Лекция 2

Статистически оценки на параметрите на разпределението

В предишната лекция разгледахме понятията случайна величина, функция на разпределение, плътност на функцията на разпределение, както и се запознахме с две конкретни разпределения на случайните величини – равномерното и нормалното. В тази лекция ще разгледаме някои статистически оценки на параметрите на разпределение.

Получените при измерванията случайни величини са независими една от друга и същевременно разпределени нормално с едни и същи параметри μ и σ^2. Ето защо те представляват един вид "реализации" на една (в повечето случаи) нормално разпределена случайна величина. Разбира се, измерванията могат да бъдат и с друго разпределение, но опитът сочи, че в химията повечето измервани величини имат нормално разпределение, затова в този материал всички изводи ще се основават на предпоставката, че измерванията са независими случайни величини, разпределени с едни и същи параметри (математическо очакване и дисперсия) и са нормално разпределени.

Всеки набор от експериментални стойности се нарича *извадка* от генералната съвкупност. Под *генерална съвкупност* се разбира извадка с неограничен обем, т.е. такава извадка, която напълно характеризира случайната величина. Като понятие генералната съвкупност напълно се припокрива с понятието разпределение, но боравенето с такъв безкраен набор от реализации на случайната величина е по-близко до сетивните представи на човека, отколкото кривата на разпределение. Много често се казва "генерална съвкупност с едикакво си разпределение", което не противоречи на изложеното тълкуване. Както беше обяснено в първия

параграф, понятието величина се асоциира с понятието стойност и най-естественото съпоставяне (досега непротиворечащо на човешката практика) е за стойност на величината да се приема математическото очакване на разпределението на получаваните случайни величини. Затова една от задачите на експериментатора е получаване на оценка на математическото очакване с помощта на извадката случайни величини. Числената стойност на дисперсията определя вероятността за появата на дадена случайна величина в определен интервал и тъй като всяка една оценка е случайна величина, то за да се определи интервалът, в който е най-вероятно да се намира математическото очакване, е необходимо да се оцени и дисперсията. Оценки на **M(X)** и **D(X)** са ***средната стойност*** и ***стандартното отклонение***.

2.1. *Средна стойност*. Тя се дефинира за N стойности по уравнението

$$\bar{X} = (X_1 + X_2 + \ldots + X_N)/N = \Sigma X_k/N \qquad (2.1)$$

Средната стойност е случайна величина, която също е разпределена нормално със собствени μ и σ^2. Математическото и очакване (прилагайки свойства на математическото очакване, вижте предишната лекция) е равно:

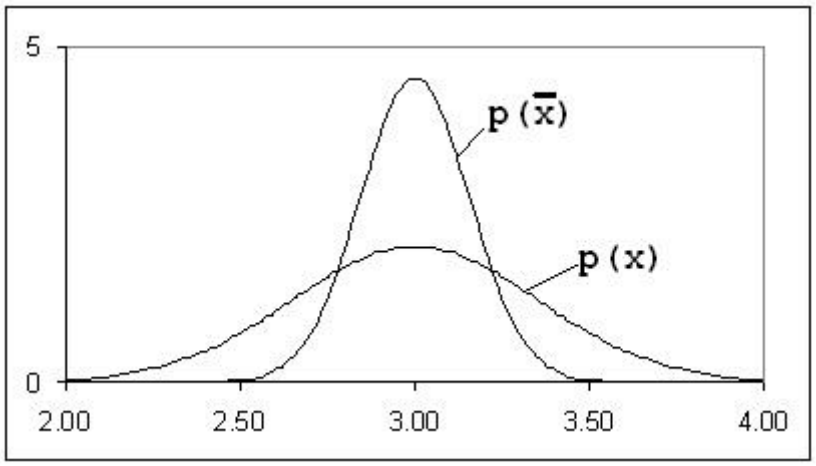
$$M(\bar{X}) = M(\Sigma X_k/N) = (1/N)\Sigma M(X_k) = (1/N)\Sigma M(X) = (1/N)NM(X) = M(X)$$

т.е. математическото очакване на средната величина на N случайни стойности съвпада с математическото им очакване. Такава статистическа оценка, за която математическото и очакване съвпада с оценяваната величина, се нарича ***неизместена оценка***. Аналогично се доказва, че дисперсията на разпределението на средните стойности е 1/N от дисперсията на случайните величини.

$$D(\bar{X}) = D(\Sigma X_k/N) = (1/N)^2\Sigma D(X_k) = (1/N)^2\Sigma D(X) = (1/N)^2ND(X) = (1/N)D(X)$$

Тук под внимание бе взето, че отделните измервания са независими случайни величини.

