

# Методы современной прикладной статистики

## 3. Критерии согласия, проверка нормальности, бутстреп

Родионов Игорь Владимирович  
vecsell@gmail.com

Весна, 2018

Пусть  $X_1, \dots, X_n$  – выборка из неизвестного распределения с функцией распределения  $F$ .

Рассмотрим гипотезу  $H_0 : F = F_0$  против альтернативы  $H_1 : F \neq F_0$ .

Критерии проверки таких гипотез называются **критериями согласия**.

# Критерий Колмогорова-Смирнова

Рассмотрим статистику

$$D_n := \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|,$$

где  $\hat{F}_n(x)$  – эмпирическая функция распределения.

## Теорема Колмогорова

Если  $F_0$  – непрерывна, то при верной гипотезе  $H_0$

$$P(\sqrt{n}D_n \leq x) \xrightarrow{n \rightarrow \infty} K(x) = \sum_{k \in \mathbb{Z}} (-1)^k e^{-2k^2 x^2}.$$

# Критерий Колмогорова-Смирнова

Таким образом, критерий Колмогорова-Смирнова выглядит так:

если  $\sqrt{n}D_n > K_{1-\alpha}$ , то отвергнуть  $H_0$ ,

где  $K_{1-\alpha}$  –  $(1 - \alpha)$ -квантиль функции распределения  $K(x)$ . Если  $\alpha$  близко к 0, то

$$K_{1-\alpha} \approx \sqrt{-\frac{1}{2} \ln \frac{\alpha}{2}}.$$

# Критерий хи-квадрат Пирсона

По-прежнему проверяем гипотезу  $H_0 : F = F_0$ . Пусть  $P_0$  – вероятностная мера (на  $\mathbb{R}$ ), соответствующая функции распределения  $F_0$ .

Разобьем  $\mathbb{R}$  на  $\approx 10$  интервалов  $\{B_i\}_{i=1}^k$  таких, что  $nP_0(B_i) \geq 5 \forall i$  ( $B_i$  выбираются заранее, вне зависимости от данных!). Обозначим  $p_i^0 = P_0(B_i)$ ,  $\mu_i = \#\{j : X_j \in B_i\}$ .

**Теорема Карла Пирсона.**

$$\hat{\chi} := \sum_{i=1}^k \frac{(\mu_i - np_i^0)^2}{np_i^0} \xrightarrow{d} \chi_{k-1}^2, \quad n \rightarrow \infty.$$

# Критерий хи-квадрат Пирсона

Критерий хи-квадрат выглядит так

*если  $\hat{\chi} > u_{1-\alpha}$ , то отвергнуть  $H_0$ ,*

где  $u_{1-\alpha}$  –  $(1 - \alpha)$ -квантиль распределения  $\chi^2_{k-1}$ .

Критерий хи-квадрат кажется более грубым, чем критерий Колмогорова, однако он незаменим при работе с дискретными данными и зачастую показывает лучшие результаты, чем критерий Колмогорова, даже в случае “непрерывных” данных.

# Другие критерии согласия

- Критерий Смирнова-Крамера-фон Мизеса ( $\omega^2$ -критерий) со статистикой

$$\omega^2 = \int (\hat{F}_n(x) - F_0(x))^2 dF_0(x);$$

- Критерий Андерсона-Дарлинга ( $\Omega^2$ -критерий) со статистикой

$$\Omega^2 = \int \frac{(\hat{F}_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x);$$

# Другие критерии согласия

- Критерий Фроцини

$$\int |\hat{F}_n(x) - F_0(x)| dF_0(x);$$

- Критерий Купера

$$\sup_{x \in \mathbb{R}} (\hat{F}_n(x) - F_0(x)) + \sup_{x \in \mathbb{R}} (F_0(x) - \hat{F}_n(x));$$

- Критерий Ватсона

$$\int \left( \hat{F}_n(x) - F_0(x) - \int (\hat{F}_n(x) - F_0(x)) dF_0(x) \right) dF_0(x).$$



# Свойства критериев согласия

- Распределение статистик критериев при верности нулевой гипотезы не зависит от  $F_0$ . Действительно, преобразование  $Y_i = F_0(X_i)$  переводит случайные величины  $X_i$  в равномерно распределенные на  $[0, 1]$ . Поэтому при верности нулевой гипотезы статистики критериев всегда сходятся к одному и тому же предельному закону.
- Все перечисленные критерии согласия являются состоятельными.

