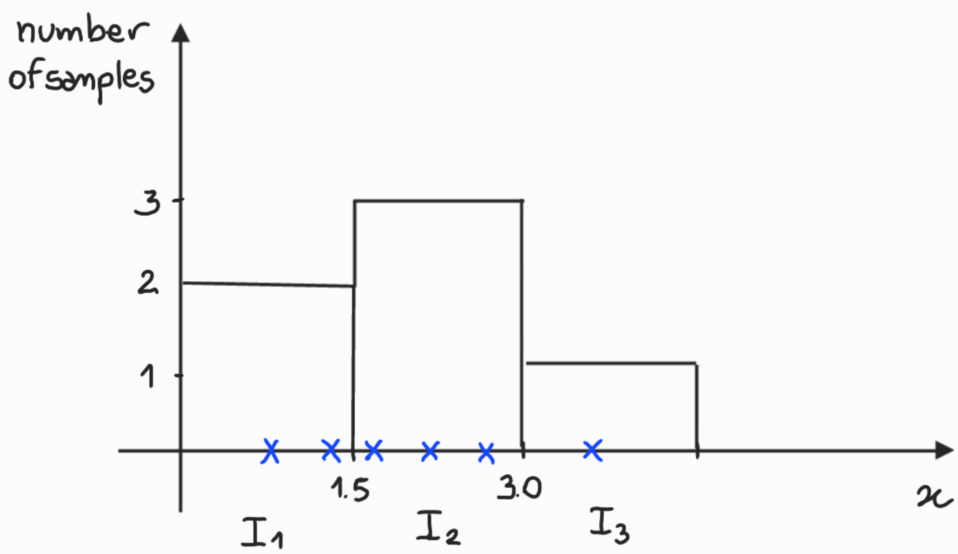
- ## Histograms and empirical pdf

Given a set of observations $\{x_1, x_2, \ldots, x_N\}$, an interval I which contains all the observations, and a partition $\{I_j\}_{j=1}^{J}$ of I (i.e. $\bigcup_{j=1}^{J} I_j = I$, $I_i \cap I_j = \emptyset$ if $i \neq j$), a histogram is a graphical representation of the number of observations falling in each interval I_j.

EXAMPLE

Consider the observations $\{0.8, 1.4, 1.6, 2.1, 2.8, 3.6\}$ and the intervals: $I_1 = [0, 1.5)$, $I_2 = [1.5, 3.0)$, $I_3 = [3.0, 4.5)$.

The corresponding histogram looks as follows:



Now, assume that $x_1, x_2, \dots, x_N$ are independent observations of the continuous random variable X.

An estimate of $P(X \in I_j)$ is obtained as:

$$\widehat{P(X \in I_j)} = \frac{N_{\{X \in I_j\}}}{N}$$

Number of samples x_i that belong to I_j

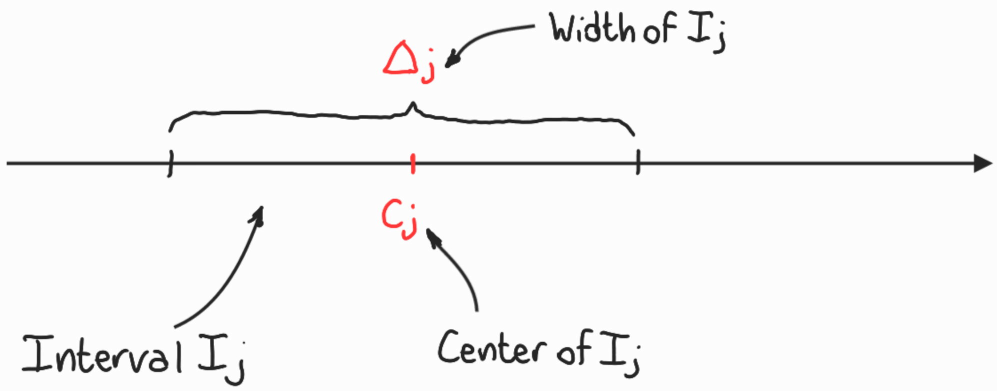
$\Downarrow$

The quantity represented in the histogram over I_j.

