doi:10.31725/0367-7966-2026-240-1-7
Известия Главной Астрономической Обсерватории в Пулкове, № 240

Простые приближения некоторых статистических функций

З. М. Малкин

Главная (Пулковская) астрономическая обсерватория РАН

Поступила в редакцию 24 февраля 2026 / Принята к публикации 18 марта 2026

Аннотация

Рассмотрены возможности упростить вычисления некоторых статистических функций, используемых для проверки статистических гипотез при обработке наблюдений: обратного нормального распределения, t-распределения Стьюдента и критерия отбраковки резко выделяющихся измерений. Для этих трех случаев предложены простые аппроксимирующие формулы для квантилей этих статистических распределений, имеющие достаточную точность для большинства практических приложений.

Ключевые слова: статистические распределения, аппроксимация, обратное нормальное распределение, t-распределение Стьюдента, исключение промахов

Введение

Статистические вычисления играют первостепенную роль при обработке измерений и наблюдений. В настоящей работе рассматриваются три функции, связанные с проверкой статистических гипотез и оценкой статистической значимости получаемых результатов обработки наблюдений. Несмотря на возможности использования многих развитых статистических пакетов, часто для обработки данных используется собственное математическое обеспечение, при разработке которого полезно иметь простые алгоритмы, обладающие достаточной для данного приложения точностью. Это особенно важно при автоматизированных и массовых вычислениях, которые также зачастую проводятся в условиях ограниченных вычислительных ресурсов и требований к высокой скорости обработки данных.

Задача упрощения статистических вычислений может решаться применением приближенных алгоритмов, основанных на аппроксимации статистических функций. В литературе можно найти много формул для аппроксимации различных статистических распределений, однако они, как правило, обладают точностью, а значит и сложностью, избыточной для большинства практических приложений. При этом нужно иметь в виду, что в этих работах авторы решают задачу аппроксимации на всем интервале определения аппроксимируемой функции, чего не требуется в практических задачах обработки наблюдений. Достаточно рассмотреть довольно узкий диапазон доверительных вероятностей 0.9–0.999, используемых в подавляющем числе прикладных задач. Это позволяет значительно упростить аппроксимирующие формулы и, соответственно, сократить время вычислений при сохранении достаточной для практики точности аппроксимации.

В работах Малкин (1993a) и Малкин (1993b) автором были предложены простые аппроксимирующие выражения для некоторых статистических распределений. Поскольку эти публикации были изложены очень сжато и к тому же сейчас мало доступны, в настоящей работе приводится их совместное изложение с уточнениями и дополнениями.

*e-mail:malkin@gaoran.ru

doi:10.31725/0367-7966-2026-240-1-7

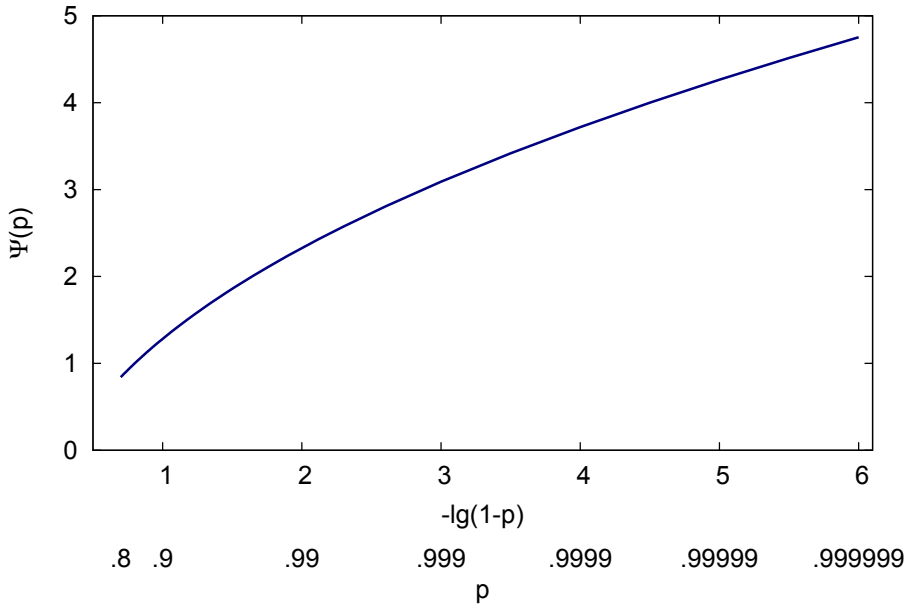

Рис. 1: Функция обратная функции нормального распределения.

