1 Точные и асимптотические нормальные доверительные интервалы

Здесь я написал небольшой конспект по классическим нормальным доверительным интервалам, который поможет вам хорошо написать контрольную =^.^=

Во всех примерах имеется ввиду, что дана выборка $X_1, \dots X_N$ из некоторого распределения, при этом мы хотим найти доверительные интервалы для среднего a или дисперсии σ^2 этого распределения (при этом мы можем иметь некоторую информацию о втором параметре, а можем и не иметь). γ везде обозначает требуемый уровень доверия.

1.1 Доверительный интервал для среднего

1.1.1 Случай известной дисперсии

Пусть нужно найти доверительный интервал для среднего в нормальной модели (т.е. известно, что исследуемое распределение нормальное), и пусть дисперсия σ^2 известна. Тогда воспользуемся следующим фактом:

$$\sqrt{N} \cdot \frac{a - \overline{X}}{\sigma} \sim \mathcal{N}(0, 1).$$

Следовательно, с вероятностью γ

$$a \in \left(\overline{X} \mp \frac{x_\gamma \cdot \sigma}{\sqrt{N}} \right),$$

где $x_\gamma = z_{(\gamma+1)/2}$ — квантиль уровня $(1+\gamma)/2$ стандартного нормального распределения.

В коде:

```
> mean(x) + qnorm(c(0.025, 0.975)) * sigma / sqrt(length(x))
```

Или (абсолютно эквивалентно, по линейности нормального распределения):

```
> qnorm(c(0.025, 0.975), mean = mean(x), sd = sigma / sqrt(length(x)))
```

1.1.2 Случай неизвестной дисперсии

Если же дисперсия неизвестна, то можно использовать ее стандартную оценку, т.е. выборочную дисперсию. Известен следующий факт:

$$\sqrt{N-1} \cdot \frac{a - \overline{X}}{\widehat{\sigma}(X)} = \sqrt{N} \cdot \frac{a - \overline{X}}{\widehat{s}(X)} \sim \text{t-Student}(df = N - 1),$$

где $\widehat{s}(X) = \sqrt{\frac{N}{N-1}} \cdot \widehat{\sigma}(X)$ — подправленное выборочное SD. Следовательно:

$$a \in \left(\overline{X} \mp \frac{t_{(\gamma+1)/2, N-1} \cdot \widehat{s}(X)}{\sqrt{N}} \right),$$

где $t_{p,N-1}$ — квантиль уровня p распределения t-Стюдента с $N-1$ степенью свободы. В R:

```
> mean(x) + qt(c(0.025, 0.975), df = length(x) - 1) * sd(x) / sqrt(length(x))
```

Также можно использовать функцию `t.test()`. Если исходное распределение является приближенно нормальным, то такой доверительный интервал также можно использовать (но он будет также приближенным).

1.1.3 Асимптотический доверительный интервал для среднего для произвольного распределения

Для случая, когда исходное распределение не является нормальным, можно использовать центральную предельной теорему:

$$\sqrt{N} \cdot \frac{a - \overline{X}}{s} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

где s — либо известное SD, либо его состоятельная оценка (например `sd(x)` — подправленное выборочное SD). Тогда модно построить следующий (асимптотический) д.и.:

$$a \in \left(\overline{X} \mp \frac{x_\gamma \cdot s}{\sqrt{N}} \right).$$

Здесь $x_\gamma = z_{(\gamma+1)/2}$ — соответствующий квантиль нормального распределения. В R:

```
> mean(x) + qnorm(c(0.025, 0.975), sd = sd(x) / sqrt(length(N)))
```

Такой доверительный интервал для среднего является наиболее стандартным и используется повсеместно. Также можно заметить, что асимптотически он эквивалентен предыдущему.

1.2 Доверительный интервал для дисперсии

1.2.1 Случай неизвестного среднего

Пусть истинное среднее неизвестно (типичный случай). Воспользуемся следующим фактом (это частный случай так называемой теоремы Фишера):

$$\frac{N \cdot \widehat{\sigma}^2(X)}{\sigma^2} = \frac{(N-1) \cdot \widehat{s}^2(X)}{\sigma^2} \sim \chi^2(df = N - 1),$$

тогда:

$$\sigma^2 \in \left(\frac{(N-1) \cdot \widehat{s}^2(X)}{\chi^2_{(1+\gamma)/2, N-1}}, \frac{(N-1) \cdot \widehat{s}^2(X)}{\chi^2_{(1-\gamma)/2, N-1}} \right),$$

где $\chi^2_{p,N-1}$ — квантили уровня p распределения хи-квадрат с $N-1$ степенью свободы.

