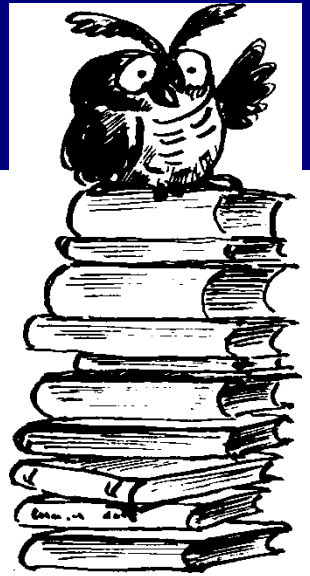
Математическая статистика



Выборка

Опр.1 Пусть X **одномерная с.в.**, а $x_1, x_2, \ldots, x_n$ значения с.в.,полученные в результате независимых испытаний. Будем называть полученные значения **выборкой из генеральной совокупности** X или **вариантами из** X.

Число n полученных значений называется **объёмом выборки**.

- **Пример 1.** Выработка у 20 рабочих за некоторый интервал времени составила (в %% от нормы):

  106, 102, 112, 102, 104,

  105, 108, 109, 97, 110,

  111, 105,103, 112, 113,

  106, 115, 118, 107, 105.

Опр. 2.
Упорядоченный (ранжированный) ряд вариант называется **вариационным рядом.**

Пример 2. Вариационный ряд выработки рабочих:

x_{min}= 97, 102, 102,103, 104, 105, 105, 105, 106, 106,
107, 108, 109, 110, 111, 112, 112, 113, 115, 118 = x_{ax}

Частота – количество одинаковых выборочных значений .
Относительная частота – частота, делённая на объём выборки.

$$w_i = n_i / n$$

Опр. 3 Статистический ряд – ряд вариант, **упорядоченный** в порядке возрастания или убывания, с **соответствующими частотами** или **относительными частотами.**

Обычно статистический ряд записывается в виде **таблицы.**

Пример 3. Статистический ряд отметок в группах потока ТН-17-7-8 на экзамене по математике в 3-м семестре:

Отметка x_i	2	3	4	5	Σ
частота n_i	8	22	15	5	50
Относительная частота $w_i = n_i / n$	0,16	0,44	0,3	0,1	1

Оценки числовых характеристик

Точечное оценивание

Оценка математического ожидания

- Пусть $m = MX$ математическое ожидание случайной величины X. В качестве оценки математического ожидания по выборочным значениям $x_1, x_2, \ldots, x_n$

- можно взять **выборочное среднее**

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx m.$$

- В случае дискретной **группировки вариант** -

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{l} n_i x_i \approx m, \quad n_1 + \ldots + n_l = n.$$

Оценка дисперсии

- В качестве оценки дисперсии $\sigma^2 = DX$ аналогично можно взять величину

$$s_b^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \approx \sigma^2.$$

выборочную дисперсию

Качество оценок. **Несмещённость выборочного среднего**

- Будем рассматривать выборочные значения как независимые случайные величины $x_1, x_2, \ldots, x_n$, имеющие одинаковое распределение со с.в X
- Вычислим математическое ожидание

$$M\,\bar{x} = M\,\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} Mx_i = \frac{nm}{n} = m.$$

- Полученное равенство $M\,\bar{x} = m$ называется свойством **несмещённости** выборочного среднего.

Дисперсия выборочного среднего

- Имеем $M(x_i - m)(x_j - m) = 0,\ i \neq j$

  В силу независимости наблюдаемых величин x_i
- Отсюда

$$D\bar{x} = M(\bar{x} - m)^2 = M\left(\frac{1}{n}\sum_{i=1}^{n} x_i - m\right)^2 =$$

$$= \frac{1}{n^2} M\left(\sum_{i=1}^{n}(x_i - m)\right)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} M(x_i - m)^2\right) =$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \Rightarrow \quad \boxed{D\bar{x} = \frac{\sigma^2}{n}, \quad \sigma[\bar{x}] = \frac{\sigma}{\sqrt{n}}}$$

