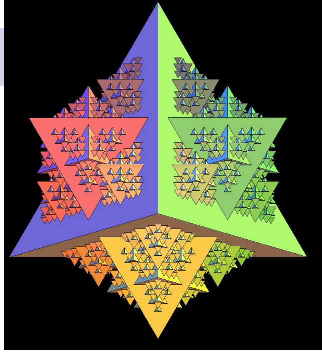
Статистические оценки параметров распределения

Точечные и интервальные оценки

- При изучении случайной величины *X*, распределенной в генеральной совокупности, часто из теоретических соображений удается установить **вид распределения** и по данным выборки **необходимо оценить (приближенно найти) его численные параметры**.
- Например, если случайная величина имеет *нормальное распределение*, то для полного его определения необходимо *оценить его математическое ожидание и среднее квадратическое отклонение*.

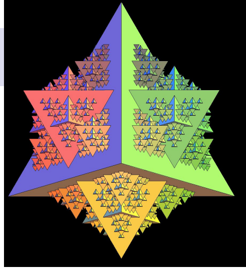


Пусть имеется генеральная совокупность объема N, $X_г$- изучаемый признак. Для изучения этого признака генеральной совокупности производится репрезентативная выборка объема n, получим выборочное распределение:

X	x_1	x_2	…	x_k
W	w_1	w_2	…	w_k

•

где $w_i = \frac{m_i}{n}, \sum_{i=1}^{k} m_i = n;$ далее, можно вычислить выборочную среднюю $\overline{X_в} = \frac{1}{n}\sum_{i=1}^{k} x_i m_i,$ выборочную дисперсию $D_в = D[X_в] = \frac{1}{n}\sum_{i=1}^{k} m_i (x_i - \overline{X_в})^2$ и др. интересующие нас характеристики. Обозначим через $Q_в$ любую из выборочных характеристик, т.к. $Q_в = \{\overline{X_в}, D_в, \ldots\}$. $Q_в$ является случайной величиной, т.к. процесс выборки носит случайный характер.



Точечной называют оценку, которая определяется одним числом.

Интервальной называют оценку, которая определяется двумя числами – концами интервала. Интервальные оценки позволяют определить точность и надежность оценок.

оценки параметров распределения

Точечная оценка неизвестного параметра

Θ – случайная функция $\Theta^*(X_1, X_2, ..., X_n)$, значение которой для любой реализации выборки принимают за приближенное значение параметра Θ,

$$\Theta \approx \Theta^*(x_1, x_2, ..., x_n).$$

Интервальная оценка неизвестного параметра

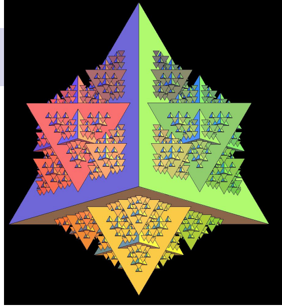
Θ – случайная функция $\Theta^*(X_1, X_2, ..., X_n)$ такая, что

$$P\left(-\varepsilon_1 < \Theta - \Theta^*(X_1, X_2, ..., X_n) < \varepsilon_2\right) =$$
$$= \gamma = 1 - \alpha$$

Точечные оценки

Несмещенные, состоятельные и эффективные оценки

- Для того, чтобы статистической оценке можно было *доверять*, она должна обладать некоторыми свойствами.



Оценки

<u>О.</u> Статистическая оценка Θ^* называется **несмещенной**, если ее математическое ожидание равно Θ, т.е. $M\left[\Theta^*\right]=\Theta$.

<u>О.</u> Статистическая оценка Θ^* называется **состоятельной**, если она сходится по вероятности к оцениваемому параметру, т.е.

$$\forall\varepsilon>0 \quad \lim_{n\to\infty} P\left(\left|\Theta^*-\Theta\right|<\varepsilon\right)=1.$$

<u>З.</u> Из неравенства Чебышева следует, что несмещенная оценка, дисперсия которой стремится к нулю при $n\to\infty$, является состоятельной.

