Statistical analysis using matlab

HY 439

Presented by: George Fortetsanakis

Roadmap

- Probability distributions
- Statistical estimation
- Fitting data to probability distributions

Continuous distributions

- Continuous random variable X takes values in subset of real numbers D⊆R
- X corresponds to measurement of some property, e.g., length, weight

- Not possible to talk about the probability of X taking a specific value

$$P(X = x) = 0$$

- Instead talk about probability of X lying in a given interval

$$P(x_1 \leq X \leq x_2) = P(X \in [x_1, x_2])$$

$$P(X \leq x) = P(X \in [-\infty, x])$$

Probability density function (pdf)

- Continuous function p(x) defined for each x∈D
- Probability of X lying in interval I⊆D computed by integral:

$$P(X \in l) = \int_{x \in l} p(x)dx$$

- Examples:

$$P(x_1 \leq X \leq x_2) = P(X \in [x_1, \mathrm{x}_2]) = \int_{x_1}^{x_2} p(x)dx$$

$$P(X \leq x) = P(X \in [-\infty, \mathrm{x}]) = \int_{-\infty}^{x} p(x)dx$$

- Important property:

$$P(X \in D) = \int_{x \in D} p(x)dx = 1$$

Cumulative distribution function (cdf)

- For each x∈D defines the probability $P(X \le x)$

$$F(x) = P(X \le x) = P(X \in [-\infty, \mathrm{x}]) = \int_{-\infty}^{x} p(x)dx$$

Important properties:

- $F(-\infty) = 0$
- $F(\infty) = 1$
- $P(x_1 \le X \le x_2) = F(x_2) - F(x_1)$

Complementary cumulative distribution function (ccdf)

$$G(x) = P(X \ge x) = 1 - P(X \le x) = 1 - F(x)$$

Exponential distribution

Probability density function

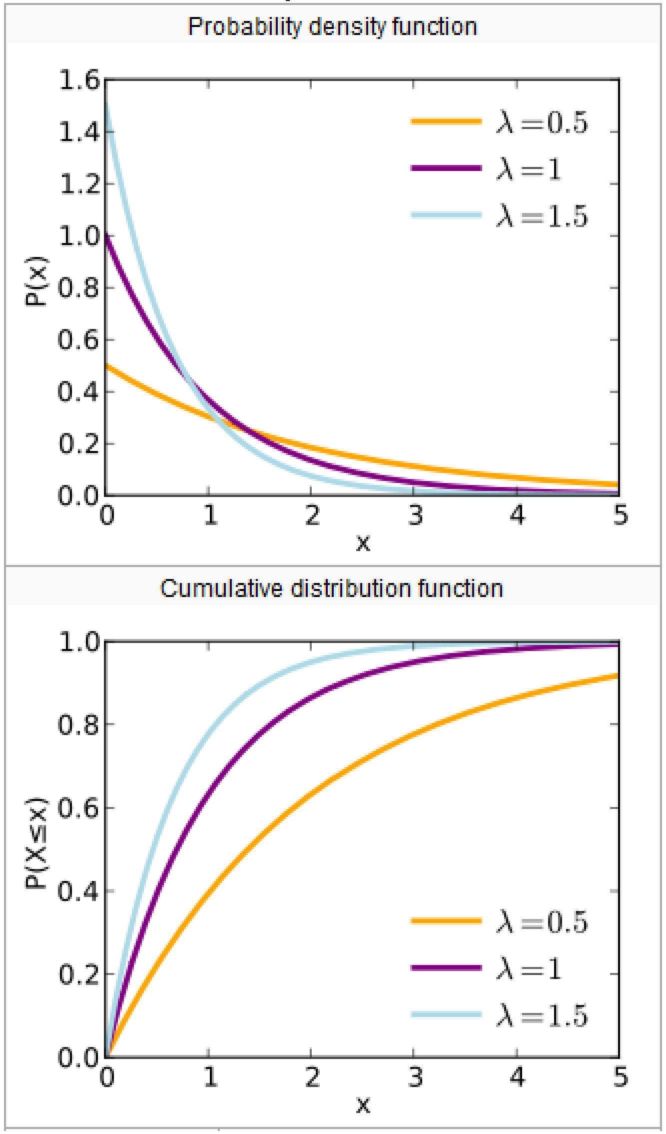
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Cumulative distribution function

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Memoryless property:

$$P(T > \tau + t \mid T \geq \tau) = P(T > t)$$

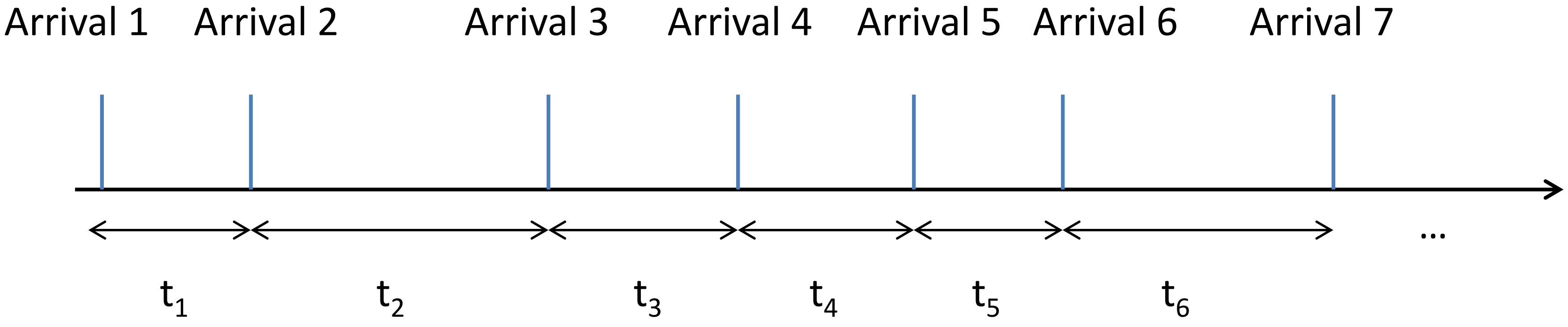


Poisson process

Random process that describes the timestamps of various events

- Telephone call arrivals
- Packet arrivals on a router

Time between two consecutive arrivals follows exponential distribution



Time intervals t_1, t_2, t_3, … are drawn from exponential distribution

Roadmap

- Probability distributions
- Statistical estimation
- Fitting data to probability distributions

Basic statistics

Suppose a set of measurements x = [x_1 x_2 … x_n]

- Estimation of mean value: $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$ (matlab m=mean(x);)
- Estimation of standard deviation: $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \hat{\mu} \right)^2}{n-1}}$ (matlab s=std(x);)

