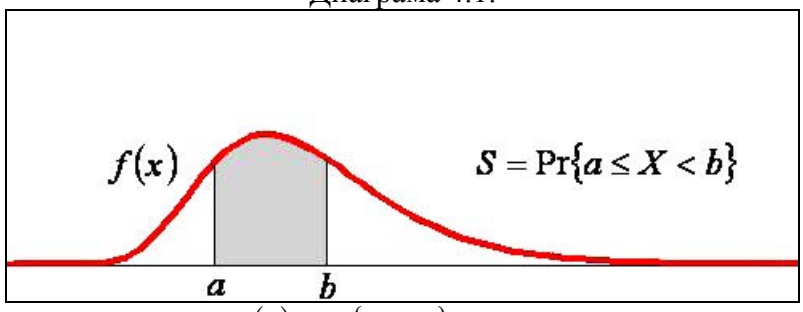
§4. Извадкови разпределения. Доверителни интервали

1. Непрекъснати величини. Разгледаното в предишния параграф нормално разпределение е частен случай на разпределение на (абсолютно) непрекъсната (случайна) величина. Тук ще говорим просто за ***непрекъснати величини***, подразбирайки онова което в теорията на вероятностите се нарича абсолютно непрекъсната случайна величина, без да навлизаме в трудните детайли. По-нататък също така ще употребяваме понятието ***вероятност*** (***probability***), разчитайки на представата, която е изградена от ежедневната употреба на това понятие. Най-общото определение за вероятност е като отношение между мярката на благоприятното към мярката на цялото, поради което вероятността $P\ (\Pr)$ е винаги число между 0 и 1. Стойността $P.100\%$ изразява вероятността в проценти. Например за $p=0.5$ имаме 50%, за $p=0.05-5\%$, за $p=0.025-2.5\%$ и т.н.

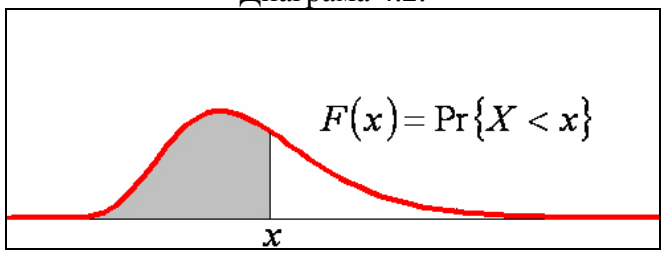
В общия случай една непрекъсната величина X се характеризира посредством своята функция на плътност (***pdf***) $f(x)$, която представлява неотрицателна функция, определена над цялата числова ос. Връзката между величината X и нейната плътност на разпределение $f(x)$ се състои в това, че за всеки две числа a и b, $(a<b)$, вероятността на събитието X да приема стойности в интервала между a и b се дава от лицето на фигурата, заградена от числовата ос на величината, графиката на $f(x)$ и двете вертикални линии през точките a и b, както е илюстрирано на диаграма 4.1.

Диаграма 4.1.



В частност функцията $F(x)=\Pr\{X<x\}$ се нарича ***функция на разпределение*** (***cumulative density function – cdf***) на величината X (диаграма 4.2).

Диаграма 4.2.



Понеже цялото лице е равно на цялата вероятност, равна на 1, то при всяко x имаме $0\le F(x)\le 1$.

В приложната статистика най-важната помощна задача се явява при известна плътност на разпределение и дадена вероятност α, $(0<\alpha<1)$, да се определи онова x, за което $F(x)=\alpha$, известна като задача за намиране ***квантил (quanlie)*** на дадено разпределение. Например нека величината Z е разпределена нормално стандартно, $Z\in N(0,1)$. Следващата таблица съдържа няколко от най-често използваните квантили, които ще означаваме чрез z_α, $\Pr(Z<z_\alpha)=\alpha$.

Таблица 4.1.

