

# Sampling Estimates of the Cumulative Distribution Function

## Contents

- [Objectives](#)
- [Estimating the cumulative distribution function](#)

```
import numpy as np
np.set_printoptions(precision=3)
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set(rc={"figure.dpi":100, 'savefig.dpi':300})
sns.set_context('notebook')
sns.set_style("ticks")
```

## Objectives

- To estimate cumulative distribution function.

## Estimating the cumulative distribution function

We would like to estimate the cumulative distribution function of  $Y = g(X)$ :

$$F(y) = p(Y \leq y) = p(g(X) \leq y).$$

The key is to observe that it can be written as an expectation using the indicator function:

$$F(y) = \mathbb{E}[1_{[-\infty, y]}(g(X))].$$

This suggests that we should consider the random variables  $1_{[-\infty, y]}(g(X_1)), 1_{[-\infty, y]}(g(X_2)), \dots$  which are independent and identically distributed. By the strong law of large numbers, we have that:

$$\bar{F}_N(y) = \frac{1}{N} \sum_{i=1}^N 1_{[-\infty, y]}(g(X_i)) \rightarrow F(y) \text{ a.s..}$$

This estimate is called the empirical CDF. Note the neat interpretation:

$$\bar{F}_N(y) = \frac{1}{N} \sum_{i=1}^N 1_{[-\infty, y]}(g(X_i)) = \frac{\text{number of } g(X_i) \leq y}{N}.$$

## Example: 1D CDF

We will continue using the 1D test function of Example 3.4 of Robert & Casella (2004). Assume that  $X \sim U([0, 1])$  and pick:

$$g(x) = (\cos(50x) + \sin(20x))^2.$$

```

# define the function here:
g = lambda x: (np.cos(50 * x) + np.sin(20 * x)) ** 2

# We will not write code for the empirical CDF as it is already
#
https://www.statsmodels.org/stable/generated/statsmodels.distributions.empirical\_distribution.ECDF.html
from statsmodels.distributions.empirical_distribution import ECDF

# Maximum number of samples to take
max_n = 10000
# Generate samples from X
x_samples = np.random.rand(max_n)
# Get the corresponding Y's
y_samples = g(x_samples)

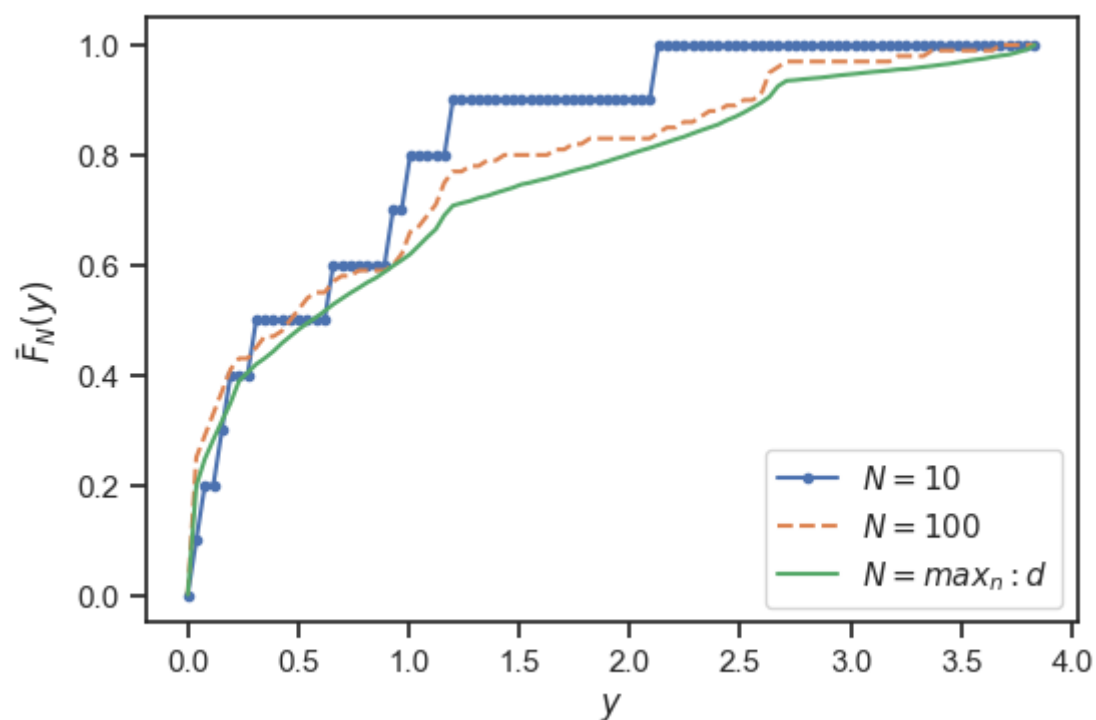
# Build ECDF with 10 samples
ecdf_10 = ECDF(y_samples[:10])

# Build ECDF with 50 samples
ecdf_100 = ECDF(y_samples[:100])

# Build ECDF with all samples
ecdf_all = ECDF(y_samples)

# Make the plot
fig, ax = plt.subplots()
# Points on which to evaluate the CDF's
ys = np.linspace(y_samples.min(), y_samples.max(), 100)
ax.plot(ys, ecdf_10(ys), "-.", label=r"$N=10$")
ax.plot(ys, ecdf_100(ys), "--", label=r"$N=100$")
ax.plot(ys, ecdf_all(ys), label=r"$N={max_n:d}$")
ax.set_xlabel("$y$")
ax.set_ylabel(r"$\bar{F}_N(y)$")
plt.legend(loc="best");