На практика това означава, че ако са проведени няколко серии от измервания на една величина и са определени техните средни стойности, то разпределението на средните стойности и това на случайната величина имат еднакви математически очаквания. Но дисперсията на $\bar{X}$ е σ^2/N, т.е. разсейването на средните стойности около стойността на величината са по-малки, и по-добре я приближават от коя да е измерена стойност. Това нагледно е представено на фигура 2.1.



Фигура 2.1. Разпределения: (1) на нормална случайна величина; (2) на средната стойност от N нейни реализации; $N = 5$.

Ако някои от измерванията се повтарят, формула (1) може да се обобщи

$$\bar{X} = (n_1X_1 + n_2X_2 + \ldots + n_NX_N)/N = \Sigma n_kX_k/N = \Sigma(n_k/N)X_k, \qquad (2.2)$$

където n_k е броят резултати със стойност X_k; $\Sigma(n_k) = N$.

Числата (n_k/N) са честотите на поява на стойностите X_k и при нарастване на N клонят към вероятностите p_k за появата на X_k, т.е. при голямо N се изпълнява равенството (вижте формула (1.1) от предишната лекция):

$$\bar{X} = \Sigma(n_k/N)X_k = \Sigma p_k X_k = M(X) \quad (2.3)$$

Следователно средната стойност клони във вероятностен смисъл към математическото очакване. Такава оценка, която при нарастване обема на извадката клони по вероятност към оценяваната стойност, се нарича ***състоятелна оценка***. Практически това означава, че при нарастване на големината на една извадка, нейната средна стойност се доближава до стойността на измерваната величина.

2.2. *Стандартно отклонение.* За N стойности стандартното отклонение се нарича величината S, чийто квадрат е равен на

$$S^2 = [(X_1 - \bar{X})^2 + \ldots + (X_N - \bar{X})]/(N-1) = [\Sigma(X_k - \bar{X})^2]/(N-1) \quad (2.4)$$

Стандартното отклонение също е случайна величина, като величината

$$\chi^2 = S^2(N-1)/\sigma^2$$

е χ^2 разпределена (за хи-разпределението ще научите в т. 2.3). S^2 е неизместена оценка на дисперсията. Логичен е въпросът защо в знаменателя на (2.4) не стои N. Изразът

$$\Sigma((X_k - \bar{X})^2/N \quad (2.5)$$

е изместена оценка на дисперсията, но и S^2 и (2.5) са състоятелни оценки на дисперсията - просто при нарастване на N се заличава разликата между N и $N-1$.

Задача 2.1. Докажете, че изразът (2.5) е изместена оценка на дисперсията, а S^2 е неизместена оценка.

Упътване: Тази задача е разбита на три части в семинар 2. Ако тя Ви затруднява прочетете и решете задачи 2.1, 2.2 и 2.3 от семинар 2. Решение на тези задачи има накрая на материала в семинар 2.

Пример 2.1. В таблица 2.1. са дадени десет серии от по десет измервания на желязо в питейна вода. Да се намерят средните стойности и стандартните отклонения.

Таблица 2.1. Десет серии от анализи на желязо в питейна вода.

