Некоторые проблемы применения критериев проверки нормальности

Лемешко Борис Юрьевич

E-mail: Lemeshko@fpm.ami.nstu.ru

При проверке отклонения распределения от нормального закона *наряду с критериями согласия* используется целый ряд специальных критериев:

1. Критерий проверки на симметричность.
2. Критерий проверки на эксцесс.
3. Критерий **Шапиро-Уилка**.
4. Критерий **Эппса-Палли**.
5. **Модифицированный** критерий **Шапиро-Уилка**.
6. Совместный критерий проверки на симметричность и нулевой коэффициент эксцесса.
7. Модификация **D’Agostino** критерия проверки на симметричность.
8. Модификация D’Agostino критерия проверки на симметричность и эксцесс.
9. Совместный критерий проверки на симметричность и нулевой коэффициент эксцесса D’Agostino.
10. Критерий **Фросини**.
11. Критерий **Гири**.
12. Критерии **Хегази-Грина**.
13. Критерий **Шпигельхалтера**.
14. Критерий **Дэвида-Хартли-Пирсона**.

Рассмотрим способность различных критериев отличать от нормального закона некоторые конкурирующие законы

В качестве гипотезы H_0 рассмотрен нормальный закон распределения

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

с параметром масштаба $\theta_1 = 1$ и параметром сдвига $\theta_0 = 0$.

В качестве альтернатив были взяты следующие гипотезы: H_1 соответствует экспоненциальному семейству распределений с плотностью

$$f(x) = \frac{\theta_2}{2\theta_1\Gamma(1/\theta_2)}\exp\left\{-\left(\frac{|x-\theta_0|}{\theta_1}\right)^{\theta_2}\right\}$$

параметром формы $\theta_2 = 4$, H_2 – распределению Лапласа с плотностью

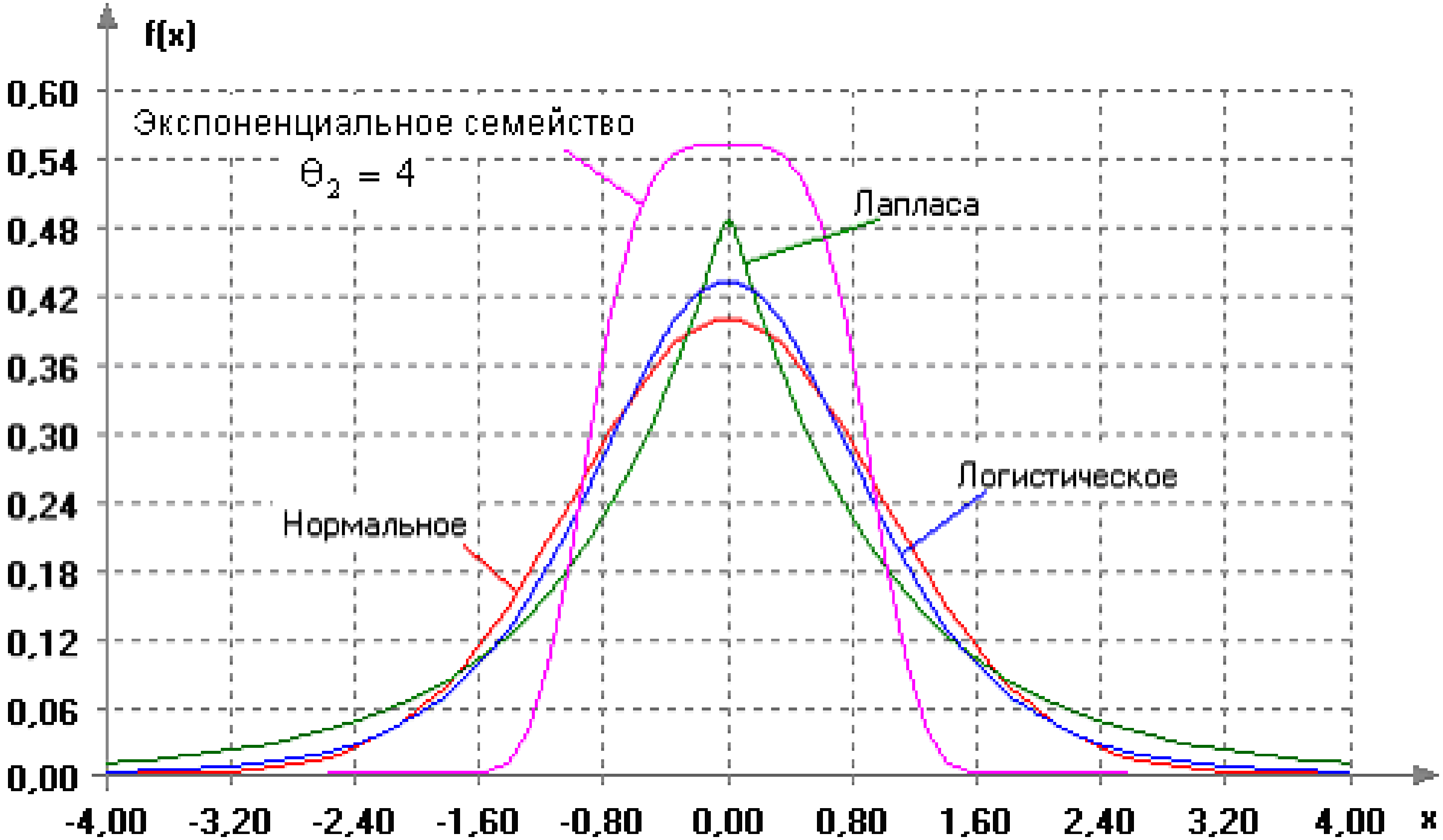
$$f(x) = \frac{1}{2\theta_1}\exp\left\{-|x-\theta_0|/\theta_1\right\}$$

и параметрами (0,1), H_3 – логистическому распределению с плотностью

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

также с параметрами (0,1).

Рассмотрим способность критериев отличать от нормального закона некоторые конкурирующие законы



Плотности распределений, соответствующих рассматриваемым гипотезам H_i

Критерий проверки на симметричность

Критерий предназначен для проверки гипотез о симметричности наблюдаемого закона (против наличия асимметрии) при объемах выборки $8 \le n \le 5000$. Для симметричных законов 3-й центральный момент $\mu_3 = E\left[(X-\mu)^3\right] = 0$. Нормированный коэффициент

$$\sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}$$

в случае симметричности закона также равен 0. Проверяется гипотеза H_0: $\sqrt{\beta_1} = 0$ против альтернативы H_1: $\sqrt{\beta_1} > 0$ (положительная асимметрия) или H_1: $\sqrt{\beta_1} < 0$ (отрицательная асимметрия).

При вычислении статистики данного критерия и критерия проверки на эксцесс оценки используемых центральных моментов (в том числе $\mu_2 = E\left[(X-\mu)^2\right] = \sigma^2$) вычисляются в соответствии с соотношением

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^j ,$$

где

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

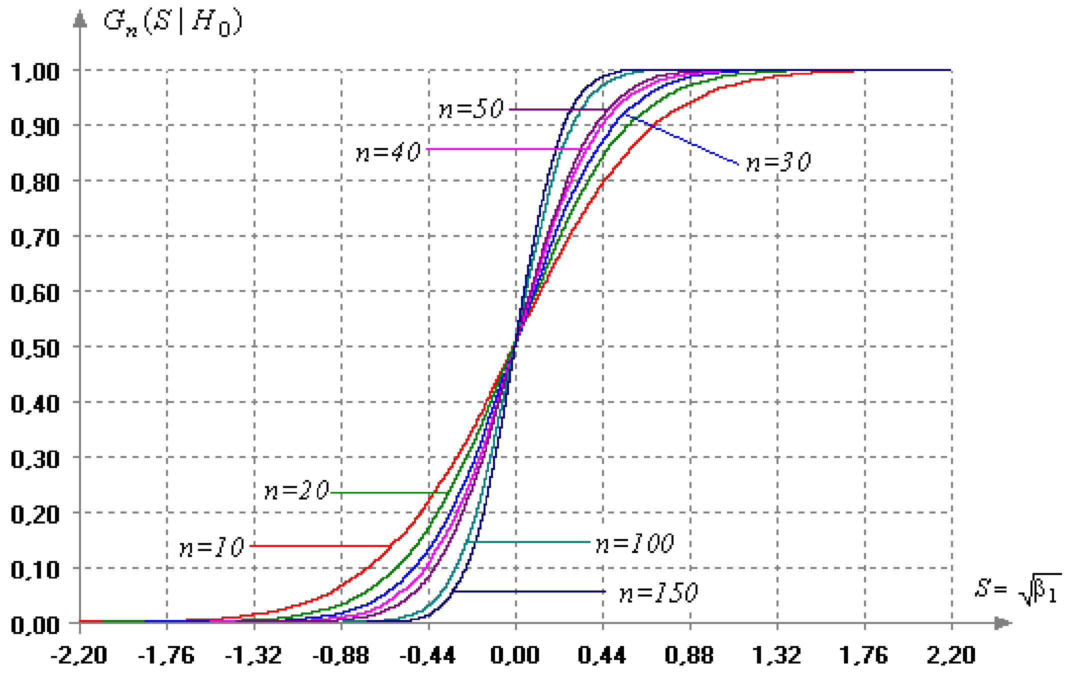


Рис. Распределения статистики критерия проверки на симметричность в зависимости от объема выборки при n=10,20,40,30,50,100,150 в случае нормального закона

Данный критерий является только критерием проверки на симметричность. Его использование полезно при проверке нормальности, но принятие гипотезы о симметричности не может служить подтверждением нормальности (условие необходимое, но не достаточное).

