Липецкий государственный технический университет

Кафедра прикладной математики

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ МАТЕМАТИЧЕСКИХ ИССЛЕДОВАНИЙ

Лекция 12

12. Бутстреп и статистическое оценивание выборочных характеристик

Составитель - Сысоев А.С., к.т.н., доц.

Липецк – 2021

Outline

12.1 Непараметрические методы статистики и ресамплинг

12.2 Складной нож и бутстреп - механизмы генерации случайных псевдовыборок

12.3 Перестановочные тесты

12.4 Бутстреп-анализ при помощи пакета boot в R

12.1 Непараметрические методы статистики и ресамплинг

Непараметрическими называют такие методы статистики, которые не зависят от какого-нибудь распределения из теоретического семейства и не используют его свойства.

В том случае, когда нет возможности получить истинные повторности наблюдений, разработаны методы, которые формируют большое количество *"псевдовыборок"*, и на их основе можно получить необходимые характеристики искомого параметра.

рандомизация, или перестановочный тест (permutation),

бутстреп (bootstrap),

метод "складного ножа" (jackknife),

кросс-проверка (cross-validation).

12.2 Складной нож и бутстреп - механизмы генерации случайных псевдовыборок

Традиционные параметрические методы позволяют оценить ошибку среднего как

$$s_m = s/\sqrt{n} = \left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$$

Оценка для ошибки среднего, вычисленная методом складного ножа

$$\hat{\sigma}_{JACK} = \left[\frac{n-1}{n}\sum_{i=1}^{n}(\tilde{x}_{(i)} - \bar{x}_{(\bullet)})^2\right]^{1/2}$$

Популярность метода "складного ножа" с его недостаточно интенсивным вычислительным подходом при анализе выборочных оценок параметров существенно снизилась в ходе развития идей **бутстрепа**, когда появилась возможность гибкой настройки и использование алгоритмов самоорганизации.

Бутстреп-процедура (bootstrap) была предложена как некоторое обобщение алгоритма "складного ножа", чтобы не уменьшать каждый раз число элементов по сравнению с исходной совокупностью.

12.2 Складной нож и бутстреп - механизмы генерации случайных псевдовыборок

Основная идея бутстрепа состоит в том, чтобы методом статистических испытаний Монте-Карло многократно извлекать повторные выборки из эмпирического распределения.

Берется конечная совокупность из n членов исходной выборки x_1, x_2, …, x_{n-1}, x_n , откуда на каждом шаге из n последовательных итераций с помощью датчика случайных чисел, равномерно распределенных на интервале [1, n], "вытягивается" произвольный элемент x_k, который снова "возвращается" в исходную выборку (т.е. может быть извлечен повторно). Можно сформировать любое, сколь угодно большое число бутстреп-выборок (обычно 5000-10000). На основе разброса значений анализируемого показателя, полученного в процессе имитации, можно построить, например, доверительные интервалы оцениваемого параметра.

Бутстреп, как и иные методы генерации повторных выборок, полезны, когда статистические выводы нельзя получить с использованием теоретических предположений (например, какие-либо предположения сделать трудно из-за недостаточного объема выборок).

В зависимости от имеющейся информации относительно статистической модели генеральной совокупности различают *непараметрический* и *параметрический бутстреп*.

НЕПАРАМЕТРИЧЕСКАЯ БУТСТРЕП-ПРОЦЕДУРА

Шаг 1: Получение большого количества повторностей – случайных наборов данных из изучаемой совокупности. При этом используется алгоритм "случайного выбора с возвращением" (random sampling with replacement), т.е. извлеченное число снова помещается в "перемешиваемую колоду" прежде чем вытягивается следующее наблюдение.

Шаг 2: Построение бутстреп-распределения оцениваемой величины.

Если простой *непараметрический бутстреп* выполняет перевыборку с учетом равной вероятности появления каждого элемента, то *стратифицированный бутстреп* учитывает соотношение частот между относительно гомогенными группами (стратами), на которые может быть разделены выборочные объекты.

ПАРАМЕТРИЧЕСКАЯ БУТСТРЕП-ПРОЦЕДУРА

Шаг 1: По выборочным данным $\{x_1, x_2, \ldots, , x_n\}$ осуществляется построение вероятностной модели и оцениваются ее параметры.

