UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

Measured Data Analysis

Renato Lo Cigno

Simulation and Performance Evaluation 2014-15

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- Whenever we take a measure we "extract" samples our of a (potentially infinite) population
- **Population:** It is the entire set of possible results of an experiment, normally it is not completely known, thus one of the goals of the experiment is to "learn and understand" the population (have insight into a problem)
- **Sample:** It is the complete outcome of the experiment, necessarily finite, possibly repeated many times
- The sample is normally raw data, and we have to manipulate it to gain insight

Good & Bad Samples

- Not all samples are equivalent to evaluate a population
- Having some a-priory knowledge on both the problem and the population can help crafting the correct experiment
- Often some pre-experiment can help gaining some insight to better design the "true" experiment

Are all balls red?

You have a black bowl with 1000 balls and you have to tell if they are all red or not

- Bad Experiment: Pick a ball, look at the color, then put it back, shake to randomize and pick the next one
- Good Experiment: Pick a ball, look at the color, set it aside, shake to randomize and pick the next one
- Why is the second experiment better?
- **Homework:** suppose 990 balls are red and 10 are blue, thus the correct answer is NO, not all balls are red.
  Compute the probability of giving the correct answer after extracting 100 balls in the bad and the good experiments

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

Given a dataset (sample) $\{x_i\}$, what is the best way of visualizing it?

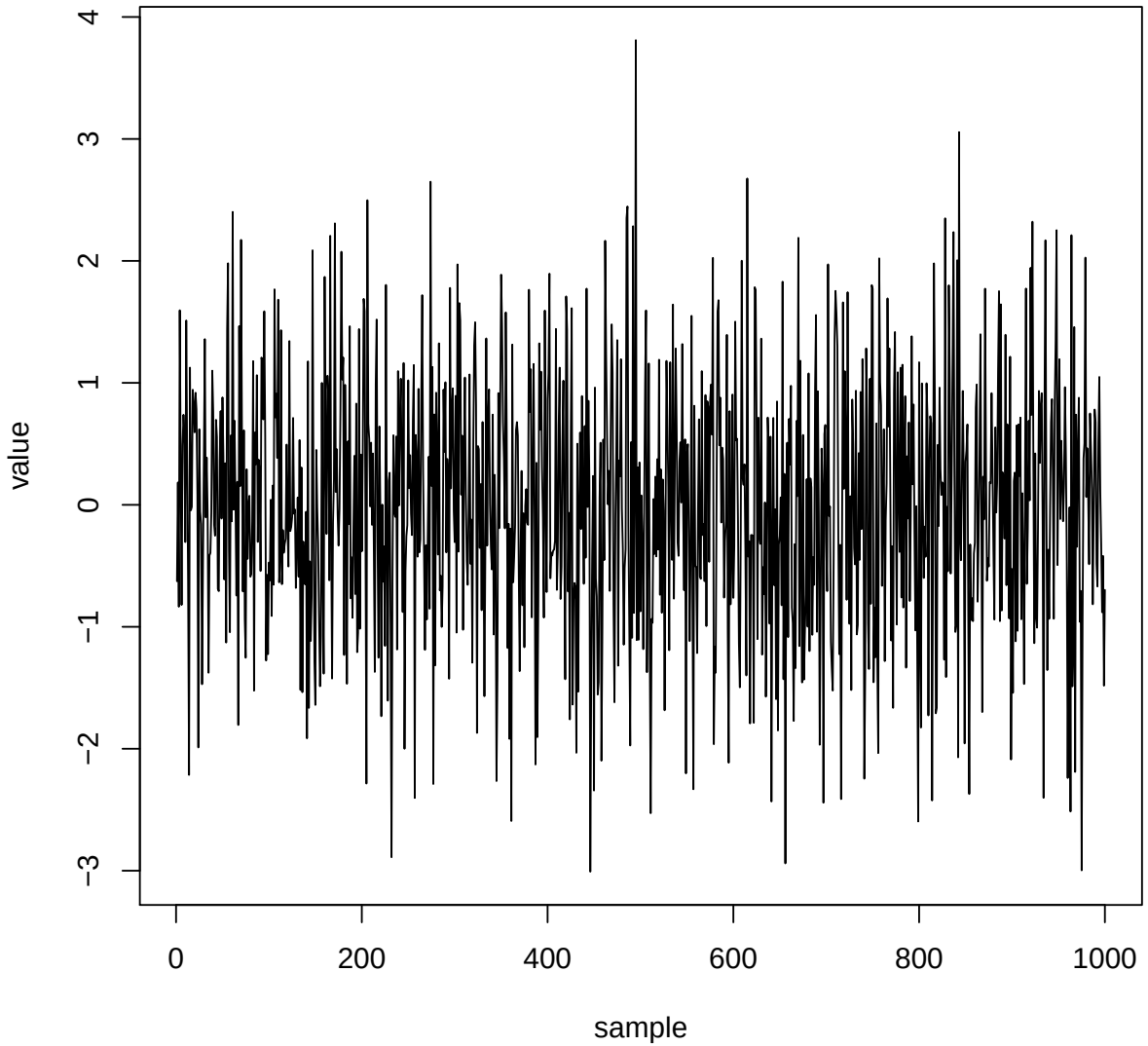
- The numbers?
- Straight plots of the indexed data?
- The plot ordered by the X_i values?
- The cumulative distribution: Experimental CDF (ECDF)?
- Or histograms of the relative frequency of data?

Numbers

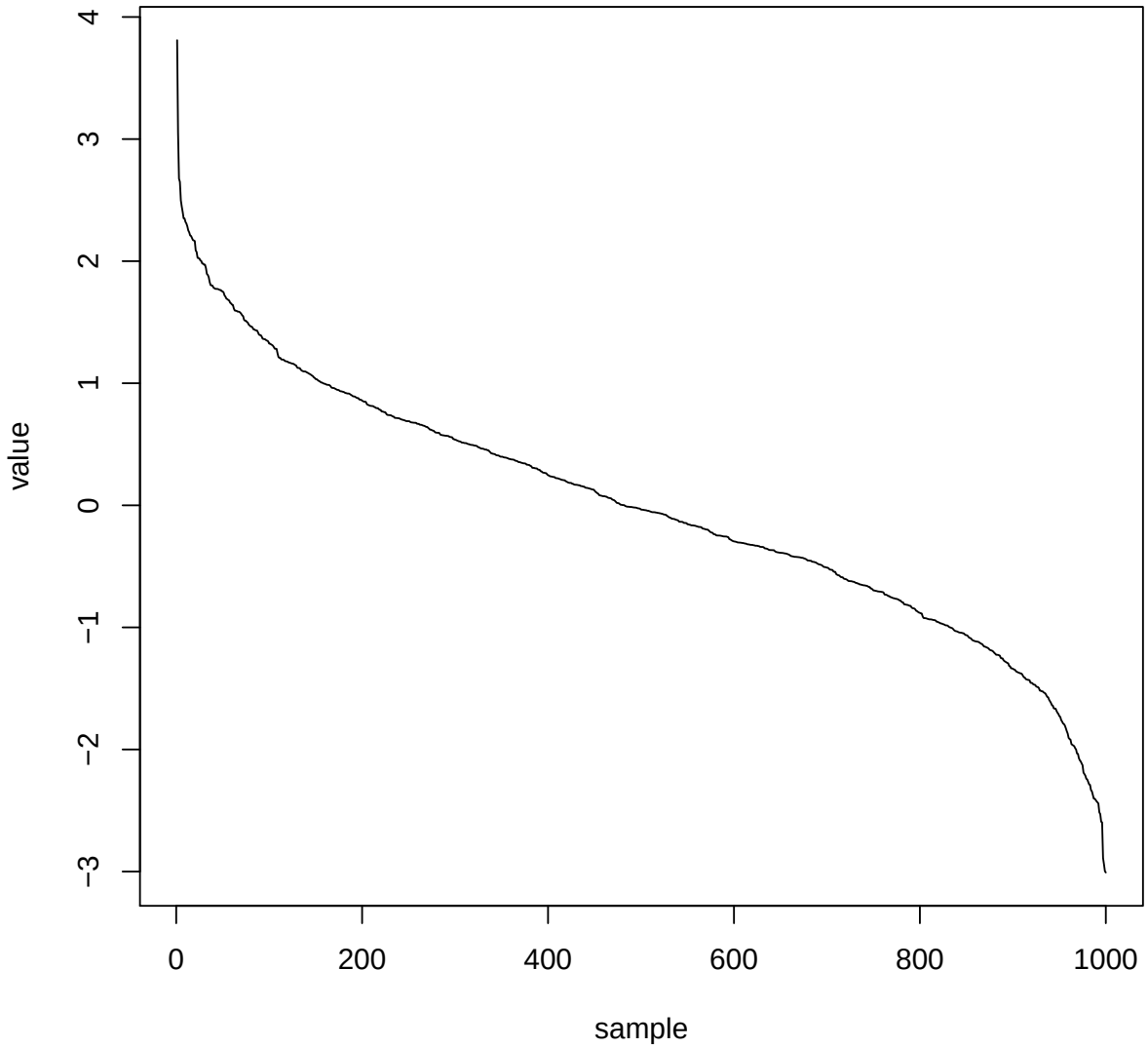
1.35976969874364
1.59431277979215
0.981219343895978
1.37933187884847
0.259440011862454
1.65806836346625
1.11261933869171
0.980250563088334
0.0783774189765842
2.08917702766969
1.54797818279134
0.550765177138121
1.67971900635026
-1.93547784711214
-0.0269931353634314
0.864489892299465
1.82027881068408
0.939980335400623
-1.26817987739339
0.701869007956606
1.38576668328979
2.05755445265121
1.09434340121316
1.43801384879194
1.6531848612294
1.13875441255562

. . . . . .

