1 Гистограмма. Доверительные интервалы.

1.1 Введение

Всякое каким-то образом выделенное множество объектов, которые могут отличаться друг от друга значением некоторой определенной характеристики (признака) $\boldsymbol{X}$, называется *генеральной совокупностью*. В математической статистике понятие генеральной совокупности трактуется как *совокупность всех мыслимых наблюдений, которые могут быть произведены при данном реальном комплексе условий*, и в этом смысле его не надо смешивать с реальными совокупностями, подлежащими статистическому изучению. Понятие генеральной совокупности аналогично понятию *случайной величины* $\boldsymbol{X}$. Число элементов генеральной совокупности называется её *объемом*.

Часть генеральной совокупности $\{X_1, X_2, ..., X_n\}$, случайным образом отобранная для наблюдений, называется случайной выборкой или, для краткости, *выборкой*. Выборку можно рассматривать как некоторый эмпирический аналог генеральной совокупности. Элементы выборки можно считать независимыми одинаково распределёнными случайными величинами X_i, поскольку они являются результатом проведения последовательности независимых испытаний с одной и той же случайной величиной $\boldsymbol{X}$.

Будем считать, что любой из элементов генеральной совокупности обладает равной возможностью быть отобранным в выборку. Элементы выборки можно считать независимыми случайными величинами. На практике исследователь работает с конкретной реализацией выборки $\{x_1, x_2, \ldots, x_n\}$, где x_i являются значениями случайных величин X_i, распределение которых совпадает с распределением признака $\boldsymbol{X}$. Число элементов выборки n называется ее объемом, а конкретные значения реализации выборки x_i – вариантами. Расположив варианты в порядке возрастания, получим *вариационный ряд*.

1.2 Выборочное среднее и выборочная дисперсия

По результатам наблюдений над выборкой можно вычислить точечные оценки неизвестных параметров распределения признака $\boldsymbol{X}$. Для неизвестного математического ожидания $E(\boldsymbol{X})$ (генеральное среднее) вычисляется точечная оценка – *выборочное среднее*. Для неизвестной

дисперсии $D(\boldsymbol{X})$ (генеральная дисперсия) вычисляется точечная оценка – *выборочная или исправленная выборочная дисперсия.* В соответствии с требованиями математической статистики эти оценки должны удовлетворять ряду критериев, основными из которых являются требования *состоятельности и несмещенности.*

Теорема 1.1. *Оценка $\hat{\Theta}_n$, вычисляемая при помощи выборки объёмом n, называется состоятельной оценкой параметра Θ, если*

$$\lim_{n\to\infty} P(|\hat{\Theta}_n - \Theta| > \epsilon) = 0,$$

для любого $\epsilon > 0$. Другими словами, вероятность отклонения оценки от истинного значения параметра можно сделать сколь угодно малой, увеличивая объем выборки.

Теорема 1.2. *Оценка $\hat{\Theta}_n$ называется несмещенной оценкой параметра Θ, если $E(\hat{\Theta}_n) = \Theta$ при любом n, т.е. отклонение $\hat{\Theta}_n$ от Θ не содержит систематической ошибки.*

Можно доказать, что выборочное среднее

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

и исправленная выборочная дисперсия

$$\bar{S_n^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2, \tag{2}$$

являются состоятельными и несмещенными оценками $E(\boldsymbol{X})$ и $D(\boldsymbol{X})$, соответственно.

Для выборочного среднего:

$$E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = E(\boldsymbol{X}).$$

Теперь покажем необходимость использования исправленной выборочной дисперсии $\bar{S_n^2}$. Для этого сначала вычислим математическое ожидание выборочной дисперсии

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

Для упрощения введем обозначение $\mu = E(x)$. Тогда

$$E(S_n^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}((x_i - \mu) - (\bar{X} - \mu)^2\right) =$$
$$= \frac{1}{n}E\left(\sum_{i=1}^{n}((x_i - \mu)^2 - 2(x_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\right) =$$
$$\frac{1}{n}E\left(\sum_{i=1}^{n}(x_i - \mu)^2\right) - 2E\left((\bar{X} - \mu)\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\right) + E\left(\frac{1}{n}\sum_{i=1}^{n}(\bar{X} - \mu)^2\right).$$