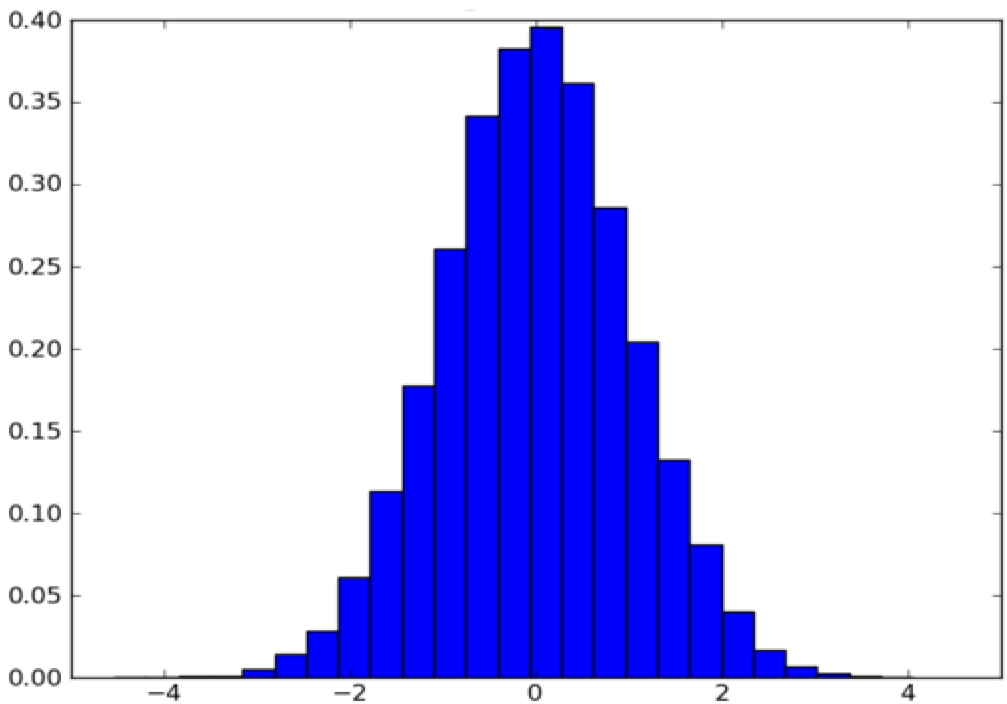
Estimate pdf

- Suppose dataset x = [x_1 x_2 … x_k]
- Can we estimate the pdf that values in x follow?

Estimate pdf

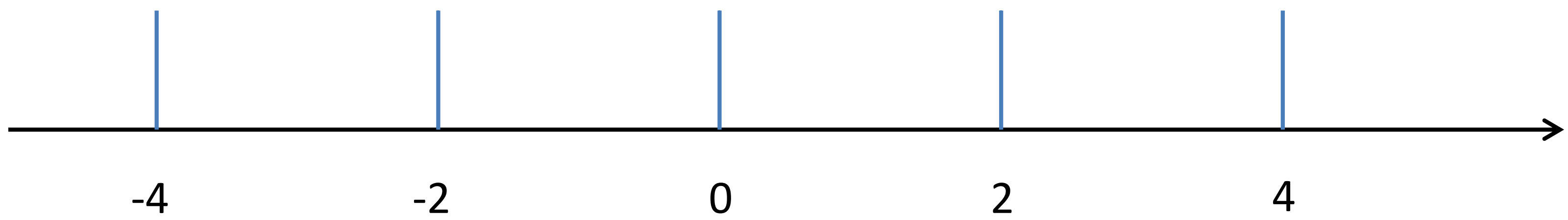
- Suppose dataset $x = [x_1\ x_2\ \dots\ x_k]$
- Can we estimate the pdf that values in x follow?

☞ **Produce histogram**

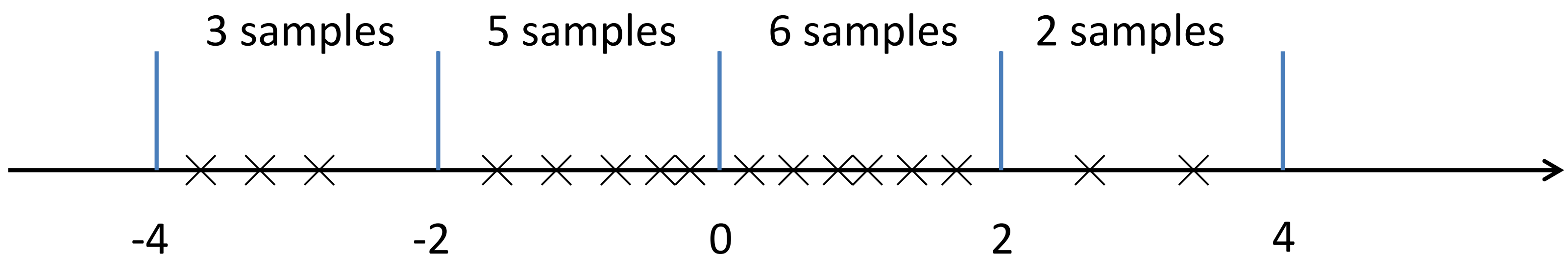


Step 1

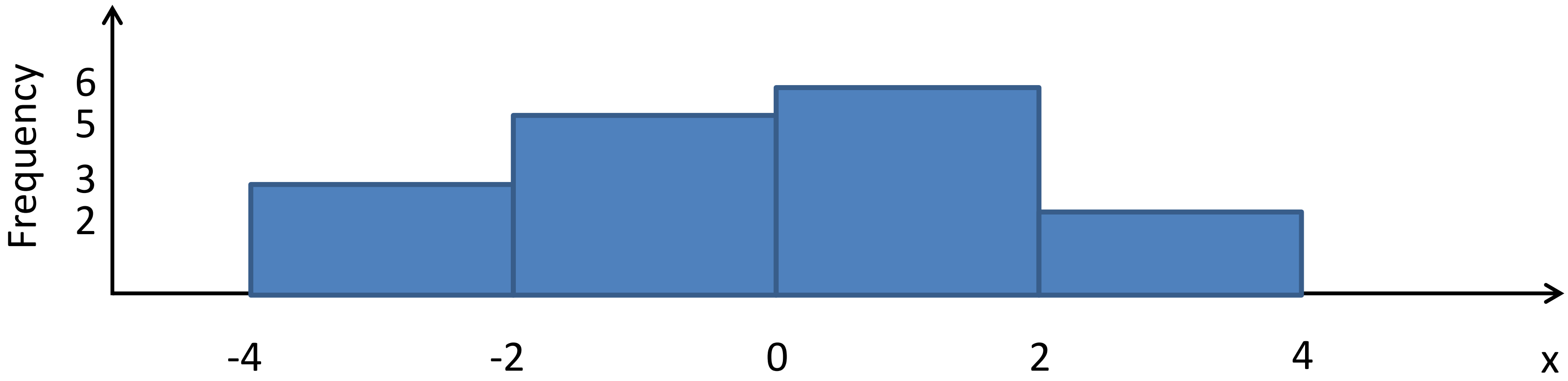
- Divide sampling space into a number of bins



