Непараметрические критерии согласия.

Проблемы применения и их решение

Лемешко Борис Юрьевич

E-mail: Lemeshko@ami.nstu.ru
http://www.ami.nstu.ru/~headrd/

2.2. Распределения статистик непараметрических критериев при простых гипотезах

2.2.1. Критерий Колмогорова

В случае простых гипотез предельные распределения статистик непараметрических критериев согласия Колмогорова, Смирнова, ω^2 Крамера-Мизеса-Смирнова и Ω^2 Андерсона-Дарлинга известны и не зависят от вида наблюдаемого закона распределения и, в частности, от его параметров. В этой связи их называют “свободными от распределения”. Это достоинство предопределило широкое использование данных критериев в различных приложениях.

Предельное распределение статистики

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x,\theta)|, \qquad (4)$$

где $F_n(x)$ – эмпирическая функция распределения, $F(x,\theta)$ – теоретическая функция распределения, n – объем выборки, было получено Колмогоровым. При $n \to \infty$ функция распределения статистики $\sqrt{n}D_n$ сходится равномерно к функции распределения Колмогорова

$$K(s) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}. \qquad (5)$$

Наиболее часто в критерии Колмогорова (Колмогорова-Смирнова) используют статистику с поправкой Большева вида

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}}, \qquad (6)$$

где

$$D_n = \max(D_n^+, D_n^-), \quad (7) \qquad D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F(x_i,\theta)\right\}, \quad (8) \qquad D_n^- = \max_{1\le i\le n}\left\{F(x_i,\theta) - \frac{i-1}{n}\right\}, \quad (9)$$

n - объем выборки, $x_1, x_2, \ldots, x_n$ - упорядоченные по возрастанию выборочные значения, $F(x,\theta)$ - функция закона распределения, согласие с которым проверяют.

Если для вычисленного по выборке значения статистики S_K^* выполняется неравенство

$$P\{S > S_{\text{K}}^*\} = 1 - K(S_{\text{K}}^*) > \alpha,$$

то нет оснований для отклонения гипотезы H_0.

2.2.2. Критерий Смирнова

В критерии Смирнова используют статистику

$$D_n^+ = \sup_{|x|<\infty}\left(F_n(x) - F(x,\theta)\right) \tag{10}$$

или статистику

$$D_n^- = -\inf_{|x|<\infty}\left(F_n(x) - F(x,\theta)\right), \tag{11}$$

значения которых вычисляют по эквивалентным соотношениям (8), (9).

Реально в критерии обычно используют статистику [3]

$$S_m = \frac{(6nD_n^+ + 1)^2}{9n}, \tag{12}$$

которая при простой гипотезе в пределе подчиняется распределению χ^2 с числом степеней свободы, равным 2.

Гипотезу H_0 не отвергают, если для вычисленного по выборке значения статистики S_m^*

$$P\{S_m > S_m^*\} = \int_{S_m^*}^{\infty} \frac{1}{2} e^{-x/2} dx = e^{-S_m^*/2} > \alpha.$$

2.2.3. Критерии ω^2

В критериях типа ω^2 расстояние между гипотетическим и истинным распределениями рассматривают в квадратичной метрике.

Проверяемая гипотеза H_0 имеет вид

$$H_0: \int_{-\infty}^{\infty} \left\{E\left[F_n(x)\right] - F(x)\right\}^2 \psi\left(F(x)\right) dF(x) = 0 \qquad (13)$$

при альтернативной гипотезе

$$H_1: \int_{-\infty}^{\infty} \left\{E\left[F_n(x)\right] - F(x)\right\}^2 \psi\left(F(x)\right) dF(x) > 0, \qquad (14)$$

где $E[\cdot]$ - оператор математического ожидания, $\psi(t)$ - заданная на отрезке $0 \le t \le 1$ неотрицательная функция, относительно которой предполагают, что $\psi(t)$, $t\,\psi(t)$, $t^2\psi(t)$ интегрируемы на отрезке $0 \le t \le 1$. Статистику критерия выражают соотношением

$$\omega_n^2\left[\psi(F)\right] = \int_{-\infty}^{\infty} \left\{E\left[F_n(x)\right] - F(x)\right\}^2 \psi\left(F(x)\right) dF(x) =$$

$$= \frac{2}{n} \sum_{i=1}^{n} \left\{ g\left[F(x_i)\right] - \frac{2i-1}{2n} f\left[F(x_i)\right] \right\} + \int_0^1 (1-t)^2 \psi(t)\, dt, \qquad (15)$$

где

$$f(t) = \int_0^t \psi(s)\, ds, \quad g(t) = \int_0^t s\psi(s)\, ds.$$

При выборе $\psi(t) \equiv 1$ для критерия ω^2 Мизеса получают статистику вида (статистику **Крамера-Мизеса-Смирнова**)

$$S_{\omega} = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left\{ F(x_i, \theta) - \frac{2i-1}{2n} \right\}^2, \qquad (16)$$

которая при простой гипотезе в пределе подчиняется закону с функцией распределения $a1(s)$, имеющей вид

$$a1(s) = \frac{1}{\sqrt{2s}} \sum_{j=0}^{\infty} \frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)} \exp\left\{ -\frac{(4j+1)^2}{16s} \right\} \times$$

$$\times \left\{ I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16s} \right] - I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16s} \right] \right\}, \qquad (17)$$

где $I_{-\frac{1}{4}}(\cdot)$, $I_{\frac{1}{4}}(\cdot)$ - модифицированные функции Бесселя,

$$I_{\nu}(z) = \sum_{k=0}^{\infty} \frac{\left(\frac{z}{2}\right)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)}, \quad |z| < \infty, \quad |\arg z| < \pi. \qquad (18)$$

При выборе $\psi(t) \equiv 1/t(1-t)$ для критерия Ω^2 Мизеса статистика приобретает вид (статистика **Андерсона-Дарлинга**)

При выборе $\psi(t) \equiv 1/t(1-t)$ для критерия Ω^2 Мизеса статистика приобретает вид (статистика **Андерсона-Дарлинга**)

$$S_\Omega = n\Omega_n^2 = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i,\theta) + \left(1-\frac{2i-1}{2n}\right)\ln(1-F(x_i,\theta))\right\}. \qquad (19)$$

В пределе эта статистика подчиняется закону с функцией распределения $a2(s)$, имеющей вид

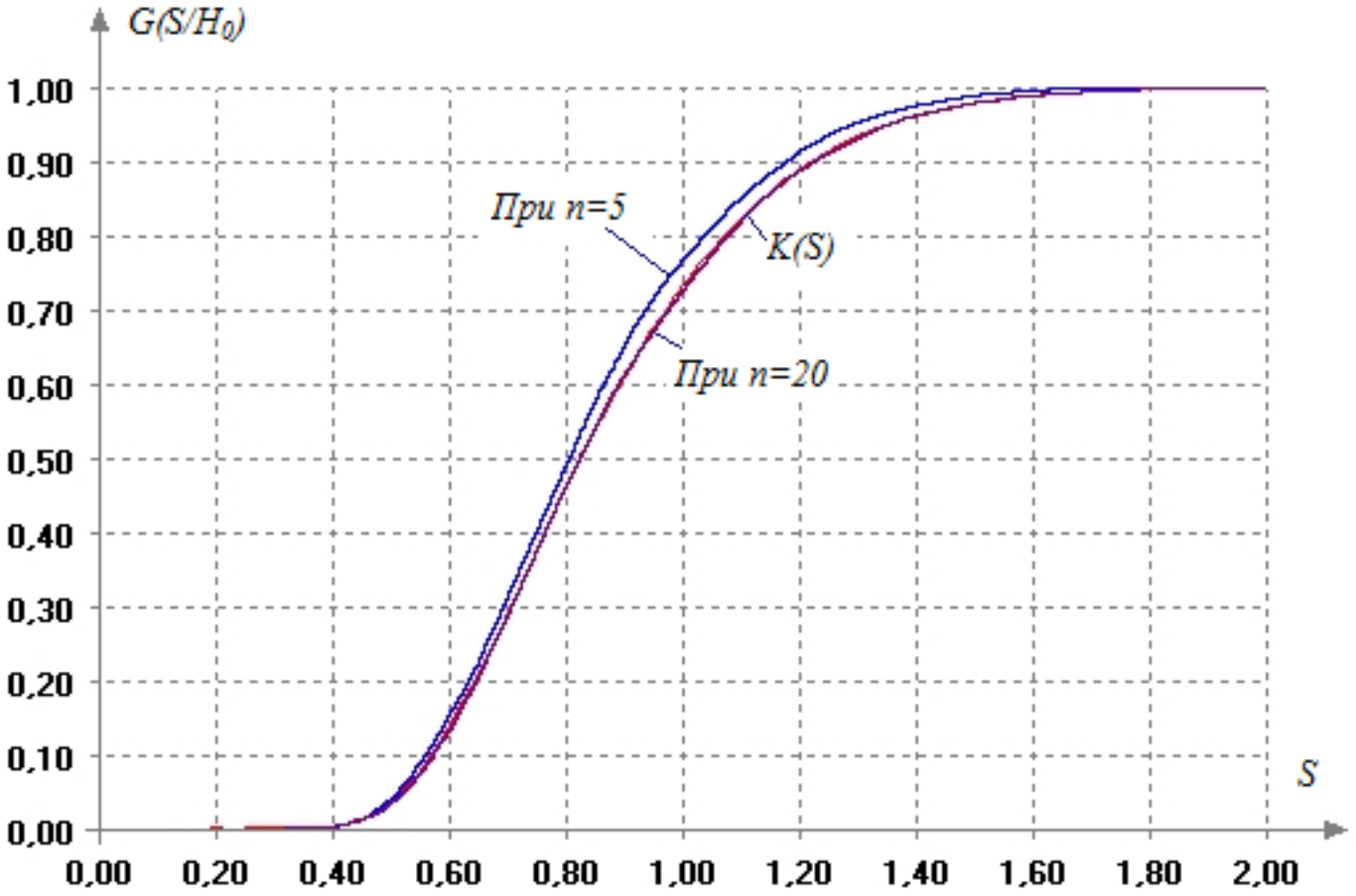
$$a2(s) = \frac{\sqrt{2\pi}}{s}\sum_{j=0}^{\infty}(-1)^j \frac{\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2\pi^2}{8s}\right\}\times$$

$$\times\int_0^{\infty}\exp\left\{\frac{s}{8(y^2+1)}-\frac{(4j+1)^2\pi^2 y^2}{8s}\right\}dy. \qquad (20)$$

Гипотезы о согласии не отвергают, если выполнены неравенства

$$P\{S_\omega > S_\omega^*\} = 1 - a1(S_\omega^*) > \alpha \quad \text{и} \quad P\{S_\Omega > S_\Omega^*\} = 1 - a2(S_\Omega^*) > \alpha.$$

Сходимость распределения статистики Колмогорова к предельному при проверке простых гипотез



2.3 Непараметрические критерии согласия при сложных гипотезах

2.3.1 Потеря критериями свойства “свободы от распределения”

При проверке сложных гипотез, когда по той же самой выборке оценивают параметры наблюдаемого закона распределения вероятностей, непараметрические критерии согласия Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса теряют свойство “свободы от распределения”. В этом случае предельные распределения статистик этих критериев будут зависеть от закона, которому подчинена наблюдаемая выборка. Более того, распределения статистик непараметрических критериев согласия зависят и от используемого метода оценивания параметров. Следует также учитывать, что распределения статистик существенно зависят от объема выборки.

Игнорирование того, что проверяют сложную гипотезу, игнорирование различия в сложных гипотезах приводят к некорректному применению непараметрических критериев согласия и, как следствие, к неверным статистическим выводам. Различия в предельных распределениях тех же самых статистик при проверке простых и сложных гипотез настолько существенны, что пренебрегать этим абсолютно недопустимо [5]-[7].

Точкой отсчета, с которой были начаты исследования предельных распределений статистик непараметрических критериев согласия при сложных гипотезах, послужила работа [8].

2.3.3. Факторы, влияющие на распределения статистик критериев при проверке сложных гипотез

Распределения статистик непараметрических критериев согласия при проверке сложных гипотез зависят от характера этой сложной гипотезы. На закон распределения статистики $G(S|H_0)$ влияют следующие факторы, определяющие “сложность” гипотезы:

– вид наблюдаемого закона распределения $F(x,\theta)$, соответствующего истинной гипотезе H_0;

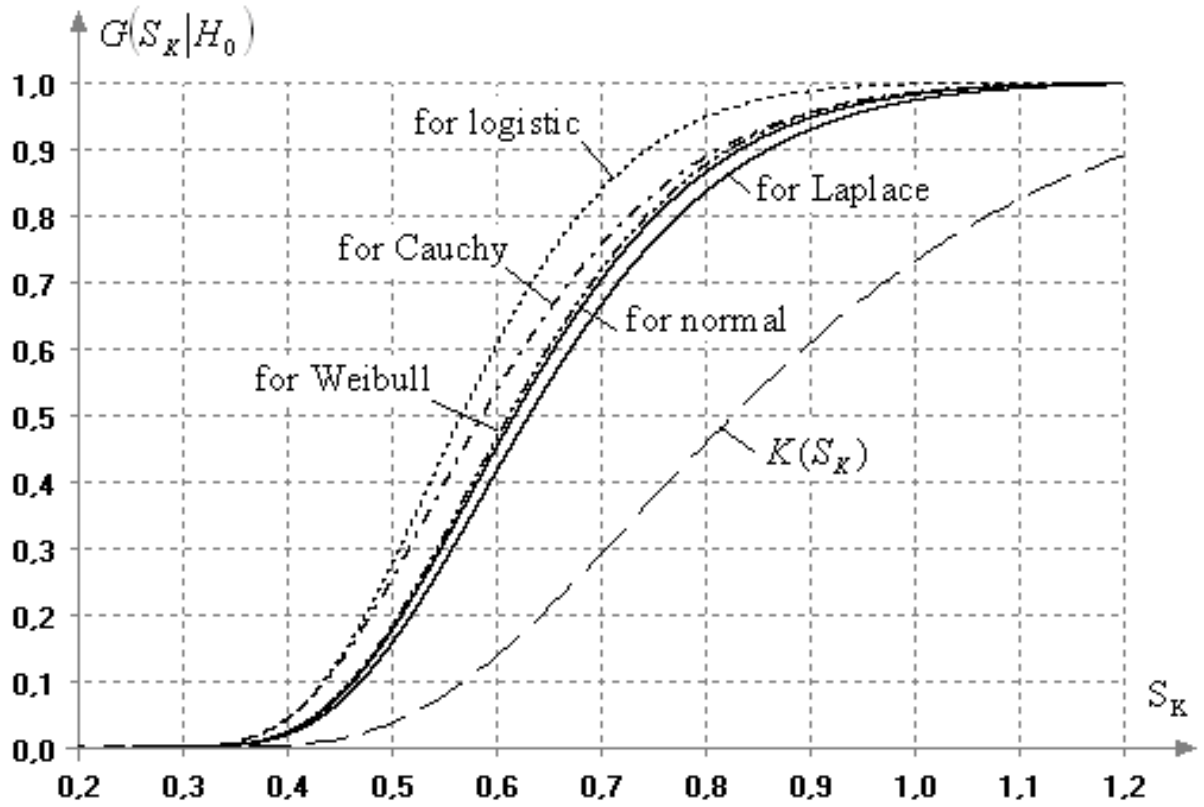
– тип оцениваемого параметра и число оцениваемых параметров;

– в некоторых ситуациях конкретное значение параметра (например, в случае гамма-распределения, бета-распределений);

– используемый метод оценивания параметров.