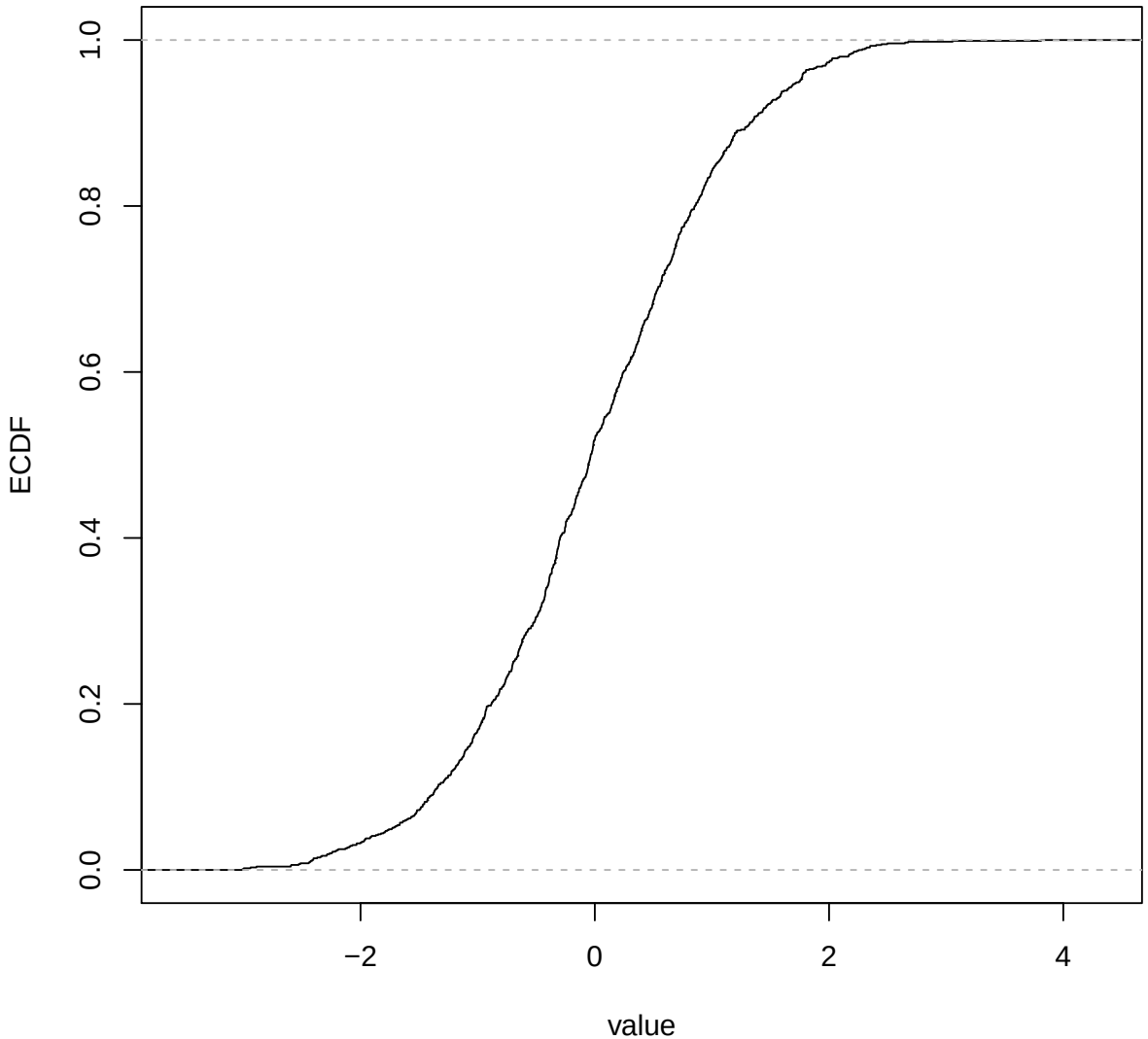
Plotting x_i versus i



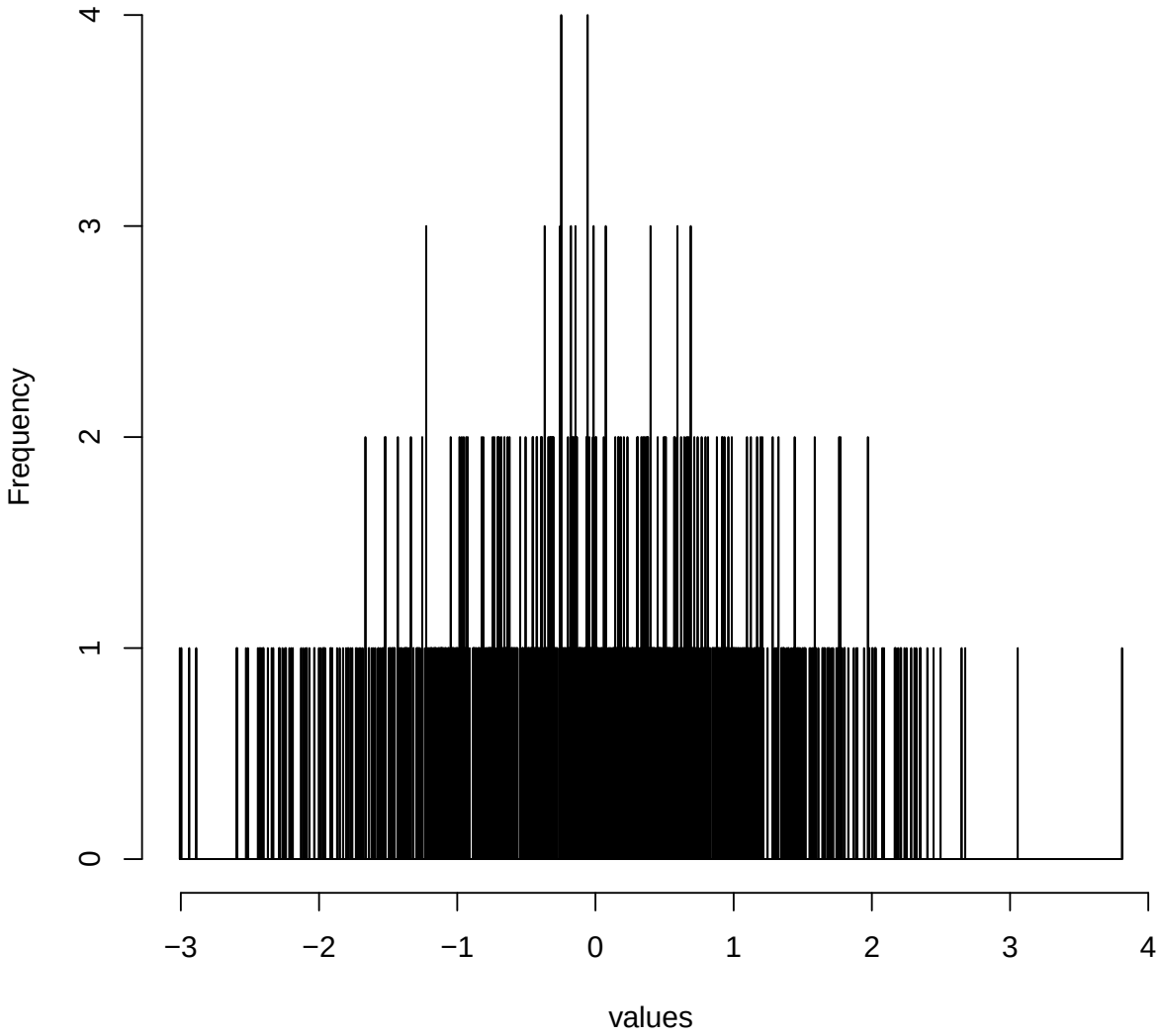
Ordering samples can give a better idea



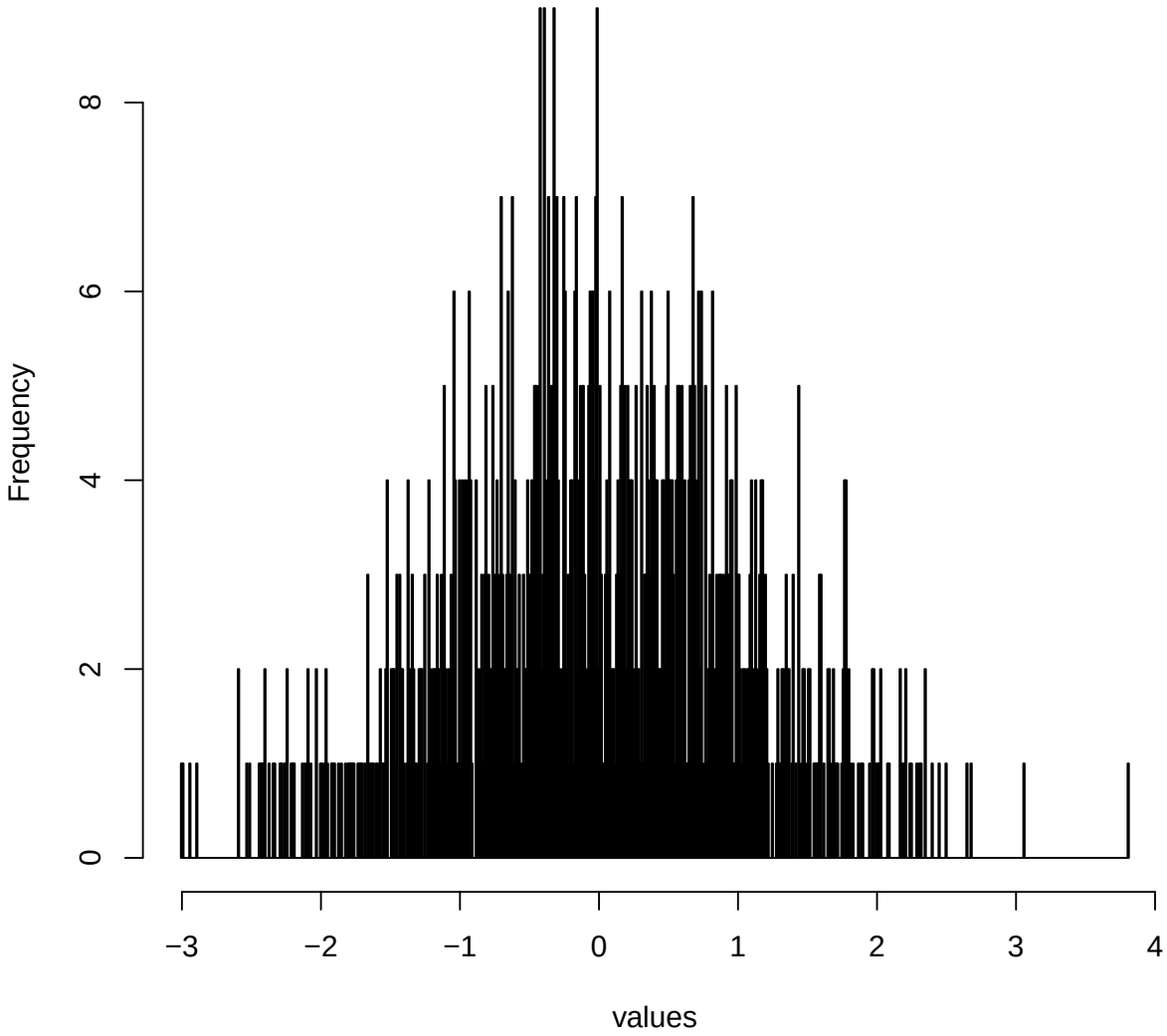
Experimental CDF



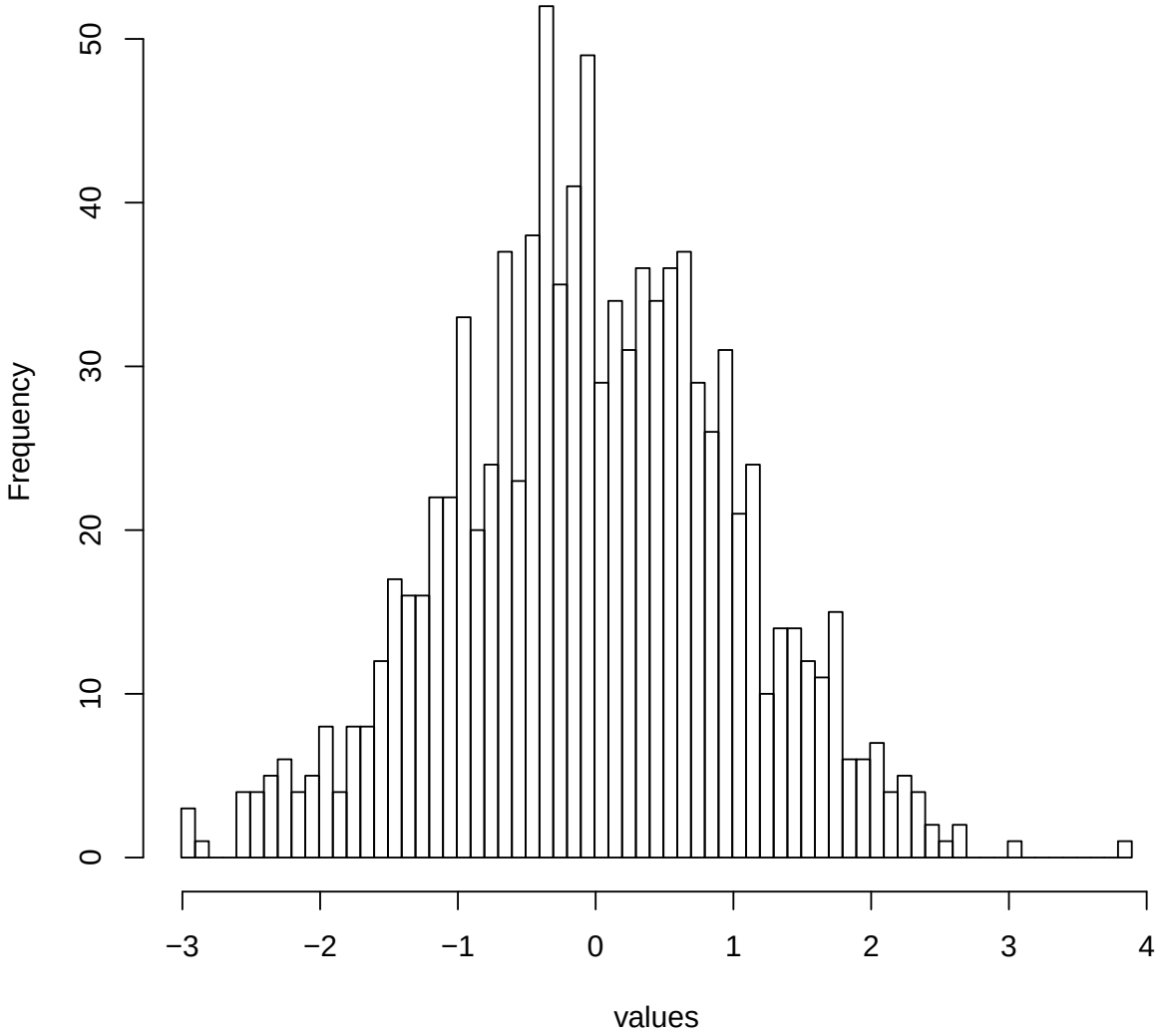
Histogram, bin=0.001



Histogram, bin=0.01



Histogram, bin=0.1



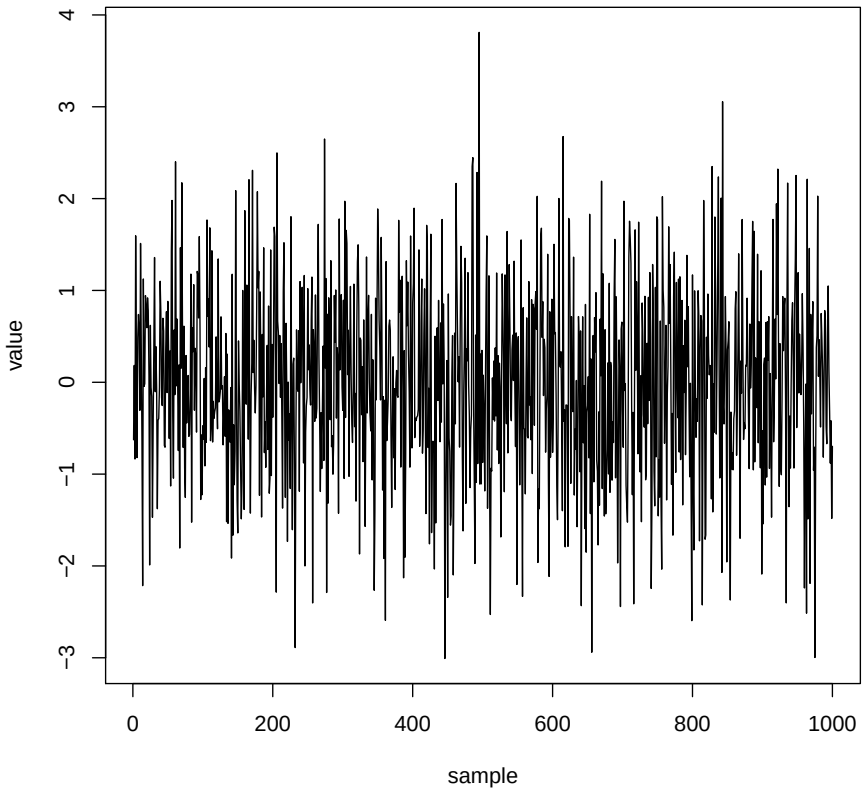
UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

Given two datasets $\{x_i\}$, say A and B, are straightforward visualizing means enough to distinguish or understand them?