α	z_α
0.005	-2.576
0.01	-2.326
0.025	-1.96
0.05	-1.645
0.95	1.645
0.975	1.96
0.99	2.236
0.995	2.576

Поради симетрията около нулата на нормалното стандартно разпределение е в сила формулата

$z_\alpha = -z_{1-\alpha}$.

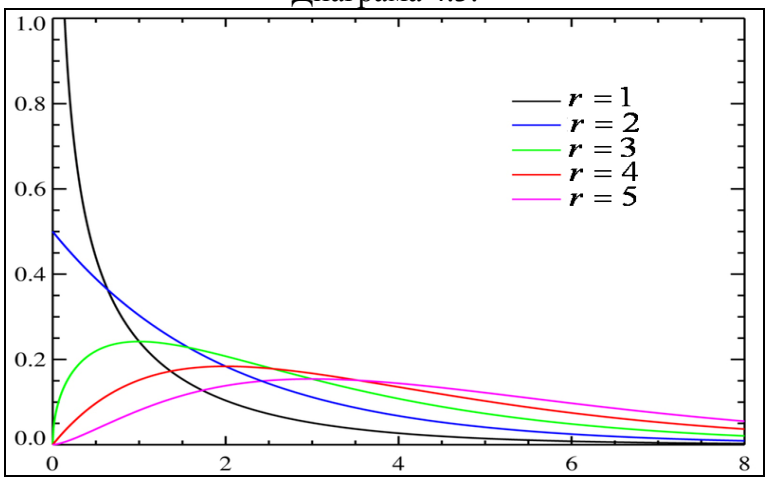
2. Основни извадкови разпределения. В този раздел ще се запознаем повърхностно с най-често използваните в приложната статистика непрекъснати разпределения, които въз основа на техния произход и начин на употреба носят името основни извадкови разпределения.

χ^2-разпределение (***chi-square distribution***). Казва се, че непрекъснатата величина X има χ^2-разпределение с r степени на свобода (***chi-square distribution with r degrees of freedom***) и се пише $X \in \chi^2(r)$, когато X има плътност

$$f(x) = \frac{x^{\frac{r}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{r}{2}} \Gamma\left(\frac{r}{2}\right)},\ x > 0,\ (\Gamma(\bullet) - \text{гама функция на Ойлер}),$$

и $f(x) = 0$ за $x \le 0$. Една величина X има $\chi^2(r)$ разпределение, когато представлява сбор от квадратите на r независими и нормално стандартно разпределени величини. На следващата диаграма са приведени няколко графики на плътност за различни r.

Диаграма 4.3.



$\chi^2(r)$ разпределението не е симетрично около средната си точка, както нормалното разпределение. Неговите квантили ще означаваме чрез $\chi^2_{\alpha;r}$, т.е. ако $X \in \chi^2(r)$, то $\Pr(X < \chi^2_{\alpha;r}) = \alpha$. Да приведем например няколко такива значения.

Таблица 4.2.

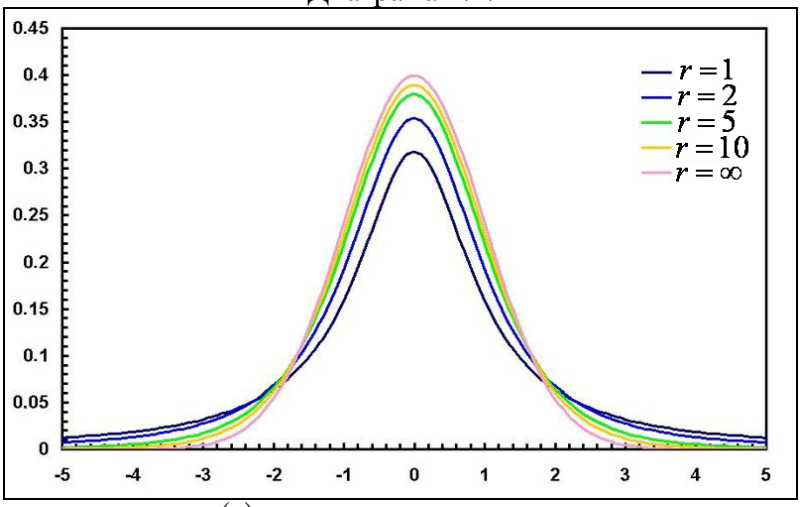
α	r	$\chi^2_{\alpha;r}$
0.05	2	$\chi^2_{0.05;2}=0.106$
0.95	2	$\chi^2_{0.95;2}=5.991$
0.025	2	$\chi^2_{0.025;2}=0.051$
0.975	2	$\chi^2_{0.975;2}=7.738$
0.005	2	$\chi^2_{0.005;2}=0.010$
0.005	2	$\chi^2_{0.995;2}=10.597$
0.05	10	$\chi^2_{0.05;10}=3.940$
0.95	10	$\chi^2_{0.95;10}=18.307$
0.025	10	$\chi^2_{0.025;10}=3.247$
0.975	10	$\chi^2_{0.975;10}=20.483$
0.005	10	$\chi^2_{0.005;10}=2.156$
0.995	10	$\chi^2_{0.995;10}=25.188$

