CDFs, Survival Functions, and Hazard Functions

10M

Cumulative Distribution Function (CDF)

The ***cumulative distribution function (CDF)***, often called the ***distribution function***, of a random variable X is the probability that X does not exceed a given value. This definition holds true for both discrete and continuous random variables. The CDF is commonly denoted by one of the following:

- $\Pr(X \leq x)$
- $F_X(x)$
- $F(x)$

Coach's Remarks

Notice the inequality used in the CDF definition is "less than or equal to" or "not exceeding", i.e. $\leq$, rather than "strictly less than", $<$.

This distinction is important for **discrete** distributions because the probability of a particular value may not be zero. In other words, for a **discrete** random variable X,

$$\begin{aligned} F(x) &= \Pr(X \leq x) \\ &= \Pr(X < x) + p(x) \\ &\neq \Pr(X < x) \end{aligned}$$

However, it is not important to make this distinction for **continuous** random variables. This is because the probability of a particular value is always zero. Therefore, for a **continuous** random variable X,

$$\begin{aligned} F(x) &= \Pr(X \le x) \\ &= \Pr(X < x) + 0 \\ &= \Pr(X < x) \end{aligned}$$

The CDF is computed as a sum for discrete random variables and as an integral for continuous random variables.

$$\begin{aligned} F(x) &= \Pr(X \le x) \\ &= \sum_{t \le x} p(t) && \text{(discrete)} \\ &= \int_{-\infty}^{x} f(t)\,\mathrm{d}t && \text{(continuous)} \end{aligned} \qquad \text{(S2.1.2.1)}$$

The definition for the **continuous** CDF implies the first derivative of the CDF returns the PDF.

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x) \qquad \text{(S2.1.2.2)}$$

All CDFs, discrete or continuous, satisfy the following properties:

- $F(-\infty) = 0$
- $F(\infty) = 1$

Survival Function

The *survival function* is the complement of the CDF. Thus, evaluating a survival function at x yields the probability that the random variable exceeds x. The survival function is commonly denoted by one of the following:

- $\Pr(X > x)$
- $S_X(x)$
- $S(x)$

Coach's Remarks

Since the survival function has a complementary relationship with the CDF, we need to pay attention to its inequality sign.

$$\begin{aligned} S(x) &= \Pr(X > x) \\ &\neq \Pr(X \geq x) \end{aligned}$$

While both inequality signs produce the same continuous event, that is not the case for discrete distributions. Using the correct inequality sign is important for discrete random variables.

The survival function formulas for discrete and continuous random variables are:

$$\begin{aligned} S(x) &= 1 - F(x) \\ &= \Pr(X > x) \\ &= \sum_{t\,>\,x} p(t) && \text{(discrete)} \\ &= \int_x^\infty f(t)\,\mathrm{d}t && \text{(continuous)} \qquad \text{(S2.1.2.3)} \end{aligned}$$

If X is **continuous**, we can use (S2.1.2.2) and the relationship between $F(x)$ and $S(x)$ to derive a relationship between the PDF and survival function.

$$\begin{aligned} f(x) &= \frac{\mathrm{d}}{\mathrm{d}x}[1 - S(x)] \\ &= -\frac{\mathrm{d}}{\mathrm{d}x}S(x) \qquad \text{(S2.1.2.4)} \end{aligned}$$

In addition, all survival functions satisfy the following properties that mirror the CDF properties:

- $S(-\infty) = 1$
- $S(\infty) = 0$

Hazard Function

The ***hazard function***, commonly known as the ***hazard rate*** or the ***force of mortality***, is the ratio of the PDF of a random variable to its survival function. The hazard function of a continuous random variable X is usually denoted by:

- $h_X(x)$
- $h(x)$

By definition,

$$\begin{aligned} h(x) &= \frac{f(x)}{S(x)} \qquad\qquad (\text{S2.1.2.5}) \\ &= \frac{-\frac{\mathrm{d}}{\mathrm{d}x} S(x)}{S(x)} \\ &= -\frac{\mathrm{d}}{\mathrm{d}x}[\ln S(x)] \end{aligned}$$

It can be interpreted as the PDF evaluated at x, adjusted for how likely the random variable is greater than x. Thus, it measures the likelihood of the random variable at x by inflating the PDF as the random variable becomes less likely to exceed x.

Just like the PDF, $h(x)$ is not a probability. Therefore, a hazard function can exceed 1.

$$h(x) \geq 0$$

Integrating the hazard function produces the ***cumulative hazard function***, which is often denoted $H_X(x)$ or $H(x)$. Substituting (S2.1.2.5) into the cumulative hazard function definition will produce a handy relationship between $H(x)$ and $S(x)$:

$$\begin{aligned} H(x) &= \int_{-\infty}^{x} h(t)\,\mathrm{d}t \\ &= \int_{-\infty}^{x} -\frac{\mathrm{d}}{\mathrm{d}t}[\ln S(t)]\,\mathrm{d}t \\ &= -\ln S(x) \end{aligned}$$

Therefore,

$$S(x) = e^{-H(x)} \tag{S2.1.2.6}$$