Лекция 1

Вводная. Некоторые понятия теории случайных величин.

Дисциплина «Методы статистической обработки гидрометеорологической информации» является фундаментальной дисциплиной, которая знакомит студентов с современными и перспективными статистическими методами обработки гидрометеорологической информации и их применениями для решения разнообразных задач метеорологии, гидрологии и океанологии. Студенты метеорологической специальности, и, в особенности, специализации «Гидрометеорологические измерения и сетевые технологии», должны изучить как общие положения теории статистической обработки и анализа гидрометеорологической информации, так и специфические подходы в использовании статистических методов для анализа информации о состоянии атмосферы и прогноза погоды.

Главной задачей дисциплины является создание у студента достаточно полного представления о современных статистических способах обработки большого потока разнообразной информации о состоянии атмосферы, в том числе в автоматизированных системах. Цели последующего применения результатов статистического анализа очень разнообразны: при прогнозировании погоды с помощью численных моделей, в климатологии, при мониторинге состояния окружающей среды, при объективном анализе метеорологических полей, при оценке репрезентативности данных наблюдений, точности измерительных приборов, при решении вопроса рационального размещения сети метеорологических станций и др.

В последние десятилетия наблюдается широкое использование аппарата теории случайных функций в геофизических науках. Основой этого является идея рассмотрения фиксированных мгновенных значений гидрометеорологических процессов и пространственных полей как отдельных реализаций некоторого случайного процесса или случайного поля. Такой подход позволяет отказаться от рассмотрения особенностей отдельных мгновенных значений гидрометеорологических полей, зависимость которых от пространственных координат, а также их временной ход носят весьма сложный и запутанный характер, и перейти к рассмотрению некоторых осредненных свойств *статистической совокупности* их реализаций.

Основным понятием в теории вероятностей является *случайная величина*. Дадим ее определение.

Случайной величиной называют такую величину, которая при проведении ряда опытов в одинаковых условиях может каждый раз принимать то или иное значение, заранее неизвестно какое именно.

Можно привести примеры случайных величин - температура воздуха в данном помещении за двадцать лет непрерывных наблюдений, относительная влажность воздуха в этом же помещении или их отклонения от нормы. В качестве случайных величин могут выступать ошибки приборов, с помощью которых производятся измерения.

Различают случайные величины дискретного типа, когда все возможные значения случайной величины можно заранее перечислить, то есть пронумеровать числами натурального ряда, и случайные величины непрерывного типа, когда все возможные значения случайной величины целиком заполняют некоторый промежуток числовой оси. К первому классу можно отнести, например, количество гроз в Петербурге за июль месяц за сто лет наблюдений. Ко второму классу – составляющие скорости ветра или их отклонения от нормы за определенный промежуток времени. Ошибки наблюдений тоже относятся к случайным величинам непрерывного типа. Случайные величины чаще всего обозначают прописными буквами A,D,X и т.д., а их возможные значения – строчными буквами.

Случайные величины определяются своими вероятностными характеристиками – законами распределения и моментами распределения. Интегральный закон распределения $F(x)$ случайной величины X можно определить как вероятность того, что случайная величина примет значение меньшее некоторого числа x

$$F(x) = P(X \lhd x), \quad (1.1)$$

где $P(x \prec x)$ означает вероятность того, что случайная величина X меньше x. Функция $F(x)$ является неубывающей функцией своего аргумента. Причем $F(-\infty) = 0,$ а $F(+\infty) = 1.$

В практической статистике интегральный закон распределения можно построить, пользуясь таким понятием как *огива*. По оси абсцисс откладываются верхние математические границы случайной величины, а по оси ординат – процентные повторяемости (или число случаев) значений случайной величины, меньших соответствующей математической границы. Каждой верхней границе соответствует одна повторяемость.

Для случайной величины непрерывного типа, функция распределения которой дифференцируема, в качестве закона распределения можно использовать производную от функции распределения

$$f(x) = F'(x) = \lim \frac{F(x + \Delta x) - F(x)}{\Delta x}, \text{ при } \Delta x \to 0 \quad (1.2)$$

которую обозначают $f(x)$ и называют дифференциальным законом распределения или плотностью распределения. Плотность распределения как производная от неубывающей функции $F(x)$ является неотрицательной функцией.

Функция распределения выражается через плотность распределения в виде интеграла

$$F(x) = \int_{-\infty}^{x} f(x)dx. \quad (1.3)$$

Так как $F(+\infty) = 1,$ то для плотности распределения выполняется условие

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

Функция распределения и плотность распределения выражаются друг через друга и, следовательно, для непрерывной случайной величины каждая из них является исчерпывающей характеристикой. График плотности распределения $f(x)$ называется кривой распределения, он наглядно представляет вид рассматриваемого распределения.

Закон распределения случайной величины является ее исчерпывающей характеристикой. Однако его не всегда удается установить, и часто используют отдельные числовые характеристики, выражающие некоторые существенные черты распределения.

Различают начальные моменты k-го порядка и центральные моменты k-го порядка.

Начальным моментом первого порядка является математическое ожидание случайной величины X и для непрерывной случайной величины выражается формулой

$$m_x = \int_{-\infty}^{+\infty} x f(x) dx \qquad (1.4).$$

Для дискретной случайной величины формула принимает вид

$$m_x = \sum_i x_i p_i \qquad (1.5),$$

где p_i - вероятность появления значения x_i в ряду наблюдений.

