Оценка доверительных интервалов бутстрап методом с использованием существенной выборки

Афанасьев Михаил Альбертович, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., профессор Ермаков М.С.
Рецензент: к.ф.-м.н., доцент Каштанов Ю.Н.

Введение

Существует задача вычисления вероятностей редких событий, которые имеют место во многих приложениях и задачах математической статистики, но при их вычислении прямым моделированием возникает две проблемы:

1. Большой объем вычислений.
2. Малейшие ошибки в качестве случайных чисел могут вызвать существенные ошибки оценок.

Наиболее распостраненный метод вычисления – *метод существенной выборки*.

Цель работы

Мы подробно остановимся на задаче оценивания малых вероятностей с помощью бутстрап метода и использования существенной выборки.

Цель данной работы -исследовать насколько точна нормальная апроксимация выборочного среднего, стьюдентизированного среднего и их бутстрап аналогов в зонах малых вероятностей, когда объем выборки мал и для распределений имеющих тяжелые хвосты.

Постановка задачи

Рассмотрим одну из простейших задач - оценки малых вероятностей выборочного среднего Пусть $\{X_i\}$ выборка, где X_i — н.о.р.с.в. из распределения с тяжелыми хвостами, например, Стьюдента (t(5), t(10), t(20)).
Наша задача состоит в оценивании следующей вероятности:

$$\mathrm{U} = P(\overline{X} - \mathrm{E}X_i) > b_n) \tag{1}$$

В нашем случае мы рассмотрим и сравним несколько вариантов:

Постановка задачи

Случай выборочного среднего $P(\sqrt{n}\frac{\overline{X}-\mathsf{E}X_i}{\sigma} > b_n)$.

Случай стьюдентизированного среднего $P(\sqrt{n}\frac{\overline{X}-\mathsf{E}X_i}{s} > b_n)$.

Аналогично для бутстрап метода:

С известной дисперсией $P(\sqrt{n}\frac{\overline{X^*}-\overline{X}}{\sigma} > b_n)$.

С неизвестной дисперсией $P(\sqrt{n}\frac{\overline{X^*}-\overline{X}}{s^*} > b_n)$.

Таким образом будут изучены вышеуказанные задачи и сравним насколько эти малые вероятности совпадают и насколько хороша нормальная аппроксимация, когда распределения имеют тяжелые хвосты. И сопоставим полученные численные результаты с аналитическими результатами в этой области.

Метод существенной выборки

Задача — оценить вероятность

$$V_n = P\left(T(\hat{F}_n) - T(F) > b_n\right), \tag{2}$$

где F – теоретическое распределение, $\hat{F}_n$ – эмпирическая фунция распределения, построенная по наблюдениям $X_i \sim F$, $1 \leqslant i \leqslant n$, T – некоторый функционал, b_n – заданное число. В дальнейшем, $T(\hat{F}_n) = \overline{X}$, $T(F) = \mathsf{E}X_i$.

Метод существенной выборки

Введем меру Q: $Q \ll F$. Обозначим $q = \dfrac{dQ}{dF}$. Промоделируем k независимых выборок с распределением Q

$$Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}, \quad 1 \leqslant i \leqslant k.$$

В качестве оценки вероятности (2) берем

$$\hat{U} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{T(\hat{Q}_n^{(i)}) - T(F) > b_n} \prod_{j=1}^{n} q^{-1}(Y_j^{(i)}), \tag{3}$$

где $\hat{Q}_n^{(i)}$ эмпирическое распределение $Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}$.

Методы решения при больших уклонениях: Крамер

Теорема (Условие сходимости по Крамеру)

Если интеграл $f(\zeta) = \int_{-\infty}^{+\infty} e^{\zeta x} F\{dx\}$ *сходится при* $|\zeta| < \zeta_0$ *и* x *меняется вместе с* n *так, что* $x = o(n^{1/6})$*, то верно соотношение*

$$\frac{1 - F_n(x)}{1 - N(x)} \to 1.$$

Крамер использовал некоторые предположения о существовании моментов, в частности, предполагал, что σ (иногда и μ_3, μ_4) исходного теоретического распределения известна и конечна.

Методы решения при больших уклонениях: Шао

Шао показал, что для стьюдентизированного среднего имеет место нормальная аппроксимация.

Теорема (Условие сходимости по Шао)

При $\mathsf{E}(X_i) = 0$ *и конечности третьего момента* $\mathsf{E}|X_i|^3 \leqslant \infty$ *имеет место следующее соотношение для* $x = o(n^{1/6})$*:*

$$\frac{P(S_n/V_n \geq x)}{1 - N(x)} \to 1,$$

где $S_n = \sum_{i=1}^{n} x_i$, $V_n = \sum_{i=1}^{n} x_i^2$.

