Как обеспечить максимальную мощность критериев согласия типа хи-квадрат

Лемешко Борис Юрьевич

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

Критерии согласия предназначены для проверки гипотез о соответствии эмпирического распределения некоторому теоретическому закону.

Различают проверку простых и сложных гипотез.

Простая проверяемая гипотеза имеет вид H_0: $F(x) = F(x, \theta)$, где $F(x, \theta)$ – функция распределения вероятностей, с которой проверяют согласие наблюдаемой выборки, а θ – известное значение параметра (скалярного или векторного).

Сложная проверяемая гипотеза может быть записана в виде H_0: $F(x) \in \{F(x, \theta),\ \theta \in \Theta\}$, где Θ – область определения неизвестного параметра θ. Отличие в процедуре применения критериев при проверке сложных гипотез и соответствующие проблемы возникают, если оценку параметра $\hat{\theta}$ теоретического распределения вычисляют по той же самой выборке, по которой проверяют согласие.

Далее мы, как правило, будем предполагать, что при проверке сложных гипотез оценка параметра $\hat{\theta}$ вычисляется по этой же выборке.

С любым критерием согласия **связана некоторая статистика** S**,** измеряющая расстояние между теоретическим законом $F(x,\theta)$ и эмпирическим законом $F_n(x)$, построенным по выборке случайных величин $x_1, x_2, \ldots, x_n$.

С проверкой статистических гипотез связывают ошибки двух видов.

Ошибка 1-го рода заключается в том, что в результате проверки отклоняется справедливая проверяемая гипотеза H_0. *Вероятность ошибки 1-го рода* обозначают α.

Ошибка 2-го рода – в признании верной гипотезы H_0, когда на самом деле справедлива некоторая конкурирующая гипотеза H_1. *Вероятность ошибки 2-го рода* обозначают β.

Процедура проверки гипотезы H_0 предполагает, что известно распределение $G(S|H_0)$ статистики S применяемого критерия при справедливости H_0.

Вероятность ошибки 1-го рода α (уровень значимости) представляет собой вероятность попадания значения статистики в критическую область: $\alpha = P\{S > S_\alpha | H_0\} = 1 - G(S_\alpha | H_0)$, где S_α – критическое значение. При проверке гипотез величина α, как правило, задается.

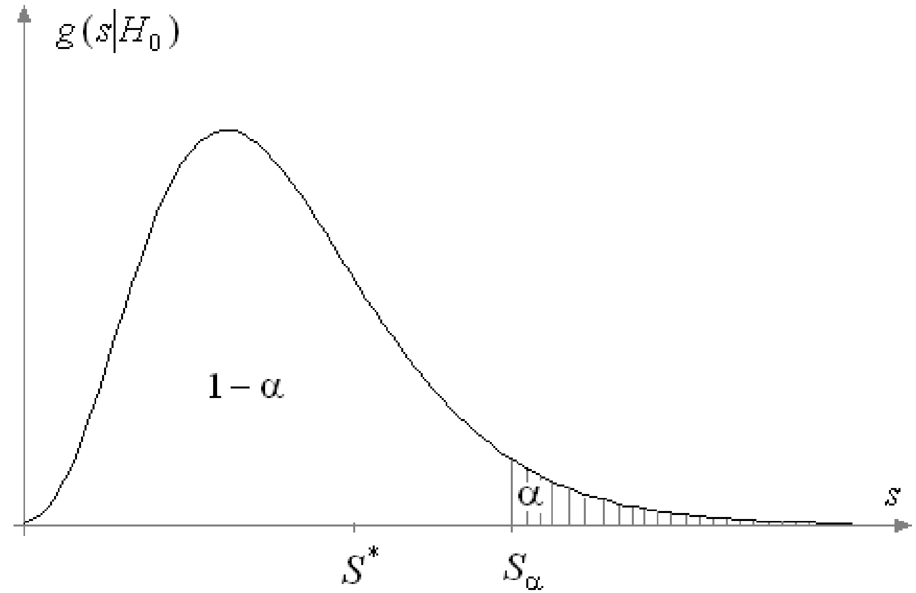


Рис. 1. Плотность распределения статистики при истинной гипотезе H_0

Если вычисленное по выборке значение статистики $S^* \le S_\alpha$, или, что то же самое, достигнутый уровень значимости $P\{S > S^* | H_0\} = 1 - G(S^* | H_0) > \alpha$, то проверяемая гипотеза H_0 не отклоняется.

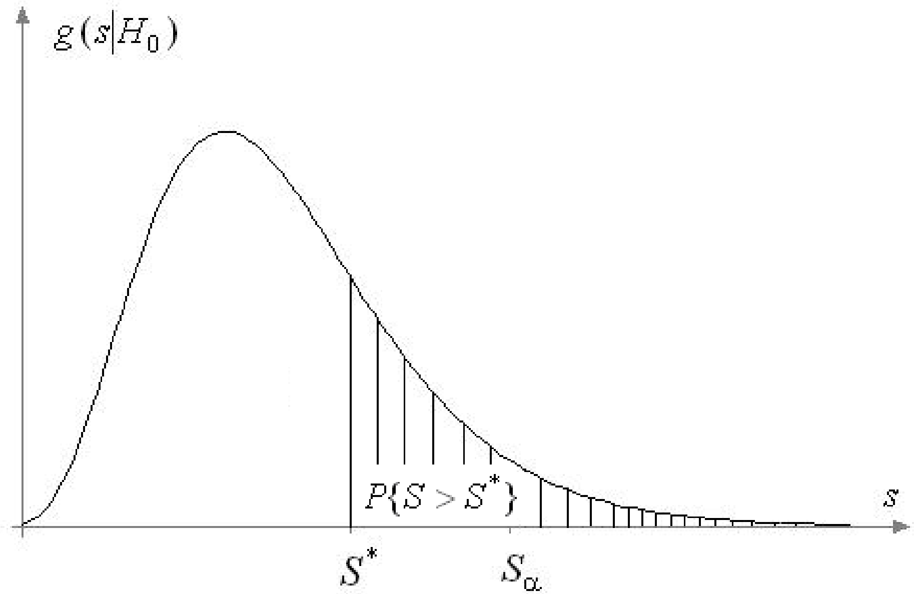


Рисунок 2 – Плотность распределения статистики при истинной гипотезе H_0

Или, что, то же самое, если достигнутый уровень значимости $P\{S > S^* | H_0\} = 1 - G(S^* | H_0) > \alpha$, то проверяемая гипотеза H_0 не отклоняется.

При задании конкурирующей гипотезы H_1 вероятность ошибки 2-го рода определяется соотношением $\beta = P\{S \le S_\alpha | H_1\} = G(S_\alpha | H_1)$, где $G(S|H_1)$ – распределение статистики критерия при справедливости H_1.

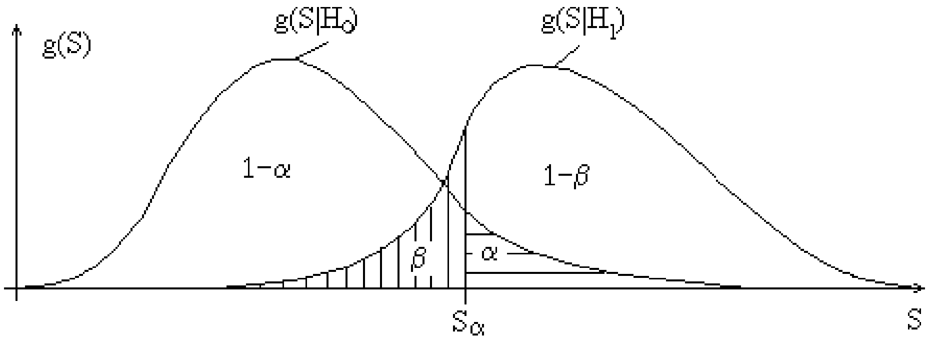


Рисунок 3 – Плотности распределения статистик при справедливости гипотез H_0 и H_1

Если критерий полностью определен, то задание α однозначно определяет величину β и наоборот. Мощность $1-\beta$ критерия при проверке гипотезы H_0 относительно H_1 представляет собой функцию, зависящую от H_0, H_1, объема выборки n и, возможно, от некоторых других факторов, связанных с построением критерия.

Чем больше мощность критерия $1-\beta$, тем лучше он различает соответствующие гипотезы.

При проведении статистического анализа, отдавая предпочтение некоторому критерию, хотелось бы иметь уверенность в том, что для заданной вероятности ошибки первого рода α гарантируется минимальная вероятность ошибки 2-го рода β, то есть, что критерий обладает наибольшей мощностью относительно интересующей нас пары конкурирующих гипотез H_0 и H_1.

Особенно важна способность критериев различать близкие конкурирующие гипотезы.

Проблемы применения критериев типа хи-квадрат

Критерий χ^2 Пирсона. Применение критериев типа χ^2 предусматривает разбиение области определения случайной величины на k интервалов с подсчетом числа наблюдений n_i, попавших в них, и вероятностей попадания в интервалы $P_i(\theta)$, соответствующих теоретическому закону. Статистика критерия согласия χ^2 Пирсона имеет вид

$$X_n^2 = n\sum_{i=1}^{k}\frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)}. \qquad \textbf{(4)}$$

В случае проверки простой справедливой гипотезы в пределе эта статистика подчиняется χ^2_{k-1}–распределению с $k-1$ степенями свободы.

Если верна конкурирующая гипотеза H_1 и выборка соответствует закону с распределением $F_1(x,\theta_1)$ с параметром θ_1, то эта же статистика в пределе подчиняется нецентральному χ^2_{k-1}-распределению с параметром нецентральности

$$\nu = n\sum_{i=1}^{k}\frac{(P_i^1(\theta_1) - P_i(\theta))^2}{P_i(\theta)},$$

где $P_i^1(\theta_1)$ – вероятность попадания в интервал при справедливой гипотезе H_1.

