- **ЛЕКЦИЯ 5**

- **Повторение пройденного**

Часть 1 - ГЛАВА 9. ЗАКОН БОЛЬШИХ ЧИСЕЛ. ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ

- При *статистическом* определении вероятности она трактуется как некоторое число, к которому стремится относительная частота случайного события. При *аксиоматическом* определении вероятность – это, по сути, аддитивная мера множества исходов, благоприятствующих случайному событию. В первом случае имеем дело с эмпирическим пределом, во втором – с теоретическим понятием меры. Совсем не очевидно, что они относятся к одному и тому же понятию. Связь разных определений вероятности устанавливает теорема Бернулли, являющаяся частным случаем закона больших чисел.

- При увеличении числа испытаний биномиальный закон стремится к нормальному распределению. Это теорема Муавра–Лапласа, которая является частным случаем центральной предельной теоремы. Последняя гласит, что функция распределения суммы независимых случайных величин с ростом числа слагаемых стремится к нормальному закону.

- Закон больших чисел и центральная предельная теорема лежат в основании математической статистики.

9.1. Неравенство Чебышева

- Пусть случайная величина ξ имеет конечные математическое ожидание $M[\xi]$ и дисперсию $D[\xi]$. Тогда для любого положительного числа ε справедливо неравенство:

$$p\left(\left|\xi - M[\xi]\right| < \varepsilon\right) > 1 - \frac{D[\xi]}{\varepsilon^2}$$

$$p\left(\left|\xi - M[\xi]\right| < \varepsilon\right) > 1 - \frac{D[\xi]}{\varepsilon^2}$$

Примечания

- Для противоположного события:

$$p\left(\left|\xi - M[\xi]\right| \geq \varepsilon\right) \leq \frac{D[\xi]}{\varepsilon^2}$$

- Неравенство Чебышева справедливо для любого закона распределения.

- Положив $\varepsilon = 3\sigma_\xi$, получаем нетривиальный факт:

$$p\left(\left|\xi - M[\xi]\right| \geq 3\sigma_\xi\right) \leq \frac{\sigma_\xi^2}{9\sigma_\xi^2} = \frac{1}{9} \approx 0{,}11$$

9.2. Закон больших чисел в форме Чебышева

- *Теорема* Пусть случайные величины $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ попарно независимы и имеют конечные дисперсии, ограниченные одной и той же постоянной $C: D[\xi_i] \le C$, где $i = 1, 2, \ldots$ Тогда для любого $\varepsilon > 0$ имеем

$$\lim_{n \to \infty} p\left(\left| \frac{1}{n} \sum_{i=1}^{n} \xi_i - \frac{1}{n} \sum_{i=1}^{n} M[\xi_i] \right| < \varepsilon \right) = 1$$

- Таким образом, закон больших чисел говорит о сходимости по вероятности среднего арифметического случайных величин (т. е. случайной величины) к среднему арифметическому их мат. ожиданий (т. е. к не случайной величине).

9.2. Закон больших чисел в форме Чебышева: дополнение

- *Теорема (Маркова)*: закон больших чисел выполняется, если дисперсия суммы случайных величин растет не слишком быстро с ростом *n:*

$$\lim_{n\to\infty}\frac{1}{n^2}D\left[\sum_{i=1}^{n}\xi_i\right]=0$$

9.3. Теорема Бернулли

- *Теорема:* Рассмотрим схему Бернулли. Пусть μ_n – число наступлений события A в n независимых испытаниях, p – вероятность наступления события A в одном испытании. Тогда для любого $\varepsilon > 0$

$$\lim_{n \to \infty} p\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) = 1$$

- Т.е. вероятность того, что отклонение относительной частоты случайного события от его вероятности *p* будет по модулю сколь угодно мало, оно стремится к единице с ростом числа испытаний *n*.

- Доказательство: Случайная величина μ_n распределена по биномиальному закону, поэтому имеем

$$M[\mu_n] = np,\ D[\mu_n] = np(1-p)$$

и тогда

$$M\left[\frac{\mu_n}{n}\right] = p,\ \ D\left[\frac{\mu_n}{n}\right] = \frac{p(1-p)}{n}.$$

Для случайной величины μ_n / n неравенство Чебышева принимает следующий вид:

$$p\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) > 1 - \frac{p(1-p)}{n\varepsilon^2}.$$

Переходя к пределу при $n \to \infty$, получаем

$$\lim_{n \to \infty} p\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) = 1$$

9.4. Характеристические функции

- *Характеристической функцией* случайной величины ξ называется функция

$$\phi_\xi(t) = M\left[\exp(it\xi)\right]$$

где $\exp(x) = e^x$.

- Таким образом, $\phi_\xi(t)$ представляет собой математическое ожидание некоторой комплексной случайной величины $\eta = \exp(it\xi)$ связанной с величиной ξ . В частности, если ξ дискретная случайная величина, заданная рядом распределения $\{x_i, p_i\}$, где $i = 1, 2,..., n$, то

$$\phi_\xi(t) = \sum_{i=1}^{n} \exp(itx_i)\, p_i$$

- Для непрерывной случайной величины ξ плотностью распределения вероятности $f_\xi(x)$

$$\phi_\xi(t) = \int_{-\infty}^{\infty} \exp(itx) f_\xi(x) \mathrm{d}x$$

Например, для случайной величины ξ, имеющей нормальный закон распределения с параметрами $a = M[\xi]$ и $\sigma = \sqrt{D[\xi]}$, характеристическая функция равна

$$\phi_\xi(t) = \exp\left(ita - \frac{t^2\sigma^2}{2}\right).$$

Соотношение (9.7) есть так называемое *прямое преобразование Фурье*. Известно, что в таком случае функцию $f_\xi(x)$ можно найти по известной характеристической функции $\phi_\xi(t)$, используя *обратное преобразование Фурье*:

$$f_\xi(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp(-itx)\phi_\xi(t)\mathrm{d}t. \qquad (9.9)$$

В силу единственности преобразования Фурье между $f_\xi(x)$ и $\phi_\xi(t)$ имеется однозначное соответствие: известной плотности распределения вероятности $f_\xi(x)$ соответствует одна и только одна характеристическая функция $\phi_\xi(t)$ и наоборот. Поскольку между $f_\xi(x)$ и $F_\xi(x)$ также существует однозначное соответствие, такое соотношение между $\phi_\xi(t)$ и $F_\xi(x)$ доказано.

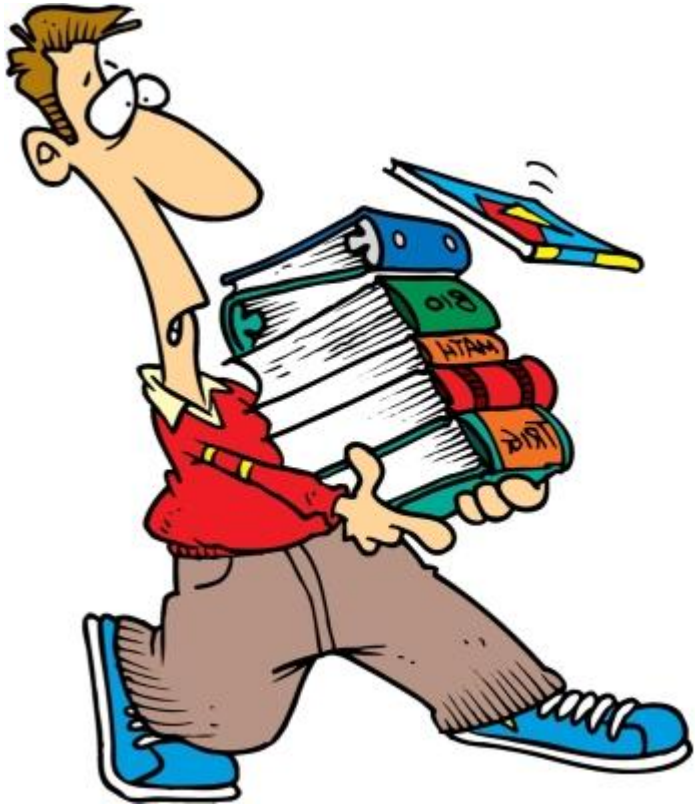
9.5. Центральная предельная теорема (теорема Ляпунова)

Группа теорем, касающихся предельных законов распределения суммы случайных величин, носит общее название *центральной предельной теоремы*. Рассмотрим ее классическую формулировку.

Теорема: Пусть $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ – бесконечная последовательность независимых одинаково распределенных случайных величин, имеющих конечные математическое ожидание a и дисперсию σ^2. Тогда при $n \to \infty$ функция распределения $\eta_n = (S_n - an)/(\sigma\sqrt{n})$, где $S_n = \sum_{i=1}^{n} \xi_i$, будет стремиться к нормальному закону с нулевым математическим ожиданием и средним квадратическим отклонением, равным единице, т. е.

$$F_{\eta_n}(x) \xrightarrow[n\to\infty]{} F_N(x, 0, 1). \qquad (9.15)$$

- **Повторили пройденное**



ОСНОВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

ЧАСТЬ II. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Эпиграф

«Существует три вида лжи: ложь, наглая ложь и статистика»

Бенджамин Дизраэли

Введение

Две основные задачи математической статистики:

- сбор и группировка статистических данных;
- разработка методов анализа полученных данных в зависимости от целей исследования.

Методы статистического анализа данных:

- оценка неизвестной вероятности события;
- оценка неизвестной функции распределения;
- оценка параметров известного распределения;
- проверка статистических гипотез о виде неизвестного распределения или о значениях параметров известного распределения.

ГЛАВА 1.
ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1.1. Генеральная совокупность и выборка

- *Генеральная совокупность* - все множество исследуемых объектов, *Выборка* – набор объектов, случайно отобранных из генеральной совокупности для исследования.
- *Объем* генеральной совокупности и *объем* выборки - число объектов в гене-ральной совокупности и выборке - будем обозначать соответственно как *N* и *n*.

- Выборка бывает *повторной*, когда каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность, и *бесповторной*, если отобранный объект в генеральную совокупность не возвращается.

Репрезентативная выборка:

- правильно представляет особенности генеральной совокупности, т.е. является *репрезентативной* (представительной).

- По закону больших чисел, можно утверждать, что это условие выполняется, если:

  1) объем выборки *n* достаточно большой;

  2) каждый объект выборки выбран случайно;

  3) для каждого объекта вероятность попасть в выборку одинакова.

- Генеральная совокупность и выборка могут быть *одномерными (однофакторными)*

  и *многомерными (многофакторными)*

1.2. Выборочный закон распределения (статистический ряд)

- Пусть в выборке объемом n интересующая нас случайная величина ξ (какой-либо параметр объектов генеральной совокупности) принимает n_1 раз значение x_1, n_2 раза – значение x_2,... и n_k раз – значение x_k. Тогда наблюдаемые значения $x_1, x_2,..., x_k$ случайной величины ξ называются *вариантами*, а $n_1, n_2,..., n_k$ – их *частотами*.

- Разность $x_{\max} - x_{\min}$ есть *размах* выборки, отношение $\omega_i = n_i / n$ – *относительная частота* варианты x_i.
- Очевидно, что

$$\sum_{i=1}^{k} n_i = n; \qquad \sum_{i=1}^{k} \omega_i = 1$$

- Если мы запишем варианты в возрастающем порядке, то получим *вариационный ряд*. Таблица, состоящая из таких упорядоченных вариант и их частот (и/или относительных частот) называется *статистическим рядом* или *выборочным законом распределения*.

x_1	x_2	x_3	...	x_k
n_1	n_2	n_3	...	n_k
ω_1	ω_2	ω_3	...	ω_k

где $x_i \le x_{i+1}$ при $i = 1,2, ..., k-1$.

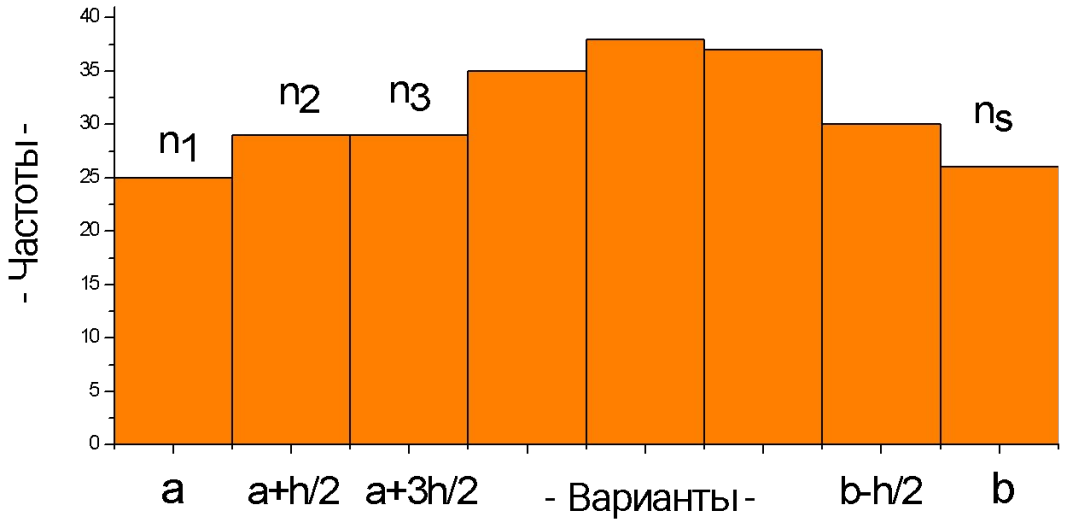
-- Аналог закона распределения дискретной случайной величины в теории вероятности

- Если вариационный ряд состоит из очень большого количества чисел или исследуется некоторый непрерывный признак, используют *группированную* выборку. Для ее получения интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько обычно равных частей (подинтервалов) длиной h. При составлении статистического ряда в качестве x_i обычно выбирают середины подинтервалов, а n_i приравнивают числу вариант, попавших в i-й подинтервал.

Пусть число подинтервалов равно s, $a = \min\{x_i\}$, $b = \max\{x_i\}$. Тогда для группированной выборки получим следующий статистический ряд:

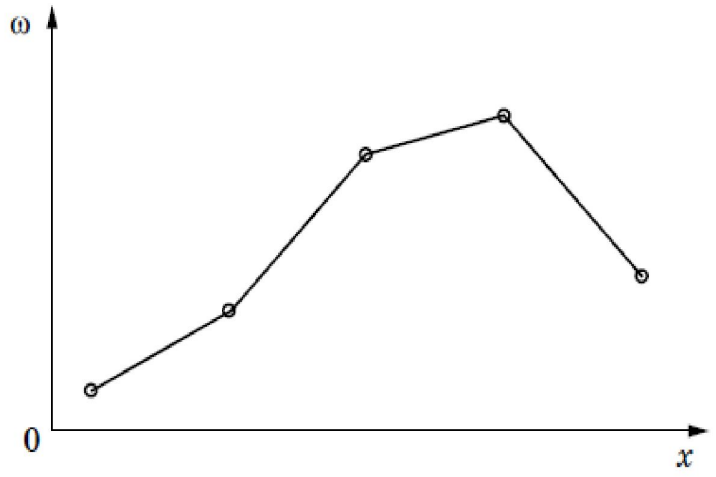
Номера подинтервалов	Границы подинтервалов	Варианты	Частоты
1	$[a, a+h]$	$a + h/2$	n_1
2	$(a+h, a+2h]$	$a + 3h/2$	n_2
…	…	…	…
s	$(b-h, b]$	$b - h/2$	n_s

где $h = (b - a)/s$, а n_i равно сумме частот вариант, попавших в i-й подинтервал.



1.3. Полигон частот, выборочная функция распределения

- Отложим значения случайной величины x_i по оси абсцисс, а значения n_i – по оси ординат. Ломаная линия, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2),..., (x_k, n_k), называется *полигоном частот*. Если вместо абсолютных значений n_i на оси ординат отложить относительные частоты ω_i, то получим *полигон относительных частот*



- По аналогии с функцией распределения дискретной случайной величины по выборочному закону распределения можно построить *выборочную (эмпирическую)* функцию распределения

$$F_n^*(x) = \sum_{x_i<x} \omega_i = \frac{1}{n} \sum_{x_i<x} n_i,$$

- где суммирование выполняется по всем частотам, которым соответствуют значения вариант, меньшие x. Заметим, что эмпирическая функция распределения зависит от объема выборки n.

