Лекция 3. Статистические методы обработки информации в нефтегазовом деле.

Составитель асс. каф. БНГС СамГТУ, магистр Никитин В.И.

2. ТЕОРИЯ ВЕРОЯТНОСТИ

2.1. Вероятность

Вероятность - числовая характеристика степени возможности наступления какого-либо определённого события в тех или иных определённых, способных повториться неограниченное число раз условиях. Вероятность отражает особый тип связей между явлениями, характерных для массовых процессов. Обычно численное значение вероятности находится с помощью определения вероятности. Рассмотрим вероятность наступления события *B*. Для этого проведём n испытаний, и зафиксируем число n_i - количество благоприятных исходов, т.е. исходов наступления события *B*. Тогда вероятность равна отношению числа n_i ,благоприятных исходов к общему числу равновозможных исходов, n. Связь вероятности P с относительной частотой события h_i зависит от общего числа испытаний n. Чем больше число n, тем реже встречаются сколько-либо значительные отклонения относительной частоты h_i от вероятности P. Таким образом, вероятность осуществления события *B* будет пределом:

$$P(B) = \lim_{n\to\infty} \frac{n_i}{n} \qquad (2.1)$$

Таким образом, каждому событию *B* соответствует некоторое неотрицательное число - его вероятность: $0 \le P \le 1$, причём для *невозможного* события $P = 0$, для достоверного $P = 1$. В соответствии с этими аксиомами падение подброшенной монеты на землю является достоверным событием, её "взлёт" - невозможное событие, а вероятности выпадения "герба" - ½, “решки”- ½, эти события – равновероятны, соответственно. Другими словами, результат падения монеты (и не только монеты.) - *случайная величина*. В повседневной жизни часто употребляется понятие вероятности в смысле возможности

наступления того или иного события, и интуитивно её указывают в процентах, что соответствует умножению (2.1) на 100%.

2.2. Распределение вероятностей случайной величины. Графические представления анализа выборки

Для дискретных случайных величин характерно то, что они не могут меняться непрерывно, т.е. плавно переходить от одного значения к другому. Поэтому значения дискретных величин изменяются скачкообразно и их значения соответствуют определённому набору. Распределение вероятностей называется дискретным, если случайная величина *X* может принимать только конкретные возможные значения $x_1, x_2, ... x_n$, которым соответствуют вероятности $P_1, P_2, ... P_n, \sum_{i=1}^{n} P_i = 1.$

Гистограммой называется ступенчатая фигура, для построения которой по оси абсцисс откладывают отрезки, изображающие частичные интервалы $(x_{i-1}; x_i)$ варьирования признака *X*, и на этих отрезках, как на основаниях, строят прямоугольники с высотами, равными частотам или частостям соответствующих интервалов. При увеличении до бесконечности размера выборки выборочные функции распределения превращаются в теоретические: гистограмма превращается в график плотности распределения.

Графически дискретный вариационный ряд изображают в виде полигона частот (в виде полигона относительных частот) следующим образом. Сначала на числовой плоскости строят точки (x_i, n_i)(точки (x_i, h_i)), где x_i — *i*-я варианта. Затем строят ломаную, соединяющую построенные точки, которую и называют *полигоном*.

Вариационные ряды графически можно изобразить в виде *кумулятивной кривой* (кривой сумм — *кумуляты*). При построении кумуляты дискретного вариационного ряда на оси абсцисс откладывают варианты x_i , а по оси ординат соответствующие им *накопленные частоты* W_i . соединяя точки (x_i, W_i) отрезками, получаем ломаную, которую называют *кумулятой*. Для получения

накопленных частот и дальнейшего построения точек (x_i, W_i) составляется расчетная табл. 2.1. Накопленные частоты вычисляются для каждого интервала по правилу:

$$W_i = W_{i-1} + h_i. \qquad (2.2)$$

При построении кумуляты интервального вариационного ряда левому концу первого интервала сопоставляется частота, равная нулю, а правому — частота этого интервала. Правому концу второго интервала соответствует накопленная частота первых двух интервалов, то есть сумма частот этих интервалов и т. д. Правая граница последнего интервала равна сумме всех частот, то есть объему *n* выборки. Для характеристики свойств статистического распределения в математической статистике вводится понятие *эмпирической функции распределения*.

