11 Cufrence and cumulative density

Recall that any sample $y = (y_k)_{k=1..N}$ determines the multance $(x, m) = (x_j, m_j)_{j=1..n}$ and the frequence $(x, f) = (x_j, f_j)_{j=1..n}$. Here we complete this collection by a **cufrence**, a **cumulative frequency sequence**, of y; it is the sequence

$$(x, F) = ((x_j)_{j=1..n}, (F_j)_{j=1..n}) = ((x_j, F_j))_{j=1..n},$$

where

$$F_j := f_1 + f_2 + \ldots + f_j = \sum_{i=1}^{j} f_i \,, j = 1..n,$$

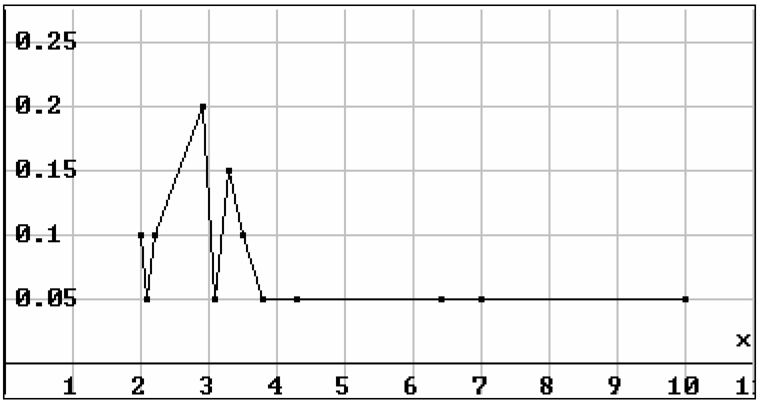
is referred to as a j-th **cumulative frequency**. Commonly, in both just introduced notions the word 'frequency' can be replaced by the word 'mass' or 'density'. Thus, for example, F_j is called a j-th **cumulative mass**, a j-th **cumulative density**.

Obviously, the description of arbitrary sample y via its cufrence (x, F) is equivalent to the description of its freqence (x, f).

Example–16. Let's deal with the payroll in the enterprise *We20*. We can easily produce its cumulative frequencies F_j by summing consecutive frequencies f_j listed in the freqence table and storing these sums in the column to the right.

j	x_j	f_j	F_j
1	2.0	0.10	0.10
2	2.1	0.05	0.15
3	2.2	0.10	0.25
4	2.9	0.20	0.45
5	3.1	0.05	0.50
6	3.3	0.15	0.65
7	3.5	0.10	0.75
8	3.8	0.05	0.80
9	4.3	0.05	0.85
10	6.4	0.05	0.90
11	7.0	0.05	0.95
12	10.0	0.05	1.00

Above: the table of the sequence y discussed in Example 16.



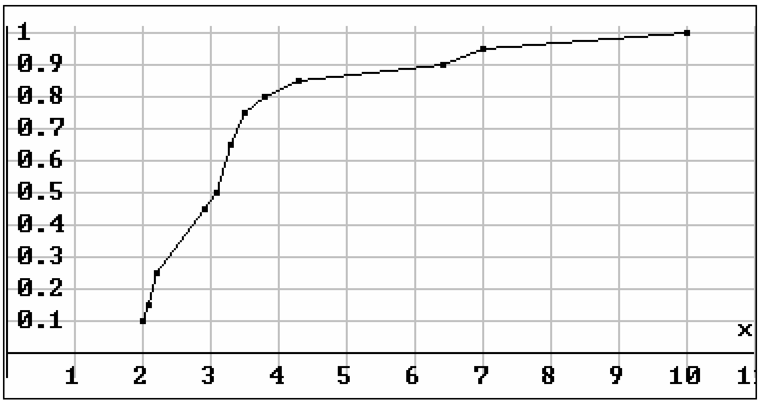


Fig.11.1. (above) The line plot of the frequence (x, f)

Fig.11.2. (below) The line plot of the cufrence (x, F)

□ *Example–17*. DRAFT VERSION

In descriptive statistics there are considered only so-called discrete distributions, i.e., distributions defined via sequences, distributions described by freqences (x, f) as well as cufrences (x, F). One can see frequencies f_j as values which are assumed by certain function (defined on an interval), let's denote it f, when its argument is taken x_j,

$$f_j = f(x_j).$$

This function f is referred to as a **density**, or a **mass function** (of considered distribution), a **PDF**, **probability mass function** (and this name is commonly used in mathematical statistics). For instance, one can say that the frequence $(x, f) = (1, 0.2; 2, 0.8)$ is induced by the function $f(x) = 0.2x^2$; really,

$$f(x_1) = 0.2 \cdot 1^2 = 0.2 = f_1,$$
$$f(x_2) = 0.2 \cdot 2^2 = 0.8 = f_2.$$

Notice that there are infinitely many functions f_i working in this way. For instance, the same frequence, namely $(x, f) = (1, 0.2; 2, 0.8)$, is produced with $f(x) = 0.6 \cdot x - 0.4$.