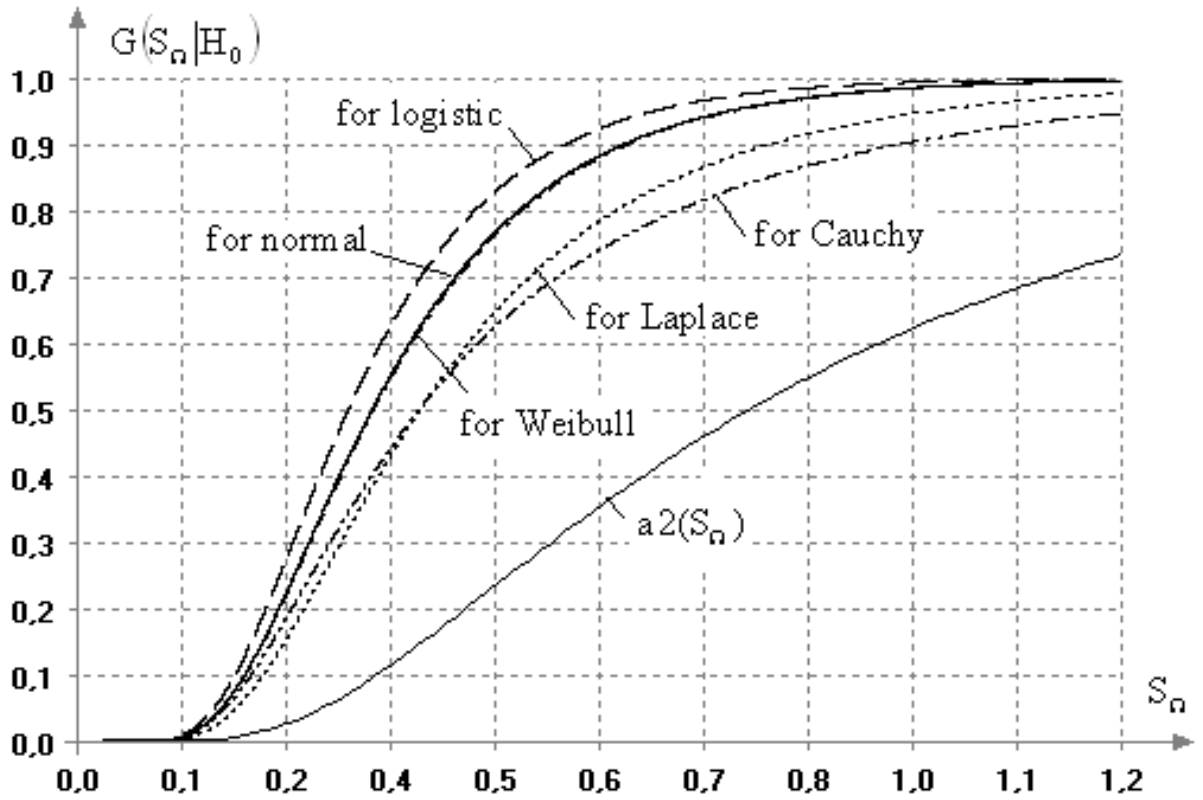
При малых объемах выборки n распределение $G(S_n|H_0)$ зависит от n. Однако существенная зависимость распределения статистики от n наблюдается только при небольших объемах выборки. Уже при $n \geq 15$–20 распределение $G(S_n|H_0)$ достаточно близко к предельному $G(S|H_0)$ и зависимостью от n можно пренебречь.

Зависимость распределений статистик при проверке сложных гипотез от вида закона



Распределения статистики Колмогорова (1) при проверке сложных гипотез с вычислением ОМП 2-х параметров закона

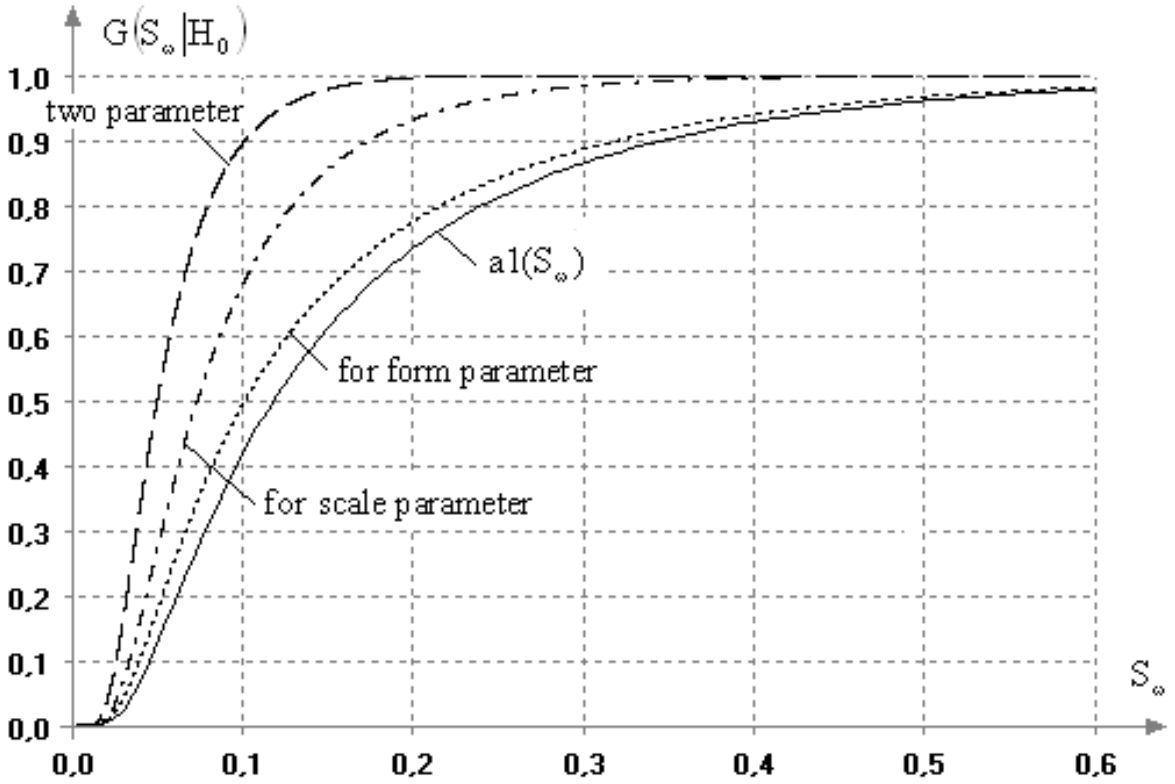
Зависимость распределений статистик при проверке сложных гипотез от вида закона



Распределения статистики Андерсона-Дарлинга при проверке сложных гипотез при вычислении ОМП 2-х параметров закона

Зависимость распределений статистик от количества и вида оцененных параметров

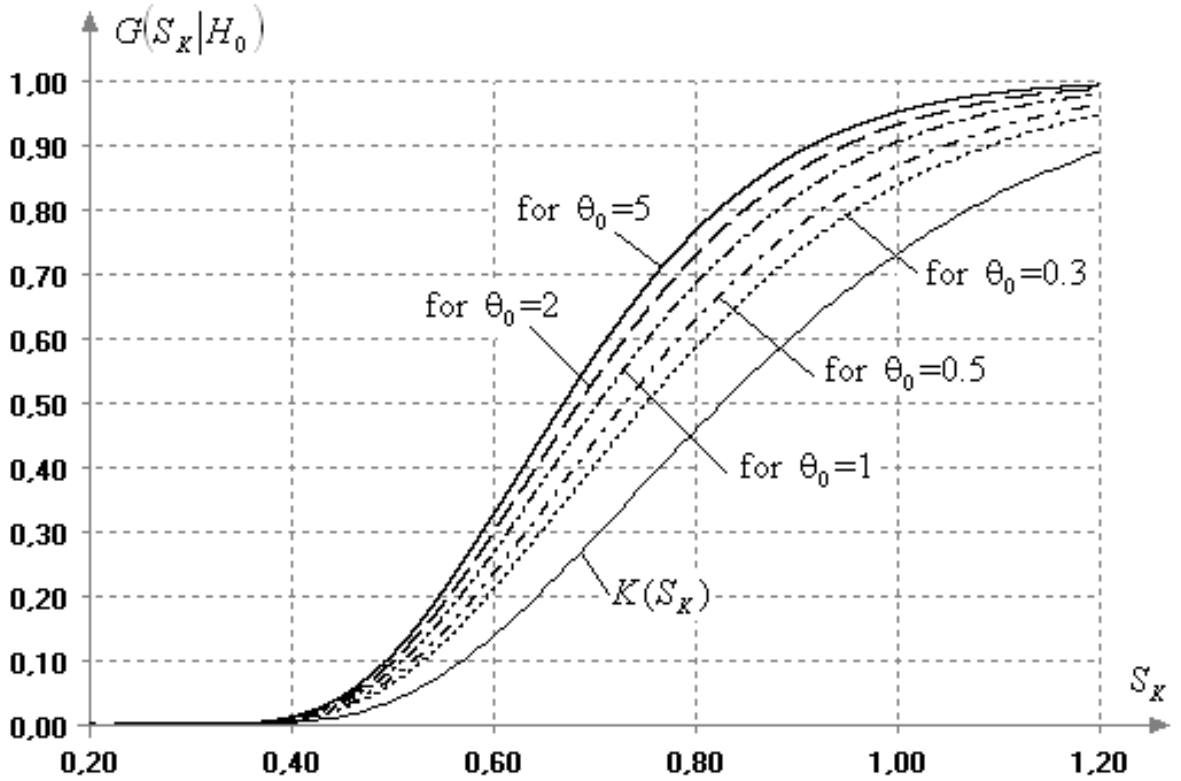
$$f(x,\theta) = \frac{\theta_0 x^{\theta_0 - 1}}{\theta_1^{\theta_0}} \exp\left\{-\left(\frac{x}{\theta_1}\right)^{\theta_0}\right\}$$



Распределения статистики Крамера-Мизеса-Смирнова при проверке сложных гипотез относительно распределения Вейбулла при вычислении ОМП различных параметров закона

Зависимость распределений статистик при проверке сложных гипотез от значения параметра (формы) гамма-распределения

$$f(x,\theta)=\frac{x^{\theta_0-1}}{\theta_1^{\theta_0}\Gamma(\theta_0)}\exp\left(-\frac{x}{\theta_1}\right)$$