Таблица 2.1

Варианты, x_i	x_1	x_2	…	x_k
Относительные частоты h_i	$h_1 = n_1/n$	$h_2 = n_2/n$	…	$h_k = n_k/n$
Накопленные относительные частоты $W_i = W_{i-1} + h_i$	$W_1 = h_1$	$W_2 = W_1 + h_2$	…	$W_k = W_{k-1} + h_k$

Расчетная таблица для построения кумулятивной кривой и эмпирической функции распределения.

$F(x)$- называется *функцией распределения* дискретной случайной величины. Если случайная величина *X* принимает конечное число дискретных значений (например, число очков на гранях игральной кости), то функция распределения вероятностей этой случайной величины представляет собой ступенчатую функцию. График следует понимать так: вероятность того, что случайная величина *X* примет значение x_1 равна P_1; вероятность того, что случайная величина *X* примет значение x_2 равна P_2; вероятность того, что

случайная величина X примет значение x_1 или x_2 равна $P_1 + P_2$ и т.д. Вероятность того, что случайная величина X примет любое значение $x_1, x_2, ... x_n$ P равна 1, это достоверное событие (Рис.2.1).

Эмпирическую функцию распределения $F_n(x)$ получают построением ступенчатой кривой относительных накопленных частот: $F_n(x)$ имеет скачки в точках соответствующих серединам интервалов.

Эмпирическая функция $F_n(x)$ служит для оценки теоретической функции распределения генеральной совокупности. Различие между ними состоит в том, что теоретическая функция $F(x)$ определяет вероятность события $X<x$, а эмпирическая функция $F_n(x)$ определяет относительную частоту этого события.

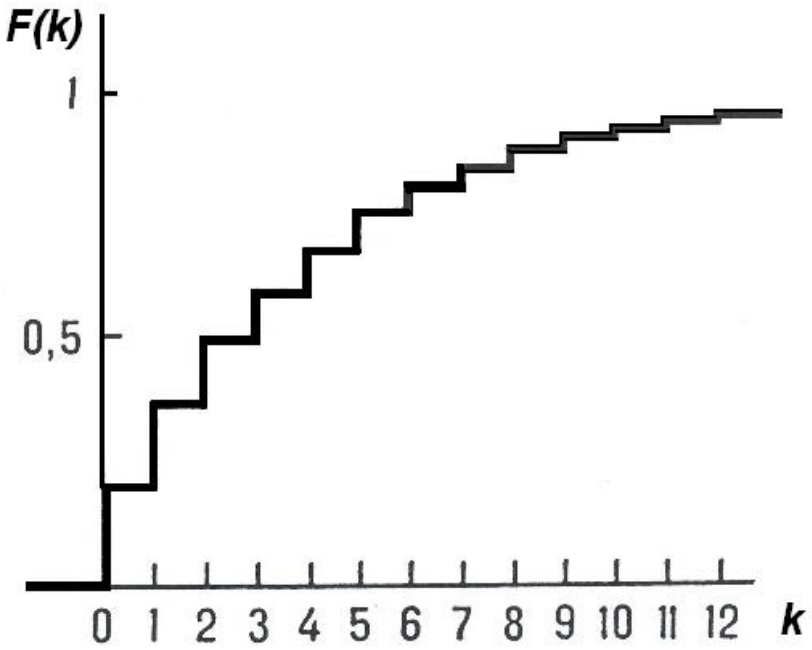

Рис.2.1. Функция распределения дискретной случайной величины.

По гистограмме и полигону частот судят о виде плотности распределения исследуемой непрерывной случайной величины или о распределении вероятностей дискретной случайной величины. Эмпирическая функция распределения дает представление о функции распределения и используется в основном в статистической проверке гипотез.

2.3. Функция распределения вероятностей непрерывной случайной величины

Функция распределения вероятностей является *монотонной* и *неубывающей.* Ордината кривой, соответствующая точке x_1, представляет собой вероятность того, что случайная величина *X* при испытании окажется меньше x_1. Ордината кривой, соответствующая точке x_2 представляет собой вероятность того, что случайная величина *X* при испытании окажется меньше x_2. Разность двух ординат, соответствующая точкам x_1 и x_2. даёт вероятность того, что значения случайной величины будут лежать в интервале между x_1 и x_2. Рис.2.2.

