

Introduction to statistics

Prof. Dr. Christoph Richard, Leonie Wicht, Anna Vandebosch and Theresa Schmid

January 6, 2024

T1. Empirical Cumulative Distribution Function

Let x_1, x_2, \dots, x_n be a data sequence with empirical cumulative distribution function $F_n(t)$ and relative interval frequencies $h_n(I)$, i.e

$$F_n : \mathbb{R} \rightarrow [0, 1], t \mapsto h_n((-\infty, t]) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}(x_i)$$

Show that for any real numbers $a < b$ we have:

$$\begin{aligned} h_n((a, b]) &= F_n(b) - F_n(a) \\ h_n(\{a\}) &= F_n(a) - F_n(a-) \\ h_n([a, b]) &= F_n(b) - F_n(a) + h_n(\{a\}) \\ h_n([a, b)) &= F_n(b-) - F_n(a-) \\ h_n((a, \infty)) &= 1 - F_n(a) \end{aligned}$$

Here $F(t-) := \lim_{x \uparrow t} F(x)$ denotes the limit from the left. If you manipulate the indicator functions, give proof of any of these rules.

Solution T1

To prove these rules, we first have to notice the following manipulations of the indicator function:

a. $\mathbb{1}_{(a, b]}(x) = \mathbb{1}_{(-\infty, b]}(x) - \mathbb{1}_{(-\infty, a]}(x)$

Proof:

$$\mathbb{1}_{(a,b]}(x) = \begin{cases} 0 & \text{if } x \notin (a,b] \\ 1 & \text{if } x \in (a,b] \end{cases}$$

$$\mathbb{1}_{(-\infty,b]}(x) - \mathbb{1}_{(-\infty,a]}(x) = \begin{cases} 0 & \text{if } x \notin (-\infty,b] \wedge x \notin (-\infty,a] \\ 1 & \text{if } x \in (-\infty,b] \wedge x \notin (-\infty,a] \\ -1 & \text{if } x \notin (-\infty,b] \wedge x \in (-\infty,a] \end{cases} \quad \vee \quad x \in (-\infty,b] \cap (-\infty,a]$$

Since $a < b$, we have

$$\begin{aligned} \mathbb{1}_{(-\infty,b]}(x) - \mathbb{1}_{(-\infty,a]}(x) &= \begin{cases} 0 & \text{if } x \notin (-\infty,b] \vee x \in (-\infty,a] \\ 1 & \text{if } x \in (-\infty,b] \wedge x \notin (-\infty,a] \end{cases} \\ &= \begin{cases} 0 & \text{if } x \notin (a,b] \\ 1 & \text{if } x \in (a,b] \end{cases} \end{aligned}$$

b. $\mathbb{1}_{\{a\}}(x) = \mathbb{1}_{(-\infty,a]}(x) - \lim_{t \uparrow a} \mathbb{1}_{(-\infty,t]}(x)$

Proof:

$$\begin{aligned} \mathbb{1}_{\{a\}}(x) &= \begin{cases} 0 & \text{if } x \notin \{a\} \\ 1 & \text{if } x \in \{a\} \end{cases} \\ \mathbb{1}_{(-\infty,a]}(x) - \lim_{t \uparrow a} \mathbb{1}_{(-\infty,t]}(x) &= \lim_{t \uparrow a} (\mathbb{1}_{(-\infty,a]}(x) - \mathbb{1}_{(-\infty,t]}(x)) \end{aligned}$$

Because of what we already showed in **a**, we have that $\mathbb{1}_{(-\infty,a]}(x) - \mathbb{1}_{(-\infty,t]}(x) = \mathbb{1}_{(t,a]}(x)$, so we can continue as follows:

$$\begin{aligned} \lim_{t \uparrow a} (\mathbb{1}_{(-\infty,a]}(x) - \mathbb{1}_{(-\infty,t]}(x)) &= \lim_{t \uparrow a} \mathbb{1}_{(t,a]}(x) \\ &= \begin{cases} 0 & \text{if } x \notin \{a\} \\ 1 & \text{if } x \in \{a\} \end{cases} \\ &= \mathbb{1}_{\{a\}}(x) \end{aligned}$$

Now that we have shown these two manipulations (**a** and **b**), we can start proving the rules.