# Свойства критериев согласия

- Однако при ложности нулевой гипотезы критерии ведут себя по-разному. К сожалению, не существует критерия, который являлся бы р.н.м. критерием в данной задаче.
- Критерий Андерсона-Дарлинга показывает наилучшие результаты на большом классе распределений.
- Критерий Колмогорова, как правило, не очень хорошо ведет себя на альтернативе, в частности, плохо различает распределения, не совпадающие на хвостах.

# Проверка сложных гипотез

Гораздо более значимой с практической точки зрения является задача определения семейства, к которому принадлежит распределение выборки (например, задача проверки нормальности данных).

Но прежде чем проверять данные на принадлежность какому-то семейству распределений, нужно определить наиболее подходящее семейство, чтобы не перебирать критерии вслепую.

Первоначальная обработка статистических данных, как правило, осуществляется с помощью методов описательной (дескриптивной) статистики, таких, как гистограмма распределения, Box-plot и QQ-plot.

Построение гистограммы выборки, её Box-plot и QQ-plot нужны для определения примерного (модельного) распределения данных, после чего применяются более точные методы и критерии для проверки адекватности выбранной модели.

# Параметры сдвига и масштаба

Пусть имеется параметрическое семейство функций распределения  $\{F_\theta(x), \theta \in \Theta\}$ .

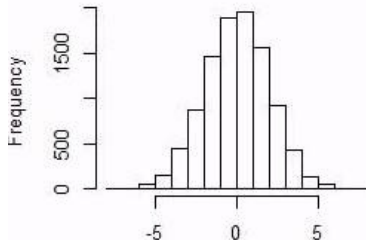
- Параметр  $\theta$  называется параметром сдвига, если  $F_\theta(x) = F_0(x - \theta)$ ;
- Параметр  $\theta$  называется параметром масштаба, если  $F_\theta(x) = F_1(x/\theta)$ ;
- Параметр, не являющийся параметром сдвига или масштаба, называется параметром формы.

К примеру, для семейства распределений  $N(a, \sigma^2)$  параметр  $a$  является параметром сдвига, а  $\sigma$  – параметром масштаба, оба параметра бета-распределения являются параметрами формы.

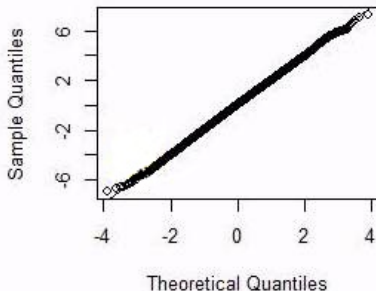
Допустим, мы хотим проверить гипотезу  $H_0 : F = F_0 \left( \frac{x-a}{\sigma} \right)$  (т.е. гипотеза о принадлежности семейству распределений с параметрами сдвига и масштаба).

QQ-plot – это график, на который нанесены точки  $\left( F_0^{-1} \left( \frac{i-0.5}{n} \right), X_{(i)} \right)$ . Если точки примерно лежат на одной прямой, то распределение данных близко к  $F_0$  (с точностью до параметров сдвига и масштаба).

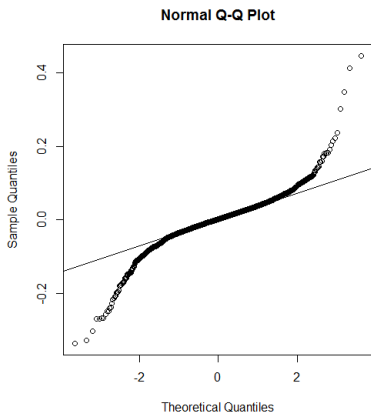
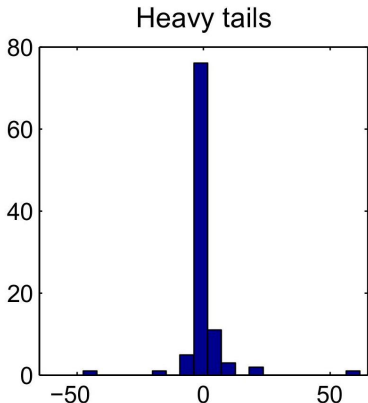
Symmetric distribution