- В отличие от функции $F_n^*(x)$,найденной для случайной величины ξ опытным путем в результате обработки статистических данных, истинную функцию распределения $F_\xi(x)$,связанную с генеральной совокупностью, называют *теоретической*. (Обычно генеральная совокупность настолько велика, что обработать ее всю невозможно, т.е. исследовать ее можно только теоретически).

- Заметим, что:

Функция $F_\xi(x)$ по определению есть вероятность события $(\xi < x)$:

$$F_\xi(x) = p(\xi < x),$$

а $F_n^*(x)$ – его относительная частота. При достаточно больших n, как следует из теоремы Бернулли (см. п. 9.3 части I), относительная частота события $(\xi < x)$, т.е. $F_n^*(x)$, должна стремиться (по вероятности) к его вероятности, т.е. к $F_\xi(x)$:

$$F_n^*(x) \xrightarrow[n\to\infty]{\text{по вер.}} F_\xi(x).$$

1.4. Свойства эмпирической функции распределения

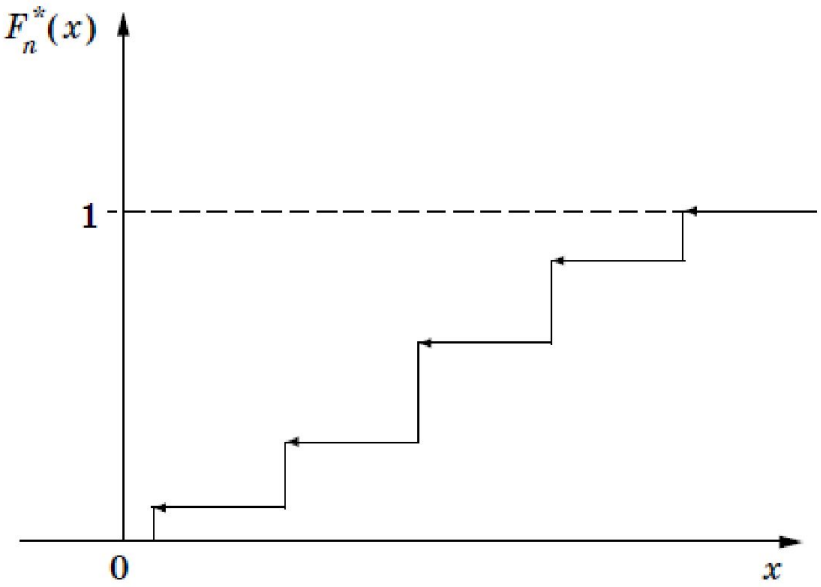
Из определения (1.1) функции $F_n^*(x)$ очевидно, что ее свойства совпадают со свойствами $F_\xi(x)$, а именно:

1) $0 \le F_n^*(x) \le 1$;

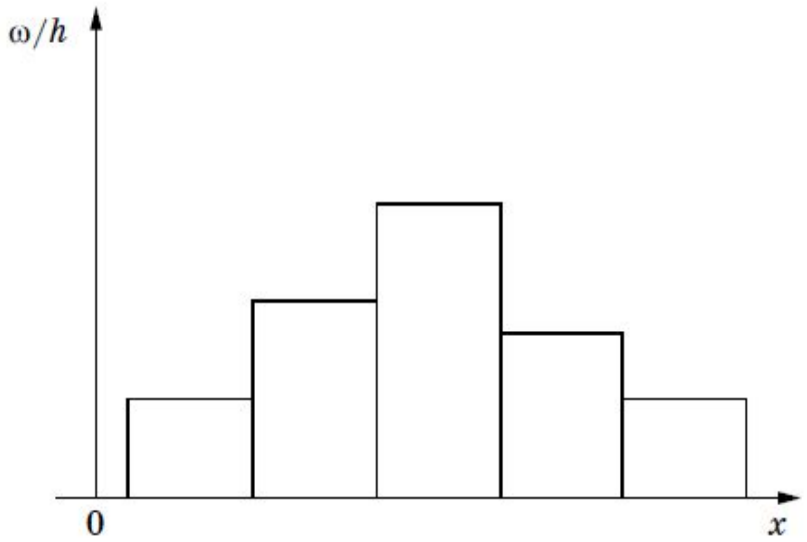
2) $F_n^*(x)$ – неубывающая функция, т.е. $F_n^*(x_2) \ge F_n^*(x_1)$ при $x_2 > x_1$;

3) Если x_1 – наименьшая варианта, то $F_n^*(x) = 0$ при $x \le x_1$; если x_k – наибольшая варианта, то $F_n^*(x) = 1$ при $x > x_k$. Соответственно $F_n^*(-\infty) = 0$ и $F_n^*(\infty) = 1$.

- Ступенчатый вид



- Еще одним графическим представлением интересующей нас выборки является *гистограмма* – ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат подинтервалы шириной h, а высотами – отрезки длиной n_i/h (гистограмма частот) или ω_i/h (гистограмма относительных частот).
- В первом случае площадь гистограммы равна объему выборки n, во втором – единице



Пример

Пример 1.1. Проведено 30 серий по 24 броска кубика, и в каждой серии отмечалось число выпадений грани с шестью очками. Получены следующие результаты: 4, 2, 5, 6, 1, 4, 3, 7, 4, 4, 2, 3, 3, 5, 6, 4, 5, 3, 2, 4, 5, 1, 8, 3, 4, 4, 5, 2, 4, 6. Рассматривая полученные данные как выборку, надо построить выборочный закон распределения для результатов данного опыта.

Решение. Случайная величина ξ – число выпадений шести очков при 24 бросках кубика. В принципе значениями ξ могли быть числа от 0 до 24. Однако в данном случае вариационный ряд включает лишь восемь элементов: 1, 2, 3, 4, 5, 6, 7 и 8. Выборочный закон распределения (статистический ряд) имеет вид:

x_i	n_i	ω_i	x_i	n_i	ω_i
1	2	2/30	5	5	5/30
2	4	4/30	6	3	3/30
3	5	5/30	7	1	1/30
4	9	9/30	8	1	1/30

ГЛАВА 2. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ВЫБОРКИ

- Задача математической статистики – по имеющейся выборке получить информацию о генеральной совокупности. Числовые характеристики репрезентативной выборки -- оценка соответствующих характеристик исследуемой случайной величины, связанной с генеральной совокупностью.

2.1. Выборочное среднее и выборочная дисперсия, эмпирические моменты

- *Выборочным средним* называется среднее арифметическое значений вариант в выборке

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} \omega_i x_i,$$

где x_i – варианты, а n_i – их частоты.

- Выборочное среднее используется для статистической оценки математического ожидания исследуемой случайной величины.

- *Выборочной дисперсией* называется величина, равная

$$D^* = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{k} \omega_i \left(x_i - \overline{x}\right)^2$$

- *Выборочным средним квадратическим отклонением* –

$$\sigma^* = \sqrt{D^*}$$

- Легко показать, что выполняется следующее соотношение, удобное для вычисления дисперсии:

$$D^* = \overline{x^2} - \overline{x}^2,$$

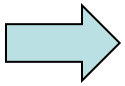
$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} n_i {x_i}^2 = \sum_{i=1}^{k} \omega_i x_i^2.$$

- Другими характеристиками вариационного ряда являются: *мода* M_0 – варианта, имеющая наибольшую частоту, и *медиана* m_e – варианта, которая делит вариационный ряд на две части, равные числу вариант.

- 2, 5, 2, 11, 5, 6, 3, 13, 5 (мода = 5)

- 2, 2, 3, 5, 5, 5, 6, 11,13 (медиана = 5)

- По аналогии с соответствующими теоретическими выражениями можно построить *эмпирические моменты*, применяемые для статистической оценки начальных и центральных моментов исследуемой случайной величины.



- По аналогии с моментами α_k и β_k теории вероятностей *начальным эмпирическим моментом* порядка *m* называется величина

$$\alpha_m^* = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^m = \sum_{i=1}^{k} \omega_i x_i^m$$

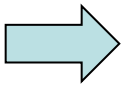
- *центральным эмпирическим моментом* порядка *m* -

$$\beta_m^* = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x})^m = \sum_{i=1}^{k} \omega_i (x_i - \overline{x})^m$$

Заметим, что первый начальный момент α_1^* – это выборочное среднее $\overline{x}$, первый центральный момент β_1^* равен нулю, а второй центральный момент β_2^* – это выборочная дисперсия D^*.

2.2. Свойства статистических оценок параметров распределения: несмещенность, эффективность, состоятельность

- После получения статистических оценок параметров распределения случайной величины ξ : выборочного среднего, выборочной дисперсии и т. д., необходимо убедиться, что они являются хорошим приближением для соответствующих параметров теоретического распределения ξ.
- Найдем условия, которые должны для этого выполняться.



Пусть A^* – статистическая оценка неизвестного параметра A теоретического распределения некоторой случайной величины ξ, например, A – дисперсия $D[\xi]$ некоторой случайной величины ξ, а A^* – ее выборочная дисперсия D^*. Выделим из генеральной совокупности l выборок одного и того же объема n и найдем по каждой из них выборочную оценку A_j^* $(j = 1, 2, \ldots, l)$ параметра A. В принципе, A^* можно рассматривать как некоторую случайную величину, принявшую значения A_1^*, A_2^* и т.д. Тогда, если математическое ожидание A^* не равно оцениваемому параметру A, мы будем получать при вычислении оценок A систематические ошибки: если $M[A^*] > A$, оценка A^* будет в среднем больше A, если $M[A^*] < A$ – меньше. Необходимым условием отсутствия систематических ошибок является требование $M[A^*] = A$.

- Статистическая оценка A^* называется *несмещенной*, если ее математическое ожидание равно оцениваемому параметру генеральной совокупности A при любом объеме выборки, т.е.

$$M[A^*] = A$$

- Если это условие не выполняется, оценка называется *смещенной*.

- Несмещенность оценки не является достаточным условием хорошего приближения статистической оценки *A** к истинному (теоретическому) значению оцениваемого параметра *A*.

- Разброс отдельных значений A_j^* относительно среднего значения $M[A^*]$ зависит от величины дисперсии $D[A^*]$. Если дисперсия велика, то значение A_j^* найденное по данным одной выборки, может значительно отличаться от оцениваемого параметра. Следовательно, для надежного оценивания дисперсия $D[A^*]$ должна быть мала. Статистическая оценка называется *эффективной*, если при заданном объеме выборки *n* она имеет наименьшую возможную дисперсию.

- К статистическим оценкам предъявляется еще требование состоятельности. Оценка называется *состоятельной*, если при $n \to \infty$ она стремится по вероятности к оцениваемому параметру. Заметим, что несмещенная оценка будет состоятельной, если при $n \to \infty$ ее дисперсия стремится к 0.

2.3. Свойства выборочного среднего

- Будем полагать, что варианты $x_1, x_2, \ldots, x_n$ являются значениями соответствующих независимых одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_n$, имеющих математическое ожидание $M[\xi_i] = a$ исперсию $D[\xi_i] = \sigma^2$да выборочное среднее можно рассматривать как случайную величину

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$$

- *Несмещенность*. Из свойств математического ожидания следует, что

$$M\left[\bar{X}\right] = M\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\right] = \frac{1}{n}M\left[\sum_{i=1}^{n}\xi_i\right] = \frac{1}{n}\sum_{i=1}^{n}M\left[\xi_i\right] = \frac{1}{n}\sum_{i=1}^{n}a = a$$

- т.е. выборочное среднее является несмещенной оценкой математического ожидания случайной величины.

- Можно также показать *эффективность* оценки по выборочному среднему математического ожидания (для нормального распределения)

- *Состоятельность*. Пусть a – оцениваемый параметр, а именно математическое ожидание генеральной совокупности $M[\xi]$, σ^2 – дисперсия генеральной совокупности $D[\xi]$. Рассмотрим неравенство Чебышева

$$p\left(\left|\xi - M[\xi]\right| \geq \varepsilon\right) \leq \frac{D[\xi]}{\varepsilon^2}$$

У нас: $\xi = \bar{X};\ M[\xi] = M\left[\bar{X}\right] = a;\ D[\xi] = D\left[\bar{X}\right] = \sigma^2 / n$

тогда $p\left(\left|\bar{X} - a\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$. При $n \longrightarrow \square$ правая часть неравенства стремится к нулю для любого ε > 0, т.е. $\lim_{n\to\infty}\left(p\left(\left|\bar{X} - a\right| \geq \varepsilon\right)\right) = 0$

и, следовательно, величина *x*, представляющая выборочную оценку, стремится к оцениваемому параметру *a* по вероятности.

- Таким образом, можно сделать вывод, что выборочное среднее является несмещенной, эффективной (по крайней мере, для нормального распределения) и состоятельной оценкой математического ожидания случайной величины, связанной с генеральной совокупностью.

- **ЛЕКЦИЯ 6**

2.4. Свойства выборочной дисперсии

- Исследуем несмещенность выборочной дисперсии D^* как оценки дисперсии случайной величины

По аналогии представим $\bar{x}$ и $\overline{x^2}$ как суммы случайных величин:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\xi_i;\ \ \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n}\xi_i^2.$$

Согласно $D^* = \overline{X^2} - \left(\bar{X}\right)^2$, поэтому

$$M\left[D^*\right] = M\left[\overline{X^2} - \left(\bar{X}\right)^2\right] = M\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right)^2\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}M\left[\xi_i^2\right] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}M\left[\xi_i\xi_j\right].$$

Разобьем двойную сумму на две, выделив суммирование по $i = j$:

$$M\left[D^*\right] = \frac{1}{n}\sum_{i=1}^{n}M\left[\xi_i^2\right] - \frac{1}{n^2}\sum_{i=1}^{n}M\left[\xi_i^2\right] - \frac{1}{n^2}\sum_{i\neq j}M\left[\xi_i\xi_j\right].$$

Поскольку ξ_i и ξ_j – независимые случайные величины, то имеем $M[\xi_i\xi_j] = M[\xi_i]\,M[\xi_j]$ Дисперсия случайной величины ξ равна $\sigma^2 = \alpha_2 - (\alpha_1)^2$, где $\alpha_1 = M[\xi_i]$ и $\alpha_2 = M[\xi_i{}^2]$ – соответственно первый и второй начальные моменты. Тогда получим

$$M\left[D^*\right] = \frac{1}{n}\sum_{i=1}^{n}\alpha_2 - \frac{1}{n^2}\sum_{i=1}^{n}\alpha_2 - \frac{1}{n^2}\sum_{i\neq j}\alpha_1^2 = \left(\frac{n-1}{n}\right)\alpha_2 - \left(\frac{n-1}{n}\right)\alpha_1^2 =$$

$$= \left(\frac{n-1}{n}\right)\left(\alpha_2 - \alpha_1^2\right) = \left(\frac{n-1}{n}\right)\sigma^2.$$

Итак, $M[D^*] \neq \sigma^2$, т.е D^* – смещенная оценка дисперсии случайной величины ξ. Однако,

$$\lim_{n\to\infty} M\left[D^*\right] = \lim_{n\to\infty}\left(\frac{n-1}{n}\right)\sigma^2 = \sigma^2,$$

что означает асимптотическую несмещенность этой оценки.

Примечания:

1. Можно предложить другую оценку дисперсии – *исправленную выборочную* дисперсию s^2, вычисляемую по формуле

$$s^2 = \frac{n}{n-1} D^* = \frac{1}{n-1} \sum_{i=1}^{k} n_i (x_i - \overline{x})^2.$$

Такая оценка будет несмещенной. Ей соответствует *исправленное выборочное* среднее квадратическое отклонение

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{k} n_i (x_i - \overline{x})^2}.$$

2. Исправленная выборочная дисперсия s^2 есть несмещенная оценка дисперсии случайной величины при неизвестном математическом ожидании. Если же математическое ожидание известно ($a = a_0$), то несмещенной оценкой дисперсии будет выборочная дисперсия

$$s_0^2 = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - a_0)^2.$$

Пример

- Найти выборочное среднее, выборочную дисперсию и среднее квадратическое отклонение, моду и исправленную выборочную дисперсию для выборки, имеющей следующий закон распределения:

x_i	n_i	ω_i	x_i	n_i	ω_i
1	2	2/30	5	5	5/30
2	4	4/30	6	3	3/30
3	5	5/30	7	1	1/30
4	9	9/30	8	1	1/30

- Решение:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{30}(2\cdot 1 + 4\cdot 2 + 5\cdot 3 + 9\cdot 4 + 5\cdot 5 + 3\cdot 6 + 1\cdot 7 + 1\cdot 8) = 3{,}97;$$

$$D^* = \overline{X^2} - \left(\bar{X}\right)^2$$

$$\overline{x^2} = \frac{1}{30}(2\cdot 1^2 + 4\cdot 2^2 + 5\cdot 3^2 + 9\cdot 4^2 + 5\cdot 5^2 + 3\cdot 6^2 + 1\cdot 7^2 + 1\cdot 8^2) = 18{,}43$$

и по формуле найдем $D^* = 18{,}43 - 15{,}76 = 2{,}67$ и $\sigma^* = \sqrt{2{,}67} = 1{,}63$.