1. $h_n((a,b]) = F_n(b) - F_n(a)$

$$\begin{aligned} h_n((a,b]) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(a,b]}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty,b]}(x_i) - \mathbb{1}_{(-\infty,a]}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty,b]}(x_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty,a]}(x_i) \\ &= F_n(b) - F_n(a) \end{aligned}$$

$$2. \quad h_n(\{a\}) = F_n(a) - F_n(a-)$$

$$\begin{aligned} h_n(\{a\}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{a\}}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, a]}(x_i) - \lim_{t \uparrow a} \mathbb{1}_{(-\infty, t]}(x_i) \\ &= F_n(a) - F_n(a-) \end{aligned}$$

$$3. \quad h_n([a, b]) = F_n(b) - F_n(a) + h_n(\{a\})$$

$$\begin{aligned} h_n([a, b]) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[a, b]}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, b]}(x_i) - \mathbb{1}_{(-\infty, a)}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, b]}(x_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, a]}(x_i) + \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{a\}}(x_i) \\ &= F_n(b) - F_n(a) + h_n(\{a\}) \end{aligned}$$

$$4. \quad h_n([a, b)) = F_n(b-) - F_n(a-)$$

$$\begin{aligned} h_n([a, b)) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[a, b)}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, b)}(x_i) - \mathbb{1}_{(-\infty, a)}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \lim_{t \uparrow b} \mathbb{1}_{(-\infty, t]}(x_i) - \lim_{t \uparrow a} \mathbb{1}_{(-\infty, t]}(x_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, b)}(x_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, a)}(x_i) \\ &= F_n(b-) - F_n(a-) \end{aligned}$$

$$5. \quad h_n((a, \infty)) = 1 - F_n(a)$$

$$\begin{aligned} h_n((a, \infty)) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(a, \infty)}(x_i) \\ &= \lim_{t \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}(x_i) - \mathbb{1}_{(-\infty, a]}(x_i) \\ &= \lim_{t \rightarrow \infty} F_n(t) - F_n(a) \\ &= 1 - F_n(a) \end{aligned}$$

T2. Convergence of cumulative distribution functions

A function $F : \mathbb{R} \rightarrow [0, 1]$ is called a cumulative distribution function, if F is monotonically increasing and right continuous, and if we have $\lim_{t \rightarrow -\infty} F(t) = 0$ and $\lim_{t \rightarrow \infty} F(t) = 1$. Let $(F_n)_{n \in \mathbb{N}}$ be a sequence of cumulative distribution functions, which converges uniformly to a function $F : \mathbb{R} \rightarrow \mathbb{R}$

1. Give the definition of uniform convergence. Recall from your calculus lecture notes the following result: Given a sequence of continuous functions that converges uniformly, then the limit function is continuous. Recall the proof of that statement.
2. Show that F is a cumulative distribution function.
3. Why is the latter result important for our approach to statistics?

Solution T2

1. We say that a sequence of functions f_n , defined on a common domain A , converges uniformly to a function f on A , if for any $\epsilon > 0$, there exists a positive integer N such that for all $n \geq N$ and for all $x \in A$ we have $|f_n(x) - f(x)| < \epsilon$.

Given a sequence of continuous functions that converges uniformly, then the limit function is continuous.

Proof:

Let $\epsilon > 0$, $x \in A$.

Uniform convergence implies that there exists a $N \in \mathbb{N}$ such that $\forall x' \in A$ we have

$$|f_N(x') - f(x')| < \frac{\epsilon}{3}$$

Since f_N is continuous, $\exists \delta > 0$ such that $\forall x' \in A$ with $|x' - a| < \delta$, $a \in A$ we have

$$|f_N(x') - f_N(a)| < \frac{\epsilon}{3}$$

Let $a \in A$ with $|x - a| < \delta$. Then, by the triangle inequality, we have

$$\begin{aligned} |f(x) - f(a)| &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(a)| + |f_N(a) - f(a)| \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &\leq \epsilon \end{aligned}$$

2. We need to show that F is
 - (a) monotonically increasing,
 - (b) right continuous, and

(c) $\lim_{t \rightarrow -\infty} F(t) = 0$ and $\lim_{t \rightarrow \infty} F(t) = 1$.

