Week 3: Cumulative distribution function, mathematical expectation

Armenak Petrosyan

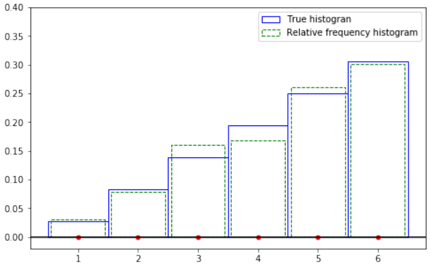
Last time

Relative frequency histogram of data

- Data $=$ values of the random variable on a sequence of trial outcomes.
- As mentioned earlier, probability of an event represents how frequent the experiment outcome terminates in the event, in a large number of repetitive trials.
- Hence, the pmf at $x \in \mathrm{Range}(X)$ can be empirically estimated using the relative frequency

$$f_{\text{emp}}(x) = \frac{\text{number of measurements in data} = x}{\text{size of data}}.$$

- Resulting relative frequency histogram will approximate the pmf histogram.



- A dice is tossed twice and the random variable is the maximum of the two tosses.
- $S = \{(i,j) :\ 1 \leq i \leq 6, 1 \leq j \leq 6\}$ and for any $s = (i,j)$,
$$X(i,j) = \max\{i,j\}.$$
- $\mathrm{Range}(X) = \{1, \ldots, 6\}$ and, for any $x \in \mathrm{Range}(X)$,
$$f(x) = \frac{2x-1}{6^2}.$$
- Data $=$ value of X on 1000 random pairs of tosses.

Tuesday class

Cumulative distribution function (cdf)

Often times we are interested in the following function.

Definition (cdf)

The function

$$F(x) = P(X \leq x), \; x \in \mathbb{R}$$

is called **cumulative distribution function** (cdf).

Example

- Let S be the set of all people in Georgia.
- Let $X(s)$ be the age of the person $s \in S$.
- Then $F(x)$ is the probability that a randomly chosen person is of age or younger than x.

Connection between pmf and cdf

Note the following: if $\mathrm{Range}(X) = \{x_1, \ldots, x_n\}$ with $x_1 < \cdots < x_n$ then

- (pmf $\to$ cdf)

$$F(x) = \begin{cases} 0 & x < x_1 \\ \sum_{i=1}^{k} f(x_i) & x_k \leq x < x_{k+1} \text{ for } k = 1, \ldots, n-1 \\ 1 & x_n \leq x \end{cases}$$

- (cdf $\to$ pmf)

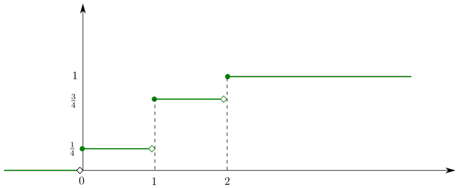
$$f(x_1) = F(x_1), \quad f(x_k) = F(x_k) - F(x_{k-1})$$

  for $k = 2, \ldots, n$.

In the double coin flap experiment, where number of heads was the random variable, we have

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4} & 0 \leq x < 1 \\ \frac{1}{4} + \frac{1}{2} & 1 \leq x < 2 \\ \frac{1}{4} + \frac{1}{2} + \frac{1}{4} & 2 \leq x \end{cases}$$

(pmf $\to$ cdf)



For any discrete random variable, its cdf is al

- a piece-wise constant function,
- non-decreasing (if $x \leq y$ then $F(x) \leq F(y)$),
- right-continuous ($\lim_{y \to x^+} F(y) = F(x)$).