Заметим, что первое слагаемое является дисперсией случайной величины $D(\boldsymbol{X}) = \sigma^2$, т.е.:

$$\frac{1}{n}E\left(\sum_{i=1}^{n}(x_i - \mu)^2\right) = \sigma^2,$$

а также

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}\mu = \bar{X} - \mu,$$

и с учетом того, что $\bar{X}$ и μ – константы, слагаемое

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

Исходя из этого

$$E(S_n^2) = \sigma^2 - 2E((\bar{X} - \mu)^2) + E\left((\bar{X} - \mu)^2\right) = \sigma^2 - E\left((\bar{X} - \mu)^2\right).$$

Так как мы выше доказали, что $E(\bar{X}) = \mu$, то слагаемое $E\left((\bar{X} - \mu)^2\right)$ представляет из себя дисперсию случайной величины $\bar{X}$, а также предполагая, что варианты вариационного ряда взаимно независимы, получим

$$E\left((\bar{X} - \mu)^2\right) = D(\bar{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = \frac{1}{n^2}D\left(\sum_{i=1}^{n}x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}D(x_i) = \frac{1}{n}\sigma^2. \tag{3}$$

Таким образом,

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

Таким образом, выборочная дисперсия является смещенной (её математическое ожидание не равно дисперсии), однако если её исправить путём домножения на $\frac{n}{n-1}$, то полученная исправленная выборочная дисперсия (2) станет несмещенной.

1.3 Гистограмма относительных частот

Если объем выборки недостаточно велик или интересующий нас признак генеральной совокупности $\boldsymbol{X}$ имеет непрерывное распределение, то варианты вариационного ряда группируются в интервалы. Таким образом строится дискретная модель распределения изучаемого признака $\boldsymbol{X}$, т.н. статистический интервальный ряд распределения.

Типичная процедура группировки выглядит так. Отрезок $[a, b]$, содержащий все варианты вариационного ряда, делится на k интервалов Затем находится частота, т.е. количество наблюдений n_i, попавших в i-й интервал. Для определенности будем полагать n_i равным числу вариант, принадлежащих интервалу $[h_{i-1}, h_i)$, варианты, попавшие на правую границу, т.е. равные h_i, включаются в следующий промежуток при всех $i < k$.

Обычно, $h_i - h_{i-1} = h$ при всех i, т.е. группировка осуществляется с шагом, равным h. Будем использовать формулу Стерджеса для выбора числа интервалов группировки

$$k = 1 + \lfloor \log_2 n \rfloor . \tag{4}$$

Например, если объем выборки n = 50, то k = 7.

Далее находим наименьший m и наибольший M элементы выборки; вычисляем размах выборки $R = M - m$, устанавливаем левую границу интервала группировки $a = m$, а шаг группировки вычисляется как

$$h = \frac{M - a}{k} \tag{5}$$

и округляется при необходимости в большую сторону, согласно размерности элементов выборки. Например, если выборка состоит из чисел 1.01, 1.03 и 3.18, то шаг округляется до сотых. С учетом этого, интервалы группировки вычисляются как

$$h_i = a + i \cdot h, i = 0, 1, ..., k, \tag{6}$$

т.е. правая граница интервала группировки b становится равной $a + kh$.

Для каждого интервала с номером i можно вычислить относительную частоту $w_i = \frac{n_i}{n}$. Наиболее информативной графической формой интервального ряда распределения является *гистограмма относительных частот* (или просто гистограмма), состоящая из прямоугольников

с основаниями (h_{i-1}, h_i), высота которых равна *плотности относительных частот*

$$\omega_i = \frac{w_i}{h_i - h_{i-1}}. \tag{7}$$

Таким образом, площадь каждого прямоугольника равна w_i, а общая сумма этих площадей равна единице.

Заметим, что площадь той части гистограммы относительных частот, что лежит между h_i и h_m $(i < m)$, равна относительному числу вариант, попавших в интервал $[h_i, h_m)$, и в соответствии с статистическим определением вероятности может быть интерпретирована как оценка вероятности $P(h_i \leq X < h_m)$, где X – признак генеральной совокупности. Следовательно, с определенными оговорками, обусловленными дискретностью модели, можно утверждать, что гистограмма относительных частот является оценкой функции плотности вероятности распределения случайной величины X.