Proof:

(a) *Monotonicity:* Let $a < b \implies \forall n \in \mathbb{N} \quad F_n(a) \leq F_n(b)$.

$$F(a) = \lim_{n \rightarrow \infty} F_n(a) \leq \lim_{n \rightarrow \infty} F_n(b) = F(b)$$

(b) *Right Continuity:*

To establish right continuity of F , we must show that for every $t \in \mathbb{R}$,

$$\lim_{s \downarrow t} F(s) = F(t).$$

Since F_n is right continuous, we have

$$\lim_{s \downarrow t} F_n(s) = F_n(t).$$

By taking the uniform limit as $n \rightarrow \infty$, we get

$$\lim_{s \downarrow t} F(s) = \lim_{s \downarrow t} \lim_{n \rightarrow \infty} F_n(s) = \lim_{n \rightarrow \infty} \lim_{s \downarrow t} F_n(s) = \lim_{n \rightarrow \infty} F_n(t) = F(t).$$

(c) *Boundary Conditions:* The uniform convergence of F_n to F also guarantees that the boundary conditions at $-\infty$ and $+\infty$ will be preserved. Specifically, we consider the limits:

$$\begin{aligned} \lim_{t \rightarrow -\infty} F(t) &= \lim_{t \rightarrow -\infty} \lim_{n \rightarrow \infty} F_n(t) = \lim_{n \rightarrow \infty} \lim_{t \rightarrow -\infty} F_n(t) \\ &= \lim_{n \rightarrow \infty} 0 = 0, \end{aligned}$$

and,

$$\begin{aligned} \lim_{t \rightarrow \infty} F(t) &= \lim_{t \rightarrow \infty} \lim_{n \rightarrow \infty} F_n(t) = \lim_{n \rightarrow \infty} \lim_{t \rightarrow \infty} F_n(t) \\ &= \lim_{n \rightarrow \infty} 1 = 1. \end{aligned}$$

These steps are justified because uniform convergence allows us to switch the order of limits for functions, and F_n satisfy the CDF boundary conditions by definition.

By confirming the monotonicity, right continuity, and boundary conditions, we have shown that F is a cumulative distribution function.

3. Importance of Result for Statistics:

ECDF mostly converges uniformly and with the above results we know that the limit is also a CDF.

So by doing experiments, we can learn something about the random mechanisms behind the experiments. This is because the ECDF converges to the true CDF for a sufficiently large number of samples (n).

T3. Symmetric cumulative distribution function

We call a continuous cumulative distribution function F symmetric in c , if $F(c+t) = 1 - F(c-t)$ for all $t \geq 0$. F is called *symmetric*, if there is a c such that F is symmetric in c .

1. Show that $F(c+t) = 1 - F(c-t)$ for all $t \geq 0$ implies the continuity of the cumulative distribution function F . (Why is it enough to prove left continuity?)
2. Assume that there is a continuous function $f : \mathbb{R} \rightarrow [0, \infty)$ such that

$$F(t) = \int_{-\infty}^t f(x)dx \quad \forall t \in \mathbb{R}$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

i.e. F has density f .

Remark: Continuity of f is not necessary for the definition. A necessary condition is measurability, a concept that is treated in measure theory.

Show that F is symmetric in c if and only if $f(c+t) = f(c-t)$ for all $t \geq 0$.

Remark: This result stays true for piecewise continuous f .

3. For a continuous cumulative distribution function F , a median of F is any real number x such that $F(x) = 1/2$. Why is a median an important parameter of a distribution? For any symmetric F , give an example of a median. May there be more than one median for a given F ?
4. Give a median of the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Is it unique?

Solution T3

1. Show that $F(c+t) = 1 - F(c-t)$ for all $t \geq 0$ implies the continuity of the cumulative distribution function F . (Why is it enough to prove left continuity?).