Exercise 1

Problem

Suppose $\mathrm{Range}(X) = \{1, 0, -2, 10, 5\}$ *and the pmf of* X *is*

$$f(1) = 0.05,\ \ f(0) = 0.1, f(-2) = 0.3, f(10) = 0.2, f(5) = 0.35.$$

Compute and draw the cdf of X*.*

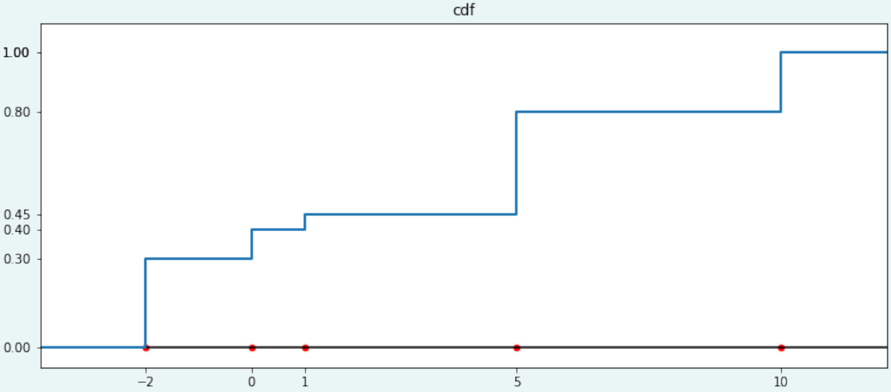
Solution

- *Arrange* $\mathrm{Range}(X)$ *is ascending order*

$$\mathrm{Range}(X) = \{-2, 0, 1, 5, 10\}.$$

- $$F(x) = \begin{cases} 0, & x < -2 \\ f(-2) = 0.3 & -2 \le x < 0 \\ f(-2) + f(0) = 0.4 & 0 \le x < 1 \\ f(-2) + f(0) + f(1) = 0.45 & 1 \le x < 5 \\ f(-2) + f(0) + f(1) + f(5) = 0.8 & 5 \le x < 10 \\ 1 & 10 < x \end{cases}.$$

Solution (cont.)

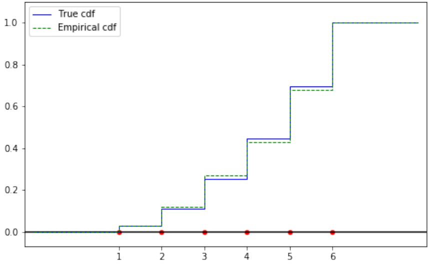


Empirical cdf

- Data = values of the random variable on a sequence of trial outcomes.
- The empirical cdf is computed as follows:

$$F_{\mathsf{emp}}(x) = \frac{\text{number of values } \leq x}{\text{size of data}}.$$

Example



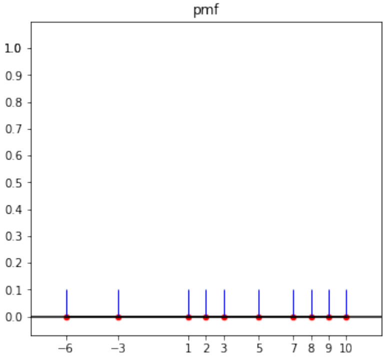
- $S = \{(i,j) : \ 1 \le i \le 6, 1 \le j \le 6\}$ and for any $s = (i,j)$, $X(i,j) = \max\{i,j\}$.
- $\mathrm{Range}(X) = \{1, \dots, 6\}$ and it can be checked that

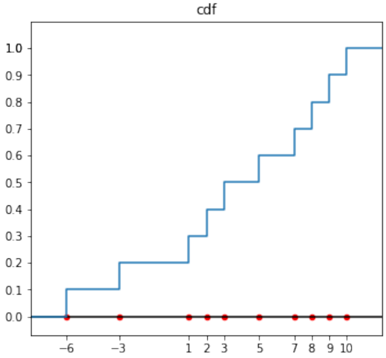
$$F(x) = \begin{cases} 0 & x < 1 \\ \displaystyle\sum_{i=1}^{k} \frac{2i-1}{6^2} = \frac{k^2}{6^2} & k \le x < k+1, \text{ for } k = 1, \dots, 5 \ . \\ 1 & 6 < x \end{cases}$$

- Data = value of X on 100 random pairs of tosses.

Definition

If $\mathrm{Range}(X) = \{x_1, \dots, x_k\}$ and $P(x_1) = \cdots = P(x_k) = \frac{1}{k}$ then we say that X has uniform distribution.





