

Введение в Бутстреп

2021

Оглавление

| | |
|---|-----------|
| 2 Точность выборочного среднего | 2 |
| 3 Случайные выборки и вероятности | 7 |
| 3.1 Введение | 7 |
| 3.2 Случайные выборки | 7 |
| 3.3 Теория вероятностей | 10 |
| 4 Эмпирическая функция распределения и принцип плагина | 16 |
| 4.1 Введение | 16 |
| 4.2 Эмпирическая функция распределения | 16 |
| 4.3 Принцип плагина | 20 |
| 5 Стандартные ошибки и оценки стандартных ошибок | 22 |
| 5.1 Введение | 22 |
| 5.2 Стандартная ошибка среднего | 22 |
| 5.3 Оценка стандартной ошибки среднего | 24 |
| 6 Бутстреп оценка стандартной ошибки | 26 |
| 6.1 Введение | 26 |
| 6.2 Бутстреп оценка стандартной ошибки | 26 |
| 6.3 Пример: коэффициент корреляции | 30 |
| 6.4 Количество бутстреп репликаций B | 31 |
| 6.5 Параметрический бутстреп | 33 |
| 7 Бутстреп и стандартные ошибки: некоторые примеры | 36 |
| 7.1 Введение | 36 |
| 7.2 Пример 1: результаты тестов | 37 |
| 7.3 Пример 2: построение кривой по данным | 45 |
| 7.4 Пример отказа бутстрепа | 54 |
| 8 Более сложные структуры данных | 56 |
| 8.1 Введение | 56 |
| 8.2 Одновыборочные задачи | 56 |
| 8.3 Двухвыборочная задача | 58 |
| 8.4 Более общие структуры данных | 60 |
| 8.5 Пример: лютеинизирующий гормон | 62 |
| 8.6 Бутстреп скользящих окон | 69 |
| 8.7 Библиографические примечания | 71 |

| | |
|--|------------|
| 9 Модели регрессии | 72 |
| 9.1 Введение | 72 |
| 9.2 Линейная регрессионная модель | 72 |
| 9.3 Пример: данные по гормонам | 74 |
| 9.4 Применение бутстрепа | 76 |
| 9.5 Бутстреп-пары против бутстреп-остатков | 78 |
| 9.6 Пример: данные о выживаемости клеток | 81 |
| 9.7 Наименьшая медиана квадратов | 84 |
| 9.8 Библиографические примечания | 85 |
| 10 Оценки смещения | 87 |
| 10.1 Введение | 87 |
| 10.2 Бутстреп оценка смещения | 87 |
| 10.3 Пример: данные об уровне гормона при ношении различных пла- стырей | 88 |
| 10.4 Улучшенная оценка смещения | 92 |
| 10.5 Оценка смещения по методу складного ножа | 95 |
| 10.6 Поправка на смещение | 98 |
| 11 Метод складного ножа | 100 |
| 11.1 Введение | 100 |
| 11.2 Определение складного ножа | 100 |
| 11.3 Пример: данные о тестировании | 102 |
| 11.4 Псевдо-значения | 104 |
| 11.5 Связь метода складного ножа и бутстрепа | 104 |
| 11.6 Отказ складного ножа | 106 |
| 11.7 Метод складного ножа с отбрасыванием d наблюдений | 107 |
| 14 Улучшенные бутстреп-доверительные интервалы | 109 |
| 14.1 Введение | 109 |
| 14.2 Пример: данные о пространственном восприятии | 109 |
| 14.3 Метод BC _a | 114 |
| 14.4 Метод ABC | 117 |
| 14.5 Пример: данные о твердости зубов | 118 |
| 15 Перестановочные тесты | 126 |
| 15.1 Введение | 126 |
| 15.2 Двухвыборочная задача | 126 |
| 15.3 Прочие тестовые статистики | 132 |
| 15.4 Связь проверки гипотез с доверительными интервалами и бут- стрепом | 135 |
| 15.5 Библиографические примечания | 139 |
| 16 Проверка гипотез с помощью бутстрепа | 140 |
| 16.1 Введение | 140 |
| 16.2 Двухвыборочная задача | 140 |
| 16.3 Связь перестановочного теста и бутстрепа | 143 |
| 16.4 Задача с одной выборкой | 144 |
| 16.5 Тестирование мультимодальности генеральной совокупности | 146 |
| 16.6 Обсуждение | 150 |

| | |
|---|------------|
| 17 Кросс-валидация и другие оценки ошибки предсказания | 152 |
| 17.1 Введение | 152 |
| 17.2 Пример: данные по гормонам | 153 |
| 17.3 Кросс-валидация | 154 |
| 17.4 C_p и другие оценки ошибки предсказания | 156 |
| 17.5 Пример: деревья классификации | 157 |
| 17.6 Бутстреп оценки ошибки предсказания | 160 |
| 17.6.1 Обзор | 160 |
| 17.6.2 Некоторые детали | 162 |
| 17.7 Бутстреп оценка 0.632 | 164 |
| 17.8 Обсуждение | 165 |

Глава 2

Точность выборочного среднего

Бутстреп - это компьютерный метод определения точности статистических оценок. Основная идея, лежащая в основе бутстрапа, очень проста и насчитывает как минимум два столетия. После ознакомления с некоторыми справочными материалами в этом отчете описывается метод бутстрапа и его применение для решения некоторых реальных задач анализа данных. В этой главе, помимо предварительного ознакомления с бутстрапом, рассматриваются некоторые фундаментальные идеи статистики. Основное внимание уделяется одному примеру статистики, для оценки точности которой не нужен компьютер: выборочное среднее. Начнем с простого примера, касающегося средних и их расчетной точности.

В таблице 2.1 показаны результаты небольшого эксперимента, в котором 7 из 16 мышей были случайным образом выбраны для получения нового лечения, а остальные 9 были отнесены к группе без лечения (контрольной). Лечение было направлено на продление выживаемости после тестовой операции. В таблице показано время выживания после операции в днях для всех 16 мышей.

Table 2.1. The mouse data. Sixteen mice were randomly assigned to a treatment group or a control group. Shown are their survival times, in days, following a test surgery. Did the treatment prolong survival?

| Group | Data | | | (Sample Size) | Mean | Estimated Standard Error |
|-------------|------|-----|-----|---------------|-------|--------------------------|
| Treatment: | 94 | 197 | 16 | | | |
| | 38 | 99 | 141 | | | |
| | 23 | | | (7) | 86.86 | 25.24 |
| Control: | 52 | 104 | 146 | | | |
| | 10 | 51 | 30 | | | |
| | 40 | 27 | 46 | (9) | 56.22 | 14.14 |
| Difference: | | | | | 30.63 | 28.93 |

Продлило ли лечение выживаемость? Сравнение средних значений для двух

групп дает предварительные основания для положительного ответа. Обозначим через x_1, x_2, \dots, x_7 продолжительность жизни в группе с лечением, соотв. $x_1 = 94, x_2 = 197, \dots, x_7 = 23$, а через y_1, y_2, \dots, y_9 продолжительность жизни в контрольной группе. Групповые выборочные средние равны

$$\bar{x} = \sum_{i=1}^7 x_i / 7 = 86.86 \quad \text{и} \quad \bar{y} = \sum_{i=1}^9 y_i / 9 = 56.22, \quad (2.1)$$

таким образом разность $\bar{x} - \bar{y}$ равна 30.63, что предполагает значительный эффект продления жизни при лечении.

Но насколько точны эти оценки? В конце концов, средние (2.1) основаны на небольших выборках, всего 7 и 9 мышей соответственно. Чтобы ответить на этот вопрос, нам нужна оценка точности выборочных средних \bar{x} и \bar{y}). Для выборочных средних и по существу только для выборочных средних формулу точности получить легко.

Расчетная стандартная ошибка среднего \bar{x} на основе n независимых наблюдений x_1, x_2, \dots, x_n , $\bar{x} = \sum_{i=1}^n x_i / n$, определяется формулой

$$\sqrt{\frac{s^2}{n}}, \quad (2.2)$$

где $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$. (Эта формула и стандартные ошибки в целом обсуждаются более подробно в главе 4.) Стандартная ошибка любой оценки определяется как квадратный корень из ее дисперсии, то есть среднеквадратичная изменчивость оценки вокруг ее математического ожидания. Это наиболее распространенная мера точности оценок. Грубо говоря, оценка отличается от своего истинного значения менее чем на одну стандартную ошибку примерно в 68% случаев и менее чем на две стандартные ошибки примерно в 95% случаев.

Если бы оценочные стандартные ошибки в эксперименте на мышах были очень малы, скажем, менее 1, тогда мы бы знали, что \bar{x} и \bar{y} были близки к их истинным значениям и что наблюдаемая разница в 30,63, вероятно, была хорошей оценкой истинного увеличения выживаемости при лечении. С другой стороны, если формула (2.2) дает большие оценочные стандартные ошибки, скажем 50, тогда оценка разности будет слишком неточной, чтобы на нее можно было полагаться.

Фактическая ситуация показана справа в Таблице 2.1. Расчетные стандартные ошибки, рассчитанные по (2.2), составляют 25,24 для \bar{x} и 14,14 для \bar{y} . Стандартная ошибка для разности $\bar{x} - \bar{y}$ равна $28,93 = \sqrt{25,24^2 + 14,14^2}$ (поскольку дисперсия разности двух независимых величин является суммой их дисперсий). Мы видим, что наблюдаемая разница 30,63 составляет всего $30,63 / 28,93 = 1,05$ стандартной ошибки разности. Читатели, знакомые с теорией проверки гипотез, сочтут это незначимым результатом, который может легко возникнуть случайно, даже если лечение действительно не имело никакого эффекта.

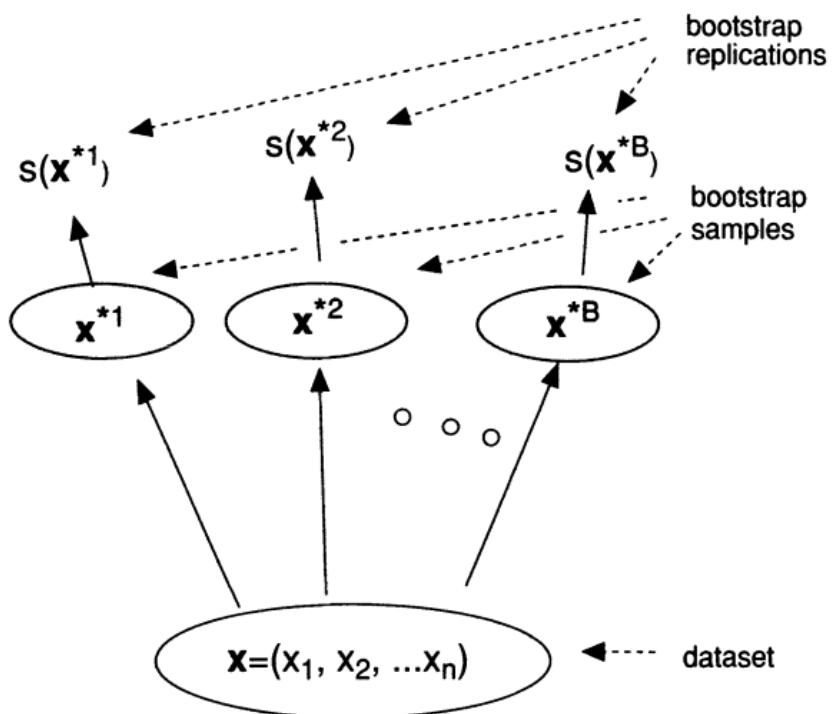
Обычно стандартные ошибки являются отличным первым шагом к критическому осмыслению статистических оценок. К сожалению, стандартные ошибки имеют серьезный недостаток: для большинства статистических оценок, отличных от среднего, не существует формулы, подобной (2.2), для получения стандартных ошибок. Другими словами, трудно оценить точность оценки, отличной от оценки среднего.

Предположим, например, что мы хотим сравнить две группы в таблице 2.1 по их медианам, а не по их средним значениям. Медианы составляют 94 для

лечения и 46 для контроля, что дает разницу в 48, что значительно больше, чем разница средних значений. Но насколько точны эти медианы? Ответы на такие вопросы - вот где вступают в игру бутстреп и другие компьютерные методы. В оставшейся части этой главы дается краткий обзор начальной оценки стандартной ошибки - метода, который будет полностью обсуждаться в следующих главах.

Предположим, что мы наблюдаем независимые данные x_1, x_2, \dots, x_n , для удобства обозначенные вектором $X = (x_1, x_2, \dots, x_n)$, по которым мы вычисляем интересующую статистику $s(X)$. Например, данные могут быть наблюдениями контрольной группы $n = 9$ в таблице 2.1, а $s(X)$ может быть средним по выборке.

Бутстреп оценка стандартной ошибки, изобретенная Эфроном в 1979 году, выглядит совершенно иначе, чем (2.2), но на самом деле они тесно связаны. Бутстреп выборка $X^* = (x_1^*, x_2^*, \dots, x_n^*)$ получается путем случайного выбора с возвращением n точек из исходных данных x_1, x_2, \dots, x_n . Например, при $n = 7$ мы можем получить $X^* = (x_5, x_7, x_5, x_4, x_7, x_3, x_1)$.



*Figure 2.1. Schematic of the bootstrap process for estimating the standard error of a statistic $s(\mathbf{x})$. B bootstrap samples are generated from the original data set. Each bootstrap sample has n elements, generated by sampling with replacement n times from the original data set. Bootstrap replicates $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \dots, s(\mathbf{x}^{*B})$ are obtained by calculating the value of the statistic $s(\mathbf{x})$ on each bootstrap sample. Finally, the standard deviation of the values $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \dots, s(\mathbf{x}^{*B})$ is our estimate of the standard error of $s(\mathbf{x})$.*

Рисунок 2.1 представляет собой схему процесса бутстрепа. Алгоритм бутстрепа начинается с генерации большого количества независимых бутстрепов выборок $X^{*1}, X^{*2}, \dots, X^{*B}$, каждая размером n . Типичные значения для B , коли-

чества бутстреп выборок, находятся в диапазоне от 50 до 200 для оценки стандартной ошибки. Каждой бутстреп выборке соответствует бутстреп репликация $s(X^{*b})$, посчитанная для X^{*b} . Если $s(X)$ - это, например, медиана выборки, то $s(X^*)$ - это медиана бутстреп выборки. Бутстреп оценка стандартной ошибки - это стандартное отклонение бутстреп репликаций

$$\hat{se}_{boot} = \left\{ \sum_{b=1}^B [s(X^{*b}) - s(\cdot)]^2 / (B - 1) \right\}^{\frac{1}{2}}, \quad (2.3)$$

где $s(\cdot) = \sum_{b=1}^B s(X^{*b}) / B$. Предположим, что $s(X) = \bar{X}$. В этом случае стандартная теория вероятностей говорит нам, что, когда B становится очень большим, формула (2.3) приближается к

$$\left\{ \sum_{i=1}^n (x_i - \bar{x})^2 / n^2 \right\}^{\frac{1}{2}}. \quad (2.4)$$

Это почти то же самое, что и формула (2.2). Мы могли бы сделать это точно таким же, умножив определение (2.3) на множитель $[n/(n - 1)]^{\frac{1}{2}}$, но в этом нет практического смысла.

Table 2.2. Bootstrap estimates of standard error for the mean and median; treatment group, mouse data, Table 2.1. The median is less accurate (has larger standard error) than the mean for this data set.

| B: | 50 | 100 | 250 | 500 | 1000 | ∞ |
|---------|-------|-------|-------|-------|-------|----------|
| mean: | 19.72 | 23.63 | 22.32 | 23.79 | 23.02 | 23.36 |
| median: | 32.21 | 36.35 | 34.46 | 36.72 | 36.48 | 37.83 |

В таблице 2.2 показаны бутстррап оценки стандартной ошибки для среднего и медианы для данных экспериментальной группы мышей из таблицы 2.1. Стандартные ошибки уменьшаются до предельных значений по мере увеличения числа бутстрраповых выборок B . Предельное значение 23,36 для среднего получается из (2.4). Формула для предельного значения 37,83 для стандартной ошибки медианы довольно сложна.

Теперь мы можем оценить точность разницы медиан между двумя группами. Описанная выше бутстреп процедура, примененная к контрольной группе, дала оценку стандартной ошибки 11,54 на основе $B = 100$ повторений ($B = \infty$ дало 9,73). Следовательно, используя $B = 100$, наблюдаемая разница в 48 имеет расчетную стандартную ошибку $\sqrt{36,35^2 + 11,54^2} = 38,14$, и, следовательно, $48/38,14 = 1,26$ стандартной ошибки. Это больше, чем наблюдаемая разница в средних, но все же незначимо.

Для большинства статистических данных у нас нет формулы для предельного значения стандартной ошибки, но на самом деле формула не нужна. Вместо этого мы используем числовой вывод бутстреп программы для некоторого удобного значения B . Легко написать бутстреп программу, которая работает для любой вычислимой статистики $s(X)$. Имея эти программы, аналитик данных может свободно использовать любую статистику, независимо от ее сложности, с уверенностью, что он или она также будет иметь разумное представление о

точности оценки. Применение бутстрепа стало доступным, поскольку компьютеры стали мощнее и дешевле. Стандартные ошибки - это простейшие меры статистической точности. В последующих главах показано, как бутстреп методы могут оценивать более сложные меры точности, такие как смещения, ошибки прогнозирования и доверительные интервалы. Бутстрепированные доверительные интервалы увеличивают вычислительную нагрузку еще в 10 раз. Результатом всех этих вычислений является увеличение количества статистических проблем, которые могут быть проанализированы, сокращение допущений анализа и устранение рутинных, но утомительных теоретических расчетов, обычно связанных с оценкой точности.

Глава 3

Случайные выборки и вероятности

3.1 Введение

Статистика - это теория накопления информации, особенно информации, поступающей постепенно. Типичная статистическая ситуация была проиллюстрирована данными по мышам в Таблице 2.1. Ни одна мышь не предоставляет много информации, поскольку индивидуальные результаты очень различаются, но семь или девять мышей, взятых вместе, начинают быть весьма информативными. Статистическая теория касается лучших способов извлечения этой информации. Теория вероятностей обеспечивает математическую основу для статистических выводов. В этой главе рассматривается простейшая вероятностная модель, используемая для моделирования случайных данных: случай, когда наблюдения представляют собой случайную выборку из одной неизвестной совокупности, свойства которой мы пытаемся узнать из наблюдаемых данных.

3.2 Случайные выборки

Проще всего визуализировать случайные выборки в терминах конечной совокупности или «вселенной» \mathbf{U} отдельных единиц U_1, U_2, \dots, U_n , любая из которых с равной вероятностью будет выбрана в одном случайному розыгрыше. В состав единиц могут входить все зарегистрированные избиратели в районе, подвергающиеся политическому обследованию, все мужчины, которые предположительно могут быть выбраны для медицинского эксперимента, все средние школы в Соединенных Штатах и т.д. У отдельных единиц есть свойства, которые нам нужны, чтобы узнать, например, о политических взглядах, времени выживания в медицине или количестве выпускников. Слишком сложно и дорого исследовать каждую единицу в \mathbf{U} , поэтому мы выбираем для наблюдения случайную выборку управляемого размера.

Случайная выборка размера n определяется как набор из n единиц u_1, u_2, \dots, u_n , выбранных случайнным образом из \mathbf{U} . В принципе, процесс выборки происходит следующим образом: устройство случайных чисел независимо выбирает целые числа j_1, j_2, \dots, j_n , каждое из которых равно любому значению от 1 до N с вероятностью $1/N$. Эти целые числа определяют, какие члены \mathbf{U} выбраны для случайной выборки, $u_1 = U_{j_1}, u_2 = U_{j_2}, \dots, u_n = U_{j_n}$. На практике процесс отбора редко бывает таким аккуратным, и совокупность \mathbf{U} может быть плохо определена, но концептуальная структура случайной выборки по-прежнему полезна для понимания статистических выводов. (Методология хорошего экспериментального дизайна, например, случайное распределение выбранных единиц в экспе-

риментальную или контрольную группы, как это было сделано в эксперименте на мышах, помогает сделать теорию случайной выборки более применимой к реальным ситуациям, подобным той, что представлена в таблице 2.1.)

Наше определение случайной выборки позволяет одной единице U_i появляться в выборке более одного раза. Мы могли бы избежать этого, настаивая на том, чтобы целые числа j_1, j_2, \dots, j_n были различными, что называется «выборкой без замены». Чуть проще разрешить повторы, то есть «выборку с заменой», как в предыдущем абзаце. Если размер случайной выборки n намного меньше, чем размер генеральной совокупности N , как это обычно бывает, вероятность повторения выборки в любом случае будет мала. См. Проблему 3.1. Случайная выборка всегда означает выборку с заменой в дальнейшем, если не указано иное.

Выбрав случайную выборку u_1, u_2, \dots, u_n , мы получаем одно или несколько представляющих интерес измерений для каждой единицы. Пусть x_i обозначает измерения для единицы u_i . Наблюдаемые данные представляют собой набор измерений x_1, x_2, \dots, x_n . Иногда мы будем обозначать наблюдаемые данные (x_1, x_2, \dots, x_n) одним символом X .

Мы можем представить себе, как проводить измерения для каждого члена U_1, U_2, \dots, U_N из \mathbf{U} , получая значения X_1, X_2, \dots, X_N . Это можно было бы назвать переписью U .

Символ \mathbf{X} будет обозначать перепись измерений (X_1, X_2, \dots, X_N) . Мы также будем называть \mathbf{X} совокупностью измерений или просто совокупностью и называть X случайной выборкой размера n из \mathbf{X} . На самом деле мы обычно не можем позволить себе провести перепись, поэтому мы взяли случайную выборку. Цель статистического вывода – сказать, что мы узнали о популяции \mathbf{X} из наблюдаемых данных X . В частности, мы будем использовать бутстреп, чтобы сказать, насколько точно статистика, вычисленная из x_1, x_2, \dots, x_n (например, медиана выборки), оценивает соответствующее количество для всей генеральной совокупности.

Table 3.1. The law school data. A random sample of size $n = 15$ was taken from the collection of $N = 82$ American law schools participating in a large study of admission practices. Two measurements were made on the entering classes of each school in 1973: LSAT, the average score for the class on a national law test, and GPA, the average undergraduate grade-point average for the class.

| School | LSAT | GPA | School | LSAT | GPA |
|--------|------|------|--------|------|------|
| 1 | 576 | 3.39 | 9 | 651 | 3.36 |
| 2 | 635 | 3.30 | 10 | 605 | 3.13 |
| 3 | 558 | 2.81 | 11 | 653 | 3.12 |
| 4 | 578 | 3.03 | 12 | 575 | 2.74 |
| 5 | 666 | 3.44 | 13 | 545 | 2.76 |
| 6 | 580 | 3.07 | 14 | 572 | 2.88 |
| 7 | 555 | 3.00 | 15 | 594 | 2.96 |
| 8 | 661 | 3.43 | | | |

В таблице 3.1 показана случайная выборка размером $n = 15$, составленная из 82 американских юридических школ. Фактически показаны два измерения, проведенные для поступающих в 1973г. для каждого учебного заведения в выборке: LSAT, средний балл класса на экзамене по национальному праву, и GPA, средний балл бакалавриата, полученный студентами. В этом случае измерение x_i на u_i , i -м члене выборки, представляет собой пару

$$x_i = (LSAT_i, GPA_i) \quad i = 1, 2, \dots, 15$$

Наблюдаемые данные x_1, x_2, \dots, x_n представляют собой набор из 15 пар чисел, показанных в таблице 3.1.

Этот пример является искусственным, потому что перепись данных X_1, X_2, \dots, X_{82} действительно была проведена. Другими словами, LSAT и GPA доступны для всей совокупности $N = 82$ школ. На Рисунке 3.1 показаны данные переписи и выборочные данные. В таблице 3.2 приведены все измерения N .

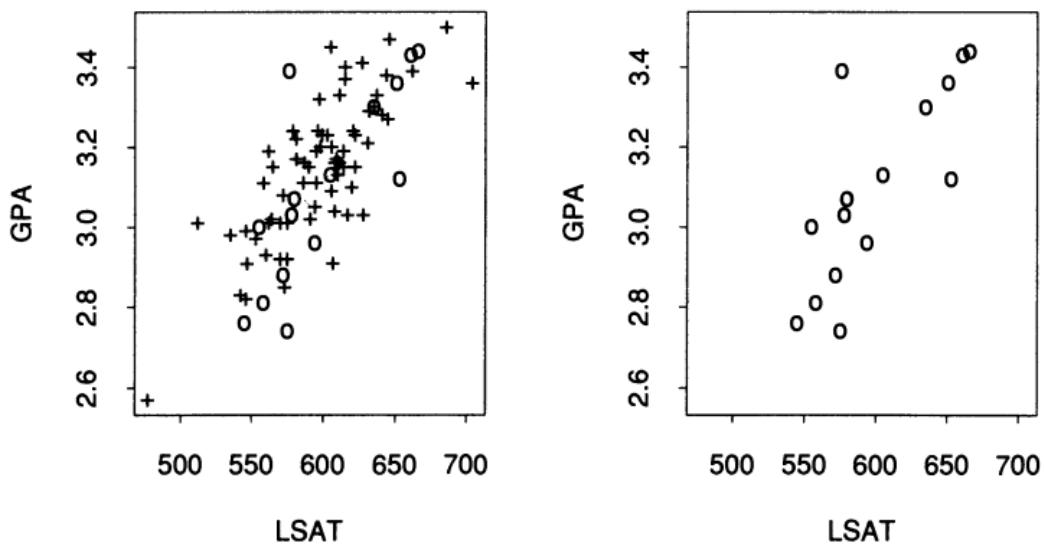


Figure 3.1. The left panel is a scatterplot of the (LSAT, GPA) data for all $N = 82$ law schools; circles indicate the $n = 15$ data points comprising the “observed sample” of Table 3.1. The right panel shows only the observed sample. In problems of statistical inference, we are trying to infer the situation on the left from the picture on the right.

В реальной статистической задаче, такой как в таблице 3.1, мы увидим только выборочные данные, из которых мы попытаемся сделать вывод о свойствах совокупности. Например, рассмотрим 15 баллов LSAT в наблюдаемой выборке. Они имеют среднее значение 600.27 с расчетной стандартной ошибкой 10.79, основанной на данных в таблице 3.1 и формуле (2.2). Вероятность того, что истинное среднее значение LSAT, среднее для всей генеральной совокупности, из которой были взяты наблюдаемые данные, составляет около 68%, находится в интервале 600.27 ± 10.79 .

Мы можем проверить этот результат, поскольку имеем дело с искусственным примером, для которого известны полные данные о населении. Среднее значение всех 82 значений LSAT составляет 597.55, оно лежит в пределах про-

гнозируемого доверительного интервала 600.27 ± 10.79 .

Table 3.2. The population of measurements (LSAT,GPA), for the universe of 82 law schools. The data in Table 3.1 was sampled from this population. The +'s indicate the sampled schools.

| school | LSAT | GPA | school | LSAT | GPA | school | LSAT | GPA |
|--------|------|------|--------|------|------|--------|------|------|
| 1 | 622 | 3.23 | 28 | 632 | 3.29 | 56 | 641 | 3.28 |
| 2 | 542 | 2.83 | 29 | 587 | 3.16 | 57 | 512 | 3.01 |
| 3 | 579 | 3.24 | 30 | 581 | 3.17 | 58 | 631 | 3.21 |
| 4+ | 653 | 3.12 | 31+ | 605 | 3.13 | 59 | 597 | 3.32 |
| 5 | 606 | 3.09 | 32 | 704 | 3.36 | 60 | 621 | 3.24 |
| 6+ | 576 | 3.39 | 33 | 477 | 2.57 | 61 | 617 | 3.03 |
| 7 | 620 | 3.10 | 34 | 591 | 3.02 | 62 | 637 | 3.33 |
| 8 | 615 | 3.40 | 35+ | 578 | 3.03 | 62 | 572 | 3.08 |
| 9 | 553 | 2.97 | 36+ | 572 | 2.88 | 64 | 610 | 3.13 |
| 10 | 607 | 2.91 | 37 | 615 | 3.37 | 65 | 562 | 3.01 |
| 11 | 558 | 3.11 | 38 | 606 | 3.20 | 66 | 635 | 3.30 |
| 12 | 596 | 3.24 | 39 | 603 | 3.23 | 67 | 614 | 3.15 |
| 13+ | 635 | 3.30 | 40 | 535 | 2.98 | 68 | 546 | 2.82 |
| 14 | 581 | 3.22 | 41 | 595 | 3.11 | 69 | 598 | 3.20 |
| 15+ | 661 | 3.43 | 42 | 575 | 2.92 | 70+ | 666 | 3.44 |
| 16 | 547 | 2.91 | 43 | 573 | 2.85 | 71 | 570 | 3.01 |
| 17 | 599 | 3.23 | 44 | 644 | 3.38 | 72 | 570 | 2.92 |
| 18 | 646 | 3.47 | 45+ | 545 | 2.76 | 73 | 605 | 3.45 |
| 19 | 622 | 3.15 | 46 | 645 | 3.27 | 74 | 565 | 3.15 |
| 20 | 611 | 3.33 | 47+ | 651 | 3.36 | 75 | 686 | 3.50 |
| 21 | 546 | 2.99 | 48 | 562 | 3.19 | 76 | 608 | 3.16 |
| 22 | 614 | 3.19 | 49 | 609 | 3.17 | 77 | 595 | 3.19 |
| 23 | 628 | 3.03 | 50+ | 555 | 3.00 | 78 | 590 | 3.15 |
| 24 | 575 | 3.01 | 51 | 586 | 3.11 | 79+ | 558 | 2.81 |
| 25 | 662 | 3.39 | 52+ | 580 | 3.07 | 80 | 611 | 3.16 |
| 26 | 627 | 3.41 | 53+ | 594 | 2.96 | 81 | 564 | 3.02 |
| 27 | 608 | 3.04 | 54 | 594 | 3.05 | 82+ | 575 | 2.74 |
| | | | 55 | 560 | 2.93 | | | |

3.3 Теория вероятностей

Статистический вывод касается обучения на опыте: мы наблюдаем случайную выборку $\mathbf{x} = (x_1, x_2, \dots, x_n)$ и хотим вывести свойства полной совокупности $\mathbf{X} = (X_1, X_2, \dots, X_N)$, которая дала образец. Теория вероятностей идет в противоположном направлении: из состава популяции \mathbf{X} мы выводим свойства случайной выборки \mathbf{x} и статистики, вычисляемой по \mathbf{x} . Статистический вывод как математическая наука был разработан почти исключительно в терминах теории вероятностей. Здесь мы кратко рассмотрим некоторые фундаментальные концепции вероятности, включая распределения вероятностей, ожидания и независимость.

В качестве первого примера пусть x представляет результат броска пра-

вильной кости, поэтому x с равной вероятностью будет 1, 2, 3, 4, 5 или 6. Мы запишем это в вероятностной нотации как

$$Prob\{x = k\} = 1/6 \quad \text{for } k = 1, 2, 3, 4, 5, 6. \quad (3.1)$$

Случайное число, такое как x , часто называется случайной величиной.

Вероятности - это идеализированные или теоретические пропорции. Мы можем представить себе пространство $(U) = \{U_1, U_2, \dots, U_N\}$ возможных бросков кубика, где U_j полностью описывает физический акт j -го броска с соответствующими результатами $\mathbf{X} = (X_1, X_2, \dots, X_N)$. Здесь N может быть очень большим или даже бесконечным. Выражение $Prob\{x = 5\} = 1/6$ означает, что случайным образом выбранный член \mathbf{X} имеет $1/6$ шанс быть равным 5, или, проще говоря, $1/6$ членов \mathbf{X} равняется 5. Обратите внимание что такие вероятности, как пропорции, никогда не могут быть меньше 0 или больше 1.

Для удобства обозначений определим частоты f_k ,

$$f_k = Prob\{x = k\}, \quad (3.2)$$

так что у справедливой кости $f_k = 1/6$ для $k = 1, 2, \dots, 6$. Распределение вероятностей случайной величины x , которую мы обозначим F , является любым полным описанием вероятностного поведения x . F также называется распределением вероятностей популяции \mathbf{X} . Здесь мы можем взять F как вектор частот

$$F = (f_1, f_2, \dots, f_6) = (1/6, 1/6, \dots, 1/6). \quad (3.3)$$

Несправедливым будет кубик, для которого F не равно $(1/6, 1/6, \dots, 1/6)$.

Некоторые распределения вероятностей возникают настолько часто, что получили специальные названия. Говорят, что случайная величина x имеет биномиальное распределение с размером n и вероятностью успеха p , что обозначается

$$x \sim Bi(n, p), \quad (3.4)$$

если его частоты

$$f_k = C_n^k p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \dots, n. \quad (3.5)$$

Здесь n – положительное целое число, p – число от 0 до 1, а C_n^k – биномиальный коэффициент $n!/[K!(n-k)!]$. На рисунке 3.2 показано распределение $F = (f_0, f_1, \dots, f_n)$ для $x \sim Bi(n, p)$, при $n = 25$ и $p = 0.25, 0.50$ и 0.90 . Мы также пишем $F = Bi(n, p)$ для обозначения ситуации (3.4).

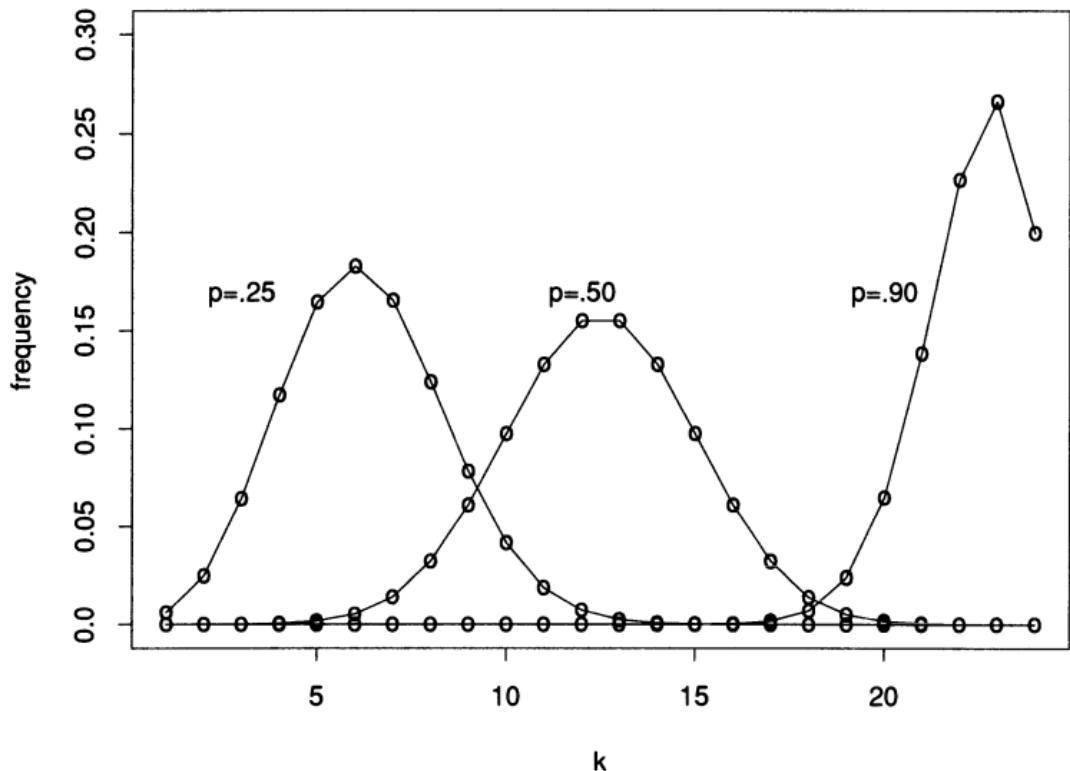


Figure 3.2. The frequencies f_0, f_1, \dots, f_n for the binomial distributions $Bi(n, p)$, $n = 25$ and $p = .25, .50$, and $.90$. The points have been connected by lines to enhance visibility.

Пусть A – набор целых чисел. Тогда вероятность того, что x принимает значение в A , или, проще говоря, вероятность A , равна

$$Prob\{x \in A\} = Prob\{A\} = \sum_{k \in A} f_k. \quad (3.6)$$

Например, если $A = \{1, 3, 5, \dots, 25\}$ и $x \sim Bi(25, p)$, то $Prob\{A\}$ – это вероятность того, что биномиальная случайная величина размера 25 и вероятность успеха p равно нечетному целому числу. Заметьте, что, поскольку f_k – это теоретическая доля раз, когда x равно k , сумма $\sum_{k \in A} f_k = Prob\{A\}$ – это теоретическая доля раз, когда x принимает свое значение в A .

Выборочное пространство x , обозначенное S_x , представляет собой набор возможных значений x . Для правильного кубика $S_x = \{1, 2, \dots, 6\}$, а $S_x = \{0, 1, 2, \dots, n\}$ для распределения $Bi(n, p)$. По определению x встречается в S_x каждый раз, то есть с теоретической пропорцией 1, поэтому

$$Prob\{S_x\} = \sum_{k \in S_x} f_k = 1. \quad (3.7)$$

Для любого распределения вероятностей целых чисел частоты f_j являются неотрицательными числами, сумма которых равна 1.

В наших примерах до сих пор пространство выборки S_x было подмножеством целых чисел. Одна из удобных особенностей вероятностных распределений заключается в том, что их можно определять в довольно общих пространствах. Рассмотрим данные юридического факультета на Рисунке 3.1. Мы могли

бы принять S_x за положительный квадрант плоскости

$$S_x = \mathbf{R}^{2+} = \{(y, z) : y > 0, z > 0\}. \quad (3.8)$$

(Сюда входят такие значения, как $x = (10^6, 10^9)$, но не повредит, если S_x будет слишком большим.) Для подмножества A из S_x мы все равно будем писать $\text{Prob}\{A\}$, чтобы указать вероятность того, что x встречается в A .

Например, мы могли бы взять

$$A = \{(y, z) : 0 < y < 600, 0 < z < 3.0\}. \quad (3.9)$$

Юридическая школа $x \in A$, если ее входной класс 1973 года имел LSAT менее 600 и средний балл менее 3,0. В этом случае мы знаем полную популяцию \mathbf{X} ; это 82 точки, указанные на левой панели рисунка 3.1 и в таблице 3.2. Из них 16 находятся в A , поэтому

$$\text{Prob}\{A\} = 16/82 = 0.195. \quad (3.10)$$

Здесь идеализированная пропорция $\text{Prob}\{A\}$ – это действительная пропорция. Только в тех случаях, когда у нас есть полная генеральная совокупность, можно напрямую оценить вероятности как пропорции.

Распределение вероятностей F по-прежнему определяется как полное описание вероятностей x . В примере с юридической школой F можно описать следующим образом: для любого подмножества A из $S_x = \mathbf{R}^{2+}$,

$$\text{Prob}\{x \in A\} = \#\{X_j \in A\}/82, \quad (3.11)$$

где $\#\{X_j \in A\}$ – это 82 точки на левой панели рисунка 3.1, которые лежат в A . Другим способом сказать, что F – это дискретное распределение, полагая вероятность (или частоту) 1/82 на каждую из указанных 82 точек.

Вероятности можно определять непрерывно, а не дискретно, как в (3.6) или (3.11). Самый известный пример – нормальное (или гауссово, или колоколообразное) распределение. Определено, что случайная величина x с действительными значениями имеет нормальное распределение со средним μ и дисперсией σ^2 , записанное

$$x \sim N(\mu, \sigma^2) \quad \text{or} \quad F = N(\mu, \sigma^2), \quad (3.12)$$

если

$$\text{Prob}\{x \in A\} = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \quad (3.13)$$

для любого подмножества A действительной прямой \mathbf{R}^1 . Интеграл в (3.13) берется по значениям $x \in A$.

Существуют версии нормального распределения с более высокой размерностью, которые включают взятие интегралов, подобных (3.13), по многомерным множествам A . Нам не понадобятся непрерывные распределения для разработки бутстрепа. Как мы увидим, одним из основных стимулов для развития бутстрепа является желание заменить теоретические вычисления компьютерными с использованием специальных распределений.

Математическое ожидание вещественной случайной величины x , обозначаемой $E(x)$, является ее средним значением, где среднее значение берется по возможным результатам x , взвешенным в соответствии с его распределением вероятностей F . Таким образом,

$$E(x) = \sum_{x=0}^n x C_n^x p^x (1-p)^{n-x} \quad \text{for } x \sim Bi(n, p), \quad (3.14)$$

и

$$E(x) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \quad for \quad x \sim N(\mu, \sigma^2). \quad (3.15)$$

Нетрудно показать, что $E(x) = np$ для $x \sim Bi(n, p)$ и $E(x) = \mu$ для $x \sim N(\mu, \sigma^2)$.

Иногда мы пишем математическое ожидание как $E_F(x)$, чтобы указать, что среднее значение берется по отношению к распределению F .

Предположим, что $r = g(x)$ – некоторая функция случайной величины x . Тогда $E(r)$, математическое ожидание r , представляет собой теоретическое среднее значение $g(x)$, взвешенное в соответствии с распределением вероятности x . Например, если $x \sim N(\mu, \sigma^2)$ и $r = x^3$, то

$$E(r) = \int_{-\infty}^{\infty} x^3 \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx. \quad (3.16)$$

Вероятности – это частный случай ожиданий. Пусть A – подмножество S_x , и возьмем $r = I_{\{x \in A\}}$, где $I_{\{x \in A\}}$ – индикаторная функция

$$I_{\{x \in A\}} = \begin{cases} 1 & if \quad x \in A \\ 0 & if \quad x \notin A \end{cases}. \quad (3.17)$$

Тогда $E(r)$ равна $Prob\{x \in A\}$ или

$$E(I_{\{x \in A\}}) = Prob\{x \in A\}. \quad (3.18)$$

Например, если $x \sim N(\mu, \sigma^2)$, тогда

$$E(r) = \int_{-\infty}^{\infty} I_{\{x \in A\}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx, \quad (3.19)$$

является $Prob\{x \in A\}$ в соответствии с (3.13).

Понятие математического ожидания как теоретического среднего является очень общим и включает случаи, когда случайная величина x не является действительной. В ситуации с юридической школой, например, нас может заинтересовать математическое ожидание соотношения LSAT и GPA. Пусть $x = (y, z)$, как в (3.8), тогда $r = y/z$, и математическое ожидание r равно

$$E(LSAT/GPA) = \frac{1}{82} \sum_{j=1}^8 2(y_j/z_j) \quad (3.20)$$

где $x_j = (y_j, z_j)$ – j -я точка в таблице 3.2. Численная оценка (3.20) дает $E(LSAT/GPA) = 190.8$.

Пусть $\mu_x = E_F(x)$ для x вещественной случайной величины с распределением F . Дисперсия x , обозначаемая σ_x^2 или просто σ^2 , определяется как ожидаемое значение $y = (x - \mu)^2$. Другими словами, σ^2 – это теоретический средний квадрат расстояния случайной величины x от ее математического ожидания μ ,

$$\sigma_x^2 = E_F(x - \mu_x)^2. \quad (3.21)$$

Дисперсия $x \sim N(\mu, \sigma^2)$ равна σ^2 ; дисперсия $x \sim Bi(n, p)$ равна $np(1 - p)$. Стандартное отклонение случайной величины определяется как квадратный корень из ее дисперсии.

Две случайные величины y и z называются независимыми, если

$$E[g(y)h(z)] = E[g(y)]E[h(z)] \quad (3.22)$$

для всех функций $g(y)$ и $h(z)$. Независимость (3.22) подразумевает, что случайный результат y не влияет на случайный результат z , и наоборот.

Чтобы убедиться в этом, пусть B и C - подмножества S_y и S_z соответственно, выборочные пространства y и z , а g и h - индикаторные функции $g(y) = I_{\{y \in B\}}$ и $h(z) = I_{\{z \in C\}}$. Обратите внимание, что

$$I_{\{y \in B\}}I_{\{z \in C\}} = \begin{cases} 1 & \text{if } y \in B \text{ and } z \in C \\ 0 & \text{otherwise.} \end{cases} \quad (3.23)$$

Итак, $I_{\{y \in B\}}I_{\{z \in C\}}$ – индикаторная функция пересечения $\{y \in B\} \cap \{z \in C\}$. Тогда в силу (3.18) и определения независимости (3.22)

$$\begin{aligned} Prob\{(y, z) \in B \cap C\} &= E(I_{\{y \in B\}}I_{\{z \in C\}}) = \\ &= E(I_{\{y \in B\}})E(I_{\{z \in C\}}) = Prob\{y \in B\}Prob\{z \in C\}. \end{aligned} \quad (3.24)$$

Глядя на рисунок 3.1, мы видим, что (3.24) не выполняется для примера юридической школы, поэтому LSAT и GPA не являются независимыми.

Независимо от того, независимы ли y и z , ожидания подчиняются простому правилу сложения

$$E[g(y) + h(z)] = E[g(y)] + E[h(z)]. \quad (3.25)$$

В общем виде

$$E\left[\sum_{i=1}^n g_i(x_i)\right] = \sum_{i=1}^n E[g_i(x_i)] \quad (3.26)$$

для любых функций g_i и любых n случайных величин x_1, x_2, \dots, x_n .

Случайная выборка с заменой гарантирует независимость: если $x = (x_1, x_2, \dots, x_n)$ – случайная выборка размера n из совокупности \mathbf{X} , то все n наблюдений x_i одинаково распределены и взаимно независимы друг от друга. Другими словами, все x_i имеют одинаковое распределение вероятностей F , и

$$E_F[g_1(x_1)g_2(x_2) \cdots g_n(x_n)] = E_F[g_1(x_1)]E_F[g_2(x_2)] \cdots E_F[g_n(x_n)] \quad (3.27)$$

для любых функций g_1, g_2, \dots, g_n . (Это почти определение того, что означает случайная выборка.) Будем писать

$$F \rightarrow (x_1, x_2, \dots, x_n) \quad (3.28)$$

чтобы указать, что $x = (x_1, x_2, \dots, x_n)$ является случайной выборкой размера n из совокупности с распределением вероятностей F . Иногда это записывается как

$$x \stackrel{\text{iid}}{\sim} F \quad i = 1, 2, \dots, n, \quad (3.29)$$

где i.i.d. означает независимый и одинаково распределенный.

Глава 4

Эмпирическая функция распределения и принцип плагина

4.1 Введение

Проблемы статистического вывода часто включают оценку некоторого свойства распределения вероятностей F на основе случайной выборки, взятой из F . Эмпирическая функция распределения, которую мы будем называть \hat{F} , представляет собой простую оценку всего распределения F . Оценка какого-то интересующего свойства F , например его среднего значения, медианы или корреляции, заключается в использовании соответствующего свойства \hat{F} . Это «принцип плагина». Как мы увидим в главе 6, метод бутстрепа является прямым применением принципа плагина.

4.2 Эмпирическая функция распределения

Пусть дана случайная выборка размера n из распределения вероятностей F

$$F \rightarrow (x_1, x_2, \dots, x_n), \quad (4.1)$$

тогда эмпирическая функция распределения \hat{F} определяется как дискретное распределение, которое ставит вероятность $1/n$ на каждое значение x_i , $i = 1, 2, \dots, n$. Другими словами, \hat{F} присваивает множеству A в пространстве выборок x его эмпирическую вероятность

$$\widehat{\text{Prob}}\{A\} = \#\{x_i \in A\}/n, \quad (4.2)$$

это доля наблюдаемой выборки $x = (x_1, x_2, \dots, x_n)$, встречающейся в A . Мы также будем писать $\text{Prob}_{\hat{F}}\{A\}$ для обозначения (4.2). Символ в шляпе « \wedge » всегда указывает на величины, рассчитанные на основе наблюдаемых данных.

Table 4.1. A random sample of 100 rolls of the die. The outcomes 1, 2, 3, 4, 5, 6 occurred 13, 19, 10, 17, 14, 27 times, respectively, so the empirical distribution is (.13, .19, .10, .17, .14, .27).

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 3 | 2 | 4 | 6 | 6 | 6 | 5 | 3 | 6 | 2 | 2 | 6 | 2 | 3 | 1 | 5 | 1 |
| 6 | 6 | 4 | 1 | 5 | 3 | 6 | 6 | 4 | 1 | 4 | 2 | 5 | 6 | 6 | 5 | 5 | 3 |
| 6 | 2 | 6 | 6 | 1 | 4 | 1 | 5 | 6 | 1 | 6 | 3 | 3 | 2 | 2 | 2 | 5 | 2 |
| 2 | 4 | 1 | 4 | 5 | 6 | 6 | 6 | 2 | 2 | 4 | 6 | 1 | 2 | 2 | 2 | 5 | 1 |
| 5 | 3 | 5 | 4 | 2 | 1 | 4 | 6 | 6 | 5 | 6 | 4 | 6 | 4 | 3 | 6 | 4 | 1 |
| 4 | 5 | 4 | 4 | 2 | 3 | 2 | 1 | 4 | 6 | | | | | | | | |

Рассмотрим выборку юридических вузов размером $n = 15$, показанную в Таблице 3.1 и на правой панели Рисунка 3.1. Эмпирическое распределение F ставит вероятность $1/15$ для каждой из 15 точек данных. Пять из 15 точек лежат в наборе $A = \{(y, z) : 0 < y < 600, 0 < z < 3.00\}$, поэтому $\widehat{Prob}\{A\} = 5/15 = 0.333$. Обратите внимание, что мы получаем другую эмпирическую вероятность для набора $\{0 < y < 600, 0 < z \leq 3.00\}$, поскольку одна из 15 точек данных имеет $GPA = 3.00$, $LSAT < 600$.

Таблица 4.1 показывает случайную выборку из $n = 100$ бросков кубика: $x_1 = 6, x_2 = 3, x_3 = 2, \dots, x_{100} = 6$. Эмпирическое распределение F ставит вероятность $1/100$ для каждого из 100 исходов. В подобных случаях, когда есть повторяющиеся значения, мы можем более экономично выразить F как вектор наблюдаемых частот \hat{f}_k , $k = 1, 2, \dots, 6$

$$\hat{f}_k = \#\{x_i = k\}/n. \quad (4.3)$$

Для данных в таблице 4.1 $\hat{F} = (0.13, 0.19, 0.10, 0.17, 0.14, 0.27)$.

Эмпирическое распределение – это список значений, принимаемых выборкой $x = (x_1, x_2, \dots, x_n)$, вместе с долей случаев, когда каждое значение встречается. Часто каждое значение, встречающееся в выборке, появляется только один раз, как в случае с данными юридических школ. Повторения, как и в случае с кубиком таблицы 4.1, позволяют сократить список. В любом случае каждой из n точек данных x_i приписывается вероятность $1/n$ эмпирическим распределением.

Очевидно ли, что мы не потеряли информацию при переходе от полного набора данных $(x_1, x_2, \dots, x_{100})$ в таблице 4.1 к сокращенному представлению в терминах частот? Нет, но это правда. Можно доказать, что вектор наблюдаемых частот $\hat{F} = (\hat{f}_1, \hat{f}_2, \dots)$ является достаточной статистикой для истинного распределения $F = (f_1, f_2, \dots)$. Это означает, что вся информация о F , содержащаяся в \mathbf{x} , также содержится в \hat{F} .

Table 4.2. Rainfall data. The yearly rainfall, in inches, in Nevada City, California, 1873 through 1978. An example of time series data.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-------|----|-----|----|----|----|----|----|----|----|----|
| 1870: | | | | 80 | 40 | 65 | 46 | 68 | 32 | 58 |
| 1880: | 60 | 61 | 60 | 45 | 48 | 63 | 44 | 66 | 39 | 35 |
| 1890: | 44 | 104 | 36 | 45 | 69 | 50 | 72 | 57 | 53 | 30 |
| 1900: | 40 | 56 | 55 | 46 | 46 | 72 | 50 | 68 | 71 | 37 |
| 1910: | 64 | 46 | 69 | 31 | 33 | 61 | 56 | 55 | 40 | 37 |
| 1920: | 40 | 34 | 60 | 54 | 52 | 20 | 49 | 43 | 62 | 44 |
| 1930: | 33 | 45 | 30 | 53 | 32 | 38 | 56 | 63 | 52 | 79 |
| 1940: | 30 | 62 | 75 | 70 | 60 | 34 | 54 | 51 | 35 | 53 |
| 1950: | 44 | 53 | 73 | 80 | 54 | 52 | 40 | 77 | 52 | 75 |
| 1960: | 42 | 43 | 39 | 54 | 70 | 40 | 73 | 41 | 75 | 43 |
| 1970: | 80 | 60 | 59 | 41 | 67 | 83 | 56 | 29 | 21 | |

Теорема достаточности предполагает, что данные были сгенерированы случайной выборкой из некоторого распределения F . Это, конечно, не всегда верно. Например, данные о мышах в Таблице 2.1 включают два распределения вероятностей, одно для лечения и одно для контроля. В таблице 4.2 показан временной ряд из 106 чисел: годовое количество осадков в Невада-Сити, Калифорния, с 1873 по 1978 год. Мы могли бы вычислить эмпирическое распределение F для этого набора данных, но оно не будет включать никакой информации временного ряда, например, если большие числа следуют за большими числами. На данный момент мы ограничиваем внимание данными, полученными путем случайной выборки из одного распределения, так называемой ситуации с одной выборкой. Это не так строго, как кажется. Например, в примере с данными о мышах мы можем применить результаты по одной выборке отдельно к экспериментальной и контрольной популяциям.

При применении статистической теории к реальным задачам ответы на интересующие вопросы обычно формулируются в терминах вероятностных распределений. Мы можем спросить, справедлива ли матрица, дающая данные в Таблице 4.1. Это эквивалентно вопросу, равно ли распределение вероятностей F кубика $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. В примере с юридической школой вопрос может заключаться в том, насколько коррелируют LSAT и GPA. В терминах F , распределение $x = (y, z) = (LSAT, GPA)$, это вопрос о значении коэффициента корреляции совокупности

$$\text{corr}(y, z) = \frac{\sum_{j=1}^{82} (Y_j - \mu_y)(Z_j - \mu_z)}{[\sum_{j=1}^{82} (Y_j - \mu_y)^2 \sum_{j=1}^{82} (Z_j - \mu_z)^2]^{\frac{1}{2}}}, \quad (4.4)$$

где (Y_j, Z_j) – j -я точка в популяции юридических школ \mathbf{X} , а $\mu_y = \sum_{j=1}^{82} Y_j / 82$, $\mu_z = \sum_{j=1}^{82} Z_j / 82$.

Когда распределение вероятностей F известно (т.е. когда у нас есть полная совокупность \mathbf{X}), ответы на такие вопросы требуют не более чем арифметических операций. Для совокупности юридических школ перепись в таблице 3.2 дает $\mu_y = 597.5$, $\mu_z = 3.13$ и

$$\text{corr}(y, z) = 0.761. \quad (4.5)$$

Это первоначальное определение «статистики». Обычно у нас нет генеральной совокупности. Поэтому нам нужен статистический вывод, более современная статистическая теория для вывода свойств F из случайной выборки \mathbf{x} .

Если бы у нас была только выборка юридических школ размером 15, таблица 3.1, мы могли бы оценить $\text{corr}(y, z)$ с помощью коэффициента корреляции выборки

$$\widehat{\text{corr}}(y, z) = \frac{\sum_{j=1}^{15} (y_j - \hat{\mu}_y)(z_j - \hat{\mu}_z)}{[\sum_{j=1}^{15} (y_j - \hat{\mu}_y)^2 \sum_{j=1}^{15} (z_j - \hat{\mu}_z)^2]^{\frac{1}{2}}}, \quad (4.6)$$

где (y_j, z_j) - j -я точка в таблице 3.1, $j = 1, 2, \dots, 15$ и $\hat{\mu}_y = \sum_{j=1}^{15} y_j / 15$, $\hat{\mu}_z = \sum_{j=1}^{15} z_j / 15$. Таблица 3.1 дает $\mu_y = 600.3$, $\mu_z = 3.09$ и

$$\widehat{\text{corr}}(y, z) = 0.776. \quad (4.7)$$

Вот еще один пример оценки плагина. Предположим, нас интересует оценка вероятности того, что результат LSAT превышает 600, то есть

$$\theta = \frac{1}{82} \sum_1^{82} I_{\{Y_i > 600\}}. \quad (4.8)$$

Поскольку 39 из 82 баллов LSAT превышают 600, $\theta = 39/82 = 0.48$. Плагин оценка θ доли баллов LSAT, превышающих 600, равна

$$\hat{\theta} = \frac{1}{15} \sum_1^{15} I_{\{y_i > 600\}}. \quad (4.9)$$

Шесть из 15 баллов LSAT превышают 600, поэтому $\hat{\theta} = 6/15 = 0.4$.

Для кубика в Таблице 4.1 у нас нет данных переписи, а есть только выборка \mathbf{x} , поэтому на любые вопросы о справедливости кубика необходимо отвечать, исходя из эмпирических частот

$$\hat{F} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_6) = (0.13, 0.19, 0.10, 0.17, 0.14, 0.27). \quad (4.10)$$

Обсуждения статистического вывода сформулированы в терминах параметров и статистики. Параметр – это функция распределения вероятностей F . Статистика – это функция выборки \mathbf{x} . Таким образом, $\text{corr}(y, z)$, (4.4), является параметром F , а $\widehat{\text{corr}}(y, z)$, (4.6), является статистикой, основанной на \mathbf{x} . Точно так же f_k – это параметр F , а \hat{f}_k – статистика, $k = 1, 2, 3, \dots, 6$.

Иногда мы будем писать параметры напрямую как функции от F , например

$$\theta = t(F). \quad (4.11)$$

Это обозначение подчеркивает, что значение параметра θ получается путем применения некоторой процедуры численной оценки $t(\cdot)$ к функции распределения F . Например, если F – это распределение вероятностей на действительной прямой, математическое ожидание можно представить как параметр

$$\theta = t(F) = E_F(x). \quad (4.12)$$

Здесь $t(F)$ или θ вычисляется через математическое ожидание, то есть среднее значение x , взвешенное в соответствии с F . Для распределения F , такого как $F = Bi(n, p)$, мы можем вычислить $t(F) = np$. Даже если F неизвестна, форма $t(F)$ сообщает нам функциональное отображение из F в θ .

4.3 Принцип плагина

Принцип плагина представляет собой простой метод оценки параметров по выборкам. Плагин оценка параметра $\theta = t(F)$ определяется как

$$\hat{\theta} = t(\hat{F}). \quad (4.13)$$

Другими словами, мы оцениваем функцию $\theta = t(F)$ распределения вероятностей F той же функцией эмпирического распределения \hat{F} , $\hat{\theta} = t(\hat{F})$. (Статистические данные, подобные (4.13), которые используются для оценки параметров, иногда называют суммарной статистикой, а также оценками и оценщиками.)

Мы уже использовали принцип плагина при оценке f_k через \hat{f}_k и при оценке $\text{corr}(y, z)$ с помощью $\widehat{\text{corr}}(y, z)$. Чтобы убедиться в этом, обратите внимание, что наша совокупность F юридических школ может быть записана как $F = (f_1, f_2, \dots, f_{82})$, где каждое f_j , вероятность j -го юридической школы, имеет значение $1/82$. Это распределение вероятностей на \mathbf{X} , 82 парах юридических школ. Коэффициент корреляции генеральной совокупности можно записать как

$$\text{corr}(y, z) = \frac{\sum_{j=1}^{82} f_j(Y_j - \mu_y)(Z_j - \mu_z)}{[\sum_{j=1}^{82} f_j(Y_j - \mu_y)^2 \sum_{j=1}^{82} f_j(Z_j - \mu_z)^2]^{\frac{1}{2}}}, \quad (4.14)$$

где

$$\mu_y = \sum_{j=1}^{82} f_j Y_j, \quad \mu_z = \sum_{j=1}^{82} f_j Z_j. \quad (4.15)$$

Установка каждого $f_j = 1/82$ дает выражение (4.4). Теперь для нашей выборки $(x_1, x_2, \dots, x_{15})$ выборочная частота \hat{f}_j – это доля точек выборки, равная X_j :

$$\hat{f}_j = \#\{x_i = X_j\}/15, j = 1, 2, \dots, 82. \quad (4.16)$$

Для выборки из Таблицы 3.1 $\hat{f}_1 = 0, \hat{f}_2 = 0, \hat{f}_3 = 0, \hat{f}_4 = 1/15$ и т.д. Теперь подставив эти значения \hat{f}_j в выражения (4.15) и (4.14), получим $\hat{\mu}_y, \hat{\mu}_z$ и $\widehat{\text{corr}}(y, z)$ соответственно. То есть $\hat{\mu}_y, \hat{\mu}_z$ и $\widehat{\text{corr}}(y, z)$ – это плагин оценки μ_y, μ_z и $\text{corr}(y, z)$.

В общем, плагин оценка математического ожидания $\theta = E_F(x)$ равна

$$\hat{\theta} = E_{\hat{F}}(x) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}. \quad (4.17)$$

Насколько хорош принцип плагина? Обычно, если единственная доступная информация о F исходит из выборки \mathbf{x} , то $\hat{\theta} = t(\hat{F})$ не может быть улучшена как оценка $\theta = t(F)$, по крайней мере, не в обычном асимптотическом ($n \rightarrow \infty$) смысле статистической теории. Например, если \hat{f}_k – плагин оценка частоты $\#\{x_i = k\}/n$, то

$$\hat{f}_k \sim Bi(n, f_k)/n. \quad (4.18)$$

В этом случае оценка \hat{f}_k является несмещенной для f_k , $E(\hat{f}_k) = f_k$, с дисперсией $f_k(1 - f_k)/n$. Это наименьшая возможная дисперсия для несмещенной оценки f_k .

Мы будем использовать бутстреп для изучения смещения и стандартной ошибки плагин оценки $\hat{\theta} = t(\hat{F})$. Достоинство бутстрепа состоит в том, что он автоматически создает смещения и стандартные ошибки, независимо от того, насколько сложным может быть функциональное сопоставление $\theta = t(F)$. Мы увидим, что сам бутстреп является применением плагин принципа.

Принцип плагина менее эффективен в ситуациях, когда имеется информация о F , отличная от той, которая предоставлена выборкой \mathbf{x} . Мы можем знать или предполагать, что F является членом параметрического семейства, например семейства многомерных нормальных распределений. Или мы можем оказаться в ситуации регрессии, когда у нас есть набор случайных выборок $\mathbf{x}(z)$ в зависимости от переменной-предиктора z . Тогда, даже если нас интересует только F_{z_0} , функция распределения для некоторого конкретного значения z_0 из z , может быть информация о F_{z_0} в других выборках $\mathbf{x}(z)$, особенно тех, для которых z близок к z_0 .

Принцип плагина и бутстреп могут быть адаптированы к параметрическим семействам и регрессионным моделям. В следующих нескольких главах мы предполагаем, что находимся в ситуации, когда у нас есть только одна случайная выборка \mathbf{x} из полностью неизвестного распределения F . Это называется непараметрической задачей с одной выборкой.

Глава 5

Стандартные ошибки и оценки стандартных ошибок

5.1 Введение

Сводная статистика, такая как $\hat{\theta} = t(\hat{F})$, часто является первым результатом анализа данных. Следующее, что мы хотим знать – это точность $\hat{\theta}$. Бутстреп обеспечивает оценки точности, используя принцип плагина для оценки стандартной ошибки сводной статистики. Это предмет следующей главы. Сначала мы обсудим оценку стандартной ошибки среднего, где принцип плагина может быть реализован явно.

5.2 Стандартная ошибка среднего

Предположим, что x – вещественная случайная величина с распределением вероятностей F . Обозначим математическое ожидание и дисперсию F символами μ_F и σ_F^2 соответственно,

$$\mu_F = E_F(x), \quad \sigma_F^2 = var_F(x) = E_F[(x - \mu_F)^2]. \quad (5.1)$$

В главе 3 эти величины назывались μ_x и σ_x^2 . Здесь мы подчеркиваем зависимость от F . Альтернативное обозначение $var_F(x)$ для дисперсии, иногда сокращенное до $var(x)$, означает то же самое, что и σ_F^2 . В дальнейшем мы иногда будем писать

$$x \sim (\mu_F, \sigma_F^2) \quad (5.2)$$

чтобы кратко обозначить математическое ожидание и дисперсию x .

Теперь пусть (x_1, \dots, x_n) будет случайной выборкой размера n из распределения F . Среднее значение выборки $\bar{x} = \sum_{i=1}^n x_i/n$ имеет математическое ожидание μ_F и дисперсию σ_F^2/n ,

$$\bar{x} \sim (\mu_F, \sigma_F^2/n). \quad (5.3)$$

Другими словами, математическое ожидание \bar{x} такое же, как ожидание одного x , но дисперсия \bar{x} равна $1/n$ дисперсии x . Это причина использования средних значений: чем больше n , тем меньше $var(\bar{x})$, поэтому большее n означает лучшую оценку μ_F .

Стандартная ошибка среднего \bar{x} , записанная как $se_F(\bar{x})$ или $se(\bar{x})$, является квадратным корнем из дисперсии \bar{x} ,

$$se_F(\bar{x}) = [var_F(\bar{x})]^{1/2} = \sigma_F / \sqrt{n}. \quad (5.4)$$

Стандартная ошибка – это общий термин для стандартного отклонения сводной статистики. Это наиболее распространенный способ индикации статистической точности. Грубо говоря, мы ожидаем, что $|\bar{x} - \mu_F|$ будет меньше одной стандартной ошибки примерно в 68% случаев и меньше двух стандартных ошибок примерно в 95% случаев.

Эти проценты основаны на центральной предельной теореме. При довольно общих условиях на F распределение \bar{x} будет приблизительно нормальным при увеличении n , что мы можем записать как

$$\bar{x} \sim N(\mu_F, \sigma_F^2/n). \quad (5.5)$$

Математическое ожидание μ_F и дисперсия σ_F^2/n в (5.5) точны, только нормальность является приблизительной. Используя (5.5), таблица нормального распределения дает

$$Prob\{|\bar{x} - \mu_F| < \frac{\sigma_F}{\sqrt{n}}\} = 0.683 \quad Prob\{|\bar{x} - \mu_F| < \frac{2\sigma_F}{\sqrt{n}}\} = 0.954, \quad (5.6)$$

как показано на Рисунке 5.1. Одно из преимуществ бутстрепа заключается в том, что не нужно полностью полагаться на центральную предельную теорему.

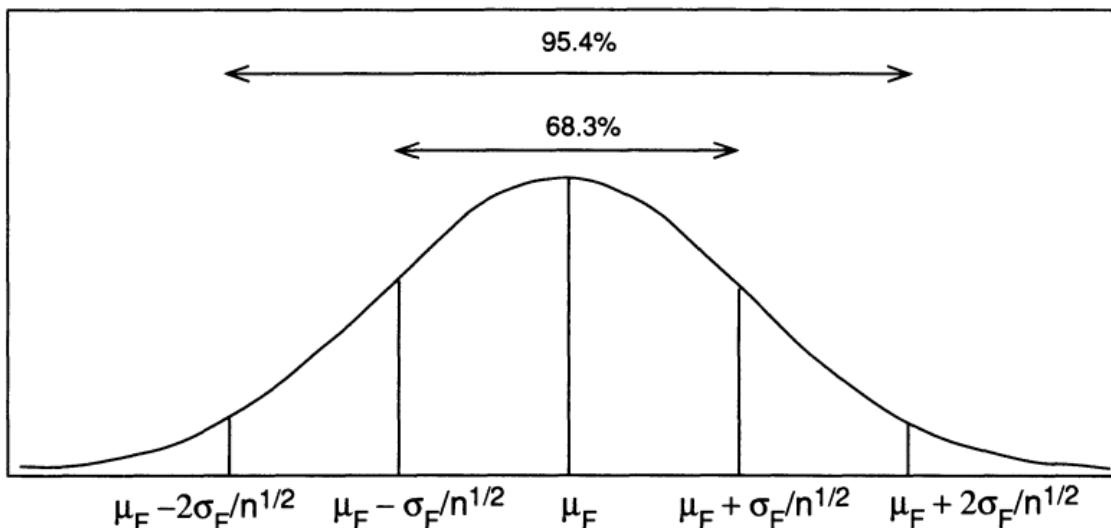


Figure 5.1. For large values of n , the mean \bar{x} of a random sample from F will have an approximate normal distribution with mean μ_F and variance σ_F^2/n .