Методы решения при больших уклонениях: Вуд

Теперь перейдем к бутстрапу

Теорема (Условия сходимости по Вуду)

Пусть $\mathsf{E}(X_i) = 0$, $\mu_3 = \mathsf{E}(X_i^3)$ *и* $\mu_4 = \mathsf{E}(X_i^4) < \infty$. *Если* $x_n \to \infty$ *и* $x_n = o(n^{1/4})$, *то*

$$G_n(x) = \hat{G}_n(x)(1 + o_p(1)).$$

Где

$$G_n(x) = P(\frac{S_n}{\sqrt{n}} > x)), \quad S_n = \sum_{i=1}^{n}(X_i - \mu),$$

$$\hat{G}_n(x) = P(\frac{S_n^*}{\sqrt{n}} > x), \quad S_n^* = \sum_{i=1}^{n}(X_i^* - \overline{X}),$$

В частности, эту теорему можно спроецировать на разные распределения с тяжелыми хвостами.

Методы решения проблем при больших уклонениях

Таблица 1: Аппроксимация сложных распредений известными

	Зона умеренных уклонений	Условие на исходное распределение и моменты
Крамер	$x_n = o(n^{1/6})$	$\int_{-\infty}^{+\infty} e^{\zeta x} F\{dx\}$ сх-ся, конечность некоторых моментов $\mu_2 < \infty$
Шао	$x_n = o(n^{1/6})$	$\mathsf{E}X_i = 0$ и $\mathsf{E}\lvert X_i\rvert^3 \leqslant \infty$
Вуд	$x_n = o(n^{1/4})$	конечность некоторых моментов $\mu_4 < \infty$, иногда $\mu_{2+\delta} < \infty$

Метод решения

Исходная мера – плотность распределения величин X_i

$$p_{x;\mu} = f(x;\mu)$$

Замена меры, смещаем на $b_n = \Delta$: $\mu_\Delta = \mu + \Delta$

$$p_{x;\mu_\Delta} = f(x;\mu+\Delta),$$

$$q(x) = \prod_{i=1}^{n} \frac{p_{i,\mu_\Delta}(x)}{p_{i,\mu}(x)}.$$

Моделируем k независимых выборок, $Y_j^{(i)}$ с плотностью $p_{j,\mu_\Delta}(x)$

$$Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_n^{(i)}, \quad 1 \leqslant i \leqslant k.$$

Оценка $\hat{\mathrm{U}}$

$$\hat{\mathrm{U}} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\{\overline{X}_\Delta^{(i)} > \mu+\Delta\}} \prod_{j=1}^{n} \frac{p(Y_j^{(i)};\mu)}{p(Y_j^{(i)};\mu+\Delta)} \tag{4}$$

Модификация бустрап метода: Джонс

Вернемся к бутстрапу. Обозначим функцию правдоподобия:

$$I_G(\mathbf{Y}) = \prod_{i=1}^{n} \prod_{j=1}^{n} g_{ij}^*,$$

где $g_{ij}^* = g_j$, если $Y_i = X_j$, и $g_{ij}^* = 1$ иначе. Из выборок $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(R)}$ получаем оценки $T_{(1)} \leqslant T_{(2)} \leqslant \cdots \leqslant T_{(R)}$.

$$S_r = \frac{1}{R} \sum_{i=1}^{r} \frac{I_F(\mathbf{Y}^{(i)})}{I_G(\mathbf{Y}^{(i)})}. \tag{5}$$

для $r = 1, 2, \ldots, R$.
$\exists! r^* : S_{r^*} \leqslant p < S_{r^*+1}$. Оценкой p-квантили для статистики $T(F)$ будет T_{r^*}. Замена меры:

$$G_n : \quad g_i = \frac{1}{n} + \frac{b_n}{sn} \left(X_i - \overline{X} \right).$$

Функцию влияния h возьмем $h(x) = x$.