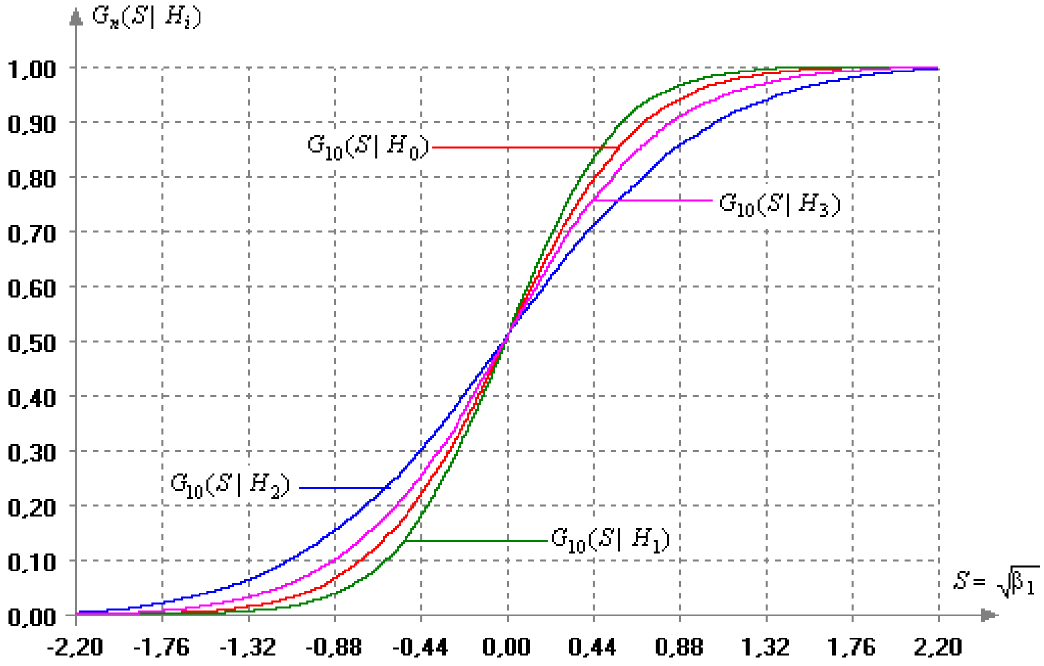


Рис. Распределения статистики критерия проверки на симметричность в зависимости от гипотез H_i при объеме выборок n=10

Критерий проверки на эксцесс

Критерий проверки на эксцесс используется при объемах выборок $8 \le n \le 5000$. Величина эксцесса представляет собой отношение центрального момента 4-го порядка $\mu_4 = E\left[(X-\mu)^4\right]$ к квадрату дисперсии

$$\beta_2 = \frac{\mu_4}{\sigma^4}.$$

В случае нормального закона $\beta_2 = 3$. Проверяется гипотеза вида H_0: $\beta_2=3$ против альтернативы H_1: $\beta_2>3$ (больший эксцесс) или H_1: $\beta_2<3$ (меньший эксцесс).

В стандарте и первоисточниках приводятся лишь таблицы процентных точек. Распределение статистики зависит от объема рассматриваемых выборок.

Недостатком критерия является сильная зависимость распределения статистики от объема выборок.

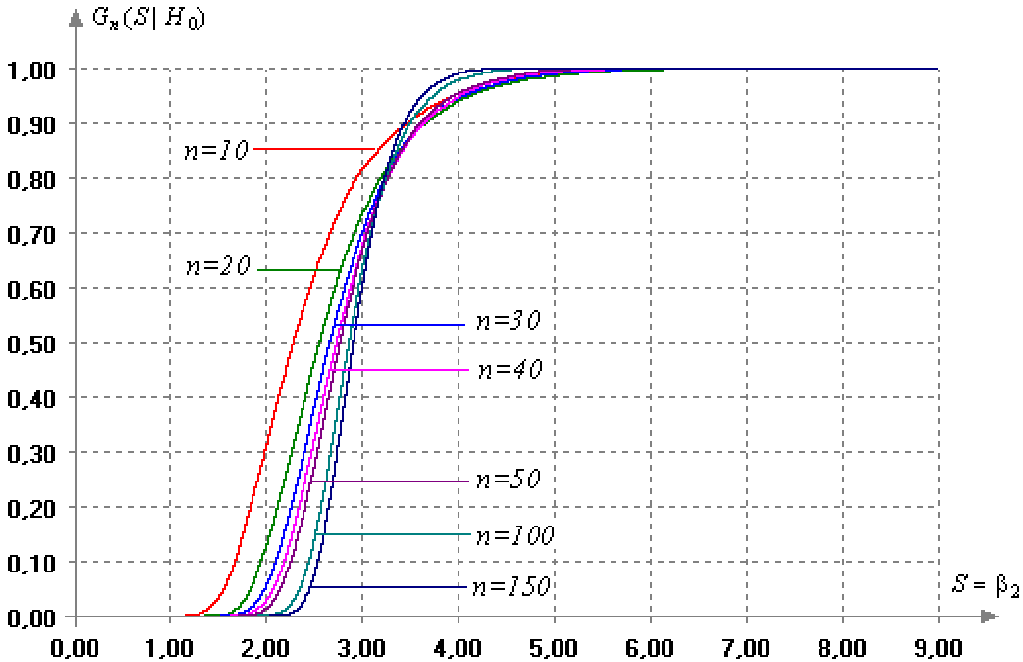


Рис. Графики распределений статистики критерия проверки на эксцесс при нормальном законе в зависимости от объема выборки при n=10,20,40,30,50,100,150

Отдельно данный критерий не позволяет принимать решение о нормальности. Но вместе с критерием симметричности данный критерий дает возможность судить о принадлежности наблюдаемой выборки нормальному закону.

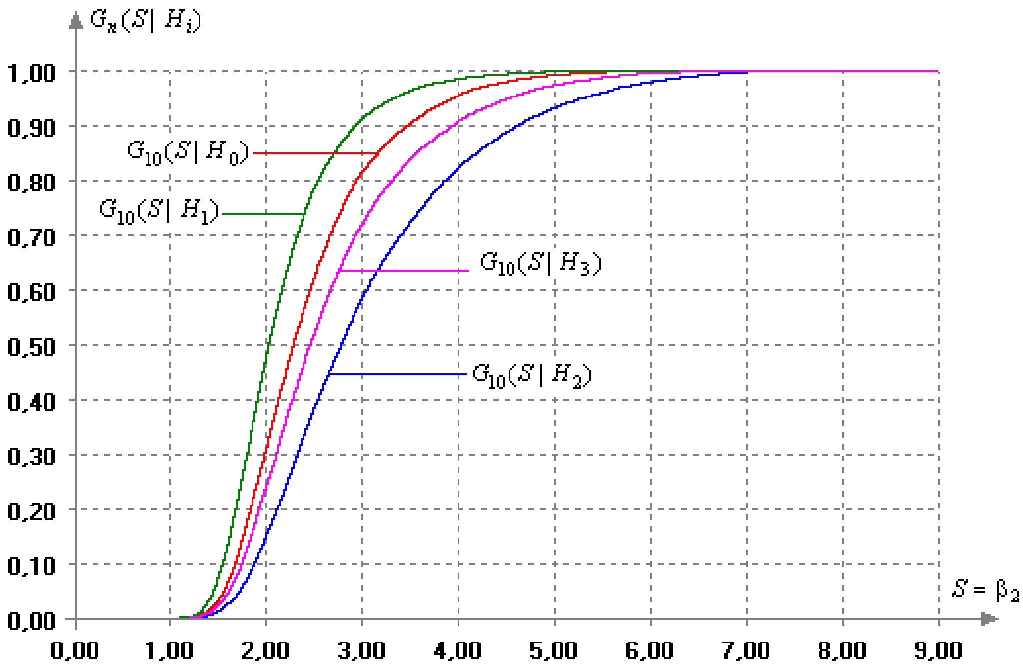


Рис. Распределения статистики критерия проверки на эксцесс в зависимости от вида наблюдаемого закона, соответствующего различным H_i, при n=10

Совместный критерий проверки на симметричность и нулевой коэффициент эксцесса

В критерии используется одномерная статистика на базе статистик $\sqrt{\beta_1}$ и β_2

$$E_p^a = \frac{n\left(\sqrt{\beta_1}\right)^2}{6} + \frac{n(\beta_2 - 3)^2}{24} = \frac{n\left(\dfrac{\mu_3}{\sigma^3}\right)^2}{6} + \frac{n\left(\dfrac{\mu_4}{\sigma^4} - 3\right)^2}{24},$$

которая асимптотически распределена как χ_2^2-распределение.