Простой пример показывает ограничения аппроксимации центральной предельной теоремы. Предположим, что F – это распределение, которое ставит вероятность только для двух исходов, 0 или 1, скажем

$$Prob\{x = 1\} = p \quad Prob\{x = 0\} = 1 - p. \quad (5.7)$$

Здесь p – параметр F , часто называемый вероятностью успеха, имеющий значение от 0 до 1. Случайная выборка $F \rightarrow (x_1, x_2, \dots, x_n)$ может рассматриваться как n независимых подбрасываний монеты с вероятностью успеха (или «орла», или $x = 1$) равной p . Тогда сумма $s = \sum_{i=1}^n x_i$ – количество успехов в n независимых бросках монеты; s имеет биномиальное распределение (3.3),

$$s \sim Bi(n, p). \quad (5.8)$$

Среднее значение $\bar{x} = s/n$ равно \hat{p} в плагин оценке p . Распределение (5.7) имеет $\mu_F = p$, $\sigma_F^2 = p(1-p)$, поэтому (5.3) дает

$$\hat{p} \sim (p, p(1-p)/n) \quad (5.9)$$

для среднего и дисперсии \hat{p} . Другими словами, \hat{p} – это несмешенная оценка p , $E(\hat{p}) = p$, со стандартной ошибкой

$$se(\hat{p}) = \left[\frac{p(1-p)}{n} \right]^{1/2}. \quad (5.10)$$

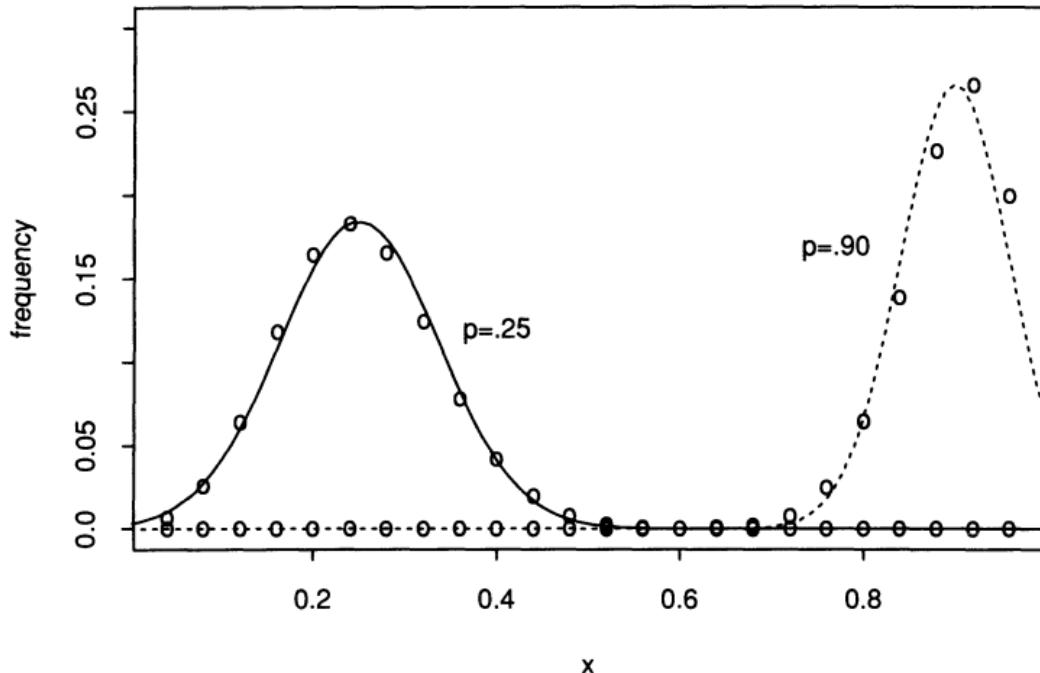


Figure 5.2. Comparison of the binomial distribution with the normal distribution suggested by the central limit theorem; $n = 25$, $p = .25$ and $p = .90$. The smooth curves are the normal densities, see problem 5.3; circles indicate the binomial probabilities (3.5). The approximation is good for $p = .25$, but is somewhat off for $p = .90$.

На рис. 5.2 показана центральная предельная теорема, работающая для биномиального распределения при $n = 25$, $p = 0.25$ и $p = 0.90$. Центральная предельная теорема дает хорошее приближение к биномиальному распределению для $n = 25$, $p = 0.25$, но несколько хуже для $n = 25$, $p = 0.9$.

5.3 Оценка стандартной ошибки среднего

Предположим, что у нас есть случайная выборка чисел $F \rightarrow x_1, x_2, \dots, x_n$, например контрольные измерения $n = 9$ для данных о мышах из таблицы 2.1. Мы вычисляем оценку \bar{x} для математического ожидания μ_F , равного 56.22 для данных о мышах, и хотим знать стандартную ошибку \bar{x} . Формула (5.4) $se_F(\bar{x}) = \sigma_F/\sqrt{n}$ включает неизвестное распределение F и поэтому не может использоваться напрямую.

На этом этапе мы можем использовать принцип плагина: мы подставляем \hat{F} вместо F в формуле $se_F(\bar{x}) = \sigma_F/\sqrt{n}$. Плагин оценка $\sigma_F = [E_F(x - \mu_F)^2]^{1/2}$ равна

$$\hat{\sigma} = \sigma_{\hat{F}} = \left\{ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right\}^{1/2}, \quad (5.11)$$

поскольку $\mu_{\hat{F}} = \bar{x}$ и $E_{\hat{F}}g(x) = \frac{1}{n} \sum_{i=1}^n g(x_i)$ для любой функции g . Это дает оценку стандартной ошибки $\hat{se}(\bar{x}) = se_{\hat{F}}(\bar{x})$,

$$\hat{se}(\bar{x}) = \sigma_{\hat{F}}/\sqrt{n} = \left\{ \sum_{i=1}^n (x_i - \bar{x})^2/n^2 \right\}^{1/2}. \quad (5.12)$$

Для контрольной группы данных о мышах $\hat{se}(\bar{x}) = 13.33$.

Формула (5.12) немного отличается от обычной оценки стандартной ошибки (2.2). Это потому, что σ_F обычно оценивается как $\bar{\sigma} = \{\sum(x_i - \bar{x})^2/(n - 1)\}^{1/2}$, а не как $\hat{\sigma}$, (5.11). Деление на $n - 1$ вместо n делает $\bar{\sigma}^2$ несмешенной для σ_F^2 . Для большинства целей $\hat{\sigma}$ так же хороша, как $\bar{\sigma}$ для оценки σ_F .

Обратите внимание, что мы использовали принцип плагина дважды: сначала для оценки математического ожидания μ_F с помощью $\mu_{\hat{F}} = \bar{x}$, а затем для оценки стандартной ошибки $se_F(\bar{x})$ с помощью $se_{\hat{F}}(\bar{x})$. Бутстреп оценка стандартной ошибки, о которой идет речь в главе 6, сводится к использованию принципа плагина для оценки стандартной ошибки произвольной статистики $\hat{\theta}$. Здесь мы видели, что если $\hat{\theta} = \bar{x}$, то этот подход приводит к (почти) обычной оценке стандартной ошибки. Как мы увидим, преимущество бутстрапа в том, что его можно применить практически к любой статистике $\hat{\theta}$, а не только к среднему значению \bar{x} .

Глава 6

Бутстреп оценка стандартной ошибки

6.1 Введение

Предположим, мы находимся в следующей общей ситуации анализа данных: была обнаружена случайная выборка $\mathbf{x} = (x_1, x_2, \dots, x_n)$ из неизвестного распределения вероятностей F , и мы хотим оценить интересующий параметр $\theta = t(F)$ на основе \mathbf{x} . Для этого мы вычисляем оценку $\hat{\theta} = s(\mathbf{x})$ из \mathbf{x} . [Обратите внимание, что $s(\mathbf{x})$ может быть плагин оценкой $t(\hat{F})$, но это не обязательно.] Насколько точна $\hat{\theta}$? Бутстреп был представлен в 1979 году как компьютерный метод оценки стандартной ошибки $\hat{\theta}$. Он имеет то преимущество, что он полностью автоматический. Самостоятельная оценка стандартной ошибки не требует теоретических вычислений и доступна независимо от того, насколько математически сложной может быть оценка $\hat{\theta} = s(\mathbf{x})$. Это описано и проиллюстрировано в этой главе.

6.2 Бутстреп оценка стандартной ошибки

Бутстреп методы зависят от понятия бутстреп выборки. Пусть \hat{F} – эмпирическое распределение, присваивающее вероятность $1/n$ на каждое из наблюдаемых значений x_i , $i = 1, 2, \dots, n$, как описано в главе 4. Бутстреп выборка является случайной выборкой размера n , набранной из \hat{F} , скажем, $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*)$,

$$\hat{F} \rightarrow (x_1^*, x_2^*, \dots, x_n^*). \quad (6.1)$$

« $*$ » указывает, что \mathbf{x}^* не является фактическим набором данных \mathbf{x} , а скорее рандомизированной или перезапущенной версией \mathbf{x} .

Есть еще один способ сказать (6.1): точки бутстреп данных $x_1^*, x_2^*, \dots, x_n^*$ являются случайной выборкой размера n , выбранной с заменой из совокупности n объектов (x_1, x_2, \dots, x_n) . Таким образом, мы могли бы иметь $x_1^* = x_7, x_2^* = x_3, x_3^* = x_3, x_4^* = x_{22}, \dots, x_n^* = x_7$. Набор бутстреп данных $(x_1^*, x_2^*, \dots, x_n^*)$ состоит из элементов исходного набора данных (x_1, x_2, \dots, x_n) , некоторые из которых появляются ноль раз, некоторые появляются один раз, некоторые появляются дважды, и так далее.

В соответствии с набором бутстреп данных \mathbf{x}^* , бутстреп репликация $\hat{\theta}$ – это

$$\hat{\theta}^* = s(\mathbf{x}^*). \quad (6.2)$$

Величина $s(\mathbf{x}^*)$ является результатом применения той же функции $s(\cdot)$ к \mathbf{x}^* , которая была применена к \mathbf{x} . Например, если $s(\mathbf{x})$ является выборочным средним значением \bar{x} , то $s(\mathbf{x}^*)$ – это среднее значение набора бутстреп данных, $\bar{x}^* = \sum_{i=1}^n x_i^*/n$.

Бутстреп оценка $se_F(\hat{\theta})$ стандартной ошибки статистики $\hat{\theta}$ представляет собой плагин оценку, которая использует эмпирическую функцию распространения \hat{F} вместо неизвестного распределения F . В частности, бутстреп оценка $se_F(\hat{\theta})$ определяется, как

$$se_{\hat{F}}(\hat{\theta}^*). \quad (6.3)$$

Другими словами, бутстреп оценка $se_F(\hat{\theta})$ является стандартной ошибкой $\hat{\theta}$ для наборов данных размером n , случайным образом выбранных из \hat{F} .

Формула (6.3) называется идеальной бутстреп оценкой ошибки $\hat{\theta}$. К сожалению, для практически любой оценки $\hat{\theta}$, кроме среднего, нет точной формулы (5.4), которая позволяет вычислить числовое значение идеальной оценки точно. Бутстреп алгоритм, описанный ниже, является вычислительным способом получения хорошего приближения к численному значению $se_{\hat{F}}(\hat{\theta}^*)$.

Легко реализовать бутстреп выборку на компьютере. Устройство выбирает случайные целые числа i_1, i_2, \dots, i_n , каждое из которых равняется любому значению между 1 и n с вероятностью $1/n$. Бутстреп выборка состоит из соответствующих членов \mathbf{x} ,

$$x_1^* = x_{i_1}, x_2^* = x_{i_2}, \dots, x_n^* = x_{i_n}. \quad (6.4)$$

Бутстреп алгоритм работает путем выбора множества независимых бутстреп выборок, оценки соответствующих бутстреп репликаций и оценки стандартной ошибки $\hat{\theta}$ через эмпирическое стандартное отклонение репликаций. Результат называется бутстреп оценкой стандартной ошибки, обозначенной \hat{se}_B , где B – количество используемых бутстреп выборок.

Алгоритм 6.1 – это более явное описание бутстреп процедуры для оценки стандартной ошибки $\hat{\theta} = s(\mathbf{x})$ из наблюдаемых данных \mathbf{x} .

Алгоритм 6.1
 Бутстреп алгоритм для оценки стандартных ошибок

1. Выберите B независимых бутстреп выборок $x^{*1}, x^{*2}, \dots, x^{*B}$, каждый из которых состоит из n точек данных, выбранных с заменой их \mathbf{x} , как в (6.1) или (6.4). [Для оценки стандартной ошибки B обычно будет в диапазоне 25-200, см. Таблицу 6.1.]
2. Оцените бутстреп репликацию, соответствующую каждой бутстреп выборке,

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}) \quad b = 1, 2, \dots, B. \quad (6.5)$$

3. Оцените стандартную ошибку $se_F(\hat{\theta})$ через выборочное стандартное отклонение B репликаций

$$\widehat{se}_B = \left\{ \sum_{b=1}^B [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2 / (B - 1) \right\}^{1/2}, \quad (6.6)$$

где $\hat{\theta}^*(\cdot) = \sum_{b=1}^B \hat{\theta}^*(b) / B$.

На рисунке 6.1 изображена схематическая диаграмма бутстреп алгоритма стандартных ошибок.

Предел \widehat{se}_B по B стремится к бесконечности – это идеальная бутстреп оценка $se_F(\hat{\theta})$,

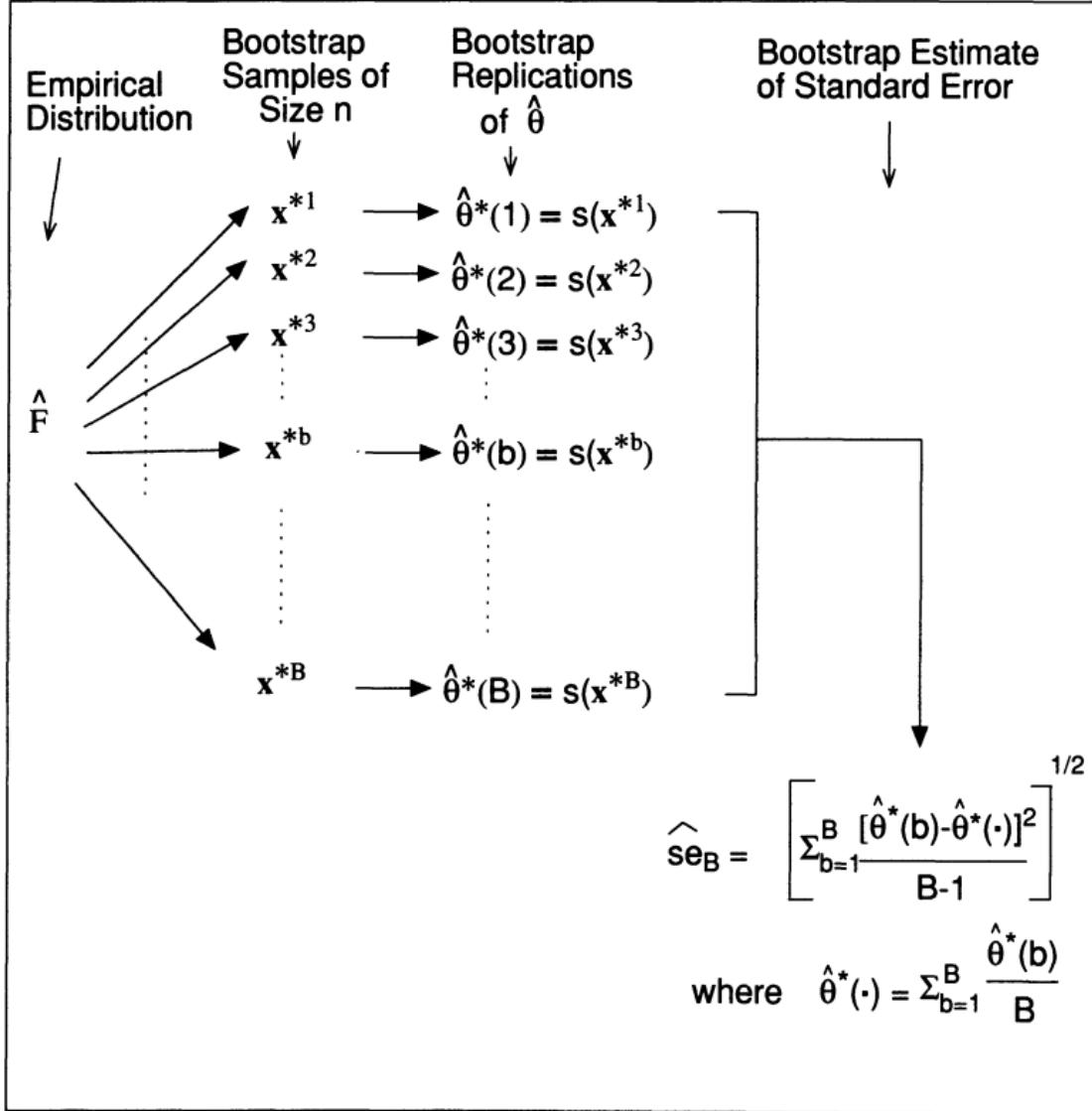
$$\lim_{B \rightarrow \infty} \widehat{se}_B = se_{\hat{F}} = se_{\hat{F}}(\hat{\theta}^*). \quad (6.7)$$

Тот факт, что \widehat{se}_B стремится к $se_{\hat{F}}$ при B , стремящимся к бесконечности, позволяет сказать, что эмпирическое стандартное отклонение приближается к стандартному отклонению совокупности с увеличением количества репликаций. «Совокупность» в этом случае является совокупностью значений $\hat{\theta}^* = s(\mathbf{x}^*)$, где $\hat{F} \rightarrow (x_1^*, x_2^*, \dots, x_n^*) = \mathbf{x}^*$.

Идеальная бутстреп оценка $se_{\hat{F}}\hat{\theta}^*$ и его приближение \widehat{se}_B иногда называют непараметрическими бутстреп оценками, потому что они основаны на \hat{F} , непараметрической оценке F . В разделе 6.5 мы обсудим параметрический бутстреп, который использует другую оценку F .

Немного об обозначениях: в (6.7) мы пишем $se_{\hat{F}}(\hat{\theta}^*)$, а не $se_{\hat{F}}(\hat{\theta})$, чтобы избежать путаницы между $\hat{\theta}$, значением $s(\mathbf{x})$ на основе наблюдаемых данных, и $\hat{\theta}^* = s(\mathbf{x}^*)$, случайной величиной на основе бутстреп выборки. Более подробное обозначение $se_{\hat{F}}(\hat{\theta}(\mathbf{x}^*))$ подчеркивает, что $se_{\hat{F}}$ является бутстрепированной стандартной ошибкой: фактические данные \mathbf{x} остаются фиксированным в (6.7); Случайность в расчете исходит из изменчивости бутстреп выборок \mathbf{x}^* для данного \mathbf{x} . Точно так же мы будем писать $E_{\hat{F}}g(\mathbf{x}^*)$, чтобы указать бутстрепированное математическое ожидание функции $g(\mathbf{x}^*)$: математическое ожидание с фиксированным \mathbf{x} (и \hat{F}) и случайным \mathbf{x}^* в соответствии с (6.1).

Figure 6.1. The bootstrap algorithm for estimating the standard error of a statistic $\hat{\theta} = s(\mathbf{x})$; each bootstrap sample is an independent random sample of size n from \hat{F} . The number of bootstrap replications B for estimating a standard error is usually between 25 and 200. As $B \rightarrow \infty$, \hat{se}_B approaches the plug-in estimate of $se_{\hat{F}}(\hat{\theta})$.



Всего существует C_n^{2n-1} различных бутстреп выборок. Обозначим их через z^1, z^2, \dots, z^m , где $m = C_n^{2n-1}$. Например, если $n = 2$, отдельными выборками являются $(x_1, x_1), (x_2, x_2)$ и (x_1, x_2) ; поскольку порядок не имеет значения, (x_2, x_1) совпадает с (x_1, x_2) . Вероятность получения одной из этих выборок при выборе с заменой может быть получена из полиномиального распределения. Обозначим вероятность j -й выборки через ω_j , $j = 1, 2, \dots, C_n^{2n-1}$. Тогда прямым способом вычисления идеальной бутстреп оценки стандартной ошибки будет использование стандартного отклонения совокупности m бутстреп значений $s(z^j)$:

$$se_{\hat{F}}(\hat{\theta}^*) = \left[\sum_{j=1}^m \omega_j \{s(z^j) - s(\cdot)\}^2 \right]^{1/2} \quad (6.8)$$

где $s(\cdot) = \sum_{j=1}^m \omega_j s(z^j)$. Сложность этого подхода заключается в том, что, если n не достаточно мало (≤ 5), число C_n^{2n-1} очень велико, что делает вычисление

(6.8) непрактичным. Отсюда необходимость в бутстррап выборках, описанных выше.

6.3 Пример: коэффициент корреляции

Мы уже видели два примера бутстррап оценок стандартной ошибки для среднего и медианного значения для экспериментальной группы данных о мышах, таблица 2.1. В качестве второго примера рассмотрим выборочный коэффициент корреляции между $y = LSAT$ и $z = GPA$ для $n = 15$ точек данных о юридических школах, таблица 3.1, $\widehat{corr}(y, z) = 0.776$. Насколько точна оценка 0.776? В таблице 6.1 показана бутстррап оценка стандартной ошибки \widehat{se}_B для B в диапазоне от 25 до 3200. Последнее значение, $\widehat{se}_{3200} = 0.132$, является нашей оценкой $se_F(\widehat{corr})$. Позже мы увидим, что \widehat{se}_{200} почти так же хороша для оценки se_F , как \widehat{se}_{3200} .

Глядя на правую часть рисунка 3.1, читатель может представить себе, как работает генерация бутстррап выборок. Выборочная корреляция по $n = 15$ исходным точкам данных составляет $\widehat{corr} = 0.776$. Бутстррап выборка состоит из 15 точек, выбранных случайным образом и заменяющих исходные 15. Корреляция бутстррап выборки представляет собой бутстррап репликацию \widehat{corr}^* , которая может быть больше или меньше, чем \widehat{corr} . Независимые повторения генерации бутстррап выборок дают бутстррап репликации $\widehat{corr}^*(1), \widehat{corr}^*(2), \dots, \widehat{corr}^*(B)$. Наконец, \widehat{se}_B – выборочное стандартное отклонение значений $\widehat{corr}^*(b)$.

Table 6.1. The bootstrap estimate of standard error for $\widehat{corr}(y, z) = .776$, the law school data of Table 3.1, $n = 15$; a run of 3200 bootstrap replications gave the tabled values of \widehat{se}_B as B increased from 25 to 3200.

| $B:$ | 25 | 50 | 100 | 200 | 400 | 800 | 1600 | 3200 |
|-------------------|------|------|------|------|------|------|------|------|
| $\widehat{se}_B:$ | .140 | .142 | .151 | .143 | .141 | .137 | .133 | .132 |

Левая панель рисунка 6.2 представляет собой гистограмму 3200 бутстррап репликаций $\widehat{corr}^*(b)$. Всегда рекомендуется просматривать бутстррап данные графически, а не полагаться полностью на одну сводную статистику, такую как \widehat{se}_B . В примере корреляции может оказаться, что несколько выпадающих значений $\widehat{corr}^*(b)$ сильно раздувают \widehat{se}_B , и в этом случае стоит использовать более надежную меру стандартного отклонения; Выводы, основанные на нормальной кривой, как в (5.6) и на рисунке 5.1, сомнительны, когда бутстррап гистограмма явно ненормальна.

В примере с юридическими школами у нас есть полная совокупность \mathbf{X} из $N = 82$ школ, Таблица 3.2. В правой части рисунка 6.2 показана гистограмма $\widehat{corr}(y, z)$ для 3200 выборок размера $n = 15$, взятых из \mathbf{X} . Другими словами, 3200 случайных выборок $\mathbf{x} = (x_1, x_2, \dots, x_{15})$ были составлены с заменой из 82 точек в \mathbf{X} , и $\widehat{corr}(\mathbf{x})$ оценивался для каждого из них. Стандартное отклонение 3200 значений $\widehat{corr}(\mathbf{x})$ составило 0.131, таким образом \widehat{se}_B является хорошей оценкой стандартной ошибки генеральной совокупности. Что еще более впечатляюще, бутстррап гистограмма слева сильно напоминает гистограмму справа. Помните, что в реальной проблеме у нас была бы только информация слева, из которой мы пытались бы вывести ситуацию справа.

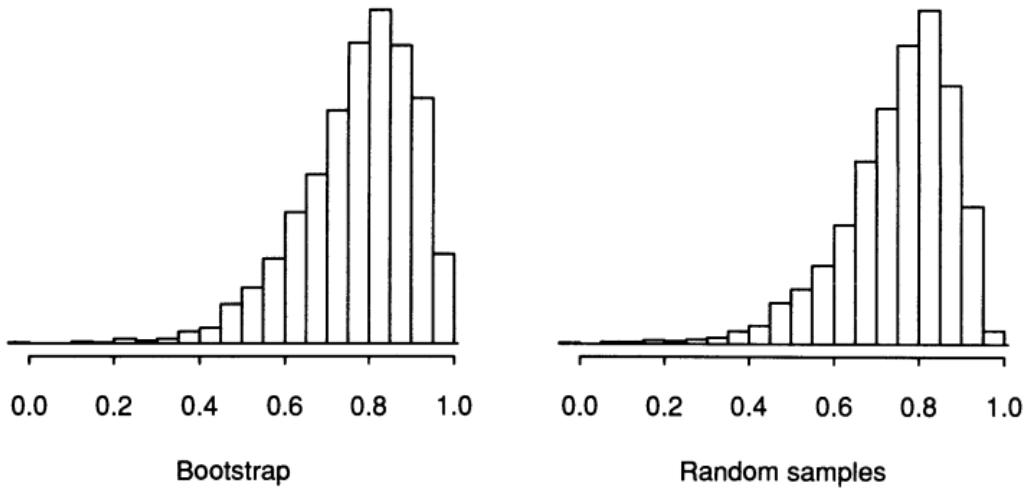


Figure 6.2. Left panel: histogram of 3200 bootstrap replications of $\widehat{\text{corr}}(\mathbf{x}^*)$, from the law school data, $n = 15$, Table 3.1. Right panel: histogram of 3200 replications $\widehat{\text{corr}}(\mathbf{x})$, where \mathbf{x} is a random sample of size n from the $N = 82$ points in the law school population, Table 3.2. The bootstrap histogram strongly resembles the population histogram. Both are notably non-normal.

6.4 Количество бутстреп репликаций B

Насколько большим мы должны взять B , количество бутстреп репликаций, используемых для оценки $\widehat{s\epsilon}_B$? Идеальная бутстреп оценка $\widehat{s\epsilon}_\infty$ использует $B = \infty$, и в этом случае $\widehat{s\epsilon}_\infty$ равно плагин оценке $se_{\hat{F}}(\hat{\theta}^*)$. Формула (5.12) дает $\widehat{s\epsilon}_\infty$ для $\hat{\theta} = \bar{x}$, но для большинства других статистических данных мы должны фактически выполнить генерацию бутстреп выборок. Время, затрачиваемое компьютером, которое в основном зависит от того, сколько времени требуется для оценки бутстреп репликаций (6.5), линейно увеличивается с B . Временные ограничения могут диктовать небольшое значение B , если $\hat{\theta} = s(\mathbf{x})$ – очень сложная функция.

Нам нужно такое же хорошее поведение от оценки стандартной ошибки, что и от оценки любой другой интересующей величины: небольшая систематическая ошибка и небольшое стандартное отклонение. Бутстреп оценка стандартной ошибки обычно имеет относительно небольшое смещение. Идеальная начальная оценка $\widehat{s\epsilon}_\infty$ имеет наименьшее возможное стандартное отклонение среди почти несмешанных оценок $se_F(\hat{\theta})$, по крайней мере, в асимптотическом ($n \rightarrow \infty$) смысле. Эти хорошие свойства вытекают из того факта, что $\widehat{s\epsilon}_\infty$ – это плагин оценка $se_{\hat{F}}(\hat{\theta})^*$. Нетрудно показать, что $\widehat{s\epsilon}_B$ всегда имеет большее стандартное отклонение, чем $\widehat{s\epsilon}_\infty$. Практический вопрос: насколько большее?

Приблизительный, но вполне удовлетворительный ответ можно сформулировать в терминах коэффициента вариации $\widehat{s\epsilon}_B$, отношения стандартного отклонения $\widehat{s\epsilon}_B$ к его математическому ожиданию. Повышенная изменчивость из-за остановки после B бутстреп репликаций, а не бесконечности, отражается

в увеличенном коэффициенте вариации,

$$cv(\hat{se}_B) = \left\{ cv(\hat{se}_\infty)^2 + \frac{E(\Delta) + 2}{4B} \right\}^{1/2}. \quad (6.9)$$

Здесь Δ – параметр, который измеряет, насколько длиннохвостым является распределение $\hat{\theta}^*$: если Δ равно нулю для нормального распределения, оно колеблется от -2 для короткохвостых распределений до произвольно больших значений, когда F длиннохвостное. На практике δ обычно не превышает 10. Коэффициент вариации в уравнении (6.9) относится к вариации как на уровне повторных выборок (бутстреп), так и на уровне исходной выборки. Идеальная оценка $\hat{se}_\infty = se_{\hat{F}}(\hat{\theta}^*)$ не идеальна. Она все еще может иметь значительную изменчивость в качестве оценки $se_F(\hat{\theta})$ из-за изменчивости \hat{F} как оценки F . Например, если x_1, x_2, \dots, x_n является случайной выборкой из нормального распределения и $\hat{\theta} = \bar{x}$, тогда $cv(\hat{se}_\infty) = 1/\sqrt{2n}$, равное 0.22 для $n = 10$. Формула (6.9) имеет важное практическое следствие: для значений $cv(\hat{se}_\infty)$ и Δ , которые могут возникнуть на практике, $cv(\hat{se}_B)$ не намного больше, чем $cv(\hat{se}_\infty)$ для $B \geq 200$.

Таблица 6.2 сравнивает $cv(\hat{se}_B)$ с $cv(\hat{se}_\infty)$ для различных вариантов B , предполагая $\Delta = 0$. Очень часто можно ожидать, что $cv(\hat{se}_\infty)$ будет не меньше, чем 0.10, и в этом случае $B = 100$ дает вполне удовлетворительные результаты.

Table 6.2. The coefficient of variation of \hat{se}_B as a function of the coefficient of variation of the ideal bootstrap estimate \hat{se}_∞ and the number of bootstrap samples B ; from formula (6.9) assuming $\Delta = 0$.

| | | $B \rightarrow$ | | | | |
|-----------------------|-----|-----------------|-----|-----|-----|----------|
| | | 25 | 50 | 100 | 200 | ∞ |
| $cv(\hat{se}_\infty)$ | .25 | .29 | .27 | .26 | .25 | .25 |
| \downarrow | .20 | .24 | .22 | .21 | .21 | .20 |
| | .15 | .21 | .18 | .17 | .16 | .15 |
| | .10 | .17 | .14 | .12 | .11 | .10 |
| | .05 | .15 | .11 | .09 | .07 | .05 |
| | .00 | .14 | .10 | .07 | .05 | .00 |

Вот два практических правила, взятых из опыта:

1. Даже небольшое количество бутстреп репликаций, скажем, $B = 25$, обычно является информативным. $B = 50$ часто бывает достаточно, чтобы дать хорошую оценку $se_F(\hat{\theta})$.
2. Очень редко для оценки стандартной ошибки требуется более $B = 200$ репликаций. (Для доверительных бутстреп интервалов требуются гораздо большие значения B .)

Аппроксимации, полученные путем случайной выборки или моделирования, называются оценками Монте-Карло. Вычислительные методы, отличные от прямого моделирования Монте-Карло, иногда могут во много раз сократить количество повторений B , необходимых для достижения заданной точности.

Между тем стоит помнить, что бутстреп данные, как и реальные данные, заслуживают внимательного изучения. В частности, отображение гистограммы бутстреп репликаций почти никогда не бывает пустой тратой времени.

6.5 Параметрический бутстреп

Может показаться странным использование бутстреп алгоритма для оценки стандартных ошибок, когда можно использовать формулу из учебника. Фактически, бутстреп выборки могут генерироваться параметрически, в этом случае результаты тесно связаны с формулами стандартных ошибок из учебников.

Параметрическая бутстреп оценка стандартной ошибки определяется как

$$se_{\hat{F}_{par}}(\hat{\theta}^*), \quad (6.10)$$

где \hat{F}_{par} – оценка F , полученная из параметрической модели данных. Здесь мы приведем простой пример, чтобы проиллюстрировать идею. Для данных о юридических школах, вместо оценки F эмпирическим распределением \hat{F} , мы могли бы предположить, что популяция имеет двумерное нормальное распределение. Разумные оценки среднего значения и ковариации этой совокупности даны как (\bar{y}, \bar{z}) и

$$\frac{1}{14} \begin{pmatrix} \sum(y_i - \bar{y})^2 & \sum(y_i - \bar{y})(z_i - \bar{z}) \\ \sum(y_i - \bar{y})(z_i - \bar{z}) & \sum(z_i - \bar{z})^2 \end{pmatrix}. \quad (6.11)$$

Обозначим двумерную нормальную популяцию с этим средним значением и ковариацией как \hat{F}_{norm} ; это пример параметрической оценки совокупности F . Используя это, параметрическая бутстреп оценка стандартной ошибки корреляции $\hat{\theta}$ является $se_{\hat{F}_{norm}}(\hat{\theta}^*)$. Как и в непараметрическом случае, идеальная параметрическая бутстреп оценка не может быть легко вычислена, за исключением тех случаев, когда $\hat{\theta}$ является средним. Поэтому мы аппроксимируем идеальную бутстреп оценку с помощью бутстреп выборок, но другим способом, чем раньше. Вместо выборок с заменой из исходных данных мы берем B выборок размера n из параметрической оценки генеральной совокупности \hat{F}_{par} :

$$\hat{F}_{par} \rightarrow (x_1^*, x_2^*, \dots, x_n^*). \quad (6.12)$$

После генерации бутстреп выборок мы действуем точно так же, как в шагах 2 и 3 бутстреп алгоритма из раздела 6.2: мы оцениваем нашу статистику для каждой бутстреп выборки, а затем вычисляем стандартное отклонение B бутстреп репликаций.

В примере с коэффициентом корреляции, предполагая двумерную нормальную совокупность, мы берем B выборок размером 15 из \hat{F}_{norm} и вычисляем коэффициент корреляции для каждой бутстреп выборки. На левой панели рисунка 6.3 показана гистограмма для $B = 3200$ бутстреп репликаций, полученных таким образом. Это очень похоже на гистограммы на рисунке 6.2. Параметрическая бутстреп оценка стандартной ошибки для этих повторений была 0.124, что близко к значению 0.131, полученному на непараметрических бутстреп выборках.

Учебная формула для стандартной ошибки коэффициента корреляции составляет $(1 - \hat{\theta}^2)/\sqrt{n - 3}$. Подставляя $\hat{\theta} = 0.776$, она дает значение 0.15 для данных о юридических школах.

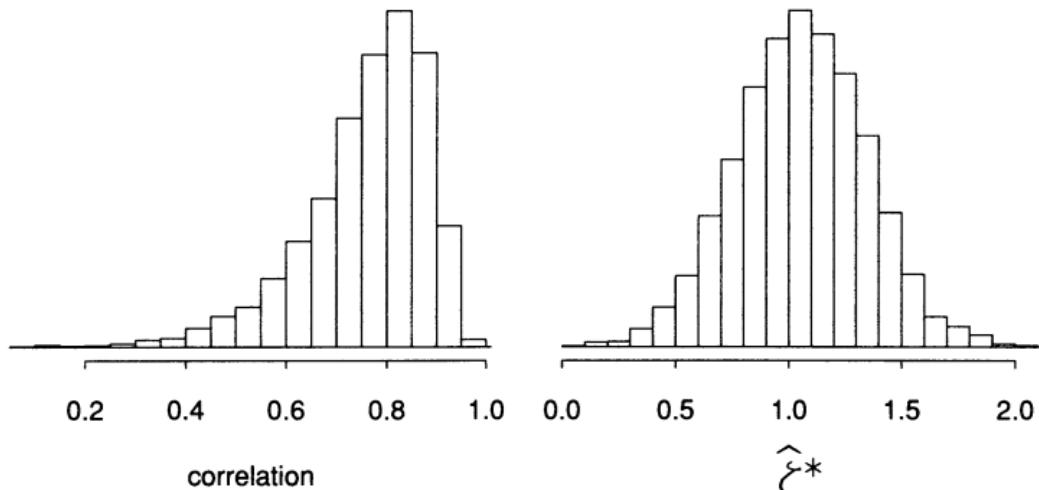


Figure 6.3. *Left panel: histogram of 3200 parametric bootstrap replications of $\widehat{\text{corr}}(\mathbf{x}^*)$, from the law school data, $n = 15$. Right panel: histogram of 3200 replications of $\hat{\zeta}$, Fisher's transformation of the correlation coefficient, defined in (6.12). The left histogram looks much like the histograms of (6.2), while the right histogram looks quite normal as predicted by statistical theory.*

Мы можем провести дальнейшее сравнение с нашим параметрическим бутстреп результатом. Учебные материалы также утверждают, что преобразование Фишера, примененное к $\hat{\theta}$

$$\hat{\zeta} = 0.5 \log \left(\frac{1 + \hat{\theta}}{1 - \hat{\theta}} \right) \quad (6.13)$$

приблизительно нормально распределено со средним $\zeta = 0.5 \log \left(\frac{1 + \theta}{1 - \theta} \right)$ и стандартным отклонением $1/\sqrt{n-3}$, где θ является коэффициентом корреляции совокупности. Исходя из этого, обычно выполняется вывод для ζ и затем преобразуется обратно, чтобы сделать вывод о коэффициенте корреляции. Чтобы сравнить это с нашим параметрическим бутстреп анализом, мы вычислили $\hat{\zeta}$ вместо $\hat{\theta}$ для каждой из наших 3200 бутстреп выборок. Гистограмма значений $\hat{\zeta}^*$ показана на правой панели рисунка 6.3 и выглядит вполне нормально. Кроме того, стандартное отклонение 3200 значений $\hat{\zeta}^*$ было 0.290, что очень близко к значению $1/\sqrt{15-3} = 0.289$.

Это соглашение выполняется в целом. Большинство формул для стандартных ошибок в учебниках являются приближениями, основанными на нормальной теории, и обычно дают ответы, близкие к параметрическому бутстрепу, который отбирает выборки из нормального распределения. Взаимосвязь между бутстрепом и традиционной статистической теорией – более сложная математическая тема.

У бутстрепа есть два несколько отличных друг от друга преимущества по сравнению с традиционными методами из учебников: 1) при использовании в непараметрическом режиме он избавляет аналитика от необходимости делать параметрические предположения о форме базовой совокупности, и 2) при использовании в параметрическом режиме он обеспечивает более точные ответы, чем формулы из учебника, и могут дать ответы на задачи, для которых не существует формул из учебника.

Большая часть этого пособия сосредоточена на непараметрическом приме-

нении бутстрепа. Параметрический бутстреп полезен в задачах, где доступны некоторые знания о форме генеральной совокупности, а также для сравнение с непараметрическим анализом. Однако основная причина использования параметрических допущений в традиционном статистическом анализе состоит в том, чтобы облегчить вывод формул для стандартных ошибок из учебников. Поскольку нам не нужны формулы в бутстреп подходе, мы можем избежать ограничительных параметрических предположений.

Глава 7

Бутстреп и стандартные ошибки: некоторые примеры

7.1 Введение

До внедрения компьютеров вычисляли стандартные ошибки используя методы математического анализа и предположения о распределении, что часто предполагало много работы на механических калькуляторах. Один такой классический результат был дан в разделе 6.5: он относится к выборочному коэффициенту корреляции $\widehat{\text{corr}}(y, z)$ (4.6). Если сделать предположение о том, что n элементов выборки (y_i, z_i) взяты из двумерного нормального распределения с функцией распределения F , тогда разумной оценкой стандартной ошибки $\widehat{\text{corr}}$ будет

$$\widehat{s}_{\text{normal}} = (1 - \widehat{\text{corr}}^2)/\sqrt{n - 3} \quad (7.1)$$

Ясно, что может последовать возражение относительно использования двумерного нормального распределения: на каком основании делается предположение о том, что F подчиняется нормальному закону? Для намётанного глаза точки на правом графике рисунка 3.1 не выглядят взятыми из нормального распределения — точка с координатами (576, 3.39) кажется слишком далёкой от остальных 14 точек. На самом деле, главная причина выбора двумерного нормального распределения — простота оценивания. Другое предположение не привело бы к такой простой оценке для $s_{\text{e}}(\widehat{\text{corr}})$.

Есть ещё одно серьёзное возражение против $\widehat{s}_{\text{normal}}$: требуется приложить серьёзные усилия для того, чтобы вывести формулу наподобие (7.1). Если выбрать чуть более сложную статистику, чем $\widehat{\text{corr}}$, или же менее стандартное распределение, то никакие математические трюки не приведут к простой формуле. Ввиду таких ограничений, до-компьютерная статистика в качестве объектов интереса рассматривала в основном небольшие классы распределений и ограниченный набор статистик. Компьютерные методы, такие как бутстреп, освобождают статистику от таких ограничений. Стандартные ошибки, равно как и другие статистические меры точности, получаются в результате процедуры автоматически, безотносительно к математической сложности.¹

¹В таком подходе есть не только плюсы. Теоретические формулы наподобие (7.1) могут помочь нам понимать ситуацию немного иначе, чем при получении численных результатов применения bootstrap. Хорошо иметь в виду, что такие методы, как bootstrap, освобождают статистику от необходимости смотреть на данные более глубоко, без страха усложнений в математике, но не менее.

Бутстреп методы оказываются очень полезными в сложных проблемах оценивания. В этой главе обсуждаются стандартные ошибки для двух таких задач: первая касается собственных значений и собственных векторов матрицы ковариаций, вторая — алгоритма *loess* приближения функций. Разъяснение данных задач требует знакомства с матричной терминологией, однако здесь это будет опущено; в любом случае эта теория не является необходимой для того, чтобы понять главную идею: что простой алгоритм бутстрепа позволяет находить стандартные ошибки для очень трудных случаев.

7.2 Пример 1: результаты тестов

В таблице 7.1 показаны данные результатов тестов из Mardia, Kent and Bibby (1979); $n = 88$ студентов сдавали пять тестов: по механике, векторному исчислению, алгебре, математическому анализу, и статистике.

На первых двух тестах не разрешалось использовать учебник, на остальных учебник был разрешён. Удобно представлять эти данные как матрица данных \mathbf{X} размерности 88×5 , где i -ая строка есть

$$\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}) \quad (7.2)$$

— пять результатов i -го студента, $i = 1, 2, \dots, 88$.

Вектор средних $\bar{\mathbf{x}} = \sum_{i=1}^{88} \mathbf{x}_i / 88$ есть вектор средних по столбцам:

$$\bar{\mathbf{x}} = (38.95, 50.59, 50.60, 46.68, 42.31). \quad (7.3)$$

Эмпирическая ковариационная матрица \mathbf{G} — это матрица 5×5 , где (j, k) -й элемент равен

$$G_{jk} = \frac{1}{88} \sum_{i=1}^{88} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad j, k = 1, 2, 3, 4, 5. \quad (7.4)$$

Заметим, что диагональные элементы G_{jj} это оценки дисперсии результатов теста j методом подстановки. Получим матрицу

$$\mathbf{G} = \begin{pmatrix} 302.3 & 125.8 & 100.4 & 105.1 & 116.1 \\ 125.8 & 170.9 & 84.2 & 93.6 & 97.9 \\ 100.4 & 84.2 & 111.6 & 110.8 & 120.5 \\ 105.1 & 93.6 & 110.8 & 217.9 & 153.8 \\ 116.1 & 97.9 & 120.5 & 153.8 & 294.4 \end{pmatrix}. \quad (7.5)$$

$$\begin{aligned} \hat{\lambda}_1 &= 679.2 & \hat{\mathbf{v}}_1 &= (.505, .368, .346, .451, .535) \\ \hat{\lambda}_2 &= 199.8 & \hat{\mathbf{v}}_2 &= (-.749, -.207, .076, .301, .548) \\ \hat{\lambda}_3 &= 102.6 & \hat{\mathbf{v}}_3 &= (-.300, .416, .145, .597, -.600) \\ \hat{\lambda}_4 &= 83.7 & \hat{\mathbf{v}}_4 &= (.296, -.783, -.003, .518, -.176) \\ \hat{\lambda}_5 &= 31.8 & \hat{\mathbf{v}}_5 &= (.079, .189, -.924, .286, .151). \end{aligned} \quad (7.6)$$

Какой интерес представляют собственные значения и векторы ковариационной матрицы? Они помогают описать структуру высокоразмерных данных (как

в случае с таблицей (7.1)) в которых описано большое число независимых величин ($n = 88$ студентов), но при этом имеются коррелированные измерения для каждого студента. Заметьте, что пять тестовых оценок высоко коррелированы. Студент, который хорошо сдал тест по механике, вероятно также хорошо сдал тест и по векторам и т.д. Очень простая модель для коррелированных оценок имеет вид

$$\mathbf{x}_i = Q_i \mathbf{v}, \quad i = 1, 2, \dots, 88. \quad (7.7)$$

Table 7.1. *The score data, from Mardia, Kent and Bibby (1979); n = 88 students each took five tests, in mechanics, vectors, algebra, analysis, and statistics; “c” and “o” indicate closed and open book, respectively.*

| # | mec (c) | vec (c) | alg (o) | ana (o) | sta (o) | # | mec (c) | vec (c) | alg (o) | ana (o) | sta (o) |
|----|------------|------------|------------|------------|------------|----|------------|------------|------------|------------|------------|
| 1 | 77 | 82 | 67 | 67 | 81 | 45 | 46 | 61 | 46 | 38 | 41 |
| 2 | 63 | 78 | 80 | 70 | 81 | 46 | 40 | 57 | 51 | 52 | 31 |
| 3 | 75 | 73 | 71 | 66 | 81 | 47 | 49 | 49 | 45 | 48 | 39 |
| 4 | 55 | 72 | 63 | 70 | 68 | 48 | 22 | 58 | 53 | 56 | 41 |
| 5 | 63 | 63 | 65 | 70 | 63 | 49 | 35 | 60 | 47 | 54 | 33 |
| 6 | 53 | 61 | 72 | 64 | 73 | 50 | 48 | 56 | 49 | 42 | 32 |
| 7 | 51 | 67 | 65 | 65 | 68 | 51 | 31 | 57 | 50 | 54 | 34 |
| 8 | 59 | 70 | 68 | 62 | 56 | 52 | 17 | 53 | 57 | 43 | 51 |
| 9 | 62 | 60 | 58 | 62 | 70 | 53 | 49 | 57 | 47 | 39 | 26 |
| 10 | 64 | 72 | 60 | 62 | 45 | 54 | 59 | 50 | 47 | 15 | 46 |
| 11 | 52 | 64 | 60 | 63 | 54 | 55 | 37 | 56 | 49 | 28 | 45 |
| 12 | 55 | 67 | 59 | 62 | 44 | 56 | 40 | 43 | 48 | 21 | 61 |
| 13 | 50 | 50 | 64 | 55 | 63 | 57 | 35 | 35 | 41 | 51 | 50 |
| 14 | 65 | 63 | 58 | 56 | 37 | 58 | 38 | 44 | 54 | 47 | 24 |
| 15 | 31 | 55 | 60 | 57 | 73 | 59 | 43 | 43 | 38 | 34 | 49 |
| 16 | 60 | 64 | 56 | 54 | 40 | 60 | 39 | 46 | 46 | 32 | 43 |
| 17 | 44 | 69 | 53 | 53 | 53 | 61 | 62 | 44 | 36 | 22 | 42 |
| 18 | 42 | 69 | 61 | 55 | 45 | 62 | 48 | 38 | 41 | 44 | 33 |
| 19 | 62 | 46 | 61 | 57 | 45 | 63 | 34 | 42 | 50 | 47 | 29 |
| 20 | 31 | 49 | 62 | 63 | 62 | 64 | 18 | 51 | 40 | 56 | 30 |
| 21 | 44 | 61 | 52 | 62 | 46 | 65 | 35 | 36 | 46 | 48 | 29 |
| 22 | 49 | 41 | 61 | 49 | 64 | 66 | 59 | 53 | 37 | 22 | 19 |
| 23 | 12 | 58 | 61 | 63 | 67 | 67 | 41 | 41 | 43 | 30 | 33 |
| 24 | 49 | 53 | 49 | 62 | 47 | 68 | 31 | 52 | 37 | 27 | 40 |
| 25 | 54 | 49 | 56 | 47 | 53 | 69 | 17 | 51 | 52 | 35 | 31 |
| 26 | 54 | 53 | 46 | 59 | 44 | 70 | 34 | 30 | 50 | 47 | 36 |
| 27 | 44 | 56 | 55 | 61 | 36 | 71 | 46 | 40 | 47 | 29 | 17 |
| 28 | 18 | 44 | 50 | 57 | 81 | 72 | 10 | 46 | 36 | 47 | 39 |
| 29 | 46 | 52 | 65 | 50 | 35 | 73 | 46 | 37 | 45 | 15 | 30 |
| 30 | 32 | 45 | 49 | 57 | 64 | 74 | 30 | 34 | 43 | 46 | 18 |
| 31 | 30 | 69 | 50 | 52 | 45 | 75 | 13 | 51 | 50 | 25 | 31 |
| 32 | 46 | 49 | 53 | 59 | 37 | 76 | 49 | 50 | 38 | 23 | 9 |
| 33 | 40 | 27 | 54 | 61 | 61 | 77 | 18 | 32 | 31 | 45 | 40 |
| 34 | 31 | 42 | 48 | 54 | 68 | 78 | 8 | 42 | 48 | 26 | 40 |
| 35 | 36 | 59 | 51 | 45 | 51 | 79 | 23 | 38 | 36 | 48 | 15 |
| 36 | 56 | 40 | 56 | 54 | 35 | 80 | 30 | 24 | 43 | 33 | 25 |
| 37 | 46 | 56 | 57 | 49 | 32 | 81 | 3 | 9 | 51 | 47 | 40 |
| 38 | 45 | 42 | 55 | 56 | 40 | 82 | 7 | 51 | 43 | 17 | 22 |
| 39 | 42 | 60 | 54 | 49 | 33 | 83 | 15 | 40 | 43 | 23 | 18 |
| 40 | 40 | 63 | 53 | 54 | 25 | 84 | 15 | 38 | 39 | 28 | 17 |
| 41 | 23 | 55 | 59 | 53 | 44 | 85 | 5 | 30 | 44 | 36 | 18 |
| 42 | 48 | 48 | 49 | 51 | 37 | 86 | 12 | 30 | 32 | 35 | 21 |
| 43 | 41 | 63 | 49 | 46 | 34 | 87 | 5 | 26 | 15 | 20 | 20 |
| 44 | 46 | 52 | 53 | 41 | 40 | 88 | 0 | 40 | 21 | 9 | 14 |

Q_i является числом, представляющим способности i -го студента, в то время как $\mathbf{v} = (v_1, v_2, v_3, v_4, v_5)$ есть фиксированный вектор из 5 чисел, определённый для всех студентов. Q_i можно рассматривать как общую оценку интеллектуальных способностей студента i (IQ). Изначально IQ были мотивированы именно

моделями чуть сложнее, чем (7.7).

Если бы модель (7.7) была верна, мы бы смогли это определить из собственных значений: только $\hat{\lambda}_1$ было бы положительным, остальные — $\hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4, \hat{\lambda}_5$ — равнялись бы нулю; также первый собственный вектор \hat{v}_1 был бы равен v . Пусть $\hat{\theta}$ есть частное наибольшего собственного значения и их суммы, то есть

$$\hat{\theta} = \hat{\lambda}_1 / \sum_{i=1}^5 \hat{\lambda}_i. \quad (7.8)$$

Модель (7.7) эквивалентна $\hat{\theta} = 1$. Конечно, мы не можем ожидать, что для таких зашумлённых данных, как оценки, модель (7.7) окажется точной, даже если модель фундаментально верна.

Рисунок 7.1 даёт стилизованную иллюстрацию этого замечания. Мы взяли только две из оценок и отобразили слева ситуацию, если бы одно число Q_i идеально отражало обе оценки. Оценки лежат на одной прямой; Q_i можно считать как расстояние вдоль прямой до каждой точки от начала координат. Рисунок справа показывает более реалистичную ситуацию. Точки не лежат вдоль прямой, но расположены близко к ней. Прямая на графике коллинеарна направлению, заданному первым собственным вектором ковариационной матрицы. Эта прямая иногда называется *прямой первой главной компоненты* и имеет следующее свойство: она минимизирует сумму квадратов расстояний между точками и прямой (в отличие от метода наименьших квадратов, который заключается в минимизации суммы квадратов вертикальных расстояний до прямой). Эти расстояния показаны на рисунке справа в виде небольших отрезков. Сложно создать такой график для всех данных об оценках: прямая главной компоненты была бы прямой в пятимерном пространстве, лежащей ближе всего к данным. Если рассмотреть проекцию каждой из точек на прямую, прямая первой главной компоненты также будет минимизировать выборочную дисперсию всех спроектированных точек.

Для данных с оценками получим

$$\hat{\theta} = \frac{679.2}{679.2 + 199.8 + \dots + 31.8} = 0.619. \quad (7.9)$$

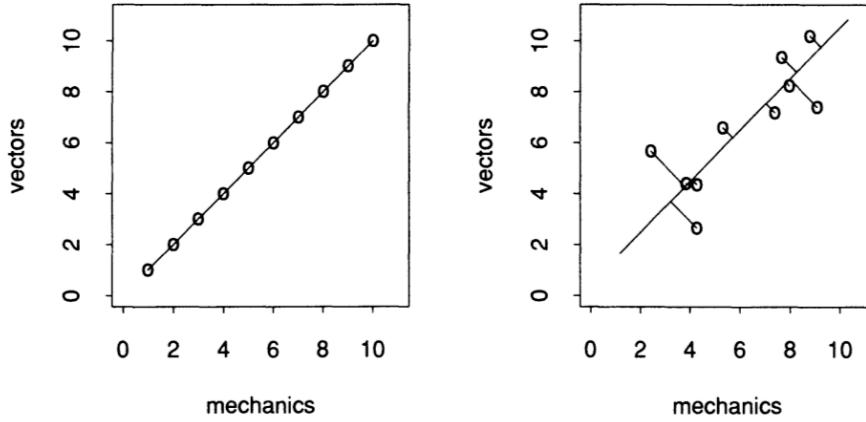


Figure 7.1. *Hypothetical plot of mechanics and vector scores. On the left, the pairs line exactly on a straight line (that is, have correlation 1) and hence a single measure captures the two scores. On the right, the scores have correlation less than one. The principal component line minimizes the sum of orthogonal distances to the line and has direction given by the largest eigenvector of the covariance matrix.*

Во многих ситуациях такое большое значение $\hat{\theta}$ можно считать достаточно любопытным, что показывает высокую степень предсказательной силы модели (7.7). Значение $\hat{\theta}$ измеряет процент дисперсии, объясняемой первой главной компонентой. Чем ближе точки лежат к прямой первой главной компоненты, тем выше значение $\hat{\theta}$. Насколько точна оценка $\hat{\theta}$? Именно для ответа на такие вопросы бутстреп и был создан. Математическая сложность вычисления $\hat{\theta}$ не важна до тех пор пока мы можем подсчитать $\hat{\theta}^*$ для любых бутстреп данных. В этом случае бутстреп выборка представлена \mathbf{X}^* — матрицей 88×5 . Строки \mathbf{x}_i^* матрицы \mathbf{X}^* есть случайная выборка размера 88 из столбцов оригинальной матрицы \mathbf{X}

$$\mathbf{x}_1^* = \mathbf{x}_{i_1}^*, \mathbf{x}_2^* = \mathbf{x}_{i_2}^*, \dots, \mathbf{x}_{88}^* = \mathbf{x}_{i_{88}}^*, \quad (7.10)$$

как в (6.4). Некоторые строки матрицы \mathbf{X} не появляются ни разу, некоторые один раз, некоторые дважды, и т.д., в итоге имеется 88 строк.

Сгенерировав матрицу \mathbf{X}^* , мы считаем её ковариационную матрицу \mathbf{G}^* по аналогии с (7.4)

$$G_{jk}^* = \frac{1}{88} \sum_{i=1}^{88} (x_{ij}^* - \bar{x}_j^*)(x_{ik}^* - \bar{x}_k^*) \quad j, k = 1, 2, 3, 4, 5. \quad (7.11)$$

Затем вычисляем собственные значения матрицы \mathbf{G}^* — $\hat{\lambda}_1^*, \hat{\lambda}_2^*, \dots, \hat{\lambda}_5^*$ — и в конце

$$\hat{\theta}^* = \hat{\lambda}_1^*/\sum_{i=1}^5 \hat{\lambda}_i^*, \quad (7.12)$$

На рисунке 7.2 изображена гистограмма $B = 200$ репликаций бутстреп оценок $\hat{\theta}^*$.

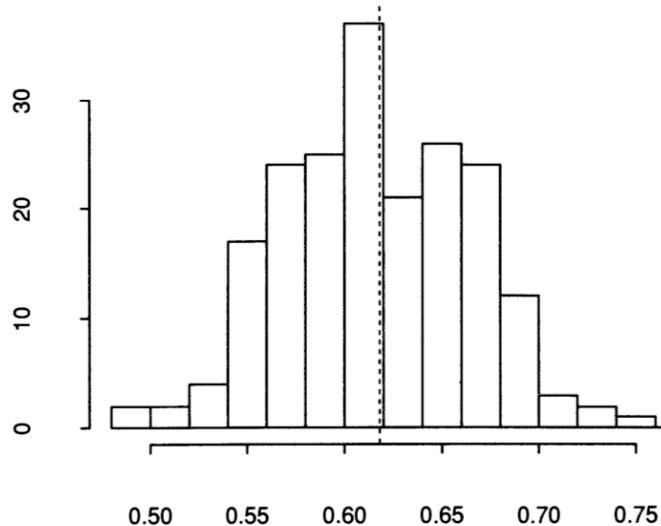


Figure 7.2. 200 bootstrap replications of the statistic $\hat{\theta} = \hat{\lambda}_1 / \sum_1^5 \hat{\lambda}_i$. The bootstrap standard error is .047. The dashed line indicates the observed value $\hat{\theta} = .619$.

Они дают следующую оценку стандартной ошибки $\hat{\theta}^*$: $\hat{se} = 0.047$. Среднее для всех 200 репликаций составило 0.625, что лишь немногим больше, чем $\hat{\theta} = 0.619$. Это означает, что $\hat{\theta}$ близка к несмешённой. Гистограмма выглядит адекватно, но $B = 200$ всё же недостаточно для того, чтобы ясно увидеть форму распределения. Некоторые квантили эмпирического распределения $\hat{\theta}^*$ показаны в таблице 7.2.

Table 7.2. Quantiles of the bootstrap distribution of $\hat{\theta}^*$ defined in (7.12)

| α | .05 | .10 | .16 | .50 | .84 | .90 | .95 |
|----------|------|------|------|------|------|------|------|
| quantile | .545 | .557 | .576 | .629 | .670 | .678 | .693 |

Стандартный доверительный интервал для настоящего значения θ (значение $\hat{\theta}$, если устремить $n \rightarrow \infty$)

$$\theta \in \hat{\theta} \pm z^{(1-\alpha)} \cdot \hat{se} \quad (\text{с вероятностью } 1 - 2\alpha) \quad (7.13)$$

где $z^{(1-\alpha)}$ есть $100(1 - \alpha)$ перцентиль стандартного нормального распределения; $z^{(.975)} = 1.960$, $z^{(.95)} = 1.645$, $z^{(.84)} = 1.000$, и т.д. Вычисление интервала основано на применении асимптотической теории, которая распространяет (5.6) на генеральные статистики $\hat{\theta}$. В нашем случае

$$\theta \in 0.619 \pm 0.047 = [0.572, 0.666] \quad \text{с вероятностью 0.683}$$

$$\theta \in 0.619 \pm 0.077 = [0.542, 0.696] \quad \text{с вероятностью 0.900.}$$

В 12–14 главах обсуждаются улучшенные бутстреп доверительные интервалы, менее зависимые от асимптотической теории нормального распределения.

Случайный вектор $\hat{\mathbf{v}}_1$, относящийся к первому собственному значению, называется первой главной компонентой \mathbf{G} . Предположим, что мы хотим выразить результаты студента одним числом, а не пятью (например, для некоторого общего оценивания). Можно показать, что наилучшая линейная комбинация оценок есть

$$y_i = \sum_{k=1}^5 \hat{v}_{1k} x_{ik}, \quad (7.14)$$

то есть линейная комбинация, использующая компоненты $\hat{\mathbf{v}}_1$ как веса. Эта линейная комбинация — «наилучшая» в смысле того, что среди всех возможных \mathbf{v} она отражает наибольшую вариативность в данных по пяти оценкам. Если же мы хотим описать успеваемость студента двумя числами, например (y_i, z_i) , вторая линейная комбинация должна выглядеть так

$$y_i = \sum_{k=1}^5 \hat{v}_{2k} x_{ik}, \quad (7.15)$$

где веса взяты из второй главной компоненты $\hat{\mathbf{v}}_2$, второго собственного значения матрицы \mathbf{G} .

Веса, заданные главными компонентами, часто дают понимание структуры многомерного набора данных. Для данных с оценками интерпретация будет следующей: первая главная компонента

$$\hat{\mathbf{v}}_1 = (0.51, 0.37, 0.35, 0.45, 0.54)$$

накладывает положительные веса примерно одинакового размера на каждый из тестов, то есть y_i условно эквивалентно взятию суммарной (или средней) оценки i -го студента. Вторая главная компонента

$$\hat{\mathbf{v}}_2 = (-0.75, -0.2, 0.08, 0.30, 0.55)$$

даёт отрицательные веса двум тестам без использования конспекта и положительные на три теста с использованием конспекта, так что z_i есть показатель *разницы* оценок между тестами с открытым и закрытым конспектом для i -го студента. (Студент с высокой оценкой z гораздо лучше справился с тестами с открытым конспектом, чем с закрытым.)

Векторы главных компонент $\hat{\mathbf{v}}_1$ и $\hat{\mathbf{v}}_2$ есть суммарные статистики, как и $\hat{\theta}$, несмотря на то, что у каждой из них есть несколько компонент. Мы можем применить бутстреп анализ для того, чтобы узнать, насколько они устойчивы. Тезже 200 бутстреп выборок, с помощью которых мы получили $\hat{\theta}^*$, дают бутстреп репликации $\hat{\mathbf{v}}_1^*$ и $\hat{\mathbf{v}}_2^*$. Они могут быть посчитаны как первые два собственных вектора \mathbf{G}^* , (7.11).

В таблице 7.3 показаны $\hat{s}\epsilon_{200}$ для каждой из компонент векторов $\hat{\mathbf{v}}_1$ и $\hat{\mathbf{v}}_2$. Первое, что можно заметить — это более высокую точность $\hat{\mathbf{v}}_1$; бутстреп стандартная ошибка компонент $\hat{\mathbf{v}}_1$ составляет менее половины ошибки $\hat{\mathbf{v}}_2$. В таблице 7.3 также указаны основанные на персентилях робастные бутстреп стандартные ошибки $\tilde{s}\epsilon_{200,\alpha}$, посчитанные для $\alpha = 0.84, 0.9, 0.95$. Для компонент $\hat{\mathbf{v}}_1$ $\tilde{s}\epsilon_{200,\alpha}$ примерно равно $\hat{s}\epsilon_{200}$.

Table 7.3. Bootstrap standard errors for the components of the first and second principal components, \hat{v}_1 and \hat{v}_2 ; $\hat{s}e_{200}$ is the usual bootstrap standard error estimate based on $B = 200$ bootstrap replications; $\tilde{s}e_{200,.84}$ is the standard error estimate $\tilde{s}e_{B,\alpha}$ of Problem 6.6, with $B = 200$, $\alpha = .84$; likewise $\tilde{s}e_{200,.90}$ and $\tilde{s}e_{200,.95}$. The values of $\hat{s}e_{200}$ for \hat{v}_{21} and \hat{v}_{25} are greatly inflated by a few outlying bootstrap replications, see Figures 7.3 and 7.4.

| | \hat{v}_{11} | \hat{v}_{12} | \hat{v}_{13} | \hat{v}_{14} | \hat{v}_{15} | \hat{v}_{21} | \hat{v}_{22} | \hat{v}_{23} | \hat{v}_{24} | \hat{v}_{25} |
|------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $\hat{s}e_{200}$ | .057 | .045 | .029 | .041 | .049 | .189 | .138 | .066 | .129 | .150 |
| $\tilde{s}e_{200,.84}$ | .055 | .041 | .028 | .041 | .047 | .078 | .122 | .064 | .110 | .114 |
| $\tilde{s}e_{200,.90}$ | .055 | .041 | .027 | .042 | .046 | .084 | .129 | .067 | .111 | .125 |
| $\tilde{s}e_{200,.95}$ | .054 | .048 | .029 | .040 | .047 | .080 | .130 | .066 | .114 | .120 |

Это не так для \hat{v}_2 , в особенности для первой и пятой координаты. На рисунке 7.3 можно увидеть, в чём проблема. На рисунке показаны эмпирические распределения для 200 бутстреп репликаций \hat{v}_{ik}^* , в отдельности для каждого из $i = 1, 2$, $k = 1, 2, \dots, 5$. Эмпирические распределения отражены ящиками с усами. Отрезок в центре ящика — медиана распределения; нижняя и верхняя сторона ящика есть соответственно 25-я и 75-я персентиль распределения; усы покрывают всё распределение за исключением некоторых выбросов (определённых по некоторому критерию), которые отмечены звёздочкой.