<u>О.</u> Статистическая оценка Θ^* называется **эффективной**, если при данном объеме выборки из всех возможных оценок она имеет наименьшую дисперсию.

Несмещенные, состоятельные и эффективные оценки

- 3. На практике не всегда удается добиться выполнения всех трех требований к оценке. Соображения практической удобности заставляют пользоваться не полностью адекватными оценками, но **необходимо представлять, каким свойством мы пренебрегаем**. Ниже, при рассмотрении конкретных оценок, эти аспекты будут обсуждаться.

Точечная оценка генерального среднего по выборочному среднему

Выберем в качестве оценки генерального среднего $M[X]=a$ среднее арифметическое случайных величин $X_1, X_2, \ldots, X_n$: $\overline{X_B} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$; для конкретной реализации выборки значения этой величины равны выборочным средним. Найдем математическое ожидание оценки $\overline{X_B}$:

$$M\left[\overline{X_B}\right] = M\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} M\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} M[X_i] = \frac{1}{n}\cdot n \cdot a = a,$$

следовательно, $\overline{X_B}$ – **несмещенная** оценка $M[X]$.

Точечная оценка генерального среднего по выборочному среднему

Если случайная величина X распределена нормально, то оценка $\overline{X_B}$ будет и **эффективной**.
На практике во всех случаях для оценки математического ожидания используется среднее арифметическое $\overline{X_B}$ (обозначается также $\overline{X}$).

Точечная оценка генеральной дисперсии по исправленной выборочной дисперсии

$$D_B = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X_B}\right)^2$$

- Можно показать, что выборочная дисперсия (среднее значение квадрата отклонения) является смещенной оценкой генеральной дисперсии.

Точечная оценка генеральной дисперсии по исправленной выборочной дисперсии

- Для получения несмещенной оценки достаточно перейти к ***исправленной выборочной дисперсии*** $s^2 = \frac{n}{n-1} D_B$

$$s^2 = \frac{n}{n-1} D_B = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X_B} \right)^2$$

