Inference

Maura Mezzetti

Department of Economics and Finance
Università Tor Vergata

Outline

What is This Course About?

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression. *Ronald A. Fisher*

Inference

Probability theory have provided us with the tools and models to reason about random experiments. These provide realizations x of random variable X.
The objective of statistical inference is to learn about aspects of the probability distribution which produced x, to allow us to make inference about the nature of the process under study.
Statistical problems involve the reverse situation respect to probability: we obtain realizations (that is, sample observations) for some unknown probability model, and we wish to estimate the values of the model parameters.

Definition

Population is the collection of all individuals, families, groups, organizations, and units that we are interested in finding out about (e.g. Students at Tor Vergata)

Sample is a subset of a larger population actually observed (e.g. students in this room)

Definition

Random sample of size n is the multiple random variable $X = (X_1, X_2, , X_n)$ whose components are mutually independent and identically distributed with marginal distribution $f(x)$, where $f(x)$ denotes the probability distribution function of the random variable X.

Observed sample denoted by $x = (x_1, x_2, \ldots, x_n)$ is the realization of random sample.

Sample space is the set of all possible outcomes and it will be denoted by Ω. It may be discrete or continuous according to whether X is discrete or continuous.

Preliminary concept

We obviously do not know from which distribution the observed sample has been generated, but, in a parametric setting, we assume that it belongs to a certain **statistical model**, i.e. a family of probability distributions indexed by a parameter θ,

$$\{f(x;\theta),\, \theta \in \Theta\},$$

where $f(x;\theta)$ (sometimes indicated as $f(x|\theta)$) denotes a *probability mass function, (pmf)* or a *probability density function (pdf)*, Θ is the *parameter space* and θ may also be a vector of parameters, $\theta = (\theta_1 \ldots, \theta_p)$.

Preliminary concept

- Provided that the statistical model holds, the knowledge of the true value of θ is equivalent to the knowledge of the true distribution that generated the data. So, our aim is **making inference** on θ on the basis of the sample.
- The most common inferential procedure is referred to as point estimation. It consists in estimating θ through a single value (point) of the parameter space (Θ).

Preliminary Concepts

Parameter is a numerical characteristic of the population (generally unknown). Generally indicated with θ, and the parameter space Θ.

Statistic (or sample statistic) is a numerical function of sample that does not depend on any unknown parameter

Estimator is a statistic used to estimate a population parameter.

Exponential Distribution

- **Definition**: A random variable X is exponential with parameter λ (shortly, $X \sim Exp(\lambda)$) if:

$$f_X(x|\lambda) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- The exponential distribution models an arrival time which can take any positive real value.
- The higher λ, the more it is concentrated near zero.
- Verify $E(X) = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$

Gamma Distribution

Definition A random variable X has the Gamma distribution with parameters (α, β), (shortly $X \sim \Gamma(\alpha, \beta)$) if:

$$f_X(x|(\alpha,\beta)) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta) & \text{if } x \geq 0, \alpha > 0, \beta > 0 \\ 0 & \text{otherwise} \end{cases}$$

- $\Gamma(\alpha)$ is a normalizing constant defined by:

$$\Gamma(n) = \int_0^\infty \lambda^n t^{n-1} exp(-\lambda t) dt$$

- The parameter α is known as the shape parameter, since it most influences the peakedness of the distribution, while the parameter β is called the scale parameter, since most of its influence is on the spread of the distribution.
- Verify $E(X) = \alpha\beta$ and $Var(X) = \alpha\beta^2$.

The Normal distribution

Definition A random variable X has the Normal or Gaussian distribution with parameters (μ, σ), (shortly $X \sim \mathcal{N}(\mu, \sigma^2)$) if:

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

- If $X \sim \mathcal{N}(0, 1)$, X is a standard normal
- If $X \sim \mathcal{N}(\mu, \sigma)$, $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- Use TABLES!
- Later other properties....

Exponential Family

A family of probability distribution function is called *an exponential family* if it can be expressed as:

$$f(x|\theta) = h(x)c(\theta)exp\left(\sum_{i=1}^{n}\omega_i(\theta)t_i(x)\right)$$

Here $h(x) \geq 0$ and $t_1(x), \ldots, t_k(x)$ are real-valued functions of the observation x (they cannot depend on θ), and $c(\theta) > 0$ and $\omega_1(\theta), \ldots, \omega_k(\theta)$ are real-valued functions of the possibly vector-valued parameter θ (they cannot depend on x). Many common distributions belong to this family: gaussian, poisson, binomial, exponential, gamma and beta. (**PROOF TO BE DONE!**)

Random Sample

Definition The random variables $(X_1, \ldots, X_n)$ are called a random sample of size n from the population $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf of each X_i is the same function $f(x)$.

