Chapter 6
Statistical Data Analysis

Vi Bảo Ngọc
0983408885 - ngocvb@lqdtu.edu.vn
Computer Science Department, Le Quy Don Technical University

Outline

1. **Distribution fitting**
2. **Kernel Density Estimation**
3. Determining confidence intervals for mean, variance, and standard deviation
4. Exploring extreme values
5. Correlating variables with correlation
6. Evaluating relationships between variables with ANOVA

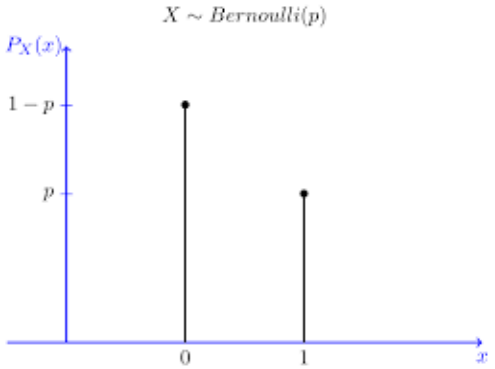
Distribution fitting

- **Distribution fitting** is the procedure of selecting a statistical distribution that best fits to a data set generated by some random process.
- In other words, if you have some random data available, and would like to know what particular distribution can be used to describe your data, then distribution fitting is what you are looking for.