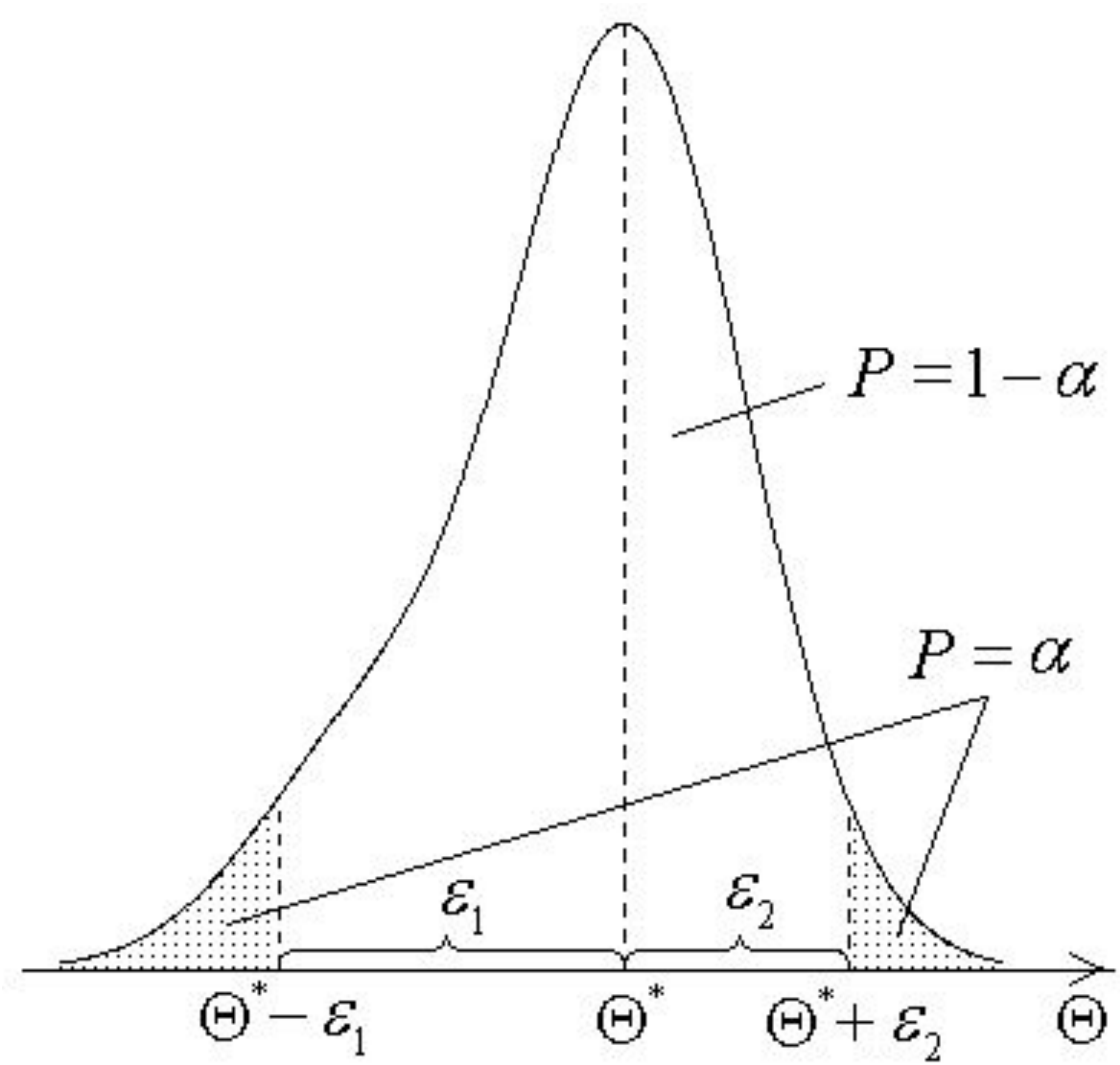
Интервальные оценки

$$P\left(\Theta^*(X_1, X_2, \ldots, X_n) - \varepsilon_1 < \Theta < \Theta^*(X_1, X_2, \ldots, X_n) + \varepsilon_2\right) = 1 - \alpha,$$

т.е., интервал $\left(\Theta^*(X_1, X_2, \ldots, X_n) - \varepsilon_1, \Theta^*(X_1, X_2, \ldots, X_n) + \varepsilon_2\right)$ заключает в себе (покрывает) неизвестный параметр Θ с вероятностью γ.

Сам интервал носит название **доверительного интервала**, величина $\gamma = 1 - \alpha$ называется **доверительной вероятностью** оценки (**надежностью**, **коэффициентом доверия**), числа $\Theta^*(X_1, X_2, \ldots, X_n) - \varepsilon_1$ и $\Theta^*(X_1, X_2, \ldots, X_n) + \varepsilon_2$ – **доверительными границами**, α – **уровнем значимости (существенности)**.

Графический смысл



Интервальные оценки

3. Если плотность распределения $\Theta^*(X_1, X_2, \ldots, X_n)$ **симметрична** относительно своей медианы, применяют следующий способ получения интервальной оценки:

$$P\left(\left|\Theta^*(X_1, X_2, \ldots, X_n) - \Theta\right| < \varepsilon\right) = \gamma = 1 - \alpha,$$

$$P\left(\Theta^*(X_1, X_2, \ldots, X_n) - \varepsilon < \Theta < \Theta^*(X_1, X_2, \ldots, X_n) + \varepsilon\right) = \gamma = 1 - \alpha,$$

т.е. нижняя и верхняя границы доверительного интервала симметричны относительно оценочной функции $\Theta^*(X_1, X_2, \ldots, X_n)$. Величина ε, характеризующая ширину доверительного интервала, называется **точностью оценки**.