Центральный момент второго порядка – это дисперсия случайной величины, которая определяет меру рассеяния.

Для непрерывной случайной величины она выражается формулой

$$d_x = \int_{-\infty}^{+\infty} (x - m_x)^2 f(x) dx, \qquad (1.6)$$

для дискретной -

$$d_x = \sum_i (x - m_x)^2 p_i \quad . \qquad (1.7)$$

Часто вместо дисперсии используют понятие среднего квадратического отклонения, которое имеет размерность самой случайной величины и равно квадратному корню из дисперсии

$$\sigma_x = \sqrt{d_x} \qquad (1.8)$$

Математическое ожидание оценивают как среднее значение по формуле

$$\hat{m}_x = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (1.8)',$$

а дисперсию – как средний квадрат

$$\hat{d}_x = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{m}_x)^2 \qquad (1.8)''.$$

Кроме центрального момента второго порядка, можно привести примеры центральных моментов третьего и четвертого порядков – ассиметрии и эксцесса.

Они рассчитываются по следующим формулам

$$As = \frac{\mu_3}{\sigma^3}, \qquad (1.9)$$

$$Ex = \frac{\mu_4}{\sigma^4} - 3. \qquad (1.10)$$

В формулах (1.9) и (1.10) As – ассиметрия, а Ex – эксцесс. μ_3, μ_4 – центральные моменты третьего и четвертого порядка соответственно. Ассиметрия является характеристикой симметричности плотности распределения случайной величины относительно математического ожидания, а эксцесс характеризует островершинность кривой плотности распределения.

Между начальными и центральными моментами распределения существуют формулы связи

$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^{\ 3} \qquad (1.10)'$$

$$\mu_4 = m_4 - 4m_3m_4 + 6m_2m_1 - 3m_1^{\ 4} \qquad (1.10)''$$

Две случайные величины, имея одинаковые математическое ожидание и дисперсию, могут обладать различными распределениями вероятностей.

Для случайной величины, которая распределена по нормальному закону, дифференциальный закон распределения имеет вид

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp(-(x - m_x)^2 / 2\sigma_x^{\ 2}). \qquad (1.11)$$

Это, так называемая кривая распределения Гаусса. Она характеризуется только двумя моментами распределения – математическим ожиданием и дисперсией. Эта кривая является симметричной относительно вертикальной прямой, проходящей через точку $x = m$ и имеет максимум в этой точке, равный $\frac{1}{\sigma\sqrt{2\pi}}$. Для кривой Гаусса асимметрия и эксцесс равны нулю. Для многих метеорологических случайных величин данный закон распределения вполне приемлем.

Сложные явления природы чаще всего бывают обусловлены совокупным воздействием ряда различных случайных величин. Например, возникновение грозы связано и с температурой воздуха у земли и на высотах, с влажностью воздуха, с энергетическими характеристиками воздушной массы и т.д. Если рассматривать перечисленные характеристики, как случайные величины, то следует, по всей видимости, исследовать не только законы их распределения, но и законы распределения целой системы случайных величин.

Систему n случайных величин $(X_1, X_2, ..., X_n)$ можно геометрически интерпретировать как $n-$ мерный случайный вектор. В качестве характеристики системы случайных величин используют $n-$мерную функцию распределения $F(x_1, x_2, ..., x_n)$, определяемую как вероятность совместного выполнения n неравенств $X_i < x_i$:

$$F(x_1, x_2, ..., x_n) = P(X_1 < x_1, X_2 < x_2 ..., X_n < x_n). \qquad (1.12)$$

Функция распределения является неубывающей функцией своих аргументов. Из-за невозможности события $X_i < -\infty$, функция распределения стремится к нулю при стремлении хотя бы одного аргумента к $-\infty$.Так как события $X_i < +\infty$ достоверны, то для получения функции распределения подсистемы случайных величин, выделенной из системы, необходимо все аргументы этой подсистемы положить равными $+\infty$.

Плотностью распределения системы или случайного вектора $(X_1, X_2 ..., X_n)$ называется смешанная частная производная от функции распределения, взятая один раз по каждому аргументу, то есть

$$f(x_1, x_{2,...,} x_n) = \frac{\partial^n F(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}, \qquad (1.13)$$

Случайные величины системы называются независимыми, если закон распределения любой ее подсистемы не зависит от того, какие значения приняли остальные величины. Функция распределения системы независимых случайных величин равна произведению функций распределения отдельных случайных величин, входящих в эту систему.

$$F(x_1, x_2, ...x_n) = F_1(x_1) F_2(x_2) ... F_n(x_n) \qquad (1.14)$$

Это условие является не только необходимым для независимых случайных величин системы, но и достаточным. Необходимое и достаточное условие независимости случайных величин можно выразить аналогичным соотношением для плотности распределения

$$f(x_1, x_2, ..., x_n) = f_1(x_1) f(x_2) ... f(x_n) \qquad (1.15)$$

Определим моменты распределения системы двух случайных величин (X,Y) Для дискретной случайной величины имеем

$$m_{k_1 k_2} = \sum \sum_j x_i^k y_j{}^k p_{i\,j} \qquad (1.16), \text{ где}$$

$$p_{ij} = P(X = x_i, Y = y_j);$$