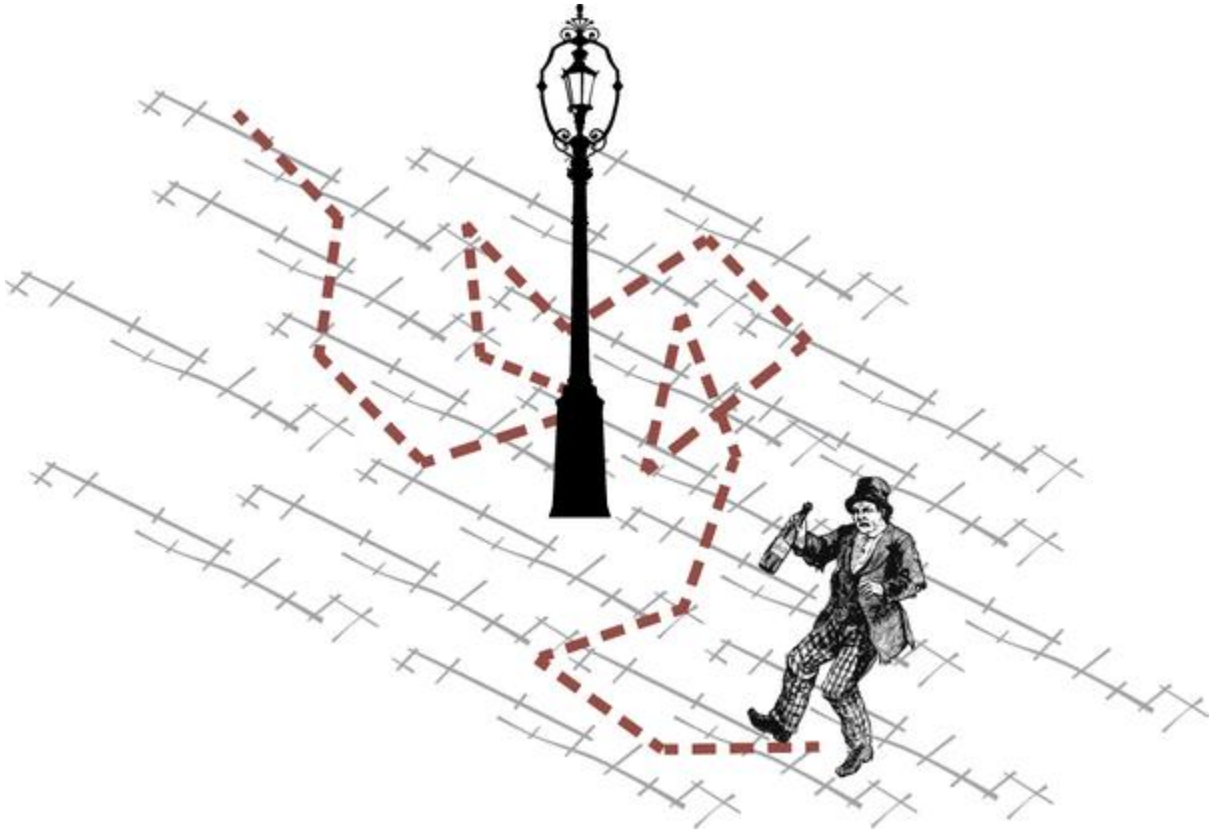
- These models have as components:
  - Nodes that represent the random variables
  - Edges that represent dependence between variables



