

Домашнее задание

8.1 Sample C.I. (Доверительный интервал для значений или оценок квантилей)

Рассмотрим следующую задачу: пусть дана выборка X_1, \dots, X_N из некоторого распределения. Требуется построить доверительный интервал для возможных новых значений X из того же распределения (т.е. фактически, оценить квантили неизвестного распределения по выборке). В реальной жизни постановка может быть следующая: по нескольким наблюдениям нужно оценить диапазон значений измеряемой величины в популяции. Рассмотрим ниже два возможных подхода.

Непараметрический подход В качестве оценок квантилей возьмем выборочные квантили X (напомню, что в R это делается функцией `quantile()`). Подход подкупает своей простотой и обоснованностью, но плохо работает для большинства распределений. В частности, для нормального распределения легко показать, что выборочные квантили тем менее устойчивы, чем ближе они к краю. Для этого мы воспользуемся асимптотической нормальностью квантилей:

$$\sqrt{N} \cdot (X^{[p]} - \xi^{(p)}) \xrightarrow{D} \mathcal{N} \left(0, \frac{p(1-p)}{\rho_{\xi}^2(\xi^{(p)})} \right),$$

где $X^{[p]}$ — выборочный, а $\xi^{(p)}$ — реальный квантиль уровня p . Немного посчитаем (для стандартного нормального):

```
> avarqn <- function(p) p * (1 - p) / dnorm(qnorm(p))^2
> avarqn(c(0.5, 0.75, 0.95, 0.975, 0.99, 0.995))
[1] 1.570796 1.856767 4.465561 7.135902 13.937053 23.794245
```

Для произвольного нормального дисперсия оценки будет в σ^2 раз больше по линейности.

При одинаковом порядке сходимости скорость сходимости отличается в разы. Таким образом, оценка квантилей на хвосте нормального распределения (да и любого распределения “с хвостом”, на самом деле) с помощью выборочных квантилей требует очень большого объема выборки (понятие об ARE подсказывает нам, что для оценки 99.5%-го квантиля требуется выборка в 15 раз большая, чем для оценки медианы с той же точностью). Нам явно нужно другое решение.

Параметрический подход Предположим, что наше распределение является нормальным. Часто это предположение оказывается верным, хотя бы приближенно. Для нормального распределения мы можем оценить параметры a и σ^2 ¹ и построить *параметрические выборочные квантили*, т.е. квантили распределения в известной модели с оцененными по выборке параметрами. В случае, когда предположение о нормальности исходного распределения верно, эти квантили оценивают реальные квантили исходного распределения:

$$\hat{q}_p(X) = \hat{a}(X) + z_p \hat{\sigma}(X) \approx \xi^{(p)},$$

где z_p — p -квантиль стандартного нормального распределения.

¹Мы можем использовать как обычные выборочные среднее и дисперсию, так и робастные оценки

Для случая оценивания параметров с помощью выборочного среднего и выборочной дисперсии, воспользовавшись теоремой о сохранении асимптотической нормальности при гладких преобразованиях, а также тем, что для нормальной выборки выборочное среднее независимо с выборочной дисперсией, можно получить следующий факт (для стандартного нормального):

$$\sqrt{N} \cdot (\hat{q}_p(X) - \xi^{(p)}) \xrightarrow{D} \mathcal{N} \left(0, 1 + \frac{z_p^2}{2} \right)$$

Для произвольного нормального дисперсия оценки будет в σ^2 раз больше по линейности.

Можно показать, что такие оценки квантилей являются асимптотически наиболее эффективными, т.е. с точки зрения ARE лучше ничего в принципе придумать нельзя.

Понятно, что недостаток параметрического подхода в том, что мы можем ошибиться при выборе параметрической модели. В общем случае исследуемое распределение является нормальным. Но на практике часто оказывается лучше использовать параметрический подход с грубой (обычно используется нормальная модель) моделью, чем непараметрический, потому что лучше грубая, но устойчивая оценка. Например, для объема выборки $N = 10$, применение выборочных квантилей затруднительно (непонятно даже, как вообще считать 2.5%-й квантиль), а параметрический подход дает вполне разумные (пусть и грубые) результаты, даже если распределение исходно не было нормальным.