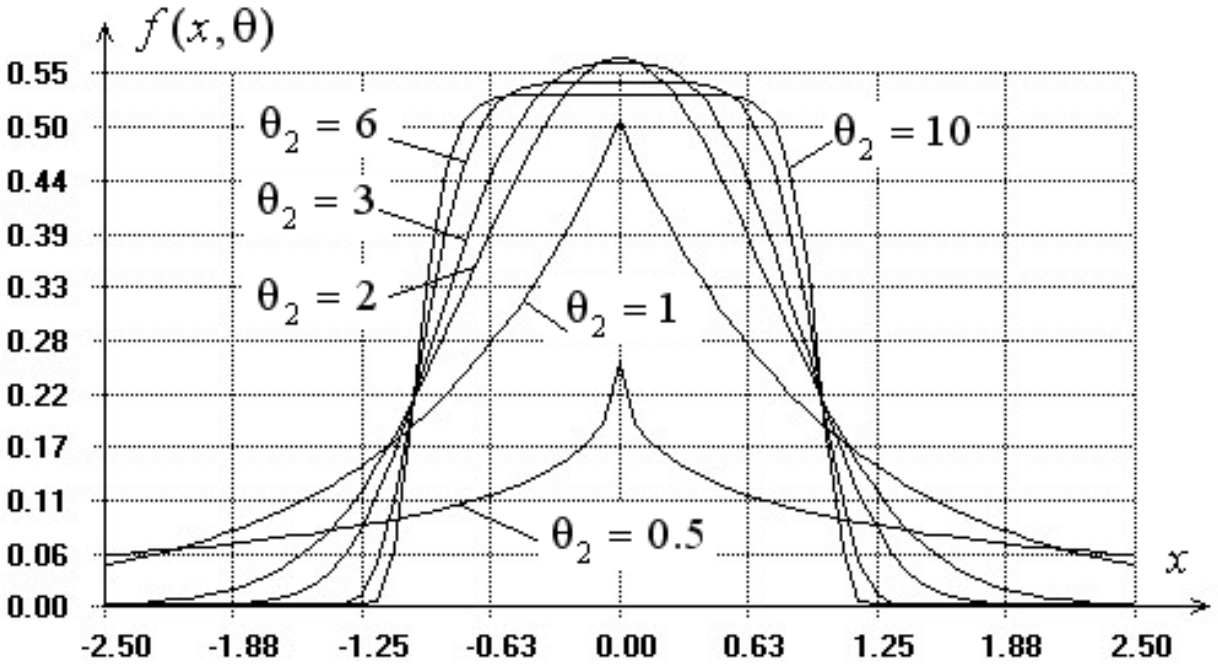


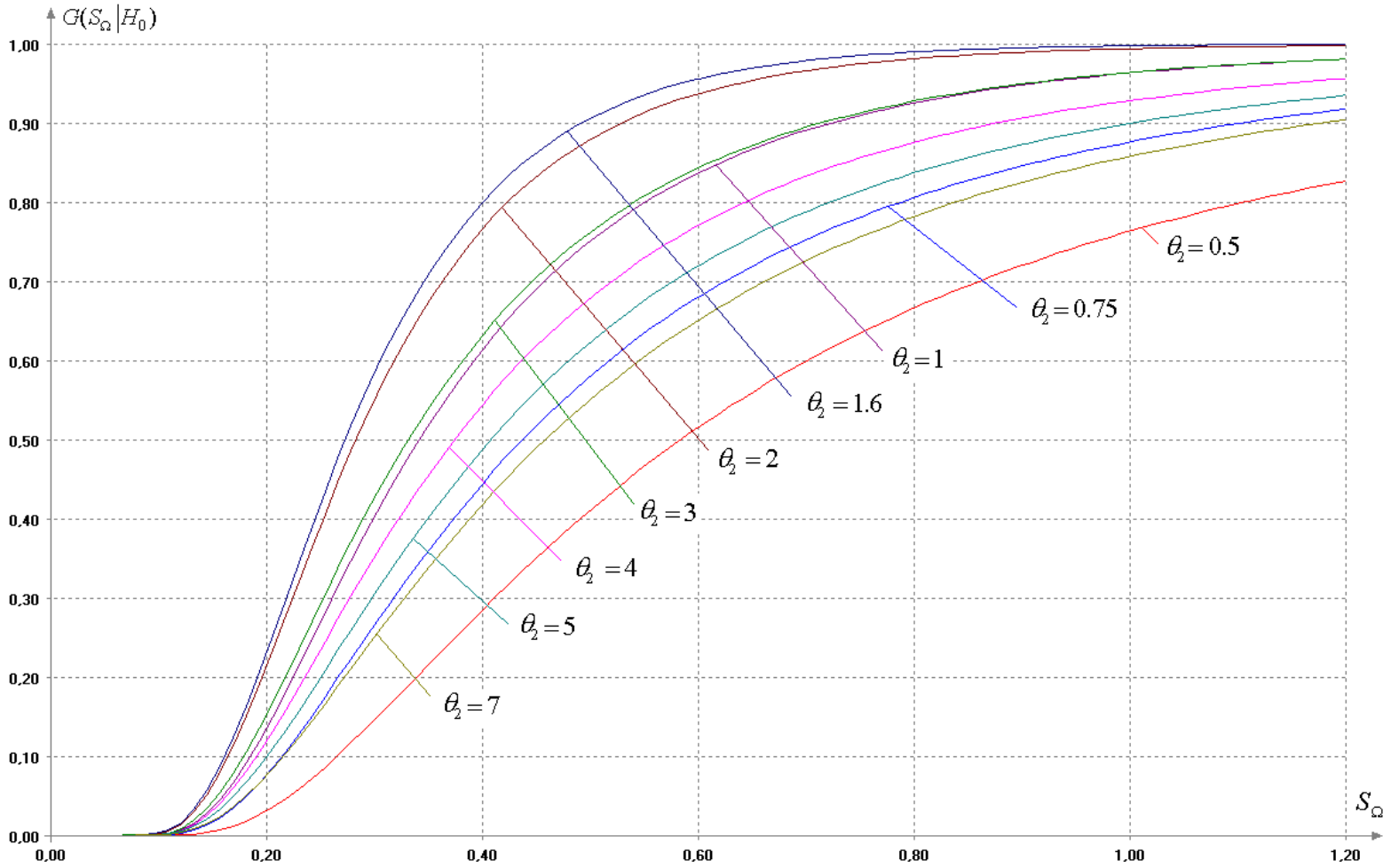
Распределения статистики Колмогорова при проверке сложных гипотез при вычислении ОМП только параметра масштаба гамма-распределения в зависимости от значения параметра формы

Интересно меняются распределения статистик непараметрических критериев согласия при проверке гипотез о согласии с распределениями семейства с плотностью

$$f(x,\boldsymbol{\theta}) = \frac{\theta_2}{2\theta_1\Gamma(1/\theta_2)}\exp\left\{-\left(\frac{|x-\theta_0|}{\theta_1}\right)^{\theta_2}\right\}. \qquad (4)$$

В зависимости от значения параметра формы θ_2 вид плотности этого закона показан на рисунке. Распределения нормальное и Лапласа являются частными случаями при значениях параметра θ_2, равных 2 и 1 соответственно.





Зависимость распределения статистики Андерсона-Дарлинга от значения параметра θ_2 распределения (4) (при вычислении ОМП трёх параметров)

Распределения $G(S|H_0)$ статистики Колмогорова хорошо аппроксимируются семейством гамма-распределений с функцией плотности

$$\gamma(\theta_0,\theta_1,\theta_2)=\frac{1}{\theta_1^{\theta_0}\Gamma(\theta_0)}(x-\theta_2)^{\theta_0-1}e^{-(x-\theta_2)/\theta_1}.$$

А распределения статистик Крамера-Мизеса-Смирнова и Андерсона-Дарлинга неплохо приближаются семейством распределений *Sb*-Джонсона

$$Sb(\theta)=\frac{\theta_1\theta_2}{(x-\theta_3)(\theta_2+\theta_3-x)}\exp\left\{-\frac{1}{2}\left[\theta_0-\theta_1\ln\frac{x-\theta_3}{\theta_2+\theta_3-x}\right]^2\right\}.$$

Распределения всех статистик при проверке сложных гипотез относительно гамма-распределения хорошо приближаются семейством бета-распределений III-го рода

$$B_3(\theta_0,\theta_1,\theta_2,\theta_3,\theta_4)=\frac{\theta_2^{\theta_0}}{\theta_3 B(\theta_0,\theta_1)}\frac{\left(\frac{x-\theta_4}{\theta_3}\right)^{\theta_0-1}\left(1-\frac{x-\theta_4}{\theta_3}\right)^{\theta_1-1}}{\left[1+(\theta_2-1)\frac{x-\theta_4}{\theta_3}\right]^{\theta_0+\theta_1}}.$$

Table. Models of limiting statistic distributions of nonparametric goodness-of-fit when MLE are used.

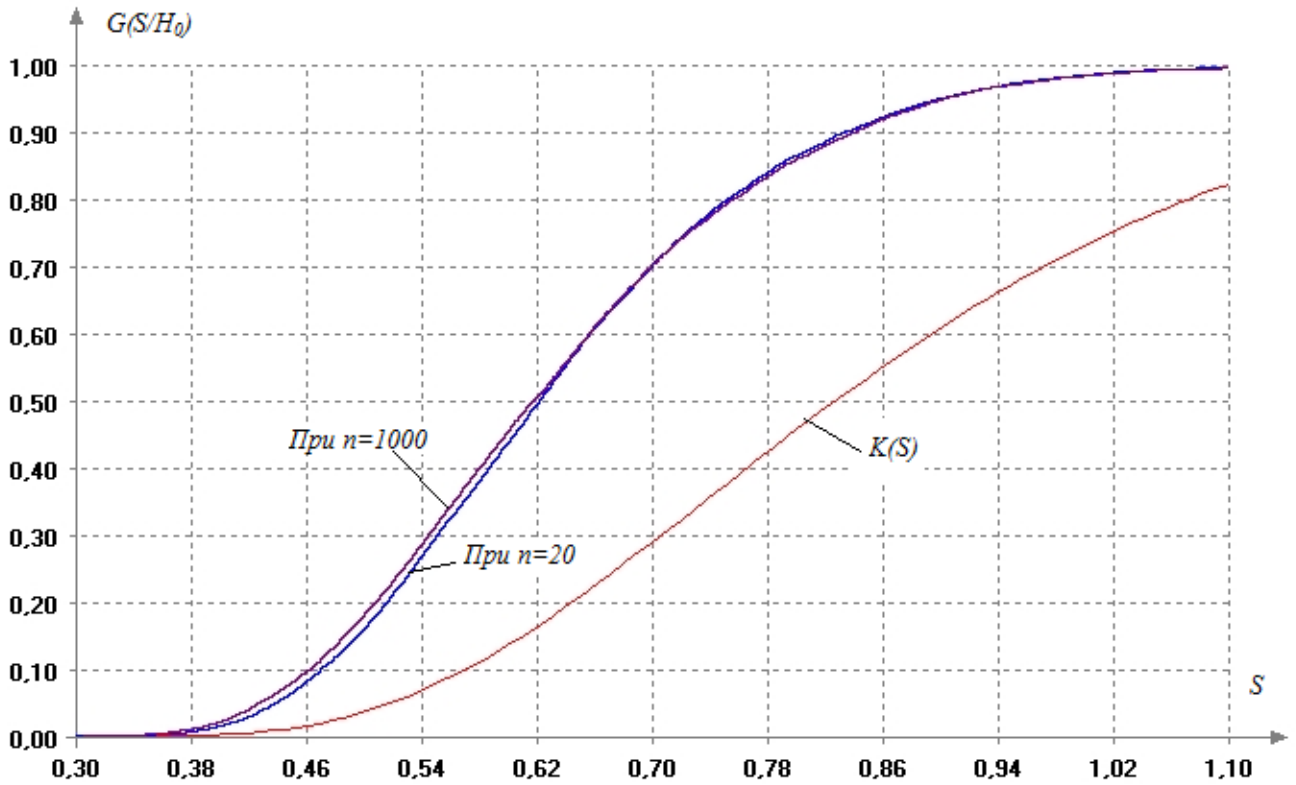
Test	Random variable distribution	Estimation of scale parameter	Estimation of shift parameter	Estimation of two parameters
Kolmogorov's	Exponential & Rayleigh	γ(5.1092; 0.0861; 0.2950)	–	–
	Seminormal	γ(4.5462; 0.1001; 0.3100)	–	–
	Maxwell	γ (5.4566; 0.0794; 0.2870)	–	–
	Laplace	γ(3.3950; 0.1426; 0.3405)	γ(6.2887; 0.0718; 0.2650)	γ(6.2949; 0.0624; 0.2613)
	Normal & Log-normal	γ(3.5609; 0.1401; 0.3375)	γ(7.5304; 0.0580; 0.2400)	γ(6.4721; 0.0580; 0.2620)
	Cauchy	γ (3.0987; 0.1463; 0.3350)	γ(5.9860; 0.0780; 0.2528)	γ(5.3642; 0.0654; 0.2600)
	Logistic	γ(3.4954; 0.1411; 0.3325)	γ (7.6325; 0.0531; 0.2368)	γ (7.5402; 0.0451; 0.2422)
	Extreme-value & Weibull	γ(3.6805; 0.1355; 0.3350) [1)]	γ(5.2194; 0.0848; 0.2920) [2)]	γ(6.6012; 0.0563; .2598)
Cramer-Mises-Smirnov's	Exponential & Rayleigh	*Sb*(3.3738;1.2145; 1.0792; 0.011)	–	–
	Seminormal	*Sb*(3.527; 1.1515; 1.5527; 0.012)	–	–
	Maxwell	*Sb*(3.353; 1.220; 0.9786; 0.0118)	–	–
	Laplace	*Sb*(3.2262; 0.9416; 2.703; 0.015)	*Sb*(2.9669; 1.2534; 0.6936; 0.01)	*Sb*(3.768; 1.2865; 0.8336; 0.0113)
	Normal & Log- normal	*Sb*(3.153; 0.9448; 2.5477; 0.016)	*Sb*(3.243; 1.315; 0.6826; 0.0095)	*Sb*(4.3950; 1.4428; 0.915; 0.009)
	Cauchy	*Sb*(3.1895; 0.9134; 2.690; 0.013)	*Sb*((2.359; 1.0732; 0.595; 0.0129)	*Sb*(3.4364; 1.0678; 1.000; 0.011)
	Logistic	*Sb*(3.264; 0.9581; 2.7046; 0.014)	*Sb*(4.0026; 1.2853; 1.00; 0.0122)	*Sb*(3.2137; 1.3612; 0.36; 0.0105)
	Extreme-value & Weibull	*Sb*(3.343; 0.9817; 2.753; 0.015) [1)]	*Sb*(3.498; 1.2236; 1.1632; 0.01) [2)]	*Sb*(3.3854; 1.4453; 0.4986; 0.007)
Anderson-Darling's	Exponential & Rayleigh	*Sb*(3.8386; 1.3429; 7.500; 0.090)	–	–
	Seminormal	*Sb*(4.2019; 1.2918; 11.500; 0.100)	–	–
	Maxwell	*Sb*(3.9591; 1.3296; 7.800; 0.1010)	–	–
	Laplace	*Sb*(4.3260; 1.0982; 27.00; 0.110)	*Sb*(3.1506; 1.3352; 4.9573; 0.096)	*Sb*(3,8071; 1,3531; 5,1809; 0,10)
	Normal & Log- normal	*Sb*(4.3271; 1.0895; 28.000; 0.120)	*Sb*(3.3085; 1.4043; 4.2537; 0.080)	*Sb*(3.5601; 1.4846; 3.0987; 0.08)
	Cauchy	*Sb*(3.7830; 1.0678; 18.0; 0.11)	*Sb*(3.4814; 1.2375; 7.810; 0.1)	*Sb*(3.290; 1.129; 5.837; 0.099)
	Logistic	*Sb*(3.516;1.054; 14.748; 0.117)	*Sb*(5.1316; 1.5681; 10.0; 0.065)	*Sb*(3.409; 1.434; 2.448; 0.095)
	Extreme-value & Weibull	*Sb*(3.512;1.064;14.496;0.125) [1)]	*Sb*(4.799; 1.402; 13.0; 0.085) [2)]	*Sb*(3.4830; 1.5138; 3.00; 0.07)

Note. [1)] - we estimated the Weibull distribution form parameter, [2)] - the Weibull distribution scale parameter.

Table. Models of limiting statistic distributions of the nonparametric goodness-of-fit when MLE are used in the case of gamma-distribution.