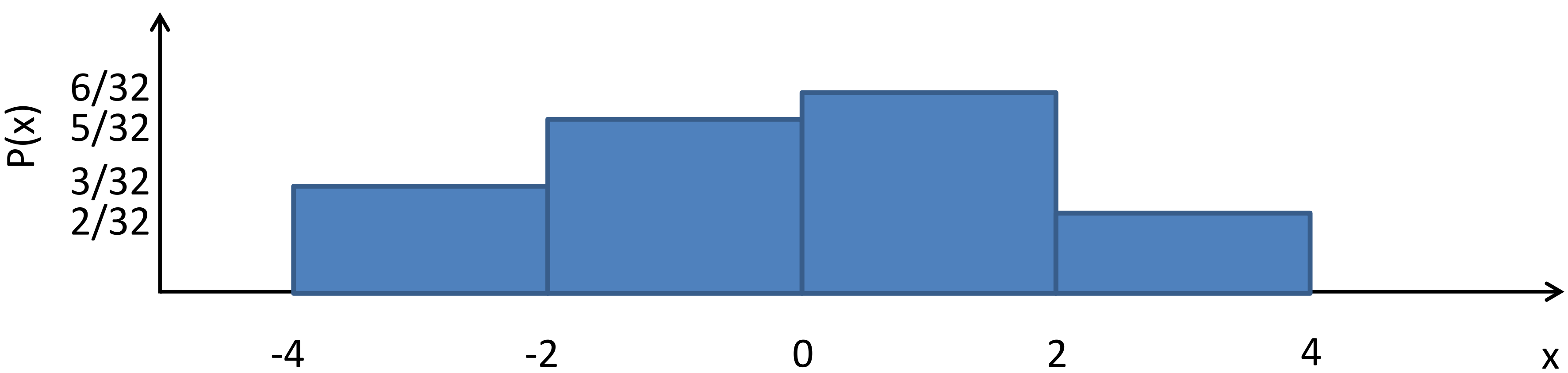
- Measure the number of samples in each bin



Step 2



- E = total area under histogram plot = 2*3 + 2*5 + 2*6 +2*2 = 32
- Normalize y axis by dividing by E



Matlab code

```
function produce_histogram(x, bins)
% input parameters
% X =[x1; x2; … xn]: a column vector containing the data x1, x2, …, xn.
% bins = [b1; b2; …bk]: A vector that Divides the sampling space in bins
% centered around the points b1, b2, …, bk.

    figure; % Create a new figure
    [f y] = hist(x, bins); % Assign your data points to the corresponding bins
    bar(y, f/trapz(y,f), 1); % Plot the histogram
    xlabel('x'); % Name axis x
    ylabel('p(x)'); % Name axis y

end
```

