ЛЕКЦИЯ 1

Предмет математической статистики. Статистические модели

Математическая статистика занимается методами сбора, описания и анализа результатов наблюдений массовых случайных явлений, экспериментов

МС базируется на понятиях и методах ТВ

Основная задача ТВ это прогноз: по построенной модели определить вероятность события, распределение СВ итд.

МС решает обратную задачу – выявляет закономерности в случайных данных

В ТВ математические модели случайных явлений опираются на понятии вероятностного пространства $(\Omega, \mathcal{A}, \mathrm{P})$. При этом считается, что вероятность определена

В МС предполагается, что вероятность в модели не известна. Известен лишь класс этих вероятностей $\mathrm{P} \in \mathcal{P}$. Если этот класс задан, то говорят, что задана *статистическая модель*

Предположим, что некоторый эксперимент повторяется n раз и в каждом эксперименте мы измеряем какую-то числовую характеристику. Можно считать, что в результате эксперимента мы наблюдаем значения некоторой СВ $\xi : \Omega \to \mathbb{R}$. Ω будем считать заданным, и оно нас не интересует.

Результатом таких экспериментов будет некоторый набор результатов измерений $X = (X_1, \ldots, X_n)$. Отметим, что величины $X_1, \ldots, X_n$ нельзя считать числами. Ведь повторив эксперимент из n опытов, мы получим совсем другой набор чисел (другие значения СВ ξ)

Это означает, что до проведения конкретного испытания $X_i, i = 1, \ldots, n$ надо считать СВ, с одинаковым распределением ξ, а $X = (X_1, \ldots, X_n)$ – конечной совокупностью случайных величин, характеризующих исход испытания

Результаты измерений, полученные в конкретных экспериментах, будем обозначать $x = (x_1, \ldots, x_n)$ и называть реализацией $X = (X_1, \ldots, X_n)$. $x_i, i = 1, \ldots, n$ представляют собой некоторые числа и СВ не являются

Определение. Совокупность наблюдаемых СВ $X = (X_1, \ldots, X_n)$ называется *выборкой*, величина $X_i, i = 1, \ldots, n$ – элемент выборки, n – объем выборки

Определение. $\mathcal{X} = \{x\}$ – множество всех значений выборки называют *выборочным пространством*

Оно может совпадать с $\mathbb{R}^n$ или быть его частью

Пример (оценивание вероятности)

Пусть многократно повторяется эксперимент, в котором наблюдается некоторое событие A (реализуется схема Бернулли). Мы хотим из случайных данных о наступлении события получить информацию о вероятности события $\mathrm{P}(A) = p$

Очевидно, что можно рассмотреть случайную величину $S_n \in \mathrm{B}_{n,p}$ – число успехов в серии из n проведенных испытаний, имеющую биномиальный закон распределения

Заметим, что $S_n = X_1 + \ldots + X_n$, где СВ $X_i \in \mathrm{B}_p$ имеет распределение Бернулли. При этом X_i независимы и одинаково распределены

Тогда о наборе $X_1, \ldots, X_n$ можно говорить как о выборке, имеющей распределение $\xi \in \mathrm{B}_p$, причем $\mathrm{E}\, X_i = p$

При этом $S_n = X_1 + \ldots + X_n$ некоторая функция от выборки (в дальнейшем мы назовем их статистиками)

Согласно ЗБЧ (т. Бернулли) $\overline{X} = \dfrac{S_n}{n} \underset{n \to \infty}{\overset{\mathrm{P}}{\to}} \mathrm{E}\, X_i = p = \mathrm{P}(A)$

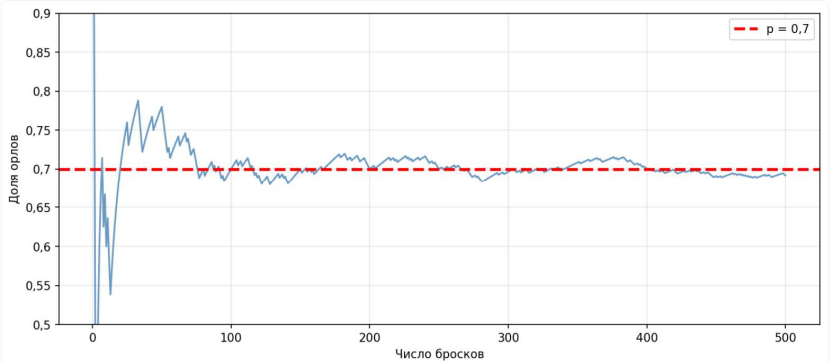

Иллюстрация 1

Замечание. Величину $\overline{X}$, являющуюся средним значением величин $X_1,\ldots,X_n$ логично считать оценкой $\mathrm{P}(A)$