Алгоритмы моделирования:Крамер, Шао

$\{\xi_{itj}\}_{t=1}^{\nu} \sim N(0,1)$, тогда сдвинутая на Δ выборка с распределением Стьюдента:

$$\eta_{ij} = \frac{\frac{1}{\sqrt{\nu}} \sum_{t=1}^{\nu} \xi_{itj}}{s_{ij}} + \Delta,$$

Тогда оценка вероятности попадания за квантиль примет вид

$$\hat{P}(\overline{X} > x_\alpha) = \hat{\mathrm{U}} = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{\overline{X}_\Delta^{(j)} > x_\alpha\}} \prod_{i=1}^{n} \frac{f_\nu(\eta_{ij})}{f_\nu(\eta_{ij} - \Delta)}, \qquad (6)$$

где f_ν - плотность, а $b_\nu^2 = \frac{\nu-1}{\nu-3}$ - дисперсия распределения Стьюдента с $\nu - 1$ степенями свободы.

$$\overline{X}_\Delta^{(j)} = \frac{1}{b_\nu \sqrt{n}} \sum_{i=1}^{k} \eta_{ij}. \quad \Delta = x_\alpha \frac{b_\nu}{\sqrt{n}}.$$

Осталось смоделировать k выборок и подставить в (6)

Алгоритмы моделирования:Бутстрап

1. Вычислим набор весов $w_i = \frac{1}{n} + \frac{b_n}{ns}\left(h(X_i) - \overline{h}\right)$.
2. Моделируем случайное $x = X_i$
   2a. Считаем вероятность попадания случайной велчисины слева от $\overline{h}$ - p_1 и p_2 справа. $p_1 + p_2 = 1$.
   2b. Моделируем β р.р на [0;1] если $\beta < p_1$ тогда с вероятностью $\frac{1}{k}$ (k- количество случайных величин слева от $\overline{h}$) берем любую из них в качестве x, случай когда $\beta > p_1$ делается аналогично.
3. $h(x) > \overline{h}$. И $h_m = (h(x)_{max} - \overline{h}), D_n = \frac{1}{n} + \frac{b_n}{ns}(h_m)$. Моделируем α р.р. на $(0; D_n)$. Если $\alpha \leqslant \frac{b_n}{s}\left(h(x) - \overline{h}\right)$, то добавляем x к выборке, иначе возвращаемся к 2b.
4. $h(x) \leqslant \overline{h}$. Моделируем α р.р. на $(0; \frac{1}{n})$. Если $\alpha \leqslant \frac{1}{n} + \frac{b_n}{ns}\left(h(x) - \overline{h}\right)$, то добавляем x к выборке, иначе возвращаемся к 2b.

Алгоритмы моделирования:Бутстрап

Пусть **X** выборка с теоретической функцией распределения F,с математическим ожиданием $\mu = 0$, дисперсией σ^2. $\mathbf{X}_i^*$ - бутстрап выборка, полученная из исходной. Нужно вычислить следующую величину:

$$\hat{\psi} = P(\sqrt{n}\frac{\overline{X^*} - \overline{X}}{\sigma} > x_\alpha).$$

Если всего по ходу бутстрапа будет промоделировано R выборок, из которых r будет удовлетворять неравенству $\sqrt{n}\frac{\overline{X_i^*} - \overline{X}}{\sigma} > x_\alpha$, то, собственно, $\frac{r}{R}$ - и есть наша вероятность.

Стьюдентизированный метод будет отличаться только тем, что по ходу бутстрапа будем вычислять не только $\overline{X^*}$ - выборочное среднее, но и $\hat{\sigma}^*$ - стандартное отклонение. И соответственно будем искать:

$$\hat{\xi} = P(\sqrt{n}\frac{\overline{X^*} - \overline{X}}{\hat{\sigma}^*} > x_\alpha).$$

Численные результаты

Изобразим полученные значения квантилей на рисунках 1 - 2.

