Univeariate analysis: Historgrams and Density plots

2024-09-02

```
params = list(
  truc= "Science des Données",
  year= 2023 ,
  curriculum= "L3 MIASHS",
  university= "Université Paris Cité",
  homepage= "https://stephane-v-boucheron.fr/courses/scidon",
  moodle= "https://moodle.u-paris.fr/course/view.php?id=13227",
  path_data = './DATA',
  country_code= '...',
  country= '...',
  datafile= '...'
  )

attach(params)
```

```
stopifnot(
  require(patchwork),
  require(glue),
  require(here),
  require(tidyverse),
  require(ggmosaic),
  require(skimr),
  require(plotly),
  require(DT),
  require(GGally),
  require(ggforce),
  require(ggfortify),
  require(vcd)
)

tidymodels::tidymodels_prefer(quiet = TRUE)

old_theme <-theme_set(theme_minimal(base_size=9, base_family = "Helvetica"))
```

- **L3 MIASHS**
- **Université Paris Cité**
- Année 2023-2024
- Course Homepage
- Moodle

Objectives

Density estimation

Histogram

A histogram is a piecewise constant density estimator.

Sliding window estimator

Let $h > 0$ be a bandwidth, let $x_1, \ldots, x_n$ be a sample, the sliding window density is defined by

$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{2h} \mathbb{I}_{[-1/2,1/2]} \left(\frac{x - x_i}{h} \right)$$

ou

$$\hat{f}_n(x) = \frac{1}{2h} \left(F_n(x+h) - F_n(x-h) \right)$$

Kernel density estimator

Simulations

Question

Simulate $N = 10$ samples of size $n = 500$ from a mixture of two Gaussian distributions $\lambda\mathcal{N}(0,1) + (1-\lambda)\mathcal{N}(\mu, \sigma^2)$.
Henceforth, λ is the *mixing* parameter. $\mathcal{N}(0,1)$ is the standard Gaussian and $\mathcal{N}(\mu, \sigma^2)$ is the non-standard Gaussian component of our *mixture* distribution,

Mixture distributions

Question

Plot regular histograms for different sample replicates.
Try different number of `bins` or `binwidths`.

Question

Repeat the above operations, but sample according the uniform distribution on $[0,1]$ but choose the breaks so that the intervals all have the same probability under the sampling distribution.

Question

Assume that you have chosen B bins.

- What is the distribution of the the number of sample points in a bin?
- What is the average number of points in a bin, what is its variance?
- Provide an upper bound on the expected maximum number of points in a bin.

Question

Assume that you have chosen B bins.
Compare the *empirical* distribution of the number of points in a bin with the theoretical distribution of the number of points in a bin.