Vorlesung Advanced Data Mining

Quantiles

Prof. Dr.-Ing. Steffen Schober

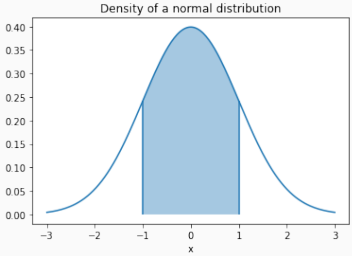
WS 23/24

Hochschule Esslingen

Density Functions

- Below we see a probability density function (PDF) $f(x)$ of a normal distribution.
- The probability to observe a value between -1 and +1 corresponds to the shaded area:

$$P(\text{value between -1 and +1}) = \int_{-1}^{+1} f(x)dx$$

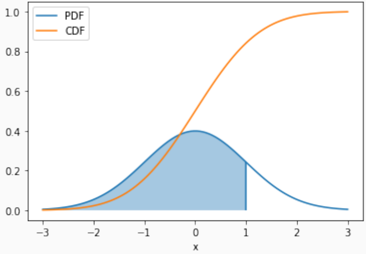


Cummulative Density Function

$$P(\text{value} \leq 1) = \int_{\infty}^{+1} f(x)dx \equiv F(1)$$

where $F(x)$ is the cummulative density function (CDF), defined as

$$F(x) := \int_{\infty}^{x} f(x)dx$$



HOCHSCHULE ESSLINGEN

Definition

The q-quantile $x_q, 0 \leq q \leq 1$ is defined by

$$F(x_q) = \int_{-\infty}^{x_q} f(x)dx = q.$$

or

$$x_q = F^{-1}(q).$$

Empirical Quantiles: Example

Given n ordered data points $x_0, x_1, ..., x_{n-1}$

- the median divides the data set in (roughly) equal parts
- for example for

x_0	x_1	x_2
1	2	3

  the median = 0.5-quantile, dentoed with $x_{0.5}$ ist given by $x_1 = 2$. The index of the median lies on *half-way*, i.e. $i = (n-1) \cdot q$
- this is similar for other quantiles ($q \neq 0.5$)
- things start to get complicated when the quantile-index is not an integer

Empirical Quantiles

Quantiles (NumPy Default)

Given n data points $x_0, x_1, \ldots, x_{n-1}$ and let $0 \leq q \leq 1$. Define the index i' as

$$i' = q \cdot (n-1)$$

Let $[\cdot]$ denote the fractional part of a number, and $\lfloor \cdot \rfloor$ denotes the floor function. Let

$$i = \lfloor i' \rfloor \quad \text{and} \quad g = [i']$$

Then the q-th quantile is given by

$$x_q = x_i + (x_{i+1} - x_i) \cdot g.$$

Implementation

HOCHSCHULE ESSLINGEN

```
def quantile_linear(x, q):
    if len(x)==1:
            return x[0]

    x.sort()
    i_ = q*(len(x)-1)
    i = int(np.floor((i_)))
    g = i_ - i
    return x[i]+ (x[i+1] - x[i])*g
```

Example

x_0	x_1	x_2	x_3	x_4	x_5
1	2	4	5	6	8

(1)

```python
x = data[0] # first row of table
print("Ours = ", quantile_linear(x, 0.25))
print("NumPy =", np.quantile(x, 0.25))
```

```
Ours =  2.5
NumPy = 2.5
```