- Точечные оценки проще в вычислении, но не позволяют установить степень достоверности оценки.
- Интегральные оценки, наряду с возможными границами значений параметра, дают вероятность, с которой **истинное** значение параметра лежит между этими (случайными) границами.
- Естественно, чем больше надежность оценки, тем шире доверительный интервал, и наоборот, так что практические вычисления являются компромиссом между точностью и надежностью оценки. Наиболее часто задают надежность 0,95; 0,99 и 0,999.

Интервальные оценки

При получении ***точечной оценки*** необходимо знать *лишь выражение для оценки* $\Theta^*(X_1, X_2, \ldots, X_n)$ как функцию данных выборки,

а для получения ***интервальной оценки*** необходимо также *знать закон распределения* $\Theta^*(X_1, X_2, \ldots, X_n)$, с помощью которого рассчитывается вероятность, вид генерального распределения и значение параметров распределения, которые и подлежат оценке.

для получения интервальной оценки

Как правило **вид генерального распределения постулируется**(нормальное распределение, равномерное распределение и т.д.). При достаточно *большом объеме выборки* **реальную функцию** распределения оценки с достаточной точностью **можно заменить асимптотической**

Значения параметров генерального распределения оценивают либо приближенно либо **точно**.

Приближенный способ

- состоит в **замене неизвестных параметров** генеральной совокупности, от которых зависит распределение , **на их точечные оценки**, *полученные в результате выборки.*
- Далее оценка строится, как если бы параметры распределения были бы известны.

Точный способ

- может быть использован лишь в том случае, **когда известен закон генерального распределения**. При этом строятся вспомогательные случайные величины, распределение которых *зависит лишь от объема выборки*.
- *В частности, при оценке среднего значения* ***нормально распределенной генеральной совокупности*** *можно использовать оценку*

$$\Theta^* = T = \frac{M[X] - \overline{X_B}}{\sqrt{\frac{n}{n-1} D_B}} \cdot \sqrt{n}$$

которая подчиняется распределению Стьюдента, зависящему только от объема выборки .

С интервальной оценкой связано решение трех типов задач

1) определение **доверительной вероятности** по *заданному доверительному интервалу и объему выборки;*

2) определение **доверительного интервала** по *заданной доверительной вероятности и объему выборки*;

3) определение **необходимого объема выборки** по *заданным доверительной вероятности и доверительному интервалу*

Интервальная оценка математического ожидания нормально распределенной случайной величины при известном σ

Пусть случайная величина X генеральной совокупности **распределена по нормальному закону**

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}},$$ где $\sigma=\sqrt{D[X]}$, $a=M[X]=\overline{X}$,

при этом ***среднее квадратическое отклонение σ считаем известным.***

Пусть сделана выборка объема n и вычислена $\overline{X_B}$, которая является точечной оценкой математического ожидания a генеральной совокупности.

Интервальная оценка математического ожидания нормально распределенной случайной величины при известном σ

Так как $\overline{X_B} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ – среднее арифметическое случайных величин $X_1, X_2, \ldots, X_n$, распределенных так же, как и признак генеральной совокупности X, **то закон распределения** $\overline{X_B}$ **также нормален,** $M\left[\overline{X_B}\right] = M[X] = a$.

Дисперсия случайной величины $\overline{X_B}$ в n раз меньше дисперсии случайной величины X: $\sigma^2\left[\overline{X_B}\right] = \dfrac{\sigma^2}{n}$.

1) Определим, с какой надежностью математическое ожидание *a* покрывается доверительным интервалом при заданной точности ε, т.е. найдем $P\left(\overline{X}_B - \varepsilon < a < \overline{X}_B + \varepsilon\right)$

Тогда, $P\left(\left|\overline{X}_в - a\right| < \varepsilon\right) = 2\Phi\left(\dfrac{\varepsilon}{\sigma\left(\overline{X}_в\right)}\right) = 2\Phi\left(\dfrac{\varepsilon\sqrt{n}}{\sigma}\right) = 2\Phi(t) = \gamma$,

где $t = \dfrac{\varepsilon\sqrt{n}}{\sigma}$, где $\Phi(t)$ – функция Лапласа.

