МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

О.Ю.Пелевин

МЕТОДИЧЕСКАЯ РАЗРАБОТКА

по курсу «Теория вероятностей и математическая статистика»

для студентов физического факультета,
специальность «Телекоммуникации»

ВАРИАЦИОННЫЕ РЯДЫ, ВЫБОРОЧНЫЕ ОЦЕНКИ, ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

Ростов-на-Дону

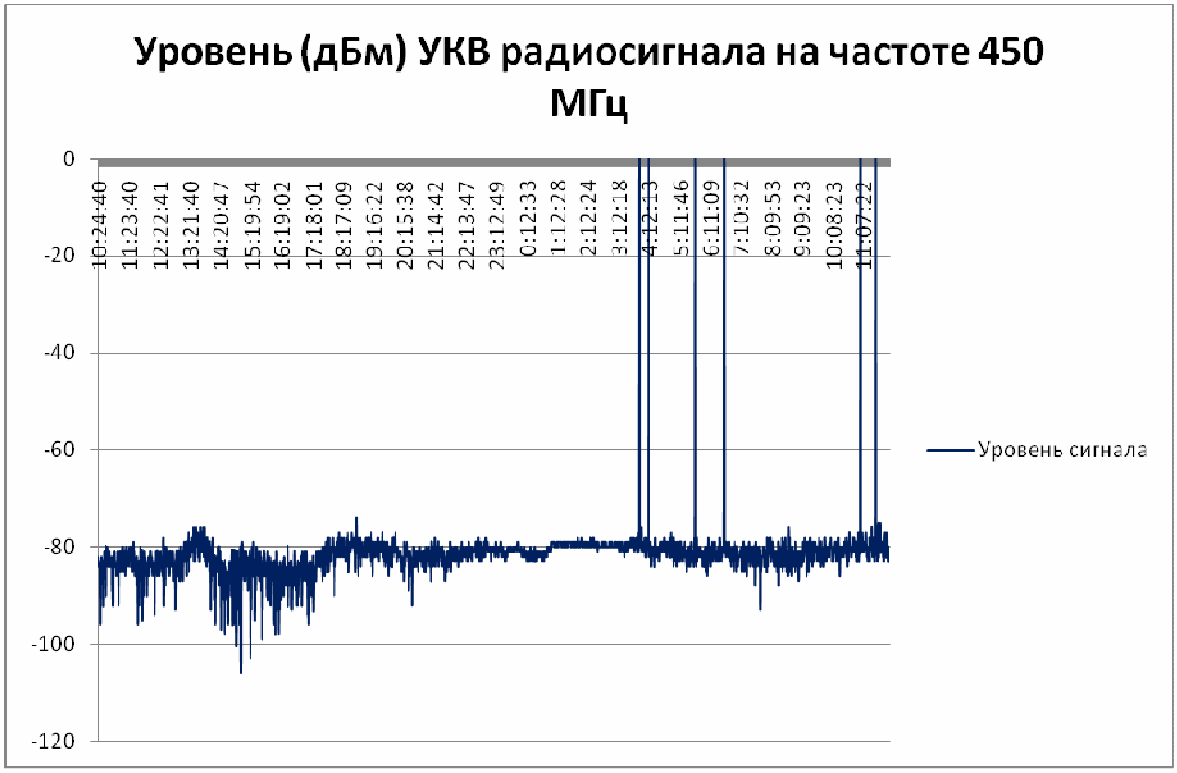
2006

Генеральная совокупность и выборка

Совокупность значений некоторого признака объекта называется генеральной совокупностью. Основная задача математической статистики – выяснение вероятностных свойств совокупности – распределения, и числовых характеристик.

Полное исследование генеральной совокупности практически невозможно или неэкономно. Обычно из генеральной совокупности делают выборку, то есть исследуют только некоторые ее объекты. С помощью выборки оценивают генеральную совокупность по вероятностным свойствам. Чтобы оценки были достоверными, выборка должна быть **представительной**, те ее вероятностные свойства должны совпадать или быть близкими к свойствам генеральной совокупности.

Представительную выборку можно получить, если выбирать объекты для исследований случайно, те гарантировать всем объектам генеральной совокупности одинаковую вероятность подвергнуться исследованию.