Hypergeometric distribution

- Suppose there are N balls in an urn, K of which are red, the rest are blue.
- n balls are selected without order and without replacement.
- S is the set of all such selections.
- For $s \in S$, let $X(s)$ be the number of red balls in s.

Definition (Hypergeometric distribution)

The pmf of the above random variable is called hypergeometric distribution with parameters (N, K, n).

Theorem

1. *The support of the hyper-geometric distribution with parameters* (N, K, n) *is equal to the set*
$$\mathrm{Range}(X) = \{\max\{0, n-(N-K)\}, \ldots, \min\{n, K\}\}.$$
2. *For any* $x \in \mathrm{Range}(X)$,
$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}.$$

Proof.

1.
   - If $n \geq N-K$ (the number of blue balls) then there must be at least $n-(N-K)$ red balls in any selection thus x is al larger than $n-(N-K)$ in this case.
   - If $n \leq N-K$, then in some selection may not be any red balls.
   - Hence the number of selection al satisfies $x \geq n-(N-K)$.
   - The number of red balls in a selection cannot be more than the number of all selected balls thus $x \leq n$.
   - The number of red balls in a selection cannot be more than the number of all possible red balls thus $x \leq K$.
   - Combining the upper and lower inequalities, we arrive at
$$\max\{0, n-(N-K)\} \leq x \leq \min\{n, K\}.$$

$\square$

Proof (cont.)

2. - The total number of selections is $\binom{N}{n}$.
   - The x red balls can be chosen in $\binom{K}{x}$.
   - The remaining $n-x$ blue balls can be chosen in $\binom{N-K}{n-x}$.
   - Using multiplication principle, the number of all n samples without order and without replacement is

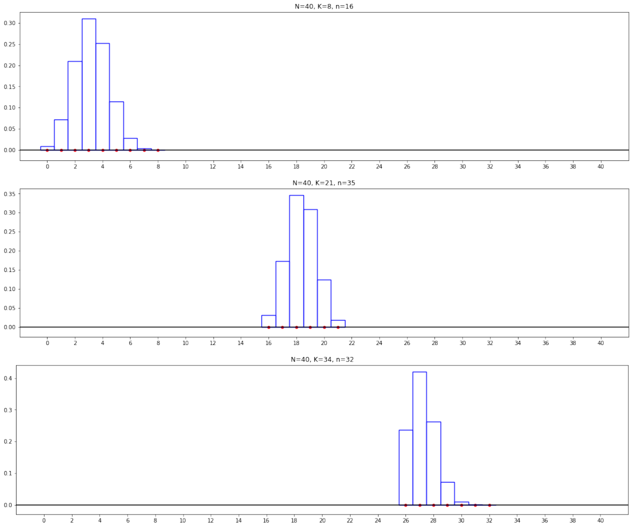
$$\binom{K}{x}\binom{N-K}{n-x}.$$

   - Thus, the probability of the event that there are exactly x balls in a random selection of n is equal to

$$f(x) := P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}.$$

$\square$

Hypergeometric distribution: the histogram



Exercise 2

Problem

There are 1000 voters in a district and 600 of them are voting for Party 1 candidate and 400 are voting for the Party 2 candidate. If an exit poll is conducted with 10 random people leaving the voting station, what is the probability that 6 of them voted for party 1 candidate?

Solution

- *Let s denote the group of people that have been polled.*
- *Let $X(s)$ denote the number of people in s that voted for Party 2 candidate.*
- *X has a hypergeometric distribution with parameters $(N = 1000, K = 600, n = 10)$.*
- *We want to compute $f(6)$.*
- *From the theorem we just proved,*

$$f(6) = \frac{\binom{600}{6}\binom{400}{4}}{\binom{1000}{10}} \approx 0.252.$$

Exercise 3

Problem (1.2-15 in the textbook)

Five cards are selected at random without replacement from a standard, thoroughly shuffled 52-card deck of playing cards. Let X equal the number of face cards (kings, queens, jacks) in the hand. Forty observations of X yielded the following data:

$$\textit{Data} = \{2,1,2,1,0,0,1,0,1,1,0,2,0,2,3,0,1,1,0,3,$$
$$1,2,0,2,0,2,0,1,0,1,1,2,1,0,1,1,2,1,1,0.\}$$