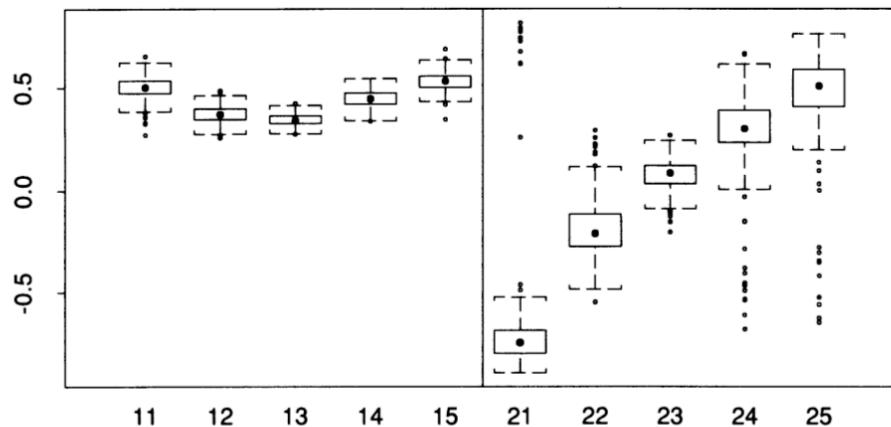


Figure 7.3. 200 bootstrap replications of the first two principal component vectors \hat{v}_1 (left panel) and \hat{v}_2 (right panel); for each component of the two vectors, the boxplot indicates the empirical distribution of the 200 bootstrap replications \hat{v}_{ik}^* . We see that \hat{v}_2 is less accurate than \hat{v}_1 , having greater bootstrap variability for each component. A few of the bootstrap samples gave completely different results than the others for \hat{v}_2 .

Можно увидеть, что большие значения $\hat{s}e_{200}$ для \hat{v}_{21} и \hat{v}_{25} вызваны несколькими выделяющимися значениями \hat{v}_{ik}^* . Приближённый доверительный интервал $\theta \in \hat{\theta} \pm z^{1-\alpha}\hat{s}e$ будет более точным, если выбрать $\tilde{s}e_{200,\alpha}$ в качестве оценки $\hat{s}e$, как минимум для умеренных значений α таких, как 0.843. Гистограмма значений v_{21}^* имеет форму нормального распределения со средним в точке -0.74 и стандартным отклонением 0.075 , с небольшим числом точек далеко от гистограммы. Это показатель того, что с малой вероятностью, порядка 1% или 2%, что \hat{v}_{21} оказывается совершенно неточной оценкой настоящего значения v_{21} . Ес-

ли же данное событие не произошло, \hat{v}_{21} вероятно находится в пределе одного или двух $\tilde{s}e_{200}$ от v_{21} .

На рисунке 7.4 показаны графики бутстреп репликаций $\hat{v}_1^*(b)$ и $\hat{v}_2^*(b)$, $b = 1, 2, \dots, 200$, соединяющие компоненты каждого вектора прямыми. Это более наглядный (хоть и менее точный) показатель вариативности \hat{v}_2 , чем предложенные ранее в таблице 7.3 и на рисунке 7.3. Три конкретных репликации, отмеченные числами 1, 2, и 3, являются выбросами на нескольких компонентах.

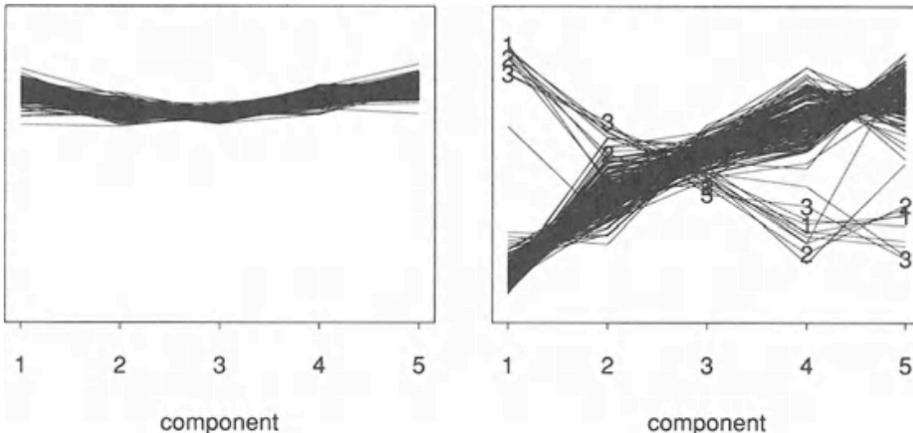


Figure 7.4. Graphs of the 200 bootstrap replications of \hat{v}_1 (left panel) and \hat{v}_2 (right panel). The numbers 1, 2, 3 in the right panel follow three of the replications $\hat{v}_2^(b)$ that gave the most discrepant values for the first component. We see that these replications were also discrepant for other components, particularly component 5.*

Читатель, которому знаком метод главных компонент, может теперь увидеть, что сложности со вторым собственным вектором объясняются проблемой единственности собственных векторов. Технически, определение собственного вектора \mathbf{v} также верно и для обратного ему вектора $-\mathbf{v}$. Алгоритм, который считает собственные числа и собственные значения, может приводить решения с разными знаками у $\hat{v}_1, \hat{v}_2, \dots$. Репликации 1 и 2 привели к матрицам \mathbf{X}^* , для которых знак \hat{v}_2^* получился обратным. Такая неопределенность обычно не важна при определении статистических особенностей оценок (хотя хорошо замечать такую неопределенность на основании результата применения бутстрепа). Если перестать учитывать 1 и 2, как происходит при оценке $\tilde{s}e_{200,\alpha}$, мы видим, что \hat{v}_2 всё равно менее точна, чем \hat{v}_1 .

7.3 Пример 2: построение кривой по данным

В этом примере мы будем оценивать функцию регрессии двумя способами, сначала с помощью стандартного метода наименьших квадратов, а затем с помощью современного метода построения кривой по данным, который называется *loess*. Мы начнём с краткого повторения теории регрессии. В главе 9 снова рассматривается задача регрессии и дан альтернативный бутстреп метод для оценки стандартных ошибок регрессии. На рисунке 7.5 показан типичный набор данных, для которого используются регрессионные методы: мужчин приняли участие в эксперименте, чтобы определить, уменьшает ли лекарство на основе

холостирамина уровень холестерина в крови. Мужчины должны были принимать по 6 пакетиков холостирамина в день, однако многие из них принимали гораздо меньше.

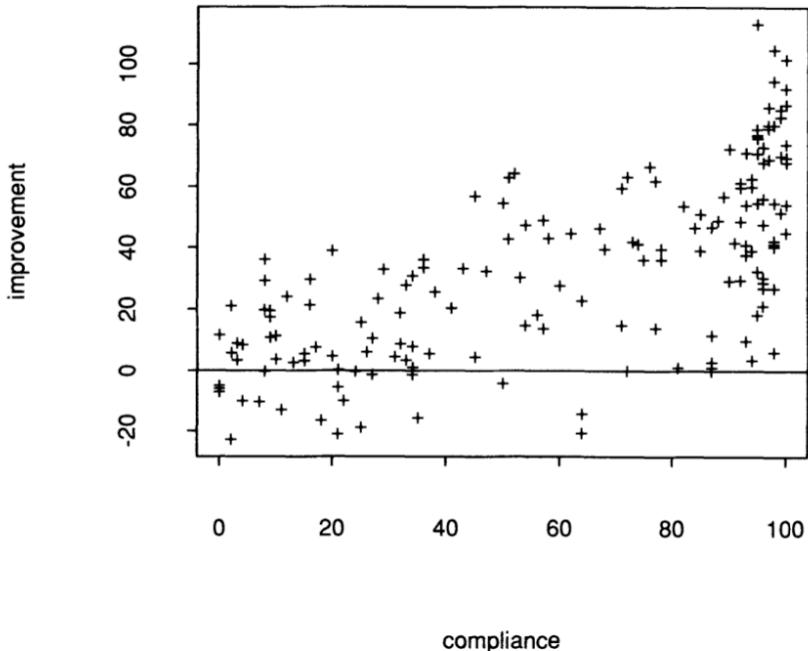


Figure 7.5. The cholestyramine data. 164 men were supposed to take 6 packets per day of the cholesterol-lowering drug cholestyramine; horizontal axis measures Compliance, in percentage of assigned dose actually taken; vertical axis measures Improvement, in terms of blood cholesterol decrease over the course of the experiment. We see that better compliers tended to have greater improvement.

Горизонтальная ось, которую мы назовём « z », измеряет *Соответствие*, то есть процент приёма от назначеннной дозы,

$$z_i = \text{процент соответствия для мужчины } i, i = 1, 2, \dots, 164.$$

Соответствие измерялось подсчётом количества пакетиков, которые вернули индивиды. Те, кто приняли все пакетики, находятся в правом краю графика; те, кто не принимал ничего — в левом. Горизонтальная ось, отмеченная « y », есть показатель *Улучшения*, уменьшение уровня холестерина в кровяной плазме за время исследования,

$$y_i = \text{уменьшение уровня холестерина в крови для индивида } i, i = 1, 2, \dots, 164.$$

Полный набор данных есть в таблице 7.4.

Table 7.4. *The cholestyramine data. 164 men were supposed to take 6 packets per day of the cholesterol-lowering drug cholestyramine. Compliance "z" is the percentage of the intended dose actually taken. Improvement "y" is the decrease in total plasma cholesterol from the beginning till the end of treatment.*

| <i>z</i> | <i>y</i> | <i>z</i> | <i>y</i> | <i>z</i> | <i>y</i> | <i>z</i> | <i>y</i> |
|----------|----------|----------|----------|----------|----------|----------|----------|
| 0 | -5.25 | 27 | -1.50 | 71 | 59.50 | 95 | 32.50 |
| 0 | -7.25 | 28 | 23.50 | 71 | 14.75 | 95 | 70.75 |
| 0 | -6.25 | 29 | 33.00 | 72 | 63.00 | 95 | 18.25 |
| 0 | 11.50 | 31 | 4.25 | 72 | 0.00 | 95 | 76.00 |
| 2 | 21.00 | 32 | 18.75 | 73 | 42.00 | 95 | 75.75 |
| 2 | -23.00 | 32 | 8.50 | 74 | 41.25 | 95 | 78.75 |
| 2 | 5.75 | 33 | 3.25 | 75 | 36.25 | 95 | 54.75 |
| 3 | 3.25 | 33 | 27.75 | 76 | 66.50 | 95 | 77.00 |
| 3 | 8.75 | 34 | 30.75 | 77 | 61.75 | 96 | 68.00 |
| 4 | 8.25 | 34 | -1.50 | 77 | 14.00 | 96 | 73.00 |
| 4 | -10.25 | 34 | 1.00 | 78 | 36.00 | 96 | 28.75 |
| 7 | -10.50 | 34 | 7.75 | 78 | 39.50 | 96 | 26.75 |
| 8 | 19.75 | 35 | -15.75 | 81 | 1.00 | 96 | 56.00 |
| 8 | -0.50 | 36 | 33.50 | 82 | 53.50 | 96 | 47.50 |
| 8 | 29.25 | 36 | 36.25 | 84 | 46.50 | 96 | 30.25 |
| 8 | 36.25 | 37 | 5.50 | 85 | 51.00 | 96 | 21.00 |
| 9 | 10.75 | 38 | 25.50 | 85 | 39.00 | 97 | 79.00 |
| 9 | 19.50 | 41 | 20.25 | 87 | -0.25 | 97 | 69.00 |
| 9 | 17.25 | 43 | 33.25 | 87 | 1.00 | 97 | 80.00 |
| 10 | 3.50 | 45 | 56.75 | 87 | 46.75 | 97 | 86.00 |
| 10 | 11.25 | 45 | 4.25 | 87 | 11.50 | 98 | 54.75 |
| 11 | -13.00 | 47 | 32.50 | 87 | 2.75 | 98 | 26.75 |
| 12 | 24.00 | 50 | 54.50 | 88 | 48.75 | 98 | 80.00 |
| 13 | 2.50 | 50 | -4.25 | 89 | 56.75 | 98 | 42.25 |
| 15 | 3.00 | 51 | 42.75 | 90 | 29.25 | 98 | 6.00 |
| 15 | 5.50 | 51 | 62.75 | 90 | 72.50 | 98 | 104.75 |
| 16 | 21.25 | 52 | 64.25 | 91 | 41.75 | 98 | 94.25 |
| 16 | 29.75 | 53 | 30.25 | 92 | 48.50 | 98 | 41.25 |
| 17 | 7.50 | 54 | 14.75 | 92 | 61.25 | 98 | 40.25 |
| 18 | -16.50 | 54 | 47.25 | 92 | 29.50 | 99 | 51.50 |
| 20 | 4.50 | 56 | 18.00 | 92 | 59.75 | 99 | 82.75 |
| 20 | 39.00 | 57 | 13.75 | 93 | 71.00 | 99 | 85.00 |
| 21 | -5.75 | 57 | 48.75 | 93 | 37.75 | 99 | 70.00 |
| 21 | -21.00 | 58 | 43.00 | 93 | 41.00 | 100 | 92.00 |
| 21 | 0.25 | 60 | 27.75 | 93 | 9.75 | 100 | 73.75 |
| 22 | -10.25 | 62 | 44.50 | 93 | 53.75 | 100 | 54.00 |
| 24 | -0.50 | 64 | 22.50 | 94 | 62.50 | 100 | 69.50 |
| 25 | -19.00 | 64 | -14.50 | 94 | 39.00 | 100 | 101.50 |
| 25 | 15.75 | 64 | -20.75 | 94 | 3.25 | 100 | 68.00 |
| 26 | 6.00 | 67 | 46.25 | 94 | 60.00 | 100 | 44.75 |
| 27 | 10.50 | 68 | 39.50 | 95 | 113.25 | 100 | 86.75 |

На рисунке видно, что мужчины, которые принимали больше хлоростира-

мина, в целом улучшили свои показатели холестерина, что и ожидалось. То, что мы видим на рисунке 7.5, или скорее то, что мы хотели бы видеть, есть увеличение среднего ответа y в то время как z увеличивается от 0 до 100%. На рисунке 7.6 показаны данные вместе с двумя графиками,

$$\hat{r}_{\text{quad}}(z) \text{ and } \hat{r}_{\text{loess}}(z). \quad (7.16)$$

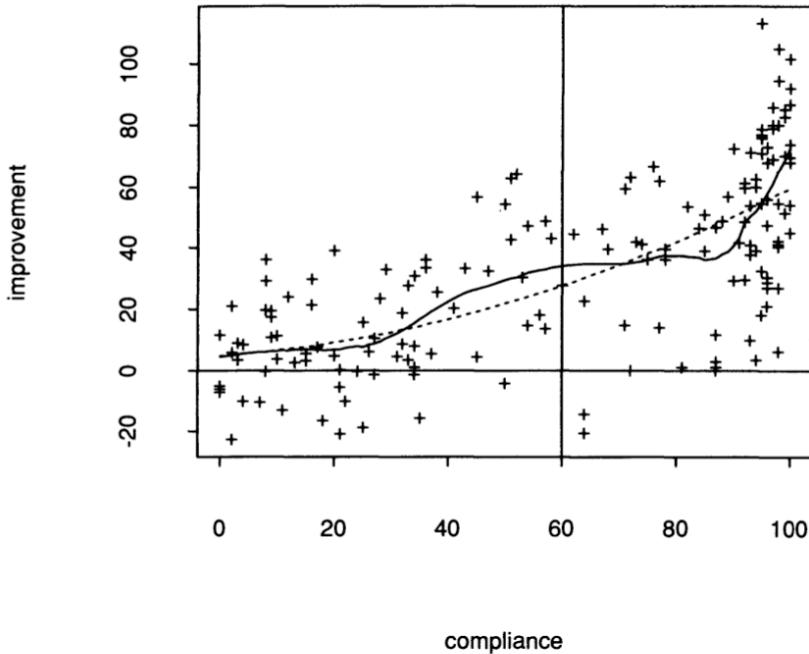


Figure 7.6. Estimated regression curves of $y = \text{Improvement}$ on $z = \text{Compliance}$. The dashed curve is $\hat{r}_{\text{quad}}(z)$, the ordinary least-squares quadratic regression of y on z ; the solid curve is $\hat{r}_{\text{loess}}(z)$, a computer-based local linear regression. We are particularly interested in estimating the true regression $r(z)$ at $z = 60\%$, the average Compliance, and at $z = 100\%$, full Compliance.

Каждый из них есть оценка кривой регрессии. Сейчас будет краткое повторение построения и оценки регрессионных кривых. По определению регрессией ответа y на независимую переменную z называется условное математическое ожидание y при некотором z ,

$$r(z) = E(y|z). \quad (7.17)$$

Предположим, что нам была доступна вся популяция \mathcal{U} мужчин, подходящих для эксперимента, и мы получили набор $\mathcal{X} = (X_1, X_2, \dots, X_N)$ оценок Соответствие-улучшение $X_j = (Z_j, Y_j)$, $j = 1, 2, \dots, N$. Далее для каждого значения z , например $z = 0\%, 1\%, 2\%, \dots, 100\%$, регрессия была бы условным математическим ожиданием (7.17),

$$r(z) = \frac{\text{сумма значений } Y_j \text{ для мужчин в } \mathcal{X} \text{ с } Z_j = z}{\text{число мужчин в } \mathcal{X} \text{ с } Z_j = z}. \quad (7.18)$$

Другими словами, $r(z)$ есть математическое ожидание Y для субпопуляции мужчин, у которых $Z = z$.

Разумеется у нас нет целой популяции \mathcal{X} . У нас имеется выборка $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{164})$, где $\mathbf{x}_i = (z_i, y_i)$, как показано на рисунке 7.5 и в таблице 7.4. Как мы можем оценить $r(z)$? Очевидная оценка методом подстановки есть

$$\hat{r}(z) = \frac{\text{сумма значений } y_j \text{ для мужчин в } \mathbf{x} \text{ с } z_j = z}{\text{число мужчин в } \mathbf{x} \text{ с } z_j = z}. \quad (7.19)$$

Можно представить себе рисование вертикальных полос шириной в 1% на рисунке 7.5 и усреднение значений на каждой полосе для получения $\hat{r}(z)$. Результаты можно увидеть на рисунке 7.7.

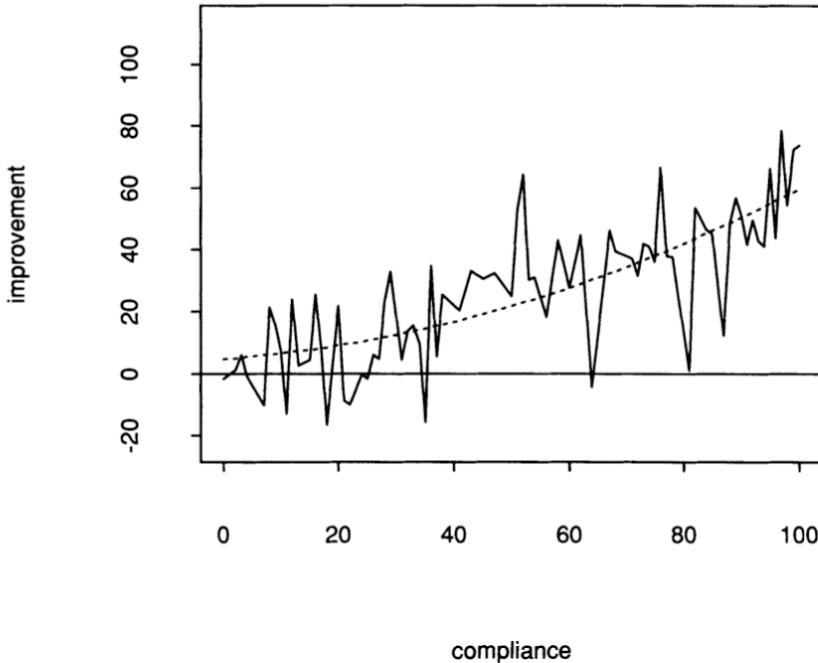


Figure 7.7. Solid curve is plug-in estimate $\hat{r}(z)$ for the regression of improvement on compliance; averages of y_i for strips of width 1% on the z axis, as in (7.19). Some strips z are not represented because none of the 164 men had $z_i = z$. The function $\hat{r}(z)$ is much rougher than we expect the population regression curve $r(z)$ to be. The dashed curve is $\hat{r}_{\text{quad}}(z)$.

Впервые нашёлся пример, для которого метод подстановки работает не очень хорошо. Оценка регрессии $\hat{r}(z)$ гораздо грубее, чем мы хотели бы для оценки популяционной регрессии $r(z)$. Проблема в том, что внутри каждой полосы шириной в 1% точек для адекватной оценки $r(z)$ недостаточно. Для некоторых полос шириной в 5% точек внутри нет вообще. Мы можем увеличить ширину промежутка, скажем, до 10% вместо 1%, но это оставит нас с небольшим числом точек для построения и, вероятно, проблема неустойчивости всё равно останется. На самом деле, имеется более элегантное и эффективное решение, которое основано на методе наименьших квадратов.

Использование метода начинается с предположения, что популяционная регрессионная функция, какая бы она не была, принадлежит семейству \mathcal{R} гладких функций, индексированных вектором параметров $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$. Для рассматриваемого примера мы ограничимся семейством квадратичных функ-

ций от z , скажем, $\mathcal{R}_{\text{quad}}$,

$$\mathcal{R}_{\text{quad}}: \quad r_{\beta}(z) = \beta_0 + \beta_1 z + \beta_2 z^2, \quad (7.20)$$

поэтому $\beta = (\beta_0, \beta_1, \beta_2)^T$. Далее мы обсудим выбор именно квадратичного семейства $\mathcal{R}_{\text{quad}}$, но на сейчас примем это как данное.

Читатель может представить себе выбор некоторого пробного значения β , к примеру, $\beta = (0, 0.75, 0.005)^T$, и построение $r_{\beta}(z)$ на рисунке 7.5. Мы хотели бы, чтобы кривая $r_{\beta}(z)$ проходила близко к нашим данным (z_i, y_i) в некотором общем смысле. Наиболее удобно для вычислений измерять близость кривой к данным с помощью суммы квадратов остатков (*Residual Squared Error*),

$$\text{RSE}(\beta) = \sum_{i=1}^n [y_i - r_{\beta}(z_i)]^2. \quad (7.21)$$

Сумма квадратов остатков получается опусканием вертикальных отрезков от каждой точки (z_i, y_i) к кривой $r_{\beta}(z_i)$, а затем суммированием квадратов их длин.

Метод наименьших квадратов, созданные Гауссом и Лежандром в начале 19 века, выбирает среди кривых в \mathcal{R} те, которые минимизируют RSE. Наилучший из них объявляется $r_{\hat{\beta}}(z)$, где $\hat{\beta}$ минимизирует $\text{RSE}(\beta)$,

$$\text{RSE}(\hat{\beta}) = \min_{\beta} \text{RSE}(\beta). \quad (7.22)$$

Кривая $\hat{r}_{\text{quad}}(z)$ на рисунке 7.6 есть $r_{\hat{\beta}}(z) = \hat{\beta}_0 + \hat{\beta}_1 z + \hat{\beta}_2 z^2$, наилучшая квадратичная функция для наших данных.

Лежандр и Гаусс обнаружили замечательную явную формулу для решения $\hat{\beta}$ задачи наименьших квадратов. Пусть \mathbf{C} есть матрица 164×3 , i -я строка которой есть

$$\mathbf{c}_i = (1, z_i, z_i^2), \quad (7.23)$$

и пусть \mathbf{y} есть вектор из 164 значений y_i . Тогда, используя стандартную матричную нотацию, имеем

$$\hat{\beta} = (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{y}. \quad (7.24)$$

Более подробно мы рассмотрим эту формулу в главе 9. Для наших целей, связанных с применением бутстрепа, нам достаточно лишь знать про то, что набор данных из n пар $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ приводит к получению квадратичной кривой $r_{\hat{\beta}}(z)$ через отображение $\mathbf{x} \rightarrow r_{\hat{\beta}}(z)$, которое описывается (7.23), (7.24) и (7.20).

Можно рассматривать $r_{\hat{\beta}}(z)$ как слаженную версию оценки по методу подстановки $\hat{r}(z)$. Предположим, что мы бы рассмотрели более широкий класс гладких функций \mathcal{R} , к примеру, класс кубических функций $\mathcal{R}_{\text{cubic}}$. В таком случае решение по методу наименьших квадратов $r_{\hat{\beta}}(z)$ стало бы ближе к данным, однако оказалось бы более «бугристым», чем квадратичное решение по методу наименьших квадратов. Если бы мы начали рассматривать полиномы всё большей степени, $r_{\hat{\beta}}$ всё больше бы походил на оценку по методу подстановки $\hat{r}(z)$. Выбор семейства квадратичных функций основан на нашем представлении о том, насколько гладкой должна быть оригинальная функция регрессии $r(z)$. Смотря на рисунок 7.7, мы явно видим, что $\hat{r}_{\text{quad}}(z)$ гораздо более гладкая, чем $\hat{r}(z)$, однако в целом соответствует $\hat{r}(z)$ как функция от z .

Легко поверить, что настоящая функция регрессии $r(z)$ есть гладкая функция от z . Сложнее поверить в то, что $r(z)$ является квадратичной от z для всех значений z . Сглаживающая функция *loess* является компромиссом между *глобальными* предположениями о форме и чисто *локальными* усреднением $\hat{r}(z)$.

Для использования *loess* нужно указать число α , которое равно части n точек, используемых при построении кривой в каждой из точек. Кривая $\hat{r}_{\text{loess}}(z)$ на рисунке 7.6 построена при выборе $\alpha = 0.3$. Для каждого из значений z значение $\hat{r}_{\text{loess}}(z)$ получается следующим образом:

1. n точек $\mathbf{x}_i = (z_i, y_i)$ упорядочиваются согласно $|z_i - z|$, а ближайшие $\alpha \cdot n$ точек с наименьшим $|z_i - z|$, запоминаются. Назовём эти точки $\mathcal{N}(z)$.²
2. Взвешенная линейная регрессия (с минимизацией наименьших квадратов)

$$\hat{r}_z(Z) = \hat{\beta}_{z,0} + \hat{\beta}_{z,1}Z \quad (7.25)$$

производится для $\alpha \cdot n$ точек в $\mathcal{N}(z)$. [То есть коэффициенты $\hat{\beta}_{z,0}$ и $\hat{\beta}_{z,1}$ выбираются как минимизирующие $\sum_{\mathbf{x}_j \in \mathcal{N}(z)} w_{z,j} [y_j - (\beta_0 + \beta_1 z_j)]^2$, где веса $w_{z,j}$ есть положительные числа, зависящие от $|z_j - z|$. Взяв

$$u_j = \frac{|z_j - z|}{\max_{\mathcal{N}(z)} |z_k - z|}, \quad (7.26)$$

веса w_j выбираются равными $(1 - u_j^3)^3$.]

3. В итоге, $\hat{r}_{\text{loess}}(z)$ назначается равным числу $\hat{r}_z(Z)$ в точке $Z = z$,

$$\hat{r}_{\text{loess}}(z) = \hat{r}_z(Z = z). \quad (7.27)$$

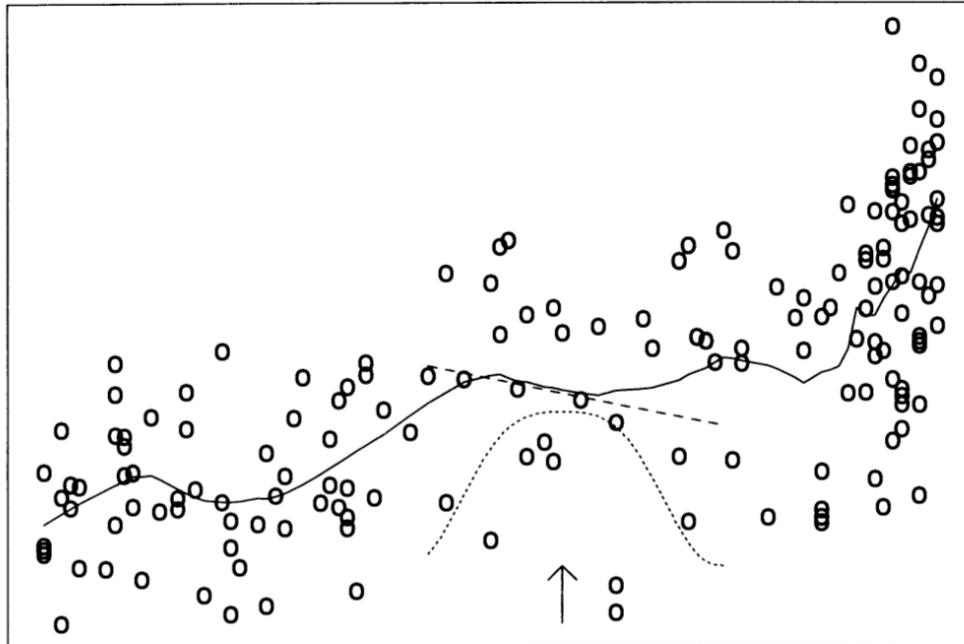


Figure 7.8. How the Loess smoother works. The shaded region indicates the window of values around the target value (arrow). A weighted linear regression (broken line) is computed, using weights given by the “tri-cube” function (dotted curve). Repeating this process for all target values gives the solid curve.

²при выборе $\alpha = 0.3, n = 164$, алгоритм выбирает в $\mathcal{N}(z)$ 49 точек

Компоненты loess сглаживания показаны на рисунке 7.8. В таблице 7.5 показано сравнение $\hat{r}_{\text{quad}}(z)$ и $\hat{r}_{\text{loess}}(z)$ в двух значениях, представляющих наибольший интерес, $z = 60\%$ и $z = 100\%$. Стандартные ошибки по бутстрепу даны для каждого из значений. Они были получены из $B = 50$ бутстреп репликаций алгоритма, показанного на рисунке 6.1.

Table 7.5. *Values of $\hat{r}_{\text{quad}}(z)$ and $\hat{r}_{\text{loess}}(z)$ at $z = 60\%$ and $z = 100\%$; also bootstrap standard errors based on $B = 50$ bootstrap replications.*

| | $\hat{r}_{\text{quad}}(60)$ | $\hat{r}_{\text{loess}}(60)$ | $\hat{r}_{\text{quad}}(100)$ | $\hat{r}_{\text{loess}}(100)$ |
|------------------------------|-----------------------------|------------------------------|------------------------------|-------------------------------|
| value: | 27.72 | 34.03 | 59.67 | 72.78 |
| $\widehat{\text{se}}_{50}$: | 3.03 | 4.41 | 3.55 | 6.44 |

В данном случае \hat{F} есть распределение, дающее вероятность $1/164$ каждому из 164 наблюдений $\mathbf{x}_i = (z_i, y_i)$. Бутстреп набор есть $\mathbf{x}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_{164}^*)$, где каждый из \mathbf{x}_i^* равен одному из 164 наблюдений с одинаковой вероятностью. Получив \mathbf{x}^* , мы вычислили $\hat{r}_{\text{quad}}^*(z)$ и $\hat{r}_{\text{loess}}^*(z)$, квадратичную и loess кривые на основе \mathbf{x}^* . В завершение мы вычислили значения $\hat{r}_{\text{quad}}^*(60)$ и $\hat{r}_{\text{loess}}^*(60)$, а также $\hat{r}_{\text{quad}}^*(100)$ и $\hat{r}_{\text{loess}}^*(100)$. $B = 50$ значений $\hat{r}_{\text{quad}}^*(60)$ имеют стандартную ошибку 3.03 и т.д., см. таблицу 7.5.

Посмотрев на результаты в таблице 7.5, можно сделать вывод о том, что оценка $\hat{r}_{\text{loess}}(z)$ значительно менее точна, чем $\hat{r}_{\text{quad}}(z)$. Это неудивительно, ведь $\hat{r}_{\text{loess}}(z)$ строится на меньшем количестве данных (размер обусловлен α), чем $\hat{r}_{\text{quad}}(z)$. Неустойчивость $\hat{r}_{\text{loess}}(z)$ очевидна по графикам на рисунке 7.9.

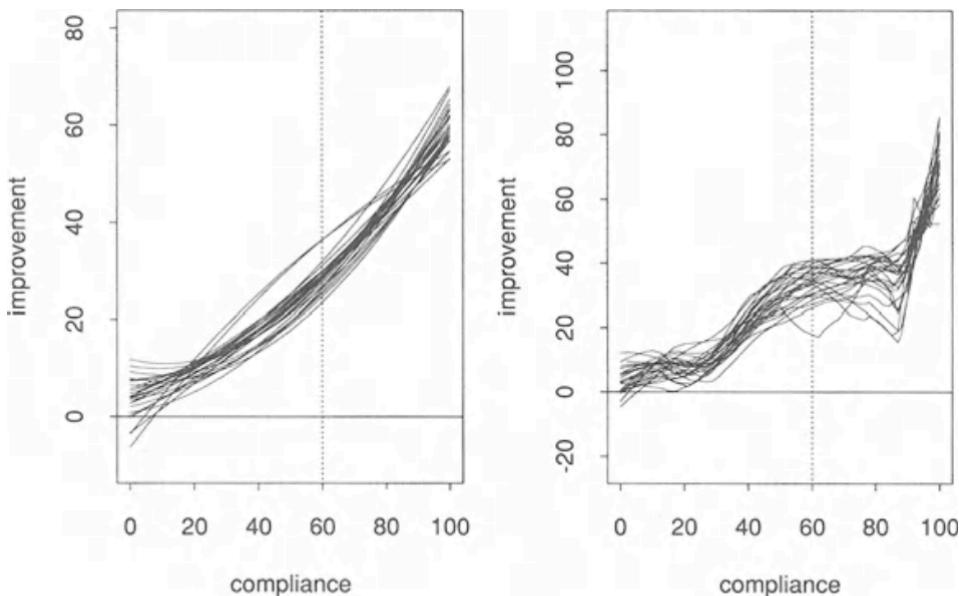
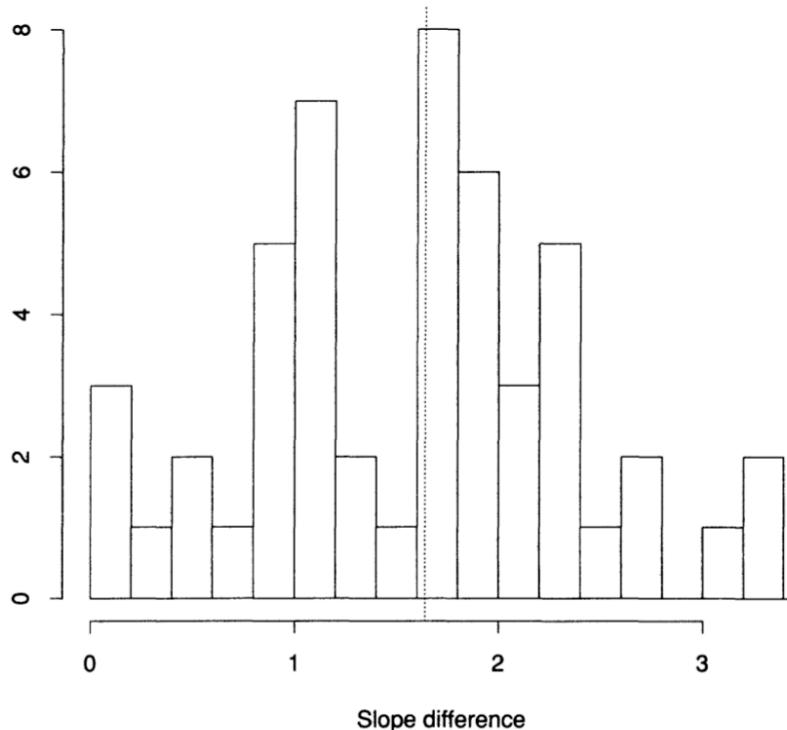


Figure 7.9. *The first 25 bootstrap replications of $\hat{r}_{\text{quad}}(z)$, left panel, and $\hat{r}_{\text{loess}}(z)$, right panel; the increased variability of $\hat{r}_{\text{loess}}(z)$ is evident.*

Полезно построить кривые бутстрепа, чтобы увидеть, сохраняются ли некоторые интересные особенности оригинальной кривой у кривых по бутстреп выборкам. Например, на рисунке 7.6 видим, что $\hat{r}_{\text{loess}}(z)$ растёт гораздо быстрее с $z = 80\%$ до $z = 100\%$, чем с $z = 60\%$ до $z = 80\%$. Разность средних углов наклона составляет

$$\begin{aligned}\hat{\theta} &= \frac{\hat{r}_{\text{loess}}(100) - \hat{r}_{\text{loess}}(80)}{20} - \frac{\hat{r}_{\text{loess}}(80) - \hat{r}_{\text{loess}}(60)}{20} \\ &= \frac{672.78 - 37.50}{20} - \frac{32.50 - 34.03}{20} = 1.84.\end{aligned}\quad (7.28)$$

Соответствующее число для \hat{r}_{quad} составляет лишь 0.17. Большинство loess кривых показывают похожий быстрый рост примерно на 80%. Ни одно из бутстреп значений $\hat{\theta}^*$ не было меньше нуля, минимум составил 0.23, большинство значений оказались больше единицы, см. рисунок 7.10.



смещение. Оба этих свойства происходят от локального характера алгоритма loess, который строит оценку $r(z)$ используя только элементы выборки в окрестности z .

Оценка $\hat{\theta} = 1.59$, построенная на \hat{r}_{loess} имеет большую вариативность, $\widehat{\text{se}}_{50} = 0.61$, однако содержание рисунка 7.10 явно намекает на то, что настоящее значение θ , каким бы оно ни было, больше, чем значение $\hat{\theta} = 0.17$, основанное на \hat{r}_{quad} . Мы рассмотрим эту проблему детальнее в главах 12–14 про бутстреп доверительные интервалы.

Таблица 7.5 намекает на то, что нам следует беспокоиться за оценки $\hat{r}_{\text{quad}}(60)$ и $\hat{r}_{\text{quad}}(100)$, которые могут быть значительно заниженными. Одним из возможных решений этой проблемы может быть выбор полиномиальных моделей более высокой размерности. Достаточно замысловатые теории построения моделей были предложены с целью определить, когда следует продолжать поиск модели в пространстве большей размерности, а когда следует остановиться. Мы глубже рассмотрим вопрос построения регрессионных моделей в главе 9, где данные примера 2 мы рассмотрим снова. Простые бутстреп оценки вариативности и неустойчивости, которые были освещены в данной главе, часто становятся полезным шагом в сторону понимания регрессионных моделей, в особенности нетрадиционных (таких как $\hat{r}_{\text{loess}}(z)$).

7.4 Пример отказа бутстрепа

Предположим, что у нас имеются данные X_1, X_2, \dots, X_n из равномерного распределения на $(0, \theta)$. Оценка $\hat{\theta}$ по методу максимума правдоподобия есть наибольшее значение выборки $X_{(n)}$. Мы сгенерирали выборку из 50 равномерно распределённых чисел на $(0, 1)$ и получили $\hat{\theta} = 0.988$. На левой части рисунка 7.11 показана гистограмма 2000 бутстреп репликаций оценки $\hat{\theta}^*$, полученных с помощью выборок из данных с возвращением. На правой части наблюдаем 2000 репликаций параметрического бутстрепа, полученных при взятии выборок из равномерного распределения на $(0, \hat{\theta})$.³ Ясно, что гистограмма слева есть плохая аппроксимация того, что мы видим на правой. Так, в случае левой гистограммы оказывается, что в 62% репликаций $\hat{\theta}^* = \hat{\theta}$. Вообще говоря, легко показать, что $\text{Prob}(\hat{\theta}^* = \hat{\theta}) = 1 - (1 - 1/n)^n \rightarrow 1 - e^{-1} \approx .632$ когда $n \rightarrow \infty$. Однако, в параметрическом случае правой гистограммы $\text{Prob}(\hat{\theta}^* = \hat{\theta}) = 0$.

³подписи parametric и nonparametric на рисунке следует поменять местами — прим.ред.

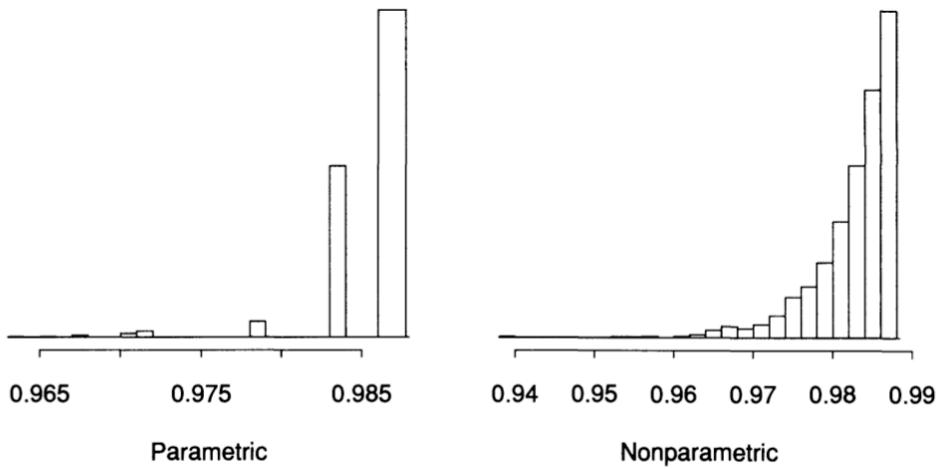


Figure 7.11. The left panel shows a histogram of 2000 bootstrap replications of $\hat{\theta}^* = X_{(n)}$ obtained by sampling with replacement from a sample of 50 uniform numbers. The right panel shows 2000 parametric bootstrap replications obtained by sampling from the uniform distribution on $(0, \hat{\theta})$.

Что не так с непараметрическим бутстрепом? Сложность возникает потому, что эмпирическая функция распределения \hat{F} не является хорошей оценкой настоящего распределения на его краях. Либо параметрические данные о F , либо некоторое сглаживание \hat{F} необходимо для того, чтобы разрешить эту проблему. Детали и ссылки об этой проблеме можно найти в Beran и Ducharme (1991, с.23). Непараметрический бутстреп может отказать и в других примерах, где θ зависит от гладкости F . К примеру, если θ есть число атомов у F , то $\hat{\theta} = n$ будет плохой оценкой $\hat{\theta}$.

Глава 8

Более сложные структуры данных

8.1 Введение

Алгоритм бутстрепа на рисунке 6.1 основан на простейшей возможной вероятностной модели для случайных данных: одновыборочная модель, в которой одно неизвестное вероятностное распределение F порождает данные \mathbf{x} путем случайной выборки

$$F \rightarrow \mathbf{x} = (x_1, x_2, \dots, x_n). \quad (8.1)$$

Отдельные элементы x_i в (8.1) сами по себе могут быть довольно сложными, возможно, в виде чисел, векторов, карт, изображений или чего-то еще, но сам вероятностный механизм прост. Многие задачи анализа данных связаны с более сложными структурами данных. Эти такие структуры как временные ряды, дисперсионный анализ, регрессионные модели, многовыборочные задачи, цензурированные данные, стратифицированная выборка и т.д. Алгоритм бутстрепа можно адаптировать к общим структурам данных, как это обсуждается здесь и в главе 9.

8.2 Одновыборочные задачи

На рис. 8.1 представлена схематическая диаграмма метода бутстрепа применительно к одновыборочным задачам. Слева — реальный мир, где неизвестное распределение F порождает наблюдаемые данные $\mathbf{x} = (x_1, x_2, \dots, x_n)$ путем генерации случайной выборки. Мы вычислили интересующую статистику из \mathbf{x} , $\hat{\theta} = s(\mathbf{x})$, и хотим узнать что-нибудь о статистическом поведении $\hat{\theta}$, возможно, о его стандартной ошибке $se_F(\hat{\theta})$.

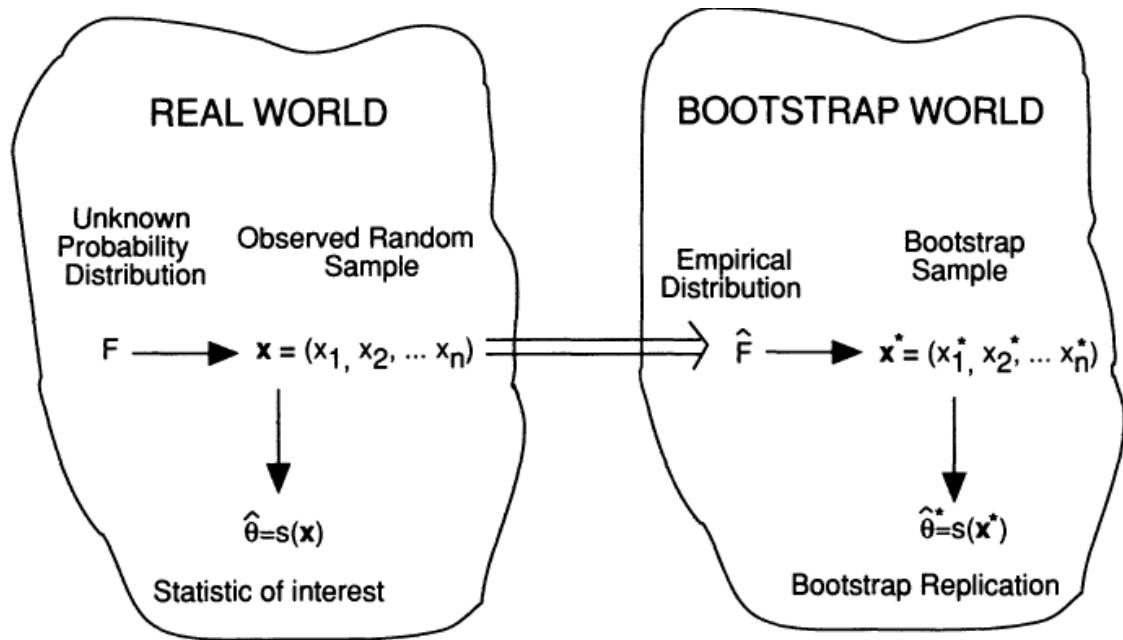


Figure 8.1. *A schematic diagram of the bootstrap as it applies to one-sample problems. In the real world, the unknown probability distribution F gives the data $\mathbf{x} = (x_1, x_2, \dots, x_n)$ by random sampling; from \mathbf{x} we calculate the statistic of interest $\hat{\theta} = s(\mathbf{x})$. In the bootstrap world, \hat{F} generates \mathbf{x}^* by random sampling, giving $\hat{\theta}^* = s(\mathbf{x}^*)$. There is only one observed value of $\hat{\theta}$, but we can generate as many bootstrap replications $\hat{\theta}^*$ as affordable. The crucial step in the bootstrap process is “ \iff ”, the process by which we construct from \mathbf{x} an estimate \hat{F} of the unknown population F .*

В правой части рисунка находится мир бутстрепа, если использовать терминологию Дэвида Фридмана. В мире бутстрепа эмпирическое распределение \hat{F} порождает бутстреп выборки $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*)$ путем генерации случайной выборки, на основе которой мы вычисляем бутстреп репликации интересующей статистики, $\hat{\theta}^* = s(\mathbf{x}^*)$. Большим преимуществом мира бутстрепа является то, что мы можем вычислить столько репликаций $\hat{\theta}^*$, сколько захотим, или, по крайней мере, столько, сколько мы можем себе позволить. Это позволяет нам делать вероятностные вычисления напрямую, например, используя наблюдаемую изменчивость $\hat{\theta}^*$ для оценки ненаблюдаемой величины $se_F(\hat{\theta})$.

Двойная стрелка на рис. 8.1 указывает на вычисление \hat{F} из F . По идеи, это решающий шаг в процессе бутстрепа, даже несмотря на то, что он прост в вычислительном отношении. Любая другая часть картины бутстрепа определяется аналогично: F порождает \mathbf{x} путем генерации случайной выборки, поэтому \hat{F} порождает \mathbf{x}^* путем генерации случайной выборки; $\hat{\theta}$ получается из \mathbf{x} через функцию $s(\mathbf{x})$, поэтому $\hat{\theta}^*$ получается из \mathbf{x}^* таким же образом. Расчеты бутстрепа для более сложных вероятностных механизмов оказываются простыми, если мы знаем, как реализовать процесс, обозначенный двойной стрелкой — оценку всего вероятностного механизма на основе данных. К счастью, это легко сделать для всех распространенных структур данных.

Чтобы облегчить изучение более сложных структур данных, мы будем ис-

пользовать обозначение

$$P \rightarrow \mathbf{x}, \quad (8.2)$$

чтобы указать, что неизвестная *вероятностная модель* P породила наблюдаемый набор данных \mathbf{x} .

8.3 Двухвыборочная задача

Чтобы понять обозначения (8.2), рассмотрим данные о мышах в таблице 2.1. Вероятностную модель P можно представить как пару распределений вероятностей F и G , первое для экспериментальной группы и второе для контрольной группы

$$P = (F, G). \quad (8.3)$$

Пусть $\mathbf{z} = (z_1, z_2, \dots, z_m)$ обозначает экспериментальные наблюдения, а $\mathbf{y} = (y_1, y_2, \dots, y_n)$ обозначает контрольные наблюдения с $n = 7$ и $m = 9$. Тогда наблюдаемые данные включают \mathbf{z} и \mathbf{y}

$$\mathbf{x} = (\mathbf{z}, \mathbf{y}). \quad (8.4)$$

Можем представить себе \mathbf{x} как 16-мерный вектор, если мы помним, что семь элементов из F , девять — из G . Отображение $P \rightarrow \mathbf{x}$ описывается следующим образом:

$$F \rightarrow \mathbf{z} \text{ независимо от } G \rightarrow \mathbf{y}. \quad (8.5)$$

Другими словами, \mathbf{z} — это случайная выборка размера 7 из F , \mathbf{y} — случайная выборка размера 9 из G , причем \mathbf{z} и \mathbf{y} взаимно независимы друг от друга. Такая постановка называется *двуихвыборочной задачей*.

В этом случае легко оценить вероятностный механизм P . Пусть \hat{F} и \hat{G} — эмпирические распределения, основанные на \mathbf{z} и \mathbf{y} соответственно. Тогда естественная оценка $P = (F, G)$ такова:

$$\hat{P} = (\hat{F}, \hat{G}). \quad (8.6)$$

После получения \hat{P} определение бутстреп выборки \mathbf{x}^* очевидно, стрелка в выражении

$$\hat{P} \rightarrow \mathbf{x}^* \quad (8.7)$$

должна означать то же самое, что и стрелка в $P \rightarrow \mathbf{x}$, (8.2). В двухвыборочной задаче (8.5) мы имеем $\mathbf{x}^* = (\mathbf{z}^*, \mathbf{y}^*)$, где

$$\hat{F} \rightarrow \mathbf{z}^* \text{ независимо от } \hat{G} \rightarrow \mathbf{y}^*. \quad (8.8)$$

Размеры выборки для \mathbf{z}^* и \mathbf{y}^* такие же, как для \mathbf{z} и \mathbf{y} соответственно.

На рисунке 8.2 показана гистограмма $B = 1400$ бутстреп репликаций статистики $\hat{\theta}$

$$\begin{aligned} \hat{\theta} &= \hat{\mu}_z - \hat{\mu}_y = \bar{z} - \bar{y} = \\ &= 86.86 - 56.22 = 30.63, \end{aligned} \quad (8.9)$$

где $\hat{\theta}$ — разность в средних значениях экспериментальной и контрольной групп в данных о мышах. Эта статистика оценивает параметр

$$\theta = \mu_z - \mu_y = E_f(z) - E_G(y). \quad (8.10)$$

Если θ действительно намного больше 0, как, по-видимому, указывает (8.9), то экспериментальная группа показывает гораздо лучший результат по сравнению с контрольной группой. Однако бутстреп оценка стандартной ошибки для $\hat{\theta} = 30.63$ это

$$\widehat{se}_{1400} = \left\{ \sum_{b=1}^{1400} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2 / 1399 \right\}^{1/2} = 26.85, \quad (8.11)$$

поэтому $\hat{\theta}$ всего лишь на 1.14 стандартной ошибки больше нуля, $1.14 = 30.63/26.85$. Обычно такой результат не считается убедительным доказательством того, что истинное значение θ больше 0.

Бутстреп репликации $\hat{\theta}^*$ были получены при помощи генератора случайных чисел для соблюдения независимости (8.8). Каждая бутстреп выборка \mathbf{x}^* вычислена следующим образом

$$\mathbf{x}^* = (\mathbf{z}^*, \mathbf{y}^*) = (z_{i_1}, z_{i_2}, \dots, z_{i_7}, y_{j_1}, y_{j_2}, \dots, y_{j_9}), \quad (8.12)$$

где (i_1, i_2, \dots, i_7) есть случайная выборка размера 7 из целых чисел $1, 2, \dots, 7$, а (j_1, j_2, \dots, j_9) есть независимо выбранная случайная выборка размера 9 из целых чисел $1, 2, \dots, 9$. Например, первая бутстреп выборка $(i_1, i_2, \dots, i_7) = (7, 3, 1, 2, 7, 6, 3)$ и $(j_1, j_2, \dots, j_9) = (7, 8, 2, 9, 6, 7, 8, 4, 2)$.

Стандартная ошибка θ может быть записана как $se_P(\hat{\theta})$, чтобы подчеркнуть тот факт, что она зависит от неизвестного вероятностного механизма $P = (F, G)$. Бутстреп оценка $se_{\hat{P}}(\hat{\theta})$ — это оценка методом подстановки

$$se_{\hat{P}}(\hat{\theta}^*) = \{ \text{var}_{\hat{P}}(\bar{z}^* - \bar{y}^*) \}^{1/2}. \quad (8.13)$$

Как и в главе 6, мы аппроксимируем идеальную бутстреп оценку $se_{\hat{P}}(\hat{\theta}^*)$ при помощи \widehat{se}_B из уравнения (6.6), в данном случае при $B = 1400$. Тот факт, что $\hat{\theta}^*$ вычисляется из двух выборок, \mathbf{z}^* и \mathbf{y}^* не влияет на определение (6.6), а именно $\widehat{se}_B = \left\{ \sum_{b=1}^B [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2 / (B - 1) \right\}^{1/2}$.

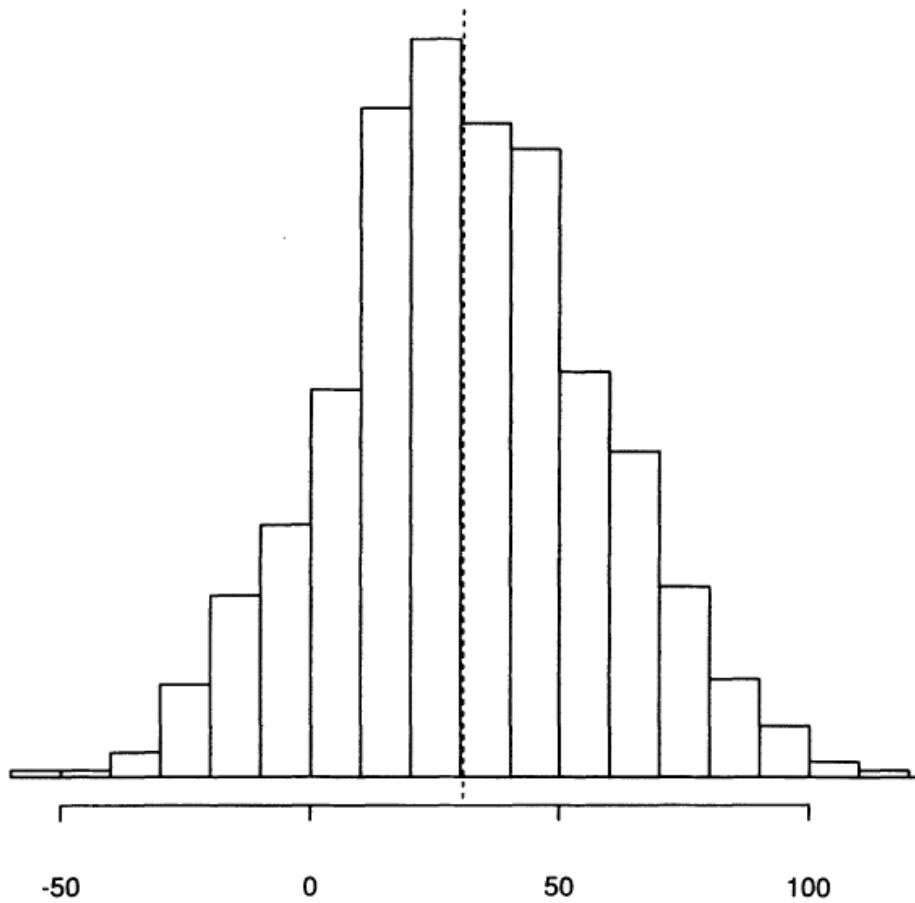


Figure 8.2. 1400 bootstrap replications of $\hat{\theta} = \bar{z} - \bar{y}$, the difference between the Treatment and Control means for the mouse data of Table 2.1; bootstrap estimated standard error was $\hat{se}_{1400} = 26.85$, so the observed value $\hat{\theta} = 30.63$ (broken line) is only 1.14 standard errors above zero; 13.1% of the 1400 $\hat{\theta}^*$ values were less than zero. This is not small enough to be considered convincing evidence that the Treatment worked better than the Control.

8.4 Более общие структуры данных

Рисунок 8.3 — это версия рисунка 8.1, в применении к общим структурам данных $P \rightarrow \mathbf{x}$. Между этими двумя рисунками нет особой концептуальной разницы, за исключением уровня обобщения. В реальном мире неизвестный вероятностный механизм P порождает наблюдаемый набор данных \mathbf{x} в соответствии с правилом построения, указанным стрелкой « \rightarrow ». В конкретных приложениях нам нужно более тщательно определять стрелку, наподобие того как в (8.5) для двухвыборочной задачи. Набор данных \mathbf{x} больше не может быть одним вектором. Он имеет форму, зависящую от структуры данных, например $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ в двухвыборочной задаче.

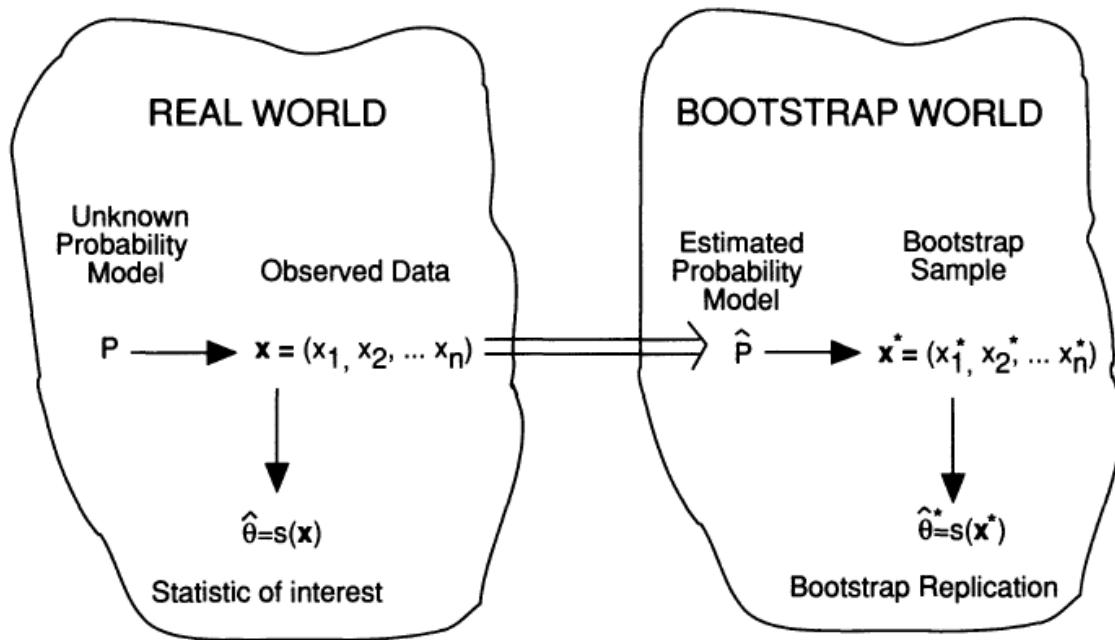


Figure 8.3. *Schematic diagram of the bootstrap applied to problems with a general data structure $P \rightarrow \mathbf{x}$. The crucial step “ \Rightarrow ” produces an estimate \hat{P} of the entire probability mechanism P from the observed data \mathbf{x} . The rest of the bootstrap picture is determined by the real world: “ $\hat{P} \rightarrow \mathbf{x}^*$ ” is the same as “ $P \rightarrow \mathbf{x}$ ”; the mapping from $\mathbf{x}^* \rightarrow \hat{\theta}^*$, $s(\mathbf{x}^*)$, is the same as the mapping from $\mathbf{x} \rightarrow \hat{\theta}$, $s(\mathbf{x})$.*

Наблюдая за \mathbf{x} , мы вычисляем интересующую статистику $\hat{\theta}$ из \mathbf{x} в соответствии с функцией $s(\cdot)$.

Бутстреп часть на рис. 8.3 описывается аналогичными понятиями и в реальном мире: стрелка в $\hat{P} \rightarrow \mathbf{x}^*$ означает то же самое, что и стрелка в $P \rightarrow \mathbf{x}$. И функция, отображающая \mathbf{x}^* в $\hat{\theta}^*$, является той же функцией $s(\cdot)$, что и отображающая \mathbf{x} в $\hat{\theta}$.

При фактическом проведении бутстреп анализа на основе рисунка 8.3 возникают две практические задачи:

(1). Нам нужно оценить весь вероятностный механизм P по наблюдаемым данным \mathbf{x} . Это шаг, обозначенный двойной стрелкой $\mathbf{x} \Rightarrow \hat{P}$. Это удивительно легко сделать для большинства знакомых структур данных. Универсального рецепта нет, но в каждом отдельном случае доступны вполне обычные конкретные решения, например, $\hat{P} = (\hat{F}, \hat{G})$ для двухвыборочной задачи. Дополнительные примеры приведены в этой и следующей главах.

(2). Нам нужно смоделировать бутстреп данные из \hat{P} в соответствии с подходящей структурой данных. Это шаг $\hat{P} \rightarrow \mathbf{x}^*$ изображен на рисунке 8.3. Этот шаг концептуально прост, будучи таким же, как $P \rightarrow \mathbf{x}$, но может потребовать некоторой осторожности при программировании, если необходима вычислительная эффективность. (Мы увидим пример данных про анализ лутенизирующего гормона ниже.) Обычно генерация бутстреп данных $\hat{P} \rightarrow \mathbf{x}^*$ требует меньше времени, чем вычисление $\hat{\theta}^* = s(\mathbf{x}^*)$.

Table 8.1. *The lutenizing hormone data.*

| period | level | period | level | period | level | period | level |
|--------|-------|--------|-------|--------|-------|--------|-------|
| 1 | 2.4 | 13 | 2.2 | 25 | 2.3 | 37 | 1.5 |
| 2 | 2.4 | 14 | 1.8 | 26 | 2.0 | 38 | 1.4 |
| 3 | 2.4 | 15 | 3.2 | 27 | 2.0 | 39 | 2.1 |
| 4 | 2.2 | 16 | 3.2 | 28 | 2.9 | 40 | 3.3 |
| 5 | 2.1 | 17 | 2.7 | 29 | 2.9 | 41 | 3.5 |
| 6 | 1.5 | 18 | 2.2 | 30 | 2.7 | 42 | 3.5 |
| 7 | 2.3 | 19 | 2.2 | 31 | 2.7 | 43 | 3.1 |
| 8 | 2.3 | 20 | 1.9 | 32 | 2.3 | 44 | 2.6 |
| 9 | 2.5 | 21 | 1.9 | 33 | 2.6 | 45 | 2.1 |
| 10 | 2.0 | 22 | 1.8 | 34 | 2.4 | 46 | 3.4 |
| 11 | 1.9 | 23 | 2.7 | 35 | 1.8 | 47 | 3.0 |
| 12 | 1.7 | 24 | 3.0 | 36 | 1.7 | 48 | 2.9 |

8.5 Пример: лютеинизирующий гормон

На рис. 8.4 показан набор данных про уровень y_t лютеинизирующего гормона для каждого из 48 отрезков времени, взятых из Diggle (1990); набор данных приведен в таблице 8.1. Это данные об уровне гормона, измеренные у здоровой женщины с 10-минутными интервалами в течение 8 часов. Лютеинизирующий гормон является одним из гормонов, регулирующих менструальный цикл, и поэтому важно понимать его суточные колебания.

Понятно, что уровни гормона не являются случайной выборкой из какого-либо распределения. На рисунке 8.4 слишком много данных. Эти данные являются примером *временного ряда*: структуры данных, для которой близкие значения временного параметра t указывают тесно связанные значения измеренной величины y_t . Для анализа временных рядов используются многие интересные вероятностные модели. Мы начнем с простейшей модели — *схемы авторегрессии первого порядка*.

Пусть μ — математическое ожидание y_t , которое предполагается одинаковым для всех моментов времени t , и определим *центрированные* измерения

$$z_t = y_t - \mu. \quad (8.14)$$

Все z_t имеют математическое ожидание 0. Схема авторегрессии первого порядка — это одна из таких схем, в которой каждый z_t является линейной комбинацией предыдущего значения z_{t-1} и независимого шума ε_t ,

$$z_t = \beta z_{t-1} + \varepsilon_t \text{ для } t = U, U + 1, U + 2, \dots, V. \quad (8.15)$$

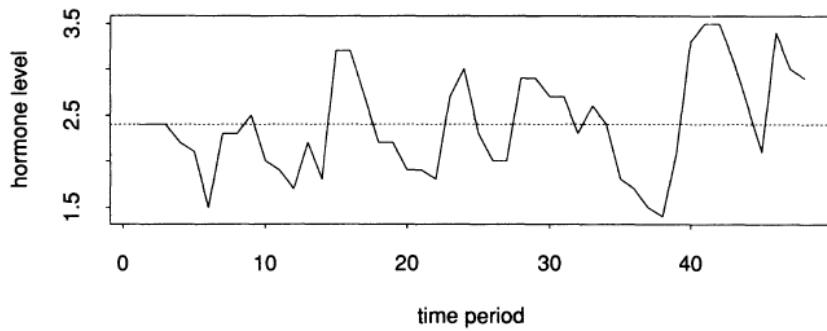


Figure 8.4. The lutenizing hormone data. Level of lutenizing hormones y_t plotted versus time period t , for t from 1 to 48. In this plot and other plots the points are connected by lines to enhance visibility. The average value $\hat{\mu} = 2.4$ is indicated by a dashed line. Table 8.1 lists the data.

Здесь β — неизвестный параметр, действительное число от -1 до 1 .

Предполагается, что шум ε_t в (8.15) является случайной выборкой из неизвестного распределения F с математическим ожиданием 0 ,

$$F \rightarrow (\varepsilon_U, \varepsilon_{U+1}, \varepsilon_{U+2}, \dots, \varepsilon_V) \quad [\mathrm{E}_F(\varepsilon) = 0]. \quad (8.16)$$

Точки U и V — это начало и конец анализируемого периода времени. Здесь у нас есть

$$U = 2 \quad \text{и} \quad V = 48. \quad (8.17)$$

Обратите внимание, что первое уравнение в (8.15) имеет вид

$$z_U = \beta z_{U-1} + \varepsilon_U, \quad (8.18)$$

поэтому нам нужно число z_{U-1} , чтобы запустить процесс авторегрессии. В нашем случае $z_{U-1} = z_1$.

Пусть мы считаем, что модель (8.15), (8.16), авторегрессионный процесс первого порядка, применима к данным лутенизирующего гормона. Как мы можем оценить значение β по данным? Один из ответов основан на методе наименьших квадратов. Прежде всего, мы оцениваем математическое ожидание μ в (8.14) по наблюдаемому среднему значению \bar{y} (это 2.4 для данных лутенизирующего гормона) и задаем

$$z_t = y_t - \bar{y} \quad (8.19)$$

для всех значений t . В дальнейшем мы проигнорируем разницу между определениями (8.14) и (8.19).

