Team Name – Machine Learners

Project Title: Statistical Change Detection for Multi-Dimensional Data

Course Instructor: P Balamurugan
Course Code: IE-506
Machine Learning Principles and Techniques

Name	Roll Number
Mayur Dhanawade	23M1512
Shivam Negi	23M1508

Contents

Hypothesis Testing

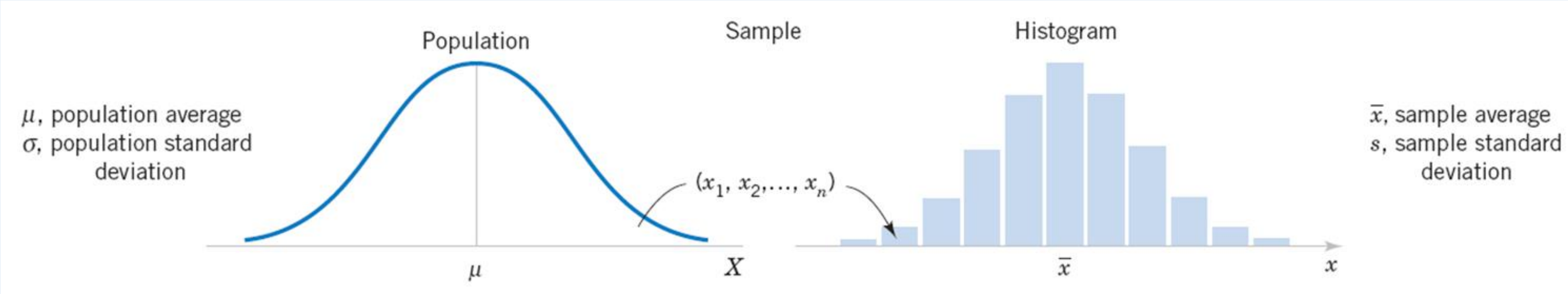
- A Hypothesis is a statement about a population parameter.
- Two Complementary Hypothesis in Hypothesis Testing:
  1. Null Hypothesis (H_0): No significant difference observed between two samples.
  2. Alternate Hypothesis (H_1 or H_a): There is significant difference.

Making Decision In Hypothesis Testing:

	Accept H_0	**Reject H_0**
H_0 is True	Correct Decision	Type I error
H_0 is False	Type II error	Correct Decision

Test Statistic

- A test statistic is a number that summarizes a data set and is used in statistical hypothesis testing. It describes how far observed data is from the null hypothesis. The null hypothesis is that there is no relationship between variables or no difference among sample groups.



Relationship between a population and a sample.

Source: www.njit.edu

Likelihood & Log Likelihood function.

Given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function $\boldsymbol{\theta}$ defined by

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \begin{cases} f(\boldsymbol{x} \mid \boldsymbol{\theta}) & \text{if } \boldsymbol{X} \text{ is continuous} \\ P_\theta(\boldsymbol{X} = \boldsymbol{x}) & \text{if } \boldsymbol{X} \text{ is discrete} \end{cases}$$

When a sample x is observed. For a given parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ if $P_{\boldsymbol{\theta}_1}(\boldsymbol{X} = \boldsymbol{x}) = L(\boldsymbol{\theta}_1 \mid \boldsymbol{x}) > L(\boldsymbol{\theta}_2 \mid \boldsymbol{x}) = P_\theta(\boldsymbol{X} = \boldsymbol{x})$ Sample would have come more likely from parameter $\boldsymbol{\theta}_1$ than $\boldsymbol{\theta}_2$. $\boldsymbol{\theta}_1$ is more plausible value of true parameter $\boldsymbol{\theta}$ than $\boldsymbol{\theta}_2$

Example: Likelihood function of Normal distribution.

Let$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be i.i.d. $\sim \mathcal{N}(\theta, 1)$.

