Density Estimation

EN5422/EV4238 | Fall 2023

w06_density_1.pdf

(Week 6 – 1/2)

Contents

1 Density Estimation Intro

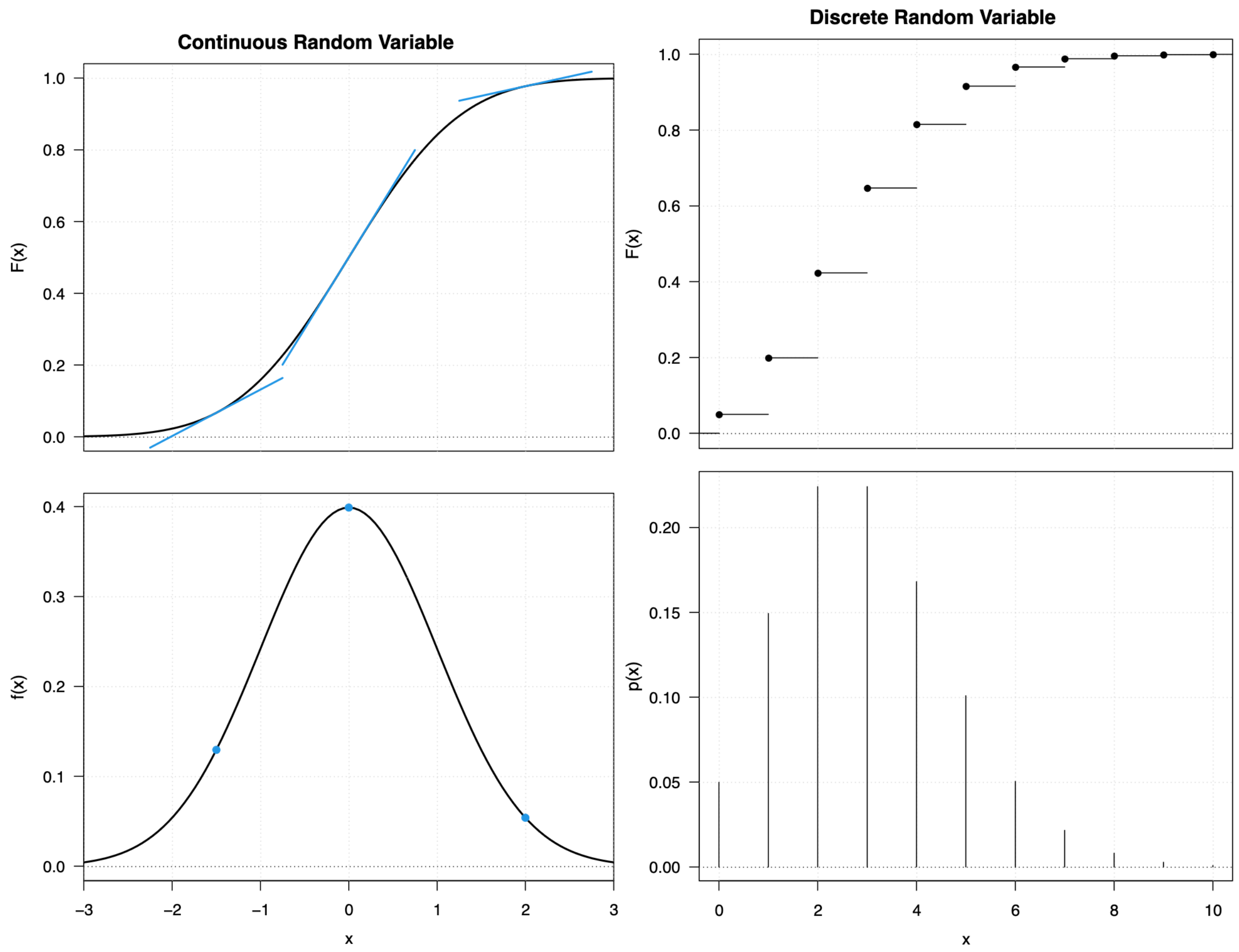
1.1 Distributions

- For many problems, so optimal decision can be formulated if we know the distribution of the relevant random variable(s).
    - The random variable(s) are the unknown or unobserved values.
- Often, only certain properties of the distribution (expected value, variance, quantiles) are needed to make decisions.
- Much of statistics is involved with estimation of the distributions or their properties.

1.1.1 Random Variables

Let *X* be a **random variable** of interest.

- The **cumulative distribution function (cdf)** is $F(x) = \Pr(X \leq x)$.
    - $F(x)$ is the probability that the random variable *X* ("big X") will take a value less than or equal to *x* ("little x").
- For *discrete* random variables, the **probability mass function (pmf)** is $f(k) = \Pr(X = k)$.
    - $f(k) \geq 0, \sum_k f(k) = 1$
    - $f(k) = F(k) - F(k-1)$
- For *continuous* random variables, the **probability density function (pdf)** is $f(x) = \frac{d}{dx} F(x)$.
    - $f(x) \geq 0, \int_{-\infty}^{\infty} f(x) = 1$



1.1.2 Parametric Distributions

A **parametric** distribution, $f(x;\boldsymbol{\theta})$ is one that is fully characterized by a set of parameters, $\boldsymbol{\theta}$. Examples include:

- □ Normal/Gaussian
  - ○ parameters: mean μ, standard deviation σ
- □ Poisson
  - ○ parameter: rate λ
- □ Binomial
  - ○ parameters: size n, probability p
- □ There are also multivariate version: Gaussian $N(\mu, \Sigma)$

If we can model (assume) the random variable follows a specific parametric distribution, then we only need to estimate the parameter(s) to have the entire distribution characterized. The parameters are often of direct interest themselves (mean, standard deviation).