1.4 Статистическая функция распределения

Кроме статистического интервального ряда распределения, признак X можно описать статистической (эмпирической) функцией распределения $\hat{F}_n(x)$, которая, в рамках дискретной модели, является аналогом $F(x) = P(X < x)$ – неизвестной функции распределения вероятностей признака X.

Статистической (эмпирической) функцией распределения $\hat{F}(x)$, построенной по случайной выборке $X_1, X_2, ..., X_n$ называется относительная частота того, что признак X примет значение меньшее заданного x. Другими словами

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(X_i, x), \tag{8}$$

где

$$\mathcal{I}(X_i, x) = \begin{cases} 1, \text{ если } X_i < x, \\ 0, \text{ иначе.} \end{cases} \tag{9}$$

$\hat{F}(x)$ – ступенчатая неубывающая функция, принимающая значения от 0 до 1.

Если данные сгруппированы в интервалы $[h_{i-1}, h_i)$, $i = 1, ..., k$ с относительными частотами w_i, то на каждом интервале функция $\hat{F}_n(x)$

определяется как накопленная сумма относительных частот

$$\hat{F}_n(x) = \begin{cases} 0, & x < h_0, \\ w_1, & h_0 \le x < h_1, \\ w_1 + w_2, & h_1 \le x < h_2, \\ \vdots & \\ \sum\limits_{i=1}^{j} w_i, & h_{j-1} \le x < h_j, j = 1, 2, ..., k. \\ w_1 + w_2 + \cdots + w_k = 1, & x \ge h_k. \end{cases} \tag{10}$$

1.5 Выборочное среднеее и выборочная дисперсия для сгруппированных данных

Если данные сгруппированы при помощи интервалов, то выборочное среднее вычисляется как

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} n_i \bar{x}_i = \sum_{i=1}^{n} w_i \bar{x}_i, \tag{11}$$

где $\bar{x}_i$ – середина интервала с номером i, которая вычисляется как

$$\bar{x}_i = h_i + h/2. \tag{12}$$

При этом исправленная выборочная дисперсия вычисляется как

$$\bar{S}_n^2 = \frac{n}{n-1} S_n^2, \tag{13}$$

где выборочная дисперсия S^2 вычисляется как

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} n_i \bar{x}_i^2 - (\bar{X})^2 = \sum_{i=1}^{n} w_i \bar{x}_i^2 - (\bar{X})^2. \tag{14}$$

Следует обратить внимание, что значения выборочного среднего и выборочной дисперсии, вычисленные по сгруппированным данным могут отличаться (например, из-за округлений при вычислении длины интервала h) от значений, полученных по формулам (1) и (2).

1.6 Пример группировки данных

Рассмотрим вариационный ряд

$$\boldsymbol{x} = \begin{array}{cccccccccc} 22 & 24 & 26 & 26 & 27 & 28 & 28 & 31 & 31 & 31 \\ 32 & 32 & 33 & 33 & 33 & 33 & 34 & 34 & 34 & 34 \\ 34 & 35 & 35 & 36 & 36 & 36 & 36 & 36 & 37 & 37 \\ 37 & 37 & 37 & 37 & 38 & 38 & 40 & 40 & 40 & 40 \\ 40 & 41 & 41 & 43 & 44 & 44 & 45 & 45 & 47 & 50 \end{array} \tag{15}$$

Объем выборки $n = 50$, поэтому, согласно (4) количество интервалов $k = 7$. Длина интервала согласно (5) для $a = 22$, $M = 50$ равна $h = 4$. С учетом этого, можно построить таблицу интервалов группировки как показано в Таблице 1. Выборочное среднее $\bar{X} = 36$, выборочная дисперсия $S^2 = 30.08$, исправленная выборочная дисперсии $\bar{S}_n^2 = 30.69$. Рисунки 1–2 показывают соответствующие гистограмму и эмпирическую функцию распределения.