Мода M_0, т.е. варианта, имеющая наибольшую частоту, равна 4.

Исправленная выборочная дисперсия $s^2 = n/(n-1)\, D^* = 2{,}76$.

ГЛАВА 3. ТОЧЕЧНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ ИЗВЕСТНОГО РАСПРЕДЕЛЕНИЯ

- Будем считать, что общий вид закона распределения нам известен и остается уточнить детали – параметры, определяющие его действительную форму. Существует несколько методов решения этой задачи, два из которых мы рассмотрим: метод моментов и метод наибольшего правдоподобия

3.1. Метод моментов

Напомним, что закон распределения непрерывной случайной величины ξ описывается плотностью распределения вероятностей $f_\xi(x, \boldsymbol{\theta})$, а дискретной величины – таблицей, связывающей значения x_i с их вероятностями $p(\xi = x_i, \boldsymbol{\theta})$. В обоих случаях $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ – вектор (набор) параметров. Например, для нормального распределения имеем два ($r = 2$) параметра: $\theta_1 = a$ и $\theta_2 = \sigma$, для распределения Пуассона – только один ($r = 1$): $\theta_1 = \lambda$.

Очевидно, что если закон распределения случайной величины зависит от параметров $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$, то и ее моменты зависят от этих же параметров, т.е. $\alpha_k = \alpha_k(\boldsymbol{\theta})$ и $\beta_k = \beta_k(\boldsymbol{\theta})$.

Можно показать, что при объеме выборки $n \to \infty$ выборочные моменты α_k^* и β_k^* сходятся по вероятности соответственно к $\alpha_k(\boldsymbol{\theta})$ и $\beta_k(\boldsymbol{\theta})$. Поэтому для достаточно больших выборок можно с большой вероятностью полагать, что для всех $k \geq 1$

$$\alpha_k(\boldsymbol{\theta}) \approx \alpha_k^*; \quad \beta_k(\boldsymbol{\theta}) \approx \beta_k^*.$$

- *Метод моментов*, развитый Карлом Пирсоном в 1894 г., основан на использовании этих приближенных равенств: моменты $\alpha_k(\theta)$, $\beta_k(\theta)$ рассчитываются теоретически по известному закону распределения с параметрами θ, а выборочные моменты α_k^*, β_k^* вычисляются по имеющейся выборке. Неизвестные параметры $\theta = (\theta_1, \theta_2, ..., \theta_r)$ определяются в результате решения системы из r уравнений, связывающих соответствующие теоретический и эмпирический моменты, например, $\alpha_k(\theta) = \alpha_k^*$ или $\beta_k(\theta) = \beta_k^*$.

- Можно показать, что оценки параметров θ, полученные методом моментов, состоятельны, их математические ожидания отличаются от истинных значений параметров на величину порядка n^{-1}, а средние квадратические отклонения являются величинами порядка $n^{-0,5}$

Пример

- Известно, что характеристика ξ объектов генеральной совокупности, являясь случайной величиной, имеет равномерное распределение, зависящее от параметров a и b:

$$f_\xi(x,a,b)=\begin{cases}0, & \text{если } x\in(-\infty,a);\\ \dfrac{1}{b-a}, & \text{если } x\in[a,b];\\ 0, & \text{если } x\in(b,\infty).\end{cases}$$

- Требуется определить методом моментов параметры a и b по известному выборочному среднему $\overline{x}$ и выборочной дисперсии $D^* = (\sigma^*)^2$

Напоминание

Начальные α_k и центральные β_k моменты определяются выражениями:

$$\alpha_k = M[\xi^k] = \int_{-\infty}^{\infty} x^k f(x)\mathrm{d}x, \ (k = 1, 2, \ldots);$$

$$\beta_k = M[\dot{\xi}^k] = \int_{-\infty}^{\infty} (x - m_\xi)^k f(x)\mathrm{d}x, \ (k = 1, 2, \ldots).$$

α_1 — мат.ожидание β_2 - дисперсия

Решение. Неизвестных параметров два: a, b. Выборочное среднее равно первому выборочному моменту $\bar{x} = \alpha_1^*$, а выборочная дисперсия – второму центральному моменту $D^* = \beta_2^*$. Поэтому в методе моментов (здесь $\boldsymbol{\theta} = (a, b)$) будем использовать два уравнения:

$$\begin{cases} \alpha_1(a,b) = \alpha_1^*; \\ \beta_2(a,b) = \beta_2^*. \end{cases} \qquad (*)$$

Первый начальный момент α_1, равный математическому ожиданию, и второй центральный момент β_2, равный дисперсии, в случае равномерного распределения имеют вид (см. п. 7.5 части I):

$$\alpha_1(a,b) = \frac{a+b}{2}; \qquad \beta_2(a,b) = \frac{(b-a)^2}{12}.$$

Подставляя соотношения в систему (*), получим

$$\begin{cases} \dfrac{a+b}{2} = \bar{x}; \\ \dfrac{(b-a)^2}{12} = D^* = \left(\sigma^*\right)^2. \end{cases}$$

Решая систему, найдем

$$\begin{cases} a = \bar{x} - \sqrt{3}\sigma^*; \\ b = \bar{x} + \sqrt{3}\sigma^*. \end{cases}$$

3.2. Метод наибольшего правдоподобия

- В основе метода лежит *функция правдоподобия* $L(x_1, x_2,..., x_n, \theta)$, являющаяся законом распределения вектора $\xi = (\xi_1, \xi_2, ..., \xi_n)$, где случайные величины ξ_i принимают значения вариант выборки, т.е. имеют одинаковое распределение. Поскольку случайные величины ξ_i независимы, функция правдоподобия имеет вид:

$$L(x_1, x_2, \ldots, x_n, \theta) = \begin{cases} \prod_{i=1}^{n} f_{\xi_i}(x_i, \theta), & \text{если } \xi_i \text{ непрерывны;} \\ \prod_{i=1}^{n} p_{\xi i}(x_i, \theta), & \text{если } \xi_i \text{ дискретны.} \end{cases}$$

- Идея *метода наибольшего правдоподобия* состоит в том, что мы ищем такие значения параметров **θ**, при которых вероятность появления в выборке значений вариант $x_1, x_2, \ldots, x_n$ является наибольшей. Иными словами, в качестве оценки параметров **θ** берется вектор $\tilde{\theta}$,при котором функция правдоподобия имеет локальный максимум при заданных $x_1, x_2, \ldots, x_n$:

$$L(x_1, x_2, \ldots, x_n, \tilde{\theta}) = \max_{\theta} \{L(x_1, x_2, \ldots, x_n, \theta)\}$$

- Оценки по методу максимального правдоподобия получаются из необходимого условия экстремума функции $L(x_1,x_2,\ldots,x_n,\boldsymbol{\theta})$ в точке $\theta=\tilde{\theta}$:

$$\begin{cases} \dfrac{\partial}{\partial\theta_1}L(x_1,x_2,\ldots,x_n,\theta)\Big|_{\theta=\tilde{\theta}}=0; \\ \dfrac{\partial}{\partial\theta_2}L(x_1,x_2,\ldots,x_n,\theta)\Big|_{\theta=\tilde{\theta}}=0; \\ \ldots \\ \dfrac{\partial}{\partial\theta_r}L(x_1,x_2,\ldots,x_n,\theta)\Big|_{\theta=\tilde{\theta}}=0. \end{cases}$$

Примечания:

- 1. При поиске максимума функции правдоподобия для упрощения расчетов можно выполнить действия, не изменяющие результата: во-первых, использовать вместо $L(x_1, x_2,..., x_n,\boldsymbol{\theta})$ *логарифмическую* функцию правдоподобия $l(x_1, x_2,..., x_n,\boldsymbol{\theta}) = \ln L(x_1, x_2,..., x_n,\boldsymbol{\theta})$; во-вторых, отбросить в выражении для функции правдоподобия не зависящие от $\boldsymbol{\theta}$ слагаемые (для l) или положительные сомножители (для L).
- 2. Оценки параметров, рассмотренные нами, можно назвать *точечными оценками*, так как для неизвестного параметра $\boldsymbol{\theta}$ определяется одна единственная точка $\tilde{\theta}$, являющаяся его приближенным значением. Однако такой подход может приводить к грубым ошибкам, и точечная оценка может значительно отличаться от истинного значения оцениваемого параметра (особенно в случае выборки малого объема).

Пример

$$f_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

Для характеристики ξ объектов генеральной совокупности, имеющей нормальную функцию распределения $F_\xi(x) = F_N(x, a, \sigma)$, методом максимального правдоподобия найти математическое ожидание a и дисперсию σ^2.

- **Решение.** В данной задаче следует оценить два неизвестных параметра: a и σ^2.
- Логарифмическая функция правдоподобия имеет вид

$$l(\mathbf{x}, a, \sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \sum_{i=1}^{n}\frac{(x_i - a)^2}{2\sigma^2}$$

- Отбросив в этой формуле слагаемое, которое не зависит от a и σ^2, составим систему уравнений правдоподобия

$$\begin{cases} \dfrac{\partial}{\partial a} l(\mathbf{x}, a, \sigma^2)\Big|_{a=\tilde{a}, \sigma^2=\tilde{\sigma}^2} = \displaystyle\sum_{i=1}^{n} \frac{x_i - \tilde{a}}{\tilde{\sigma}^2} = 0, \\ \dfrac{\partial}{\partial \sigma^2} l(\mathbf{x}, a, \sigma^2)\Big|_{a=\tilde{a}, \sigma^2=\tilde{\sigma}^2} = \displaystyle\sum_{i=1}^{n} \frac{(x_i - \tilde{a})^2}{2\left(\tilde{\sigma}^2\right)^2} - \frac{n}{2\tilde{\sigma}^2} = 0 \end{cases}$$

- Решая, получаем:

$$\begin{cases} \tilde{a} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i; \\ \tilde{\sigma}^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \tilde{a})^2 \end{cases}$$

ГЛАВА 4. ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ ИЗВЕСТНОГО РАСПРЕДЕЛЕНИЯ

- Задачу оценивания параметра известного распределения можно решать путем построения интервала, в который с заданной вероятностью попадает истинное значение параметра. Такой метод оценивания называется *интервальной оценкой*.
- Обычно в математике для оценки $\tilde{\theta}$ параметра θ строится неравенство

$$\left|\tilde{\theta} - \theta\right| < \delta \qquad (*)$$

- где число δ характеризует точность оценки: чем меньше δ, тем лучше оценка.

Статистические методы, однако, позволяют говорить только о том, что неравенство (*) выполняется с некоторой вероятностью. Поэтому «хорошей» оценкой здесь нужно считать построение достаточно узкого интервала (т.е. с достаточно малым δ), в который параметр θ попадает с достаточно большой вероятностью γ:

$$p\left(\left|\tilde{\theta}-\theta\right|<\delta\right)=p\left(\tilde{\theta}-\delta<\theta<\tilde{\theta}+\delta\right)=\gamma. \qquad (4.2)$$

С соотношением (4.2) связаны следующие термины:

1) γ – вероятность, с которой выполняется неравенство, называется *надежностью (доверительной вероятностью)* оценки $\tilde{\theta}$ параметра θ;

2) $\alpha = (1-\gamma)$ – вероятность противоположенного события, которую называют *уровнем значимости*;

3) интервал $\left(\tilde{\theta}-\delta<\theta<\tilde{\theta}+\delta\right)$, в который попадает неизвестный параметр θ с заданной надежностью γ, является *доверительным интервалом*.

4.1. Оценивание математического ожидания нормально распределенной величины при известной дисперсии

- Пусть исследуемая случайная величина ξ распределена по нормальному закону с известным средним квадратическим отклонением σ и неизвестным математическим ожиданием a. Требуется по значению выборочного среднего $\overline{x}$ оценить математическое ожидание ξ.
- Как и ранее, будем рассматривать получаемое выборочное среднее $\overline{x}$ как значение случайной величины $\overline{X}$, а значения вариант выборки $x_1, x_2, \ldots, x_n$ – соответственно как значения одинаково распределенных независимых случайных величин $\xi_1, \xi_2, \ldots, \xi_n$дая из которых имеет мат. ожи-дание a и среднее квадратическое отклонение σ.

- Имеем:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\xi_i; \quad M\left[\bar{X}\right] = \frac{1}{n}\sum_{i=1}^{n}M[\xi_i] = a;$$

$$\sigma\left[\bar{X}\right] = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}D[\xi_i]} = \frac{\sigma}{\sqrt{n}}. \qquad (1)$$

При достаточно большом n согласно центральной предельной теореме случайная величина $\left(n\bar{X} - nM[\bar{X}]\right) / n\sigma[\bar{X}]$ имеет практически нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Тогда вероятность попадания этой величины в интервал $(-x, x)$ равна

$$p\left(\left|\frac{\bar{X} - M[\bar{X}]}{\sigma[\bar{X}]}\right| < x\right) = 2\Phi(x), \qquad (2)$$

где $\Phi(x)$ – функция Лапласа.

Пусть $\tilde{x}_\gamma$ – число, при котором $\Phi(\tilde{x}_\gamma) = \gamma/2$. Тогда для данной выборки, где величина $\overline{X}$ принимает значение $\overline{x}$, из формулы (2) с учетом (1) после несложных преобразований получим

$$\begin{cases} p\left(\overline{x} - \dfrac{\sigma\tilde{x}_\gamma}{\sqrt{n}} < a < \overline{x} + \dfrac{\sigma\tilde{x}_\gamma}{\sqrt{n}}\right) = 2\Phi(\tilde{x}_\gamma), \\ 2\Phi(\tilde{x}_\gamma) = \gamma. \end{cases} \qquad (*)$$

Система $(*)$ связывает надежность γ и доверительный интервал для математического ожидания, равный

$$\left(\overline{x} - \frac{\sigma\tilde{x}_\gamma}{\sqrt{n}}, \quad \overline{x} + \frac{\sigma\tilde{x}_\gamma}{\sqrt{n}}\right).$$

4.2. Оценивание математического ожидания нормально распределенной величины при неизвестной дисперсии

Пусть исследуемая случайная величина ξ распределена по нормальному закону с неизвестными математическим ожиданием a и средним квадратическим отклонением σ.

Используя величины $\xi_1, \xi_2, \ldots, \xi_n$, определенные в п. 4.1, введем случайную величину Z, принимающую значения исправленной выборочной дисперсии s^2:

$$Z = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - a)^2.$$

Рассмотрим случайную величину

$$t_n = \frac{\overline{X} - a}{\sqrt{Z/n}},$$

где $\overline{X}$ – случайная величина, определенная в формулах (4.3), a – неизвестное математическое ожидание, n – объем выборки.