Test	Value of the form parameter	Estimation of scale parameter	Estimation of form parameter	Estimation of two parameters
Kolmogorov's	**0.3**	B_3(6.3045; 5.9555; 3.0350; 1.3170; 0.281)	B_3(6.4536; 5.7519; 3.3099; 1.6503; 0.280)	B_3(6.9705; 5.6777; 3.6297; 1.5070; 0.270)
	0.5	B_3(6.9356; 5.0081; 4.3582; 1.8470; 0.280)	B_3(6.3860; 5.9685; 3.1228; 1.6154; 0.280)	B_3(6.4083; 5.9339; 3.2063; 1.4483; 0.2774)
	1.0	B_3(6.7187; 5.3740; 3.7755; 1.6875; 0.282)	B_3(6.1176; 6.4704; 2.6933; 1.5501; 0.280)	B_3(5.6031; 6.1293; 2.7065; 1.3607; 0.2903)
	2.0	B_3(5.8359; 22.6032; 2.1921; 4.00; 0.282)	B_3(6.1387; 6.5644; 2.6021; 1.4840; 0.280)	B_3(5.8324; 6.1446; 2.7546; 1.3280; 0.2862)
	3.0	B_3(5.9055; 24.4312; 2.0996; 4.00; 0.282)	B_3(6.1221; 6.6131; 2.5536; 1.4590; 0.280)	B_3(6.0393; 6.1276; 2.8312; 1.3203; 0.2827)
	4.0	B_3(5.9419; 27.1264; 1.9151; 4.00; 0.282)	B_3(6.0827; 6.7095; 2.4956; 1.4494; 0.280)	B_3(6.1584; 6.1187; 2.8748; 1.3170; 0.2807)
	5.0	B_3(5.8774; 30.0692; 1.7199; 4.00; 0.282)	B_3(6.0887; 6.7265; 2.4894; 1.4432; 0.280)	B_3(6.1957; 6.1114; 2.8894; 1.3140; 0.2801)
Cramer-Mises-Smirnov's	**0.3**	B_3(3.2722; 1.9595; 16.1768; 0.750; 0.013)	B_3(3.0247; 3.2256; 11.113; 0.7755; 0.0125)	B_3(2.3607; 4.0840; 7.0606; 0.6189; 0.0145)
	0.5	B_3(3.2296; 2.1984; 14.3153; 0.700; 0.013)	B_3(3.0143; 3.3504; 10.095; 0.7214; 0.0125)	B_3(2.7216; 3.9844; 7.4993; 0.5372; 0.013)
	1.0	B_3(3.1201; 2.5460; 11.1200; 0.600; 0.013)	B_3(2.9928; 3.4716; 8.8275; 0.6346; 0.0125)	B_3(3.0000; 3.8959; 7.3247; 0.4508; 0.012)
	2.0	B_3(2.9463; 3.1124; 9.1160; 0.600; 0.013)	B_3(2.9909; 3.5333; 8.2010; 0.5786; 0.0125)	B_3(3.0533; 3.9402; 7.1173; 0.4246; 0.0118)
	3.0	B_3(2.8840; 3.3796; 8.4342; 0.600; 0.013)	B_3(2.9737; 3.5528; 7.8843; 0.5549; 0.0125)	B_3(3.0703; 3.9618; 7.034; 0.4163; 0.0117)
	4.0	B_3(2.8522; 3.5285; 8.1044; 0.600; 0.013)	B_3(2.9677; 3.5426; 7.7632; 0.5418; 0.0125)	B_3(3.0967; 3.9539; 7.064; 0.4122; 0.0116)
	5.0	B_3(2.8249; 3.6280; 7.8756; 0.6000; 0.013)	B_3(2.9638; 3.5465; 7.6558; 0.5334; 0.0125)	B_3(4.4332; 3.6256; 10.552; 0.4098; 0.0084)
Anderson- Darling's	**0.3**	B_3(3.3848; 2.8829; 14.684; 6.0416; 0.1088)	B_3(3.1073; 3.7039; 8.6717; 4.3439; 0.1120)	B_3(4.5322; 4.060; 10.0718; 2.9212; 0.078)
	0.5	B_3(5.0045; 2.9358; 18.8524; 5.2436; 0.077)	B_3(3.1104; 3.7292; 8.0678; 4.0132; 0.1120)	B_3(5.0079; 4.056; 10.0292; 2.5872; 0.073)
	1.0	B_3(5.0314; 3.1848; 15.4626; 4.3804; 0.077)	B_3(3.1149; 3.7919; 7.4813; 3.6770; 0.1120)	B_3(5.0034; 4.1093; 9.1610; 2.3427; 0.073)
	2.0	B_3(4.9479; 3.3747; 13.0426; 3.8304; 0.077)	B_3(3.0434; 4.1620; 7.1516; 3.8500; 0.1120)	B_3(4.9237; 4.2091; 8.6643; 2.2754; 0.073)
	3.0	B_3(5.0367; 3.4129; 12.9013; 3.6867; 0.077)	B_3(3.0565; 3.9092; 6.7844; 3.3972; 0.1120)	B_3(4.9475; 4.2070; 8.6686; 2.2512; 0.073)
	4.0	B_3(4.9432; 3.5038; 12.2240; 3.6302; 0.077)	B_3(3.0531; 3.9437; 6.7619; 3.3993; 0.1120)	B_3(4.9274; 4.2279; 8.5573; 2.2390; 0.073)
	5.0	B_3(4.8810; 3.5762; 11.7894; 3.6051; 0.077)	B_3(3.0502; 3.9640; 6.7510; 3.4024; 0.1120)	B_3(4.9207; 4.2432; 8.4881; 2.2314; 0.073)

Сходимость распределения статистики Колмогорова к предельному при проверке сложных гипотез



При проверке сложной гипотезы о согласии с нормальным законом при вычислении ОМП 2-х параметров

Выводы

1. **Надо быть очень внимательным к факту, какая гипотеза проверяется, простая или сложная.**
2. **Если сложная и гипотеза проверяется по той же выборке, по которой проверяется согласие, то нельзя использовать классические результаты (предельные распределения и процентные точки).**
3. **Следует помнить, что предельные распределения статистик непараметрических критериев согласия зависят от ряда факторов (типа оцениваемого параметра, количества оцениваемых параметров, возможно, от конкретных значений параметров формы, от метода оценивания параметров).**
4. **Если модель предельного распределения статистики критерия для конкретной ситуации неизвестна, всегда можно воспользоваться технологией компьютерного моделирования и построить модель распределения статистики для этой ситуации.**
5. **Остаются вопросы, связанные с мощностью критериев, предназначенных для проверки одних и тех же гипотез (?).**

Анализ мощности критериев согласия при близких альтернативах

Отдавая при проведении статистического анализа данных предпочтение некоторому критерию, экспериментатор хотел бы иметь уверенность в том, что для заданной вероятности ошибки первого рода α гарантируется минимальная вероятность ошибки 2-го рода β. Другими словами, хотелось бы отдать предпочтение критерию, наиболее мощному относительно интересующей нас пары альтернатив H_0 и H_1.

Информация, содержащаяся в различных источниках, о преимуществах в определенных ситуациях того или иного критерия согласия неоднозначна и зачастую противоречива. Результаты исследования асимптотической мощности критериев, например [1-4], трудно использовать вследствие ограниченных объемов выборок, с которыми приходится иметь дело практику. Рекомендации различных авторов носят субъективный характер, отражают сложившиеся стереотипы, базируются на конкретных частных примерах и ограниченном опыте практического применения.

Исследования мощности затруднены отсутствием результатов, связанных с аналитическим представлением функций распределения $G(S|H_1)$ для конкретных критериев согласия при проверке сложных гипотез, в частности, для непараметрических критериев и для критериев типа χ^2 при оценивании параметров по точечным выборкам (по негруппированным наблюдениям).

Цель исследований, результаты которых приводятся ниже, заключалась в сравнительном анализе мощности наиболее часто используемых критериев согласия на некоторых парах достаточно близких конкурирующих гипотез H_0 и H_1. Интерес представляет способность критериев различать именно близкие гипотезы, так как распознавание отличия в далеких законах распределения, как правило, не составляет проблем.

Рассматриваемые альтернативы. **Результаты сравнительного анализа мощности критериев согласия в работе иллюстрируются на двух парах альтернатив. Первую пару составили нормальный и логистический законы: проверяемой гипотезе H_0 соответствовал нормальный закон с плотностью**

$$f(x)=\frac{1}{\theta_0\sqrt{2\pi}}\exp\left\{-\frac{(x-\theta_1)^2}{2\theta_0^2}\right\},$$

а конкурирующей гипотезе H_1 – логистический с функцией плотности

$$f(x)=\frac{\pi}{\theta_0\sqrt{3}}\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\Bigg/\left[1+\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\right]^2$$

и параметрами $\theta_0=1$, $\theta_1=0$. В случае простой гипотезы H_0 параметры нормального закона имеют те же значения. Эти два закона близки и трудно различимы с помощью критериев согласия.

Вторую пару составили: H_0 – распределение Вейбулла с плотностью

$$f(x)=\frac{\theta_0(x-\theta_2)^{\theta_0-1}}{\theta_1^{\theta_0}}\exp\left\{-\left(\frac{x-\theta_2}{\theta_1}\right)^{\theta_0}\right\}$$

и параметрами $\theta_0=2$, $\theta_1=2$, $\theta_2=0$; H_1 – гамма-распределение с плотностью

$$f(x)=\frac{1}{\theta_1\Gamma(\theta_0)}\left(\frac{x-\theta_2}{\theta_1}\right)^{\theta_0-1}e^{-(x-\theta_2)/\theta_1}$$

и параметрами $\theta_0=3.12154$, $\theta_1=0.557706$, $\theta_2=0$, при которых гамма-распределение наиболее близко к данному распределению Вейбулла.

В работе исследовалась мощность при проверке простых и сложных гипотез H_0 против простой альтернативы H_1.

Мощность критериев в случае пары альтернатив “нормальный-логистический”

О близости распределений нормального и логистического, соответствующих конкурирующим гипотезам H_0 и H_1, свидетельствует рис. 1.

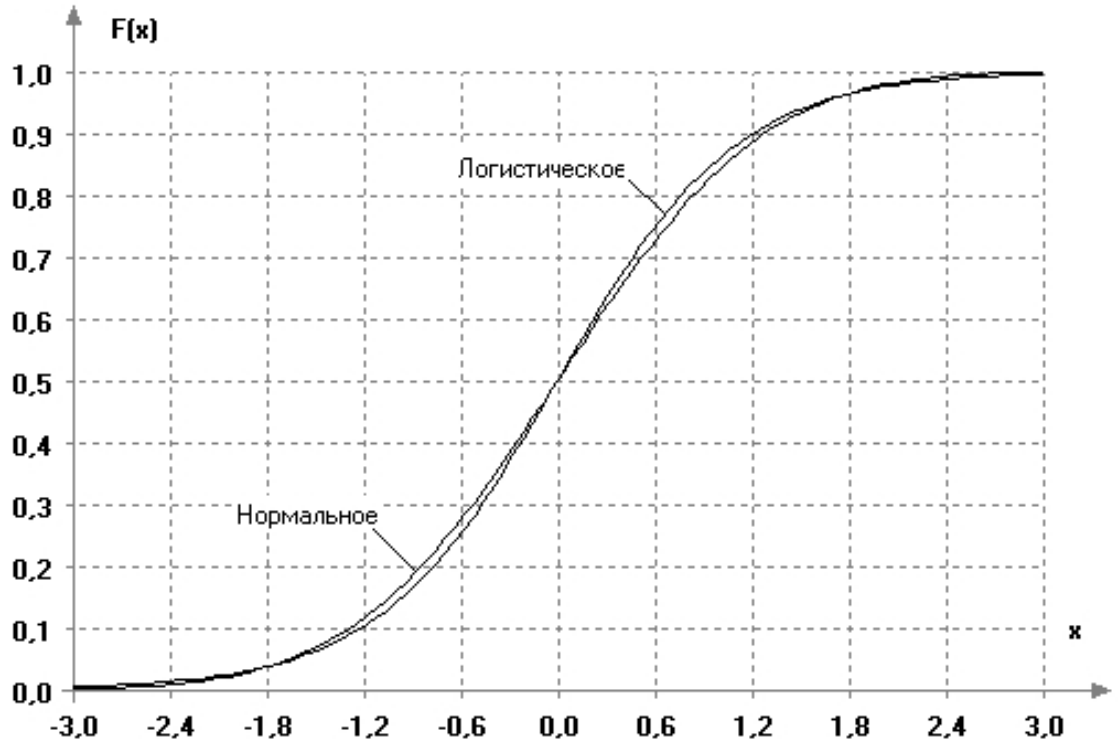


Рис. 1. Нормальное и логистическое распределения, соответствующие H_0 и H_1

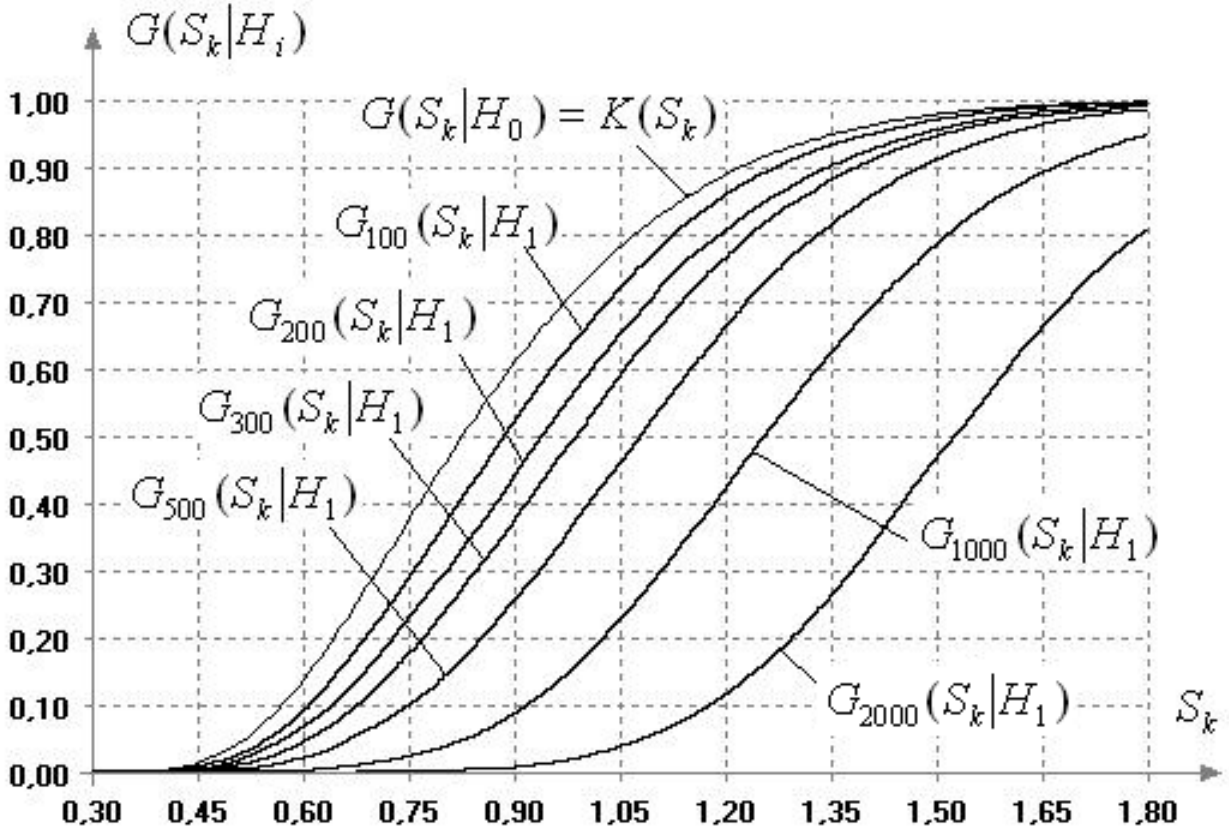


Рис. 2. Распределения статистики (1) типа Колмогорова $G(S_k|H_0) = K(S_k)$ и $G_n(S_k|H_1)$ при проверке простой гипотезы H_0 о согласии с нормальным законом при альтернативе H_1