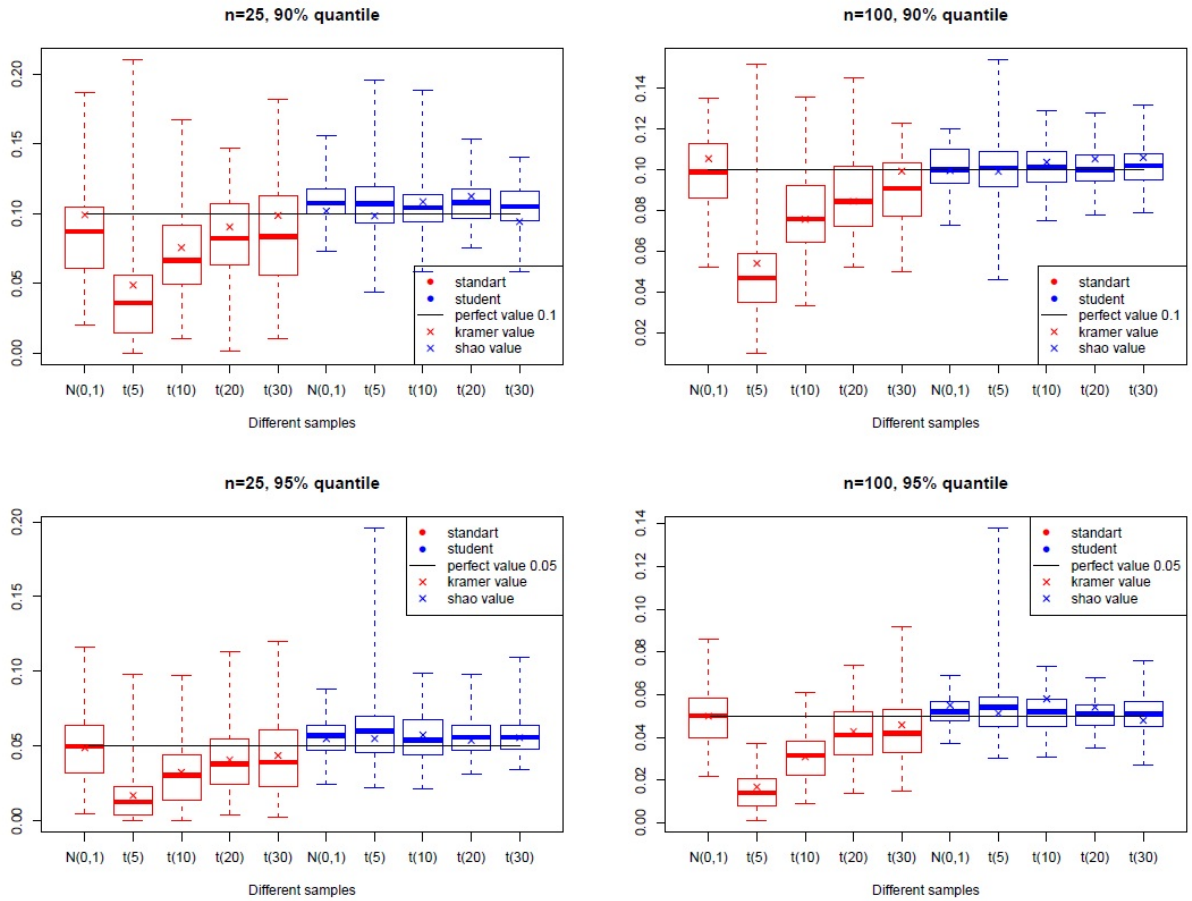

Рис. 1: Разброс вероятностей на хвостах.