В коде:

```
> var(x) * (length(x) - 1) / qchisq(c(0.975, 0.025), df = length(x) - 1)
```

Если же среднее известно (как было сказано, это задача оценки погрешности прибора при калибровке), то можно использовать следующий (существенно более очевидный, чем предыдущий, кстати говоря) факт:

$$\frac{N \cdot \widehat{\sigma^*}^2(X)}{\sigma^2} \sim \chi^2(df = N),$$

где $\widehat{\sigma^*}^2(X) = \overline{(X - a)^2}$ — выборочная дисперсия при известном среднем. Тогда:

$$\sigma^2 \in \left(\frac{N \cdot \widehat{s^*}^2(X)}{\chi^2_{(1+\gamma)/2, N}}, \frac{N \cdot \widehat{s^*}^2(X)}{\chi^2_{(1-\gamma)/2, N}} \right),$$

В коде:

```
> mean((x - a)^2) * length(x) / qchisq(c(0.975, 0.025), df = length(x))
```

Что же делать в случае, когда исходное распределение ненормальное? Можно воспользоваться следующим утверждением (асимптотическая нормальность выборочных моментов):

$$\sqrt{N} \cdot \left(\widehat{M_k}(X) - M_k\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, M_{2k} - M_k^2),$$

где M_k — k-й момент распределения X, $\widehat{M_k}(X)$ — выборочный k-й момент распределения X. Это утверждение верно при условии, что M_{2k} момент существует и конечен (иначе порядок сходимости выборочного момента будет хуже, чем стандартный) и $M_{2k} - M_k^2 > 0$ (иначе выборочный момент будет сверхсходиться).

Напомню определение *эксцесса:*

$$\gamma_4 = \frac{M_4}{\sigma^4} - 3 = \frac{M_4}{M_2^2} - 3.$$

Таким образом:

$$\sqrt{N} \cdot (\sigma^2 - \widehat{\sigma}^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma), \text{где } \Sigma = M_4 - M_2^2 = (\gamma_4 + 3)M_2^2 - M_2^2 = (\gamma_4 + 2)M_2^2.$$

Итого (подставив вместо неизвестного M_2 оценку дисперсии):

$$\sigma^2 \in \left(\widehat{\sigma}^2(X) \mp x_\gamma \sqrt{\frac{\gamma_4 + 2}{N}} \cdot \widehat{\sigma}^2(X)\right).$$

Если точное значение эксцесса неизвестно, можно использовать и вместо него состоятельную оценку (например, выборочный эксцесс). Вычислить значение выборочного эксцесса в R можно с помощью функции `kurtosis()` из пакета `e1071` (`library(e1071)`, нужно предварительно установить через `install.packages("e1071")`).

```
> library(e1071)
> var(x) + qnorm(c(0.025, 0.975), sd = sqrt((kurtosis(x)+2)/length(x)) * var(x))
```

2 Доверительные интервалы для пропорции (доли)

TBD

3 Доверительный интервал для значений или оценка квантилей

Рассмотрим следующую задачу: пусть дана выборка $X_1, \ldots, X_N$ 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:

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

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:

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

где z_p — p-квантиль стандартного нормального распределения.

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):

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

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.

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

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.

Задание 3.1 (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 $sd(x)/\sqrt{M}$)

4 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 MAD 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 $\mu_N = \frac{a - \mathrm{median}(X)}{\mathrm{mad}(X)}$, а для масштаба $\kappa_N = \frac{s}{\mathrm{mad}(X)}$. 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:

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

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 асимтотически 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 точностью[2]. Подставив оценки в формулы для доверительных интервалов, мы получим приближенные (но сколь угодно точные) доверительные интервалы, пригодные для практического использования.

[2]Тут надо вспомнить предыдущий раздел — чем ближе квантиль к краю, тем хуже он оценивается