Normal Q-Q Plot

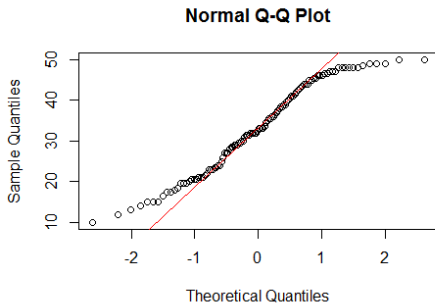
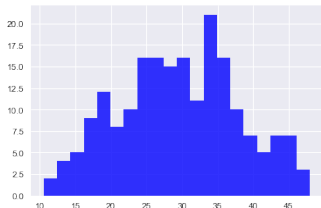


На рисунке: QQ-plot относительно нормального распределения (Normal QQ-plot) для выборки из нормального закона (среднее – 0, дисперсия  $\approx 6.5$ ). Так же QQ-plot может выглядеть для выборки из распределения Стьюдента с большим числом степеней свободы.

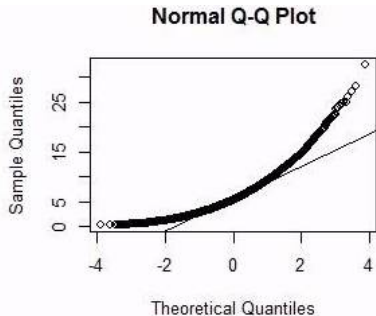
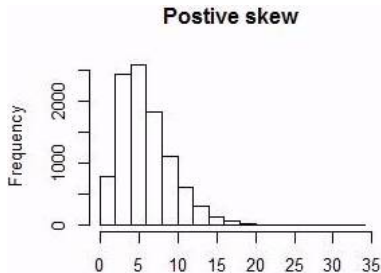


На рисунке: Normal QQ-plot для выборки с тяжелыми (относительно нормального закона) хвостами (heavy tails).  
Примеры: *Exp*, *Pareto*, *Gamma*, *LN*, *Cauchy* и тд.

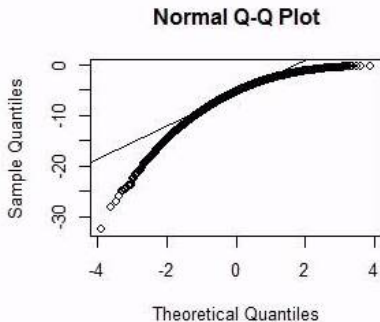
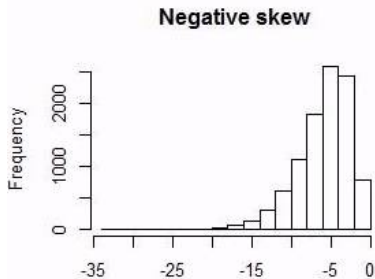




На рисунке: Normal QQ-plot для выборки с легкими (относительно нормального закона) хвостами (light tails).  
Примеры:  $Weibull(\alpha)$ ,  $\alpha > 2$ ,  $Gumbel$ , распределения с ограниченным носителем.



На рисунке: Normal QQ-plot для выборки из правостороннего распределения (right-skewed). Примеры: *Gamma*, *Poisson*, *Beta*( $\alpha, \beta$ ),  $\alpha < \beta$ .



На рисунке: Normal QQ-plot для выборки из левостороннего распределения (left-skewed). Примеры: *Inverse-Gamma*,  $Beta(\alpha, \beta)$ ,  $\alpha > \beta$ .

# Почему работает QQ-plot

Объясним, почему точки  $(F_0^{-1}(\frac{i-0.5}{n}), X_{(i)})$ ,  $i = 1, \dots, n$  должны лежать примерно на одной прямой, если  $\{X_i\}_{i=1}^n$  – выборка из  $F_0(\frac{x-a}{\sigma})$ .

Пусть  $\{Y_i\}_{i=1}^n$  – выборка из  $F_0(x)$ , тогда по теореме о выборочной квантили

$$Y_{(i)} - F_0^{-1}\left(\frac{i}{n}\right) \xrightarrow{P} 0, \quad n \rightarrow \infty,$$

т.е. при достаточно больших  $n$   $Y_{(i)} \approx F_0^{-1}\left(\frac{i-0.5}{n}\right)$ . Но  $\frac{X_{(i)}-a}{\sigma} \stackrel{d}{=} Y_{(i)}$ , откуда  $\frac{X_{(i)}-a}{\sigma} \stackrel{d}{=} Y_{(i)} \approx F_0^{-1}\left(\frac{i-0.5}{n}\right)$ , т.е. точки  $(F_0^{-1}(\frac{i-0.5}{n}), X_{(i)})$  должны лежать около прямой  $y = \frac{x-a}{\sigma}$ .

