Замечание Пусть 1) $\vec{X} = (X_1, \dots, X_n)$ – случ. выб. из ген. сов-ти X.
2) $F(t)$ – ф-я распр. с. в. X.

Тогда
$$F_{X_{(1)}}(t) = P\{X_{(1)} < t\} = 1 - P\{X_{(1)} \geq t\} =$$
$$= 1 - P\{X_1 \geq t, X_2 \geq t, \dots, X_n \geq t\} =$$

$$F_{X_{(n)}}(t) = P\{X_{(n)} < t\} = P\{X_1 < t, \dots, X_n < t\} =$$
$$= \{X_i, i = \overline{1;n}, \text{ - незав}\} = P\{X_1 < t\} \cdot \ldots \cdot P\{X_n < t\} =$$
$$= F_{X_1}(t) \cdot \ldots \cdot F_{X_n}(t) = \left\{ \begin{matrix} X_i \sim X \\ F_i = F \end{matrix} \right\} = [F(t)]^n$$

$$= 1 - P\{X_1 \geq t\} \cdot P\{X_2 \geq t\} \cdot \ldots \cdot P\{X_n \geq t\} =$$
$$= 1 - (1 - P\{X_1 < t\}) \cdot \ldots \cdot (1 - P\{X_n < t\}) = \{F_i = F\} =$$
$$= 1 - (1 - F(t))^n$$

15.03.24. Лекция 4.

Среди элементов реализации выборки $\vec{x} = (x_1, \dots, x_n)$ могут ~~найтись~~ быть одинаковые.
Это м.б., когда ген. сов-ть X имеет дискретное распределение или в случае, когда измеряемая величина искажается ограниченным шкалы.
В таком случае не исключено что в реализации отражается одинаковые значений.

Тогда отберем только уникальные значения $\vec{x} = (x_1, \dots, x_n)$ и выпишем соотв. мн-во $z_1, \dots, z_m$, $m \leq n$.

Упорядочим полученное мн-во и получим
$$z_{(1)} < \dots < z_{(m)}$$

Пусть n_i – количество раз, с которым $z_{(i)}$ появляется в реализации $\vec{x} = (x_1, \dots, x_n)$, $i = \overline{1;m}$.

Опр. Тогда таблица

$z_{(1)}$	...	$z_{(m)}$
n_1	...	n_m

наз. статистический рядом $\vec{x} = (x_1, \dots, x_n)$

n_i, $i = \overline{1;m}$, наз. частотой значения $z_{(i)}$.

$\frac{n_i}{n}$ наз. относительной частотой значение $z_{(i)}$, $i = \overline{1;m}$

Замечание Относительно начальной $\vec{x}$ никакие значения не "потеряны" (однако потеряли порядок).
Таблица имеет все необх. сведения об эксперим-те

(3) Эмпирическая функция распределения.

Пусть $\vec{x} = (x_1, \dots, x_n)$ — реализация выборки $\vec{X} = (X_1, \dots, X_n)$.

Обозначим $l(t, \vec{x})$ — число элементов $\vec{x}$, которые меньше t $(t \in \mathbb{R})$.

Опр Эмпирической ф-ей распределения наз. отображение

$$F_n : \mathbb{R} \to \mathbb{R},$$

заданное формулой

$$F_n(t) = \frac{l(t, \vec{x})}{n}.$$

Замечание: 1) Функция F_n обладает всеми св-вами ф-ии распределения.
При этом она явл. кусочно-постоянной и изменяется скачками в точках $x_{(i)}$, $i = \overline{1, n}$.

