

Приложение

Метод генерации неполных месячных порций данных с использованием неравномерной рандомизации

Задача формирования выборочной совокупности значений случайной переменной путем неравномерной рандомизации состоит в следующем. Пусть имеется общая совокупность значений размером n (в условиях поставленной задачи $n=28,29\dots31$). Из этой общей совокупности извлекается выборка размером $k < n$. Причем i -й член общей совокупности имеет разную, отличную от 0, вероятность π_i ($i=1\dots n$) включения в выборку k . Рандомизация может производиться либо с повторением (выборка размера k формируется k отборами из полной совокупности, в результате чего отдельные его члены могут быть отобраны несколько раз), либо без повторения. Исходя из поставленной задачи, следует использовать вариант рандомизации без повторений, в результате чего формируется выборка размером k , содержащая k уникальных единиц полной совокупности, т.е. k уникальных номеров суток в месяце.

Как правило, отличие в вероятностях включения π_i связана с разным «размером» или весом i -й единицы полной совокупности. Размер можно выразить некоторым положительным числом, пропорциональным вероятности π_i . Поскольку отсутствует какая-либо другая априорная информация, в качестве такого положительного числа в настоящей работе использовалась плотность нормального распределения с параметрами: m = некоторый номер суток в месяце, $\sigma=10$ суток. Соответственно, сутки с номерами близкими к m имеют больше шансов попасть в выборку, тем самым, моделируется нерегулярность возникновения пропусков вразбивку. Существует множество алгоритмов неравномерного отбора, отличающихся стратегией формирования выборки с учетом заданных вероятностей включения π_i . Подробный обзор существующих алгоритмов неравномерной рандомизации представлены в книге (Tillé, 2006). В настоящих расчетах использовался один их простых методов рандомизации – обобщенный алгоритм Мидзуно. Как показали предварительные расчеты, другие методы, такие как, например, Тиле и метод систематического отбора, дают примерный схожий результат.

Для каждого «полного» месяца равномерным датчиком случайных чисел генерировалось 100 номеров суток в месяце. По каждому из этих номеров центрировалась дифференциальная кривая нормального распределения со стандартным отклонением $\sigma=10$ суток и рассчитывались вероятности включения π_i . Далее из полной совокупности производился отбор значений в выборку размером k с использованием алгоритма Мидзуно (Tillé, 2006). В результате такой «двойной» рандомизации генерировались разнообразные последовательности пропусков вразбивку, неравномерно распределенных в месяце.

Формула расчета стандартной погрешности σ_k и способ статистического оценивания

В результате применения t-теста к L парам «полных» и «неполных» выборок можно сформировать последовательность $x^{(L)}$, каждый член x_i которой является бинарной случайной величиной, принимающей значения 1 (значимое отклонение в средних – «успех») с вероятностью p_k или 0 (незначимое отклонение – «неудача») с вероятностью $(1-p_k)$. Число «успехов» $r = \sum_{i=1}^L x_i$ в последовательности $x^{(L)}$ подчиняется биномиальному закону распределения.

$$Bin(r|p_k, L) = \binom{L}{r} p_k^r (1-p_k)^{L-r}. \quad (\text{П.1})$$

Таким образом, погрешность статистической оценки p_k определяется свойствами биномиальной случайной величины. В дальнейшем индекс k будет опущен для простоты представления формул.

Формула (П.1) справедлива в случае, если вероятность p является неслучайной величиной и достоверно известна. Однако, требовать этого, в действительности, нельзя, – строго говоря, сама p имеет некоторое распределение. В силу отсутствия дополнительной информации в роли такого распределения может выступить двухпараметрическое β -распределение $p \sim Beta(a, b)$, которое часто применяется к величинам типа $p \in (a, b)$ (MacKay, 2003). Форма β -распределения регулируется параметрами a, b .

Если предположить наличие сериальной коррелированности в ряду результатов t-теста, то в первом приближении связанность членов последовательности может аппроксимироваться простой авторегрессионной моделью 1-го порядка, параметр ϕ которой численно равен коэффициенту корреляции между смежными членами ряда и определен на отрезке $\phi \in (0, 1)$. Аналогично p , выборочная оценка параметра ϕ должна иметь некоторое априорное распределение, которое может аппроксимироваться β -распределением $\phi \sim Beta(\varepsilon, \zeta)$. Предполагая, что выборочные параметры могут быть найдены из данных неравномерной рандомизации, рассмотрим формулу для расчета дисперсии оценки r/L .