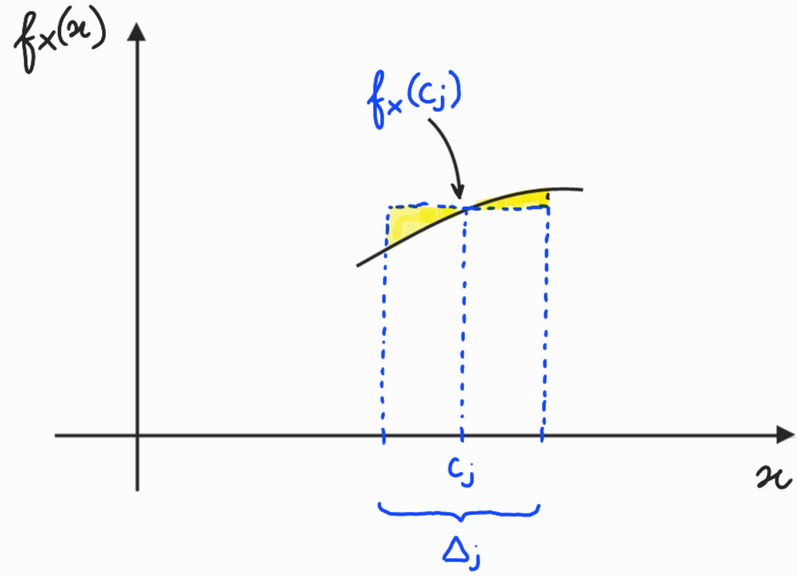
On the other hand,

$$P(X \in I_j) = \int_{I_j} f_X(x)\,dx$$

where $f_X(x)$ is the pdf of X. Let Δ_j and c_j be the width and the center of I_j, respectively.



If Δ_j is chosen sufficiently small, so that $f_X(x)$ can be considered almost constant over I_j,



then we can approximate the integral $\int_{I_j} f_X(x)dx$ (area below the black curve) with the area of the rectangle with basis Δ_j and height $f_X(c_j)$:

$$\int_{I_j} f_X(x)dx \simeq f_X(c_j)\Delta_j$$

Hence:

$$\frac{N_{\{X\in I_j\}}}{N} = \widehat{P(X\in I_j)} \simeq P(X\in I_j) = \int_{I_j} f_X(x)dx \simeq f_X(c_j)\Delta_j$$

$$\Rightarrow \quad f_X(c_j)\Delta_j \simeq \frac{N_{\{X\in I_j\}}}{N}$$

$$\Rightarrow \quad \boxed{f_X(c_j) \simeq \frac{N_{\{X\in I_j\}}}{N\Delta_j}}$$

Estimate of the pdf of X at the point c_j

The estimated values of $f_X(c_1), \dots, f_X(c_J)$ can be finally interpolated to generate the empirical pdf of X.

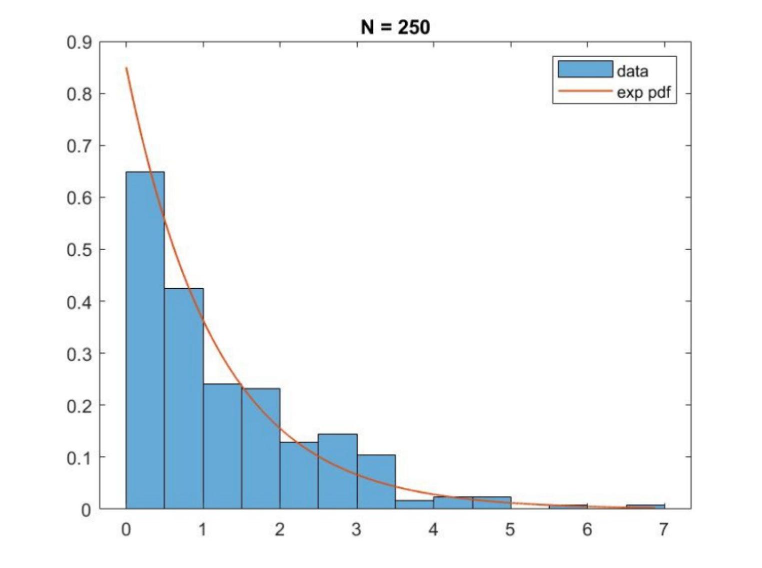
EXAMPLE

We consider data sets of $N = 250$, $N = 5000$ and $N = 100000$ independent observations of the random variable

$$X \sim \mathrm{Exp}\left(\frac{1}{\lambda}\right), \quad \text{where } \lambda = 0.85.$$