Stochastic processes

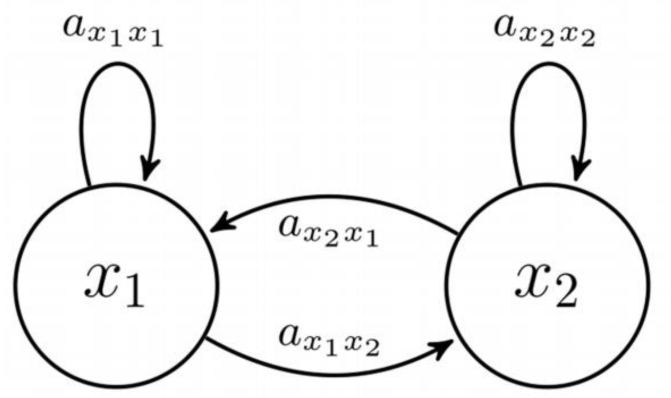
- A stochastic or random process refers to a collection of random variables that are associated or are indexed by another variable
  - i.e. A variable depend on a position or time
- Most of the sciences use stochastic processes
  - Physics
  - Biology
  - Engineering
- E.g. Random walks or Brownian motion

Random walk



Markov models

- A PGM that models transitions over the dependent variable
- Transition graph:



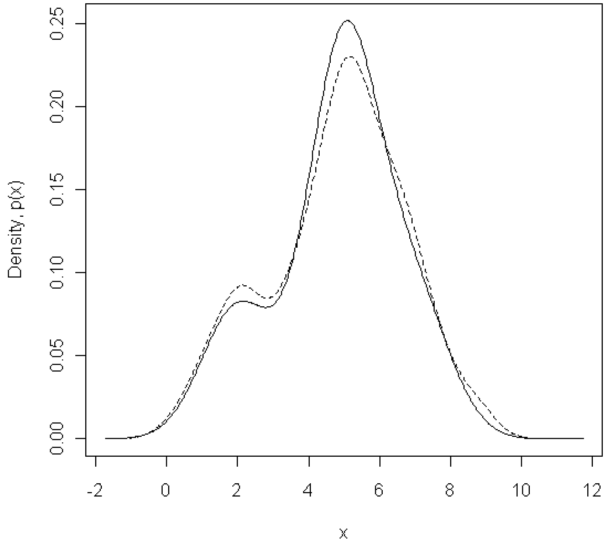
- Markov assumption:
  - The future state only depends in the present one

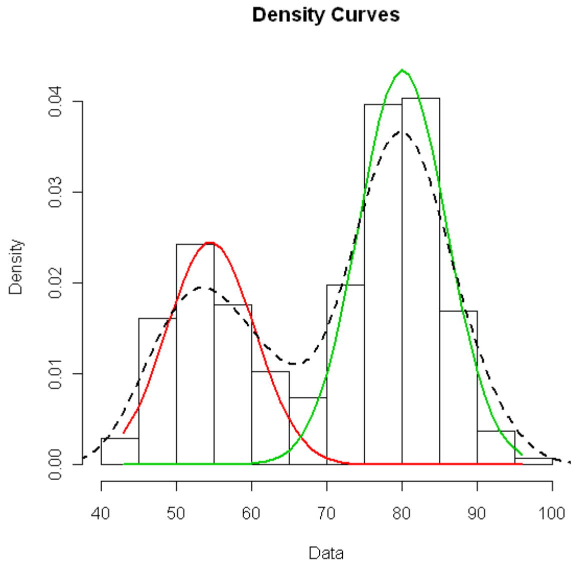
Stochastic models with hidden variables

- A hidden or latent variable is used to represent unmeasurable variables
- They can be used as wildcards to represent a priori information
- Also they are used to simplify equations and to solve probabilistic dependences that are analytically unsolvable

Stochastic models with hidden variables

- Finite Mixture Models
- Every distribution can be modelled as a finite mixture of normal distributions