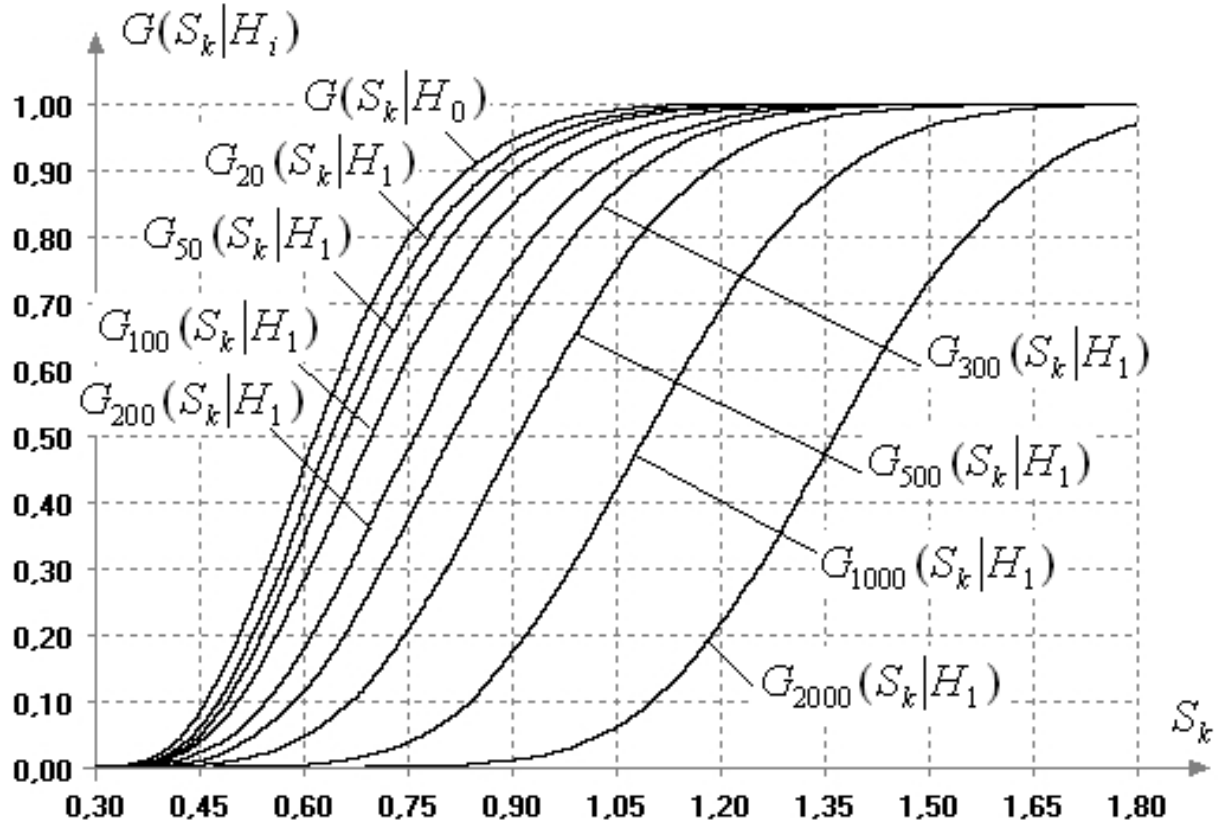


Рис.3. Распределения статистики (1) типа Колмогорова $G(S_k|H_0)$ и $G_n(S_k|H_1)$ при проверке сложной гипотезы H_0 о согласии с нормальным законом в случае использования ОМП при альтернативе H_1

Мощность непараметрических критериев сильно зависит от используемого метода оценивания

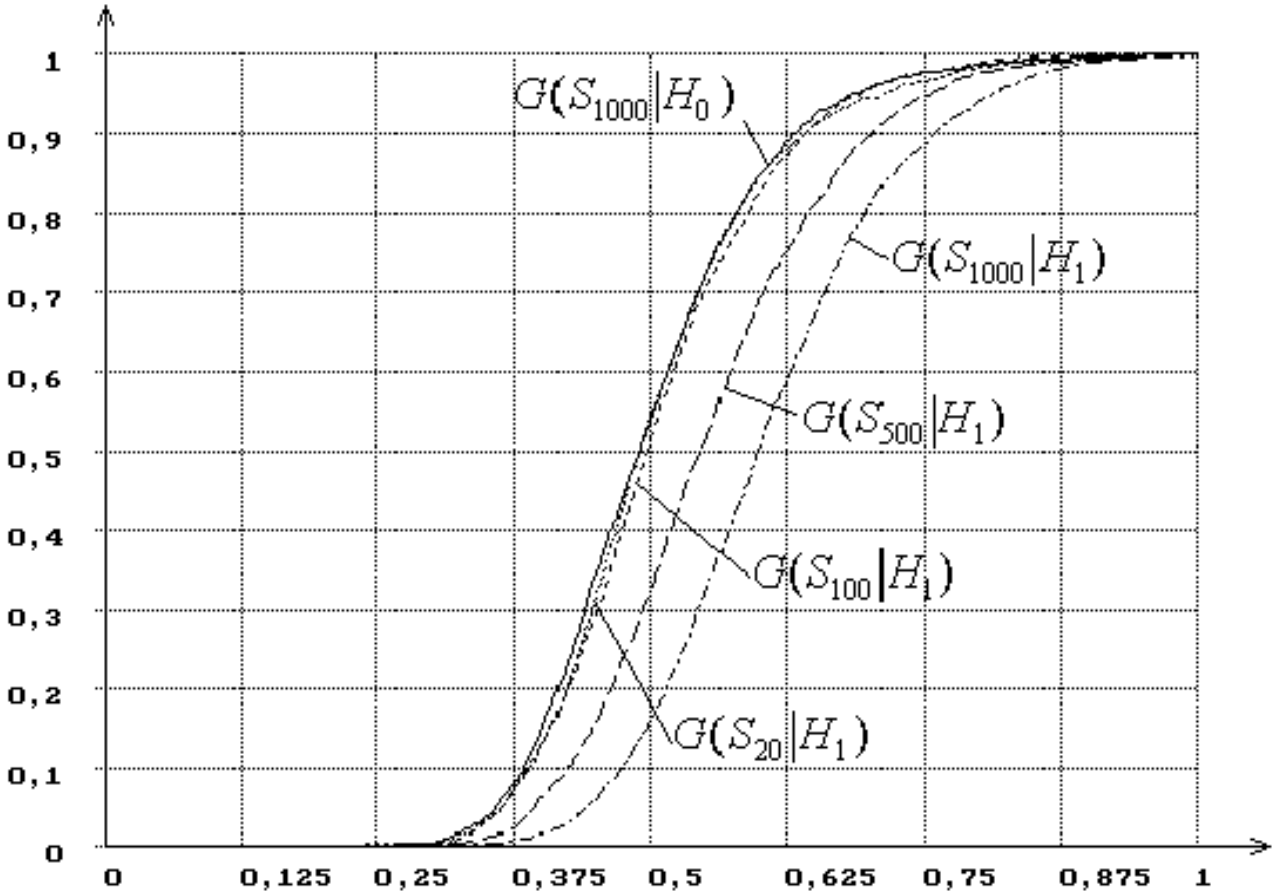


Рис. 3а. Распределения статистики (1) типа Колмогорова $G(S_k|H_0)$ и $G_n(S_k|H_1)$ при проверке сложной гипотезы H_0 о согласии с нормальным законом в случае использования *MD*-оценок, минимизирующих статистику (1), при альтернативе H_1 (логистическое распределение)

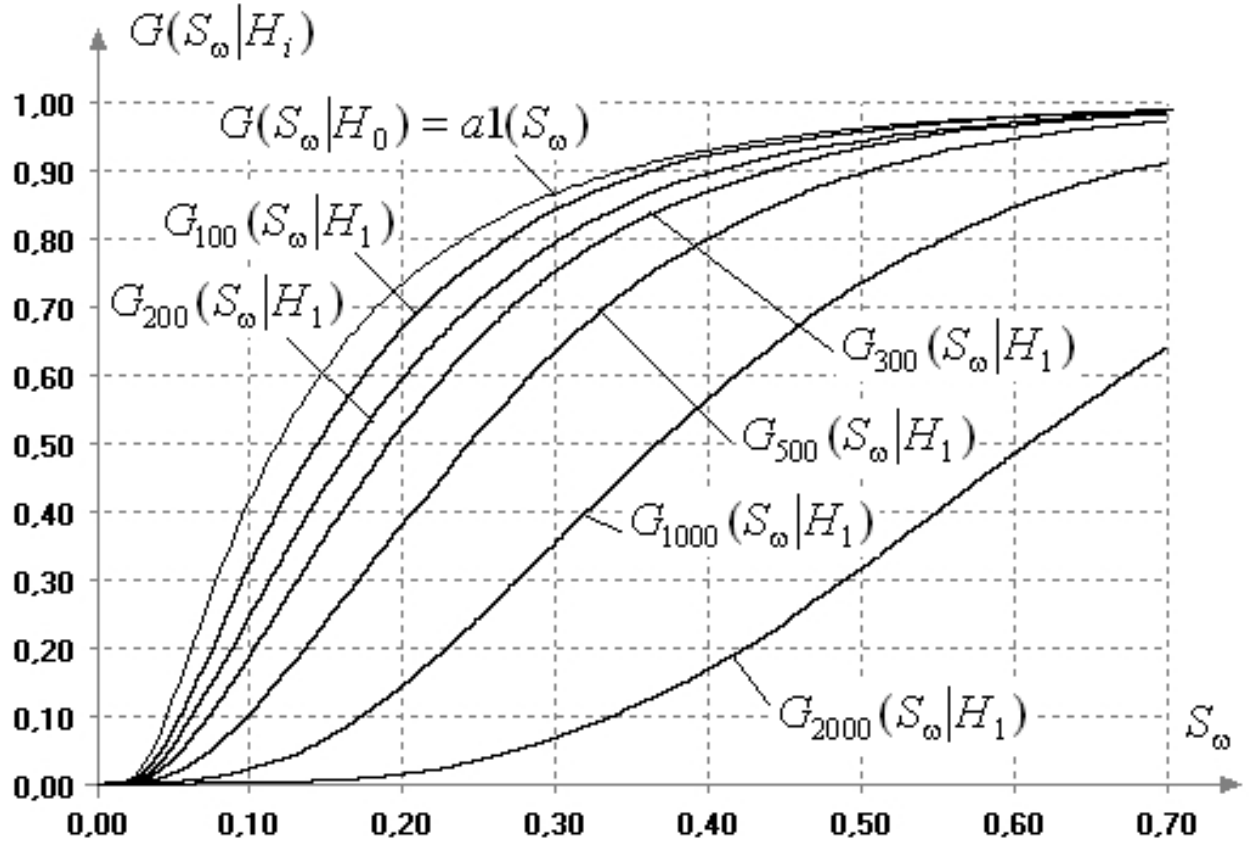


Рис. 4. Распределения статистики (2) типа ω^2 Крамера-Мизеса-Смирнова $G(S_\omega|H_0)=a1(S_\omega)$ и $G_n(S_\omega|H_1)$ при проверке простой гипотезы H_0 о согласии с нормальным законом при альтернативе H_1