Если $r/L, p, \phi$ являются случайными величинами и определены в одном и том же вероятностном пространстве, то полная дисперсия (r/L) может быть рассчитана по формуле (Bowsher, Swain, 2012):

$$\mathbf{D}\left(\frac{r}{L}\right) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L} \middle| p, \phi\right)\right] + \mathbf{E}\left[\mathbf{D}\left(\mathbf{E}\left[\frac{r}{L} \middle| p, \phi\right] \middle| p\right)\right] + \mathbf{D}\left[\mathbf{E}\left(\frac{r}{L} \middle| p\right)\right],$$

откуда немедленно следует:

$$\mathbf{D}\left(\frac{r}{L}\right) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L} \middle| p, \varphi\right)\right] + \mathbf{E}[\mathbf{D}(p|p)] + \mathbf{D}(p) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L} \middle| p, \varphi\right)\right] + \mathbf{D}(p). \quad (\text{П.2})$$

В первое слагаемое в правой части ур.(П.2) входит дисперсия выборочного среднего при наличии сериальной корреляции. При большом объеме выборки она рассчитывается по формуле (Wilks, 1997):

$$\mathbf{D}\left(\frac{r}{L} \middle| p, \varphi\right) = V \cdot \frac{s_x^2}{L} = \frac{1}{L} \cdot p(1-p) \cdot V.$$

Здесь множитель V учитывает соответствующее смещение выборочной дисперсии среднего и связан с коэффициентом авторегрессии 1-го порядка φ :

$$V = 1 + 2 \sum_{j=1}^{L-1} \frac{j}{L} \varphi^{L-j}.$$

Продолжив преобразования (П.2) с учетом β -распределений p, φ , получим

$$\begin{aligned} \mathbf{D}\left(\frac{r}{L}\right) &= \mathbf{E}\left[\frac{1}{L} \cdot p(1-p) \cdot V\right] + \mathbf{D}(p) = \frac{1}{L} \cdot \mathbf{E}[p - p^2] \cdot \mathbf{E}(V) + \mathbf{D}(p) = \\ &= \frac{1}{L} \cdot [\mathbf{E}(p) - \mathbf{D}(p) - \mathbf{E}^2(p)] \cdot \left(1 + \frac{2}{L} \sum_{j=1}^{L-1} j \frac{\Gamma(\varepsilon + \zeta) \Gamma(\varepsilon + L - j)}{\Gamma(\varepsilon) \Gamma(\varepsilon + \zeta + L - j)}\right) + \mathbf{D}(p). \end{aligned}$$

Поскольку формулы для расчета математического ожидания и дисперсии величин, подчиняющихся β -распределению, известны (Джонсон и др., 2010), то после очевидных упрощений приведем окончательный вид расчетной формулы для дисперсии r/L :

$$\mathbf{D}\left(\frac{r}{L}\right) = \frac{ab}{L(a+b)^2(a+b+1)} \cdot \left[(a+b) \left(1 + \frac{2}{L} \sum_{j=1}^{L-1} j \frac{\Gamma(\varepsilon + \zeta) \Gamma(\varepsilon + L - j)}{\Gamma(\varepsilon) \Gamma(\varepsilon + \zeta + L - j)} \right) + L \right]. \quad (\text{П.3})$$

Формула (П.3) позволяет оценить стандартную погрешность расчета $\sigma = \sqrt{\mathbf{D}\left(\frac{r}{L}\right)}$ доли «успешных» событий t -теста, а, значит, и вероятности значимого отклонения «неполного» среднего значения.

Точечная оценка параметров распределений, используемых для расчета стандартной погрешности σ при k пропусках, может быть дана с помощью метода максимального правдоподобия применительно к сумме результатов испытаний t -теста r . Для этого, вообще говоря, необходимо построить функцию правдоподобия на пространстве некоторой совокупности выборочных значений $\tilde{r}$. С этой целью $x^{(L)}$ был поделен на N последовательностей $x^{(n)}_l$ ($l=1..N$) меньшей длины $n < L$ (в расчетах n было принято равным 200 для обоих типов пропусков, а $N = \lfloor L/n \rfloor$).

Если принять во внимание законы распределения величин r и p , то безусловным распределением r будет:

$$f(r) = \int_0^1 f(r|p) \cdot f(p) dp = \int_0^1 \text{Bin}(r|p, n) \cdot \text{Beta}(p|a, b) dp,$$

т.е. бета-биномиальное распределение:

$$f(r) = BB(r|n, a, b) = \frac{\Gamma(a+b)\Gamma(a+r)\Gamma(n-r+b)\Gamma(n+1)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)\Gamma(r+1)\Gamma(n-r+1)} \quad (\text{П.4})$$

На основе функции (П.4) строится искомая функция правдоподобия

$$LK(a, b) = \prod_{l=1}^N BB(r_l|n, a, b),$$

максимизация которой позволяет получить оценки параметров $(\hat{a}, \hat{b}) = \arg \max_{a, b} LK(a, b)$.

По набору последовательностей $x^{(n)}_l$ может быть дана точечная оценка параметров $(\hat{\varepsilon}, \hat{\zeta})$ β -распределения коэффициента авторегрессии φ тем же методом максимального правдоподобия. Для этого достаточно построить функционал:

$$LK(\varepsilon, \zeta) = \prod_{l=1}^N \text{Beta}(\varphi_l|\varepsilon, \zeta),$$

где φ_l – коэффициента сериальной корреляции при единичном лаге, оцененный по последовательности $x^{(n)}_l$.

Поиск максимума функций правдоподобия и определение необходимых параметров может быть выполнен градиентным методом, например, методом Ньютона-Рафсона. В заключении стоит отметить, что в расчетах предпочтительней пользоваться логарифмом функции правдоподобия $\ln LK$.