В случае проверки сложной гипотезы, при справедливости H_0 и при условии, что оценки параметров находятся в результате минимизации статистики (4) по этой же самой выборке, статистика X_n^2 асимптотически распределена как χ^2_{k-r-1}, где r – число оцененных по выборке параметров.

Ошибки и неверные действия при использовании критериев типа хи-квадрат

Анализ примеров “неудачного” применения критериев типа χ^2 позволяет выделить две группы причин, которые могут приводить к неверным статистическим выводам.

Во-первых, это часто совершаемые принципиальные ошибки, при которых использование в качестве предельного χ^2_{k-r-1}-распределения оказывается неправомерным.

Во-вторых, действия, использующие возможности критерия не наилучшим образом.

В первом случае возрастает вероятность ошибки первого рода α (отклонить верную проверяемую гипотезу), во втором – вероятность ошибки второго рода β (принять проверяемую гипотезу при справедливости альтернативы).

Первоначально предполагалось, что в случае проверки сложных гипотез и оценивании по выборке параметров наблюдаемого закона использование в качестве предельных χ^2_{k-r-1}-распределений справедливо лишь при определении оценок минимизацией статистики X_n^2.

Позднее было доказано, что статистика X_n^2 подчиняется χ^2_{k-r-1}-распределению и в том случае, если используются оценки максимального правдоподобия (ОМП) по группированным наблюдениям.

Наши исследования методами статистического моделирования распределений данной статистики при проверке сложных гипотез и использовании ОМП по группированным наблюдениям (при конечных объемах выборок) также подтвердили хорошее согласие получаемых эмпирических распределений статистики с χ^2_{k-r-1}-распределениями. Кроме того, наши исследования показали, что есть все основания использовать χ^2_{k-r-1}-распределения в качестве предельных распределений статистики X_n^2 и в том случае, если параметры сдвига и масштаба наблюдаемых законов случайных величин будут находиться в виде линейных комбинаций выборочных квантилей (L-оценок и оптимальных L-оценок).

Следует полагать, что применение χ^2_{k-r-1}-распределений в качестве предельных распределений оказывается оправданным и при использовании ряда других оценок, предусматривающих группирование наблюдений, в частности, при нахождении оценок в результате минимизации модифицированной статистики X_n^2

$$\mathrm{mod}\, X_n^2 = n\sum_{i=1}^{k}\frac{\left(n_i - nP_i(\theta)\right)^2}{n_i},$$

где n_i заменяется на 1, если $n_i = 0$, в результате минимизации расстояния Хеллингера

$$H_D = \arccos\sum_{i=1}^{k}\sqrt{\left(n_i / n\right)P_i(\theta)},$$

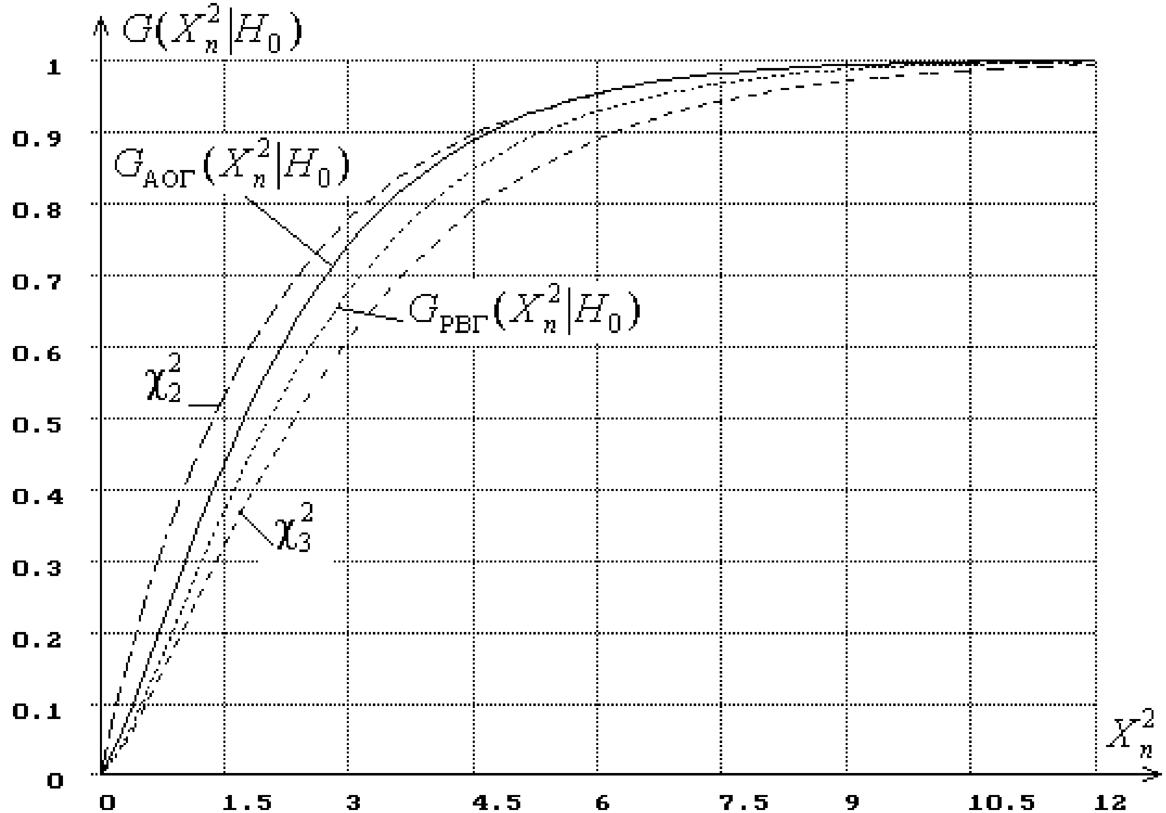
в результате минимизации дивергенции Кульбака-Лейблера (информации Кульбака-Лейблера)

$$S_{KL} = \sum_{i=1}^{k} P_i(\theta)\ln\left[P_i(\theta)/\left(n_i / n\right)\right].$$

Если же оценки параметров искать по точечным выборкам (по исходным негруппированным наблюдениям), то *предельные распределения статистики* X_n^2 *не являются* χ^2_{k-r-1}-*распределениями*.

При вычислении оценок максимального правдоподобия (ОМП) по негруппированным данным эта же статистика распределена как сумма независимых слагаемых $\chi^2_{k-r-1} + \sum_{j=1}^{r} \lambda_j \xi_j^2$, где $\xi_1,\ldots,\xi_r$ – стандартные нормальные случайные величины, независимые одна от другой и от χ^2_{k-r-1}, а $\lambda_1,\ldots,\lambda_r$ – некоторые числа между 0 и 1.

Более того, распределения статистики X_n^2 становятся зависящими от того, как разбивается область определения случайной величины на интервалы.



Распределения $G(X_n^2|H_0)$ при асимптотически оптимальном группировании (АОГ) и при разбиении на интервалы равной вероятности (РВГ) в случае проверки согласия с нормальным распределением с оцениванием двух его параметров и числе интервалов $k = 5$.

При оценивании параметров нормального закона по группированной выборке статистика X_n^2 подчинялась бы в данном случае χ_2^2-распределению. При оценивании по негруппированной распределения статистики $G_{АОГ}(X_n^2|H_0)$ и $G_{РВГ}(X_n^2|H_0)$ существенно отличаются от χ_2^2-распределения. Игнорирование этого факта **приводит к увеличению вероятности ошибок первого рода.**

Критерий Никулина. В данном случае рассматривалась статистика, предложенная Никулиным [Никулин, 1973]. Критерий предусматривает оценивание неизвестных параметров распределения $F(x,\theta)$ методом максимального правдоподобия по негруппированным данным. При этом вектор вероятностей попадания в интервалы $\mathbf{P}=(P_1,\ldots,P_k)^{\mathrm{T}}$ предполагается заданным, и граничные точки интервалов определяют по соотношениям $x_i(\theta)=F^{-1}(P_1+\ldots+P_i)$, $i=\overline{1,(k-1)}$. Предложенная статистика имеет вид

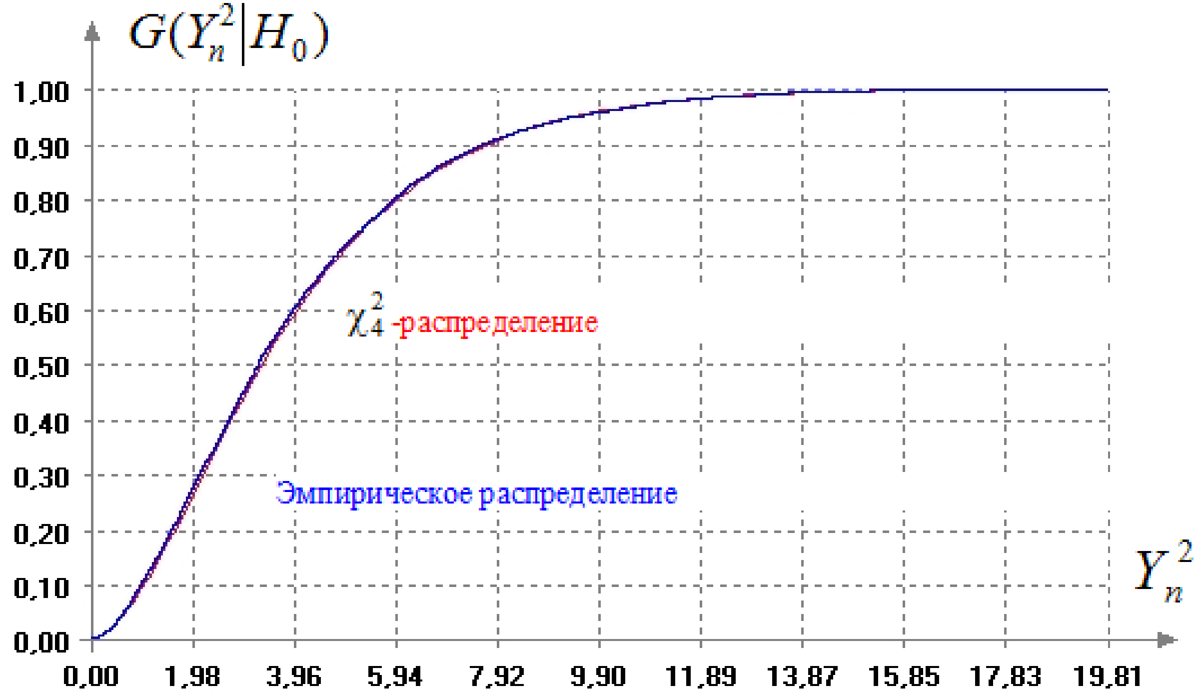
$$Y_n^2 = X_n^2 + n^{-1}a^{\mathrm{T}}(\theta)\boldsymbol{\Lambda}(\theta)a(\theta), \tag{5}$$