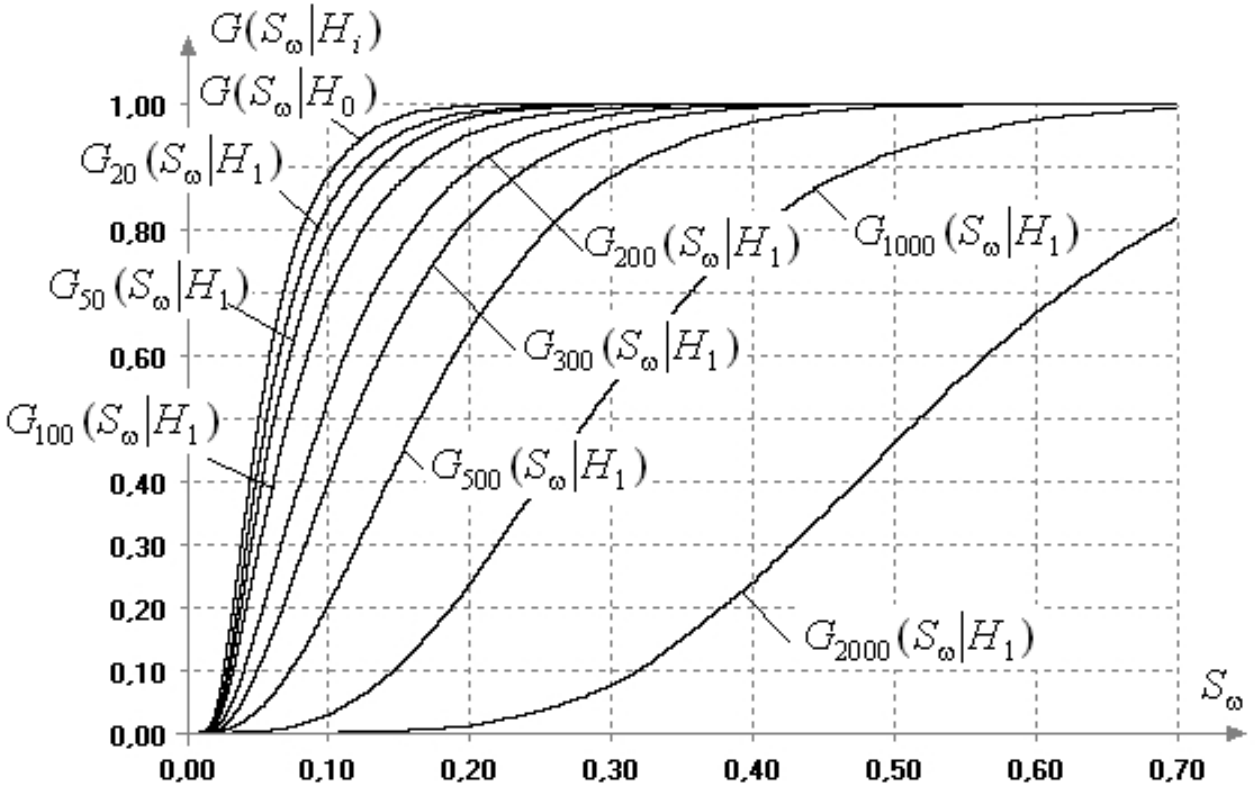
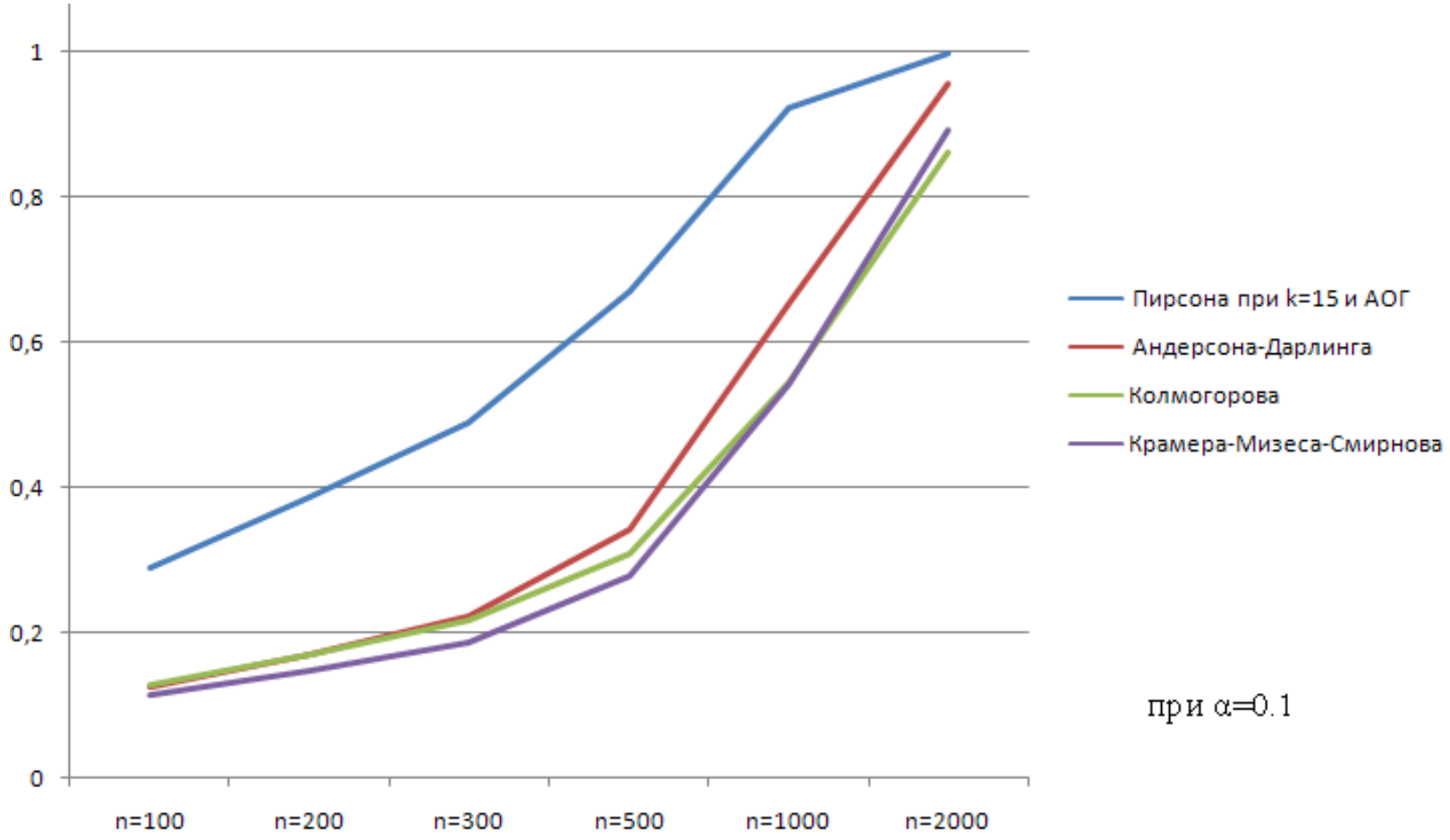
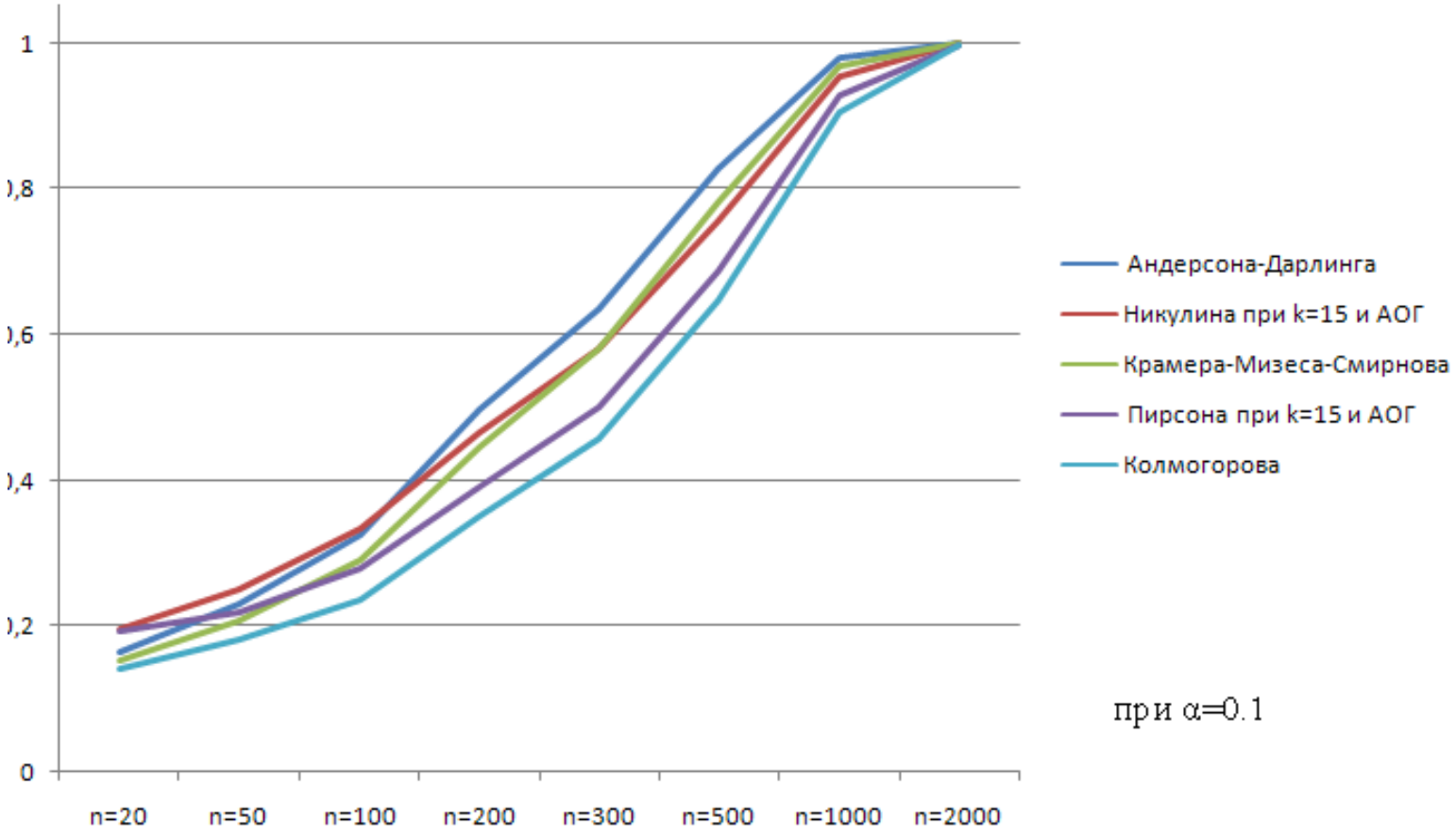


Рис. 5. Распределения статистики (2) типа ω^2 Крамера-Мизеса-Смирнова $G(S_\omega|H_0)$ и $G_n(S_\omega|H_1)$ при проверке сложной гипотезы H_0 о согласии с нормальным законом в случае использования ОМП при альтернативе H_1



Мощность критериев согласия при проверке **простой гипотезы** H_0 (нормальное распределение) против конкурирующей гипотезы H_1 (логистическое) в зависимости от объема выборки



Мощность критериев согласия при проверке **сложной гипотезы** H_0 (нормальное распределение) против конкурирующей гипотезы H_1 (логистическое) в зависимости от объема выборки при α=0.1

Выводы по мощности критеиев согласия

Для случая проверки *простых гипотез* можно упорядочить критерии по мощности следующим образом:

X_n^2 **Пирсона (АОГ)** $\succ$ Ω^2 **Андерсона-Дарлинга** $\succ$ ω^2 **Мизеса** $\succ$ **=Колмогорова**

Такая шкала справедлива при использовании в критерии χ^2 **Пирсона АОГ, при котором минимизируются потери в информации Фишера. При очень близких гипотезах может быть:**

Колмогорова $\succ$ ω^2 **Мизеса.**

При проверке *сложных гипотез* градация по мощности оказывается существенно иной:

Ω^2 **Андерсона-Дарлинга** $\succ$ ω^2 **Мизеса** $\succ$ Y_n^2 **(АОГ)** $\succ$ X_n^2 **Пирсона (АОГ)** $\succ$ **Колмогорова.**

При очень близких гипотезах может быть:

Ω^2 **Андерсона-Дарлинга** $\succ$ Y_n^2 **(АОГ)** $\succ$ ω^2 **Мизеса** $\succ$ X_n^2 **Пирсона (АОГ)** $\succ$ **Колмогорова.**

Указанные выводы носят интегрированный характер. Такое упорядочение не является жёстким. Как видно из таблиц с приведенными значениями мощности, иногда критерий имеет преимущества по мощности при одних значениях α **и объемах выборок** n **и уступает при других значениях** α **и** n**.**

Надо иметь в ввиду, что мощность критериев типа χ^2 **(Пирсона и Никулина) зависит не только от гипотез** H_0**,** H_1 **и объема выборок** n**, но при заданных** H_0 **и** H_1 **– от способа группирования и числа интервалов.**

В случае, если нас интересует способность критериев различать конкретную пару конкурирующих гипотез (конкретную пару законов), мощность критериев типа χ^2 можно максимизировать, подобрав оптимальное число интервалов и оптимальное расположение граничных точек интервалов.

Воздействовать каким-то образом на мощность непараметрических критериев согласия возможности нет.

На следующих рисунках иллюстрируются результаты сравнительного анализа мощности критериев согласия относительно 2-х конкурирующих гипотез: проверяемой гипотезе H_0 соответствует нормальный закон с плотностью

$$f(x) = \frac{1}{\theta_0\sqrt{2\pi}}\exp\left\{-\frac{(x-\theta_1)^2}{2\theta_0^2}\right\},$$

а конкурирующей гипотезе H_1 – логистическое распределение с плотностью

$$f(x) = \frac{\pi}{\theta_0\sqrt{3}}\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\Bigg/\left[1+\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\right]^2$$

с параметрами в том и другом случае $\theta_0 = 1,\ \theta_1 = 0$.