- Известно, что случайная величина t_n, заданная таким образом, имеет *распределение Стьюдента* с $k = n - 1$ степенями свободы. Плотность распределения вероятностей такой величины есть

$$f_t(x, n-1) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}$$

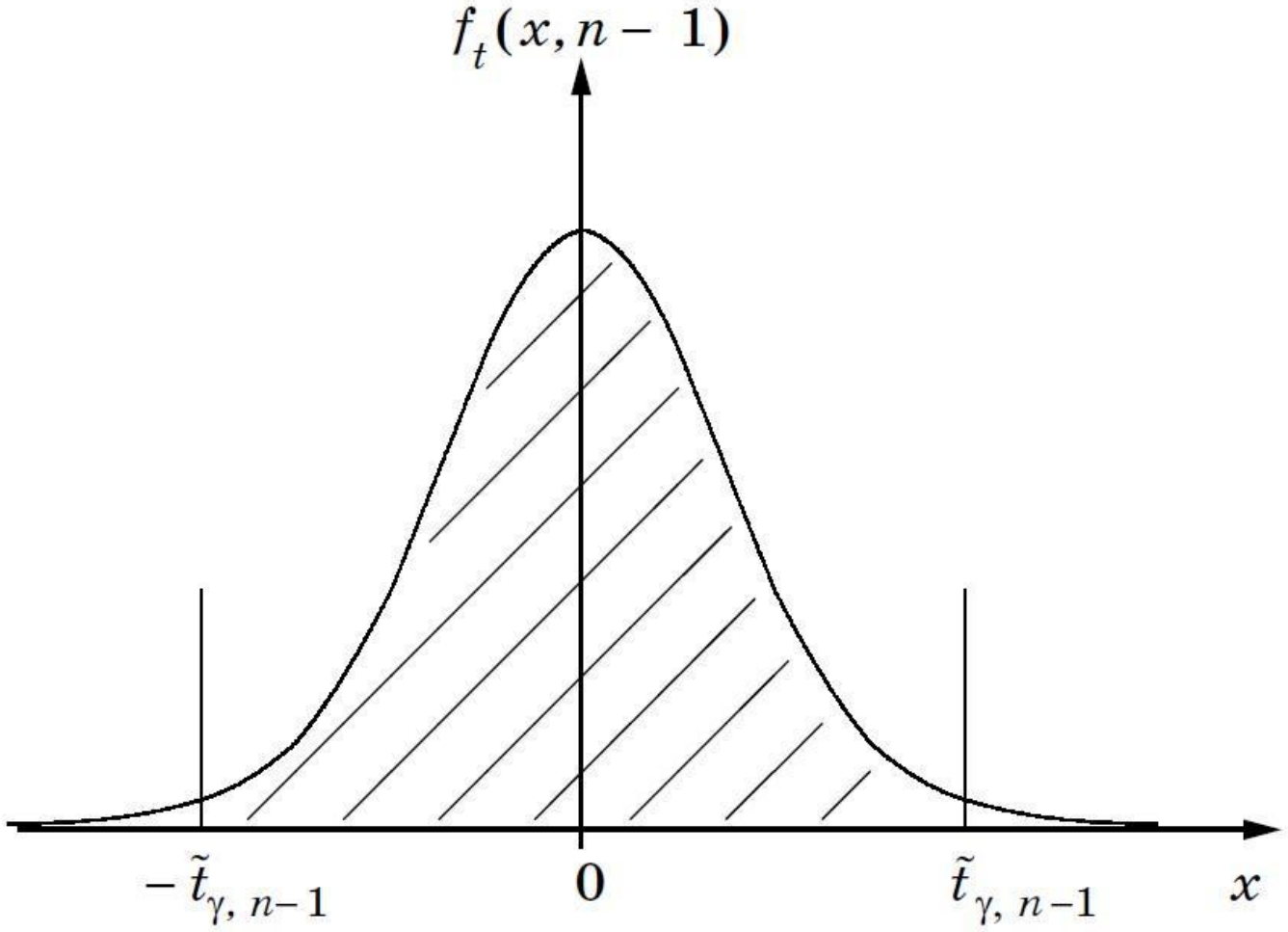
где $\Gamma(x)$ – гамма-функция

Пусть число $\tilde{t}_{\gamma,n-1}$ определяется следующим соотношением, где учтена четность функции $f_t(x,\, n-1)$:

$$p\left(\left|t_n\right|<\tilde{t}_{\gamma,n-1}\right)=\int\limits_{-\tilde{t}_{\gamma,n-1}}^{\tilde{t}_{\gamma,n-1}} f_t(x,n-1)\,\mathrm{d}x=2\int\limits_{0}^{\tilde{t}_{\gamma,n-1}} f_t(x,n-1)\,\mathrm{d}x=\gamma.$$

Отсюда для данной выборки, где случайная величина $\bar{X}$ принимает значение выборочного среднего $\bar{x}$, а Z – исправленной выборочной дисперсии s^2, получим

$$p\left(\bar{x}-\frac{s\tilde{t}_{\gamma,n-1}}{\sqrt{n}}<a<\bar{x}+\frac{s\tilde{t}_{\gamma,n-1}}{\sqrt{n}}\right)=\gamma.$$



Плотность распределения Стьюдента с $n-1$ степенями свободы

В зависимости от имеющихся таблиц параметр $\tilde{t}_{\gamma,n-1}$ находят следующими способами:

1) как квантиль распределения Стьюдента с $n-1$ степенями свободы, т.е. как число, для которого

$$p\left(t_n < \tilde{t}_{\gamma,n-1}\right) = \gamma + \frac{1-\gamma}{2} = \frac{1+\gamma}{2},$$

и тогда $\tilde{t}_{\gamma,n-1} = t_{(1+\gamma)/2,n-1}$, где $t_{(1+\gamma)/2,n-1}$ – квантиль порядка $(1+\gamma)/2$, определяемый по таблицам квантилей данного распределения

2) как критическую точку распределения Стьюдента для уровня значимости $(1-\gamma)/2$, т.е. как число, для которого

$$p\left(t_n > \tilde{t}_{\gamma,n-1}\right) = 1 - \frac{1+\gamma}{2} = \frac{1-\gamma}{2},$$

и тогда $\tilde{t}_{\gamma,n-1} = t_1((1-\gamma)/2, n-1)$, где $t_1((1-\gamma)/2, n-1)$ – критическая точка с уровнем значимости $(1-\gamma)/2$ для односторонней области, определяемая по таблицам с подобными данными;

3) как критическую точку распределения Стьюдента для уровня значимости $1-\gamma$:

$$p\left(|t_n| > \tilde{t}_{\gamma,n-1}\right) = 1-\gamma,$$

и тогда $\tilde{t}_{\gamma,n-1} = t_2(1-\gamma, n-1)$, где $t_2(1-\gamma, n-1)$ – критическая точка с уровнем значимости $(1-\gamma)$ для двусторонней области, определяемая по таблицам с соответствующими данными.

- **Примечание.** При большом числе степеней свободы *k* распределение Стьюдента стремится к нормальному распределению с нулевым математическим ожиданием и единичной дисперсией. Поэтому при $k \geq 30$ доверительный интервал можно на практике находить по формулам

$$\begin{cases} p\left(\bar{x} - \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}} < a < \bar{x} + \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}}\right) = 2\Phi(\tilde{x}_\gamma), \\ 2\Phi(\tilde{x}_\gamma) = \gamma. \end{cases}$$

4.3. Оценивание среднего квадратического отклонения нормально распределенной величины

- Пусть исследуемая случайная величина ξ распределена по нормальному закону с математическим ожиданием a и неизвестным средним квадратическим отклонением σ.
- Рассмотрим два случая: с известным и неизвестным математическим ожиданием.

4.3.1. Частный случай известного математического ожидания

- Пусть известно значение $M[\xi] = a$ и требуется оценить только σ или дисперсию $D[\xi] = \sigma^2$. Напомним, что при известном мат. ожидании несмещенной оценкой дисперсии является выборочная дисперсия $D^* = (\sigma^*)^2$
- Используя величины $\xi_1, \xi_2, \ldots, \xi_n$, определенные выше, введем случайную величину Y, принимающую значения выборочной дисперсии D^*:

$$Y = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - a)^2$$

- Рассмотрим случайную величину

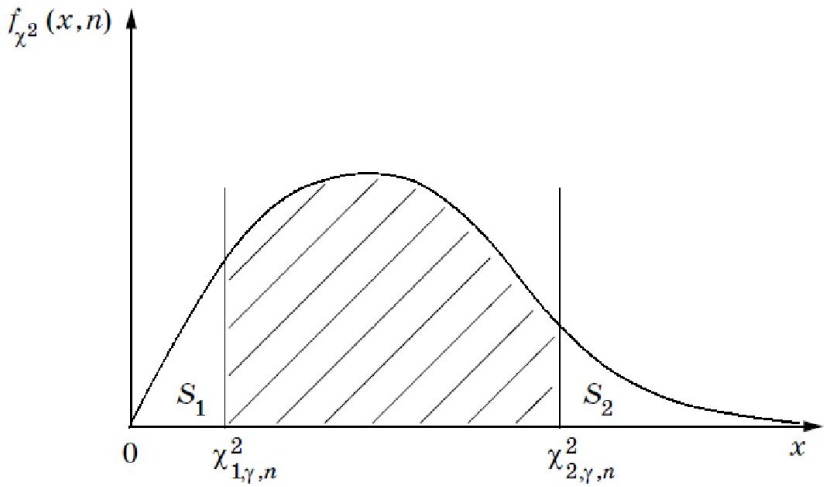
$$H_n = \frac{nY}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{\xi_i - a}{\sigma} \right)^2$$

- Стоящие под знаком суммы случайные величины $(\xi_i - a)/\sigma$ имеют нормальное распределение с плотностью $f_N(x, 0, 1)$. Тогда H_n имеет *распределение* χ^2 с n степенями свободы как сумма квадратов n независимых стандартных ($a = 0$, $\sigma = 1$) нормальных случайных величин.

- Определим доверительный интервал из условия

$$p\left(\tilde{\chi}^2_{1,\gamma,n} < H_n < \tilde{\chi}^2_{2,\gamma,n}\right) = \int_{\tilde{\chi}^2_{1,\gamma,n}}^{\tilde{\chi}^2_{2,\gamma,n}} f_{\chi^2}(x,n)\,\mathrm{d}x = \gamma$$

- где $f_{\chi^2}(x,n)$ – плотность распределения χ^2 и γ – надежность (доверительная вероятность). Величина γ численно равна площади заштрихованной фигуры на рис.



Плотность распределения χ^2 с n степенями свободы

Если для нахождения $\tilde{\chi}^2_{1,\gamma,n}$ и $\tilde{\chi}^2_{2,\gamma,n}$ используются таблицы квантилей распределения χ^2 (например, табл. 3 приложения), то значения $\tilde{\chi}^2_{1,\gamma,n}$ и $\tilde{\chi}^2_{2,\gamma,n}$ определяются по формулам:

$$p\left(H_n < \tilde{\chi}^2_{1,\gamma,n}\right) = \frac{1-\gamma}{2}; \quad p\left(H_n < \tilde{\chi}^2_{2,\gamma,n}\right) = \gamma + \frac{1-\gamma}{2} = \frac{1+\gamma}{2}, \quad (4.16)$$

т.е. $\tilde{\chi}^2_{1,\gamma,n} = \chi^2_{(1-\gamma)/2,n}$ и $\tilde{\chi}^2_{2,\gamma,n} = \chi^2_{(1+\gamma)/2,n}$, где $\chi^2_{\alpha,n}$ есть квантиль порядка α распределения χ^2 с n степенями свободы.

В случае, когда $\tilde{\chi}^2_{1,\gamma,n}$ и $\tilde{\chi}^2_{2,\gamma,n}$ находятся по таблицам критических точек, следует пользоваться соотношениями:

$$p\left(H_n > \tilde{\chi}^2_{1,\gamma,n}\right) = \frac{1+\gamma}{2}; \quad p\left(H_n > \tilde{\chi}^2_{2,\gamma,n}\right) = \frac{1-\gamma}{2}. \quad (4.17)$$

По известным $\tilde{\chi}^2_{1,\gamma,n}$ и $\tilde{\chi}^2_{2,\gamma,n}$ легко вычислить интервал, вероятность попадания дисперсии $D[\xi] = \sigma^2$ в который равна γ. Для данной выборки, где Y принимает значение выборочной дисперсии $D^* = (\sigma^*)^2$, из формулы (4.15) следует, что

$$p\left(\tilde{\chi}^2_{1,\gamma,n} < \frac{n\left(\sigma^*\right)^2}{\sigma^2} < \tilde{\chi}^2_{2,\gamma,n}\right) = p\left(\frac{n\left(\sigma^*\right)^2}{\tilde{\chi}^2_{2,\gamma,n}} < D[\xi] < \frac{n\left(\sigma^*\right)^2}{\tilde{\chi}^2_{1,\gamma,n}}\right) = \gamma,$$

т.е. доверительный интервал для дисперсии равен

$$\left(\frac{n\left(\sigma^*\right)^2}{\tilde{\chi}^2_{2,\gamma,n}}, \frac{n\left(\sigma^*\right)^2}{\tilde{\chi}^2_{1,\gamma,n}}\right).$$

Для среднего квадратического отклонения σ имеем

$$p\left(\sqrt{\frac{n}{\tilde{\chi}^2_{2,\gamma,n}}}\sigma^* < \sigma < \sqrt{\frac{n}{\tilde{\chi}^2_{1,\gamma,n}}}\sigma^*\right) = \gamma,$$

и доверительный интервал соответственно равен

$$\left(\sqrt{\frac{n}{\tilde{\chi}^2_{2,\gamma,n}}}\sigma^*, \sqrt{\frac{n}{\tilde{\chi}^2_{1,\gamma,n}}}\sigma^*\right).$$

4.3.2. Частный случай неизвестного математического ожидания

- На практике чаще всего встречается ситуация, когда неизвестны оба параметра нормального распределения: математическое ожидание a и среднее квадратическое отклонение σ.
- В этом случае построение доверительного интервала основывается на теореме Фишера, из кот. следует, что случайная величина $H_n = \frac{(n-1)Z}{\sigma^2}$
- (где случайная величина $Z = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - a)^2$) принимающая значения несмещенной выборочной дисперсии s^2, имеет распределение χ^2 с $n-1$ степенями свободы.

Рассмотрим равенство

$$p\left(\tilde{\chi}^2_{1,\gamma,n-1} < H_n < \tilde{\chi}^2_{2,\gamma,n-1}\right) = \gamma.$$

Для данной выборки, когда величина Z принимает значение s^2, после несложных алгебраических преобразований получим

$$p\left(\frac{(n-1)s^2}{\tilde{\chi}^2_{2,\gamma,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\tilde{\chi}^2_{1,\gamma,n-1}}\right) = \gamma.$$

Значения $\tilde{\chi}^2_{1,\gamma,n-1}$ и $\tilde{\chi}^2_{2,\gamma,n-1}$ определяются по таблицам квантилей или критических точек распределения χ^2 с $n-1$ степенью свободы

4.4. Оценивание математического ожидания случайной величины для произвольной выборки

- Интервальные оценки математического ожидания $M[\xi]$, полученные для нормально распределенной случайной величины ξ, являются, вообще говоря, непригодными для случайных величин, имеющих иной вид распределения. Однако есть ситуация, когда для любых случайных величин можно пользоваться подобными интервальными соотношениями, – это имеет место при выборке большого объема ($n >> 1$).

- Как и выше, будем рассматривать варианты $x_1, x_2, \ldots, x_n$ как значения независимых, одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_n$, имеющих математическое ожидание $M[\xi_i] = m_\xi$ и дисперсию $D[\xi_i] = \sigma_\xi^2$, а полученное выборочное среднее $\bar{x}$ как значение случайной величины $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$
- Согласно центральной предельной теореме величина $\bar{X}$ имеет асимптотически нормальный закон распределения с математическим ожиданием m_ξ и дисперсией σ_ξ^2 / n.

- Поэтому, если известно значение дисперсии случайной величины ξ, то можно пользоваться приближенными формулами

$$\begin{cases} p\left(\overline{x} - \dfrac{\sigma_\xi \tilde{x}_\gamma}{\sqrt{n}} < m_\xi < \overline{x} + \dfrac{\sigma_\xi \tilde{x}_\gamma}{\sqrt{n}}\right) \approx 2\Phi(\tilde{x}_\gamma) \\ 2\Phi(\tilde{x}_\gamma) = \gamma, \end{cases}$$

- Если же значение дисперсии величины ξ неизвестно, то при больших n можно использовать формулу

$$\begin{cases} p\left(\overline{x} - \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}} < m_\xi < \overline{x} + \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}}\right) \approx 2\Phi(\tilde{x}_\gamma) \\ 2\Phi(\tilde{x}_\gamma) = \gamma, \end{cases}$$

- где s – исправленное ср.-кв. отклонение

- **Лекция 7**

- **Повторение пройденного**

ГЛАВА 4. ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ ИЗВЕСТНОГО РАСПРЕДЕЛЕНИЯ

- Задачу оценивания параметра известного распределения можно решать путем построения интервала, в который с заданной вероятностью попадает истинное значение параметра. Такой метод оценивания называется *интервальной оценкой*.
- Обычно в математике для оценки $\tilde{\theta}$ параметра θ строится неравенство

$$\left|\tilde{\theta} - \theta\right| < \delta \qquad (*)$$

- где число δ характеризует точность оценки: чем меньше δ, тем лучше оценка.

