Course:	Mathematical Statistics
Term:	Fall 2018
Instructor:	Gordan Žitković

Lecture 3

Cumulative distribution functions and derived quantities

When we talk about the distribution of a discrete random variable, we write down its pmf (or a distribution table), and when the variable is continuous, we give its pdf. There are other ways of expressing the same information; depending on the context, these other ways can be much more useful or effective.

3.1 Cumulative distribution functions (cdf)

Definition 3.1.1. For a random variable Y, discrete or continuous, we define its **cumulative distribution function (cdf)** $F_Y : \mathbb{R} \to [0,1]$ by

$$F_Y(y) = \mathbb{P}[Y \leq y],\ y \in \mathbb{R}.$$

The first, obvious, advantage of the cdf is that it can be used for both discrete and continuous random variables. Since it is defined as a probability of an event, $F_Y(y)$ can be computed (at least in principle) from the distribution table in the discrete case

$$F_Y(y) = \sum_{u \in \mathcal{S}_Y, u \leq y} p_Y(u),$$

or from the pdf (in the continuous case):

$$F_Y(y) = \int_{-\infty}^{y} f_Y(u)\, du. \tag{3.1.1}$$

As we shall see in the examples, going the other way in the discrete case is possible, but the formula is a bit clumsy. The continuous case is nicer because one could use the fundamental theorem of calculus to conclude that

$$f_Y(y) = \tfrac{d}{dy} F_Y(y) \text{ for } y \in \mathbb{R},$$

at least for those y where f_Y is a continuous function.

We know that the pdf f_Y of any random variable Y must be nonnegative and integrate to 1. In a similar way, any cdf will have the following properties:

1. $0 \leq F_Y(u) \leq 1$,
2. F_Y is nondecreasing, and
3. $\lim_{u \to \infty} F_Y(u) = 1$ and $\lim_{u \to -\infty} F_Y(u) = 0$.

Example 3.1.2.

1. **Bernoulli.** Let Y be a Bernoulli random variable $B(p)$. To find an expression for F_Y, we first note that

$$F_Y(y) = 0 \text{ for } y < 0.$$

This follows directly from the defintion - Y takes values 0 or 1, so $\mathbb{P}[Y \leq y] = 0$, as soon as $y < 0$. Similarly,

$$F_Y(y) = 1 \text{ for } y \geq 1.$$

What happens in the middle? For any $y \in [0, 1)$, the only way for $Y \leq y$ to be true is if $Y = 0$. Therefore,

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[Y = 0] = q \text{ for } y \in [0, 1).$$

A picture makes it even easier to grasp:

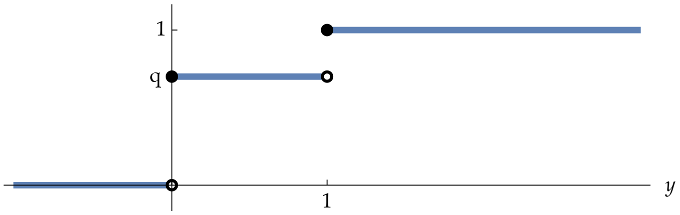

Figure 1. The cumulative distribution function (CDF) for the Bernoulli $B(p)$ distribution.

2. **Discrete with finite support.** Let Y be a discrete random variable with a *finite* support $\mathcal{S}_Y = \{y_1, \dots, y_n\}$ and let its distribution table be given by

y_1	y_2	$\dots$	y_n
p_1	p_2	$\dots$	p_n

Following the same reasoning as in the Bernoulli case, we get the following expression for the cdf

$$F_Y(y) = \begin{cases} 0, & y < y_1, \\ p_1, & y_1 \leq y < y_2, \\ p_1 + p_2, & y_2 \leq y < y_3, \\ \dots & \\ p_1 + p_2 + \dots + p_{n-1} & y_{n-1} \leq y < y_n, \\ 1, & y \geq y_n. \end{cases}$$

Again, a picture is easier to parse:

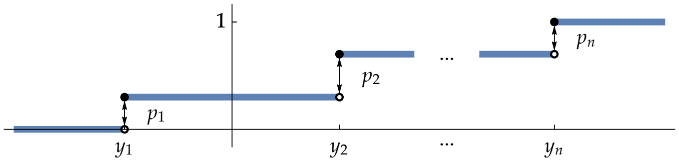

Figure 2. The cumulative distribution of a discrete distribution with support $\{y_1, \dots, y_n\}$ and the associated probabilities $\{p_1, \dots, p_n\}$.