Таблица 1.: Таблица интервалов группировки

i	h_{i-1}	h_i	$\bar{x}_i$	n_i	w_i	$\hat{F}_n(x)$	ω_i
1	22	26	24	2	0.04	0.04	0.010
2	26	30	28	5	0.10	0.14	0.025
3	30	34	32	9	0.18	0.32	0.045
4	34	38	36	18	0.36	0.68	0.090
5	38	42	40	9	0.18	0.86	0.045
6	42	46	44	5	0.10	0.96	0.025
7	46	50	48	2	0.04	1.00	0.010

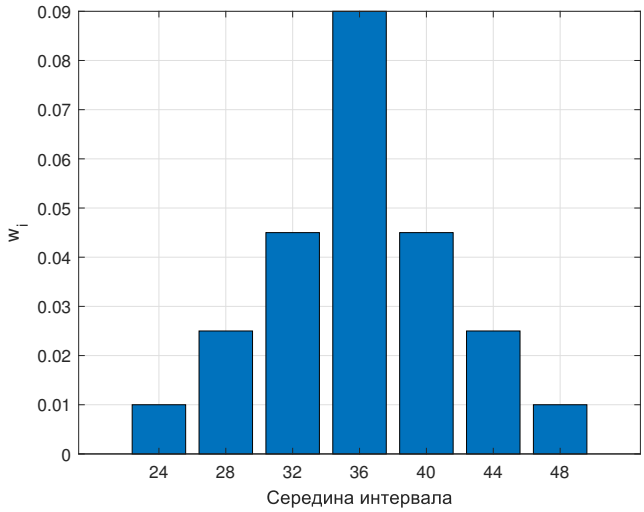

Рис. 1.: Гистограмма

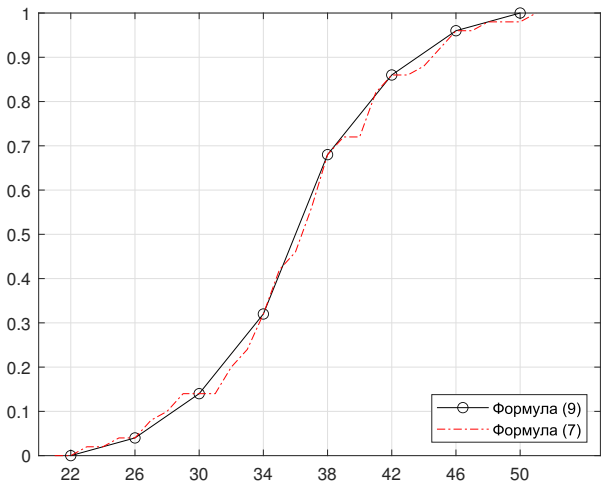

Рис. 2.: Эмпирическая функция распределения $\hat{F}_n(x)$

1.7 Доверительные интервалы для генерального математического ожидания и генеральной дисперсии

Выборочное среднее $\bar{X}$ является случайной величиной, значение которой зависит от того, какой вариационный ряд из генеральной совокупности используется для его вычисления. Как мы уже доказали, $E(\bar{X}) = \mu$. Однако в практических задачах может быть необходимо определить, в каком интервале $\bar{X} \pm \epsilon_\gamma$ будет находиться μ относительно текущего значения $\bar{X}$ с некоторой доверительной вероятностью γ. Данный доверительный интервал определяется из равенства:

$$P(|\bar{X} - \mu| < \epsilon_\gamma) = \gamma.$$

Из неравенства Берри-Эссеена следует, что отклонение $\bar{X}$ от нормального распределения пропорционально $\frac{1}{\sqrt{n}}$, а дисперсия выборочной дисперсии для нормального распределения $D(\bar{S}_n^2) = \frac{2\sigma^4}{n-1}$. Поэтому если предполагается, что исследуемое распределение близко к нормальному, то при $n > 40$ используется эвристика, согласно которой распределение $\bar{X}$ можно приближенно считать нормальным с параметрами $\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, и $\sigma^2 \approx \bar{S}_n^2$. Теперь перейдём к величине