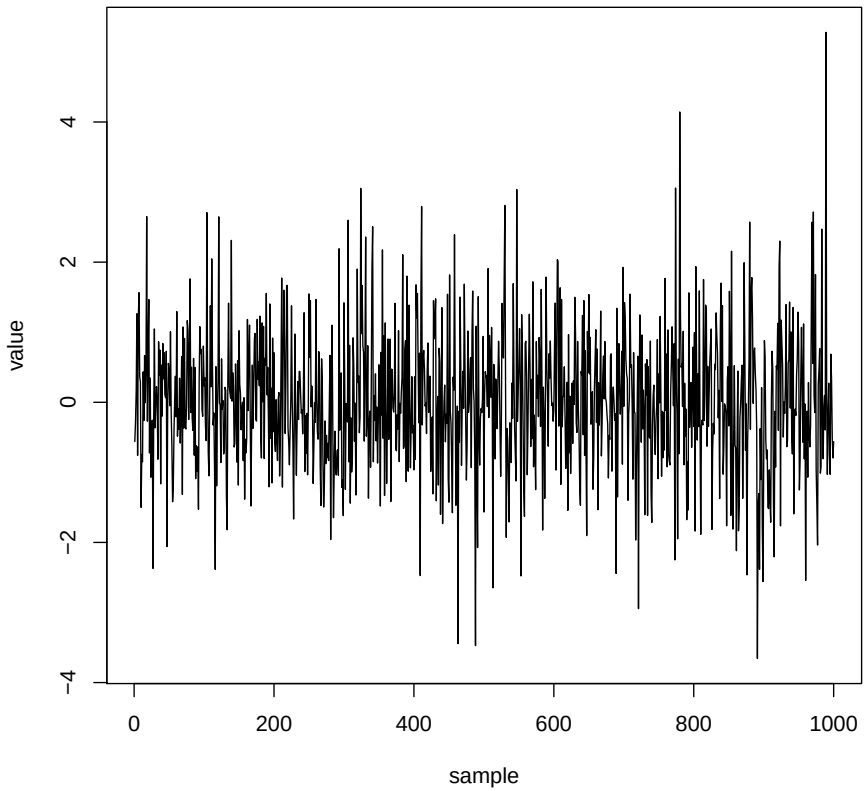
- Straight plots of the indexed data?
- The plot ordered by the X_i values?
- The cumulative distribution: Experimental CDF (ECDF)?
- Or histograms of the relative frequency of data?

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

Can we distinguish different distributions?



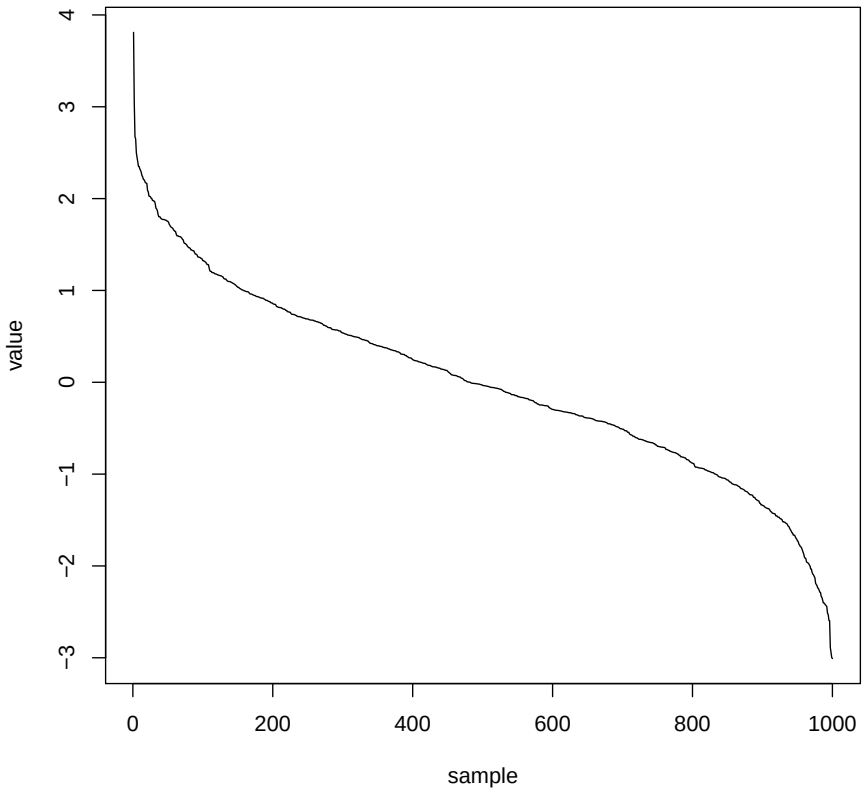
Distribution A



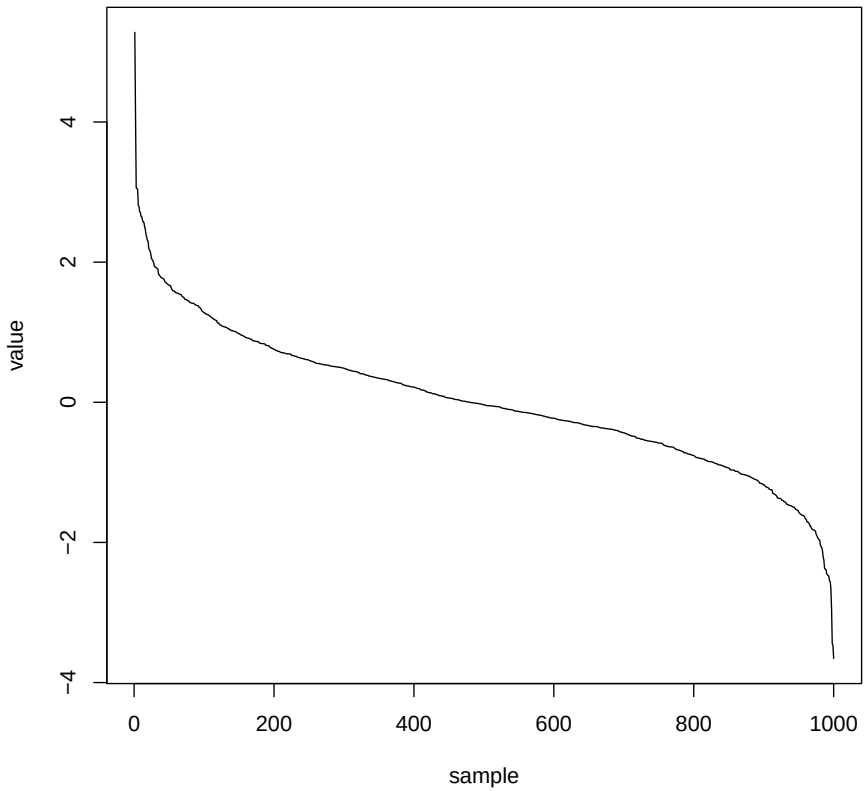
Distribution B

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

Can we distinguish different distributions?



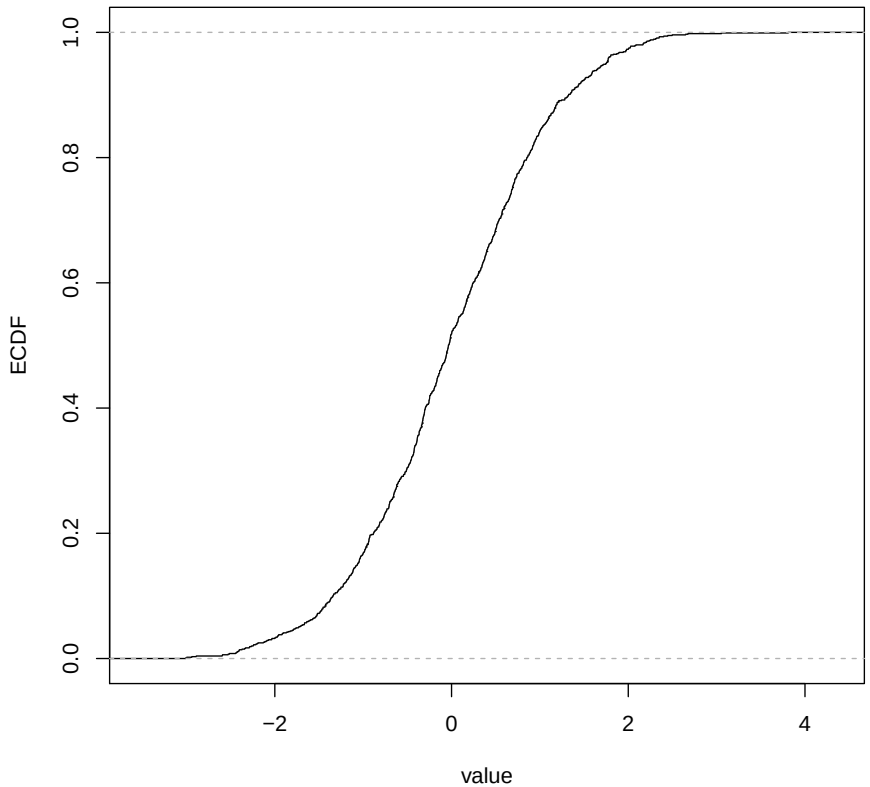
Distribution A



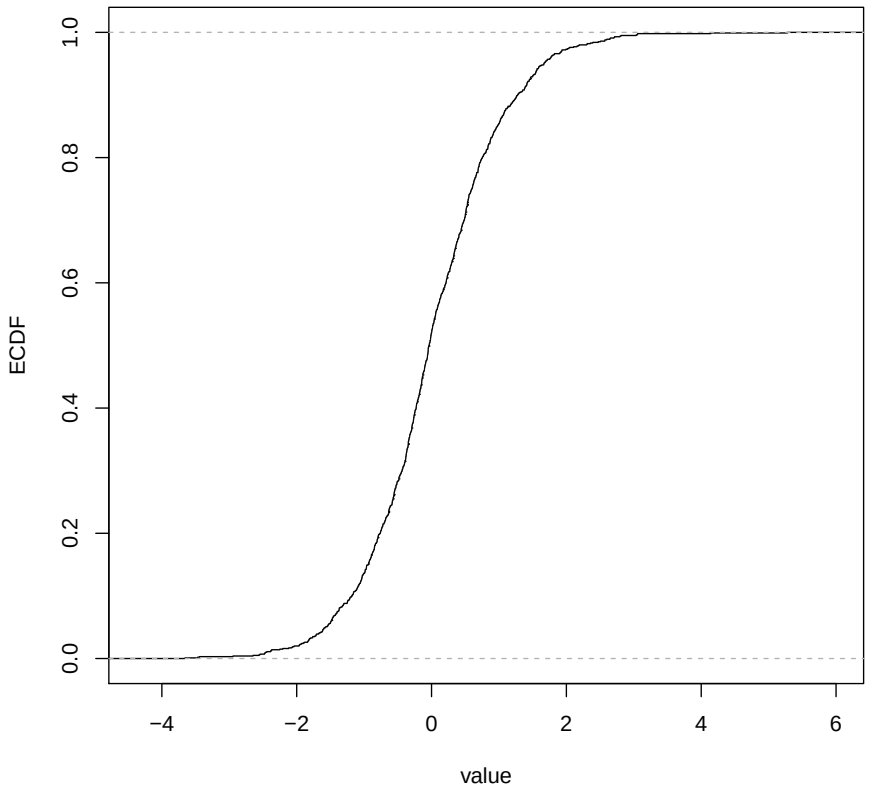
Distribution B

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

Can we distinguish different distributions?