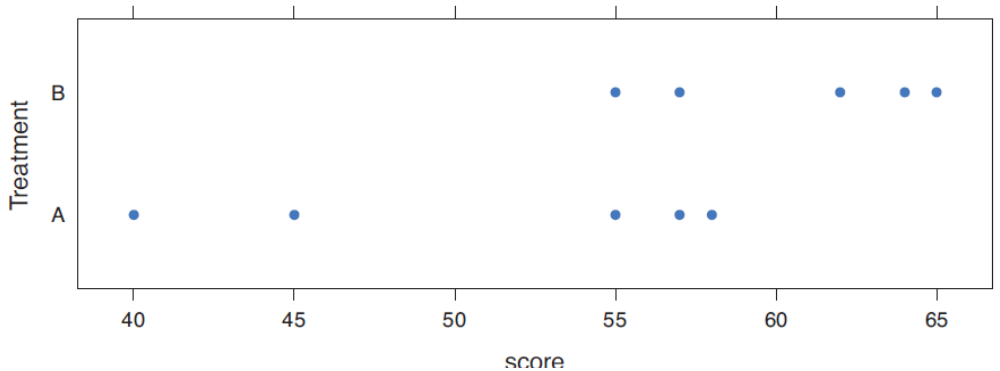
Шаг 2: Случайным образом из подобранного распределения генерируются n элементов $\{x^*_1, x^*_2, \ldots, x^*_n\}$ и бутстреп-повторность, полученная такой имитацией, используется для расчета значения статистики $t^* = T(x^*)$.

Шаг 3: Шаг 2 выполняется B раз и формируется бутстреп-распределение анализируемой статистики $\{t^*_1, t^*_2, \ldots, t^*_j, \ldots, t^*_B\}$.

12.3 Перестановочные тесты

Задача: Десять объектов были случайно подвергнуты одному из двух воздействий (A или B), а затем были зарегистрированы значения результирующей переменной (`score`). Достаточны ли эти данные для того, чтобы заключить, что подобные воздействия приводят к разным результатам?

Воздействие A	Воздействие B
40	57
57	64
45	55
55	62
58	65



1. Вычислить выборочную t-статистику как при параметрическом подходе, t_0.
2. Поместить все 10 значений в одну группу.
3. Случайно поместить пять значений в группу A и пять – в группу B.
4. Вычислить новую t-статистику.
5. Повторить шаги 3–4 для всех возможных способов размещения пяти значений в группу A и пяти – в группу B. Всего существует 252 способа.
6. Расположить 252 значения t-статистики в порядке возрастания. Это эмпирическое распределение, основанное на выборках.
7. Если t_0 не входит в центральные 95% значений эмпирического распределения, то следует с вероятностью 95% отвергнуть нулевую гипотезу о равенстве средних значений в двух группах.

12.3 Перестановочные тесты

Тест	Функция пакета `coin` *
Перестановочный тест для двух и *k* выборок	`oneway_test(y ~ A)`
Перестановочный тест для двух и *k* выборок с группирующим фактором	`oneway_test(y ~ A | C)`
Тест ранговых сумм Вилкоксона-Манна-Уитни	`wilcox_test(y ~ A)`
Тест Краскела-Уоллиса	`kruskal_test(y ~ A)`
Хи-квадрат тест Пирсона	`chisq_test(A ~ B)`
Тест Кохрана-Мантеля-Гензеля	`cmh_test(A ~ B | C)`
Критерий линейной зависимости (Linear-by-linear association test)	`lbl_test(D ~ E)`
Тест Спирмена	`spearman_test(y ~ x)`
Тест Фридмана	`friedman_test(y ~ A | C)`
Ранговый тест Вилкоксона	`wilcoxsign_test(y1 ~ y2)`

* `y` и `x` – числовые переменные, `A` и `B` – категориальные переменные, `C` – категориальная группирующая переменная, `D` и `E` – упорядоченные факторы, `y1` и `y2` – парные числовые переменные.

12.3 Перестановочные тесты

```
название_функции (formula, data= , distribution= ),
```

• `formula` описывает взаимосвязь между переменными, которую нужно проверить;

• `data` задает таблицу с данными;

• `distribution` указывает, как должно быть создано эмпирическое распределение для нулевой гипотезы. Возможные значения: `exact`, `asymptotic` и `approximate`.

Тесты на независимость для двух и k выборок

Независимость в таблицах сопряженности

Независимость между числовыми переменными

Тесты для двух и k зависимых выборок

12.4 Бутстреп-анализ при помощи пакета boot в R

В общем случае бутстреп-анализ состоит из трех этапов:

1. Написать функцию, которая вычисляет нужную статистику или нужные статистики. Если имеется одна статистика (например, медиана), функция должна возвращать число. Если есть набор статистик (например, набор регрессионных коэффициентов), функция должна возвращать вектор.

2. Применить функцию `boot()` к функции, чтобы создать бутстреп-повторности статистики или статистик.

