Data Science II

- Introduction to Data Visualization -

Visualizing Cumulative Distribution Functions and Q-Q Plots

Prof. Dr. Eduard Kromer
Summer Semester 2024
University of Applied Sciences Landshut

Visualizing Cumulative Distribution Functions and Q-Q Plots

Cumulative Distribution Function (CDF)

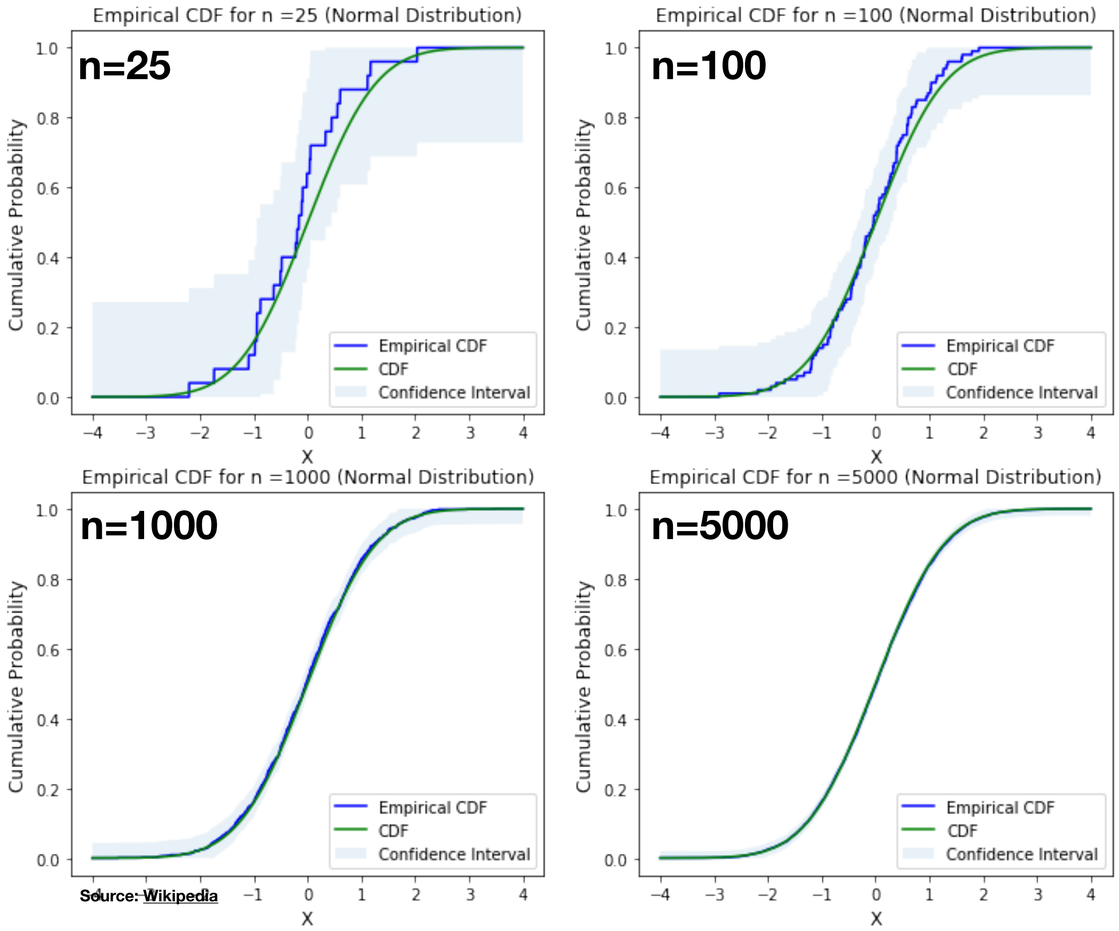
- the CDF of a real-valued random variable X, evaluated at x, is the probability of the event $\{X \leq x\}$; it is the function given by