A random samples $(X_1, \ldots, X_n)$ of size n from a population distribution with pdf $f(x)$, by definition of independence, we can write the joint pdf as

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

With some abuse of notation we will often denote $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ simply by $f(x_1, \ldots, x_n)$ or $f(\underline{x})$.

Random Sample

In empirical experiment it will be very unusual to observe only one value of one random variable. Usually n values are observed, and this different values will be considered as observations on different random variables, one for each observed value.

$X_1, X_2, \ldots, X_n$ are called i.i.d random variables:

- i: independent
- i: identically
- d: distributed

What is the distribution of a random sample?

Let $X_1, \ldots, X_n$ be a random sample from an exponential (β) population, corresponding, for example, to the times until failure (measured in year) for n identical circuit boards that are put on test and used until they fail. What is the probability that *all* the boards last more than 2 years?

$$\begin{aligned} P(X_1 > 2, \quad \ldots \quad , X_n > 2) = & \\ &= \int_2^\infty \ldots \int_2^\infty \left(\prod_{i=1}^n \frac{1}{\beta} \exp\left(-x_i/\beta\right) \right) dx_1 \ldots dx_n \\ &= \exp\left(-2n/\beta\right) \end{aligned}$$

What is the distribution of a random sample?

Let $X_1, \ldots, X_n$ be a random sample from an exponential (β) population, corresponding, for example, to the times until failure (measured in year) for n identical circuit boards that are put on test and used until they fail.
Define the variable $Y_i = I(X_i > 2)$

- What is the distribution of Y_i?
- What is the distribution of $\sum_{i=1}^{n} Y_i$?

What is the distribution of a random sample?

Let X be the height (in centimeters) of an Italian University student, and assume $X \sim N\left(170, \sigma^2 = 100\right)$. Let $X_1, \ldots, X_{10}$ be a random sample of 10 students from X, what is the probability all of them will be smaller than 180 centimeters?

$$\begin{aligned} P(X_1 < 180, \ldots, X_{10} < 180) &= P(X < 180)^{10} \\ &= (0.8413)^{10} = 0.1777 \end{aligned}$$

What is the distribution of a random sample?

Let X be a Bernoulli random variable with $\pi = 0.4$, assuming value 1 if a person surveyed smokes. Let $X_1, \ldots, X_5$ be a random sample of 5 students from X, what is the probability three of them smoke?

$$\begin{aligned} P\left(\sum_{i=1}^{5} X_i = 3\right) &= \binom{5}{3} 0.4^3 0.6^2 \\ &= 0.2304 \end{aligned}$$

Sample Statistics

- An observed sample is used to make inference on the parameter of interest θ. However, the sample usually consists in a long list of numbers that may be hard interpret. So, to *summarize this information*, we usually make use of a sample statistic.
- A **statistic** is any real-valued function of the random sample $t(X) = t(X_1, \ldots, X_n)$
- The most common sample statistics are:
  - Sample mean: $\bar{X} = \frac{\sum_i X_i}{n}$
  - Sample variance $S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n}$
  - Order statistics: $Y_1 \leq Y_2 \leq \ldots \leq Y_n$

Sampling Distribution

- Just as the observations vary from sample to sample, so does any value computed from a sample. *Capturing this uncertainty is a key component of statistical analysis*.
- With this concept of uncertainty in mind, we now define:
- *A* **statistic** *is any quantity calculated from sample data. Prior to obtaining data, there is uncertainty as to which value of the statistic will occur. Therefore a statistic is a random variable.*
- Since a statistic is a random variable it has a probability distribution. The probability distribution of a statistic is often called the sampling distribution, to stress that it reflects how the statistic varies in value across the possible samples that could be selected.