Статистические методы, однако, позволяют говорить только о том, что неравенство (*) выполняется с некоторой вероятностью. Поэтому «хорошей» оценкой здесь нужно считать построение достаточно узкого интервала (т.е. с достаточно малым δ), в который параметр θ попадает с достаточно большой вероятностью γ:

$$p\left(\left|\tilde{\theta}-\theta\right|<\delta\right)=p\left(\tilde{\theta}-\delta<\theta<\tilde{\theta}+\delta\right)=\gamma. \qquad (4.2)$$

С соотношением (4.2) связаны следующие термины:

1) γ – вероятность, с которой выполняется неравенство, называется *надежностью (доверительной вероятностью)* оценки $\tilde{\theta}$ параметра θ;

2) $\alpha = (1-\gamma)$ – вероятность противоположенного события, которую называют *уровнем значимости*;

3) интервал $\left(\tilde{\theta}-\delta<\theta<\tilde{\theta}+\delta\right)$, в который попадает неизвестный параметр θ с заданной надежностью γ, является *доверительным интервалом*.

4.1. Оценивание математического ожидания нормально распределенной величины при известной дисперсии

- Пусть исследуемая случайная величина ξ распределена по нормальному закону с известным средним квадратическим отклонением σ и неизвестным математическим ожиданием a. Требуется по значению выборочного среднего $\overline{x}$ оценить математическое ожидание ξ.
- Как и ранее, будем рассматривать получаемое выборочное среднее $\overline{x}$ как значение случайной величины $\overline{X}$, а значения вариант выборки $x_1, x_2, \ldots, x_n$ – соответственно как значения одинаково распределенных независимых случайных величин $\xi_1, \xi_2, \ldots, \xi_n$дая из которых имеет мат. ожи-дание a и среднее квадратическое отклонение σ.

- Имеем:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\xi_i; \quad M\left[\bar{X}\right] = \frac{1}{n}\sum_{i=1}^{n}M[\xi_i] = a;$$

$$\sigma\left[\bar{X}\right] = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}D[\xi_i]} = \frac{\sigma}{\sqrt{n}}. \qquad (1)$$

При достаточно большом n согласно центральной предельной теореме случайная величина $\left(n\bar{X} - nM[\bar{X}]\right) / n\sigma[\bar{X}]$ имеет практически нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Тогда вероятность попадания этой величины в интервал $(-x, x)$ равна

$$p\left(\left|\frac{\bar{X} - M[\bar{X}]}{\sigma[\bar{X}]}\right| < x\right) = 2\Phi(x), \qquad (2)$$

где $\Phi(x)$ – функция Лапласа.

Пусть $\tilde{x}_\gamma$ – число, при котором $\Phi(\tilde{x}_\gamma) = \gamma/2$. Тогда для данной выборки, где величина $\overline{X}$ принимает значение $\overline{x}$, из формулы (2) с учетом (1) после несложных преобразований получим

$$\begin{cases} p\left(\overline{x} - \dfrac{\sigma\tilde{x}_\gamma}{\sqrt{n}} < a < \overline{x} + \dfrac{\sigma\tilde{x}_\gamma}{\sqrt{n}}\right) = 2\Phi(\tilde{x}_\gamma), \\ 2\Phi(\tilde{x}_\gamma) = \gamma. \end{cases} \qquad (*)$$

Система $(*)$ связывает надежность γ и доверительный интервал для математического ожидания, равный

$$\left(\overline{x} - \frac{\sigma\tilde{x}_\gamma}{\sqrt{n}}, \quad \overline{x} + \frac{\sigma\tilde{x}_\gamma}{\sqrt{n}}\right).$$

4.2. Оценивание математического ожидания нормально распределенной величины при неизвестной дисперсии

Пусть исследуемая случайная величина ξ распределена по нормальному закону с неизвестными математическим ожиданием a и средним квадратическим отклонением σ.

Используя величины $\xi_1, \xi_2, \ldots, \xi_n$, определенные в п. 4.1, введем случайную величину Z, принимающую значения исправленной выборочной дисперсии s^2:

$$Z = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - a)^2.$$

Рассмотрим случайную величину

$$t_n = \frac{\overline{X} - a}{\sqrt{Z/n}},$$

где $\overline{X}$ – случайная величина, определенная в формулах (4.3), a – неизвестное математическое ожидание, n – объем выборки.

- Известно, что случайная величина t_n, заданная таким образом, имеет *распределение Стьюдента* с $k = n - 1$ степенями свободы. Плотность распределения вероятностей такой величины есть

$$f_t(x, n-1) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}$$

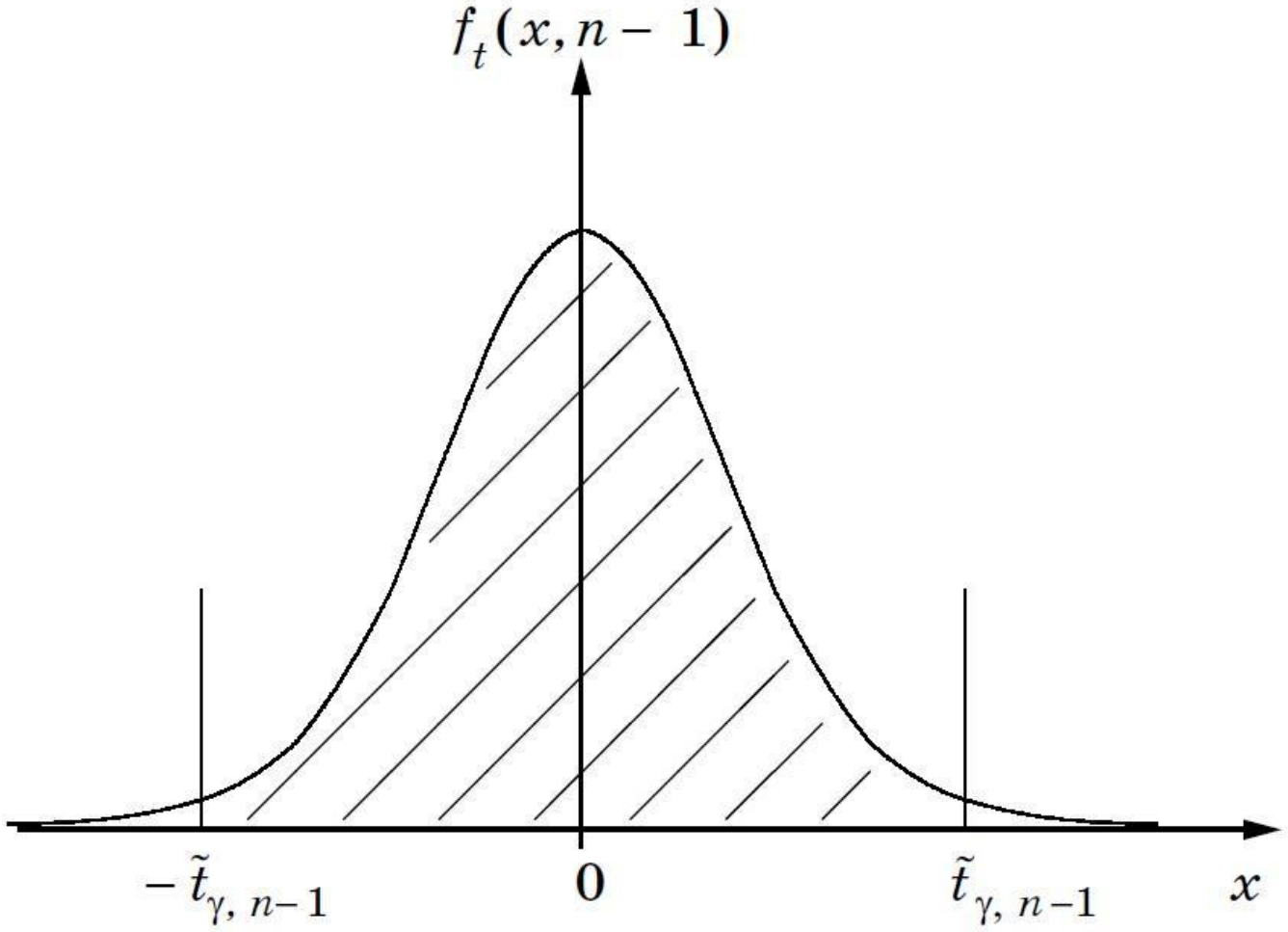
где $\Gamma(x)$ – гамма-функция

Пусть число $\tilde{t}_{\gamma,n-1}$ определяется следующим соотношением, где учтена четность функции $f_t(x,\, n-1)$:

$$p\left(\left|t_n\right|<\tilde{t}_{\gamma,n-1}\right)=\int\limits_{-\tilde{t}_{\gamma,n-1}}^{\tilde{t}_{\gamma,n-1}} f_t(x,n-1)\,\mathrm{d}x=2\int\limits_{0}^{\tilde{t}_{\gamma,n-1}} f_t(x,n-1)\,\mathrm{d}x=\gamma.$$

Отсюда для данной выборки, где случайная величина $\bar{X}$ принимает значение выборочного среднего $\bar{x}$, а Z – исправленной выборочной дисперсии s^2, получим

$$p\left(\bar{x}-\frac{s\tilde{t}_{\gamma,n-1}}{\sqrt{n}}<a<\bar{x}+\frac{s\tilde{t}_{\gamma,n-1}}{\sqrt{n}}\right)=\gamma.$$



Плотность распределения Стьюдента с $n-1$ степенями свободы

В зависимости от имеющихся таблиц параметр $\tilde{t}_{\gamma,n-1}$ находят следующими способами:

1) как квантиль распределения Стьюдента с $n-1$ степенями свободы, т.е. как число, для которого

$$p\left(t_n < \tilde{t}_{\gamma,n-1}\right) = \gamma + \frac{1-\gamma}{2} = \frac{1+\gamma}{2},$$

и тогда $\tilde{t}_{\gamma,n-1} = t_{(1+\gamma)/2,n-1}$, где $t_{(1+\gamma)/2,n-1}$ – квантиль порядка $(1+\gamma)/2$, определяемый по таблицам квантилей данного распределения

2) как критическую точку распределения Стьюдента для уровня значимости $(1-\gamma)/2$, т.е. как число, для которого

$$p\left(t_n > \tilde{t}_{\gamma,n-1}\right) = 1 - \frac{1+\gamma}{2} = \frac{1-\gamma}{2},$$

и тогда $\tilde{t}_{\gamma,n-1} = t_1((1-\gamma)/2, n-1)$, где $t_1((1-\gamma)/2, n-1)$ – критическая точка с уровнем значимости $(1-\gamma)/2$ для односторонней области, определяемая по таблицам с подобными данными;

3) как критическую точку распределения Стьюдента для уровня значимости $1-\gamma$:

$$p\left(|t_n| > \tilde{t}_{\gamma,n-1}\right) = 1-\gamma,$$

и тогда $\tilde{t}_{\gamma,n-1} = t_2(1-\gamma, n-1)$, где $t_2(1-\gamma, n-1)$ – критическая точка с уровнем значимости $(1-\gamma)$ для двусторонней области, определяемая по таблицам с соответствующими данными.

- **Примечание.** При большом числе степеней свободы *k* распределение Стьюдента стремится к нормальному распределению с нулевым математическим ожиданием и единичной дисперсией. Поэтому при $k \geq 30$ доверительный интервал можно на практике находить по формулам

$$\begin{cases} p\left(\bar{x} - \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}} < a < \bar{x} + \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}}\right) = 2\Phi(\tilde{x}_\gamma), \\ 2\Phi(\tilde{x}_\gamma) = \gamma. \end{cases}$$

4.3. Оценивание среднего квадратического отклонения нормально распределенной величины

- Пусть исследуемая случайная величина ξ распределена по нормальному закону с математическим ожиданием a и неизвестным средним квадратическим отклонением σ.
- Рассмотрим два случая: с известным и неизвестным математическим ожиданием.

4.3.1. Частный случай известного математического ожидания

- Пусть известно значение $M[\xi] = a$ и требуется оценить только σ или дисперсию $D[\xi] = \sigma^2$. Напомним, что при известном мат. ожидании несмещенной оценкой дисперсии является выборочная дисперсия $D^* = (\sigma^*)^2$
- Используя величины $\xi_1, \xi_2, \ldots, \xi_n$, определенные выше, введем случайную величину Y, принимающую значения выборочной дисперсии D^*:

$$Y = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - a)^2$$

- Рассмотрим случайную величину

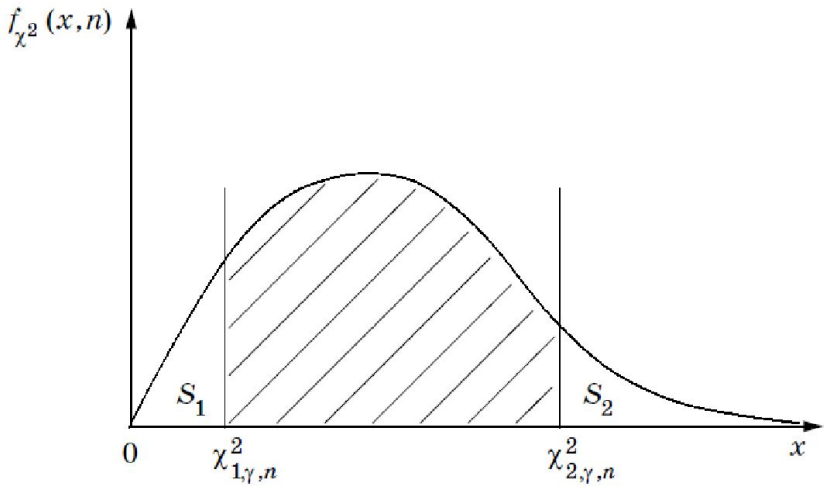
$$H_n = \frac{nY}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{\xi_i - a}{\sigma} \right)^2$$

- Стоящие под знаком суммы случайные величины $(\xi_i - a)/\sigma$ имеют нормальное распределение с плотностью $f_N(x, 0, 1)$. Тогда H_n имеет *распределение* χ^2 с n степенями свободы как сумма квадратов n независимых стандартных ($a = 0$, $\sigma = 1$) нормальных случайных величин.

- Определим доверительный интервал из условия

$$p\left(\tilde{\chi}^2_{1,\gamma,n} < H_n < \tilde{\chi}^2_{2,\gamma,n}\right) = \int_{\tilde{\chi}^2_{1,\gamma,n}}^{\tilde{\chi}^2_{2,\gamma,n}} f_{\chi^2}(x,n)\,\mathrm{d}x = \gamma$$

- где $f_{\chi^2}(x,n)$ – плотность распределения χ^2 и γ – надежность (доверительная вероятность). Величина γ численно равна площади заштрихованной фигуры на рис.



Плотность распределения χ^2 с n степенями свободы

Если для нахождения $\tilde{\chi}^2_{1,\gamma,n}$ и $\tilde{\chi}^2_{2,\gamma,n}$ используются таблицы квантилей распределения χ^2 (например, табл. 3 приложения), то значения $\tilde{\chi}^2_{1,\gamma,n}$ и $\tilde{\chi}^2_{2,\gamma,n}$ определяются по формулам:

$$p\left(H_n < \tilde{\chi}^2_{1,\gamma,n}\right) = \frac{1-\gamma}{2}; \quad p\left(H_n < \tilde{\chi}^2_{2,\gamma,n}\right) = \gamma + \frac{1-\gamma}{2} = \frac{1+\gamma}{2}, \quad (4.16)$$

т.е. $\tilde{\chi}^2_{1,\gamma,n} = \chi^2_{(1-\gamma)/2,n}$ и $\tilde{\chi}^2_{2,\gamma,n} = \chi^2_{(1+\gamma)/2,n}$, где $\chi^2_{\alpha,n}$ есть квантиль порядка α распределения χ^2 с n степенями свободы.