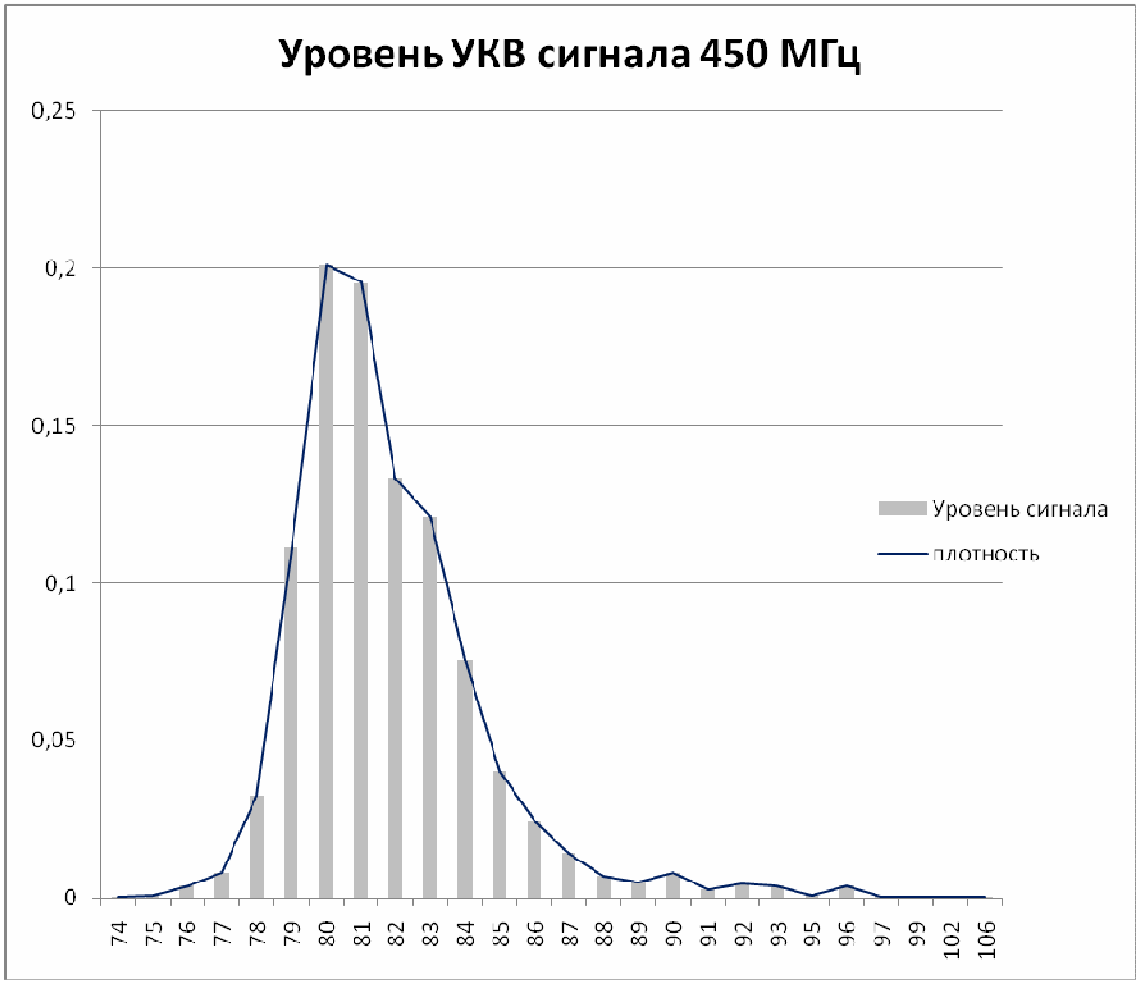
Вариационный ряд

Выбор объекта из генеральной совокупности и измерение значения признака называется статистическим наблюдением. Упорядоченные по возрастанию или убыванию результаты наблюдений наглядно представляются в виде вариационных рядов или таблиц, где в

первом столбце записываются всевозможные значения (варианты) x_i генеральной совокупности (или случайной величины), а во втором ряду – числа n_i/n – относительные частоты (частости) появления i-го значения.

Очевидно, что если точки гистограммы соединить плавной кривой, то эта кривая в первом приближении будет представлять график плотности вероятности случайной величины X.