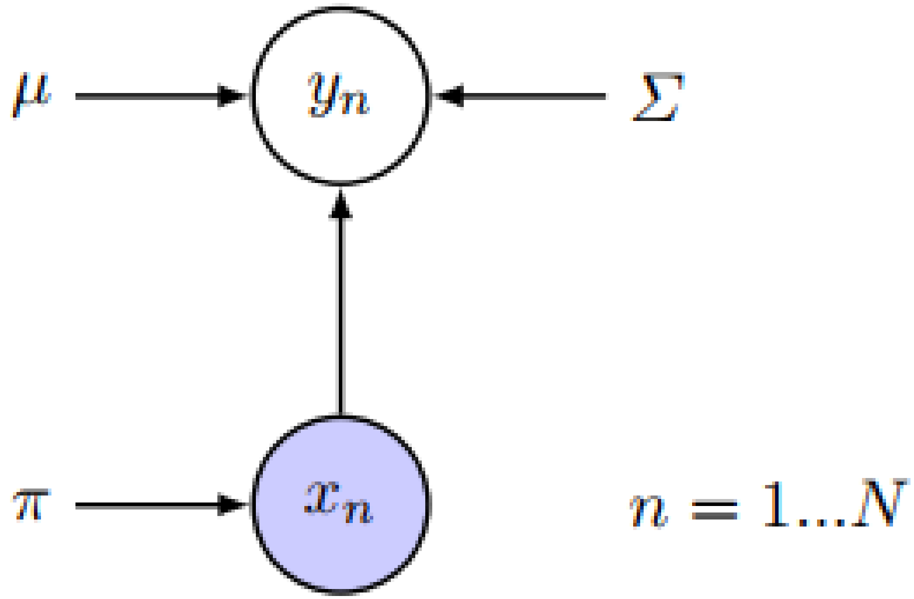




Stochastic models with hidden variables

- Gaussian Mixture Model is a parametric latent variable model

- GMMs are often used for clustering or modeling multimodal data.

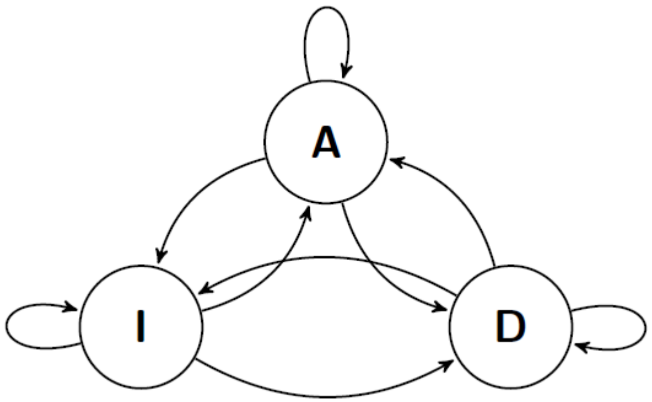


Example

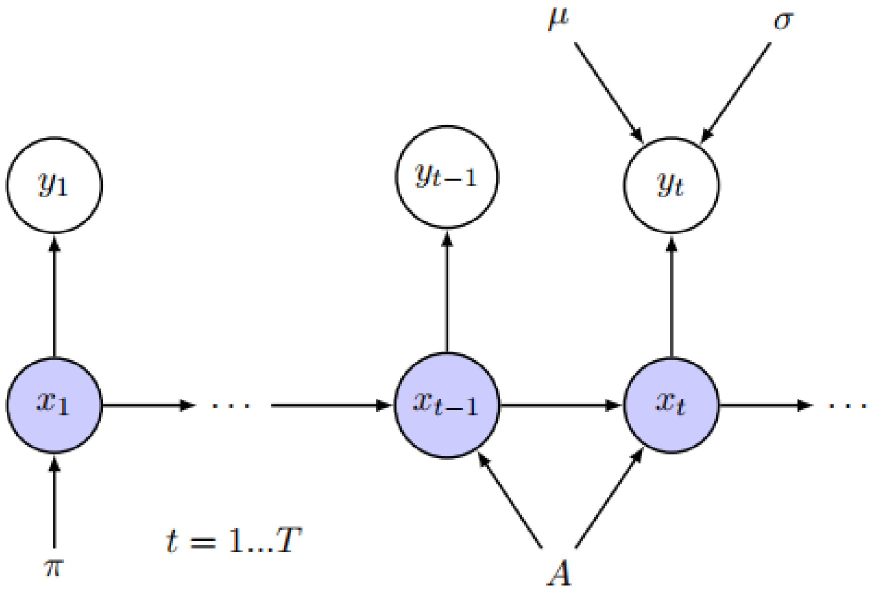
- Database: faithful
- Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA

Stochastic models with hidden variables

- Hidden Markov models (HMMs) are parametric latent variable models that are often used for modeling time series data
- The model stochastic processes

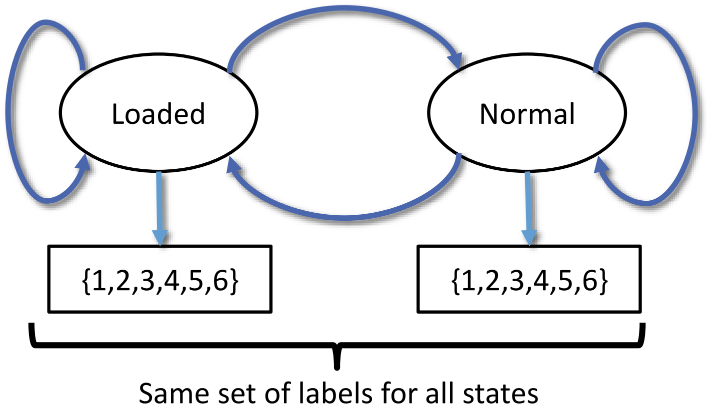


State transition example with 3 states



Hidden Markov models - Applications

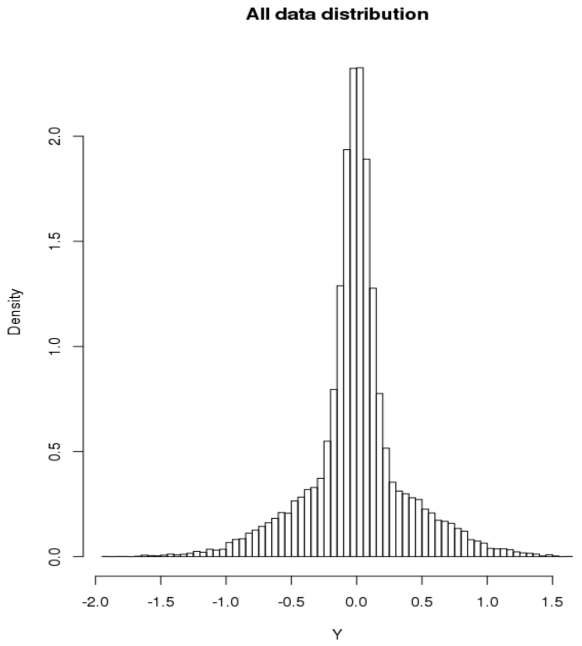
- The observation sequence is a set of nominal categorical values.
- Example
  - Loaded die Values= {1,2,3,4,5,6} and sequence = {1,2,1,1,4,,...,3}
    - The hidden variable Loaded die or Normal die. N = 2
    - The number of Possible symbols to be observed are 6. M = 6