```



Let's now use the empirical CDF to find the probability of that  $Y$  takes specific values. For example, let's find the probability that  $Y$  is between 1 and 3. We have:

$$p(1 \leq Y \leq 3) = F(3) - F(1) \approx \bar{F}_N(3) - \bar{F}_N(1).$$

Let's calculate this numerically for various choices of  $N$ :

```

# Estimate of the probability with 10 samples:
p_Y_in_set_10 = ecdf_10(3.0) - ecdf_10(1.0)
print(f"N = 10:\tp(1 <= Y <= 3) ~= {p_Y_in_set_10:.2f}")
# Estimate of the probability with 100 samples:
p_Y_in_set_100 = ecdf_100(3.0) - ecdf_100(1.0)
print(f"N = 100:\tp(1 <= Y <= 3) ~= {p_Y_in_set_100:.2f}")
# Estimate of the probability with all 10000 samples:
p_Y_in_set_all = ecdf_all(3.0) - ecdf_all(1.0)
print(f"N = 1000:\tp(1 <= Y <= 3) ~= {p_Y_in_set_all:.2f}")




```

```

N = 10:      p(1 <= Y <= 3) ~= 0.20
N = 100:    p(1 <= Y <= 3) ~= 0.31
N = 1000:   p(1 <= Y <= 3) ~= 0.33

```

## Questions

- Why is the empirical CDF for small  $N$  discontinuous? 
  - How do you know how many samples you need? For now, just think about it on your own. We will give the answer in lecture 10. 
  - Use the best empirical CDF we have constructed so far to find the probability of that  $Y$  is in  $[0.5, 2]$  or  $[3, 4]$ , i.e., find  $p(0.5 \leq Y \leq 2 \text{ or } 3 \leq Y \leq 4)$ . 
- Because the number of samples ( $N$ ) is small. As a result, the possible values for  $F_{\text{bar}N}(y)$  will be a small set of fractions. It is not a coincidence that for  $N = 10$  there are 10 discontinuities in the  $F_{\text{bar}N}(y)$  plot
- So  $F_{\text{bar}N}(y)$  is continuous and changes very little when  $N$  is increased further

---

By Ilias Bilionis (ibilion[at]purdue.edu)

© Copyright 2021.