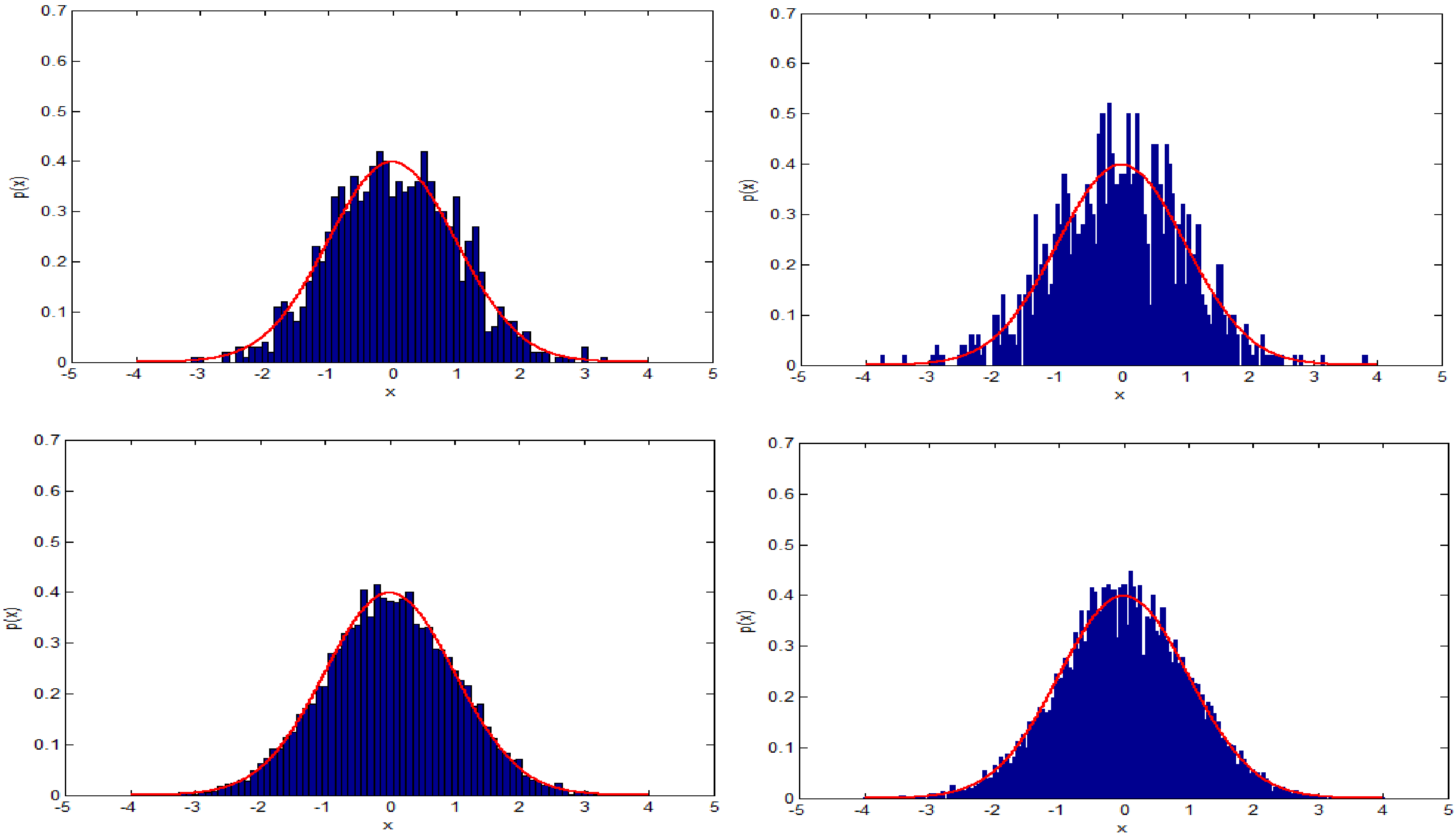
Histogram examples

Bin spacing 0.1

Bin spacing 0.05

1000 samples

10000 samples

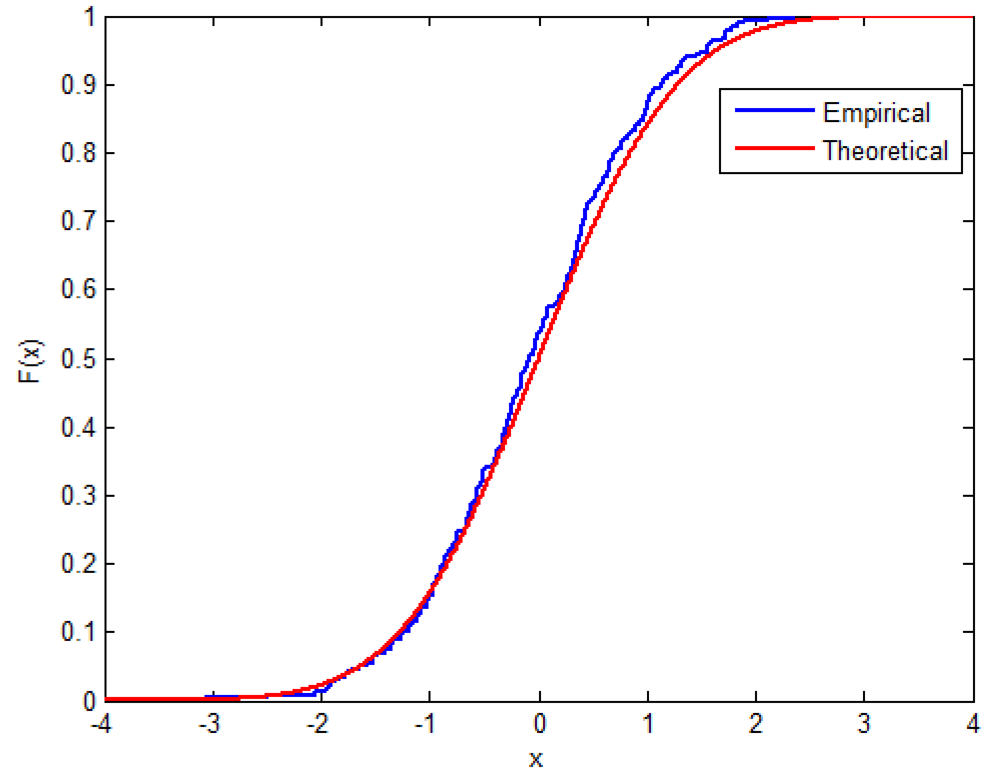


Empirical cdf

How can we estimate the cdf that values in x follow?

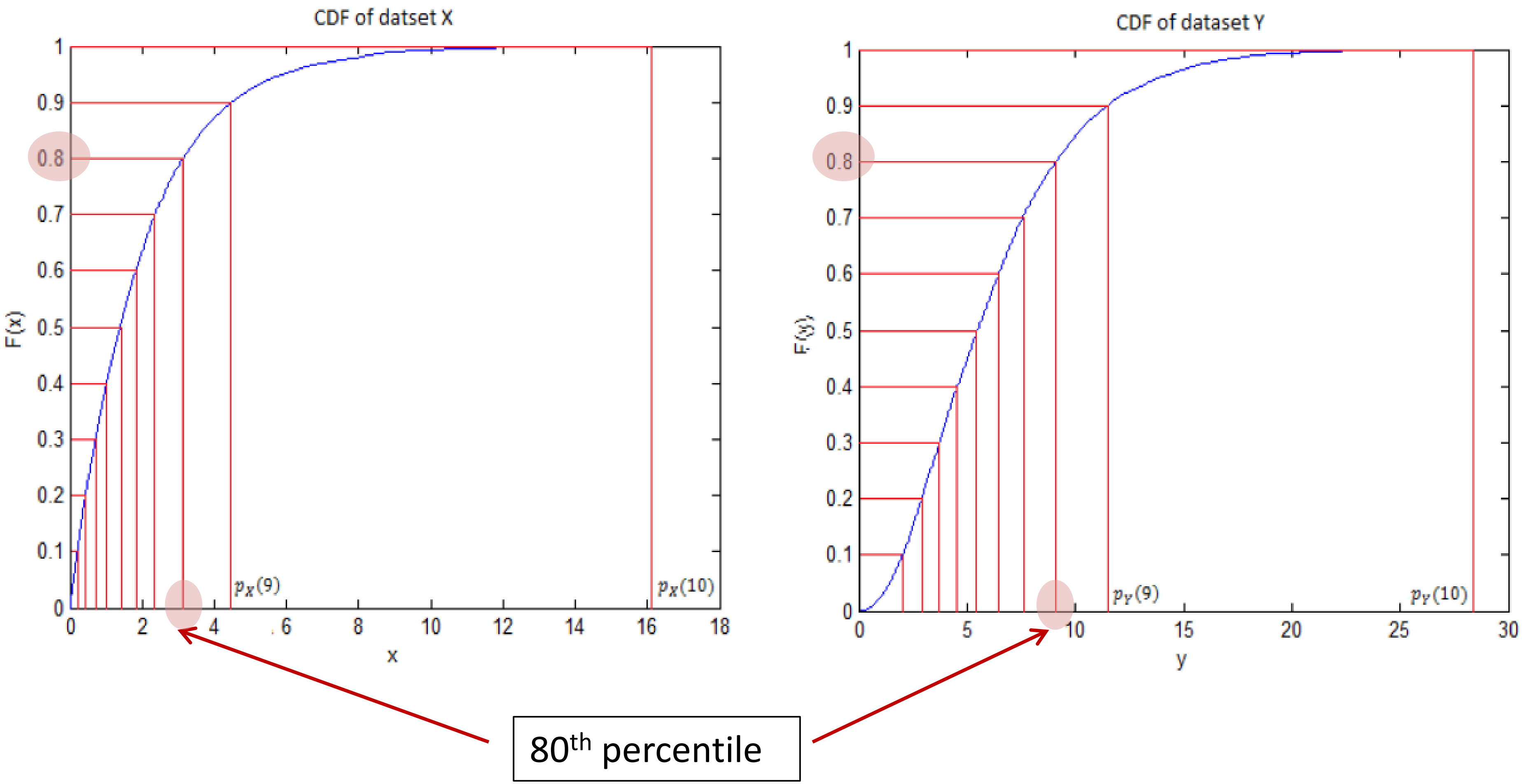
☞ **Use matlab function ecdf(x)**

Empirical cdf estimated with 300 samples from normal distribution



Percentiles

- Values of variable below which a certain percentage of observations fall
- 80th percentile is the value, below which 80 % of observations fall.



Estimate percentiles

Percentiles in matlab: p = prctile(x, y);

- y takes values in interval [0 100]
- 80^{th} percentile: p = prctile(x, 80);

Median: the 50^{th} percentile

- med = prctile(x, 50); or
- med = median(x);

Why is median different than the mean?