Proof:

Since we know F is a CDF, we know F has to also be right continuous (because all CDFs are right continuous). Therefore we only need to show that F is also left continuous at all points in its domain to prove that F is continuous.

We have to show that for any $x \in \mathbb{R}$ and any sequence $\{x_n\}$ converging to x from the left ($x_n \uparrow x$ as $n \rightarrow \infty$), we have $\lim_{n \rightarrow \infty} F(x_n) = F(x)$.

(a) **First case:** $x > c$

Let $t > 0$ such that $x = c + t$. Define t_n such that $x_n = c + t_n$. Since $x_n \uparrow x$, then $t_n \uparrow t$ as $n \rightarrow \infty$

$$\begin{aligned}
\lim_{n \rightarrow \infty} F(x_n) &= \lim_{n \rightarrow \infty} F(c + t_n) \\
&\stackrel{(1)}{=} \lim_{n \rightarrow \infty} (1 - F(c - t_n)) \\
&= 1 - \lim_{n \rightarrow \infty} (F(c - t_n)) \\
&\stackrel{(2)}{=} 1 - F(c - t) \\
&\stackrel{(1)}{=} F(c + t) \\
&= F(x)
\end{aligned}$$

(1): F is symmetric in c .

(2): F right continuity: $c - t_n$ is a sequence converging to $c - t$ from the right.

(b) **Second case:** $x < c$

Same as before, but with $x = c - t$.

Let $t > 0$ such that $x = c - t$. Define t_n such that $x_n = c - t_n$. Since $x_n \uparrow x$, then $t_n \downarrow t$ as $n \rightarrow \infty$

$$\begin{aligned}
\lim_{n \rightarrow \infty} F(x_n) &= \lim_{n \rightarrow \infty} F(c - t_n) \\
&\stackrel{(1)}{=} \lim_{n \rightarrow \infty} (1 - F(c + t_n)) \\
&= 1 - \lim_{n \rightarrow \infty} (F(c + t_n)) \\
&\stackrel{(2)}{=} 1 - F(c + t) \\
&\stackrel{(1)}{=} F(c - t) \\
&= F(x)
\end{aligned}$$

(1): F is symmetric in c .

(2): F right continuity: $c + t_n$ is a sequence converging to $c + t$ from the right.

(c) **Third case:** $x = c$ and $F(c) = 1 - F(c)$

s_n is a sequence converging to c from the left. Let t_n be a sequence such that $s_n = c - t_n$. Since $s_n \uparrow c$, then $t_n \downarrow 0$ as $n \rightarrow \infty$

$$\begin{aligned}
\lim_{n \rightarrow \infty} F(s_n) &= \lim_{n \rightarrow \infty} F(c - t_n) \\
&\stackrel{(1)}{=} 1 - \lim_{n \rightarrow \infty} F(c + t_n) \\
&\stackrel{(2)}{=} 1 - F(c) \\
&= F(c)
\end{aligned}$$

- (1): F is symmetric in c .
(2): F right continuity: $c + t_n$ is a sequence converging to c from the right.

2. Assume that there is a continuous function $f : \mathbb{R} \rightarrow [0, \infty)$ such that

$$F(t) = \int_{-\infty}^t f(x)dx \quad \forall t \in \mathbb{R}$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

i.e. F has density f .

Show that F is symmetric in c if and only if $f(c + t) = f(c - t)$ for all $t \geq 0$.

Proof:

F is symmetric in c

$$\begin{aligned} \Leftrightarrow \quad \forall t \geq 0 \quad &: \quad F(c + t) = 1 - F(c - t) \\ \Leftrightarrow \quad \forall t \geq 0 \quad &: \quad \int_{-\infty}^{c+t} f(x)dx = 1 - \int_{-\infty}^{c-t} f(x)dx = \int_{c-t}^{\infty} f(x)dx \\ \Leftrightarrow \quad \forall t \geq 0 \quad &: \quad \int_{-\infty}^{c+t} f(x)dx = \int_{c-t}^{\infty} f(x)dx \\ \Leftrightarrow \quad \forall t \geq 0 \quad &: \quad \int_{-\infty}^t f(c+x)dx = - \int_t^{-\infty} f(c-x)dx \\ \Leftrightarrow \quad \forall t \geq 0 \quad &: \quad \int_{-\infty}^t f(c+x)dx = \int_{-\infty}^t f(c-x)dx \\ \Leftrightarrow \quad \forall t \geq 0 \quad &: \quad f(c-t) = \lim_{h \rightarrow 0} \frac{F(c-t+h) - F(c-t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{1 - F(c+t-h) - (1 - F(c+t))}{h} \\ &= \lim_{h \rightarrow 0} \frac{F(c+t) - F(c+t-h)}{h} \\ &= f(c+t) \end{aligned}$$