Предположим, что b — это любое предположение об истинном значении β в (8.15). Определим остаточную квадратичную ошибку для этого предположения как

$$\mathrm{RSE}(b) = \sum_{t=U}^V (z_t - b z_{t-1})^2. \quad (8.20)$$

Используя (8.15) и тот факт, что $\mathrm{E}_F(\varepsilon) = 0$, легко показать, что $\mathrm{RSE}(b)$ имеет математическое ожидание

$$\mathrm{E}(\mathrm{RSE}(b)) = (b - \beta)^2 \mathrm{E}\left(\sum_{t=U}^V z_{t-1}^2\right) + (V - U + 1) \mathrm{var}_F(\varepsilon).$$

Оно минимизируется, когда b равно истинному значению β . Мы убедились, что $\text{RSE}(b)$ должен достичь своего минимума где-то рядом с истинным значением β .

Учитывая данные временного ряда, мы можем рассчитать $\text{RSE}(b)$ как функцию b , и выбрать минимизирующее значение, которое будет нашей оценкой β

$$\text{RSE}(\hat{\beta}) = \min_b \text{RSE}(b). \quad (8.21)$$

Данные про лутенизирующий гормон имеют следующую оценку наименьших квадратов

$$\hat{\beta} = 0.586. \quad (8.22)$$

Насколько точна оценка $\hat{\beta}$? Чтобы ответить на этот вопрос, мы можем использовать общую бутстреп процедуру, показанную на рис. 8.3. Вероятностный механизм P , описанный в (8.15), (8.16), имеет два неизвестных элемента, β и F , скажем, $P = (\beta, F)$. (Здесь мы рассматриваем μ в (8.14) как известную и равную \bar{y} .) Данные \mathbf{x} состоят из наблюдений y_t и соответствующих им периодов времени t . Мы знаем, что правило построения $P \rightarrow \mathbf{x}$ описывается формулами (8.15)–(8.16). Интересующая статистика $\hat{\theta}$ равна $\hat{\beta}$, поэтому отображения $s(\cdot)$ неявно задаются (8.21).

Остается один шаг, прежде чем мы сможем применить бутстреп: шаг двойной стрелки $\mathbf{x} \Rightarrow \hat{P}$, в котором $P = (\beta, F)$ оценивается по данным. Теперь β уже была оценена с помощью $\hat{\beta}$, (8.21), поэтому нам нужно только оценить распределение отклонений F . Если бы мы знали β , то мы могли бы вычислить $\varepsilon_t = z_t - \beta z_{t-1}$ для каждого t и оценить F по эмпирическому распределению значений ε_t . Мы не знаем β , то мы можем использовать оценочное значение $\hat{\beta}$, чтобы вычислить *отклонения [approximate disturbances]*

$$\hat{\varepsilon}_t = z_t - \hat{\beta} z_{t-1} \text{ для } t = U, U + 1, U + 2, \dots, V. \quad (8.23)$$

Пусть $T = V - U + 1$, количество слагаемых в (8.23); $T = 47$ для выбора (8.17). Очевидная оценка F — это \hat{F} , эмпирическое распределение отклонений

$$\hat{F} : \text{вероятность } 1/T \text{ для } \hat{\varepsilon}_t \text{ при } t = U, U + 1, \dots, V. \quad (8.24)$$

На рис. 8.5 показана гистограмма отклонений $\hat{\varepsilon}_t = z_t - \hat{\beta} z_{t-1}$ с $T = 47$ для схемы авторегрессии первого порядка, примененной к данным о лутенизирующем гормоне в промежутке от 2 до 48 лет.

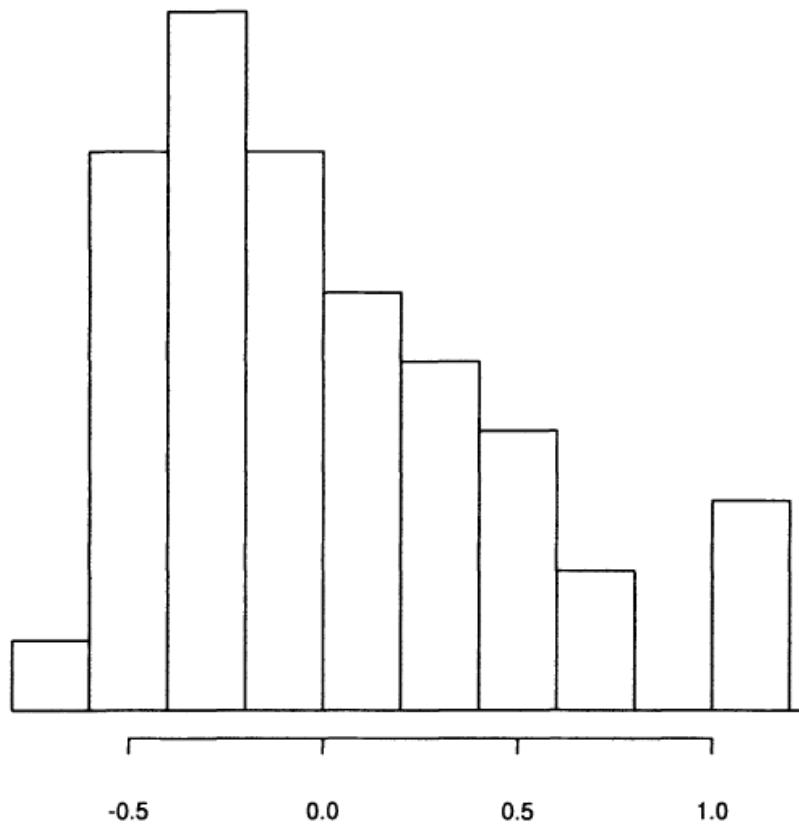


Figure 8.5. Histogram of the 47 approximate disturbances $\hat{\epsilon}_t = z_t - \hat{\beta}z_{t-1}$, for $t = 2$ through 48; $\hat{\beta}$ equals 0.586 the least-squares estimate for the first-order autoregressive scheme. The distribution is long-tailed to the right. The disturbances averaged 0.006, with a standard deviation of 0.454, and so are nearly centered at zero.

Мы видим, что распределение \hat{F} не является нормальным и имеет длинный хвост справа. Распределение имеет среднее значение 0.006 и стандартное отклонение 0.454. Не случайно, что среднее значение \hat{F} близко к 0. Если бы это было не так, мы могли бы соблюдать определение $E_F(\varepsilon) = 0$ в (8.16) путём центрирования \hat{F} ; то есть изменением каждой вероятностной точки в (8.23) от $\hat{\epsilon}_t$ до $\hat{\epsilon}_t - \bar{\varepsilon}$, где $\bar{\varepsilon} = \sum_{t=U}^V \hat{\epsilon}_t / T$.

Теперь мы готовы провести бутстреп анализ точности оценки $\hat{\beta} = 0.586$. Набор бутстреп данных $\hat{P} \rightarrow \mathbf{x}^*$ генерируется в соответствии с определениями (8.15)–(8.16), за исключением $\hat{P} = (\hat{\beta}, \hat{F})$, заменённого на $P = (\beta, F)$. Начнем с начального значение $z_1 = y_1 - \bar{y}$, которое считается фиксированной константой (как размер выборки n в одновыборочной задаче). Бутстреп временной ряд z_t^* вычисляется рекурсивно

$$\begin{aligned}
 z_2^* &= \hat{\beta}z_1 + \varepsilon_2^* \\
 z_3^* &= \hat{\beta}z_2^* + \varepsilon_3^* \\
 z_4^* &= \hat{\beta}z_3^* + \varepsilon_4^* \\
 &\vdots \\
 z_{48}^* &= \hat{\beta}z_{47}^* + \varepsilon_{48}^*. \tag{8.25}
 \end{aligned}$$

Бутстреп остатки ε_t^* представляют собой случайную выборку из \hat{F} ,

$$\hat{F} \rightarrow (\varepsilon_1^*, \varepsilon_2^*, \dots, \varepsilon_{48}^*). \quad (8.26)$$

Другими словами, каждый ε_t^* равняется любому из T отклонений (8.23) с вероятностью $1/T$.

Процесс бутстрепа (8.25)–(8.26) был запущен $B = 200$ раз, что дало 200 бутстреп временных рядов. Каждый из них дал бутстреп репликацию $\hat{\beta}^*$ для оценки $\hat{\beta}$ методом наименьших квадратов, (8.21). На рисунке 8.6 показана гистограмма из 200 значений $\hat{\beta}^*$. Оценка бутстреп стандартной ошибки для $\hat{\beta}$ составляет $s\hat{e}_{200} = 0.116$. Гистограмма имеет довольно нормальную форму.

В схеме авторегрессии первого порядка каждый z_t зависит от своих предшественников только через значение z_{t-1} (Этот вид зависимости известен как *марковский процесс первого порядка*.) Схема авторегрессии второго порядка расширяет зависимость обратно до z_{t-2} ,

$$z_t = \beta_1 z_{t-1} + \beta_2 z_{t-2} + \varepsilon_t \quad \text{для } t = U, U+1, U+2, \dots, V. \quad (8.27)$$

Здесь $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ — двумерный вектор неизвестных параметров. ε_t — независимые отклонения, как в (8.16). Согласно (8.18) исходные уравнения следующие

$$\begin{aligned} z_U &= \beta_1 z_{U-1} + \beta_2 z_{U-2} + \varepsilon_U \\ z_{U+1} &= \beta_1 z_U + \beta_2 z_{U-1} + \varepsilon_{U+1}, \end{aligned} \quad (8.28)$$

поэтому нам нужны числа z_{U-2} и z_{U-1} для начала. Теперь $U = 3$, $V = 48$ и $T = V - U + 1 = 46$.

Метод наименьших квадратов непосредственно приводит к оценке вектора $\boldsymbol{\beta}$. Пусть \mathbf{z} является T -мерным вектором $(z_U, z_{U+1}, \dots, z_V)^T$, и пусть \mathbf{Z} — матрица $T \times 2$ с первым столбцом $(z_{U-1}, z_U, \dots, z_{V-1})^T$, вторым столбцом $(z_{U-2}, z_{U-1}, z_U, \dots, z_{V-2})^T$. Тогда оценка $\boldsymbol{\beta}$ методом наименьших квадратов следующая

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{z}. \quad (8.29)$$

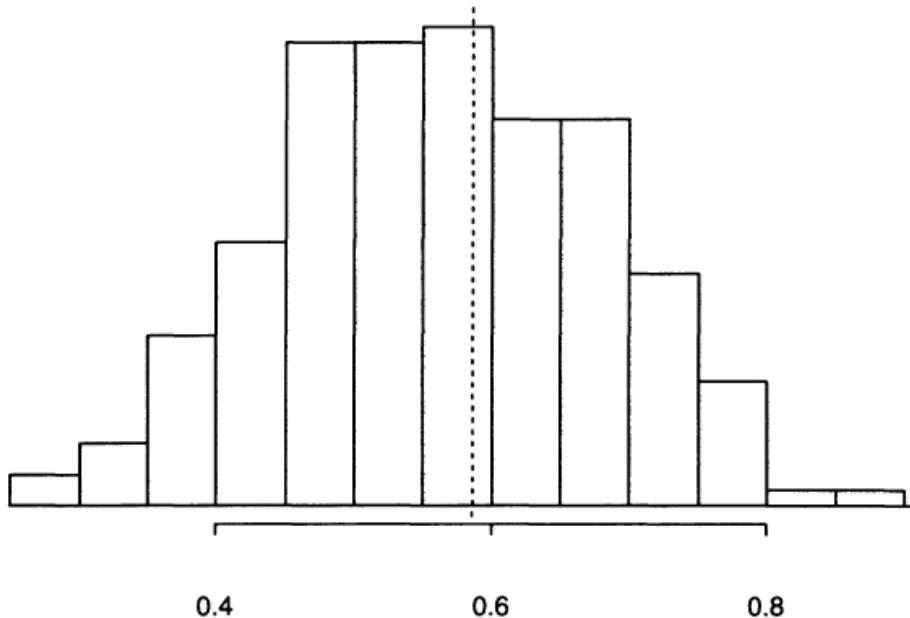


Figure 8.6. Histogram of $B = 200$ bootstrap replications of $\hat{\beta}$, the first-order autoregressive parameter estimate for the lutenizing hormone data; from (8.25), (8.26); the bootstrap estimate of standard error is $\hat{se}_{200} = 0.116$. The broken line is drawn at the observed value $\hat{\beta} = 0.586$.

Для данных о лутенизирующем гормоне схема авторегрессии второго порядка имела следующие оценки наименьших квадратов

$$\hat{\beta} = (0.771, -0.222)^T. \quad (8.30)$$

На рис. 8.7 показаны гистограммы с $B = 200$ бутстреп репликациями из двух компонентов из $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T$. Стандартные ошибки бутстрепа равны

$$\hat{se}_{200}(\hat{\beta}_1) = 0.147, \quad \hat{se}_{200}(\hat{\beta}_2) = 0.149. \quad (8.31)$$

Обе гистограммы приближенно имеют форму нормального распределения.

Схема авторегрессии второго порядка при $\beta_2 = 0$ является схемой авторегрессии первого порядка. При выполнении анализа точности для схемы второго порядка мы проверяем, не отклоняется ли $\hat{\beta}_2$ от 0 менее чем на 2 стандартные ошибки, что обычно интерпретируется как незначительное отличие $\hat{\beta}_2$ от 0. Здесь $\hat{\beta}_2$ — это примерное отклонение на 1.5 стандартные ошибки от 0, и в этом случае у нас нет убедительных доказательств того, что схема авторегрессии первого порядка не дает разумного представления данных о лютеинизирующем гормоне.

Знаем ли мы наверняка, что схема первого порядка дает хорошее представление о ряде лутенизирующих гормонов? Мы не можем дать окончательный от-

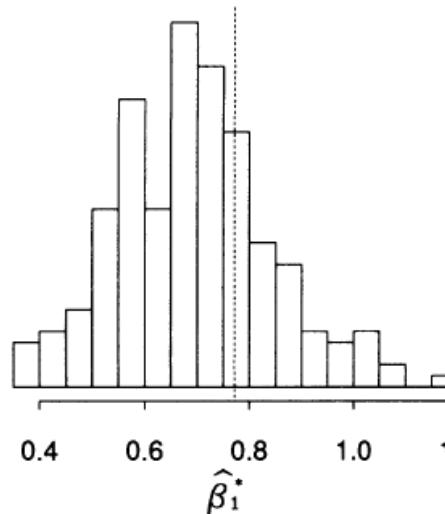


Figure 8.7. $B = 200$ bootstrap second-order autoregressive parameter estimate for the hormone data. As in the other

вет на этот вопрос, не рассматривая еще более общие модели, такие как схемы авторегрессии более высоких порядков. Приблизительный ответ может быть получен путем сравнения бустреп временных рядов с фактическими рядами на рис. 8.4. На рисунке 8.8 на левых графиках показаны первые четыре бустреп набора из первой схемы, правые четыре графика отображают реализации, полученные путем выборки с повторением из исходного временного ряда. Исходные данные на рис. 8.4 очень похожи на реализации левых графиков и совсем не похожи на реализации правых графиков.

Дальнейший анализ показывает, что модель AR(1) обеспечивает разумное соответствие этим данным. Однако нам потребуется более длительный временной ряд, чтобы эффективно различать разные модели для этого гормона.

В общем, стоит помнить, что математические модели представляют собой удобные упрощенные представления сложных явлений реального мира и иногда не совсем корректны. Часто необходим некоторый компромисс между усложнением модели и научными потребностями исследования. Методы бустрепа особенно полезны, если существует потребность в сложных моделях, поскольку математическая сложность не является препятствием для анализа точности бустрепа.

8.6 Бутстреп скользящих окон

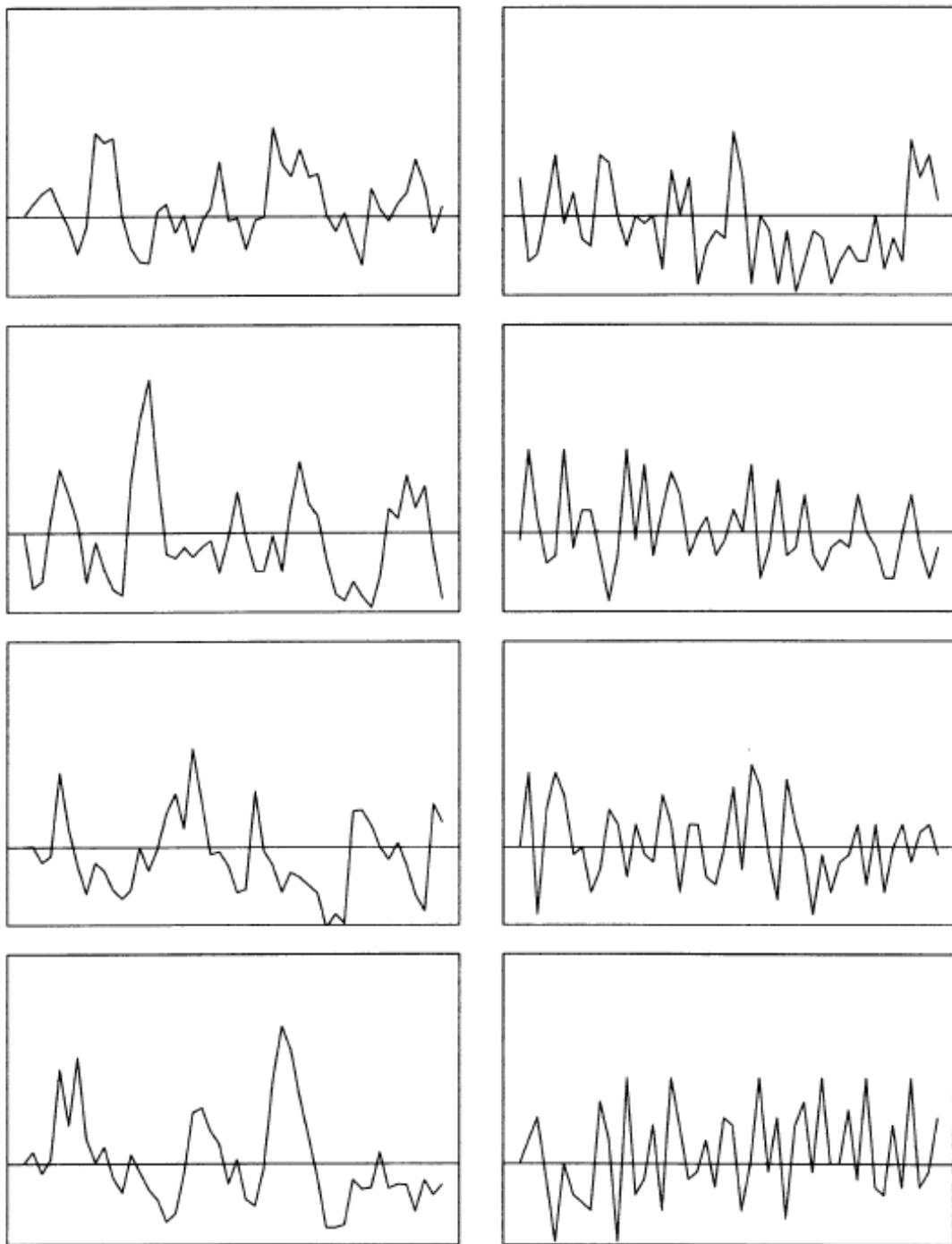


Figure 8.8. Left panel: the first four bootstrap replications of the luteinizing hormone data from the first-order autoregressive scheme, $y_t^* = z_t^* + 2.4$, (8.25), (8.26). Right panel: four bootstrap replications obtained by sampling with replacement from the original time series. The values from the first-order scheme look a lot more like the actual time series in Figure 8.4.

В этом разделе мы кратко опишем другой метод применения бутстрепа к вре-

менным рядам. Вместо подгонки модели и последующей выборки из остатков этот метод использует подход, более близкий к подходу, используемому для одновыборочных задач. Идея проиллюстрирована на рисунке 8.9. Исходный временной ряд представлен черными кружками. Чтобы сгенерировать бутстреп реализацию временного ряда (белые кружки), мы выбираем длину окна («3» на диаграмме) и рассматриваем все возможные смежные окна этой длины. Мы составляем выборку с заменой из этих окон и объединяем их вместе, чтобы сформировать бутстреп временные ряды. Выбирается ровно столько окон, сколько необходимо для получения серии примерно такой же длины, что и исходная. Если длина окна равна l , то выберем k окон так, чтобы $n \approx k \cdot l$.

Для иллюстрации мы выполним эти действия для данных о лютенизирующем гормоне. Интересующей статистикой была оценка $\hat{\beta}$ методом наименьших квадратов у AR(1). Мы выбрали длину окна 3 и использовали бутстреп скользящих окон для генерации бутстреп выборки для данных про лютенизирующий гормон. Типичная реализация бутстрепа показана на рисунке 8.10, и она очень похожа на исходный временной ряд. Затем мы подогнали модель AR(1) к этому бутстреп временному ряду и оценили коэффициент $\hat{\beta}^*$ у AR(1). Весь этот процесс был повторен $B = 200$ раз. (Обратите внимание, что модель AR(1) используется здесь для оценки β , но не используется при генерации бутстреп реализаций временного ряда.) В результате стандартная ошибка бутстрепа составила $\hat{s}_e_{200}(\hat{\beta}) = 0.120$. Это примерно то же самое, что и значение 0.116, полученное из AR(1) выборок, сгенерированных в предыдущем разделе. Увеличение размера окна до 5 привело к уменьшению этого значения до 0.103.

Чем оправдан бутстреп скользящих окон? Как мы видели ранее, мы не можем просто создать повторную выборку из отдельных наблюдений, так как это разрушило бы корреляцию, на которой мы хотим сфокусировать внимание. (Использование размера окна, равного единице, соответствует выборке с возвращением, и дает 0.139 для оценки стандартной ошибки.) С бутстрепом скользящих окон идея состоит в том, чтобы выбрать размер окна l достаточно большим, чтобы наблюдения, отстоящие более чем на l единиц времени, были почти независимыми. Выбирая окна длиной l , мы сохраняем корреляцию, присутствующую в наблюдениях, отстоящих менее чем на l единиц.

Бутстреп скользящих окон имеет то преимущество, что он менее «зависит от модели», чем подход бутстреп остатков, использовавшийся ранее. Как мы видели, последний метод зависит от модели, которая соответствует исходному временному ряду (например, модель AR(1) или AR(2)). Однако выбор размера окна l может быть весьма важным, и эффективные методы для этого еще не разработаны.

В задаче регрессии, обсуждаемой в следующей главе, мы сталкиваемся с различными методами бутстрепа, аналогичными подходам для временных рядов, которые мы обсуждали здесь.

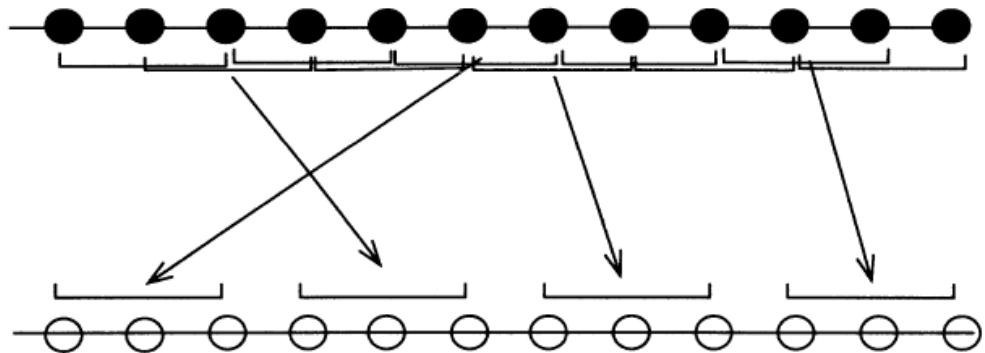


Figure 8.9. A schematic diagram of the moving blocks bootstrap for time series. The black circles are the original time series. A bootstrap realization of the time series (white circles) is generated by choosing a block length ("3" in the diagram) and sampling with replacement from all possible contiguous blocks of this length.

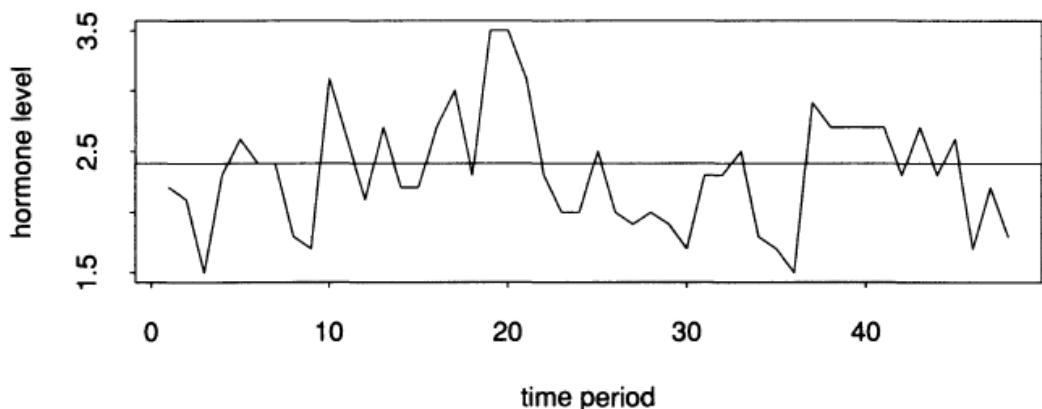


Figure 8.10. A bootstrap realization of the lutenizing hormone data, using the moving blocks bootstrap with block length equal to 3.

8.7 Библиографические примечания

Анализ временных рядов описан во многих книгах, включая Бокс and Дженкинс (1970), Чатфилд (1980) и Дигл (1990). Применение бутстрепа к временным рядам обсуждается в Эффрон and Тибштранни (1986); Метод скользящих окон и связанные с ним техники можно найти у Карлштейн (1986), Киинч (1989), Луи и Сингх (1992) и Политис и Романо (1992).

Глава 9

Модели регрессии

9.1 Введение

Регрессионные модели являются одними из самых полезных и наиболее часто используемых статистических методов. Они предлагают относительно простой анализ сложных ситуаций, когда мы пытаемся отсортировать влияние многих возможных объясняющих переменных на зависимую переменную. В главе 7 мы используем алгоритм одновыборочного бутстрепа, алгоритм для анализа точности регрессионного анализа данных холостиармина из таблицы 7.4. Здесь мы более критически смотрим на задачу регрессии. Рассмотрен общий бутстреп алгоритм, показанный на рис. 8.3, что приводит к несколько иному бутстреп анализу для задач регрессии.

9.2 Линейная регрессионная модель

Мы начнем с классической модели линейной регрессии, или линейной модели, восходящей к Лежандру и Гауссу в начале 19 века. Набор данных \mathbf{x} для модели линейной регрессии состоит из n точек $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, где каждый \mathbf{x}_i сам по себе является парой, скажем

$$\mathbf{x}_i = (\mathbf{c}_i, y_i). \quad (9.1)$$

Здесь \mathbf{c}_i — это $1 \times p$ вектор $\mathbf{c}_i = (c_{i1}, c_{i2}, \dots, c_{ip})$, называемый *вектором признаков* или *предиктором*, а y_i — действительное число, называемое *ответом*.

Пусть μ_i указывает условное ожидание i -го ответа y_i с учетом предиктора \mathbf{c}_i ,

$$\mu_i = E(y_i | \mathbf{c}_i) \quad (i = 1, 2, \dots, n). \quad (9.2)$$

Ключевое предположение в линейной модели состоит в том, что μ_i представляется собой линейную комбинацию компонентов предиктора \mathbf{c}_i ,

$$\mu_i = \mathbf{c}_i \boldsymbol{\beta} = \sum_{j=1}^p c_{ij} \beta_j. \quad (9.3)$$

Вектор параметров, или *параметр регрессии*, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ неизвестен, обычна цель регрессионного анализа состоит в том, чтобы вывести $\boldsymbol{\beta}$ из наблюдаемых данных $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$. В квадратичной регрессии (7.20) для данных холостиармина ответ y_i — это улучшение для i -го человека, признак \mathbf{c}_i — это вектор $(1, z_i, z_i^2)$ и $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$. Примечание: «Линейность» в линейной регрессии относится к линейной форме математического ожидания (9.3).

Нет никакого противоречия в том, что линейная модель (7.20) является квадратичной функцией z .

Вероятностная структура линейной модели обычно выражается как

$$y_i = \mathbf{c}_i\beta + \varepsilon_i \quad \text{для } i = 1, 2, \dots, n. \quad (9.4)$$

Предполагается, что ошибка ε_i в (9.4) является случайной выборкой из неизвестного распределения ошибок F с математическим ожиданием 0,

$$F \rightarrow (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n) = \boldsymbol{\varepsilon} \quad [\mathbb{E}_F(\boldsymbol{\varepsilon}) = 0]. \quad (9.5)$$

Заметим, что (9.4), (9.5) влекут

$$\begin{aligned} \mathbb{E}(y_i|\mathbf{c}_i) &= \mathbb{E}(\mathbf{c}_i\beta + \varepsilon_i|\mathbf{c}_i) = \mathbb{E}(\mathbf{c}_i\beta|\mathbf{c}_i) + \mathbb{E}(\varepsilon_i|\mathbf{c}_i) \\ &= \mathbf{c}_i\beta, \end{aligned} \quad (9.6)$$

что является предположением о линейности (9.3). Здесь мы использовали тот факт, что математическое ожидание $\mathbb{E}(\varepsilon_i|\mathbf{c}_i)$ совпадает с безусловным ожиданием $\mathbb{E}(\varepsilon_i) = 0$, поскольку ε_i выбираются независимо от \mathbf{c}_i .

Мы хотим оценить вектор коэффициентов β из наблюдаемых данных $(\mathbf{c}_1, y_1), (\mathbf{c}_2, y_2), \dots, (\mathbf{c}_n, y_n)$. Пробное значение β , скажем \mathbf{b} , дает *остаточную квадратичную ошибку*

$$\text{RSE}(\mathbf{b}) = \sum_{i=1}^n (y_i - \mathbf{c}_i\mathbf{b})^2, \quad (9.7)$$

как в уравнении (7.21). Оценка методом наименьших квадратов β — это значение $\hat{\beta}$ из \mathbf{b} , которое минимизирует $\text{RSE}(\mathbf{b})$,

$$\text{RSE}(\hat{\beta}) = \min_{\mathbf{b}} [\text{RSE}(\mathbf{b})]. \quad (9.8)$$

Пусть \mathbf{C} — матрица размерности $n \times p$ с i -й строкой \mathbf{c}_i , а \mathbf{y} — вектор $(y_1, y_2, \dots, y_n)^T$. Тогда оценка методом наименьших квадратов является решением следующего уравнения

$$\mathbf{C}^T \mathbf{C} \hat{\beta} = \mathbf{C}^T \mathbf{y} \quad (9.9)$$

Table 9.1. *The hormone data. Amount in milligrams of anti-inflammatory hormone remaining in 27 devices, after a certain number of hours of wear. The devices were sampled from 3 different manufacturing lots, called A, B, and C. Lot C looks like it had greater amounts of remaining hormone, but it also was worn the least number of hours. A regression analysis clarifies the situation.*

| lot | hrs | amount | lot | hrs | amount | lot | hrs | amount | |
|-------|-----|--------|------|-----|--------|------|-----|--------|------|
| A | 99 | 25.8 | B | 376 | 16.3 | C | 119 | 28.8 | |
| A | 152 | 20.5 | B | 385 | 11.6 | C | 188 | 22.0 | |
| A | 293 | 14.3 | B | 402 | 11.8 | C | 115 | 29.7 | |
| A | 155 | 23.2 | B | 29 | 32.5 | C | 88 | 28.9 | |
| A | 196 | 20.6 | B | 76 | 32.0 | C | 58 | 32.8 | |
| A | 53 | 31.1 | B | 296 | 18.0 | C | 49 | 32.5 | |
| A | 184 | 20.9 | B | 151 | 24.1 | C | 150 | 25.4 | |
| A | 171 | 20.9 | B | 177 | 26.5 | C | 107 | 31.7 | |
| A | 52 | 30.4 | B | 209 | 25.8 | C | 125 | 28.5 | |
| mean: | | 150.6 | 23.1 | | 233.4 | 22.1 | | 111.0 | 28.9 |

и задается формулой

$$\hat{\boldsymbol{\beta}} = (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{y}. \quad (9.10)$$

9.3 Пример: данные по гормонам

В таблице 9.1 показан небольшой набор данных, который является подходящим для регрессионного анализа. Медицинское устройство для непрерывной доставки противовоспалительного гормона было протестировано на 27 пациентах. Переменная ответа y_i — это количество гормона, оставшееся в устройстве после ношения,

y_i = оставшееся количество гормона в устройстве i , $i = 1, 2, \dots, 27$.

Есть две переменные—предикторы,

z_i = количество часов ношения i -го устройства

и

L_i = производственная партия устройства i .

Тестируемые устройства были случайным образом выбраны из трех различных производственных партий, названных A , B и C .

Левая часть рисунка 9.1 представляет собой диаграмму рассеяния 27 точек $(z_i, y_i) = (\text{часы}_i, \text{число}_i)$ с символом L_i , используемым в качестве графического сивола. Мы видим, что более длительное время ношения приводит к меньшему количеству оставшегося гормона, как и следовало ожидать. Мы можем количественно оценить это наблюдение с помощью регрессионного анализа.

Рассмотрим модель, в которой математическое ожидание y является линейной функцией z ,

$$\mu_i = E(y_i|z_i) = \beta_0 + \beta_1 z_i \quad i = 1, 2, \dots, 27. \quad (9.11)$$

Эта модель игнорирует L_i : она имеет форму (9.3) с векторами признаков размерности $p = 2$,

$$\mathbf{c}_i = (1, z_i). \quad (9.12)$$

Вектор неизвестных параметров $\boldsymbol{\beta}$ был помечен (β_0, β_1) вместо (β_1, β_2) , так что индексы соответствуют степеням z , как в (7.20). Нормальные уравнения (9.10) дают оценку по методу наименьших квадратов

$$\hat{\boldsymbol{\beta}} = (34.17, -0.0574)^T. \quad (9.13)$$

Линия регрессии, оцененная методом наименьших квадратов

$$\hat{\mu}_i = \mathbf{c}_i \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 z_i \quad (9.14)$$

изображена на правой части рисунка 9.1. Среди всех возможных линий, которые можно было нарисовать, эта линия минимизирует сумму квадратов 27 вертикальных расстояний от точек до линии.

Насколько точен оценочный вектор параметров $\hat{\boldsymbol{\beta}}$? Ответ дает чрезвычайно полезная формула, также восходящая к Лежандру и Гауссу. Пусть \mathbf{G} — матрица скалярных произведений $p \times p$,

$$\mathbf{G} = \mathbf{C}^T \mathbf{C}, \quad (9.15)$$

матрица с элементом $g_{hj} = \sum_{i=1}^n c_{ih}c_{ij}$ в строке h , столбце j . Пусть σ_F^2 будет дисперсией ошибок в модели (9.4),

$$\sigma_F^2 = \text{var}_F(\varepsilon). \quad (9.16)$$

Тогда стандартная ошибка j -го компонента $\hat{\beta}_j$, квадратного корня из его дисперсии, равна

$$\text{se}(\hat{\beta}_j) = \sigma_F \sqrt{G^{jj}} \quad (9.17)$$

где G^{jj} — j -й диагональный элемент обратной матрицы \mathbf{G}^{-1} .

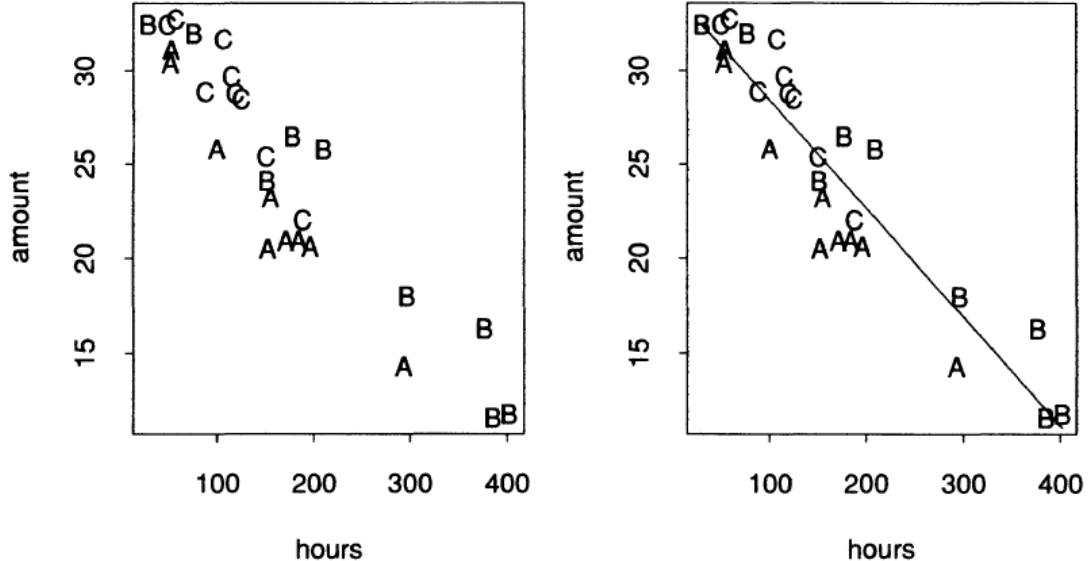


Figure 9.1. Scatterplot of the hormone data points $(z_i, y_i) = (hours_i, amount_i)$, labeled by lot. It is clear that longer hours of wear result in lower amounts of remaining hormone. The right panel shows the least-squares regression of y_i on z_i : $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 z_i$, where $\hat{\beta} = (34.17, - .0574)$.

Последняя формула является обобщением формулы (5.4) для стандартной ошибки выборочного среднего, $\text{se}_F(\bar{x}) = \sigma_F / \sqrt{n}$, см. задачу 9.1. На практике σ_F оценивается по формуле, аналогичной (5.11),

$$\hat{\sigma}_F = \left\{ \sum_{i=1}^n (y_i - \mathbf{c}_i \hat{\beta})^2 / n \right\}^{1/2} = \{\text{RSE}(\hat{\beta})/n\}^{1/2} \quad (9.18)$$

или версией $\hat{\sigma}_F$ с корректированным смещением,

$$\bar{\sigma}_F = \{\text{RSE}(\hat{\beta})/(n-p)\}^{1/2}. \quad (9.19)$$

Соответствующие оценочные стандартные ошибки для компонентов $\hat{\beta}$ равны

$$\hat{\text{se}}(\hat{\beta}_j) = \hat{\sigma}_F \sqrt{G^{jj}} \quad \text{или} \quad \overline{\text{se}}(\hat{\beta}_j) = \hat{\sigma}_F \sqrt{G^{jj}}. \quad (9.20)$$

Связь между $\hat{\text{se}}(\hat{\beta}_j)$ и $\overline{\text{se}}(\hat{\beta}_j)$ такая же, как между формулами (5.12) и (2.2) для среднего.

Table 9.2. Results of fitting model (9.11) to the hormone data

| | Estimate | \hat{s}_e | \bar{s}_e |
|-----------------|----------|-------------|-------------|
| $\hat{\beta}_0$ | 34.17 | .83 | .87 |
| $\hat{\beta}_1$ | -.0574 | .0043 | .0045 |

Table 9.3. Results of fitting model (9.21) to the hormone data.

| | Estimate | \hat{s}_e | \bar{s}_e |
|-----------------|----------|-------------|-------------|
| $\hat{\beta}_A$ | 32.13 | .69 | .75 |
| $\hat{\beta}_B$ | 36.11 | .89 | .97 |
| $\hat{\beta}_C$ | 35.60 | .60 | .66 |
| $\hat{\beta}_1$ | -.0601 | .0032 | .0035 |

Большинство программ для линейной регрессии с библиотеками обычно выводят результат $\hat{s}_e(\hat{\beta}_j)$ вместе с оценкой $\hat{\beta}_j$ методом наименьших квадратов. Применение такой программы к модели (9.11) для данных по гормону дает результаты в таблице 9.2.

Глядя на правую часть рисунка 9.1, большинство точек для партии A лежат ниже подобранной линии регрессии, в то время как большинство точек для партий B и C лежат выше этой линии. Это говорит о неточности модели (9.11). Если бы модель была точной, можно было бы ожидать, что примерно половина каждой партии будет лежать выше, а половина ниже установленной линии. Выражаясь обычной терминологией, похоже, что в данных присутствует эффект партии.

В нашу линейную модель легко включить эффект партии. Мы предполагаем, что условное математическое ожидание y при заданных L и z имеет вид

$$E(y|L, z) = \beta_L + \beta_1 z. \quad (9.21)$$

Здесь β_L равно одному из трех возможных значений: $\beta_A, \beta_B, \beta_C$, в зависимости от партии устройства. Это похоже на модель (9.11), за исключением того, что (9.21) допускает разные точки пересечения для каждой партии, а не одну точку пересечения β_0 из (9.11). Анализ модели (9.21) методом наименьших квадратов дал результаты в таблице 9.3.

Обратите внимание, что $\hat{\beta}_A$ на несколько стандартных ошибок меньше чем $\hat{\beta}_B$ и $\hat{\beta}_C$, что указывает на то, что устройства в партии A содержат значительно меньше гормона.

9.4 Применение бутстрепа

Пока ни один из расчетов не требует бутстрепа. Однако полезно выполнить бутстреп-анализ для модели линейной регрессии. Оказывается, оценки стандартной ошибки бутстрепа такие же, как $\hat{s}_e(\hat{\beta}_j)$, (9.20). Убедившись, что бутстреп

дает разумные ответы в случае, который мы можем проанализировать математически, мы можем продолжить применять бутстреп к более общим моделям регрессии, которые не имеют математического решения: где функция регрессии нелинейна по параметрам β , и где мы используем методы подбора, отличные от метода наименьших квадратов.

Вероятностная модель $P \rightarrow \mathbf{x}$ для линейной регрессии, как описано в (9.4), (9.5), состоит из двух компонентов:

$$P = (\beta, F), \quad (9.22)$$

где β — вектор параметров коэффициентов регрессии, а F — распределение вероятностей ошибок. Общий алгоритм бустрепа на рис. 8.3 требует, чтобы мы оценили P . У нас уже есть доступная $\hat{\beta}$ — оценка методом наименьших квадратов для β . Как мы можем оценить F ? Если предположить, что β известно, мы могли бы вычислить ошибки $\varepsilon_i = y_i - \mathbf{c}_i\beta$ для $i = 1, 2, \dots, n$ и оценить F по их эмпирическому распределению. Мы не знаем β , но можем использовать $\hat{\beta}$ для вычисления аппроксимаций ошибок

$$\hat{\varepsilon}_i = y_i - \mathbf{c}_i\hat{\beta}, \text{ для } i = 1, 2, \dots, n. \quad (9.23)$$

($\hat{\varepsilon}_i$ также называют *остатками*.) Очевидная оценка F — это эмпирическое распределение $\hat{\varepsilon}_i$,

$$\hat{F} : \text{вероятность } 1/n \text{ для } \hat{\varepsilon}_i \text{ при } i = 1, 2, \dots, n. \quad (9.24)$$

Обычно \hat{F} будет иметь математическое ожидание 0, как требуется в (9.5).

Имея $\hat{P} = (\hat{\beta}, \hat{F})$, мы знаем, как рассчитать бутстреп наборы данных для модели линейной регрессии, в этом случае вероятностный механизм $\hat{P} \rightarrow \mathbf{x}^*$ должен означать то же самое, что и вероятностный механизм $P \rightarrow \mathbf{x}$, дающий фактический набор данных \mathbf{x} , см. (9.4), (9.5). Чтобы сгенерировать \mathbf{x}^* , мы сначала выбираем случайную выборку бустреп ошибок

$$\hat{F} \rightarrow (\varepsilon_1^*, \varepsilon_2^*, \dots, \varepsilon_n^*) = \boldsymbol{\varepsilon}^*. \quad (9.25)$$

Каждый ε_i^* равен любому из n значений $\hat{\varepsilon}_j$ с вероятностью $1/n$. Затем бустреп ответы y_i^* генерируются согласно (9.4),

$$y_i^* = \mathbf{c}_i\hat{\beta} + \varepsilon_i^* \text{ для } i = 1, 2, \dots, n. \quad (9.26)$$

Читатель должен убедиться, что формулы (9.24), (9.25), (9.26) есть то же самое, что (9.4), (9.5), за исключением того, что $\hat{P} = (\hat{\beta}, \hat{F})$ заменяет $P = (\beta, F)$. Обратите внимание, что $\hat{\beta}$ — фиксированная величина в (9.26), имеющая одинаковые значения для всех i .

Бутстреп набор данных \mathbf{x}^* представляет из себя $(\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_n^*)$, где $\mathbf{x}_i^* = (\mathbf{c}_i, y_i^*)$. Может показаться странным, что векторы признаков \mathbf{c}_i для бутстреп данных такие же, как и для фактических данных. Это происходит потому, что мы рассматриваем \mathbf{c}_i как фиксированные величины, а не как случайные. (Во всех наших примерах размер выборки n трактовался одинаково.) Этот момент дополнительно обсуждается ниже.

Бустреп оценка $\hat{\beta}^*$ методом наименьших квадратов является минимизатором квадратичной остаточной ошибки для бустреп данных,

$$\sum_{i=1}^n (y_i^* - \mathbf{c}_i\hat{\beta}^*)^2 = \min_{\mathbf{b}} (y_i^* - \mathbf{c}_i\mathbf{b})^2. \quad (9.27)$$

Нормальные уравнения (9.10), примененные к бутстреп данным, дают

$$\hat{\beta}^* = (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{y}^*. \quad (9.28)$$

В этом случае нам не нужны симуляции Монте–Карло, чтобы вычислить бутстреп стандартные ошибки для компонентов $\hat{\beta}^*$. Несложный расчет дает выражение в явной форме для $\text{se}_{\hat{F}}(\hat{\beta}_j^*) = \widehat{\text{se}}_\infty(\hat{\beta}_j)$, идеальной оценки бутстреп стандартной ошибки:

$$\begin{aligned} \text{var}(\hat{\beta}^*) &= (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \text{var}(\mathbf{y}^*) \mathbf{C} (\mathbf{C}^T \mathbf{C})^{-1} = \\ &= \hat{\sigma}_F^2 (\mathbf{C}^T \mathbf{C})^{-1}, \end{aligned} \quad (9.29)$$

поскольку $\text{var}(\mathbf{y}^*) = \hat{\sigma}_F^2 \mathbf{I}$, где \mathbf{I} — единичная матрица. Следовательно

$$\widehat{\text{se}}_\infty(\hat{\beta}_j) = \hat{\sigma}_F \sqrt{G_{jj}}. \quad (9.30)$$

Другими словами, бутстреп оценка стандартной ошибки для $\hat{\beta}_j$ такая же, как и обычная оценка $\widehat{\text{se}}(\hat{\beta}_j)$, (9.20).

9.5 Бутстреп-пары против бутстреп-остатков

Читатель, возможно, заметил интересный факт, теперь у нас есть два различных способа построения бутстреп регрессионной модели. Метод, описанный в главе 7, выбирает пары $\mathbf{x}_i = (\mathbf{c}_i, y_i)$, так что бутстреп набор данных \mathbf{x}^* имел форму

$$\mathbf{x}^* = \{(\mathbf{c}_{i_1}, y_{i_1}), (\mathbf{c}_{i_2}, y_{i_2}), \dots, (\mathbf{c}_{i_n}, y_{i_n})\}, \quad (9.31)$$

для i_1, i_2, \dots, i_n в случайной выборке целых чисел от 1 до n . Обсуждаемый в этой главе метод (9.24), (9.25), (9.26) можно назвать «бутстрепом остатков». Он создает бутстреп наборы данных в форме

$$\mathbf{x}^* = \{(\mathbf{c}_1, \mathbf{c}_1 \hat{\beta} + \hat{\varepsilon}_{i_1}), (\mathbf{c}_2, \mathbf{c}_2 \hat{\beta} + \hat{\varepsilon}_{i_2}), \dots, (\mathbf{c}_n, \mathbf{c}_n \hat{\beta} + \hat{\varepsilon}_{i_n})\}. \quad (9.32)$$

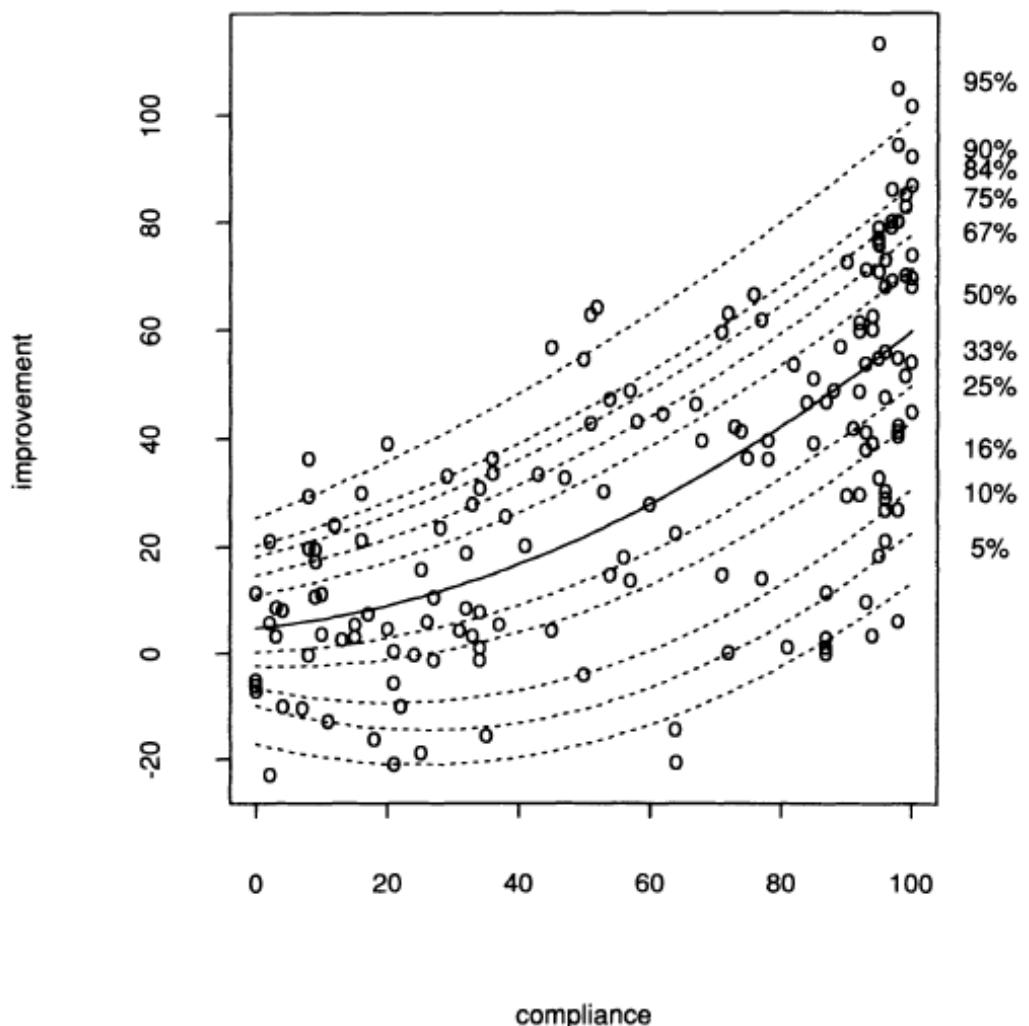


Figure 9.2. Regression percentiles for the cholestyramine data of Figure 7.5; for example the curve labeled “75%” approximates the conditional 75th percentile of the Improvement y given the Compliance z , plotted as a function of z . The percentile curves are twice as far apart at $z = 100$ as at $z = 0$. The linear regression model (9.4), (9.5) can't be correct for this data set. (Regression percentiles calculated using asymmetric maximum likelihood, Efron, 1991.)

Какой бутстреп метод лучше? Ответ зависит от того, насколько мы доверяем модели линейной регрессии (9.4). Эта модель говорит, что разность между y_i и его средним значением $\mu_i = \mathbf{c}_i\boldsymbol{\beta}$ не зависит от \mathbf{c}_i ; он имеет одинаковое распре-

деление « F » независимо от \mathbf{c}_i . Это — сильное предположение, которое может оказаться неверным, даже если модель математического ожидания $\mu_i = \mathbf{c}_i\boldsymbol{\beta}$ верна. Это не соответствует данным про холостирамин на рис. 7.4.

На рисунке 9.2 показаны *процентили регрессии* для данных про холостирамин. Например, кривая, обозначенная «75%», аппроксимирует условный 75-й процентиль показателя улучшения y как функцию показателя соответствия z . Вблизи любого заданного значения z около 75% нанесенных на график точек лежат ниже кривой. Модель (9.4), (9.5) предсказывает, что эти кривые будут находиться на одинаковом расстоянии друг от друга для всех значений z . Вместо этого кривые расходятся по мере увеличения z , находясь вдвое дальше друг от друга при $z = 100$, чем при $z = 0$. Другими словами, ошибки ε_i в (9.4) стремятся быть вдвое больше при $z = 100$, чем при $z = 0$.

Бутстреп-пары менее чувствительны к предположениям, чем бутстреп-остатки. Стандартная оценка ошибки, полученная с помощью бутстреп-пар (9.31), дает разумные результаты, даже если (9.4), (9.5) полностью неверны. Единственное предположение, стоящее за (9.31), состоит в том, что исходные пары $\mathbf{x}_i = (\mathbf{c}_i, y_i)$ были случайным образом выбраны из некоторого распределения F , где F — распределение на $(p + 1)$ -мерных векторах (\mathbf{c}, y) . Даже если (9.4), (9.5) верны, нет ничего плохого в бутстреп-парах, как показано в (9.31); можно показать, что ответ (9.31) приближается к ответу (9.32) по мере увеличения числа пар n . Простая модель для данных о гормонах (9.12) была повторно проанализирована методом бутстреп-пар. $B = 800$ бутстреп-репликаций дали

$$\hat{s}e_{800}(\hat{\beta}_0) = 0.77 \quad \hat{s}e_{800}(\hat{\beta}_1) = 0.0045, \quad (9.33)$$

что не сильно отличается от результатов в таблице 9.2.

Можно привести и обратный аргумент. Модель (9.4), (9.5) не обязательно должна выполняться идеально, чтобы бутстреп остатков, как в (9.32), давал разумные результаты. Более того, различия в распределении ошибок, как и в данных о холостирамине, могут быть включены в модель (9.4), (9.5), что приведет к более подходящей версии бутстреп-остатков; см. модель (9.42). Возможно, наиболее важным моментом здесь является то, что бутстреп не является однозначно определенной концепцией. Рисунок 8.3 может быть реализован по-разному для одной и той же задачи, в зависимости от того, как интерпретируется вероятностная модель $P \rightarrow \mathbf{x}$.

Когда мы осуществляем бутстреп остатков, бутстреп наборы данных $\mathbf{x}^* = \{(\mathbf{c}_1, y_1^*), (\mathbf{c}_2, y_2^*), \dots, (\mathbf{c}_n, y_n^*)\}$ имеют векторы признаков $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n$ в точности такие же, как и для фактического набора данных \mathbf{x} . Это кажется неестественным для данных о гормонах, где \mathbf{c}_i включает z_i , затраченное количество часов, которое является такой же случайной величиной, как и переменная ответа y_i , оставшееся количество гормона.

Даже когда признаки генерируются случайным образом, есть причины проводить анализ так, как если бы они были фиксированными. Коэффициенты регрессии имеют большую стандартную ошибку, когда признаки имеют меньшее стандартное отклонение. Рассматривая признаки как фиксированные константы, мы получаем стандартную ошибку, которая отражает точность, связанную с выборкой фактически наблюдаемых признаков. Однако, как показывает (9.33), разница между \mathbf{c}_i фиксированной и \mathbf{c}_i случайной обычно не сильно влияет на оценку стандартной ошибки.

9.6 Пример: данные о выживаемости клеток

Бывают случаи в регрессии, когда признаки более естественно считать фиксированными, а не случайными. Данные по выживаемости клеток в таблице 9.4 показывают такую ситуацию. Радиолог провел эксперимент с 14 бактериальными пластиинами. Пластиинки подвергали воздействию различных доз радиации и измеряли долю выживших клеток. Как и следовало ожидать, более высокие дозы приводят к меньшей выживаемости. Знак вопроса после ответа на пластине 13 отражает некоторую неуверенность в этом результате, выраженную исследователем.

Исследователя интересовал регрессионный анализ с переменной предиктором

$$\text{доза}_i = z_i \quad i = 1, 2, \dots, 14 \quad (9.34)$$

и переменной ответом

$$\log(\text{пропорция выживания})_i = y_i \quad i = 1, 2, \dots, 14. \quad (9.35)$$

Были доступны две различные теоретические модели радиационного поражения, одна из которых предсказывала линейную регрессию

$$\mu_i = E(y_i|z_i) = \beta_1 z_i, \quad (9.36)$$

а другая квадратичную регрессию

$$\mu_i = E(y_i|z_i) = \beta_1 z_i + \beta_2 z_i^2. \quad (9.37)$$

В (9.36) или (9.37) нет пересекающих членов β_0 , потому что мы знаем, что нулевая доза дает коэффициент выживаемости 1, $y = \log(1) = 0$.

В таблице 9.5 показаны оценки по методу наименьших квадратов ($\hat{\beta}_1, \hat{\beta}_2$) и их оценочные стандартные ошибки $\text{se}(\hat{\beta}_j)$, (9.20). Представлены два анализа методом наименьших квадратов, один с данными для всех 14 пластиин, другой за исключением сомнительной пластины 13. В обоих анализах оцененный коэффициент квадратичной регрессии $\hat{\beta}_2$ является положительным. Является ли это отличие значимым? Другими словами, можем ли мы заключить, что $\hat{\beta}_2$ останется положительным, если будет исследовано гораздо больше пластиин? Отношение

Table 9.4. The Cell Survival data. Fourteen cells were exposed to 14 different levels of radiation. The observed response was the proportion of cells which survived the radiation exposure. This proportion was considered somewhat uncertain by the investigator.

| plate number | dose (rads/100) | survive prop |
|-----------------|--------------------|-----------------|
| 1 | 1.175 | 0.4400 |
| 2 | 1.175 | 0.5500 |
| 3 | 2.350 | 0.1600 |
| 4 | 2.350 | 0.1300 |
| 5 | 4.700 | 0.0400 |
| 6 | 4.700 | 0.0196 |
| 7 | 4.700 | 0.0612 |
| 8 | 7.050 | 0.0050 |
| 9 | 7.050 | 0.0032 |
| 10 | 9.400 | 0.0011 |
| 11 | 9.400 | 0.00015 |
| 12 | 9.400 | 0.00019 |
| 13 | 14.100 | 0.00700 |
| 14 | 14.100 | 0.00000 |

$\hat{\beta}_2/\hat{s}e(\hat{\beta}_2)$ помогает ответить на этот вопрос.

Отношение составляет 2.46 для анализа, основанного на всех 14 пластинах, что обычно считается убедительным доказательством того, что $\hat{\beta}_2$ значительно больше нуля. Если верить этому результату, то квадратичная модель (9.37) сильно предпочтительнее модели (9.36), которая имеет $\beta_2 = 0$.

Однако удаление сомнительной пластины 13 из анализа снижает $\hat{\beta}_2/\hat{s}e(\hat{\beta}_j)$ только до 0.95, что является незначимым результатом. Вывод заключается не в том, что β_2 *обязательно* равен нулю, а в том, что он легко может быть равен нулю: если $\beta_2 = 0$, и если $(\beta_2) \doteq 0.0091$, как в строке 2 таблицы 9.5, то это вовсе не удивительно, что значение $\hat{\beta}_2$ такое же большое или больше наблюдаемого значения 0.0086. Таким образом, у нас нет убедительных доказательств для отказа от линейной модели в пользу квадратичной модели.

Статистика — это наука о сборе информации по крупицам с целью получения высоконформативных сложных результатов. Статистики настораживаются, когда видят, что один элемент выборки, особенно подозрительный, доминирует в ответе на важный вопрос. Действительная критика регрессии по методу наименьших квадратов состоит в том, что один удаленный элемент, такой как пластина 13, может иметь слишком большое влияние на подобранный кривую регрессии. Это показано на рисунке 9.3, на котором построена кривая регрессии методом наименьших квадратов как с пластиной 13, так и без нее. Мощный эф-

Table 9.5. Estimated regression coefficients and standard errors for the quadratic model (9.37) applied to the cell survival data. Least squares estimates (9.10) were obtained using all 14 plates (line 1), excluding plate 13 (line 2). Estimated standard errors for lines 1 and 2 are $\bar{se}(\hat{\beta}_j)$, (9.20). The estimated standard errors for the least median of squares regression (all 14 plates), line 3, were obtained from a bootstrap analysis, $B = 400$. The quadratic coefficient looks significantly different in line 1, but not in lines 2 or 3. Line 4 gives the standard error for the least median of squares estimate, based on resampling residuals for the model (9.42).

| | $\hat{\beta}_1$ | (\hat{se}) | $\hat{\beta}_2$ | (\hat{se}) |
|-----------------------------|-----------------|----------------|-----------------|----------------|
| 1. Least Squares, 14 plates | -1.05 | (.159) | .0341 | (.0143) |
| 2. Least Squares, 13 plates | -0.86 | (.094) | .0086 | (.0091) |
| 3. Least Median of Squares | -0.83 | (.272) | .0114 | (.0362) |
| 4. (Resampling residuals) | | (.141) | | (.0160) |

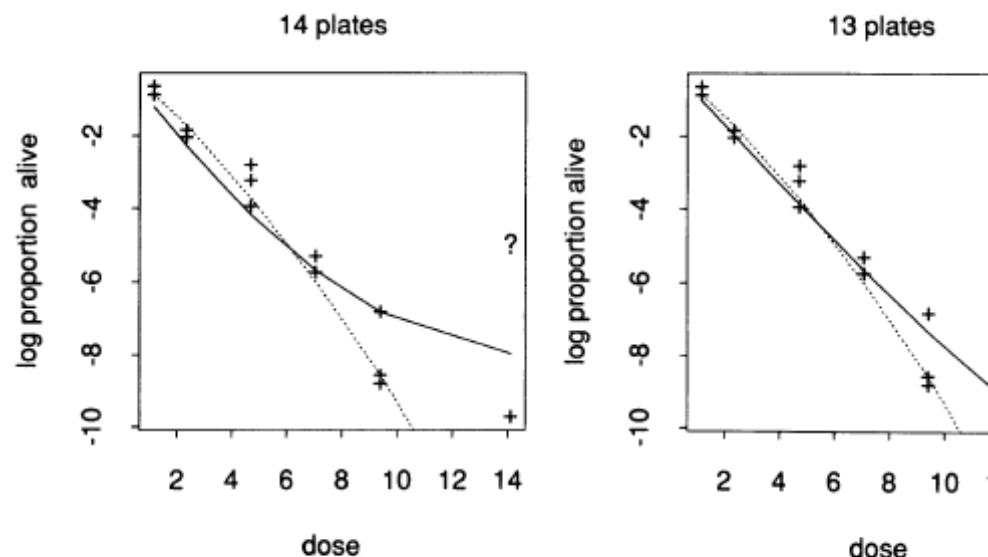


Figure 9.3. Scatterplot of the cell survival data; solid line is the quadratic regression $\hat{\beta}_1 z + \hat{\beta}_2 z^2$ obtained by least-squares. Dashed line is quadratic regression fit by method of least median of squares (LMS). Left panel: all 14 plates; Right panel: thirteen plates, excluding the questionable point from plate 13. Plate 13, marked "?" in the left panel, has a large effect on the fitted least-squares curve. The questionable point has no effect on the LMS curve.

фект точки «?» очевиден.

Даже если бы исследователь не подвергал сомнению достоверность пластинки 13, мы бы предпочли, чтобы наши подогнанные кривые не зависели так сильно от отдельных элементов выборки.

9.7 Наименьшая медиана квадратов

Наименьшая медиана квадратов регрессии, сокращенно LMS, является менее чувствительным методом подбора, чем метод наименьших квадратов. Единственное различие между методом наименьших квадратов и LMS — это выбор критерия соответствия. Чтобы обосновать критерий, давайте разделим остаточную квадратичную ошибку (9.7) на размер выборки, получив среднеквадратичные остатки

$$\frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{c}_i \mathbf{b})^2. \quad (9.38)$$

Минимизация (9.38), очевидно, то же самое, что минимизация (9.7). Средние выборок чувствительны к влияющим значениям, а медианы — нет. Следовательно, чтобы сделать (9.38) менее чувствительным, мы можем заменить среднее значение на медиану, получив *медиану квадратичных остатков*

$$\text{MSR}(\mathbf{b}) = \text{med}(y_i - \mathbf{c}_i \mathbf{b})^2. \quad (9.39)$$

Оценка LMS $\boldsymbol{\beta}$ — это значение $\hat{\boldsymbol{\beta}}$, минимизирующее $\text{MSR}(\mathbf{b})$,

$$\text{MSR}(\hat{\boldsymbol{\beta}}) = \min_{\mathbf{b}} [\text{MSR}(\mathbf{b})]. \quad (9.40)$$

Обратите внимание, что разница между методом наименьших квадратов и LMS заключается не в выборе модели, которая остается (9.3), а в том, как мы измеряем расхождения между моделью и наблюдаемыми данными. $\text{MSR}(\mathbf{b})$ менее чувствителен, чем $\text{RSE}(\mathbf{b})$, к удаленным точкам данных. Это можно увидеть на рис. 9.3, где, по-видимому, очень мало различий между квадратичной LMS-аппроксимацией с точкой «?» или без нее. На самом деле разницы нет. Расчетные коэффициенты регрессии равны $(\hat{\beta}_1, \hat{\beta}_2) = (-0.81, 0.0088)$ в обоих случаях.

Можно показать, что *порог* (breakdown) оценки LMS составляет примерно 50%. Порог оценщика — это наименьшая часть данных, которая может иметь сколь угодно большое влияние на его значение. Другими словами, оценщик имеет порог α , если по крайней мере $t = \alpha \cdot n$ точек данных будут «плохими», прежде чем он разделит. Высокий порог — это хорошо, при этом 50% — это наибольшее значение, которое имеет смысл (если $\alpha > 50%$, неясно, какие из них являются хорошими, а какие плохими). Например, среднее значение выборки имеет порог $1/n$, поскольку, изменяя только одно значение данных, мы можем заставить среднее значение выборки принимать любое значение. Медиана выборки имеет порог 50%, что отражает тот факт, что она менее чувствительна к отдельным значениям. Оценщик регрессии наименьших квадратов наследует чувствительность среднего и имеет порог $1/n$, в то время как оценщик наименьших средних квадратов, как и медиана, имеет порог примерно 50%.

Насколько точны LMS оценки $\hat{\beta}_1, \hat{\beta}_2$? Нет четкой формулы, подобной (9.20) для стандартных ошибок LMS. (Нет четкой формулы для самих оценок LMS.

Они вычисляются с использованием алгоритма выборки с возвращением.) Стандартные ошибки в таблице 9.5 были получены методами бутстрепа. Стандартные ошибки в строке 3 основаны на парах выборок без возвращения, как в разделе 7.3. Бутстреп набор данных имеет форму $\mathbf{x}^* = ((\mathbf{c}_1^*, y_1^*), (\mathbf{c}_2^*, y_2^*), \dots, (\mathbf{c}_n^*, y_n^*))$, как в (9.31), где $\mathbf{c}_i = (z_i, z_i^2)$. После генерации \mathbf{x}^* была получена бутстреп репликация $\hat{\beta}^*$ для вектора регрессии LMS как минимизатор медианы квадратичных остатков для бутстреп данных, то есть минимизатор по \mathbf{b} для

$$\text{med}(y_i^* - \mathbf{c}_i^* \mathbf{b})^2 \quad (9.41)$$

$B = 400$ репликаций бутстрепа дают оценочные стандартные ошибки в строке 3 таблицы 9.5. Обратите внимание, что $\hat{\beta}_2$ не намного больше нуля.

