Learning New Distributions (1st Assignment)

Pablo Dellacassa

Statistical Foundations for Data Science

October 26, 2020

Table of contents

Summary of the basic distributions: Uniform

Continuous uniform distribution or rectanguar distribution

Definition. A continuous random variable X has a uniform distribution on the interval $[a, b]$ if its Probability Density Function (PDF) f is given by the expression:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

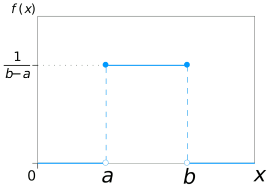
We denote this distribution by $U(b,a)$.

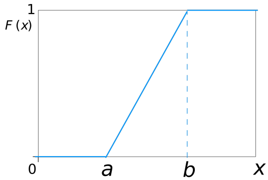
- **Highlights**: One of the simplest probability distributions in statistics.
- **Other formulas**:
  $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{b-a}{12}$ see demostrations

Summary of the basic distributions: Uniform

Below is shown the interpretation of equation (1) on the right. Graphically, the PDF is portrayed as a rectangle where $b - a$ is the base and $\frac{1}{b-a}$ the height. As the distance between a and b increases, the density at any particular value within the distribution boundaries decreases. Since the PDF integrates to 1, the height of the probability density function decreases as the base length increases. On the left is the Cumulative Distribution Function (CDF) shown, which behavior is:

$$\mathrm{F(x)} = \begin{cases} 0 & \text{for} \quad x < a \\ \frac{x-a}{b-a} & \text{for} \quad a \le x \le b \\ 1 & \text{for} \quad x \ge b \end{cases}$$





Bernoulli distribution

Definition. A discrete random variable X has a *Bernoulli distribution* with parameter p if its probability mass function f over possible outcomes k is:

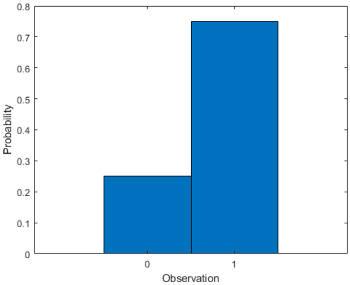
$$P(k;p) = \begin{cases} q = 1\text{-}p & \text{if } \; k{=}0 \\ p & \text{if } \; k{=}1 \end{cases} \tag{2}$$

- **Highlights**: It is used to model an experiment with only two possible outcomes, often referred to as "success" and "failure", usually encoded as 1 and 0. The Bernoulli distribution is a special case of the binomial distribution.
- **Other formulas**:

$$E(X) = p \quad \text{and} \quad V(X) = pq = p(1-p)$$ see demostrations

Summary of the basic distributions: Bernouilli

Below is shown the interpretation of equation (2) on the right. On the left is the CDF shown, which is: $f(k;p) = \begin{cases} 0 & \text{if} \quad k < 0 \\ 1-p & \text{if} \quad 0 \leq k < 1 \\ 1 & \text{if} \quad k \geq 1 \end{cases}$



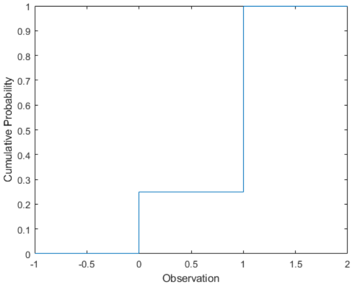


Table of contents

Logistic distribution: introduction and review

The logistic distribution is a continuous distribution function. Both its PDF and CDF have been used in many different areas such as logistic regression, logit models, neural networks. It has been used in the physical sciences, sports modeling, and recently in finance. The logistic distribution is very similar to normal but it has heavier tails (higher kurtosis) so it is more consistent with the underlying data and provides better insight into the likelihood of extreme events.

Logistic Distribution Statistics	
Notation	$Logistic(\mu, s)$
Parameter	$0 \leq \mu \leq \infty$ $s > 0$
Distribution	$0 \leq x \leq \infty$
Pdf	$\frac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left(1+\exp\left(-\frac{x-\mu}{s}\right)\right)^2}$
Cdf	$\frac{1}{1+\exp\left(-\frac{x-\mu}{s}\right)}$
Mean	μ
Variance	$\frac{1}{3}s^2\pi^2$
Skewness	0
Kurtosis	6/5

Table of contents

Logistic distribution: formulas, graphics, and parameters

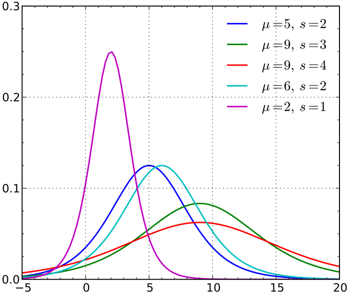
PDF: μ is the location and s is the scale parameters, μ makes the PDF slide along x axis and the scale makes the PDF fat or skinny. The PDF of the logistic distribution is given by:

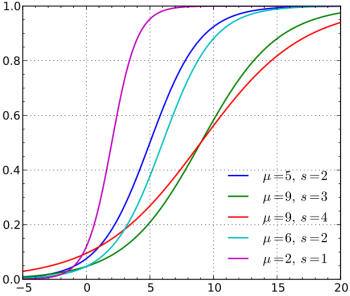
$$f(x;\mu,s) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2} \qquad -\infty < x < \infty$$

CDF: the logistic distribution receives its name from its cumulative distribution function, which is an instance of the family of logistic functions:

$$F(x;\mu,s) = \frac{1}{1+e^{-(x-\mu)/s}}$$

Logistic distribution: formulas, graphics, and parameters





Other formulas:

$$E(X) = \mu \quad \text{and} \quad V(X) = \frac{\pi^2 \sigma^2}{3} \quad \text{see demostrations}$$

Table of contents

Examples of application: *hydrology*

- In hydrology the distribution of long duration river discharge and rainfall (e.g., monthly and yearly totals, consisting of the sum of 30 respectively 360 daily values) is often thought to be almost normal according to the central limit theorem. The normal distribution, however, needs a numeric approximation. As the logistic distribution, which can be solved analytically, is similar to the normal distribution, it can be used instead.

This picture on the left illustrates an example of fitting the logistic distribution to ranked October rainfalls (that are almost normally distributed). It shows the 90% confidence belt based on the binomial distribution. The rainfall data are represented by plotting positions as part of the cumulative frequency analysis.

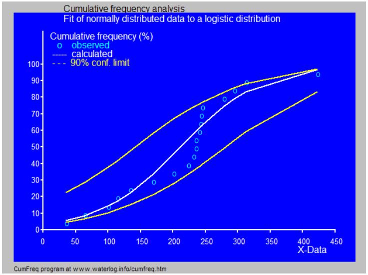


Table of contents

Python notebooks

GO!