$$Y = \frac{\mu - \bar{X}}{\bar{S}_n/\sqrt{n}},$$

которая $Y \approx \mathcal{N}(0,1)$. Для этой случайной величины необходимо найти такое ϵ_γ, что $P(-\epsilon_\gamma \le Y \le \epsilon_\gamma) = \gamma$ (см. рисунок 3). Тогда, с учетом симметричности $\Phi(z)$, т.е. $\Phi(-z) = 1 - \Phi(z)$, получим

$$P(-\epsilon_\gamma \le Y \le \epsilon_\gamma) = \int_{-\epsilon_\gamma}^{\epsilon_\gamma} f(x)dx = \Phi(\epsilon_\gamma) - \Phi(-\epsilon_\gamma) = 2\Phi(\epsilon_\gamma) - 1 = \gamma,$$

откуда

$$\epsilon_\gamma = \Phi^{-1}\left(\frac{\gamma+1}{2}\right). \tag{16}$$

Возвращаясь к исходной величине $\bar{X}$, получим

$$-\epsilon_\gamma \le \frac{\mu - \bar{X}}{\bar{S}_n/\sqrt{n}} \le \epsilon_\gamma,$$

откуда

$$-\frac{\epsilon_\gamma \bar{S}_n}{\sqrt{n}} + \bar{X} \le \mu \le \frac{\epsilon_\gamma \bar{S}_n}{\sqrt{n}} + \bar{X}, \tag{17}$$

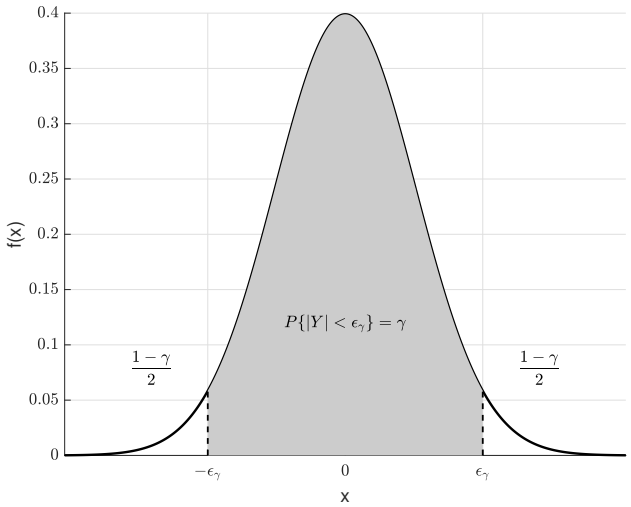

Рис. 3.: Выбор доверительного интервала для нормального распределения

что является доверительным интервалом для математического ожидания μ с доверительной вероятностью γ.

Пример 1.1. В качестве примера, вычислим доверительный интервал для генерального математического ожидания для вариационного ряда (15) для доверительной вероятности $\gamma = 0.95$. Согласно (16), $\epsilon_\gamma = 1.96$. С учетом того, что $\bar{X} = 36$ и $\bar{S}_n^2 = 30.69$, из (16) получим что $\mu \in [34.46, 37.54]$.

Если $n < 40$ (малая выборка), то предположение о том, что $\bar{X}$ является нормально распределённой случайной величиной справедливо только тогда, когда сами величины x_i распределены нормально (сумма нормальных случайных величин распределена нормально). Кроме этого, для данного случая дисперсией выборочной дисперсии даже для нормального распределения $D(\bar{S}_n^2) = \frac{2\sigma^4}{n-1}$ нельзя пренебрегать, т.е. мы не можем считать, что $\sigma^2 \approx \bar{S}_n^2$.

Таким образом, если при малой выборке мы предполагаем, что генеральная совокупность имеет нормальное распределение, то для нахождения доверительного интервала для выборочного среднего нужно использовать тот факт, что величина $t_k = Y$ имеет t-распределение Стьюдента с $k = n - 1$ степенями свободы. Оно не зависит от неизвестных параметров распределения случайной величины, а зависит только от числа k, и напоминает нормальное распределение, а при $k \to \infty$ приближается к нему. С учетом симметричности функции распределения