$$F_X(x) = P(X \leq x)$$

- therefore we have

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

Empirical CDF



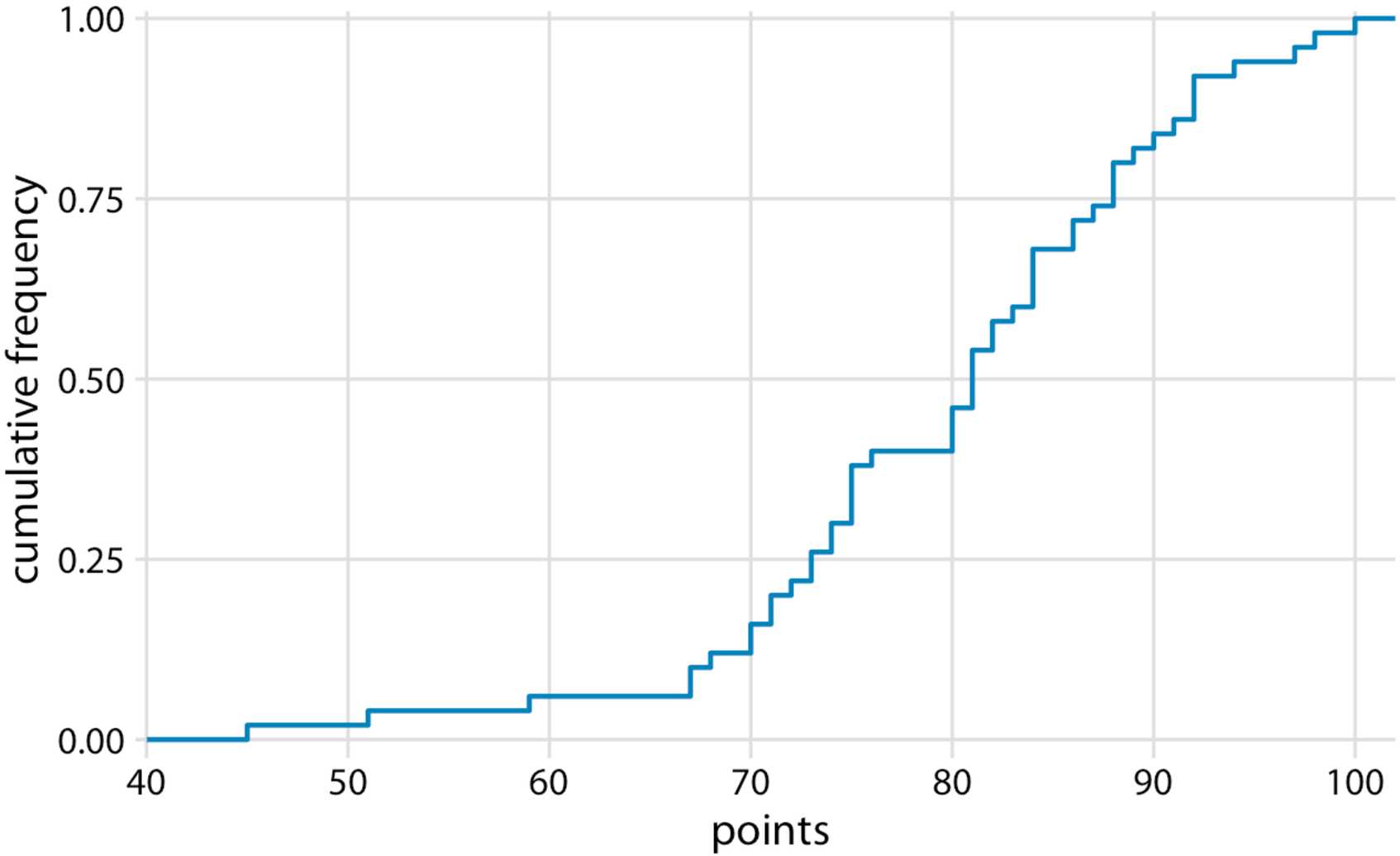
Source: Wikipedia

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}}$$ for observed values $x_1, \ldots, x_n$ in the sample