Note

More details about some common parametric distributions can be found in the PyMC document.

1.1.3 Non-Parametric Distributions

A distribution can also be estimated using **non-parametric** methods (e.g., histograms, kernel methods, splines). These approaches do not enforce a parametric family (which is essentially a type of prior knowledge), but let the data *fully* determine the shape of the density/pmf/cdf. As you might imagine more data is required for these methods to work well. Non-parametric approaches are excellent for exploratory data analysis, but can also be very useful for other types of modeling (e.g., classification, anomaly detection). **All in all, non-parametric methods aim to estimate the distribution directly from the data without making strong assumptions.** non-parametric methods offer flexibility by avoiding strong distributional assumptions. While they can require more data and can be computationally intensive, their adaptability makes them valuable tools in a wide range of applications.

Note

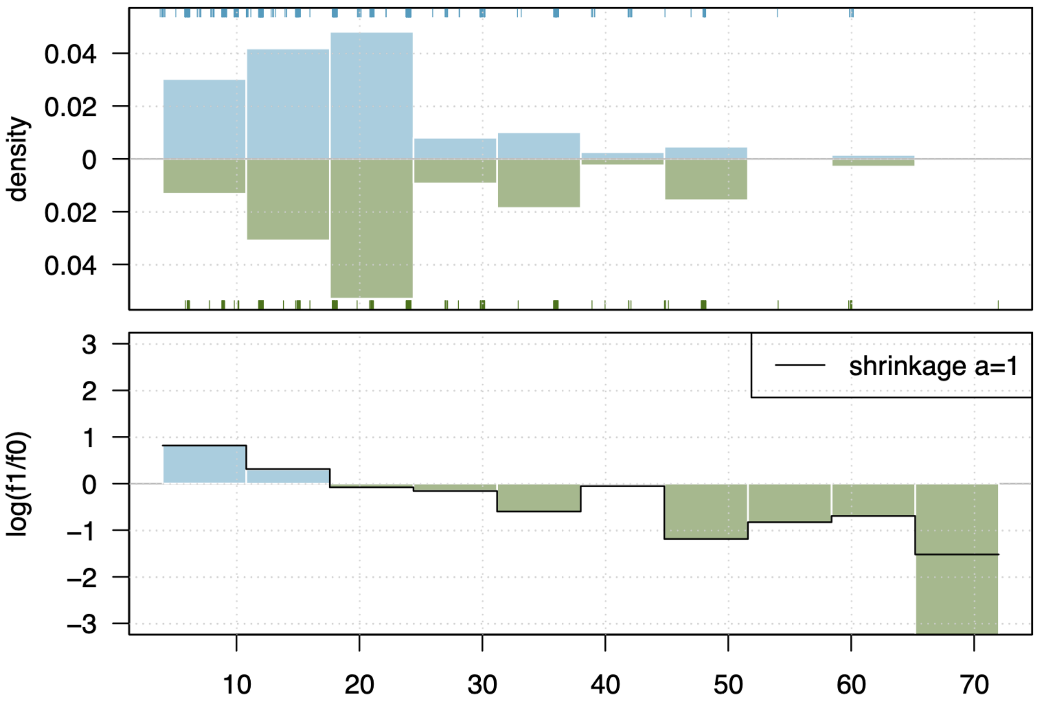
Because non-parametric methods don't make strong assumptions about the underlying distribution, they often require more data to provide accurate estimates. With too little data, a non-parametric estimate might be overly influenced by noise or random variations.

1.2 Example: Default Classification

Density estimation can be useful in *classification problems*, where the goal is to determine which class a new observation belongs to.

Below are two *histogram* density estimates; one for customers of a German bank that have good credit (blue) and the other for customers who defaulted (green). If a new customer is observed to have $X = 5$, then the evidence favors them having good credit because $X = 5$ is more likely under customers with good credit.

The bottom plot shows the corresponding log density ratio, which can help the bank make a decision on the customer’s credit-worthiness.