В случае, когда $\tilde{\chi}^2_{1,\gamma,n}$ и $\tilde{\chi}^2_{2,\gamma,n}$ находятся по таблицам критических точек, следует пользоваться соотношениями:

$$p\left(H_n > \tilde{\chi}^2_{1,\gamma,n}\right) = \frac{1+\gamma}{2}; \quad p\left(H_n > \tilde{\chi}^2_{2,\gamma,n}\right) = \frac{1-\gamma}{2}. \quad (4.17)$$

По известным $\tilde{\chi}^2_{1,\gamma,n}$ и $\tilde{\chi}^2_{2,\gamma,n}$ легко вычислить интервал, вероятность попадания дисперсии $D[\xi] = \sigma^2$ в который равна γ. Для данной выборки, где Y принимает значение выборочной дисперсии $D^* = (\sigma^*)^2$, из формулы (4.15) следует, что

$$p\left(\tilde{\chi}^2_{1,\gamma,n} < \frac{n\left(\sigma^*\right)^2}{\sigma^2} < \tilde{\chi}^2_{2,\gamma,n}\right) = p\left(\frac{n\left(\sigma^*\right)^2}{\tilde{\chi}^2_{2,\gamma,n}} < D[\xi] < \frac{n\left(\sigma^*\right)^2}{\tilde{\chi}^2_{1,\gamma,n}}\right) = \gamma,$$

т.е. доверительный интервал для дисперсии равен

$$\left(\frac{n\left(\sigma^*\right)^2}{\tilde{\chi}^2_{2,\gamma,n}}, \frac{n\left(\sigma^*\right)^2}{\tilde{\chi}^2_{1,\gamma,n}}\right).$$

Для среднего квадратического отклонения σ имеем

$$p\left(\sqrt{\frac{n}{\tilde{\chi}^2_{2,\gamma,n}}}\sigma^* < \sigma < \sqrt{\frac{n}{\tilde{\chi}^2_{1,\gamma,n}}}\sigma^*\right) = \gamma,$$

и доверительный интервал соответственно равен

$$\left(\sqrt{\frac{n}{\tilde{\chi}^2_{2,\gamma,n}}}\sigma^*, \sqrt{\frac{n}{\tilde{\chi}^2_{1,\gamma,n}}}\sigma^*\right).$$

4.3.2. Частный случай неизвестного математического ожидания

- На практике чаще всего встречается ситуация, когда неизвестны оба параметра нормального распределения: математическое ожидание a и среднее квадратическое отклонение σ.
- В этом случае построение доверительного интервала основывается на теореме Фишера, из кот. следует, что случайная величина $H_n = \frac{(n-1)Z}{\sigma^2}$
- (где случайная величина $Z = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - a)^2$) принимающая значения несмещенной выборочной дисперсии s^2, имеет распределение χ^2 с $n-1$ степенями свободы.

Рассмотрим равенство

$$p\left(\tilde{\chi}^2_{1,\gamma,n-1} < H_n < \tilde{\chi}^2_{2,\gamma,n-1}\right) = \gamma.$$

Для данной выборки, когда величина Z принимает значение s^2, после несложных алгебраических преобразований получим

$$p\left(\frac{(n-1)s^2}{\tilde{\chi}^2_{2,\gamma,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\tilde{\chi}^2_{1,\gamma,n-1}}\right) = \gamma.$$

Значения $\tilde{\chi}^2_{1,\gamma,n-1}$ и $\tilde{\chi}^2_{2,\gamma,n-1}$ определяются по таблицам квантилей или критических точек распределения χ^2 с $n-1$ степенью свободы

4.4. Оценивание математического ожидания случайной величины для произвольной выборки

- Интервальные оценки математического ожидания $M[\xi]$, полученные для нормально распределенной случайной величины ξ, являются, вообще говоря, непригодными для случайных величин, имеющих иной вид распределения. Однако есть ситуация, когда для любых случайных величин можно пользоваться подобными интервальными соотношениями, – это имеет место при выборке большого объема ($n >> 1$).

- Как и выше, будем рассматривать варианты $x_1, x_2, \ldots, x_n$ как значения независимых, одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_n$, имеющих математическое ожидание $M[\xi_i] = m_\xi$ и дисперсию $D[\xi_i] = \sigma_\xi^2$, а полученное выборочное среднее $\bar{x}$ как значение случайной величины $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$
- Согласно центральной предельной теореме величина $\bar{X}$ имеет асимптотически нормальный закон распределения с математическим ожиданием m_ξ и дисперсией σ_ξ^2 / n.

- Поэтому, если известно значение дисперсии случайной величины ξ, то можно пользоваться приближенными формулами

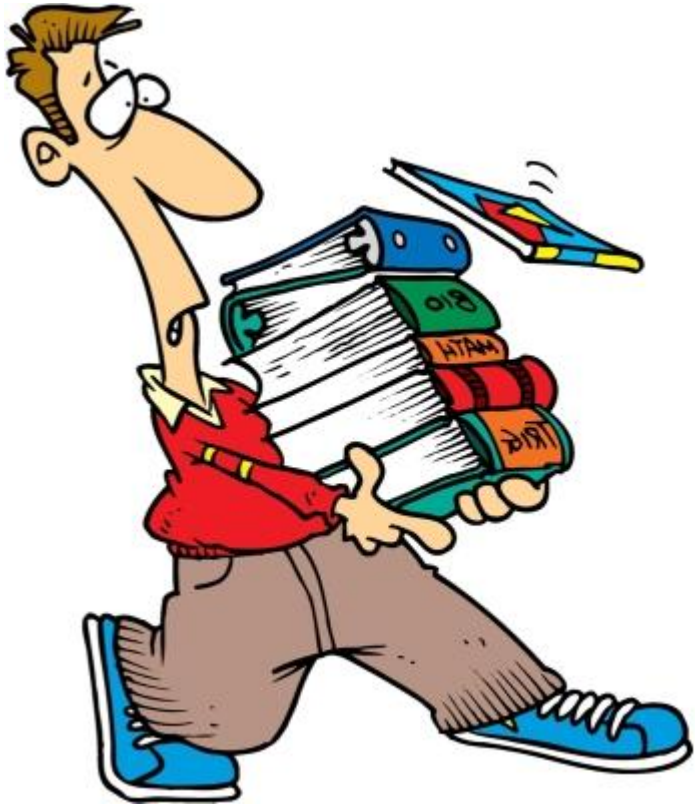
$$\begin{cases} p\left(\overline{x} - \dfrac{\sigma_\xi \tilde{x}_\gamma}{\sqrt{n}} < m_\xi < \overline{x} + \dfrac{\sigma_\xi \tilde{x}_\gamma}{\sqrt{n}}\right) \approx 2\Phi(\tilde{x}_\gamma) \\ 2\Phi(\tilde{x}_\gamma) = \gamma, \end{cases}$$

- Если же значение дисперсии величины ξ неизвестно, то при больших n можно использовать формулу

$$\begin{cases} p\left(\overline{x} - \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}} < m_\xi < \overline{x} + \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}}\right) \approx 2\Phi(\tilde{x}_\gamma) \\ 2\Phi(\tilde{x}_\gamma) = \gamma, \end{cases}$$

- где s – исправленное ср.-кв. отклонение

- **Повторили пройденное**



ГЛАВА 5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

- *Статистической гипотезой* называют гипотезу о виде неизвестного распределения или о параметрах известного распределения случайной величины.
- Проверяемая гипотеза, обозначаемая обычно как H_0, называется *нулевой* или *основной* гипотезы. Дополнительно используемая гипотеза H_1, противоречащая гипотезе H_0, называется *конкурирующей* или *альтернативной*.
- *Статистическая проверка* выдвинутой нулевой гипотезы H_0 состоит в ее сопоставлении с выборочными данными. При такой проверке возможно появление ошибок двух видов:
- а) *ошибки первого рода* – случаи, когда отвергается правильная гипотеза H_0;
- б) *ошибки второго рода* – случаи, когда принимается неверная гипотеза H_0.

- Вероятность ошибки первого рода будем называть *уровнем значимости* и обозначать как α.
- Основной прием проверки статистических гипотез заключается в том, что по имеющейся выборке вычисляется значение *статистического критерия* – некоторой случайной величины T, имеющей известный закон распределения. Область значений T, при которых основная гипотеза H_0 должна быть отвергнута, называют *критической*, а область значений T, при которых эту гипотезу можно принять, – *областью принятия гипотезы*.

Процесс проверки гипотезы H_0 состоит в следующем:

1) определяют статистический критерий T и по имеющейся выборке вычисляют его эмпирическое (наблюдаемое) значение T^*;

2) выбирают уровень значимости α и по известному закону распределения T вычисляют его критическое значение T_α. Оно делит область возможных значений T на две части: критическую область и область принятия гипотезы (например, области $T > T_\alpha$ и $T \le T_\alpha$);

3) если значение T^* попадает в область принятия гипотезы, то гипотезу H_0 можно принять; если в критическую область, то гипотезу H_0 следует отвергнуть.

5.1. Проверка гипотез о параметрах известного распределения

- **5.1.1. *Проверка гипотезы о математическом ожидании нормально распределенной случайной величины***

- Пусть случайная величина ξ имеет нормальное распределение.
- Требуется проверить предположение о том, что ее математическое ожидание равно некоторому числу a_0. Рассмотрим отдельно случаи, когда дисперсия ξ известна и когда она неизвестна.

- В случае известной дисперсии $D[\xi] = \sigma^2$, как и в п. 4.1, определим случайную величину $\overline{X}$, принимающую значения выборочного среднего $\overline{x}$. Гипотеза H_0 изначально формулируется как $M[\xi] = a_0$. Поскольку выборочное среднее является несмещенной оценкой $M[\xi]$, то гипотезу H_0 можно представить как

$$M\left[\overline{X}\right] = a_0$$

Для проверки этой гипотезы выберем следующий статистический критерий:

$$U = \frac{\bar{X} - a_0}{\sigma\left[\bar{X}\right]} = \frac{\left(\bar{X} - a_0\right)\sqrt{n}}{\sigma},$$

где учтено, что $\sigma\left[\bar{X}\right] = \sqrt{D[\bar{X}]}$ и в нашем случае $D\left[\bar{X}\right] = \sigma^2 / n$

Отметим, что если нулевая гипотеза справедлива, то при достаточно большом объеме выборки n согласно центральной предельной теореме случайная величина U должна иметь практически нормальное распределение с $M[U] = 0$ и $D[U] = 1$

Конфигурация критической области определяется видом конкурирующей гипотезы H_1, которая может состоять в том, что $M\left[\bar{X}\right] > a_0$, $M\left[\bar{X}\right] < a_0$ или $M\left[\bar{X}\right] \neq a_0$.

1. Если $M\left|\overline{X}\right| > a_0$, критическая область является *правосторонней*. На рис. подобная критическая область заштрихована.

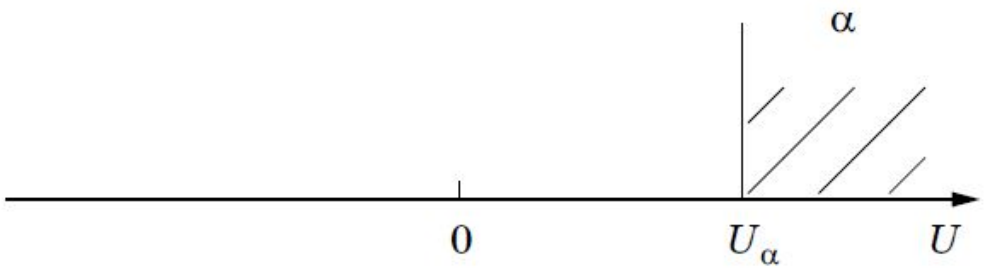
Поскольку можно считать, что случайная величина U распределена нормально, то критическое значение U_α можно найти из условия $p(U > U_\alpha) = \alpha$. Имеем $p(U > U_\alpha) = 1 - F_U(U_\alpha) = 1 - (\Phi(U_\alpha) + 1/2) = 1/2 - \Phi(U_\alpha)$, где $\Phi(x)$ – функция Лапласа, и получаем

$$\Phi(U_\alpha) = \frac{1}{2} - \alpha,$$

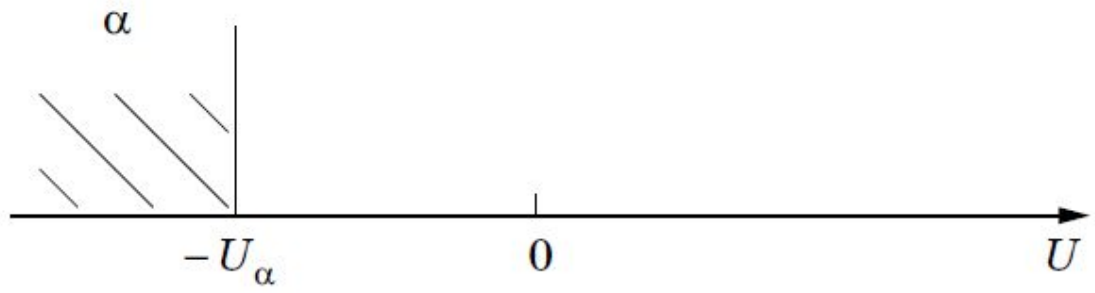
т.е. U_α равно квантили нормального распределения $x_{1-\alpha}$. Остается по выборке найти значение

$$U^* = \frac{(\overline{x} - a_0)\sqrt{n}}{\sigma}$$

и сравнить его с U_α. Если $U^* < U_\alpha$, то нулевую гипотезу H_0 можно принять, а если $U^* > U_\alpha$, то ее следует отвергнуть.



2. Если $M[\bar{X}] < a_0$, критическая область является *левосторонней* и U_α определяется из условия $p(U < -U_\alpha) = \alpha$.



Левосторонняя критическая область

Поскольку $p(U < -U_\alpha) = F_U(-U_\alpha) = \Phi(-U_\alpha) + 1/2 = 1/2 - \Phi(U_\alpha)$, то опять получаем

$$\Phi(U_\alpha) = \frac{1}{2} - \alpha. \qquad (5.3)$$

Здесь нулевую гипотезу H_0 можно принять, если $U^* > -U_\alpha$, и следует отвергнуть, если $U^* < -U_\alpha$.

3. Если $M[\bar{X}] \neq a_0$, критическая область удовлетворяет условию $p(|U| > U_\alpha) = \alpha$ и является *двусторонней*, поскольку состоит из двух частей: $U < -U_\alpha$ и $U > U_\alpha$ (рис. 5.3).

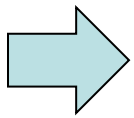
Вероятность попадания критерия U в каждую из половин критической области равна $\alpha/2$. Поэтому U_α определяется из соотношения $p(U > U_\alpha) = \alpha/2$. Поскольку $p(U > U_\alpha) = 1/2 - \Phi(U_\alpha)$, то получаем

$$\Phi(U_\alpha) = \frac{1-\alpha}{2}. \qquad (5.4)$$

Здесь нулевую гипотезу H_0 можно принять, если $|U^*| < U_\alpha$, и следует отвергнуть, если $|U^*| > U_\alpha$.

В случае неизвестной дисперсии величины ξ в качестве статистического критерия выберем

$$T = \frac{\left(\bar{X} - a_0\right)}{\sqrt{Z/n}}, \qquad (5.5)$$



$$T = \frac{\left(\bar{X} - a_0\right)}{\sqrt{Z / n}}, \qquad (5.5)$$

где случайные величины $\bar{X}$ и Z, определенные в п. 4.2, принимают для данной выборки значения соответственно выборочного среднего $\bar{x}$ и исправленной (несмещенной) выборочной дисперсии s^2.

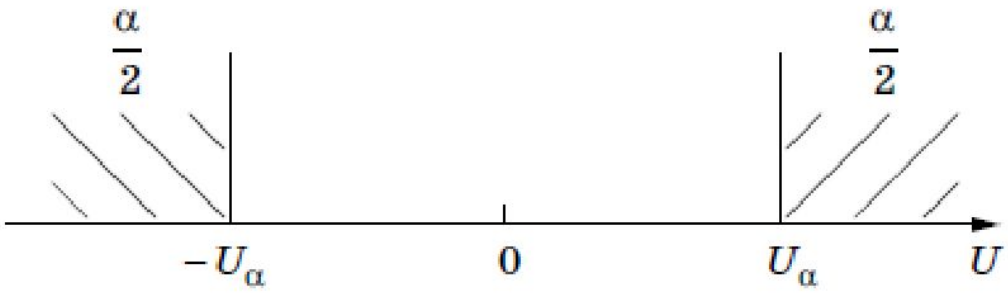


Рис. 5.3. Двусторонняя критическая область

Случайная величина T имеет распределение Стьюдента с $k = n - 1$ степенями свободы (см. подробнее п. 6.2).

Если конкурирующая (альтернативная) гипотеза H_1 базируется на неравенстве $M[X] > a_0$, то критическая область является правосторонней (см. п. 4.2), и согласно формуле (4.10) критическое значение равно $T_\alpha = \tilde{t}_{1-\alpha,n-1} = t_{1-\alpha/2,n-1}$, где $t_{1-\alpha/2,n-1}$ – квантиль порядка $(1-\alpha/2)$ распределения Стьюдента с $n-1$ степенями свободы, которую можно найти по соответствующим таблицам.

Эмпирическое значение критерия для данной выборки равно

$$T^* = \frac{(\bar{x} - a_0)\sqrt{n}}{s}. \qquad (5.6)$$

Нулевую гипотезу H_0 можно принять, если $|T^*| < T_\alpha$, и следует отвергнуть, если $|T^*| > T_\alpha$.