Зная σ, ε и n, можно найти по таблице значений функции Лапласа надежность γ оценки $\overline{X}_B$ математического ожидания а.

2) По выборочному значению математического ожидания $\overline{X}_B$ и известному σ найти доверительный интервал, который с заданной надежностью γ покрывает математическое ожидание *a* генеральной совокупности. Это и есть задача получения интервальной оценки $\overline{X}_B$

Используя *таблицы значений функции Лапласа*, **по** γ **определяют** $t=\frac{\varepsilon\sqrt{n}}{\sigma}$, отсюда $\varepsilon=\frac{t\sigma}{\sqrt{n}}$.

Таким образом получают **искомый доверительный интервал** $\left(\overline{X}_B-\frac{t\sigma}{\sqrt{n}},\overline{X}_B+\frac{t\sigma}{\sqrt{n}}\right)$,

с надежностью γ покрывающий неизвестный параметр a; точность оценки $\varepsilon=\frac{t\sigma}{\sqrt{n}}$.

3) По заданным σ, ε и γ, используя $2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)=\gamma$ соотношение , найти объем выборки n.

Решая уравнение $2\Phi(t)=\gamma$, по γ находим t,

а затем из $t=\frac{\varepsilon\sqrt{n}}{\sigma}$ находим минимальный

объем выборки $n=\left(\frac{t\sigma}{\varepsilon}\right)^2$.

Пример:

Измерение массы 50 случайно отобранных после изготовления деталей дало $\overline{X_B^*}$ = 10 г. Есть основания полагать, что генеральная дисперсия σ^2 = 0,09.

1: *Определить с вероятностью P = 0,95 доверительные границы для средней массы деталей $\overline{X}$ во всей партии.*

На основании **теоремы Ляпунова** можно исходить из предположения ***о нормальном распределении массы деталей***.

Дано n = 50, $\overline{X_B^*}$ = 10 г, σ^2 = 0,09 и P = 0,95.

Из равенства $2\Phi(t) = 0,95$ по таблицам функции Лапласа находим

t = 1,96, откуда $\varepsilon = \dfrac{t\sigma}{\sqrt{n}} = \dfrac{1,96 \cdot 0,3}{\sqrt{50}} = 0,083$.

Таким образом, получаем, что с вероятностью 0,95 средняя масса содержится в промежутке [9,917; 10,083].

Пример:

Измерение массы 50 случайно отобранных после изготовления деталей дало $\overline{X_B^*} = 10$ г. Есть основания полагать, что генеральная дисперсия $\sigma^2 = 0{,}09$.

2: *Определить при тех же условиях, с какой доверительной вероятностью можно гарантировать ошибку выборки, не превышающую 0,05.*

По величине $\varepsilon = 0{,}05$ вычисляем $t = \dfrac{\varepsilon\sqrt{n}}{\sigma} = \dfrac{0{,}05 \cdot \sqrt{50}}{0{,}3} = 1{,}1785$.

По таблицам функции Лапласа $\Phi = 2(1{,}1785) \approx 0{,}76$.

Пример:

Измерение массы 50 случайно отобранных после изготовления деталей дало $\overline{X_B^*} = 10$ г. Есть основания полагать, что генеральная дисперсия $\sigma^2 = 0{,}09$.

3: *Определить объем выборки, при котором указанная предельная ошибка $\varepsilon = 0{,}05$ гарантируется с вероятностью $P = 0{,}95$.*

Из $P = 0{,}95$ находим $t = 1{,}96$, откуда

$$n = \left(\frac{t\sigma}{\varepsilon}\right)^2 = \left(\frac{1{,}96 \cdot 0{,}3}{0{,}05}\right)^2 \approx 138{,}2976 \approx 140$$