Замечание. Качество такого оценивания мы обсудим позже

Как известно распределение вероятностей случайных величин полностью определяется функцией распределения

$$F(x_1,\ldots,x_n)=\mathrm{P}(X_1<x_1,\ldots,X_n<x_n)$$

Поэтому статистическую модель можно дополнить еще и семейством возможных функций распределения, которому принадлежит неизвестная функция распределения выборки $\mathcal{F}$. Оно, в основном, и определяет статистическую модель

В дальнейшем мы будем рассматривать такие модели экспериментов, в которых компоненты выборки независимые случайные величины с одинаковым законом распределения как у некоторой СВ ξ $F(x)=F_\xi(x)$. Тогда из независимости получаем, что $F(x_1,\ldots,x_n)=F(x_1)\cdot\ldots\cdot F(x_n)$

Определение. Множество значений ξ с функцией распределения $F_\xi(x)$ называют *генеральной совокупностью*, из которой извлекается выборка

Т. о. генеральная совокупность – случайная величина ξ, а выборка – n -мерная СВ, компоненты которой независимы и одинаково распределены как ξ. Такие выборки называют *простыми*

Определение. Если функции распределения класса $\mathcal{F}$ заданы с точностью до параметра $\theta\in\Theta$, то модель называется *параметрической* и обозначается $\mathcal{F}_\theta$

Порядковые статистики и вариационный ряд

Пусть $X=(X_1,\ldots,X_n)$ выборка из распределения F и $x=(x_1,\ldots,x_n)$ ее реализация в некотором эксперименте

Каждой реализации можно поставить в соответствие упорядоченную последовательность $x_1^*\le\ldots\le x_n^*$

Определение. Упорядоченные по возрастанию элементы выборки называют *вариационным рядом реализации*

Соответствующую элементу вариационного ряда случайную величины X_k^* называют k-ой порядковой статистикой. Для них выполняется $X_1^*\le\ldots\le X_n^*$. Эта последовательность называется *вариационным рядом выборки*

Эмпирическая функция распределения

Как уже говорилось, основная задача статистики – это получение некоторой информации о СВ по результатам эксперимента. Поэтому возникает естественный вопрос: а нельзя ли что-то понять про $F_\xi(x)$ по результатам проведенного эксперимента?

Пусть $X=(X_1,\ldots,X_n)$ выборка из генеральной совокупности $F_\xi(x)$

Попробуем построить по ней некоторое приближение этой функции

Определение. Эмпирической частотой $\nu_n(x)$ выборки X, называется случайная функция от x, равная числу элементов выборки, значения которых меньше x

Теорема. В модели $\mathcal{F}_\xi$ эмпирическая частота имеет биномиальное распределение $\mathrm{B}_{n,F_\xi(x)}$

Доказательство:

Рассмотрим испытание, успех которого наступление события $\{\xi<x\}$. Число элементов выборки, значения которых $<x$ равно числу успехов в n независимых испытаниях, т. е. имеет биномиальное распределение в n опытах – это первый параметр распределения.

Поскольку $F_\xi(x)=\mathrm{P}(X<x)$ это вероятность успеха в каждом опыте, то – это второй параметр $\mathrm{B}_{n,F_\xi(x)}$

Замечание. Так же, как и в случае с оценкой вероятности величину $\nu_n(x)$ можно представить в виде суммы независимых СВ с распределением Бернулли $\mathrm{B}_{F_\xi(x)}$. Каждое слагаемое в которой соответствует событию $X_k^*<x$ (X_k^* это k-ый по величине элемент в выборке)

Теорема. В модели $\mathcal{F}_\xi$ $F_{X_k^*}(x)=\sum_{i=k}^{n}C_n^i\left(F_\xi(x)\right)^i\left(1-F_\xi(x)\right)^{n-i}$

Доказательство:

$F_{X_k^*}(x) = \mathrm{P}\left(X_k^* < x\right) = \mathrm{P}\left(\nu_n(x) \geq k\right)$, если k-ый по величине элемент $< x$, то и все предыдущие $< x$. Поэтому как минимум k элементов выборки удовлетворяют условию $X_i < x$

$$\mathrm{P}\left(\nu_n(x) \geq k\right) \underset{\text{по замеч.}}{=} \sum_{i=k}^{n} \mathrm{P}\left(\nu_n(x) = i\right) \underset{\text{по теор}}{=} \sum_{i=k}^{n} C_n^i \left(F_\xi(x)\right)^i \left(1 - F_\xi(x)\right)^{n-i}$$