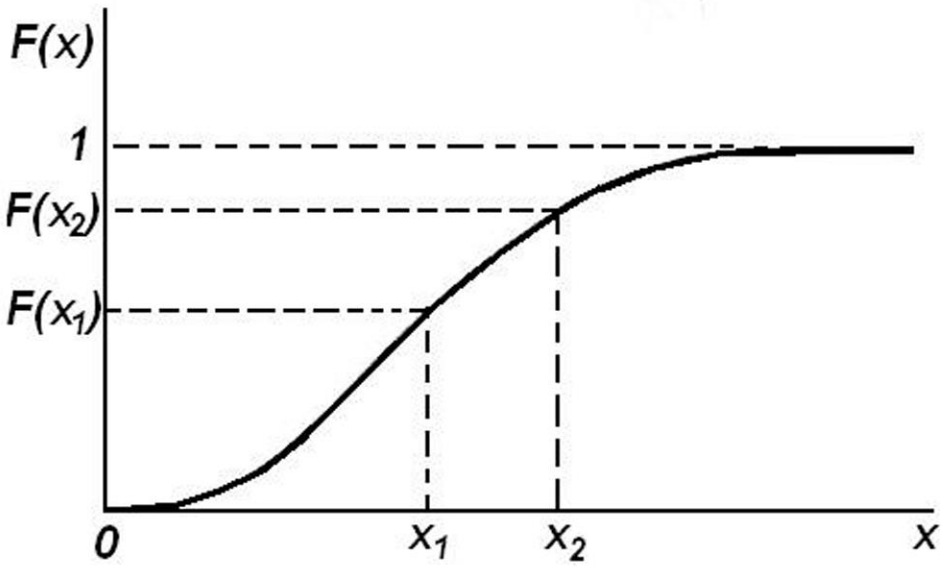

Рис.2.2. Теоретическая функция распределения непрерывной случайной величины.

2.4. Плотность распределения вероятностей

Плотность вероятностей - функция, характеризующая вероятность того, что значения случайной величины *X* будут заключены в том или ином интервале. Другими словами, плотность вероятностей характеризует вероятность попадания случайной величины *X* в интервал $(x_{i-1}; x_i)$; плотность

случайных величин X в том или ином интервале. Плотность вероятностей $p(x)$ есть всегда действительная неотрицательная функция. По сути плотность вероятностей аналогична гистограмме распределения, но оперирует бесконечно малыми величинами:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} p(x)dx = F(x_2) - F(x_1) \tag{2.3}$$

$$p(x) = F'(x) = dF/dx \tag{2.4}$$

Естественно, что $P(x_1 < X < x_2)$ - доля от единицы, это площадь под кривой плотности вероятностей в интервале $(x_1; x_2)$.

Плотность вероятностей случайной величины ***X*** **(**Рис.2.3). Заштрихована область, площадь которой соответствует вероятности попадания случайной величины в интервал $(x_1; x_2)$ (общая площадь под кривой плотности вероятностей равна 1).

