Empirical Distributions

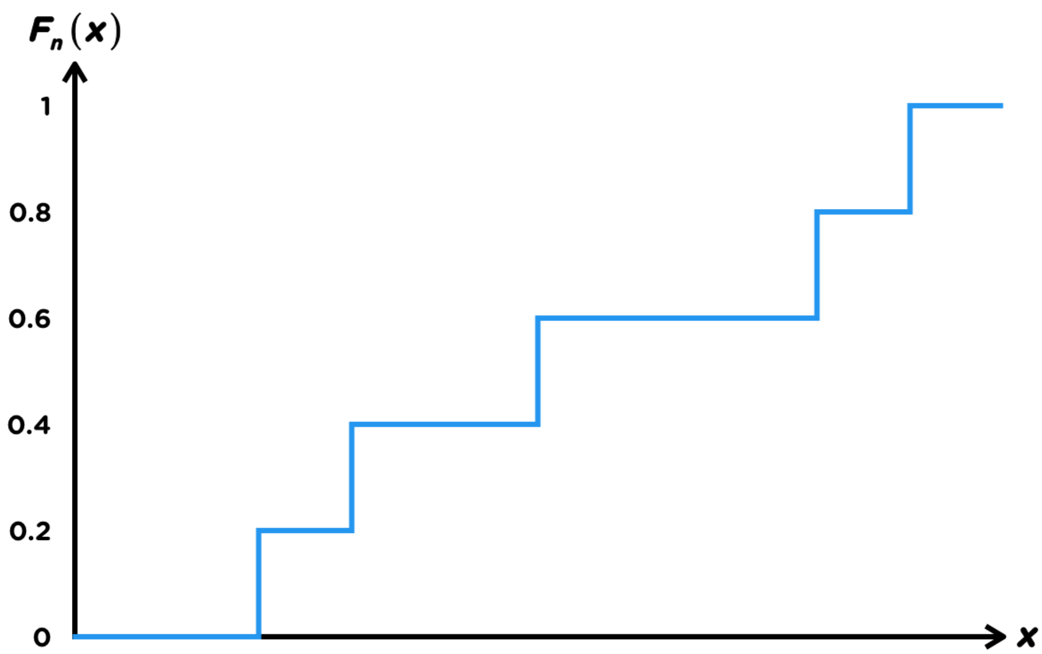
15M

The *empirical distribution* is a discrete distribution based on a sample of size n that assigns probability $\frac{1}{n}$ to each data point.

Let $x_1, x_2, \ldots, x_n$ be a sample of size n. The *empirical distribution function* is the CDF of the empirical distribution. It is calculated as the proportion of observations no more than x out of n total observations.

$$F_n(x) = \frac{\text{Number of observations} \le x}{n} \tag{S2.1.6.1}$$

Like other discrete distributions, the empirical distribution function is a step function. Here is an example:



We can calculate the *empirical 100p^{th} percentile* the same way we calculate the 100p^{th} percentile for discrete distributions. Because all n observations are equally likely, the empirical 100p^{th} percentile reduces to the $\lceil np \rceil^{\text{th}}$ order statistic of the sample, where $\lceil \cdot \rceil$ is the ceiling or round-up function.

$$\pi_p = x_{(\lceil np \rceil)}$$

Coach's Remarks

Recall that the k^{th} order statistic, i.e. $x_{(k)}$, is the k^{th} smallest observation. For example, given a sample {5, 0, 3, 2, 5}, the order statistics are

- $x_{(1)} = 0$
- $x_{(2)} = 2$
- $x_{(3)} = 3$
- $x_{(4)} = 5$
- $x_{(5)} = 5$

In general, the empirical expected value can be calculated using (S2.1.4.1).

$$\mathrm{E}[g(X)] = \frac{\sum_{i=1}^{n} g(x_i)}{n}$$

The empirical 1st raw moment is called the ***sample mean*** and is denoted as $\bar{x}$.

$$\mathrm{E}[X] = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (\text{S2.1.6.2})$$

Then, the empirical 2nd central moment is called the ***biased sample variance***.

$$\begin{aligned}\mathrm{Var}[X] &= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n} \\ &= \frac{\sum_{i=1}^{n} x_i^2}{n}-\bar{x}^2\end{aligned} \qquad \text{(S2.1.6.3)}$$

The equivalence of the two forms is provided in the appendix at the end of this section.

In contrast, the ***unbiased sample variance*** has a divisor of $(n-1)$ and is denoted as s^2.

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1} \\ &= \frac{n}{n-1}\cdot\mathrm{Var}[X]\end{aligned} \qquad \text{(S2.1.6.4)}$$

Coach's Remarks

Whether to use the biased or the unbiased sample variance is a big source of confusion for many students. Here is a rule of thumb:

- Use the **biased** sample variance when calculating the variance of the empirical distribution.
- Use the **unbiased** sample variance when estimating the population variance, particularly when the estimation method is left unspecified.

Throughout this exam, we mostly encounter the second case.

Coach's Remarks

We typically reserve σ^2 to denote the population variance. If N represents the number of observations in a population, then

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

where μ is the population mean.

Note that the empirical distribution treats the sample data as though it is the entire population. Thus, the variance formula of the empirical distribution, i.e.

$$\text{Var}[X] = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

closely resembles the formula for σ^2. Despite the similarity, the biased sample variance is conceptually distinct from σ^2.