Analogously, for every cufrence $(x, F) = ((x_j, F_j))_{j=1..n}$ there exists the function (defined on the whole real axis) F such that

$$F_j = F(x_j);$$

this function F is defined as follows

$$F(x) = 0 \text{ if } x < x_1,$$
$$F(x) = F_j \text{ for } x \in \langle x_j, x_{j+1}) \text{ and } j=1..n,$$
$$F(x) = 1 \text{ when } x \geq x_n,$$

and it is referred to as a **CDF**, **cumulative density function**, or **cumulative mass function**.

Both PDF and CDF are of fundamental importance in mathematical statistics, they serve to define so-called theoretical distributions [1]]. An exemplary theoretical distribution is the binomial distribution, and we deal with it below.

The line plot presented above, Fig.11.2, is one of possible visualizations of a cufrence (x, F). Another visualization of any cufrence (x, F) – in fact, commonly used in statistics – is to plot CDF, i.e., to draw horizontal segments, for $j = 1..n-1$ the j-th horizontal segment is drawn on the level F_j above/over the interval $\langle x_j, x_{j+1})$, and these $n-1$ segments are completed by two semilines: that

[1]] In some sense CDF is even more essential that PDF, namely in terms of the probability it can not be clearly interpreted what PDF of a continuous distribution is, but it is easy to interpret what CDF is. Although here we do not go in this subject, let's mention that PDF f describes the probability to have a value x_j (this probability is equal to $f_j = f(x_j)$), and CDF F provides the probability to have a value less than x_j (this probability is F_j). This is denoted as

$$f_j = \Pr\{ X = x_j \},\ F_j = \Pr\ \{ X \leq x_j \},$$

where X is a random variable (the notion 'random variable' will be defined later).

laying on the horizontal axis *Ox* and covered by equation $y = 0$ for $x < x_1$, and that described by $y = 1$ for $x \geq x_n$. So, the graph produced in this way is composed of n+1 horizontal segments, the most left one is unbounded from the left, and the most right one is extended to the infinity. This visualization may be called a **segmental chart/plot**, see figures below.

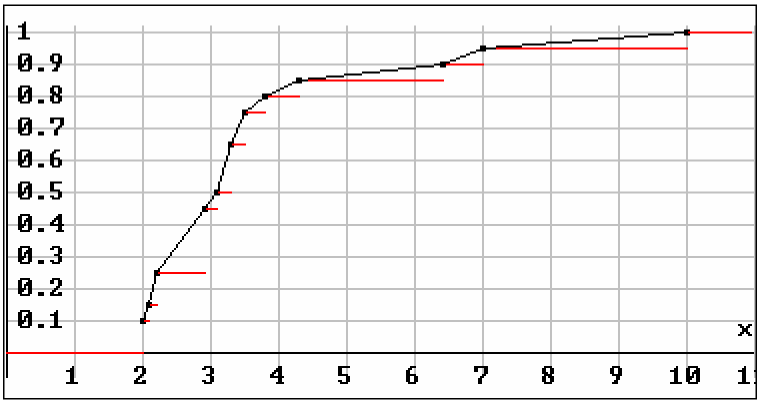


Fig.11.3. The line plot and the segmental plot of the cufrence (x, F)

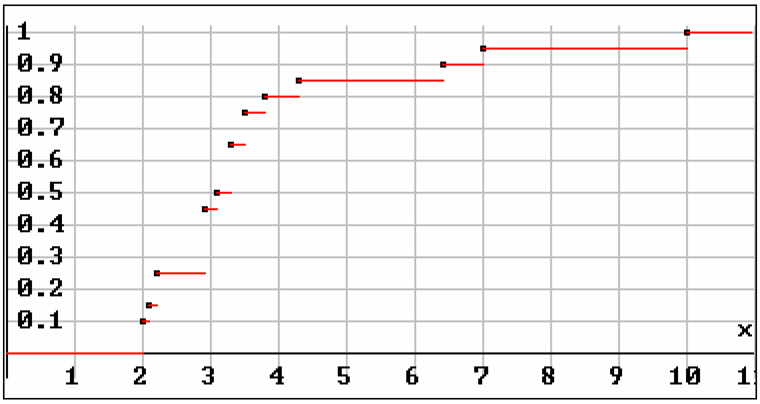


Fig.11.4.
The segmental plot of the cufrence (x, F), the plot of CDF F; for instance $F(x) = 0.85$ if $4.3 \leq x < 6.4$.