Но, **к сожалению**, распределение статистики критерия **медленно сходится** к χ_2^2-распределению.

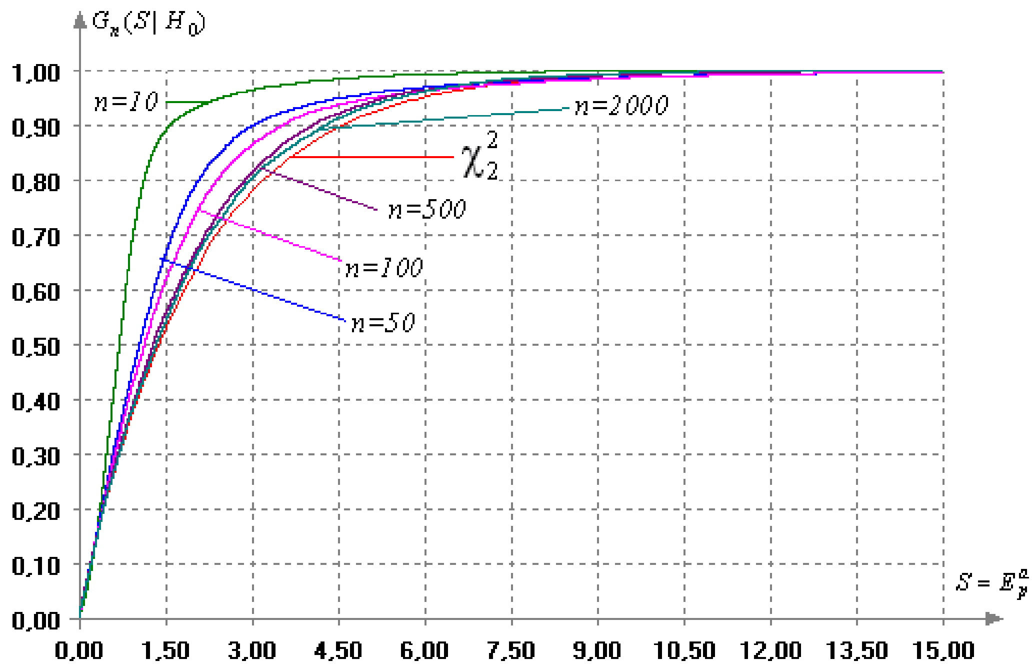


Рис. Сходимость распределений статистики E_p^a к асимптотическому χ_2^2-распределению

Критерий Шапиро - Уилка

Статистика критерия имеет вид

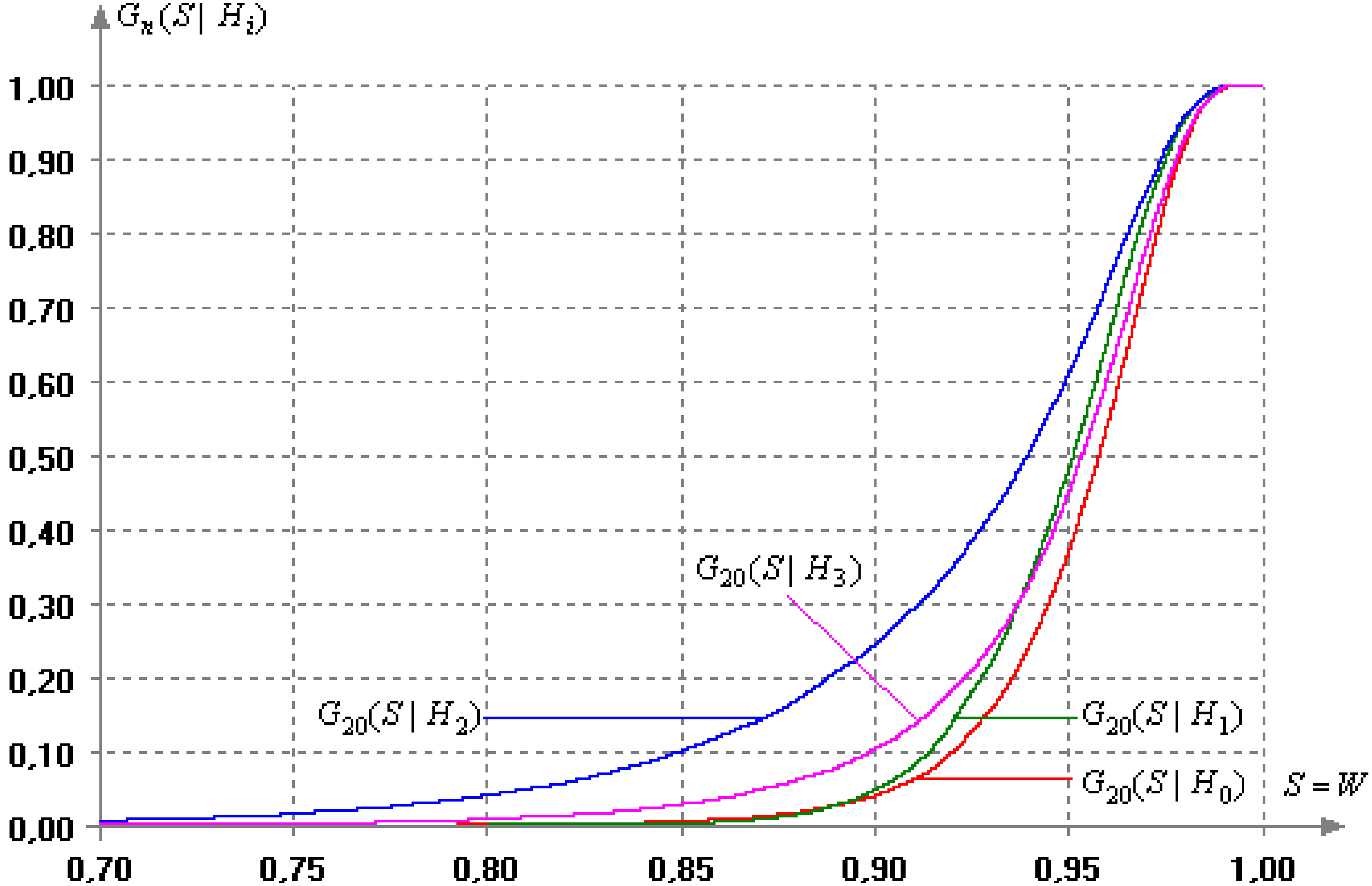
$$W = S^2 \Big/ \sum_{i-1}^{n} \left(X_i - \bar{X} \right)^2 ,$$

где

$$S = \sum_{k} a_k \left[X_{(n+1-k)} - X_{(k)} \right] ,$$

коэффициенты a_k приведены в стандарте и первоисточниках.

Гипотеза о нормальности отвергается при малых значениях статистики W .



