Math 408 - Mathematical Statistics

Lecture 34. Summarizing Data

April 24, 2013

Agenda

- Methods Based on the CDF
  - The Empirical CDF
    - Example: Data from Uniform Distribution
    - Example: Data from Normal Distribution
  - Statistical Properties of the eCDF
  - The Survival Function
    - Example: Data from Exponential Distribution
  - The Hazard Function
    - Example: The Hazard Function for the Exponential Distribution
- Summary

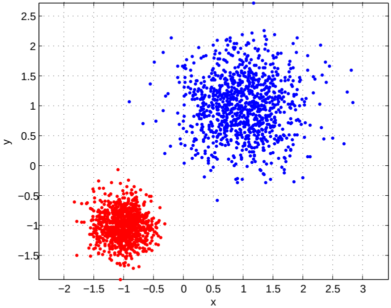
Describing Data

In the next few Lectures we will discuss methods for describing and summarizing data that are in the form of one or more samples. These methods are useful for revealing the structure of data that are initially in the form of numbers.

Example: the arithmetic mean $\overline{x} = (x_1 + \ldots + x_n)/n$ is often used as a summary of a collection of numbers $x_1, \ldots, x_n$: it indicates a "typical value".

Example:

- $x = (1.5147, 1.7223, 1.063, 1.4916, \ldots)$
- $y = (0.7353, 0.0781, 0.276, 1.5666, \ldots)$



Empirical CDF

Suppose that $x_1, \dots, x_n$ is a batch of numbers.

Remark: We use the word

- "sample" when $X_1, \dots, X_n$ is a collection of random variables.
- "batch" when $x_1, \dots, x_n$ are fixed numbers (realization of sample).

Definition

The **empirical cumulative distribution function** (eCDF) is defined as

$$\boxed{F_n(x) = \frac{1}{n}(\#x_i \leq x)}$$

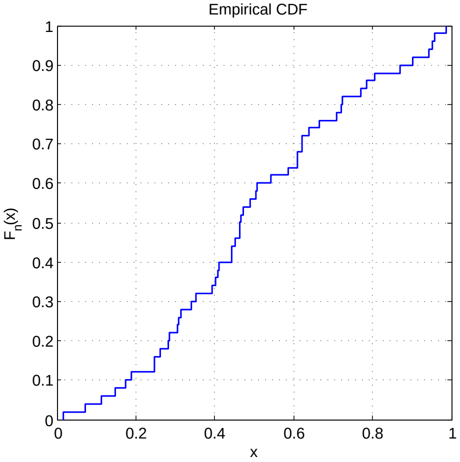
Denote the ordered batch of numbers by $x_{(1)}, \dots, x_{(n)}$.

- If $x < x_{(1)}$, then $F_n(x) = 0$
- If $x_{(1)} \leq x < x_{(2)}$, then $F_n(x) = 1/n$
- If $x_{(k)} \leq x < x_{(k+1)}$, then $F_n(x) = k/n$

The eCDF is the "data analogue" of the CDF of a random variable

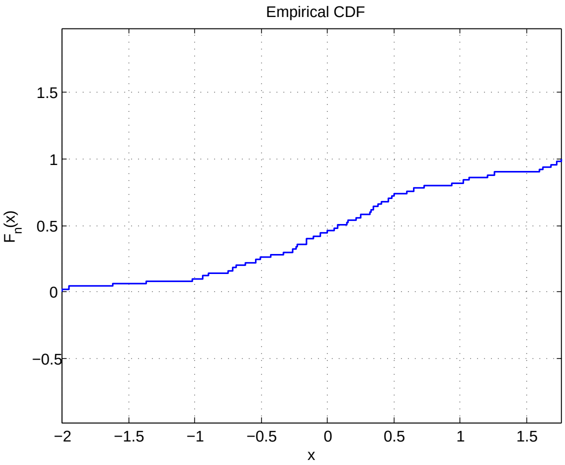
Example: Data from Uniform Distribution

- Let $(X_1, \ldots, X_n) \sim U[0, 1]$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 50$
  - $(x_1, \ldots, x_n) = (0.24733, 0.3527, 0.18786, 0.49064, \ldots)$



Example: Data from Normal Distribution

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0,1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 50$
  - $(x_1, \ldots, x_n) = (-0.23573, 0.45952, -0.93808, -0.62162, \ldots)$



Statistical Properties of the eCDF

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution F. Then the eCDF can be written as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i),$$

where

$$I_{(-\infty,x]}(X_i) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{if } X_i > x \end{cases}$$

The random variables $I_{(-\infty,x)}(X_1), \ldots, I_{(-\infty,x)}(X_n)$ are independent Bernoulli random variables:

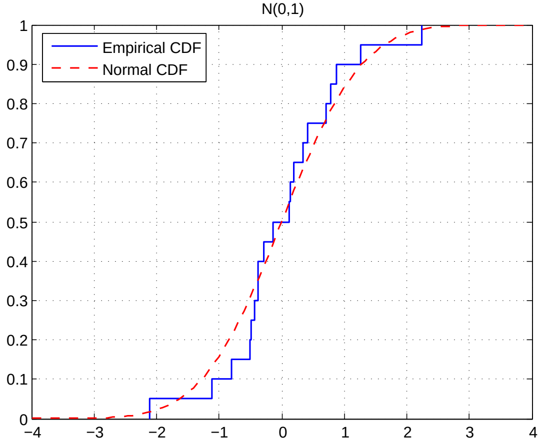
$$I_{(-\infty,x)}(X_i) = \begin{cases} 1, & \text{with probability } F(x) \\ 0, & \text{with probability } 1 - F(x) \end{cases}$$

Thus, $nF_n(x)$ is a binomial random variable: $nF_n(x) \sim \mathrm{Bin}(n, F(x))$

