

играет. Чтобы объяснить почему, нам придется еще раз вернуться к параметрическим методам. Они, как мы видели на примере параметрической t-статистики, требуют расчета дисперсии выборки. Из каждого значения в выборке вычитается среднее по выборке, а полученный результат возводится в квадрат. Сумма всех этих квадратов есть дисперсия. Выполнив эти простые расчеты, вы обнаружите, что величина дисперсии почти целиком зависит от значений, намного отличающихся от среднего. Даже у больших выборок дисперсия обычно лишь на 2% зависит от средней трети значений и на 98% — от других двух третей наибольших и наименьших значений выборки. Когда размер выборки не превышает 12 объектов, дисперсия определяется всего двумя экстремальными точками — наибольшим и наименьшим значениями выборки.

Данный не требующий расчетов метод позволяет получить 90-процентный CI, лишь чуть-чуть более широкий, чем при использовании t-статистики, без связанных с ней проблем. Вспомним пример, в котором пять руководителей отделов продаж указали, что тратят на общение с отстающими торговыми представителями 1, 6, 12, 12 и 7 часов в неделю. С помощью t-статистики мы установили, что верхняя граница интервала — 13, но знаем, что по другим соображениям она должна быть ниже, и с помощью нашего не требующего расчетов метода получим интервал 1–12. Мы знаем, что 12 — вполне возможное значение верхней границы интервала, так как это одно из значений нашей выборки. Если отобрать еще шесть руководителей с затратами времени 4, 5, 10, 7, 9 и 10 часов в неделю, то выборка составит 11 объектов. Таблица показывает, что при таком размере выборки границами CI, близкого к 90-процентному, служат третья наибольшая и наименьшая ее значения. С учетом этого получаем 90-процентный CI, составляющий 5–11 часов в неделю. А t-статистика в этом (довольно редком) случае даст нам несколько более широкий интервал 4,5–11,3 часа в неделю.

Важно отметить, что использованный мной непараметрический метод в отличие от t-статистики определяет 90-процентный CI для медианы, а не для среднего значения. Медианой генеральной совокупности называют такое значение, выше которого располагается ровно одна половина ее значений, а ниже — другая. Среднее генеральной совокупности — это сумма всех ее значений, деленная на размер. При смещенном (асимметричном) распределении генеральной совокупности медиана не совпадает со средним значением. Однако если допустить, что распределение близко к симметричному, то медиана и среднее совпадут. В этом случае наша таблица позволит определить 90-процентный CI и для медианы, и для среднего значения.

В некоторых случаях данное допущение оказывается натяжкой, но вообще-то в параметрической статистике мы делаем гораздо более сомнительные допущения. В параметрической статистике мы обязаны

придать графику нашего распределения вполне определенную форму. А оценивая медиану по таблице 9.2, мы не делаем никаких допущений о распределении значений генеральной совокупности. Оно может быть и нерегулярным — горбатым (camel-back) (как график распределения населения США по возрасту, форма которого объясняется произошедшим после войны демографическим взрывом), и равномерным (как график распределения выигрышей при игре в рулетку). Таблица 9.2 позволяет определить диапазон значений медианы и в том, и в другом случаях. Но если распределение к тому же симметрично, неважно, равномерное оно, нормальное, горбатое или типа «бабочки» (bow-tie), то таблица годится и для определения диапазона среднего значения.

## ПРИСТРАСТНЫЙ ОТБОР МЕТОДОВ ВЫБОРОЧНОГО ОБСЛЕДОВАНИЯ

Как обычный работник измерил бы популяцию рыб, обитающих в озере? Этот вопрос я всегда задаю участникам своих семинаров. Обычно в ответ слышу: «Осушил бы озеро». По мнению, например, среднего бухгалтера или даже менеджера среднего звена по ИТ, «измерить» означает «пересчитать». Поэтому когда речь заходит о численности (популяции) рыб, такие люди полагают, что их просят назвать точный итог, а не просто уменьшить неопределенность. С этой мыслью они и предлагают осушить озеро и, несомненно, сумели бы организовать дело так, чтобы каждая мертвая рыбешка была подобрана, брошена в кузов грузовика и сосчитана вручную. Возможно, кто-то пересчитал бы рыбу в грузовике еще раз и осмотрел бы дно осушенного озера, чтобы убедиться в точности подсчетов. Затем они сообщили бы, что всего в озере обитали ровно 22 573 рыбы, так что прошлогодние усилия по пополнению рыбных запасов озера не пропали даром. Правда, теперь вся эта рыба погибла.

А вот если поручить биологам измерить численность рыбной популяции в озере, то уж они не слутают слова «измерить» и «пересчитать». Взамен они, скорее всего, воспользуются методом, состоящим в выпуске пойманной рыбы и повторной ловле. Сначала биологи поймают и пометят некое число, скажем 1000, особей и снова выпустят их в озеро. После того как меченая рыба перемещается с немеченым, они отлавливают еще некое число особей. Допустим, поймали опять 1000 рыб, из которых 50 меченых. Это означает, что помечено 5% всех имеющихся в озере рыб. Зная число первоначально меченых рыб — 1000, биологи делают вывод: в озере около 20 тыс. рыбин (1000 — это 5% от 20 000).

Такого рода выборка подчиняется так называемому биномиальному распределению, но для больших чисел можно считать такое рас-