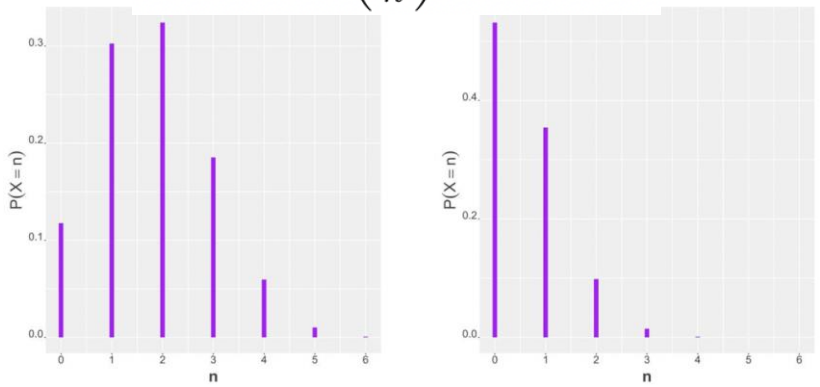
Distribution fitting

- **Some popular distributions:**

$$P(x) = p^x(1-p)^{1-x}.$$

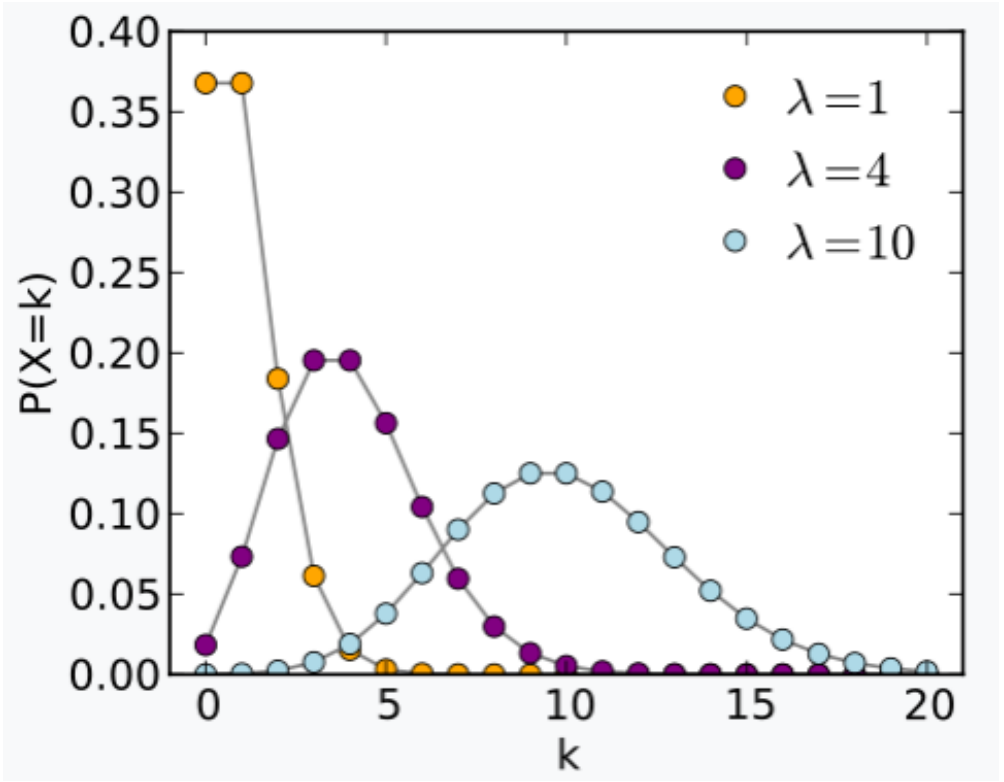


$$P(X = n) = \binom{N}{n} p^n (1-p)^{N-n}.$$



N = 6, p = 0.3 N = 6, p = 0.1

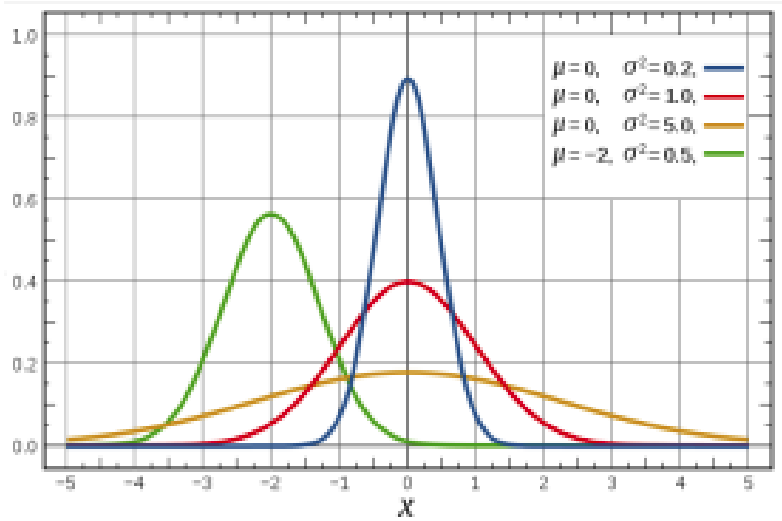
$$P(X = n) = \frac{\lambda^n \exp(-\lambda)}{n!}.$$



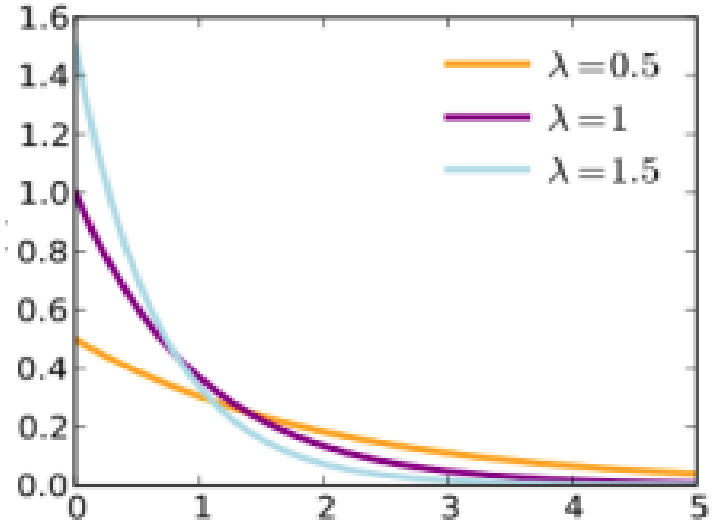
Distribution fitting

- **Some popular distributions:**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$