- $\mathbb{E}[F_n(x)] = F(x)$
- $\mathbb{V}[F_n(x)] = \frac{1}{n} F(x)(1 - F(x))$
- $\mathbb{V}[F_n(x)] \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty$

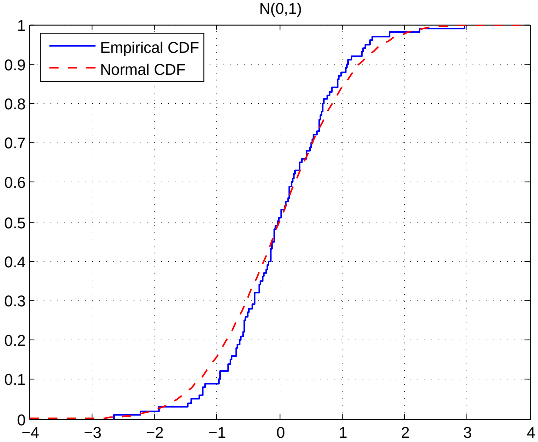
Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 20$



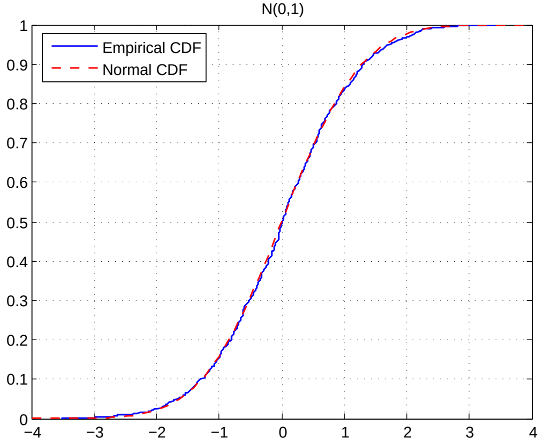
Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 100$



Example: Convergence of the eCDF to the CDF

- Let $(X_1, \ldots, X_n) \sim \mathcal{N}(0,1)$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 1000$



The Survival Function

The **survival function** is equivalent to the CDF and is defined as

$$\boxed{S(t) = \mathbb{P}(T > t) = 1 - F(t)}$$

In applications where the data consists of times until failure or death (and are thus nonnegative), it is often customary to work with the survival function rather than the CDF, although the two give equivalent information.
Data of this type occur in

- medical studies
- reliability studies

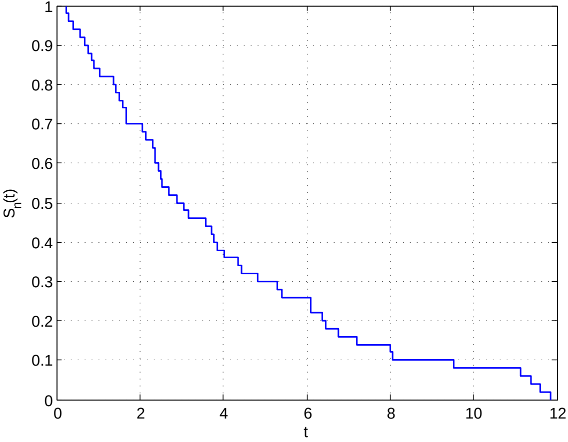
$$\boxed{S(t) = \text{ Probability that the lifetime will be longer than } t}$$

The data analogue of $S(t)$ is the **empirical survival function**:

$$\boxed{S_n(t) = 1 - F_n(t)}$$

Example: Data from Exponential Distribution

- Let $(X_1, \ldots, X_n) \sim \mathrm{Exp}(\beta)$, $\beta = 5$
- Let $(x_1, \ldots, x_n)$ is a particular realization of $(X_1, \ldots, X_n)$, $n = 50$
  - $(x_1, \ldots, x_n) = (4.4356, 1.684, 11.376, 4.8357, \ldots)$



The Hazard Function

Let T is a random variable (time) with the CDF F and PDF f.

Definition

The **hazard function** is defined as

$$\boxed{h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}}$$

- The **hazard function** may be interpreted as the instantaneous death rate for individuals who have survived up to a given time: if an individual is alive at time t, the probability that individual will die in the time interval $(t, t + \epsilon)$ is

$$\mathbb{P}(t \leq T \leq t + \epsilon | T \geq t) \approx \frac{\epsilon f(t)}{1 - F(t)}$$

- If T is the lifetime of a manufactured component, it maybe natural to think of $h(t)$ as the age-specific failure rate. It may also be expressed as

$$h(t) = -\frac{d}{dt} \log S(t)$$

Example: Hazard Function for the Exponential Distribution

Let $T \sim \mathrm{Exp}(\beta)$, then

- $f(t) = \frac{1}{\beta} e^{-t/\beta}$
- $F(t) = 1 - e^{-t/\beta}$
- $S(t) = e^{-t/\beta}$
- $h(t) = \frac{1}{\beta}$

The instantaneous death rate is constant.

If the exponential distribution were used as a model for the lifetime of a component, it would imply that the probability of the component failing did not depend on its age.

Typically, a hazard function is U-shaped:

- the rate of failure is high for very new components because of flaws in the manufacturing process that show up very quickly,
- the rate of failure is relatively low for components of intermediate age,
- the rate of failure increases for older components as they wear out.

Summary

- The empirical cumulative distribution function (eCDF) is

$$F_n(x) = \frac{1}{n}(\#x_i \leq x)$$

- The survival function is equivalent to the CDF and is defined as

$$S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

- The data analogue of $S(t)$ is the empirical survival function:

$$S_n(t) = 1 - F_n(t)$$

- The hazard function is

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

  - may be interpreted as the instantaneous death rate for individuals who have survived up to a given time