Смещённость оценки выборочной дисперсии

- Справедливы равенства

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-m+m-\bar{x})^2=$$

$$=\sum_{i=1}^{n}(x_i-m)^2+2(\bar{x}-m)\sum_{i=1}^{n}(x_i-m)+$$

$$+\sum_{i=1}^{n}(\bar{x}-m)^2=\sum_{i=1}^{n}(x_i-m)^2+2(\bar{x}-m)n(\bar{x}-m)+$$

$$+n(\bar{x}-m)^2=\sum_{i=1}^{n}(x_i-m)^2-n(\bar{x}-m)^2$$

Смещённость оценки s_b^2

- Отсюда

$$Ms_b^2 = M\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 =$$

$$= \frac{1}{n}M\left(\sum_{i=1}^{n}(x_i - m)^2 - n(\bar{x} - m)^2\right) =$$

$$= \frac{n\sigma^2}{n} - \frac{\sigma^2}{n} = \frac{(n-1)\sigma^2}{n}$$

- Т.е.

$$\boxed{Ms_b^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2}$$

Несмещённая оценка дисперсии

- Для получения несмещенной оценки дисперсии следует взять **выборочную дисперсию** в виде

$$s^2 = \frac{n}{n-1} s_b^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 =$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

- Итак, **несмещенная выборочная дисперсия** имеет вид

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Формулы для вычисления выборочной дисперсии

- Справедливо равенство

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2\right).$$

- В случае дискретной группировки

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{l} n_i x_i^{\,2} - n\bar{x}^2\right), \quad n_1 + \ldots + n_l = n.$$

Состоятельность выборочного среднего

- Оценка $\hat{\theta}_n$ параметра θ, найденная по выборке объёма n называется состоятельной оценкой, если

$$\lim_{n\to\infty} P(|\theta_n - \theta| < \varepsilon) = 1.$$

- Применим **теорему Чебышёва к выборочной средней** $\eta_n = \bar{x}$

$$\lim_{n\to\infty} P(|\bar{x} - m| \geq \varepsilon) = 0 \Rightarrow \lim_{n\to\infty} P(|\bar{x} - m| < \varepsilon) = 1.$$

- Таким образом выборочное среднее является **состоятельной оценкой.**

- Состоятельность выборочной дисперсии. Так как

$$Ms^2 = \sigma^2,$$

то применив второе неравенство Чебышёва, получаем оценку

$$P(|s^2 - Ms^2| \geq \varepsilon) = P(|s^2 - \sigma^2| \geq \varepsilon) \leq \frac{Ds^2}{\varepsilon^2}.$$

- Можно показать, что дисперсия выборочной дисперсии $Ds^2 \to 0, \quad n \to \infty.$

(индекс n в последовательности $s^2 = s_n^2$ опущен)

- Отсюда следует

$$P(|s^2 - \sigma^2| \geq \varepsilon) \leq \frac{Ds^2}{\varepsilon^2} \to 0, \quad n \to \infty,$$

т.е. **состоятельность оценки дисперсии**.

Эффективность оценок

- Оценка $\hat{\theta}_n$ называется эффективной оценкой в некотором классе оценок Θ, если при любом фиксированном n оценка имеет минимальную дисперсию в классе оценок Θ .
- Можно доказать, что среди линейных оценок

$$\hat{m} = \sum_{i=1}^{n} c_i x_i$$

- математического ожидания, дисперсия $D\hat{m}$ принимает минимальное значение при $c_i = \frac{1}{n}$, т.е. выборочная средняя $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

является эффективной оценкой в классе линейных оценок.

Интервальное оценивание

Определение

- Пусть $\boldsymbol{X}$ случайная величина, а θ неизвестная числовая характеристика случайной величины. **Случайный** интервал $(\boldsymbol{z}_1, \boldsymbol{z}_2)$, который с заданной вероятностью $\gamma = \mathbf{1} - \alpha$ покрывает параметр θ называется **доверительным интервалом с уровнем доверия (надёжностью)**

$$\gamma = \mathbf{1} - \alpha.$$

- Параметр α называется **уровнем значимости.**
- Рассмотрим задачу построения доверительного интервала для математического ожидания.