Численные результаты

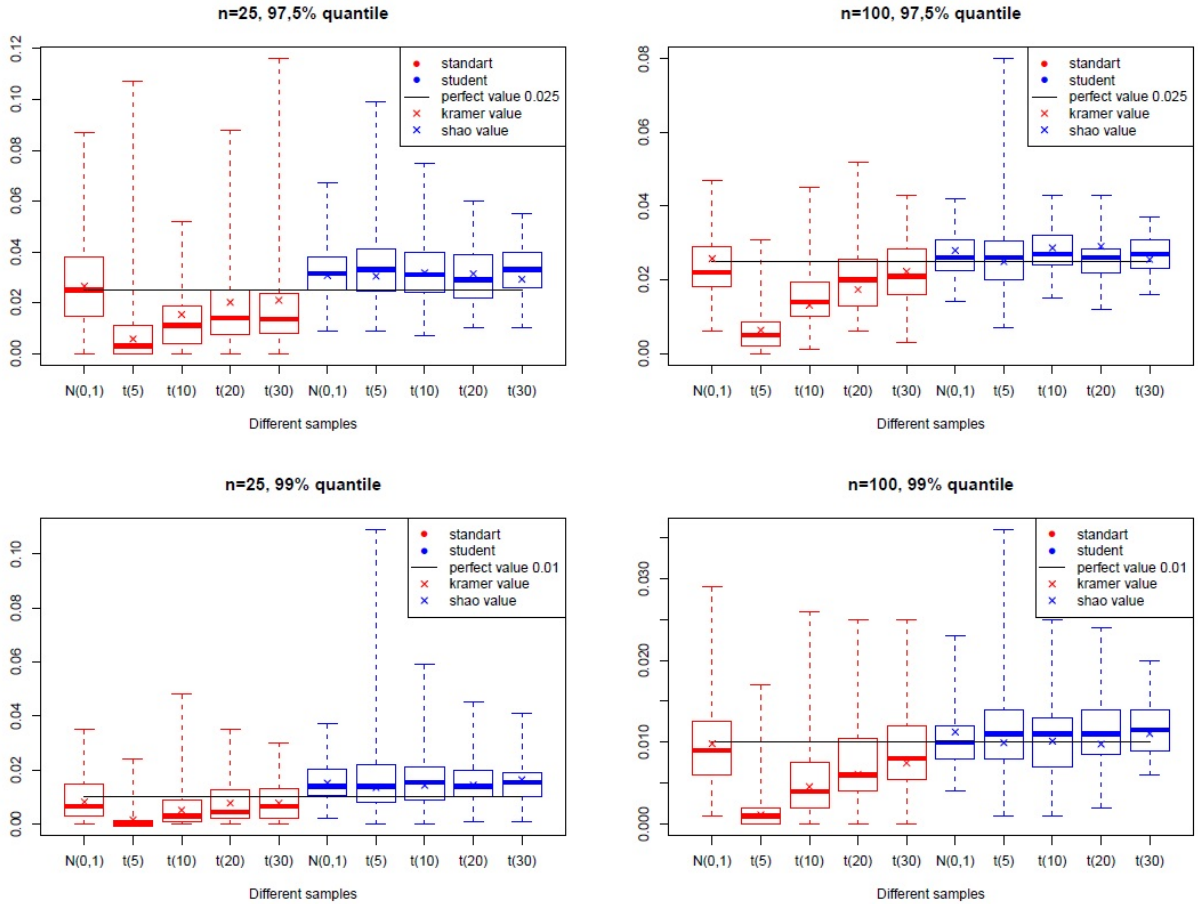

Рис. 2: Разброс вероятностей на хвостах.

Численные результаты

Модернизированный бутстрап в сравнении с обычным (Рис. 3).
Исходная выборка из распределения $t(10)$ размера $n = 25$.

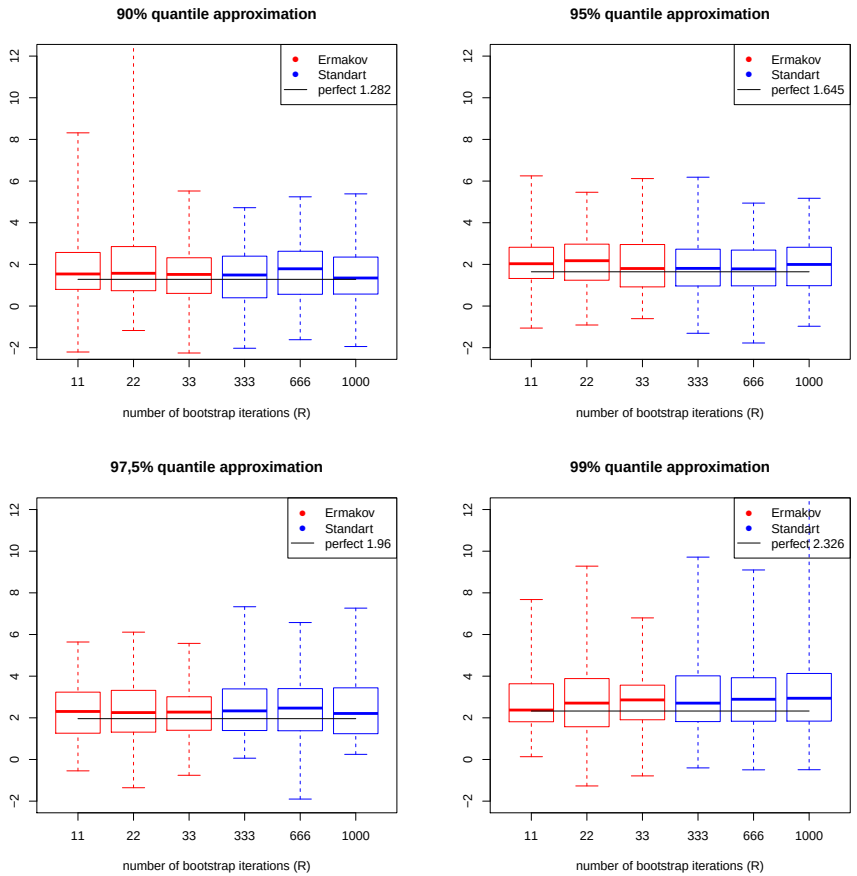

Рис. 3: Разброс квантилей.

Численные результаты: зависимость от n

Изобразим разброс значений квантилей, полученных по алгоритму на рисунке 4.