# Модификация критериев согласия

Один из методов проверки принадлежности распределения выборки какому-либо семейству распределений заключается в применении критериев согласия.

Пусть семейство функций распределения  $\{F(x; a, b)\}$  параметризовано параметрами сдвига и масштаба, т.е.  $F(x; a, b) = F_0(\frac{x-a}{b})$ , где  $F_0(x) = F(x; 0, 1)$ .

Пусть  $\hat{a}$  и  $\hat{b}$  – состоятельные оценки параметров  $a$  и  $b$  соответственно.

Тогда для проверки гипотезы  $H_0 : F \in \{F(x; a, b)\}$  можно воспользоваться критерием согласия, где вместо  $F_0$  в статистике критерия нужно подставить  $F_0(\frac{x-\hat{a}}{\hat{b}})$ .

# Критерий Лиллиефорса

В частности, если для семейства нормальных распределений  $\{N(a, \sigma^2)\}$  в качестве критерия согласия взять критерий Колмогорова-Смирнова со статистикой  $D_n = \sup_x |\hat{F}_n(x) - F_0(x)|$ , а в качестве оценок параметров  $a$  и  $\sigma^2$  – выборочное среднее и выборочную дисперсию соответственно, то получим **критерий Лиллиефорса** проверки нормальности выборки.

*Замечание.* Предельное распределение статистик модифицированных критериев согласия при верности нулевой гипотезы могут существенно зависеть не только от вида семейства распределений, но и от вида выбранных оценок параметров.

# Критерий Лиллиефорса

Таблица квантилей уровня  $1 - \alpha$  распределений Лиллиефорса и Колмогорова, при  $n > 30$  можно пользоваться ими.

Критические значения $\lambda_{\alpha 0}$ для распределений:	Уровень значимости $\alpha$			
	0,20	0,10	0,05	0,01
Распределение Колмогорова	1,073	1,224	1,358	1,627
Распределение Лиллиефорса	0,736	0,805	0,886	1,031

# Обобщенный критерий хи-квадрат

Обобщенный критерий хи-квадрат позволяет проверять гипотезы общего вида

$H_0 : F \in \mathcal{P}_0 = \{P_\theta, \theta \in \Theta\}$ , где  $\dim(\Theta) = d$ ,  
против альтернативы  $H_0 : F \notin \mathcal{P}_0$ .

Как и ранее, разобьем  $\mathbb{R}$  на несколько интервалов  $\{B_i\}_{i=1}^k$ ,  $k > d$ , и обозначим  $\mu_i = \#\{j : X_j \in B_i\}$ ,  
 $p_i(\theta) = P_\theta(B_i)$ , потребовав  $\forall \theta \in \Theta$  выполнения  
 $p_i(\theta) > c > 0$ .



# Обобщенный критерий хи-квадрат

## Теорема Фишера.

Пусть  $\Theta$  – открытое множество в  $\mathbb{R}^d$ , предположим, что матрица частных производных  $\left\| \frac{\partial p_i(\theta)}{\partial \theta_j} \right\|$  имеет ранг  $d$  для всех  $\theta \in \Theta$ . Пусть  $\hat{\theta}$  – ОМП по “выборке”  $\mu_1, \dots, \mu_k$ , т.е.

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \frac{n!}{\mu_1! \dots \mu_k!} \prod_{i=1}^k p_i^{\mu_i}(\theta),$$

или  $\hat{\theta}$  – оценка по минимуму хи-квадрат, т.е.

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \sum_{i=1}^k \frac{(\mu_i - np_i(\theta))^2}{np_i(\theta)}.$$

Тогда при условии верности  $H_0$  выполнено

$$\chi(\hat{\theta}) = \sum_{i=1}^k \frac{(\mu_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \xrightarrow{d} \chi_{k-d-1}^2, n \rightarrow \infty.$$

## Критерий Шапиро-Уилка

Проверяется гипотеза  $H_0 : F \in \{N(a, \sigma^2)\}$  против альтернативы  $H_1 : F \notin \{N(a, \sigma^2)\}$ . Статистика критерия

$$W = \frac{(\sum_{i=1}^n a_i X_{(i)})^2}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

где  $(a_1, \dots, a_n) = \frac{m^T V^{-1}}{\|m^T V^{-1}\|^2}$ ,  $(m_1, \dots, m_n)$  – мат. ожидания порядковых статистик выборки размера  $n$  из  $N(0, 1)$ , а  $V$  – матрица ковариаций этих порядковых статистик.