1. *Determine the pmf of* X.
2. *Draw a probability histogram for this distribution.*
3. *Determine the relative frequencies of 0, 1, 2, 3, and superimpose the relative frequency histogram on your probability histogram.*

Solution

1. - X *has Hypergeometric distribution with parameters* $(N = 52, K = 12, n = 5)$.
   - $\mathrm{Range}(X) = \{0, \ldots, 5\}$ *and*

$$f(x) = \frac{\binom{12}{x}\binom{40}{5-x}}{\binom{52}{5}}.$$

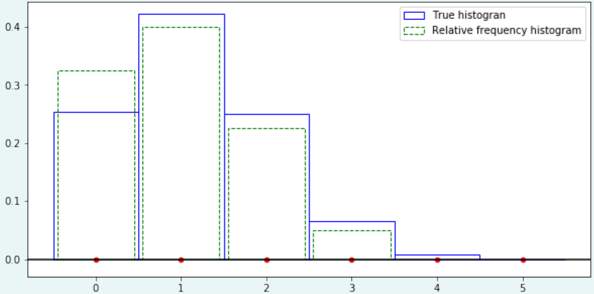
Solution

2. *Numerically can be checked that*

$$f(0) \approx 0.2532 \qquad f(1) \approx 0.4220 \qquad f(2) \approx 0.2509$$
$$f(3) \approx 0.0660 \qquad f(4) \approx 0.0076 \qquad f(5) \approx 0.0003$$

3. *Data* $= \{0\ (\times 13), 1\ (\times 16), 2\ (\times 9), 3\ (\times 2)\}$ *hence*

$$f_{freq}(0) = \frac{13}{40} \approx 0.3250 \qquad f_{freq}(1) = \frac{16}{40} \approx 0.400 \qquad f_{freq}(2) = \frac{9}{40} \approx 0.225$$
$$f_{freq}(3) = \frac{2}{40} \approx 0.0500 \qquad f_{freq}(4) = 0.0000 \qquad f_{freq}(5) = 0.0000$$



Thursday class

Example

A local government is conducting a survey to understand the average household size under its jurisdiction. They randomly select n households and record the number of people. The estimated average size is

$$\text{Average} = \frac{\text{\# of all people in the surveyed households}}{n} =$$

$$\frac{1 \times (\text{\# of households with 1 people}) + 2 \times (\text{\# of households with 2 people}) + \cdots}{n} =$$

$$1 \times \frac{\text{\# of households with 1 people}}{n} + 2 \times \frac{\text{\# of households with 2 people}}{n} + \cdots$$

- Let S denote the set of all households.
- Let $X(s)$ be the number of people in household s.
- Let $f(n)$ be the pmf of X for $n = 1, 2, \ldots$.
- Compare above sum to

$$1 \cdot f(1) + 2 \cdot f(2) + \cdots$$

in terms of relative frequencies.

Mathematical expectation

Definition (Mathematical expectation)

The **mathematical expectation** or the **expected value** of the random variable X is called the number

$$E[X] = \sum_{x \in \mathrm{Range}(X)} x f(x)$$

assuming the sum is absolutely convergent:

$$\sum_{x \in \mathrm{Range}(X)} |x| f(x) < \infty.$$

X is a discrete random variable, so $\mathrm{Range}(X)$ can be enumerated: $\mathrm{Range}(X) = \{x_1, x_2, \dots\}$. Then we understand

$$\sum_{x \in \mathrm{Range}(X)} x f(x) := \lim_{N \to \infty} \sum_{i=1}^{N} x_i f(x_i)$$

for one such enumeration.

- Without absolute convergence, $E[X]$ is not well defined – it will depend on the order of enumeration of $\mathrm{Range}(X)$ (*Rieman series theorem*).
- If the sum above is not absolutely convergent, we say that the **expected value of X does not exist**.

Example

If $\mathrm{Range}(X)$ is finite, the mathematical expectation of X always exists.