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

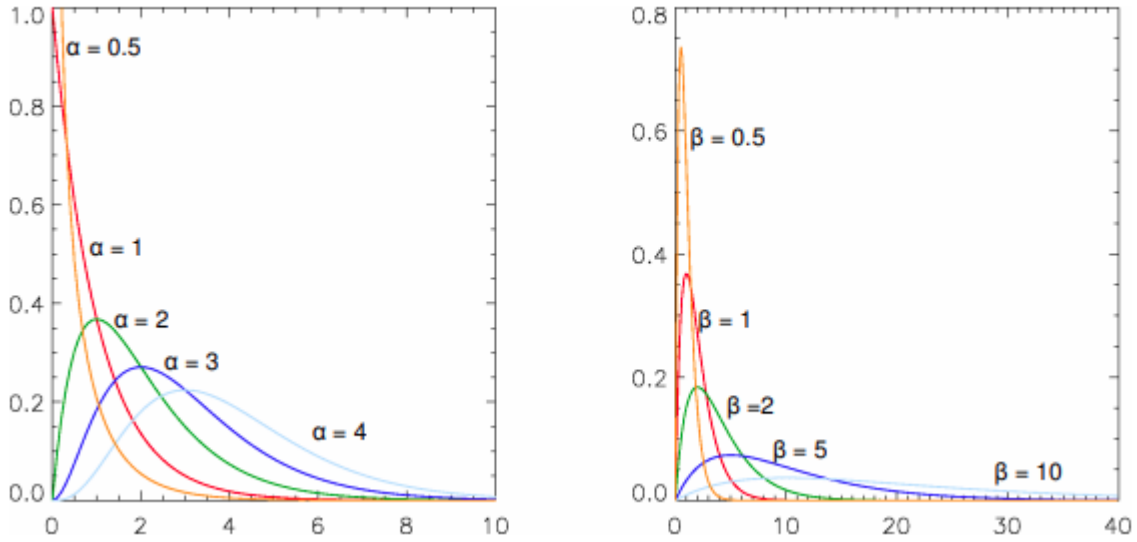


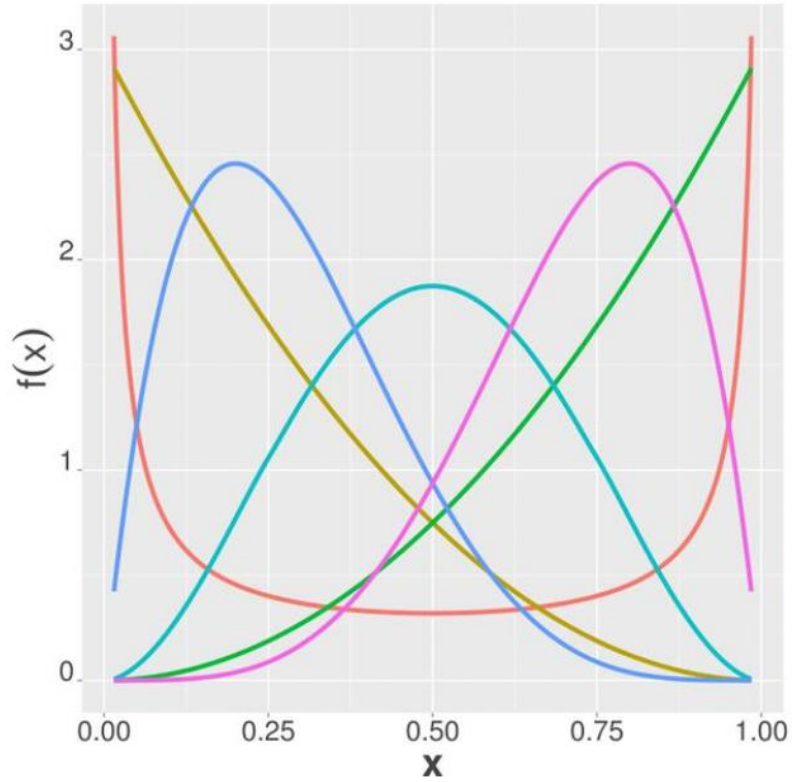
Distribution fitting

- **Some popular distributions:**

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{-\alpha-1} \exp\left(-x/\beta\right) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in]0,1[.$$





Distribution fitting

- **Which Distribution Should I Choose?**
  - You cannot "just guess" and use any other particular distribution without testing several alternative models as this can result in analysis errors.
  - In most cases, you need to fit two or more distributions, compare the results, and select the most valid model.
  - The "candidate“ distributions you fit should be chosen depending on the nature of your probability data.
    - For example, if you need to analyze the time between failures of technical devices, you should fit non-negative distributions such as Exponential or Weibull, since the failure time cannot be negative.
  - You can also apply some other identification methods based on properties of your data.
    - For example, you can build a histogram and determine whether the data are symmetric, left-skewed, or right-skewed, and use the distributions which have the same shape.

Distribution fitting

- **Which Distribution Should I Choose?**
  - To actually fit the "candidate" distributions you selected, you need to employ statistical methods allowing to estimate distribution parameters based on your sample data.
  - After the distributions are fitted, it is necessary to determine how well the distributions you selected fit to your data. This can be done using the specific goodness of fit tests or visually by comparing the empirical (based on sample data) and theoretical (fitted) distribution graphs. As a result, you will select the most valid model describing your data.