- Suppose dataset x = [1 100 100]: mean = 201/3=67, median = 100

Roadmap

- Elements of probability theory
- Probability distributions
- Statistical estimation
- Fitting data to probability distributions

Problem definition

Dataset D={x_1, x_2, …, x_k} collected from an experiment

Families of distributions: $S = \{P_1(x \mid \boldsymbol{\theta}_1), P_2(x \mid \boldsymbol{\theta}_2), \ldots, P_N(x \mid \boldsymbol{\theta}_N)\}$

- Gaussian: $\boldsymbol{\theta}_i = (\mu, \sigma)$
- Exponential: $\boldsymbol{\theta}_i = \lambda$
- Generalized pareto: $\boldsymbol{\theta}_i = (\kappa, \sigma, \theta)$

☞ Which family of distributions better describes the dataset D?

Step 1: Maximum likelihood estimation

- For each family i determine parameter $\boldsymbol{\theta}_{\mathbf{i}}^{*}$ that better **fits** the data
- Maximize likelihood of obtaining the data with respect to $\boldsymbol{\theta}_{\mathrm{i}}$

$$\boldsymbol{\theta}_{\mathbf{i}}^{*} = \arg\max_{\boldsymbol{\theta}_{\mathbf{i}}} p(D \mid \boldsymbol{\theta}_{\mathbf{i}})$$

Likelihood function

$$= \arg\max_{\boldsymbol{\theta}_{\mathbf{i}}} p(x_1, x_2, \ldots, x_k \mid \boldsymbol{\theta}_{\mathbf{i}})$$

$$= \arg\max_{\boldsymbol{\theta}_{\mathbf{i}}} \prod_{j=1}^{k} p(x_j \mid \boldsymbol{\theta}_{\mathbf{i}})$$

Due to independence of samples

$$= \arg\max_{\boldsymbol{\theta}_{\mathbf{i}}} \sum_{j=1}^{k} \ln\left(p(x_j \mid \boldsymbol{\theta}_{\mathbf{i}})\right)$$

Example: exponential distribution

- Probability density function

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

- Define the log-likelihood function

$$l(\lambda) = \sum_{i=1}^{k} \ln(\lambda e^{-\lambda x_i}) = \sum_{i=1}^{k} \ln(\lambda) - \sum_{i=1}^{k} \lambda x_i = k \ln(\lambda) - \lambda \sum_{i=1}^{k} x_i$$

- Set derivative equal to 0 to find maximum

$$\frac{dl(\lambda)}{d\lambda} = 0 \Rightarrow \frac{k}{\lambda} - \sum_{i=1}^{k} x_i = 0 \Rightarrow \lambda^* = \frac{k}{\sum_{i=1}^{k} x_i}$$

Reform question

After MLE: instead of families we have specific distributions

$$P_1(x\,|\,\boldsymbol{\theta}_{\mathbf{1}}^{*}), P_2(x\,|\,\boldsymbol{\theta}_2^{*}), \ldots, P_N(x\,|\,\boldsymbol{\theta}_{\mathrm{N}}^{*})$$

☞ Which distribution better describes the data?

Choose most appropriate distribution based on:

- Q-Q plots
- Kullback–Leibler divergence

Method of Q-Q plots

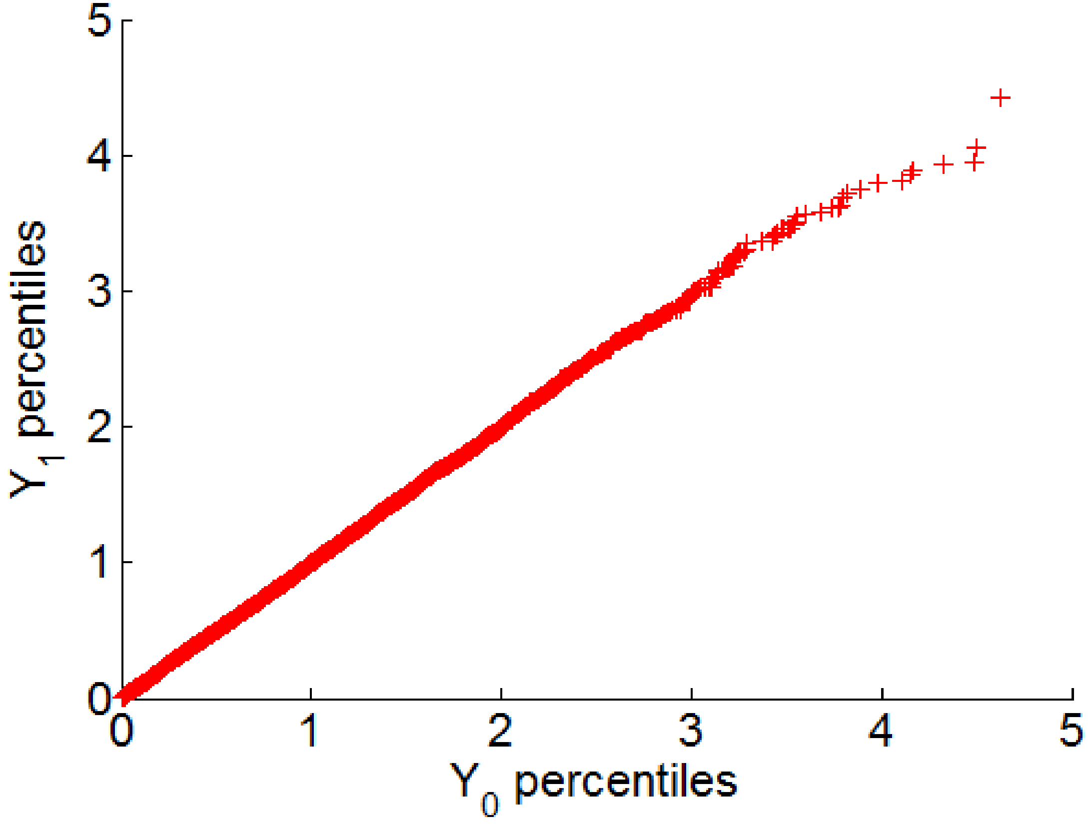
Checks how well a probability distribution $P_i(x \mid \boldsymbol{\theta}_i^*)$ describes the data