Example (Problem 2.2-6 in the textbook)

- Let $S = \mathbb{N}$
- Take $X(n) = n^2$, for every $n \in \mathbb{N}$.
- $\mathrm{Range}(X) = \{1^2, 2^2, 3^2, \dots\}$
- Take $f(x) = \frac{6}{\pi^2 x}$, for every $x \in \mathrm{Range}(X)$.
- $\sum\limits_{x \in \mathrm{Range}(X)} f(x) = \sum\limits_{n=1}^{\infty} \frac{6}{\pi^2 n^2} = 1$ so this is a pmf.
- Notice that

$$E[X] = \sum_{x \in \mathrm{Range}(X)} f(x) = \sum_{n=1}^{\infty} n^2 f(n^2) = \frac{6}{\pi^2} \sum_{n=1}^{\infty} 1 = \infty.$$

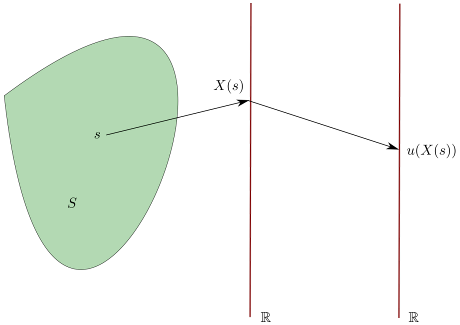
- Thus the mathematical expectation of X does not exist.

Change of a random variable

- Let $X : S \to \mathbb{R}$ be a random variable on the set of outcomes S.
- Let $u : \mathbb{R} \to \mathbb{R}$ be any function (e.g. $u(x) = x^2$).
- Then the composition function

$$Y = u(X), \quad u(X) : S \to \mathbb{R}$$

will be a new random variable.



Example

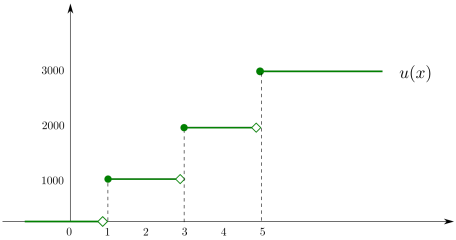
A local government planning a crisis relief to the population during a pandemic. The proposed plan is to provide 1000$ to households with 1-2 people, \$2000 to households with 3-4 people, and \$3000 for households with 5 or more people. Let us find the expected value of the amount a household will receive as part of this relief plan.

On one hand

- Let S denote the set of all households.
- Let $Y(s)$ be the amount of help the household s receives.
- Let $f_Y(y)$ be the pmf of Y at $y \in \text{Range}(Y) = \{1000, 2000, 3000\}$.
- Then
$$E[Y] = 1000 \cdot f_Y(1000) + 2000 \cdot f_Y(2000) + 3000 \cdot f_Y(3000).$$

On the other hand

- Let $X(s)$ be the number of people in household s.
- Let $f_X(n)$ be the pmf of X for $n = 1, 2, \ldots$.
- Then $Y = u(X)$ for
$$u(x) = \begin{cases} 0 & x < 1 \\ 1000 & 1 \le x < 3 \\ 2000 & 3 \le x < 5 \\ 3000 & 5 \le x \end{cases}.$$



- Notice that

$$f_Y(1000) = f_X(1) + f_X(2)$$
$$f_Y(2000) = f_X(3) + f_X(4)$$
$$f_Y(3000) = f_X(5) + f_X(6) + \cdots$$

- Consequently,

$$\begin{aligned} E[Y] &= 1000 \cdot [f_X(1) + f_X(2)] + 2000 \cdot [f_X(3) + f_X(4)] + 3000 \cdot [f_X(5) + f_X(6) + \cdots] \\ &= u(1) \cdot f_X(1) + u(2) \cdot f_X(2) + \cdots = \\ &= \sum_{x \in \mathrm{Range} X} u(x) f_X(x). \end{aligned}$$

Theorem

Let X be a random variable and $u : \mathbb{R} \to \mathbb{R}$ be a function. If

$$\sum_{x \in \mathrm{Range}(X)} |u(x)| f(x) < \infty.$$

Then the expected value of the random variable $Y = u(X)$ exists and

$$E[Y] = \sum_{x \in \mathrm{Range}(X)} u(x) f(x).$$

- **We do not prove this theorem but you can use it.**
- The idea of proof: similar to the example above combined with Fubini's theorem.

