Липецкий государственный технический университет

Кафедра прикладной математики

Индивидуальное домашнее задание №4
по курсу «Организация обработки данных»
Вариант 2

Студент	________________	Богомолов Е.А.
	подпись, дата	фамилия, инициалы
Группа		
Руководитель		
	________________	Левина Л.В.
ученая степень, ученое звание	подпись, дата	фамилия, инициалы

Липецк 2022 г.

Содержание

Задание кафедры

Дана выборка – 44 значения:

1. Для первых 10 элементов найти:

   - $\hat{M}(x), \hat{D}(x)$ – обычным способом.
   - $\widetilde{M}(x), \widetilde{D}(x)$ – по методу бутстреп.
   - $\hat{D}\widetilde{M}(x), \hat{D}\widetilde{D}(x)$.

2. Для следующих 20 элементов:

   - $\hat{M}(x), \hat{D}(x)$ – обычным способом.
   - $\widetilde{M}(x), \widetilde{D}(x)$ – по методу бутстреп.
   - $\hat{D}\widetilde{M}(x), \hat{D}\widetilde{D}(x)$.

3. Для всей выборки найти $M(x), D(x)$

1. Краткая теоритическая справка

В статистике и анализе данных бутстрапом называют статистическую процедуру, основанную на выборке с замещением для определения точности (смещения) выборочных оценок дисперсии, среднего, стандартного отклонения, доверительных интервалов и других структурных характеристик совокупности.

Метод разработан и впервые опубликован в 1972 году Бредли Эфроном.

В основе идеи бутстрапа лежит оценка структурных характеристик генеральной совокупности на основе перевыборки (resampling) из выборки. Иными словами, перевыборка по отношению к выборке рассматривается как выборка по отношению к генеральной совокупности.

Алгоритм работы метода следующий:

- Из генеральной совокупности формируется случайная выборка из N(t) наблюдений.
- К выборке применяется случайная перевыборка с возвратом (псевдовыборка) того же объема, но в которую некоторые наблюдения могут попасть несколько раз, а другие не попасть совсем.
- Процедура перевыборки повторяется достаточно много раз (несколько десятков, сотен или даже тысяч), и для каждого случая вычисляется среднее.
- Из полученного набора средних значений вычисляется среднее и рассматривается как среднее всей генеральной совокупности.

Важнейшим преимуществом бутстрапа являются: простота реализации; отсутствие необходимости гипотез о параметрах распределения данных; возможность оценивания многих статистических характеристик (среднего, дисперсии, стандартного отклонения, доверительных интервалов, квантилей, коэффициентов корреляции и др.).

К недостатку метода можно отнести использование малореалистичного предположения о независимости перевыборок и значительные вычислительные затраты при их многократном построении.

Метод оказывается особенно полезным, когда теоретическое распределение данных неизвестно или объем выборки мал для прямой статистической оценки.

В анализе данных бутстрап используется для оценки точности аналитических моделей.

2. Ход работы

2.0..1 10 значений

$$\widehat{M}(x)=\frac{1}{10}\sum_{u=1}^{10}x_u\,,\ \widehat{D}(x)=\frac{1}{10-1}\sum_{u=1}^{10}(x_u-M)^2$$

Рисунок 1 – Формула для пункта «а»

$$\mathrm{M}_{\mathrm{i}}^{*}=\frac{1}{10}\cdot\sum_{\mathrm{j}=1}^{10}\mathrm{V}_{\mathrm{ij}}\mathrm{X}_{\mathrm{ij}}\quad M^{*}=\frac{1}{10}\cdot\sum_{i=1}^{10}M_i^{*}\quad \Delta=M^{*}-\widehat{M}\quad \mathrm{D}_{\mathrm{i}}^{*}=\frac{1}{9}\cdot\sum_{\mathrm{j}=1}^{10}\mathrm{V}_{\mathrm{ij}}(\mathrm{X}_{\mathrm{j}}-\mathrm{M}_{\mathrm{i}})^2$$

$$D^{*}=\frac{1}{10}\sum_{i=1}^{10}D_i^{*}\quad \Delta=D^{*}-\widehat{D}$$

Рисунок 2 – Формула для пункта «б»

Шаг 1	1	2	3	4	5	6	7	8	9	10
X1	1	5	0	3	0	0	1	3	2	1
X2	4	0	2	1	1	4	2	0	2	1
X3	3	2	1	0	5	2	0	5	2	3
X4	1	0	1	0	0	1	1	0	1	1
X5	0	1	1	4	1	0	1	0	1	0
X6	0	0	1	0	1	1	1	0	0	0
X7	0	0	1	1	0	1	1	0	0	0
X8	0	0	1	0	1	1	0	0	1	0
X9	0	1	0	1	0	0	1	0	1	3
X10	1	1	2	0	1	0	2	2	0	1
Шаг 2										
M*i	-0.027	-0.542	0.044	0.189	-0.270	0.390	0.172	-0.789	-0.045	-0.007
M*	-0.089	Погрешность	-0.180							