При верной гипотезе  $H_0$  статистика  $W$  имеет табличное распределение, значения вектора  $(a_1, \dots, a_n)$  также табулированы.

**Критерий Харке-Бера** тоже используется для проверки выборки на нормальность. Рассмотрим статистику

$$JB = \frac{n}{6}((Sk)^2 + \frac{1}{4}(Ku)^2),$$

где  $Sk = \frac{m_3}{m_2^{3/2}}$ ,  $Ku = \frac{m_4}{m_2^2} - 3$  и  $m_j = \frac{1}{n} \sum_i (X_i - \bar{X})^j$  – выборочный центральный момент.

При  $n > 2000$   $JB \approx \chi^2(2)$ , при меньших  $n$  следует искать квантили в таблицах. Модификацией критерия Харке-Бера является довольно точный комбинированный  $K^2$ -критерий.

# Проверка нормальности

Сравнение критериев проверки  
нормальности распределения случайных величин

Наименование критерия (раздел)	Характер альтернативного распределения					Ранг
	асимметричное		симметричное		≈ нормальное	
	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 \approx 3$	
Критерий Шапиро-Уилка (3.2.2.1)	1	1	3	2	2	1
Критерий $K^2$ (3.2.2.16)	7	8	10	6	4	2
Критерий Дарбина (3.1.2.7)	11	7	7	15	1	3
Критерий Д'Агостино (3.2.2.14)	12	9	4	5	12	4
Критерий $\alpha_4$ (3.2.2.16)	14	5	2	4	18	5
Критерий Васичека (3.2.2.2)	2	14	8	10	10	6
Критерий Дэвида-Хартли-Пирсона (3.2.2.10)	21	2	1	9	1	7
Критерий $\chi^2$ (3.1.1.1)	9	20	9	8	3	8
Критерий Андерсона-Дарлинга (3.1.2.4)	18	3	5	18	7	9
Критерий Филлибена (3.2.2.5)	3	12	18	1	9	10
Критерий Колмогорова-Смирнова (3.1.2.1)	16	10	6	16	5	11
Критерий Мартинеса-Иглевица (3.2.2.14)	10	16	13	3	15	12
Критерий Лина-Мудхолкара (3.2.2.13)	4	15	12	12	16	13
Критерий $\alpha_3$ (3.2.2.16)	8	6	21	7	19	14
Критерий Шпигельхальтера (3.2.2.11)	19	13	11	11	8	15
Критерий Саркади (3.2.2.12)	5	18	15	14	13	16
Критерий Смирнова-Крамера-фон Мизеса (3.1.2.2)	17	11	20	17	6	17
Критерий Локка-Спурье (3.2.2.7)	13	4	19	21	17	18
Критерий Оя (3.2.2.8)	20	17	14	13	14	19
Критерий Хегази-Грина (3.2.2.3)	6	19	16	19	21	20
Критерий Муроты-Такеучи (3.2.2.17)	15	21	17	20	20	21

Кобзарь, 3.2.2.19, табл. 80.

Допустим, мы хотим оценить параметр  $\theta$  с помощью асимптотически нормальной оценки  $\hat{\theta}_n$ ,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d_\theta} N(0, \sigma^2(\theta)), n \rightarrow \infty.$$

Но реальная дисперсия  $\hat{\theta}_n$  при фиксированном  $n$  может сильно отличаться от  $\sigma^2(\theta)/n$ , т.е. вероятность покрытия асимптотическим доверительным интервалом истинного значения  $\theta$  может быть гораздо ниже заявленной.

Пусть  $X_1, \dots, X_n$  – выборка из неизвестной ф.р.  $F$ . Тогда возьмем в качестве нового выборочного пространства значения  $(X_1, \dots, X_n)$  и положим каждому значению вероятность  $1/n$ .

Полученное дискретное распределение будет иметь функцию распределения  $\hat{F}_n$ . Бутстреп – это семплирование одной или нескольких выборок из  $\hat{F}_n$ . Поскольку  $\hat{F}_n$  при больших  $n$  близка к  $F$ , то бутстрепная выборка будет похожа на выборку из ф.р.  $F$ .