1 Аппроксимирующие выражения

В этои разделе приведены простые аппроксимирующие выражения для трех статистических распределений и для нескольких уровней значимости чаще всего применяющихся для обработки измерительных и наблюдательных данных. В качестве исходных и тестовых данных для аппроксимации использовались сборники таблиц Большев, Смирнов (1983) и Оуэн (1973).

1.1 Обратное нормальное распределение

Обратная функция стандартного нормального распределения (с нулевым средним и единичной дисперсией) $x = N^{-1}(p; 0, 1) = \Psi(p)$ возвращает такое значение x, что $p = N(x; 0, 1)$. Здесь $N^{-1}(x; 0, 1) = \Phi(x)$ – интегральная функция стандартного нормального распределения. Вид функции $\Psi(p)$ приведен на рис. 1. Строго говоря, x определятся как корень уравнения

$$p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \,. \tag{1}$$

Величина $\Psi(p)$ является квантилем стандартного нормального распределения для доверительной вероятности p, что соответствует уровню значимости $\alpha_1 = (1 - p)$ для односторонней области или $\alpha_2 = 2(1 - p)$ для двусторонней области. Часто уровень значимости обозначают Q и выражают в процентах. Обычно используются значения уровня значимости от 0.1% до 10%.

Функция $\Psi(p)$ часто встречается в статистических вычислениях как самостоятельно, так и как составляющая других статистических функций, поэтому этот раздел изложен наиболее подробно. Значения этой функции могут быть получены как из точного решения уравнения (1), так и по аппроксимирующим формулам, различные варианты которых нетрудно найти в литературе. Однако обычно такие формулы предназначены для аппроксимации $\Psi(p)$ на всей области её определения и, как следствие, достаточно сложны. Для основных применяющихся при обработке данных значений доверительной вероятности p можно предложить простую и достаточно точную для практики аппроксимирующую функцию

$$\Psi(p) = a_1 + a_2\, t + a_3 \sqrt{t - a_4}\,, \tag{2}$$

где $t = -\ln(1 - p)$.

Таблица 1: Результаты аппроксимации функции $\Psi(p)$ по формуле (2).

Интервал p	a_1	a_2	a_3	a_4	Максимальная ошибка
0.95 – 0.999	−0.87350465	−0.02104348	1.61639568	−0.44533427	0.00005
0.9 – 0.999	−0.92337495	−0.02522121	1.64201371	−0.40330687	0.00010
0.8 – 0.9999	−0.95495887	−0.02695222	1.65576265	−0.37514736	0.0007
0.8 – 0.99999	−0.92270803	−0.02326696	1.63600922	−0.39742660	0.0011
0.8 – 0.999999	−0.88998754	−0.01991532	1.61689621	−0.42100939	0.0012
0.8 – 0.99999999	−0.84935143	−0.01629260	1.59450774	−0.45174214	0.0024

Оптимальные коэффициенты $a_1 \dots a_4$ формулы (2) зависят от заданного интервала значений p для которого желательно получить наилучшую точность аппроксимации. При этом до $p = 0.999$ включительно использовались таблицы Большев, Смирнов (1983), а для $p > 0.999$ использовались таблицы Оуэн (1973). Результаты вычислений для пяти интервалов p приведены в табл. 1.

В каждой строке табл. 1 приведены наборы коэффициентов $a_1 \dots a_4$, обеспечивающие наилучшую точность аппроксимации на интервале, указанном в первой колонке. В последней колонке приведена максимальная ошибка аппроксимации для данного интервала, под которой понимается абсолютная величина разности между значениями, вычисленными по (2), и точными значениями. Пользователь может выбрать строку (набор коэффициентов) соответствующую его задачам. Первый из этих интервалов представляется наиболее полезным для практики.