t*-разпределение на Стюдънт.** Казва се, че непрекъснатата величина X има разпределение на Стюдънт с r степени на свобода (Student's t-distribution with r degrees of freedom***) и се пише $X \in t(r)$, когато X има плътност

$$f(x)=\frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\,\Gamma\left(\frac{r}{2}\right)\left(1+\frac{x^2}{r}\right)^{\frac{r+1}{2}}}.$$

Степените на свобода (***degrees of freedom – df***) представляват параметър на това разпределение. Разпределението на Стюдънт е симетрично около нулата и неговият външен вид наподобява твърде много нормалното стандартно разпределение. При достатъчно голямо r, например $r>120$, се приема, че разпределението на Стюдънт е на практика идентично с нормалното стандартно разпределение. На следващата диаграма са приведени няколко графики на плътност за различни r.

Диаграма 4.4.



Една величина X има $t(r)$ разпределение, когато представлява частно

$$X = \frac{Z}{\sqrt{\chi^2(r)/r}},$$

където Z и $\chi^2(r)$ са две независими величини, разпределени съответно $N(0,1)$ и $\chi^2(r)$.

Квантилите на $t(r)$ разпределението ще означаваме чрез $t_{\alpha;r}$, т.е. ако $X \in t(r)$, то $\Pr(X < t_{\alpha;r}) = \alpha$. Да приведем например няколко такива значения.

Таблица 4.3.

α	r	$t_{\alpha;r}$
0.95	10	$t_{0.95;10} = 1.812$
0.95	20	$t_{0.95;20} = 1.725$
0.95	30	$t_{0.95;30} = 1.697$
0.975	10	$t_{0.975;10} = 2.228$
0.975	20	$t_{0.975;20} = 2.086$
0.975	30	$t_{0.975;30} = 2.042$
0.99	10	$t_{0.99;10} = 2.764$
0.99	20	$t_{0.99;20} = 2.528$
0.99	30	$t_{0.99;30} = 2.457$

И тук поради симетрията на $t(r)$ разпределението около нулата е в сила формулата

$$t_{\alpha;r} = -t_{1-\alpha;r},$$

от която например въз основа на таблица 4.3 можем да пресметнем

$$t_{0.025;10} = t_{1-0.025;10} = t_{0.975;10} = -2.228.$$

F*-разпределение** (F*-distribution**). Казва се, че непрекъснатата величина X има F-разпределение на Фишер със степени на свобода m и n (***Fisher's F-distribution with degrees of freedom*** m ***and*** n) и се пише $X \in F(m,n)$, когато X има плътност

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)\sqrt{\frac{(mx)^m n^n}{(mx+n)^{m+n}}}}{x\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)},\ x > 0,$$

и $f(x) = 0$ за $x \le 0$. Една величина X има $F(m,n)$ разпределение, когато представлява частно

$$X = \frac{\chi^2(m)/m}{\chi^2(n)/n},$$

където $\chi^2(m)$ и $\chi^2(n)$ са две независими χ^2-разпределени величини със съответните степени на свобода. F-разпределението не е симетрично относно средната си точка и неговите форми наподобяват формите на χ^2-разпределението. Квантилите на $F(m,n)$ разпределението ще означаваме с $F_{\alpha;m,n}$, т.е. ако $X \in F(m,n)$, то $\Pr(X < F_{\alpha;m,n}) = \alpha$.