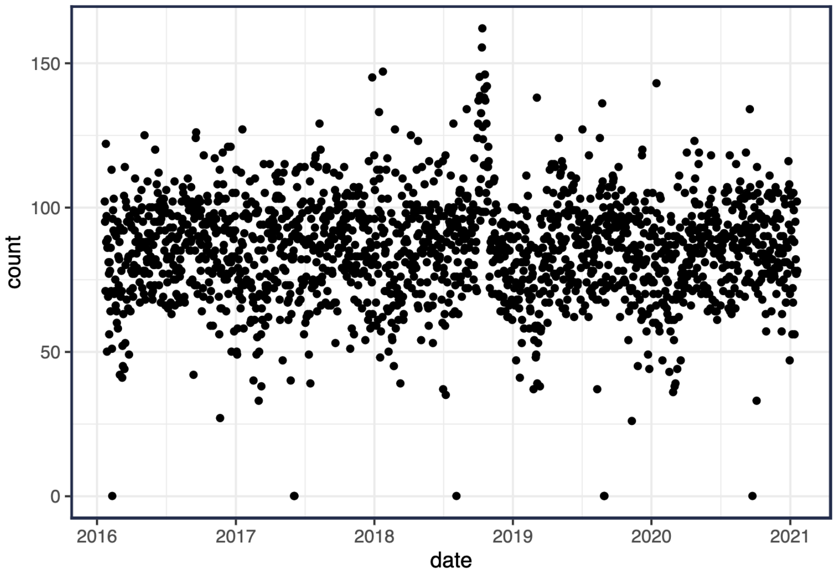
1.3 Example: Association Analysis

Density estimation can be useful in *association analysis*, where the goal is to find the regions with unusually high density (bump-hunting).

1.4 Example: Disease Outbreak Detection

Density estimation can be useful in *anomaly detection systems,* where the goal is to (often quickly) determine the time point when observations starting coming from a new or different distribution.

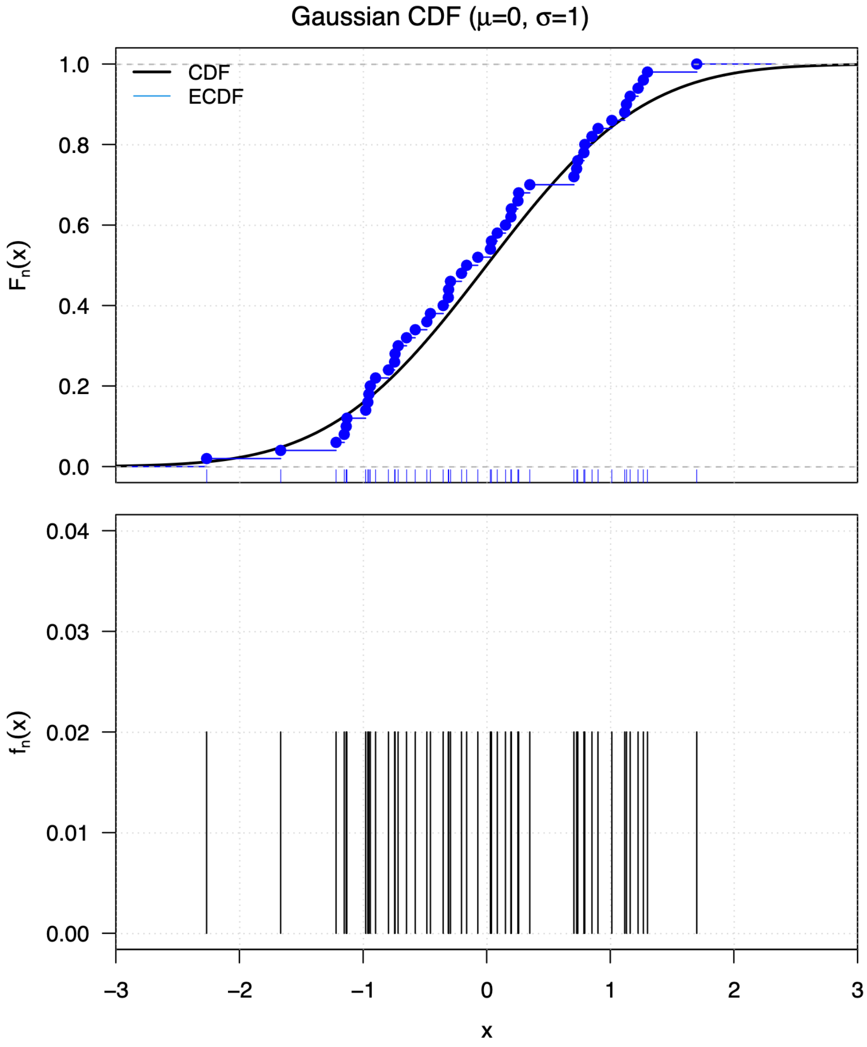
Below is simulated disease outbreak data representing the number of cases of some reported symptoms in an Emergency Department. If we can estimate the distribution of the baseline, or normal counts, on each day, then we will be able to flag an anomaly whenever the observations become *unlikely*.



1.5 Estimation

- These problems would be relatively easy to solve if we knew the exact distributions of the random variables of interest.
- Unfortunately, this is usually never the case (but of course: flipping coins, drawing cards, and playing with urns is different).
- We must use data to estimate the aspects/parameters of a distribution necessary to make good decisions.
    - And it is important to be mindful of the resulting uncertainty (bias, variance) in our estimation.

1.5.1 Empirical CDF and PDF



Data: $n = 50$ from $X \sim N(0, 1)$