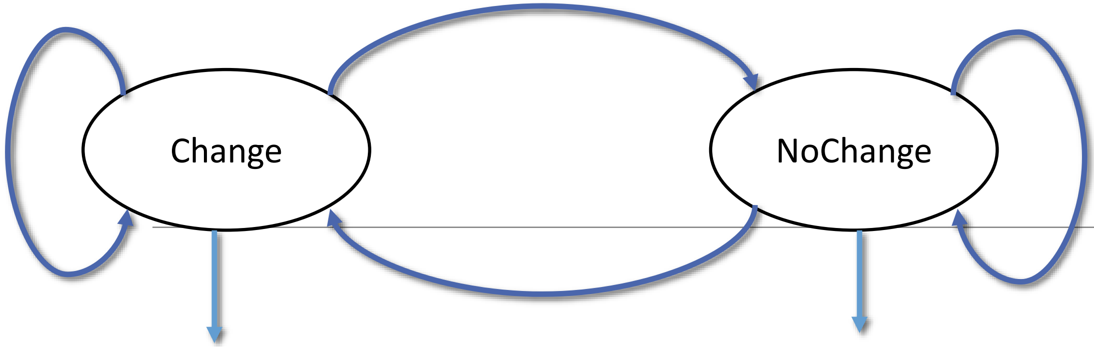


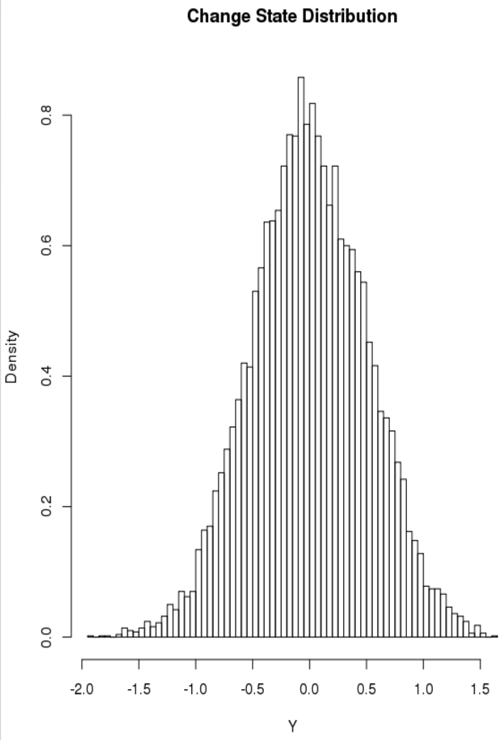
Hidden Markov models - Applications

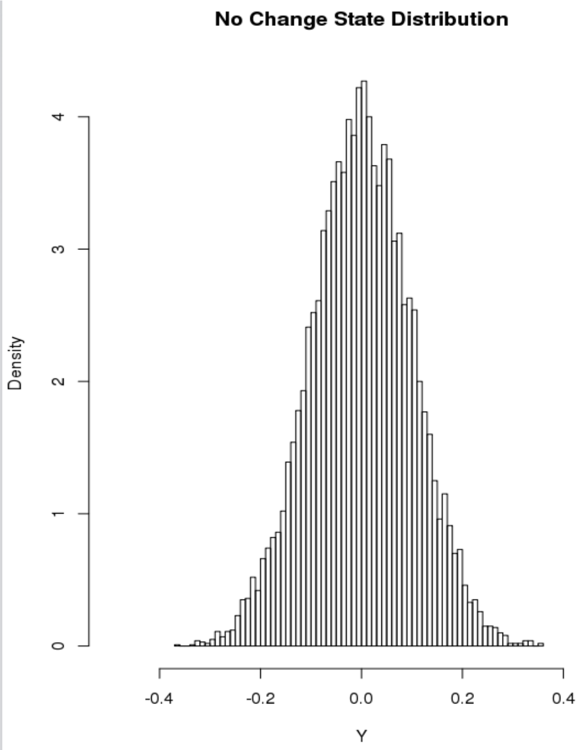
- The observation sequence is a set of continuous values.
  - The observation comes from a Multivariate Normal Distribution
    - Parameters: Mean, Variance Covariance Matrix per State
    - If M = 1, then it is an Univariate Normal Distribution