Практически, при программной реализации этого алгоритма целесообразно возвращать точные (табличные) значения $\Psi(p)$ для наиболее часто используемых значений доверительной вероятности: 0.8, 0.9, 0.95, 0.975, 0.98, 0.99, 0.995, 0.9975, 0.999, 0.9995 и 0.9999. Можно также заметить, что вычисление $\ln(1-p)$ не вызывает существенной потери точности для указанных интервалов p при применении арифметики с двойной точностью.

В работе Brophy (1985) проведено сравнение нескольких алгоритмов для аппроксимации обратной функции нормального распределения и только один из них (Odeh, Evans, 1974) обеспечивает точность лучше формулы (2), а именно $1 \cdot 10^{-6}$. При этом этот алгоритм реализован намного более сложной формулой

$$z = y - \frac{c_1\, y^4 + c_2\, y^3 + c^3\, y^2 + c_4\, y + c_5}{c_6\, y^4 + c_7\, y^3 + c_8\, y^2 + c_9\, y + c_{10}}\,, \quad y = \sqrt{-2\,\log(x)}\,, \tag{3}$$

а его точность представляется избыточной для обычной практики обработки наблюдений. В литературе можно найти и ещё более точные (и, соответственно, ещё более сложные) алгоритмы для вычисления квантилей нормального распределения, например, Wichura (1988), но они тем более вряд ли имеют интерес для практического использования.

1.2 Распределение Стьюдента

Другим важным для статистического анализа данных является распределение Стьюдента (или t-распределение), которое используется, например, для проверки статистической значимости коэффициентов корреляции, построения доверительных интервалов, сравнении средних двух выборок и т.д. Это распределение зависит от одного параметра – числа степеней свободы m и имеет вид, показанный на рис. 2.

Число степеней свободы при применении критериев, основанных на распределении Стьюдента, зависит от конкретного приложения t-критерия. Например, при проверке гипотезы о равенстве двух выборочных средних число степеней свободы равно $m = n_1 + n_2 - 2$, где n_1 и n_2 – размеры выборок, а при проверке гипотезы о равенстве нулю коэффициента корреляции число степеней свободы равно $m = n - 2$.

doi:10.31725/0367-7966-2026-240-1-7

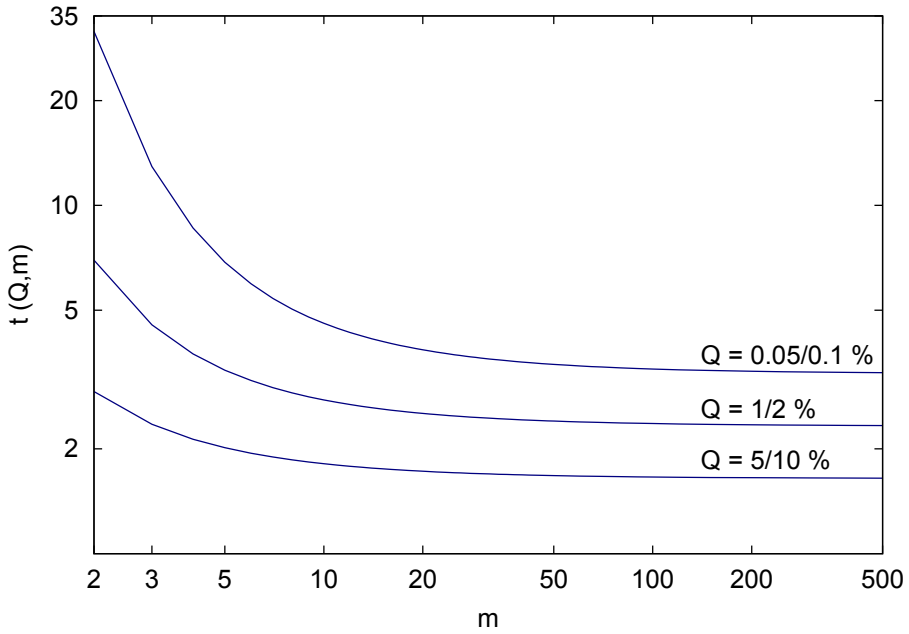

Рис. 2: Примеры t-распределения Стьюдента для трех уровней значимости: на надписях у графиков первое число соответствует односторонней области, второе –двусторонней области.