No на серия	**Съдържание на желязо в питейна вода (в μg/ml)**									
1	6.23	6.29	6.17	6.17	6.15	6.29	6.31	6.02	6.33	6.24
2	6.20	6.14	6.17	6.09	6.09	6.22	6.26	6.09	6.15	6.32
3	6.28	6.36	6.24	6.40	6.27	6.19	6.36	6.24	6.12	6.16
4	6.21	6.34	6.20	6.16	6.21	6.20	6.10	6.36	6.21	6.32
5	6.25	6.24	6.23	6.30	6.46	6.36	6.31	6.31	6.34	6.24
6	6.07	6.39	6.24	6.19	6.23	6.23	6.24	6.29	6.30	6.15
7	6.05	6.43	6.24	6.25	6.25	6.34	6.26	6.33	6.44	6.17
8	6.26	6.33	6.35	6.28	6.18	6.14	6.13	6.30	6.27	6.15
9	6.25	6.25	6.40	6.32	6.20	6.29	6.30	6.37	6.36	6.13
10	6.30	6.17	6.30	6.27	6.03	6.39	6.22	6.11	6.19	6.15

Решение: Когато се изчислява на ръка или с калкулатор е целесъобразно да се смята само със стойностите, които се променят от резултат на резултат (в случая само с десетите и стотните, а шестицата не се взима под внимание). Като се приложи **(2.1)**, се получава за <u>първата серия</u>:

$$\bar{x} = (23+29+17+17+15+29+31+2+33+24)/10 = 22 \text{ /в стотни/},$$

т.е. $\bar{x} = 6.22$.

Аналогично по **(2.4)** се получава:

$$s^2 = [(23-22)^2+(29-22)^2+(17-22)^2+(17-22)^2+(15-22)^2 + +(29-22)^2+(31-22)^2+(2-22)^2+(33-22)^2+(24-22)^2]/9 = = 81$$

или $s^2 = 81$ / в стотни /, т.е. $s = 0.09$.

За десетте серии се получават следните резултати:

No на серията	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$\bar{\mathbf{X}}$	6.22	6.17	6.26	6.23	6.30	6.23	6.28	6.24	6.29	6.21
S	0.09	0.08	0.09	0.08	0.07	0.09	0.12	0.08	0.08	0.11

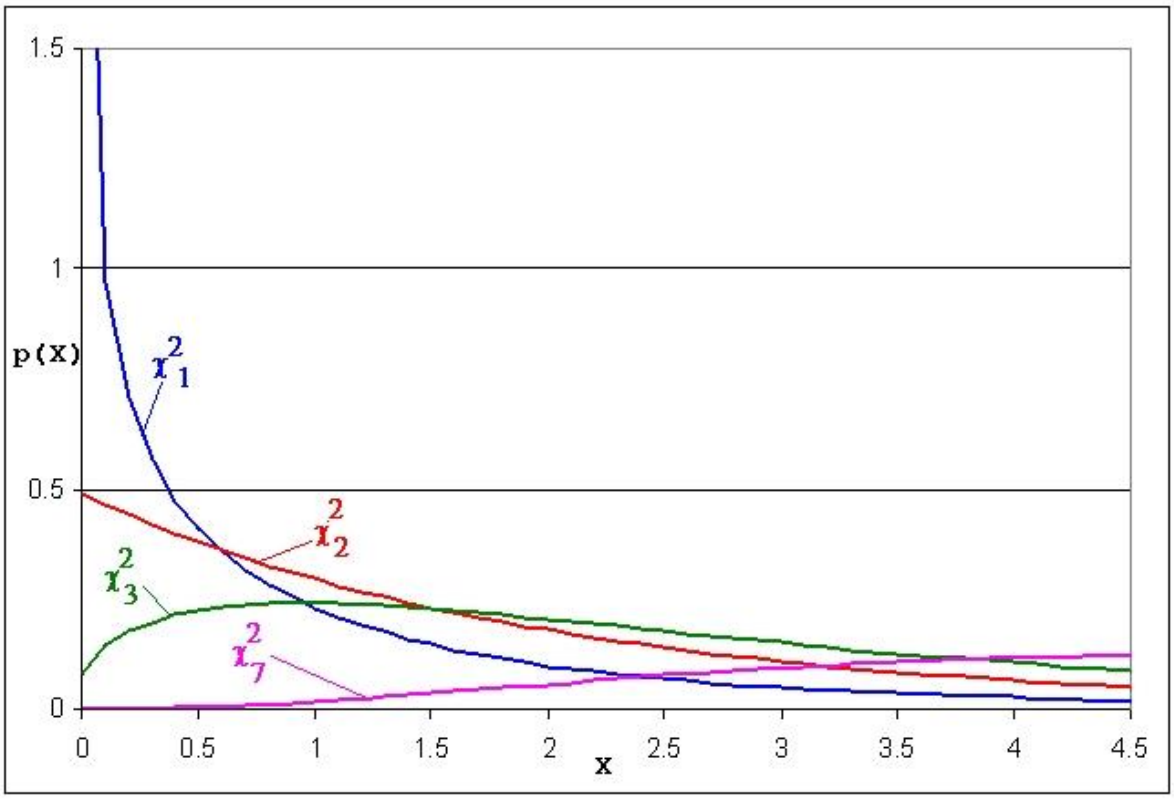
Една по-точна оценка за резултата (математическото очакване) е средното на десетте средни стойности, а за дисперсията средното на квадратите на десетте стандартни отклонения (само при равен брой на измерванията във всички серии, вижте следващите лекции). Съответно се получава $\bar{\mathbf{X}} = 6.25$ и $\mathbf{S = 0.099 = 0.1}$.