Рисунок 3 – Формирование подвыборок и вычисление математического ожидания для 10 повторений

Шаг 3										
	0.701	0.10364154346956	0.8240813869054	1.10972859199146	0.35334059477424	1.57421277538205	1.07322925041049	0.00558594877107	0.670430653620605	0.73435708386909
	0.84243652232103	2.05396365858626	0.71793489524429	0.49233112219846	1.3471666486594	0.25042349087066	0.51715067526791	2.82361130724311	0.876660266674512	0.80667935540567
	0.27304134497126	5.200405184501E-05	0.35172799446759	0.54569819561844	0.0782331572589	0.88351291929598	0.52019449568242	0.05759206991591	0.254092225664367	0.29400652801416
	0.00026081184259	0.249173213797372	0.00751388865388	0.05397721348284	0.05138473126503	0.18798274264458	0.04616533027927	0.55706428235176	5.3282375936466E-06	0.00128448951447
	0.11903628892311	0.740182436971448	0.07534110445641	0.01659873128275	0.34556531241614	0.00524230103596	0.02140510679825	1.22663027955861	0.132113587092527	0.10583717207793
	0.61605062522997	1.69054896838836	0.51030436041439	0.32342950428315	1.05620963202903	0.13503334006556	0.34360395254539	2.39446528692391	0.64536623050558	0.58552923391995
	1.9381813647153	3.63858685307488	1.74676613220366	1.3829889968278	2.67328221677883	0.95016822041673	1.42438091495984	4.64274451690505	1.98991590289421	1.8837444933974
	0.37580649102018	0.009546927532683	0.4672591933495	0.68759067681587	0.13704722883395	1.0618285954596	0.65892523291789	0.02234620825877	0.353516590452577	0.40033546489092
	0.96722719352044	2.24639931972688	0.83346688246396	0.58874437432807	1.50383319523285	0.3204205236617	0.61585700800368	3.04849558786044	1.00387386674182	0.92888535822898
	1.56341750848822	0.540290785089914	1.74477672799737	2.15076082565751	1.01512732989833	2.78151363162089	2.09982840608954	0.23799751369617	1.51759973977385	1.61304492315985
D*i	0.71706072218271	0.449464918582192	0.89825954666986	0.6510716882733	0.47692021534793	0.56708141118456	0.97317774730131	0.08674591369834	0.565763962715943	0.75822683451983
D*	0.614	Погрешность	-0.192							

Рисунок 4 – Вычисление дисперсии для 10 повторений

M*	-0.125	Погрешность	-0.216

Рисунок 5 – Математическое ожидание для 30 повторений

D*	0.591	Погрешность	-0.215

Рисунок 6 – Дисперсия для 30 повторений

M*	-0.085	Погрешность	-0.177

Рисунок 7 – Математическое ожидание для 50 повторений

D*	0.622	Погрешность	-0.184

Рисунок 8 – Дисперсия для 50 повторений

Оценки	
n=10	
(M*i-M)^2	0.01
DM	0.16
(D*i-D)^2	0.01
DD	0.10
n=30	
(M*i-M)^2	0.56
DM	0.17
(D*i-D)^2	0.30
DD	0.10
n=50	
(M*i-M)^2	0.01
DM	0.14
(D*i-D)^2	0.00

M*	0.09	Погрешность	-0.04

Рисунок 10 – Математическое ожидание для 10 повторений

D*	0.56	Погрешность	-0.27

Рисунок 11 – Дисперсия для 10 повторений

M*	0.10	Погрешность	-0.03

Рисунок 12 – Математическое ожидание для 30 повторений

D*	0.58	Погрешность	-0.26

Рисунок 13 – Дисперсия для30 повторений

M*	0.10	Погрешность	-0.73

Рисунок 14 – Математическое ожидание для 50 повторений

D*	0.58	Погрешность	-0.25

Рисунок 15 – Дисперсия для 50 повторений

Оценки	
n=10	
(M*i-M)^2	0.07
DM	0.03
(D*i-D)^2	0.00
DD	0.13
n=30	
(M*i-M)^2	0.04
DM	0.03
(D*i-D)^2	0.37
DD	0.11
n=50	
(M*i-M)^2	0.03
DM	0.04
(D*i-D)^2	0.09
DD	0.12

Рисунок 16 – Оценки для математического ожидания и дисперсии

M
0.09
0.13
0.16
D
0.81
0.83
0.77

Рисунок 17 – Математическое ожидание и дисперсия для всей выборки

Выводы

Из материала, представленного в отчёте, можно сделать вывод, что метод бутстреп в сравнении с методом джекнайф имеет большую погрешность, но применение данного метода позволит легче интерпретировать полученные оценки.