Visualizing CDFs

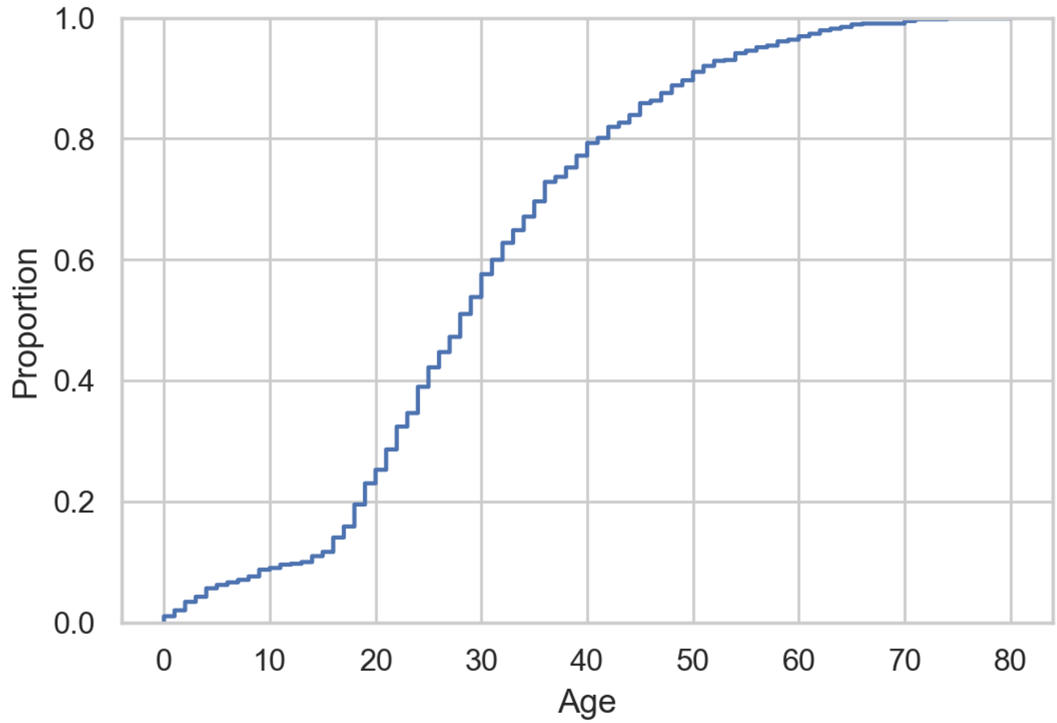
- histograms and density plots share the limitation that the resulting figure depends to a substantial degree on parameters the user has to choose (e.g. bin width, bandwidth)
- there are visualizations that require no arbitrary parameter choices:
  - empirical cumulative distribution functions
  - quantile-quantile plots (Q-Q Plots)
- they are less intuitive, but show all of the data at once

Visualizing CDFs



ecdfplot

- read the documentation of the `seaborn.ecdfplot` function and use it to generate a visualization of the empirical CDF of the ‘Age’ column in the Titanic data set