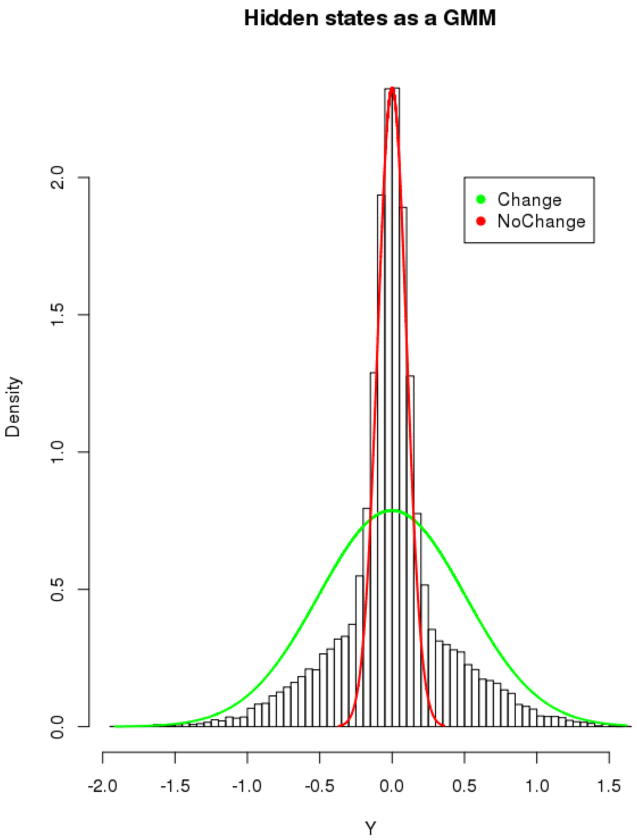
- Example
  - Temperature change and Y = {0.1, 1.0,…,0.5}
    - The hidden variable may be $X_t \in \{\,Change, NoChange\}$
    - The number of Hidden States is 2. N = 2
    - The dimensionality of the Y_t vector is 1. M = 1