Простое, но обеспечивающее достаточную для практического применения точность, выражение для вычисления квантилей t-распределения имеет вид:

$$t(Q,m) = a_1 + \frac{a_2}{m + a_3}\,, \tag{4}$$

где Q – уровень значимости, m – число степеней свободы. Значения коэффициентов формулы (4) приведены в табл. 2. В первых двух колонках приведены уровни значимости в процентах для односторонней и двусторонней области. В колонках M_1, M_2 и M_3 приведены значения M такие, что при $m > M$ обеспечивается точность вычисления (абсолютная величина разности с точным значением) квантилей t-распределения не хуже 0.05, 0.01 и 0.001 соответственно.

Описанный здесь алгоритм аппроксимации t-распределения тестировался на табл. 3.2 из книги Большев, Смирнов (1983), которая содержит данные для $m = 1 \ldots 500$, что и определят пределы m, для которых справедлива указанная ошибка аппроксимации. В то же время известно, что при $m \to \infty$ распределение Стьюдента стремится к нормальному распределению, так что

$$\begin{array}{rcll} t(Q,\infty) & = & \Psi(1-Q/100) & \text{для односторонней области,} \\ t(Q,\infty) & = & \Psi(1-Q/200) & \text{для двусторонней области.} \end{array} \tag{5}$$

Эти значения приведены в последней колонке табл. 2. С другой стороны, при $m \to \infty$ аппроксимирующая формула (4) превращается в $t(Q,\infty) = a_1$. Сравнение колонок a_1 и $t(Q,\infty)$ табл. 2 показывает, что эти значения очень близки между собой, особенно для варианта коэффициентов, приведенных во второй строке для каждого Q. Таким образом, формула (4) применима и при $m > 500$ при сохранении указанной ошибки аппроксимации.

1.3 Исключение выбросов

При обработке наблюдений обычно возникает проблема отбраковки резко выделяющихся измерений. Одна из известных процедур статистического выявления грубых измерений (промахов) заключается в следующем. Пусть в результате обработки ряда наблюдений, содержащего n измерений, получены оценки определяемой величины $\hat{x}$ и её дисперсии:

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad s = \frac{1}{n-1}\sum_{i=1}^{n} v_i^2\,, \quad v_i = x_i - \hat{x}\,. \tag{6}$$

doi:10.31725/0367-7966-2026-240-1-7

Таблица 2: Результаты аппроксимации распределения Стьюдента по формуле (4).

Q_1, %	Q_2, %	a_1	a_2	a_3	M_1	M_2	M_3	$t(Q,\infty)$
10	20	1.2815	0.8483	-0.6407	2	3		1.2816
		1.2815	0.8476	-0.6505			5	
5	10	1.6448	1.5285	-0.8798	3	4		1.6449
		1.6448	1.5249	-0.9050			5	
2.5	5	1.9598	2.3848	-1.1072	3	4		1.9600
		1.9599	2.3759	-1.1457			6	
1	2	2.3259	3.7626	-1.3982	4	5		2.3263
		2.3263	3.7396	-1.4587			7	
0.50	1	2.5750	4.9793	-1.6092	5	6		2.5758
		2.5757	4.9356	-1.6932			7	
0.25	0.5	2.8055	6.3402	-1.8126	5	6		2.8070
		2.8068	6.2630	-1.9258			8	
0.10	0.2	3.0873	8.3566	-2.0689	5	6		3.0902
		3.0897	8.2103	-2.2253			9	
0.05	0.1	3.2860	10.0454	-2.2535	6	6		3.2905
		3.2898	9.8193	-2.4492			10	

Предположим, что k-ое измерение имеет максимальную невязку v_k и необходимо проверить, не является ли это измерение промахом, подлежащим исключению. Вопрос об исключении решается положительно, если выполняется условие

$$\frac{|v_k|}{s} > \zeta(Q,n)\,, \tag{7}$$

где Q – уровень значимости. Вид функции $\zeta(Q,n)$ для трех уровней значимости показан на рис. 3.