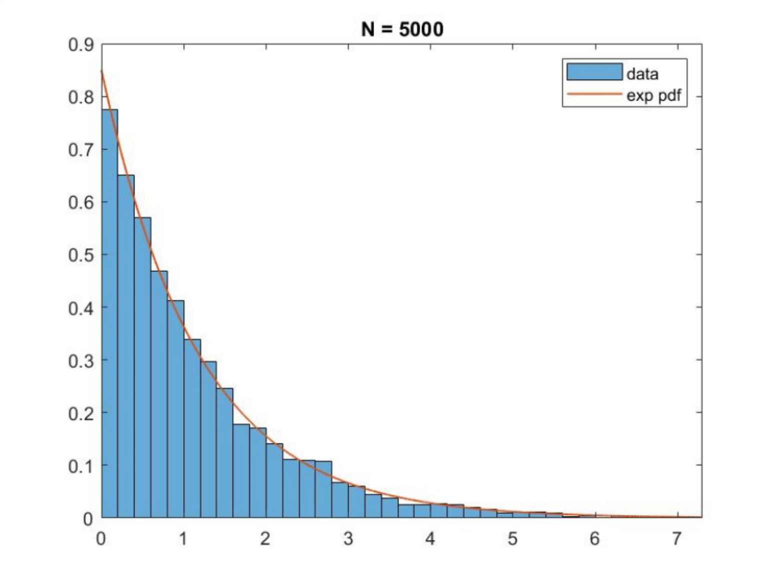
For each data set, we represent the empirical pdf with a bar plot, where the bar height is the estimated value of the pdf of X at the center of the corresponding interval.

We also represent the true pdf (red curve).



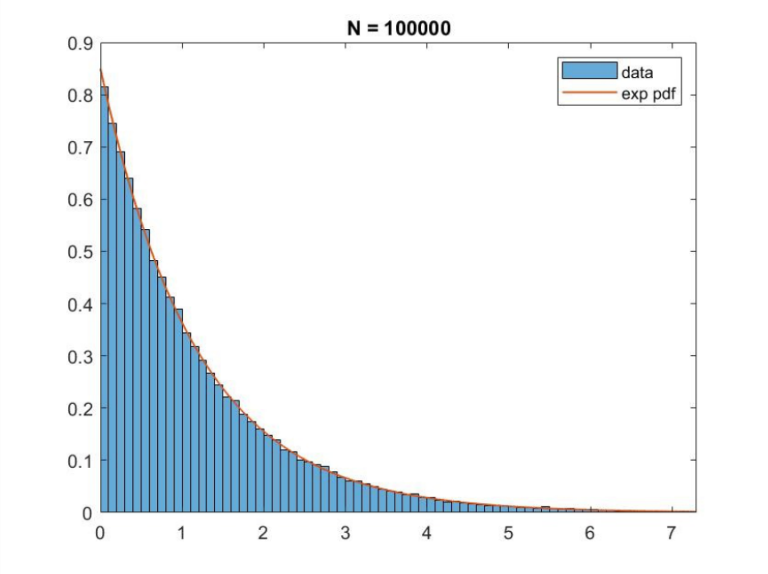
<u>$N = 250$</u>

The number of samples is small. The width Δ_j cannot be made too small, because we need a sufficient number of samples in each interval. The approximation is rough.



<u>$N = 5000$</u>

More samples make it possible to decrease Δ_j. Better approximation.

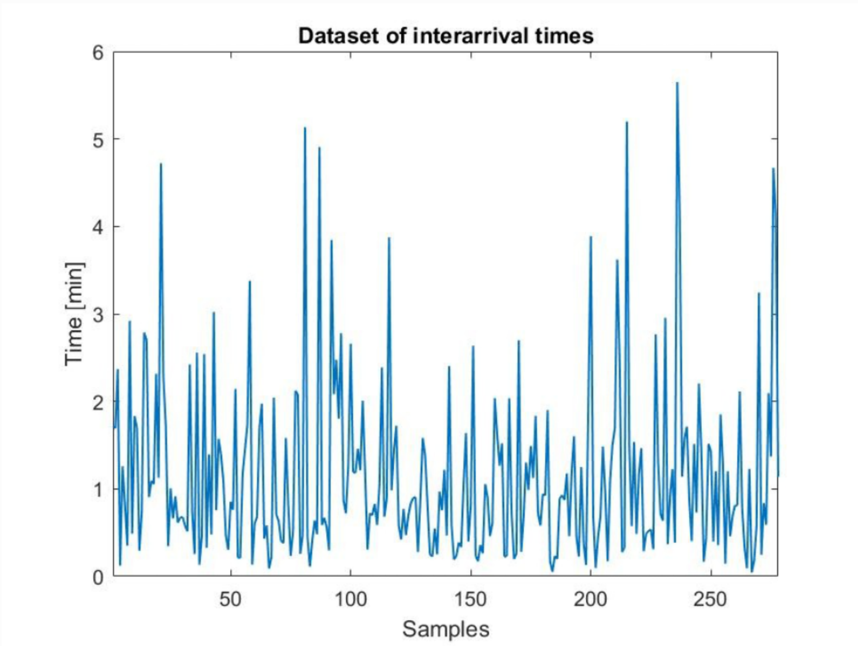


$N = 100000$

Very good approximation with small Δ_j and large number of samples.

