Элементы математической статистики

Лекция №2

§2. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ВЫБОРОЧНОГО (ЭМПИРИЧЕСКОГО) РАСПРЕДЕЛЕНИЯ

Наиболее распространенными способами графического представления эмпирических данных (выборки) являются ***гистограмма***, ***полигон частот*** и ***эмпирическая функция распределения*** (накопленные относительные частоты).

Пусть $x_{\min}$ и $x_{\max}$ – соответственно наименьшее и наибольшее значения вариант выборки. Величина $R = x_{\max} - x_{\min}$ называется ***размахом*** выборки. Размах делится на число интервалов K (интервальная группировка), которое можно вычислить по одной из следующих формул:

$$K \approx \sqrt{n},$$

$$K < 5 \cdot \lg n$$

$$K \approx 1 + 3,322 \cdot \lg n$$

округлив до целого числа.

Обычно предполагают, что количество интервалов должно удовлетворять условию $5 \le K \le 20$.

Ширина каждого интервала d вычисляется по формуле $d = \frac{R}{K}$. После разбиения на интервалы определяют:

- абсолютные частоты m_i, $i = 1,..,K$, где m_i − количество элементов выборки, попавших в i − й интервал (элемент, попавший на границу интервала, относят к какому-нибудь выбранному интервалу, например, левому, или правому; если на границу интервала попадает много элементов выборки, то их делят пополам между левым и правым интервалами);
- относительные частоты $h_i = \frac{m_i}{n}$;

- относительные накопленные частоты $\sum_{j=1}^{i}\frac{m_j}{n}=\sum_{j=1}^{i}h_j, i=1,\dots,K;;$
- середины интервалов u_i.

Все полученные результаты сводятся в таблицу.

Номер интервала	Границы интервалов	m_i	h_i	$\sum_{j=1}^{i}h_j$	u_i
1		m_1	h_1	h_1	z_1
2		m_2	h_2	h_1+h_2	z_2
…		…	…	…	…

При этом $\sum_{i=1}^{K}m_i=n, \sum_{i=1}^{K}h_i=1$.

2.1. Гистограмма и полигон частот

Гистограмма строится следующим образом. На оси абсцисс откладываются интервалы, и на каждом из них строится прямоугольник, площадь которого равна относительной частоте, соответствующей этому интервалу, т.е. высота прямоугольника (ордината) равна $\frac{h_i}{d} = \frac{m_i}{nd}$, так что полная площадь гистограммы равна 1. (Гистограмма является эмпирическим аналогом плотности распределения). Так как множители $\frac{1}{nd}$ (или $\frac{1}{d}$) можно рассматривать как масштабные, то по оси ординат можно откладывать частоты m_i или h_i (правда, в этом случае площадь всех прямоугольников будет равна nd или d).

Полигон частот - ломаная линия, которая получается, если из середины каждого интервала u_i восстановить перпендикуляр высотой h_i (или m_i) и соединить вершины этих перпендикуляров. Полигон частот чаще используют при дискретной группировке.

2.2. Эмпирическая функция распределения

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

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. Кроме того, эмпирическая функция распределения используется для определения эмпирических (выборочных) квантилей. **Квантилем** порядка p $(0 < p < 1)$ называется величина x_p, определяемая из соотношения $P\{\xi < x_p\} = F(x_p) = p$. **Выборочный квантиль** x_p определяется из соотношения $F_n(x_p) = p$. Задавая p, можно по графику эмпирической функции распределения оценить, например, значение x_p, которое исследуемая величина не превзойдет с вероятностью p, либо, задавая x_p, по тому же графику оценить соответствующую вероятность p. По $F_n(x)$ можно оценить медиану из соотношения $F_n(\hat{M}_e) = 0{,}5$.

Пример 1. Построение гистограммы, полигона частот и эмпирической функции распределения по данным примера 4 из §1.