Algorithm

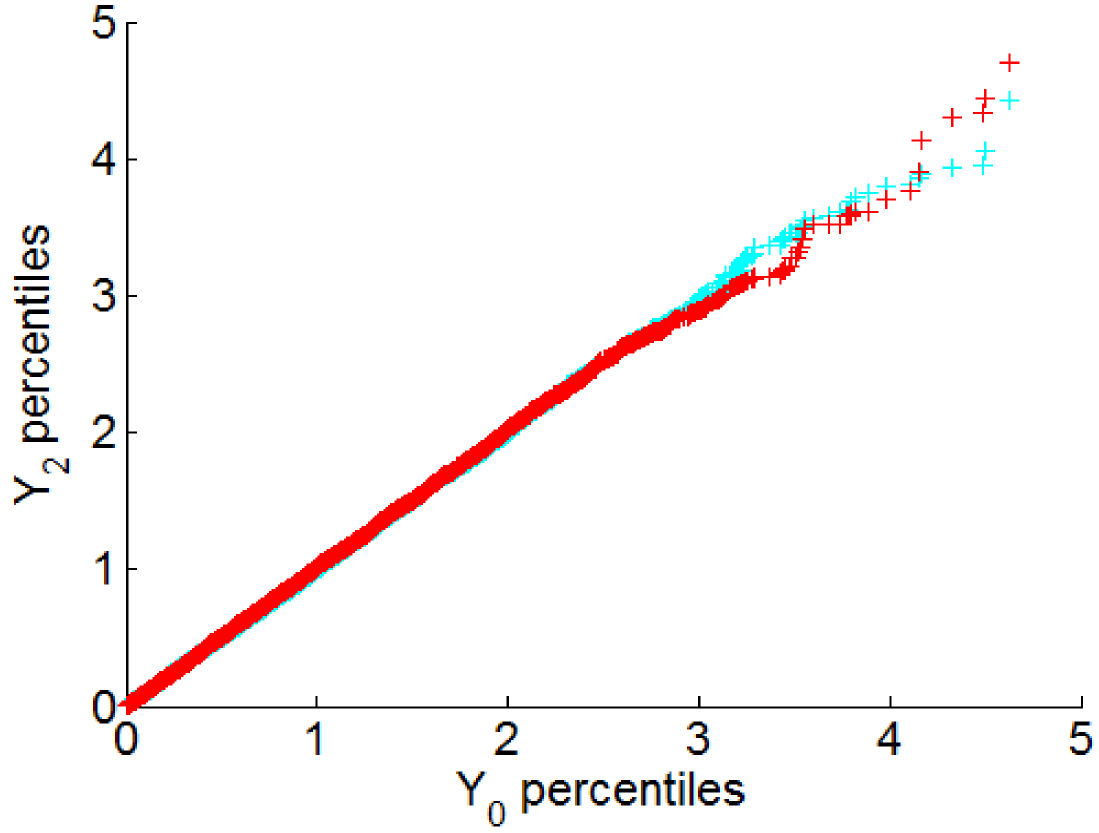
1. Draw random datasets Y_0, Y_1, Y_2, …, Y_M from distribution $P_i(x \mid \boldsymbol{\theta}_i^*)$
2. Compute percentiles of these datasets at predefined set of points
3. Compute percentiles of experimental dataset D at the same points
4. Plot percentiles of Y_0 against percentiles of each of Y_1, Y_2, .., Y_M
5. Plot percentiles of Y_0 against percentiles of dataset D

If plot of step 5 is in the area defined by plots in step 4 the distribution describes the data well

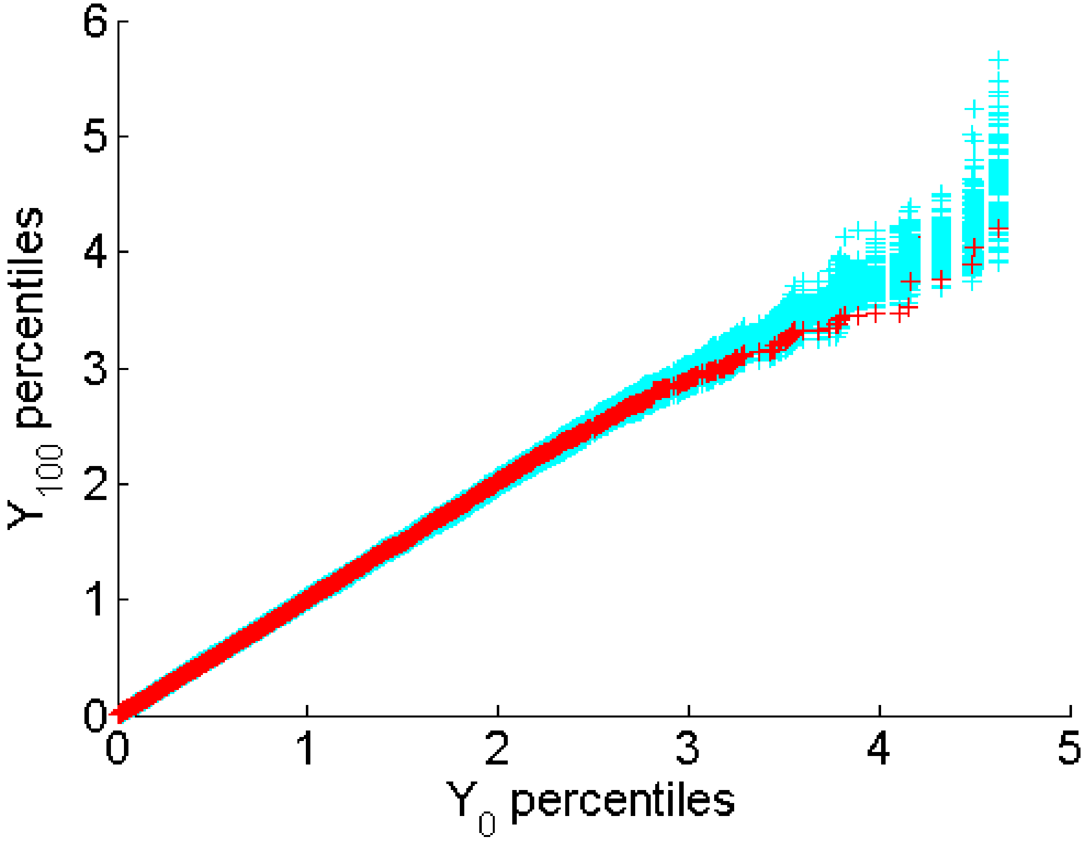
Plot percentiles of Y_0 vs. percentiles of Y_1