3. Пресмятане на квантилите. Пресмятането на квантилите за споменатите, а също така и за други разпределения, представлява сложна изчислителна задача, която

се решава автоматично от програмните среди за статистическа обработка. Преди време, когато такава възможност не е била достъпна, за решаването на тази задача са били използвани обширни таблици, които обикновено съпровождат по-старите учебници по статистика. По-нататък ще се убедим, че тази промяна има не само технически характер, но и внася съществено нов елемент в процедурите за проверка на статистически хипотези.

4. Доверителни интервали. Основната идея за доверителните интервали ще покажем върху сравнително елементарна ситуация, свързана с анализ на една метрична променлива, което обаче добре показва главните характеристики на подхода.

При детерминирането на основния статистически модел за една метрична величина X (който се състои в предположението за нормалност на величината X) приведохме точкови оценки за популационното средно μ и популационната дисперсия σ^2, посредством извадковото средно $\bar{x}$ и извадковата дисперсия s^2. Техниката на доверителните интервали позволява да се направи нещо повече. Да разгледаме отначало задачата за доверителен интервал за популационното средно μ. В математическата статистика се доказва, че ако $X \in N(\mu, \sigma^2)$, то величината

$$t_{emp} = \sqrt{n}\,\frac{\bar{x} - \mu}{s}$$

се подчинява на разпределение на Стюдънт с $n-1$ степени на свобода, $t_{emp} \in t(n-1)$. Да изберем едно достатъчно малка вероятност α, $(0 < \alpha < 1)$, обикновено $\alpha < 0.10$, която ще наречем ***ниво на значимост*** (***significance level***). Типичните стойности са $\alpha = 0.10, \alpha = 0.05$, $\alpha = 0.01$ или $\alpha = 0.001$. Стойността по ***подразбиране*** във всичките програмни среди за статистическа обработка е $\alpha = 0.05\,(5\%)$. В такъв случай вероятността $\gamma = 1 - \alpha$ се нарича ***ниво на доверие (confidence level)***. При такъв избор на α стойността на γ се получава обикновено по-голяма от $0.9\,(90\%)$. Тогава следвайки специфични математически съображения се получава, че $(1-\alpha)100\%$ доверителен интервал за популационното средно се получава интервал с краища

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2};n-1},$$

или в явен вид интервала

(4.1) $$\left[\bar{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2};n-1}, \bar{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2};n-1}\right].$$

За този интервал се твърди, че с $(1-\alpha)100\%$ ***доверителна вероятност*** (с ниво на доверие $\gamma = 1 - \alpha$) съдържа популационното средно μ, което твърдение трябва да се разбира по следния начин. Ако направим по случаен начин известен брой извадки с един и същ обем n, то (приблизително) в $(1-\alpha)100\%$ случаите популационното средно ще се намира вътре във въпросния доверителен интервал.

При голям брой степени на свобода имаме $t_{\alpha;n} \approx z_\alpha$, което дава възможност в този случай да опростим формулата (4.1) до

$$\left[\bar{x} + \frac{s}{\sqrt{n}} z_{\frac{\alpha}{2}}, \bar{x} - \frac{s}{\sqrt{n}} z_{\frac{\alpha}{2}}\right],$$

което дава така наречения ***асимптотичен*** доверителен интервал.