Стьюдента $\mathcal{F}_{(T,k)}(z)$, т.е. $\mathcal{F}_{(T,k)}(-z) = 1 - \mathcal{F}_{(T,k)}(z)$, получим

$$P(-\epsilon_\gamma \leq Y \leq \epsilon_\gamma) = \mathcal{F}_{(T,k)}(\epsilon_\gamma) - \mathcal{F}_{(T,k)}(-\epsilon_\gamma) = 2\mathcal{F}_{(T,k)}(\epsilon_\gamma) - 1 = \gamma,$$

откуда

$$\epsilon_\gamma = \mathcal{F}^{-1}_{(T,k)}\left(\frac{\gamma+1}{2}\right). \tag{18}$$

Пример 1.2. В качестве примера для доверительной вероятности $\gamma = 0.95$ вычислим доверительный интервал для генерального математического ожидания для вариационного ряда

$$x = \{17.58, 18.86, 17.34, 16.39, 17.65, 18.54, 19.08, 18.75, 17.45, 15.63\}. \tag{19}$$

Согласно (18) $\epsilon_\gamma = 2.2622$. С учетом того, что $\bar{X} = 17.73$ и $\bar{S}^2_n = 1.25$, то из (16) получим что $\mu \in [16.93, 18.53]$.

Для поиска доверительного интервала для генеральной дисперсии при малой выборке из нормального распределения используется тот факт, что величина

$$\chi^2 = \frac{(n-1)\bar{S}^2_n}{\sigma^2}$$

имеет распределение Хи-квадрат со $k = n - 1$ степенями свободы. Основное отличие от нормального распределения и t-распределения Стьюдента заключается в том, что распределение Хи-квадрат не является симметричным (см. рисунок 4).

Поэтому поиск доверительного интервала необходимо искать в виде

$$P(\chi^2_1 \leq \chi^2 \leq \chi^2_2) = \gamma,$$

где $\chi^2_1 \neq -\chi^2_2$. С учетом этого, выбираем границы интервала χ^2_1 и χ^2_2 так, чтобы $P(\chi^2 \geq \chi^2_2) = P(\chi^2 < \chi^2_1) = \frac{1-\gamma}{2}$. Откуда $P(\chi^2 \geq \chi^2_2) = 1 - P(\chi^2 < \chi^2_2) = \frac{1+\gamma}{2}$. С учетом этого, мы сначала вычисляем величины

$$\chi^2_1 = \mathcal{F}^{-1}_{(\chi^2,k)}\left(\frac{1-\gamma}{2}\right), \chi^2_2 = \mathcal{F}^{-1}_{(\chi^2,k)}\left(\frac{1+\gamma}{2}\right), \tag{20}$$

где $\mathcal{F}(.)$ – функция распределения Хи-квадрат. Тогда

$$\chi^2_1 \leq \frac{(n-1)\bar{S}^2_n}{\sigma^2} \leq \chi^2_2,$$

тогда доверительный интервал для генеральной дисперсии записывается как

$$\frac{(n-1)\bar{S}^2_n}{\chi^2_2} \leq \sigma^2 \leq \frac{(n-1)\bar{S}^2_n}{\chi^2_1}. \tag{21}$$

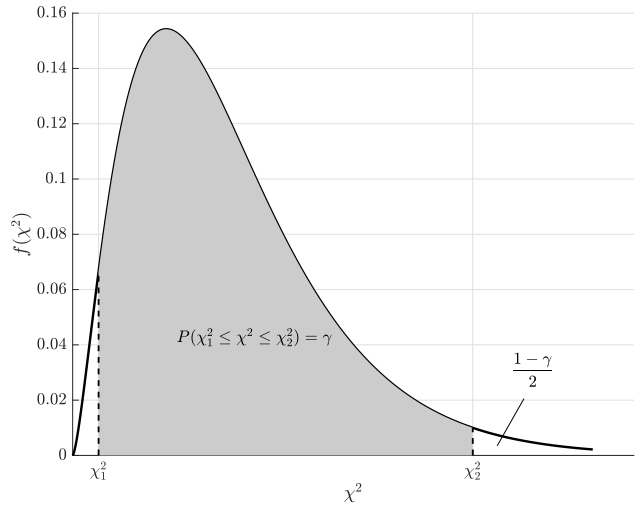

Рис. 4.: Выбор доверительного интервала для распределения Хи-квадрат

Пример 1.3. В качестве примера для $\gamma = 0.95$ вычислим доверительный интервал для генеральной дисперсии вариационного ряда (19). Согласно (20) получим $\chi_2^2 = 2.7004$, $\chi_2^2 = 19.023$, С учетом того, что $\bar{X} = 17.73$ и $\bar{S}_n^2 = 1.25$, то из (21) получим что $\sigma \in [0.59, 4.17]$.