Distribution fitting

- **Goodness-of-Fit Tests:**
  - The chi-square test is used to test if a sample of data came from a population with a specific distribution.
  - Another way of looking at that is to ask if the frequency distribution fits a specific pattern.
  - Two values are involved, an observed value, which is the frequency of a category from a sample, and the expected frequency, which is calculated based upon the claimed distribution.
  - The idea is that if the observed frequency is really close to the claimed (expected) frequency, then the square of the deviations will be small. The square of the deviation is divided by the expected frequency to weight frequencies. A difference of 10 may be very significant if 12 was the expected frequency, but a difference of 10 isn't very significant at all if the expected frequency was 1200.

Distribution fitting

- **Techniques of fitting:**
  - Parametric methods, by which the parameters of the distribution are calculated from the data series. The parametric methods are – method of moments, method of L-moments and Maximum likelihood method
  - Regression method, using a transformation of the cumulative distribution function so that a linear relation is found between the cumulative probability and the values of the data, which may also need to be transformed, depending on the selected probability distribution.

Distribution fitting

- Method of Moments
  - The method of moments involves equating sample moments with theoretical moments

1. $E(X^k)$ is the k^{th} (theoretical) moment of the distribution (**about the origin**), for $k = 1, 2, \ldots$
2. $E\left[(X-\mu)^k\right]$ is the k^{th} (theoretical) moment of the distribution (**about the mean**), for $k = 1, 2, \ldots$
3. $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ is the k^{th} sample moment, for $k = 1, 2, \ldots$
4. $M_k^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$ is the k^{th} sample moment about the mean, for $k = 1, 2, \ldots$

  - Basic idea:

1. Equate the first sample moment about the origin $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ to the first theoretical moment $E(X)$.
2. Equate the second sample moment about the origin $M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ to the second theoretical moment $E(X^2)$.
3. Continue equating sample moments about the origin, M_k, with the corresponding theoretical moments $E(X^k)$, $k = 3, 4, \ldots$ until you have as many equations as you have parameters.
4. Solve for the parameters.

Distribution fitting

- Method of Moments:
  - Example: Let $X_1, X_2, \ldots, X_n$ be Bernoulli random variables with parameter . What is the method of moments estimator of p?
  - Solution:
    - The first theoretical moment about the origin is:

$$E(X_i) = p$$

    - Equating the first theoretical moment about the origin with the corresponding sample moment, we get:

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Distribution fitting

- Maximum Likelihood Estimation:
  - Suppose we have a random sample $X_1, X_2, \ldots, X_n$ for which the probability density (or mass) function of each is $f(x_i;\theta)$
  - Then, the joint probability mass (or density) function is:

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

  - the "**likelihood function**" L(θ) as a function of θ, and find the value of θ that maximizes it.

Distribution fitting

- Maximum Likelihood Estimation:
  - Example: Example: Let $X_1, X_2, \ldots, X_n$ be Bernoulli random variables with parameter . What is the method of moments estimator of p?
  - Solution:
    - The probability mass function of each X_i.

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

    - The likelihood function

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

Distribution fitting

- Maximum Likelihood Estimation:
  - Solution: (cont)
    - The likelihood function $L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$
    - The natural logarithm of the likelihood function is:

$$\log L(p) = (\sum x_i)\log(p) + (n - \sum x_i)\log(1-p)$$

    - The derivative of log-likelihood and setting it to 0 then find the p to maximize L

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} \overset{SET}{\equiv} 0$$

$$\rightarrow \quad (\sum x_i)(1-p) - (n - \sum x_i)p = 0 \quad \rightarrow \quad \hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Distribution fitting