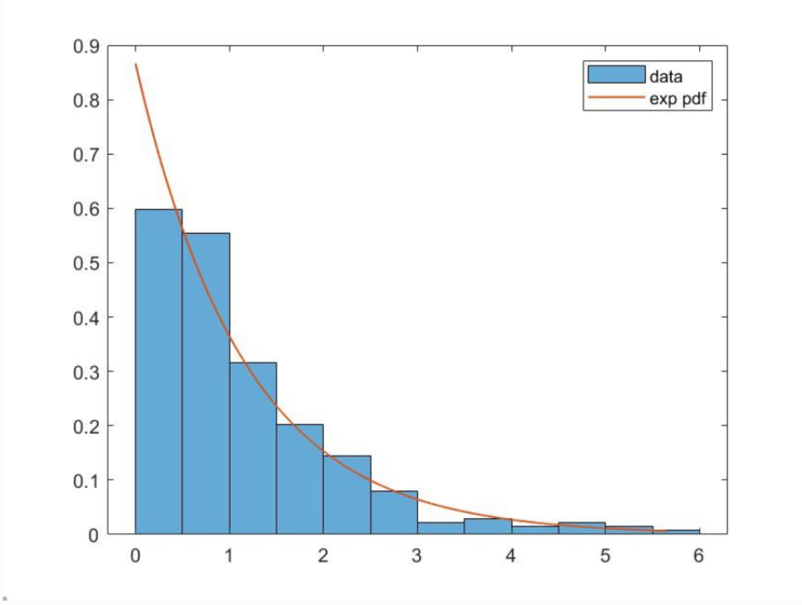
EXAMPLE

Data set of $N = 278$ interarrival times of customers at the Chicago branch of the Foster Bank.



Random variable X = interarrival time

As before, we represent the empirical pdf of X with a bar plot.



In this case, we don't have the true pdf.

For a comparison, we plot the pdf of a random variable

$$Z \sim Exp\left(\frac{1}{\lambda}\right)$$

where λ is the inverse of the mean value of the data set:

$$\frac{1}{\lambda} \simeq 1.1541 \text{ min} \Rightarrow \lambda \simeq 0.8665 \text{ min}^{-1}$$

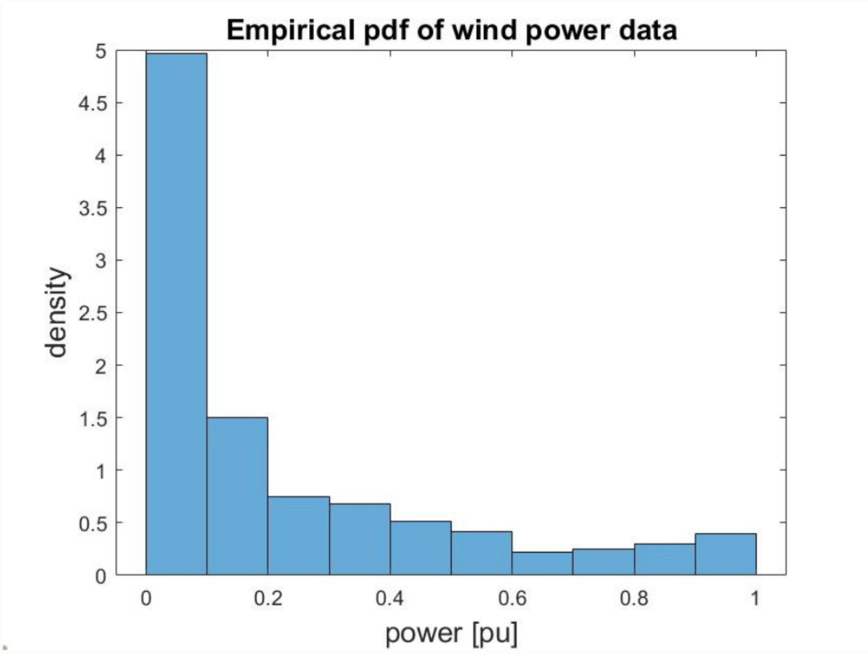
From this visual comparison, it cannot be excluded that the data set was drawn from an exponential distribution.

REMARK

In many real applications, the arrival process is well approximated by a Poisson process.

EXAMPLE

The figure shows the empirical pdf for the wind power data set.



It is apparent in this case that the data were not generated by an exponential distribution (see the tail on the right).

=> Histogram-like representations of the empirical pdf can be used for a quick, preliminary analysis of the data distribution.