3. Использовать функцию `boot.ci()`, чтобы вычислить доверительные интервалы для статистики или статистик, созданных на этапе 2.

Основная функция для бустреп-анализа `boot()`

```
bootobject <- boot(data=, statistic=, R=, ...)
```

Параметр	Описание
`data`	Вектор, матрица или таблица данных
`statistic`	Функция, которая создает *k* статистик, которые будут подвергнуты бутстреп-анализу (*k*=1, если есть только одна статистика)
`R`	Число бутстреп-выборок
`...`	Дополнительные параметры функции, которая создает нужную статистику или нужные статистики

12.4 Бутстреп-анализ при помощи пакета boot в R

```
boot.ci(bootobject, conf=, type= )
```

Параметр	Описание
bootobject	Объект, создаваемый функцией `boot()`
`conf`	Требуемый доверительный интервал (по умолчанию `conf=0.95`)
`type`	Тип вычисляемого доверительного интервала. Возможны следующие значения: `"norm"`, `"basic"`, `"stud"`, `"perc"`, `"bca"` и `"all"` (по умолчанию `type="all"`)

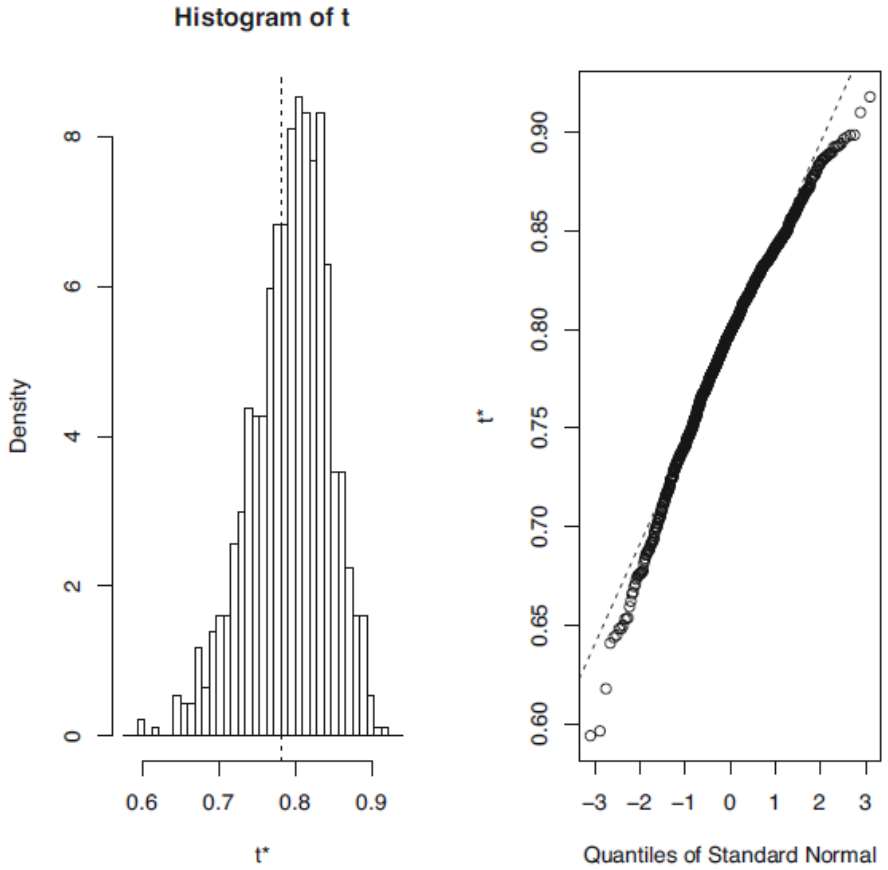
БУТСТРЕП-АНАЛИЗ ДЛЯ ОДНОЙ СТАТИСТИКИ

Используется множественная регрессия для предсказания расхода топлива по весу машины и по рабочему объему цилиндров двигателя.

```
rsq <- function(formula, data, indices) {
         d <- data[indices,]
         fit <- lm(formula, data=d)
         return(summary(fit)$r.square)
}

library(boot)
set.seed(1234)
results <- boot(data=mtcars, statistic=rsq,
                R=1000, formula=mpg~wt+disp)
```

БУТСТРЕП-АНАЛИЗ ДЛЯ ОДНОЙ СТАТИСТИКИ (ПРОДОЛЖЕНИЕ)




```
> boot.ci(results, type=c("perc", "bca"))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
CALL :
boot.ci(boot.out = results, type = c("perc", "bca"))
Intervals :
Level     Percentile            BCa

 95%   ( 0.6838,  0.8833 )   ( 0.6344,  0.8549 )
 Calculations and Intervals on Original Scale
 Some BCa intervals may be unstable
```