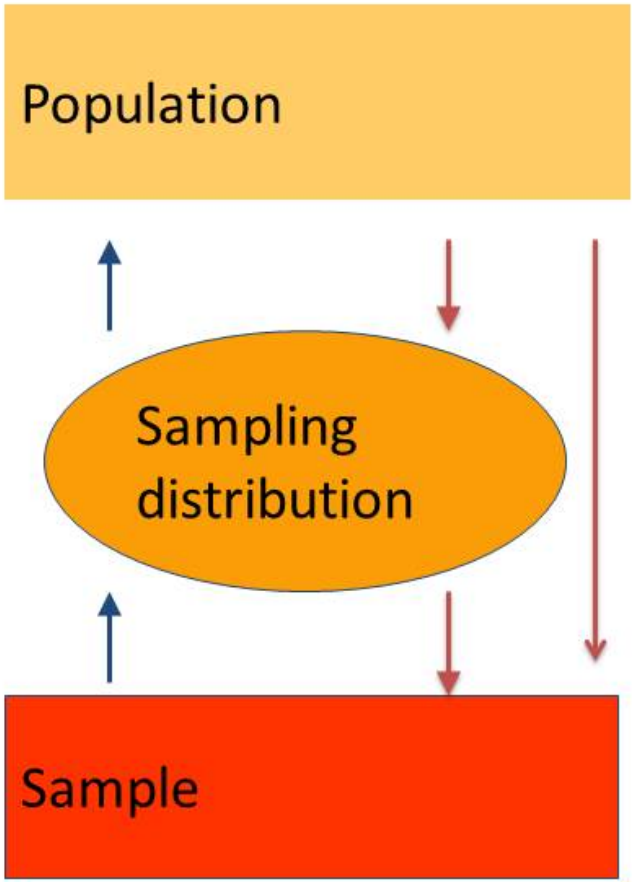
Sampling Distribution

- Since the random sample is a multiple random variable, any statistic is a random variable with a distribution that is usually referred to as **sampling distribution**.
- Formally, the sampling distribution of $Y = t(\mathbf{X})$ may be represented by the distribution function:

$$F(y) = P(t(\mathbf{X}) \leq y)$$

- The sampling distribution is often difficult to derive, but its moments may be easily computed when the statistic of interest is the sample mean or the sample variance.

Probability/Inference



- Sample Distribution: empirical, known (shape, central tendency and dispersion)
- Population Distribution: unknown (purpose of inference is collecting informations on population distribution)
- Sampling Distribution (not empirical, theoretical)

Example

Consider the random variable X, with probability distribution function given by:

x	1	2	3
p(X=x)	0.2	0.3	0.5

Suppose we draw two numbers X_1 and X_2 independently, according to $P(X = x)$, and are interested in the mean $\bar{X} = (X_1 + X_2)/2$. Independence means we can compute $P(X_1 = x_1 \cap X_2 = x_2)$ as the product of the marginal probabilities. The small range of X enables us to evaluate all pairs (x_1, x_2), and the corresponding mean.

Example

x_1	x_2	$\bar{x}$	$p(x_1, x_2)$
1	1	1	0.04
1	2	1.5	0.06
1	3	2	0.1
2	1	1.5	0.06
2	2	2	0.09
2	3	2.5	0.15
3	1	2	0.1
3	2	2.5	0.15
3	3	3	0.25

Example

The following distribution for the sample mean is obtained:

$\bar{x}$	1	1.5	2	2.5	3
$P(\bar{X} = \bar{x})$	0.04	0.12	0.29	0.3	0.25

Note that the expected value of $\bar{X}$ is 2.3 and is equal to expected value of X, $\mu = 2.3$ and the theoretical variance is $\sigma^2 = 0.61$ and variance of sample mean is exactly σ^2/n.
In this case, on average, the sample mean gives the population value, and the variability of the mean is less than the variability of the original distribution.

Example

Population - unknown

Let X a r.v with distribution

x	1	2	3
P(X=x)	0.2	0.3	0.5

Random sample

X_1, X_2 i.i.d is a random sample of size $n = 2$ from the population X

Observed sample

x_1, x_2 is the realization of the random sample.

Example

Population - unknown

Let X a r.v with distribution

x	1	2	3
P(X=x)	0.2	0.3	0.5

Statistic

The mean $\bar{X} = \frac{X_1+X_2}{2}$

Sample space

The small range of X enables us to evaluate all pairs (x_1, x_2), and the corresponding mean.

Theoretical \ Empirical

Population

For the population $E(X) = 2.3,\ Var(X) = 0.61$

Statistics

For the statistics $E(\bar{X}) = 2.3,\ Var(\bar{X}) = 0.305$

Empirical?

see matlab

Sample Mean

Definition Let $X_1, X_2, \ldots, X_n$ be a random sample from a population X, the sample mean is defined as $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

Theorem Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$, then

1. $E\left(\bar{X}\right) = \mu$
2. $Var\left(\bar{X}\right) = \frac{\sigma^2}{n}$

(Proof to do as exercise)

Sample Mean

Theorem Let $X_1, X_2, \ldots, X_n$ be a random sample from a population $X \sim N(\mu, \sigma^2)$, the sample mean is Normally distributed:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

EXAMPLE: Sample Mean

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a population $X \sim Bernoulli(\pi)$, what is the distribution of the the sample mean?
- The distribution of the sample sum is Binomial with parameter (n, π)
- The distribution of sample mean is proportional to the distribution of a Binomial.