3. **Uniform.** The cdf of the uniform distribution $U(l, r)$ will no longer have "jumps". In fact, that is the reason behind calling continuous distributions continuous. Here, we use the expression (3.1.1) and integrate the pdf f_Y of the uniform distribution from $-\infty$ to y. As above, $F_Y(y) = 0$ for $y < l$ because $f_Y(y) = 0$ for $y < l$ and integration of 0 yields 0. To see what is going on between l and r, we pick $y \in [l, r]$ and note that

$$\int_{-\infty}^{y} f_Y(u)\, du = \int_{l}^{y} f_Y(u)\, du = \int_{l}^{y} \tfrac{1}{r-l} \mathbf{1}_{[l,r]}(y)\, du = \tfrac{1}{r-l} \int_{l}^{y} du = \tfrac{y-l}{r-l}.$$

Finally, for $y > r$, we have $F_Y(y) = 1$. Alternatively, we could have used the definition of F_Y to conclude directly that

$$F_Y(y) = \mathbb{P}[Y \leq y] = \begin{cases} 0, & y < l, \\ \frac{y-l}{r-l}, & y \in [l, r], \\ 1, & y > l. \end{cases}$$

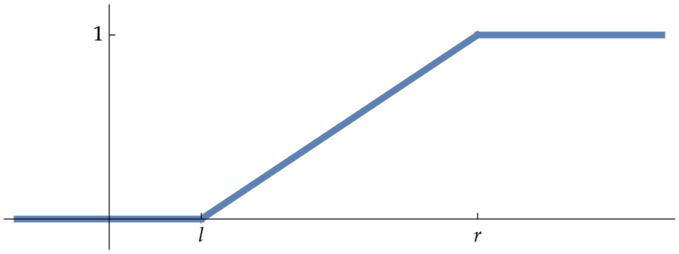

Figure 3. The cumulative distribution of a uniform $U(l, r)$ distribution.

4. **Normal Distribution.** The CDF of the normal distribution $N(\mu, \sigma)$

$$F_Y(y) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u-\mu)^2}{2\sigma^2}}\, du$$

does not have an explicit expression in terms of elementary functions (not even for $\mu = 0$ and $\sigma = 1$). That is why you had to use tables (or software) to compute various probabilities associate to the normal in your probability class. Using mathematical software, one can evaluate this integral numerically, and the resulting picture is given below:

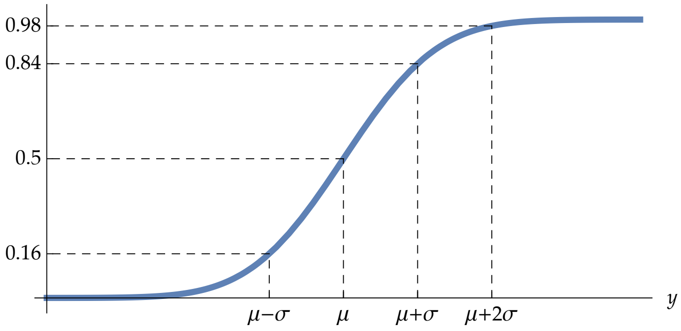

Figure 4. The cumulative distribution of a normal $N(\mu, \sigma)$ distribution.