12.4 Бутстреп-анализ при помощи пакета boot в R

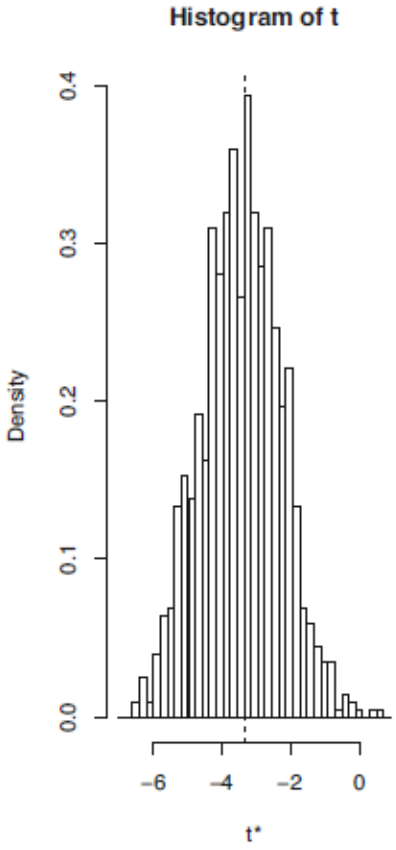
БУТСТРЕП-АНАЛИЗ ДЛЯ НЕСКОЛЬКИХ СТАТИСТИК

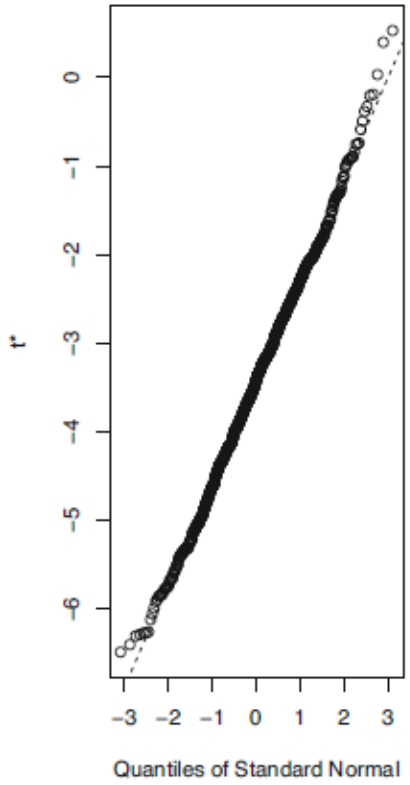
Расчет доверительных интервалов для трех коэффициентов регрессионной модели (свободный член, вес автомобиля и объем цилиндров).

```
bs <- function(formula, data, indices) {
        d <- data[indices,]
        fit <- lm(formula, data=d)
        return(coef(fit))
}
library(boot)
set.seed(1234)
results <- boot(data=mtcars, statistic=bs,
                R=1000, formula=mpg~wt+disp)
print(results)
ORDINARY NONPARAMETRIC BOOTSTRAP
Call:
boot(data = mtcars, statistic = bs, R = 1000, formula = mpg ~
    wt + disp)
Bootstrap Statistics :
    original   bias    std. error
t1*  34.9606  0.137873     2.48576
t2*  -3.3508 -0.053904     1.17043
t3*  -0.0177 -0.000121     0.00879
```

12.4 Бутстреп-анализ при помощи пакета boot в R

БУТСТРЕП-АНАЛИЗ ДЛЯ НЕСКОЛЬКИХ СТАТИСТИК (ПРОДОЛЖЕНИЕ)






```
> boot.ci(results, type="bca", index=2)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = results, type = "bca", index = 2)

Intervals :
Level       BCa
95%   (-5.66, -1.19 )
Calculations and Intervals on Original Scale

> boot.ci(results, type="bca", index=3)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = results, type = "bca", index = 3)

Intervals :
Level       BCa
95%   (-0.0331,  0.0010 )
Calculations and Intervals on Original Scale
```

Список литературы

Мастицкий С. Э., Шитиков В. К. (2014) Статистический анализ и визуализация данных с помощью R. - Электронная книга, 400 с

Кабаков Р. К. (2014) R в действии. Анализ и визуализация данных на языке R Издательство: ДМК Пресс, 580 с.

Шитиков В. К., Розенберг Г. С. (2012) Рандомизация, бутстреп и методы Монте-Карло. Примеры статистического анализа данных по биологии и экологии. Тольятти: Ин-т экологии Волжского бассейна.