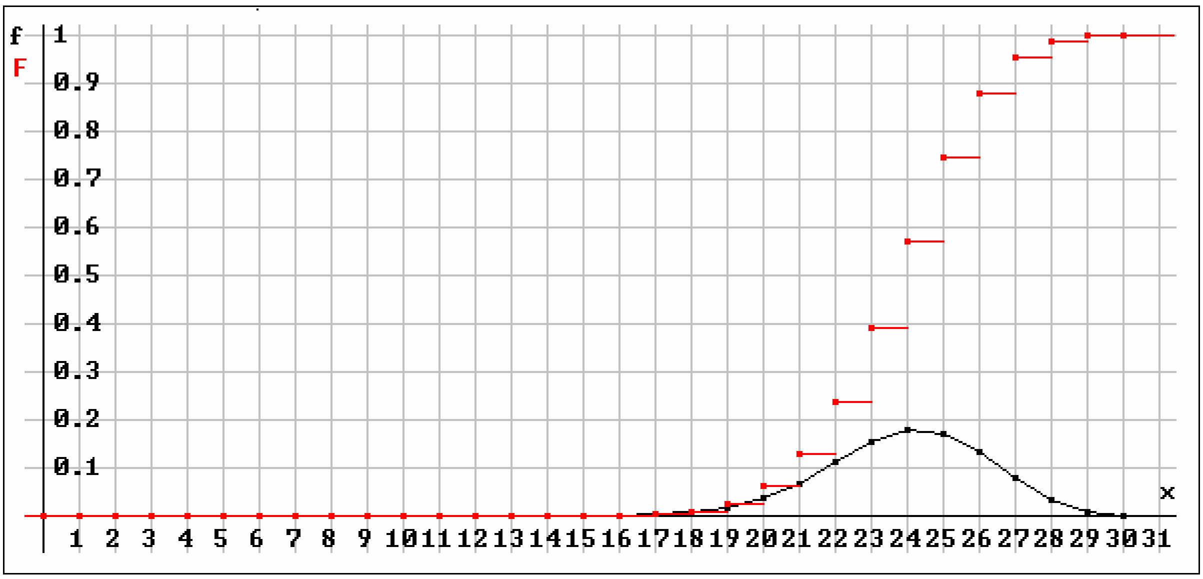


Fig.11.5. The line plot of the Binomial(30, 0.8) density f and the (segmental) plot of Binomial(30, 0.8) cumulative density F; for instance, $f(x_{23}) = f(23) = 0.15382$, $F(x) = 0.39303$ for $x \in$ <23, 24)

j	f_j	F_j
0	$1.07374 \cdot 10^{-21}$	$1.07374 \cdot 10^{-21}$
1	$1.28849 \cdot 10^{-19}$	$1.29922 \cdot 10^{-19}$
2	$7.47324 \cdot 10^{-18}$	$7.60316 \cdot 10^{-18}$
3	$2.79001 \cdot 10^{-16}$	$2.86604 \cdot 10^{-16}$
4	$7.53302 \cdot 10^{-15}$	$7.81963 \cdot 10^{-15}$
5	$1.56687 \cdot 10^{-13}$	$1.64506 \cdot 10^{-13}$
6	$2.61145 \cdot 10^{-12}$	$2.77595 \cdot 10^{-12}$
7	$3.58141 \cdot 10^{-11}$	$3.85901 \cdot 10^{-11}$
8	$4.11862 \cdot 10^{-10}$	$4.50453 \cdot 10^{-10}$
9	$4.02710 \cdot 10^{-9}$	$4.47755 \cdot 10^{-9}$
10	$3.38276 \cdot 10^{-8}$	$3.83052 \cdot 10^{-8}$
11	0.000000246019	0.000000284324
12	0.00000155812	0.00000184244
13	0.00000862960	0.0000104720
14	0.0000419152	0.0000523872
15	0.000178838	0.000231225
16	0.000670643	0.000901869
17	0.00220917	0.00311104
18	0.00638207	0.00949312
19	0.0161231	0.0256162
20	0.0354708	0.0610871
21	0.0675636	0.128650
22	0.110558	0.239209
23	0.153820	0.393030
24	0.179457	0.572487
25	0.172279	0.744766
26	0.132522	0.877289
27	0.0785318	0.955821
28	0.0336564	0.989477
29	0.00928455	0.998762
30	0.00123794	1

Binomial(30, 0.8) distribution: values of frequencies $f_j = f(j) = C_{30,j} \cdot 0.8^j \cdot 0.2^{30-j}$ ($C_{n,j}$ denotes the j-th Newton coefficient) and of cumulative densities F_j
DRAFT VERSION