2.3. χ^2-разпределение с n степени на свобода. Това е разпределение на случайна величина $\chi^2_n = y_1^2 + y_2^2 + \ldots + y_n^2$, където y_k, $k = 1 \ldots n$ са случайни величини, разпределени стандартно (нормално с $\mu = 0$ и $\sigma^2 = 1$) и са <u>независими</u>. Цялото число **n** се нарича степени на свобода и обикновено се означава с **f** .

Плътността на разпределението се дава с уравнение (2.6) и е представена на фигура 2.2. Тя е с несиметрична форма, която силно зависи от степените свобода **n**. При голям брой степени свобода χ^2 разпределението преминава в нормално разпределение.

$$p(X) = \begin{cases} \{1/[2^{n/2}\,\Gamma(n/2)]\}\,[X^{n/2-1}\,e^{-X/2}], & \text{за } X \geq 0 \\ 0, & \text{за } X < 0 \end{cases} \qquad (2.6)$$

където **Γ()** е така наречената гама функция.



Фигура 2.2. Плътност на χ^2-разпределението при няколко степени на свобода, **f = 1**, **f = 2**, **f = 3**, **f = 7**.

Ясно се вижда от формула (2.6), че плътността на **χ^2-разпределението** е тъждествено равна на нула при X < 0.

Ако $\bar{X}$ е средната стойност на **N** измервания, а σ^2 е дисперсията на разпределението (генералната съвкупност) на тези измервания, то случайната величина

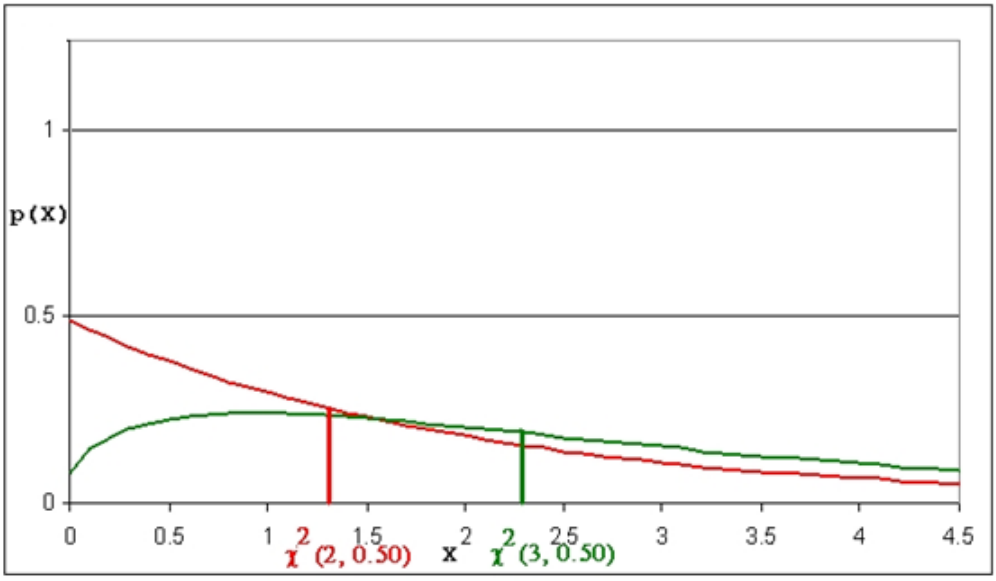
$$\Sigma(X_k - \bar{X})^2/\sigma^2 = (N - 1)\ S^2 / \sigma^2$$

е χ^2 разпределена с **N−1** степени свобода.