Distribution of $Y = X_1 + \ldots + X_n$

Bernoulli

Let $X_i, i = 1, 2, \ldots, n$, a random sample from a Bernoulli population X with parameter π then

$$Y = X_1 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

has a Binomial distribution with parameters (n, π)

Distribution of $Y = X_1 + \ldots + X_n$

Gaussian

Let X_i, $i = 1, 2, \ldots, n$ a random sample from a Gaussian distribution with parameters μ and σ^2 then

$$Y = X_1 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

has a Gaussian distribution with parameters $n\mu$ and $n\sigma^2$

Distribution of $Y = X_1 + \ldots + X_n$

Poisson

Let $X_i, i = 1, 2, \ldots, n$, a random sample from a Poisson population X with parameter λ then

$$Y = X_1 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

has a Poisson distribution with parameter $n\lambda$

Distribution of $Y = X_1 + \ldots + X_n$

Exponential

Let X_i, $i = 1, 2, \ldots, n$ a random sample from a Exponential population X with parameters λ then

$$Y = X_1 + \ldots + X_n = \sum_{i=1}^{n} X_i$$

has a Gamma distribution with parameter $(n, 1/\lambda)$

Sampling from any distribution

Theorem If X is a continuous random variable with cumulative distribution function F_X, then the random variable $Y = F_X(X)$ has a uniform distribution on $[0, 1]$.

$$
\begin{aligned}
F(y) &= P(Y \leq y) = P(F_X(X) \leq y) \\
F(y) &= P(X \leq F_X^{-1}(y)) \\
F(y) &= F(F_X^{-1}(y)) \\
F(y) &= y
\end{aligned}
$$

Interpretation of Limit Theorems

- One interpretation of the statement "the event A has probability p" is that, if we keep repeating the experiment, the proportion of times that A occurs should converge to p.
- After giving the axiomatic definition of probability, we now can recover this property as a Theorem.
- We call Limit Theorems statements that involve infinitely many random variables, and that do not depend on the outcome of any finite set of them.

Markov and Chebishev Inequalities

Proposition Let X be a positive random variable (i.e. $P(X \geq 0) = 1$) with finite expectation. Then for all $a > 0$ we have that

$$P(X > a) \leq \frac{E(X)}{a}$$

Corollary Let X be a random variable with finite variance. Then, for all $a > 0$ we have that:

$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2}$$

The underlying idea is that from information about expectations of functions of X, you can get information about the distribution of X itself.

The Weak Law of Large Numbers- WLLN

- **Proposition** Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables independent, identically distributed with finite expectation μ. Then for every $\varepsilon > 0$, we have that:

$$lim_{n \to \infty} P\left(\left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \mu \right| \geq \varepsilon \right) = 0$$

- The Weak Law of Large Numbers basically says that, no matter what ε one chooses, for sufficiently large n the probability that the sample mean is farther than ε from the theoretical mean will be very small.

Convergence of Random Variables

Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of random variables:

- Y_n converges to Y **almost surely** (with probability 1) if:

$$P\left(lim_{n\to\infty} Y_n = Y\right) = 1$$

- Y_n converges to Y **in probability** if for all $\varepsilon > 0$:

$$lim_{n\to\infty} P\left(|Y_n - Y| \geq \varepsilon\right) = 0$$

- Y_n converges **in law** (or converges **in distribution**) to Y if, for all t where F_Y is continuous (if Y has a density for all t):

$$lim_{n\to\infty} F_{Y_n}(t) = F_Y(t)$$

Convergence of Random Variable

- Almost sure convergence is very strong: it says that it is (almost) impossible that Y_n does not converge to Y. No matter what, Y_n will always converge to Y , even though it may take a long time.
- Convergence in probability says that, for any confidence level ε eventually the probability that Y_n is further than ε from Y will converge to zero. But it does not say what happens for any sample $\omega = (x_1, x_2, \ldots, x_n)$.

Convergence of Random Variable

- Convergence in law says that for n big, the distribution of Y_n is very similar to that of Y , so we may well approximate any probability $P(Y_n \in [a, b])$ with the probability $P(Y \in [a, b])$.
- Convergence almost surely implies convergence in probability, which implies convergence in law. No other implication is true in general.