Нормирование выборочного среднего

- Рассмотрим выборку из генеральной совокупности такой что математическое ожидание $\boldsymbol{a}$, дисперсия σ^2 Тогда, выборочное среднее $\bar{\boldsymbol{x}}$ по центральной предельной теореме имеет приближённо нормальное распределение с параметрами $\boldsymbol{a}, \sigma / \sqrt{\boldsymbol{n}}$. Поэтому случайная величина

$$\boldsymbol{U} = \frac{(\bar{\boldsymbol{x}} - \boldsymbol{a})}{\sigma / \sqrt{\boldsymbol{n}}}$$

имеет приближённо нормальное распределение $\boldsymbol{N}(\mathbf{0},\mathbf{1})$.

Доверительный интервал для математического ожидания при **известной дисперсии**

- По заданной вероятности $\gamma = \mathbf{1} - \alpha$ найдём δ такое, что справедливо неравенство

$$\boldsymbol{P}(\left|\bar{\boldsymbol{x}} - \boldsymbol{a}\right| < \delta) = \mathbf{1} - \alpha,$$

- или равносильное

$$\boldsymbol{P}(\left|\frac{\bar{\boldsymbol{x}} - \boldsymbol{a}}{\sigma / \sqrt{\boldsymbol{n}}}\right| < \frac{\delta}{\sigma / \sqrt{\boldsymbol{n}}}) = \mathbf{1} - \alpha.$$

- Т.е. $\boldsymbol{P}(\left|\boldsymbol{U}\right| < \frac{\delta}{\sigma / \sqrt{\boldsymbol{n}}}) = \mathbf{1} - \alpha \Rightarrow \mathbf{2}\boldsymbol{\Phi}(\frac{\delta}{\sigma / \sqrt{\boldsymbol{n}}}) = \mathbf{1} - \alpha.$

Доверительный интервал для математического ожидания при **известной дисперсии**

- Пусть $u_{\frac{1-\alpha}{2}}$ такое, что $2\Phi(u_{\frac{1-\alpha}{2}}) = 1-\alpha.$
- Тогда $u_{\frac{1-\alpha}{2}} = \dfrac{\delta}{\sigma/\sqrt{n}}.$ Откуда $\delta = u_{\frac{1-\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}.$
- Итак, получаем неравенство с вероятностью $\gamma = 1-\alpha$

$$\left|\bar{x} - a\right| < u_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

- Или равносильное

$$\boxed{\bar{x} - u_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}} < a < \bar{x} + u_{\frac{1-\alpha}{2}} \frac{\sigma}{\sqrt{n}},}$$

- которое определяет **доверительный интервал для** a.

Доверительный интервал для математического ожидания при **неизвестной дисперсии**

- Рассматривается случайная величина $t(n-1)$:

$$t(n-1) = \frac{(\bar{x} - a)}{s/\sqrt{n}}$$

- распределение которой, имеет **распределение Стьюдента с** $n-1$ **степенью свободы**. Критические значения $t_{n,1-\alpha}$ такие , что

$$P(|t(n)| < t_{n,1-\alpha}) = 1 - \alpha$$

- можно найти в соответствующей **таблице**.

Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,95
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Доверительный интервал для математического ожидания при **неизвестной дисперсии**

- Пусть $t_{n-1,1-\alpha}$ такое, что $P(|t(n-1)| < t_{n-1,1-\alpha}) = 1-\alpha.$

Тогда

$$P\left(\left|\frac{\bar{x}-a}{s/\sqrt{n}}\right| < t_{n-1,1-\alpha}\right) = 1-\alpha.$$

Решая неравенство $\left|\frac{\bar{x}-a}{s/\sqrt{n}}\right| < t_{n-1,1-\alpha}$ относительно a,

получаем **доверительный интервал для** a, **надёжностью** $\gamma = 1-\alpha$

$$\bar{x} - t_{n-1,1-\alpha}\frac{s}{\sqrt{n}} < a < \bar{x} + t_{n-1,1-\alpha}\frac{s}{\sqrt{n}},$$

Замечание

- Распределение Стьюдента при $n > 30$ близко к стандартному нормальному распределению. Поэтому равносильны равенства

$$P\left(\left|\frac{\bar{x}-a}{s/\sqrt{n}}\right| < \frac{\delta}{s/\sqrt{n}}\right) = 1-\alpha. \qquad 2\Phi\left(\frac{\delta}{s/\sqrt{n}}\right) = 1-\alpha.$$

- Отсюда $u_{\frac{1-\alpha}{2}} = \dfrac{\delta}{s/\sqrt{n}}$, где $\Phi(u_{\frac{1-\alpha}{2}}) = \dfrac{1-\alpha}{2}$.