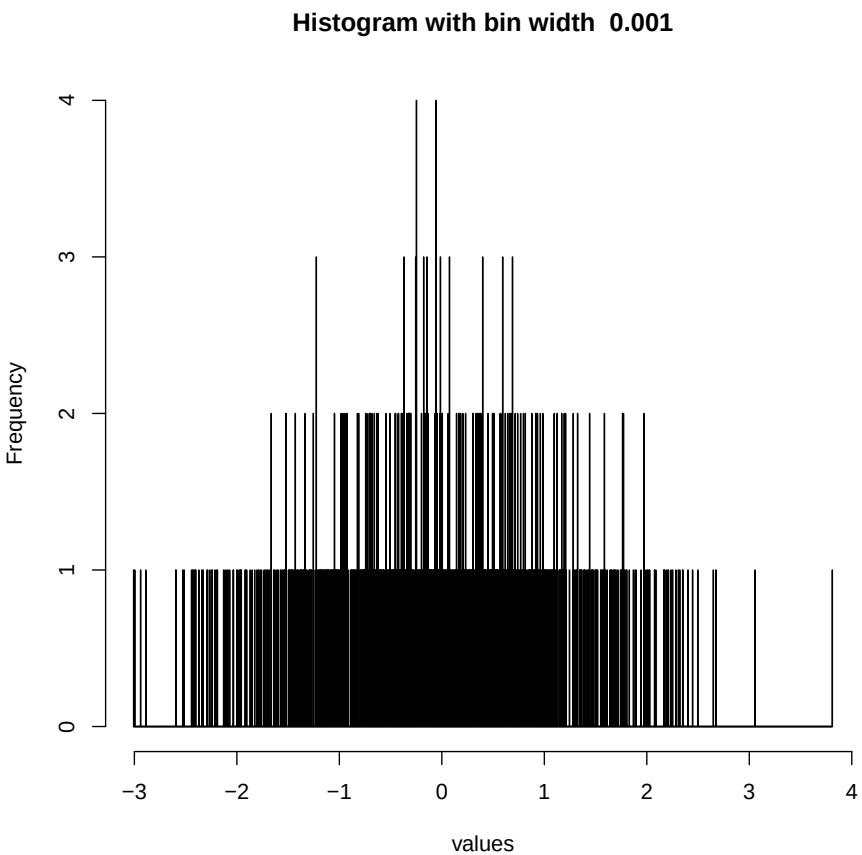
Distribution A



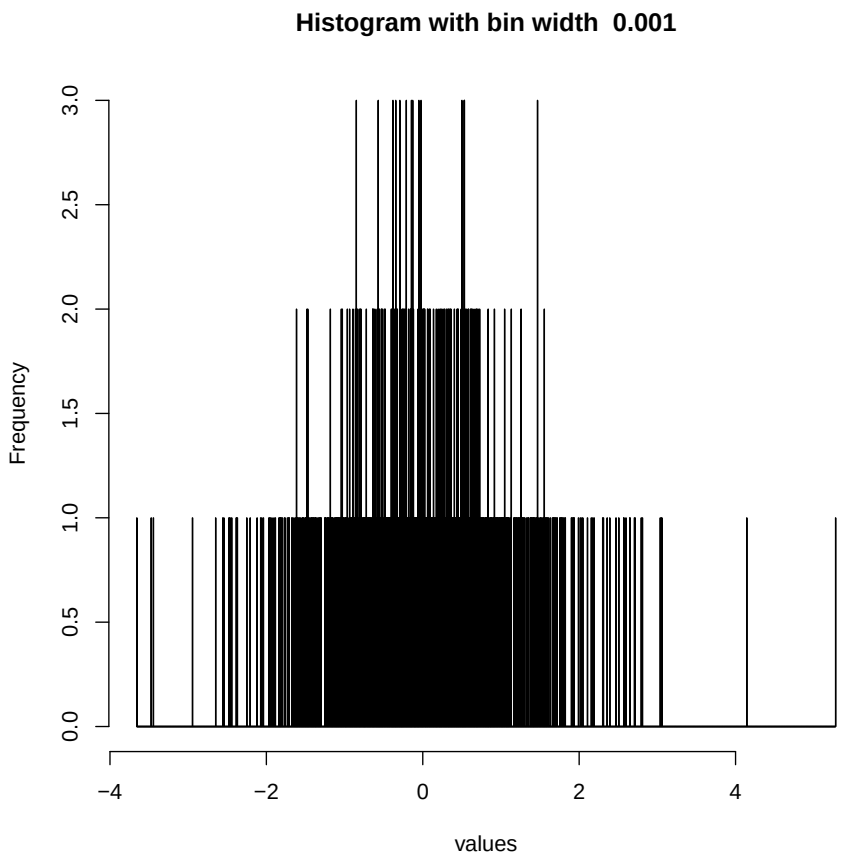
Distribution B

Histogram, bin=0.001

Can we distinguish different distributions?



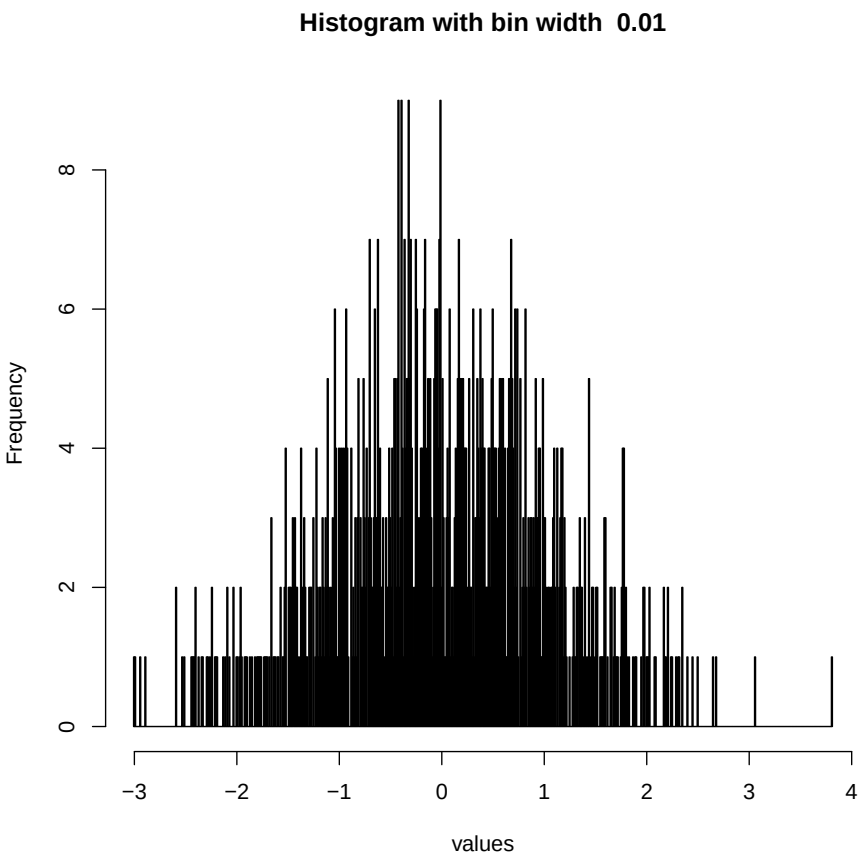
Distribution A



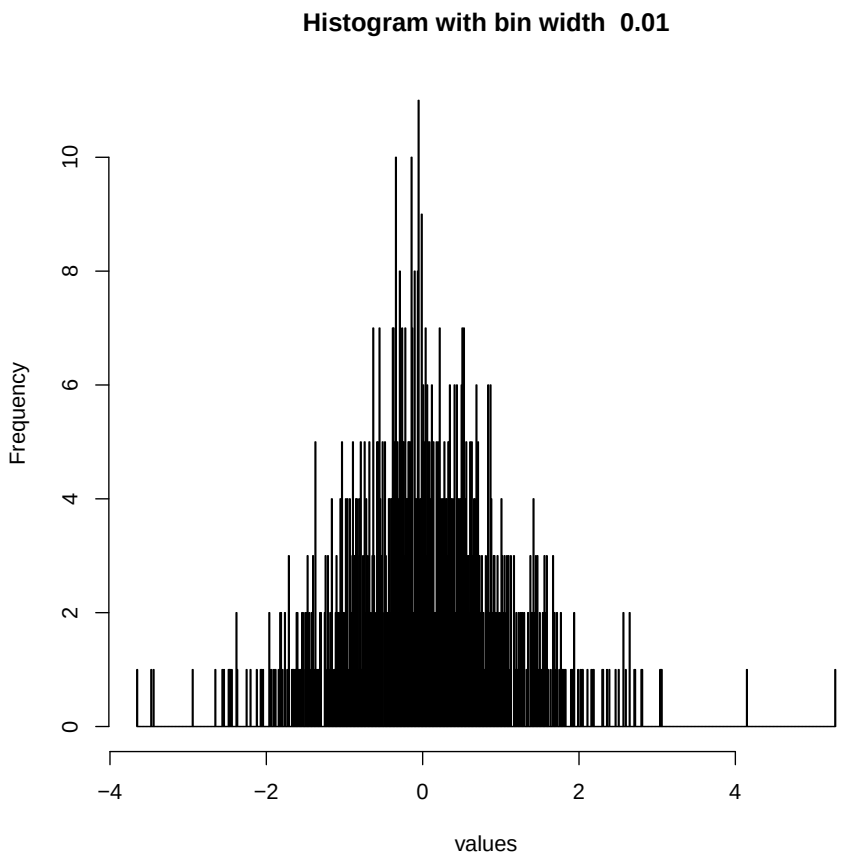
Distribution B

Histogram, bin=0.01

Can we distinguish different distributions?



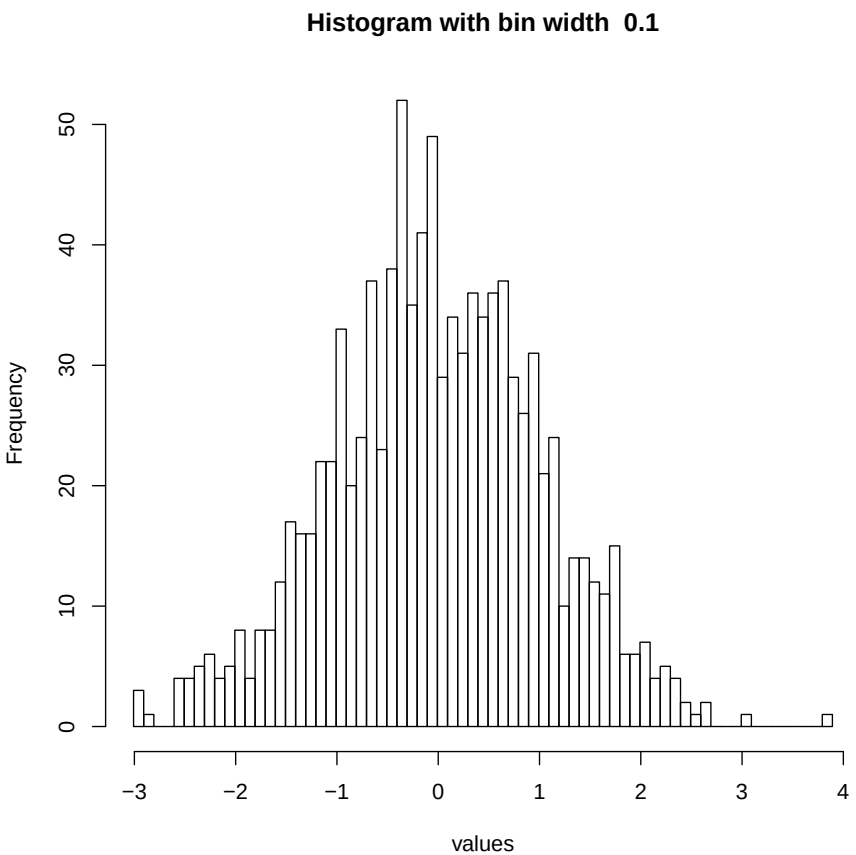
Distribution A



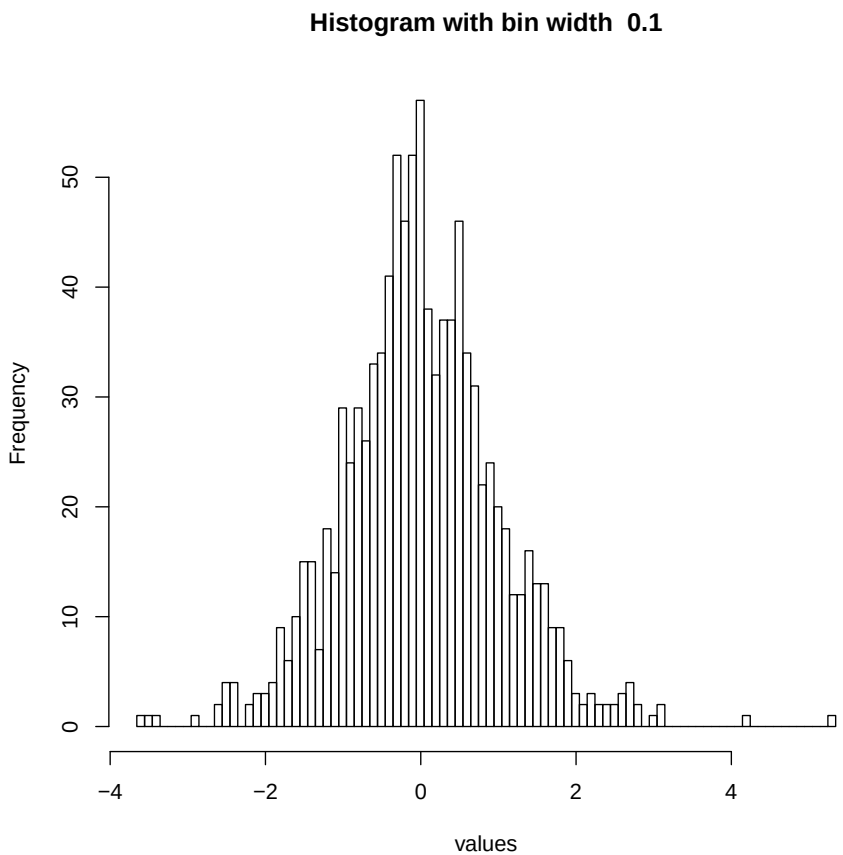
Distribution B

Histogram, bin=0.1

Can we distinguish different distributions?