где X_n^2 вычисляется по формуле (4); матрица $\boldsymbol{\Lambda}(\theta)=\left\| J(\theta_l,\theta_j)-\sum_{i=1}^{k}\frac{w_{\theta_l i}w_{\theta_j i}}{p_i}\right\|^{-1}$, элементы и размерность которой определяются оцениваемыми компонентами вектора параметров θ; $J(\theta_l,\theta_j)=\int\left(\frac{\partial f(x,\theta)}{\partial\theta_l}\frac{\partial f(x,\theta)}{\partial\theta_j}\right)f(x,\theta)dx$ – элементы информационной матрицы по негруппированным данным; компоненты вектора $a(\theta)$ имеют вид $a_{\theta_l}=w_{\theta_l 1}n_1/P_1+\ldots+w_{\theta_l k}n_k/P_k$;

$$w_{\theta_l i}=-f\left[x_i(\theta),\theta\right]\frac{\partial x_i(\theta)}{\partial\theta_l}+f\left[x_{i-1}(\theta),\theta\right]\frac{\partial x_{i-1}(\theta)}{\partial\theta_l}.$$

Замечательным фактом, отличающим эти критерии, является то, что статистика критерия при справедливости проверяемой гипотезы в пределе подчиняется χ^2_{k-1}-распределению независимо от числа параметров закона, оцененных методом максимального правдоподобия. Во-вторых, мощность критерия, как правило, выше мощности критерия χ^2 Пирсона.

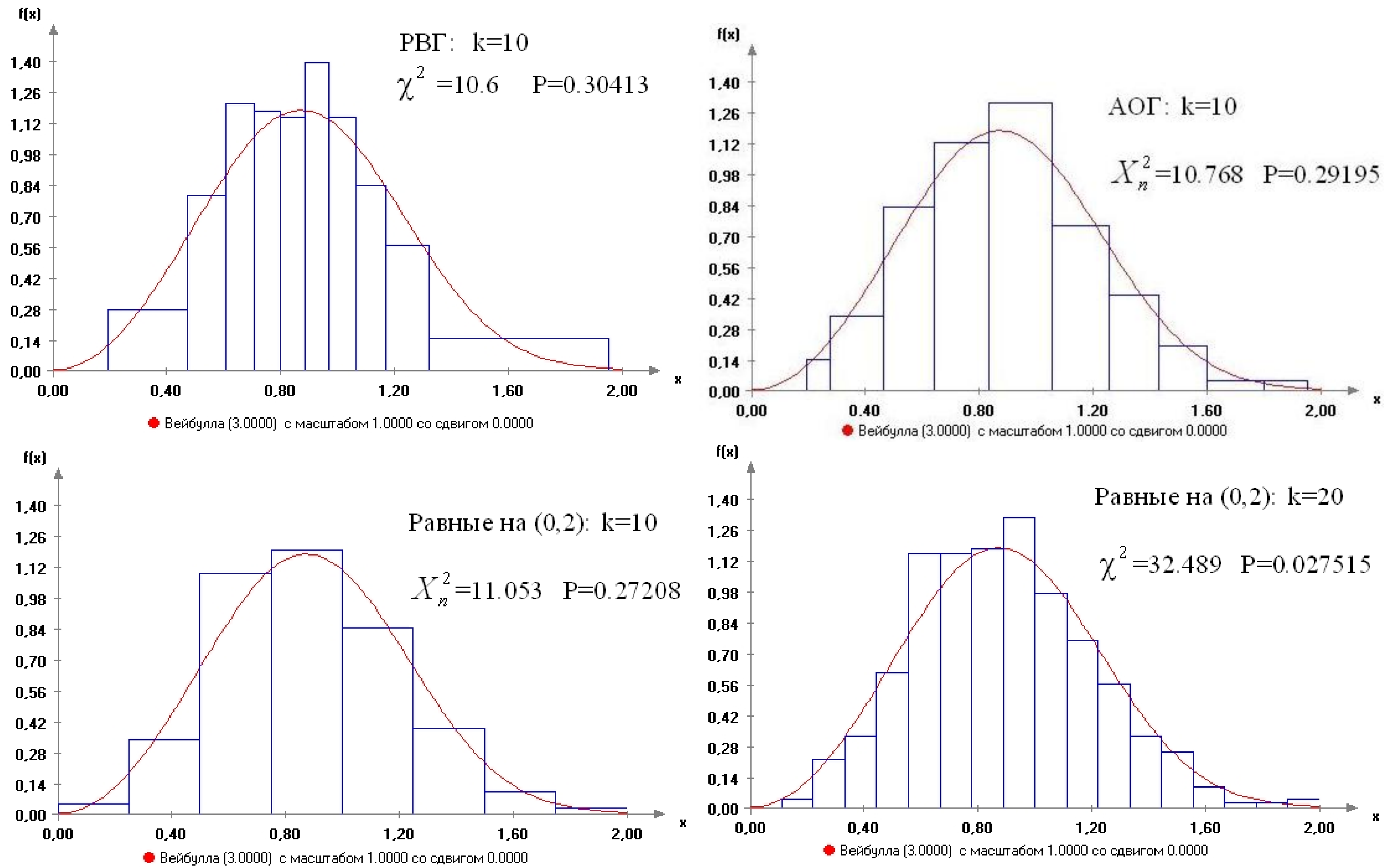
Этот рисунок иллюстрирует близость смоделированного распределения статистики Никулина теоретическому χ^2_{k-1}-распределению



При 5 интервалах и вычислении 2-х параметров логистического распределения

Факторы, влияющие но мощность критериев типа хи-квадрат

1. Способ группирования, способ разбиения области определения случайной величины на интервалы.
2. Выбор числа интервалов.



Результаты применения критерия Пирсона при проверке согласия с распределением Вейбулла при различных способах разбиения на интервалы (n=500, простая гипотеза,выбока сгенерирована в соответствии с распределением Вейбулла)

Зависимость мощности от способа группирования

Практики чаще всего разбивают область, которой принадлежит выборка, на интервалы равной длины, теоретики предпочитают разбиение на интервалы равной вероятности.

С позиции мощности критериев, то есть способности различать, в том числе, близкие альтернативы эти подходы не являются оптимальными.

Как упоминалось, если верна конкурирующая гипотеза H_1 и выборка соответствует закону с распределением $F_1(x,\theta_1)$ с параметром θ_1, то статистика критерия Пирсона при проверке простой гипотезы в пределе подчиняется нецентральному χ^2_{k-1}-распределению с параметром нецентральности

$$\nu = n\sum_{i=1}^{k}\frac{(P_i^1(\theta_1)-P_i(\theta))^2}{P_i(\theta)},$$

где $P_i^1(\theta_1)$ – вероятность попадания в интервал при справедливой гипотезе H_1.

В случае проверки сложной гипотезы, при справедливости H_0 и при условии, что оценки параметров находятся в результате минимизации статистики (4) по этой же самой выборке, статистика X_n^2 асимптотически распределена как χ^2_{k-r-1}, где r – число оцененных по выборке параметров.

Можно показать, разлагая $P_i(\theta_1)$ в соотношении (3) в ряд Тейлора при малых $\delta\theta = \theta_1 - \theta$ и пренебрегая членами высшего порядка, что

$$\nu \approx N \sum_{i=1}^{k} \frac{\left[P_i(\theta) + \nabla^{\mathrm{T}} P_i(\theta)\delta\theta - P_i(\theta)\right]^2}{P_i(\theta)} = N \sum_{i=1}^{k} \frac{\delta\theta^{\mathrm{T}} \nabla P_i(\theta) \nabla^{\mathrm{T}} P_i(\theta)\delta\theta}{P_i(\theta)} =$$

$$= N\delta\theta^{\mathrm{T}} \left(\sum_{i=1}^{k} \frac{\nabla P_i(\theta) \nabla^{\mathrm{T}} P_i(\theta)}{P_i(\theta)} \right) \delta\theta = N\delta\theta^{\mathrm{T}} \mathbf{J}_{\Gamma}(\theta)\delta\theta,$$

где

$$\mathbf{J}_{\Gamma}(\theta) = \sum_{i=1}^{k} \frac{\nabla P_i(\theta) \nabla^{\mathrm{T}} P_i(\theta)}{P_i(\theta)}$$

– информационная матрица Фишера по группированным данным. Мощность критерия χ^2 Пирсона представляет собой неубывающую функцию от ν. Матрица потерь информации, вызванных группированием, $\Delta\mathbf{J} = \mathbf{J}(\theta) - \mathbf{J}_{\Gamma}(\theta)$, где $\mathbf{J}(\theta)$ – информационная матрица Фишера по негрупированным наблюдениям, является неотрицательно определенной, и, следовательно, $\delta\theta^{\mathrm{T}} \Delta\mathbf{J} \delta\theta \ge 0$. Так как $\delta\theta^{\mathrm{T}} \mathbf{J}_{\Gamma}(\theta)\delta\theta = \delta\theta^{\mathrm{T}} \mathbf{J}(\theta)\delta\theta - -\delta\theta^{\mathrm{T}} \Delta\mathbf{J}\delta\theta$, то очевидно, что с ростом потерь информации падает и мощность критерия при близких конкурирующих гипотезах.