Weak Law of Large Number as a Convergence Results

We may now restate the WLLN as follows:
Proposition (WLLN) Let $X_1, X_2, \ldots$ be a sequence IID random variables with finite expectation μ. Then the sample means converge to μ in probability.
Under the same assumptions a stronger result holds true:
Proposition (SLLN) Let $X_1, X_2, \ldots$ be a sequence IID random variables with finite expectation μ. Then the sample means converge to μ almost surely.
Define : $S_n = X_1 + X_2 + \ldots + X_n$ and $M_n = \frac{S_n}{n}$

Central Limit Theorem

- The Law of Large Numbers says that sample means M_n tends to a constant, and this is easy to check with finite variances.
- Let $X_1, X_2, \ldots$ be a sequence of IID r.v. with expectation μ and variance σ^2, and consider the quantities:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma}(M_n - \mu)$$

- It is easy to check that Z_n have all expectation 0 and variance 1, so if variance passes to the limit (and in this case it does), they cannot possibly converge to a constant.
- Where are they converging then? Do they really converge? In which sense?

Central Limit Theorem

Theorem Let $X_1, X_2, \ldots$ be a sequence of IID r.v. with expectation μ and variance σ^2. Then, the random variable:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma}(M_n - \mu)$$

converge in law to a standard normal $N(0, 1)$.
The Law of Large Numbers and the Central Limit Theorem require specific integrability assumptions: finite expectation for the Law of Large Numbers, and finite variance for the Central Limit Theorem. When these assumptions are not satisfied, the conclusions may not be true.

Sampling from the normal distribution

Theorem Let $X_1, X_2, \ldots$ be a sequence of IID Gaussian r.v. with expectation μ and variance σ^2. Then,

- $\bar{X}$ and S^2 are independent random variables,
- $\frac{(n-1)S^2}{\sigma^2}$ has a Chi-square distribution with $(n-1)$ degrees of freedom.

Given $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

Empirical Distribution Function

Definition If we have a random sample $X_1, X_2, \ldots, X_n$ of size n then the *Empirical Distribution Function*, $\hat{F}_n(x)$ is the cdf of the distribution that puts mass $1/n$ at each data point X_i. Thus by definition

$$\hat{F}_n(x) = \sum_{i=1}^{n} \frac{I(X_i \leq x)}{n}$$

$\hat{F}_n(x)$ shows the fraction of observations with a value smaller or equal than x.

Properties of $\hat{F}$

- At any fixed value of x, $E(\hat{F}(x)) = F(x)$
- $Var(\hat{F}(x)) = \frac{1}{n}F(x)(1 - F(x))$
- Note that these two facts imply that

$$\hat{F}(x) \longrightarrow^{P} F(x)$$

- An even stronger proof of convergence is given by the Glivenko-Cantelli Theorem:

$$sup_x|\hat{F}(x) - F(x)| \longrightarrow_{a.s.} 0$$

Order Statistics

Given a random sample $X_1, X_2, \ldots, X_n$, the sample **order statistics** are the sample values placed in ascending order

$$\begin{aligned} X_{(1)} &= min_{1 \leq i \leq n} X_i \\ X_{(2)} &= second\ smallest\ X_i \\ \ldots &= \ldots \\ X_{(n)} &= max_{1 \leq i \leq n} X_i \end{aligned}$$

Order Statistics

Remarks:

- Order statistics are random variables themselves (as functions of a random sample).
- Order statistics satisfy

$$X_{(1)} \leq \ldots \leq X_{(n)}$$

- Though the samples $X_1, X_2, \ldots, X_n$ are independently and identically distributed, the order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$ are never independent because of the order restriction.
- We will study their marginal distributions and joint distributions.

Distributions of Order Statistics - Continuous Case

Assume $X_1, X_2, \ldots, X_n$ are from a continuous population with cdf $F(x)$ and pdf $f(x)$. Then

- The $n-th$ order statistic, or the sample maximum, $X(n)$ had the pdf
$$f_{X_{(n)}}(x) = n\left[F(X)\right]^{n-1} f(x)$$
- The first order statistic, or the sample minimum, $X(1)$ had the pdf
$$f_{X_{(1)}}(x) = n\left[1 - F(X)\right]^{n-1} f(x)$$
- More generally, the $j-th$ order statistic $X(j)$ had the pdf
$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)\left[F(X)\right]^{j-1}\left[1 - F(X)\right]^{n-j}$$

Distributions of Order Statistics -Uniform Case

Assume $X_1, X_2, \ldots, X_n$ are from a continuous population $U(a, b)$ and pdf $f(x)$. Then

- The $n - th$ order statistic, or the sample maximum, $X(n)$ had the pdf

$$f_{X_{(n)}}(x) = \frac{n}{b-a}\left[\frac{x-a}{b-a}\right]^{n-1}$$

- The first order statistic, or the sample minimum, $X(1)$ had the pdf

$$f_{X_{(1)}}(x) = \frac{n}{b-a}\left[\frac{b-x}{b-a}\right]^{n-1}$$