Likelihood function for Normal distribution is $L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{(x_i-\theta)^2}{2}\right\}$

Log Likelihood function:

$\log L(\boldsymbol{\theta} \mid \boldsymbol{x}$ instead of $L(\boldsymbol{\theta} \mid \boldsymbol{x})$, called log likelihood function

Normal Log Likelihood: $\log L(\theta \mid \boldsymbol{x}) = n \log \dfrac{1}{\sqrt{2\pi}} - \dfrac{\sum_{i=1}^{n} (x_i - \theta)^2}{2}$

Likelihood Ratio Test

Likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid \boldsymbol{x})}{\sup_{\theta \in \{\Theta_0, \Theta_0^c\}} L(\theta \mid \boldsymbol{x})}$$

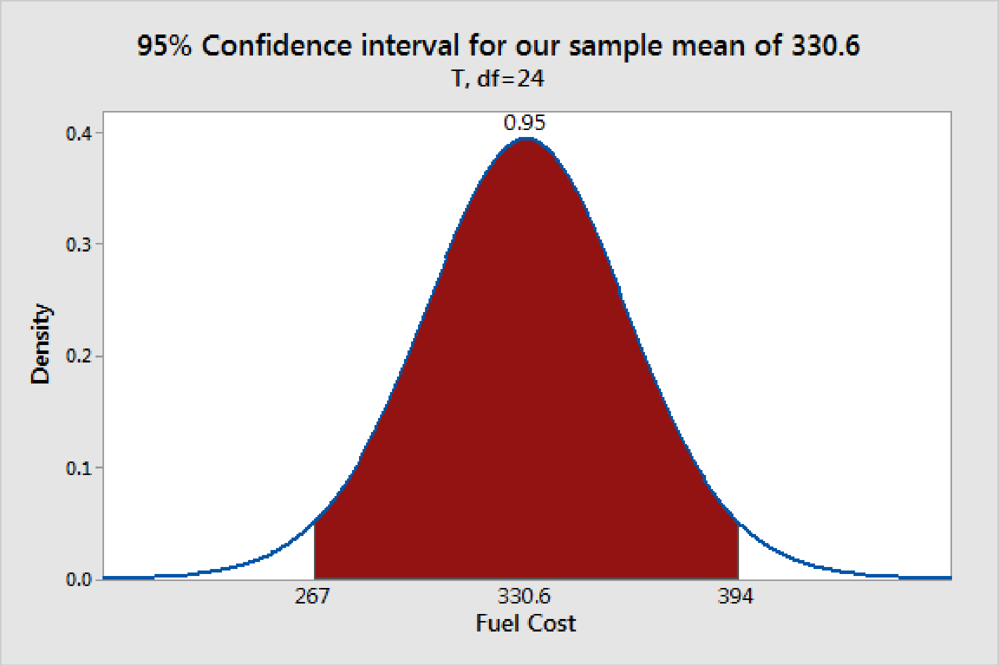
A Likelihood Ratio Test (LRT) is any any test that has the reject region of the form $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$ for some $c \in (0, 1)$.

p – value and confidence level

- **P-value**: Measures evidence against the null hypothesis. Small p-values suggest rejecting the null hypothesis.
- **Confidence Interval**: Provides an estimate of a parameter along with a range of likely values. The confidence level specifies our confidence in the interval.

- For example:

The shaded range of sample means [267, 394] covers 95% of this sampling distribution. This range is the 95% confidence interval for our sample data. We can be 95% confident that the population mean for fuel costs fall between 267 and 394.