Если число опытов увеличивать и выбирать более мелкие группы в статистической совокупности, то гистограмма будет все более приближаться к плотности вероятности случайной величины. На нижеследующем рисунке в качестве примера показана гистограмма значений уровня радиосигнала, построенная по его временному ряду за время порядка одних суток.



Эмпирическая функция распределения

Каждая генеральная совокупность имеет свою функцию распределения, которая обычно неизвестна. По выборке можно найти эмпирическую функцию распределения $F^*(x)$, где

на основании теоремы Бернулли вместо вероятностей p_i берутся частости n_i/n. Процесс нахождения эмпирической функции распределения аналогичен процессу нахождения функции распределения дискретной случайной величины. Значениями эмпирической функции распределения являются так называемые накопленные частости.

Числовые характеристики выборки

Выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

x_i – элементы выборки, n- объем выборки.

Если составлен вариационный ряд, то

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i$$

m – количество вариантов, x_i – варианты случайной величины, n- объем выборки.

Выборочная дисперсия

$$\overline{S^2} = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2$$

$$\overline{S^2} = \frac{1}{n}\sum_{i=1}^{m} x_i^2 n_i - \left(\overline{x}\right)^2$$

Стандартное отклонение

$$\overline{S} = \sqrt{\overline{S^2}}$$

Мода

Если вариационный ряд составлен по значениям генеральной совокупности, то модой выборки называется значение, имеющее максимальную частоту.

Медиана

Медианой выборки является значение серединного элемента вариационного ряда.

Оценки параметров распределений и их свойства

Одним из основных предметов для статистических выводов является построение вероятностных моделей и, далее, получение оценок параметров этих моделей (математического ожидания и дисперсии).

Существует два типа оценок – точечные и доверительные.

Точечная оценка определяется одним числом, которое должно быть как можно ближе к оцениваемому параметру.

Доверительной называют статистическую оценку, которая определяется двумя числами – концами интервала, покрывающего оцениваемый параметр.

За оценку математического ожидания и дисперсии можно принять соответствующие выборочные характеристики.

Выборочное среднее (арифметическое)

Рассмотрим последовательность независимых испытаний. Проводя каждое последовательное i-е испытание, мы наблюдаем реализацию случайной величины X_i. После n испытаний мы получаем выборочные значения n случайных величин $X_1 \ldots X_n$, каждая из которых имеет одно и то же распределение. Выборочное значение определяется как численное усреднение наблюдений:

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

Здесь, в первую очередь, необходимо отметить, что выборочное среднее – это функция случайных величин X_i, и следовательно, само является случайной величиной, в отличие от математического ожидания случайной величины, которое является числом.

Поэтому для выборочного среднего можно найти математическое ожидание и дисперсию.

По известным свойствам математического ожидания и дисперсии получаем:

$$M\left[\overline{X}\right] = \frac{1}{n}\sum_{i=1}^{n} M\left[X_i\right] = \frac{1}{n}\left(M\left[X\right]n\right) = M\left[X\right]$$

$$D\left[\overline{X}\right] = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D\left[X_i\right] = \frac{1}{n^2} nD\left[X_i\right] = \frac{D\left[X\right]}{n}$$

При неограниченном росте n дисперсия выборочного среднего убывает, и при $n \to \infty$ X_n сходится к $M[X]$.

Свойства оценок

Выборочной оценкой неизвестного параметра θ будем называть произвольную функцию

$\tilde{\theta}_n = f(x_1,...,x_n)$

Состоятельность

Оценку $\tilde{\theta}_n$ параметра θ называют состоятельной, если $\tilde{\theta}_n$ при $n \to \infty$ сходится по вероятности к θ. Другими словами,

для $\forall \varepsilon, \eta$ существует такое N, что при $n>N$

$$P\left(\left|\tilde{\theta}_n - \theta\right| < \varepsilon\right) > 1-\eta, \eta \to \infty \quad \text{при } n \to \infty$$

Смещение

Оценку $\tilde{\theta}_n$ параметра θ называют несмещенной, если ее математическое ожидание равно оцениваемому параметру, то есть

$$M\left[\tilde{\theta}_n\right] = \theta$$

Это свойство означает, что если пользоваться несмещенной оценкой, то в одних случаях можно завысить искомый параметр, в других – занизить, но в среднем мы будем «попадать в цель».