1.8 Визуальное сравнение с нормальным распределением

Одной из задач математической статистики является визуальная проверка близости изучаемого распределения признака генеральной совокупности к нормальному[1].

Предположим, что рассматриваемый признак X распределён по нормальному закону с плотностью распределения

$$f(x) = \frac{1}{2\pi\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{22}$$

где неизвестные математическое ожидание и дисперсия заменены на их соответствующие оценки, т.е. $\mu = \bar{X}$, $\sigma = \bar{S}$. Для построения гистограммы нормального распределения можно использовать тот факт, что

[1] Отметим, что математическая статистика располагает более формальными процедурами проверки гипотезы о нормальности распределения признака X по так называемым критериями согласия, которые рассматриваются в Лабораторной работе №4.

вероятность того, что распределённая по нормальному закону случайная величина X находится в интервале $[a, b]$, $a < b$ определяется как

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \tag{23}$$

где $\Phi(.)$ – функция Лапласа. С учетом этого, для примера из Таблицы 1 мы можем вычислить соответствующие значения $\frac{p_i}{h}$, где

$$p_i = P(h_{i-1} \leq X \leq h_i) = \Phi\left(\frac{h_i - \bar{X}}{\bar{S}}\right) - \Phi\left(\frac{h_{i-1} - \bar{X}}{\bar{S}}\right). \tag{24}$$

При этом нужно учесть, что нормальное распределение определено на всем интервале значений случайной величины, т.е. чтобы сумма вероятностей по всем интервалам была равна единице мы также должны добавить два крайних интервала $[-\infty, h_1]$ и $[h_k, \infty]$. С учетом того, что $\Phi(-\infty) = 0$ и $\Phi(\infty) = 1$ получим, что для крайне левого интервала

$$P(-\infty < X \leq h_1) = \Phi\left(\frac{h_1 - \bar{X}}{\bar{S}}\right), \tag{25}$$

а для крайне правого интервала

$$P(h_k \leq X < \infty) = 1 - \Phi\left(\frac{h_k - \bar{X}}{\bar{S}}\right). \tag{26}$$

Таблица 2.: Таблица интервалов для нормального распределения

i	h_{i-1}	h_i	$\Phi\left(\frac{h_i-\bar{X}}{\bar{S}}\right)$	$\Phi\left(\frac{h_{i-1}-\bar{X}}{\bar{S}}\right)$	p_i	$\frac{p_i}{h}$
1	-∞	22	0.0058	0.0000	0.0058	0.0014
2	22	26	0.0355	0.0058	0.0298	0.0074
3	26	30	0.1394	0.0355	0.1039	0.0260
4	30	34	0.3591	0.1394	0.2196	0.0549
5	34	38	0.6409	0.3591	0.2819	0.0705
6	38	42	0.8606	0.6409	0.2196	0.0549
7	42	46	0.9645	0.8606	0.1039	0.0260
8	46	50	0.9942	0.9645	0.0298	0.0074
9	50	∞	1.0000	0.9942	0.0058	0.0014

На рисунке 5 показано сравнение эмпирической гистограммы и гистограммы нормального распределения с параметрами $\mu = \bar{X}$, $\sigma = \bar{S}$

для вариационного ряда (15). На рисунке 5 показано сравнение эмпирической и соответствующей нормальной функций распределения. Как можно отметить, распределение эмпирических данных похоже на нормальное распределение.