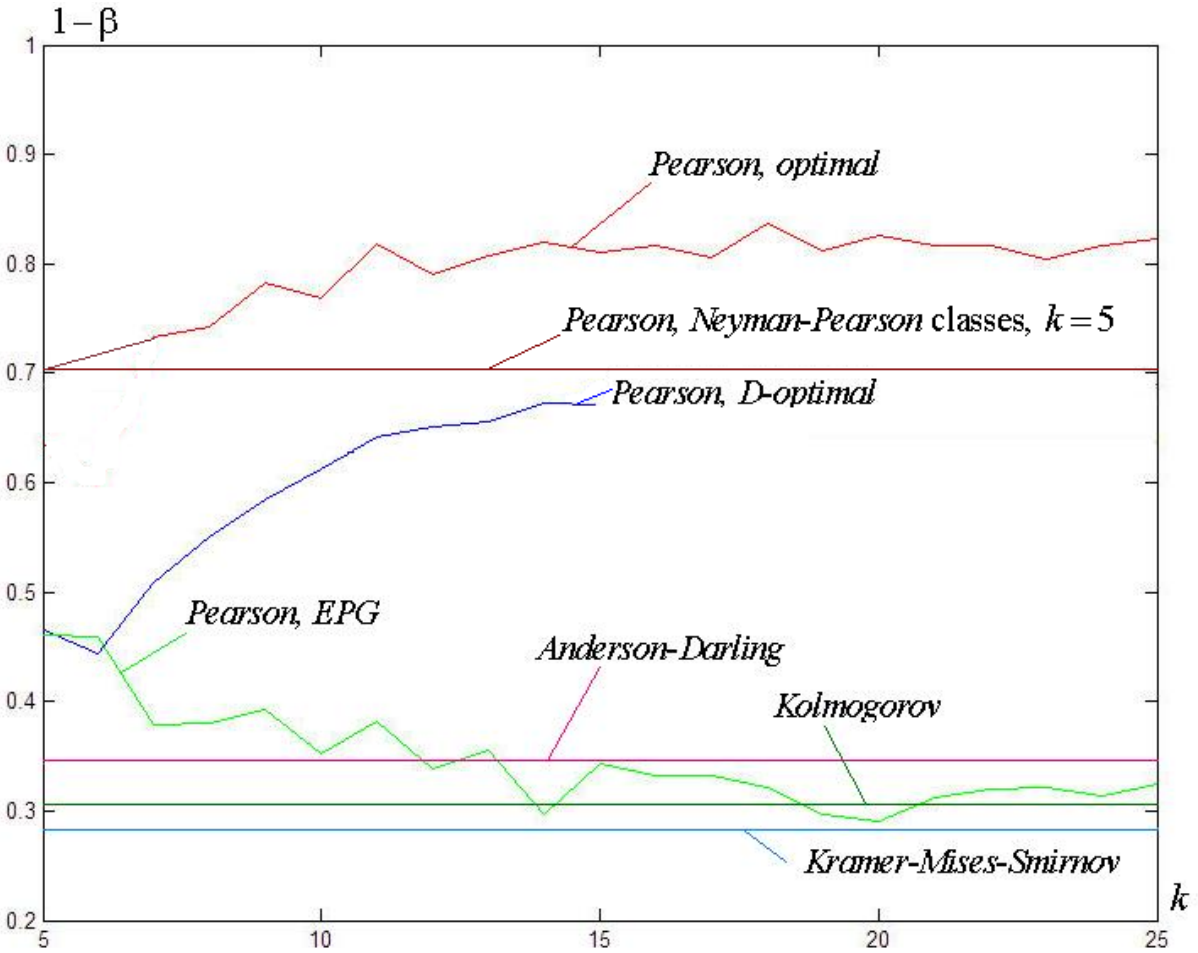


Fig.36. Мощность критериев согласия при проверке **простой** гипотезы H_0 (normal distribution) против конкурирующей H_1 (logistic distribution), $\alpha = 0.1$, $n = 500$. Мощность критерия χ^2 Пирсона зависит от k.

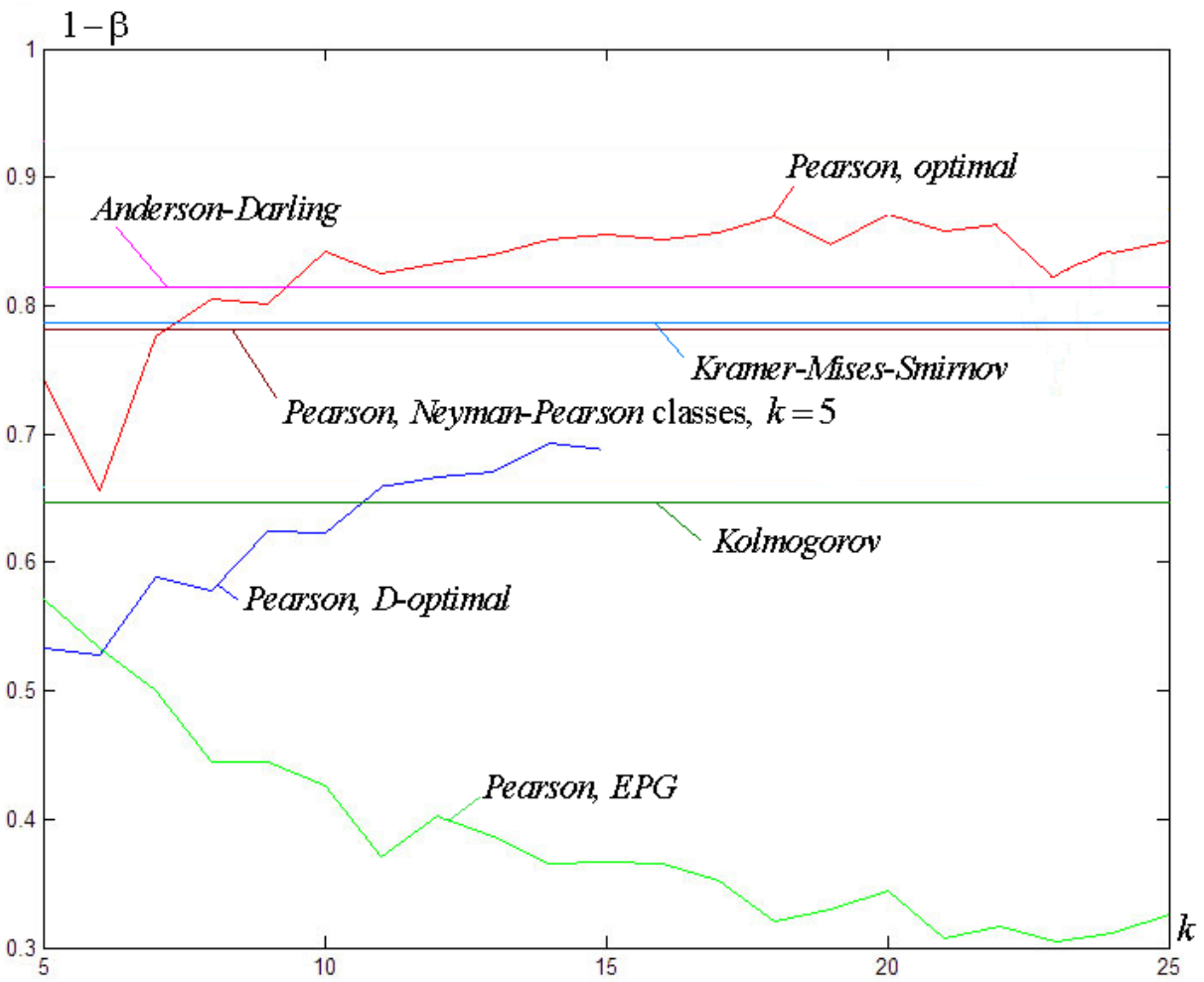
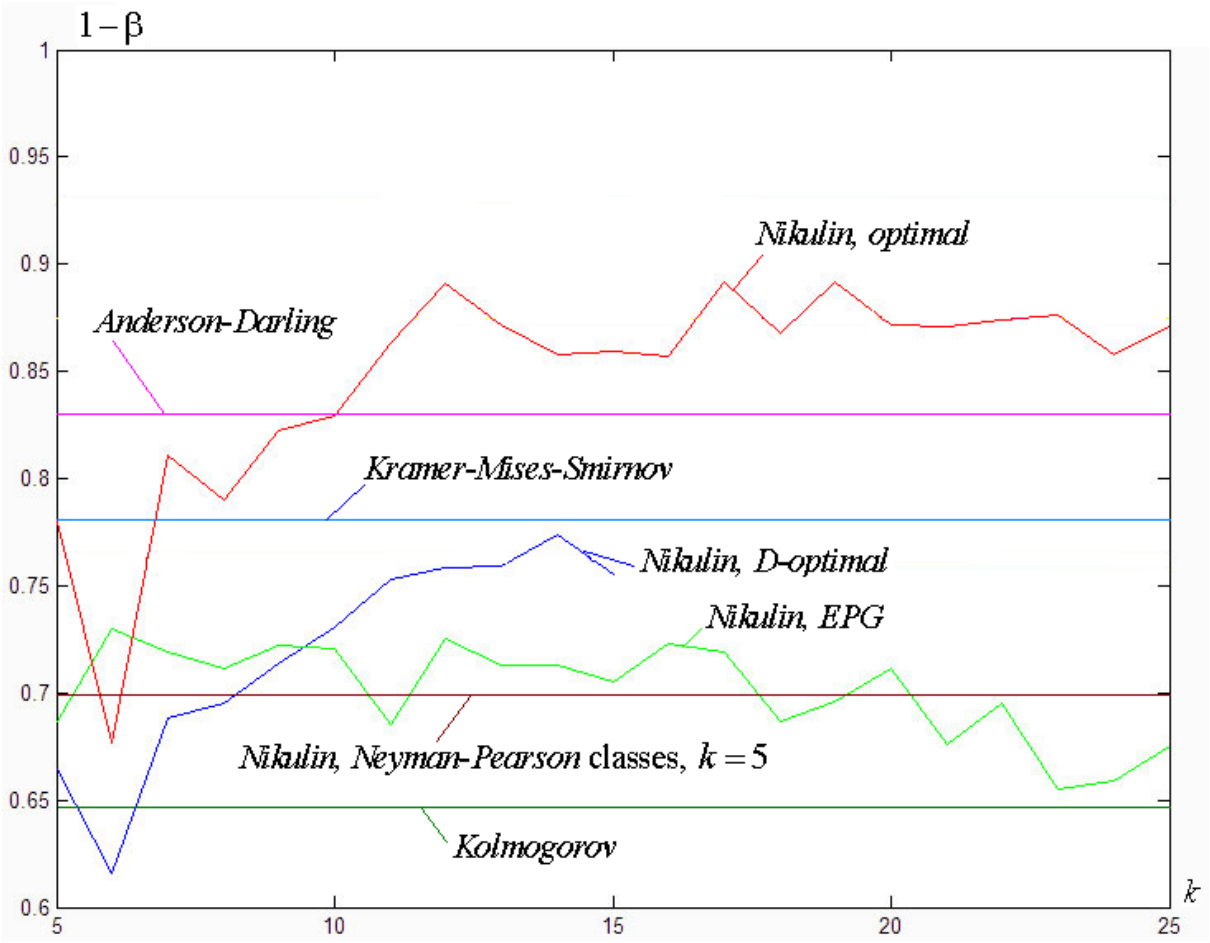


Fig. 37. Мощность критериев согласия при проверке **сложной** гипотезы H_0 (normal distribution) против конкурирующей H_1 (logistic distribution), $\alpha = 0.1$, $n = 500$. Мощность критерия χ^2 Пирсона зависит от k.



Мощность критериев согласия при проверке **сложной** гипотезы H_0 (normal distribution) против конкурирующей H_1 (logistic distribution), $\alpha = 0.1$, $n = 500$. Мощность критерия χ^2 Никулина зависит от k.

Критерий Купера

В работе [20] Купером предложен критерий типа Колмогорова, статистика V_n которого определяется соотношением

$$V_n = \sup_{-\infty<x<\infty} \left\{F_n(x) - F(x,\theta)\right\} - \inf_{-\infty<x<\infty} \left\{F_n(x) - F(x,\theta)\right\}$$

и используется в виде

$$V_n = D_n^+ + D_n^-, \tag{2.18}$$

где D_n^+, D_n^- определяются соотношениями (2.5), (2.6). $i = \overline{1,n}$, n – объем выборки, x_i – здесь и далее элементы вариационного ряда, построенного по выборке (упорядоченная по возрастанию выборка).

Существенным недостатком критерия со статистикой (2.18) является сильная зависимость распределения $G(V_n|H_0)$ статистики от объема выборки n. Таблицы процентных точек для случая проверки простых гипотез по критерию со статистикой $\sqrt{n}V_n$ можно найти в работах [4, 54]. Купером [20] в качестве предельного распределения $G\left(\sqrt{n}V_n|H_0\right)$ статистики $\sqrt{n}V_n$ дана следующая функция распределения [54]:

$$G\left(s|H_0\right) = 1 - \sum_{m=1}^{\infty} 2(4m^2s^2 - 1)e^{-2m^2s^2}. \tag{2.19}$$

В [52] для модификации статистики

$$V = V_n\left(\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}}\right), \tag{2.20}$$

распределение которой в меньшей степени чем распределение $\sqrt{n}V_n$ зависит от n, приведены процентные точки, которые представлены во 2-й строке таблицы 1. Зависимостью распределения статистики (23) от объема выборки можно пренебречь при $n \geq 20$, так как отклонение реального распределения статистики от предельного незначительно и практически не влияет на результаты статистического вывода.

В [80] предложено применять в критерии Купера статистику в следующем виде

$$V_n^{mod} = \sqrt{n}(D_n^+ + D_n^-) + \frac{1}{3\sqrt{n}}, \tag{2.21}$$

где идея использования поправки вытекает из выражения для статистики критерия согласия Смирнова [65, с. 81]. Зависимостью распределения статистики (24) от объема выборки можно практически пренебречь при $n \geq 30$.

Критерий Ватсона

Статистика критерия Ватсона [56, 57] имеет вид

$$U_n^2 = n\int_{-\infty}^{\infty}\left\{F_n(x) - F(x,\theta) - \int_{-\infty}^{\infty}\left(F_n(y) - F(y,\theta)\right)dF(y,\theta)\right\}^2 dF(x,\theta)$$

и используется в следующей удобной для расчетов форме

$$U_n^2 = \sum_{i=1}^{n}\left(F(x_i,\theta) - \frac{i-\frac{1}{2}}{n}\right)^2 - n\left(\frac{1}{n}\sum_{i=1}^{n}F(x_i,\theta) - \frac{1}{2}\right)^2 + \frac{1}{12n}. \qquad (2.23)$$

Процентные точки статистики U_n^2 при проверке простой гипотезы можно найти в [57, 48]. Предельное распределение $G\left(U_n^2 \middle| H_0\right)$ статистики U_n^2 приведено в [56, 57] в виде

$$G\left(s \middle| H_0\right) = 1 - 2\sum_{m=1}^{\infty}(-1)^{m-1}e^{-2m^2\pi^2 s}. \qquad (2.24)$$

Критерии Жанга

В диссертации Жанга [58] и в последующих работах [59, 60, 61] были предложены непараметрические критерии согласия, статистики которых имеют вид:

$$Z_K = \max_{1\le i\le n}\left(\left(i-\frac{1}{2}\right)\log\left\{\frac{i-\frac{1}{2}}{nF(x_i,\theta)}\right\}+\left(n-i+\frac{1}{2}\right)\log\left[\frac{n-i+\frac{1}{2}}{n\{1-F(x_i,\theta)\}}\right]\right), \tag{2.27}$$

$$Z_A = -\sum_{i=1}^{n}\left[\frac{\log\{F(x_i,\theta)\}}{n-i+\frac{1}{2}}+\frac{\log\{1-F(x_i,\theta)\}}{i-\frac{1}{2}}\right], \tag{2.28}$$

$$Z_C = \sum_{i=1}^{n}\left[\log\left\{\frac{[F(x_i,\theta)]^{-1}-1}{(n-\frac{1}{2})/(i-\frac{3}{4})-1}\right\}\right]^2 \tag{2.29}$$

Справедливость утверждений автора о более высокой мощности предлагаемых критериев по сравнению с критериями Колмогорова, Крамера-Мизеса-Смирнова и Андерсона-Дарлинга была подтверждена проведенными исследованиями [69, 80, 21, 104].

