ГЕНЕРИРОВАНИЕ ДЕМОГРАФИЧЕСКОЙ ВЫБОРКИ С ЗАДАННЫМИ ХАРАКТЕРИСТИКАМИ С ПОМОЩЬЮ СЛУЧАЙНЫХ ЧИСЕЛ

Дубровина Н.

к.э.н., PhD, доцент

Высшая Школа Экономики и Менеджмента Публичной Администрации в Братиславе (Словакия)

Гурьянова Л.

д.э.н., профессор

Харьковский национальный экономический университет им. С. Кузнеца (Украина)

Гуляшова И.

Трнавский университет (Словакия)

Имитационное моделирование и вычислительные эксперименты являются важной составляющей исследования случайных процессов в социальной экономике, в социологии, в медицине и организации здравоохранения, когда в силу различных факторов полная информация об исследуемом процессе не известна или не доступна [1, 3] (thesis01-897.html#thesis01-897-ref), а необходимо воспроизвести свойства и характеристики случайного процесса с заданными параметрами или данными [2, 4] (thesis01-897.html#thesis01-897-ref). В приведенном исследовании с помощью простых генераторов случайных чисел, имеющихся в различных программах, надстройках, статистических пакетах [2, 4] (thesis01-897.html#thesis01-897-ref), и ряда гипотез о свойствах случайного процесса, социально-демографических характеристиках выборки [2, 3] (thesis01-897.html#thesis01-897-ref), получена случайная выборка, содержащая 300 наблюдений, которая характеризуется следующими показателями: пол, возраст,

состояние здоровья и вероятность заболевания некоторой болезнью, зависящей от пола, возраста и состояния здоровья, для которой возможно получить экспертные оценки в заданных точках.

Представим кратко основную технологию получения случайной выборки и ряд предположений, которые использованы в данном исследовании. Первое предположение состояло в том, что достаточно сложное распределение случайной величины, имеющей заданные характеристики и отличающееся от нормального распределения, можно представить в виде суммы более простых распределений (например, равномерного распределения, для которого реализованы генераторы случайных чисел в Excel), взятых с определенными весами или пропорциями, или с цензурированными определенными интервалами значений генерируемой случайной величины.

Введем следующие обозначения, которые использованы в данном эксперименте: x – случайная величина, округленная до целых значений, характеризующая возраст от 0 до 100 лет (процент долгожителей в обычных выборках крайне малый, поэтому верхнюю возрастную границу можно принять за 100 лет); y – случайная величина, принимающая значения 0 – для женщин и 1 – для мужчин; z – случайная величина, принимающая дискретные значения: “-1” – для случаев хорошего здоровья; “0” – в случаях нормального, обычного, здоровья и “1” – для случаев плохого здоровья. Величина z связана с переменной h (уровень здоровья) следующим образом, $z = -h$.

Далее рассмотрим функцию для оценки вероятности заболевания некоторой болезнью с учетом влияния трех факторов: возраста x, пола y и величины z, противоположной состоянию здоровья. Предположим, что на основании экспертной информации известны следующие факты: с увеличением возраста вероятность заболевания резко возрастает, вероятность данного заболевания у новорожденных практически равна 0, а у лиц в группе от 95 до 100 лет эта вероятность равна 1; мужчины болеют чаще, чем женщины; для лиц, достигших 50 лет вероятность заболевания равна 0,5; вероятность заболевания также зависит от состояния

здоровья, при наличии плохого состояния здоровья вероятность заболевания увеличивается, а при хорошем состоянии здоровья, наоборот, уменьшается. Были выбраны следующие значения переменных: x принимает значения ”0” – новорожденные, “50” – лица среднего возраста; “100” – лица старшего возраста и долгожители. Переменная y принимала два значения: 0 – для женщин и 1 – для мужчин; переменная z принимала три значения: “-1” – для случаев хорошего здоровья; “0” – в случаях нормального, обычного, здоровья и “1” – для случаев плохого здоровья. Таким образом, общее количество точек, представляющих возможные комбинации вариантов значений переменных x, y и z составляет 18 $(3 \cdot 2 \cdot 3)$. Для этих 18 точек – узлов функции вероятности p мы с помощью экспертов оценили вероятности и представили оценку вероятности p в виде нелинейной функции, зависящей от трех указанных переменных x, y и z, которая учитывает общее и отдельное взаимодействие трех указанных переменных. Были получены оценки параметров для данной функции, значения которой находятся в пределах от 0 до 1, и эти значения учитывают факты, представленные на основе предварительной экспертной информации. Как видно из табл. 1 (thesis01-897.html#tab897-1), данная функция достаточно хорошо аппроксимирует результаты экспертных оценок, коэффициент детерминации составляет более 0,99, т.е. 99% дисперсии значений p объясняется представленной зависимостью.