С помощта на горния резултат можем да правим истински съдържателни статистически анализи. Да разгледаме следния илюстративен пример (***EX04IQ***). Нека е извършено изследване за нивото на интелигентност посредством IQ-тест за

интелигентност на група от $n=64$ обучаеми в отговаряща на теста възраст, случайно подбрани от специализирано учебно заведение. При IQ-тестовете за интелигентност е известно, че нормата за среден резултат е $\mu_0=100$, което естествено приемаме за средно за онази популация, към която е стандартизиран тестът. Нека след обработка на резултатите от извадката са получени стойности $\bar{x}=106.547$ и $s=9.976$. В този случай имаме $\bar{x}=106.547>100=\mu_0$, т.е. средното от извадката надвишава нормата за средно, но е възможно този резултат да се дължи на известна игра на случайността. На базата на този факт можем да вземем решение, че средните постижения на учениците от специализираното учебно заведение са по-високи от нормата, което решение обаче включва съществен елемент на неопределеност. Тук техниката на доверителните интервали предлага по-добър критерий. Да изберем $\alpha=0.05$ и да пресметнем $(1-\alpha)100\%=95\%$ доверителен интервал по формулата (4.1). Получаваме интервала

$$\left[106.547-\frac{9.976}{\sqrt{64}}1.998,\ 106.547+\frac{9.976}{\sqrt{64}}1.998\right]=[104.055,109.039],\ \left(t_{0.025;63}=-1.998\right)$$

което показва, че нормата (популационното средно) $\mu_0=100$ не се съдържа в този интервал, а по-точно лежи вляво от него. Последният факт вече представлява убедително доказателство, че средните постижения в тази група надвишават съществено нормата за средно. При $\alpha=0.01$ имаме $t_{0.005;63}=-2.656$, откъдето получаваме следния $(1-\alpha)100\%=99\%$ доверителен интервал

$$\left[106.547-\frac{9.976}{\sqrt{64}}2.656,\ 106.547+\frac{9.976}{\sqrt{64}}2.656\right]=[103.235,109.859],$$

което потвърждава направеният извод с по-висока сигурност.

По-нататък ще дадем друго решение на разисквания проблем чрез техниката на проверка на статистически хипотези, което решение в основните си характеристики е идентично с приведеното.

Доверителни интервали за дисперсията се пресмятат въз основа на факта, че ако $X\in N(\mu,\sigma^2)$, то величината

$$\chi^2_{emp}=(n-1)\frac{s^2}{\sigma^2}$$

се подчинява на χ^2-разпределение с $n-1$ степени на свобода, $\chi^2_{emp}\in\chi^2(n-1)$. Следвайки специфични статистически съображение, последният факт позволява да пресметнем $(1-\alpha)100\%$ доверителен интервал по формулата

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}},\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right].$$

За хипотетичният пример с IQ-теста, при $\alpha=0.05$ имаме $\chi^2_{0.025;63}=42.950$ и $\chi^2_{0.975;63}=86.830$, което дава следният 95% доверителен интервал

$$\left[\frac{(64-1)(9.976)^2}{86.830},\frac{(64-1)(9.976)^2}{42.950}\right]=[72.207,145.979],$$

за популационната дисперсия.

По-нататък в настоящия лекционен курс ще приведем начини за пресмятане на доверителни интервали в по-сложни ситуации.

5. Защо точно $\alpha = 0.05\,(5\%)$? В примера от предишния раздел избрахме да пресметнем 95% доверителен интервал, въз основа на което направихме заключението за съществено различие между средното за постиженията от групата и нормата за средно. Ако бяхме избрали по-голяма стойност за α, то доверителният интервал щеше да се получи по-къс и следователно щяха да се увеличат шансовете да направим същия извод, но от друга страна изводът щеше да се получи по-малко достоверен. В този случай централният въпрос е къде е компромисната бариерна стойност, която разделя между естественото желание на даден експериментатор да потвърди съществуването на някакъв ефект и достоверността на извода, че такъв ефект съществува. ***Общоприетата*** такава бариерна стойност е $\alpha = 0.05$, което няма значение за математиката и представлява въпрос на договореност в средите на изследователите в областта на социалните (а също така и на другите) науки. В следващата лекция, посветена на проверка на статистически хипотези, стойността на нивото на значимост α ще придобие по-ясни очертания, запазвайки първоначалния си смисъл.