В приложение 2 са дадени интегралните граници на χ^2-разпределението $\chi^2(n,\alpha)$, за които функциите на разпределение **P** имат определени стойности **P = 1 − α**. В приложението тези стойности на **P** са 0.01, 0.05, 0.10, 0.50, 0.90, 0.95 и 0.99 при степените свобода **n** от 1 до

20. Числата 0.01, 0.05 и т.н. се наричат *статистическа сигурност* `P`, защото със съответната вероятност (статистическа сигурност) може да се твърди, че случайната величина заема стойности в интервала, определен от нула и интегралните граници. Т.е. интеграл от плътността на χ^2-разпределението от нула до $\chi^2(n,\alpha)$ е равен на $P = 1 - \alpha$.

Тези получени като $\alpha = 1 - P$ се наричат *статистическа значимост*. На фигура 2.3 са показани интегралните граници за две χ^2-разпределения с различни степени на свобода.



Фигура 2.3. Интегрални граници на χ^2-разпределения със степени свобода, съответно 2 и 3 и статистическа сигурност $P = 0.50$. От приложение 2 се вижда, че те са съответно $\chi^2(2,0.50) = 1.39$ и $\chi^2(3,0.50) = 2.27$.

Пример 2.2.

а) Да се пресметне вероятността случайната величина χ^2_5 да заема стойности по-малки от 4.3.

б) Да се намери интервалът $(0, X)$, в който случайната величина χ^2_9 заема стойности с вероятност 0.90.

Решение:

а) От приложение 2 се вижда, че за степени свобода 5 от интегралните граници най-близка до 4.3 е 4.35. Следователно

$$\int_{-\infty}^{4.35} p_{\chi_5^2}(x)dx \approx \int_{-\infty}^{4.3} p_{\chi_5^2}(x)dx = 0.50$$

б) Търси се X, за което

$$\int_{-\infty}^{X} p_{\chi_9^2}(x)dx = 0.90$$

Фактически долната граница на горния интеграл е нула, понеже плътността на χ^2-разпределението е нула за $x < 0$.

От същата таблица за степени свобода 9 и статистическа сигурност $P = 0.90$ се намира интегрална граница, $\chi^2(9,0.10)$ равна на 14.7. Интервалът е $(0,14.7)$.

Пример 2.3. Да се намери вероятността отношението s^2/σ^2 да е в интервала (0,1), където s е стандартното отклонение от десет измервания, а σ^2 е дисперсията на тяхното разпределение.

Решение: Условието $0 < s^2/\sigma^2 < 1$ е еквивалентно на $0 < 9\,s^2/\sigma^2 < 9$, а величината $9s^2/\sigma^2$ е χ^2_9 разпределена (s е станадартното отклонение на десетте резултата), т.е. търси се вероятността случайната величина χ^2_9 да е в интервала $(0,9)$. От приложение 2 за $f = 9$ се вижда, че 9 е между интегралните граници 8.34 (за $P = 0.50$) и 14.7 (за $P = 0.90$), но по-близко до първото число, т.е. търсената вероятност е около 0.50.