Example S2.1.6.1

You are given a sample of size 6:

$$1 \qquad 4 \qquad 5 \qquad 8 \qquad 8 \qquad 10$$

Calculate

1. the empirical distribution function evaluated at 6.
2. the empirical 75$^{\text{th}}$ percentile.
3. the skewness of the empirical distribution.
4. the unbiased sample variance.

Solution to (1)

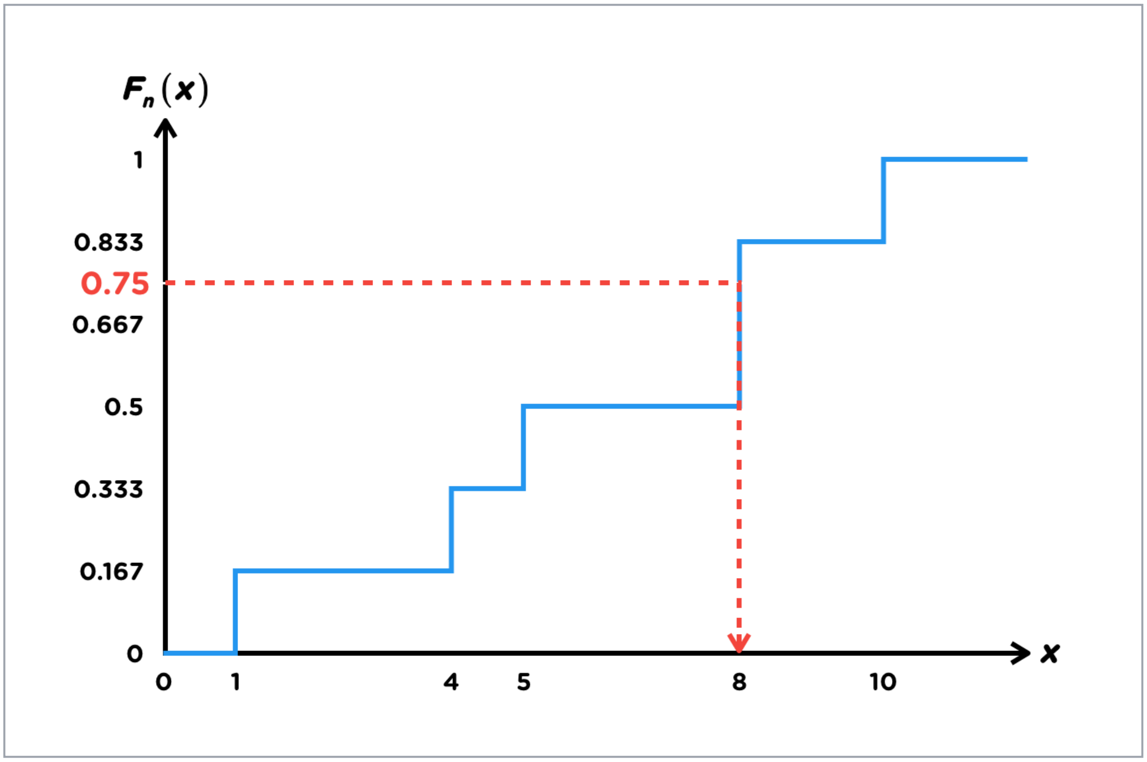
Out of the 6 observations, 3 are at or below 6. Thus,

$$\begin{aligned} F_6(6) &= \frac{\text{Number of observations} \le 6}{6} \\ &= \mathbf{\frac{3}{6}} \end{aligned}$$

Solution to (2)

The empirical 75th percentile is the $\lceil 6(0.75) \rceil = \lceil 4.5 \rceil = 5^{\text{th}}$ order statistic, which is $\mathbf{8}$.

Here is a visual representation:



Solution to (3)

The empirical distribution's mean is

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{1+4+5+8+8+10}{6} \\ &= 6\end{aligned}$$

The empirical distribution's variance is

$$\begin{aligned}\mathrm{Var}[X] &= \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2 \\ &= \frac{1^2+4^2+5^2+8^2+8^2+10^2}{6} - 6^2 \\ &= 9\end{aligned}$$

Using $(S2.1.4.1)$, the 3rd central moment is

$$\begin{aligned}\mu_3 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n} \\ &= \frac{(1-6)^3+(4-6)^3+(5-6)^3+(8-6)^3+(8-6)^3+(10-6)}{6} \\ &= -9\end{aligned}$$

Thus, the skewness of the empirical distribution is

$$\begin{aligned}\frac{\mu_3}{\left(\sqrt{\mathrm{Var}[X]}\right)^3} &= \frac{-9}{\left(\sqrt{9}\right)^3} \\ &= -\frac{\mathbf{1}}{\mathbf{3}}\end{aligned}$$

Solution to (4)

The unbiased sample variance is

$$
\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \\
&= \frac{(1-6)^2+(4-6)^2+(5-6)^2+(8-6)^2+(8-6)^2+(10-6)}{6-1} \\
&= \mathbf{10.8}
\end{aligned}
$$

Alternatively, we can also calculate the unbiased sample variance by scaling the biased sample variance.

$$
\begin{aligned}
s^2 &= \frac{n}{n-1}\cdot \mathrm{Var}[X] \\
&= \frac{6}{6-1}\cdot 9 \\
&= \mathbf{10.8}
\end{aligned}
$$