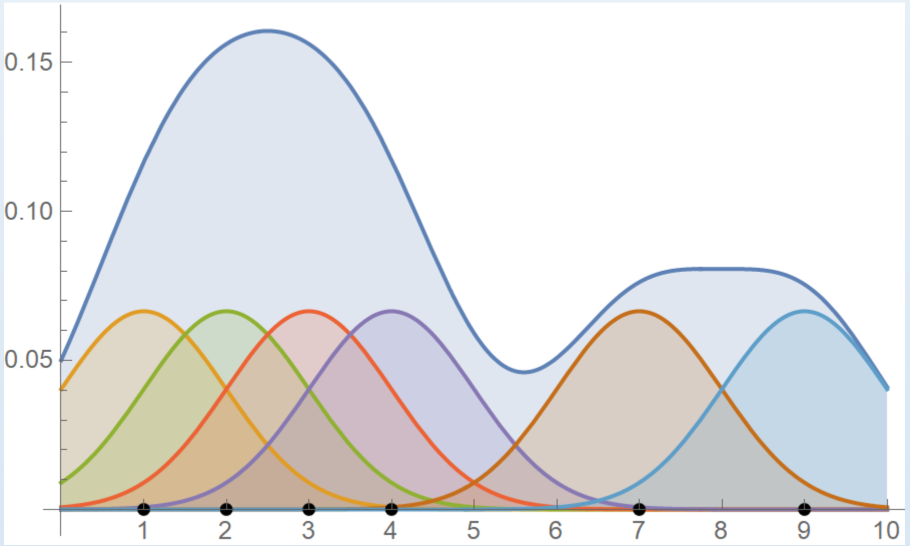


Source : statisticsbyjim.com

Kernel Density Estimator for One Variable.

- The kernel function is evaluated for each datapoint separately, and these partial results are summed to form the KDE.
- Kernel function One of the most common kernels is the Gaussian kernel i.e