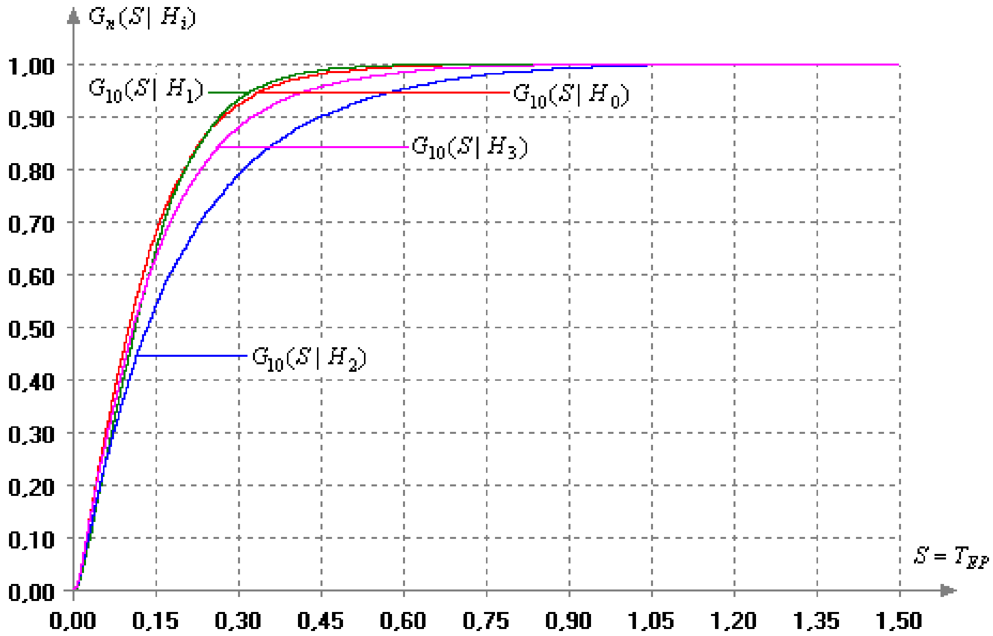
Распределения статистики критерия Шапиро-Уилка в зависимости от наблюдаемого закона при n=20

Критерий Эппса-Палли

Критерий Эппса-Палли базируется на сравнении эмпирической и теоретической характеристических функций. В стандарте предусмотрено его применение при $8 \le n \le 200$. По наблюдаемой выборке $X_1, X_2, ..., X_n$ вычисляют статистику критерия

$$T_{EP} = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n}\sum_{k=2}^{n}\sum_{j=1}^{k-1}\exp\left\{-\frac{\left(X_j - X_k\right)^2}{2\hat{\mu}_2}\right\} - \sqrt{2}\sum_{j=1}^{n}\exp\left\{-\frac{\left(X_j - \overline{X}\right)^2}{4\hat{\mu}_2}\right\},$$

где $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Выборка может быть неупорядочена, порядок наблюдений произволен, но он должен быть неизменным в течение всех проводимых вычислений. Гипотезу о нормальности отвергают при больших значениях статистики.



Распределения статистики критерия Эппса-Палли в зависимости от наблюдаемого закона при n=10

Эти рисунки говорят о смещенности критериев относительно некоторых альтернатив.

Модифицированный критерий Шапиро-Уилкса

Для h последовательных выборок объемом n каждая, отобранных из одной совокупности, подсчитывается значение W_j $(j=\overline{1,h})$ в соответствии с выражением

$$W_j = S_j^2 \Big/ \sum_{i-1}^{n} \left(X_i - \overline{X}\right)^2 ,$$

где $S_j = \sum_k a_k \left[X_{(n=1-k)} - X_{(k)}\right]$, индекс k изменяется от 1 до $n/2$ или от 1 до $(n-1)/2$ при четном и нечетном n соответственно.

Для совместного критерия вычисляют соответствующие значения C_j по формуле:

$$C_j = \gamma(n) + \delta(n) v_j ,$$

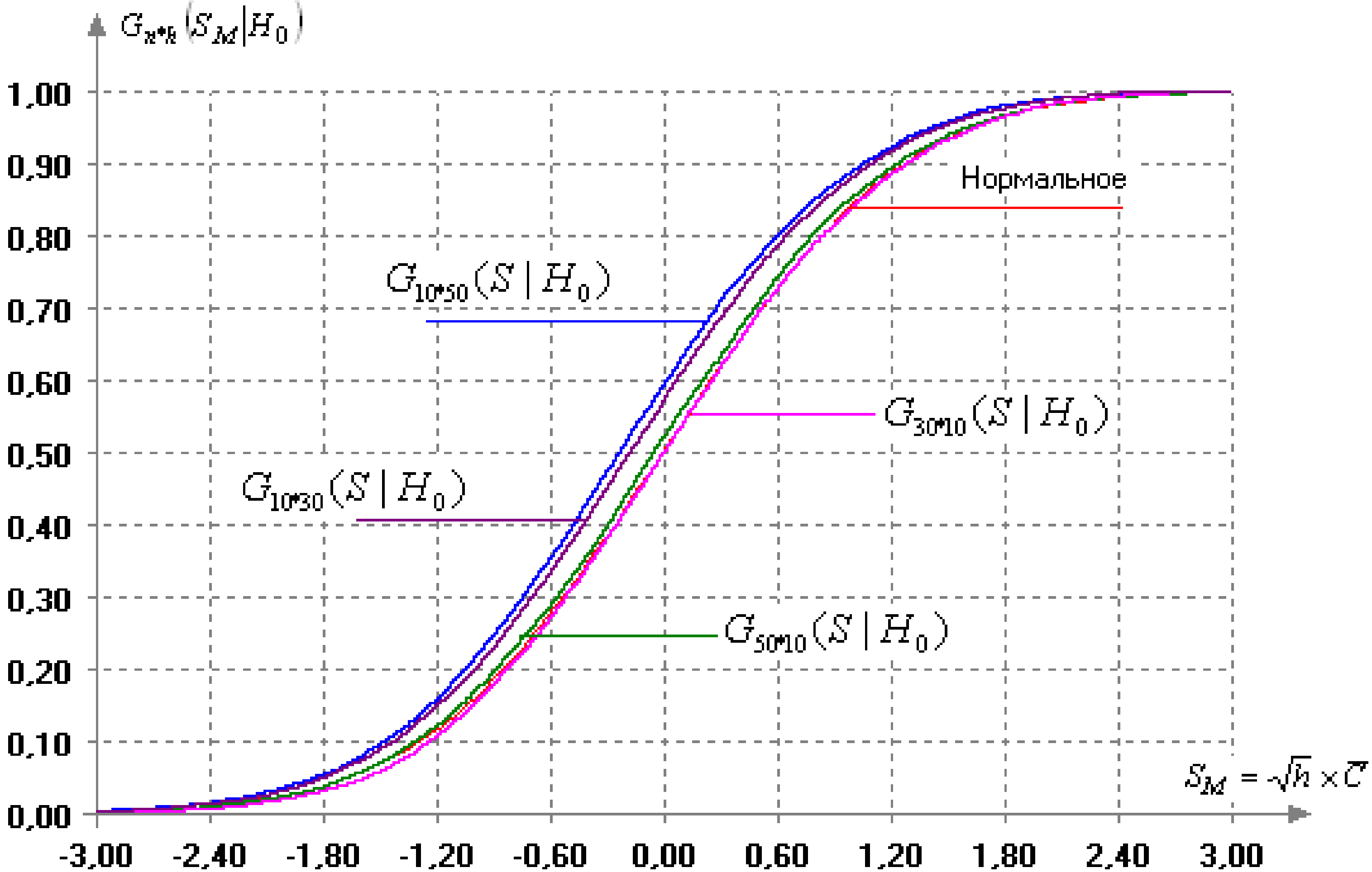
где $v_j = \ln\left\{\frac{W_j - E(n)}{1-W_j}\right\}$. Коэффициенты $\gamma(n), \delta(n)$ и $E(n)$ для преобразования W_j и C_j табулированы.

Если основное распределение вероятностей нормальное, величины C_j приблизительно подчиняются нормальному распределению. Среднее арифметическое значение переменное C_j равно $\overline{C} = \frac{1}{h}\sum_{j=1}^{h} C_j$ и статистикой критерия является величина

$$S_M = \sqrt{h} \times \overline{C} ,$$

которая должна подчиняться стандартному нормальному закону.

Распределения статистики модифицированного критерия Шапиро-Уилка плохо сходятся к «предельному»



Условные распределения $G_{n*h}\left(S_M \mid H_0\right)$ статистики модифицированного критерия Шапиро-Уилка при различных комбинациях n и h

Модификация D'Agostino критерия проверки на симметричность

Предложена модификация критерия проверки симметричности, в которой распределение статистики подчиняется стандартному нормальному закону. Преобразование коэффициента асимметрии $\sqrt{\beta_1}$ в стандартную нормальную величину z_1 осуществляется с помощью следующих соотношений:

$$b = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$

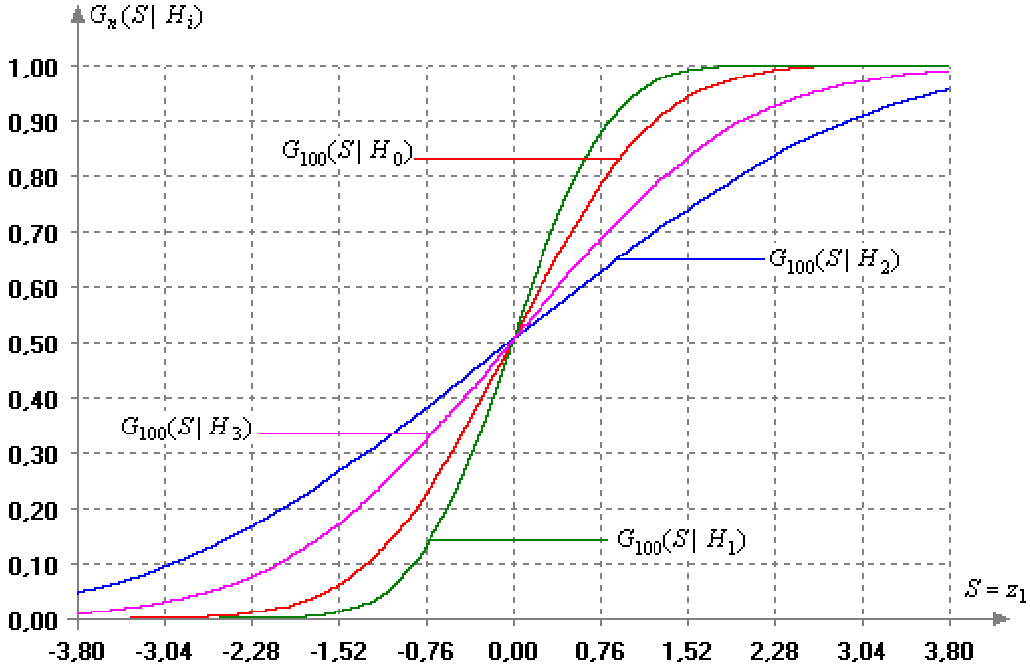
$$\omega^2 = -1 + \left\{2(b-1)\right\}^{1/2},$$

$$\delta = \frac{1}{\left\{\log(\sqrt{\omega^2})\right\}^{1/2}},$$

$$y = \sqrt{\beta_1}\left\{\frac{\omega^2 - 1}{2} \cdot \frac{(n+1)(n+3)}{6(n-2)}\right\}^{1/2},$$

$$z_1 = \delta \log\left\{y + (y^2 + 1)^{1/2}\right\}.$$

Это критерий проверки симметричности (только), но он удобней, так как опирается на стандартное нормальное распределение.



Условные распределения статистики z_1 в зависимости от вида наблюдаемого закона, соответствующего различным H_i, при n=100

Модификация D'Agostino критерия проверки отклонения от нормального закона

При помощи следующих преобразований статистики критериев симметричности и эксцесса преобразуются в статистику, приближенно распределенную по стандартному нормальному закону:

$$\delta = (n-3)(n+1)(n^2+15n-4),$$

$$a = \frac{(n-2)(n+5)(n+7)(n^2+27n-70)}{6\delta},$$

$$c = \frac{(n-7)(n+5)(n+7)(n^2+2n-5)}{6\delta},$$

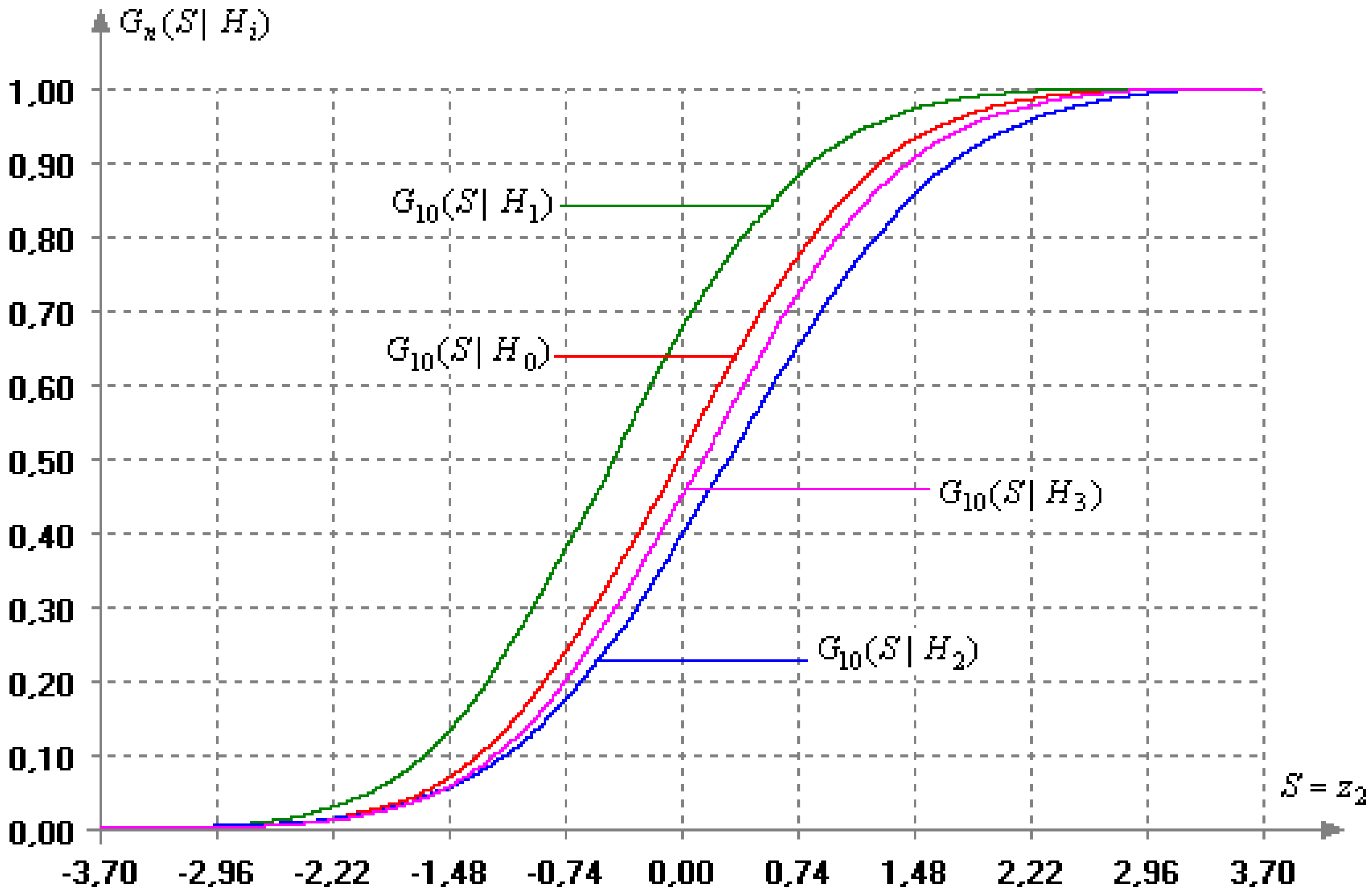
$$k = \frac{(n+5)(n+7)(n^3+37n^2+11n-313)}{12\delta},$$

$$\alpha = a + \beta_1 c,$$

$$\chi = (\beta_2 - 1 - \beta_1)2k,$$

$$z_2 = \left\{\left(\frac{\chi}{2\alpha}\right)^{1/3} - 1 + \frac{1}{9\alpha}\right\}(9\alpha)^{1/2}, \qquad (1)$$

где β_1 - оценка коэффициента симметрии, β_2 - оценка коэффициента эксцесса



Условные распределения статистики z_2 D`Agostino в зависимости от вида наблюдаемого закона при n=10

Данный критерий отличает все альтернативы, но сходимость распределения статистики к стандартному нормальному закону хуже.

Критерий Фросини

Статистика критерия Фросини имеет вид

$$B_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left| \Phi(z_i) - \frac{i-0.5}{n} \right|,$$

где элементы выборки x_i, $i = \overline{1,n}$, упорядочены по возрастанию, $z_i = \dfrac{x_i - \overline{x}}{s}$,

$\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, $\Phi(z_i)$ – функция распределения стандартного нормального закона $N(0,1)$.

Применение критерия осложняется тем, что условные распределения $G(B_n | H_0)$ статистики критерия Фросини при справедливости проверяемой гипотезы H_0 зависят от объемов выборок n.

С ростом n распределения $G(B_n|H_0)$ статистики смещаются вправо и достаточно быстро сходятся к некоторому предельному распределению. При объемах выборок $n > 100$ распределения статистики уже существенно не меняются.