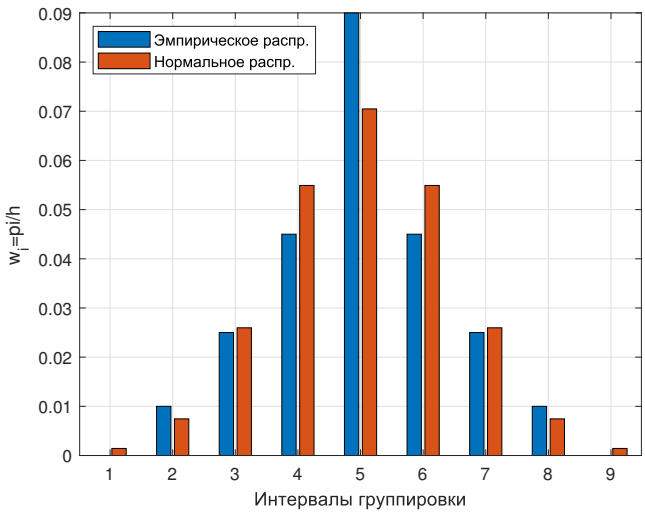

Рис. 5.: Сравнение гистограм эмпирического и соответствующего нормального распределения

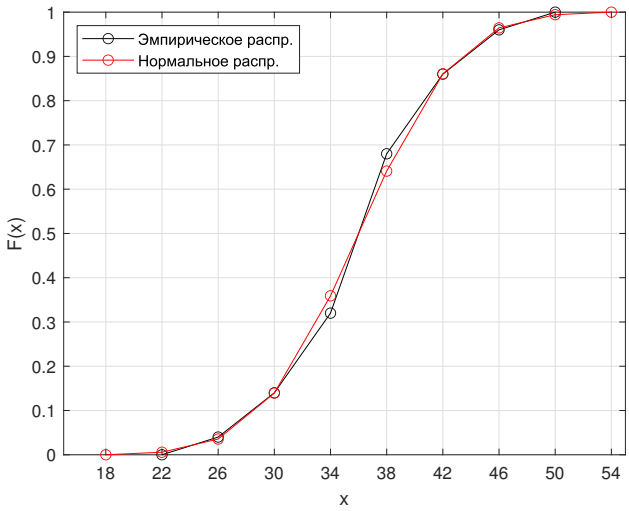

Рис. 6.: Сравнение эмпирической и соответствующей нормальной функций распределения

1.9 Ход выполнения работы

На языке программирования Python для заданных данных:

1. Построить гистограмму относительных частот.
2. Построить эмпирическую функцию распределения.
3. Вычислить выборочное среднее, выборочную дисперсию и исправленную выборочную дисперсию.
4. Вычислить доверительный интервал для генерального математического ожидания для доверительных вероятностей $\gamma \in \{0.9, 0.95, 0.99\}$. Для первых десяти членов вариационного ряда вычислить доверительный интервал для генерального математического ожидания и генеральной дисперасии для доверительных вероятностей $\gamma \in \{0.9, 0.95, 0.99\}$.
5. Сравнить визуально полученные эмпирические гистограмму и функцию распределения с соответствующим нормальным распределением.
6. Сделать выводы о визуальной схожести эмпирического и нормального распределений.
7. В отчете должны быть предоставлены аналоги Таблиц 1–2 и Рисунков 5–6.
8. Для выполнения работы разрешается использовать базовые вычислительные операции библиотеки *numpy*, такие как сложение, вычитание, деление и т.д. При этом
   (a) Для вычисления значения функции Лапласа $\Phi(x)$ можно использовать функцию ошибки $erf(x)$ через функцию *math.erf(x)*. При этом нужно учесть, что
   $$\Phi(x) = \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right).$$
   (b) Для вычисления значения функции Лапласа $\Phi^{-1}(x)$ можно использовать обратную функцию ошибки $erf^{-1}(x)$ через функцию *math.erfinv(x)*. При этом нужно учесть, что
   $$\Phi^{-1}(x) = \sqrt{2} erf^{-1}\left(2\Phi(x) - 1\right).$$
   (c) $\mathcal{F}^{-1}_{(T,k)}(x)$ вычисляется как *scipy.stats.t.ppfr(x,k)*.
   (d) $\mathcal{F}^{-1}_{(\chi^2,k)}(x)$ вычисляется как *scipy.stats.chi2.ppfr(x,k)*.