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$



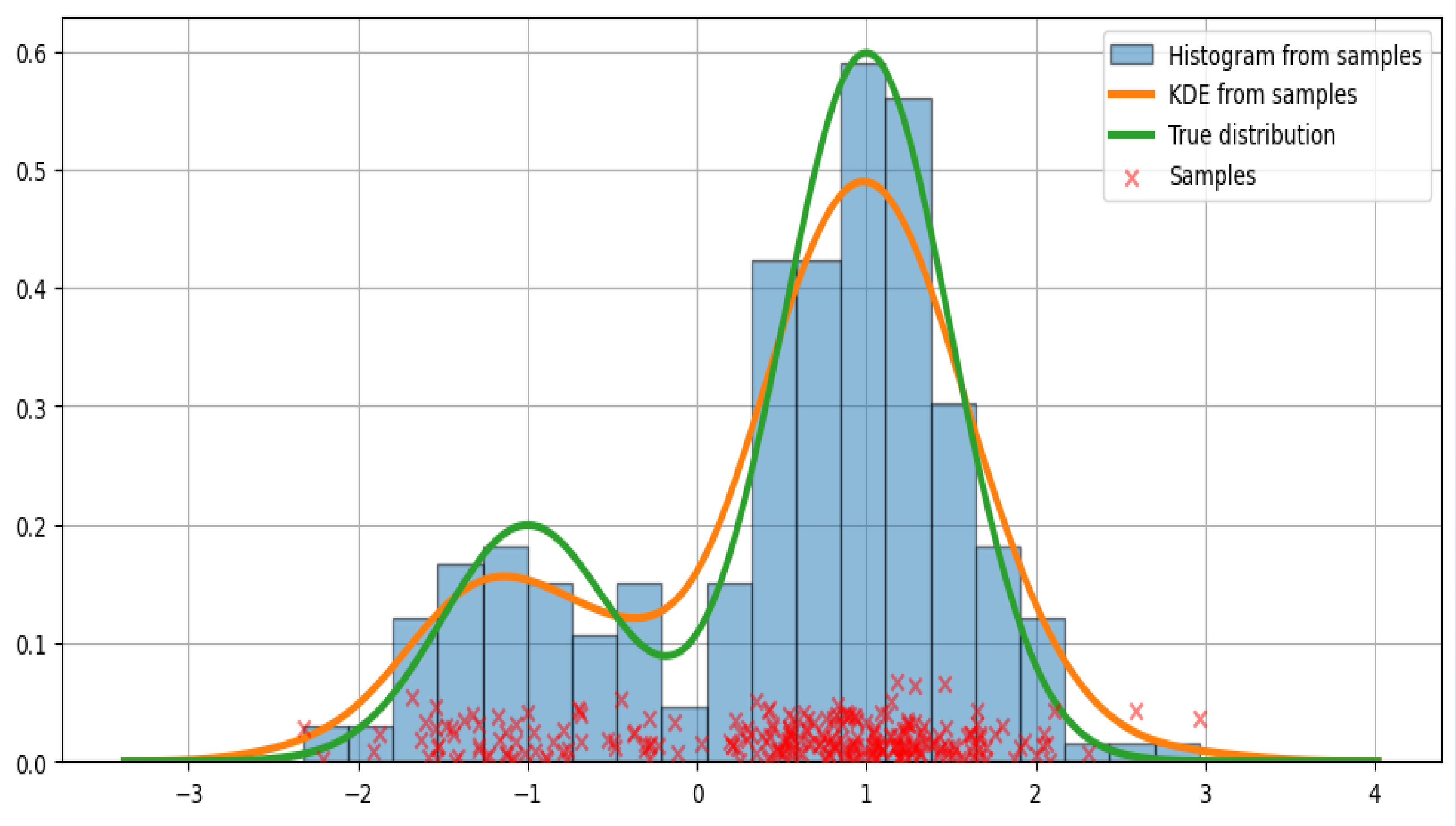
Source: ekamperi.github.io

- A kernel K typically satisfies the following conditions.

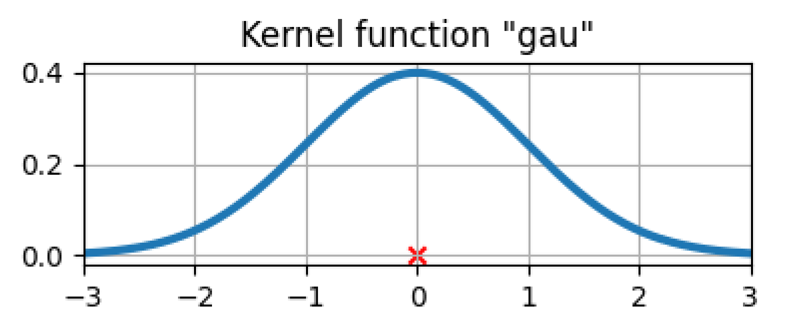
1. Symmetry such that $K(u) = K(-u)$.
2. Normalization such that $\int_{-\infty}^{\infty} K(u)du = 1$.
3. Monotonically decreasing such that $K'(u) < 0$ when $u > 0$.
4. Expected value equal to zero such that $\mathrm{E}[K] = 0$.

Kernel Functions

- Types of kernel functions:
  1. Gaussian Kernel.
  2. Rectangular Kernel.
  3. Triangular Kernel.
  4. Uniform Kernel.
  5. Bi weight Kernel.
  6. Cosine Kernel.
- But we use Gaussian Kernel function.



Source : www.statsmodels.org



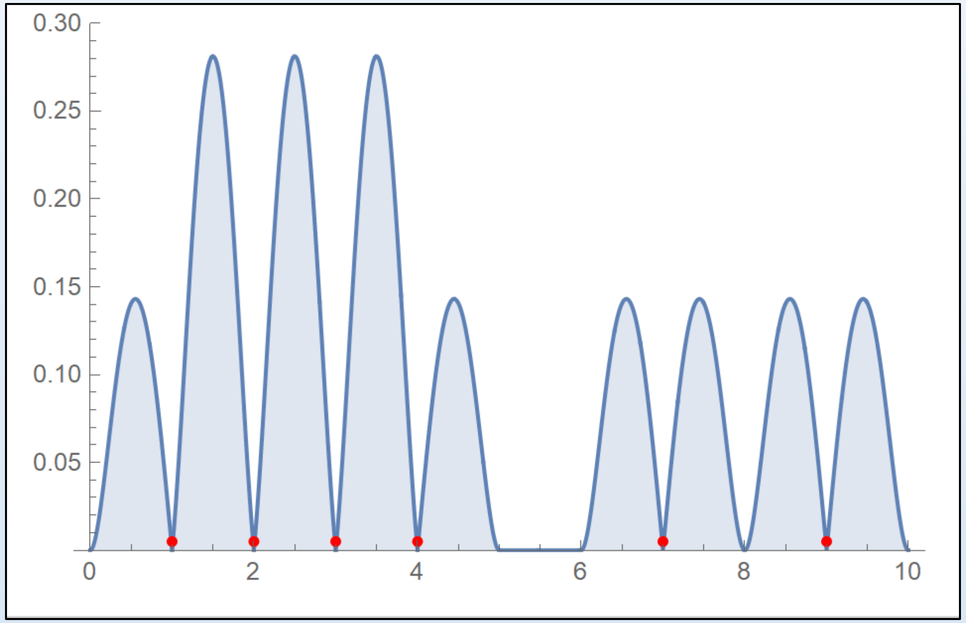
Resource : www.statsmodels.org

Animation of How Scaling factor (h) Works.