- Доверительный интервал имеет вид

$$\bar{x} - u_{\frac{1-\alpha}{2}}\frac{s}{\sqrt{n}} < a < \bar{x} + u_{\frac{1-\alpha}{2}}\frac{s}{\sqrt{n}}.$$

Пример 4

- Найти доверительный интервал средней выработки рабочих в примере 1 **с уровнем доверия** $\gamma = \mathbf{1} - \alpha = \mathbf{0.9}$ ($\alpha = \mathbf{0,1}$ –**уровнем значимости**).
- Вычислим $\bar{x} = \mathbf{107,5}; \quad s = \mathbf{5}$
- Найдём по таблице $t_{\mathbf{19;0,9}} \approx \mathbf{1,73}$
- Доверительный интервал имеет вид

$$\mathbf{107,5} - \mathbf{1,73}\frac{\mathbf{5}}{\sqrt{\mathbf{20}}} < a < \mathbf{107,5} + \mathbf{1,73}\frac{\mathbf{5}}{\sqrt{\mathbf{20}}}.$$

- Итак, с вероятностью 0,9 получаем интервал

$$\mathbf{105,6} < a < \mathbf{109,4}.$$

Замечание

- Если в примере 4 для нахождения доверительного интервала использовать вместо распределения Стьюдента стандартное нормальное распределение, т.е. использовать формулу

$$\bar{x} - u_{\frac{1-\alpha}{2}} \frac{s}{\sqrt{n}} < a < \bar{x} + u_{\frac{1-\alpha}{2}} \frac{s}{\sqrt{n}},$$

где $u_{\frac{0,9}{2}} = 1,64.$

- Доверительный интервал получается несколько **уже**

$$105,7 < a < 109,3.$$

Графическое изображение вариационных рядов.

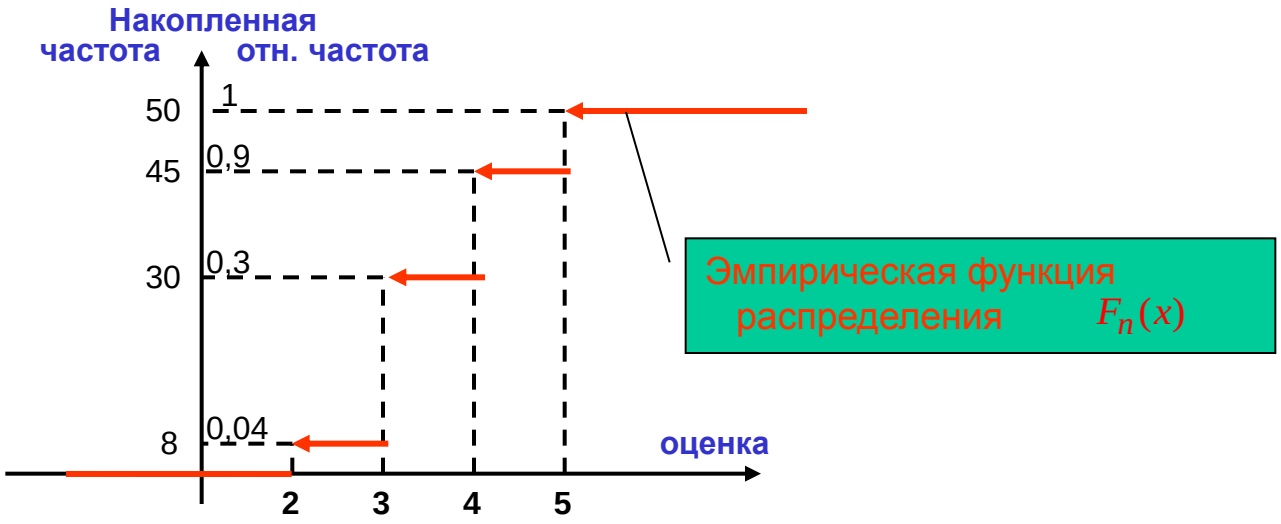
Эмпирическая функция распределения

- Рассмотрим задачу нахождения функции распределения с.в. X по полученной выборке $x_1, x_2, \ldots, x_n$.