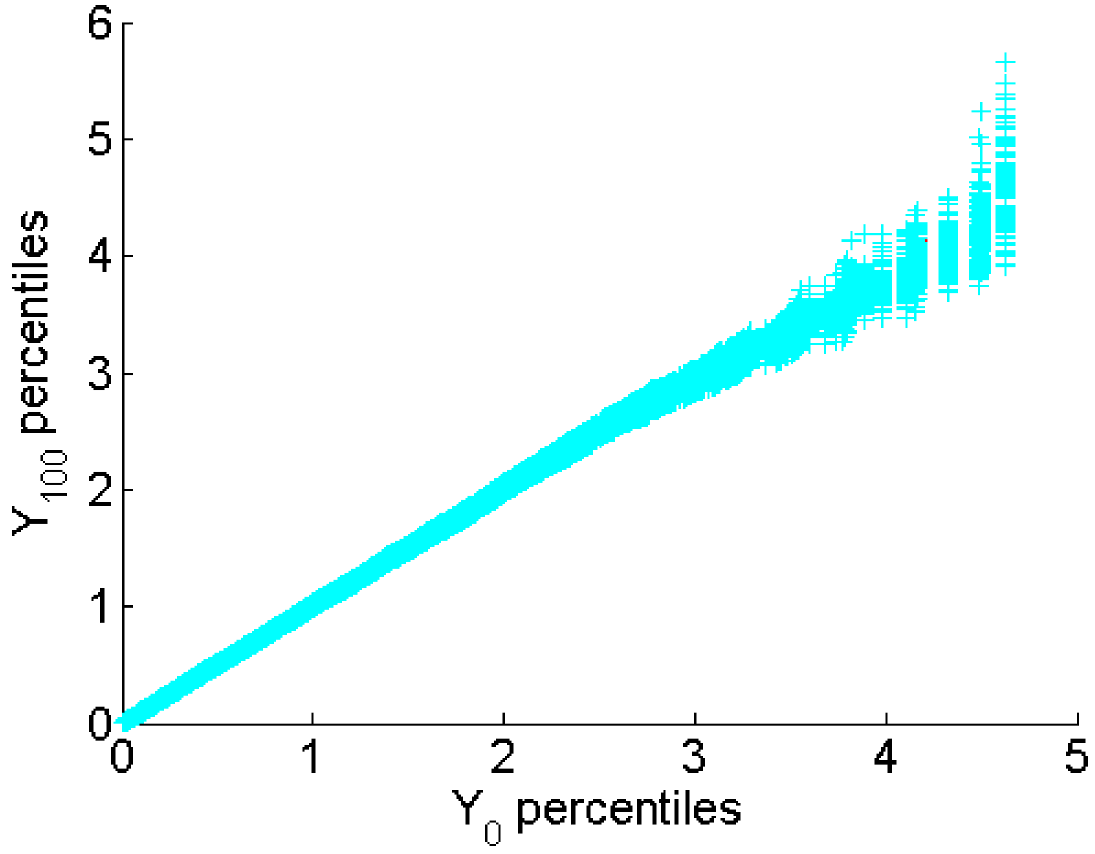
Plot percentiles of Y_0 vs. percentiles of Y_2



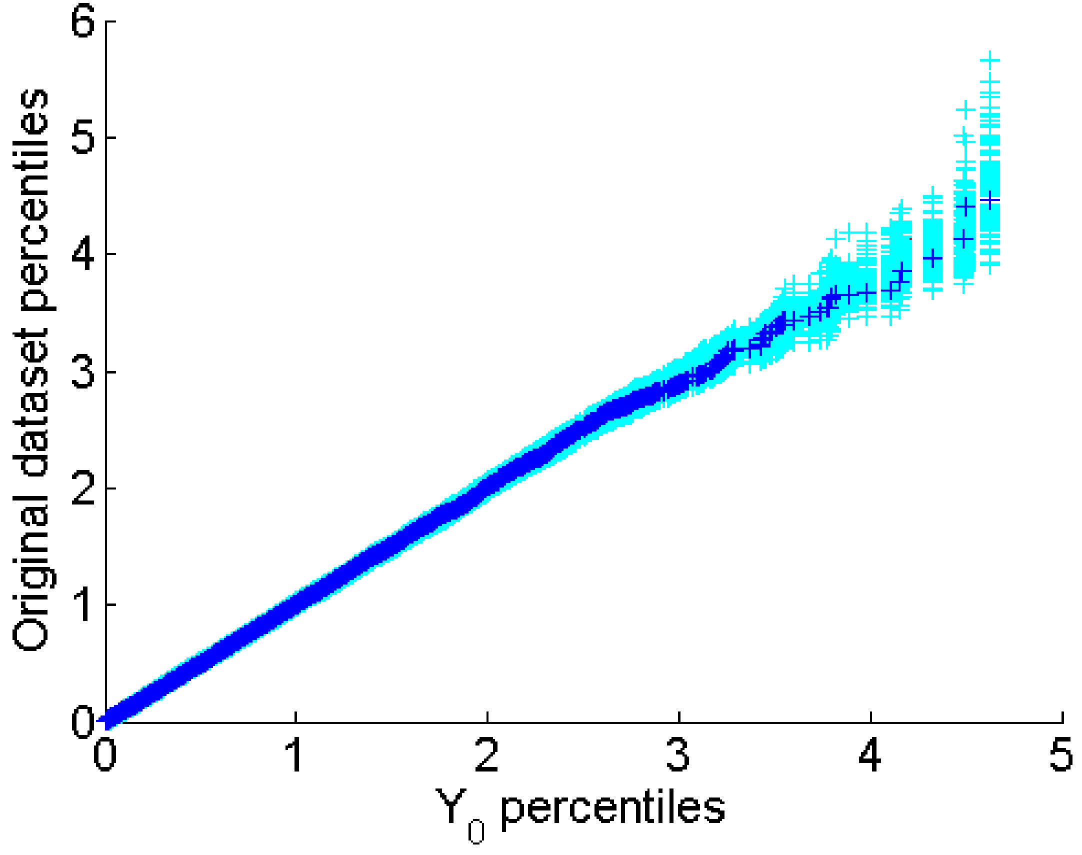
Plot percentiles of Y_0 vs. percentiles of Y_{100}