$$x_{\min} = 133,5,$$

$$x_{\max} = 147,0,$$

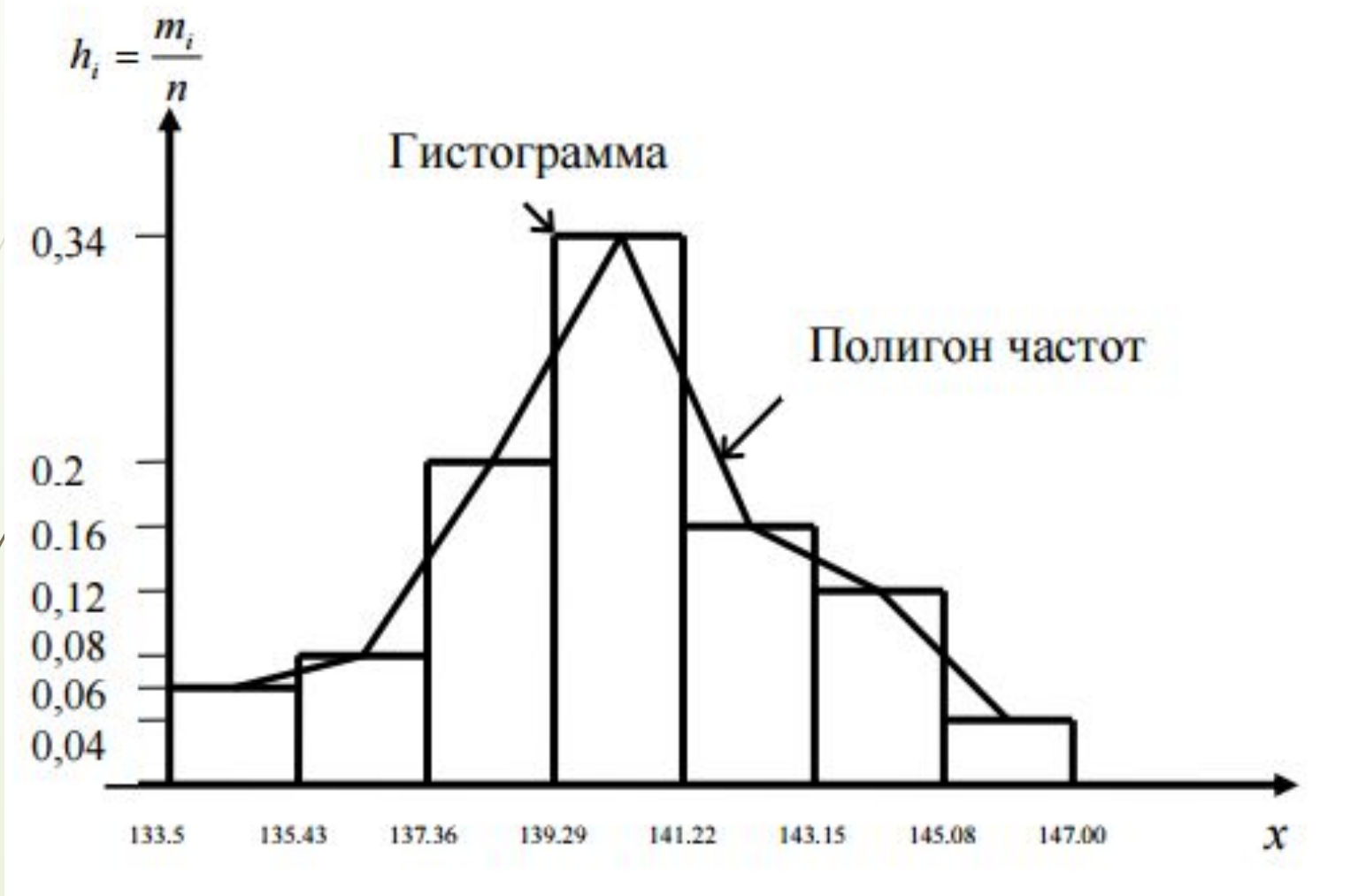
$$R = 13,5,$$

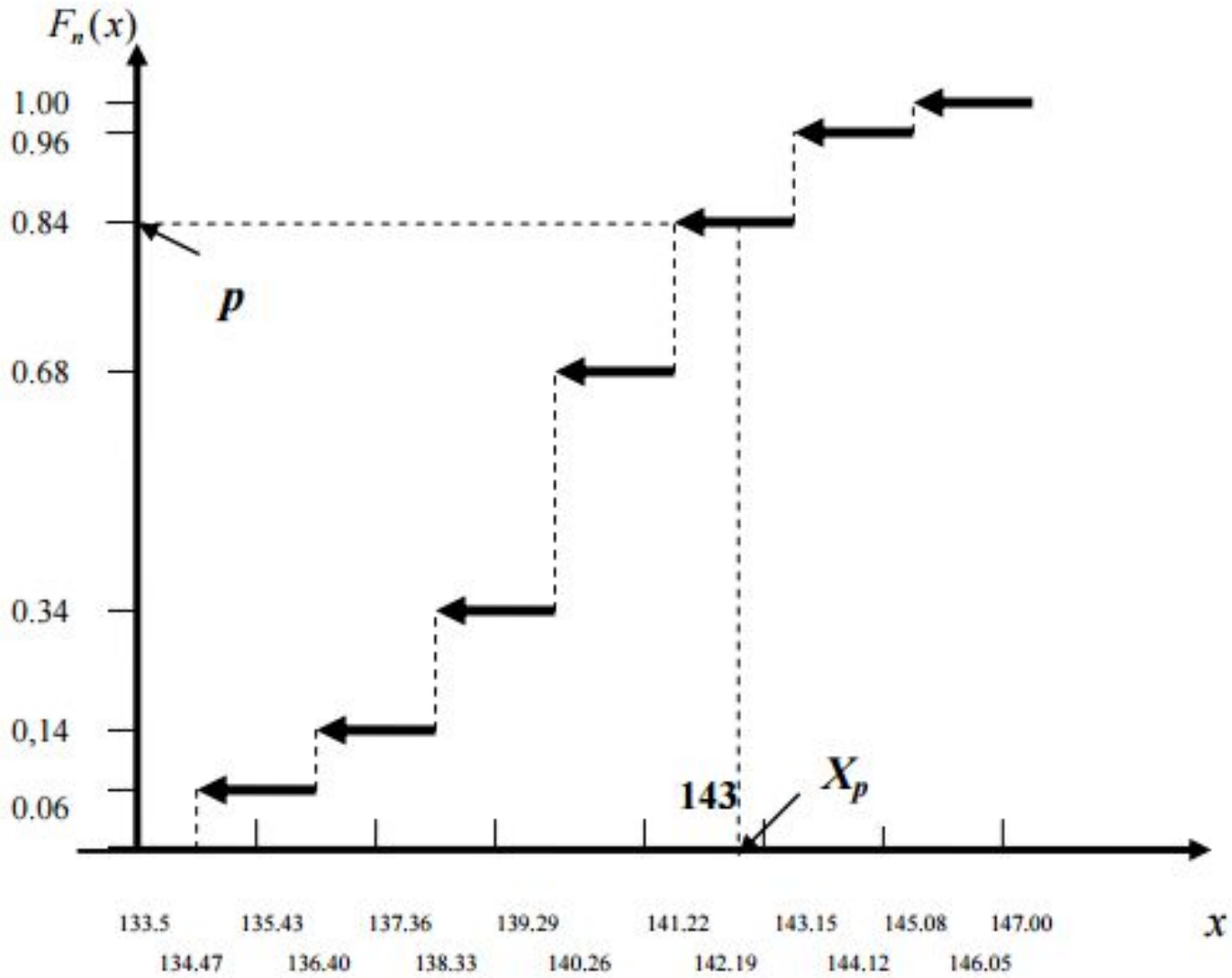
$$n = 50.$$

Количество интервалов: $K \approx \sqrt{n},\ K = 7$.

Ширина интервала: $d = \dfrac{R}{K} = 1,93$.

Номер интервала	Границы	m_i	$h_i = m_i / n$	$\sum_{j=1}^{i} h_j$	u_i
1	133,50 – 135,43	3	0,06	0,06	134,47
2	135,43 – 137,36	4	0,08	0,14	136,40
3	137,36 – 139,29	10	0,20	0,34	138,33
4	139,29 – 141,22	17	0,34	0,68	140,26
5	141,22 – 143,15	8	0,16	0,84	142,19
6	143,15 – 145,08	6	0,12	0,96	144,12
7	145,08 – 147,00	2	0,04	1	146,05


$h_i = \frac{m_i}{n}$
Гистограмма
Полигон частот
0,34
0.2
0.16
0,12
0,08
0,06
0,04
133.5
135.43
137.36
139.29
141.22
143.15
145.08
147.00
x



Эмпирическая функция распределения

2.3. Задачи

2.1. Построить гистограмму и эмпирическую функцию распределения по данным задачи 1.3. Оценить вероятность того, что скорость превысит 80 км/час.

2.2. Построить полигон частот и эмпирическую функцию распределения для распределения 45 пар мужской обуви, проданных магазином за день:

39, 41, 40, 42, 41, 40, 42, 44, 40, 43, 42, 41, 43, 39, 42, 41, 42, 39, 41, 37, 43, 41, 38, 43, 42, 41, 40, 41, 38, 44, 40, 39, 41, 40, 42, 40, 41, 42, 40, 43, 38, 39, 41, 41, 42.

Оценить по эмпирической функции распределения медиану.

2.3. Через каждый час измерялось напряжение в электросети. При этом были получены следующие значения (в вольтах):

227, 219, 215, 230, 232, 223, 220, 222, 218, 219, 222, 221, 227, 226, 226, 209, 211, 215, 218, 220, 216, 220, 221, 225, 224, 212, 217, 219, 220.

Построить гистограмму, полигон частот, эмпирическую функцию распределения; оценить вероятность того, что напряжение не превосходит 220 В.