Distribution A



Distribution B

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- It is clear that straightforward visualization can help, but is it is not enough to understand characteristics of the population given a sample from an experiment
- Histograms are in general more informative than other plots
- ECDFs can help comparing different samples "quickly"
- They are "quick & dirty" means to have a first insight in the problem and in devising better data collection for a deeper analysis
- Computing parameters and functions of the sample, as the means, variance, etc. is a further step to understand our measures

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

Given a sample $\{x_i\}$ we want to gain insight in the population that generated it

- The population can be normally described with a SP $\{X(t,s)|s \in S,\ t \in T\}$
- Insight is given by parameters of the population as the mean, variance, etc.
- Let be θ the parameter to be evaluated
- We are interested in computing an estimate $\hat{\Theta}(\{x_i\})$, which is representative of the real function $\Theta(\{X(t)\})$ that compute θ
- We call the estimator $\hat{\Theta}$ **unbiased** if

$$E[\hat{\Theta}(\{x_i\})] = \theta$$

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- The mean of the sample is

$$\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- $\overline{X}$ is an unbiased estimator of the population mean μ (if the SP representing the population is wide-sense stationary and μ exists), i.e.

$$E[\overline{X}] = \mu$$

- Homework: Prove that that $\bar{X}$ is unbiased

- Thanks to the linearity of the average operator we can compute $\overline{X}$ in *batches* splitting the sample of dimension n in k smaller subsets

$$\overline{X} = \frac{1}{k}\sum_{i=1}^{k}\left[\frac{k}{n}\sum_{j=1}^{n/k} x_{(ki+j)}\right] = \frac{1}{k}\sum_{i=1}^{k}\left[\frac{k}{n}\overline{X_i}\right]$$

- This "trick" can greatly reduce numerical problems
- This method is very useful to compute the "reliability" of the estimation, i.e., compute confidence intervals and levels because the $\overline{X_i}$ are by construction Gaussian RVs with good approximation if k/n is sufficiently large

Accuracy of $\overline{X}$

What is the accuracy of the estimator $\overline{X}$ as the sample size increases?

- Let's compute the variance of $\overline{X}$

$$\begin{aligned} \mathrm{Var}[\overline{X}] &= \sum_{i=1}^{n} \mathrm{Var}[X_i/n] = \frac{n\,\mathrm{Var}[X_i]}{n_2} \\ &= \frac{\mathrm{Var}[X]}{n} = \frac{\sigma^2}{n} \end{aligned}$$

- The quality of the estimation improves hyperbolically with the sample size
- As usual σ must exist and be finite

Variance of a dataset

- We define the variance of a dataset $\{x_i\}$ of size n

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

- Why $n-1$? Basically because we have used one degree or freedom to estimate $\overline{X}$, if we can use the true mean μ then we should use n to have an unbiased estimator ...
  but μ is normally not known ...

Proof that S^2 is unbiased

Expanding the square binomial we have

$$
\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{X} + \overline{X}^2) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2\right) - \frac{2n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\overline{X} + \frac{n}{n-1}\overline{X}^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\overline{X}^2
\end{aligned}
$$

Proof that S^2 is unbiased

Taking the average of S^2 results in

$$E[S^2] = \frac{1}{n-1}\sum_{i=1}^{n} E[X_i^2] - \frac{n}{n-1}E[\overline{X}^2] \quad (1)$$

but we also have that

$$E[X_i^2] = \mathrm{Var}[X_i] + (E[X_i])^2 = \sigma^2 + \mu^2 \quad (2)$$

$$E[\overline{X}^2] = \mathrm{Var}[\overline{X}] + (E[\overline{X}])^2 = \frac{\sigma^2}{n} + \mu^2 \quad (3)$$

and substituting (2) and (3) in (1)

$$E[S^2] = \frac{1}{n-1}n(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu\right) = \sigma^2 \quad (4)$$

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- If the population is known to be finite of size M, and assuming sampling without replacement the variance of a dataset $\{x_i\}$ of size n is

$$S^2 = \frac{1 - \frac{1}{M}}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2$$

- Also this estimator is unbiased, but we do not prove it ...

Additional parameters

- Mean and variance already tell much about a system
- But how can we distinguish between different populations with the same mean and variance?
- Functions of higher moments can help
- The third moment estimates how asymmetric is a distribution (at least for mono-modal ones)
- The fourth moment is a good estimator of how "peaked" is our population around the mean
- There are many functions of third and fourth moment
  ... normally all called skewness and kurtosis ... we give here two definitions taken from the NIST Statistical Handbook
  `http://www.itl.nist.gov/div898/handbook/index.htm`

Skewness of a dataset

- We define skewness $\mathrm{Sk}(\cdot)$ of a sample $\{x_i\}$

$$\mathbf{G} = \mathrm{Sk}(\{x_i\}) = \frac{1}{nS^3}\sum_{i=1}^{n}(x_i - \overline{X})^3$$

- S here should be computed as $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$

  to ensure that $\mathbf{G}$ is normalized to 1, but the difference with the computation of S with $n-1$ at denominator is marginal if n is sufficiently large

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- We define kurtosis $\mathrm{Ku}(\cdot)$ of a sample $\{x_i\}$

$$\mathsf{K} = \mathrm{Ku}(\{x_i\}) = \frac{1}{nS^4}\sum_{i=1}^{n}(x_i - \overline{X})^4 - 3$$

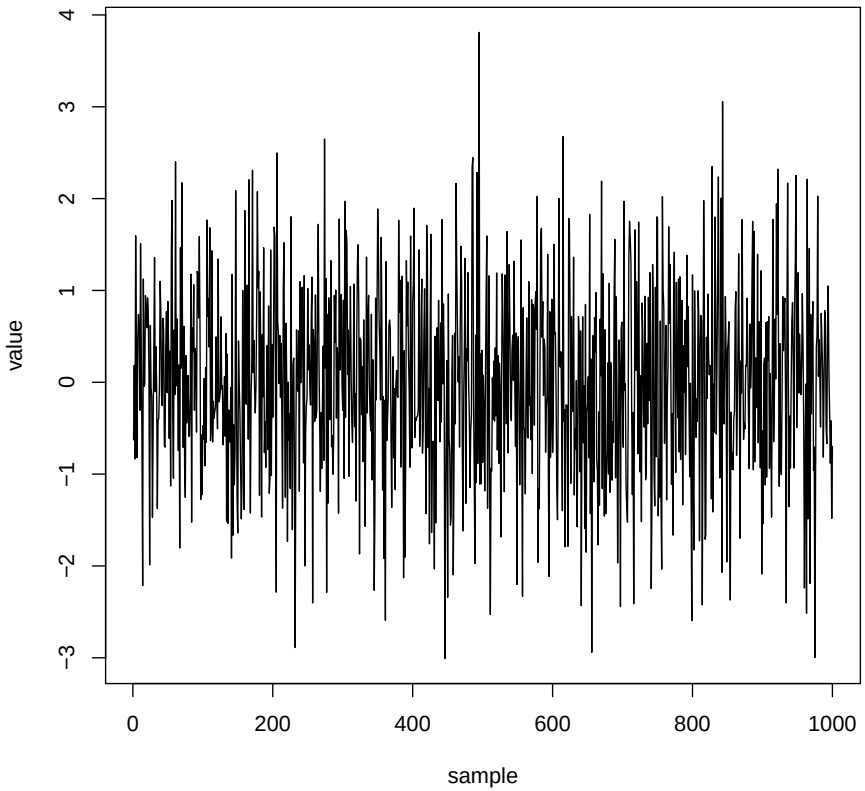
- Also in this case S should be computed with $n-1$ at denominator for normalization reasons
- The "-3" is a normalization: the non-normalized kurtosis of a Gaussian with $\sigma = 1$ is exactly 3, thus with this definition of kurtosis we have a quick comparison with a normal distribution

Properties of G and K

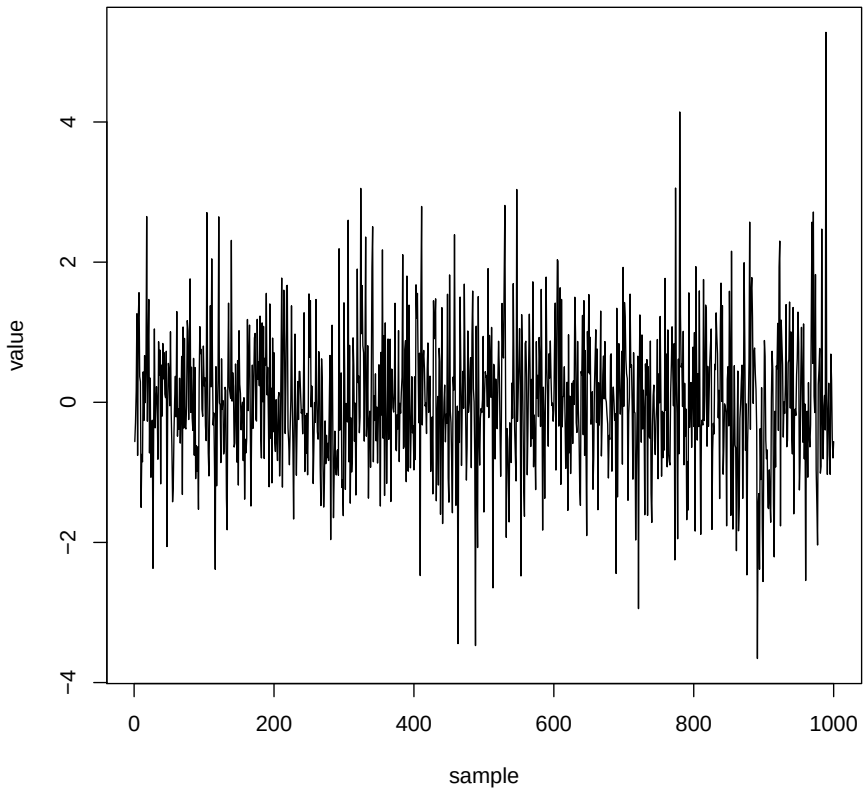
- A dataset with $\mathbf{G}\{x_i\} > 0$ indicates that the mode is larger than the mean (skewness right)
- A dataset with $\mathbf{G}\{x_i\} < 0$ indicates that the mode is smaller than the mean (skewness left)
- A dataset with $\mathbf{K}\{x_i\} > 0$ indicates that the population (distribution) is "peaked" (compared to a Gaussian with $\sigma = 1$), i.e., it is more concentrated around the mode
- A dataset with $\mathbf{K}\{x_i\} < 0$ indicates that the that population "flat", i.e., it is less concentrated around the mode than a standard Gaussian
- Notice that with the same variance a **more peaked distribution has a slower decay** of the distribution tails

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

How can we distinguish these two distributions?



Distribution A



Distribution B

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

	Distribution A	Distribution B	
$\overline{X}$	+0.954	+1.064	?
S^2	+1.043	+1.021	??
G	-0.035	+0.026	???
K	+0.017	+1.393	**!!!!!!**

A and B datasets differs for some shape parameter related to the fourth and possibly higher moments, with set B being more peaked, thus also with longer tails . . . which however may be difficult to see with just 1000 samples

A is Gaussian (maybe) . . . and B?

- Indeed even with 4 parameters it is not possible to identify with precision the population distribution
- We can make "educated guess" and then make some hypothesis testing (coming later)
- For the time being accept that it is a logistic distribution with $\mu = 1$ and $\sigma = 1$

$$f_X(x) = \frac{e^{-\frac{x-\mu}{s}}}{s(1+e^{-\frac{x-\mu}{s}})^2}$$

- And the logistic distribution is **very** different form the Gaussian as it has longer tails; $\sigma^2 = \frac{\pi^2 s^2}{3}$

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- The analysis so far is fine and correct, but tells us nothing about the memory of the underlying process
- We can use the autocorrelation function . . . but how we compute it on a dataset $\{x_i\}$?
- Let's assume for the time being that the underlying process is wide-sense stationary and recall that

$$R(\tau) = E[X(t) \cdot X(t+\tau)]$$

- If we let τ sweep all the samples, then we have just products of samples of RVs . . . too noisy!

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

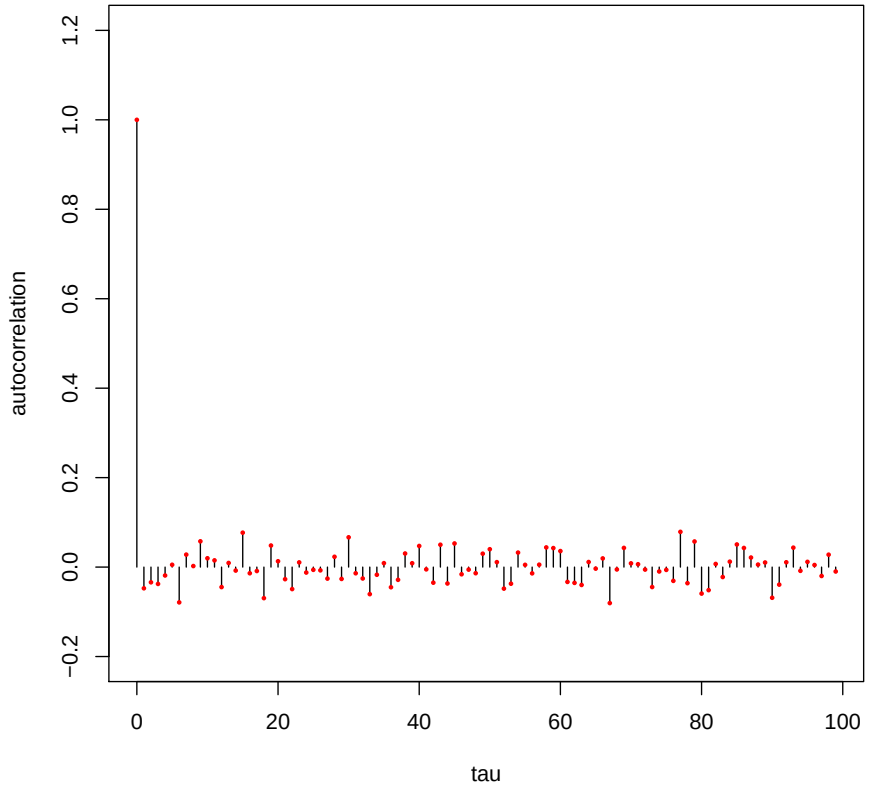
Given a dataset $\{x_i\}$ of size n

- Limit τ variation to a reasonable limited value $\tau_{\max} << n$, which is the "window" where we estimate the autocorrelation
- Then we can evaluate the *sample covariance* as the average of all the $n - \tau_{\max}$ possible couples of samples at distance $0 \leq \tau \leq \tau_{\max}$