p-Quantile

- for a real-valued random variable X the real number x_p is a p-Quantile of X if:

$$P(X \leq x_p) \geq p \quad \text{and} \quad P(X \geq x_p) \geq 1 - p$$

- the p-quantiles of X are the p-quantiles of its distribution

Empirical p-Quantiles

For p with $0 < p < 1$

$$x_p = \begin{cases} x_{\lfloor np+1 \rfloor}, & \text{if } np \notin \mathbb{N} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}), & \text{if } np \in \mathbb{N} \end{cases}$$

is the empirical p-quantile of the observed values $x_1, \ldots, x_n$

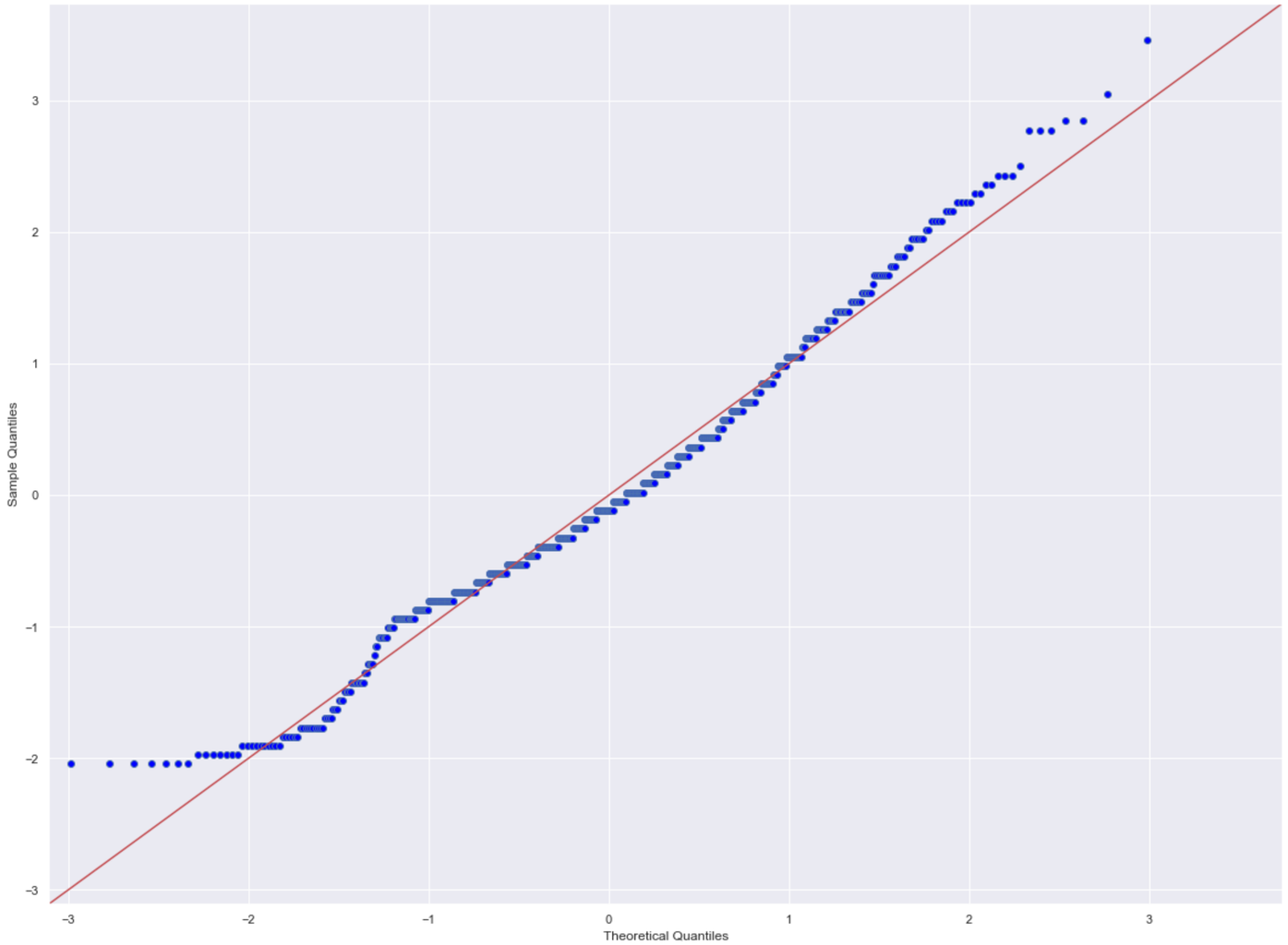
- computing quantiles with numpy:

```
np.quantile(a=data, q=0.5)
```