2) Если все эл-ты $\vec{x}$, т.е. x_k, $k = \overline{1, n}$, различны, то ф-ю распределения можно записать так:

$$F_n(t) = \begin{cases} 0, & t \le x_{(1)}; \\ i/n, & t \in (x_{(i)}; x_{(i+1)}], \quad i = 1, \dots, n \\ 1, & t > x_{(n)} \end{cases}$$

3) Эмпирическая ф. распределения F_n позволяет расс-ть элементы реализации $\vec{x} = (x_1, \dots, x_n)$ как значение некоторой дискретной случ. величины $\tilde{X}$, распределение которой задано таблицей статистического ряда:

$\tilde{X}$	$x_{(1)}$	...	$x_{(n)}$
p	$1/n$	...	$1/n$

Закон распределения $\tilde{X}$ явл. дискретным, при этом он связан с распределением исх. ген. сов-ти X.
Эта связь будет раскрыта ниже.

(4) Выборочная функция распределения

Пусть $\vec{X} = (X_1, \dots, X_n)$ — выборка объема n из ген. сов-ти X.

Опр Выборочной функцией распр-я наз. ф-я:

$$\hat{F}_n(t, \vec{X}) = \frac{l(t, \vec{X})}{n}$$

— случайная ф-я

Замечание: Зафиксируем $t \in \mathbb{R}$ и рассмотрим св-ва данной ф-ии.

1°. Функция $\hat{F}_n(t, \vec{X})$ явл. случайной величиной, которая может принимать значения из ряда $0, 1/n, \ldots, (n-1)/n, 1$.

2°. Обозначим $p = P\{X < t\}$ (*)

В силу определения выборки:

$P\{X_k < t\} = p, \ \forall k = 1, \ldots, n$

$$P\{\hat{F}_n(t, \vec{X}) = \tfrac{k}{n}\} = P\{\ell(t, \vec{x}) = k\} = \left\{\begin{array}{l}\text{схема исп. Бернулли} \\ \text{успех } X_k < t, \\ \text{неудача } X_k \geq t, \\ \text{эксп-т — есть ли } [illegible] \end{array}\right\} = C_n^k \cdot p^k \cdot (1-p)^{n-k}$$

Т.о. для фикс. t $\ell(t, \vec{x})$ расп-на по бин. закону: $\ell(t, \vec{x}) \sim B(n, p)$, а с.в. $\hat{F}_n(t, \vec{X})$ есть подвергнутая линейному преобразованию биномиальная с.в. $B(n,p)$ так, чтобы значения лежали в $[0, 1]$.

Th о сходимости выборочной ф-ии расп-я

Пусть 1) X — ген. совокупность, F_X — ее ф-я расп-я

2) $\hat{F}_n$ — выборочная ф-я распределения X.

Тогда для любого фикс. $t \in \mathbb{R}$

$$\hat{F}_n(t, \vec{X}) \xrightarrow{P} F_X(t)$$

Док-во:

1) $p = F_X(t)$ (в силу данного выше опр-я (*))

2) То, что $\ell(t, \vec{X})$ расп-на по бин. з-ну $B(n, p)$, означает, что $\hat{F}_n(t, \vec{X})$ — относительная частота успеха в серии из n испытаний.

В соотв. с ЗБЧ в форме Бернулли:

$$\hat{F}_n(t, \vec{X}) \xrightarrow{P} p = F_X(t) \quad \text{чтд}$$

в силу произвольности t

Поскольку эмпирическая функция $F_n(t)$ явл. реализацией ф-ии выборочной $\hat{F}_n(t, \vec{X})$, доказанная Th устанавливает связь дискретного распределения с.в. $\tilde{X}$, которое непоср. можно исследовать (получено в экспе.) с ген. совокупностью X.

Th устанавливает вероятностную меру близости данных ф-ий, позволяет принять распределение $\tilde{X}$ за прибл. расп. X.

5 Интервальный статистический ряд.

Если размер выборки велик ($n > 50$), то данные группируют не только в виде статистического ряда, но и в виде интервального статистического ряда.