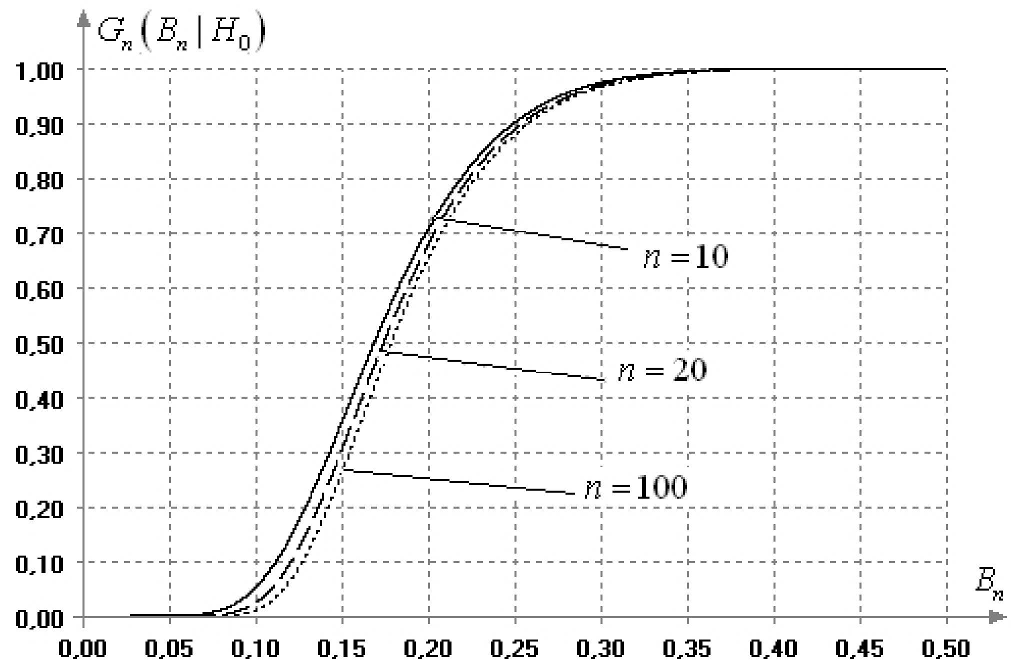


Рис. Зависимость распределения $G(B_n \mid H_0)$ статистики B_n от объема выборки

У критерия Фросини отсутствует недостаток, свойственный при малых n ($n \leq 20$) критериям Шапиро-Уилка и Эппса-Палли по отношению к конкурирующей гипотезе H_1.

Критерии Хегази-Грина

Хегази и Грин в [5] предложили критерии со статистиками

$$T_1 = \frac{1}{n}\sum_{i=i}^{n}\left|z_i - \eta_i\right|,$$

$$T_2 = \frac{1}{n}\sum_{i=1}^{n}\{z_i - \eta_i\}^2 ,$$

где $z_i = \dfrac{x_{(i)} - \bar{x}}{s}$, $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, η_i – математическое ожидание i-й порядковой статистики стандартного нормального закона, которое можно найти из соотношения $\eta_i = \Phi^{-1}\left(\dfrac{i}{n+1}\right)$. Проверяемая гипотеза отклоняется при больших значениях статистик. Подчеркнем, что в статистике должна использоваться именно несмещенная оценка дисперсии, это не учтено в [Кобзарь].

Распределения этих статистик очень сильно зависят от объема выборки.

При $n = 10$ и $n = 20$ критерий ***практически не различает гипотезы*** H_0 и H_1 (аналогично критериям Шапиро-Уилка и Эппса-Палли он оказывается смещённым).

Критерий Гири

Гири в работах [7, 11, 12] рассмотрел критерий проверки отклонения от нормального закона, основанный на статистике

$$d = \frac{1}{ns}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|,$$

где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Критерий является двусторонним, и гипотеза о нормальности не отклоняется, если $d_{\alpha/2} \le d \le d_{1-\alpha/2}$.

Критерий Гири с простой статистикой достаточно уверенно демонстрирует высокую мощность к различным конкурирующим гипотезам.

Однако распределения статистики существенно зависят от объема выборок.

Автором критерия утверждается, что статистика критерия при $n \ge 50$ распределена асимптотически нормально. Выражения для математического ожидания и дисперсии асимптотического закона представлены, например, в [Кобзарь].

Однако на самом деле распределения статистики асимметричны и плохо аппроксимируются нормальным законом, **тем более** с указанными в [Кобзарь] параметрами.

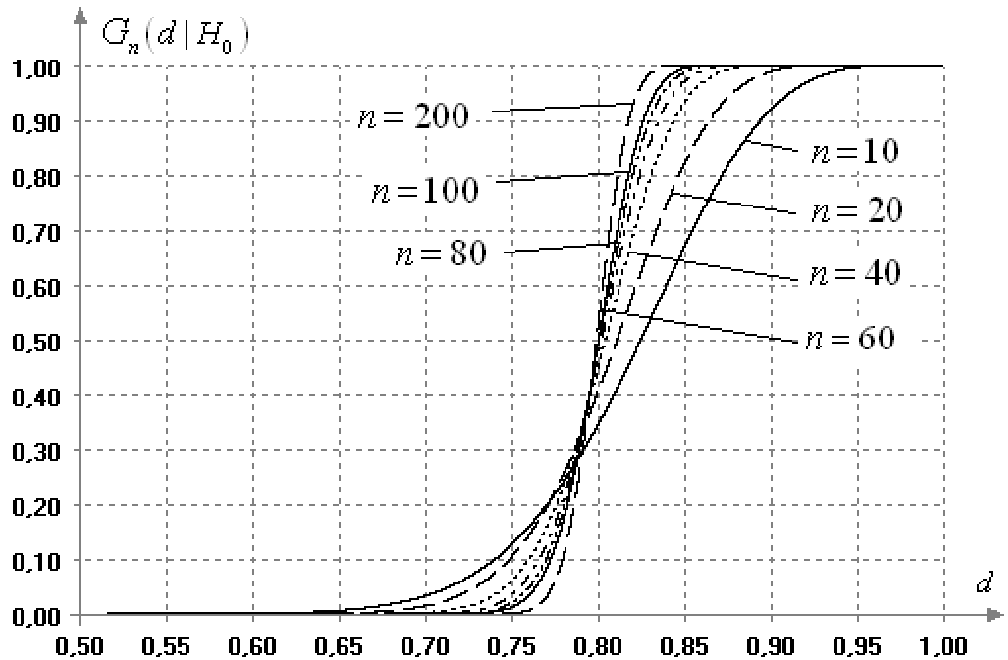


Рис. Зависимость распределений статистики d от объема выборки

Критерий Дэвида-Хартли-Пирсона

В критерии Дэвида-Хартли-Пирсона [8] рассматривается отношение размаха выборки к выборочному стандартному отклонению, и его статистика имеет вид:

$$U = \frac{R}{s},$$

где $R = x_{\max} - x_{\min}$ - размах выборки, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ - несмещённая оценка дисперсии.

Критерий двусторонний: гипотеза о нормальности распределения отвергается, если $U < U_{\alpha/2}$ или $U > U_{1-\alpha/2}$.

Критерий Дэвида-Хартли-Пирсона уступает остальным рассмотренным критериям.

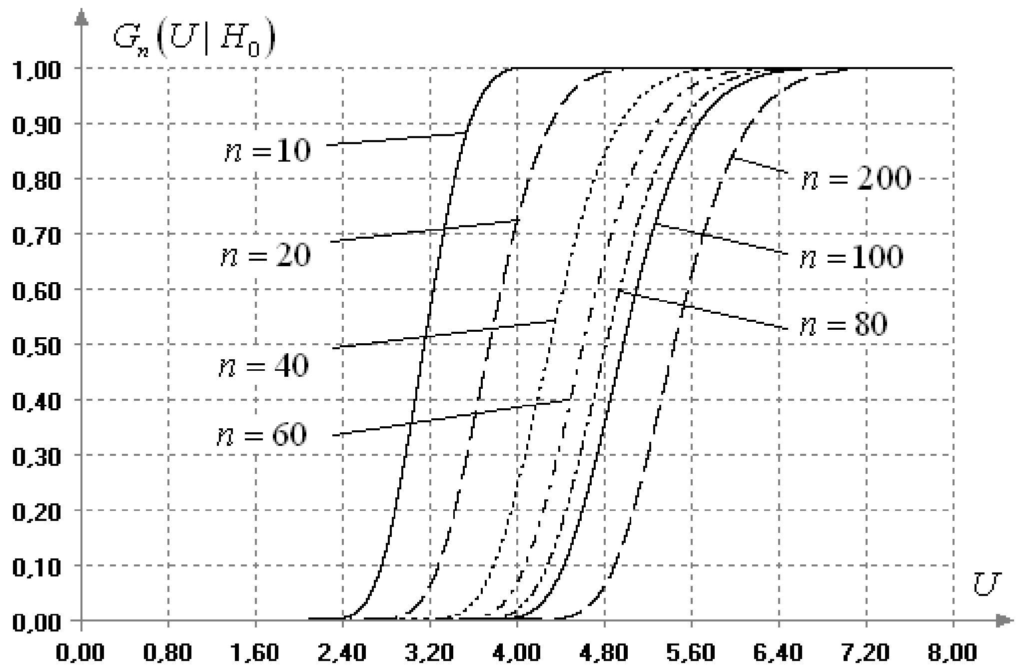


Рис. Зависимость распределений статистики U от объема выборки

Критерий Шпигельхальтера

Его статистика базируется на комбинации статистик критериев Гири и Дэвида, Хартли и Пирсона и имеет вид

$$T' = \{(C_n U)^{-(n-1)} + g^{-(n-1)}\}^{\frac{1}{n-1}},$$

где $C_n = \frac{1}{2n}(n!)^{\frac{1}{n-1}}$, U – статистика критерия Дэвида-Хартли-Пирсона, $g = \frac{d}{\sqrt{(n-1)/n}}$, d – статистика критерия Гири.