Эффективность

Несмещенную оценку $\tilde{\theta}_n$, которая имеет наименьшую выборочную дисперсию среди всех возможных несмещенных оценок параметра θ, вычисленных по выборкам одного и того же объема, называют эффективной оценкой.

Выборочное среднее $\overline{X_n}$, вычисляемое по n независимым наблюдениям над случайной величиной X, которая имеет математическое ожидание μ и ограниченную дисперсию $\sigma^2 < \infty$ является несмещенной и состоятельной оценкой математического ожидания. В случае нормального распределения эта оценка еще и эффективна.

Если случайная выборка состоит из n независимых наблюдений над случайной величиной X, которая имеет математическое ожидание μ и ограниченную дисперсию $\sigma^2 < \infty$ то выборочная дисперсия является смещенной состоятельной оценкой дисперсии.

Несмещенная и состоятельная оценка дисперсии (исправленная выборочная дисперсия) равна

$$S^{*2} = \frac{n}{n-1}\overline{S^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Оценки $S^{*2}, \overline{S^2}$ не являются эффективными.

При больших *n* $S^{*2}, \overline{S^2}$ очень близки. Однако при *n=1* $\overline{S^2} = 0$, а S^{*2}, напротив, не определена. Поскольку дисперсия – это мера разброса случайной величины, то невозможно оценить этот разброс только по данным одного наблюдения. Поэтому оценка $\overline{S^2} = 0$ полностью нелогична. С другой стороны, оценка по двум наблюдениям

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{\left(X_1 - X_2\right)^2}{2}$$

дает четкое представление о разбросе значений *X*.

Доверительные интервалы

При малом объеме выборки точечная оценка может значительно отличаться от оцениваемого параметра. Чтобы избежать грубых ошибок при малом объеме выборки вводят интервальную оценку, которая определяется двумя числами – границами интервала.

Надежность

Надежностью (доверительной вероятностью) оценки параметра θ по $\tilde{\theta}_n$ называют вероятность γ, с которой осуществляется неравенство $\left|\theta - \tilde{\theta}_n\right| < \delta$

$$P\left(\left|\tilde{\theta}_n - \theta\right| < \delta\right) = \gamma$$

На практике обычно стараются сделать так, чтобы надежность γ была как можно ближе к 1:

$\gamma=1-\alpha$, где α называется уровнем значимости. Интервал ($\tilde{\theta}_n$-δ, $\tilde{\theta}_n$+ δ) называют доверительным интервалом для неизвестного параметра θ, соответствующий доверительной вероятности (надежности) γ.

Построение доверительного интервала для математического ожидания нормально распределенной генеральной совокупности при известной дисперсии σ^2

Пусть выборка $X_1,\ldots,X_n$ состоит из *n* независимых нормально распределенных с параметрами μ и σ случайных величин, причем σ известно, а величину μ оцениваем выборочным средним:

$$\mu \approx \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Оценим точность этого приближенного равенства, то есть укажем доверительные пределы $\tilde{\theta}_n$, в которых практически достоверно лежит число μ.

Величина $\overline{X}$ распределена (по центральной предельной теореме) нормально с математическим ожиданием μ и среднеквадратичным отклонением $\frac{\sigma}{\sqrt{n}}$.

Нормированное отклонение $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ распределено также нормально с параметрами 0 и 1.

Поэтому вероятность любого отклонения может быть вычислена по формуле

$P(\left|\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\right| < \delta)$=2Φ(δ)-1 или

$P(\left|\overline{X}-\mu\right| < \frac{\delta\sigma}{\sqrt{n}})$=2Φ(δ)-1,

где Φ(δ) – функция Лапласа.

Зададим надежность γ так, чтобы событие с вероятностью γ можно было считать практически достоверным, то есть

$P(\left|\overline{X}-\mu\right| < \frac{\delta\sigma}{\sqrt{n}})=\gamma$

Теперь мы можем найти δ как корень уравнения

$2\Phi(\delta)-1 = \gamma$

Пусть δ_γ - корень этого уравнения. Тогда находим, что

$$P((\overline{X}-\delta_\gamma\frac{\sigma}{\sqrt{n}}) < \mu < (\overline{X}+\delta_\gamma\frac{\sigma}{\sqrt{n}})) = \gamma$$

Таким образом, с вероятностью γ интервал со случайными концами

$$\overline{X}-\delta_\gamma\frac{\sigma}{\sqrt{n}}, \overline{X}+\delta_\gamma\frac{\sigma}{\sqrt{n}}$$

покрывает неизвестное значение $\mu=M[X]$. Этот интервал является доверительным интервалом для μ, соответствующим надежности γ. Получена так называемая классическая оценка.