  Обозначим через $\mu_n(x)$ **число** вариант **меньших** x
- Опр. **Эмпирической функцией распределения** $F_n(x)$ называется отношение

$$F_n(x) = \frac{\mu_n(x)}{n}.$$

  т.е это отношение числа значений признака, попавших в интервал $(-\infty, x)$, к объёму выборки.
- Величина $\mu_n(x)$ **называется накопленной частотой**.
- Замечание . Для дискретного ряда эмпирическая функция распределения – разрывная ступенчатая функция (как у дискретной случайной величины).



эмпирическая функция распределения оценок студентов в примере 3.

Отметка	2	3	4	5	Σ
частота	8	22	15	5	50

Теорема Колмогорова - Гливенко

- **Теорема.** Если функция распределения $F(x)$ с.в. X непрерывна, то при $n \to \infty$ справедливо равенство

$$\lim_{n\to\infty} P(\sqrt{n}(\max|F_n(x) - F(x)|) < z) = K(z),$$

где

$$K(z) = \begin{cases} 0, & z \le 0, \\ \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, & z > 0. \end{cases}$$

- **функция Колмогорова**, значения которой табулированы.

Следствие

- Теорема Колмогорова – Гливенко позволяет указать с заданной вероятностью γ **точность приближения функции распределения эмпирической функцией**. Действительно, если z_γ -такое значение аргумента функции $K(z)$, что $K(z_\gamma) = \gamma,$ то

$$P(\sqrt{n}(|F_n(x) - F(x)|) < z_\gamma) \approx \gamma$$

- при $n \to \infty$. Таким образом при **больших** n с вероятностью γ справедлива оценка

$$|F_n(x) - F(x)| < \frac{z_\gamma}{\sqrt{n}}.$$

Группировка вариант (разбиение по интервалам).

Если объём выбоки велик, то вариационный ряд преобразуется группировкой вариант в интервалы **– интервальный ряд**.

Формула Стерджеса*. Рекомендуемое число *m* интервалов разбиения

$$m \approx 1 + \log_2 n \approx 1 + 3{,}322 \cdot \lg n, \qquad \text{либо} \qquad m = \sqrt{n}$$

(*n* – общее число наблюдений)

Замечание . **Левую границу** интервала будем включать в интервал разбиения.

Результат группировки представляют рядом **интервалов вариант и соответствующих частот.**

**Herbert Arthur Sturges*, 1882—1958

Пример 4

- Рассмотрим ранжированный ряд выработки рабочих (Пример 2), n=20 . Имеем по формуле Стерджеса

$$\boldsymbol{m} = \mathbf{1} + \log_2 \mathbf{20} \approx \mathbf{1} + \mathbf{3,32} \lg \mathbf{20} \approx \mathbf{1} + \mathbf{4,32} = \mathbf{5,32}$$

- Вычислим длину интервала

$$\boldsymbol{h} = \frac{\boldsymbol{x}_{\max} - \boldsymbol{x}_{\min}}{\boldsymbol{m}} = \frac{\mathbf{118} - \mathbf{97}}{\mathbf{5,32}} \approx \mathbf{4}$$

- *Для удобства вычислений* **расширим интервал**, содержащий выборку, положив

$$\frac{\boldsymbol{h}}{\mathbf{2}} \le \boldsymbol{a} = \mathbf{96} \le \boldsymbol{x}_{\min}, \quad \boldsymbol{b} = \mathbf{120} > \boldsymbol{x}_{\max}.$$

Интервальная таблица (интервальный ряд) примера 2

номер интервала, i	выработка x	частота (кол-во рабочих), n_i	Относительная частота $w_i = n_i/n$	Накопленная (кумулятивная) частота $\mu_n(x)$	Накопленная (кумулятивная) относительная частота $\nu_n(x)$
1	[96 – 100)	1	0,05	1	0,05
2	[100 – 104)	3	0,15	4	0,2
3	[104 – 108)	7	0,35	11	0,6
4	[108 – 112)	4	0,2	15	0,75
5	[112 – 116)	4	0,2	19	0,95
6	[116 – 120)	1	0,05	20	1
	$\sum$	**20**	**1,0**		

xmin= 97, 102, 102,103, 104, 105, 105, 105, 106, 106, 107, 108, 109, 110, 111, 112, 112, 113, 115, 118 =x*max*

Графическое изображение вариационных рядов.