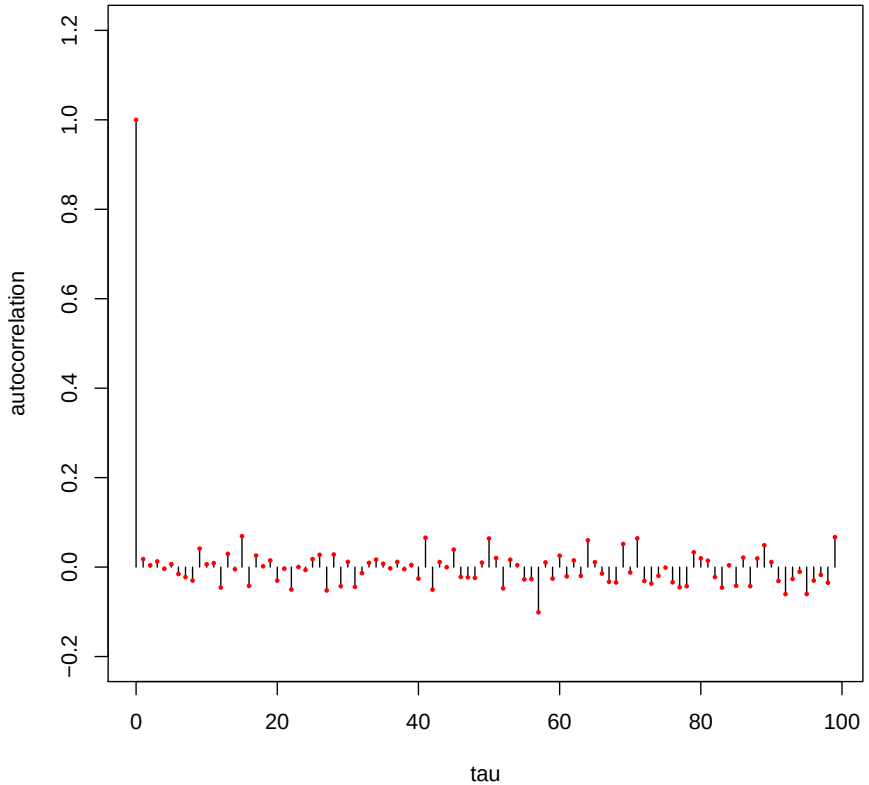
$$\text{Covs}(x_i, x_{i+\tau}) = \frac{1}{n - \tau_{\max}} \sum_{j=1}^{n-\tau_{\max}} (x_j - \overline{X}) \cdot (x_{j+\tau} - \overline{X})$$

- and normalizing with respect to S^2 we obtain the sample normalized autocovariance $R'(\tau) = \dfrac{\text{Covs}(x_i, x_{i+\tau})}{S^2}$

Autocorrelation of A & B datasets



Distribution A



Distribution B

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- We can state that both datasets A and B derive from SPs that are independent
- The variations of $R'(\tau)$ around 0 for $\tau > 0$ are random variations that decay as n grows
- To ensure perfect normalization of the autocorrelation function it is customary to compute it as

$$R'(\tau) = \frac{\text{Covs}(x_i, x_{i+\tau})}{\text{Covs}(x_i^2)}$$

- Another test that is often useful is verifying that $\overline{X}$ is time independent, e.g., with a sliding window or batch means

Is independence common?

- One might ask what are the systems and SPs that contain memory
- We already know that Markovian processed do have memory
- An example (CSDT because we sample) is the position estimated by a GPS receiver
- Each sample contains additive noise which is roughly Gaussian, but it is added on the previous estimate of the position not to the true position
- Hence the position error contains memory, which can be successfully modeled as a Markov-Gauss process

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

- The position error of a GPS receiver is successfully modeled by the following process
  - Indeed one independent process for (x, y, z), with the vertical error (z) larger than the horizontal error, but we model only one

$$\{X_t : t \in T\}; \quad f_{X_t|X_{t-1}}(x) = e^{\frac{\Delta_t}{T}} x_{t-1} + N(0, \sigma_n)$$

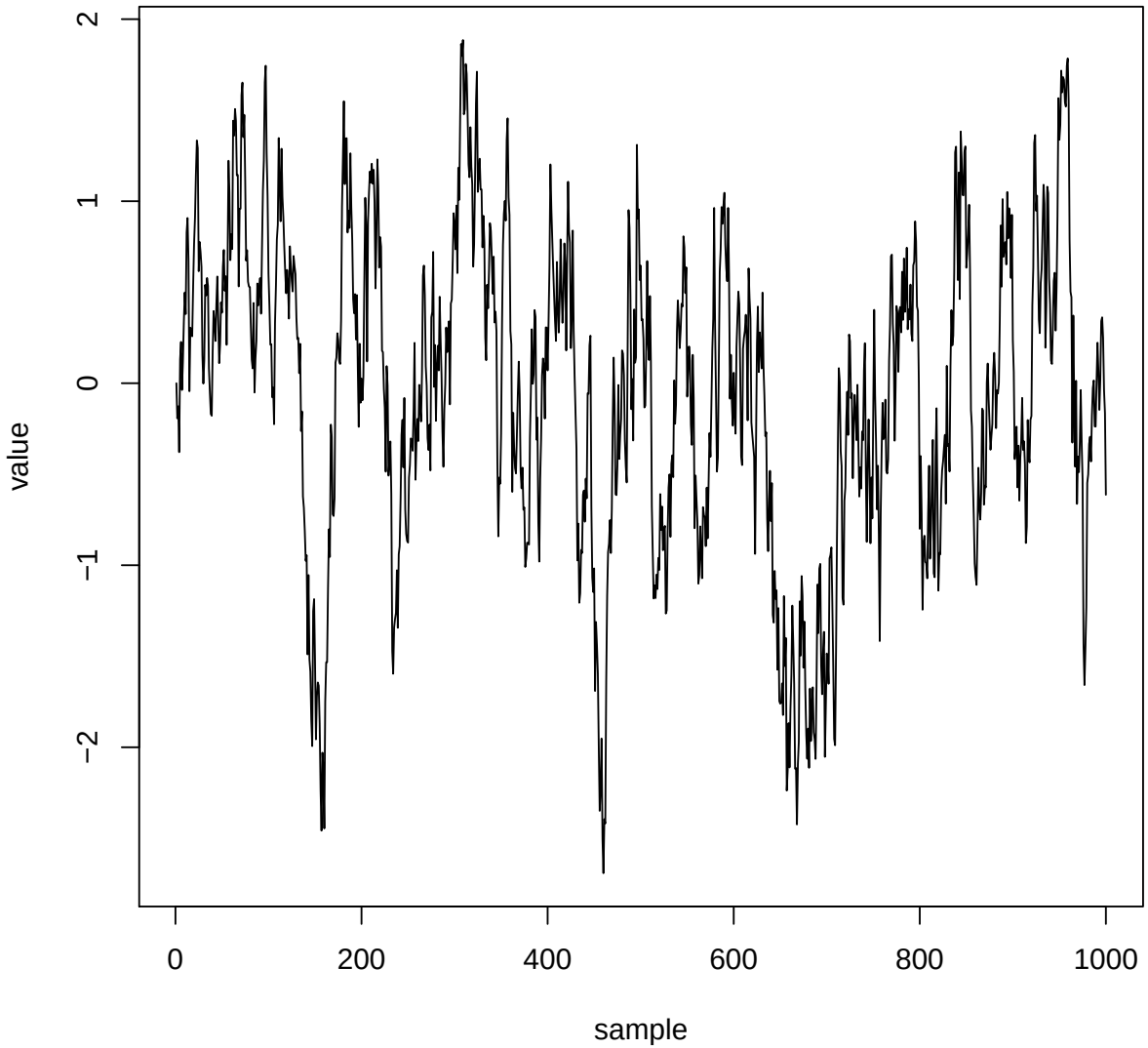
  where Δ_t is the sampling time, T the actual memory of the process and $N(0, \sigma_n)$ a Gaussian RV σ_n depends on the quality of the receiver, but also on Δ_t
- The mean of the process is $\mu = 0$, the variance is

$$\sigma^2 = \frac{\sigma_n^2}{1 - e^{-2\frac{\Delta_t}{T}}}$$

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$\sigma_n = 1$
$T = 0.2\text{s}$
$\Delta_t = 0.01\text{s}$

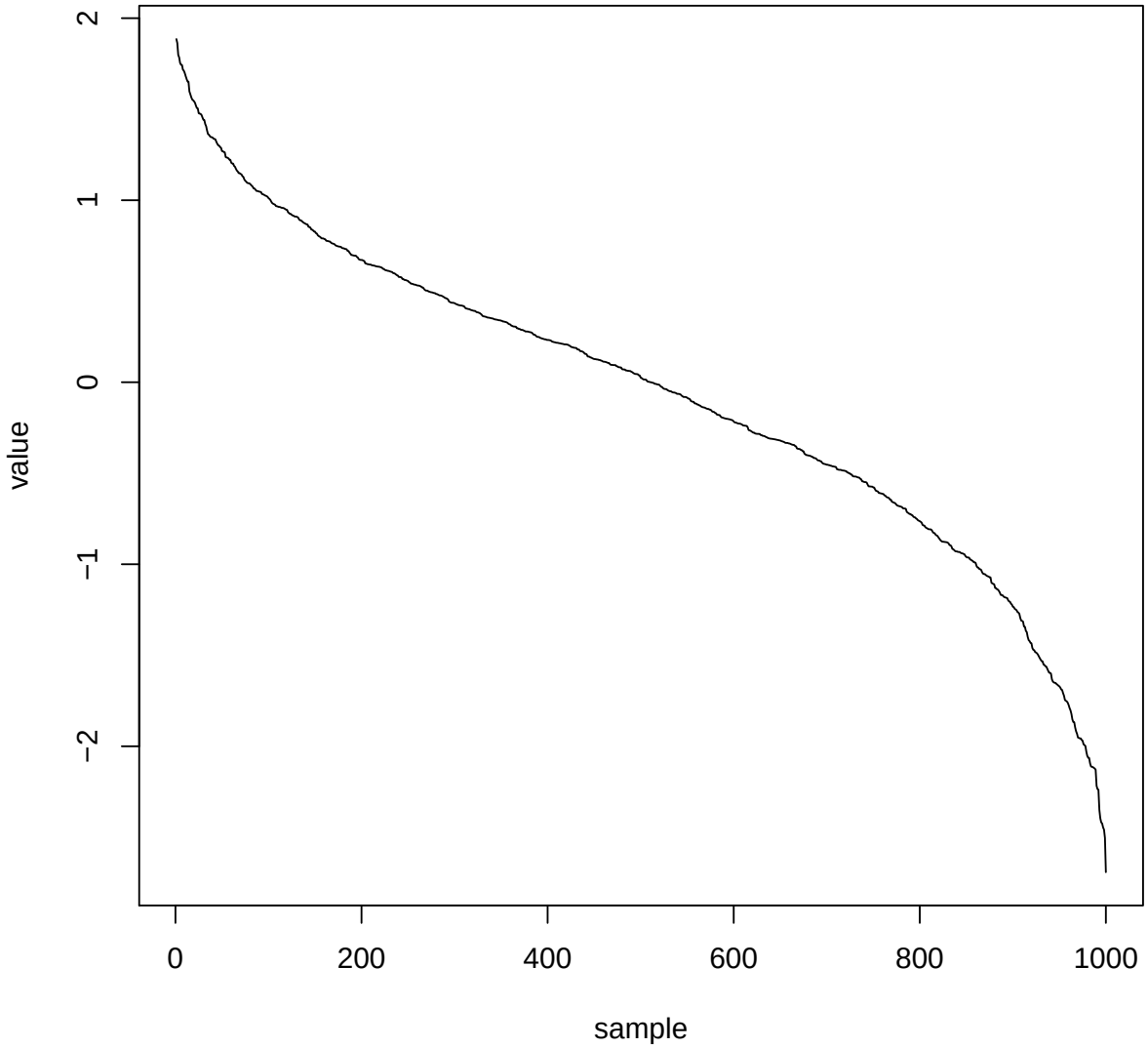
Visualization of the samples



UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$\sigma_n = 1$
$T = 0.2\text{s}$
$\Delta_t = 0.01\text{s}$