5. **Exponential distribution.** The integration in the computation of the cdf F_Y of an exponentially-distributed random variable $Y \sim E(\tau)$ can be performed quite easily and completely explicitly. First

of all, for $y < 0$, we clearly have $F_Y(y) = 0$. For $y > 0$, we compute

$$F_Y(y) = \int_{-\infty}^{y} \tfrac{1}{\tau} e^{-u/\tau} \mathbf{1}_{[0,\infty)}(u)\, du = \int_0^y \tfrac{1}{\tau} e^{-u/\tau}\, du = 1 - e^{-y/\tau},\ y > 0.$$

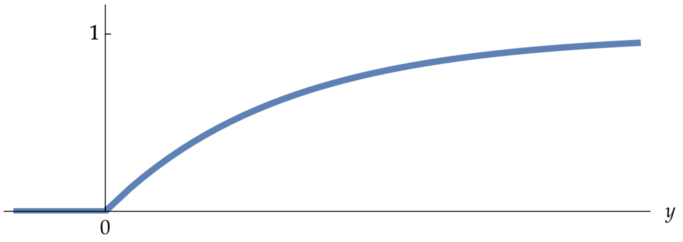

Figure 5. CDF of the exponential distribution $E(\tau)$.

3.2 Quantiles

The notion of a quantile is familiar to almost everyone, even if you have not learned it formally in a class. You don know what "top 1%" means, right? The formal definition is easy once we have the notion of a cdf at our disposal:

Definition 3.2.1. For $\alpha \in (0, 1)$, we define the α**-quantile** of the distribution of the random Y as the number $q_Y(\alpha) \in \mathbb{R}$ with the property that

$$F_Y(q_Y(\alpha)) = \alpha, \quad \textit{i.e.,} \quad \mathbb{P}[Y \leq q_Y(\alpha)] = \alpha.$$

Caveat: The way we defined above, the quantile $q_Y(\alpha)$ may not need to exist for all α. This can be remedied by adopting a more careful definition, but, since we will not have to deal with this problem in these notes - and whenever we need quantiles, they will happily exist - we simply ignore it. If you want to think about this a bit more, try to figure out which quantiles of the Bernoulli distribution actually exist, i.e., for which α can we find a number q such that $\mathbb{P}[Y \leq q] = \alpha$, when Y is Bernoulli. Is such a q uniquely determined?

Example 3.2.2. Normal quantiles. In practice, one finds quantiles by inverting the CDF; graphically this amounts to finding α on the vertical axis, and then finding a value q on the horizontal axis such that $F_Y(q) = \alpha$. For example, Figure 4. in Example (3.1.2), part 4., above, reveals that, for $Y \sim N(\mu, \sigma)$, we have (approximately)

$$q_Y(0.16) = \mu - \sigma,\ q_Y(0.5) = \mu,\ q_Y(0.84) = \mu + \sigma \text{ and } q_Y(0.98) = \mu + 2\sigma.$$

This is very much related to the well-known $68 - 95 - 99.7$-rule.

3.3 Survival and hazard functions

Survival and hazard functions are especially important for an area of statistics called the survival analysis, but are also a part of the vocabulary of general statistics.

Definition 3.3.1. Let Y be a random variable with cdf F_Y.

1. The **survival function** $S_Y(y)$ of Y is defined by

$$S_Y(y) = 1 - F_Y(y) \text{ for } y \in \mathbb{R}.$$

2. If Y is continuous, the **hazard function** $h_Y(y)$ is given by

$$h_y(y) = \frac{f_Y(y)}{S_Y(y)} \text{ for } y \text{ with } F_Y(y) < 1.$$

These quantities have natural interpretations when Y is thought of as a lifetime (of a particle, bulb, bacterium, individual, etc.). Fixing, for convenience, the interpretation that Y is the age at death of an individual, we have

1. $S_Y(y)$ is the probability that the individual will survive at least y years.
2. $h_Y(y)\Delta y$ is the (conditional) probability that the individual will die some time in the (small) interval $[y, y + \Delta y]$, given that it has survived until y.

Example 3.3.2. Let Y be an exponential random variable with parameter τ. Then

$$S_Y(y) = e^{-y/\tau} \text{ and } h_Y(y) = \tfrac{1}{\tau} \text{ for } y \geq 0.$$

In words, exponentially-distributed lifetimes have constant hazard functions - "the probability of dying in the next Δy is constant and does not depend on the age y." For comparison, Figure 6 below features some real data about humans where the hazard rate is far from constant.