Признаками в данных выживаемости клеток были фиксированные числа, установленные исследователем: она выбрала дозы

$$1.175, 1.175, 2.35, \dots, 14.100,$$

чтобы провести хороший эксперимент по различению линейной и квадратичной моделей радиационной выживаемости. Это заставляет нас больше интересоваться бутстреп-остатками (9.32), нежели бутстреп-парами. Тогда бутстреп наборы данных \mathbf{x}^* будут иметь те же векторы признаков $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n$, которые исследователь намеренно использовал в эксперименте.

Модель (9.4), (9.5) не совсем подходит для данных о выживаемости клеток. Глядя на рисунок 9.3, мы видим, что зависимая переменная y_i более рассеяна при больших значениях z . Это похоже на ситуацию с холостиарином на рис. 9.2, за исключением того, что у нас недостаточно точек для построения хороших процентилей регрессии. Грубо говоря, мы будем предполагать, что ошибки линейной модели линейно возрастают с дозой z . Это равносильно замене (9.4) на

$$y_i = \mathbf{c}_i \boldsymbol{\beta} + z_i \varepsilon_i \text{ для } i = 1, 2, \dots, 14. \quad (9.42)$$

Мы по-прежнему предполагаем, что $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ — случайная выборка из некоторого распределения F , (9.5). Для модели квадратичной регрессии $\mathbf{c}_i = (z_i, z_i^2)$.

Модель вероятности для (9.42), как и раньше, равна $P = (\boldsymbol{\beta}, F)$; $\boldsymbol{\beta}$ было оценено при помощи LMS, $\hat{\boldsymbol{\beta}} = (-0.83, 0.0114)$. Затем F было оценено с помощью \hat{F} , эмпирического распределения величин $(y_i - \mathbf{c}_i \hat{\boldsymbol{\beta}} / z_i)$, $i = 1, 2, \dots, 14$.

Строка 4 таблицы 9.5 сообщает о бутстреп стандартных ошибках для оценок медиан квадратичных остатков $\hat{\beta}_1$ и $\hat{\beta}_2$, полученных из $B = 200$ бутстреп репликаций, с бутстреп-остатками в модели (9.42). Стандартные ошибки заметно меньше, чем при бутстрепе пар. (Но недостаточно мал, чтобы сделать $\hat{\beta}_2$ значимо отличным от нуля.) Стандартные ошибки в строке 4 следует рассматривать с осторожностью, поскольку данные модели (9.42) лишь делают слабое предположение. Самым важным в представлении модели было проиллюстрировать, как бутстреп остатков может быть выполнен в ситуациях, более сложных, чем (9.4).

9.8 Библиографические примечания

Регрессия обсуждается в большинстве текстов по элементарной статистике, и есть много книг, посвященных этой теме, в том числе Драпер and Смит (1981) и

Вайсберг (1980). Бутстреп регрессионных моделей обсуждается на более глубоком математическом уровне в работах Фридмана (1981), Шорака (1982), Бикеля и Фридмана (1983), Вебера (1984), Ву (1986) и Шао (1988). Фридман и Петерс (1984), Петерс и Фридман (1984a, 1984b) рассмотрели некоторые практические аспекты. Русью (1984) вводит оценку наименьшей медианы квадратов. Эфрон (1991) обсуждает оценку процентилей регрессии.

Глава 10

Оценки смещения

10.1 Введение

Мы сосредоточились на стандартной ошибке как на показателе точности оценки $\hat{\theta}$. Существуют и другие пригодные меры статистической точности (или статистической ошибки), измеряющие различные аспекты поведения оценок $\hat{\theta}$. В этой главе речь идет о смещении, разнице между ожидаемым значением оценки $\hat{\theta}$ и оцениваемой величиной θ . Алгоритм бутстрепа легко адаптируется для получения оценок смещения, ровно как и для получения оценок стандартной ошибки. Также вводится оценка смещения по методу складного ножа, хотя мы откладываем полное обсуждение метода складного ножа до главы 11. Можно использовать оценку смещения для исправленной оценки смещения. Однако это может быть опасной практикой, о чем говорится в конце главы.

10.2 Бутстреп оценка смещения

Для начала предположим, что мы вернулись к ситуации с непараметрической выборкой, как в главе 6. Неизвестное распределение вероятностей F дает набор $x = (x_1, x_2 \dots, x_n)$ путем случайной выборки $F \rightarrow x$. Мы хотим оценить вещественный параметр $\theta = t(F)$. Пока возьмем за оценку любую статистику $\hat{\theta} = s(x)$, как показано на рисунке 6.1. Позже нас будет особенно интересовать оценка плагина $\hat{\theta} = t(\hat{F})$.

Смещение $\hat{\theta} = s(x)$ как оценка θ определяется как разница между математическим ожиданием $\hat{\theta}$ и значением параметра θ ,

$$\text{bias}_F = \text{bias}_F(\hat{\theta}, \theta) = E_F[s(x)] - t(F). \quad (10.1)$$

Большое смещение обычно является нежелательным аспектом поведения оценки. Мы смирились с тем фактом, что $\hat{\theta}$ является непостоянной оценкой θ , но обычно мы не хотим, чтобы изменчивость была исключительно низкой или высокой. Несмешенные оценки те, для которых $E_F(\hat{\theta}) = \theta$, играют важную роль в статистической теории и практике. Они способствуют хорошему чувству научной объективности в процессе оценки. Оценки плагина $\hat{\theta} = t(\hat{F})$ необязательно являются несмешенными, но они, как правило, имеют небольшие смещения по сравнению с величиной их стандартных ошибок. Это одна из хороших черт принципа плагина.

Мы можем использовать бутстреп чтобы вычислить смещение любой оценки $\hat{\theta} = s(x)$. Бутстреп оценка смещения определяется как оценка смещения,

которую мы получаем, подставляя \hat{F} вместо F в 10.1,

$$\text{bias}_{\hat{F}} = E_{\hat{F}}[s(x^*)] - t(\hat{F}). \quad (10.2)$$

Здесь $t(\hat{F})$ — оценка плагина θ может отличаться от $\hat{\theta} = s(x)$. Другими словами, $\text{bias}_{\hat{F}}$ — это оценка плагина bias_F независимо от того, является ли $\hat{\theta}$ оценкой плагина θ или нет. Обратите внимание, что \hat{F} используется дважды при переходе от 10.1 к 10.2: при замене F в $t(F)$ и при замене F в $E_F[s(x)]$.

Если $s(x)$ — среднее значение, а $t(F)$ — среднее значение генеральной совокупности, легко показать, что $\text{bias}_{\hat{F}} = 0$. Это имеет смысл, потому что среднее — это несмещенная оценка среднего для генеральной совокупности, то есть $\text{bias}_F = 0$. Однако обычно статистика имеет некоторую систематическую ошибку, и $\text{bias}_{\hat{F}}$ дает оценку этой систематической ошибки. Простым примером является выборочная дисперсия $s(x) = \sum_1^n (x_i - \bar{x})^2 / n$, погрешность которой равна $(-1/n)$ дисперсии генеральной совокупности. В этом случае легко показать, что $\text{bias}_{\hat{F}} = (-1/n^2) \sum_1^n (x_i - \bar{x})^2$.

Для большинства статистик, которые возникают на практике, идеальная бутстреп оценка $\text{bias}_{\hat{F}}$ должна быть аппроксимирована моделированием Монте-Карло. Мы генерируем независимую бутстреп выборку $x^{*1}, x^{*2}, \dots, x^{*B}$ как на рисунке 6.1, вычисляем бутстреп репликации $\hat{\theta}^*(b) = s(x^{*b})$ и аппроксимируем бутстреп математическое ожидание $E_{\hat{F}}[s(x^*)]$ средним

$$\hat{\theta}^*(\cdot) = \sum_{b=1}^B \hat{\theta}^*(b)/B = \sum_{b=1}^B s(x^{*b})/B. \quad (10.3)$$

Бутстреп оценка смещения, основанная на B репликах $\widehat{\text{bias}}_B$, есть 10.2 с заменой $E_{\hat{F}}[s(x^*)]$ на $\hat{\theta}^*(\cdot)$,

$$\widehat{\text{bias}}_B = \hat{\theta}^*(\cdot) - t(\hat{F}). \quad (10.4)$$

Обратите внимание, что алгоритм, показанный на рисунке 6.1, точно применяется к вычислению 10.4, за исключением того, что на последнем шаге мы вычисляем $\hat{\theta}^*(\cdot) - t(\hat{F})$, а не $\widehat{s\epsilon}_B$. Конечно, мы можем вычислить как $\widehat{s\epsilon}_B$, так и $\widehat{\text{bias}}_B$ с помощью того же набора бутстреп репликаций.

10.3 Пример: данные об уровне гормона при ношении различных пластырей

Исторически сложилось так, что статистики очень беспокоились о возможных смещениях в оценках соотношений. Данные об уровне гормона в таблице 10.1 представляют удобный пример. Восемь субъектов носили медицинские пластыри, предназначенные для введения в кровоток определенного природного гормона. У каждого испытуемого измеряли уровень гормона в крови после ношения трех разных пластырей: пластиря с плацебо, не содержащего гормона, «старого» пластиря, произведенного на более старом заводе, и «нового» пластиря, произведенного на недавно открывшемся заводе. Первые три столбца таблицы показывают три измерения показателей крови для каждого субъекта.

Table 10.1. The patch data. Eight subjects wore medical patches designed to increase the blood levels of a certain natural hormone. Each subject had his blood levels of the hormone measured after wearing three different patches: a placebo patch, which had no medicine in it, an “old” patch which was from a lot manufactured at an old plant, and a “new” patch, which was from a lot manufactured at a newly opened plant. For each subject, $z = \text{oldpatch} - \text{placebo measurement}$, and $y = \text{newpatch} - \text{oldpatch measurement}$. The purpose of the experiment was to show that the new plant was producing patches equivalent to those from the old plant. Chapter 25 has an extended analysis of this data set.

| subject | placebo | oldpatch | newpatch | old-plac. | new-old |
|---------|---------|----------|----------|-----------|---------|
| 1 | 9243 | 17649 | 16449 | 8406 | -1200 |
| 2 | 9671 | 12013 | 14614 | 2342 | 2601 |
| 3 | 11792 | 19979 | 17274 | 8187 | -2705 |
| 4 | 13357 | 21816 | 23798 | 8459 | 1982 |
| 5 | 9055 | 13850 | 12560 | 4795 | -1290 |
| 6 | 6290 | 9806 | 10157 | 3516 | 351 |
| 7 | 12412 | 17208 | 16570 | 4796 | -638 |
| 8 | 18806 | 29044 | 26325 | 10238 | -2719 |
| mean: | | | | 6342 | -452.3 |

Целью эксперимента с пластырем было показать биоэквивалентность. Пластиры, изготовленные на старом заводе, уже были одобрены для продажи Управлением по санитарному надзору за качеством пищевых продуктов и медикаментов (FDA). Пластиры с нового завода не потребовали полного нового исследования FDA. Их одобрили бы к продаже, если бы можно было доказать, что они биоэквивалентны тем, что были изготовлены на старом заводе. Критерий биоэквивалентности FDA заключается в том, что ожидаемая эффективность новых пластырей соответствует ожидаемой эффективности старых пластырей в том смысле, что

$$\frac{|E(\text{new}) - E(\text{old})|}{E(\text{old}) - E(\text{placebo})} \leq .20. \quad (10.5)$$

Другими словами, FDA хочет, чтобы новое лекарство соответствовало старому в пределах 20% количества гормона, которое старый препарат добавляет к «плацебо» уровню крови.

Пусть θ параметр

$$\theta = \frac{|E(\text{new patch}) - E(\text{old patch})|}{E(\text{old patch}) - E(\text{placebo patch})}. \quad (10.6)$$

В главах 12–14 рассматриваются доверительные интервалы для θ , подход, который приводит к полному ответу на вопрос о биоэквивалентности: «действитель-

но ли $|\theta| \leq 0.20?$.»¹ Здесь мы рассматриваем только смещение и стандартную ошибку оценки плагина $\hat{\theta}$.

Нас интересуют две статистики, z_i и y_i , полученные для каждого из восьми субъектов,

$$z = \text{oldpatch measurement} - \text{placebo measurement} \quad (10.7)$$

и

$$y = \text{newpatch measurement} - \text{oldpatch measurement}. \quad (10.8)$$

Предполагая, что пары $x_i = (z_i, y_i)$ получены путем случайной выборки из неизвестного двумерного распределения F , $F \rightarrow x = (x_1, x_2 \dots x_8)$, тогда θ в 10.6 это параметр

$$\theta = t(F) = \frac{E_F(y)}{E_F(z)}. \quad (10.9)$$

В этом случае $t(\cdot)$ является функцией, которая принимает распределение вероятностей F на парах $x = (z, y)$ и выдает отношение математических ожиданий. Оценка плагина θ равна

$$\hat{\theta} = t(\hat{F}) = \frac{\bar{y}}{\bar{z}} = \frac{\sum_{i=1}^8 y_i/8}{\sum_{i=1}^8 z_i/8}, \quad (10.10)$$

которую мы возьмем за нашу оценку $\hat{\theta} = s(X)$. Обратите внимание, что ничто в этих формулировках не предполагает, что z и y независимы друг от друга. Последние два столбца таблицы 10.1 показывают z_i и y_i для восьми испытуемых. Значение $\hat{\theta}$ равно

$$\hat{\theta} = \frac{-452.3}{6342} = -0.0713. \quad (10.11)$$

Мы видим, что $|\hat{\theta}|$ значительно меньше 0.20, так что есть некоторая надежда на выполнение условия биоэквивалентности FDA.

На рисунке 10.1 показана гистограмма $B = 400$ бутстреп репликаций $\hat{\theta}$, полученных как в (6.1–6.2): бутстреп выборки $x^* = (x_1^*, x_2^*, \dots, x_8^*) = (x_{i_1}, x_{i_2}, \dots, x_{i_8})$ дают бутстреп репликации

$$\hat{\theta}^* = \frac{\bar{y}^*}{\bar{z}^*} = \frac{\sum_{j=1}^8 y_{i_j}/8}{\sum_{j=1}^8 z_{i_j}/8}. \quad (10.12)$$

¹ В главе 25 представлен расширенный анализ биоэквивалентности этого набора данных.

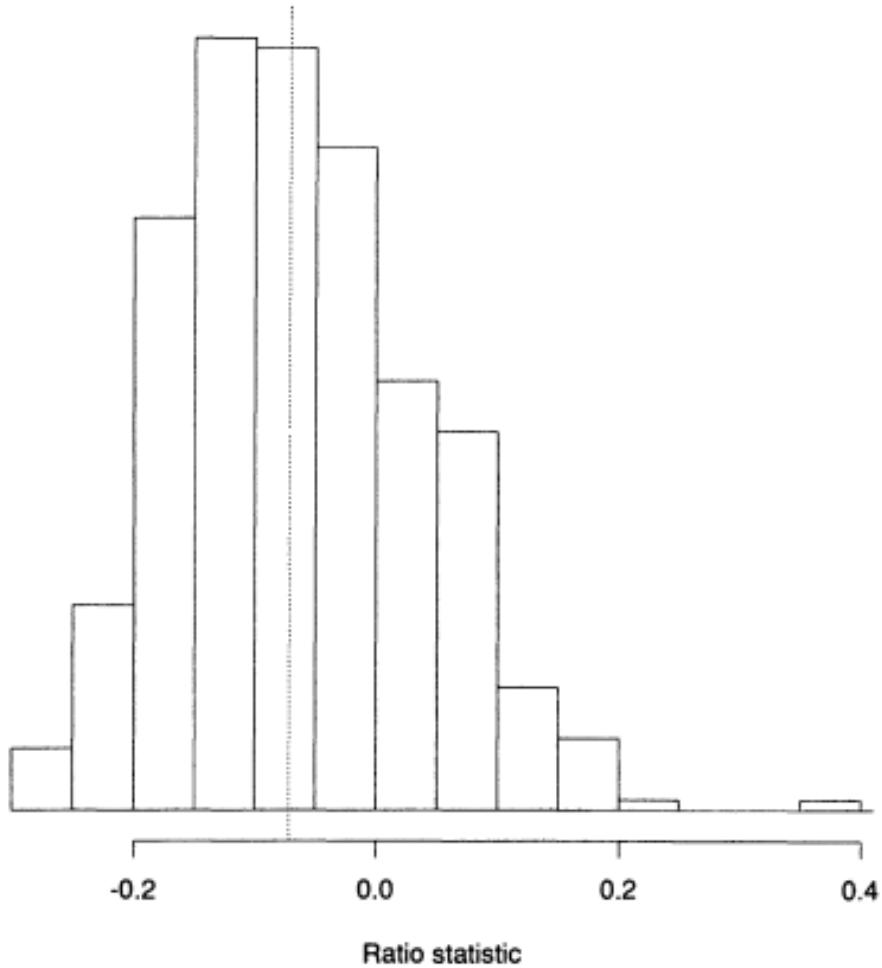


Figure 10.1. $B = 400$ bootstrap replications of the ratio statistic (10.10), $\hat{\theta} = \bar{y}/\bar{z}$, for the patch data of Table 10.1. The dashed line indicates $\hat{\theta} = -.0713$. The 400 replications had standard deviation $\widehat{se}_{400} = .105$ and mean $\hat{\theta}^*(\cdot) = -.0670$, so the bootstrap bias estimate was $\widehat{bias}_{400} = .0043$.

Для 400 повторов стандартное отклонение выборки составило $\widehat{se}_{400} = 0.105$, а среднее значение выборки $\hat{\theta}^*(\cdot) = -0.0670$. Бутстреп оценка смещения составляет

$$\widehat{bias}_{400} = -0.0670 - (-0.0713) = 0.0043. \quad (10.13)$$

Это вычисление основано на формуле 10.4, с использованием того факта, что в данном случае $\hat{\theta} = t(\hat{F})$.

Отношение оцененного смещения к стандартной ошибке $\widehat{bias}_{400}/\widehat{se}_{400} = 0.041$ мало, что указывает на то, что в этом случае нам не нужно беспокоиться о смещении $\hat{\theta}$. Как правило, смещение меньшее чем 0.25 стандартных ошибок можно игнорировать, если мы не пытаемся провести аккуратные вычисления доверительных интервалов. Среднеквадратичная ошибка оценки $\hat{\theta}$ для θ есть $\sqrt{E_F[(\hat{\theta} - \theta)^2]}$ мера точности, которая учитывает как смещение, так и стандартную ошибку. Можно показать, что значение среднеквадратичной ошибки равно

$$\sqrt{E_F[(\hat{\theta} - \theta)^2]} = \sqrt{se_F(\hat{\theta})^2 + bias_F(\hat{\theta}, \theta)^2} =$$

$$= \text{se}_F(\hat{\theta}) \cdot \sqrt{1 + \left(\frac{\text{bias}_F}{\text{se}_F} \right)^2} \doteq \text{se}_F(\hat{\theta}) \cdot \left[1 + \left(\frac{\text{bias}_F}{\text{se}_F} \right)^2 \right]. \quad (10.14)$$

Если $\text{bias}_F = 0$, то среднеквадратичная ошибка равна её минимальному значению se_F . Если $|\text{bias}_F/\text{se}_F| < 0.25$, тогда среднеквадратичная ошибка не превосходит se_F больше, чем примерно на 3.1%.

Мы знаем, что $B = 400$ бутстреп репликаций обычно более чем достаточно для получения хорошей оценки стандартной ошибки. Достаточно ли этого, чтобы получить хорошую оценку смещения? Ответ в данном конкретном случае — нет. Помните, что в определении идеальной бутстреп оценки смещения $\widehat{\text{bias}}_\infty = \text{bias}_{\hat{F}}$ 10.2, bias_B 10.4 заменяет $E_{\hat{F}}(\hat{\theta}^*)$ на $\hat{\theta}^*(\cdot)$. По распределению бутстреп репликаций мы можем сказать, насколько хорошо $\hat{\theta}^*(\cdot)$ оценивает $E_{\hat{F}}(\hat{\theta}^*)$. Применение 5.6 дает

$$\text{Prob}_{\hat{F}} \left\{ |\hat{\theta}^*(\cdot) - E_{\hat{F}}\{\hat{\theta}^*\}| < 2 \frac{\widehat{\text{se}}_B}{\sqrt{B}} \right\} = \text{Prob}_{\hat{F}} \left\{ |\widehat{\text{bias}}_B - \widehat{\text{bias}}_\infty| < 2 \frac{\widehat{\text{se}}_B}{\sqrt{B}} \right\} \doteq 0.95, \quad (10.15)$$

где $\widehat{\text{se}}_B$ — бутстреп оценка стандартной ошибки. Для бутстреп данных на рисунке 10.1 с $\widehat{\text{se}}_B = 0.105$ и $B = 400$, мы получаем

$$\text{Prob}_{\hat{F}} \left\{ |\widehat{\text{bias}}_B - \widehat{\text{bias}}_\infty| < 0.0105 \right\} \doteq 0.95, \quad (10.16)$$

большой диапазон погрешности по сравнению с рассчитанным значением $\widehat{\text{bias}}_{400} = 0.0043$.

Граница ошибки 0.0105 в 10.16 достаточно мала, чтобы показать, что смещение здесь не является большой проблемой: так как $\widehat{\text{bias}}_{400} = 0.0043$, мы, вероятно, имеем $|\widehat{\text{bias}}_\infty| < 0.0043 + 0.0105 = 0.0148$ и поэтому $|\text{bias}|/\widehat{\text{se}} < 0.0148/0.106 = 0.14$. Что довольно меньше, чем эмпирическое граница 0.25. Однако нам все еще может быть интересно узнать $\widehat{\text{bias}}_\infty$ или хорошее приближение к нему и вычислениям 10.16, показывающим что $\text{bias}_{400} = 0.0043$, нельзя доверять. Мы могли бы просто увеличить B (смотри задачу 10.5), но в этом нет необходимости.

10.4 Улучшенная оценка смещения

Оказывается, что существует лучший метод, чем 10.4, для аппроксимации $\widehat{\text{bias}}_\infty = \text{bias}_{\hat{F}}$ из B бутстреп репликаций. Улучшенный метод применяется, когда $\hat{\theta}$ — это плагин оценка $t(\hat{F})$ для $\theta = t(F)$. Мы описываем метод здесь и даем объяснение, почему он работает, в главе 23.

Нам нужно определить понятие вектора повторной выборки (???resampling vector). Пусть P_j^* указывает долю бутстреп наблюдений $x^* = (x_1^*, x_2^*, \dots, x_n^*)$, которая равна j -ому исходному наблюдению,

$$P_j^* = \#\{x_j^* = x_j\}/n, \quad j = 1, 2, \dots, n. \quad (10.17)$$

Вектор повторной выборки

$$P^* = (P_1^*, P_2^*, \dots, P_n^*) \quad (10.18)$$

имеет неотрицательные компоненты в сумме дающие единицу. Например, третья бутстреп выборка для данных пластиря была

$$X^* = (x_1, x_6, x_6, x_5, x_7, x_1, x_3, x_8)$$

и соответствующий вектор повторной выборки

$$P^* = (2/8, 0, 1/8, 0, 1/8, 2/8, 1/8, 1/8).$$

Бутстреп репликацию $\hat{\theta}^* = s(x^*)$ можно рассматривать как функцию вектора повторной выборки P^* . Например, если $\hat{\theta} = \bar{y}/\bar{z}$ в 10.10,

$$\hat{\theta}^* = \bar{y}^*/\bar{z}^* = \frac{\sum_{i=1}^8 P_j^* y_i / 8}{\sum_{i=1}^8 P_j^* z_i / 8}. \quad (10.19)$$

(Обратите внимание, что исходные данные x считаются фиксированными в этом определении; единственными случайными величинами являются P_j^* .) Для $\hat{\theta} = t(\hat{F})$, плагин оценки θ , запишем

$$\hat{\theta}^* = T(P^*) \quad (10.20)$$

чтобы определить $\hat{\theta}^*$ как функцию вектора повторной выборки.² Формула 10.19 определяет $T(\cdot)$ для $\hat{\theta} = \bar{y}/\bar{z}$.

Пусть P^0 обозначает вектор длины n , все элементы которого равны $1/n$,

$$P^0 = (1/n, 1/n, \dots, 1/n). \quad (10.21)$$

Значение $T(P^0)$ — это значение $\hat{\theta}^*$, когда каждый элемент $P_j^* = 1/n$, то есть когда каждая точка исходных данных x_j встречается ровно один раз в бутстреп выборке x^* . Это означает, что $x^* = x$, за исключением, возможно, перестановок порядка, в котором появляются элементы x_1, x_2, \dots, x_n . Но статистика вида $\hat{\theta} = t(\hat{F})$ не меняется, когда элементы $x = (x_1, x_2, \dots, x_n)$ переупорядочиваются, потому что F не изменяется. Другими словами,

$$T(P^0) = \hat{\theta} = t(\hat{F}), \quad (10.22)$$

наблюдаемое выборочное значение статистики. (Это легко проверить в 10.19.)

B бутстреп выборок $x^{*1}, x^{*2}, \dots, x^{*B}$ приводят к соответствующим векторам повторной выборки $P^{*1}, P^{*2}, \dots, P^{*B}$, каждый вектор P^{*b} имеет форму 10.18. Определим \bar{P}^* как среднее значение этих векторов

$$\bar{P}^* = \sum_{i=1}^B P^{*b} / B. \quad (10.23)$$

Согласно 10.22 мы можем записать бутстреп оценку смещения 10.4 в виде

$$\widehat{\text{bias}}_B = \hat{\theta}^*(\cdot) - T(P^0). \quad (10.24)$$

Улучшенная бутстреп оценка смещения, которую мы обозначим как $\overline{\text{bias}}_B$, равна

$$\overline{\text{bias}}_B = \hat{\theta}^*(\cdot) - T(\bar{P}^*). \quad (10.25)$$

Для рисунка 10.1 четыреста векторов повторной выборки усреднены до $\bar{P}^* = (0.1178, 0.1187, 0.1$. Это приводит к

$$T(\bar{P}^*) = \frac{\sum_{i=1}^8 \bar{P}_j^* y_i}{\sum_{i=1}^8 \bar{P}_j^* z_i} = -0.0750 \quad (10.26)$$

²Мы обозначаем плагин статистику двумя способами $\hat{\theta} = s(x) = t(\hat{F})$. Аналогично бутстреп репликации обозначаются $\hat{\theta}^* = s(x^*) = T(P^*)$. Три функции $s(\cdot), t(\cdot)$ и $T(\cdot)$ представляют одну и ту же статистику, но рассматриваются как функция в трех разных пространствах.

и

$$\overline{\text{bias}}_{400} = -0.0670 - (-0.0750) = 0.0080, \quad (10.27)$$

в отличие от $\widehat{\text{bias}}_{400} = 0.0043$.

Обе оценки bias_B и $\widehat{\text{bias}}_B$ сходятся к $\widehat{\text{bias}}_\infty = \text{bias}_{\hat{F}}$, идеальной бутстреп оценке смещения, когда B стремится к бесконечности. Для $\overline{\text{bias}}_B$ сходимость проходит намного быстрее, поэтому мы назвали её улучшенной. Более быстрая сходимость очевидна на рисунке 10.2, на котором рассмотрены $\widehat{\text{bias}}_B$ и $\overline{\text{bias}}_B$ для B , равного 25, 50, 100, 200, 400, 800, 1670, 3200. Предельное значение $\widehat{\text{bias}}_\infty$ было аппроксимировано $\widehat{\text{bias}}_{100000} = 0.0079$, показанным пунктирной горизонтальной линией. $\widehat{\text{bias}}_B$ плавно и быстро приближается к пунктирной линии, в то время как $\overline{\text{bias}}_B$ все еще довольно изменчива даже для $B = 3200$.

В главе 23 обсуждаются улучшенные вычислительные бутстреп методы. Там будет показано, что $\overline{\text{bias}}_B$ сводится к использованию $\widehat{\text{bias}}_{CB}$, где C — большая константа, часто 50 или больше. Задача 10.7 предлагает одну причину превосходства $\overline{\text{bias}}_B$.

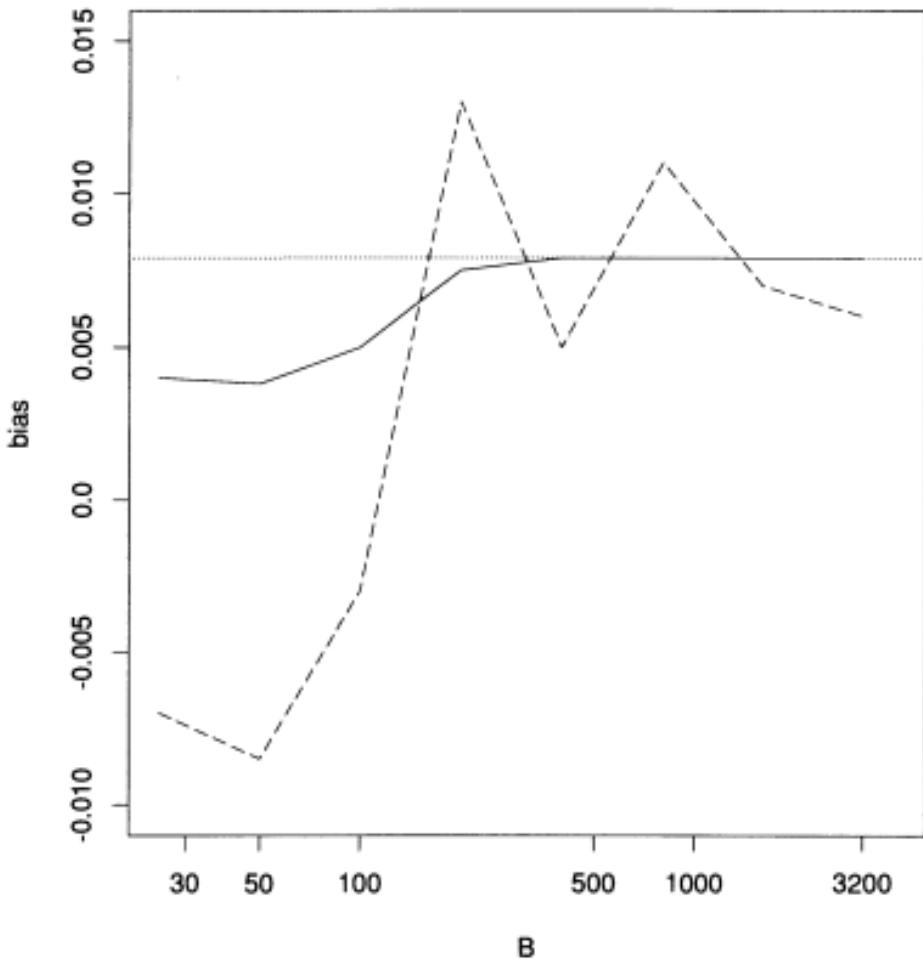


Figure 10.2. The bootstrap bias estimate $\widehat{\text{bias}}_B$ broken line, and the better bootstrap bias estimate $\overline{\text{bias}}_B$, solid line, for $B = 25, 50, 100, \dots, 3200$; log scale for B ; dotted line is $\widehat{\text{bias}}_{100,000} = .0079$. We see that $\overline{\text{bias}}_B$ converges much faster than $\widehat{\text{bias}}_B$ to the limiting ideal bootstrap estimate $\widehat{\text{bias}}_\infty = \text{bias}_{\hat{F}}$.

10.5 Оценка смещения по методу складного ножа

Складной нож был оригинальным компьютерным методом оценки смещения и стандартных ошибок. Оценка смещения методом складного ножа, которая кратко обсуждается здесь и более подробно в главе 11, была предложена Моррисом Кенуйем в середине 1950-х годов. При наличии набора данных $x = (x_1, x_2, \dots, x_n)$, i -я реализация складного ножа $x_{(i)}$ определяется как x с удаленной i -й точкой наблюдений,

$$x_{(i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n), \quad (10.28)$$

для $i = 1, 2, \dots, n$. I -я репликация складного ножа $\hat{\theta}_{(i)}$ статистики $\theta = s(x)$ это $s(\cdot)$, вычисленная для $x_{(i)}$, предположим

$$\hat{\theta}_{(i)} = s(x_{(i)}) \quad \text{для } i = 1, 2, \dots, n. \quad (10.29)$$

Для плагин статистики $\hat{\theta} = t(\hat{F})$, $\hat{\theta}_{(i)}$ равна $t(\hat{F}_{(i)})$, где $\hat{F}_{(i)}$ — эмпирическое распределение $n - 1$ точек в $x_{(i)}$.

Оценка смещения складного ножа определяется как

$$\widehat{\text{bias}}_{\text{jack}} = (n - 1)(\hat{\theta}_{(.)} - \hat{\theta}), \quad (10.30)$$

где

$$\hat{\theta}_{(.)} = \sum_{i=1}^n \hat{\theta}_{(i)}/n. \quad (10.31)$$

Эта формула применяется только к плагин статистике $\hat{\theta} = t(\hat{F})$. Формула не работает, если $t(\hat{F})$ — негладкая статистика, такая как медиана, но для гладкой статистики, такой как $\hat{\theta} = \bar{y}/\bar{z}$ (тех статистик, для которых функция $T(P^*)$ в 10.20 дважды дифференцируема) она дает оценку смещения с помощью всего n повторных вычислений функции $t(\cdot)$. Это сравнивается с B повторными вычислениями для будстреп оценок, где B должно быть не менее 200 даже для $\widehat{\text{bias}}_B$.

Table 10.2. Jackknife values for the patch data

| $\hat{\theta}_{(1)}$ | $\hat{\theta}_{(2)}$ | $\hat{\theta}_{(3)}$ | $\hat{\theta}_{(4)}$ | $\hat{\theta}_{(5)}$ | $\hat{\theta}_{(6)}$ | $\hat{\theta}_{(7)}$ | $\hat{\theta}_{(8)}$ |
|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| -.0571 | -.1285 | -.0215 | -.1325 | -.0507 | -.0840 | -.0649 | -.0222 |

Для данных об уровне гормона статистика отношения $\hat{\theta} = \bar{z}/\bar{y} = -0.0713$ (формула 10.10), репликации складного ножа показаны в таблице 10.2. Это приводит к $\hat{\theta}_{(.)} = -0.0702$, и

$$\widehat{\text{bias}}_{\text{jack}} = 7\{-0.0702 - (-0.0713)\} = 0.0080. \quad (10.32)$$

Не случайно, что $\widehat{\text{bias}}_{\text{jack}}$ так близко согласуется с идеальной будстреп оценкой $\widehat{\text{bias}}_\infty = \widehat{\text{bias}}_{\hat{F}}$. В главе 20 показано, что $\widehat{\text{bias}}_{\text{jack}}$ — это приближение плагин оценки смещения рядом Тейлора второго порядка. Важно помнить следующее: все три оценки смещения, bias_B , $\widehat{\text{bias}}_B$ и $\widehat{\text{bias}}_{\text{jack}}$ пытаются аппроксимировать одну и ту же идеальную оценку $\text{bias}_{\hat{F}}$. В главе 20 обсуждается инфинитезимальный складной нож — еще один способ приближенного определения смещения.

Мы также увидим аппроксимации, отличные от \widehat{se}_B , для идеальной оценки стандартной ошибки $se_{\hat{F}}$ (хотя здесь сложнее улучшить прямое приближение Монте-Карло \widehat{se}_B). Во всех методах численной аппроксимации работает только один принцип оценки, подстановка \hat{F} вместо F в любую меру точности, которую мы хотим оценить. Реализация этого принципа численно эффективным способом — важная тема, но современные компьютеры настолько мощны, что даже неэффективные способы обычно достаточно хороши, чтобы дать пригодные ответы.

Идеальная оценка $bias_{\hat{F}}$ имеет недостатки. Если позволить $B \rightarrow \infty$, изменчивость смещения \widehat{bias}_B из-за выборки Монте-Карло устраняется. Однако остается вариабельность $bias_{\infty} = bias_{\hat{F}}$ из-за случайности \hat{F} как оценки F . Другими словами, у нас все еще есть обычные ошибки, связанные с оценкой любого параметра по выборке.

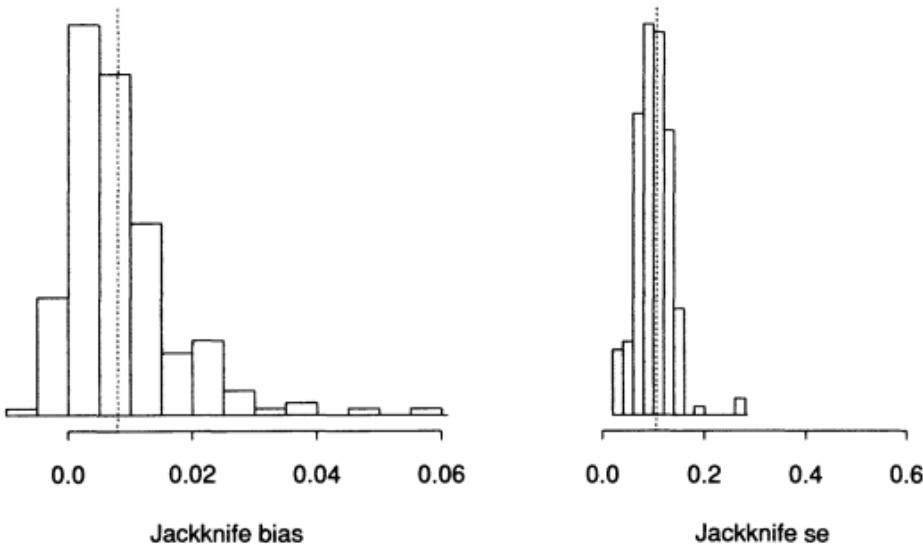


Figure 10.3. *Left panel:* 200 bootstrap replications of the jackknife bias estimate (10.30) for $\hat{\theta} = \bar{y}/\bar{z}$, patch data; dashed line indicates actual estimate $\widehat{bias}_{jack} = .0080$; estimated coefficient of variation for \widehat{bias}_{jack} equals .96; \widehat{bias}_{jack} has low accuracy. *Right panel:* the corresponding 200 bootstrap replications of the jackknife standard error estimate for $\hat{\theta}$, (10.34); dashed line indicates actual estimate $\widehat{se}_{jack} = .106$; scale has been chosen so that 0 and dashed lines match left panel; estimated coefficient of variation is .33; \widehat{se}_{jack} is about 3 times more accurate than \widehat{bias}_{jack} .

Мы могли бы использовать бутстреп для вычисления изменчивости идеальной бутстреп оценки $bias_{\hat{F}}$, как показано на рисунке 6.1, за исключением практических трудностей вычисления статистики $s(x) = bias_{\hat{F}}$. Вместо этого давайте рассмотрим более простую статистику $s(x) = \widehat{bias}_{jack}$, которая для $\hat{\theta} = \bar{y}/\bar{z}$ обычно близка к $bias_{\hat{F}}$. Статистика $s(x) = \widehat{bias}_{jack}$ является сложной функцией от x , требующей сначала вычисления $\hat{\theta}$, затем $\hat{\theta}_{(i)}$ и, наконец, 10.30, но мы все ещё можем использовать бутстреп для оценки стандартной ошибки $s(x)$.

$B = 200$ бутстреп выборок размера $n = 8$ были сгенерированы из данных об уровне гормона, и для каждой выборки была рассчитана оценка смещения по методу складного ножа для статистики отношения, скажем, \widehat{bias}_{jack}^* . Левая

часть рисунка 10.3 представляет собой гистограмму из 200 значений $\widehat{\text{bias}}_{\text{jack}}^*$.

Ясно, что статистика $s(x) = \widehat{\text{bias}}_{\text{jack}}$ сильно варьируется. 200 реплик $s(x^*)$ имели стандартное отклонение 0.0081 и среднее 0.0084, что давало оценку коэффициента вариации

$$\widehat{\text{cv}}(\widehat{\text{bias}}_{\text{jack}}) = 0.0081/0.0084 = 0.96. \quad (10.33)$$

Десять процентов значений $\widehat{\text{bias}}_{\text{jack}}^*$ были меньше нуля и 16% больше $2 \cdot \widehat{\text{bias}}_{\text{jack}} = 0.0160$.

Нет ничего плохого ни в $\widehat{\text{bias}}_{\text{jack}}$, ни в $\widehat{\text{bias}}_F$. Проблема в том, что $n = 8$ точек данных недостаточно для точного определения смещения статистики отношения в этой ситуации. Рисунок 10.3 поясняет это. Вычисления смещения не были пустой тратой времени. Мы достаточно уверены, что истинное смещение $\hat{\theta} = \bar{y}/\bar{z}$, каким бы оно ни было, находится где-то между -0.005 и 0.025. Бутстреп стандартная ошибка $\hat{\theta}$ была 0.105, поэтому отношение абсолютного смещения к стандартной ошибке, вероятно, меньше 0.25. Вычисления 10.14 показывает, что в данном случае систематическая ошибка не вызывает особого беспокойства.

Этот расчет предполагает другое беспокойство. Возможно, бутстреп оценка стандартной ошибки $\widehat{s}\epsilon_{200} = 0.105$ тоже ненадежна. Теоретически мы могли бы провести бутстреп $\widehat{s}\epsilon_{200}$, чтобы выяснить это, но это сложно с вычислительной точки зрения. Однако существует оценка стандартной ошибки по методу складного ножа, предложенная Джоном Тьюки в конце 1950-х годов, которая требует меньше вычислений, чем $\widehat{s}\epsilon_{200}$:

$$\widehat{s}\epsilon_{\text{jack}} = \left[\frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 \right]^{1/2}. \quad (10.34)$$

Эта формула, которая применяется к гладко определенной статистике, такой как $\hat{\theta} = \bar{y}/\bar{z}$, обсуждается в главе 11. Оказывается, это альтернатива $\widehat{s}\epsilon_B$ численной аппроксимации идеальной бутстреп оценки $\widehat{s}\epsilon_\infty = s\epsilon_F(\hat{\theta}^*)$. Статистика отношения данных об уровне гормона 10.2 дает

$$\widehat{s}\epsilon_{\text{jack}} = 0.106, \quad (10.35)$$

почти то же самое, что и $\widehat{s}\epsilon_{200}$. Мы увидим, что $\widehat{s}\epsilon_{\text{jack}}$ не всегда является хорошим приближением к $\widehat{s}\epsilon_\infty$, но для $\hat{\theta} = \bar{y}/\bar{z}$ это вполне приемлемо.

Те же 200 бутстреп выборок, использованные для обеспечения репликации $\widehat{\text{bias}}_{\text{jack}}$ на рисунке 10.3, также дали бутстреп репликации $\widehat{s}\epsilon_{\text{jack}}$. Гистограмма 200 бутстреп значений $\widehat{s}\epsilon_{\text{jack}}$, показанная на правой части рисунка 10.3, указывает на существенную изменчивость, но не такую большую, как для $\widehat{\text{bias}}_{\text{jack}}$. Гистограмма имеет среднее значение 0.099 и стандартное отклонение 0.033, что дает выборочный коэффициент вариации

$$\widehat{\text{cv}}(\widehat{s}\epsilon_{\text{jack}}) = 0.33, \quad (10.36)$$

только треть $\widehat{\text{cv}}(\widehat{\text{bias}}_{\text{jack}})$. На самом деле стандартную ошибку обычно легче оценить, чем смещение, а также она является более важным фактором, определяющим вероятностные характеристики оценки $\hat{\theta}$.

Мы обсудили оценку $\text{bias}_F(\hat{\theta}, \theta)$, уравнение 10.1. Бутстреп процедуру оценки смещения, которая сводится к подстановке \hat{F} вместо F в bias_F , можно обобщить: 1) мы можем рассмотреть общие вероятностные механизмы $P \rightarrow x$, как

на рисунке 8.3. (Обратите внимание, что здесь « P » означает нечто иное, чем вектор повторной выборки P^* , 10.18.) 2) Мы можем рассмотреть общие меры смещения, $\text{Bias}_P(\hat{\theta}, \theta)$, например, медианное смещение

$$\text{Bias}_P(\hat{\theta}, \theta) = \text{median}_P(\hat{\theta}(x)) - \theta(P). \quad (10.37)$$

На рисунке 10.4 показана схема. Идеальная бутстреп оценка $\text{Bias}_P(\hat{\theta}, \theta)$ это плагин оценка

$$\text{Bias}_P(\hat{\theta}^*, \theta(\hat{P})). \quad (10.38)$$

Здесь $\hat{P} \rightarrow x^*$ бутстреп данные; $\hat{\theta}^* = s(x^*)$ бутстреп репликация $\hat{\theta} = s(x)$; и $\theta(\hat{P})$ — значение интересующего параметра $\theta = t(P)$, когда $P = \hat{P}$, механизм оценки вероятности. (Мы не можем писать $\theta(\hat{P}) = \hat{\theta}$, поскольку $t(\cdot)$ может быть другой функцией, отличной от $s(\cdot)$, смотри задачу 10.10.) Для медианного смещения 10.37

$$\text{Bias}_{\hat{P}}(\hat{\theta}^*, \theta(\hat{P})) = \text{median}_{\hat{P}}(\hat{\theta}(x^*)) - \theta(\hat{P}). \quad (10.39)$$

Обычно $\text{Bias}_{\hat{P}}$ нужно аппроксимировать методами Монте-Карло. Усовершенствованные методы, такие как $\overline{\text{bias}}_B$ и $\widehat{\text{bias}}_{\text{jack}}$, обычно недоступны для общих мер смещения, таких как 10.37.

10.6 Поправка на смещение

Зачем нам нужно оценивать смещение $\hat{\theta}$? Обычная причина — исправить $\hat{\theta}$, чтобы она стала менее смещенной. Если $\widehat{\text{bias}}$ является оценкой смещения $\text{bias}_F(\hat{\theta}, \theta)$, то очевидной оценкой с поправкой на смещение является

$$\bar{\theta} = \hat{\theta} - \widehat{\text{bias}} \quad (10.40)$$

Принимая $\widehat{\text{bias}}$ равным $\widehat{\text{bias}}_B = \hat{\theta}^*(\cdot) - \hat{\theta}$, получаем

$$\bar{\theta} = 2\hat{\theta} - \hat{\theta}^*(\cdot). \quad (10.41)$$

(Существует тенденция, неправильная тенденция думать о самой $\hat{\theta}^*(\cdot)$ как о скорректированной на смещение оценке. Обратите внимание, что 10.41 утверждает, что если $\hat{\theta}^*(\cdot)$ больше $\hat{\theta}$, то исправленная оценка $\bar{\theta}$ должна быть меньше $\hat{\theta}$.) Положим $\widehat{\text{bias}} = 0.0080$ для статистики отношения данных об уровне гормона, равной как $\overline{\text{bias}}_{400}$, так и $\widehat{\text{bias}}_{\text{jack}}$, исправленная оценка отношения θ равна

$$\bar{\theta} = -0.0713 - 0.0080 = -0.0793. \quad (10.42)$$

На практике исправление смещения может быть опасно. Даже если $\bar{\theta}$ менее смещена, чем $\hat{\theta}$, она может иметь значительно большую стандартную ошибку. Еще раз, это можно проверить с помощью бутсрепа. Для статистики отношения данных об уровне гормона 200 бутстреп репликаций $\bar{\theta} = \hat{\theta} - \widehat{\text{bias}}_{\text{jack}}$ сравнивались с соответствующими репликациями $\hat{\theta}$. Бутстреп оценки стандартной ошибки $\bar{\theta}$ и $\hat{\theta}$ были почти идентичны, поэтому в этом случае исправление смещения не было опасным.

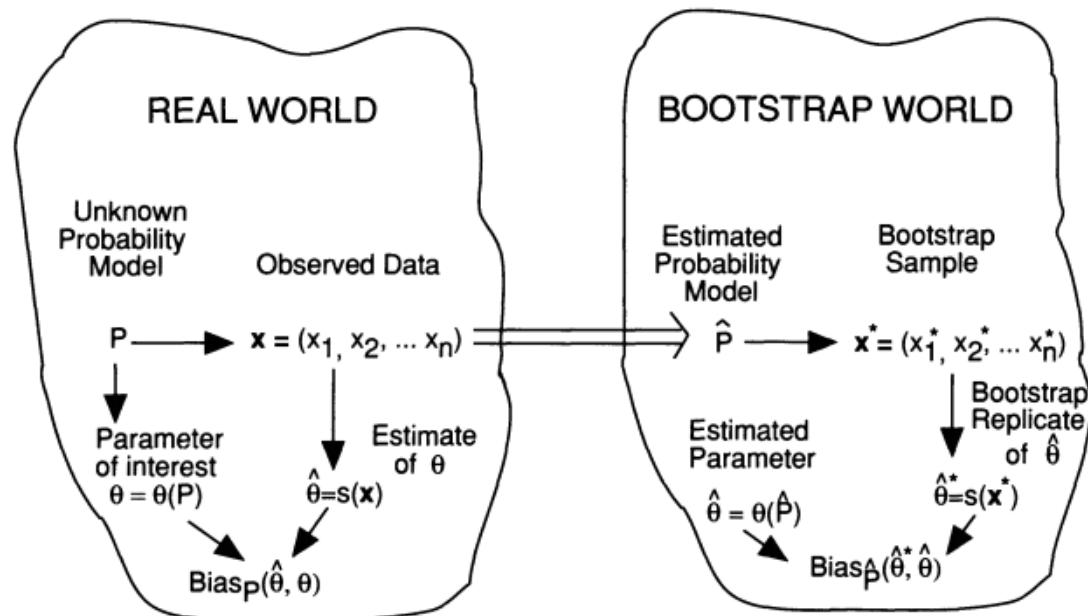


Figure 10.4. Diagram of bootstrap bias estimation in a general framework, an extension of Figure 8.3. $B_{\hat{P}}(\hat{\theta}^*, \theta(\hat{P}))$ is a general bias measure. Usually $Bias_{\hat{P}}(\hat{\theta}^*, \theta(\hat{P}))$ must be approximated by Monte Carlo methods.

Подводя итог, можно сказать, что оценка смещения обычно интересна и целесообразна, но точное использование оценки смещения часто проблематично. Систематические ошибки оценить труднее, чем стандартные ошибки, как показано на рисунке 10.3. Прямое исправление смещения 10.40 может быть опасно для использования на практике из-за большой изменчивости $\widehat{\text{bias}}$. Исправление смещения может вызвать большее увеличение стандартной ошибки, что, в свою очередь, приводит к большей среднеквадратичной ошибке (уравнение 10.14). Если $\widehat{\text{bias}}$ мало по сравнению с предполагаемой стандартной ошибкой $\widehat{\text{se}}$, то безопаснее использовать $\widehat{\theta}$, чем $\bar{\theta}$. Если $\widehat{\text{bias}}$ велико по сравнению с $\widehat{\text{se}}$, то это может указывать на то, что статистика $\hat{\theta} = s(x)$ не является подходящей оценкой параметра θ .

Оценка ошибки предсказания — одна из важных задач, в которой полезно исправление смещения. Смещение очевидной оценки велико по сравнению со стандартной ошибкой, и его можно эффективно уменьшить, добавив поправочный член. Подробности приведены в главе 17.

Глава 11

Метод складного ножа

11.1 Введение

В главе 10 мы упоминаем складной нож — метод оценки смещения и стандартной ошибки оценки. Складной нож появился раньше бутстрепа и имеет близкое сходство с ним. В этой главе мы подробно исследуем метод складного ножа. Некоторые из представленных здесь идей получили дальнейшее развитие в главах 20 и 21.

11.2 Определение складного ножа

Предположим, у нас есть выборка $x = (x_1, x_2, \dots, x_n)$ и оценка $\hat{\theta} = s(x)$. Мы хотим оценить смещение и стандартную ошибку $\hat{\theta}$. Складной нож фокусируется на выборках, которые не учитывают одно наблюдение за раз:

$$x_{(i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \quad (11.1)$$

для $i = 1, 2, \dots, n$, так называемых выборках складного ножа. I -ая выборка складного ножа состоит из набора данных с удаленным i -м наблюдением. Пусть

$$\hat{\theta}_{(i)} = s(x_{(i)}) \quad (11.2)$$

будет i -й репликацией складного ножа $\hat{\theta}$.

Оценка смещения по методу складного ножа определяется как

$$\widehat{\text{bias}}_{\text{jack}} = (n - 1)(\hat{\theta}_{(.)} - \hat{\theta}), \quad (11.3)$$

где

$$\hat{\theta}_{(.)} = \sum_{i=1}^n \hat{\theta}_{(i)} / n. \quad (11.4)$$

Оценка стандартной ошибки по методу складного ножа определяется как

$$\widehat{\text{se}}_{\text{jack}} = \left[\frac{n-1}{n} \sum (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 \right]^{1/2}. \quad (11.5)$$

Откуда берутся эти формулы? Начнем с $\widehat{\text{se}}_{\text{jack}}$. Вместо того, чтобы смотреть на все (или некоторые) наборы данных, которые могут быть получены путем выборки с заменой из x_1, x_2, \dots, x_n , складной нож рассматривает n фиксированных выборок $x_{(1)}, \dots, x_{(n)}$, полученные удалением по одному наблюдению

за раз. Подобно бутстреп оценке стандартной ошибки, формула для $\widehat{s}_{\text{jack}}$ выглядит как стандартное отклонение выборки этих n значений, за исключением того, что первый коэффициент равен $(n-1)/n$ вместо $1/(n-1)$ или $1/n$. Конечно, $(n-1)/n$ намного больше, чем $1/(n-1)$ или $1/n$. Интуитивно этот «коэффициент увеличения» (??? "inflation factor") необходим, потому что отклонения складного ножа

$$(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 \quad (11.6)$$

имеют тенденцию быть меньше, чем бутстреп отклонения

$$[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2, \quad (11.7)$$

поскольку типичная выборка метода складного ножа больше похожа на исходные данные x , чем типичная бутстреп выборка.

Точный вид множителя $(n-1)/n$ получается путем рассмотрения частного случая $\hat{\theta} = \bar{x}$. Тогда легко показать, что

$$\widehat{s}_{\text{jack}} = \left\{ \sum_1^n (x_i - \bar{x})^2 / \{(n-1)n\} \right\}^{1/2}, \quad (11.8)$$

(задача 11.1). То есть коэффициент $(n-1)/n$ — это именно то, что нужно, чтобы сделать $\widehat{s}_{\text{jack}}$ равным несмешенной оценке стандартной ошибки среднего. Коэффициент $[(n-1)/n]^2$ приведет плагин оценке

$$\left\{ \sum_1^n (x_i - \bar{x})^2 / n^2 \right\}^{1/2}, \quad (11.9)$$

но это существенно не отличается от несмешенной оценки, если только n не мало. Соглашение о том, что $\widehat{s}_{\text{jack}}$ использует множитель $(n-1)/n$, несколько произвольно.

Аналогичным образом, оценка смещения по методу складного ножа 11.3 кратна среднему значению отклонений складного ножа

$$\hat{\theta}_{(i)} - \hat{\theta}, \quad i = 1, 2, \dots, n. \quad (11.10)$$

Величины 11.10 иногда называют величинами влияния складного ножа. Обратите внимание на множитель $(n-1)$ в 11.3. Это коэффициент увеличения, аналогичный тому, который появляется при оценке по методу складного ножа стандартной ошибки. Чтобы вывести его, мы не можем обратиться к частному случаю $\hat{\theta} = \bar{x}$, потому что \bar{x} несмешенная, а $\hat{\theta}_{(.)} - \hat{\theta}$, как и должно быть, равно нулю (задача 11.7). Поскольку этот случай не говорит нам, каким должен быть старший фактор, мы вместо этого рассматриваем в качестве нашего тестового примера выборочную дисперсию

$$\hat{\theta} = \sum_1^n (x_i - \bar{x})^2 / n. \quad (11.11)$$

Она имеет смещение $-1/n$ дисперсий генеральной совокупности, а множитель $(n-1)$ перед $\hat{\theta}_{(.)} - \hat{\theta}$ делает $\widehat{\text{bias}}_{\text{jack}}$ равным $-1/n$, умноженному на $\sum(x_i - \bar{x})^2 / (n-1)$, несмешенной оценке дисперсии генеральной совокупности (задача 11.8).

11.3 Пример: данные о тестировании

Давайте применим оценку стандартной ошибки по методу складного ножа к набору данных о результатах тестов 88 студентов, приведенному в таблице 7.1. Напомним, что представляющая интерес статистика — это отношение наибольшего собственного значения ковариационной матрицы к сумме собственных значений, как указано в (7.8)

$$\hat{\theta} = \hat{\lambda}_1 / \sum_1^5 \hat{\lambda}_i. \quad (11.12)$$

Чтобы применить метод складного ножа, мы удаляем по одному каждый случай (строку) в таблице 7.1 и вычисляем $\hat{\theta}$ для каждого набора данных размером 87. На верхней части рисунка 11.1 показана гистограмма 88 значений складного ножа $\hat{\theta}_{(i)}$.

Мы также вычислили 88 бутстреп значений $\hat{\theta}$. Обратите внимание, что разброс гистограммы, полученной с помощью метода складного ножа, намного меньше, чем разброс бутстреп гистограммы, показанной на нижней части рисунка (мы принудительно задаем одну и ту же горизонтальную шкалу во всех гистограммах). Это иллюстрирует тот факт, что наборы данных складного ножа в среднем более похожи на исходный набор данных, чем бутстреп наборы данных. По средне на рисунке показана гистограмма «занышенных» значений складного ножа

$$\sqrt{87}(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}) \quad (11.13)$$

с разрывом в точке среднего складного ножа $\hat{\theta}_{(.)}$. С этим коэффициентом увеличения гистограмма складного ножа похожа на бутстреп гистограмму, показанную на нижней части рисунка. Величина $\widehat{s\epsilon}_{\text{jack}}$ оказывается равной 0.049, что лишь немного больше, чем значение 0.047 для бутстреп оценки, полученной в главе 7.

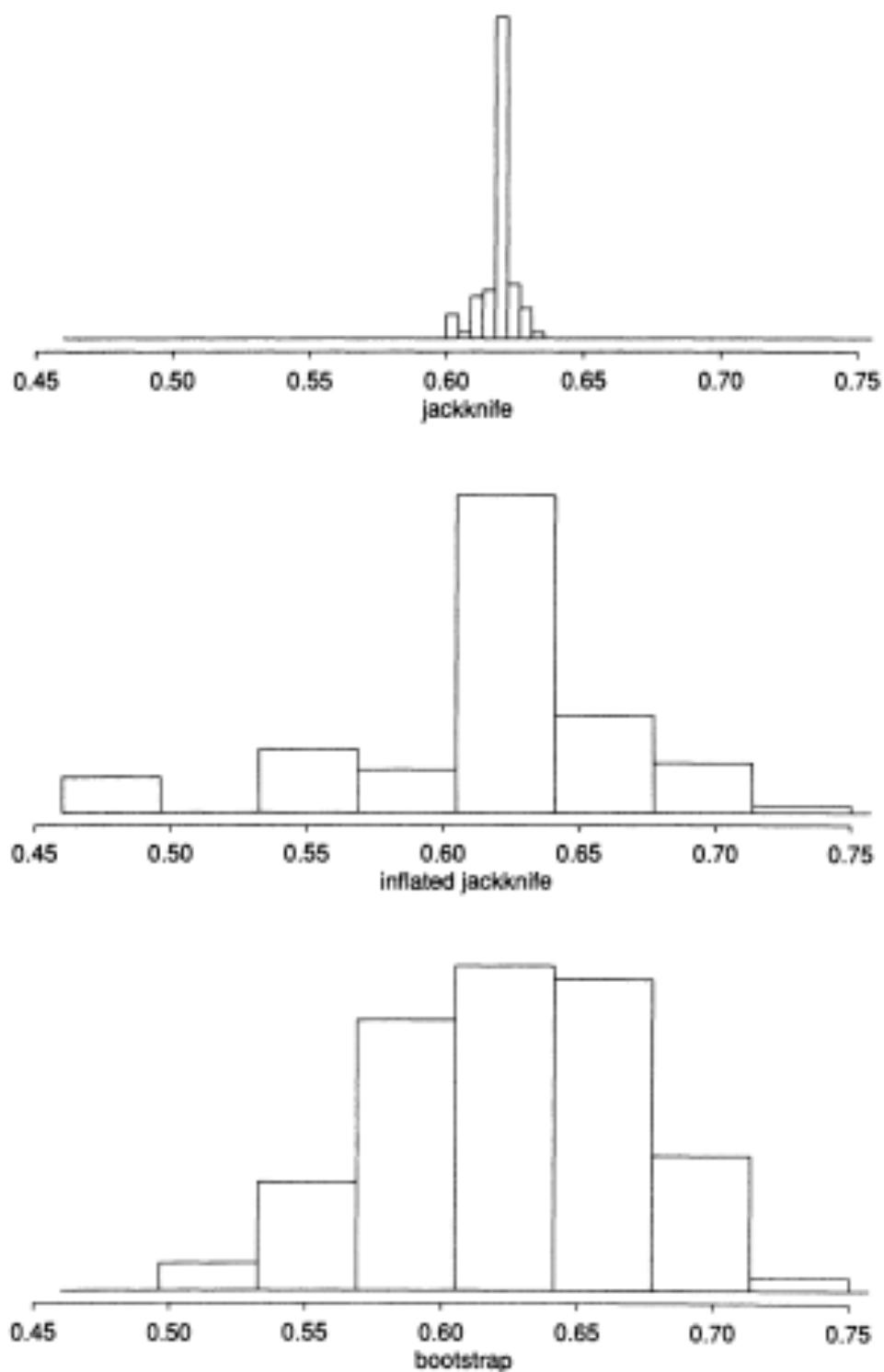


Figure 11.1. Histogram of the 88 jackknife values for the score data of Table 7.1 (top panel); jackknife values inflated by a factor of $\sqrt{87}$ from their mean (middle panel); 88 bootstrap values for the same problem (bottom panel).

11.4 Псевдо-значения

Другой способ думать о складном ноже — это псевдо-значения

$$\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}. \quad (11.14)$$

Обратите внимание, что в частном случае $\hat{\theta} = \bar{x}$, мы имеем $\tilde{\theta}_i = x_i$, i -е значение данных. Кроме того, для любой $\hat{\theta}$ формула для \widehat{se}_{jack} может быть выражена как

$$\widehat{se}_{jack} = \left\{ \sum_1^n (\tilde{\theta}_i - \tilde{\theta})^2 / \{(n-1)n\} \right\}^{1/2}, \quad (11.15)$$

где $\tilde{\theta} = \sum \tilde{\theta}_i / n$. Это похоже на оценку стандартной ошибки среднего для «данных» $\tilde{\theta}_i$, $i = 1, 2, \dots, n$. Идея, лежащая в основе 11.14, состоит в том, что псевдо-значения должны действовать так, как если бы они были n независимыми значениями.

Что произойдет, если мы попытаемся продолжить эту идею и использовать псевдо-значения для построения доверительного интервала? Один из разумных подходов — сформировать интервал

$$\tilde{\theta} \pm t_{n-1}^{(1-\alpha)} \widehat{se}_{jack}, \quad (11.16)$$

где $t_{n-1}^{(1-\alpha)}$ — $(1 - \alpha)$ -й процентиль распределения t с $n - 1$ степенями свободы. Оказывается, этот интервал работает не очень хорошо: в частности, он ненамного лучше, чем более грубые интервалы, основанные на теории о нормальном распределении. Для построения доверительного интервала необходимы более совершенные подходы, как описано в главах 12–14. Хотя псевдо-значения интересуют, неясно, являются ли они хороший способ думать о складном ноже. Мы не будем здесь их рассматривать.

11.5 Связь метода складного ножа и бутстрепа

Что лучше, бутстреп или складной нож? Поскольку для этого требуется вычисление $\hat{\theta}$ только для n наборов данных складного ножа, складной нож будет легче вычислить, если n будет меньше, чем, скажем, 100 или 200 репликаций, используемых бутстрепом для оценки стандартной ошибки. Однако, рассматривая только n выборок складного ножа, складной нож использует только ограниченную информацию о статистике $\hat{\theta}$, и, таким образом, можно предположить, что складной нож менее эффективен, чем бутстреп. Фактически оказывается, что складной нож можно рассматривать как приближение к бутстрепу. Это объясняется в задачах 11.4 и 11.5 и в главе 20. Вот суть идеи: рассмотрим линейную статистику, то есть статистику, которую можно записать в виде

$$\hat{\theta} = s(x) = \mu + \frac{1}{n} \sum_1^n \alpha(x_i), \quad (11.17)$$

где μ — константа, а $\alpha(\cdot)$ — функция. Среднее — это простой пример линейной статистики, для которой $\mu = 0$ и $\alpha(x_i) = x_i$. Теперь для такой статистики, оказывается, что оценка по методу складного ножа и бутстреп оценка стандартных ошибок совпадают, за исключением незначительного множителя $\{(n-1)n\}^{1/2}$, используемого в определении складного ножа. Это именно

то, что мы нашли для $\hat{\theta} = \bar{x}$: складной нож дает оценку стандартной ошибки $\left\{ \sum_1^n (x_i - \bar{x})^2 / \{(n-1)n\} \right\}^{1/2}$ в то время как бутстреп приводит к этому значению, умноженному на $\{(n-1)n\}^{1/2}$. Неудивительно, что для линейной статистики нет потери информации при использовании складного ножа, поскольку знание линейной статистики для n наборов данных складного ножа $x_{(i)}$ определяет значение $\hat{\theta}$ для любого бутстреп набора данных x^* (задача 11.3).

При нелинейной статистике происходит потеря информации. Складной нож линейно аппроксимирует бутстреп: то есть он соглашается с бутстрепом (за исключением множителя $\{(n-1)n\}^{1/2}$) для некоторой линейной статистики вида 11.17, которая приближает $\hat{\theta}$. Детали этой интересной взаимосвязи приведены в задачах 11.5 и 11.6 и в главе 20. С практической точки зрения, эти результаты показывают, что точность оценки стандартной ошибки по методу складного ножа зависит от того, насколько $\hat{\theta}$ близка к линейности. Для сильно нелинейных функций складной нож может быть неэффективным, а иногда и опасным.

На рисунке 11.2 показаны результаты исследования этой неэффективности на конкретном примере. Мы сгенерировали 200 выборок размером 10 из двумерной нормальной совокупности с нулевым средним, единичной дисперсией и корреляцией 0.7. Ящики с усами слева показывают оценки, полученные по методам бутстреп и складного ножа, стандартной ошибки для $\hat{\theta} = \bar{x}$, а справа — для коэффициента корреляции. Горизонтальные линии показывают истинную стандартную ошибку $\hat{\theta}$ в каждом случае. В обоих случаях бутстреп и складной нож демонстрируют небольшое смещение при оценке стандартной ошибки. Вариабельность оценки складного ножа немного больше, чем у бутстрепа для среднего (линейная статистика), но значительно больше для коэффициента корреляции (нелинейная статистика). По этой причине в последнем случае предпочтительнее использовать бутстреп. Задача 11.13 рассматривает бутстреп и метод складного ножа для другой нелинейной статистики.