(1) F is symmetric in c .

3. For a continuous cumulative distribution function F , a median of F is any real number x such that $F(x) = 1/2$.

(a) Why is a median an important parameter of a distribution?

The median is an important parameter of a distribution because it is the value that divides the distribution in two equal parts. Half of the data points are less

than or equal to the median. The other half are greater than or equal to the median. So you know more or less where the "middle" of the data points is. Another reason why the median is an important parameter of a distribution is because it is robust against outliers (not like the mean).

- (b) For any symmetric F , give an example of a median.

If there is a c such that F is symmetric in c , then c is a median of F .

$$\begin{aligned} F(c) &= 1 - F(c) \\ 2F(c) &= 1 \\ F(c) &= \frac{1}{2} \end{aligned}$$

- (c) May there be more than one median for a given F ?

Yes, if F is piecewise constant at $\frac{1}{2}$, then every point in the interval where F is constant at $\frac{1}{2}$ is a median of F . For example, the CDF of the Bernoulli distribution: Every $x \in [0, 1)$ is a median of F .

- (d) Give a median of the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Is it unique?

The median of the normal distribution is the same as the mean μ , as the PDF and CDF of the normal distribution are symmetric around μ .

The median is unique, as the CDF is strictly monotonically increasing. (There exists only one c such that the symmetry condition is fulfilled.)

T4: Continuum approximation of the geometric distribution

We consider the following probability mass function:

$$f_p : \mathbb{N} \rightarrow \mathbb{R}_+, \quad k \mapsto \mathbb{P}(\{k\}) = p(1-p)^{k-1}$$

where $p \in (0, 1)$. The corresponding distribution is the geometric distribution with parameter p . Moreover let g_λ be the following probability density function:

$$g_\lambda : \mathbb{R} \rightarrow \mathbb{R}_+, \quad x \mapsto \lambda e^{-\lambda x} 1_{[0, \infty)}(x)$$

where $\lambda > 0$. The corresponding distribution is the exponential distribution with parameter λ .

(i) Determine the cumulative distribution function F_p of the geometric distribution and the cumulative distribution function G_λ of the exponential distribution.

(ii) Let $(p_n)_{n \in \mathbb{N}}$ be a sequence in $(0, 1)$ such that $\lim_{n \rightarrow \infty} n \cdot p_n = \lambda$. Prove that

$$\lim_{n \rightarrow \infty} F_{p_n}(nt) = G_\lambda(t), \quad \forall t \in \mathbb{R}$$

(iii) Explain in your own words what the above result means when modelling rare events on an interval by coin tossing. Give an explicit example from economics, physics or biology.

Hint: If $(x_m)_{m \in \mathbb{N}}, (y_m)_{m \in \mathbb{N}}$ are sequences in \mathbb{R} that converge to $x \in \mathbb{R}$ and 1, respectively, then the following holds:

$$\lim_{m \rightarrow \infty} \left(1 + \frac{x_m}{m}\right)^{m \cdot y_m} = e^x$$

Solution T4

1. Determine the cumulative distribution function F_p of the geometric distribution and the cumulative distribution function G_λ of the exponential distribution.

$$\begin{aligned} F_p(x) &= \sum_{k=1}^{\lfloor x \rfloor} f_p(k) \mathbb{1}_{[1, \infty)}(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1 - (1-p)^{\lfloor x \rfloor} & \text{for } x \geq 1 \end{cases} \\ &= \begin{cases} 0 & \text{for } x < 0 \\ 1 - (1-p)^{\lfloor x \rfloor} & \text{for } x \geq 0 \end{cases} \end{aligned}$$

$$\sum_{k=1}^n p(1-p)^{k-1} = \sum_{k=0}^{n-1} p(1-p)^k \stackrel{(1)}{=} p \frac{1 - (1-p)^n}{1 - (1-p)} = 1 - (1-p)^n$$

(1) Geometric series

Let's go through the solution step by step to understand how to determine the cumulative distribution function (CDF) F_p for the geometric distribution.