Аналогичным образом рассматриваются и остальные варианты выбора гипотезы H_1 (см. п. 4.2).

5.1.2. Сравнение дисперсий нормально распределенных случайных величин

- Пусть имеются две нормально распределенные случайные величины ξ_1 и ξ_2 Для них по независимым выборкам объемом n_1 и n_2 соответственно получены исправленные выборочные дисперсии $s^2_{\xi_1}$ и $s^2_{\xi_2}$ дем считать, что $s^2_{\xi_1} > s^2_{\xi_2}$. Требуется при заданном уровне значимости α проверить нулевую гипотезу H_0 о равенстве дисперсий рассматриваемых случайных величин.

- Учитывая несмещенность исправленных выборочных дисперсий, нулевую гипотезу можно записать следующим образом: $M[Z_{\xi_1}] = M[Z_{\xi_2}],$

  где случайная величина $Z_\xi = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - M[\xi])^2$

  принимает значения исправленной выборочной дисперсии величины ξ и аналогична случайной величине Z, рассмотренной в п. 4.2.
- В качестве статистического критерия выберем случайную величину $F = \frac{Z_{\xi_1}}{Z_{\xi_2}}$

  принимающую значение отношения бóльшей выборочной дисперсии к меньшей.

- Случайная величина *F* имеет распределение Фишера – Снедекора с числом степеней свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1 – объем выборки, по которой вычислена бóльшая исправленная дисперсия $s^2_{\xi_1}$, а n_2 – объем второй выборки, по которой найдена меньшая дисперсия $s^2_{\xi_2}$.
- Рассмотрим два вида конкурирующих гипотез $H_1 \left(s^2_{\xi_1} > s^2_{\xi_2} \text{ и } s^2_{\xi_1} \neq s^2_{\xi_2} \right)$

1. Если $s^2_{\xi_1} > s^2_{\xi_2}$, строят одностороннюю (правостороннюю) критическую область, исходя из условия

$$p(F > F(\alpha, k_1, k_2)) = \alpha\,.$$

Критическую точку $F(\alpha, k_1, k_2)$ находят по таблице значений таких точек для распределения Фишера – Снедекора с числами степеней свободы k_1 и k_2 при уровне значимости α. Затем для данных выборок вычисляют эмпирическое значение критерия

$$F^* = \frac{s^2_{\xi_1}}{s^2_{\xi_2}}$$

и сравнивают его с $F(\alpha, k_1, k_2)$. Если $F^* < F(\alpha, k_1, k_2)$, то нулевая гипотеза принимается, если $F^* > F(\alpha, k_1, k_2)$, то отвергается.

2. Если $s^2_{\xi_1} \neq s^2_{\xi_2}$, строят двустороннюю критическую область из условия, что $p\big((F < F_1) \cup (F > F_2)\big) = \alpha$.

Оказывается, что наибольшая мощность (вероятность попадания критерия в критическую область при справедливости конкурирующей гипотезы) достигается, когда вероятность попадания критерия в каждый из двух интервалов критической области равна $\alpha/2$:

$$p(F < F_1) = \frac{\alpha}{2}; \qquad p(F > F_2) = \frac{\alpha}{2}.$$

Критическая точка $F_2 = F(\alpha/2, k_1, k_2)$ находится по таблице критических точек для распределения Фишера – Снедекора с числами степеней свободы k_1 и k_2 при уровне значимости $\alpha/2$. Критическую точку F_1 можно и не отыскивать. Действительно, если вероятность попадания критерия в «правую часть» критической области равна $\alpha/2$, то и вероятность попадания в «левую часть» также равна $\alpha/2$. Так как эти события несовместны, то вероятность попадания рассматриваемого критерия во всю двустороннюю критическую область будет равна

$$\frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

Таким образом, если $F^* < F(\alpha/2, k_1, k_2)$, то нулевая гипотеза принимается, а при $F^* > F(\alpha/2, k_1, k_2)$ – отвергается.

5.1.3. Сравнение математических ожиданий независимых случайных величин

- Сначала рассмотрим случай нормального распределения случайных величин с известными дисперсиями, а затем на его основе – более общий случай произвольного распределения величин при достаточно больших независимых выборках.
- Пусть случайные величины ξ_1 и ξ_2 независимы и распределены нормально, и пусть их дисперсии $D[\xi_1]$ и $D[\xi_2]$ известны. (Например, они могут быть найдены из какого-то другого опыта или рассчитаны теоретически). Извлечены выборки объемом n_1 и n_2 соответственно. Пусть $\bar{x}_1$ и $\bar{x}_2$ – выборочные средние для этих выборок. Требуется по выборочным средним при заданном уровне значимости α проверить гипотезу о равенстве математических ожиданий рассматриваемых случайных величин

$$M[\xi_1] = M[\xi_2].$$

- Введем случайные величины $\bar{X}_1$ и $\bar{X}_2$, принимающие значения выборочных средних $\bar{x}_1$ и $\bar{x}_2$ ветственно. Поскольку выборочные средние – это несмещенные оценки математических ожиданий, нулевую гипотезу H_0 можно записать в следующем виде:

$$M[\bar{X}_1] = M[\bar{X}_2]$$

- В качестве статистического критерия для проверки H_0 возьмем случайную величину

$$U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{D[\xi_1]}{n_1} + \frac{D[\xi_2]}{n_2}}}$$

Можно показать, что при $n_1, n_2 \to \infty$ сумма $(D[\xi_1]/n_1 + D[\xi_2]/n_2)$ стремится к $D[\bar{X}_1 - \bar{X}_2]$. Тогда, если нулевая гипотеза (5.9) справедлива, то согласно центральной предельной теореме с ростом объемов выборок распределение величины U стремится к нормальному с нулевым математическим ожиданием и единичной дисперсией.

Критическую область строят в зависимости от вида конкурирующей гипотезы H_1 $\left(M[\xi_1] \neq M[\xi_2],\ M[\xi_1] > M[\xi_2],\ M[\xi_1] < M[\xi_2]\right)$.

1. Если $M[\xi_1] \neq M[\xi_2]$, строят двустороннюю критическую область. Как и в п. 5.1.1, критическая точка U_α находится из условия

$$\Phi(U_\alpha) = \frac{1-\alpha}{2}.$$

Затем по имеющимся выборкам вычисляют эмпирическое значение критерия

$$U^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{D[\xi_1]}{n_1} + \dfrac{D[\xi_2]}{n_2}}}$$

и сравнивают его с теоретическим значением U_α:

а) если $|U^*| < U_\alpha$, нет оснований отвергать нулевую гипотезу;

б) если $|U^*| > U_\alpha$, нулевую гипотезу отвергают.

2. Если $M[\xi_1] > M[\xi_2]$, строят правостороннюю критическую область. На практике такая ситуация возникает тогда, когда из априорных соображений можно ожидать, что математическое ожидание одной случайной величины должно быть больше, чем математическое ожидание другой.

Критическая точка U_α определяется из условия, аналогичного соотношению (5.2),

$$\Phi(U_\alpha) = \frac{1}{2} - \alpha. \qquad (5.12)$$

Вывод делается по результатам сравнения U^* и U_α:

а) если $U^* < U_\alpha$, нулевую гипотезу можно принять;

б) если $U^* > U_\alpha$, ее следует отвергнуть.

3. Если $M[\xi_1] < M[\xi_2]$, строят левостороннюю критическую область. Критическая точка U_α определяется, как и в п. 5.1.1, из условия

$$\Phi(U_\alpha) = \frac{1}{2} - \alpha. \qquad (5.13)$$

Нулевую гипотезу H_0 можно принять, если $U^* > -U_\alpha$, и следует отвергнуть, если $U^* < -U_\alpha$.

Сравним математические ожидания двух произвольно распределенных случайных величин ξ_1 и ξ_2 при больших ($n > 30$) независимых выборках.

В этом случае в силу центральной предельной теоремы выборочные средние распределены приближенно по нормальному закону вне зависимости от распределения самих случайных величин. Поскольку выборочные дисперсии являются при этом достаточно хорошими оценками дисперсий случайных величин, последние можно считать приближенно известными, т.е. $D[\xi_1] \approx D^*[\xi_1]$, $D[\xi_2] \approx D^*[\xi_2]$. Тогда можно полагать, что статистический критерий

$$\tilde{U} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{D^*[\xi_1]}{n_1} + \dfrac{D^*[\xi_2]}{n_2}}} \qquad (5.14)$$

имеет распределение, близкое к нормальному, причем:

а) если справедлива нулевая гипотеза $M[\xi_1] = M[\xi_2]$, то $M[\tilde{U}] \approx 0$;

б) если рассматриваемые выборки независимы, то $D[\tilde{U}] \approx 1$.

Дальнейшая проверка статистических гипотез проводится так же, как для критерия U.

5.2. Проверка гипотез о виде закона распределения случайной величины. Критерий Пирсона

- Надежное предположение о распределении случайной величины, связанной с генеральной совокупностью, можно иногда сделать из априорных соображений, основываясь на условиях эксперимента, и тогда предположения о параметрах распределения исследуются, как показано ранее. Однако весьма часто возникает необходимость проверить выдвинутую гипотезу о законе распределения.
- Статистические критерии, предназначенные для таких проверок, обычно называются *критериями согласия*.

- Известно несколько критериев согласия. Достоинством критерия Пирсона является его универсальность. С его помощью можно проверять гипотезы о различных законах распределения.
- *Критерий Пирсона* основан на сравнении частот, найденных по выборке (эмпирических частот), с частотами, рассчитанными с помощью проверяемого закона распределения (теоретическими частотами).
- Обычно эмпирические и теоретические частоты различаются. Следует выяснить, случайно ли расхождение частот или оно значимо и объясняется тем, что теоретические частоты вычислены исходя из неверной гипотезы о распределении генеральной совокупности.
- Критерий Пирсона, как и любой другой, отвечает на вопрос, есть ли согласие выдвинутой гипотезы с эмпирическими данными при заданном уровне значимости.

5.2.1. Проверка гипотезы о нормальном распределении

- Пусть имеется случайная величина ξ и сделана выборка достаточно большого объема *n* с большим количеством различных значений вариант. Требуется при уровне значимости α проверить нулевую гипотезу H_0 о том, что случайная величина ξ распределена нормально.
- Для удобства обработки выборки возьмем два числа α и β: $\alpha < x_1, \beta > x_n$ и разделим интервал [α, β] на s подинтервалов. Ьудем считать, что значения вариант, попавших в каждый подинтервал,приближенно равны числу, задающему середину подинтервала. Подсчитав число вариант, попавших в каждый интервал, составим группированную выборку с вариантами: x_1, x_2, ..., x_s и их частотами n_1, n_2, ..., n_s, где $x_j = (b_j + a_j)/2$ – середина j-го подинтервала $(a_j, b_j]$; n_j – количество вариант, попавших в этот подинтервал, т.е. эмпирическая частота.

По полученным данным можно вычислить выборочное среднее $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{s} n_j x_j$$

и выборочное среднее квадратическое отклонение σ^*:

$$\sigma^* = \sqrt{D^*} = \sqrt{\frac{1}{n}\sum_{j=1}^{s} n_j (x_j - \bar{x})^2}\,.$$

Найдем количество вариант, которое должно оказаться в каждом подинтервале при выборке объемом n в предположении, что случайная величина ξ распределена по нормальному закону с параметрами $M[\xi] = \bar{x}$, $D[\xi] = (\sigma^*)^2$.

Вероятность того, что значение такой величины попадет в j-й подинтервал, равна

$$p_j = \Phi\left(\frac{b_j - \bar{x}}{\sigma^*}\right) - \Phi\left(\frac{a_j - \bar{x}}{\sigma^*}\right), \qquad (5.15)$$

где a_j и b_j – границы подинтервала, $\Phi(x)$ – функция Лапласа. Умножив вероятности p_j на объем выборки n, найдем теоретические частоты $n'_j = np_j$.

Выберем статистический критерий, равный

$$H = \sum_{j=1}^{s} \frac{(\eta_j - n'_j)^2}{n_j}, \qquad (5.16)$$

где η_j – случайная величина, принимающая для данной выборки значение частоты n_j. Можно показать, что вне зависимости от реального закона распределения случайной величины ξ распределение H при $n \to \infty$ стремится к распределению χ^2 с числом степеней свободы $k = s - 1 - r$, где r – число параметров предполагаемого распределения, оцениваемых по выборке. В случае нормального распределения, характеризуемого двумя параметрами, $k = s - 3$.

Для выбранного критерия построим правостороннюю критическую область, определяемую условием

$$p(H > \chi^2(\alpha, k)) = \alpha,$$

где α – уровень значимости. Критическая точка $\chi^2(\alpha, k) = \chi^2_{1-\alpha,k}$, $\chi^2_{1-\alpha,k}$ – квантиль порядка $(1 - \alpha)$ распределения χ^2 с числом степеней свободы $k = s - 3$, которая определяется из соответствующих таблиц.

Значение критерия H^* для данной выборки равно

$$H^* = \sum_{j=1}^{s} \frac{(n_j - n'_j)^2}{n_j}. \qquad (5,17)$$

Если $H^* < \chi^2(\alpha, k)$, то нулевую гипотезу H_0 о том, что случайная величина ξ распределена нормально, можно принять с уровнем значимости α. Если $H^* > \chi^2(\alpha, k)$, то гипотезу H_0 необходимо отвергнуть.

Примечание. Объем выборки n должен быть достаточно большим, во всяком случае, не менее 50. Каждый подинтервал должен содержать не менее 5–8 вариант. Если в подинтервале слишком мало точек, то его следует объединить с соседним.

- **ГЛАВА 6. ВАЖНЕЙШИЕ РАСПРЕДЕЛЕНИЯ И ИХ КВАНТИЛИ**

6.1. Нормальное распределение

- По определению нормально распределенная случайная величина ξ имеет плотность распределения вероятностей

$$f_\xi(x) = f_N(x, a, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

- где a и σ являются параметрами.

- *Квантилью* порядка α (0 < α < 1) непрерывной случайной величины ξ называется такое число x_α, для которого выполняется равенство $p(\xi < x_\alpha) = \alpha$
- Квантиль $x_{1/2}$ называется *медианой* случайной величины ξ, *квантили* $x_{1/4}$ и $x_{3/4}$ – ее *квартилями*, а $x_{0,1}, x_{0,2}, \ldots, x_{0,9}$ – *децилями*.
- Для стандартного нормального распределения (a = 0, σ = 1) и, следовательно,

$$p(\xi < x_\alpha) = F_N(x_\alpha, 0, 1) = \frac{1}{2} + \Phi(x_\alpha)$$

- где $F_N(x, a, \sigma)$ – функция распределения нормально распределенной случайной величины, а $\Phi(x)$ – функция Лапласа.
- Квантиль стандартного нормального распределения x_α для заданного α можно найти из соотношения

$$\Phi(x_\alpha) = \alpha - \frac{1}{2}$$

6.2. Распределение Стьюдента

- Если $\xi_0, \xi_1, \xi_2, \ldots, \xi_n$ – независимые случайные величины, имеющие нормальное распределение с нулевым математическим ожиданием и единичной дисперсией, то распределение случайной величины

$$t_n = \frac{\xi_0}{\sqrt{\frac{(\xi_1^2 + \ldots + \xi_n^2)}{n}}}$$

- называют *распределением Стьюдента* с n степенями свободы (W.S. Gosset).

Плотность распределения вероятностей t_n равна (рис. 6.1)

$$f_t(x,n)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

где $\Gamma(x)$ – гамма-функция, имеющая следующее определение и значения:

$$\Gamma(x)=\int_0^\infty z^{x-1}\exp(-z)\,\mathrm{d}z;\quad \Gamma(n+1)=1\cdot 2\cdot 3\cdots n=n!;$$

$$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi};\quad \Gamma\left(n+\frac{1}{2}\right)=\frac{1\cdot 3\cdot 5\cdots(2n-1)}{2^n}\Gamma\left(\frac{1}{2}\right).$$

Однопараметрическая функция $f_t(x, n)$ является четной: $f_t(-x, n) = f_t(x, n)$.