Потери от группирования можно уменьшить, решая задачу асимптотически оптимального группирования и подбирая граничные точки так, чтобы $\mathbf{J}_{\Gamma}(\theta)$ стремилась к информационной матрице по негруппированным данным $\mathbf{J}(\theta)$. В случае скалярного параметра эта задача сводится к максимизации количества информации Фишера о параметре по группированной выборке

$$\max_{x_0 < x_1 < \ldots < x_{k-1} < x_k} \sum_{i=1}^{k} \left(\frac{\partial \ln P_i(\theta)}{\partial \theta} \right)^2 P_i(\theta).$$

А в случае вектора параметров в качестве критериев оптимальности могут быть выбраны различные функционалы от информационной матрицы Фишера. Наиболее естественно максимизировать определитель информационной матрицы, т.е. решать задачу

$$\max_{x_0 < x_1 < \ldots < x_{k-1} < x_k} \det \mathbf{J}_{\Gamma}(\theta).$$

Таблица А.24

Оптимальные граничные точки интервалов группирования при проверке сложных гипотез по критериям типа χ^2 (при оценивании математического ожидания нормального распределения, параметра сдвига логарифмически нормальных распределений) и соответствующие значения относительной асимптотической информации *A*

k	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	A
2	0.0	–	–	–	–	–	–	–	–	0.6336
3	–0.6120	0.6120	–	–	–	–	–	–	–	0.8098
4	–0.9816	0.0	0.9816	–	–	–	–	–	–	0.8825
5	–1.2444	–0.3823	0.3823	1.2444	–	–	–	–	–	0.9201
6	–1.4468	–0.6589	0.0	0.6589	1.4468	–	–	–	–	0.9420
7	–1.6108	–0.8744	–0.2803	0.2803	0.8744	1.6108	–	–	–	0.9560
8	–1.7479	–1.0499	–0.5005	0.0	0.5005	1.0499	1.7479	–	–	0.9655
9	–1.8655	–1.1976	–0.6812	–0.2218	0.2218	0.6812	1.1976	1.8655	–	0.9721
10	–1.9682	–1.3246	–0.8338	–0.4047	0.0	0.4047	0.8338	1.3246	1.9682	0.9771

Таблица А.25

Оптимальные вероятности (частоты) при проверке сложных гипотез по критериям типа χ^2 (при оценивании математического ожидания нормального распределения, параметра сдвига логарифмически нормальных распределений) и соответствующие значения относительной асимптотической информации *A*

k	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	A
2	0.5000	0.5000	–	–	–	–	–	–	–	–	0.6336
3	0.2703	0.4594	0.2703	–	–	–	–	–	–	–	0.8098
4	0.1631	0.3369	0.3369	0.1631	–	–	–	–	–	–	0.8825
5	0.1067	0.2444	0.2978	0.2444	0.1067	–	–	–	–	–	0.9201
6	0.0740	0.1810	0.2450	0.2450	0.1810	0.0740	–	–	–	–	0.9320
7	0.0536	0.1373	0.1987	0.2208	0.1987	0.1373	0.0536	–	–	–	0.9560
8	0.0403	0.1066	0.1615	0.1916	0.1916	0.1615	0.1066	0.0403	–	–	0.9655
9	0.0310	0.0845	0.1323	0.1644	0.1756	0.1644	0.1323	0.0845	0.0310	–	0.9721
10	0.0245	0.0662	0.1095	0.1406	0.1572	0.1572	0.1406	0.1095	0.0662	0.0245	0.9771

Таблица А.26

Оптимальные граничные точки интервалов группирования при проверке сложных гипотез по критериям типа χ^2 (при оценивании стандартного отклонения нормального распределения, масштабного параметра логарифмически нормальных распределений) и соответствующие значения относительной асимптотической информации *A*

k	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	A
2	–1.5750	–	–	–	–	–	–	–	–	0.3042
	1.5750	–	–	–	–	–	–	–	–	0.3042
3	–1.4821	1.4821	–	–	–	–	–	–	–	0.6522
4	–2.0249	–1.1865	1.4520	–	–	–	–	–	–	0.7358
	–1.4520	1.1855	2.0249	–	–	–	–	–	–	0.7358
5	–1.9956	–1.1401	1.1401	1.9958	–	–	–	–	–	0.8244
6	–2.3269	–1.6190	–0.9837	1.1190	1.9821	–	–	–	–	0.8588
	–1.9821	–1.1190	0.9837	1.6190	2.3269	–	–	–	–	0.8588
7	–2.3130	–1.6002	–0.9558	0.9558	1.6002	2.3130	–	–	–	0.8943
8	–2.5488	–1.9105	–1.3848	–0.8548	0.9400	1.5897	2.3053	–	–	0.9117
	–2.3053	–1.5897	–0.9400	0.8548	1.3848	1.9105	2.5488	–	–	0.9117
9	–2.5408	–1.9003	–1.3715	–0.8355	0.8355	1.3715	1.9003	2.5408	–	0.9294
10	–2.7223	–2.1286	–1.6602	–1.2260	–0.7634	0.8233	1.3632	1.8939	2.5358	0.9394
	–2.5358	–1.8939	–1.3632	–0.8233	0.7634	1.2260	1.6602	2.1286	2. 7223	0.9394

Таблица А.27

Оптимальные вероятности (частоты) при проверке сложных гипотез по критериям типа χ^2 (при оценивании стандартного отклонения нормального распределения, масштабного параметра логарифмически нормальных распределений) и соответствующие значения относительной асимптотической информации *A*

k	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	A
2	0.0576	0.9424	–	–	–	–	–	–	–	–	0.3042
	0.9424	0.0576	–	–	–	–	–	–	–	–	0.3042
3	0.0692	0.8616	0.0692	–	–	–	–	–	–	–	0.6522
4	0.0214	0.0965	0.8089	0.0732	–	–	–	–	–	–	0.7358
	0.0732	0.8089	0.0965	0.0214	–	–	–	–	–	–	0.7358
5	0.0230	0.1041	0.7458	0.1041	0.0230	–	–	–	–	–	0.8244
6	0.0100	0.0427	0.1099	0.7058	0.1079	0.0237	–	–	–	–	0.8588
	0.0237	0.1079	0.7058	0. 1099	0.0427	0.0100	–	–	–	–	0.8588
7	0.0104	0.0444	0.1148	0.6608	0.1148	0.0444	0.0104	–	–	–	0.8943
8	0.0054	0.0226	0.0550	0.1133	0.6300	0.1177	0.0454	0.0106	–	–	0.9117
	0.0106	0.0454	0.1177	0.6300	0.1133	0.0550	0.0226	0.0054	–	–	0.9117
9	0.0055	0.0232	0.0564	0.1166	0.5966	0.1166	0.0564	0.0232	0.0055	–	0.9294
10	0.0032	0.0134	0.0318	0.0617	0.1125	0.5722	0.1188	0.0573	0.0235	0.0056	0.9394
	0.0056	0.0235	0.0573	0.1188	0.5722	0.1126	0.0617	0.0318	0.0314	0.0032	0.9394

Таблица А.28

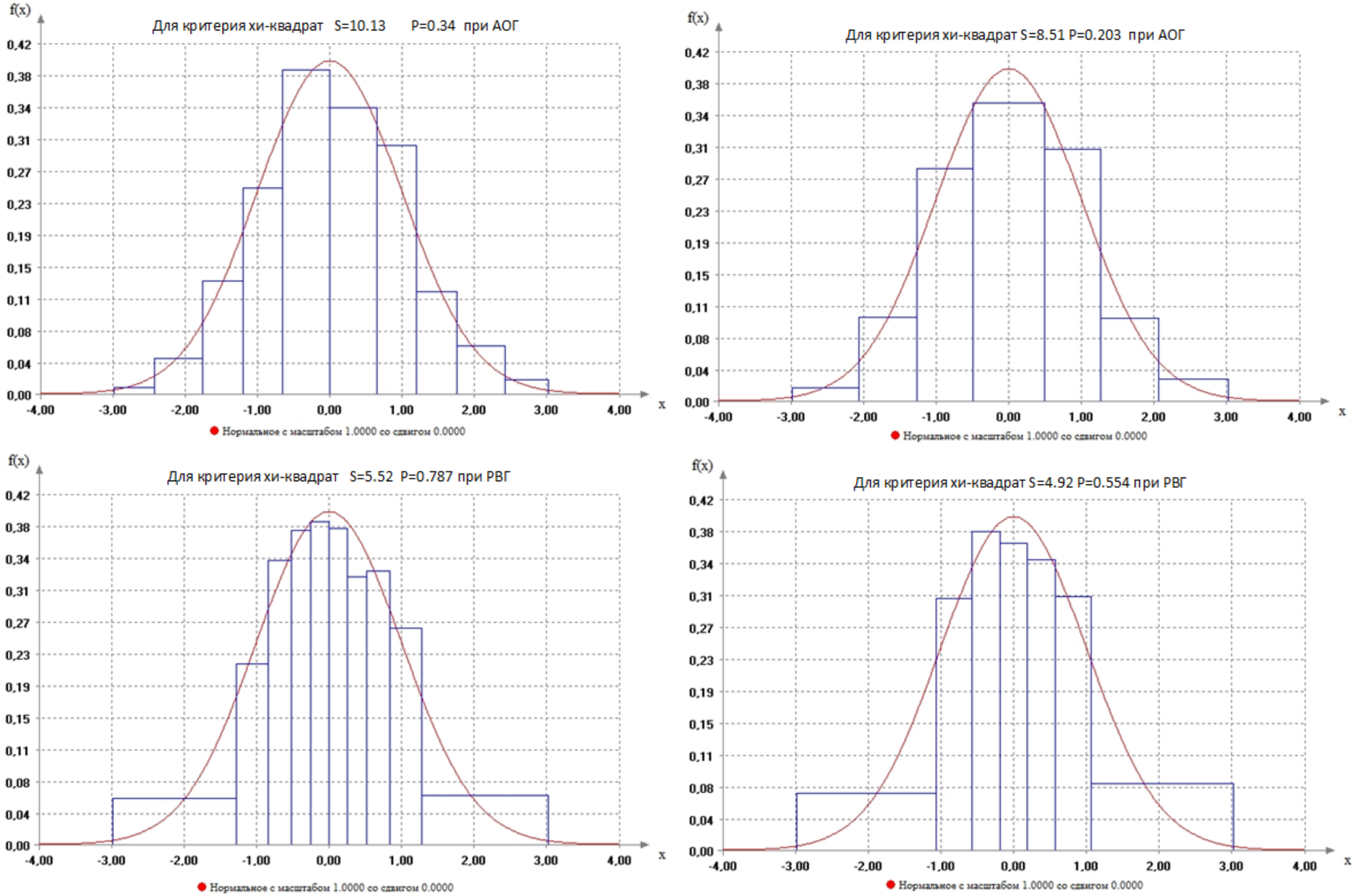
Оптимальные граничные точки интервалов группирования при проверке простых и сложных гипотез по критериям типа χ^2 (при оценивании двух параметров нормального распределения, двух параметров логарифмически нормальных распределений) и соответствующие значения относительной асимптотической информации *A*