# Бутстрепная оценка дисперсии

Пусть имеется некоторая статистика

$T_n(X) = T_n(X_1, \dots, X_n)$ , найдем оценку её дисперсии с помощью бутстрепа.

Семплируем  $m$  бутстрепных выборок

$\{X_{i,1}^*\}_{i=1}^n, \dots, \{X_{i,m}^*\}_{i=1}^n$  из совокупности  $\{X_1, \dots, X_n\}$  и найдем  $T_{n,j}^* = T_n(X_{1,j}^*, \dots, X_{n,j}^*) \forall j \in \{1 \dots m\}$ .

Тогда оценка методом бутстрепа дисперсии статистики  $T_n(X)$  будет равна

$$v_{boot}(T) = \frac{1}{m} \sum_{j=1}^m \left( T_{n,j}^* - \frac{1}{m} \sum_{j=1}^m T_{n,j}^* \right)^2.$$

# Нормальный интервал

Предположим, что  $T_n(X)$  оценивает некий параметр  $\theta$ , также предположим, что распределение  $T_n(X)$  близко к нормальному.

Тогда можно предложить следующий доверительный интервал уровня доверия  $1 - \alpha$  для параметра  $\theta$  :

$$\left( T_n + z_{\alpha/2} \sqrt{v_{boot}(T)}, T_n + z_{1-\alpha/2} \sqrt{v_{boot}(T)} \right),$$

где  $z_\gamma$  —  $\gamma$ -квантиль  $N(0, 1)$ .



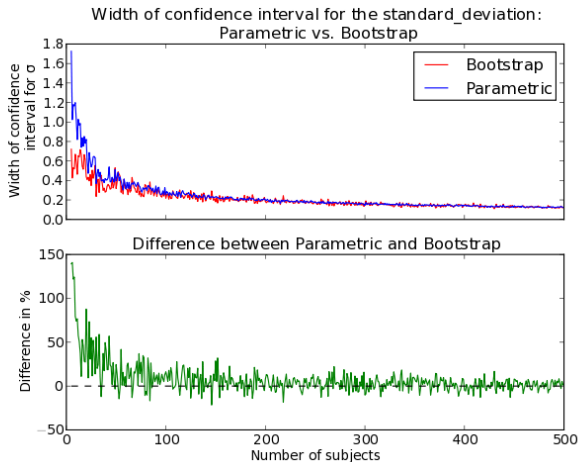
# Центральный интервал

Заметим, что поскольку  $T_n(X)$  является оценкой для  $\theta$ , а  $T_{n,1}^*$  — для  $T_n(X)$ , то  $T_{n,1}^* - T_n$  оценивает  $T_n - \theta$ .

Образует вариационный ряд совокупности  $\{T_{n,1}^*, \dots, T_{n,m}^*\} : T_{(1)}^* \leq T_{(2)}^* \dots \leq T_{(m)}^*$ .

Тогда с большой вероятностью  $T_n(X) - \theta$  лежит в интервале  $(T_{([m\alpha])}^* - T_n, T_{([m(1-\alpha)])}^* - T_n)$ , т.е. получаем следующий доверительный интервал для  $\theta$

$$\left(2T_n - T_{([m(1-\alpha)])}^*, 2T_n - T_{([m\alpha])}^*\right).$$



Сравнение длин асимптотического и бутстрепного доверительного интервала для оценки стандартного отклонения.

Предшественник бутстрепа – метод складного ножа. Пусть имеется выборка  $(X_1, \dots, X_n)$  и статистика  $T_n(X) = T_n(X_1, \dots, X_n)$ . Обозначим

$$\begin{aligned}\hat{T}_1 &= T_{n-1}(X_2, \dots, X_n); \\ \hat{T}_2 &= T_{n-1}(X_1, X_3, \dots, X_n); \dots \\ \hat{T}_n &= T_{n-1}(X_1, \dots, X_{n-1}).\end{aligned}$$

Если  $T_n$  – оценка параметра  $\theta$ , то для оценки смещения  $Bias = T_n - \theta$  можно использовать оценку

$$(n-1) \left( \frac{1}{n} \sum_{i=1}^n \hat{T}_i - T_n \right),$$

Также можно оценить дисперсию  $T_n$  следующим образом

$$v_{jack} = \frac{m-1}{m} \sum_{i=1}^n \left( \hat{T}_i - \frac{1}{n} \sum_{i=1}^n \hat{T}_i \right)^2.$$

# Finita!