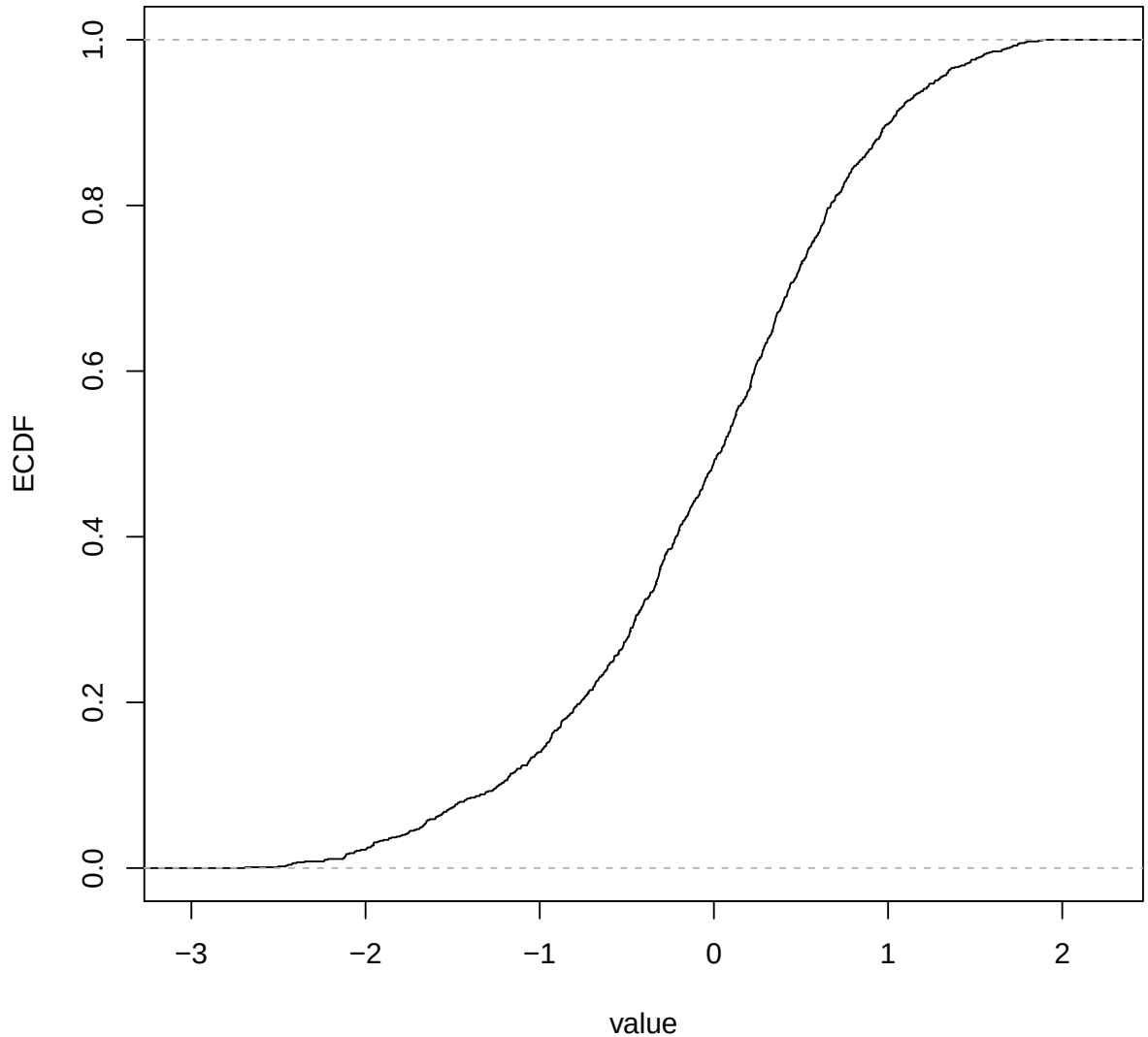
Ordered samples



UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$\sigma_n = 1$
$T = 0.2\text{s}$
$\Delta_t = 0.01\text{s}$

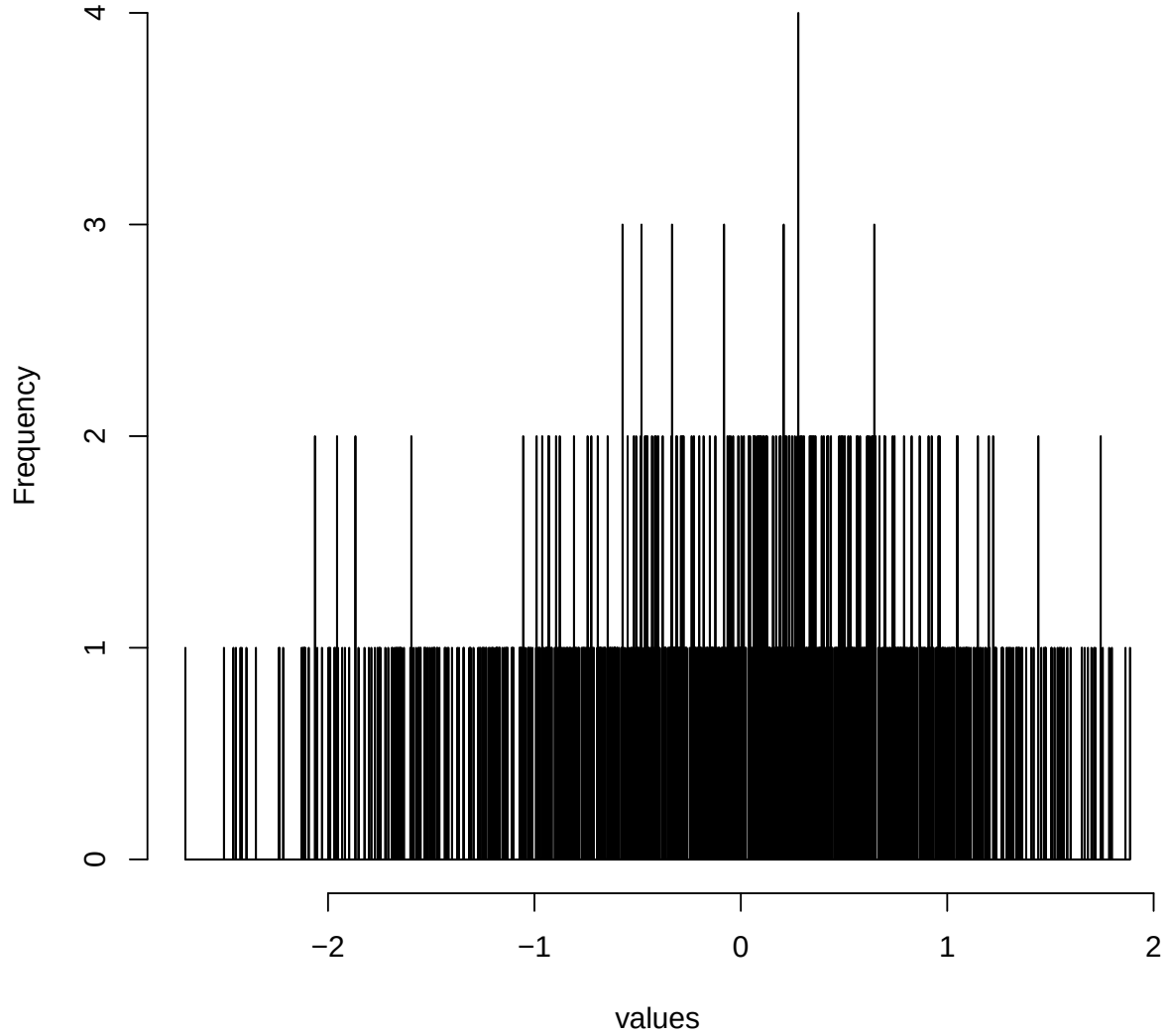
ECDF



UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$\sigma_n = 1$
$T = 0.2$s
$\Delta_t = 0.01$s

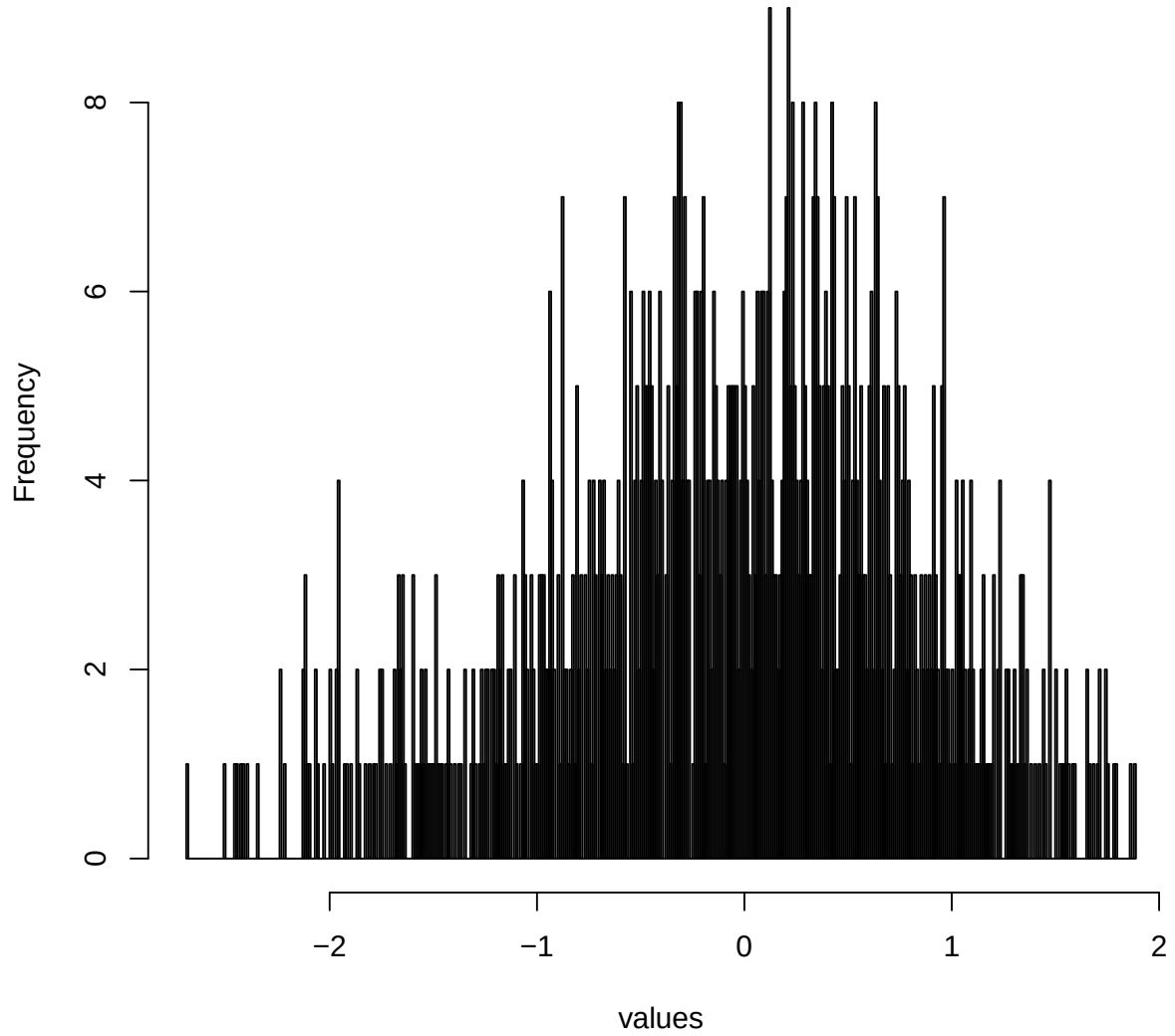
Histogram:
bin $= 0.001$



UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$\sigma_n = 1$
$T = 0.2$s
$\Delta_t = 0.01$s

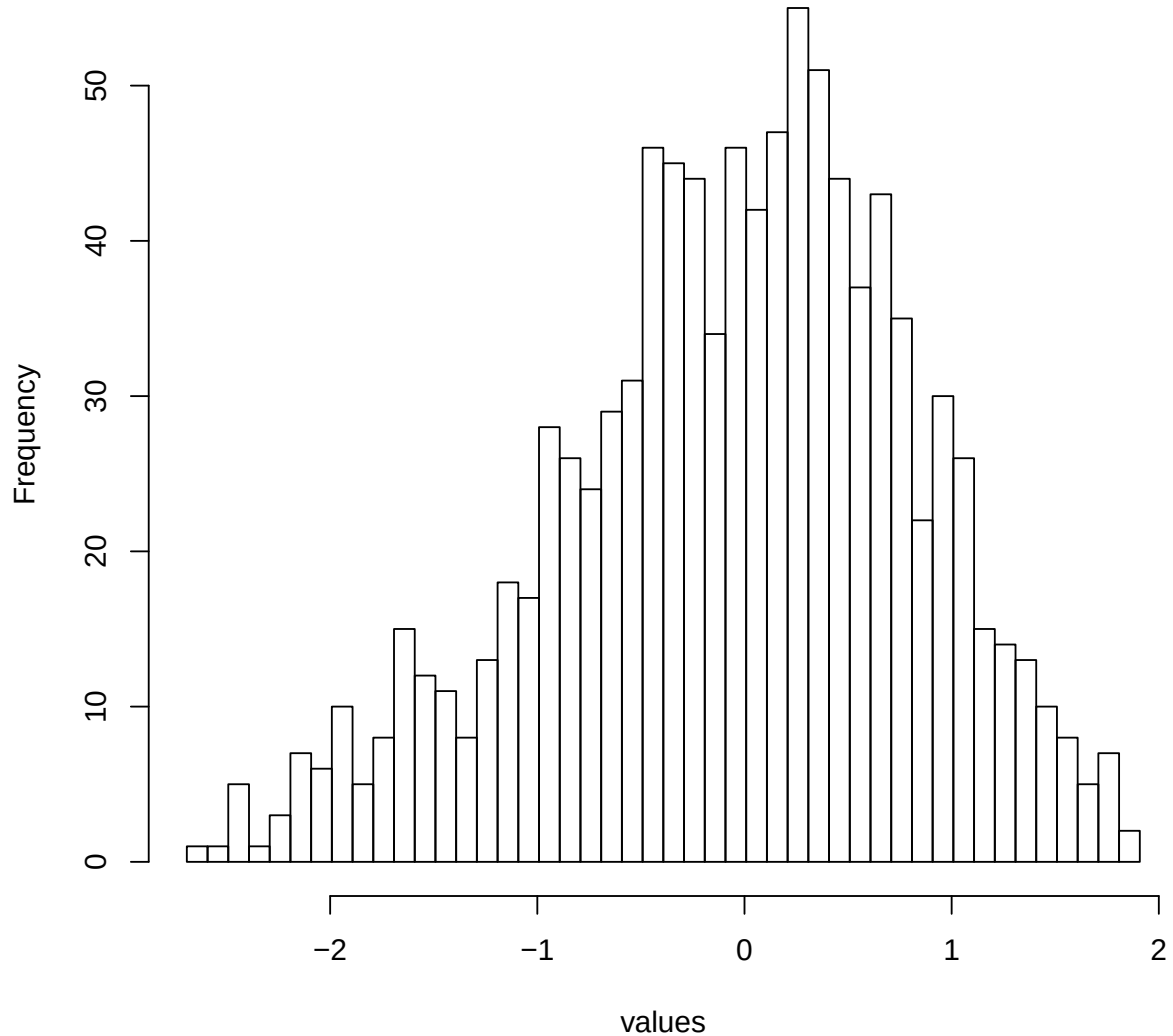
Histogram:
bin $= 0.01$



UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$\sigma_n = 1$
$T = 0.2\text{s}$
$\Delta_t = 0.01\text{s}$