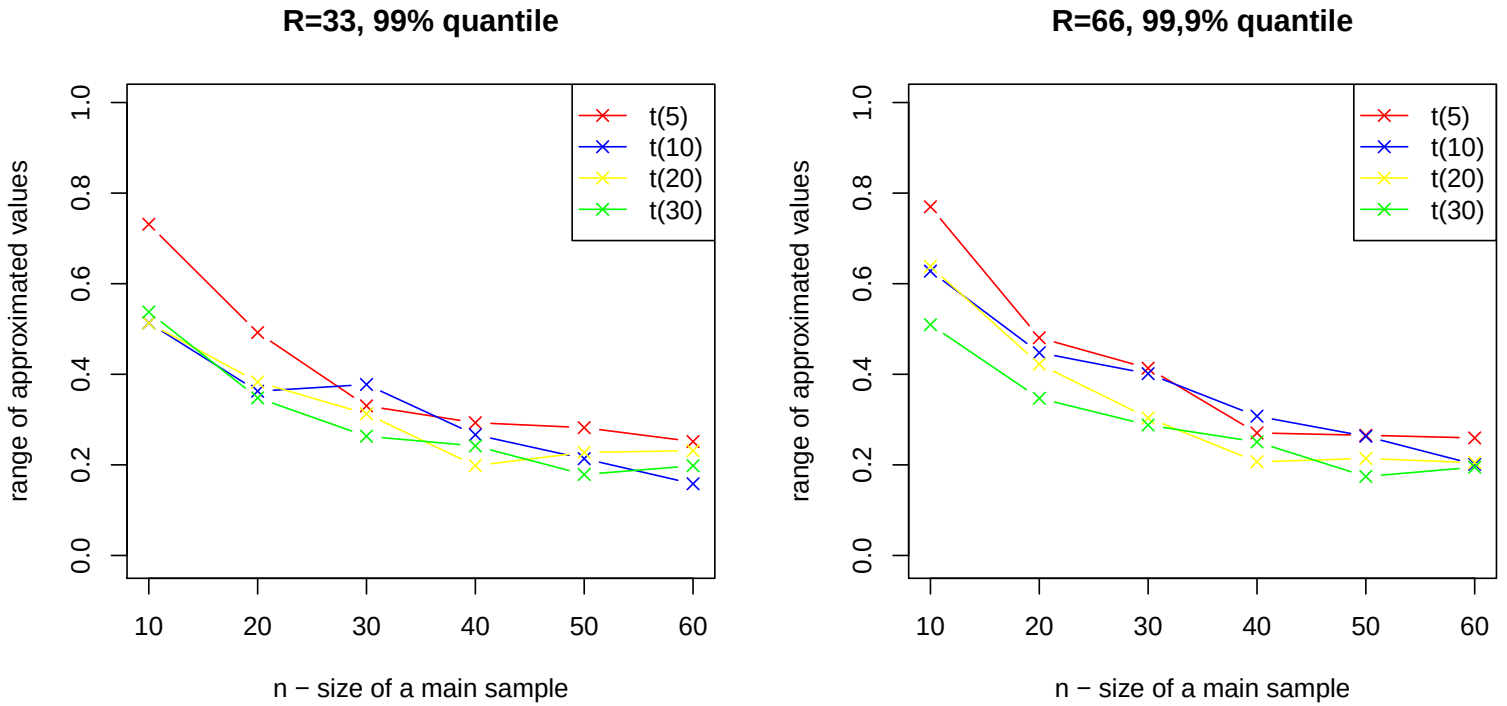

Рис. 4: Разброс аппроксимаций квантилей при различных исходных распределениях.

Численные результаты: зависимость от распределения

Изобразим полученные оценки на рисунках 5 - 6.