Если в коде, то это все выглядит так:

```
> N <- 100
> x <- rnorm(N, mean = 3, sd = 5)

# Real quantiles
> qnorm(c(0.025, 0.975), mean = 3, sd = 5)
[1] -6.79982 12.79982

# Non-parametric
> quantile(x, probs = c(0.025, 0.975))
      2.5%      97.5%
-5.356409 11.910011

# Parametric
> qnorm(c(0.025, 0.975), mean = mean(x), sd = sd(x))
[1] -6.480048 11.398967

# ARE
> avarqn <- function(p) p * (1 - p) / dnorm(qnorm(p))^2
> avarqp <- function(p) 1 + qnorm(p)^2/2
> ARE <- function(p) avarqn(p) / avarqp(p)
> curve(ARE)
```

Хотя параметрический подход является более точным при одиночном испытании этого, конечно, увидеть нельзя, однако, это можно продемонстрировать с помощью моделирования.

Задача (Построение доверительного интервала для выборки). Для нормального распределения и для распределения t-Стюдента со степенями свободы 5 и 20 для объемов выборки 10, 100, 1000, 10000 построить доверительные интервалы для значения (95%-е и 99%-е) непараметрически и параметрически с нормальной моделью (для Стюдента тоже надо использовать нормальную модель, так мы изучим свойства параметрических д.и. при неточной модели). Исследовать с помощью моделирования эмпирический уровень доверия и длину доверительного интервала, а также СКО (RMSE) самих оценок (истинные значения квантилей известны). Число независимых серий M возьмите 1000–10000.

При выводе оценок (например, длины интервала или его уровня доверия) ограничивайте число цифр, выводите только реально известные. Рядом с оценкой в скобках желательно писать оценку ее СКО (напомню, что если оценка считается как среднее, то ее СКО можно оценить как $sd(x)/\sqrt{M}$).

8.2 МССІ: Монте-Карло доверительные интервалы

Пусть дана выборка X_1, \dots, X_N из обобщенного распределения Коши, т.е. распределения с плотностью

$$f(x) = \frac{1}{\pi s \left(1 + \frac{(x-a)^2}{s^2}\right)}$$

и нам нужно оценить параметры a и s и построить доверительные интервалы.

Как мы узнали раньше, выборочная медиана и выборочный МАД являются хорошими оценками параметров сдвига и масштаба, но как построить на их основе доверительные интервалы?

Попробуем использовать стандартный подход к построению д.и. Рассмотрим некоторую функцию от выборки, которая содержит явно неизвестный параметр, но при этом не зависит от него. Например, для сдвига можно рассмотреть $\mu_N = \frac{a - \text{median}(X)}{\text{mad}(X)}$, а для масштаба $\kappa_N = \frac{s}{\text{mad}(X)}$. Это случайные величины, их распределение не зависит от a, s , а зависит только от длины выборки N .

Преобразовав неравенства, мы получим доверительные интервалы для a и s :

$$\begin{aligned} a &\in \left(\text{median}(X) + \mu_{(1-\gamma)/2, N} \cdot \text{mad}(X), \text{median}(X) + \mu_{(1+\gamma)/2, N} \cdot \text{mad}(X) \right); \\ s &\in \left(\kappa_{(1-\gamma)/2, N} \cdot \text{mad}(X), \kappa_{(1+\gamma)/2, N} \cdot \text{mad}(X) \right), \end{aligned}$$

где $\mu_{p, N}$ и $\kappa_{p, N}$ — квантили уровня p для распределений μ_N и κ_N соответственно.

К сожалению, мы не знаем точного вида этих распределений (хотя и можно показать, что они асимптотически нормальные), а значит такие д.и. на практике неприменимы.

Но, хоть мы и не знаем точный вид этих распределений, можем их моделировать, явно моделируя выборки длины N с известными параметрами и вычисляя значения μ и κ . С помощью моделирования мы можем оценить квантили интересующих нас уровней с любой наперед заданной точностью². Подставив оценки в формулы для доверительных интервалов, мы получим приближенные (но сколь угодно точные) доверительные интервалы, пригодные для практического использования.

Задача. Для обобщенного распределения Коши построить Монте-Карло доверительные интервалы для параметров сдвига и масштаба на основе выборочной медианы и

²Тут надо вспомнить предыдущее задание — чем ближе квантиль к краю, тем хуже он оценивается

выборочного MAD для уровня $\gamma = 0.95$ для выборок длины $N = 10, 100, 1000, 10000$. Использовать два способа оценивания квантилей — непараметрический (через выборочные квантили) и параметрический с нормальной моделью, моделируя для оценивания квантилей $M = 10000$ реализаций μ_N, κ_N . Проверить, являются ли полученные доверительные интервалы честными, т.е. соответствует ли реальный уровень значимости для них заданному.