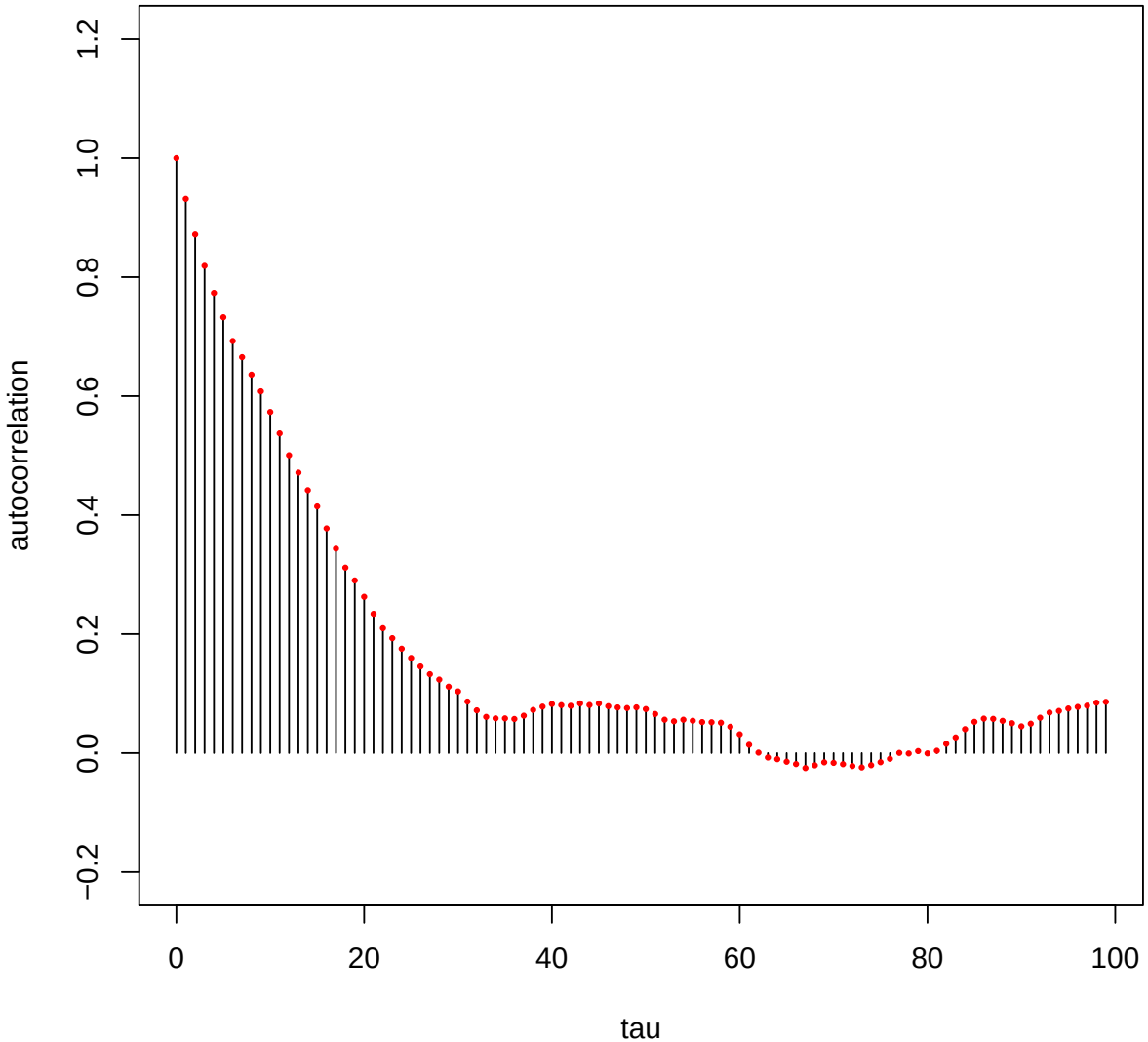
Histogram:
bin $= 0.1$



UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$\sigma_n = 1$
$T = 0.2\text{s}$
$\Delta_t = 0.01\text{s}$

Autocorrelation



What is a Confidence Interval?

- Once we have estimated some parameter, for instance the mean $\overline{X}$, of a dataset, what is the confidence we have in this estimation, how much is it representative of the real value?
- We know that if the estimator is unbiased, then "on average" our estimation is correct
- We also know that if we have an estimator $\hat{\sigma}$ of the population standard deviation σ then

$$\text{Var}[\overline{X}] = \frac{\hat{\sigma}}{n}$$

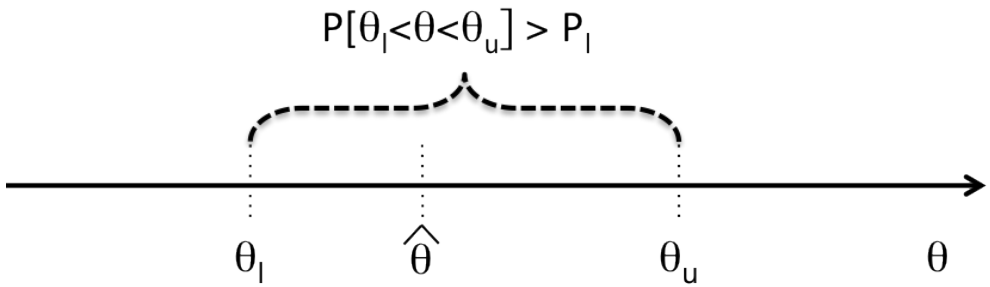
  where n is the number of samples
- This is a very strong knowledge . . . let's see why

What is a Confidence Interval?

- We define **confidence interval** around the estimated value $\hat{\theta}$ the interval (θ_l, θ_u) such that the true value θ falls within the interval (θ_l, θ_u) with a given probability $\mathbf{P}_l$ that we call the **confidence level**

$$\mathbf{P}[\theta_l \leq \theta \leq \theta_u \,|\, \hat{\theta}] \geq \mathbf{P}_l$$

- Often (θ_l, θ_u) is expressed as a fraction (percentage) of $\hat{\theta}$ around $\hat{\theta}$, assuming symmetry (which is not necessarily true)
- E.g., a confidence interval of $\pm 5\%$ with a confidence level $\mathbf{P}_l = 99\%$



Chebychev inequality

- Relates the probability of the outcome of any RV to fall within a given boundary as a function of the RV variance
- Gives by definition a symmetric interval ϵ, as it is a function of a single parameter
- It states that given and RV X with mean μ and standard deviation σ

$$\mathbf{P}[\mu - \epsilon < X < \mu + \epsilon] \geq 1 - \frac{\sigma^2}{\epsilon^2}$$

- In other words it relates the probability of an outcome being farther than a given amount from the average value as a function of the RV variability (σ) and the amount itself

- It is possible to express ϵ as a function of σ: $\epsilon = k\sigma$

$$\mathbf{P}[\mu - k\sigma < X < \mu + k\sigma] \geq 1 - \frac{\sigma^2}{k^2\sigma^2} = 1 - \frac{1}{k^2}$$

- Clearly this inequality can be used to state confidence in an estimation . . . it expresses both a confidence interval and a confidence level, but to have a high level with a small interval it is necessary to have a very small σ

- $\overline{X}$ can in itself be interpreted as an RV
  - Other measures from the same population would yield different values of $\overline{X}$
- We also know that the variance of the RV $\overline{X}$ is $\mathrm{Var}[\overline{X}] = \frac{\sigma^2}{n}$ (n is the size of the sample)
- Thus we can estimate it using the estimate s of σ computed on the sample and rewrite the Chebycheff inequality as

$$\mathbf{P}[\mu - ks < X < \mu + ks] \geq 1 - \frac{1}{k^2}$$

- Letting $\epsilon = ks;\ k = \frac{\epsilon}{s} \simeq \frac{n\epsilon}{\sigma}$

$$\mathbf{P}[\mu - \epsilon < X < \mu + \epsilon] \geq 1 - \frac{s^2}{\epsilon^2} \simeq 1 - \frac{\sigma^2}{n\epsilon^2}$$

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

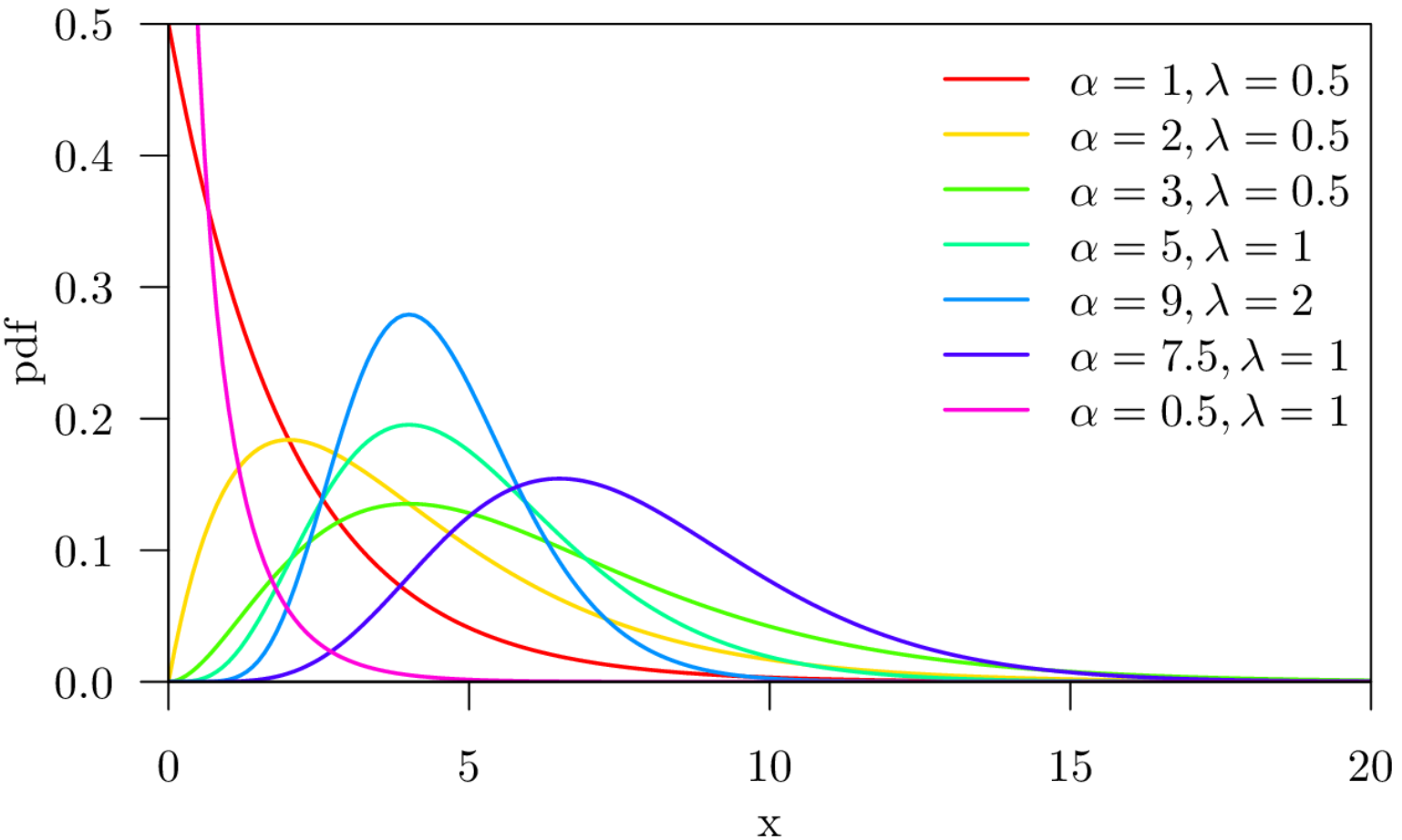
It is a distribution of the exponential family with the following pdf

$$f_X(x) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)};\ \alpha, \lambda, x > 0$$

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$$

UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$f_X(x)$ in linear scale for various λ, α



UNIVERSITY OF TRENTO
Department of Information Engineering and Computer Science

$f_X(x)$ in log scale for various λ, α