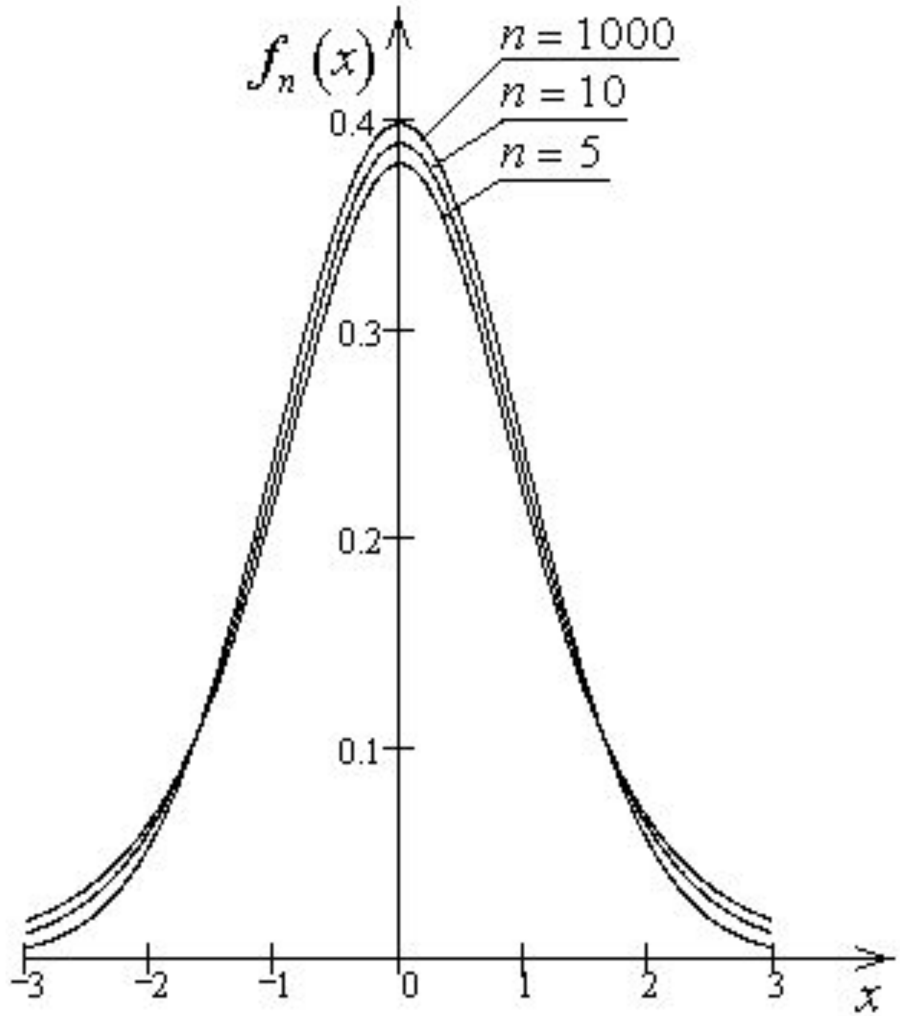
Доверительный интервал для оценки неизвестного математического ожидания нормально распределенной случайной величины при неизвестном σ.

Для решения этой задачи используется закон распределения случайной величины $T=\{t\}$, $t = \dfrac{\overline{X}_в - a}{s_в / \sqrt{n}}$,