Construct envelope

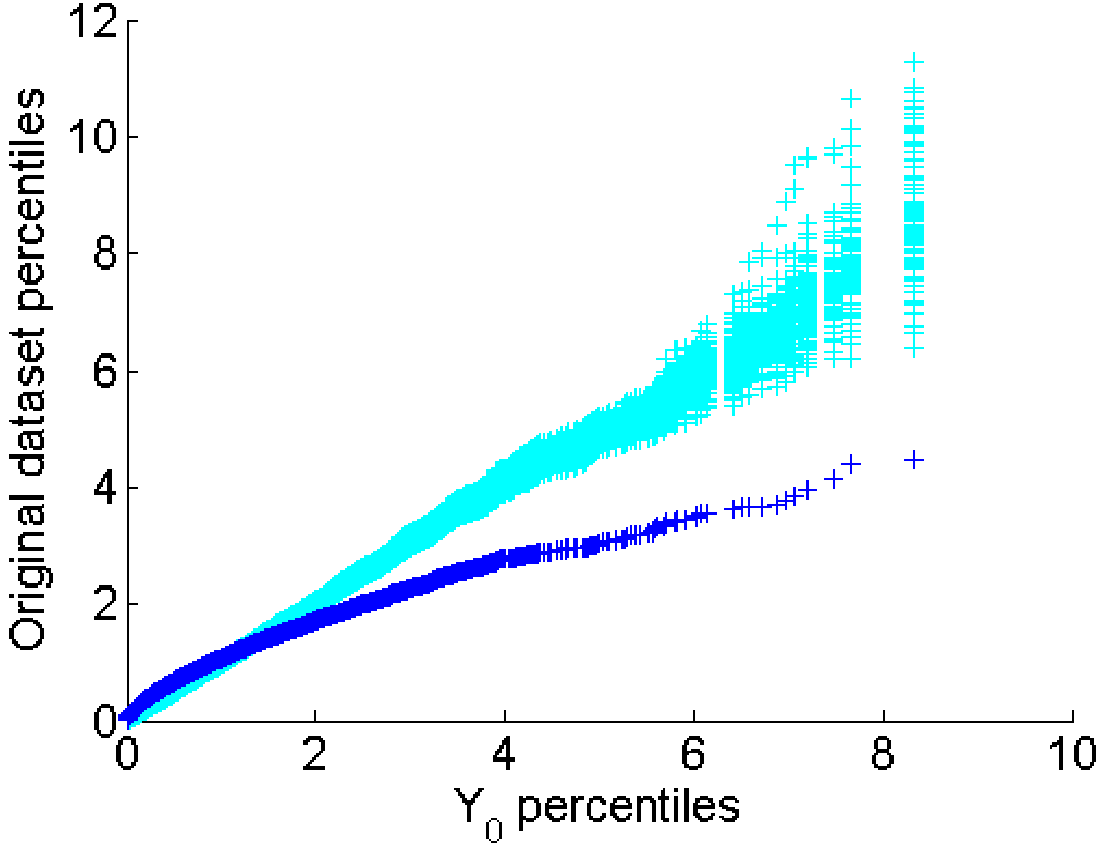


Plot percentiles of Y_0 vs. percentiles of D



☞ **Good fitting**: The blue curve of original percentiles lies in the envelope

Plot percentiles of Y_0 vs. percentiles of D



☞ **Bad fitting**: The blue curve of original percentiles lies outside the envelope

Method of Kullback–Leibler divergence

Non-symmetric metric of difference between distributions P and Q

Discrete distributions

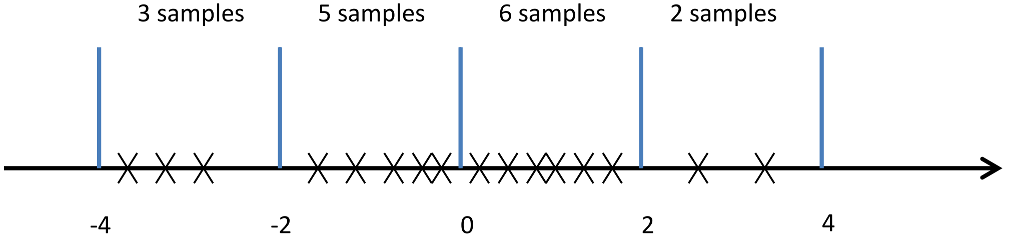
$$D_{KL}(P \parallel Q) = \sum_{i=1}^{N} p(i) \log \frac{p(i)}{q(i)}$$

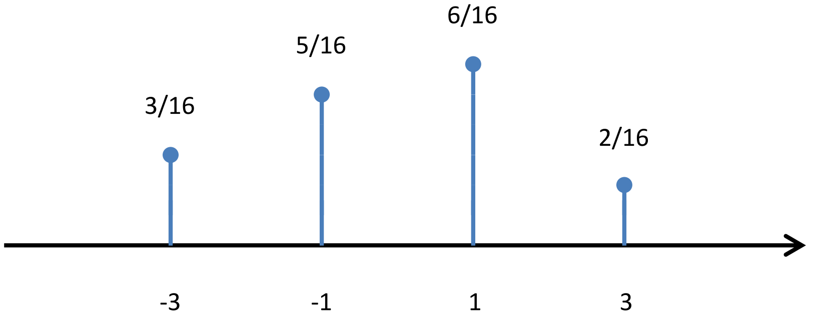
Continuous distributions

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Algorithm

1. Discretize the empirical pdf of the Dataset D





2. Discretize all distributions $P_1(x\,|\,\boldsymbol{\theta}_1^*), P_2(x\,|\,\boldsymbol{\theta}_2^*), \ldots, P_N(x\,|\,\boldsymbol{\theta}_N^*)$
3. Compute KL divergence of theoretical distributions with dataset D
4. Choose the distribution with the lowest KL divergence

Online material

http://www.csd.uoc.gr/~hy439/schedule09.html

- Tutorials → Statistics

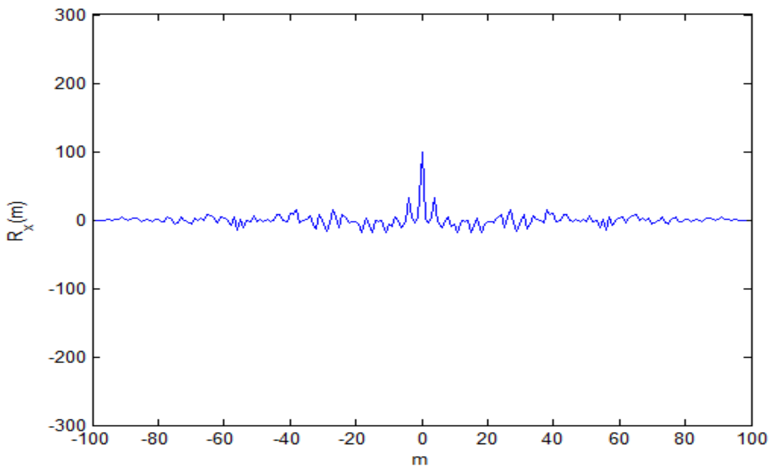
Cross correlation

xcorr(x, y): estimates the cross correlation between two time series x and y

$$R_{xy}(m) = E[x_{n+m} y_n] = E[x_n y_{n-m}]$$

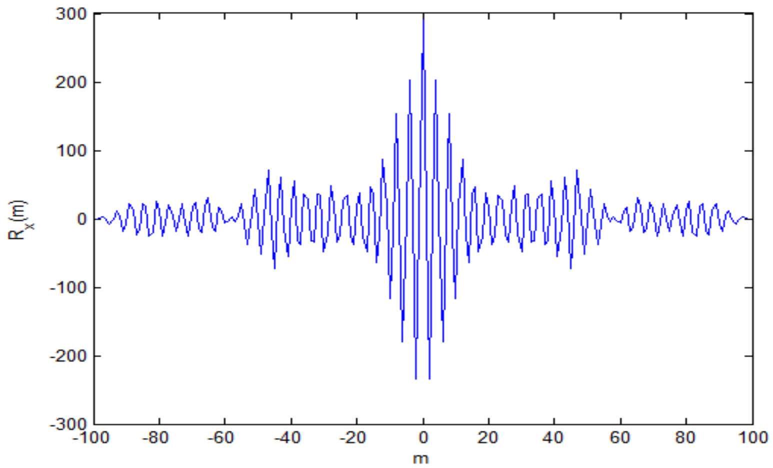
The larger the absolute value of the cross correlation the larger the correlation of the two variables

White noise



No correlation

Output of IIR filter



Some correlation