Exercise 4

Problem (2.2-5 in the textbook)

Let the random variable X be the number of days that a certain patient needs to be in the hospital. Suppose X has the pmf

$$f(x) = \frac{5-x}{10}, \quad x = 1, 2, 3, 4.$$

If the patient is to receive $200 from an insurance company for each of the first two days in the hospital and $100 for each day after the first two days, what is the expected payment for the hospitalization?

Solution

- $X(s) =$ *number of days the patient* S *spends in the hospital.*
- $u(x) =$ *amount of money received for staying in hospital for* x *days:*

$$u(1) = 200, \quad u(2) = 400, \quad u(3) = 500, \quad u(4) = 600.$$

- $Y(s)=$ *amount of money patient* s *receives.*
-

$$E[Y] = \sum_x u(x) f(x) = 200\frac{4}{10} + 400\frac{3}{10} + 500\frac{2}{10} + 600\frac{1}{10} = 360.$$

Properties of expectation

Theorem

1. *For constant random variable $X \equiv c$ (this notation means $X(s) = c$ for all $s \in S$)*

$$E[c] = c.$$

2. *For any number c and any random variable X whose expected value exists,*

$$E[c \cdot X] = c \cdot E[X].$$

3. *For any numbers c_1, c_2, any functions u_1, u_2, and any random variable X,*

$$E[c_1 u_1(X) + c_2 u_2(X)] = c_1 E[u_1(X)] + c_2 E[u_2(X)].$$

(under the assumption that all the above expected values exist).

Proof

1. $\mathrm{Range}(X) = \{c\}$ and $f(c) = P(S) = 1$ so

$$E[c] = \sum_{x \in \mathrm{Range}(X)} x f(x) = c f(c) = c.$$

cont.

2. Take $u(x) = cx$, then

$$E[c \cdot X] = \sum_{x \in \text{Range}(X)} u(x)f(x) = \sum_{x \in \text{Range}(X)} cxf(x) = c \sum_{x \in \text{Range}(X)} f(x) = cE[X].$$

3. $u(x) = c_1u_1(x) + c_2u_2(x)$. Then

$$\begin{aligned}
E[c_1u_1(X) + c_2u_2(X)] &= E[u(X)] \\
&= \sum_{x \in \text{Range}(X)} u(x)f(x) \\
&= \sum_{x \in \text{Range}(X)} [c_1u_1(x) + c_2u_2(x)]f(x) = \\
&= c_1 \sum_{x \in \text{Range}(X)} u_1(x)f(x) + c_2 \sum_{x \in \text{Range}(X)} u_2(x)f(x) \\
&= c_1E[u_1(X)] + c_2E[u_2(X)].
\end{aligned}$$

$\square$

- Condition 3. extends to multiple function $u_1, \ldots, u_k$ by induction

$$E[c_1 u_1(X) + \cdots + c_k u_k(X)] = c_1 E[u_1(X)] + \cdots + c_k E[u_k(X)].$$

- Due to properties 2. and 3., we say that $X \mapsto E[X]$ correspondence is a linear functional.
- The line function $l(z) = az$ has similar properties; the name is derived from there.

Example

- Put $u(x) = (x-b)^2$.
- Notice
$$g(b) = E[(X-b)^2] = E[X^2 - 2bX + b^2] = E[X^2] - 2bE[X] + b^2.$$
- Let us compute the minimum of the function $g(b)$:
$$\frac{\partial g}{\partial b}(b) = 2b - 2E[X] = 0$$
hence the minimum is at $b = E[X]$.

Intuitively: expected value is the "center" of the histogram where it concentrates (the point all of them are simultaneously close to).

Hypergeometric distribution

Theorem

Let X be a hypergeometric distribution with parameters (N, K, n). Then

$$E[X] = \frac{nK}{N}.$$

Intuitively: red balls are the $\frac{K}{N}$ part of all balls. We are doing n selections so expect to get on average $n\frac{K}{N}$ red balls.