

Lesson 9. Input Data Analysis - Continuous Distributions

1 Continuous random variables and distributions: review

- A random variable is **continuous** if it can take on a continuum of values
- Let X be a continuous random variable
- The **cumulative distribution function (cdf)** F_X of X is

$$F_X(a) = \Pr\{X \leq a\}$$

- The **probability density function (pdf)** p_X of X is

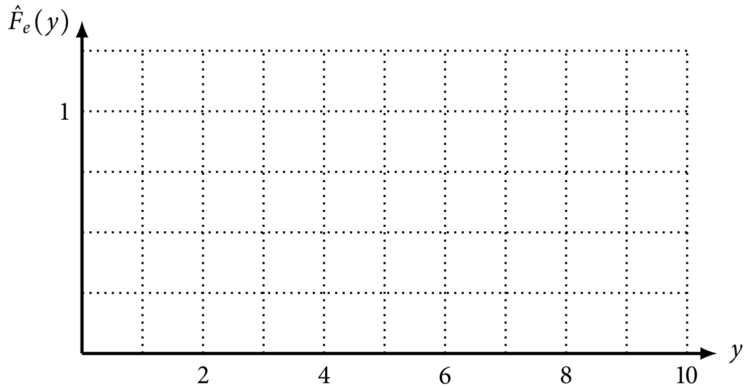
- Another way the pdf and cdf of a continuous random variable are related:

2 The empirical cdf

- Let $Y_0, \ldots, Y_{n-1}$ be n independent and identically distributed (iid) random variables with cdf F_Y
- Let $y_0, \ldots, y_{n-1}$ be observations of $Y_0, \ldots, Y_{n-1}$
- In words, $F_Y(a) = \Pr\{Y \leq a\} \approx$

- The **empirical cdf** is
  - Note that $F_e(a)$ is a random variable for any fixed value of a
- The **observed empirical cdf** is

Example 1. Let $n = 4$. Suppose the observations of Y_0, Y_1, Y_2, Y_3 are $y_0 = 3$, $y_1 = 1$, $y_2 = 8$, $y_3 = 4$. Plot the observed empirical cdf $\hat{F}_e$.



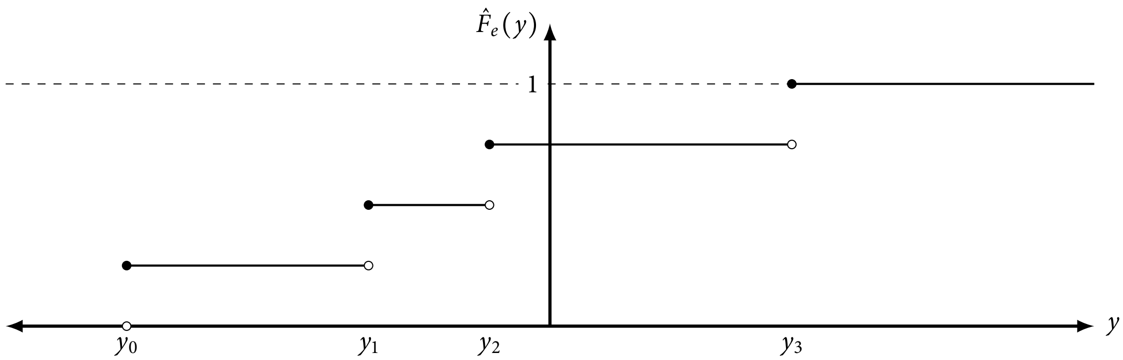
- Let $y_{(0)}, y_{(1)}, \dots, y_{(n-1)}$ be the observations $y_0, \dots, y_{n-1}$ sorted from smallest to largest

$\Rightarrow$ $\hat{F}_e\left(y_{(i)}\right) =$ for $i = 0, 1, \dots, n-1$.

3 Kolmogorov-Smirnov goodness-of-fit test

- Let $Y_0, \dots, Y_{n-1}$ be n iid continuous random variables
- Let $y_0, \dots, y_{n-1}$ be observations of $Y_0, \dots, Y_{n-1}$
- Let X be the proposed continuous random variable with cdf F_X
- The **Kolmogorov-Smirnov (K-S) goodness-of-fit test** compares the empirical cdf of the Y_j's with the cdf of the proposed random variable X
- Question: Do the Y_j's share the same distribution as X?
- Null hypothesis H_0: for any Y_j,
- The **test statistic** is
- The **observed test statistic** is

- The p-**value** is $\Pr\{D \geq d\}$
  - $\sqrt{n}D$ follows a **Kolmogorov distribution**
  - Important caveat: $\sqrt{n}D$ does not follow a Kolmogorov distribution if the proposed distribution of X depends on estimates based on the observations $y_0, \dots, y_{n-1}$
  - e.g. if you propose X as an exponential random variable, but guess the mean based on $y_0, \dots, y_{n-1}$
  - There are ways around this, some quick-and-dirty, some more rigorous
- How do we compute d? Do we really need to consider all values of x?



- So, we can compute the observed test statistic as