Проверяемая гипотеза о принадлежности анализируемой выборки нормальному закону по критерию Шпигельхальтера отклоняется при больших значениях статистики T'.

Зависимость распределения статистики T' от объема выборки в случае справедливости проверяемой гипотезы иллюстрирует следующий рисунок.

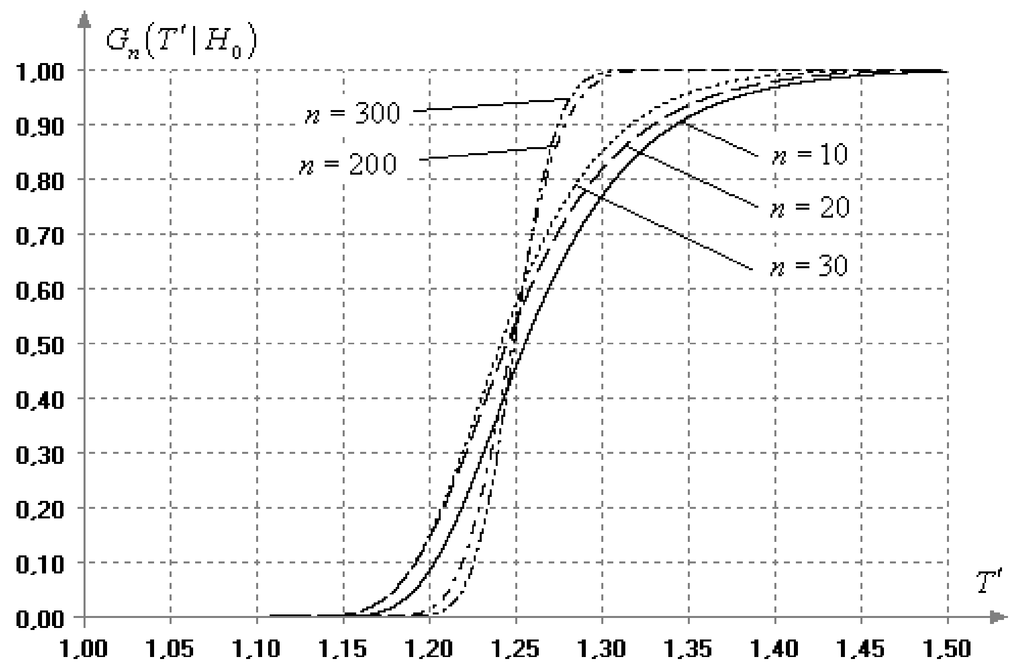


Рис. Зависимость распределения статистики T' от объема выборки в случае нормального закона

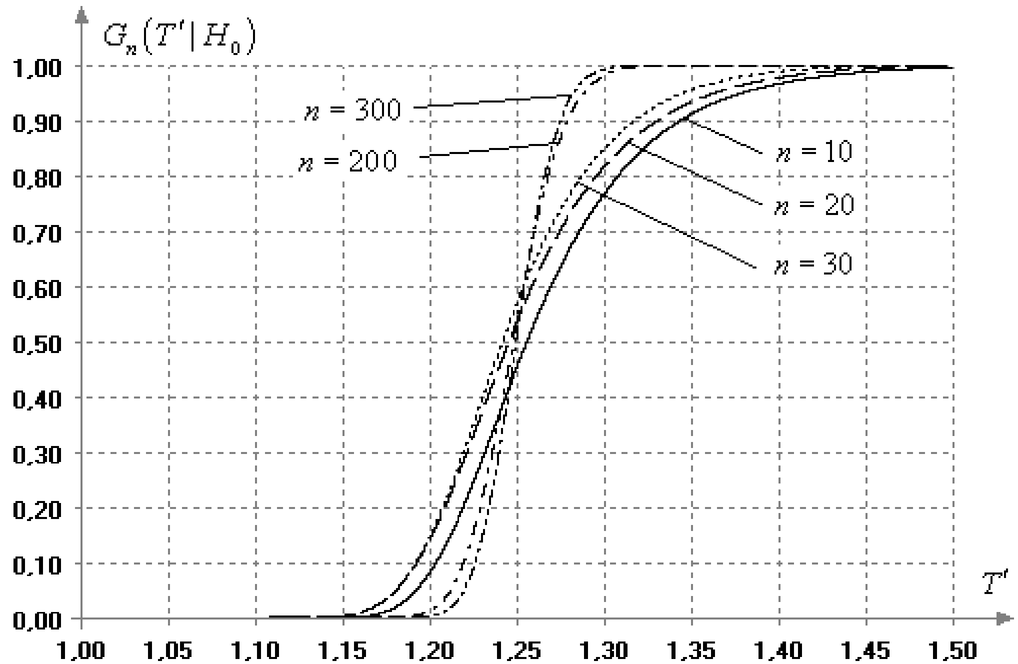


Рис. Зависимость распределения статистики T' от объема выборки в случае нормального закона

Однако данный критерий **имеет** очень **существенный недостаток**: критерий способен отличить от нормального закона далеко не все конкурирующие распределения. В частности, это касается конкурирующей гипотезы H_1.

Мощность любого корректно построенного критерия должна увеличиваться с ростом n. В принципе, так и происходит с мощностью критерия Шпигельхальтера по отношению к конкурирующим гипотезам H_2 и H_3.

Совсем другая картина наблюдается относительно гипотезы H_1.

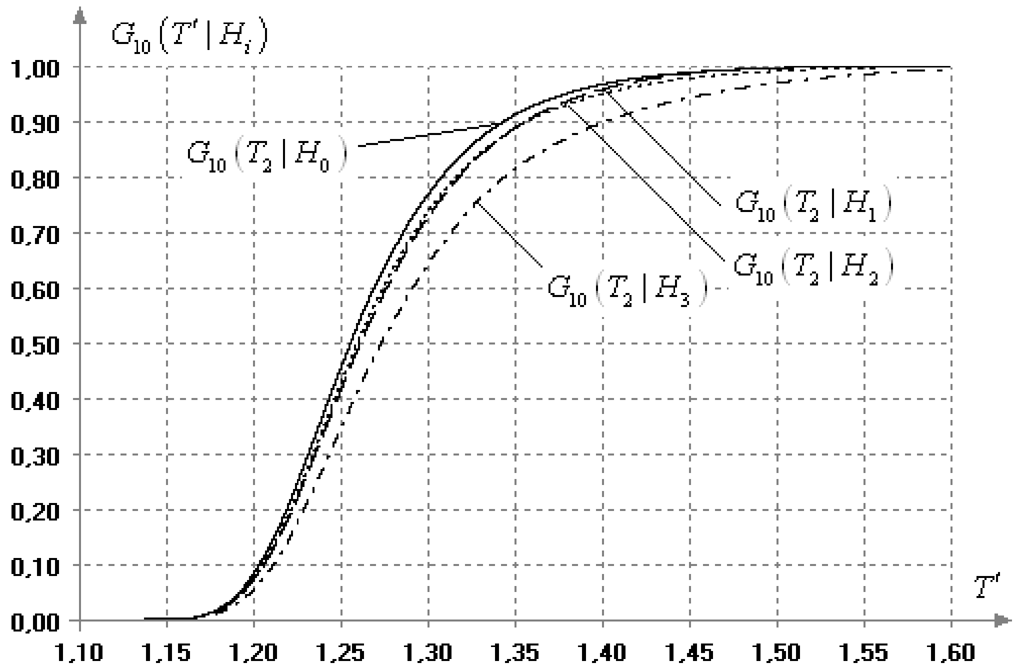


Рис. Условные распределения $G(T' \mid H_i)$ статистики T' при объеме выборок $n = 10$