В настоящей работе представлены достаточно простые и точные для практических приложений выражения для приближенного вычисления статистики $\zeta(Q,n)$. Для аппроксимации использована таблица 4.8в из Большев, Смирнов (1983) с двумя отличиями. Во-первых, таблица была продолжена до $n = 500$ с помощью формулы Большев, Смирнов (1983), стр. 60 и аппроксимации $\Psi(p)$, описанной в разделе 1.1. Во-вторых, таблица 4.8в из Большев, Смирнов (1983) приведена для случая вычисления дисперсии как $s = \frac{1}{n}\sum v_i^2$. Поэтому она была пересчитана для случая вычисления выборочной дисперсии как $s = \frac{1}{n-1}\sum v_i^2$ путем умножения значений из Большев, Смирнов (1983) на $\sqrt{\frac{n-1}{n}}$.

Аппроксимация функции $\zeta(Q,n)$ производится по формуле

$$\zeta(Q,n) = \Psi(p)\cdot\left(a_1 + a_2\,n + \frac{a_3}{a_4+n}\right)\,, \tag{8}$$

где $p = 1 - Q/200n$, а Q выражено в процентах. Значения коэффициентов $a_1 \ldots a_4$ для разных значений Q приведены в табл. 3. Для каждого Q в таблице приведены два варианта коэффициентов. В верхней строке приведены оптимальные коэффициенты для $n = 6 \ldots 500$, в нижней – для $n = 6 \ldots 100$. Ошибка аппроксимации (абсолютная величина разности между приближенными значениями, вычисленными по (8), и точными значениями) меньше 0.007 в первом случае и 0.003 во втором случае.

doi:10.31725/0367-7966-2026-240-1-7

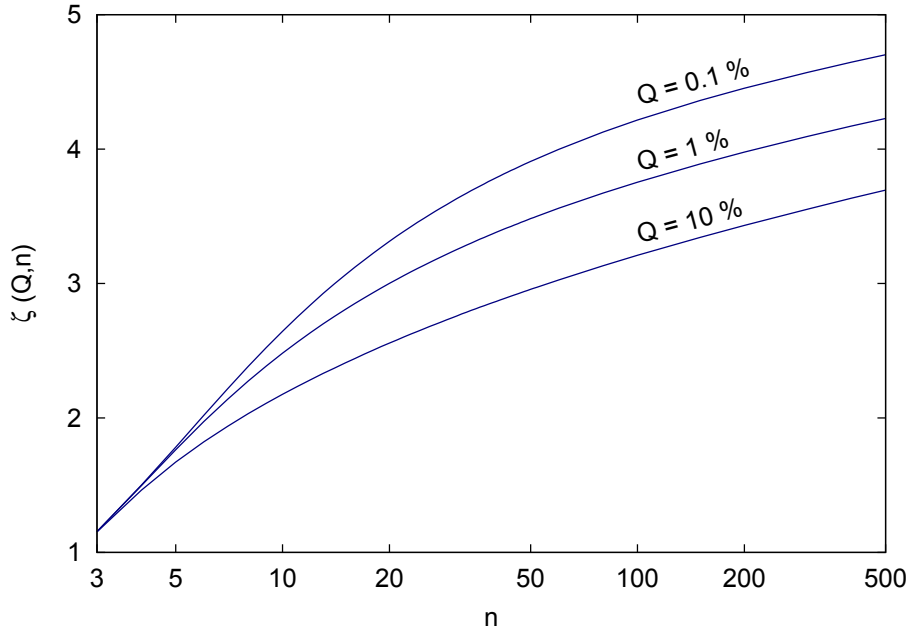

Рис. 3: Примеры распределения $\zeta(Q, n)$ для трех уровней значимости.

Таблица 3: Результаты аппроксимации функции $\zeta(Q, n)$ по формуле (8).