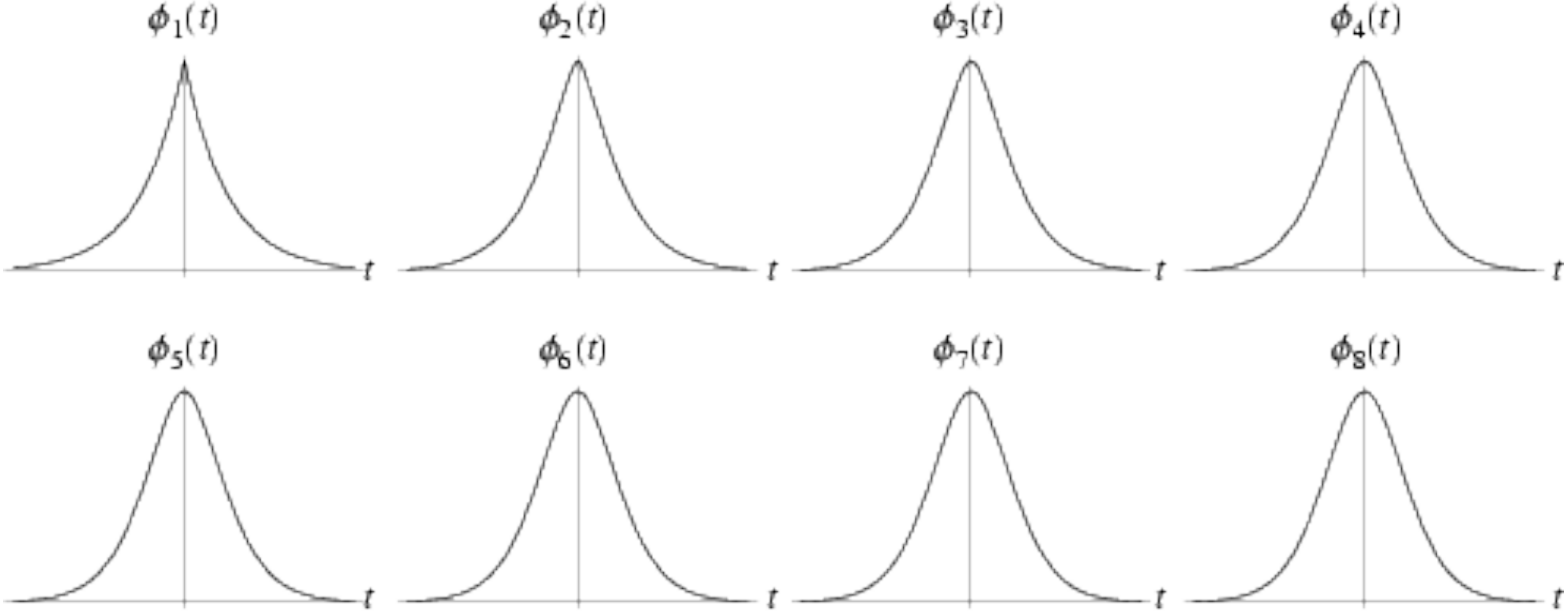
где $s_в = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}_в\right)^2}$ **несмещенная оценка среднего квадратичного отклонения**.

График $f_n(t)$ похож на график плотности вероятности нормального закона, но $f_n(t)$ определяется только объемом выборки n и не зависит от неизвестных параметров a и σ

$$\lim_{n\to\infty} f_n(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$



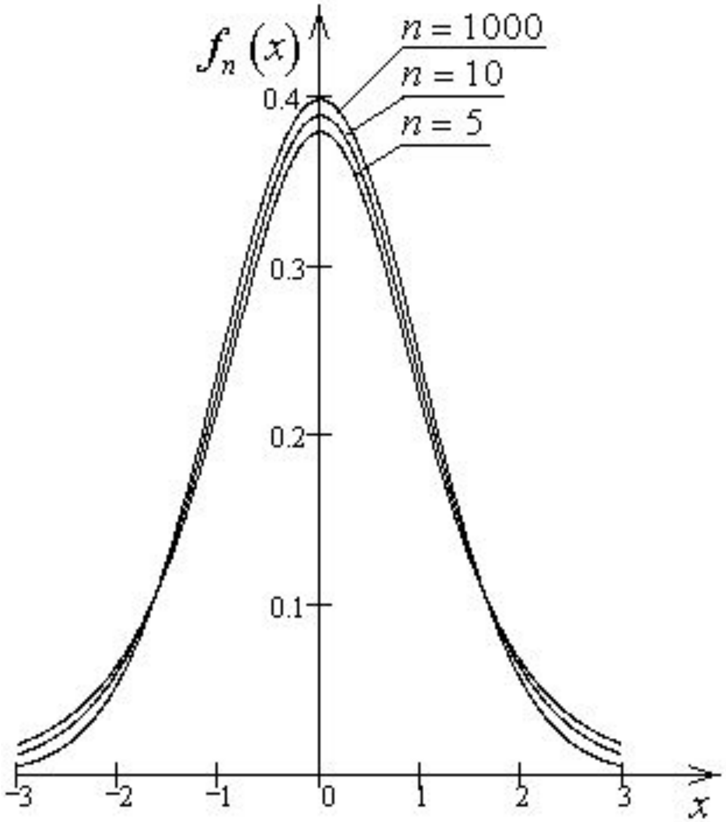
. Так как при $n \to \infty$ распределение Стьюдента стремится к нормальному, то практически при $n > 30$ можно пользоваться вместо распределения Стьюдента нормальным распределением