k	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	A
3	–1.1106	1.1106	–	–	–	–	–	–	–	–	–	–	–	–	0.4065
4	–1.3834	0.0	1.3834	–	–	–	–	–	–	–	–	–	–	–	0.5527
5	–1.6961	–0.6894	0.6894	1.6961	–	–	–	–	–	–	–	–	–	–	0.6826
6	–1.8817	–0.9970	0.0	0.9970	1.8817	–	–	–	–	–	–	–	–	–	0.7557
7	–2.0600	–1.2647	–0.4918	0.4918	1.2647	2.0600	–	–	–	–	–	–	–	–	0.8103
8	–2.1954	–1.4552	–0.7863	0.0	0.7863	1.4552	2.1954	–	–	–	–	–	–	–	0.8474
9	–2.3188	–1.6218	–1.0223	–0.3828	0.3828	1.0223	1.6218	2.3188	–	–	–	–	–	–	0.8753
10	–2.4225	–1.7578	–1.2046	–0.6497	0.0	0.6497	1.2046	1.7578	2.4225	–	–	–	–	–	0.8960
11	–2.5167	–1.8784	–1.3602	–0.8621	–0.3143	0.3143	0.8621	1.3602	1.8784	2.5167	–	–	–	–	0.9121
12	–2.5993	–1.9028	–1.4914	–1.0331	–0.5334	0.0	0.5334	1.0331	1.4914	1.9028	2.5993	–	–	–	0.9247
13	–2.6746	–2.0762	–1.6068	–1.1784	–0.7465	–0.2669	0.2669	0.7465	1.1784	1.6068	2.0762	2.6746	–	–	0.9348
14	–2.7436	–2.1609	–1.7092	–1.3042	–0.9065	–0.4818	0.0	0.4818	0.9065	1.3042	1.7092	2.1609	2.7436	–	0.9430
15	–2.8069	–2.2378	–1.8011	–1.4150	–1.0435	–0.6590	–0.2325	0.2325	0.6590	1.0435	1.4150	1.8011	2.2378	2.8069	0.9498

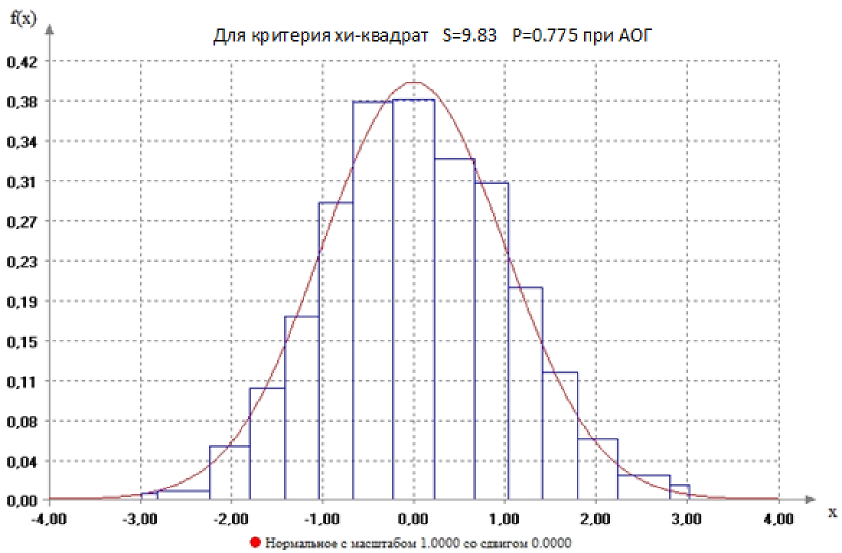
Таблица А.29

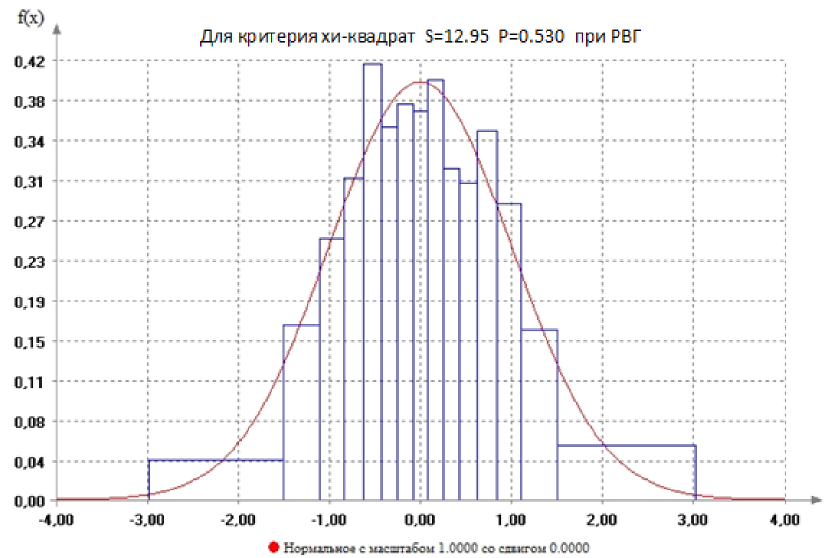
Оптимальные вероятности (частоты) при проверке простых и сложных гипотез по критериям типа χ^2 (при оценивании двух параметров нормального распределения, двух параметров логарифмически нормальных распределений) и соответствующие значения относительной асимптотической информации *A*

k	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	A
3	0.1334	0.7332	0.1334	–	–	–	–	–	–	–	–	–	–	–	–	0.4065
4	0.0833	0.4167	0.4167	0.0833	–	–	–	–	–	–	–	–	–	–	–	0.5527
5	0.0449	0.2004	0.5094	0.2004	0.0449	–	–	–	–	–	–	–	–	–	–	0.6826
6	0.0299	0.1295	0.3406	0.3406	0.1295	0.0299	–	–	–	–	–	–	–	–	–	0.7557
7	0.0197	0.0833	0.2084	0.3772	0.2084	0.0833	0.0197	–	–	–	–	–	–	–	–	0.8103
8	0.0141	0.0587	0.1431	0.2841	0.2841	0.1431	0.0587	0.0141	–	–	–	–	–	–	–	0.8474
9	0.0102	0.0422	0.1009	0.1976	0.2982	0.1976	0.1009	0.0422	0.0102	–	–	–	–	–	–	0.8753
10	0.0077	0.0317	0.0748	0.1438	0.2420	0.2420	0.1438	0.0748	0.0317	0.0077	–	–	–	–	–	0.8960
11	0.0059	0.0243	0.0567	0.1074	0.1823	0.2468	0.1823	0.1074	0.0567	0.0243	0.0059	–	–	–	–	0.9121
12	0.0047	0.0190	0.0442	0.0829	0.1392	0.2100	0.2100	0.1392	0.0829	0.0442	0.0190	0.0047	–	–	–	0.9247
13	0.0037	0.0152	0.0352	0.0652	0.1085	0.1670	0.2104	0.1670	0.1085	0.0652	0.0352	0.0152	0.0037	–	–	0.9348
14	0.0030	0.0124	0.0283	0.0524	0.0862	0.1327	0.1850	0.1850	0.1327	0.0862	0.0524	0.0283	0.0124	0.0030	–	0.9430
15	0.0025	0.0101	0.0232	0.0427	0.0698	0.1066	0.1532	0.1838	0.1532	0.1066	0.0698	0.0427	0.0232	0.0101	0.0025	0.9498



Результаты применения критерия Пирсона при проверке согласия со стандартным нормальным законом при различных способах разбиения на интервалы (n=1000, простая гипотеза,выбока сгенерирована в соответствии со стандартным нормальным законом)