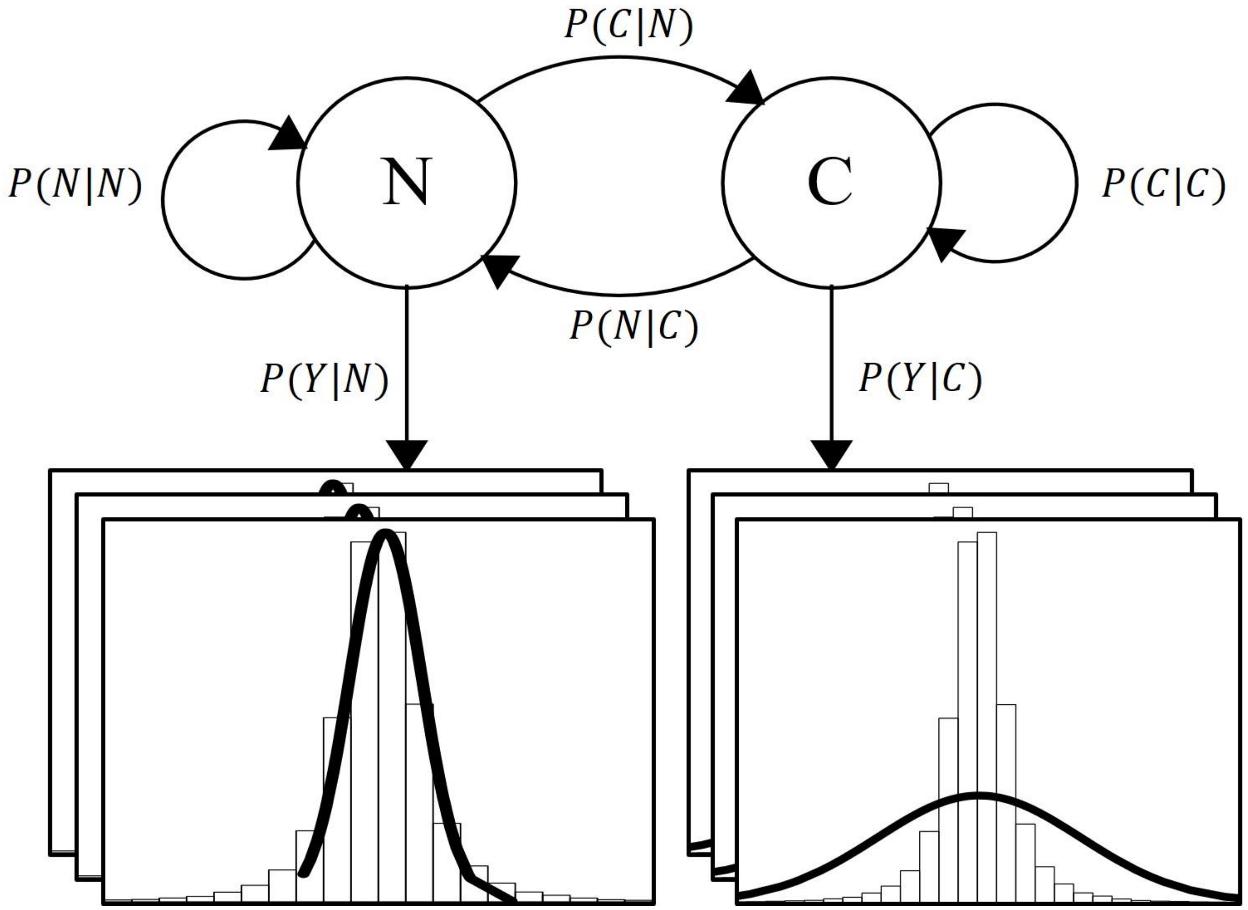

Change
NoChange


Change State Distribution
Density
Y


No Change State Distribution
Density
Y

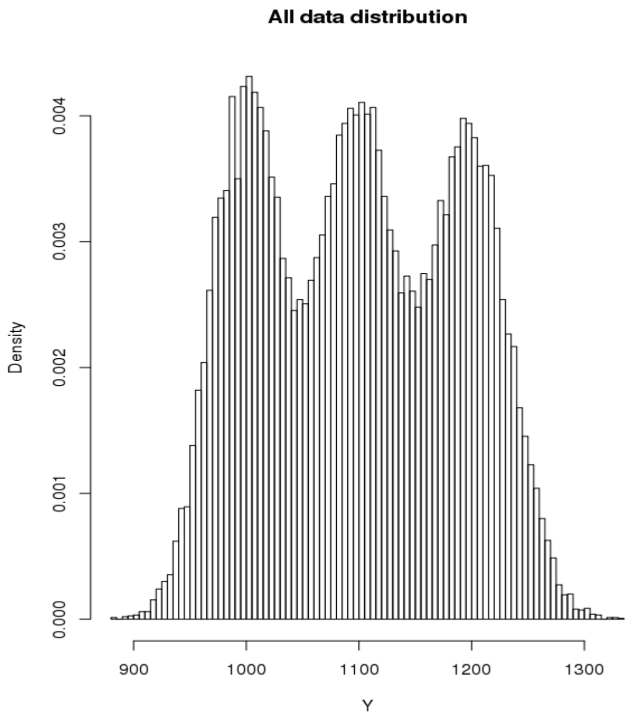

Hidden states as a GMM
Change
NoChange
Density
Y

Hidden Markov models - Applications



Hidden Markov models - Applications

- The observation sequence is a set of discrete values (Counts) .
  - The observation comes from a Poisson Distribution
    - Parameters: Lambda



- Example
  - People in a bank queau and Y = {100,120,…,200}
    - The hidden variable may be $X_t \in \{\, Holiday, Normal, PayDay\}$
    - The number of Hidden States is 3. N = 3

Hidden Markov models

- Learning
  - NP-hard algorithm
  - Expectation-Maximization algorithm
  - Structure fixed –> Parameter estimation
- Inference:
  - Decoding – Hidden states visited
  - Evaluation
  - Data generation

HW

- R code:
  - With the Dataset.csv, filtered by "Drug use disorders" and "Deaths per 100 000 population (standardized rates)“ apply a statistical test to see if the deaths in 2014 are significantly different than in 2003. (50%)
    - Justify the answer and the use of the statistical test

HW

- R Code:
  - Dataset: iris
    - It gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.
  - Use a GMM algorithm to cluster values into 3 classes
  - Independent variables: Sepal length and width, petal length and width
  - Return Confusion matrix of predicted class versus real class (Species)
  - Plot the answer (+10)
    - Independent variables or transformation of independent variables colored by the predicted clas