Step 1: Understanding the Probability Mass Function (PMF)

The geometric distribution's PMF is given by:

$$f_p(k) = p(1 - p)^{k-1} \quad \text{for } k \in \mathbb{N},$$

where p is the probability of success on each trial.

Step 2: Defining the Cumulative Distribution Function (CDF)

The CDF $F_p(x)$ of the geometric distribution is the sum of the PMF values up to a certain point x . By definition, we have:

$$F_p(x) = \sum_{k=1}^{\lfloor x \rfloor} f_p(k) \quad \text{for } x \geq 1,$$

where $\lfloor x \rfloor$ denotes the floor function, which gives the greatest integer less than or equal to x .

Step 3: Applying the Indicator Function

The indicator function $\mathbb{1}_{[1, \infty)}(x)$ is equal to 1 for $x \geq 1$ and equal to 0 for $x < 1$. This function is used to "activate" the PMF sum starting from $x = 1$, as there are no probabilities to sum for $x < 1$ in the geometric distribution.

The CDF then reads:

$$F_p(x) = \sum_{k=1}^{\lfloor x \rfloor} f_p(k) \mathbb{1}_{[1, \infty)}(x).$$

Step 4: Handling the CDF for $x < 1$

Since we cannot have negative probabilities and the smallest natural number is 1, the CDF $F_p(x)$ is defined as 0 for $x < 1$, as no trials would have occurred (geometrically speaking) before $x = 1$.

Step 5: Defining Cases for the CDF

The CDF $F_p(x)$ can then be expressed in a piecewise manner as:

$$F_p(x) = \begin{cases} 0 & \text{for } x < 1, \\ 1 - (1 - p)^{\lfloor x \rfloor} & \text{for } x \geq 1, \end{cases}$$

reflecting the distinct behavior depending on whether x is before or after the start of the distribution.

Step 6: Expanding the CDF for $x < 0$

For $x < 0$, the CDF remains 0 since we cannot have trials at negative counts. The piecewise definition could equivalently start from $x = 0$:

$$F_p(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 - (1 - p)^{\lfloor x \rfloor} & \text{for } x \geq 0. \end{cases}$$

Step 7: Understanding the Geometric Series

The sum of the geometric series is given by:

$$\sum_{k=1}^n p(1-p)^{k-1} = \sum_{k=0}^{n-1} p(1-p)^k = p \frac{1 - (1-p)^n}{1 - (1-p)} = 1 - (1-p)^n,$$

where $n = \lfloor x \rfloor$ for our CDF.

This equation (1) is an algebraic manipulation of the geometric series sum that takes advantage of the closed-form formula for the sum of a finite geometric series

Indeed, there is a discrepancy between those steps that needs clarification. Let's analyze the transition from step 5 to step 6 more critically.

The cumulative distribution function $F_p(x)$ for the geometric distribution is supposed to return the sum of the probabilities from the first trial up to the largest integer less than or equal to x (denoted by $\lfloor x \rfloor$). Specifically, the geometric distribution is only defined for positive integers k , meaning the distribution actually starts at $k = 1$.

For the geometric distribution, the CDF $F_p(x)$ must be 0 for all $x < 1$, because no event (no "success" in the sequence of Bernoulli trials) can occur before the first trial. Hence, the original CDF in step 5 is correct when it states:

$$F_p(x) = \begin{cases} 0 & \text{for } x < 1, \\ 1 - (1 - p)^{\lfloor x \rfloor} & \text{for } x \geq 1. \end{cases}$$

The discrepancy arises in the transition to step 6, where the CDF is written to have a switch point at $x = 0$ rather than at $x = 1$. This would imply that there's some probability mass for x in the interval $[0, 1)$, which is not correct for the geometric distribution—it must be 0 for any $x < 1$.