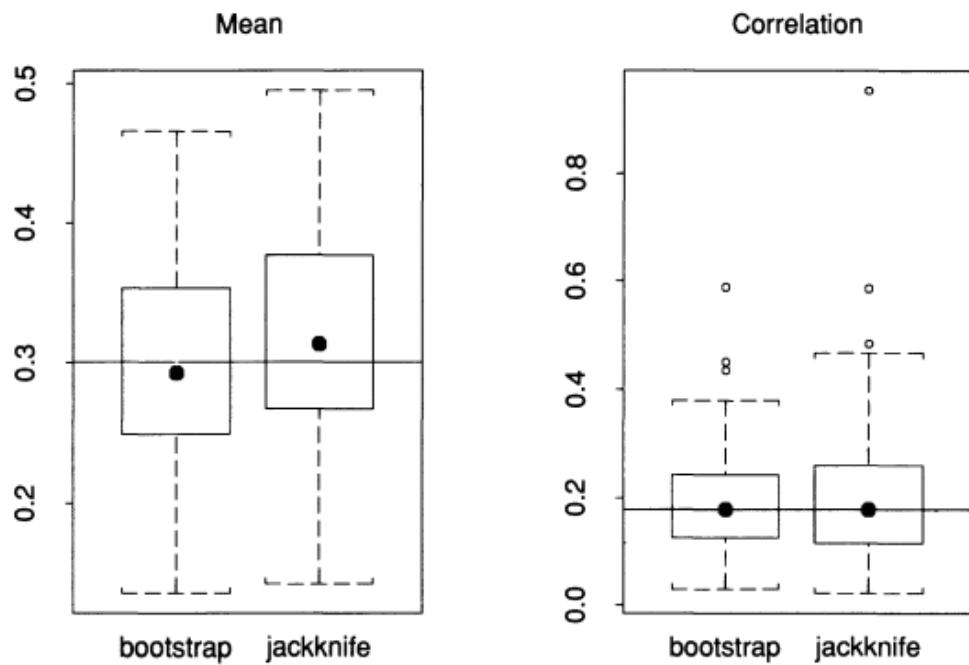


Figure 11.2. Bootstrap and jackknife estimates of standard error for two different statistics $\hat{\theta}$, for samples of size 10 from a bivariate normal population with correlation .7. On the left $\hat{\theta} = \bar{x}$; on the right $\hat{\theta}$ is the sample correlation. Boxplots indicate the distribution of standard error estimates over 100 simulated samples.

Точно так же можно показать, что оценка смещения складного ножа является приближением к начальной оценке смещения. Приближение в терминах квадратичной (а не линейной) статистики, которая имеет вид

$$\hat{\theta} = s(x) = \mu + \frac{1}{n} \sum_{1 \leq i \leq n} \alpha(x_i) + \frac{1}{n^2} \sum_{1 \leq i < j \leq n} \beta(x_i, x_j). \quad (11.18)$$

Простым примером квадратичной статистики является выборочная дисперсия 11.11. Раскрывая ее, мы обнаруживаем, что ее можно выразить в форме уравнения 11.18 (задача 11.9). Для такой статистики, если мы знаем значение $\hat{\theta}$ для x , а также $x_{(i)}, i = 1, 2, \dots, n$, мы можем вывести значение $\hat{\theta}$ для любого бутстреп набора данных. Как показано в задачах 11.10–11.11, оценки смещения складного ножа и бутстрепа по существу совпадают для квадратичной статистики.

11.6 Отказ складного ножа

Подводя итог, можно сказать, что складной нож часто обеспечивает простое и хорошее приближение к бутстрепу для оценки стандартных ошибок и смещения. Однако, как вкратце упоминалось в главе 10, складной нож может с треском выйти из строя, если статистика $\hat{\theta}$ не является «гладкой». Интуитивно идея гладкости заключается в том, что небольшие изменения в наборе данных вызывают только небольшие изменения в статистике. Простым примером

негладкой статистики является медиана. Чтобы понять, почему медиана не является гладкой, рассмотрим 9 упорядоченных значений из контрольной группы данных о мышах (таблица 2.1):

$$10, 27, 31, 40, 46, 50, 52, 104, 146. \quad (11.19)$$

Медиана этих значений равна 46. Теперь предположим, что мы начинаем увеличивать значение 4-го по величине значения $x = 40$. Медиана не меняется вообще, пока x не станет больше 46, а затем после этого медиана будет равна x , пока x не превысит 50. Это означает, что медиана не является дифференцируемой (или гладкой) функцией от x .

Это отсутствие гладкости приводит к тому, что оценка стандартной ошибки по методу складного ножа несовместима с медианой. Для данных о мышах значения складного ножа для медианы¹ равны

$$48, 48, 48, 48, 45, 43, 43, 43, 43. \quad (11.20)$$

Обратите внимание, что встречаются только 3 различных значения, что является следствием недостаточной гладкости медианы и того факта, что наборы данных складного ножа отличаются от исходного набора данных только на одно наблюдение. Итоговая оценка se_{jack} составляет 6.68. Для данных о мышах бутстреп оценка стандартной ошибки на основе бутстреп выборок объема $B = 100$ составляет 9.58, что значительно больше, чем значение складного ножа, равное 6.68. При $n \rightarrow \infty$, можно показать, что se_{jack} противоречива, то есть не может сходиться к истинной стандартной ошибке. С другой стороны, бутстреп рассматривает наборы данных, которые менее похожи на исходный набор данных, чем наборы данных складного ножа, и, следовательно, согласованы с медианой.

11.7 Метод складного ножа с отбрасыванием d наблюдений

Есть способ исправить несоответствие складного ножа негладкой статистике. Вместо того чтобы исключать по одному наблюдению за раз, мы не учитываем d наблюдений, где $n = r \cdot d$ для некоторого целого числа r . Можно показать, что если $n^{1/2}/d \rightarrow \infty$ и $n - d \rightarrow \infty$, то складной нож «с отбрасыванием» согласован с медианой. Грубо говоря, нужно исключить более $d = \sqrt{n}$, но менее n наблюдений, чтобы добиться согласованности в оценке стандартной ошибки складным ножом. Пусть $\hat{\theta}_{(s)}$ обозначает $\hat{\theta}$, примененную к набору данных с удаленным подмножеством s . Формула для оценки стандартной ошибки складным ножом с отбрасыванием d наблюдений:

$$\left\{ \frac{r}{\binom{n}{d}} \sum (\hat{\theta}_{(s)} - \hat{\theta}_{(.)})^2 \right\}^{1/2}, \quad (11.21)$$

где $\hat{\theta}_{(.)} = \sum \hat{\theta}_{(s)} / \binom{n}{d}$ и сумма ведется по всем подмножествам s размера $n - d$, выбранным без замены из x_1, x_2, \dots, x_n .

В нашем примере с $n = 9$ мы можем выбрать $d = 4 > \sqrt{9}$, и вычисление складного ножа delete-d включает в себя нахождение медианы для

$$\binom{9}{4} = 126 \quad (11.22)$$

¹Медиана четного числа точек данных — это среднее двух значений из середины.

выборок, соответствующих одновременному исключению 4 наблюдений. Это дает оценку стандартной ошибки 7.16, что несколько ближе к бутстреп значению 9.58, чем значение складного ножа с удалением одного элемента, которое равно 6.68.

Если n велико и $\sqrt{n} < d < n$, количество выборок складного ножа $\binom{n}{d}$ может быть очень значительным. Вместо вычисления $\hat{\theta}$ для всех этих подмножеств можно вместо этого охватить случайную выборку подмножеств, что, в свою очередь, сделает складной нож delete-d больше похожим на бутстреп. Текущая работа над складным ножом delete-d представляет собой возрождение исследований складного ножа.

Функция складного ножа на языке S описана в приложении.

Глава 14

Улучшенные бутстреп-доверительные интервалы

14.1 Введение

Одной из основных целей теории бутстрепа является автоматическое создание хороших доверительных интервалов. «Хорошо» означает, что бутстреп интервалы должны быть близки к точным доверительным интервалам в тех особых ситуациях, когда статистическая теория дает точный ответ, и должны иметь надежные вероятности покрытия в любых ситуациях. Ни метод бутстреп-*t* главы 12, ни метод процентиелей главы 13 не соответствуют этим критериям. Бутстреп-*t* интервалы имеют хорошие теоретические вероятности покрытия, но на практике имеют тенденцию быть неустойчивыми. Процентильные интервалы более устойчивы, но имеют менее удовлетворительные свойства покрытия.

В этой главе обсуждается улучшенная версия процентильного метода, называемого BC_a (аббревиатура, bias-corrected and accelerated). Интервалы BC_a являются существенным улучшением по сравнению с процентильными интервалами как в теоретическом плане, так и на практике. Они близки к приведенным выше критериям качества, хотя точность их покрытия все еще может быть неустойчивой для небольших размеров выборки. (Возможны улучшения, как показано в главе 25.) Простой компьютерный алгоритм под названием `bcanon` производит интервалы BC_a , затрачивая для этого немного больше усилий, чем для процентильных интервалов. Мы также обсудим метод под названием ABC (аббревиатура, approximate bootstrap confidence intervals) который значительно уменьшает объем вычислений, необходимых для интервалов BC_a . Глава заканчивается применением этих методов к реальной задаче.

14.2 Пример: данные о пространственном восприятии

Следующий пример, основанный на данных о пространственном восприятии, показывает необходимость улучшения процентильного метода и метода бутстреп-*t*. Каждый из двадцати шести детей с неврологическими дефектами проходил два теста на пространственное восприятие, тест «A» и тест «B». Эти данные показаны в таблице 14.1 и представлены графически на рис. 14.1.

Table 14.1. *Spatial Test Data; n = 26 children have each taken two tests of spatial ability, called A and B.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 48 | 36 | 20 | 29 | 42 | 42 | 20 | 42 | 22 | 41 | 45 | 14 | 6 |
| B | 42 | 33 | 16 | 39 | 38 | 36 | 15 | 33 | 20 | 43 | 34 | 22 | 7 |
| | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| A | 0 | 33 | 28 | 34 | 4 | 32 | 24 | 47 | 41 | 24 | 26 | 30 | 41 |
| B | 15 | 34 | 29 | 41 | 13 | 38 | 25 | 27 | 41 | 28 | 14 | 28 | 40 |

Предположим, что мы хотим найти 90% доверительный интервал для $\theta = \text{var}(A)$, дисперсии результата теста «A».

Оценка θ по методу подстановки основана на $n = 26$ парах $x_i = (A_i, B_i)$ из таблицы 14.1.

$$\hat{\theta} = \sum_{i=1}^n (A_i - \bar{A})^2 / n = 171.5, \quad (\bar{A} = \sum_{i=1}^n A_i / n) \quad (14.1)$$

Следует заметить, что это немного меньше обычной несмешенной оценки θ ,

$$\bar{\theta} = \sum_{i=1}^n (A_i - \bar{A})^2 / (n - 1) = 178.4 \quad (14.2)$$

Оценка методом подстановки $\hat{\theta}$ смещена вниз. Метод ВС_a автоматически делает поправку на смещение в оценке по методу подстановки, что является одним из его достоинств перед процентильным методом.¹ Гистограмма 2000 бутстреп репликаций $\hat{\theta}^*$ показана на левой панели рисунка 14.2. Репликации получены таким же образом, как и в случае рисунка 6.1: если $\mathbf{x} = (x_1, x_2, \dots, x_{26})$ представляет из себя исходный набор данных таблицы 14.1, где $x_i = (A_i, B_i)$, $1, 2, \dots, 26$, тогда бутстреп выборка $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_{26}^*)$ есть случайная выборка размера 26 с возвращением из набора $\{x_1, x_2, \dots, x_{26}\}$; бутстреп репликация $\hat{\theta}^*$ есть дисперсия A компонент \mathbf{x}^* , где $x_i^* = (A_i^*, B_i^*)$,

$$\hat{\theta} = \sum_{i=1}^n (A_i^* - \bar{A}^*)^2 / n, \quad (\bar{A}^* = \sum_{i=1}^n A_i^* / n). \quad (14.3)$$

¹ Для рассуждений в этой части, а также для алгоритмов `bcanon` и `abcanon`, предполагаем, что статистика имеет форму $\hat{\theta} = t(\hat{F})$ (получена методом подстановки)

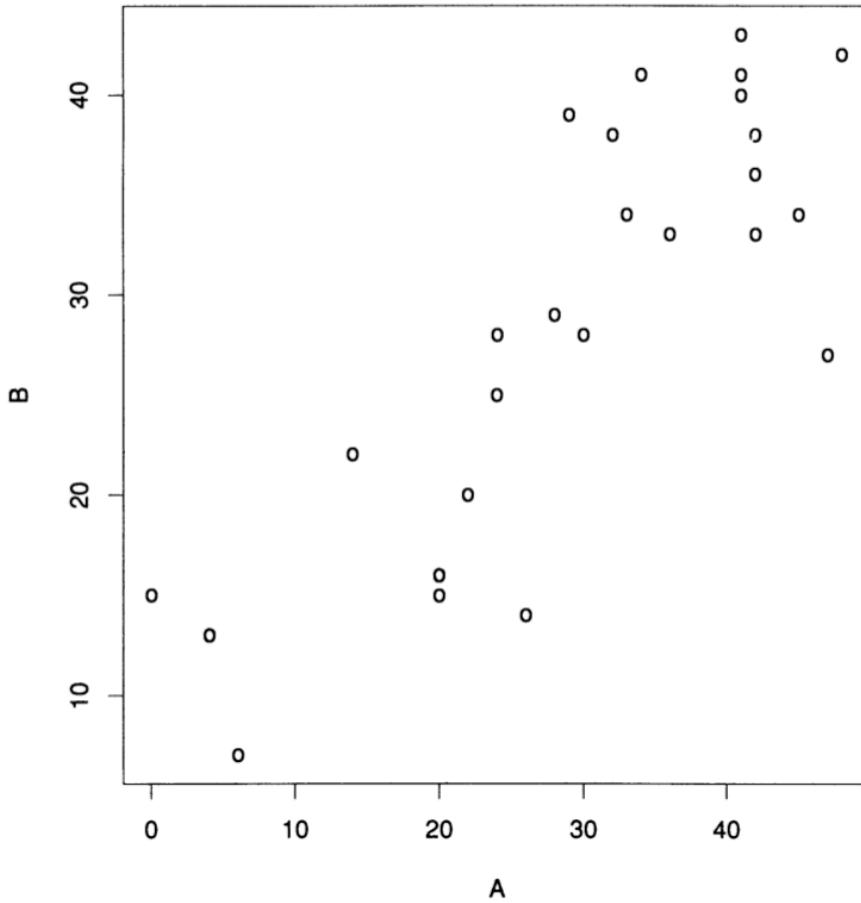


Figure 14.1. The spatial test data of Table 14.1.

$B = 2000$ бутстреп выборок \mathbf{x} дают 2000 бутстреп репликаций $\hat{\theta}^*$ на рис.14.2.² Это так называемые *непараметрические* бутстреп репликации, которые мы уже рассматривали в предыдущих частях. Далее мы также обсудим *параметрические* бутстреп репликации, а именно предположим нормальную модель для данных. Согласно обозначениям главы 6, непараметрическая бутстреп выборка генерируется случайным выбором из \widehat{F} ,

$$\widehat{F} \rightarrow \mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*), \quad (14.4)$$

где \widehat{F} есть эмпирическая функция распределения, для которого вероятность каждого из x_i равна $1/n$.

В верхней части таблицы 14.2 показаны пять различных приближенных 90% непараметрических доверительных интервалов для θ :

- стандартный интервал $\hat{\theta} \pm 1.645\hat{\sigma}$;
- бутстреп оценка стандартной ошибки;
- процентильный интервал $(\hat{\theta}^{*(0.05)}, \hat{\theta}^{*(0.95)})$, построенный на основе левой гистограммы на рис. 14.2;
- BC_a и ABC интервалы, которые обсуждаются в следующих двух разделах;
- бутстреп-t интервалы из главы 12.

²легко проверить, что нам не нужны вторые компоненты x_i^* для этих вычислений

Table 14.2. Top: five different approximate 90% nonparametric confidence intervals for $\theta = \text{var}(A)$; in this case the standard and percentile intervals are nearly the same; the BC_a and ABC intervals are longer, and asymmetric around the point estimate $\hat{\theta} = 171.5$. Bottom: parametric normal-theory intervals. In the normal case there is an exact confidence interval for θ . Notice how much better the exact interval is approximated by the BC_a and ABC intervals. Bottom line: the bootstrap-*t* intervals are nearly exact in the parametric case, but give too large an upper limit nonparametrically.

| Nonparametric | | | | |
|---------------------|-------|-------|--------|-------|
| method | 0.05 | 0.95 | length | shape |
| standard | 98.8 | 233.6 | 134.8 | 1.00 |
| percentile | 100.8 | 233.9 | 133.1 | 0.88 |
| BC _a | 115.8 | 259.6 | 143.8 | 1.58 |
| ABC | 116.7 | 260.9 | 144.2 | 1.63 |
| bootstrap- <i>t</i> | 112.3 | 314.8 | 202.5 | 2.42 |

| Parametric (Normal-Theory) | | | | |
|----------------------------|-------|-------|--------|-------|
| method | 0.05 | 0.95 | length | shape |
| standard | 91.9 | 251.2 | 159.3 | 1.00 |
| percentile | 95.0 | 248.6 | 153.6 | 1.01 |
| BC _a | 114.6 | 294.7 | 180.1 | 2.17 |
| ABC | 119.3 | 303.4 | 184.1 | 2.52 |
| exact | 118.4 | 305.2 | 186.8 | 2.52 |
| bootstrap- <i>t</i> | 119.4 | 303.6 | 184.2 | 2.54 |

Каждый из интервалов $(\hat{\theta}_{\text{lo}}, \hat{\theta}_{\text{up}})$ определяется своей длиной и формой (shape):

$$\text{length} = \hat{\theta}_{\text{up}} - \hat{\theta}_{\text{lo}}, \quad \text{shape} = \frac{\hat{\theta}_{\text{up}} - \hat{\theta}}{\hat{\theta} - \hat{\theta}_{\text{lo}}}.$$
 (14.5)

«Форма» есть показатель асимметричности интервала относительно оценки $\hat{\theta}$. Показатель формы больший, чем 1, означает, что расстояние между $\hat{\theta}$ и $\hat{\theta}_{\text{up}}$ больше, чем расстояние между $\hat{\theta}$ и $\hat{\theta}_{\text{lo}}$. Стандартные интервалы симметричны относительно $\hat{\theta}$, откуда shape = 1 по определению. Точные интервалы, когда они существуют, чаще всего оказываются асимметричными. Построенные стандартные интервалы оказываются ошибочными во-многом из-за их «врожденной» симметрии.

Для рассматриваемого набора данных стандартные и процентильные интервалы практически совпадают. Оба несколько отличаются от BC_a и ABC интервалов, которые оказались более длинными и асимметричными вправо относительно $\hat{\theta}$. Общий результат, приведенный в разделе 13.2, говорит о том, что интервалы BC_a и ABC лучше, однако мы не можем утверждать об этом однозначно, так как для таких сравнений не существует «золотого стандарта.»

В то же время, мы можем получить «золотой стандарт», если рассмотрим задачу оценивания $\text{var}(A)$ в рамках параметрического подхода (предположив гауссовость³). Для этого мы предположим, что результаты тестов $x_i = (A_i, B_i)$ есть случайная выборка из двумерного нормального распределения F_{norm} ,

$$F_{\text{norm}} \rightarrow \mathbf{x} = (x_1, x_2, \dots, x_n). \quad (14.6)$$

При выбранном параметрическом подходе мы можем построить точный доверительный интервал для $\theta = \text{var}(A)$. Этот интервал, названный «точным» (*exact*) в таблице 14.2, является «золотым стандартом» для оценки различных приближенных интервалов в параметрических условиях. Выборка гауссова параметрического бутстрепа получается генерацией выборок из двумерного нормального распределения $\widehat{F}_{\text{norm}}$, которое наилучшим образом соответствует данным \mathbf{x} вместо эмпирического распределения \widehat{F} , то есть

$$\widehat{F}_{\text{norm}} \rightarrow \mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*). \quad (14.7)$$

Получив \mathbf{x}^* , бутстреп репликация $\widehat{\theta}^*$ будет равна

$$\sum_1^n (A_i^* - \bar{A}^*)^2 / n,$$

как в (14.3).

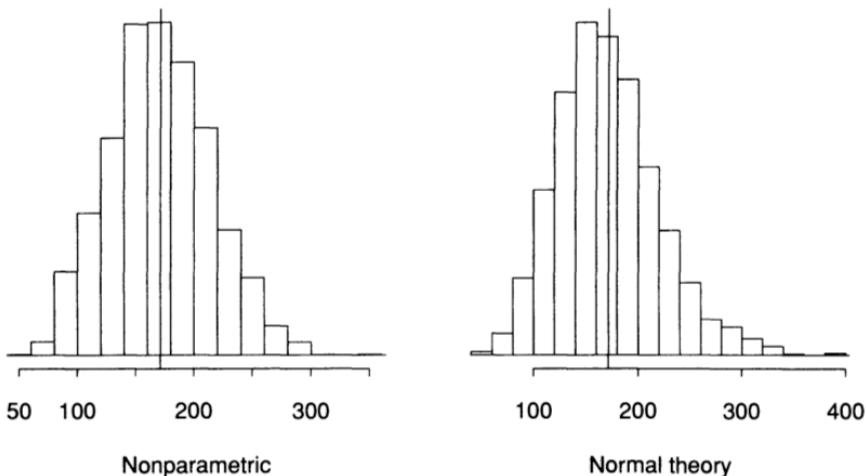


Figure 14.2. Left panel: 2000 nonparametric bootstrap replications of the variance $\widehat{\theta}$, (14.2); Right panel: 2000 normal-theory parametric bootstrap replications of $\widehat{\theta}$. A solid vertical line is drawn at $\widehat{\theta}$ in each histogram. The parametric bootstrap histogram is long-tailed to the right. These histograms are used to form the percentile and BC_a intervals in Table 14.2.

Правая гистограмма на рис. 14.2 — гистограмма 2000 репликаций параметрического бутстрепа. Если сравнить ее с непараметрической версией, гистограмма *normal theory* имеет длинный хвост справа, а также шире, при этом $\widehat{\sigma} = 47.1$, если сравнить со стандартной ошибкой в непараметрическом случае — 41.0.

³На самом деле, не похоже, что исходные данные распределены нормально. Однако это не запрещает провести сравнительный анализ методов, аппроксимирующих точный интервал в *предположении*, что данные распределены нормально. Все же, если сравнивать параметрические и непараметрические интервалы, то последние оказываются более предпочтительными для данного набора данных.

Если обратить внимание на нижнюю часть таблицы 14.2, то можно увидеть, что интервалы по методам BC_a и ABC оказываются близкими к точному «золотому стандарту». И это не просто случайность или частный случай. На самом деле, теория бутстрепа, приведенная кратко в разделе 14.3, говорит о том, что мы можем ожидать успешные результаты от BC_a и ABC.

Бутстреп- t интервалы для θ показаны в нижних частях таблицы 14.2. Они основаны на 1000 бутстреп репликаций статистики (по аналогии с t -статистикой) $(\hat{\theta} - \theta)/\widehat{se}$, со знаменателем, «взятым» из стандартной статистической теории,

$$\widehat{se} = \left[\frac{U_4 - U_2^2}{26} \right]^{1/2} \quad (U_h = \sum_{i=1}^{26} (A_i - \bar{A})^h / 26). \quad (14.8)$$

Результирующие интервалы, как в (12.19), оказываются практически точными в случаях нормальной теории. Однако верхний предел непараметрического интервала кажется слишком большим, хотя об этом сложно утверждать в отсутствии непараметрического «золотого стандарта». На данном уровне развития метод бутстреп- t не может быть рекомендован к использованию в непараметрической постановке.

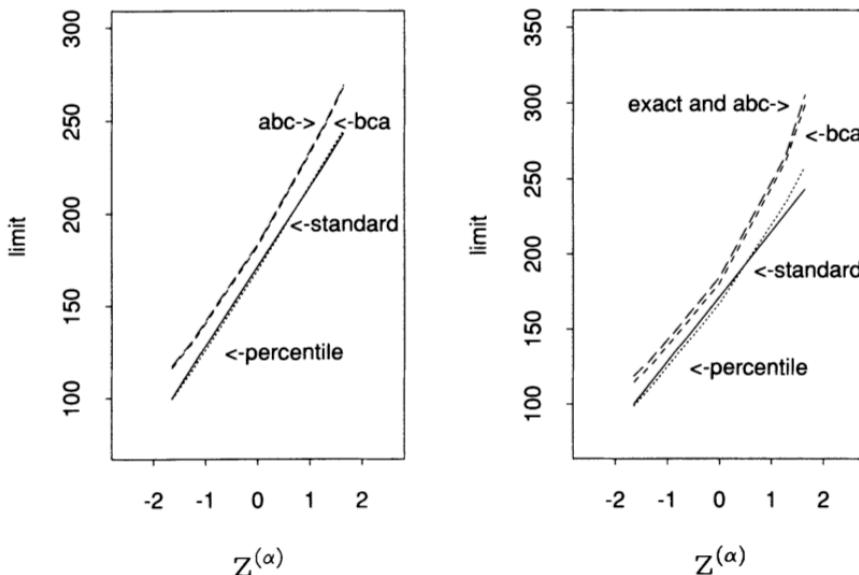


Figure 14.3. A comparison of various approximate confidence intervals for $\theta = \text{var}(A)$, spatial test data; interval endpoint $\hat{\theta}[\alpha]$ is plotted versus $\Phi^{-1}(\alpha) = z^{(\alpha)}$. Left panel: nonparametric intervals. Right panel: normal-theory parametric intervals. In the parametric case we can see that the BC_a and ABC endpoints are close to the exact answer.

14.3 Метод BC_a

В этом разделе описано построение BC_a интервалов. Они оказываются более сложными в описании, чем процентильные интервалы, однако в применении они так же просты. Алгоритм `bcanon`, данный в приложении, строит непараметрические BC_a интервалы адаптивно.

Пусть $\hat{\theta}^{*(\alpha)}$ обозначает $100 \cdot \alpha$ процентиль B бутстреп репликаций

$$\hat{\theta}^*(1), \hat{\theta}^*(2), \dots, \hat{\theta}^*(B),$$

как в (13.5). Процентильный интервал $(\hat{\theta}_{lo}, \hat{\theta}_{up})$ предполагаемого покрытия $1 - 2\alpha$, получается напрямую из этих процентилей, то есть

$$\text{процентильный метод: } (\hat{\theta}_{lo}, \hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}).$$

Например, пусть $B = 2000$ и $\alpha = 0.05$; тогда процентильный интервал $(\hat{\theta}^{*(0.05)}, \hat{\theta}^{*(0.95)})$ будет интервалом, покрывающим упорядоченные значения $\hat{\theta}^*(b)$ от 100го до 1900го.

Границы интервала BC_a также даются процентилями бутстреп распределения, однако они необязательно совпадают с интервалом на основе (14.8). Используемые процентили зависят от двух чисел \hat{a} и \hat{z}_0 , которые определены как *ускорение* (acceleration) и *поправка смещения* (bias-correction), соответственно. Далее мы опишем получение чисел \hat{a} и \hat{z}_0 , но сначала дадим определение границ интервала BC_a .

BC_a интервал с предполагаемым покрытием $1 - 2\alpha$ задается парой значений

$$BC_a : (\hat{\theta}_{lo}, \hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}), \quad (14.9)$$

где

$$\begin{aligned} \alpha_1 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right), \\ \alpha_2 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right). \end{aligned} \quad (14.10)$$

Здесь $\Phi(\cdot)$ есть функция стандартного нормального распределения, а $z^{(\alpha)}$ есть 100α процентиль стандартного нормального распределения. К примеру, $z^{(0.95)} = 1.645$ и $\Phi(1.645) = 0.95$.

Формула (14.10) выглядит сложно, однако её легко вычислить. Заметим, что если приравнять \hat{a} и \hat{z}_0 к нулю, то

$$\alpha_1 = \Phi(z^\alpha) = \alpha \quad \text{и} \quad \alpha_2 = \Phi(z^{1-\alpha}) = 1 - \alpha, \quad (14.11)$$

откуда можно увидеть, что в таком случае BC_a интервал (14.9) совпадает с процентильным интервалом (13.4). Ненулевые значения \hat{a} или \hat{z}_0 изменяют процентили, используемые для вычисления границ BC_a . Такие изменения исправляют некоторые недостатки стандартного и процентильного методов, что объясняется в 22 главе. Непараметрические BC_a интервалы из таблицы 14.2 построены на значениях

$$(\hat{a}, \hat{z}_0) = (0.061, 0.146), \quad (14.12)$$

что приводит к значениям (согласно (14.10))

$$(\alpha_1, \alpha_2) = (0.110, 0.985). \quad (14.13)$$

В данном случае 90% BC_a интервал есть $(\hat{\theta}^{*(0.110)}, \hat{\theta}^{*(0.985)})$, интервал, расположенный между 220-ым и 1970-ым упорядоченными значениями 2000 чисел $\hat{\theta}^*(b)$.

Как вычисляются \hat{a} и \hat{z}_0 ? Значение поправки смещения \hat{z}_0 получается напрямую из доли бутстреп репликаций, меньших исходной оценки $\hat{\theta}$,

$$\hat{z}_0 = \Phi^{-1} \left(\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B} \right), \quad (14.14)$$

где $\Phi^{-1}(\cdot)$ есть обратная функция к функции распределения стандартного нормального закона.⁴ У левой гистограммы на рисунке 14.2 1116 из 2000 значений $\hat{\theta}^*$ оказались меньше, чем $\hat{\theta} = 171.5$, откуда $\hat{z}_0 = \Phi^{-1}(0.558) = 0.146$. Грубо говоря, \hat{z}_0 оценивает медианное смещение $\hat{\theta}^*$, то есть степень различия между медианой $\hat{\theta}^*(b)$ и $\hat{\theta}$ в «нормальной» шкале. Мы получим \hat{z}_0 , если ровно половина из всех значений $\hat{\theta}^*(b)$ окажется меньшими или равными $\hat{\theta}$.

Есть несколько способов вычисления ускорения \hat{a} . Проще всего описать его с помощью значений по методу складного ножа статистики $\hat{\theta} = s(\mathbf{x})$. Пусть $\mathbf{x}_{(i)}$ — исходная выборка с удаленным наблюдением x_i , также обозначим $\hat{\theta}_{(i)} := s(\mathbf{x}_{(i)})$ и $\hat{\theta}_{(\cdot)} = \sum_{i=1}^n \hat{\theta}_{(i)}/n$, согласно рассуждениям из начала главы 11. Простое выражение для ускорения

$$\hat{a} = \frac{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6 \left\{ \sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2 \right\}^{3/2}}. \quad (14.15)$$

У статистики $s(\mathbf{x}) = \sum_{i=1}^n (A_i - \bar{A})^2/n$, из (14.2), значение \hat{a} для набора данных о тестах на пространственное восприятие составляет $\hat{a} = 0.061$. Как \hat{a} , так и \hat{z}_0 вычисляются автоматически реализацией непараметрического алгоритма BC_a . Величина \hat{a} называется *ускорением* из-за того, что она описывает скорость изменения стандартной ошибки $\hat{\theta}$ относительно истинного значения параметра θ . Стандартная нормальная аппроксимация — $\hat{\theta} \sim N(\theta, se^2)$ — предполагает, что стандартная ошибка $\hat{\theta}$ одинакова для всех θ . Однако часто это предположение нереалистично, и константа ускорения \hat{a} делает поправку. Например, в текущем примере, где $\hat{\theta}$ есть дисперсия, в контексте теории нормального распределения ясно, что $se\hat{\theta} \sim \theta$. Фактически, \hat{a} есть скорость изменения стандартной ошибки $\hat{\theta}$ относительно истинного значения параметра θ , измеренная в «нормальной» шкале. Не является очевидным то, почему формула (14.15) должна привести к оценке ускорения стандартной ошибки: некоторые разъяснения этого результата можно найти у Efron (1987).

У метода BC_a есть два важных теоретических преимущества. Во-первых, этот метод сохраняет отображения,⁵ как в формуле (13.10). Это означает, что граничные точки интервала BC_a отображаются корректно при замене интересующего параметра θ на некоторую функцию от него. Например, BC_a интервалы для $\sqrt{\text{var}(A)} = \sqrt{\theta}$ получаются взятием квадратных корней из граничных точек BC_a в таблице 14.2. Свойство сохранения интервала при отображении оберегает от сомнений, которые имеют место при выборе масштаба для бутстреп-т интервала, как в разделе 12.6. BC_a автоматически выбирает наилучшую шкалу.

Второе преимущество метода BC_a заключается в его точности. Доверительный интервал $(\hat{\theta}_{\text{lo}}, \hat{\theta}_{\text{up}})$ уровня $1 - 2\alpha$ должен иметь вероятность α *непокрытия* истинного значения θ сверху или снизу, то есть

$$\text{Prob} \left\{ \theta < \hat{\theta}_{\text{lo}} \right\} \doteq \alpha \quad \text{и} \quad \text{Prob} \left\{ \theta > \hat{\theta}_{\text{up}} \right\} \doteq \alpha \quad (14.16)$$

Можно оценить качество приближенных доверительных интервалов на основании того, насколько они удовлетворяют (14.16). Можно показать, что интервалы

⁴то есть $\Phi^{-1}(0.95) = 1.645$

⁵Данное утверждение будет строго верным, если принять другое определение \hat{a} , основанное на конечных разностях, как в главе 22. На практике это различие оказывается несущественным

BC_a имеют второй порядок точности. Это означает, что отклонение от (14.16) сходится к нулю со скоростью $1/n$, (с увеличением размера выборки n) то есть

$$\text{Prob} \left\{ \theta < \hat{\theta}_{\text{lo}} \right\} = \alpha + \frac{c_{\text{lo}}}{n} \quad \text{и} \quad \text{Prob} \left\{ \theta > \hat{\theta}_{\text{up}} \right\} = \alpha + \frac{c_{\text{up}}}{n} \quad (14.17)$$

для двух констант c_{lo} и c_{up} . Стандартный и процентильный методы имеют лишь *первый порядок точности*, поэтому ошибки оказываются на порядок выше:

$$\text{Prob} \left\{ \theta < \hat{\theta}_{\text{lo}} \right\} = \alpha + \frac{c_{\text{lo}}}{\sqrt{n}} \quad \text{и} \quad \text{Prob} \left\{ \theta > \hat{\theta}_{\text{up}} \right\} = \alpha + \frac{c_{\text{up}}}{\sqrt{n}}, \quad (14.18)$$

где константы c_{lo} и c_{up} могут отличаться от тех, которые были ранее. Разница между первым и вторым порядком точности имеет не только теоретический характер. Она также приводит к улучшенной аппроксимации точных границ тогда, когда они существуют, как в таблице 14.2.

Метод бутстреп- t имеет второй порядок точности, однако не обладает свойством сохранения отображения. Процентильный метод обладает, однако не имеет второй порядок точности; как и стандартный метод. BC_a метод обладает обоими преимуществами. На текущий момент метод BC_a рекомендуется к универсальному использованию, в особенности для непараметрических задач. Нельзя сказать, что метод идеален или не может быть модифицирован: в разделе 25.6 главы 25 используется дополнительное применение бутстрапа для улучшения результатов, полученных с помощью BC_a и ABC методов.

Стандартный вызов функции `bcanon` имеет вид

$$\text{bcanon}(\mathbf{x}, \text{nboot}, \text{theta}), \quad (14.19)$$

где \mathbf{x} — данные, nboot — число бутстреп репликаций, theta — вид статистики $\hat{\theta}$. Больше подробностей — в приложении.

14.4 Метод ABC

Главный недостаток метода BC_a заключается в необходимости проведения большого числа итераций. В главе 19 показано, что для удовлетворительного уменьшения ошибки Монте-Карло требуется не менее $B = 1000$ репликаций. Метод ABC (*approximate bootstrap confidence*) представляет из себя метод, который оценивает границы интервалов аналитически, без использования репликаций Монте-Карло. Данная аппроксимация обычно оказывается достаточно неплохой, что видно из результатов в таблице 14.2. (Разница между граничными точками BC_a и ABC объясняется вариативностью Монте-Карло при вычислении BC_a интервала. Увеличение B до 10000 параметрических репликаций дает BC_a интервал (118.4, 303.8), практически совпадающий с ABC интервалом.)

Метод ABC описан в главе 22. Его работа заключается в аппроксимации результатов бутстреп моделирования используя разложения Тейлора. Для этого необходимо, чтобы оцениваемая статистика $\hat{\theta} = s(\mathbf{x})$ была гладкой по \mathbf{x} . Пример негладкой статистики — выборочная медиана. Для большинства часто встречающихся статистик метод ABC оказывается весьма удовлетворительным. (Контрпример приведен в разделе 14.5.) Построенные по ABC интервалы — как и граничные точки по методу BC_a — сохраняют отображения и имеют второй порядок точности. Для построения оценок по методу ABC в таблице 14.2 потребовалось всего 3% вычислительных затрат, необходимых для построения BC_a интервала.

Непараметрические границы по методу АВС в таблице 14.2 были получены из алгоритма `abcnon`, приведенном в приложении. Для использования этого алгоритма статистика $\hat{\theta}$ должна быть представлена в специальной форме (*resampling form*). Как будет показано в главе 20, эта форма важна для развития теории бутстреп методов. Форма была определена в разделе 10.4. Фиксируя исходную выборку $\mathbf{x} = (x_1, x_2, \dots, x_n)$, запишем бутстреп значение $\hat{\theta}^* = s(\mathbf{x}^*)$ как функцию вектора повторной выборки \mathbf{P}^* , то есть

$$\hat{\theta}^* = T(\mathbf{P}^*). \quad (14.20)$$

Вектор $\mathbf{P}^* = (P_1^*, P_2^*, \dots, P_n^*)$ состоит из долей

$$P_i^* = N_i^*/n = \frac{\#\{x_j^* > x_i\}}{n} \quad (i = 1, 2, \dots, n). \quad (14.21)$$

Статистика $\hat{\theta}^* = \sum_{i=1}^n (A_i^* - \bar{A}^*)^2/n$, (14.3), может быть представлена в виде (14.20) следующим образом

$$\hat{\theta}^* = \sum_{i=1}^n P_i^*(A_i - \bar{A}^*)^2, \text{ где } \bar{A}^* = \sum_{i=1}^n P_i^* A_i. \quad (14.22)$$

Функция $T(\mathbf{P}^*)$ из (14.20) есть необходимая форма статистики, которая используется в алгоритме АВС. Напомним, определен и следующий вектор повторной выборки

$$\mathbf{P}^0 = (1/n, 1/n, \dots, 1/n) \quad (14.23)$$

для которого выполняется $T(\mathbf{P}^0) = \hat{\theta}$, исходное значение статистики. Алгоритм `abcnon` требует, чтобы $T(\mathbf{P}^*)$ была гладкой для \mathbf{P}^* в окрестности \mathbf{P}^0 . Это происходит естественным образом, как в (14.22), для статистик по методу подстановки $\hat{\theta} = t(\hat{F})$.

Типичный вызов функции `abcnon` имеет вид

$$\text{abcnon}(\mathbf{x}, \mathbf{tt}), \quad (14.24)$$

где \mathbf{x} — данные, \mathbf{tt} — статистика $\hat{\theta}^*$ в специальной форме. Больше информации — в приложении.

Подводя итоги, АВС интервалы сохраняют отображения, имеют второй порядок точности, а также служат хорошим приближением ВС_a интервалов для крупного класса гладких статистик $\hat{\theta}^* = s(\mathbf{x}^*)$. Для реализации `abcnon` алгоритма АВС требуется, чтобы статистика была приведена в специальной форме $\hat{\theta}^* = T(\mathbf{P}^*)$. В то же время, удобная и простая реализация, а также серьезные вычислительные преимущества указывают на пригодность данного подхода.

14.5 Пример: данные о твердости зубов

⁶ Мы завершаем эту главу рассмотрением более сложного примера, который покажет как возможности, так и ограничения непараметрических ВС_a и АВС доверительных интервалов.

В таблице 14.3 можно увидеть данные о твердости зубов. Тринадцать человек, попавших в некоторые происшествия потеряли от одного до четырех

⁶Материал этого раздела является продвинутым, поэтому он может быть пропущен при первом прочтении

здоровых зубов. Твердость удалённых зубов была оценена деструктивным исследованием, что в стандартных условиях неосуществимо. «Твердость» в последнем столбце таблицы 14.3 — измеренный для каждого пациента показатель средней твердости зубов (в логарифмической шкале).

Table 14.3. *The tooth data.* Thirteen accident victims have had the strength of their teeth measured, right column. It is desired to predict tooth strength from measurements not requiring destructive testing. Four such variables have been measured for each subject: the pair labeled (D_1, D_2) , are difficult to obtain, the pair labeled (E_1, E_2) are easy to obtain. Do the Easy variables predict strength as well as the Difficult ones?

| patient | D_1 | D_2 | E_1 | E_2 | strength |
|---------|--------|--------|-------|-------|----------|
| 1 | -5.288 | 10.091 | 12.30 | 13.08 | 36.05 |
| 2 | -5.944 | 10.001 | 11.41 | 12.98 | 35.51 |
| 3 | -5.607 | 10.184 | 11.76 | 13.19 | 35.35 |
| 4 | -5.413 | 10.131 | 12.09 | 12.75 | 35.95 |
| 5 | -5.198 | 8.835 | 10.72 | 11.73 | 34.64 |
| 6 | -5.598 | 9.837 | 11.74 | 12.80 | 33.99 |
| 7 | -6.120 | 10.052 | 11.10 | 12.87 | 34.60 |
| 8 | -5.572 | 9.900 | 11.85 | 12.72 | 34.62 |
| 9 | -6.056 | 9.966 | 11.78 | 13.06 | 35.05 |
| 10 | -5.010 | 10.449 | 12.91 | 13.15 | 35.85 |
| 11 | -6.090 | 10.294 | 11.63 | 12.97 | 35.53 |
| 12 | -5.900 | 10.252 | 11.91 | 13.15 | 34.86 |
| 13 | -5.620 | 9.316 | 10.89 | 12.25 | 34.75 |

Исследователи хотели предсказать твердость зубов используя переменные, которые не требуют разрушения зубов и могут быть измерены на рутинных осмотрах. В таблице 14.3 показаны данные о четырех таких переменных — D_1 , D_2 , E_1 , E_2 . Пару (D_1, D_2) трудно и дорого получить, а пару (E_1, E_2) — легко и дешево. Исследователи задались следующим вопросом: насколько хорошо «простые» переменные (E_1, E_2) предсказывают твердость зубов в сравнении с «труднодоступными» (D_1, D_2) .

Данный вопрос можно формализовать используя линейные модели, как в главах 7 и 9. Каждая строка x_i матрицы данных из таблицы 14.3 состоит из пяти чисел: двух D , двух E , а также значения твердости, то есть

$$x_i = (d_{i1}, d_{i2}, e_{i1}, e_{i2}, y_i) \quad (i = 1, 2, \dots, 13). \quad (14.25)$$

Пусть \mathbf{D} — матрица, использующая только переменные D для предсказания y_i с помощью линейной регрессии (включая сдвиг), то есть \mathbf{D} — матрица 13×3 с i -й строкой вида

$$(1, d_{i1}, d_{i2}). \quad (14.26)$$

Оценка y_i по методу наименьших квадратов на основе переменных D имеет вид

$$\hat{y}_i(D) = \hat{\beta}_0(D) + \hat{\beta}_1(D)d_{i1} + \hat{\beta}_2(D)d_{i2}, \quad (14.27)$$

где вектор $\widehat{\beta}(D) = (\widehat{\beta}_0(D), \widehat{\beta}_1(D), \widehat{\beta}_2(D))$ есть решение задачи наименьших квадратов (9.28), то есть

$$\widehat{\beta}(D) = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{y}, \quad (14.28)$$

где $y = (y_1, y_2, \dots, y_{13})$. RSE(D) (residual squared error) есть сумма квадратов ошибок между предсказаниями $\widehat{y}_i(D)$ и наблюдениями y_i для $n = 13$ пациентов

$$RSE(D) = \sum_{i=1}^N (y_i - \widehat{y}_i(D))^2. \quad (14.29)$$

Меньшие значения RSE являются индикатором хорошего качества предсказания; наилучшее возможное значение RSE = 0 достигается на идеальном предсказании для каждого из пациентов.

Аналогичным образом мы можем предсказывать y_i , используя только переменные E и в результате вычислить

$$RSE(E) = \sum_{i=1}^n (y_i - \widehat{y}_i(E))^2. \quad (14.30)$$

Вопрос исследователя о том, насколько переменные двух разных типов, D и E , сравнимы по качеству предсказания, может быть переформулирован как вопрос о сравнении RSE(D) и RSE(E). Практическая в этом случае статистика имеет вид

$$\widehat{\theta} = \frac{1}{n} [RSE(E) - RSE(D)]. \quad (14.31)$$

Положительное значение $\widehat{\theta}$ означало бы, что переменные E хуже чем переменные D в предсказании твердости. (Если бы число измерений E и D не совпадают, то статистика $\widehat{\theta}$ должна быть преобразована). Были получены значения RSE(D) = 2.761 и RSE(E) = 3.130, что приводит к

$$\widehat{\theta} = 0.0285. \quad (14.32)$$

Это свидетельствует о том, что переменные D являются более предпочтительными для предсказания, так как $\widehat{\theta}$ больше нуля, однако мы не можем быть уверены в этом, пока не оценим статистическую изменчивость $\widehat{\theta}$. Для этого мы используем методы BC_a и ABC. Рисунок 14.4 указывает на то, что ситуация может оказаться «пограничной», так как предсказанные значения $\widehat{y}_i(D)$ и $\widehat{y}_i(E)$ близки для каждого из наблюдений. Также следует заметить, что разница между RSE(D) и RSE(E) составляет около 10% от самих значений RSE. Поэтому даже если эта разница статистически значима, она может быть не сильно важной в практическом плане. Построение доверительного интервала позволит ответить как на вопрос значимости, так и на вопрос важности этой разницы.

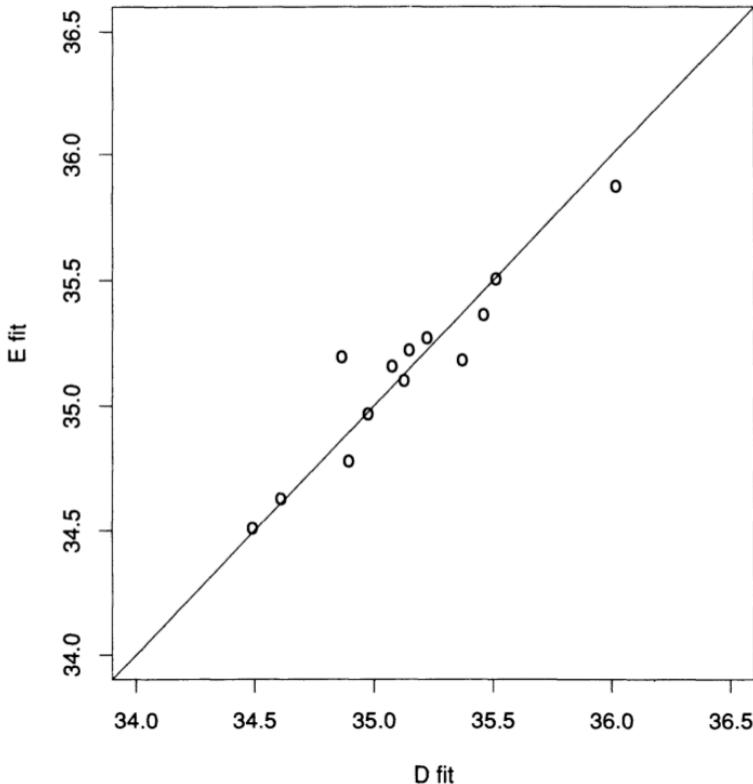


Figure 14.4. The least-squares predictions $\hat{y}(D)$, horizontal axis, versus $\hat{y}_i(E)$, vertical axis, for the 13 patients in Table 14.3. The 45° line is shown for reference. The two sets of predictions appear quite similar.

В левой части рисунка 14.5 — гистограмма 2000 репликаций непараметрического бутстрэпа статистики RSE разности $\hat{\theta}$, (14.31). Пусть $\mathbf{x} = (x_1, x_2, \dots, x_{13})$ представляет из себя матрицу данных из таблицы 14.3, где x_i есть i -ый столбец матрицы, (14.25). Непараметрическая бутстрэп выборка $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_{13}^*)$ есть матрица, состоящая из строк, взятых с возвращением из совокупности $\{x_1, x_2, \dots, x_{13}\}$. Это эквивалентно следующей записи:

$$\hat{F} \rightarrow \mathbf{x}^* = (x_1^*, x_2^*, \dots, x_{13}^*), \quad (14.33)$$

где \hat{F} есть эмпирическая функция распределения, которая задает вероятность $1/13$ выбора каждой из строк x_i .

Следуя определениям из (14.25)–(14.30) бутстрэп матрица \mathbf{x}^* приводит к \mathbf{y}^* , \mathbf{D}^* , $\hat{\beta}(D)^*$, $\hat{y}_i(D)^*$, а затем

$$RSE(D)^* = \sum_{i=1}^{13} (y_i^* - \hat{y}_i(D)^*)^2, \quad (14.34)$$

и аналогично $RSE(E)^* = \sum_{i=1}^{13} (y_i^* - \hat{y}_i(E)^*)^2$. Бутстрэп репликация $\hat{\theta}$ будет иметь вид

$$\hat{\theta}^* = \frac{1}{13} [RSE(E)^* - RSE(D)^*]. \quad (14.35)$$

Как и всегда, $\hat{\theta}^*$ вычисляется с помощью того же алгоритма, что и исходная оценка $\hat{\theta}$. Меняется лишь матрица данных \mathbf{x} на \mathbf{x}^* и вектор \mathbf{y} на \mathbf{y}^* .

Бутстрэп гистограмма содержит информацию, которая нам необходима для ответа на вопросы о значимости и важности $\hat{\theta}$. Даже без построения довери-

тельных интервалов можно получить некоторые ответы. Бутстреп оценка стандартной ошибки (6.6) равна

$$\widehat{s}e_{2000} = 0.0311. \quad (14.36)$$

Это означает, что $\widehat{\theta} = 0.0285$ менее чем на одну стандартную ошибку отстоит от нуля, откуда можно сделать вывод о том, что нам не следует ожидать серьезных оснований отвергнуть гипотезу о том, что истинное значение θ равно 0. С другой стороны, оценка смещена вниз (62% значений $\widehat{\theta}^*$ оказываются меньшими, чем $\widehat{\theta}^*$). Это указывает на то, что уровень значимости окажется больше $0.18 = 1 - \Phi(0.0285/0.0311)$, в условиях нормальной аппроксимации $\widehat{\theta} \sim N(\theta, 0.0311^2)$.

Бутстреп гистограмма указывает на то, что θ оказывается не более 0.10. Насколько существенна эта разница? Для этого нужно понять что именно измеряет параметр θ . Если F есть истинное пятимерное распределение вектора (d_1, d_2, e_1, e_2, y) , то

$$\begin{aligned} \theta_D &= \min_{\beta_D} E_F[y - (\beta_{D_0} + \beta_{D_1}d_1 + \beta_{D_2}d_2)]^2, \\ \theta_E &= \min_{\beta_E} E_F[y - (\beta_{E_0} + \beta_{E_1}e_1 + \beta_{E_2}e_2)]^2 \end{aligned} \quad (14.37)$$

— есть истинное значение ошибок предсказания при использовании переменных D и E , соответственно. Параметр θ , соответствующий $\widehat{\theta}$ есть

$$\theta = \theta_E - \theta_D. \quad (14.38)$$

Оценка θ_D по методу подстановки — $\widehat{\theta}_D = \text{RSE}(D)/13 = 0.212$. Наша гипотеза о том, что $\theta \leq 0.10$ приводит к

$$\frac{\theta_E - \theta_D}{\theta_D} = \frac{\theta_E - \theta_D}{\widehat{\theta}_D} < \frac{0.10}{0.212} = 0.47. \quad (14.39)$$

Можно сделать вывод о том, что переменные Е вероятно не лучше, чем переменные D для предсказания твердости, и вероятно не более чем на 50% хуже.

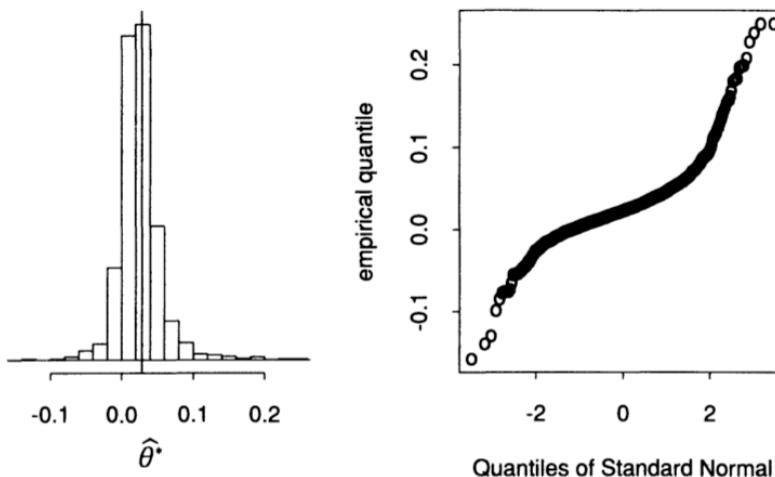


Figure 14.5. *Left panel:* 2000 nonparametric bootstrap replications of the RSE difference statistic $\widehat{\theta}$, (14.31); bootstrap standard error estimate is $\widehat{s}e_{2000} = .0311$; 1237 of the 2000 $\widehat{\theta}^*$ values are less than $\widehat{\theta} = .0285$, so $\widehat{z}_0 = .302$. *Right panel:* quantile-quantile plot of the $\widehat{\theta}^*$ values. Their distribution has much heavier tails than a normal distribution.

В первом столбце таблицы 14.4 указаны доверительные интервалы для θ по методу BC_a на основе 2000 репликаций непараметрического бутстрэпа.

Доверительные границы $\widehat{\theta}[\alpha]$ даны для восьми значений уровня значимости. Доверительные интервалы получаются взятием пар вида $(\widehat{\theta}[\alpha], \widehat{\theta}[1 - \alpha])$ (например, $(\widehat{\theta}[0.05], \widehat{\theta}[0.95]) = 90\%$ интервал). Формулы (14.14) и (14.15) приводят к малому значению ускорения и большой поправке на смещение — $\hat{a} = 0.040$ и $\hat{z}_0 = 0.47$.

Заметим, что непараметрическая граница для 0.05 положительна, $\widehat{\theta}[0.05] = 0.004$. Как говорилось ранее, это связано с большим значением поправки на смещение. Если бы BC_a метод был точным, мы могли бы утверждать, что нулевая гипотеза $\theta = 0$ отвергается при односторонней критической области и уровне значимости 0.05. Метод не точный, поэтому следует быть осторожнее с выводами. Непараметрические BC_a интервалы часто оказываются немного короткими, в особенности в случаях малой выборки (как в этом примере). Если бы проверка гипотезы имела критическое значение, то возможно добиться улучшения уровня значимости посредством калибровке, как в разделе 25.6.

Table 14.4. Bootstrap confidence limits for θ , (14.31); limits $\widehat{\theta}[\alpha]$ given for significance levels $\alpha = .025, .05, \dots, .975$, so central 90% interval is $(\widehat{\theta}[.05], \widehat{\theta}[.95])$. Left panel: nonparametric bootstrap (14.33); Center panel: normal theory bootstrap (14.7); Right panel: linear model bootstrap, (9.25), (9.26). BC_a limits based on 2000 bootstrap replications for each of the three models; ABC limits obtained from the programs `abcnon` and `abcpars` in the Appendix (assuming normal errors for the linear model case); values of \hat{a} and \hat{z}_0 vary depending on the details of the program used. The ABC limits are much too short in the nonparametric case because of the very heavy tails of the bootstrap distribution shown in Figure 14.5. Notice that in the nonparametric case the bootstrap estimate of standard error is nearly twice as big as the estimate used in the ABC calculations.

| α | nonparametric | | normal theory | | linear model | |
|-------------|---------------|-------|---------------|--------|--------------|--------|
| | BC_a | ABC | BC_a | ABC | BC_a | ABC |
| 0.025 | -0.002 | 0.004 | -0.010 | -0.010 | -.031 | -0.019 |
| 0.05 | 0.004 | 0.008 | -0.004 | -0.004 | -.020 | -0.012 |
| 0.1 | 0.010 | 0.012 | 0.004 | 0.003 | -.008 | -0.004 |
| 0.16 | 0.015 | 0.016 | 0.010 | 0.010 | .000 | 0.003 |
| 0.84 | 0.073 | 0.053 | 0.099 | 0.092 | .070 | 0.067 |
| 0.9 | 0.095 | 0.061 | 0.113 | 0.111 | .083 | 0.079 |
| 0.95 | 0.155 | 0.072 | 0.145 | 0.139 | .098 | 0.094 |
| 0.975 | 0.199 | 0.085 | 0.192 | 0.167 | .118 | 0.108 |
| \hat{s}_e | .0311 | .0170 | .0349 | .0336 | .0366 | .0316 |
| \hat{a} | .040 | .056 | .062 | .062 | 0 | 0 |
| \hat{z}_0 | .302 | .203 | .353 | .372 | .059 | .011 |

Для проверки непараметрических интервалов было построено еще 2000 бутстрэп выборок, в данном случае основанных на нормальной модели; предполагается, что строки x_i матрицы данных были получены моделированием из пя-

тимерного нормального распределения F_{norm} . Далее было найдено наилучшее аппроксимирующее данные распределение $\widehat{F}_{\text{norm}}$, а затем сгенерированы выборки \mathbf{x}^* из $\widehat{F}_{\text{norm}}$ как в (14.7). Гистограмма на основе 2000 бутстреп репликаций параметрического бутстрепа показана на левой стороне рисунка 14.6; она похожа на гистограмму из рисунка 14.5 (за исключением того, что в данном случае хвосты менее тяжелые).

BC_a интервалы вычислены так же, как и ранее, используя формулы (14.9) и (14.10). Формула поправки на смещение (14.14) также не изменилась. Параметр ускорения \hat{a} вычислен с помощью параметрической версии формулы (14.15), взятой из алгоритма `abctrag` построения параметрических ABC интервалов. В данном случае параметрические BC_a границы, указанные в центральной части таблицы 14.4, не сильно отличаются от своего непараметрического аналога. В то же время разница оказывается достаточно большой, такой, что гипотеза $\theta = 0$ уже не отвергается на уровне 0.05 (при выборе одностороннего доверительного интервала).

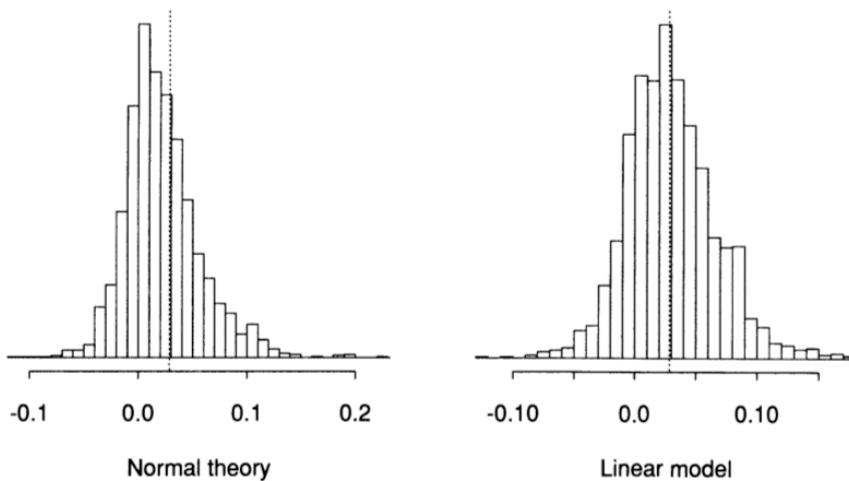


Figure 14.6. Bootstrap replications giving the normal theory and linear model BC_a confidence limits in Table 14.4. Left panel: normal theory; Right panel: linear model. A broken line is drawn at the parameter estimate.

Непохоже, что исходные данные распределены нормально. Однако причиной использования бутстрепа с предположением нормальности заключается в малой размерности данных, $n = 13$. Для слишком малых выборок бутстреп даже с неподходящей параметрической моделью может оказаться более успешным, уменьшив дисперсию результатов в ущерб допустимого смещения. В данном же примере результаты по двум методам весьма похожи.

В главе 9 рассматриваются модели линейной регрессии. Можем использовать модель линейной регрессии для того, чтобы предложить другой вариант бутстреп анализа разностной статистики $\hat{\theta}$. Используя обозначения из (14.25), пусть \mathbf{c}_i представляет из себя вектор

$$\mathbf{c}_i = (1, d_{1i}, d_{2i}, e_{1i}, e_{2i}); \quad (14.40)$$

рассмотрим линейную модель (9.4), (9.5)

$$y_i = \mathbf{c}_i \beta + \varepsilon_i \quad (1, 2, \dots, 13). \quad (14.41)$$

Бутстреп выборки $\mathbf{y}^* = (y_1^*, y_2^*, \dots, y_{13}^*)$ получены из остатков повторных выборок (как в (9.25) и (9.26)). Бутстреп репликации $\hat{\theta}^*$ остаются такими же, как в

(14.35). Следует заметить, что вычисление $\hat{y}_i(D)^*$ и $\hat{y}_i(E)^*$ несколько отличается.

На правой стороне рисунка 14.6 показано бутстреп распределение на основе 2000 репликаций $\hat{\theta}^*$. Хвосты у этой гистограммы легче, чем хвосты на рисунке 14.5. Это отражено и в таблице 14.4 — соответствующие интервалы стали более узкими. Несмотря на это, гипотеза $\theta = 0$ отвергается реже, чем раньше: при уровне $\alpha = 0.16$. Это происходит из-за того, что $\hat{\theta}$ уже не выглядит смещенной вниз, \hat{z}_0 равен 0.059, а не 0.302, как в непараметрическом случае.

Доверительные интервалы и проверка гипотез являются «деликатными» инструментами статистической теории выводов. В этой связи они сильнее зависят от выбора модели, чем обычные стандартные ошибки. В особенности эти соображения верны для случаев, когда размер выборки мал. Исследование зависимостей между пятью переменными на основе 13 наблюдений очевидно является задачей с малым размером выборки. Даже если бы BC_a интервалы были точными (а это не так), то изменение модели приводило бы к другим доверительным интервалам, что видно из таблицы 14.4.

В таблице 14.4 показаны ABC интервалы для трёх различных вариантов выбора модели. Результаты были получены с использованием реализаций программ `abcsnop` и `abcspar` из приложения. Непараметрические ABC интервалы оказываются слишком короткими в данном случае. Это происходит из-за необычно тяжелых хвостов у распределения непараметрического бустстрепа. Если говорить на языке статистики, то метод ABC может исправить асимметрию распределения, но не его экспесс (это все, что требуется для достижения точности второго порядка). Асимптотическая точность метода ABC не гарантирует его успешность на малых выборках.

Стандартные ошибки $\hat{\theta}$ даны для каждого из шести столбцов в таблице 14.4. Приведенные для BC_a — это классические бутстреп стандартные ошибки. Стандартные ошибки для ABC получены с помощью дельта метода из главы 21 (похожего на метод вычисления стандартной ошибки по методу складного ножа, (11.5)). Стандартная ошибка BC_a более чем в два раза превышает ошибку по методу ABC в непараметрическом случае, что свидетельствует о том, что интервалы ABC будут слишком короткими. (Большая BC_a стандартная ошибка стала заметна уже после первых 100 бутстреп репликаций.) Обычно аппроксимация по методу ABC работает удовлетворительно как в таблице 14.2. Однако в любом случае может оказаться полезной проверка стандартных ошибок используя небольшое число бутстреп репликаций (например, 100).

Глава 15

Перестановочные тесты

15.1 Введение

Перестановочные методы — это трудоемкий компьютерный статистический метод, появившийся еще до возникновения компьютеров. Идея была предложена Р.А. Фишером в 1930-х годах, скорее как теоретический аргумент в пользу t -критерия Стьюдента, нежели чем самостоятельный полезный статистический метод. Современные вычислительные мощности делают перестановочные тесты практическими для повседневного использования. Основная идея привлекательно проста и свободна от математических предположений. Существует тесная связь с бутстрепом, который обсуждается далее в этой главе.

15.2 Двухвыборочная задача

Основное применение перестановочных тестов, и единственное, что мы обсуждаем здесь — это двухвыборочная задача (8.3)–(8.5), в которой мы наблюдаем две независимые случайные выборки $\mathbf{z} = (z_1, z_2, \dots, z_n)$ и $\mathbf{y} = (y_1, y_2, \dots, y_m)$, взятые из возможно различных распределений вероятностей F и G ,

$$\begin{aligned} F \rightarrow \mathbf{z} &= (z_1, z_2, \dots, z_n) \text{ независимо от} \\ G \rightarrow \mathbf{y} &= (y_1, y_2, \dots, y_m). \end{aligned} \tag{15.1}$$

Наблюдая за \mathbf{z} и \mathbf{y} , мы хотим проверить *нулевую гипотезу* H_0 об отсутствии разницы между F и G

$$H_0 : F = G. \tag{15.2}$$

Равенство $F = G$ означает, что F и G присваивают равные вероятности всем множествам, $\text{Prob}_F\{A\} = \text{Prob}_G\{A\}$ для A любых подмножеств общего пространства выборок z и y . Если H_0 истинно, то нет никаких различий между вероятностным поведением случайного z или случайного y .

Проверка гипотез — полезный инструмент для ситуаций, подобных ситуации с данными о мышах, таблица 2.1. Мы наблюдали небольшое количество данных, $n = 7$ экспериментальных измерений и $m = 9$ контрольных. Разница в средних

$$\hat{\theta} = \bar{z} - \bar{y} = 30.63 \tag{15.3}$$

побуждает нас поверить в то, что распределение экспериментальных данных F дает более продолжительное время выживания, чем распределение контрольных данных G . На самом деле эксперимент был разработан, чтобы продемонстрировать именно этот результат.

В этой ситуации *нулевая гипотеза* (15.2) о том, что $F = G$, играет роль адвоката дьявола. Если мы не можем решительно отвергнуть возможность того, что H_0 истинно (как окажется в случае с данными о мышах), то мы не сможем успешно продемонстрировать превосходство лечения над не лечением. *Проверка гипотезы*, примером которой является перестановочный тест, представляет собой формальный способ решить, отвергают ли данные H_0 решительно.

Проверка гипотез начинается с *тестовой статистики* $\hat{\theta}$, разности средних (15.3). Для удобства мы предположим, что если нулевая гипотеза H_0 не верна, мы ожидаем увидеть большие значения $\hat{\theta}$, чем если бы H_0 была верна. Если лечение работает лучше, чем не лечение в эксперименте с мышами, как и предполагалось, то мы ожидаем, что $\hat{\theta} = \bar{z} - \bar{y}$ будет большим. Нам не нужно количественно определять, что означает «большой», чтобы проводить проверку гипотезы. Все, что мы говорим, это то что, чем большее значение $\hat{\theta}$, которое мы наблюдаем, тем сильнее доказательства против H_0 . Конечно, в других ситуациях мы могли бы выбрать меньшие значения вместо больших, чтобы представить более убедительные доказательства. Возможны и более сложные варианты, см. (15.26).