Построение доверительного интервала для математического ожидания нормально распределенной генеральной совокупности при неизвестной дисперсии σ^2

Ранее мы получили, что нормированное отклонение $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ имеет нормальное распределение с параметрами 0 и 1. Как поступить, когда дисперсия σ неизвестна и является случайной величиной, определяемой по выборке?

Распределение χ^2

Пусть $\chi^2 = \sum_{i=1}^{n} x_i^2$, где X_i – независимые случайные величины, распределенные по стандартному нормальному закону (с нулевым математическим ожиданием и единичной дисперсией). Тогда можно показать, что величина χ^2 имеет следующее распределение:

$\mathrm{K}_n(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{n/2-1} e^{-x/2}$ - распределение Пирсона, или χ^2.

Здесь

$\Gamma(p) = \int_0^{\infty} e^{-z} z^{p-1} dz$ - гамма-функция.

Отметим, что $M[\chi^2]=n$, $D[\chi^2]=2n$.

При $n \to \infty$

$P(\chi^2_n<x) \to \Phi(\sqrt{2x}) - \Phi(\sqrt{2n-1})$,

поэтому таблицы распределения χ^2_n приводятся обычно лишь для небольших значений n (до 30).

Можно показать, что в выборке из нормальной генеральной совокупности с характеристиками μ и σ смещенная выборочная дисперсия в виде параметра $\frac{n\overline{S^2}}{\sigma^2}$ следует закону χ^2_{n-1} (с $n-1$ степенями свободы). При этом выборочное среднее и выборочная дисперсия взаимно независимы. Таким образом, мы можем исключить неизвестную нам дисперсию σ, переходя к отношению

$$t = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\sqrt{n-1}\frac{1}{\sqrt{\frac{n\overline{S^2}}{\sigma^2}}} = \frac{\overline{X}-\mu}{\overline{S}}\sqrt{n-1}$$

Теперь требуется определить распределение величины t.

t – это показатель Стьюдента и в общем виде определяется следующим образом:

$t = \frac{z\sqrt{k}}{\sqrt{v}}$, где величины z и v независимы, z распределена нормально стандартно, а v следует закону χ^2_{k-1} (с $k-1$ степенями свободы). При таких условиях плотность вероятности величины t имеет вид

$$s_k(x) = \mathrm{B}_k(1+\frac{x^2}{k})^{-\frac{k+1}{2}}, \text{ где } \mathrm{B}_k = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\sqrt{\pi k}} \quad -$$

закон Стьюдента с k степенями свободы.

При увеличении n распределение $s_k(x)$ стремится к нормальному закону. Справедливо утверждение:

$$\lim_{k\to\infty} s_k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Таким образом, мы получили, что величина $\frac{\overline{X}-\mu}{\overline{S}}\sqrt{n-1}$ распределена по закону Стьюдента с $n-1$ степенями свободы.

Теперь мы можем найти вероятность

$P(\left|\frac{\overline{X}-\mu}{\overline{S}}\sqrt{n-1}\right| < \delta_{\gamma,n-1}) = \gamma$, где γ-надежность.

Здесь δ зависит уже не только от γ, но и от n.

Раскрывая модуль, запишем вероятность в следующем виде:

$$P(\overline{X} - \delta_{\gamma,n-1}\frac{\overline{S}}{\sqrt{n-1}} < \mu < \overline{X} + \delta_{\gamma,n-1}\frac{\overline{S}}{\sqrt{n-1}}) = \gamma$$

Следовательно, доверительный интервал, отвечающий надежности γ, будет иметь вид:

$$(\overline{X} - \delta_{\gamma,n-1}\frac{\overline{S}}{\sqrt{n-1}}, \overline{X} + \delta_{\gamma,n-1}\frac{\overline{S}}{\sqrt{n-1}})$$

Напомним, что в данном выражении фигурирует среднеквадратичное отклонение, соответствующее смещенной дисперсии.