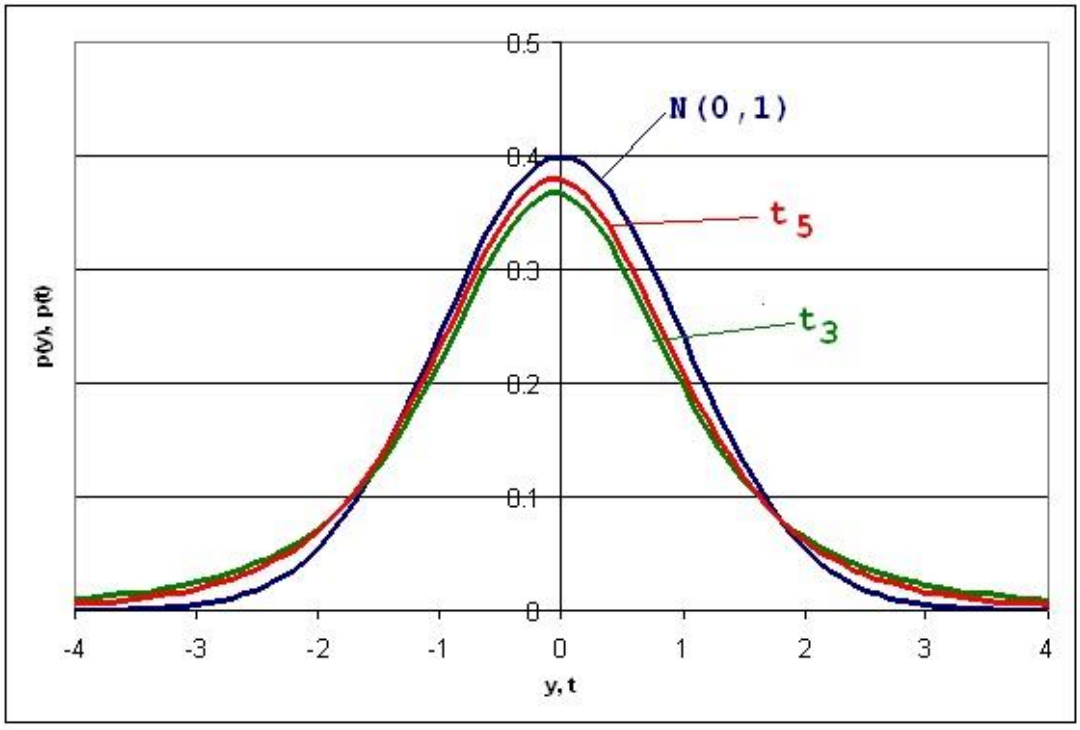
2.4. Разпределение на Стюдънт (t-разпределение). Разпределението на Стюдънт е разпределение на случайна величина $t_n = Y/X$, където Y е

стандартно разпределена случайна величина, а X е χ^2-разпределена случайна величина, със степени свобода, равни на **n**. Обикновено степените свобода се отбелязват с **f**. Плътността на t-разпределението се дава с уравнение (2.7):

$$p(t) = \Gamma[(n+1)/2]/[\mathrm{Sqrt}(n\pi)\ \Gamma(n/2)](1 + t^2/n)^{(n+1)/2}, \quad (2.7)$$

където Sqrt() означава корен квадратен, а Γ() е гама функцията.

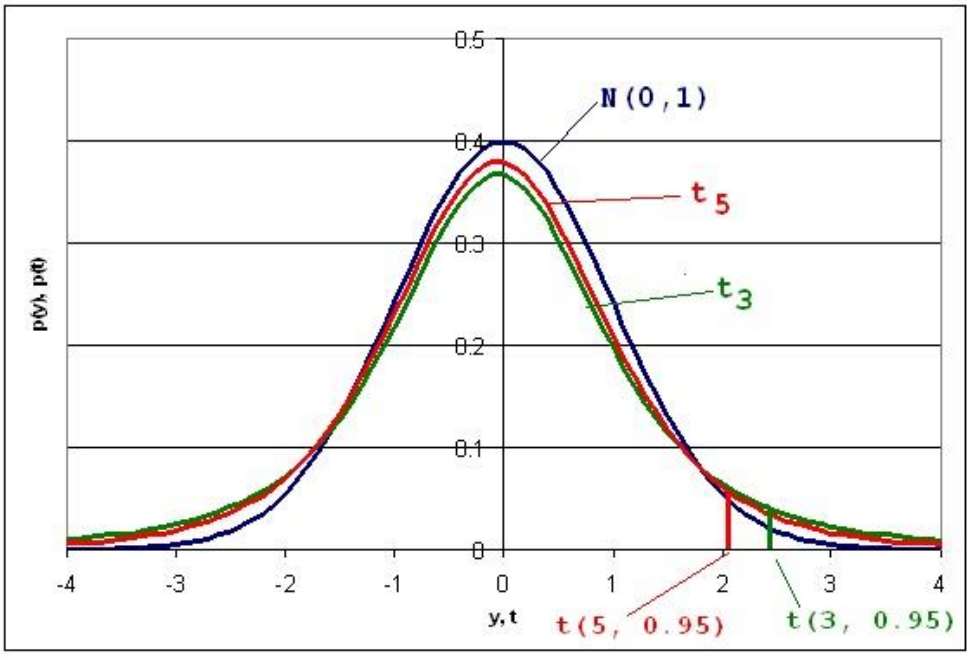
Плътността на t-разпределението е камбановидна крива, симетрична относно правата **t = 0**, която е по-малко стръмна от тази на стандартното разпределение и при нарастване на степените свобода **f** клони към нея. На фигура 2.4 са дадени няколки криви на плътността на t_f-разпределението при различни **f**, както и плътността на стандартното разпределение.