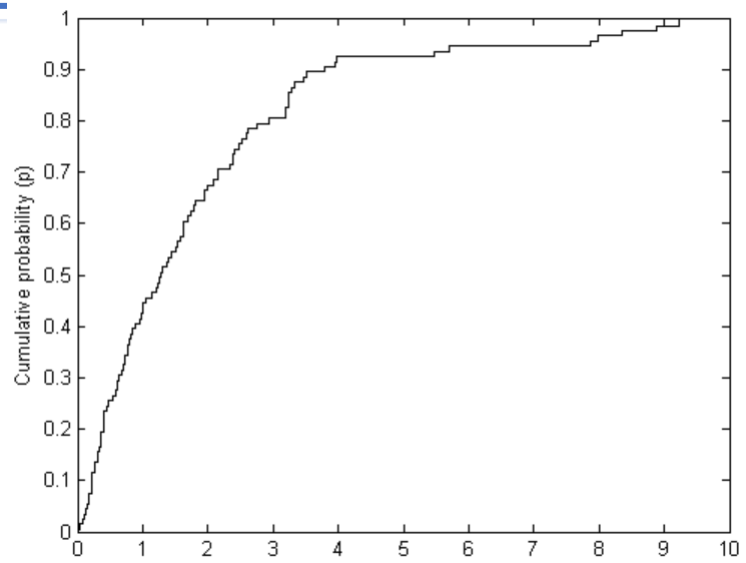
- Regression method:
  - Transformation of the cumulative distribution function so that a linear relation is found between the cumulative probability and the values of the data.
  - Example: Exponential distribution

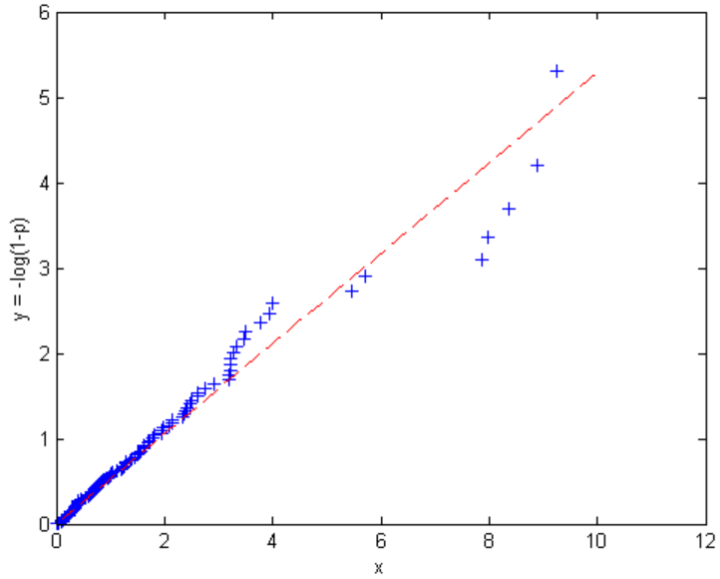
$$F(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-\lambda x} & x > 0. \end{cases}$$

$$x = -\frac{1}{\lambda}\log\left(1 - F(x)\right) = -\mu\log\left(1 - p\right)$$

x axis: data

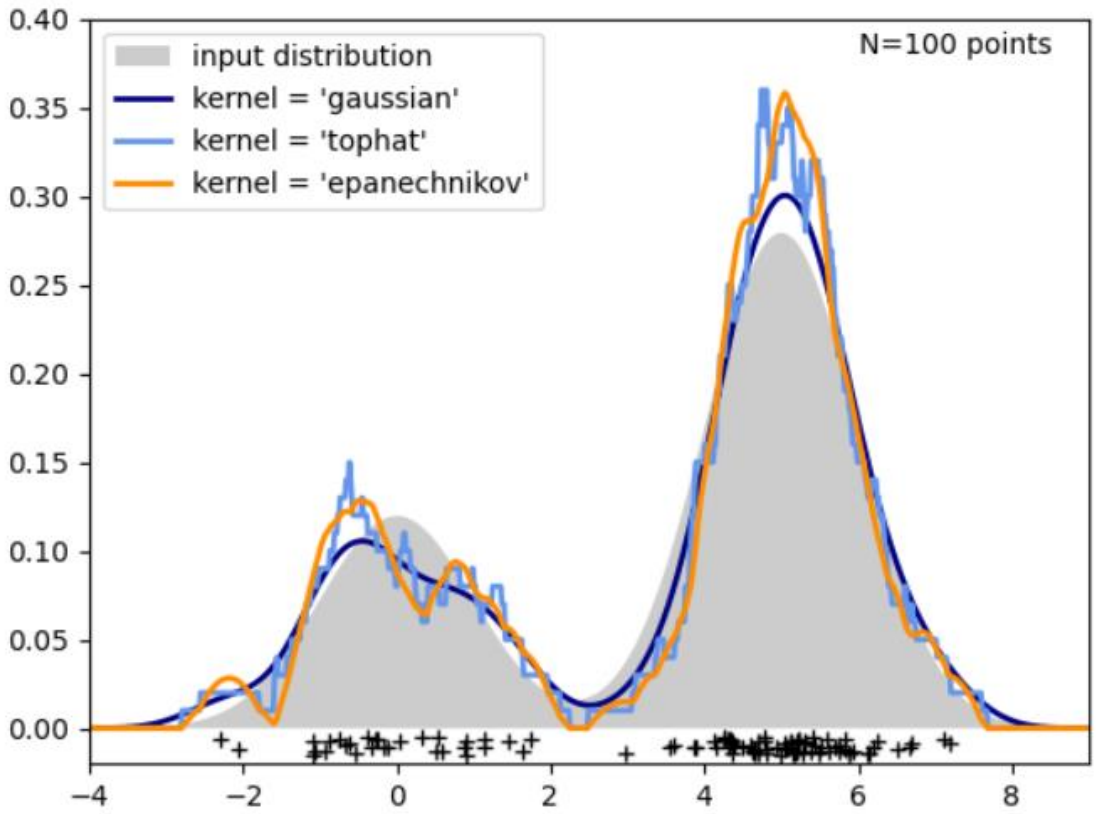
y axis: -log(1-p), p is computed by empirical cumulative distribution function (ECDF) of the data