Значение $\delta_{\gamma,n-1}$ определяется из уравнения

$S_k(\delta_{\gamma,n-1)}) - S_k(\delta_{\gamma,n-1}) = \gamma$ или с учетом $S(-x) = 1 - S(x)$

$2\,S_k(\delta_{\gamma,n-1}) = 1+\gamma$, где

$S_k(x) = \int_{-\infty}^{x} s_k(z)dz$ - интегральная функция распределения Стьюдента, $k=n-1$.

Тогда $\delta_{\gamma,n-1}$ – это корень уравнения

$S_k(\delta_{\gamma,n-1})=(1+\gamma)/2$. Так, для надежности γ=0,99 получим

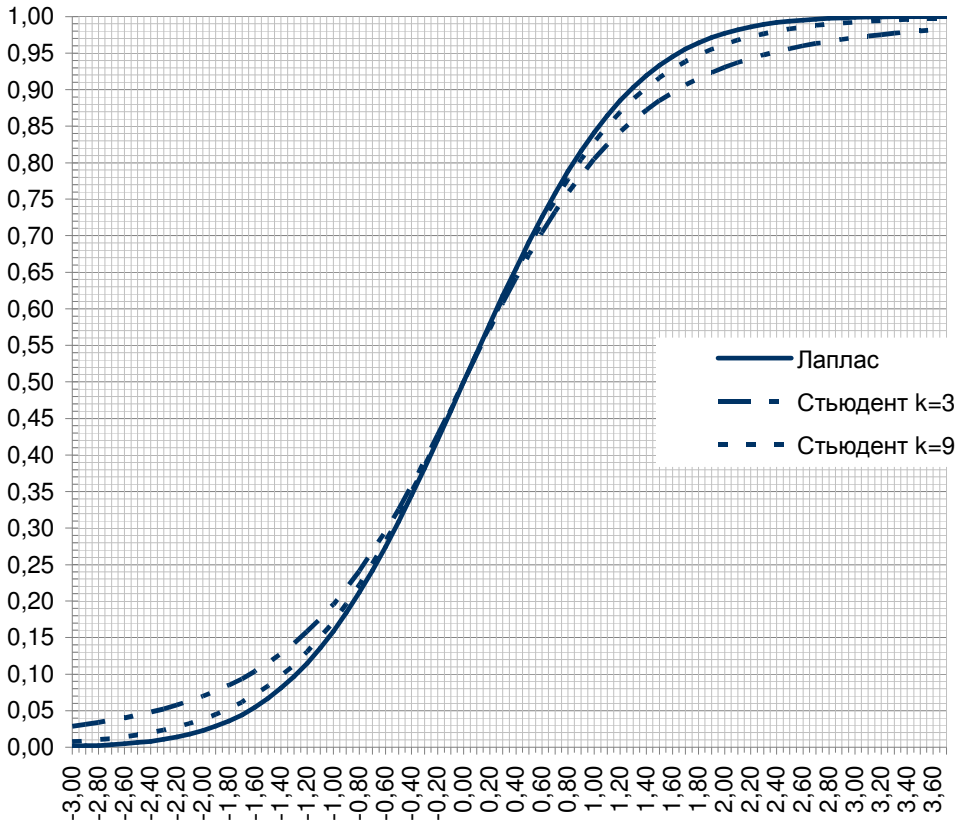
$S_k(\delta_{\gamma,n-1})=(1+0,99)/2=0,995$.

В таблицу сведены некоторые значения корней данного уравнения для двух значений надежности и трех значений степеней свободы:

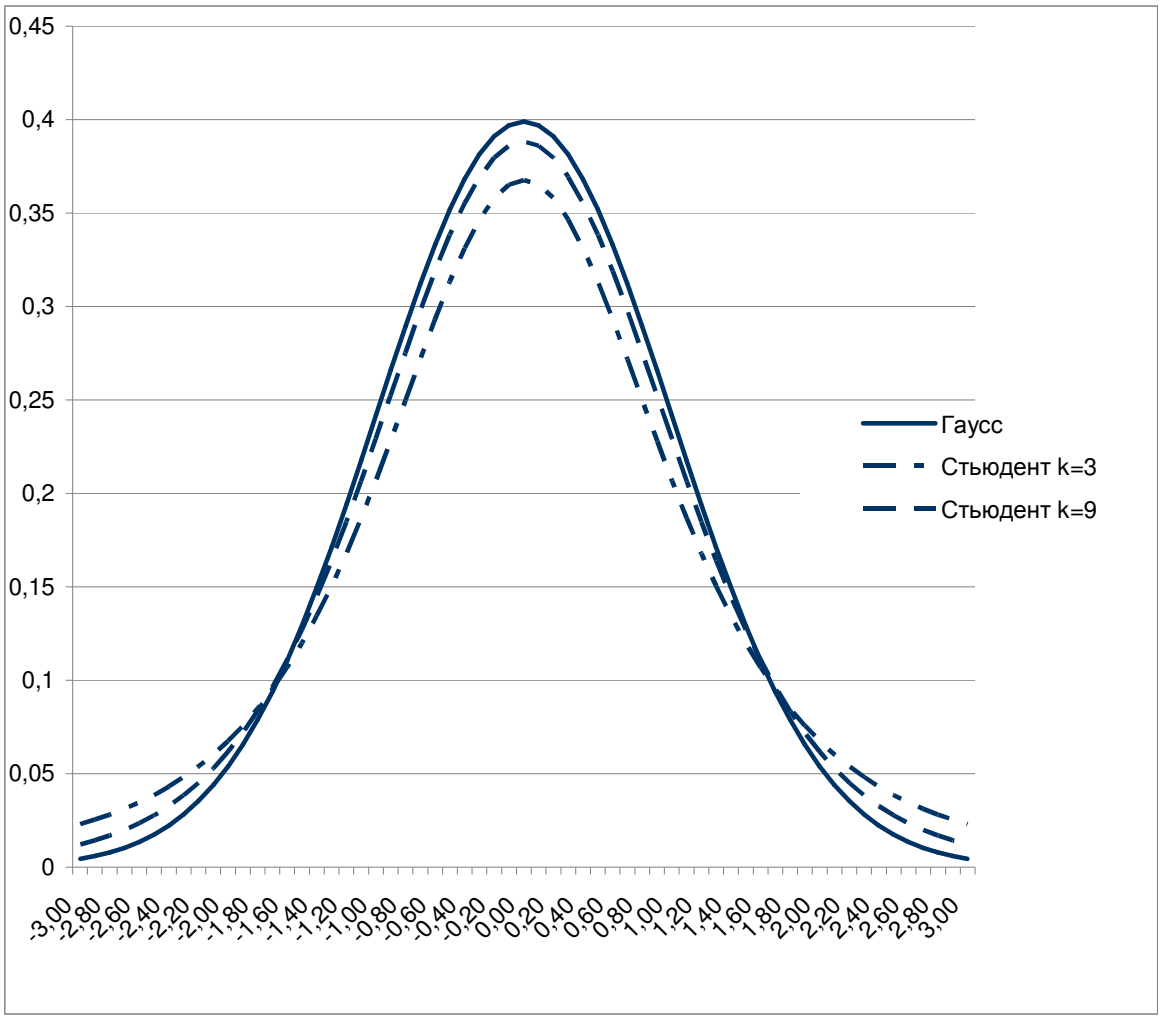
k	$\delta_{0.9,k}$	$\delta_{0.99,k}$
3	2.353	5.841
5	2.015	4.032
9	1.833	3.250

Таким образом, видно, что при увеличении надежности доверительный интервал расширяется, а при уменьшении – сокращается.

При малом числе k оценки, полученные при использовании распределения Стьюдента, заметно отличаются от оценок, основанных на использовании нормального распределения. Так, для нормального распределения $\delta_{0.99}$=2,576.



Интегральные функции распределения Лапласа и Стьюдента



Дифференциальные функции распределения Лапласа и Стьюдента

Литература

1. Н.В.Смирнов, И.В.Дунин-Барковский. Курс теории вероятностей и математической статистики (для технических приложений). – Москва, Наука, 1969.

2. Л.З.Румшиский. Элементы теории вероятностей. – Москва, Наука, 1976.

3. R.D.Yates, D.J.Goodman Probability and Stochastic Processes (A friendly introduction for electrical and computer engineers). John Wiley, NY, 2005.