So, the step 6 transition is unwarranted; the CDF should remain as:

$$F_p(x) = \begin{cases} 0 & \text{for } x < 1, \\ 1 - (1 - p)^{\lfloor x \rfloor} & \text{for } x \geq 1, \end{cases}$$

This accurately states that the probability (cumulative) is 0 before the first trial (for $x < 1$), and it begins to increase from the first trial onwards ($x \geq 1$). There is no need to redefine the regions for $x < 0$ versus $x < 1$ since the former is simply a subset of the latter, and the CDF value in both regions is 0.

but please notice that $\lfloor x \rfloor = 0$ for any $x \in [0, 1)$. If $\lfloor x \rfloor = 0$, then it means that $1 - (1 - p)^{\lfloor x \rfloor} = 0$ even for $x \in [0, 1)$. I think that's why we can write

$$F_p(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 - (1 - p)^{\lfloor x \rfloor} & \text{for } x \geq 0. \end{cases}$$

instead of

$$F_p(x) = \begin{cases} 0 & \text{for } x < 1, \\ 1 - (1 - p)^{\lfloor x \rfloor} & \text{for } x \geq 1. \end{cases}$$

What do you think?. Is that correct?

You are correct in pointing out that for x in the interval $[0, 1)$, the floor function $\lfloor x \rfloor$ yields 0. Therefore, $1 - (1 - p)^{\lfloor x \rfloor}$ indeed equals 0 for any x in this interval because $(1 - p)^0$ equals 1, resulting in $1 - 1 = 0$.

Given this accurate observation, it is indeed correct that we can express the cumulative distribution function $F_p(x)$ for the geometric distribution as:

$$F_p(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 - (1 - p)^{\lfloor x \rfloor} & \text{for } x \geq 0. \end{cases}$$

This piecewise function is equivalent to the prior expression because $F_p(x)$ is 0 for any x in the interval $[-\infty, 1)$. The switch point at $x = 0$ is acceptable, as it aligns with the fact that there cannot be any accumulated probability mass below $x = 1$.

Therefore, it is not only acceptable but also correct. It may initially seem somewhat non-intuitive, but the floor function ensures that probabilities are only counted starting from the first success, which, given the context of the geometric distribution, cannot happen before the first trial.

In summary, the statement that redefines the switch point at $x = 0$ is correct and appropriately represents the geometric distribution.

There's something I still do not understand.

I know that the geometric distribution's PMF is given by:

$$f_p(k) = p(1 - p)^{k-1} \quad \text{for } k \in \mathbb{N},$$

where p is the probability of success on each trial.

I also know that the CDF $F_p(x)$ of the geometric distribution is the sum of the PMF values up to a certain point x . By definition, we have:

$$F_p(x) = \sum_{k=1}^{\lfloor x \rfloor} f_p(k) \quad \text{for } x \geq 1,$$

where $\lfloor x \rfloor$ denotes the floor function, which gives the greatest integer less than or equal to x .

I also understand that the indicator function $\mathbb{1}_{[1, \infty)}(x)$ is equal to 1 for $x \geq 1$ and equal to 0 for $x < 1$. This function is used to "activate" the PMF sum starting from $x = 1$, as there are no probabilities to sum for $x < 1$ in the geometric distribution.

The CDF then reads:

$$F_p(x) = \sum_{k=1}^{\lfloor x \rfloor} f_p(k) \mathbb{1}_{[1, \infty)}(x).$$

What I really do not understand is why the CDF $F_p(x)$ can be expressed in a piecewise manner as:

$$F_p(x) = \begin{cases} 0 & \text{for } x < 1, \\ 1 - (1 - p)^{\lfloor x \rfloor} & \text{for } x \geq 1, \end{cases}$$

If $f_p(k) = p(1 - p)^{k-1}$, here does the $1 - (1 - p)^{\lfloor x \rfloor}$ come from?