Однако использование критериев со статистиками (2.27) – (2.29) осложняет сильная зависимость распределений статистик от объема выборки n. Естественно, зависимость от n сохраняется и в случае проверки сложных гипотез.

При проверке простых гипотез можно воспользоваться таблицами процентных точек, приводимых автором, что, в принципе, не очень удобно, так как не позволяет оценить достигаемый уровень значимости. При проверке сложных гипотез применение данных критериев связано с дополнительными препятствиями. Обойти эти препятствия можно за счет использования **интерактивного режима** исследования распределений статистик применяемых критериев.

Результаты анализа мощности позволяют упорядочить рассматриваемые непараметрические критерии Колмогорова (K), Крамера-Мизеса-Смирнова (KMS), Андерсона-Дарлинга (AD), Купера (V_n), Ватсона (U_n^2), Жанга (Z_C , Z_A и Z_K) по мощности следующим образом:

– при проверке простых гипотез –

$$Z_C \succ Z_A \succ Z_K \succ U_n^2 \succ V_n \succ AD \succ KMS \succ \approx K \ ;$$

– при проверке сложных гипотез –

$$Z_A \succ \approx Z_C \succ Z_K \approx AD \succ KMS \succ U_n^2 \succ V_n \succ K \ .$$

Мощность непараметрических критериев при проверке сложных гипотез при тех же объемах выборок n всегда существенно выше, чем при проверке простых.

Применение критериев согласия для анализа больших выборок

Вопросы применения статистических методов к анализу больших массивов данных (Big Data) в последние годы вызывают большой интерес. В приложениях всё чаще приходится сталкиваться с необходимостью анализа гигантских объёмов накапливаемых данных. Возникают потребности извлечения и использования закономерностей, в том числе вероятностных, скрытых в этих данных.

При попытках применения для анализа больших данных классического аппарата прикладной математической статистики, как правило, встречаются со специфическими проблемами, ограничивающими возможности их корректного применения. Например, сталкиваются с тем, что хорошо зарекомендовавшие себя методы и алгоритмы становятся неэффективными из-за “проклятия размерности”. Одни популярные критерии проверки гипотез оказываются не приспособленными для анализа выборок даже порядка тысячи наблюдений. Другие, которые формально можно использовать при объёмах выборок $n \to \infty$, всегда приводят к отклонению даже справедливой проверяемой гипотезы H_0. Такие проблемы характерны для многих критериев, в том числе для непараметрических критериев согласия. И связаны они не только с ростом вычислительных затрат.

То, что очень часто информация о законе распределения $G(S|H_0)$ статистики критерия ограничена лишь узкими рамками таблицы критических значений, совсем не ограничивает возможность корректного применения критерия при объемах выборок за рамками этой таблицы. Для этого достаточно лишь воспользоваться интерактивным режимом для моделирования и последующего использования $G_N(S_n|H_0)$.

Основная причина, препятствующая корректному применению множества классических критериев проверки статистических гипотез, заключается в следующем. Как правило, объёмы выборок в Big Data (принадлежащие некоторому непрерывному закону распределения) практически неограничены, но сами данные представлены с ограниченной точностью (округлены с некоторым Δ). По сути, “нарушается предположение” о том, что наблюдается непрерывная случайная величина.

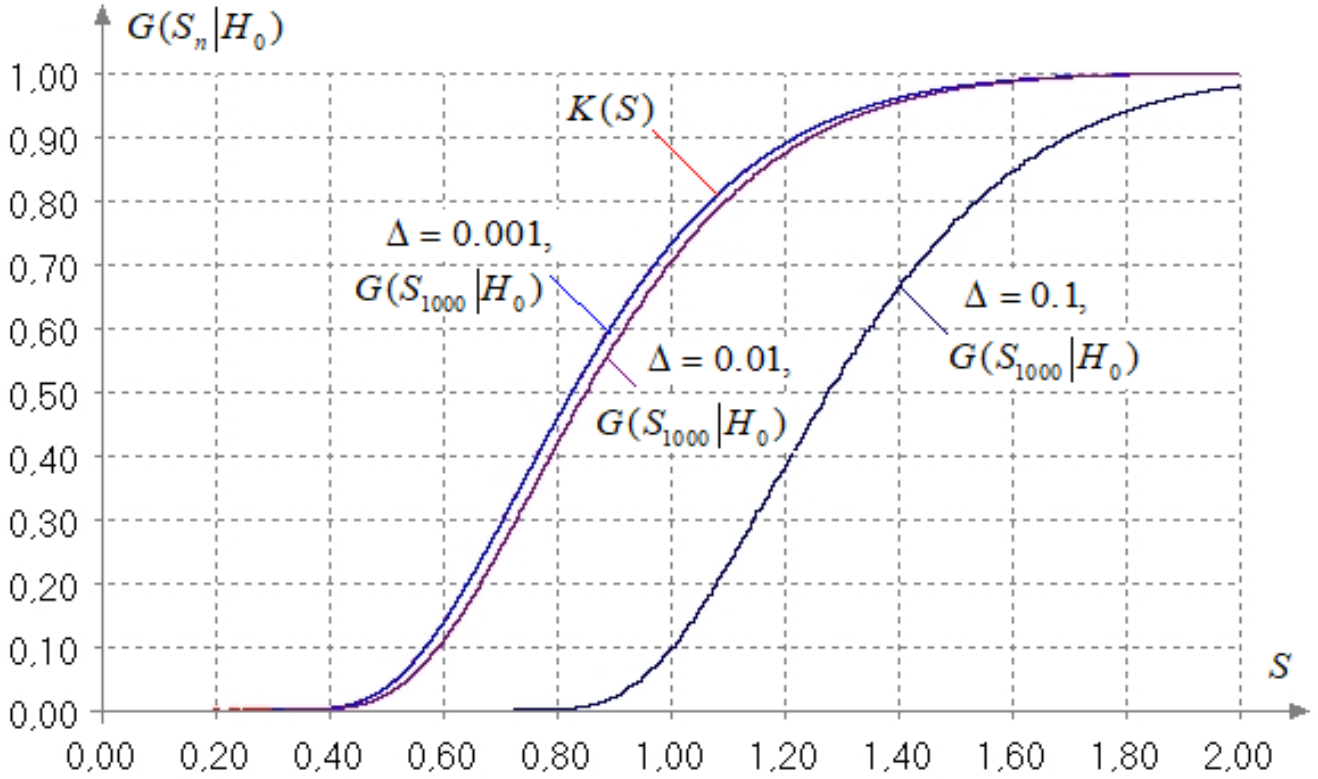
Допустим, для критерия существует предельное распределение статистики $G(S|H_0)$. Эмпирическое распределение $F_n(x)$, соответствующее выборке непрерывных случайных величин (без округления), при $n \to \infty$ сходится к функции распределения $F(x)$ этой случайной величины. Эмпирическое распределение $G_N(S_n|H_0)$ статистики, строящейся по выборке непрерывной случайной величины при $n \to \infty$ (и $N \to \infty$) сходится к предельному $G(S|H_0)$.

Пусть теперь наблюдаемые данные округляются с некоторым Δ. Тогда, начиная с некоторого n, зависящего от вида $F(x)$, от области определения случайной величины и от Δ, $\max|F_n(x) - F(x)|$ перестанет уменьшаться, а распределение $G_N(S_n|H_0)$ – станет с ростом n отклоняться от предельного $G(S|H_0)$ (чем больше Δ, тем при меньшем n).

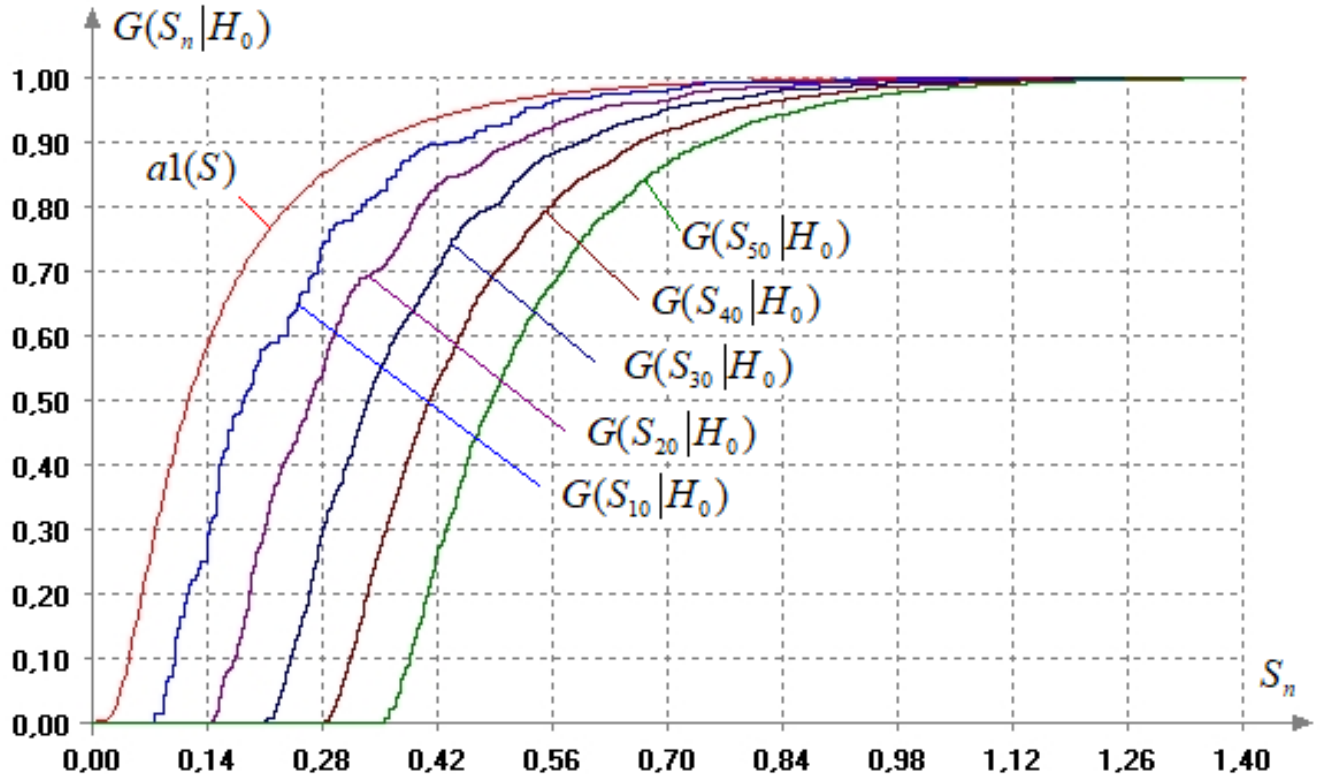
Описанная проблема имеет место не только в случае анализа выборок очень большого объёма. Она типична для многих задач, возникающих в приложениях, когда данные (измерения) фиксируются с существенной степенью округления, из-за чего в выборках оказывается относительно много повторяющихся наблюдений. В таких ситуациях реальные распределения $G(S_n|H_0)$ статистик критериев (при данной степени округления Δ) могут быть далёкими от предельных $G(S|H_0)$ распределений и существенно отличающимися от $G(S_n|H_0)$, имеющими место в ситуации без округления измерений.

Рассмотрим поведение распределений $G(S_n|H_0)$ статистик критериев на примере проверки согласия со стандартным нормальным законом.

При округлении с точностью до 1 в выборках, принадлежащих $\mathrm{N}(0,1)$, может появляться **9** уникальных значений, при округлении с точностью $\Delta = 0.1$ – порядка **86** уникальных значений, с точностью $\Delta = 0.01$ – порядка **956**, с точностью $\Delta = 0.001$ – порядка **9830**.



Эмпирические распределения $G(S_{1000}|H_0)$ статистики критерия Колмогорова при проверке согласия со стандартным нормальным законом в зависимости от Δ



Эмпирические распределения $G(S_n|H_0)$ статистики критерия Крамера–Мизеса–Смирнова при проверке согласия со стандартным нормальным законом в зависимости от n при $\Delta = 1$

При анализе очень больших выборок для того, чтобы при использовании соответствующих критериев для вычисления **p**-value можно было использовать предельное распределение статистики $G(S|H_0)$ (имеющее место при проверке простой или сложной гипотезы), рекомендуется применять критерий не ко всему массиву Big Data, а **извлекать для анализа из этого массива выборки ограниченного объема**. То есть, применять критерий при таких n, при которых для данной степени округления Δ реальное распределение статистики $G(S_n|H_0)$ ещё практически не отличается от $G(S|H_0)$.

Возможность корректного применения критерия в условиях ограниченных объёмов выборок и существенных округлений решается интерактивным моделированием распределения статистики $G_N(S_n|H_0)$ критерия при Δ и n, соответствующих условиям получения анализируемых данных, что и реализуется в ISW.

В ISW при моделировании псевдослучайных (*непрерывных*) величин используются числа двойной точности (с плавающей точкой), что обеспечивает представление данных с 15–17 значимыми десятичными цифрами в диапазоне примерно от 10^{-308} до 10^{308}. Такая точность при моделировании позволяет, с одной стороны, подтверждать имеющиеся теоретические и асимптотические закономерности, а с другой – является критерием точности программной реализации соответствующего критерия, когда подтверждается соответствие $G_N(S_n|H_0)$ известному теоретическому закону $G(S|H_0)$. Но возможно и моделировании псевдослучайных величин с заданным округлением Δ.

Спасибо за внимание!