Вид распределения стьюдента

- Зависимость кривой распределения от параметров в примере к пакету Зависимость кривой распределения от параметров в примере к пакету

  Mathematica



Распределение стьюдента

Очевидно, что $P\left(|t|<t_\gamma\right)=\int\limits_{-t_\gamma}^{t_\gamma} f_n(t)dt=\gamma$. Функция $f_n(t)$ четная, поэтому можно записать $\gamma=2\int\limits_0^{t_\gamma} f_n(t)dt$. Это равенство позволяет найти t_γ по заданным n и γ. Во всех книгах по математической статистике имеются таблицы, по которым можно найти t_γ, зная n и γ. Учитывая, что $t=\frac{\overline{X}_\text{в}-a}{s_\text{в}/\sqrt{n}}$, можно записать $P\left(|t|<t_\gamma\right)=P\left(\frac{\left|\overline{X}_\text{в}-a\right|}{s_\text{в}/\sqrt{n}}<t_\gamma\right)=\gamma$, или $P\left(\left|\overline{\ \ }a-\ \ \right|<\frac{t_\gamma s_\text{в}}{\sqrt{n}}\right)=\gamma$. Таким образом, определив t_γ по таблице Стьюдента, можно найти доверительный интервал, покрывающий с надежностью γ математическое ожидание a генеральной совокупности.

Приложение 4. Критические точки t - распределения Стьюдента
(k - число степеней свободы)

k	Уровень значимости α (двусторонняя критическая область)								
	0,25	**0,2**	**0,1**	**0,05**	**0,02**	**0,01**	**0,005**	**0,002**	**0,001**
1	2,414	3,078	6,314	12,706	31,821	63,656	127,32	318,29	636,58
2	1,604	1,886	2,920	4,303	6,965	9,925	14,089	22,328	31,600
3	1,423	1,638	2,353	3,182	4,541	5,841	7,453	10,214	12,924
4	1,344	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	1,301	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	1,273	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	1,254	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	1,240	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	1,230	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	1,221	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	1,214	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	1,209	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	1,204	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	1,200	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	1,197	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	1,194	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	1,191	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	1,189	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	1,187	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	1,185	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	1,183	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	1,182	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	1,180	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	1,179	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	1,178	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	1,177	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	1,176	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	1,175	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	1,174	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,660
30	1,173	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	1,167	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
50	1,164	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
60	1,162	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460

Пример:

Случайная величина X имеет нормальное распределение. С надежностью $\gamma = 0{,}95$ найти доверительный интервал для оценки неизвестного математического ожидания a, если при объеме выборки n=36 получено $\overline{x}_в$=24, $s = 3$.

3:

По таблицам распределения Стьюдента при $n = 36$ и $\gamma = 0{,}95$ находим $t_\gamma = 2{,}03$. Тогда точность оценки $\frac{t_\gamma s}{\sqrt{n}} = \frac{2{,}03 \cdot 3}{\sqrt{36}} = 1{,}015$.

Доверительный интервал для a $\left(\overline{X}_в - 1{,}015; \overline{X}_в + 1{,}015\right) = \left(22{,}985; 25{,}015\right)$.

В результате студент должен уметь:

по данным выборки получать точечные и интервальные оценки параметров распределения при построении математических моделей случайных событий , в том числе и с использованием предельных теорем теории вероятностей.