K_h is scaled version of the kernel i.e

$$K_h = \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

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



Source: ekamperi.github.io

Where:

$\hat{f}(x)$ is the estimated density function.
n is the number of data points.
K is the kernel function.
h is the bandwidth.
x_i are the data points.

Kernel Density Estimator For Multi Dimension

- This function is the prototype of a multidimensional Gaussian Kernel i.e K(x) .

$$K(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} \det(\mathbf{H})^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{x}}$$

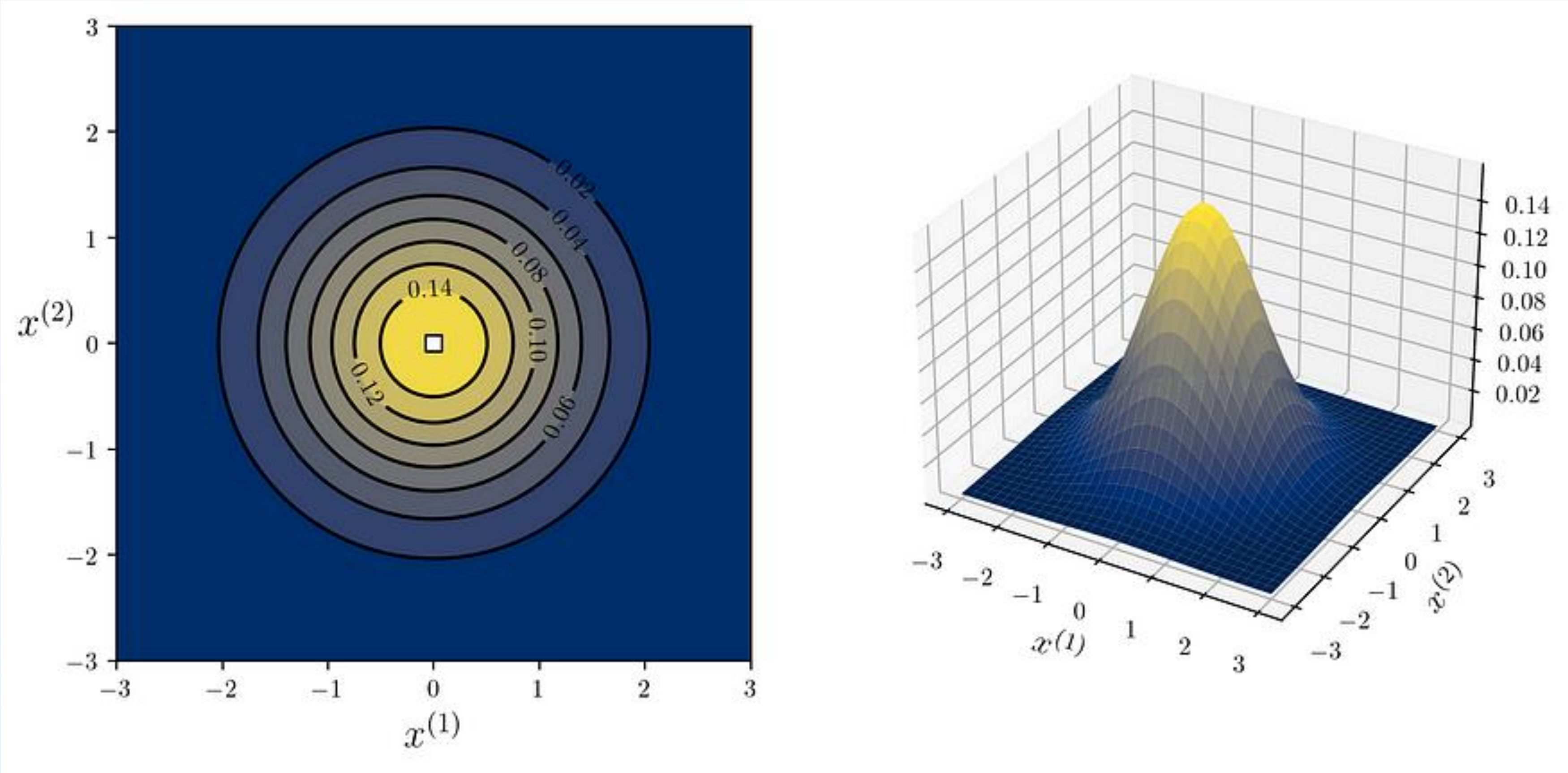
$$\hat{f}(\mathbf{x}, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)$$