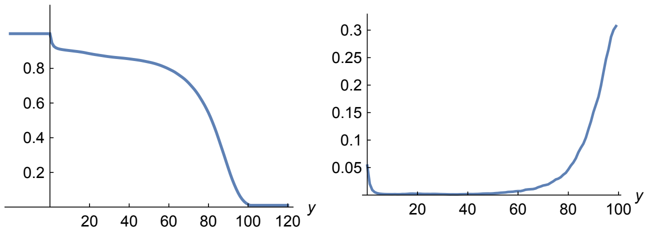

Figure 6. The survival (left) and the hazard (right) functions of the empirical distribution of the ages of death of all female individuals born in the US in 1917.

3.4 Problems

Problem 3.4.1. Two (unbiased, independent) coins are tossed, and the total number of heads is denoted by Y. Write an expression for the CDF of Y and sketch its graph.

Problem 3.4.2. Which of the following pairs of functions *could* be the pdf and the cdf (respectively) of some probability distribution:

(a) $f(x) = x^2$, $F(x) = \frac{1}{3}x^3$

(b) $f(x) = \cos(x)$, $F(x) = \sin(x)$.

(c) $f(x) = 2e^{-2x}\mathbf{1}_{\{x>0\}}$, $F(x) = (1 - e^{-2x})\mathbf{1}_{\{x>0\}}$.

(d) $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, $F(x) = 1 - e^{-x^2}$.

(e) $f(x) = \mathbf{1}_{\{x>0\}}$, $F(x) = x\mathbf{1}_{\{x>0\}}$.

Problem 3.4.3. Let Y be a random variable with CDF F_Y, and let $q_Y : (0,1) \to \mathbb{R}$ be its quantile function (we assume it exists for each $\alpha \in (0,1)$). What is the relationship between the graphs of F_Y and q_Y, i.e., how do you get one from the other?

Problem 3.4.4. Let Y be a continuous random variable with the density f_Y given by

$$f_Y(y) = cy^2(1-y)\mathbf{1}_{[0,1]}(y),$$

for an appropriate constant c.

1. Sketch the graph of f and find the value of the constant c.
2. Compute the cumulative distribution function (cdf) F_Y and the survival function S_Y, of Y.
3. What is the domain of the hazard function? Compute the hazard function h_Y itself.
4. Find the mode of Y
5. Compute the $\frac{5}{16}$-th quantile of Y. (*Note:* Guess and verify.)

Problem 3.4.5. Let Y be a random variable with the pdf

$$f_Y(y) = 2y\mathbf{1}_{\{0 \le y \le 1\}}.$$

Compute the hazard function h_Y of Y.

Problem 3.4.6. Let Y be a uniform random variable on the interval $[0,100]$. The hazard function h_Y of the distribution of Y is given by

(a) $\frac{1}{y}\mathbf{1}_{\{y>0\}}$ for $y \in (-\infty, 100)$

(b) $\frac{1}{100-y}\mathbf{1}_{\{y>0\}}$ for $y \in (-\infty, 100)$

(c) $\mathbf{1}_{\{y<0\}} + \frac{100-y}{100}\mathbf{1}_{\{0 \le y \le 100\}}$ for $y \in (-\infty, 100]$

(d) $(100-y)\mathbf{1}_{\{y \in [0,100)\}}$ for $y \in [0,\infty)$

(e) none of the above

Problem 3.4.7. The expected lifetime of a bulb is h (in hours). Assuming that the bulb lifetimes are exponentially distributed, compute

1. the probability that the bulb is still functional at time h
2. the half-life of the bulb, i.e., a number t^* such that the probability that the bulb is still functional after t^* hours is exactly $1/2$.

Problem 3.4.8. Compute the α-quantile $q_Y(\alpha)$ for $\alpha = 0.75$ where Y is the uniform distribution $U(4,8)$ on $[4,8]$.