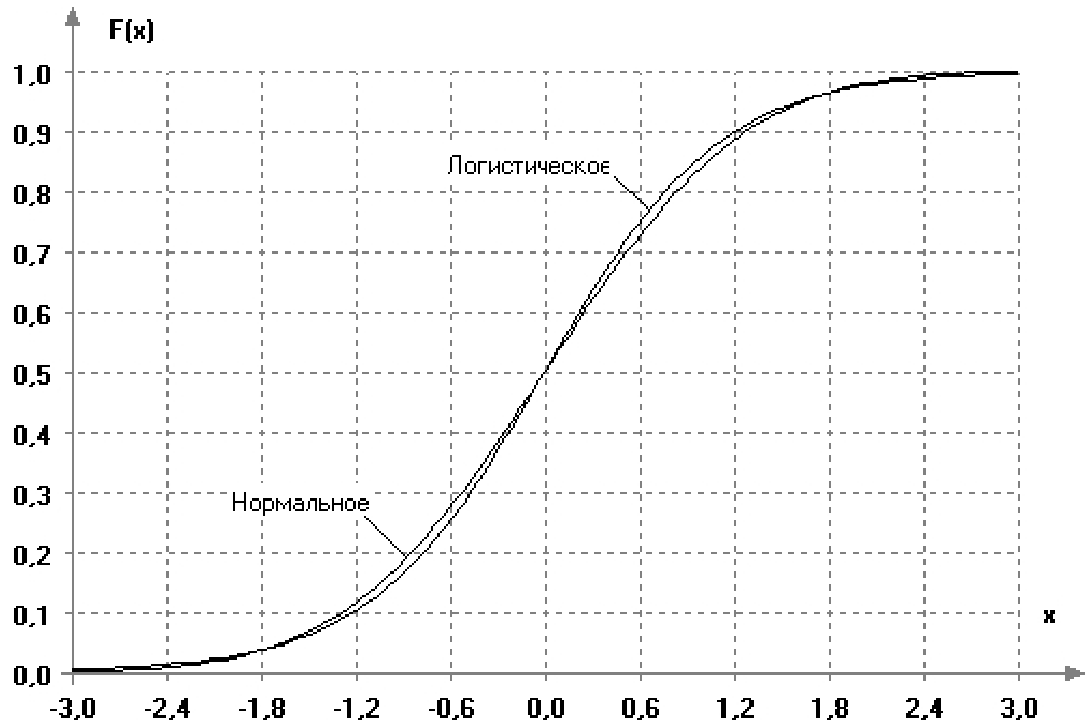




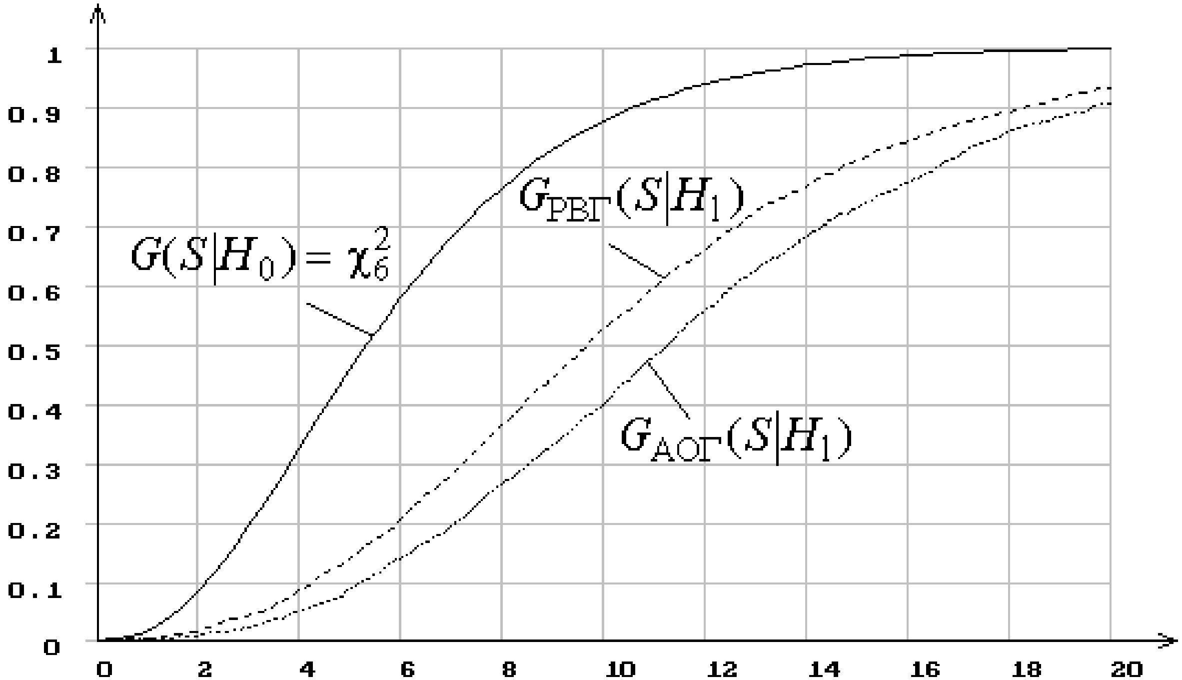
15 интервалов

Результаты применения критерия Пирсона при проверке согласия со стандартным нормальным законом при различных способах разбиения на интервалы (n=1000, простая гипотеза,выбока сгенерирована в соответствии со стандартным нормальным законом)

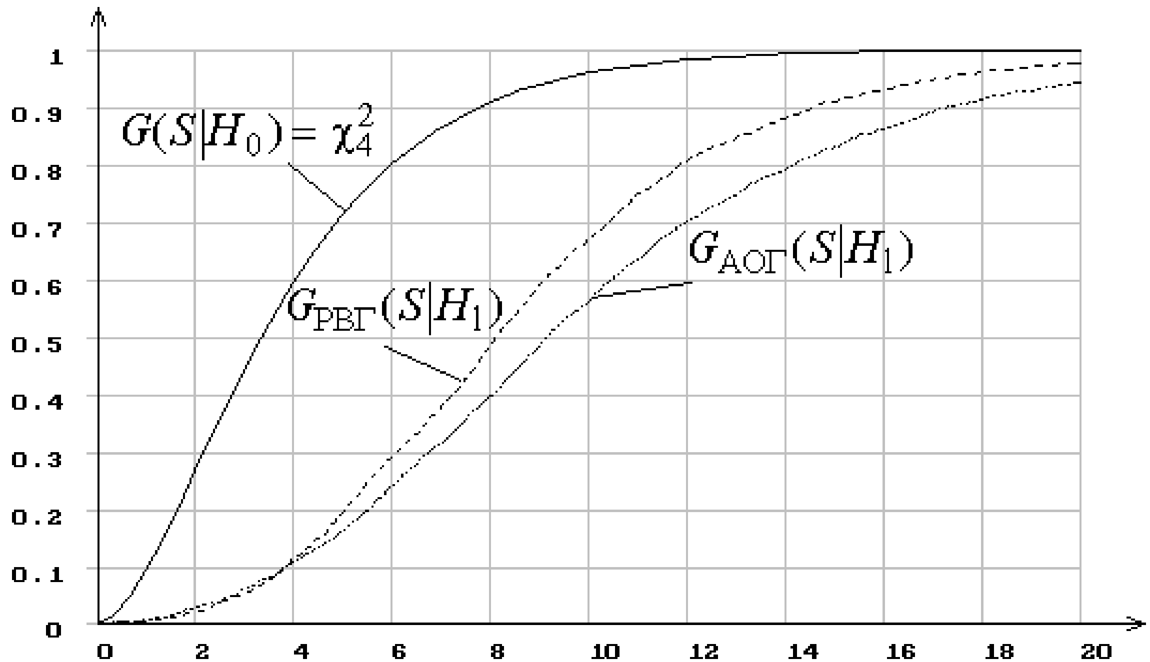
Рассмотрим два близких распределения, нормальное и логистическое, соответствующие конкурирующим гипотезам H_0 и H_1. Эти два закона достаточно близки и трудноразличимы с помощью критериев согласия.



Нормальное и логистическое распределения, соответствующие H_0 и H_1



Распределения статистики χ^2 Пирсона при проверке простой гипотезы H_0.
Число интервалов группирования – 7.



Распределения статистики χ^2 Пирсона при проверке сложной гипотезы H_0.
Число интервалов группирования – 7.

Применение асимптотически оптимального группирования обеспечивает максимальную мощность критериев типа χ^2 относительно близких конкурирующих гипотез!

Р 50.1.033-2001. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа хи-квадрат. – М.: Изд-во стандартов. 2002. – 87 с.

Зависимость мощности от числа интервалов

Исторически было предложено много формул для определения числа интервалов.

Во многих источниках, можно найти упоминание эвристической формулы Старджесса для определения “оптимального” числа интервалов

$$k = \log_2 N + 1 = 3{,}3 \lg N + 1.$$

Для определения “оптимального” числа интервалов рекомендуют формулу Брукса и Каррузера

$$k = 5 \lg N.$$

Рекомендуют соотношение

$$k = \sqrt{N}.$$

Для равновероятных интервалов их число устанавливают порядка

$$k \approx 4 \sqrt[5]{2} (N / t)^{0.4},$$

где t – квантиль стандартного нормального распределения для заданного уровня значимости. В ряде работ приводят модификации данной формулы. Предлагают значение

$$k = 4 \lg N,$$

и её дальнейшее развитие

$$k = 5 \lg N - 5.$$

Предлагают соотношение

$$k = \frac{4}{\aleph} \lg \frac{N}{10},$$

где $\aleph = 1 \Big/ \sqrt{\mu_4 / \sigma^4}$ – значение контрэксцесса, μ_4 – четвертый центральный момент, σ – среднее квадратичное отклонение.