Фигура 2.4. Плътност на станадартното (синята крива) и t-разпределението при степени свобода **f = 3** и **f = 5**.

Пример за t_{N-1} разпределена величина е $t_{N-1} = (\bar{X} - \mu)Sqrt(N)/S$, където $\bar{X}$ е средната величина на N резултата, S е тяхното стандартно отклонение, а μ е математическото им очакване; резултатите X_k са нормално разпределени. Забележете, че степените свобода са равни на броя измервания минус едно, f = N - 1.

В приложение 3 са дадени интегралните граници t(f, P) на t-разпределението, т.е. решенията на F(t(f, P)) = P, за P равно на 0.50, 0.75, 0.90, 0.95, 0.98 и 0.99, както и F'(t(f, P')) = P', съответно за P' = 0.75, 0.875, 0.95, 0.975, 0.99 и 0.995. F(t) е функцията на t-разпределението, която е равна на интеграл от плътността в граници от -ω до t, а F'(t) = 2F(t) - 1 и е интеграл от плътността в граници от -t до t. За първото решение (F(X) = число) се казва, че е при едностранна постановка на въпроса, а за второто (F'(X) = число) при двустранна постановка на въпроса. Тези интегрални граници зависят от степените свобода f, което ясно се вижда на фигура 2.5.



Фигура 2.5. Интегрални граници за F(t) = 0.95 при степени свобода f =

3 и f = 5. Така записано (еф без прим) означава, че работим при едностранна постановка на въпроса. От приложение 3 намираме съответно числата t(3, 0.95) = 2.35 и t(5, 0.95) = 2.01.

Пример 2.4. Да се намери вероятността t_9-разпределена случайна величина да заема стойности:

а) по-малки от 2.8

б) в интервала (2.8, 2.8)

Решение: От приложение 3 се намира, за f = 9 и интегрална граница 2.82, че статистическата сигурност е 0.99 за едностранна постановка на въпроса и 0.98 за двустранна постановка, т.е. съответните вероятности са около 0.99 и 0.98 (Работи се с 2.82, което е близко по стойност до 2.80).

Пример 2.5. Проведени са десет измервания на олово в детски храни. Те са със средната стойност $\bar{x}$ = 2.81 мкг/мл и със станадартно отклонение s = 0.3 мкг/мл. Да се намери интервалът, за който $\bar{x}$ е среда и вероятността истинската стойност да е в него е 0.98.

Решение: Величината $t_9 = (\bar{x} - \mu)\mathrm{Sqrt}(10)/S$ е t_9 разпределена, със степени свобода f = 9. От приложение 3 се вижда, че за f = 9 и статистическа сигурност 0.98 при двустранна постановка на въпроса интегралната граница е 2.82 (вижте предишния пример!), т.е. вероятността t_9 разпределена величина да е в интервала (-2.82, 2.82) е 0.98. Приложено това към задачата дава $P(2.82 < (\bar{x} - \mu)\,\mathrm{Sqrt}(10)/S < 2.82) = 0.98$, т.е. истинската стойност е с вероятност 0.98 в интервала

$$[\bar{x} - 2.82S/\mathrm{Sqrt}(10),\ \bar{x} + 2.82S/\mathrm{Sqrt}(10)]$$

При заместване на конкретните стойности $\bar{x}$ = 2.81 и s = 0.3 се получава, че съдържанието на олово е в интервала (2.54, 3.08) с вероятност 0.98.