Определение. Эмпирической функцией распределения $F_n(x)$, соответствующей выборке X, называют случайную функцию от x:

$F_n(x) = \frac{\nu_n}{n}$, где ν_n – число элементов выборки, значения которых меньше x

Замечание. Формула $\frac{\nu_n}{n}$ фактически оценивает вероятность соответствующего события

Замечание. Для каждой реализации $x = (x_1, \ldots, x_n)$ функция $F_n(x)$ однозначно определена, обладает всеми свойствами функции распределения (проверить самостоятельно). Это кусочно постоянная функция, возрастает в точках x_k (скачок). При условии, что все x_k различны, можно записать, что

$$F_n(x) = \begin{cases} 0, x \leq x_1^* \\ \frac{k}{n}, x_k^* < x \leq x_{k+1}^*, k = 1, \ldots, n-1 \\ 1, x > x_n^* \end{cases}$$

Пример

Дана выборка $X = (0,11,2,3,9,2,8,6,3,4,8,7,5,9,4,8,6)$

Построим вариационный ряд, вычислим оценки вероятностей и построим эмпирическую функцию распределения

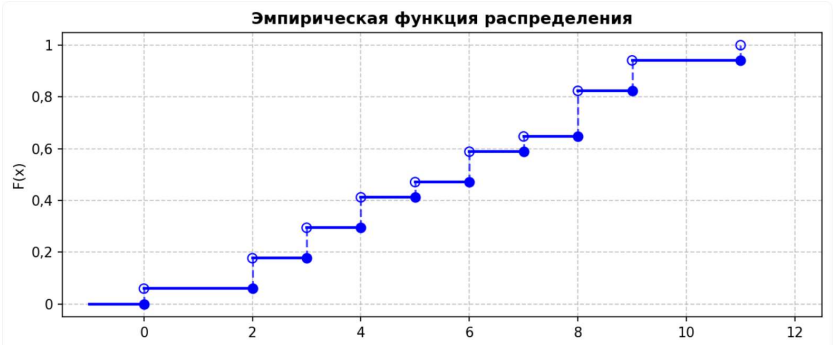

Иллюстрация 2

Свойства эмпирической функции распределения:

1. $\mathrm{E}\, F_n(x) = F_\xi(x)$

$$\mathrm{E}\, F_n(x) = \mathrm{E}\left(\frac{\nu_n(x)}{n}\right) = \frac{\mathrm{E}\,\nu_n(x)}{n} \underset{\text{бином.р. E=}np}{=} \frac{nF_\xi(x)}{n} = F_\xi(x)$$

2. $\mathrm{D}\, F_n(x) = \frac{F_\xi(x)\left(1 - F_\xi(x)\right)}{n}$. Аналогично

$$\mathrm{D} = npq \Rightarrow \mathrm{D}\,\nu_n(x) = nF_\xi(x)\left(1 - F_\xi(x)\right)$$

3. Асимптотическая нормальность

$$\sqrt{n}\left(F_n(x) - F_\xi(x)\right) \sim \mathrm{N}_{0, \sqrt{F_\xi(x)\left(1 - F_\xi(x)\right)}}$$

$F_n(x) = \frac{\nu_n}{n} \Rightarrow \nu_n = nF_n(x)$ по замечанию можно представить в виде суммы независимых СВ с распределением Бернулли $\mathrm{B}_{F_\xi(x)}$. Тогда по ЦПТ справедливо

$$\frac{n\left(F_n(x) - F_\xi(x)\right)}{\sqrt{nF_\xi(x)\left(1 - F_\xi(x)\right)}} \xrightarrow{d} \mathrm{N}_{0,1} \Rightarrow \frac{\sqrt{n}\left(F_n(x) - F_\xi(x)\right)}{\sqrt{F_\xi(x)\left(1 - F_\xi(x)\right)}} \xrightarrow{d} \mathrm{N}_{0,1}$$

4. Сходимость по вероятности $F_n(x) \xrightarrow{p} F_\xi(x)$

Из ЗБЧ в форме Бернулли, т. к. $F_n(x)$ частота события $X < x$, а $F_\xi(x)$ его вероятность

Сформулируем без доказательства еще одну теорему о сходимости эмпирической функции распределения