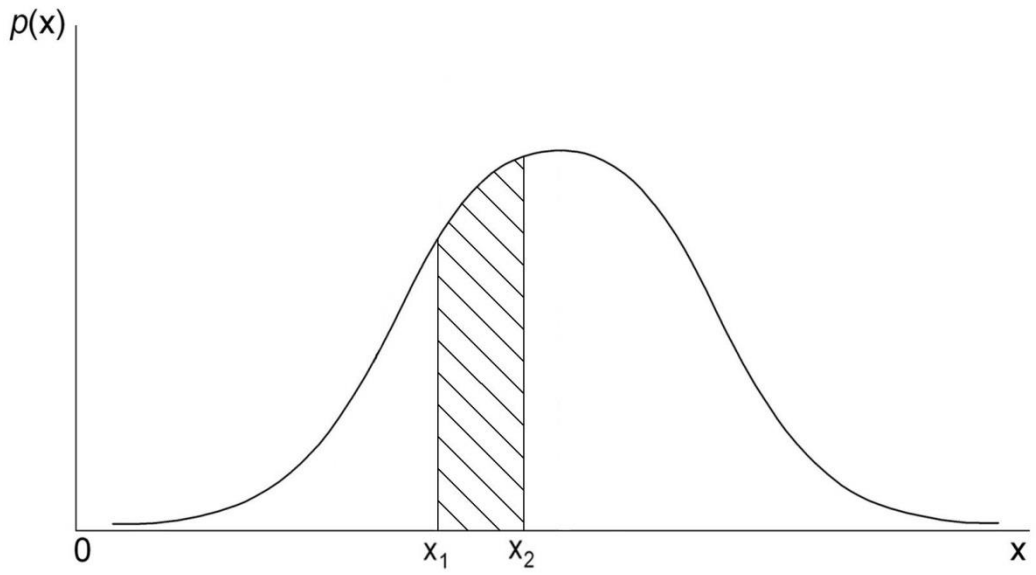

Рис.2.3. Плотность вероятностей случайной величины X

ПРИМЕР 2. (Построение гистограммы, полигона, кумуляты)

Найти эмпирическую функцию распределения значений механической скорости бурения, представленной в вариационном ряду Примера 1. Примите количество интервалов K=5.Построить гистограмму, полигон и кумуляту.

Решение

Воспользуемся интервальным вариационным рядом, полученным в Примере 1. Дополним имеющуюся таблицу накопленными относительными частотами, согласно формуле (2.2).

(Интервалы-варианты) Скорость проходки, м/ч	$0.65-0.69$	$0.69-0.73$	$0.73-0.77$	$0.77-0.81$	$0.81-0.85$
Частоты n_i	4	19	32	15	10
Середины интервалов x_i^*	0.67	0.71	0.75	0.79	0.83
Относительные частоты $h_i = n_i/n$	0.05	0.2375	0.4	0.1875	0.125
Накопленные отн. частоты $W_i = W_{i-1} + h_i$	0.05	0.2875	0.6875	0.875	1

Используя накопленные относительные частоты построим эмпирическую функцию распределения. Заметим, что при этом следует пользовать не интервальным представлением вариационного ряда, а его представлением через середины интервалов. Таким образом, в аналитическом виде $F_n(x)$ записывается следующим образом:

$$F_n(x) = \begin{cases} 0, x \in (-\infty, 0.67] \\ 0.05, x \in (0.67, 0.71] \\ 0.29, x \in (0.71, 0.75] \\ 0.69, x \in (0.75, 0.79] \\ 0.88, x \in (0.79, 0.83] \\ 1, x \in (0.83, +\infty) \end{cases}$$