Гистограмма – ступенчатая фигура, в основании которой лежат интервалы значений X, а высоты равны соответствующим частотам n_i/h (или относительным частотам w_i/h). Где h длина интервала.

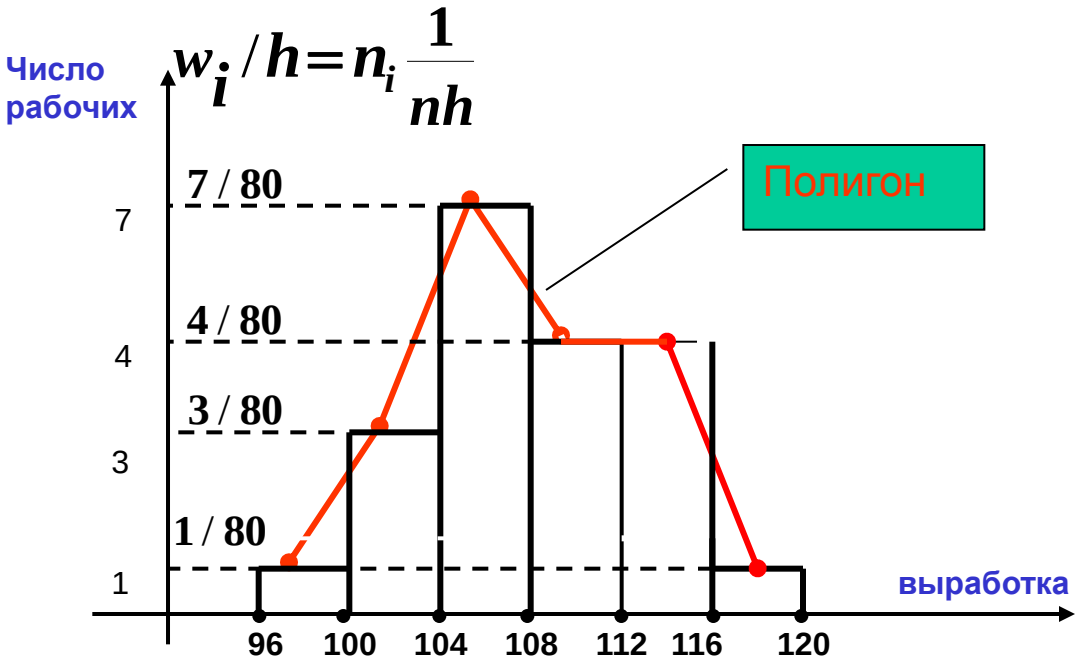
Гистограмма даёт представление о плотности распределения с.в. X .

Действительно, относительная частота w_i **значений, попавших в полуинтервал** $[z_i, z_{i+1})$ равна $w_i = F_n(z_{i+1}) - F_n(z_i)$

А в силу теоремы Колмогорова-Гливенко $F_n(z_i) \approx F(z_i)$ при больших n

$$w_i \approx F(z_{i+1}) - F(z_i) = \int_{z_i}^{z_{i+1}} p(x)dx = p(\xi_i)(z_i - z_{i+1}) = p(\xi_i)h,$$

Т.е. $$p(\xi_i) = \frac{w_i}{h}.$$



Гистограмма относительных частот распределения рабочих по выработке в примере 1.

Замечание 1. Для интервального вариационного ряда полигон получается при соединении **середин** оснований прямоугольников.

Замечание 2. **Площадь фигуры, ограниченная гистограммой относительных частот равна 1.**