Q, %	a_1	a_2	a_3	a_4
10	$9.88545 \cdot 10^{-1}$	$2.01128 \cdot 10^{-5}$	-1.68729	1.63739
	$9.81392 \cdot 10^{-1}$	$9.79867 \cdot 10^{-5}$	-1.51368	0.96360
5	$9.88424 \cdot 10^{-1}$	$2.04021 \cdot 10^{-5}$	-1.99297	1.45970
	$9.81622 \cdot 10^{-1}$	$9.41882 \cdot 10^{-5}$	-1.82875	0.93060
2	$9.88432 \cdot 10^{-1}$	$2.03774 \cdot 10^{-5}$	-2.41798	1.52820
	$9.81751 \cdot 10^{-1}$	$9.28241 \cdot 10^{-5}$	-2.25551	1.09606
1	$9.88913 \cdot 10^{-1}$	$1.90686 \cdot 10^{-5}$	-2.76375	1.72925
	$9.82771 \cdot 10^{-1}$	$8.45343 \cdot 10^{-5}$	-2.61076	1.36652
0.5	$9.89111 \cdot 10^{-1}$	$1.88814 \cdot 10^{-5}$	-3.10307	1.95188
	$9.83396 \cdot 10^{-1}$	$7.86601 \cdot 10^{-5}$	-2.95706	1.63688
0.2	$9.90087 \cdot 10^{-1}$	$1.63353 \cdot 10^{-5}$	-3.58659	2.37111
	$9.85744 \cdot 10^{-1}$	$5.91809 \cdot 10^{-5}$	-3.46890	2.14099
0.1	$9.90843 \cdot 10^{-1}$	$1.44077 \cdot 10^{-5}$	-3.95667	2.71055
	$9.87049 \cdot 10^{-1}$	$5.12540 \cdot 10^{-5}$	-3.85096	2.51786

2 Заключение

В настоящей работе представлены несколько простых выражений для аппроксимации трех статистических распределений, использующихся при обработке наблюдательных данных, в частности, для проверки статистических гипотез. Предлагаемые алгоритмы могут уступать по точности другим более изощренным методам аппроксимации, но их достоинством является вычислительная простота, а значит и скорость вычислений, что может быть полезным, а иногда и критическим, при массовых вычислениях, особенно в реальном времени.

Конечно, автор не претендует на оптимальное решение поставленной задачи упрощения и ускорения статистических вычислений и надеется, что заинтересованные читатели и пользователи смогут предложить более удачные аппроксимирующие решения как для рассмотренных в настоящей работе, так и для других статистических функций.

На сайте ГАО РАН[1] размещено несколько функций на языке Fortran, реализующих алгоритмы, рассмотренные в настоящей работе.

Благодарности

Автор благодарен рецензенту за полезные замечания по первоначальному варианту работы.

Список литературы

Малкин, З. М. (1993a). Простое вычисление квантилей t-распределения. *Астрон. цирк.*, 1554, 43.

— (1993b). Об исключении резко выделяющихся измерений. *Астрон. цирк.*, 1555, 33—34.

Большев, Л. Н., Н. В. Смирнов (1983). *Таблицы математической статистики.* М. : Наука. Физматлит – 416 С.

Оуэн, Д. Б. (1973). *Сборник статистических таблиц.* М.: Вычислительный центр АН СССР – 586 С.

Brophy, A. L. (1985). Approximation of the inverse normal distribution function. *Behavior Research Methods, Instruments, & Computers*, 17, 415—417.

Odeh, R. E., J. O. Evans (1974). Algorithm AS 70: The percentage points of the normal distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23, 96—97.

Wichura, M. J. (1988). Algorithm AS 241: The Percentage Points of the Normal Distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 37, 477—484.

Simple approximations of some statistical functions

Z.M. Malkin

Central Astronomical Observatory at Pulkovo of RAS

Received 24 February 2026 / Accepted 18 March 2026

Abstract

Possibilities are considered to simplify the calculation of some statistical functions used to test statistical hypotheses when processing observations: the inverse normal distribution, the Student's t-distribution, and the criterion for rejecting outliers. For these three cases, simple approximation expressions are proposed for the quantiles of these statistical distributions, which are accurate enough for most practical applications.

Key words: statistical distributions, approximation, inverse normal distribution, Student's t-distribution, outlier elimination

[1] https://www.gaoran.ru/english/as/soft/