Where:

$$\mathbf{x}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{x} = \begin{bmatrix} x^{(1)} & x^{(2)} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}^{-1} \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix} \quad \mathbf{x}_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(d)} \end{bmatrix}$$

$\hat{f}(x, H)$ is the estimated density function.
n is the number of data points.
K is the kernel function.
H covariance matrix of order $d \times d$.
x_i The vector x has a total of d dimensions (features) with the superscript representing the index of the features.

Kernel Density Estimator For Multi Dimension

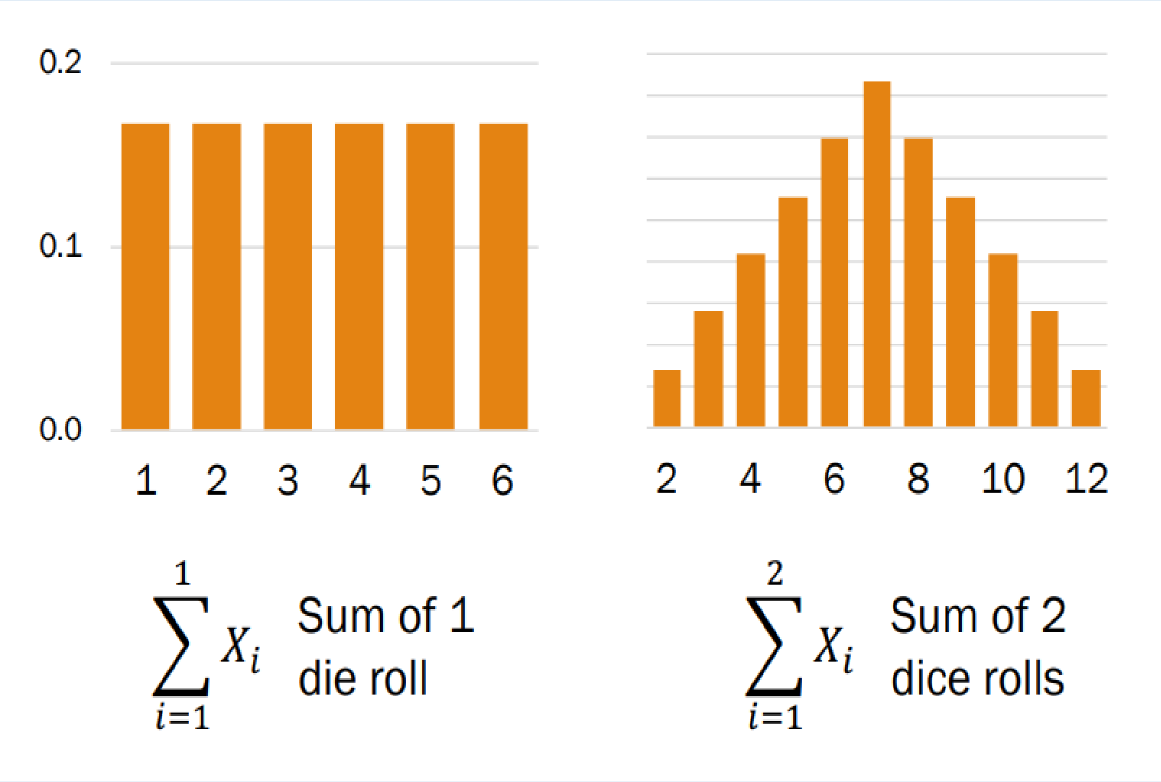


Source: towardsdatascience.com

Central Limit Theorem (CLT)