Kernel Density Estimation

- In statistics, ***Kernel Density Estimation*** is a mathematic process of finding an estimate probability density function of a random variable



Kernel Density Estimation

- The pdf of distribution is estimated as:

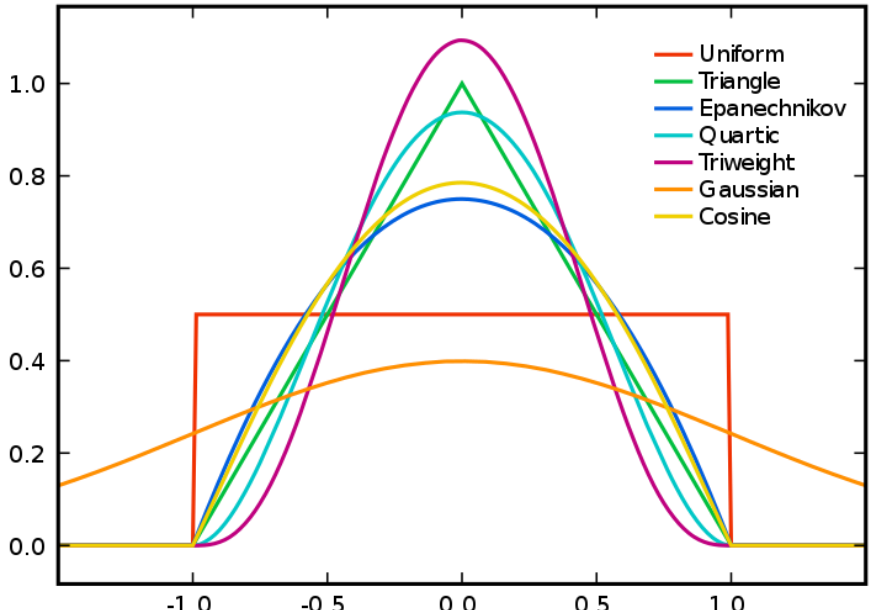
$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where:

- K is the kernel function, a function with properties similar to a PDF
- h is the bandwidth, which controls the smoothing

Kernel Density Estimation

- The kernel function typically exhibits the following properties:
  1. Symmetry such that $K(u) = K(-u)$.
  2. Normalization such that $\int_{-\infty}^{\infty} K(u)\, du = 1$.
  3. Monotonically decreasing such that $K'(u) < 0$ when $u > 0$.
  4. Expected value equal to zero such that $\mathrm{E}[K] = 0$.



Kernel Density Estimation

- KDE and generative classification

1. Split the training data by label.

2. For each set, fit a KDE to obtain a generative model of the data. This allows you for any observation x and label y to compute a likelihood $P(x \mid y)$.

3. From the number of examples of each class in the training set, compute the *class prior*, $P(y)$.

4. For an unknown point x, the posterior probability for each class is $P(y \mid x) \propto P(x \mid y)P(y)$. The class which maximizes this posterior is the label assigned to the point.

Reference

- https://online.stat.psu.edu/stat415/
- Idris, Ivan. *Python data analysis cookbook*. Packt Publishing Ltd, 2016., Chapter 3.