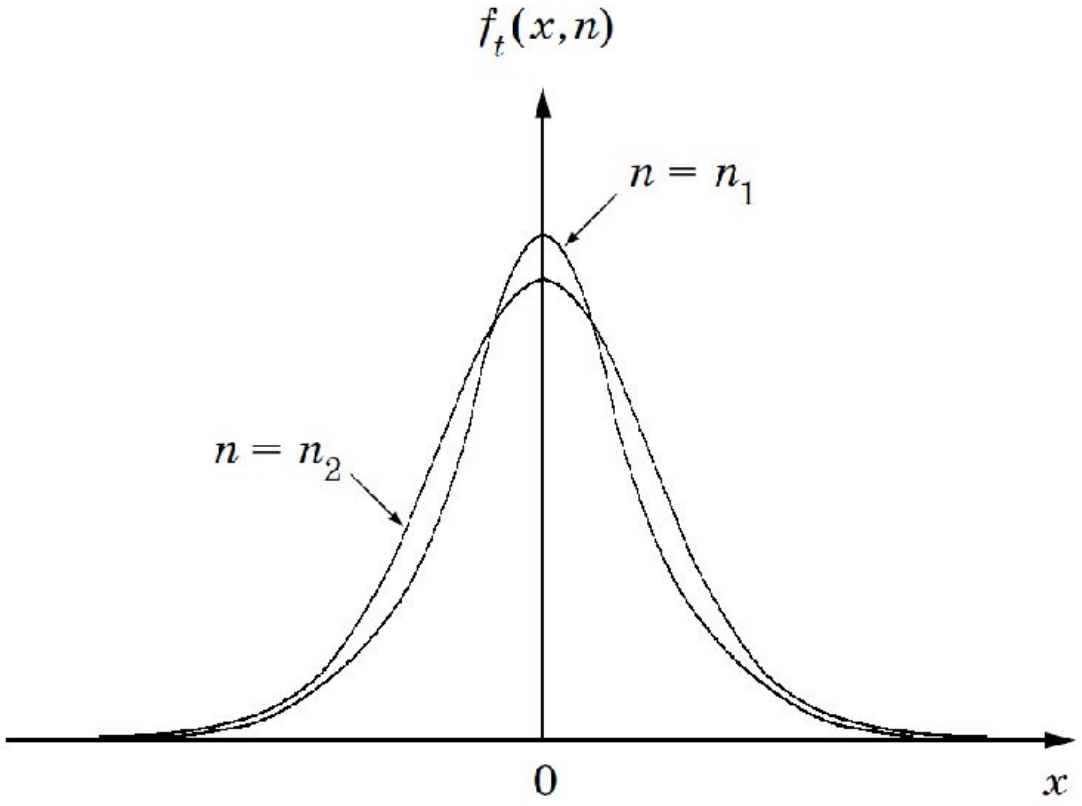


Рис. 6.1. Плотности распределения Стьюдента при $n_1 > n_2$

Случайная величина t_n имеет моменты только при $k < n$, причем начальные моменты равны

$$\alpha_k[t_n] = M[t_n^k] = \begin{cases} 0, & \text{если } k \text{ нечетно;} \\ \dfrac{\Gamma\left(\dfrac{k+1}{2}\right)\Gamma\left(\dfrac{n-k}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\dfrac{n}{2}\right)}\sqrt{n^k}, & \text{если } k \text{ четно.} \end{cases}$$

Математическое ожидание и дисперсия равны соответственно: $M[t_n] = 0$ и $D[t_n] = \dfrac{n}{n-2}$, если $n > 2$.

При $n \to \infty$ распределение Стьюдента стремится к нормальному с $f_N(x, 0, 1)$, т.е.

$$\lim_{n\to\infty} f_t(x,n) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right).$$

По определению квантиль порядка α распределения Стьюдента с n степенями свободы есть такое число $t_{\alpha,n}$, что вероятность $p(t_n < t_{\alpha,n}) = \alpha$ или эквивалентно

$$\int_{-\infty}^{t_{\alpha,n}} f_t(x,n)\,\mathrm{d}x = \int_{-\infty}^{t_{\alpha,n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\mathrm{d}x = \alpha.$$

Очевидно, что вследствие симметрии распределения Стьюдента $t_{1-\alpha,\,n} = -t_{\alpha,n}$.

При больших n ($n \geq 30$) выполняется приближенное равенство $t_{1-\alpha,\,n} \approx x_\alpha$, где x_α – квантиль порядка α нормального распределения (см. п. 6.1). Более точное выражение для $t_{\alpha,n}$ при больших n дает формула

$$t_{\alpha,n} = \frac{x_\alpha}{\sqrt{\left(1-\frac{1}{4n}\right)^2 - \frac{x_\alpha^2}{2n}}}. \qquad (6.6)$$

6.3. Распределение χ^2

- Если $\xi_1, \xi_2, \ldots, \xi_n$ – независимые случайные величины, имеющие нормальное распределение с нулевым математическим ожиданием и единичной дисперсией, то распределение случайной величины

$$H_n = \sum_{j=1}^{n} \xi_j^2$$

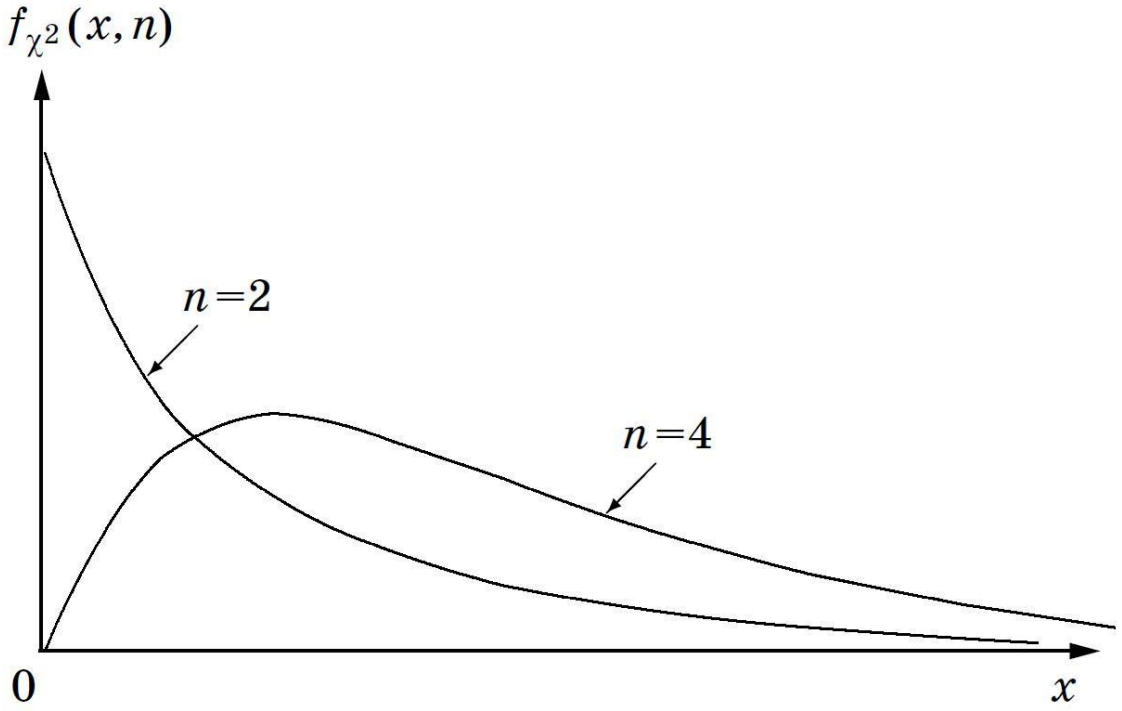
называют *распределением* χ^2 с n степенями свободы. Обычно и для самой случайной величины H_n используется тот же символ, т.е. вместо H_n пишут χ^2.

Плотность распределения вероятностей H_n (или χ^2) является однопараметрической функцией, равной (рис. 6.2)

$$f_{\chi^2}(x,n)=\frac{\left(\frac{x}{2}\right)^{\frac{n}{2}-1}}{2\Gamma\left(\frac{n}{2}\right)}\exp\left(-\frac{x}{2}\right),$$

где $\Gamma(x)$ – гамма-функция

Заметим, что распределение χ^2 устойчиво относительно суммирования.



Математическое ожидание и дисперсия H_n равны соответственно $M[H_n]=n$ и $D[H_n]=2n$.

В силу центральной предельной теоремы при большом числе степеней свободы распределение случайной величины H_n можно рассматривать как нормальное с параметрами: $a=n$; $\sigma^2=2n$. Другими словами, функция распределения $(H_n-n)/\sqrt{2n}$ стремится к $F_N(x, 0, 1)$ при $n\to\infty$.

Квантилью порядка α распределения χ^2 называется такое число $\chi^2_{\alpha,n}$, для которого справедливо равенство $p(H_n<\chi^2_{\alpha,n})=\alpha$.

Значения квантилей даны в виде таблиц в учебниках и специальных справочниках

Для приближенного вычисления квантилей при больших n ($n\geq 30$) используют асимптотическую нормальность распределения χ^2. Поскольку при $n\to\infty$ функция распределения $(H_n-n)/\sqrt{2n}$ стремится к нормальной, то

$$p\left(\frac{H_n-n}{\sqrt{2n}}<x_\alpha\right)\approx\alpha.$$

где x_α – квантиль порядка α нормального распределения.

$$p(H_n < \chi^2_{\alpha,n}) = p\left(\frac{H_n - n}{\sqrt{2n}} < \frac{\chi^2_{\alpha,n} - n}{\sqrt{2n}}\right) = \alpha,$$

что дает

$$\frac{\chi^2_{\alpha,n} - n}{\sqrt{2n}} \approx x_\alpha .$$

Таким образом, при больших n справедливо приближенное выражение

$$\chi^2_{\alpha,n} \approx n + x_\alpha \sqrt{2n} .$$

Можно получить и другие приближенные формулы, точность которых возрастает с увеличением n:

$$\chi^2_{\alpha,n} \approx \frac{1}{2}(x_\alpha + \sqrt{2n-1})^2 ;$$

$$\chi^2_{\alpha,n} \approx n\left(1 - \frac{2}{9n} + x_\alpha\sqrt{\frac{2}{9n}}\right)^3.$$

Критическая точка $\chi^2(\alpha, n)$ определяется соотношением $p(H_n > \chi^2(\alpha, n)) = \alpha$ и связана с квантилью соотношением $\chi^2(\alpha, n) = \chi^2_{1-\alpha,\, n}$.

- **ГЛАВА 7. ПРИМЕР СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ВЫБОРКИ**

- Будем считать максимальную дневную температуру в Санкт-Петербурге 1 сентября случайной величиной ξ. Генеральная совокупность – это данные Гидрометеослужбы о такой температуре в разные годы. Сделана следующая выборка из генеральной совокупности (ºС):

14	16	16	18	13	13	16	16	25	10
19	14	15	17	12	16	13	19	16	17
17	12	15	19	14	15	17	19	18	14
26	21	17	15	19	18	18	13	15	18
17	17	16	18	16	21	15	13	20	14

- Рассмотрим некоторые задачи, на которые разбивается статистическая обработка выборки, направленная на определение свойств данной случайной величины

Пример 7.1. Для приведенной выборки построить вариационный ряд и выборочный закон распределения ξ. Найти выборочное среднее $\overline{x}$, выборочную дисперсию D^* и исправленную выборочную дисперсию s^2.

Решение. Вариационный ряд, построенный по данной выборке, извлеченной из генеральной совокупности, включает 13 вариант: 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 25, 26° С.

Выборочный закон распределения оказывается следующим:

x_i	n_i	ω_i	x_i	n_i	ω_i
10	1	0,02	18	6	0,12
12	2	0,04	19	5	0,10
13	5	0,10	20	1	0,02
14	5	0,10	21	2	0,04
15	6	0,12	25	1	0,02
16	8	0,16	26	1	0,02
17	7	0,14			

Выборочное среднее равно

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{13} n_i x_i = \frac{1}{50}(10 + 2\cdot 12 + \ldots + 26) = \frac{822}{50} = 16{,}44\ ^\circ\text{C}.$$

Найдем

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{13} n_i x_i^2 = \frac{1}{50}\{(10)^2 + 2(12)^2 + \ldots + (26)^2\} =$$

$$= \frac{13966}{50} = 279{,}32\ \text{град}^2$$

и получим выборочные дисперсии:

$$D^* = (\sigma^*)^2 = \overline{x^2} - (\overline{x})^2 = 279{,}32 - (16{,}44)^2 = 9{,}0464\ \text{град}^2;$$

$$s^2 = \frac{n}{n-1}(\sigma^*)^2 = \frac{50}{49}9{,}0464 = 9{,}231\ \text{град}^2.$$

Пример 7.2. Построить с надежностью $\gamma = 0{,}90$ доверительный интервал для математического ожидания случайной величины ξ.

Решение. Будем считать, что рассматриваемая выборка достаточна велика $(n >> 1)$ и применима формула (4.23):

$$\begin{cases} p\left(\overline{x} - \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}} < M[\xi] < \overline{x} + \dfrac{s\,\tilde{x}_\gamma}{\sqrt{n}}\right) \approx 2\Phi(\tilde{x}_\gamma); \\ 2\Phi(\tilde{x}_\gamma) = \gamma, \end{cases}$$

в которой, учитывая полученные выше результаты, $\overline{x} = 16{,}44^\circ\text{C}$, $s = \sqrt{9{,}231} \approx 3{,}038^\circ\text{C}$, а $2\Phi(\tilde{x}_\gamma) = 0{,}9$ и $\tilde{x}_\gamma = 1{,}65$.

Тогда доверительный интервал для математического ожидания $M[\xi]$ с надежностью (доверительной вероятностью) $\gamma = 0{,}9$ определяется следующим образом:

$$p\left(16{,}44 - \frac{3{,}038 \cdot 1{,}65}{\sqrt{50}} < M[\xi] < 16{,}44 + \frac{3{,}038 \cdot 1{,}65}{\sqrt{50}}\right) =$$

$$= p\left(15{,}73 < M[\xi] < 17{,}15\right) = 0{,}9\,.$$

Пример 7.3. Построить с надежностью $\gamma = 0{,}90$ доверительный интервал для дисперсии $D[\xi]$ случайной величины ξ в предположении, что она имеет нормальное распределение.

Решение. Здесь можно воспользоваться формулой (4.21):

$$p\left(\frac{(n-1)s^2}{\tilde{\chi}^2_{2,\gamma,n-1}} < D[\xi] < \frac{(n-1)s^2}{\tilde{\chi}^2_{1,\gamma,n-1}}\right) = \gamma.$$

Поскольку в приложении дана таблица квантилей распределения χ^2, то для определения значений $\tilde{\chi}^2_{1,\gamma,n-1}$ и $\tilde{\chi}^2_{2,\gamma,n-1}$ будем использовать соотношения (4.16), которые дают

$$\tilde{\chi}^2_{1,\gamma,n-1} = \chi^2_{(1-\gamma)/2,n-1},\ \tilde{\chi}^2_{2,\gamma,n-1} = \chi^2_{(1+\gamma)/2,n-1},$$

где $\chi^2_{\alpha,n}$ есть квантиль порядка α распределения χ^2 с n степенями свободы. В нашем случае требуется найти квантили $\chi^2_{0,05;49}$ и $\chi^2_{0,95;49}$. Поскольку табл. 3 приложения не содержит этих квантилей, но объем выборки достаточно велик ($n > 30$), для вычисления квантилей можно применить асимптотическую формулу (6.11):

$$\chi^2_{\alpha,n} \approx n\left(1-\frac{2}{9n}+x_\alpha\sqrt{\frac{2}{9n}}\right)^3.$$

Квантили нормального распределения равны: $x_{0,05} = -1,65$; $x_{0,95} = 1,65$ и для $n = 50$, $\gamma = 0,9$ имеем

$$\tilde{\chi}^2_{1,\gamma,n-1} = \chi^2_{0,05;\,49} \approx 49\left(1-\frac{2}{9\times 49}-1,65\sqrt{\frac{2}{9\times 49}}\right)^3 = 33,89;$$

$$\tilde{\chi}^2_{2,\gamma,n-1} = \chi^2_{0,95;\,49} \approx 49\left(1-\frac{2}{9\times 49}+1,65\sqrt{\frac{2}{9\times 49}}\right)^3 = 66,40.$$

Таким образом доверительный интервал для дисперсии случайной величины ξ, если предположить, что она имеет нормальное распределение, с надежностью (доверительной вероятностью) $\gamma = 0,9$ определяется следующим образом:

$$p\left(\frac{49\times 9,231}{66,40} < D[\xi] < \frac{49\times 9,231}{33,89}\right) =$$

$$= p\left(6,81 < D[\xi] < 13,35 \text{ град}^2\right) = 0,90.$$

Конец