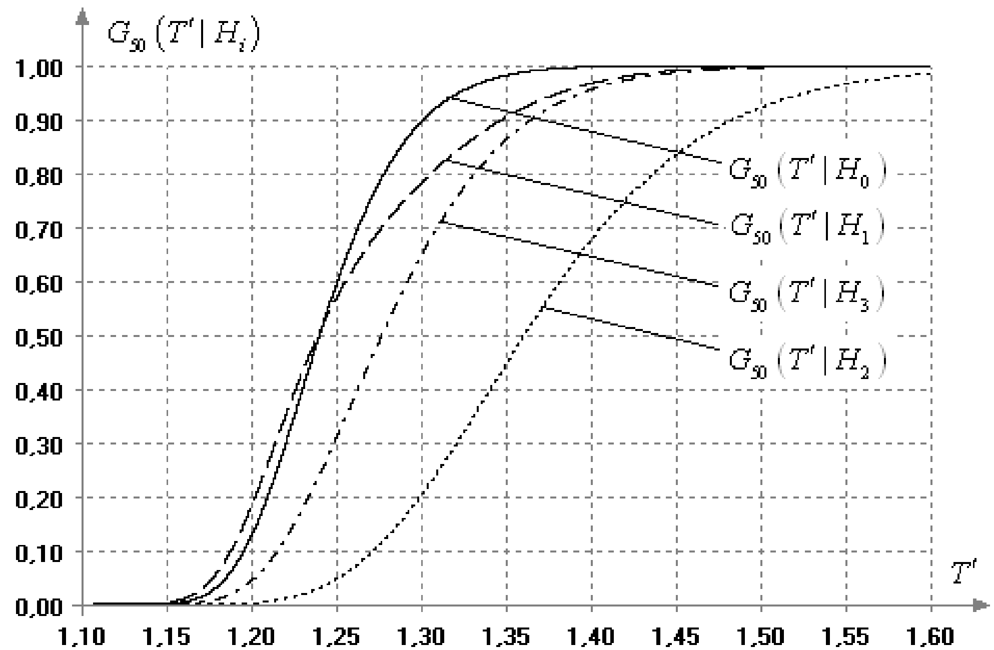


Рис. Условные распределения $G(T' \mid H_i)$ статистики T' при объеме выборок $n = 50$

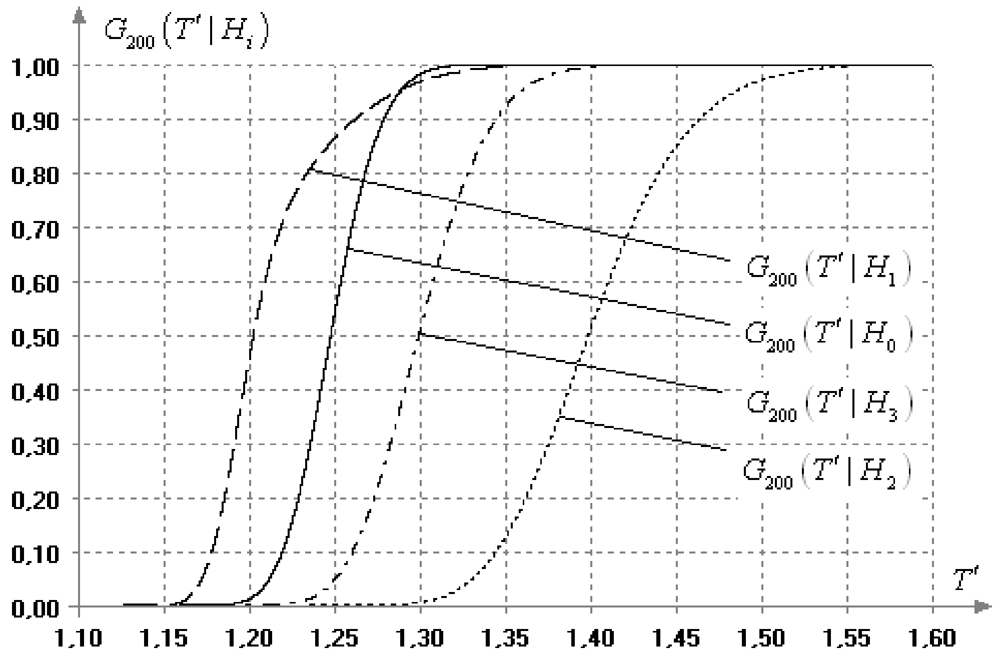


Рис. Условные распределения $G(T' | H_i)$ статистики T' при объеме выборок $n = 200$

Выводы

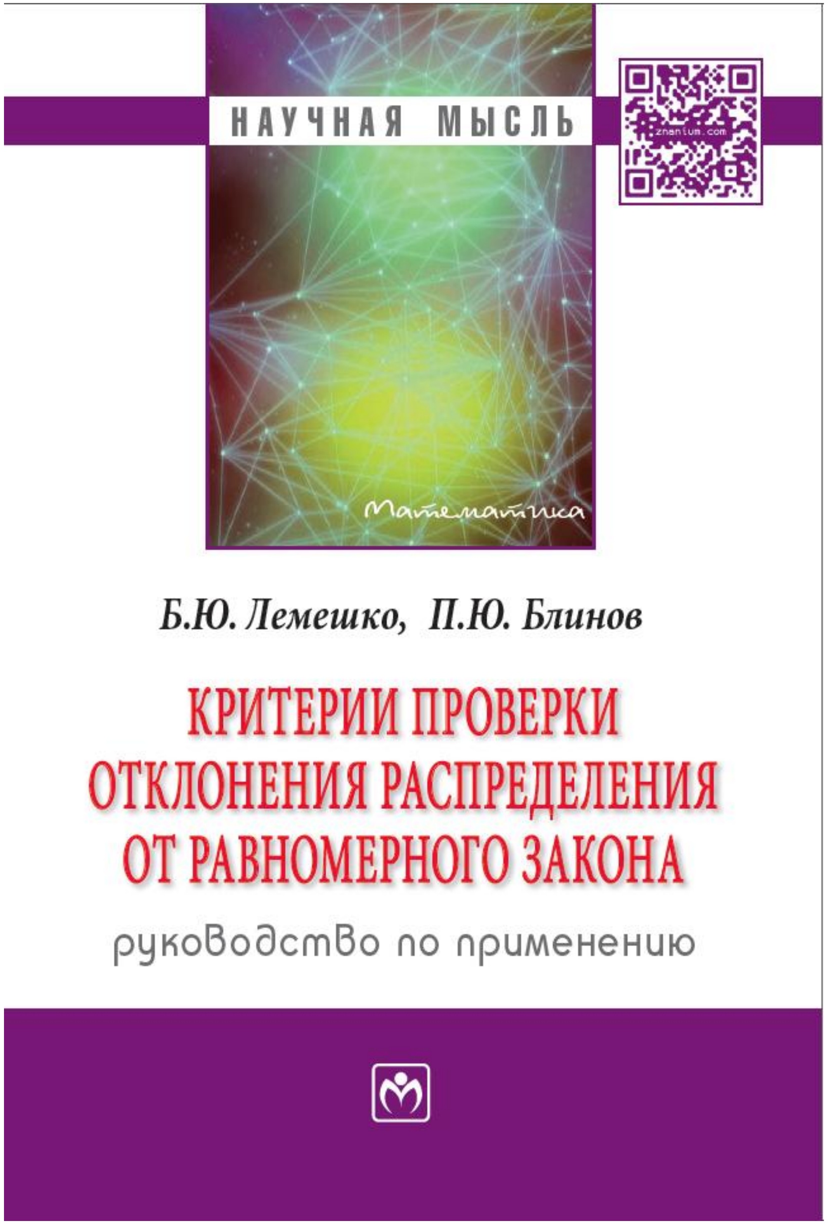
Рассмотренные критерии по мощности относительно конкурирующих гипотез H_1, H_2, H_3 можно проранжировать следующим образом:

Гири $\succ$ Шпигельгалтера[1)] $\succ$ Хегази-Грина (T_2)[2)] $\succ$ Хегази-Грина (T_1)[3)] $\succ$ Эппса-Палли $\succ$ Дэвида-Хартли-Пирсона $\succ$ Шапиро-Уилка $\succ$ Фросини.

Критерий Д`Агостино со статистикой z_2 относительно рассмотренных конкурирующих гипотез H_1 и H_3 показал себя наиболее мощным. В приведенном выше ряду предпочтения он просится на первое место. Однако по отношению к более далекой гипотезе H_2 он уступает по мощности остальным критериям.

Следует отметить, что относительно наиболее близкой конкурирующей гипотезы H_3 критерии согласия Ω^2 Андерсона-Дарлинга и критерий типа χ^2 Никулина в условиях проверки сложной гипотезы лишь не многим уступают по мощности критериям со статистикой z_2, Хегази-Грина, Шпигельгалтера и Гири, превосходя остальные критерии проверки отклонения от нормального закона.


НАУЧНАЯ МЫСЛЬ
znanium.com
Статистика
Б.Ю. Лемешко
КРИТЕРИИ ПРОВЕРКИ ОТКЛОНЕНИЯ РАСПРЕДЕЛЕНИЯ ОТ НОРМАЛЬНОГО ЗАКОНА
руководство по применению


НАУЧНАЯ МЫСЛЬ
znanium.com
Математика
Б.Ю. Лемешко, П.Ю. Блинов
КРИТЕРИИ ПРОВЕРКИ ОТКЛОНЕНИЯ РАСПРЕДЕЛЕНИЯ ОТ РАВНОМЕРНОГО ЗАКОНА
руководство по применению


НЕПАРАМЕТРИЧЕСКИЕ
КРИТЕРИИ СОГЛАСИЯ
Б.Ю. Лемешко