Теорема. Пусть $F_n(x)$ – эмпирическая функция распределения, построенная по выборке $X=(X_1,\ldots,X_n)$ из распределения $F_\xi(x)$ и $F(x)$ – соответствующая теоретическая функция распределения. Тогда для $\forall x\in\mathbb{R}\ \ \mathrm{P}\left(\lim\limits_{n\to\infty}\mathrm{D}_n=0\right)=1$, где $\mathrm{D}_n(x)=\sup\limits_{x\in\mathbb{R}}\left|F_n(x)-F(x)\right|$

Замечание. Поскольку эмпирическая ФР близка к теоретической, выборочные характеристики можно рассчитывать как обычные характеристики дискретной СВ, принимающей значения $X_1,\ldots,X_n$ с равными вероятностями $\frac{1}{n}$. Чуть позже мы этим воспользуемся

Гистограмма и полигон частот

Как известно, для описания непрерывных СВ используется еще и плотность распределения f_ξ. Попробуем оценить ее по выборке

Пусть дана реализация $x=(x_1,\ldots,x_n)$ выборки из F_ξ, а $x_1^*\le\ldots\le x_n^*$ ее вариационный ряд. Обозначим $x_1^*=a_0, x_n^*=a_k$. Разобьём отрезок $[a_0;a_k]$ на подинтервалы $[a_0;a_1),[a_1;a_2),\ldots,[a_{k-1};a_k]$. Подсчитаем число n_j элементов выборки, попавших в $[a_{j-1};a_j)$ и вычислим частоту попадания. В этом случае говорят, что данные *сгруппированы*

На каждом подинтервале строят прямоугольник, площадь которого равна $\frac{n_j}{n}$, а высота $\frac{n_j}{nh_j}$, где $h_j=a_j-a_{j-1}$

Замечание. Как правило используют разбиение отрезка на интервалы одинаковой длины

Определение. Полученную при построении фигуру называют *гистограммой*

Поскольку площадь прямоугольника равна частоте, то по ЗБЧ (т. Бернулли) сходится по вероятности к вероятности попадания в соответствующий интервал. Если СВ ξ непрерывная с плотностью f_ξ, то огибающая гистограммы является статистическим аналогом теоретической плотности

Точность приближения можно улучшить, если применять на подинтервалах кусочно-линейные функции

Определение. Ломаная линия, проходящая через середины верхних границ прямоугольников гистограммы, называется *полигоном частот*

Замечание. Увеличение числа интервалов улучшает приближение

Пример

Для той же выборке $X=(0,11,2,3,9,2,8,6,3,4,8,7,5,9,4,8,6)$ построим группированную выборку, гистограмму и полигон частот

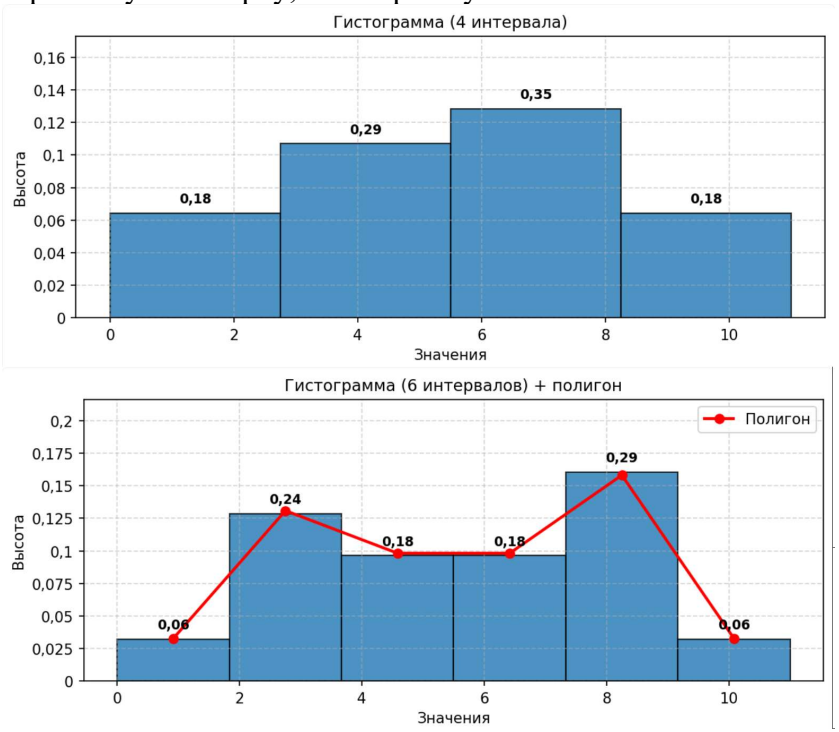

Иллюстрация 3