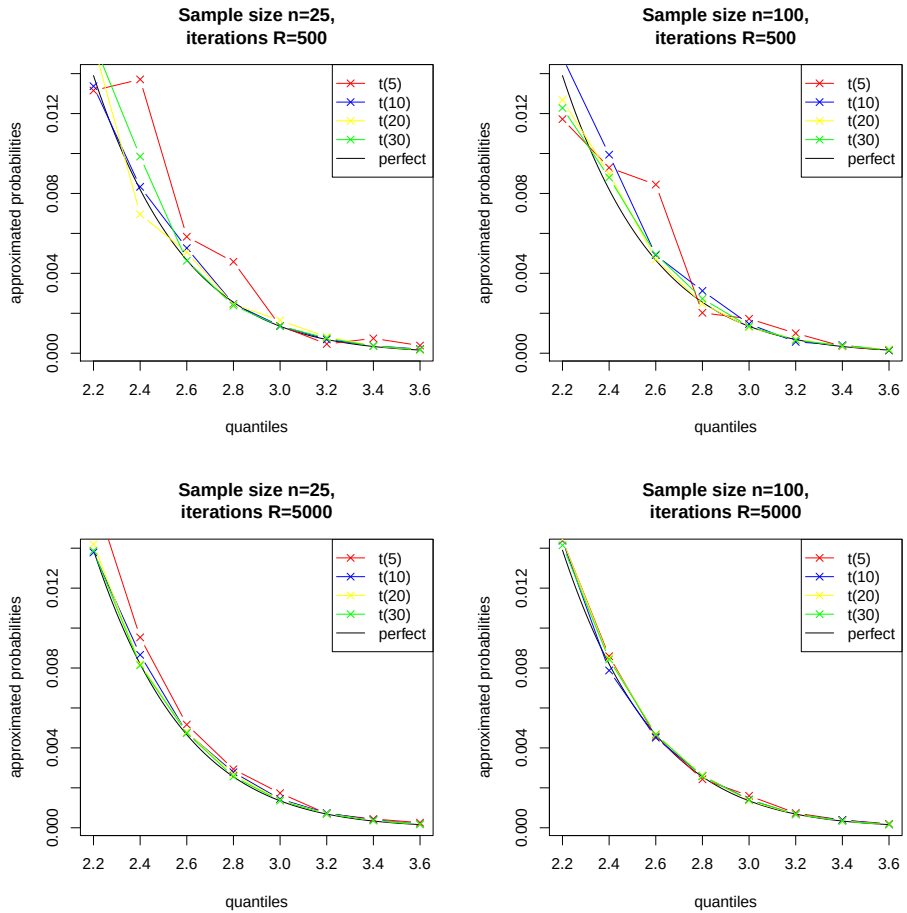

Рис. 5: Результат для моделирования выборочного среднего с известной диспресией.

Численные результаты: зависимость от распределения

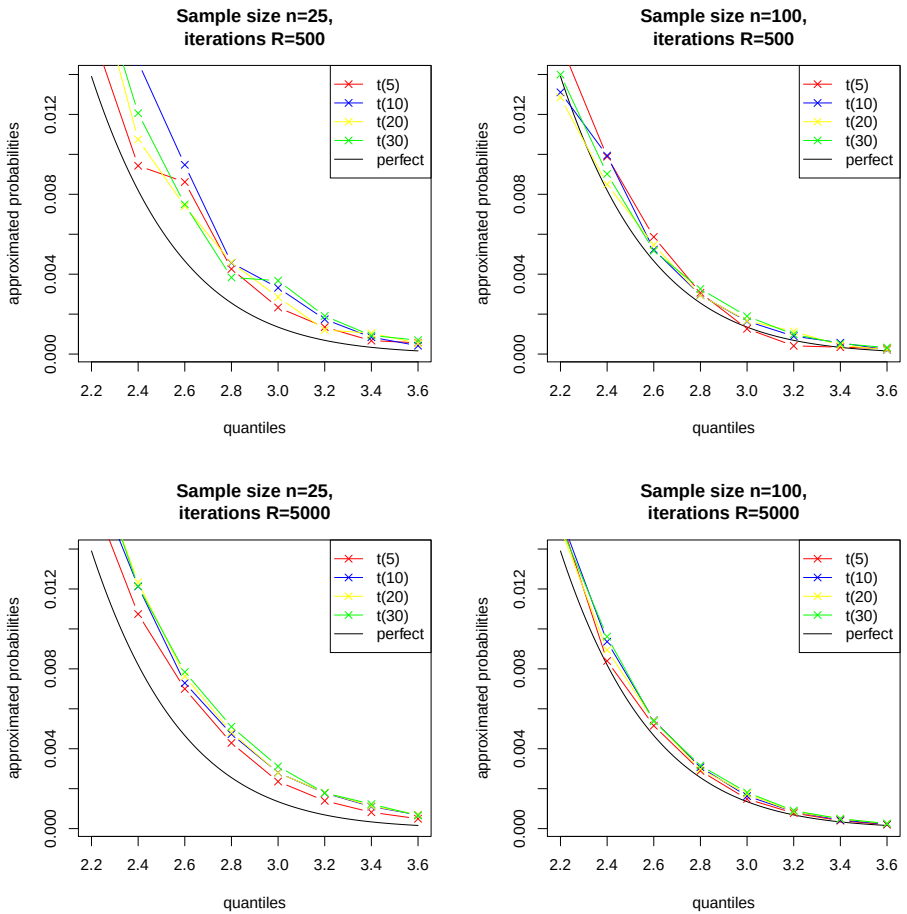

Рис. 6: Результат для стьюдентизированного среднего с выборочной дисперсией.