Наблюдаемый $\hat{\theta}$ (достигнутый уровень значимости теста, сокращенно ASL), определяется как вероятность наблюдения, по крайней мере, такого большого значения, когда нулевая гипотеза верна,

$$\text{ASL} = \text{Prob}_{H_0}\{\hat{\theta}^* \geq \hat{\theta}\}. \quad (15.4)$$

Чем меньше значение ASL, тем сильнее доказательства против H_0 , как подробно описано ниже. Величина $\hat{\theta}$ в (15.4) зафиксирована на своем наблюдаемом значении; случайная величина $\hat{\theta}^*$ имеет распределение нулевой гипотезы, распределение $\hat{\theta}$, если H_0 истинно. Как и раньше, в обозначениях со звездочкой указывается различие между фактическим наблюдением $\hat{\theta}$ и гипотетическим $\hat{\theta}^*$, сгенерированным в соответствии с H_0 .

Проверка гипотезы H_0 состоит из вычисления ASL и проверки того, является ли оно слишком маленьким в сравнении с определенными стандартными пороговыми значениями. Формально мы выбираем малую вероятность α , например 0.05 или 0.01, и *отклоняем* H_0 , если ASL меньше α . Если ASL больше α , то мы *принимаем* H_0 , что означает, что экспериментальные данные не отвергают решительно нулевую гипотезу (15.2) об отсутствии разницы между F и G . Менее формально, мы наблюдаем ASL и оцениваем доказательства против H_0 в соответствии со следующими приблизительными соглашениями:

| | |
|----------------------|--|
| $\text{ASL} < 0.1$ | слабо веские доказательства против H_0 ; |
| $\text{ASL} < 0.05$ | достаточно веские доказательства против H_0 ; |
| $\text{ASL} < 0.025$ | веские доказательства против H_0 ; |
| $\text{ASL} < 0.01$ | очень веские доказательства против H_0 . (15.5) |

Традиционная проверка гипотез для данных о мышах может начинаться с предположения, что F и G являются нормальными распределениями с возможно разными средними значениями

$$F = N(\mu_T, \sigma^2), \quad G = N(\mu_C, \sigma^2). \quad (15.6)$$

Нулевая гипотеза $H_0 : \mu_T = \mu_C$. При H_0 , $\hat{\theta} = \bar{z} - \bar{y}$ имеет нормальное распределение

ление со средним 0 и дисперсией $\sigma^2 \left[\frac{1}{n} + \frac{1}{m} \right]$,

$$H_0 : \hat{\theta} \sim N \left(0, \sigma^2 \left(\frac{1}{n} + \frac{1}{m} \right) \right), \quad (15.7)$$

Наблюдая $\hat{\theta}$, ASL представляет собой вероятность того, что случайная величина $\hat{\theta}^*$, распределенная, как в (15.7), превышает $\hat{\theta}$,

$$\text{ASL} = \text{Prob} \left\{ Z > \frac{\hat{\theta}}{\sigma \sqrt{1/n + 1/m}} \right\} = 1 - \Phi \left(\frac{\hat{\theta}}{\sigma \sqrt{1/n + 1/m}} \right), \quad (15.8)$$

где Φ — функция распределения стандартной нормальной переменной Z .

Мы не знаем σ . Стандартная оценка, основанная на (15.6)

$$\bar{\sigma} = \left\{ \left[\sum_{i=1}^n (z_i - \bar{z})^2 + \sum_{j=1}^m (y_j - \bar{y})^2 \right] / [n + m - 2] \right\}^{1/2}, \quad (15.9)$$

что равно 54.21 для данных о мышах. Подставляя $\bar{\sigma}$ в (15.8) и помня, что $\hat{\theta} = 30.63$ дает

$$\text{ASL} = 1 - \Phi \left(\frac{30.63}{54.21 \sqrt{1/9 + 1/7}} \right) = 0.131. \quad (15.10)$$

При этом вычислении $\bar{\sigma}$ рассматривается как фиксированная константа. Т-критерий Стьюдента, учитывающий случайность в $\bar{\sigma}$, дает

$$\text{ASL} = \text{Prob} \left\{ t_{14} > \frac{30.63}{54.21 \sqrt{1/9 + 1/7}} \right\} = 0.141, \quad (15.11)$$

t_{14} указывает на статистику t с 14 степенями свободы. Тест Стьюдента основан на тестовой статистике $\hat{\theta} / [\bar{\sigma} \sqrt{1/n + 1/m}]^{1/2}$ вместо $\hat{\theta}$. Эта статистика имеет распределение t_{n+m-2} при нулевой гипотезе. В этом случае ни (15.10), ни (15.11) не позволяют нам отвергнуть нулевую гипотезу H_0 согласно (15.5) при всех стандартных уровнях значимости.

Основная практическая трудность при проверке гипотез заключается в вычислении ASL (15.4). Мы написали $\text{Prob}_{H_0} \{ \hat{\theta}^* > \hat{\theta} \}$, как если бы нулевая гипотеза H_0 задавала единственное распределение, из которого мы можем вычислить вероятность того, что $\hat{\theta}^*$ превысит $\hat{\theta}$. В большинстве задач нулевая гипотеза (15.2), $F = G$, оставляет нам семейство возможных распределений нулевых гипотез, а не только одно распределение. В нормальном случае (15.6), например, семейство нулевой гипотезы (15.7) включает в себя все нормальные распределения с математическим ожиданием 0. Чтобы фактически вычислить ASL, мы должны были либо аппроксимировать дисперсию нулевой гипотезы, как в (15.10), либо использовать Метод Стьюдента (15.11). Метод Стьюдента прекрасно решает проблему, но он применим только в случаях нормального распределения (15.6).

Перестановочный метод Фишера — это разумный способ вычисления ASL для общей нулевой гипотезы $F = G$. Перед тем, как мы углубимся в подробности, приведем простое его описание. Если нулевая гипотеза верна, любое время выживания для любой из мышей может быть одинаково хорошим при любом

из вариантов лечения. Итак, мы объединяем все $m + n$ наблюдений из обеих групп вместе, затем берем выборку размера m без возвращения, чтобы представить первую группу; остальные n наблюдений составляют вторую группу. Мы вычисляем разницу между средними значениями групп и затем повторяем этот процесс много раз. Если исходная разница в выборочных средних выходит за пределы средних 95% распределения различий, двусторонний перестановочный тест отклоняет нулевую гипотезу на уровне 5%.

Перестановочные тесты основаны на *порядковом статистическом представлении* данных $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ из двухвыборочной задачи. В таблице 15.1 показано порядковое статистическое представление для данных о мышах из таблицы 2.1. Все 16 времен выживания были объединены и отсортированы от наименьшего к наибольшему. В нижней строке приведены ранжированные значения в диапазоне от наименьшего значения 10 до наибольшего значения 197. К какой группе принадлежит каждая точка данных, « z » для экспериментальной или « y » для контрольной группы, показано в верхней строке. Во второй строке показаны ранги с 1 по 16. Мы видим, например, что 11-е наименьшее значение в объединенном наборе данных имело место в экспериментальной группе и равнялось 94. Таблица 15.1 содержит ту же информацию, что и таблица 2.1, но организована таким образом, чтобы облегчить сравнение относительных размеров значений экспериментальной и контрольной групп.

Пусть N равно объединенному размеру выборки $n+m$, и пусть $\mathbf{v} = (v_1, v_2, \dots, v_N)$ будет объединенным и упорядоченным вектором значений; $N = 16$ и $\mathbf{v} = (10, 16, 23, \dots, 197)$ для данных о мышах.

Table 15.1. Order statistic representation for the mouse data of Table 2.1. All 16 data points have been combined and ordered from smallest to largest. The group code is “z” for Treatment and “y” for Control. For example, the 5th smallest of all 16 data points equals 31, and occurs in the Control group.

| | | | | | | | | |
|--------|----|----|----|----|----|----|----|----|
| group: | y | z | z | y | y | z | y | y |
| rank: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| value: | 10 | 16 | 23 | 27 | 31 | 38 | 40 | 46 |

| | | | | | | | | |
|---------|----|----|----|----|-----|-----|-----|-----|
| group: | y | y | z | z | y | z | y | z |
| rank: | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| value : | 50 | 52 | 94 | 99 | 104 | 141 | 146 | 197 |

Также пусть $\mathbf{g} = (g_1, g_2, \dots, g_N)$ будет вектором, который указывает, к какой группе принадлежит каждое упорядоченное наблюдение, верхняя строка в таблице 15.1. Вместе \mathbf{v} и \mathbf{g} передают ту же информацию, что и $\mathbf{x} = (\mathbf{z}, \mathbf{y})$.

Вектор \mathbf{g} состоит из n значений z и m значений y . Есть

$$C_N^n = \frac{N!}{n!m!} \quad (15.12)$$

возможных \mathbf{g} векторов, соответствующих всем возможным способам разбиения N элементов на два подмножества размера n и m . Перестановочные тесты зависят от следующего важного результата:

Лемма о перестановке. При $H_0 : F = G$ вектор \mathbf{g} с вероятностью $1/C_N^n$ равняется любому из своих возможных значений.

Другими словами, все перестановки z и y равновероятны, если $F = G$. Мы можем думать о тестовой статистике $\hat{\theta}$ как о функции от \mathbf{g} и \mathbf{v} , скажем

$$\hat{\theta} = S(\mathbf{g}, \mathbf{v}). \quad (15.13)$$

Например, $\hat{\theta} = \bar{z} - \bar{y}$ можно выразить как

$$\hat{\theta} = \frac{1}{n} \sum_{g_i=z} v_i - \frac{1}{m} \sum_{g_i=y} v_i, \quad (15.14)$$

где $\sum_{g_i=z} v_i$ обозначает сумму v_i по значениям $i = 1, 2, \dots, N$, имеющим $g_i = z$.

Пусть \mathbf{g}^* обозначает любой из C_N^n возможных векторов состоящих из n значений z и m значений y и определяет *перестановочные репликации* значения $\hat{\theta}$,

$$\hat{\theta}^* = \hat{\theta}(\mathbf{g}^*) = S(\mathbf{g}^*, \mathbf{v}). \quad (15.15)$$

Имеется C_N^n перестановочных репликаций $\hat{\theta}^*$. Распределение, которое имеет вероятность $1/C_N^n$ для каждого из них, называется *перестановочным распределением* $\hat{\theta}$ или *перестановочным распределением* $\hat{\theta}^*$. Перестановочный ASL определяется как перестановочная вероятность того, что $\hat{\theta}^*$ превышает $\hat{\theta}$,

$$ASL_{\text{perm}} = \left\{ \text{Prob}_{\text{perm}} \hat{\theta}^* \geq \hat{\theta} \right\} = \#\{\hat{\theta}^* \geq \hat{\theta}\} / C_N^n. \quad (15.16)$$

Два определения ASL_{perm} в (15.16) идентичны из-за леммы о перестановке.

На практике ASL_{perm} обычно аппроксимируется методами Монте–Карло в соответствии с алгоритмом 15.1.

Перестановочный алгоритм очень похож на бутстреп алгоритм, показанный на рис. 6.1. Основное отличие состоит в том, что отбор проб осуществляется без повторения, а не с повторением.

Алгоритм 15.1

Вычисление двухвыборочной перестановочной тестовой статистики.

1. Выберите B независимых векторов $\mathbf{g}^*(1), \mathbf{g}^*(2), \dots, \mathbf{g}^*(B)$, каждый из которых состоит из n значений z и m значений y и каждый выбирается случайным образом из множества всех C_N^n возможных таких векторов. [B обычно составляет не менее 1000; см. таблицу (15.3).]
2. Оцените перестановочные репликации $\hat{\theta}$, соответствующие каждому перестановочному вектору

$$\hat{\theta}^*(b) = S(\mathbf{g}^*(b), \mathbf{v}), \quad b = 1, 2, \dots, B. \quad (15.17)$$

3. Оцениваем ASL_{perm}

$$\widehat{ASL}_{\text{perm}} = \#\{\hat{\theta}^*(b) \geq \hat{\theta}\} / B. \quad (15.18)$$

На верхней левой части рисунка 15.1 показана гистограмма для $B = 1000$ перестановочных репликаций разности средни $\hat{\theta} = \bar{z} - \bar{y}$, (15.3); 132 из 1000 репликаций $\hat{\theta}^*$ превышают $\hat{\theta} = 30.63$, поэтому это подтверждает наш предыдущий вывод о том, что данные в таблице 2.1 не гарантируют отказ от гипотезы $F = G$

$$\widehat{ASL}_{\text{perm}} = 132 / 1000 = 0.132. \quad (15.19)$$

Перестановочный ASL близок к t -критерию ASL, (15.11), даже несмотря на то, что нет никаких предположений о нормальности, *подчеркивающих* $\widehat{ASL}_{\text{perm}}$. Это не случайно, хотя очень небольшая разница между (15.19) и (15.11) отчасти случайна. Фишер продемонстрировал тесную теоретическую связь между перестановочным тестом на основе $\bar{z} - \bar{y}$ и критерием Стьюдента. Его основной целью при введении перестановочных тестов было поддержание использования теста Стьюдента в нестандартных приложениях.

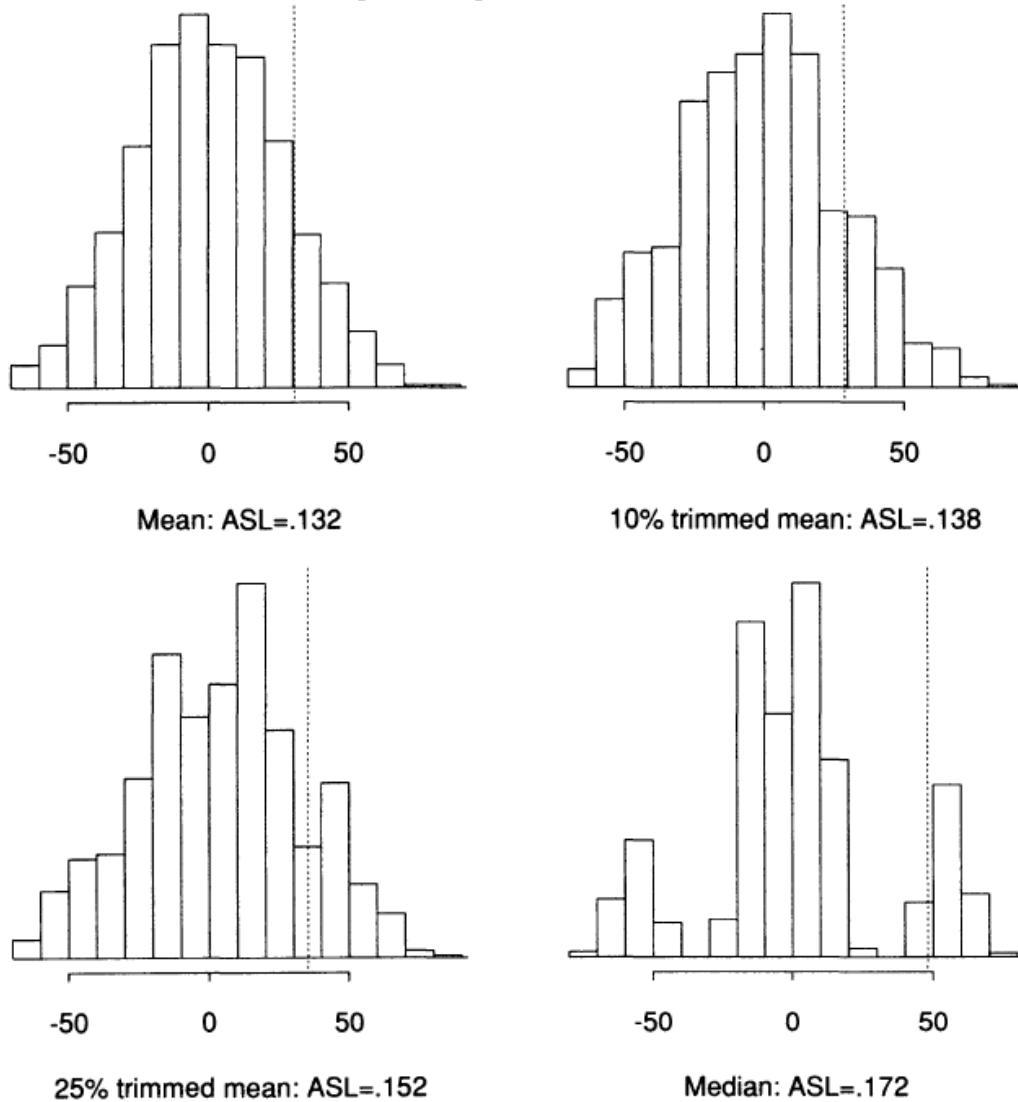


Figure 15.1. Permutation distributions for four different test statistics $\hat{\theta}$, mouse data, Table 2.1; dashed line indicates observed value of $\hat{\theta}$; $\widehat{ASL}_{\text{perm}}$ leads to non-rejection of the null hypothesis for all four statistics. Top left: $\hat{\theta} = \bar{z} - \bar{y}$, difference of means, Treatment-Control groups. Top right: $\hat{\theta}$ equals the difference of 15% trimmed means. Bottom left: difference of 25% trimmed means. Bottom right: difference of medians.

Сколько требуется перестановочных репликаций? Для удобного обозначения пусть $A = ASL_{\text{perm}}$ и $\hat{A} = \widehat{ASL}_{\text{perm}}$. Тогда $B \cdot \hat{A}$ равно количеству значений $\hat{\theta}^*(b)$, превышающих наблюдаемое значение $\hat{\theta}$, и поэтому имеет биномиальное

распределение, как в задаче 3.6,

$$B \cdot \hat{A} \sim \text{Bin}(B, A); \quad \text{E}(\hat{A}) = A; \quad \text{var}(\hat{A}) = \frac{A(1 - A)}{B}. \quad (15.20)$$

Table 15.2. $[(1 - A)/A]^{1/2}$ as a function of A .

| $A:$ | .5 | .25 | .1 | .05 | .025: |
|----------------------|------|------|------|------|-------|
| $[(1 - A)/A]^{1/2}:$ | 1.00 | 1.73 | 3.00 | 4.36 | 6.24 |

(Помните, что $\hat{\theta}$ является фиксированной величиной в (15.18), только $\hat{\theta}^*$ является случайным.) Коэффициент вариации \hat{A} равен

$$\text{cv}_B(\hat{A}) = \left[\frac{(1 - A)/A}{B} \right]^{1/2}. \quad (15.21)$$

Величина $[(1 - A)/A]^{1/2}$ становится больше по мере уменьшения A , как показано в таблице 15.2.

Предположим, мы требуем, чтобы $\text{cv}_B(\hat{A})$ было 0.10, что означает, что мы не хотим, чтобы ошибка Монте–Карло влияла на нашу оценку ASL_{perm} более чем на 10%. В таблице 15.3 указано количество требуемых перестановочных репликаций B .

Читателя может беспокоить особенность перестановочного тестирования: перестановочные репликации $\hat{\theta}^* = S(\mathbf{g}^*, \mathbf{v})$ изменяют часть исходных данных, но оставляют другую часть неизменной. Почему мы должны пересчитать \mathbf{g} , но не \mathbf{v} ? В статистической литературе приводятся некоторые веские теоретические доводы, но основная причина — практическая. «Условие на \mathbf{v} », то есть сохранение \mathbf{v} фиксированным в перестановочном процессе сводит двухвыборочную ситуацию к единственному распределению при нулевой гипотезе $F = G$. Это суть леммы о перестановке. Величина $\text{ASL}_{\text{perm}} = \text{Prob}_{\text{perm}} \hat{\theta}^* > \hat{\theta}$ четко определена, хотя, возможно, ее трудно вычислить, потому что $\text{Prob}_{\text{perm}}$ относится к уникальному распределению вероятностей. Величина $\text{ASL} = \text{Prob}_{H_0} \hat{\theta}^* > \hat{\theta}$ не определена должным образом, потому что не существует единого распределения Prob_{H_0} .

Самым большим достоинством перестановочного тестирования является ее точность. Если $H_0 : F = G$ истинно, вероятность того, что ASL_{perm} будет меньше 0.05, составляет почти 5%. В общем виде

$$\text{Prob}_{H_0}\{\text{ASL}_{\text{perm}} < \alpha\} = \alpha \quad (15.22)$$

для любого значения α от 0 до 1, за исключением небольших расхождений, вызванных дискретностью перестановочного распределения. Это важно, потому что интерпретирующая шкала (15.5) во многих областях применения понимается буквально.

15.3 Прочие тестовые статистики

Перестановочные тесты точны при применении любой тестовой статистики $\hat{\theta}$. Верхняя правая часть рисунка 15.1 отображает разницу 15% средних

$$\hat{\theta} = \bar{z}_{0.15} - \bar{y}_{0.15}. \quad (15.23)$$

Table 15.3. Number of permutations required to make $\widehat{ASL} \leq .10$, as a function of the achieved significance level.

| $ASL_{perm}:$ | .5 | .25 | .1 | .05 | .025 |
|---------------|-----|-----|-----|------|------|
| $B:$ | 100 | 299 | 900 | 1901 | 3894 |

Нижняя левая часть рисунка отображает разницу 25% средних значений, а нижняя правая часть — разницу медиан. Однаковое число $B = 1000$ перестановочных векторов \mathbf{g}^* используется во всех четырех частях рисунка, изменяется только статистика $\hat{\theta}^* = S(\mathbf{g}^*, \mathbf{v})$. Все четыре значения \widehat{ASL}_{perm} , 0.132, 0.138, 0.152 и 0.172 согласуются с принятием нулевой гипотезы $F = G$.

Тот факт, что каждое $\hat{\theta}$ приводит к точному ASL_{perm} , не означает, что все $\hat{\theta}$ являются одинаково хорошей тестовой статистикой. «Точность» означает, что эти ASL_{perm} не будут иметь тенденцию быть обманчиво маленькими, когда H_0 истинно, как указано в (15.22). Однако, если H_0 ложно, если экспериментальная группа дает действительно лучшие результаты, чем контрольная, тогда мы хотим, чтобы ASL_{perm} был маленьким. Это свойство статистического теста называется мощностью. Наказанием за выбор плохой тестовой статистики $\hat{\theta}$ является малая мощность — мы не получаем большой вероятности отклонить H_0 , когда оно ложно. Мы скажем немного больше о выборе $\hat{\theta}$ в обсуждении бутстрепа, который завершает эту главу.

Глядя на Таблицу 2.1, эти две группы, кажется, различаются больше по дисперсии, чем по средним. Отношение оценочных дисперсий составляет около 2.5

$$\hat{\sigma}_z^2/\hat{\sigma}_y^2 = 2.48. \quad (15.24)$$

Является ли эта разница значимой или это просто артефакт небольшого размера выборки?

Мы можем ответить на этот вопрос с помощью перестановочного теста. На рисунке 15.2 показано 1000 перестановочных репликаций

$$\hat{\theta} = \log(\hat{\sigma}_z^2/\hat{\sigma}_y^2). \quad (15.25)$$

(Логарифм не влияет на перестановочные результаты). 152 из 1000 значений $\hat{\theta}^*$ превышают $\hat{\theta} = \log(2.48) = 0.907$, что дает $\widehat{ASL}_{perm} = 0.152$. И снова нет оснований отвергать нулевую гипотезу $F = G$. Обратите внимание, что мы *могли* отклонить H_0 с помощью этой $\hat{\theta}$, даже если мы не отклонили ее с помощью $\hat{\theta} = \bar{z} - \bar{y}$. Этот $\hat{\theta}$ измеряет отклонения от H_0 другим способом, нежели $\hat{\theta}$ на рисунке 15.1.

Статистика $\log(\hat{\sigma}_z^2/\hat{\sigma}_y^2)$ существенно отличается от $\bar{z} - \bar{y}$. Лечение было разработано для увеличения времени выживания, поэтому мы ожидаем, что $\bar{z} - \bar{y}$ будет больше нуля, если лечение работает, то есть если H_0 ложно. С другой стороны, у нас нет априорных оснований полагать, что $\hat{\theta} = \log(\hat{\sigma}_z^2/\hat{\sigma}_y^2)$ будет больше нуля, а не меньше нуля, если H_0 ложно. Другими словами, нас бы так же интересовал результат $\hat{\theta} = -\log(2.48)$, как и результат $\hat{\theta} = \log(2.48)$.

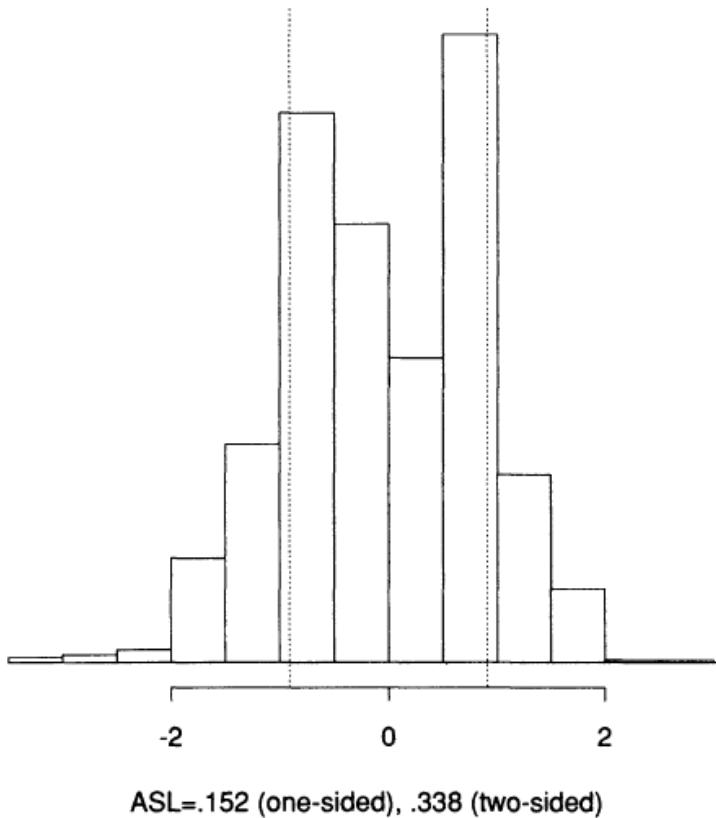


Figure 15.2. $B = 1000$ permutation replications of the log variance ratio $\hat{\theta} = \log(\hat{\sigma}_z^2/\hat{\sigma}_y^2)$ for the mouse data of Table 2.1; 152 of the 1000 replications gave $\hat{\theta}^*$ greater than the observed value $\hat{\theta} = .907$; 338 of the 1000 replications gave either $\hat{\theta}^*$ or $-\hat{\theta}^*$ greater than .907. The dashed lines indicate $\hat{\theta}$ and $-\hat{\theta}$.

В этой ситуации обычно вычисляется *двусторонний* ASL, а не *односторонний* ASL (15.4). Это делается путем сравнения абсолютного значения $\hat{\theta}^*$ с абсолютным значением $\hat{\theta}$

$$\widehat{ASL}_{\text{perm}}(\text{двусторонний}) = \#\{|\hat{\theta}^*(b)| > |\hat{\theta}|\}/B. \quad (15.26)$$

Точно так же мы подсчитываем случаи, когда либо $\hat{\theta}^*$, либо $-\hat{\theta}^*$ превосходят $|\hat{\theta}|$. Двусторонний ASL всегда больше одностороннего ASL, что дает меньше оснований для отклонения H_0 . Двусторонний тест по своей сути более консервативен. Для данных о мышах, статистика (15.25) дала двусторонний ASL в размере 0.338.

Идея теста значимости может быть сформулирована следующим образом: мы ранжируем все возможные наборы данных \mathbf{x} в соответствии с тем, насколько сильно они противоречат нулевой гипотезе H_0 ; тогда мы отклоняем H_0 , если \mathbf{x} входит в 5% (или 10%, или 1% и т.д., как в (15.5)) набора данных, которые наиболее сильно противоречат H_0 . Определение ASL в (15.4) сводится к измерению противоречия в соответствии с размером $\hat{\theta}(\mathbf{x})$, большие значения $\hat{\theta}$ подразумевают большие доказательства против H_0 . Иногда, однако, мы считаем, что большие отрицательные значения $\hat{\theta}$ также хороши, как большие положительные значения для дискредитации H_0 . Так было в (15.25). Если так, нам необходимо принять это во внимание при определении 5% набора данных,

которые наиболее сильно противоречат H_0 . Это точка в определении (15.26) двустороннего ASL.

Есть много других ситуаций, в которых нам нужно быть осторожными при ранжировании свидетельств против H_0 . Предположим, например, что мы запускаем четыре перестановочных теста, показанных на рисунке 15.1, и решаем выбрать тот, у которого наименьшее значение ASL, в данном случае $\widehat{\text{ASL}} = 0.132$. Тогда мы действительно ранжируем доказательства по \mathbf{x} против H_0 согласно статистике

$$\hat{\varphi}(\mathbf{x}) = \min_k \{\widehat{\text{ASL}}_k\}, \quad (15.27)$$

где $\widehat{\text{ASL}}_k$ — перестановочный ASL для k -й статистики $\hat{\theta}$ для $k = 1, 2, 3, 4$. Малые значения $\hat{\varphi}$ сильнее противоречат H_0 . *Неверно*, что, наблюдая $\hat{\theta} = 0.132$, перестановочный ASL основанный на $\hat{\varphi}$ равняется 0.132. Более 13,2% перестановок будут иметь $\hat{\varphi}^* < 0.132$ из-за минимизации в определении (15.27).

Вот как вычислить правильный перестановочный ASL для $\hat{\varphi}$, используя все 4000 перестановочных репликаций $\hat{\theta}_k^*(b)$ на рисунке 15.1, $k = 1, 2, 3, 4, b = 1, 2, \dots, 1000$. Для каждого значения k и b определяют

$$A_k^*(b) = \frac{1}{1000} \sum_{i=1}^B I_{\{\theta_k^*(i) \geq \hat{\theta}_k^*(b)\}}, \quad (15.28)$$

где $I_{\{\cdot\}}$ — индикаторная функция. Итак, $A_k^*(b)$ — это доля значений $\hat{\theta}^*$, превышающих $\hat{\theta}_k^*(b)$. Тогда пусть

$$\hat{\varphi}^*(b) = \min_k \{A_k^*(b)\}. \quad (15.29)$$

Это не очевидно, но верно, что $\hat{\varphi}^*(b)$ являются подлинными перестановочными репликациями, (15.27), поэтому перестановочные ASL для $\hat{\varphi}$ есть

$$\widehat{\text{ASL}}_{\text{perm}} = \#\{\hat{\varphi}^*(b) \leq \hat{\varphi}\}/1000. \quad (15.30)$$

На рисунке 15.3 показана гистограмма 1000 значений $\hat{\varphi}^*(b)$. 167 из 1000 значений меньше $\hat{\varphi} = 0.132$, что дает перестановочный ASL = 0.167.

15.4 Связь проверки гипотез с доверительными интервалами и бутстрепом

Существует тесная связь между проверкой гипотез и доверительными интервалами. Предположим, что наблюдаемое значение интересующей статистики $\hat{\theta}$ больше нуля. Выберем α так, чтобы $\hat{\theta}_{lo}$, нижний предел доверительного интервала $1 - 2\alpha$ для θ , в точности равнялся 0.

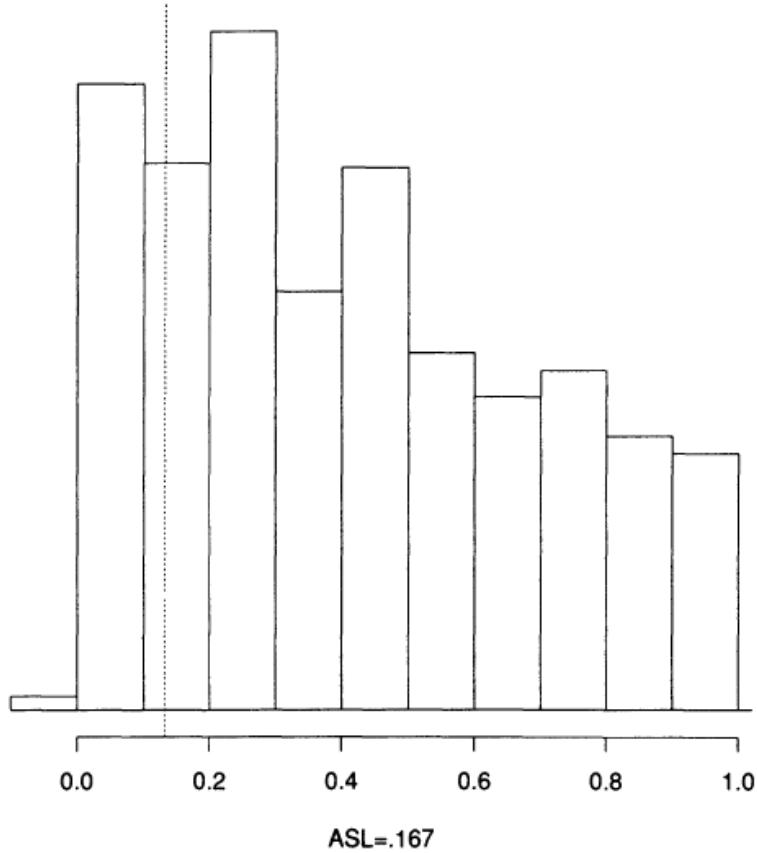


Figure 15.3. *Permutation distribution for the minimum ASL statistic (15.27); based on the 1000 permutations used in Figure 15.1; dashed line indicates $\hat{\phi} = .132$; 167 of the 1000 $\hat{\phi}^*$ values are less than .132, so $\widehat{ASL}_{\text{perm}} = .167$.*

Тогда $\text{Prob}_{\theta=0}\{\hat{\theta}^* \geq \hat{\theta}\} = \alpha$ согласно (12.13). Однако если $\theta = 0$ — нулевая гипотеза, как в примере с данными о мышах, то определение (15.4) дает $\text{ASL} = \alpha$. Например, если 0.94 доверительный интервал $[\hat{\theta}_{lo}, \hat{\theta}_{up}]$ имеет $\hat{\theta}_{lo} = 0$, тогда ASL наблюдаемого значения $\hat{\theta}$ должен быть равен 0.03 (поскольку $0.94 = 1 - 2 \cdot 0.03$).

Другими словами, мы можем использовать доверительные интервалы для вычисления ASL . Имея это в виду, на рисунке 15.4 показано бутстреп распределение двух статистик, которые мы можем использовать для формирования доверительных интервалов для разницы между экспериментальной и контрольной группами, разности средних $\hat{\theta}_0 = \bar{z} - \bar{y}$ (левая часть рисунка), и 0.25 отсеченных разностей средних $\hat{\theta}_{0.25} = \bar{z}_{0.25} - \bar{y}_{0.25}$ (правая часть рисунка). Какое значение α сделает нижний предел бутстреп доверительного интервала равным нулю? Для бутстреп метода процентилей, примененного к статистике $\hat{\theta}^*$, ответ составляет

$$\alpha_0 = \#\{\hat{\theta}^*(b) < 0\}/B, \quad (15.31)$$

доля бутстреп репликаций меньше нуля. (Тогда $\hat{\theta}_{lo} = \hat{\theta}_{\alpha_0}^* = 0$ согласно (13.5).) Согласно предыдущему абзацу ASL θ равно α_0 , скажем

$$\widehat{ASL}_{\%} = \#\{\hat{\theta}^*(b) < 0\}/B. \quad (15.32)$$

$B = 1000$ бутстреп репликации, показанные на рис. 15.4, дали

$$\widehat{\text{ASL}}\%(\hat{\theta}_0) = 0.132 \quad \text{и} \quad \widehat{\text{ASL}}\%(\hat{\theta}_{0.25}) = 0.180. \quad (15.33)$$

Обратите внимание, насколько эти результаты похожи на $\widehat{\text{ASL}}_{\text{perm}}(\hat{\theta}_0) = 0.132$, $\widehat{\text{ASL}}_{\text{perm}}(\hat{\theta}_{0.25}) = 0.152$, рисунок 15.1.

Для доверительных интервалов BC_α для главы 14 расчет ASL дает

$$\widehat{\text{ASL}}_{\text{BC}_\alpha} = \Phi^{-1} \left(\frac{\omega_0 - \hat{z}_0}{1 + \hat{\alpha}(\omega_0 - \hat{z}_0)} \right), \quad (15.34)$$

где

$$\omega_0 = \Phi^{-1}(\alpha_0) \quad (15.35)$$

и константа коррекции смещения \hat{z}_0 аппроксимируется согласно формуле (14.14). Эта формула дала $\hat{z}_0 = -0.040$ для $\hat{\theta}_0$ и $\hat{z}_0 = 0.035$ для $\hat{\theta}_{0.25}$.

Константа ускорения $\hat{\alpha}$ определяется двухвыборочной версией формулы (14.15). Пусть $\hat{\theta}_{z,(i)}$ будет значением $\hat{\theta}$, когда мы исключаем z_i , и $\hat{\theta}_{y,(i)}$ будет значением $\hat{\theta}$, когда мы исключаем y_i . Пусть

$$\hat{\theta}_{z,(\cdot)} = \sum_1^n \hat{\theta}_{z,(i)} / n,$$

$$\hat{\theta}_{y,(\cdot)} = \sum_1^m \hat{\theta}_{y,(i)} / m,$$

$$U_{z,i} = (n-1)(\hat{\theta}_{z,(\cdot)} - \hat{\theta}_{z,(i)}),$$

$$U_{y,i} = (m-1)(\hat{\theta}_{y,(\cdot)} - \hat{\theta}_{y,(i)}).$$

Тогда

$$\hat{\alpha} = \frac{1}{6} \frac{[\sum_{i=1}^n U_{z,i}^3 / n^3 + \sum_{i=1}^m U_{y,i}^3 / m^3]}{[\sum_{i=1}^n U_{z,i}^2 / n^2 + \sum_{i=1}^m U_{y,i}^2 / m^2]^{3/2}}, \quad (15.36)$$

n и m — длины \mathbf{z} и \mathbf{y} . Формула (15.36) дает $\hat{\alpha} = 0.06$ и $\hat{\alpha} = -0.01$ для $\hat{\theta}_0$ и $\hat{\theta}_{0.25}$ соответственно. Тогда

$$\widehat{\text{ASL}}_{\text{BC}_a}(\hat{\theta}_0) = 0.147 \quad \text{и} \quad \widehat{\text{ASL}}_{\text{BC}_a}(\hat{\theta}_{0.25}) = 0.167 \quad (15.37)$$

согласно (15.34).

Вот некоторые моменты, о которых следует помнить при сравнении рисунков 15.1 и 15.4:

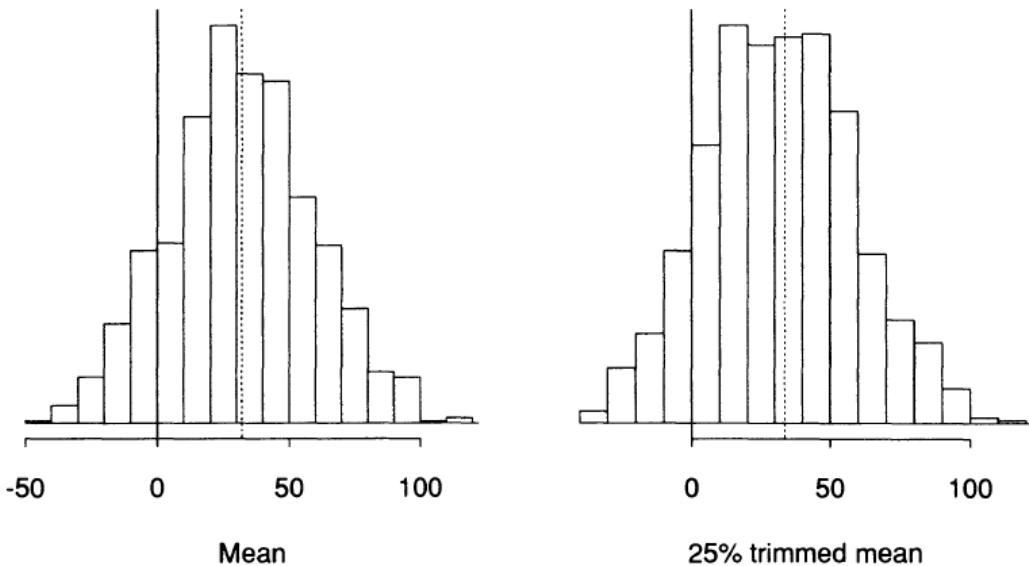


Figure 15.4. $B = 1000$ bootstrap replications of the mean difference $\hat{\theta}$ for the mouse data, left panel, and the difference of the .25 trimmed means $\hat{\theta}_{.25}$, right panel; dashed lines are the observed estimates $\hat{\theta}_0 = 30.23$ and $\hat{\theta}_{.25} = 33.66$; 132 of the 1000 $\hat{\theta}_0^*$ values were less than zero; 180 of the 1000 $\hat{\theta}_{.25}^*$ values were less than zero.

- Перестановочный ASL является точным, тогда как бутстреп ASL является приблизительным. Однако на практике оба метода часто дают очень похожие результаты, как в данном случае.
- Гистограммы бутстрепа центрированы около 0, в то время как гистограммы перестановок сосредоточены около 0. В этом смысле, ASL_{perm} измеряет, насколько далека наблюдаемая оценка $\hat{\theta}$ от 0, в то время как бутстреп ASL измеряет, насколько далек 0 от $\hat{\theta}$. BC_a -метод, относящийся к методу процентиля (15.34) по сравнению с (15.31), предназначен для согласования этих двух способов измерения статистического «расстояния».
- Бутстреп ASL проверяет нулевую гипотезу $\theta = 0$, а перестановочный ASL проверяет $F = G$. Последний более особенный, чем первый, и иногда может показаться нереалистичным. Для данных о мышах мы могли бы проверить гипотезу о том, что средние значения двух групп равны, $\theta_0 = 0$, даже не будучи уверенными, например, в том, что два распределения имеют одинаковую дисперсию. Это скорее теоретическое возражение, нежели практическое возражение против перестановочных тестов, которые обычно работают достаточно хорошо, даже если $F = G$ является далеко не самой разумной нулевой гипотезой.
- Стандартное отклонение перестановочного распределения *не* является надежной оценкой стандартной ошибки для $\hat{\theta}$ (оно не предназначено для этого), в то время как бутстреп стандартное отклонение является. В таблице 15.4 показаны стандартные отклонения перестановочного распределения и бутстреп распределения данных о мышах

Table 15.4. Standard deviations of the mouse data permutation and bootstrap distributions for $\hat{\theta}_0 = \bar{z} - \bar{y}$, $\hat{\theta}_{.15} = \bar{z}_{.15} - \bar{y}_{.15}$, $\hat{\theta}_{.25} = \bar{z}_{.25} - \bar{y}_{.25}$, and $\hat{\theta}_{.5} = \bar{z}_{.5} - \bar{y}_{.5}$.

| | $\hat{\theta}_0$ | $\hat{\theta}_{.15}$ | $\hat{\theta}_{.25}$ | $\hat{\theta}_{.5}$ |
|--------------|------------------|----------------------|----------------------|---------------------|
| permutation: | 27.9 | 28.6 | 30.8 | 33.5 |
| bootstrap: | 27.0 | 29.9 | 33.4 | 40.8 |

для $\hat{\theta}_0 = \bar{z} - \bar{y}$, $\hat{\theta}_{.15} = \bar{z}_{.15} - \bar{y}_{.15}$, $\hat{\theta}_{.25} = \bar{z}_{.25} - \bar{y}_{.25}$, $\hat{\theta}_{.5} = \bar{z}_{.5} - \bar{y}_{.5}$. Бутстреп числа показывают более быстрое увеличение стандартной ошибки при увеличении пропорции обрезки от 0 до 0.5.

- Комбинация точечной оценки и доверительного интервала обычно более информативна, чем просто проверка гипотезы сама по себе. В эксперименте на мышах значение $\widehat{ASL}_{\text{perm}} = 0.132$ говорит нам только о том, что мы не можем исключить $\theta = 0$. На левой части рисунка 15.4 показано, что истинное среднее лежит между -14.5 и 73.8 с достоверностью 0.90, метод BC_a. По опыту авторов, проверка гипотез, как правило, используется чрезмерно, а доверительные интервалы — недостаточно в статистических приложениях.

Перестановочные методы, как правило, применимы только к узкому кругу задач. Однако когда они применяются, как при тестировании $F = G$ в задаче с двумя выборками, они дают точные ответы без параметрических предположений. Первоначально бутстреп распределение называлось «комбинированным распределением». Оно было разработано, чтобы расширить возможности перестановочного тестирования для большего охвата статистических задач, где нечего переставлять. Когда есть что-то, что нужно переставить, как на рис. 15.1, это хорошая идея, даже если действуются и другие методы, такие как бутстреп. В следующей главе мы обсудим задачи, для которых метод перестановки не может быть применен, но бутстреп проверка гипотезы все еще может быть использована.

15.5 Библиографические примечания

Перестановочные тесты описаны во многих книгах. Исчерпывающий обзор дан Эджингтоном (1987). Норин (1989) дает введение в перестановочные тесты и связывает их с бутстрепом.

Глава 16

Проверка гипотез с помощью бутстрепа

16.1 Введение

В главе 15 описывается перестановочный тест, полезный инструмент для проверки гипотез. В конце предыдущей главы мы связали проверку гипотез с доверительными интервалами и, в частности, показали, как можно использовать бутстреп доверительный интервал чтобы гарантировать достижение уровня значимости для проверки гипотез. В этой главе мы описываем бутстреп методы, которые предназначены непосредственно для проверки гипотез. Мы увидим, что бутстреп тесты дают те же результаты, что и перестановочные тесты, в случае когда оба могут быть применены. Бутстреп тесты имеют более широкое применение, хотя и менее точны.

16.2 Двухвыборочная задача

Начнем с задачи с двумя выборками, описанной в предыдущей главе. Имеются выборки \mathbf{z} и \mathbf{y} из, возможно, различных распределений F и G , и мы хотим проверить нулевую гипотезу $H_0 : F = G$. Бутстреп проверка гипотез основана на тестовой статистике, как и проверка гипотез с помощью перестановочных тестов. В предыдущей главе тестовая статистика обозначалась как $\hat{\theta}$. Чтобы подчеркнуть, что тестовая статистика не обязательно должна быть оценкой параметра, обозначим ее здесь как $t(\mathbf{x})$. В примере с данными о мышах $t(\mathbf{x}) = \bar{z} - \bar{y}$, разница средних наблюдаемых значений 30.63. Мы стремимся достигнуть уровень значимости

$$\text{ASL} = \text{Prob}_{H_0} \{t(\mathbf{x}^*) \geq t(\mathbf{x})\} \quad (16.1)$$

как в (15.4). Величина $t(\mathbf{x})$ фиксируется на своем наблюдаемом значении, а случайная величина \mathbf{x}^* имеет распределение, заданное нулевой гипотезой H_0 . Назовем это распределение F_0 . Теперь вопрос состоит в том, что такое F_0 ? В перестановочном teste в предыдущей главе мы зафиксировали порядковую статистику \mathbf{v} и определили F_0 как распределение возможных перестановок рангов \mathbf{g} . С другой стороны, бутстреп проверка гипотез использует оценку F_0 , полученную методом подстановки. Обозначим объединенную выборку \mathbf{x} и пусть ее эмпирическое распределение будет \hat{F}_0 , полагая вероятность каждого элемента \mathbf{x} равной $1/(n+m)$. Под H_0 , \hat{F}_0 обеспечивает непараметрическую оценку общей

популяции, которая порождает как \mathbf{z} , так и \mathbf{y} . Алгоритм 16.1 показывает, как вычисляется ASL.

Алгоритм 16.1

1. Получить B выборок размера $n + m$ с возвращением из \mathbf{x} . Обозначить первые n наблюдений \mathbf{z}^* , а оставшиеся m наблюдений \mathbf{y}^* .
2. Оценить $t(\cdot)$ для каждой выборки,

$$t(\mathbf{x}^{*b}) = \bar{\mathbf{z}}^* - \bar{\mathbf{y}}^*, \quad b = 1, 2, \dots, B. \quad (16.2)$$

3. Аппроксимировать ASL_{boot} с помощью

$$\widehat{\text{ASL}}_{\text{boot}} = \# \{ t(\mathbf{x}^{*b}) \geq t_{\text{obs}} \} / B, \quad (16.3)$$

где $t_{\text{obs}} = t(\mathbf{x})$ наблюданное значение статистики.

Обратите внимание, что единственное различие между этим алгоритмом и перестановочным алгоритмом в уравнениях (15.17) и (15.18) состоит в том, что выборки осуществляются с возвращением, а не без него. Неудивительно, что он дает очень похожие результаты (левая часть рисунка 16.1). Была сгенерирована тысяча бутстреп выборок, и 120 имели $t(\mathbf{x}^*) > 30.63$. Значение $\widehat{\text{ASL}}_{\text{boot}}$ составляет $120/1000 = 0.12$ в отличие от 0.152 в перестановочном тесте.

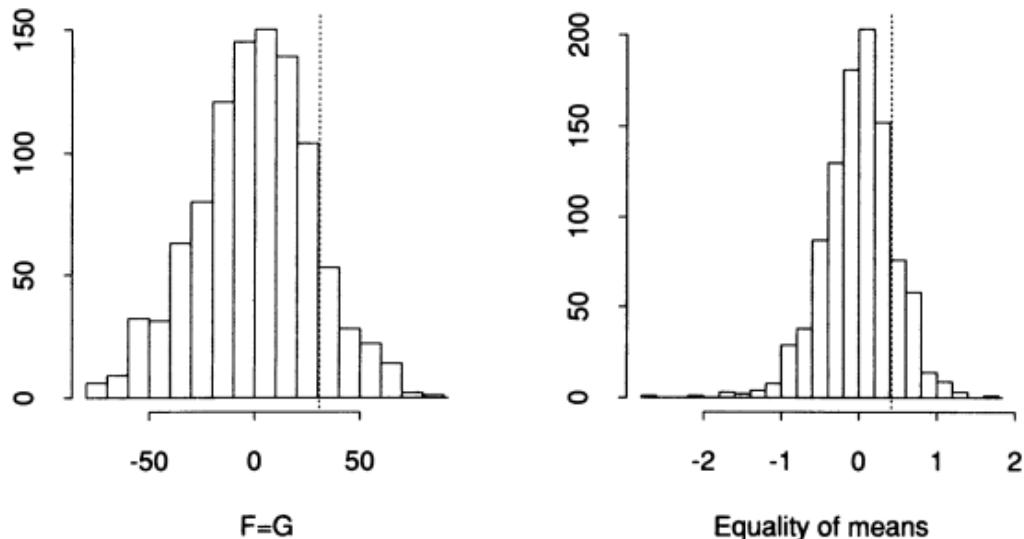


Figure 16.1. Histograms of bootstrap replications for the mouse data example. The left panel is a histogram of bootstrap replications of $\bar{z} - \bar{y}$ for the test of $H_0 : F = G$, while the right panel is a histogram of bootstrap replications of the studentized statistic (16.5) for the test of equality of means. The dotted lines are drawn at the observed values (30.63 on the left, .416 on the right). In the left panel, $\widehat{\text{ASL}}_{\text{boot}}$ (the bootstrap estimate of the achieved significance level) equals .120, the proportion of values greater than 30.63. In the right panel, $\widehat{\text{ASL}}_{\text{boot}}$ equals .152.

Более точный критерий можно получить за счет использования стьюдентизированной статистики. В приведенном выше тесте вместо $t(\mathbf{x}) = \bar{z} - \bar{y}$ мы могли бы использовать

$$t(\mathbf{x}) = \frac{\bar{z} - \bar{y}}{\bar{\sigma} \sqrt{1/n + 1/m}}, \quad (16.4)$$

где $\bar{\sigma} = \left\{ \left[\sum_{i=1}^n (z_i - \bar{z})^2 + \sum_{j=1}^m (y_j - \bar{y})^2 \right] / [n + m - 2] \right\}^{1/2}$. Это двухвыборочная статистика t , описанная в главе 15. Наблюдаемое значение $t(\mathbf{x})$ составило 1.12. Повторение вышеупомянутого бутстрап алгоритма с использованием $t(\mathbf{x}^*)$, определенной 16.4, привело к 134 значениям из 1000, превышающих 1.12, и, следовательно, $\text{ASL}_{\text{boot}} = 0.134$. В этом вычислении мы использовали точно такой же набор бутстрап выборок, которые дали значение 0.12 для $\widehat{\text{ASL}}_{\text{boot}}$ на основе $t(\mathbf{x}) = \bar{z} - \bar{y}$. В отличие от перестановочного теста, где мы показали, что стьюдентизация не влияет на ответ, стьюдентизация действительно приводит к другому значению $\widehat{\text{ASL}}_{\text{boot}}$. Однако в этом конкретном бутстрап подходе для задачи с двумя выборками разница обычно довольно мала.

Алгоритм 16.1 проверяет нулевую гипотезу о том, что две популяции идентичны, то есть $F = G$. Что, если бы мы хотели проверить только то, равны ли их средние значения? Один из подходов — использовать двухвыборочную t -статистику 16.4. При нулевой гипотезе и в предположении нормально распределенных популяций с равными дисперсиями она имеет распределение Стьюдента с $n + m - 2$ степенями свободы. Она использует объединенную оценку стандартной ошибки $\bar{\sigma}$. Если мы не хотим предполагать, что дисперсии в двух популяциях равны, мы могли бы основывать тест на

$$t(\mathbf{x}) = \frac{\bar{z} - \bar{y}}{\sqrt{\hat{\sigma}_1^2/n + \hat{\sigma}_2^2/m}}, \quad (16.5)$$

где $\hat{\sigma}_1^2 = \sum_1^n (z_i - \hat{z})^2 / (n - 1)$, $\hat{\sigma}_2^2 = \sum_1^m (y_i - \hat{m})^2 / (m - 1)$. При генеральной совокупности имеющей нормальное распределение величина 16.5 больше не имеет t -распределение Стьюдента, и поэтому был предложен ряд приближенных решений. В литературе это известно как проблема Беренса-Фишера.

Предположение о равной дисперсии привлекательно для t -критерия, поскольку оно упрощает форму результирующего распределения. При рассмотрении бутстрап теста для сравнения двух средних нет веских причин предполагать равные дисперсии, и, следовательно, мы не делаем этого предположения. Для продолжения нам потребуются оценки F и G , в которых используется только предположение об общем среднем значении. Допустим, что \bar{x} будет средним значением объединенной выборки, мы можем преобразовать обе выборки так, чтобы они имели среднее значение \bar{x} , а затем ресэмплировать каждую совокупность по отдельности. Процедура подробно описана в алгоритме 16.2.

Результаты этого показаны на правой части рисунка 16.1. Значение $\widehat{\text{ASL}}_{\text{boot}}$ было $152/1000 = 0.152$.

Вычисление бутстреп статистики для проверки гипотезы равенстве средних

1. Пусть \hat{F} задает одинаковую вероятность для точек $\tilde{z}_i = z_i - \bar{z} + \bar{x}$, $i = 1, 2, \dots, n$, а \hat{G} задает одинаковую вероятность для точек $\tilde{y}_i = y_i - \bar{y} + \bar{x}$, $i = 1, 2, \dots, m$, где \bar{z} и \bar{y} — групповое среднее, а \bar{x} — средние по объединенной выборке.
2. Сформировать B бутстреп выборок $(\mathbf{z}^*, \mathbf{y}^*)$, где \mathbf{z}^* выборка с возвращением из $\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_n$ и \mathbf{y}^* выборка с возвращением из $\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_m$.
3. Вычислить $t(\cdot)$, определенное в 16.5, для каждой выборки

$$t(\mathbf{x}^{*b}) = \frac{\bar{z}^* - \bar{y}^*}{\sqrt{\hat{\sigma}_1^{2*}/n + \hat{\sigma}_2^{2*}/m}}, \quad b = 1, 2, \dots, B. \quad (16.6)$$

4. Аппроксимировать ASL_{boot} с помощью

$$\widehat{\text{ASL}}_{\text{boot}} = \# \{ t(\mathbf{x}^{*b}) \geq t_{\text{obs}} \} / B, \quad (16.7)$$

где $t_{\text{obs}} = t(\mathbf{x})$ наблюдаемое значение статистики.

16.3 Связь перестановочного теста и бутстрепа

Предыдущий пример иллюстрирует некоторые важные различия между перестановочным тестом и бутстреп проверкой гипотез. Перестановочный тест использует особую симметрию, которая существует при выполнении нулевой гипотезы, для создания перестановочного распределения статистики критерия. Например, в задаче с двумя выборками при проверке гипотезы $F = G$ все перестановки порядковой статистики объединенной выборки равновероятны. В результате этой симметрии ASL из перестановочного теста является точным: в задаче с двумя выборками ASL_{perm} представляет собой точную вероятность получения такой экстремальной статистики теста, как наблюдаемая, с фиксированными значениями данных объединенной выборки.

Напротив, бутстреп явно оценивает вероятностный механизм при нулевой гипотезе, а затем генерирует выборку из него для оценки ASL. Оценка ASL_{boot} не интерпретируется как точная вероятность, но, как и все бутстреп оценки, точность гарантирована только тогда, когда размер выборки стремится к бесконечности. С другой стороны, бутстреп проверка гипотез не требует особой симметрии, которая требуется для проверки перестановочным методом, и поэтому может применяться гораздо более широко. Например, в задаче с двумя выборками тестирование гипотезы перестановочным методом может проверять только нулевую гипотезу $F = G$, в то время как бутстреп может проверять равенство средних и дисперсии или равенство средних при возможно неравных дисперсиях.

16.4 Задача с одной выборкой

В качестве нашего второго примера рассмотрим задачу с одной выборкой, включающую только мышей, подвергшихся лечению. Предположим, что другие исследователи провели эксперименты, аналогичные нашим, но с гораздо большим количеством мышей, и они наблюдали среднюю продолжительность жизни 129.0 дней для мышей подвергшихся лечению. Мы могли бы захотеть проверить, было ли среднее значение для группы лечения в таблице 2.1 также равно 129.0:

$$H_0 : \mu_z = 129.0. \quad (16.8)$$

Можно использовать одновыборочную версию нормального теста. Предполагая генеральную совокупность имеющую нормальное распределение, при нулевой гипотезе

$$\bar{z} \sim N(129.0, \sigma^2/n), \quad (16.9)$$

где σ — стандартное отклонение продолжительности лечения. Наблюдая $\bar{z} = 86.9$, ASL представляет собой вероятность того, что случайная величина \bar{z}^* , распределенная согласно 16.9, меньше наблюдаемого значения 86.9

$$ASL = \Phi\left(\frac{86.9 - 129.0}{\sigma/\sqrt{n}}\right), \quad (16.10)$$

где Φ — функция стандартного нормального распределения.

Поскольку σ неизвестно, подставим оценку

$$\bar{\sigma} = \left\{ \sum_1^n (z_i - \bar{z})^2 / (n - 1) \right\}^{1/2} = 66.8, \quad (16.11)$$

в 16.10, получим

$$ASL = \Phi\left(\frac{-42.1}{66.8/\sqrt{7}}\right) = 0.05. \quad (16.12)$$

T -критерий Стьюдента дает несколько больший ASL

$$ASL = \text{Prob}\left\{ t_6 < \frac{-42.1}{66.8/\sqrt{7}} \right\} = 0.07. \quad (16.13)$$

Таким образом, есть несущественные(???) доказательства того, что у мышей подвергшихся лечению в нашем исследовании среднее время выживания составляет менее 129.0 дней. Двусторонние значения ASL равны 0.10 и 0.14 соответственно.

Обратите внимание, что двухвыборочный перестановочный тест не может использоваться для этой задачи. Если бы у нас были доступны все времена для пролеченных мышей (а не только их среднее значение 129.0), мы могли бы провести двухвыборочный перестановочный тест на эквивалентность двух популяций. Однако у нас нет данных о всех временах, а мы знаем только их среднее значение и хотим проверить $H_0 : \mu_z = 129.0$.

Напротив, можно использовать бутстреп. Мы основываем бутстреп проверку гипотезы на распределении тестовой статистики

$$t(\mathbf{z}) = \frac{\bar{z} - 129.0}{\bar{\sigma}/\sqrt{7}} \quad (16.14)$$

при выполнении нулевой гипотезы $\mu_z = 129.0$. Наблюданное значение

$$\frac{86.9 - 129.0}{66.8/\sqrt{7}} = -1.67. \quad (16.15)$$

Но каково пригодное нулевое распределение? Нам нужно распределение F , которое оценивает совокупность продолжительности лечения при H_0 . Прежде всего отметим, что эмпирическое распределение \hat{F} не является подходящей оценкой для F , потому что оно не подчиняется H_0 . То есть среднее значение \hat{F} не равно нулевому значению 129.0. Каким-то образом нам нужно получить оценку генеральной совокупности, которая имеет среднее значение 129.0. Простой способ — преобразовать эмпирическое распределение \hat{F} так, чтобы оно имело желаемое среднее значение.¹ Другими словами, мы используем в качестве оценочного нулевого распределения эмпирическое распределение значений

$$\tilde{z}_i = z_i - \bar{z} + 129.0 = z_i + 42.1 \quad (16.16)$$

для $i = 1, 2, \dots, 7$. Мы делаем выборку $\tilde{z}_1^*, \dots, \tilde{z}_7^*$ с возвращением из $\tilde{z}_1, \dots, \tilde{z}_7$, и для каждой бутстреп выборки вычисляем статистику

$$t(\tilde{\mathbf{z}}^*) = \frac{\tilde{z}^* - 129.0}{\tilde{\sigma}^*/\sqrt{7}}, \quad (16.17)$$

где $\tilde{\sigma}^*$ — стандартное отклонение бутстреп выборки. Всего в 100 из 1000 выборок имеют $t(\tilde{\mathbf{z}}^*)$ меньше чем -1.67 , и, следовательно, достигнутый уровень значимости составляет $100/1000 = 0.10$ в отличие от 0.05 и 0.07 для нормального и t -теста соответственно.

Обратите внимание, что наш выбор нулевого распределения предполагает, что возможные распределения времени лечения, поскольку среднее время меняется, являются просто преобразованными версиями друг друга. Такое семейство распределений называется семейством преобразований(???). Это предположение также присутствует в нормальном и t -тестах; но в этих тестах мы предполагаем, что генеральные совокупности имеют нормальное распределение. В любом случае, может быть разумным логарифмировать время выживания перед проведением анализа, потому что зарегистрированное время жизни с большей вероятностью удовлетворит предположению о семействе преобразований или нормальному семейству.

Существует другой, но эквивалентный метод бутстрепа для задачи с одной выборкой. Мы делаем выборку с возвращением из (не подвергшихся преобразованию) значений z_1, z_2, \dots, z_7 и вычисляем статистику

$$t(\mathbf{z}^*) = \frac{\bar{z}^* - 129.0}{\bar{\sigma}^*/\sqrt{7}}, \quad (16.18)$$

где $\bar{\sigma}^*$ — стандартное отклонение бутстреп выборки. Эта статистика такая же, как 16.17, поскольку

$$\tilde{z}^* - 129.0 = (\bar{z}^* - \bar{z} + 129.0) - 129.0 = \bar{z}^* - \bar{z}$$

и стандартные отклонения также равны. Это также показывает эквивалентность между тестированием гипотезы с помощью бутстрепа для задачи с одной выборкой и бутстреп доверительным интервалом, описанным в главе 12.

¹Другой метод обсуждается в задаче 16.5.

Этот интервал основан на процентилях статистики 16.18 вычисленной по бутстреп выборке из z_1, z_2, \dots, z_7 , точно так же, как указано выше. Следовательно, бутстреп- t доверительный интервал состоит из тех значений μ_0 , которые не отклоняются бутстреп критерием для проверки гипотезы, описанным выше. Эта общая связь между доверительными интервалами и проверками гипотез более подробно описана в разделе 12.3.

16.5 Тестирование мультимодальности генеральной совокупности

Наш второй пример гораздо более экзотичен. Это случай, когда не существует простой теории для тестов, основанных на нормальном распределении, и нельзя использовать перестановочный тест, но можно эффективно использовать бутстреп. Данные представляют собой толщину в миллиметрах 485 марок, напечатанных в 1872 году. Выпуск марок того года считался «филателистической смесью», то есть напечатанным на более чем одном типе бумаги. Исторический интерес представляет определение того, сколько различных типов бумаги использовалось.

Гистограмма данных показана в верхнем левом углу рисунка 16.2. Эта выборка является частью большой коллекции марок 1872 года, и мы можем представить себе распределение измерений толщины для всей генеральной совокупности. Мы ставим статистический вопрос: сколько мод у этой генеральной совокупности? Мода определяется как локальный максимум или «скачок» плотности популяции. Количество мод указывает на количество различных типов бумаги, используемых при печати.

Из гистограммы на рисунке 16.2 видно, что популяция может иметь 2 или более мод. Однако это трудно сказать, потому что гистограмма не гладкая. Чтобы получить более гладкую оценку, мы можем использовать *оценку плотности ядра Гаусса*. Обозначая данные как x_1, x_2, \dots, x_n , оценка плотности ядра Гаусса определяется выражением

$$\hat{f}(t; h) = \frac{1}{nh} \sum_1^n \phi\left(\frac{t - x_i}{h}\right), \quad (16.19)$$

где $\phi(t)$ — плотность стандартного нормального распределения $(1/\sqrt{2\pi}) \exp(-t^2/2)$. Параметр h называется размером окна и определяет степень сглаживания, применяемого к данным. Чем больше значение h , тем более гладкая оценка плотности.

Мы можем думать о 16.19 как о суммировании n маленьких гауссовых кривых плотности с центрами в каждой точке x_i , каждая из которых имеет стандартное отклонение h ; рисунок 16.3 иллюстрирует это.

На верхней правой части рисунка 16.2 показана итоговая оценка плотности с использованием $h = 0.003$; имеется 2 или 3 моды. Однако, изменяя h , мы можем получить большее или меньшее количество мод. Внизу слева и справа показаны оценки плотности, полученные с использованием $h = 0.008$ и $h = 0.001$ соответственно. На левом есть одна мода, а на правом как минимум 7! Очевидно, что вывод, который мы делаем из наших данных, сильно зависит от значения h , которое мы выбираем.