Q-Q Plots

- q-q plots are a useful visualization when we want to determine to what extent the observed data points follow a given distribution
- they are based on ranking the data and visualizing the relationship between ranks and actual values
  - ranks are not visualized directly
  - ranks are used to predict where a given data point should fall if the data were distributed according to a specified reference distribution (most commonly, a normal distribution is used)

Visualizing Q-Q Plots

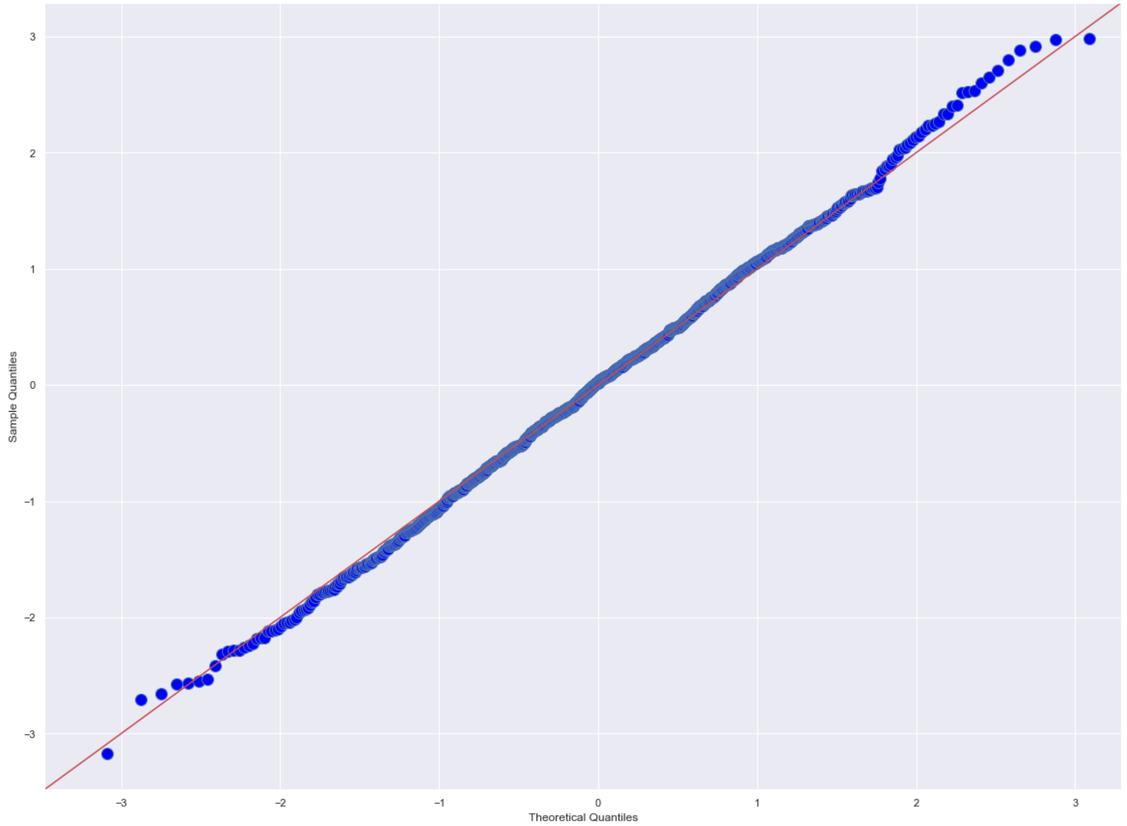


Visualizing Q-Q Plots with statsmodels

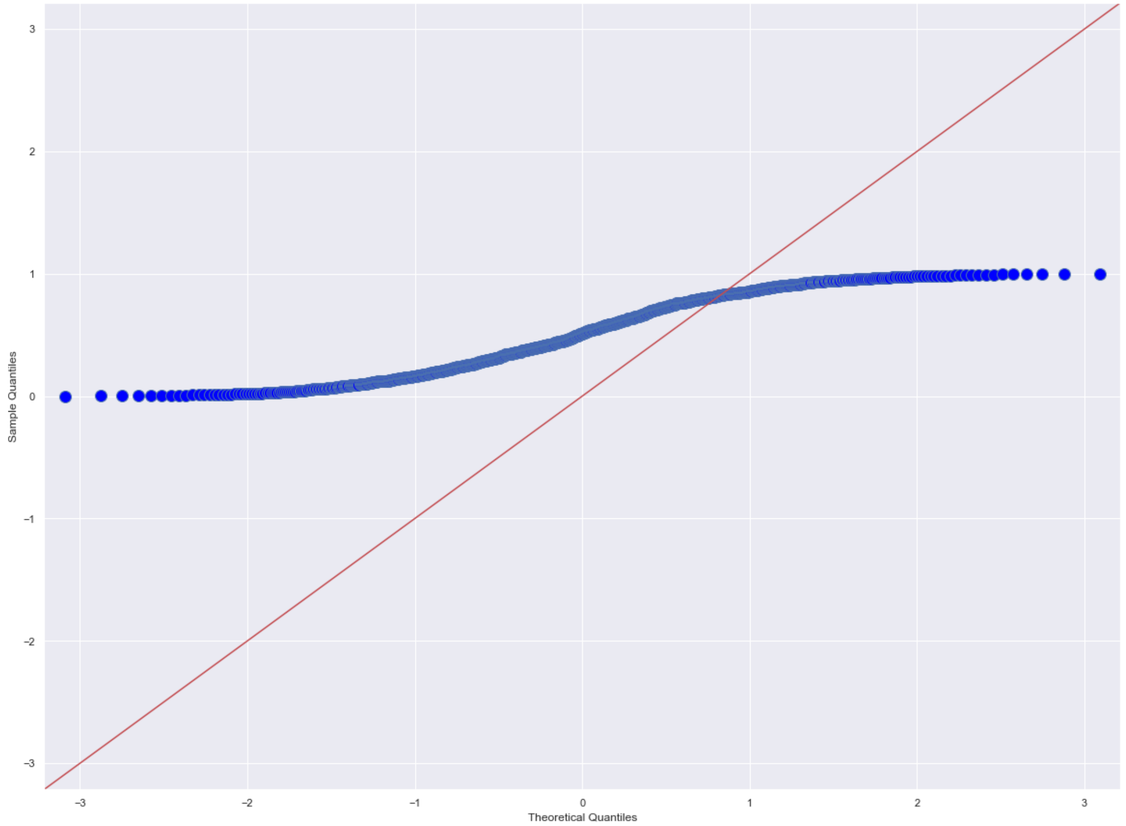
- read the documentation of `statsmodels.graphics.gofplots.qqplot`
- generate a q-q plot for the ‘Age’ column of the Titanic dataset using the normal distribution as reference distribution
- sample 1000 data points from the standard normal distribution and generate the corresponding q-q plot
- sample 1000 data points form the uniform distribution (range $[0,1]$) and generate the correspoding q-q plot using the normal distribution as reference distribution
- compare the 3 Q-Q plots — what do you notice?

Visualizing Q-Q Plots with statsmodels

Standard Normal Distribution



Uniform Distribution



HOCHSCHULE LANDSHUT