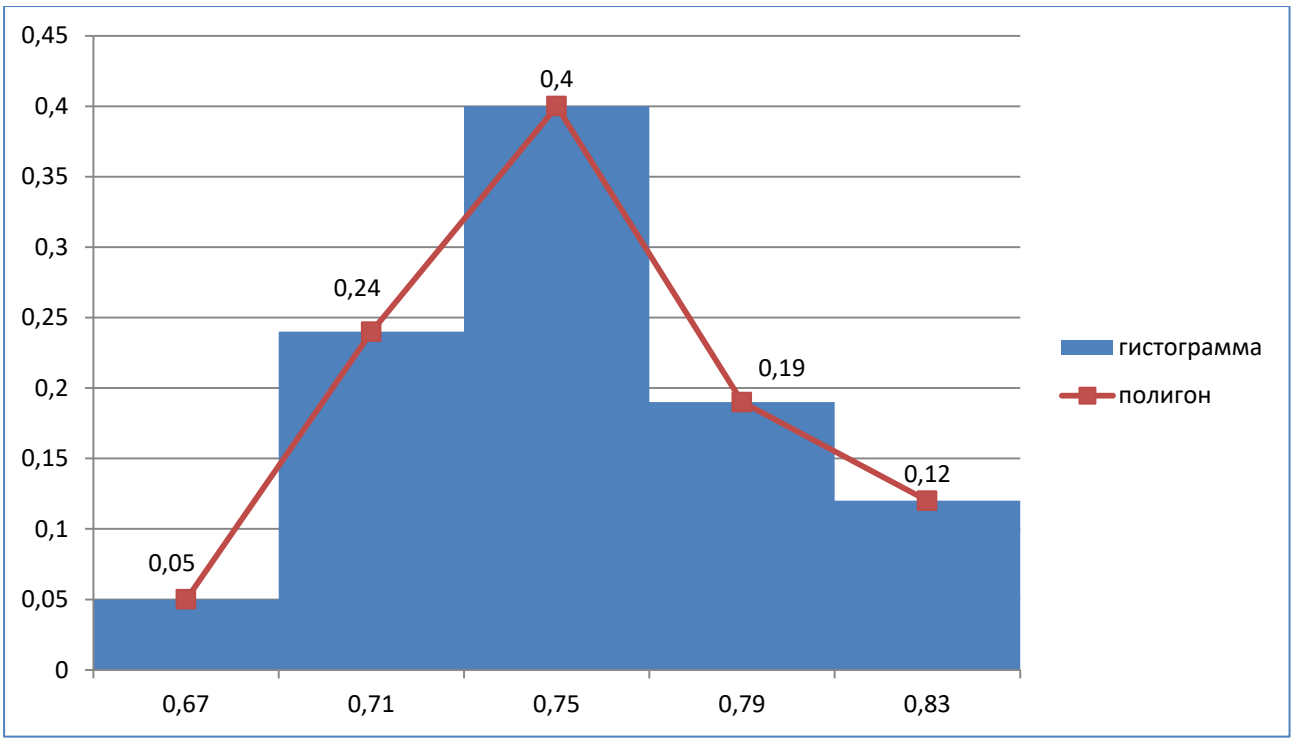

Рис.4. Гистограмма и полигон частот

На рис.4 Изображена гистограмма и полигон относительных частот. График был построен при помощи электронных таблиц Microsoft Office.

Построим кумуляту и эмпирическую функцию распределения. Для интервальных данных ломаная линия начинается с точки, абсцисса которой равна началу первого интервала, а ордината – накопленной частоте, равной нулю. Другие точки этой ломаной соответствуют концам интервалов и накопленным частотам. Эмпирическая функция распределения строится согласно её аналитической записи и имеет скачки в серединах интервалов. Диаграмма значений функции распределения и кумулята представлены на рис.5.

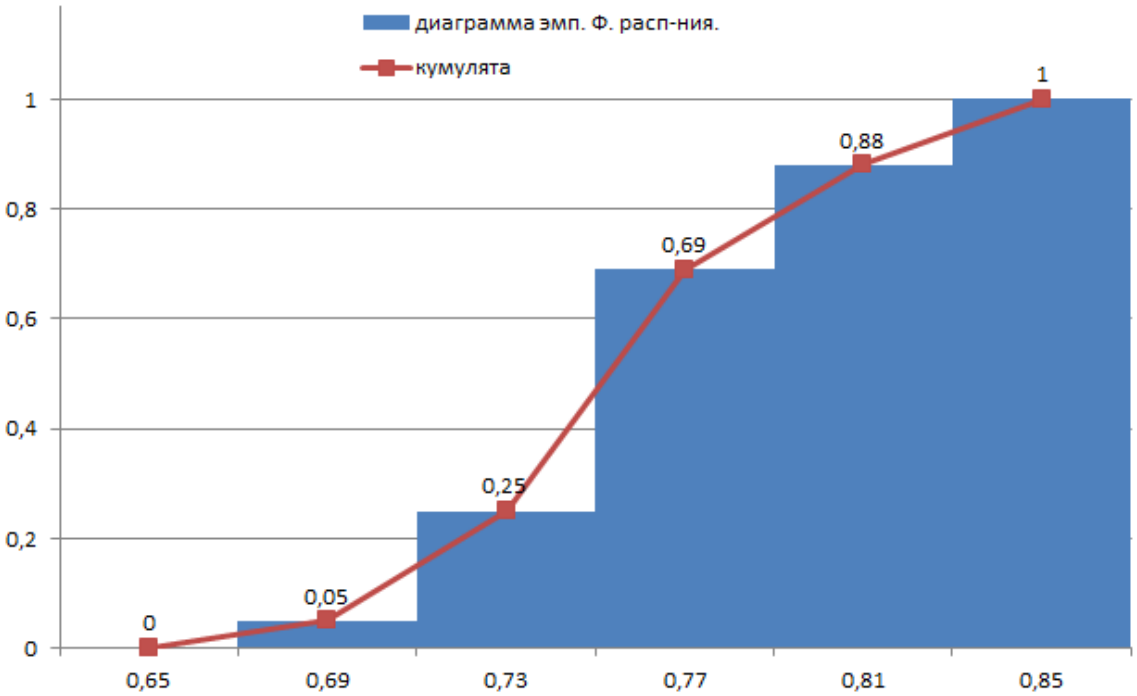

Рис.5. Кумулята и диаграмма эмпирической функции распределения.