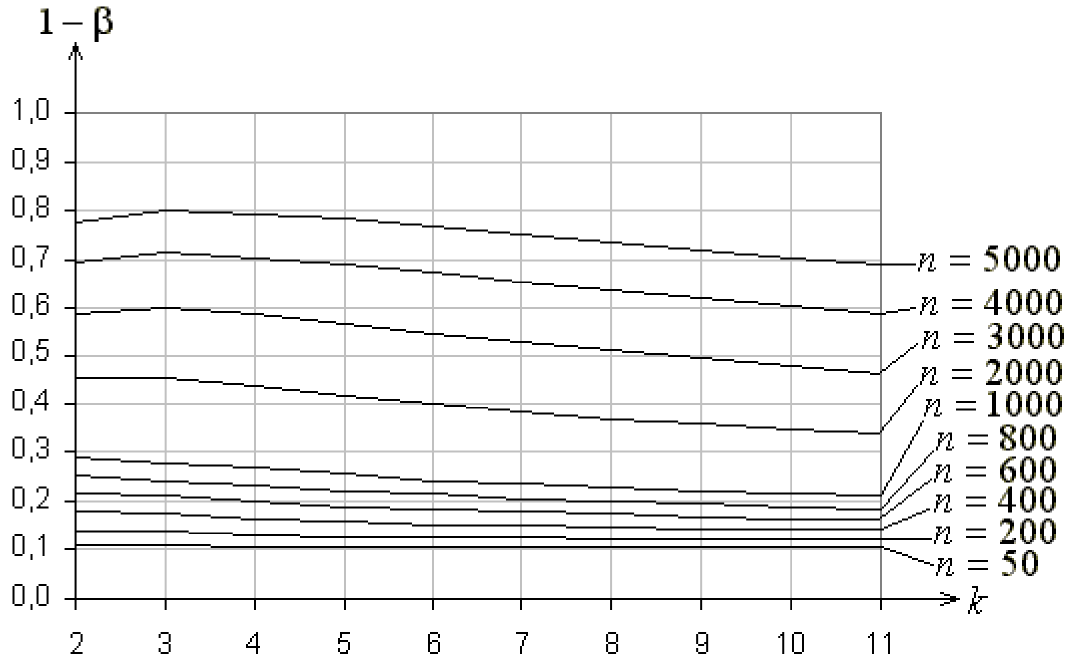
При больших объемах выборок N разброс значений k, задаваемых различными формулами, достаточно велик. Поэтому на практике при выборе числа интервалов больше руководствуются тем, чтобы в интервалы попадало число наблюдений не менее 5-10. Так, например, в рекомендациях ВНИИМетрологии [34] в зависимости от N предлагают следующие значения k:

N	k
40–100	7–9
100–500	8–12
500–1000	10–16
1000–10000	12–22

Все вышеперечисленные рекомендации опирались на предположение, что k следует выбирать таким образом, чтобы вид гистограммы был как можно ближе к плавной кривой плотности распределения генеральной совокупности. Ченцовым Н.Н. показано, что уклонение гистограммы от плотности распределения в лучшем случае имеет порядок $1/\sqrt[3]{N}$, достигаемый при числе интервалов k порядка $\sqrt[3]{N}$.

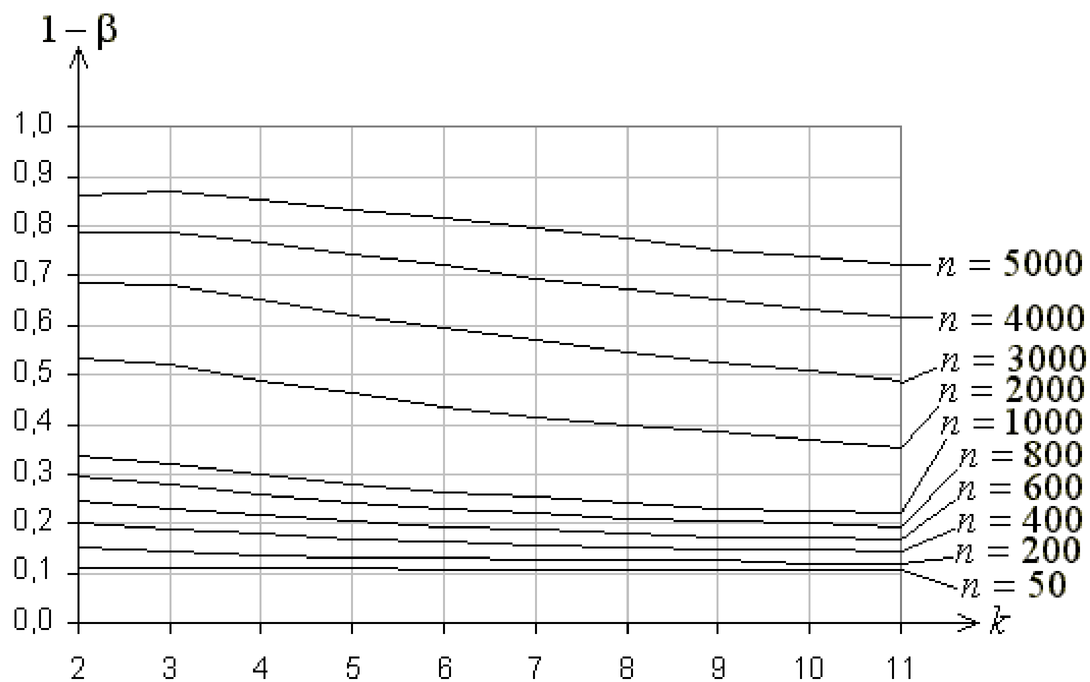
Очевидно, что “оптимальное” значение k зависит не только от объема выборки, но и от вида закона распределения и от способа группирования.

На рисунке в зависимости от числа k равновероятных интервалов при различных n представлены функции мощности критерия χ^2 Пирсона при проверке простой гипотезы о согласии с экспоненциальным законом (H_0: $f(x)=\theta\exp\{-\theta x\}$ при $\theta=1$; H_1: $f(x)=\theta\exp\{-\theta x\}$ при $\theta=1.05$).



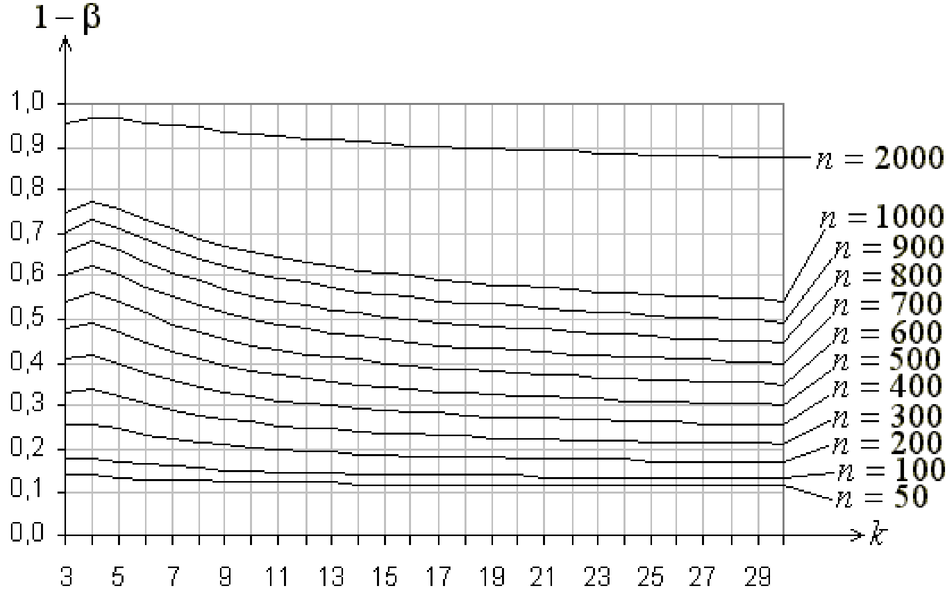
Функции мощности критерия χ^2 Пирсона при проверке простой гипотезы
о согласии с экспоненциальным законом при равновероятном группировании

Аналогичные функции при использовании асимптотически оптимального группирования. И в том и в другом случае с ростом k мощность падает, но в случае асимптотически оптимального группирования она выше, чем при равновероятном.

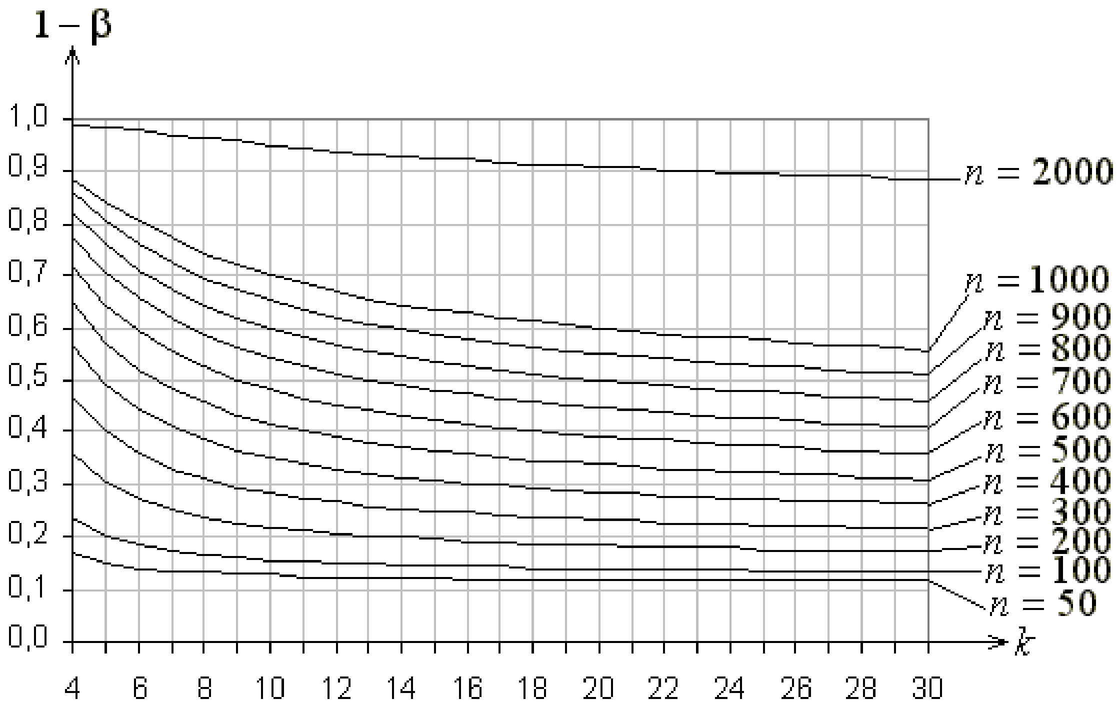


Функции мощности критерия χ^2 Пирсона при проверке простой гипотезы о согласии с экспоненциальным законом при асимптотически оптимальном группировании

Здесь и на следующем рисунке представлены функции мощности критерия χ^2 Пирсона при проверке простых и сложных гипотез о согласии с нормальным законом H_0: $f(x) = \frac{1}{\theta_1\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta_0)^2}{2\theta_1^2}\right\}$ при логистическом законе в качестве альтернативы H_1: $f(x) = \frac{\pi}{\theta_1\sqrt{3}} \exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\} \Big/ \left[1+\exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\right]^2$ при значениях параметров $\theta_0 = 0$, $\theta_1 = 1$.