- Sum of dice rolls: Roll n independent dice. Let Xi be outcome of roll i. Xi ’s are i.i.d.
- The sum of n i.i.d. random variables is normally distributed with mean nμ and variance nσ^2.

According to CLT: $\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right)$



Source: [web.stanford.edu]

Source: [youtu.be/SCN]

About Data Set

- It contains oceanographic and surface meteorological readings taken from a series of buoys positioned throughout the equatorial Pacific.

Observation	Year	Month	Day	Date	Latitude	Longitude	Zonal Winds	Meridional Winds	Humidity	Air Temp	Sea Surface Temp
4060	93	5	9	930509	-0.02	-109.96	-2.1	2.1	81.2	26.8	27.02
4061	93	5	10	930510	-0.02	-109.96	-3.4	1.4	84.2	26.95	26.91
4062	93	5	11	930511	-0.02	-109.96	-3.8	2.2	84.9	26.98	26.78
4063	93	5	12	930512	-0.02	-109.96	-3	1.5	86.9	26.93	26.74
4064	93	5	13	930513	-0.02	-109.96	-4.5	1.9	87.6	27.01	26.82

Rows = 93,935, Columns = 12 Source: www.kaggle.com/datasets/uciml/el-nino-dataset

About Code Availability

- There are no code links available in our Research Paper, but three Algorithms are available in our Research Paper.

Algorithm 1 LearnBandwidth(S_1, $MaxIteration$, ϕ)

$t = 0$
while $t < MaxIteration$ **do**
    Compute density $p(x_j|\omega_i)$ for all i, j by Equation (5)
    Compute soft membership $P(\omega_i|x_j)$ for all i, j by Equation (4)
    Compute bandwidth Σ_{x_i} for all i by Equation (3)
    Compute the objective function L_t by Equation (2)
    if $\frac{L_t - L_{t-1}}{L_{t-1}} < \phi$ **then**
        break
    end if
    $t++$
end while

Algorithm 2 EstVar(V, S_2, $estSize$, β)

Initialize the array Est
for $t = 1$ to $estSize$ **do**
    Bootstrap resample R from S_2, where $|R| = |S_2|$
    $Est[t] \leftarrow \frac{|S_2|}{|S_2|-1} \mathbf{Var}[f(R)]$
end for
$V \leftarrow [estSize \times (1-\beta)]^{th}$ percentile of Est
$\hat{\mathbf{Var}}[\Delta] \leftarrow (|S'| + \frac{|S'|^2}{|S_2|})V$

Algorithm 3 CriticalValue(p, $stepSize$)

$M = \frac{p}{stepSize} - 1$, M is the number of (α, β) pairs
$\alpha_i = i \times stepSize, i = 1, 2, \ldots, M$
$\beta_i = p - \alpha_i, i = 1, 2, \ldots, M$
Let C be an array of critical values
for $i = 1$ to M **do**
    Run EstVar(V, S_2, $estSize$, β_i) to obtain $\hat{\mathbf{Var}}[\Delta]$
    Find c such that $P(\Delta \leq c) = \alpha_i$, $C[i] = c$.
end for
$C_{max} \leftarrow$ largest value in array C.

Links For Various Resources

- Hypothesis Testing: towardsdatascience.com , youtube.com/playlist
- Kernel Density Estimator: www.statsmodels.org , towardsdatascience.com , ekamperi.github.io
- Test Statistic & Central Limit Theorem: youtube.com/playlist , web.stanford.edu
- Likelihood Function, Log Likelihood function, Likelihood Ratio Test: youtube.com/playlist

Next Steps Are

- Implementing Three Algorithms.
- Testing on Different Data Sets.

Thank You