Численные результаты для стьюдентизированного среднего с помощью бутстрап метода

Изобразим полученные оценки вероятностей на рисунке 7 ($n = 50$).

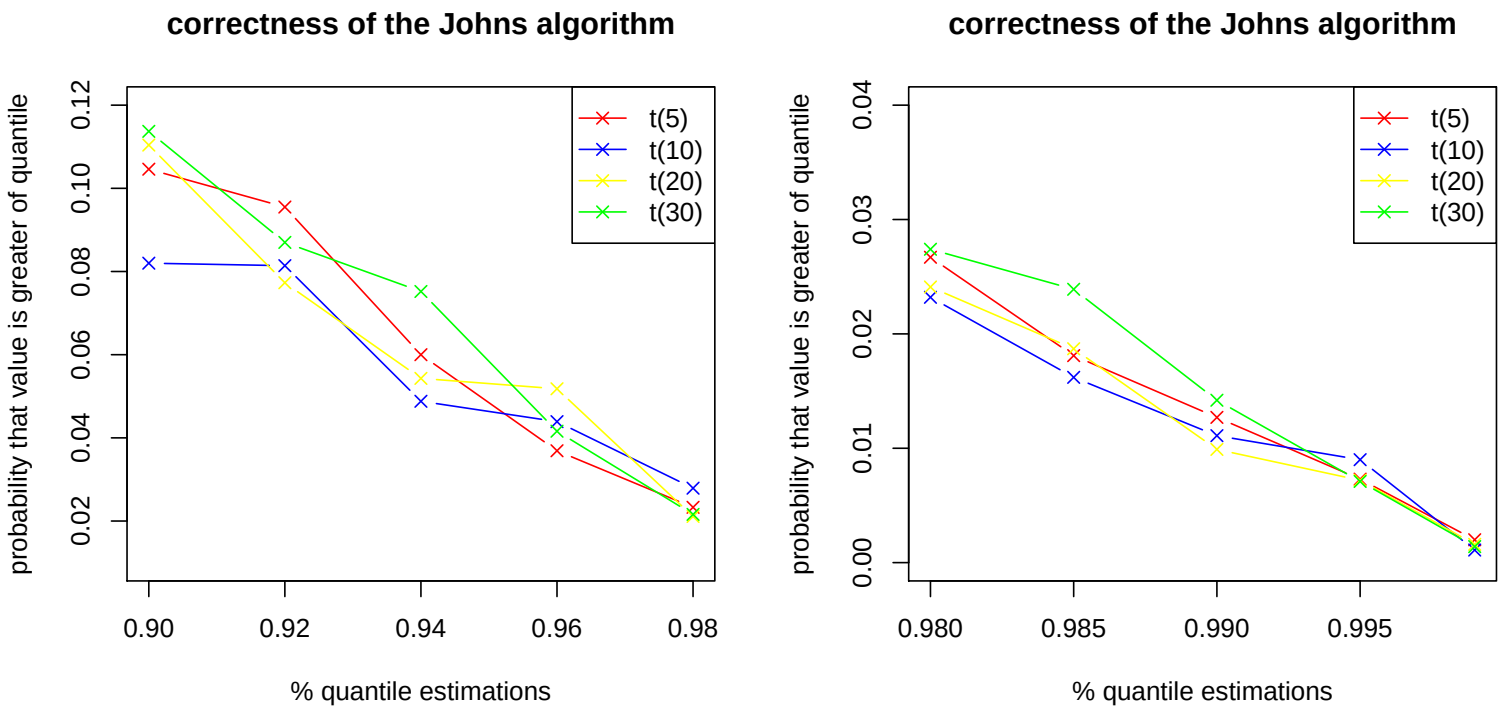

Рис. 7: Метод Джонса для различных малых вероятностей.

Заключение

1. Чем хвосты распределения тяжелее, тем более неточными получаются результаты при стандартных методах.
2. Моделирование показало, что стьюдентизированная оценка действительно имеет лучшую нормальную аппроксимацию, чем среднее нормированное на стандартное отклонение.
3. Вуд показал, что в случае бутстрапа нормальная аппроксимация выборочного среднего имеет место в более широкой зоне вероятностей умеренных уклонений. Это подтверждают результаты моделирования.
4. Джонс модернизировали бутстрап так, что он стал работать в несколько раз быстрее и точнее чем банальные алгоритмы. Мы показали, что этот метод прекрасно работает для вычисления малых вероятностей.
5. Оказывается, что можно точно оценивать квантили очень малых вероятностей (< 0.025) с помощью последнего метода.