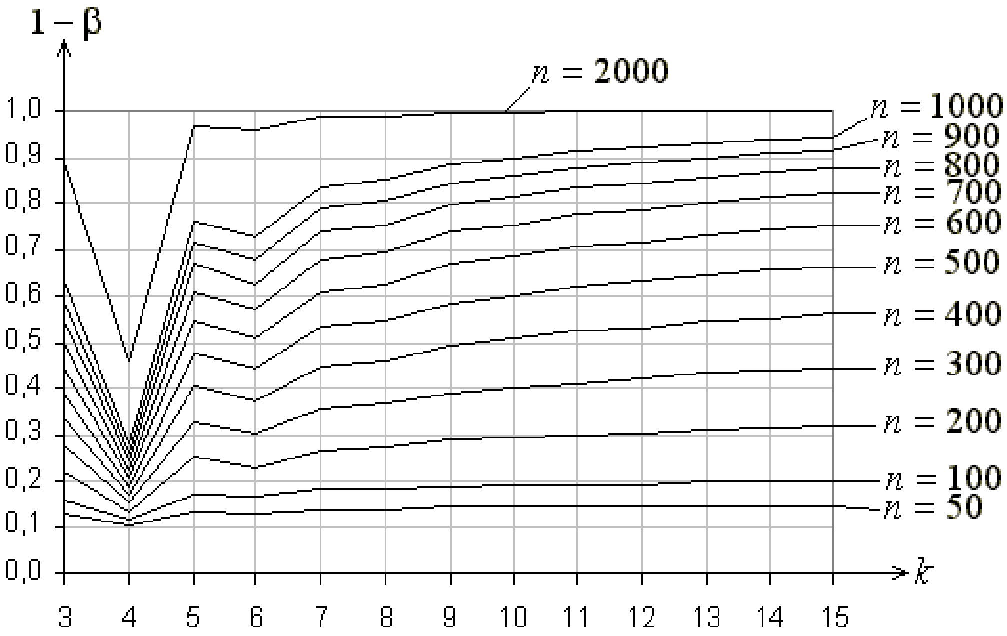


Функции мощности критерия χ^2 Пирсона при проверке простой гипотезы о согласии с нормальным законом при **равновероятном** группировании и альтернативе, соответствующей логистическому закону

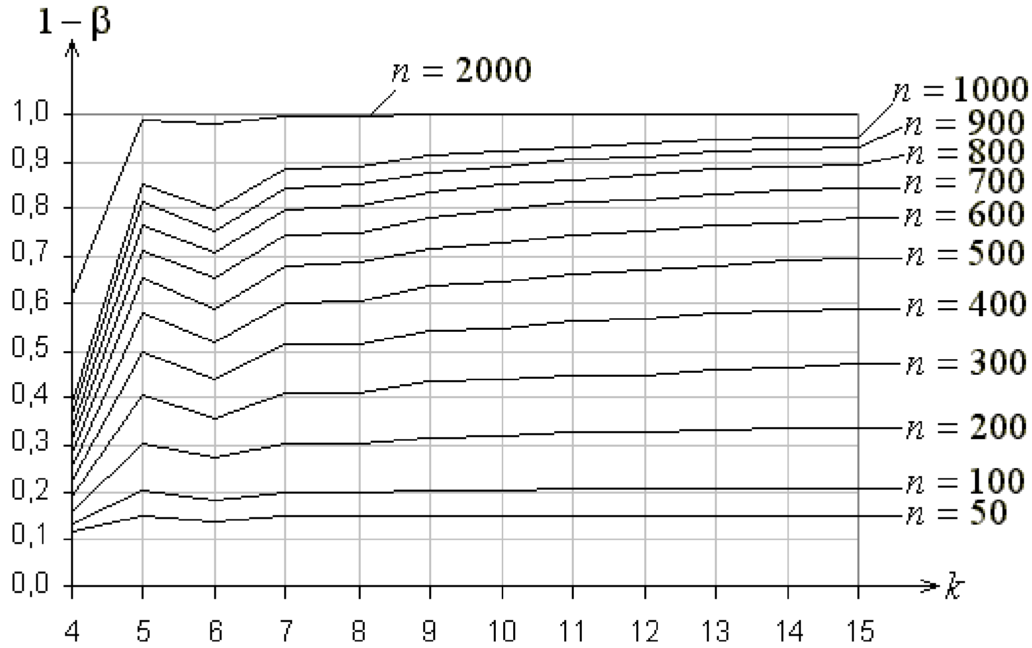


Функции мощности критерия χ^2 Пирсона при проверке сложной гипотезы о согласии с нормальным законом при **равновероятном** группировании и альтернативе, соответствующей логистическому закону

Аналогично при **асимптотически оптимальном** группировании

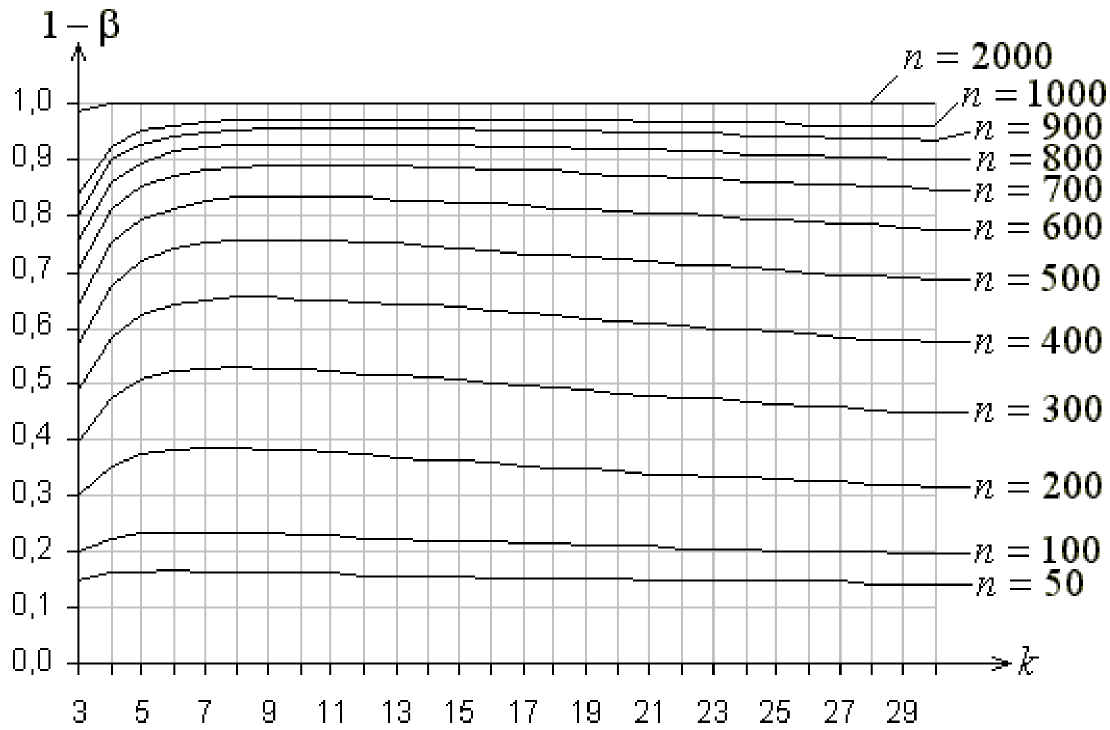


Функции мощности критерия χ^2 Пирсона при проверке **простой** гипотезы
о согласии с нормальным законом при асимптотически оптимальном группировании
и альтернативе, соответствующей логистическому закону



Функции мощности критерия χ^2 Пирсона при проверке **сложной** гипотезы о согласии с нормальным законом при **асимптотически оптимальном** группировании и альтернативе, соответствующей логистическому закону

Функция мощности критерия типа χ^2 Никулина, как следует из рисунка на области значений k, содержащей максимальное значение мощности, является выпуклой вверх функцией.



Функции мощности критерия типа χ^2 Никулина при проверке сложной гипотезы о согласии с нормальным законом при **равновероятном** группировании и альтернативе, соответствующей логистическому закону

Выводы

Таким образом, выбирая число интервалов в критериях типа χ^2, следует понимать, что увеличение их числа не всегда приводит к росту мощности критерия!

В то же время, если заданы конкретная конкурирующая гипотеза и объем выборки ***n***, можно выбрать оптимальное число интервалов ***k*** и способ группирования так, чтобы максимизировать мощность критерия.