Кумулятивная кривая – кривая накопленных относительных частот

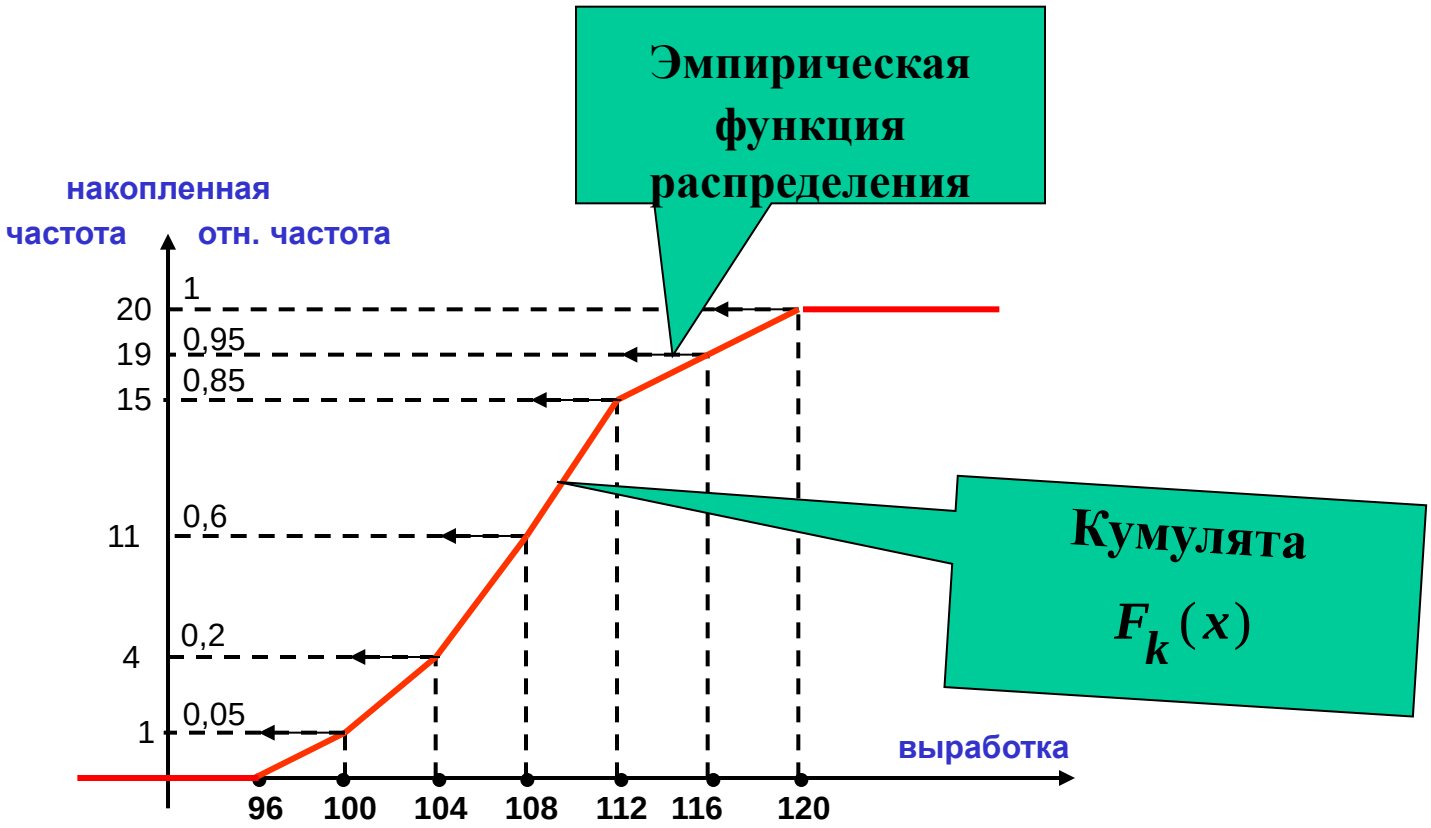
- для интервального ряда – ломаная, соединяющая точки, абсциссы которых отвечают концам интервалов, а ординаты равны накопленным частотам

$$\nu_n(x_i) = \frac{\mu_n(x_i)}{n}$$

Пусть $y = z(x)$ - уравнение гистограммы относительных частот, тогда функция

$$F_k(x) = \int_{-\infty}^{x} z(x)dx$$

является кумулятой (кумулятивной кривой) .



Кумулята и эмпирическая функция распределения рабочих по выработке в примере 1.

Задание по статистике

- Найти несмещённые оценки математического ожидания и дисперсии.
- Найти интервальные оценки математического ожидания с уровнем доверия $\gamma = 0{,}95$.
- Сгруппировать данные , подобрав длину интервала. Построить интервальную таблицу (интервальный ряд)
- Построить гистограмму относительных частот и кумулятивную кривую. Сделать вывод о законе распределения генеральной совокупности.