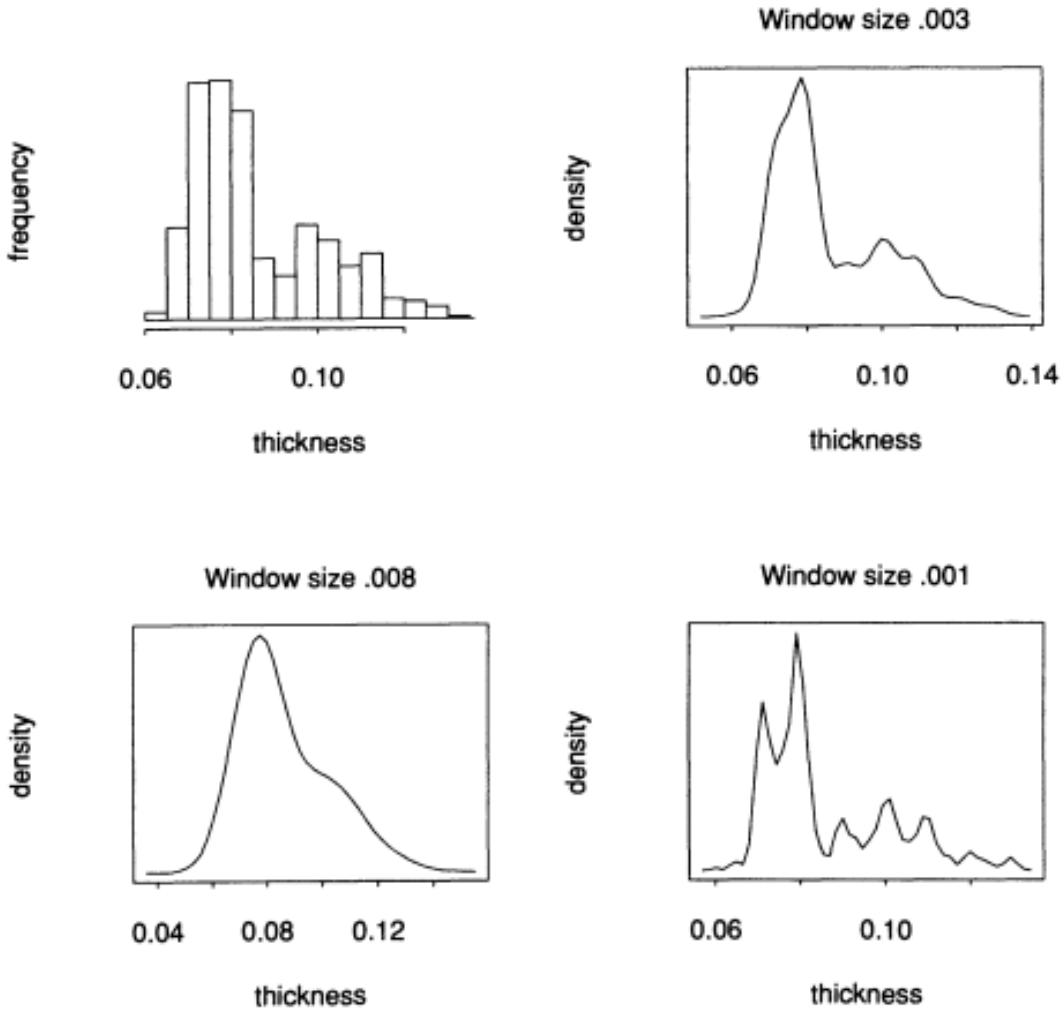


Figure 16.2. *Top left panel shows histogram of thicknesses of 485 stamps. Top right and bottom panels are Gaussian kernel density estimates for the same sample, using window size .003 (top right), .008 (bottom left) and .001 (bottom right).*

Если мы подойдем к проблеме с точки зрения проверки гипотез, есть естественный способ выбрать h . Нам понадобится следующий важный результат, который мы сформулируем без доказательства: с увеличением h количество мод в оценке плотности ядра Гаусса не увеличивается. Это проиллюстрировано для данных о толщине марок на рисунке 16.4.

Теперь рассмотрим проверку

$$H_0 : \text{число мод} = 1 \quad (16.20)$$

против альтернативы количество мод > 1 . Поскольку количество мод уменьшается с увеличением h , существует наименьшее значение h такое, что $\hat{f}(t; h)$ имеет одну моду. Назовем его \hat{h}_1 . Глядя на рисунок 16.4, можно сказать, что $\hat{h}_1 \approx 0.0068$.

Кажется разумным использовать $\hat{f}(t; h)$ в качестве оценки нулевого распределения для нашего тестирования H_0 . В некотором смысле это оценка плотности, наиболее близкая к нашим данным, которая согласуется с H_0 . Под «ближайшая» мы подразумеваем, что она использует наименьшее сглаживание (наименьшее значение h) среди всех оценок с одной модой.

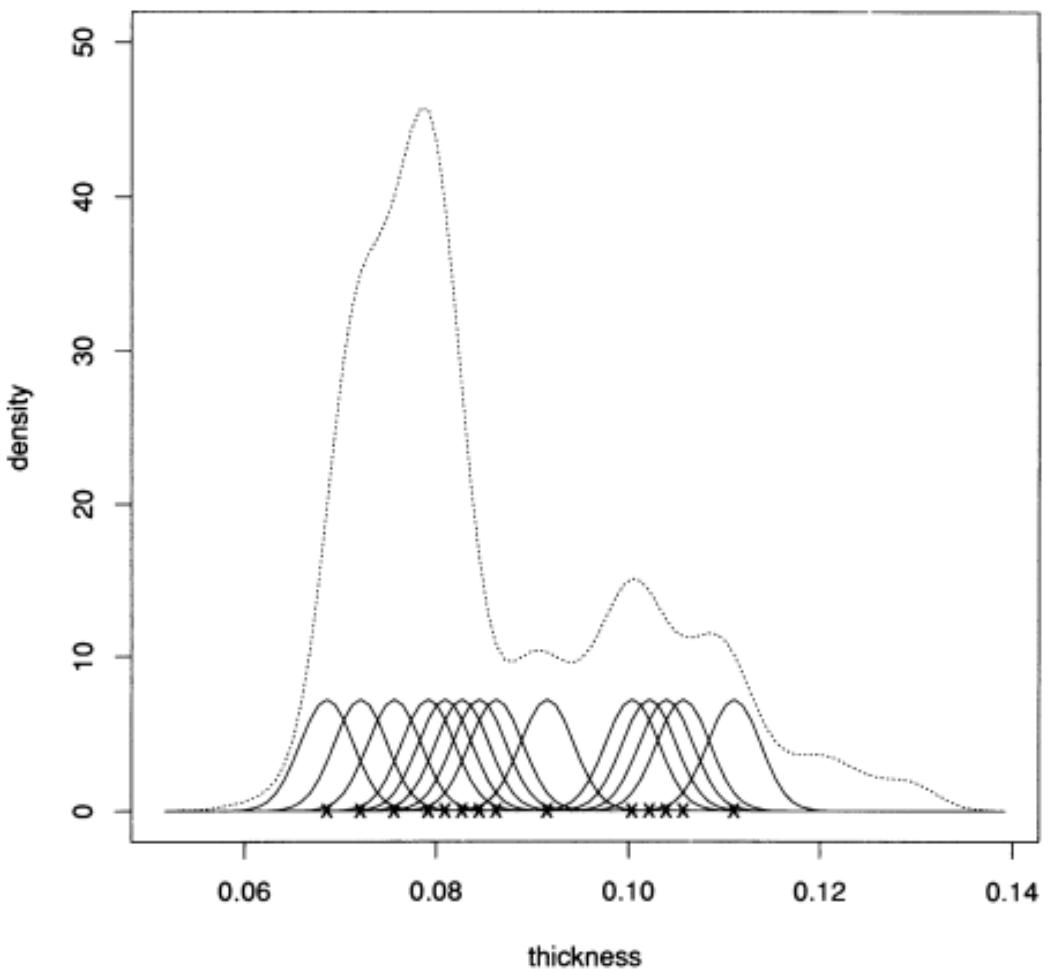


Figure 16.3. Illustration of a Gaussian kernel density estimate. A small Gaussian density is centered at each data value (marked with an “x”) and the density estimate (broken line) at each value is determined by adding up the values of all the Gaussian densities at that point. For the stamp data there are actually 485 little Gaussian densities used (one for each point); for clarity we have shown only a few.

Есть одна небольшая поправка, которую мы вносим в $\hat{f}(\cdot; \hat{h}_1)$. Формула 16.19 искусственно увеличивает дисперсию оценки, поэтому мы масштабируем ее так, чтобы дисперсия была равна дисперсии выборки. Обозначим измененную оценку через $\hat{g}(\cdot; \hat{h}_1)$.

Наконец, нам нужно выбрать тестовую статистику. Естественный выбор — это \hat{h}_1 , наименьший размер окна, дающий оценку плотности с одной модой. Большое значение \hat{h}_1 указывает на то, что необходимо выполнить большое сглаживание, чтобы получить оценку с одной модой, и, следовательно, является свидетельством против H_0 .

Собирая все это вместе, бутстреп проверка гипотезы H_0 : число мод = 1 основана на достижении уровня значимости

$$\text{ASL}_{\text{boot}} = \text{Prob}_{\hat{g}(\cdot; \hat{h}_1)} \left\{ \hat{h}_1^* > \hat{h}_1 \right\}. \quad (16.21)$$

Здесь \hat{h}_1 зафиксировано на наблюдаемом значении 0.0068; бутстреп выборка $x_1^*, x_2^*, \dots, x_n^*$ берется из $\hat{g}(\cdot; \hat{h}_1)$, а \hat{h}_1^* — наименьшее значение h , дающее оценку плотности с одной модой по бутстреп выборке $x_1^*, x_2^*, \dots, x_n^*$.

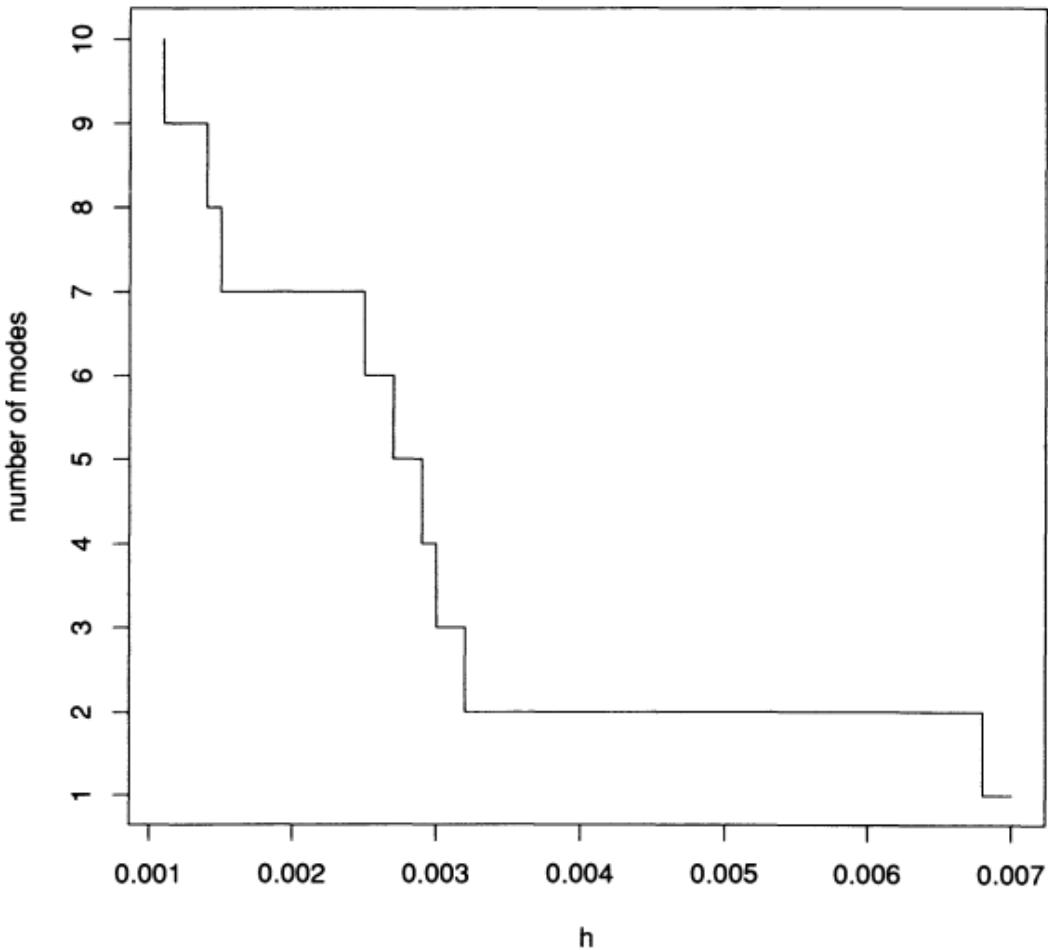


Figure 16.4. Stamp data: number of modes in the Gaussian kernel density estimate as a function of the window size h .

Чтобы аппроксимировать ASL_{boot} , нам нужно получить бутстреп выборки из масштабированной оценки плотности $\hat{g}(\cdot; \hat{h}_1)$. То есть, вместо того, чтобы производить выборку с возвращением, мы получаем выборку с помощью гладкой оценки плотности генеральной совокупности. Это называется гладким бутстрепом. Из-за удобной формы оценки ядра Гаусса легко получить выборку из $\hat{g}(\cdot; \hat{h}_1)$. Выберем $y_1^*, y_2^*, \dots, y_n^*$ с возвращением из x_1, x_2, \dots, x_n и положим

$$x_i^* = \bar{y}^* + \left(1 + \hat{h}_1^2/\hat{\sigma}^2\right)^{-1/2} \left(y_i^* - \bar{y}^* + \hat{h}_1 \epsilon_i\right); \quad i = 1, 2, \dots, n, \quad (16.22)$$

где \bar{y}^* — среднее значение $y_1^*, y_2^*, \dots, y_n^*$, $\hat{\sigma}^2$ — оценка(plug estimate?) дисперсии, а ϵ_i — стандартные нормальные случайные величины. Множитель $\left(1 + \hat{h}_1^2/\hat{\sigma}^2\right)^{-1/2}$ масштабирует оценку так, чтобы его дисперсия была приблизительно равна $\hat{\sigma}^2$. Краткое изложение шагов показано в алгоритме 16.3.

Мы выполнили этот процесс с $B = 500$. Из 500 бутстреп выборок ни одна не имела $\hat{h}_1^* > 0.0068$, поэтому $\text{ASL}_{\text{boot}} = 0$. Мы повторили процесс для H_0 : число мод = 2, 3, …, и таблица 16.1 показывает итоговые p -значения. Интерпретируя эти результаты последовательно, начиная с числа мод = 1, мы отвергаем гипотезу об одной моде, но не отвергаем гипотезу двух модах. На этом процесс рассуждений во многих случаях заканчивался. Если бы мы были готовы принять более экзотические гипотезы, то из таблицы 16.1 также можно было бы предположить, что в генеральной совокупности может быть 7 мод.

Алгоритм 16.3

Вычисление статистики для бутстреп проверки гипотезы о мультимодальности

1. Получить B выборок размера n из $\hat{g}(\cdot; \hat{h}_1)$ используя 16.22.
2. Для каждой бутстреп выборки вычислить наименьшую ширину окна, которая дает оценку плотности с одной модой. Обозначим B значений \hat{h}_1^* через $\hat{h}_1^*(1), \dots, \hat{h}_1^*(B)$.
3. Аппроксимировать ASL_{boot} с помощью

$$\widehat{\text{ASL}}_{\text{boot}} = \# \left\{ \hat{h}_1^*(b) \geq \hat{h}_1 \right\} / B. \quad (16.23)$$

Table 16.1. *P-values for stamp example.*

| number of modes(m) | \hat{h}_m | P-value |
|------------------------|-------------|---------|
| 1 | .0068 | .00 |
| 2 | .0032 | .29 |
| 3 | .0030 | .06 |
| 4 | .0029 | .00 |
| 5 | .0027 | .00 |
| 6 | .0025 | .00 |
| 7 | .0015 | .46 |
| 8 | .0014 | .17 |
| 9 | .0011 | .17 |

16.6 Обсуждение

Как показывают примеры в этой главе, при проведении бутстреп проверки гипотез мы должны выбрать две величины:

- (a) Тестовую статистику $t(\mathbf{x})$.
- (b) Нулевое распределение \hat{F}_0 для данных подчиняющихся H_0 .

Учитывая это, мы генерируем B -бутстреп значений $t(\mathbf{x}^*)$ при \hat{F}_0 и оцениваем достигнутый уровень значимости как

$$\widehat{\text{ASL}}_{\text{boot}} = \# \left\{ t(\mathbf{x}^{*b}) \geq t(\mathbf{x}) \right\} / B. \quad (16.24)$$

Как показывает пример о толщине марок, иногда выбор $t(\mathbf{x})$ и \hat{F}_0 не очевиден. Сложность выбора \hat{F}_0 состоит в том, что в большинстве случаев H_0 является

сложной гипотезой. В примере о толщине марок H_0 относится ко всем возможным плотностям с одной модой. Хорошим выбором для \hat{F}_0 является распределение, которое подчиняется H_0 и наиболее разумно для наших данных; этот выбор делает тест консервативным, то есть с меньшей вероятностью тест ошибочно отвергнет нулевую гипотезу. В примере с марками мы проверили унимодальность путем генерации выборок из унимодального распределения, которое в большинстве случаев является почти бимодальным. Другими словами, мы использовали наименьшее возможное значение для \hat{h}_1 , и это делает вероятность в 16.21 максимально большой.

Выбор тестовой статистики $t(\mathbf{x})$ будет определять мощность теста, то есть вероятность того, что мы отклоним H_0 , в случае когда она не верна. В примере с марками, если фактическая плотность генеральной совокупности является бимодальной, но плотность ядра Гаусса не аппроксимирует ее точно, тогда тест, основанный на ширине окна \hat{h}_1 , не будет иметь высокой мощности.

Бутстреп тесты полезны в ситуациях, когда альтернативная гипотеза не уточняется. В случаях, когда существует параметрическая альтернативная гипотеза, могут быть предпочтительны методы правдоподобия или байесовские методы.

Глава 17

Кросс-валидация и другие оценки ошибки предсказания

17.1 Введение

До сих пор наше обсуждение было сосредоточено на нескольких статистических показателях точности: стандартная ошибка, смещение и доверительный интервал. Это меры точности параметров модели. Ошибка предсказания — это величина, которая показывает, насколько хорошо модель предсказывает значение отклика для будущего наблюдения. Она часто используется для выбора модели, поскольку разумно выбрать модель с наименьшей ошибкой предсказания.

Кросс-валидация — это стандартный инструмент для оценки ошибки предсказания. В связи с увеличением доступной вычислительной мощности эта идея (предшествующая бутстрепу) стала набирать популярность в последние годы. В этой главе мы обсудим, кросс-валидацию, бутстреп и некоторые другие методы, которые оценивают ошибки предсказания.

В регрессионных моделях ошибка предсказания — это математическое ожидание квадрата разности истинного значения и предсказания, которое было получено с помощью модели:

$$\text{PE} = \mathbb{E}(y - \hat{y})^2. \quad (17.1)$$

(The expectation refers to repeated sampling from the true population.)???Математическое ожидание относится к повторной выборке из генеральной совокупности. Ошибка предсказания также возникает в задаче классификации, когда ответом является один из классов k . Например, в политическом опросе возможными ответами могут быть республиканец, демократ или независимый кандидат. Обычно, в задачах классификации ошибка определяется как вероятность неправильной классификации

$$\text{PE} = \text{Prob}(\hat{y} \neq y), \quad (17.2)$$

она также называется долей неправильно классифицированных наблюдений. Методы, описанные в этой главе, применимы как для определения ошибок предсказания, так и для других ошибок. Мы начнем с интуитивного описания методов, а затем рассмотрим их более подробно в разделе 17.6.2.

17.2 Пример: данные по гормонам

Давайте снова посмотрим на пример данных по гормонам из 9 главы. Для удобства повторно изобразим данные на рисунке 17.1. Напомним, что переменная отклика y_i — это количество противовоспалительного гормона, оставшегося в устройстве после z_i часов его ношения. Устройства были случайным образом выбраны из трех партий A , B и C , которые обозначены буквами на графике. В девятой главе мы строили модель регрессии по данным каждой партии

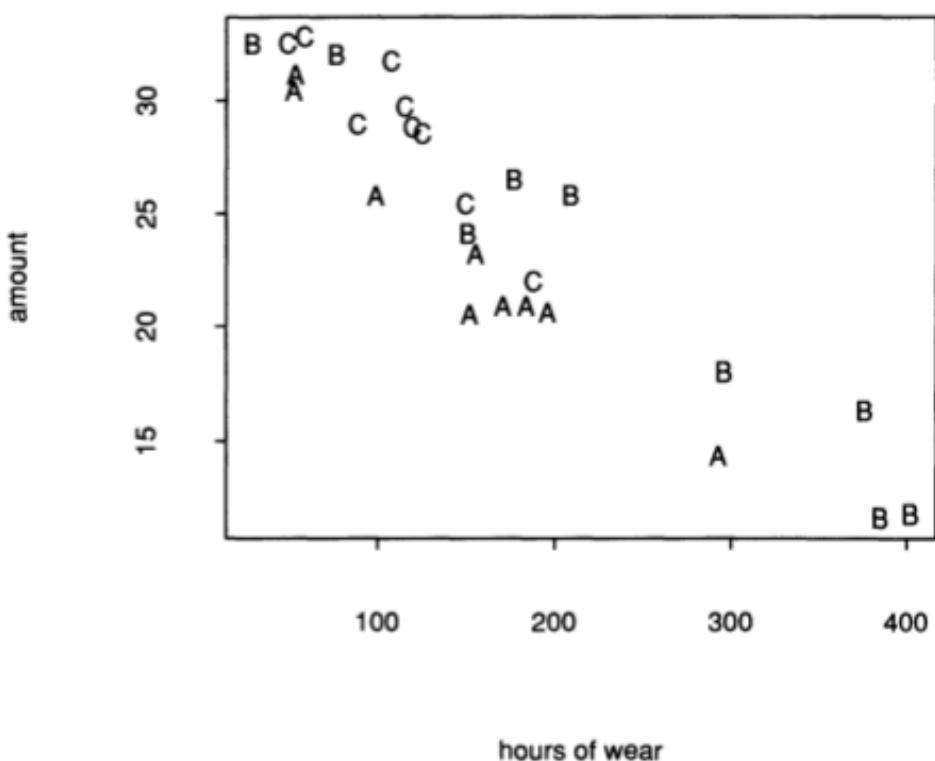


Figure 17.1. Hormone data. Plot shows the amount of hormone remaining for a device versus the hours of wear. The symbol represents the lot number.

с разными свободными, но одинаковыми угловыми коэффициентами. Оценки приведены в таблице 9.3 на 110 странице.

Здесь мы обсудим два вопроса: 1) Насколько хорошо модель предсказывает количество оставшегося гормона для нового устройства? 2) Прогнозирует ли эта модель лучше (или хуже), в случае, когда у нас одна линия регрессии? Чтобы ответить на первый вопрос, мы могли бы посмотреть на среднюю остаточную ошибку по всем $n = 27$ наблюдениям.

$$\frac{\text{RSE}}{n} = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n} = 2.20, \quad (17.3)$$

но это может быть слишком «оптимистично», и истинная ошибка предсказания, возможно, будет недооценена. Дело в том, что мы используем для оценки модели те же данные, что и для ее обучения, используя оценки параметров,

которые подобраны на этом же наборе данных. Другими словами, тестовая выборка, иногда называемая обучающей выборкой, такая же, как и исходная. Полученные таким образом оценки для ошибок предсказания называют явными ошибками ("apparent error" estimates)????.

Способ исправить (17.3) — это заменить деление на n на деление на $n-p$, где p — количество признаков. Это помогает сделать обычную оценку остаточной дисперсии несмещенной $\hat{\sigma}^2 = \sum \frac{(y_i - \hat{y}_i)^2}{n-p}$. Мы увидим, что для решения проблемы необходимы более серьезные исправления.

17.3 Кросс-валидация

Чтобы получить более реалистичную оценку ошибки предсказания, мы хотели бы иметь тестовую выборку, отличную от нашей обучающей выборки. В идеале это могли бы быть новые данные из той же генеральной совокупности, из которой была взята наша первоначальная выборка. В нашем примере это количество часов, которое носилось устройство и количество оставшегося в нем гормона, скажем для m новых устройств. Если бы у нас были эти новые данные, $(z_1^0, y_1^0), \dots, (z_m^0, y_m^0)$, мы вычислили бы значения предсказаний \hat{y}_i^0 из таблицы (9.3):

$$\hat{y}_i^0 = \hat{\beta}_j^0 + \hat{\beta}_1^0 z_i^0, \quad (17.4)$$

(где $j = A, B$ или C в зависимости от партии) и среднюю сумму квадратов разности:

$$\sum_1^m \frac{(y_i^0 - \hat{y}_i^0)^2}{m}. \quad (17.5)$$

Эта величина оценивает, насколько в среднем наш прогноз \hat{y}_i^0 отличается от фактического значения y_i^0 .

Обычно дополнительные данные недоступны по причинам трудозатратности (logistics???) или стоимости. Чтобы решить эту проблему, кросс-валидация использует одну часть имеющихся данных для построения модели, а другую часть для ее тестирования. При больших объемах данных обычной практикой является разделение данных на две равные части. Когда объем данных не большой, как в случае с данными о гормонах, k -fold кросс-валидация позволяет более эффективно использовать доступные данные. Процедура показана в алгоритме 17.1.

Algorithm 1 k -fold кросс-валидация

1. Разделите данные на K частей примерно равного размера.
 2. Постройте модель по $K-1$ частям данных и вычислите ошибку предсказания построенной модели на предсказании для k -й части данных, которая не участвовала в построении модели.
 3. Выполните эти два шага для $k = 1, 2, \dots, K$ и сложите K оценок ошибок предсказания.
-

Далее рассмотрим k -fold кросс-валидацию подробнее. Предположим, мы разбиваем данные на K частей. Пусть $k(i)$ — это часть, которая содержит i -ое наблюдение. Обозначим через $\hat{y}_i^{-k(i)}$ предсказание для i -го наблюдения, которое было вычислено по данным с удалением $k(i)$ -ой части. Тогда оценка

ошибки предсказания полученная с помощью кросс-валидации будет равна:

$$CV = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{(k(i))})^2. \quad (17.6)$$

Часто выбирается $k = n$, что приводит к «leave-one-out» кросс-валидации. Строим модель без учета i -го наблюдения, а затем используем ее для его предсказания, обозначим результат предсказания $\hat{y}_i^{(i)}$. Мы делаем так для каждого наблюдения, а затем вычисляем среднюю сумму квадратов полученную с помощью кросс-валидации $CV = \sum_{i=1}^n \frac{(y_i - \hat{y}_i^{(i)})^2}{n}$.

Мы использовали leave-one-out кросс-валидацию к данным по гормонам: значение CV оказалось равным 3.09. Для сравнения, средняя остаточная квадратичная ошибка (17.3) равна 2.20, таким образом, она занижает значение ошибки предсказания примерно на 29%. На рисунке 17.2 показан график обычных остатков $y_i - \hat{y}_i$ (кружки) и график остатков полученных с помощью кросс-валидации $y_i - \hat{y}_i^{(i)}$ (звездочки). Обратите внимание, что для каждого наблюдения, значения остатков полученных с помощью кросс-валидации или больше (по абсолютной величине) обычных остатков или равны им.

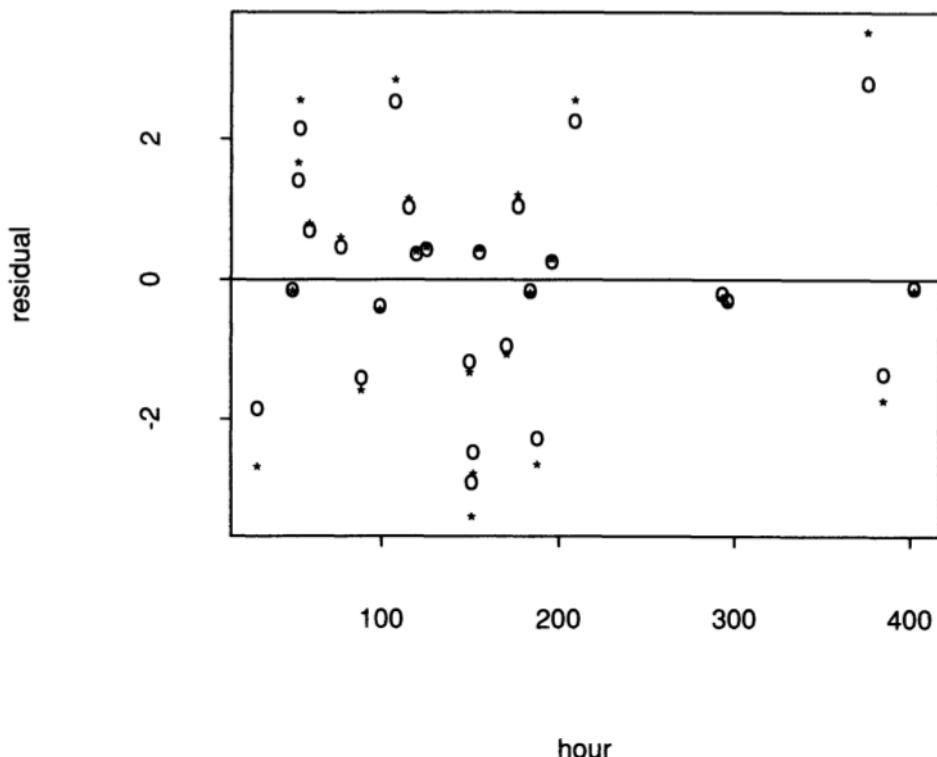


Figure 17.2. Plot of residuals (circles) and cross-validated residuals (stars) for hormone data.

Далее мы можем посмотреть на разбивку CV по партиям: средние значения равны 2.09, 4.76 и 2.43 для партии A , B и C , соответственно. Следовательно, сложнее предсказать значения для устройств из партии B , нежели для устройств из партий A и C .

Кросс-валидация, как было описано выше, требует обучения всей модели

n раз. В общем, это неизбежно, но для аппроксимации методом наименьших квадратов возможен более рациональный способ.

17.4 C_p и другие оценки ошибки предсказания

Есть и другие способы оценки ошибки предсказания, и все они основаны на остаточной стандартной ошибке RSE. Последняя часть этой главы описывает бутстреп подход. Простая аналитическая мера — (adjusted)??? скорректированная остаточная стандартная ошибка:

$$\frac{\text{RSE}}{(n - 2p)}, \quad (17.7)$$

где p обозначает количество регрессоров в модели. Эта корректировка RSE/n позволит учесть (fitting)??? подгонку, чем больше p тем выше степень коррекции. Обратите внимание, что $\text{RSE}/(n - 2p)$ является более серьезной коррекцией RSE, чем несмешенная оценка дисперсии $\text{RSE}/(n - p)$.

Другая оценка (одна из форм) — это « C_p » статистика:

$$C_p = \frac{\text{RSE}}{n} + \frac{2p\hat{\sigma}^2}{n}. \quad (17.8)$$

Здесь $\hat{\sigma}^2$ — оценка остаточной дисперсии, разумным было бы положить ее равной $\text{RSE}/(n - p)$. (При вычислении C_p статистики для ряда моделей $\hat{\sigma}^2$ вычисляется один раз на основе $\text{RSE}/(n - p)$ для некоторой фиксированной большой модели). Данная статистика является частным случаем информационного критерия Акаике (AIC) для общих моделей. Это позволяет скорректировать RSE/n так, чтобы получить приблизительно несмешенную оценку для ошибки предсказания: $E(C_p) \approx \text{PE}$.

Неявно эти поправки учитывают еще и тот факт, что одни и те же данные используются как для построения модели, таки и для ее оценки с помощью остаточной стандартной ошибки. Заметим что « p » в знаменателе и второй член в C_p — это штрафы, учитывающие степень подгонки. A simple argument shows that the ad- justed residual squared error and C_p statistic are equivalent to a first order of approximation (Problem 17.4.) (Простой аргумент показывает, что скорректированная остаточная стандартная ошибка и C_p статистика эквивалентны первому порядку аппроксимации.)???

Также C_p статистика подобна критерию Шварца или BIC (байесовский информационный критерий).

$$\text{BIC} = \frac{\text{RSE}}{n} + \frac{\log n \cdot p\hat{\sigma}^2}{n}. \quad (17.9)$$

Если сравнить C_p и BIC, то можно заметить, что в последнем на месте «2» стоит $\log n$, а следовательно, назначается больший штраф, чем в C_p , так как, $n > e^2$. Таким образом, при сравнении моделей с помощью BIC будут выбираться модели с меньшим числом регрессоров, чем при сравнении с помощью C_p . Можно показать, что BIC является (consistent)???? непротиворечивым критерием в том смысле, что он выбирает правильную модель при $n \rightarrow \infty$. Это не относится к скорректированной RSE или C_p .

В примере с гормоном $\text{RSE} = 59.27$, $\hat{\sigma}^2 = 2.58$ и $p = 4$ следовательно, $\text{RSE}/(n - 2p) = 3.12$, $C_p = 2.96$, $\text{BIC} = 3.45$, по сравнению со значениями CV, равным 3.09.

Зачем использовать кросс-валидацию, когда есть более простые альтернативы? Основная причина заключается в том, что для задач построения модели методом более сложным чем метод наименьших квадратов, количество параметров « p » не является известным. В отличие от кросс-валидации, скорректированная RSE, C_p статистика и BIC требуют знания p . Как и бутстрап, в простых задачах кросс-валидация стремится давать ответы, аналогичные тем что дают стандартные методы, и ее преимущество заключается в ее применимости к более сложным ситуациям. Ниже приведен пример деревьев классификации.

Второе преимущество кросс-валидации — это ее надежность. C_p статистика и BIC требуют приблизительно правильной модели для получения оценки $\hat{\sigma}^2$. Кросс-валидации не требует этого и будет хорошо работать, даже если оцениваемые модели далеки от правильности.

Теперь, ответим на второй вопрос, который обсуждался выше, касательно модели с одинаковыми угловыми коэффициентами, но с отдельными линиями регрессии для каждой партии и простой моделью, которая определяет одну общую линию регрессии для всех партий сразу. Таким же образом, как описано выше, мы можем вычислить сумму квадратов с помощью кросс-валидации для модели с одной линией регрессии. Это значение равно 5.89, что немного больше, чем значение равное 3.27, полученное для случая, когда для каждой партии допускается свое значение свободного члена. Это неудивительно, учитывая статистически значимые различия между свободными членами в таблице 9.3. Но кросс-валидация полезна, поскольку она дает количественную оценку цены, которую исследователь заплатил бы, если бы он не внес поправку на номер партии устройства.

17.5 Пример: деревья классификации

В качестве примера, который демонстрирует пользу кросс-валидации, приведем современную статистическую процедуру, которая называется «деревья классификации». В эксперименте, разработанном для получения информации о причинах возникновения язв двенадцатиперстной кишки, каждой из 745 крыс вводили один из 56 модельных алкилнуклеофилов. Позднее у каждой крысы было проведено вскрытие на предмет развития язвы двенадцатиперстной кишки, и результат был классифицирован как 1, 2 или 3 в порядке возрастания степени тяжести. Было 535 результатов первого класса, 90 — второго класса и 120 результатов третьего класса. Было измерено шестьдесят семь характеристик этих соединений, и цель анализа состояла в том, чтобы установить, какие из характеристик были связаны с развитием язв двенадцатиперстной кишки.

Рассмотрим довольно популярный, но вычислительно затратный алгоритм CART (для деревьев классификации и регрессии) Бреймана, Фридмана, Олшена и Стоуна (1984). Дерево классификации для наших данных, найденное с помощью алгоритма CART показано на рисунке 17.3.

В каждом узле дерева задается «да-нет» вопрос, и те данные, для которых дан ответ «да», присваиваются левой ветви, а остальные правой. Листья дерева, показанного на рис. 17.3, называются «конечными узлами». На основе ответов на вопросы каждое наблюдение присваивается одному из конечных узлов. Например, крыса, с дипольным моментом ≤ 3.56 и точкой плавления > 98.1 , попадает в левый, а затем конечный вправый узел, обозначенный «[13, 7, 41]». Три числа, которые изображены под каждым конечным узлом, указывают на количество наблюдений в каждом классе в этом узле, например, для

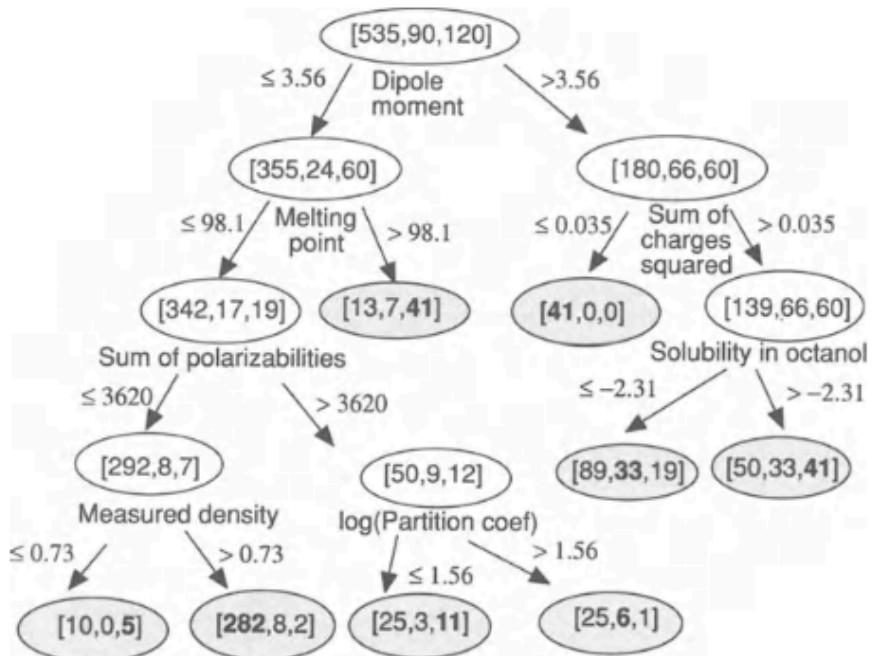


Figure 17.3. CART tree. Classification tree from the CART analysis of data on duodenal ulcers. At each node of the tree a question is asked, and data points for which the answer is “yes” are assigned to the left branch and the others to the right branch. The shaded regions are the terminal nodes, or leaves, of the tree. The numbers in square brackets are the number of observations in each of the three classes present at each node. The bold number indicates the predicted class for the node. In this particular example, five penalty points are charged for misclassifying observations in true class 2 or 3, and one penalty point is charged for misclassifying observations in class 1. The predicted class is the one resulting in the fewest number of penalty points.

конечного узла с числами «[13, 7, 41]» верно, что в этом конечном узле имеется 13 наблюдений класса 1, 7 наблюдений класса 2 и 41 наблюдение из класса 2.

Прежде чем обсуждать, как именно это дерево было построено с помощью алгоритма CART, давайте посмотрим, как оно используется для классификации. Каждому конечному узлу назначается класс (1,2 или 3). Самый очевидный способ назначить классы конечным узлам — это использовать правило большинства и назначить узлу тот класс, который является в нём наиболее многочисленным. Используя правило большинства, узел «[13,7,41]» будет отнесен к классу 3, а все остальные конечные узлы будут отнесены к классу 1. В этом исследовании было решено, что будет в пять раз хуже если мы неправильно классифицируем животное, у которого была тяжелая или умеренная язва, нежели если мы ошибемся с классификацией животного с легкой степенью язвы. Следовательно, если наблюдения из класса 2 и 3 будут классифицированы неверно, то будет начислено пять штрафных баллов, в случае же когда неверно классифицируются наблюдения из класса 1 начисляется один штрафной балл. Прогнозируемый класс — это тот класс, который дает наименьшее количество штрафных баллов. На рисунке 17.3 на каждом конечном узле жирным шрифтом выделен спрогнозированный класс, например, для узла в левом нижнем углу — «[10,0,5]» жирным шрифтом выделена цифра 5, следовательно, узел соотносится с классом 3.

Можем сделать следующие выводы по поводу дерева. Верхний («корневой») узел раскололся по дипольному моменту. Высокий дипольный момент указывает на наличие электроотрицательных групп. Происходит разделение на классы 1 и 2: отношение класса 2 к классу 1 в правом разветвлении равно $66/180$, что более чем в 5 раз больше, чем соотношение этих классов в левом разветвлении, равное $24/355$. Однако наблюдения класса 3 делятся поровну, по 60 на каждой стороне разветвления. Если при этом сумма квадратов атомных зарядов мала, то CART определяет, что все наблюдения относятся к классу 1. Следовательно, ионизация является основным определяющим фактором биологического действия для наблюдений с высокими дипольными моментами. Двигаясь дальше вниз по правой части дерева, растворимость в октаноле (частично) отделяет класс 3 от класса 2. Высокая растворимость октанола, вероятно, отражает способность проникать через мембранны и проникать в центральную нервную систему.

На левой стороне корневого узла соединения с низким дипольным моментом и высокой температурой плавления оказались отнесены классу 3. Соединения в этом конечном узле связаны с цистеамином. Наблюдения с низкими температурами плавления и высокой поляризуемостью, все тиолы в этом исследовании, были классифицированы как класс 2 или 3. Химические вещества с низкой поляризуемостью и высокой плотностью относятся к классу 1. Эти химические вещества имеют высокую молекулярную массу и объем, а также этот конечный узел содержит наибольшее количество наблюдений. На стороне разветвления с низкой плотностью все амины с короткой цепью.

В терминологии, упомянутой ранее, набор данных из 745 наблюдений называется обучающей выборкой. Легко вычислить частоту ошибочной классификации для каждого класса, если применить к обучающей выборке дерево, показанное на рис. 17.3. Если посмотреть на конечные узлы, которые предсказывают классы 2 или 3, количество ошибок для класса 1 составляет $13 + 89 + 50 + 10 + 25 + 25 = 212$, поэтому очевидный процент ошибочной классификации для класса 1 составляет $212/535 = 39.6\%$. Аналогичным образом, явные коэффициенты ошибочной классификации для классов 2 и 3 составляют 56.7% и 18.3%. «Явный» является здесь важным уточнением, поскольку частота ошибочной классификации для обучающей выборки может быть сильно смещена в сторону понижения по той же причине, по которой квадрат остаточной ошибки слишком оптимистичен для регрессии.

Как CART строит дерево, которое изображено на рис. 17.3? CART - это полностью автоматическая процедура, которая выбирает переменные и значения, по которым данные лучше всего делятся на классы. Например, «дипольный момент ≤ 3.56 » — был выбран лучшим признаком для разделения данных на классы. CART выбрал как переменную по которой происходит разделение — «дипольный момент», так и значение, которое равно 3.56. После того как было выбрано первое правило для разделения, новые правила разделения выбираются для каждой из двух полученных после первого разделения групп, и этот процесс повторяется.

Вместо того, чтобы остановить алгоритм, когда дерево достигает некоторого разумного размера, CART использует более эффективный подход: создается большое дерево, которое затем обрезается снизу. Это более эффективный подход для обнаружения взаимосвязей, в которых участвует несколько переменных.

Возникает важный вопрос: какого размера должно быть дерево? Если мы

построили бы очень большое дерево у которого было бы только одно наблюдение в каждом конечном узле, то очевидная вероятность ошибочной классификации составила бы 0%. Однако, это дерево не справится с предсказанием результатов для новой выборки. Причина в том, что дерево будет ориентировано на обучающую выборку — это «переобучение».

Дерево наилучшего размера — это такое дерево, которое дает наименьшую ошибку классификации на новых данных. Таким образом, если бы у нас был второй набор данных (тестовая выборка), мы могли бы применить к нему деревья разных размеров, а затем выбрать то дерево, у которого самое маленько количество ошибок классификации.

В большинстве ситуаций у нас нет дополнительных данных, поэтому воспользуемся кросс-валидацией. Leave-one-out кросс-валидация не подойдет для нашего случая, потому что результирующие деревья не будут достаточно отличаться от исходного дерева. Опыт показывает, что гораздо лучше разделить данные на 10 групп равного размера, построить дерево для 90% данных, а затем оценить уровень ошибочной классификации на оставшихся 10% данных. Это делается по очереди для каждой из 10 групп, и общий коэффициент ошибочной классификации вычисляется за 10 прогонов. Лучшим размером дерева считается тот размер дерева, который дает наименьшую вероятность ошибочной классификации. Этот размер будет использоваться для построения окончательного дерева на всех данных. Важнейшей особенностью кросс-валидации является разделение данных для построения и оценки деревьев: каждая десятая часть данных выступает в качестве тестовой выборки для остальных девяти частей.

Кросс-валидация позволяет не только получить оценку для лучшего размера дерева, но также дает реалистичную оценку уровня ошибочной классификации для окончательного дерева. Вычисленные выше явные ошибки часто оказываются ниже, потому что обучающая выборка используется как для построения, так и для оценки дерева. Для дерева на рис. 17.3 частота ошибочной классификации с проверкой, с помощью кросс-валидации, была примерно на 10% выше, чем частота явных ошибок. Уровень ошибки вычисленный с помощью кросс-валидации обеспечивает точную оценку того, насколько эффективно дерево будет классифицировать новую выборку.

17.6 Бутстреп оценки ошибки предсказания

17.6.1 Обзор

В следующих двух разделах мы посмотрим, как можно использовать бутстреп для оценки ошибки предсказания. Для точной формулировки потребуются некоторые обозначения. Прежде чем перейти к ним обсудим основные идеи. Простейший бутстреп подход заключается в следующем: генерируется B бутстреп выборок для каждой из которых оценивается модель, затем каждая из моделей применяется к исходной выборке и оценивается ошибка предсказания, таким образом, получаем B оценок ошибки предсказания. Общая оценка ошибки предсказания — это среднее значение этих B оценок. В качестве примера в левом столбце таблицы 17.1 приведены десять оценок ошибок предсказания («err») для десяти бутстреп выборок для данных о гормонах, которые были описаны в разделе 17.2. Их среднее значение равно 2.52, для сравнения, среднее значение для RSE/n равно 2.20.

Table 17.1. Bootstrap estimates of prediction error for hormone data of Chapter 9. In each row of the table a bootstrap sample was generated by sampling with replacement from the hormone data, and the model specified in equation (9.21) was fit. The left column shows the resulting prediction error when this model is applied to the original data. The average of the left column (=2.52) is the simple bootstrap estimate of prediction error. The center column is the prediction error that results when the model is applied to the bootstrap sample, the so-called “apparent error.” It is unrealistically low. The difference between the first and second columns is the “optimism” in the apparent error, given in the third column. The more refined bootstrap estimate adds the average optimism (=0.82) to the average residual squared error (=2.20), giving an estimate of 3.02.

| | $\text{err}(\mathbf{x}^*, \hat{F})$ | $\text{err}(\mathbf{x}^*, \hat{F}^*)$ | $\text{err}(\mathbf{x}^*, \hat{F}) - \text{err}(\mathbf{x}^*, \hat{F}^*)$ |
|------------|-------------------------------------|---------------------------------------|---|
| sample 1: | 2.30 | 1.47 | 0.83 |
| sample 2: | 2.56 | 3.03 | -0.47 |
| sample 3: | 2.30 | 1.65 | 0.65 |
| sample 4: | 2.43 | 1.76 | 0.67 |
| sample 5: | 2.44 | 2.00 | 0.44 |
| sample 6: | 2.67 | 1.17 | 1.50 |
| sample 7: | 2.68 | 1.23 | 1.45 |
| sample 8: | 2.39 | 1.55 | 0.84 |
| sample 9: | 2.86 | 1.76 | 1.10 |
| sample 10: | 2.54 | 1.37 | 1.17 |
| AVERAGE: | 2.52 | 1.70 | 0.82 |

Такой простой бутстреп подход не очень хорош, но его легко можно улучшить. Посмотрим на второй столбец таблицы 17.1 он отображает ошибку предсказания для модели которая применяется к той же бутстреп выборки на которой и была обучена. Неудивительно, что значения во втором столбце в среднем ниже, чем в первом. Улучшенная бутстреп оценка основана на разнице между первым и вторым столбцом. Она называется «оптимизм» — это величина, на которую средняя остаточная ошибка (или «частота явных ошибок») занижает истинную ошибку предсказания. Общая оценка «оптимизма» — это среднее значение в разностях между первым и вторым столбцом, в этом примере это значение равно 0.82.

После того как «оптимизм» оценка получена, она складывается с явной частотой ошибок, так и получается улучшенная оценка ошибки предсказания. Здесь $2.20 + 0.82 = 3.02$. Конечно, 10 бутстреп выборок — это слишком мало. Если использовать 200 бутстреп выборок, то бутстреп оценка будет равна 2.77, а (optimism)??? оценка 0.80, тогда улучшенная оценка ошибки прогноза будет равна $2.20 + 0.80 = 3.00$. По сути, тут происходит коррекция смещения явной частоты ошибок подобно тому, как мы сделали это в 10 главе.

17.6.2 Некоторые детали

Улучшение простого бутстреп подхода происходит посредством устранения разницы между строками таблицы 17.1, что во многом аналогично устраниению (block effects)??? в двухфакторном дисперсионном анализе. Чтобы лучше понять обоснование бутстреп процедуры, нам нужно мыслить в терминах вероятностных моделей.

В седьмой и девятой главах мы описываем два бутстреп метода регрессионных моделей. Второй метод, на котором мы остановимся, обрабатывает данные $\mathbf{x}_i = (\mathbf{c}_i, y_i), i = 1, 2, \dots, n$ как независимые и одинаково распределенные из многомерного распределения F . Напомним, что \mathbf{c}_i может быть вектором, в данных по гормону \mathbf{c}_i — это номер партии и время ношения i -го устройства. Назовем всю выборку \mathbf{x} . Таким же образом можно сформулировать проблему классификации с помощью y_i , указывающему на класс, которому принадлежит i -ое наблюдение. Обсуждение, которое мы приведем далее, носит довольно общий характер и охватывает проблемы как регрессии, так и классификации.

Предположим, мы оцениваем модель на основе наших данных, для прогноза y при $\mathbf{c} = \mathbf{c}_0$, обозначим ее:

$$\eta_{\mathbf{x}}(\mathbf{c}_0). \quad (17.10)$$

Мы предполагаем, что $\eta_{\mathbf{x}}(\mathbf{c}_0)$ может быть выражено как статистика подстановки, то есть $\eta_{\mathbf{x}}(\mathbf{c}_0) = \eta_{\mathbf{x}}(\mathbf{c}_0, \hat{F})$ для некоторой функции η , где F — эмпирическая функция распределения данных. Если мы решаем задачу регрессии, как в примере с гормоном, то $\eta_{\mathbf{x}}(\mathbf{c}_0) = \mathbf{c}_0 \hat{\beta}$, где $\hat{\beta}$ — это оценка параметра регрессии методом наименьших квадратов. В задаче классификации $\eta_{\mathbf{x}}(\mathbf{c}_0)$ — это предсказанный класс для наблюдения с $\mathbf{c} = \mathbf{c}_0$.

Пусть $\mathcal{Q}[y, \eta]$ обозначает степень ошибки между правильным ответом y и предсказанием η . В регрессии мы часто выбираем $\mathcal{Q}[y, \eta] = (y - \eta)^2$. В классификации $\mathcal{Q}[y, \eta] = I_{y \neq \eta}$, то есть $\mathcal{Q}[y, \eta] = 1$, если $y \neq \eta$ и 0 в противном случае.

Ошибка предсказания для $\eta_{\mathbf{x}}(\mathbf{c}_0)$ определяется как:

$$\text{err}(\mathbf{x}, F) \equiv E_{0F} \{ \mathcal{Q}[Y_0, \eta_{\mathbf{x}}(\mathbf{C}_0)] \}. \quad (17.11)$$

E_{0F} обозначает математическое ожидание нового наблюдения (\mathbf{C}_0, Y_0) из распределения F . Обратите внимание, что E_{0F} не усредняется по набору данных \mathbf{x} , который считается фиксированным. Явная частота ошибок равна:

$$\text{err}(\mathbf{x}, \hat{F}) = E_{0\hat{F}} \{ \mathcal{Q}[Y_0, \eta_{\mathbf{x}}(\mathbf{c}_i)] \} = \frac{1}{n} \sum_1^n \mathcal{Q}[y_i, \eta_{\mathbf{x}}(\mathbf{c}_i)], \quad (17.12)$$

потому что E_{0F} просто усредняет n наблюдаемых случаев (\mathbf{c}_0, y_0) . В случае регрессии, когда $\mathcal{Q}[y, \eta] = I_{y \neq \eta}$, получаем: $\text{err}(\mathbf{x}, \hat{F}) = \sum_1^n [y_i - \eta_{\mathbf{x}}(\mathbf{c}_i)]^2 / n$. В то время как для классификации при $\mathcal{Q}[y, \eta] = I_{y \neq \eta}$, оценка ошибки будет равна $\{\#\eta_{\mathbf{x}}(\mathbf{c}_i) \neq y_i\} / n$ — это есть коэффициент ошибочной классификации по исходному набору данных.

K-fold кросс-валидация из раздела 17.3 также может быть выражена в этих обозначениях. Обозначим через $k(i)$ часть, содержащую i -ое наблюдение, тогда $\eta_{\mathbf{x}}^{-k(i)}(\mathbf{c})$ прогнозируемое значение \mathbf{c} , вычисленное с удалением $k(i)$ -ой части данных. Тогда оценка истинной частоты ошибок при кросс-валидации будет:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{Q}[y_i, \eta_{\mathbf{x}}^{-k(i)}(\mathbf{c}_i)]. \quad (17.13)$$

Чтобы построить бутстреп оценку ошибки предсказания, мы применяем метод подстановки к уравнению (17.11). Пусть $\mathbf{x}^* = \{(\mathbf{c}_1^*, y_1^*), (\mathbf{c}_2^*, y_2^*), \dots, (\mathbf{c}_n^*, y_n^*)\}$ — это бутстреп выборка. Тогда оценка $\text{err}(\mathbf{x}, \hat{F})$, полученная методом подстановки равна:

$$\text{err}(\mathbf{x}^*, \hat{F}) = \frac{1}{n} \sum_{i=1}^n \mathcal{Q}[y_i, \eta_{\mathbf{x}^*}(\mathbf{c}_i)]. \quad (17.14)$$

В этом выражении $\eta_{\mathbf{x}^*}(\mathbf{c}_i)$ — это прогнозируемое значение при $\mathbf{c} = \mathbf{c}_i$, основанное на модели, которая была оценена на бутстреп выборке \mathbf{x}^* .

Мы могли бы использовать $\text{err}(\mathbf{x}^*, \hat{F})$ в качестве нашей оценки, но она включает только одну бутстреп выборку и, следовательно, слишком неустойчива. Вместо этого мы должны сосредоточиться на *средней* ошибке прогноза:

$$E_F[\text{err}(\mathbf{x}, F)], \quad (17.15)$$

где E_F обозначает математическое ожидание по набору данных \mathbf{x} с наблюдениями $\mathbf{x}_i \sim F$. Бутстреп оценка равна:

$$E_{\hat{F}}[\text{err}(\mathbf{x}^*, \hat{F})] = E_{\hat{F}} \sum_{i=1}^n \frac{\mathcal{Q}[y_i, \eta_{\mathbf{x}^*}(\mathbf{c}_i)]}{n}. \quad (17.16)$$

???Intuitively, the underlying idea is much the same as in Figure 8.3: in the "bootstrap world" the bootstrap sample is playing the role of the original sample, while the original sample is playing the role of the underlying population F . Во многом основная идея такая же, как на рис. 8.3: в «бутстреп мире» бутстреп выборка играет роль исходной выборки, в то время как исходная выборка лежит в основе распределения F .

Выражение (17.16) будет идеальной бутстреп оценкой, в случае бесконечно-го числа бутстреп выборок. В случае же когда у нас конечное число B бутстреп выборок мы аппроксимируем это следующим образом. Пусть $\eta_{\mathbf{x}^{*b}}(\mathbf{c}_i)$ — это прогнозируемое значение в \mathbf{c}_i , полученное с помощью модели, которая подобрана на b -ой бутстреп выборке, $b = 1, 2, \dots, B$. Тогда наше приближение $E_{\hat{F}}[\text{err}(\mathbf{x}^*, \hat{F})]$ равно:

$$\hat{E}_{\hat{F}}[\text{err}(\mathbf{x}^*, \hat{F})] = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n \frac{\mathcal{Q}[y_i, \eta_{\mathbf{x}^{*b}}(\mathbf{c}_i)]}{n}. \quad (17.17)$$

Для регрессии формулу можно расписать как: $\sum_{i=1}^n \mathcal{Q}[y_i, \eta_{\mathbf{x}^{*b}}(\mathbf{c}_i)]/n = \sum_{i=1}^n [y_i - \eta_{\mathbf{x}^{*b}}(\mathbf{c}_i)]^2/n$. Это соответствует значениям в левом столбце таблицы 17.1, и их среднее значение, равное (2.52), соотносится с формулой (17.17).

Более совершенный бутстреп подход оценивает смещение $\text{err}(\mathbf{x}, \hat{F})$ как оценку $\text{err}(\mathbf{x}, F)$, а затем корректирует $\text{err}(\mathbf{x}, \hat{F})$, вычитая из него это оцененное смещение. Средний «оптимизм» определяем как:

$$\omega(F) \equiv E_F[\text{err}(\mathbf{x}, F) - \text{err}(\mathbf{x}, \hat{F})]. \quad (17.18)$$

Это среднее разности между истинной и явной ошибкой предсказания по набору данных \mathbf{x} с наблюдениями $\mathbf{x}_i \sim F$. Обратите внимание, что $\omega(F)$ в большинстве случаев будет положительна, потому что явная частота ошибок имеет тенденцию занижать ошибку предсказания. Бутстреп оценка $\omega(F)$ выводится с помощью метода подстановки:

$$\omega(\hat{F}) \equiv E_{\hat{F}}[\text{err}(\mathbf{x}^*, \hat{F}) - \text{err}(\mathbf{x}^*, \hat{F}^*)]. \quad (17.19)$$

Здесь $\omega(\widehat{F}^*)$ — это эмпирическая функция распределения бутстреп выборки \mathbf{x}^* . Аппроксимация для случая не бесконечного числа бутстреп выборок выражается как:

$$\widehat{\omega}(\widehat{F}) = \frac{1}{B \cdot n} \left\{ \sum_{b=1}^B \sum_{i=1}^n \mathcal{Q}[y_i, \eta_{\mathbf{x}^{*b}}(\mathbf{c}_i)] - \sum_{b=1}^B \sum_{i=1}^n \mathcal{Q}[y_{ib}^*, \eta_{\mathbf{x}^{*b}}(\mathbf{c}_i^*)] \right\}. \quad (17.20)$$

В приведенном выше уравнении $\eta_{\mathbf{x}^{*b}}(\mathbf{c}_i^*)$ — это прогнозируемое значение в \mathbf{c}_i^* , полученное с помощью модели, которая была оценена на b -ой бутстреп выборке, $b = 1, 2, \dots, B$, а y_{ib}^* — правильное значение i -го наблюдения для b -ой бутстреп выборки. В таблице 17.1 это оценивается по средней разности второго и третьего столбца, и равно 0.82. Окончательная оценка ошибки предсказания — это явная ошибка плюс смещение в сторону уменьшения явной ошибки, которая находится по формуле (17.20). Окончательная оценка ошибки предсказания:

$$\text{err}(\mathbf{x}, \widehat{F}) + \omega(\widehat{F}), \quad (17.21)$$

что аппроксимируется формулой $\frac{1}{n} \sum_{i=1}^n \mathcal{Q}[y_i, \eta_{\mathbf{x}}(\mathbf{c}_i)] + \widehat{\omega}(\widehat{F})$. В нашем примере это равно $2.20 + 0.82 = 3.02$.

Обратим внимание, что $\omega(\widehat{F})$, и $\widehat{E}[\text{err}(\mathbf{x}^*, \widehat{F})]$ не фиксируют \mathbf{x} (как указано в определении 17.11), а измеряют среднее значение по наборам данных, которые были взяты из распределения \widehat{F} . Улучшенная оценка, которая описывается формулой (17.21), лучше чем простая оценка (17.17), потому что она использует наблюдаемое \mathbf{x} в первом члене $\text{err}(\mathbf{x}, \widehat{F})$, а усреднение входит только в поправочный член $\omega(\widehat{F})$.

17.7 Бутстреп оценка 0.632

Простую бутстреп оценку (17.17) можно записать иначе:

$$E_{\widehat{F}}[\text{err}(\mathbf{x}^*, \widehat{F})] = \frac{1}{B \cdot n} \sum_{b=1}^B \sum_{i=1}^n \mathcal{Q}[y_i, \eta_{\mathbf{x}^{*b}}(\mathbf{c}_i)]. \quad (17.22)$$

Мы можем рассматривать уравнение (17.22) как оценку ошибки прогноза для каждой точки данных (\mathbf{c}_i, y_i) , а затем усреднение ошибки по $i = 1, 2, \dots, n$. Теперь для каждого наблюдения (\mathbf{c}_i, y_i) можно разделить бутстреп выборки на те, которые содержат наблюдение (\mathbf{c}_i, y_i) , и те, которые его не содержат. Ошибка предсказания для наблюдения (\mathbf{c}_i, y_i) , будет больше для бутстреп выборки, которая его не содержит, поскольку такая бутстреп выборка в некотором смысле «далее» от наблюдения (\mathbf{c}_i, y_i) . Идея, лежащая в основе бутстреп оценки 0.632, состоит в том, чтобы использовать ошибку предсказания для корректировки «оптимизма» в явной частоте ошибок только для таких ситуаций.

Пусть ϵ_0 будет средней частотой ошибок, полученной из бутстреп наборов данных, не содержащих прогнозируемое наблюдение (ниже мы приводим подробные сведения об оценке ϵ_0). Как и раньше, $\text{err}(\mathbf{x}, \widehat{F})$ — это явная частота ошибки. Разумным выглядит использование нескольких ϵ_0 — $\text{err}(\mathbf{x}, \widehat{F})$ как «оптимизм» оценку $\text{err}(\mathbf{x}, \widehat{F})$. «оптимизм» оценка 0.632 определяется как:

$$\widehat{\omega}^{0.632} = 0.632[\epsilon_0 - \text{err}(\mathbf{x}, \widehat{F})]. \quad (17.23)$$

Прибавление этой оценки к $\text{err}(\mathbf{x}, \hat{F})$ дает оценку ошибки предсказания 0.632:

$$\begin{aligned}\widehat{\text{err}}^{0.632} &= \text{err}(\mathbf{x}, \hat{F}) + 0.632[\epsilon_0 - \text{err}(\mathbf{x}, \hat{F})] = \\ &= 0.368 \cdot \text{err}(\mathbf{x}, \hat{F}) + 0.632 \cdot \epsilon_0.\end{aligned}\tag{17.24}$$

Аргумент «0.632» возникает из теоретического факта, который показывает что бутстреп выборки, используемые при вычислении ϵ_0 , в среднем дальше, чем типичный тестовый образец, примерно в $1/0.632$ раза. Корректировка в (17.23) исправляет это, поэтому $\widehat{\text{err}}^{0.632}$ примерно несмещено для истинной частоты ошибки. Мы не будем приводить здесь теоретические факты, но отметим, что значение 0.632 возникает потому, что это приблизительная вероятность того, что данное наблюдение появится в бутстреп выборке размера n .

С учетом В бутстреп выборок, мы можем оценить ϵ_0 как:

$$\epsilon_0 = \frac{1}{n} \sum_{i=1}^n \sum_{b \in C_i} \frac{\mathcal{Q}[y_i, \eta_{\mathbf{x}^{*b}}(\mathbf{c}_i)]}{B_i},\tag{17.25}$$

где C_i — это набор индексов бутстреп выборок, не содержащих i -ое наблюдение, а B_i — количество таких бутстреп выборок. В таблице 17.2 показаны номера наблюдений, которые появляются в каждой из 10 бутстреп выборок таблицы 17.1. Например, наблюдение 5 не появляется в бутстреп выборках 3, 4, 8 и 9. В обозначениях уравнения (17.25) $C_i = (3, 4, 8, 9)$. Таким образом, мы будем использовать только эти четыре бутстреп выборки для оценки ошибки предсказания для $i = 5$ наблюдения в уравнении (17.25).

В нашем примере ϵ_0 равно 3.63. Неудивительно, что она больше, чем явная ошибка, равная 2.20, поскольку это средняя ошибка предсказания для наблюдений, которых нет в бутстреп выборке, которая используется для их предсказания. Следовательно, оценка ошибки предсказания 0.632 равна $0.368 \cdot 2.20 + 0.632 \cdot 3.63 = 3.10$, что близко к значению 3.00, которое было получено ранее с помощью улучшенного бутстреп подхода.

Интересно отметить, что средняя ошибка предсказания для наблюдений, которые присутствовали в бутстреп выборке, которая использовалась для их предсказания, составляла 3.08, однако, это значение не используется при построении оценки 0.632.

17.8 Обсуждение

Оценки ошибки предсказания, которые были описаны в этой главе, являются значительными улучшениями явной частоты ошибок. Несколько, какой из этих конкурирующих методов лучше. Асимптотически методы одинаковы, но для выборок небольшого размера могут вести себя по разному. Имитационные эксперименты показывают, что кросс-валидация дает примерно несмещенную оценку, но она может иметь большой (variability)??? разброс. Простой бутстреп метод имеет меньший разброс, но может быть сильно смещен вниз. Усовершенствованный бутстреп подход хоть и дает лучший результат, но все еще смещен вниз. В немногочисленных исследованиях, которые были проведены на сегодняшний день, оценка 0.632 показала лучший результат среди всех методов, но нам нужно больше доказательств, прежде чем давать какие-либо рекомендации.

Функции языка S для вычисления оценки ошибки предсказания с помощью кросс-валидации и бутстрепа описаны в приложении.

Table 17.2. *The observation numbers appearing in each of the 10 bootstrap samples of Table 17.1.*

| Bootstrap sample | | | | | | | | | | |
|------------------|----|----|----|----|----|----|----|----|----|--|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| 1 | 16 | 25 | 1 | 14 | 15 | 14 | 23 | 6 | 5 | |
| 5 | 5 | 4 | 7 | 10 | 24 | 7 | 17 | 26 | 9 | |
| 23 | 16 | 12 | 12 | 2 | 12 | 1 | 15 | 10 | 3 | |
| 11 | 24 | 16 | 7 | 8 | 18 | 6 | 9 | 9 | 3 | |
| 11 | 11 | 14 | 14 | 13 | 15 | 11 | 6 | 27 | 26 | |
| 24 | 14 | 27 | 25 | 5 | 23 | 21 | 22 | 10 | 4 | |
| 15 | 17 | 24 | 1 | 1 | 9 | 22 | 9 | 23 | 25 | |
| 10 | 26 | 7 | 22 | 7 | 8 | 5 | 22 | 7 | 21 | |
| 27 | 11 | 23 | 26 | 1 | 7 | 27 | 3 | 3 | 20 | |
| 26 | 27 | 18 | 4 | 6 | 9 | 25 | 8 | 7 | 15 | |
| 4 | 20 | 14 | 26 | 25 | 25 | 25 | 7 | 9 | 14 | |
| 2 | 10 | 13 | 15 | 25 | 9 | 23 | 26 | 4 | 5 | |
| 5 | 26 | 2 | 9 | 19 | 6 | 22 | 2 | 18 | 7 | |
| 24 | 26 | 27 | 6 | 20 | 22 | 8 | 17 | 11 | 25 | |
| 1 | 22 | 14 | 26 | 5 | 18 | 6 | 17 | 19 | 20 | |
| 27 | 22 | 8 | 7 | 20 | 25 | 23 | 22 | 20 | 16 | |
| 8 | 21 | 3 | 21 | 17 | 2 | 11 | 27 | 21 | 17 | |
| 17 | 21 | 6 | 10 | 25 | 26 | 4 | 22 | 17 | 23 | |
| 9 | 26 | 17 | 17 | 4 | 7 | 22 | 8 | 3 | 12 | |
| 4 | 16 | 27 | 14 | 11 | 21 | 17 | 15 | 11 | 8 | |
| 14 | 14 | 11 | 13 | 21 | 14 | 25 | 24 | 2 | 26 | |
| 14 | 20 | 25 | 18 | 12 | 15 | 7 | 16 | 12 | 19 | |
| 13 | 14 | 8 | 22 | 16 | 24 | 16 | 3 | 8 | 15 | |
| 22 | 23 | 25 | 25 | 24 | 4 | 3 | 19 | 22 | 3 | |
| 8 | 13 | 19 | 24 | 9 | 14 | 27 | 27 | 8 | 9 | |
| 2 | 13 | 26 | 7 | 9 | 27 | 18 | 23 | 1 | 15 | |
| 3 | 16 | 25 | 1 | 18 | 5 | 8 | 3 | 14 | 23 | |