Для этого выбирают некоторое число $m \in \mathbb{N}$ - количество интервалов, а отрезок $J = [x_{(1)}, x_{(n)}]$ разбивают на m равновеликих промежутков длины

$$\Delta = \frac{x_{(n)} - x_{(1)}}{m}$$

Интервалы определяются равенством:

$$J_i = [x_{(1)} + (i-1)\Delta ; x_{(1)} + i\Delta), \quad i = 1, \dots, m-1,$$

$$J_m = [x_{(1)} + (m-1)\Delta ; x_{(1)} + m\Delta]$$

Опр. Интервальным статистическим рядом называют таблицу:

J_1	...	J_m
n_1	...	n_m

где n_i - число элементов в реализации $\vec{x} = (x_1, \dots, x_n)$, попавших в промежуток J_i, $i = 1, \dots, m$.

! Информация частично потеряна ⇒ важно правильно выбрать m.

! Замечание: 1) Очевидно, что сумма $\sum_{i=1}^{m} n_i = n$

2) При выборе числа m используют формулу $m = [\log_2 n] + 1$

или $m = [3{,}322 \lg(n)] + 1$

Формула носит название "Правило Стерджесса" (1926 г.)

2^N - число всех экспериментов

$2^N = C_N^0 + C_N^1 + \dots + C_N^N$

$16 = 2^4$

⇒ 1, 4, 6, 4, 1 - 5 интервалов, каждый успех в свою "клетку"

т.е. отражается суть ЦПТ

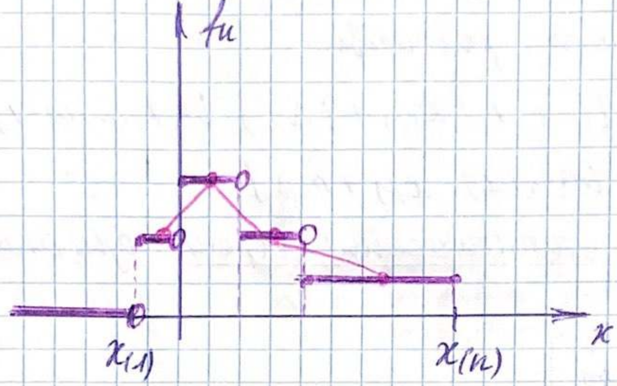
⑥ Эмпирическая плотность

Предположим, что для данной реализации $\vec{x}$ построен интервальный статистический ряд.

Опр. Эмпирической плотностью распределения наз. ф-ю:

$$f_n(t) = \begin{cases} n_i / (n \cdot \Delta), & t \in J_i, \ i = 1, \dots, m \\ 0, & t \notin J \end{cases}$$

Опр. График ф-ии $f_n(t)$ называют гистограммой.



! Замечание: Функция $f_n(t)$ является статистическим аналогом плотности распределения с. в.

- Проверим условие нормировки

$$\int_{-\infty}^{+\infty} f_n(t)\,dt = \int_{-\infty}^{x_{(1)}} f_n(t)\,dt + \int_{x_{(1)}}^{x_{(2)}} f_n(t)\,dt + \dots + \int_{x_{(n)}}^{+\infty} f_n(t)\,dt =$$

$$= \sum_{i=1}^{m} \int_{J_i} f_n(t)\,dt = \sum_{i=1}^{m} \int_{J_i} \frac{n_i}{n\Delta}\,dt = \frac{1}{n} \sum_{i=1}^{m} \frac{n_i}{\Delta} \int_{J_i} dt =$$

$$= \frac{1}{n} \sum_{i=1}^{m} n_i = 1 \text{ - выполняется}$$

- Аналогично тому, как было показано, что $F_n(t, \vec{X}) \xrightarrow{P} F_X(t)$, можно показать, что при больших n $f_n(t) \approx f_X(t)$ – интегралы
ступенчатая
$\int_{J_i} f_n(t)\,dt \to \int_{J_i} f_X(t)\,dt$

⑦ Полигон частот

Предположим, что по $\vec{x}$ построена гистограмма.

Опр. Полигоном частот наз-ют ломаную, звенья которой соединяют середины верхних сторон прямоугольников гистограммы