Таблица 1

Результаты реализации теста Гренджера

Model: p=a1*xyz+a2*xy+a3*xz+a4*yz+a5*x+a6*y+a7*z
Dep. var: P Loss: (OBS-PRED)**2
Final loss: .017891667 R=.99586 Variance explained: 99.173%

	A1	A2	A3	A4	A5	A6
Estimate	0,00015	0,0006	0,00025	-0,0125	0,008267	0,066667

Источник: результаты расчетов авторов в пакете Statistica

Построенная функция p затем использовалась для определения исходов заболеваемости для случайной выборки с значениями величин x, y и z. Переменная исходов была задана дискретным образом и принимала два значения: 1, если $p \geq 0,5$ и 0 – в противном случае.

Представим полученные результаты генерирования случайной демографической выборки с заданными характеристиками. В начале мы получили выборку случайной величины из 100 наблюдений, равномерно распределенную на интервале от 0 до 100. Это была базовая выборка, для которой были сгенерированы случайным образом значения y (пол) и z (характеристика состояния здоровья). Для генерирования переменной y использовалось случайная величина с равномерным распределением от 0 до 1, значения которой потом округлялись до ближайшего целого, т.е. 0 или 1. Значения переменной z генерировались с помощью двух случайных величин z^- , равномерно распределенной на интервале от -1 до 1, и z^+ , равномерно распределенной на интервале от 0 до 1. Полученные значения z^- и z^+ округлялись до ближайших целых, т.е. до -1,

0 и 1 соответственно и находилась алгебраическая сумма этих значений. Затем указанные случайные значения для переменных x, y и z подставлялись в функцию p и находились ее значения, которые затем и определяли исход заболевания.

Помимо базовой выборки были построены аналогичным образом еще две выборки с по 100 случаев каждая, при этом для первой случайной выборки переменная возраста x имела равномерное распределение от 0 до 20 (молодое поколение, т.е. дети дошкольного и младшего школьного возраста, учащиеся средних и старших классов, студенты первых курсов колледжей и вузов), а вторая выборка содержала значения переменной возраста x с равномерным распределением от 15 до 65 лет, т.е. экономически активное население. После того, как были генерированы две вспомогательные выборки, общая выборка была получена объединением базовой выборки и двух указанных выборок, где границы переменных возраста были цензурированы, т.е. искусственно ограничены в соответствии с гипотезами о том, что доли молодого и экономически активного населения должны быть гораздо выше, чем доля пожилого населения и лиц старшего возраста.

В табл. 1 (thesis01-897.html#tab897-2) представлены результаты обработки случайной выборки, содержащей 300 наблюдений, с учетом описанного выше подхода генерирования демографической выборки с заданными характеристиками. Как видно из результатов, представленных в табл. 2 (thesis01-897.html#tab897-2), средний возраст составляет 33 года, минимальное и максимальное значения доверительного 95% интервала оценок значений среднего возраста составляют 30 лет и 36 лет соответственно; граница нижнего квартиля составляет 13 лет, а верхнего – 52 года, медиана составляет 24 года. Данное распределение не симметрично и отличается от нормального закона распределения, что подтверждается и формальными критериями согласия.

Таблица 2

Статистические характеристики выборки

Var	Mean	Confid. -95%	Confid. +95%	Median	Min	Max	Lowe Quar
X	33,32	30,35	36,28	24	0	98	13
P	0,3178	0,2921	0,3435	0,247	0,0054	0,9998	0,134

Источник: результаты расчетов авторов в пакете Statistica

С точки зрения распределения частот выборки по определенным социально-возрастным группам были получены следующие результаты: в интервале возрастной группы [0;5] лет содержалось 35 наблюдений или 11,67% от объема выборки; в интервале (5;15] лет было 55 наблюдений или 18,33%; в интервале (15;20] лет содержалось 47 наблюдений или 15,67%; в интервале (20;65] лет размешалось 118 наблюдений или 39,33% и в интервале (65;100] лет содержалось 45 наблюдений или 15%.

В данной выборке количество мужчин составило 147 человек или 49%, а количество женщин – 153 человек или 51%. Результаты распределения состояния здоровья в выборке были следующими: в 68 случаях (22,67% наблюдений) переменная z принимала значение “-1”, т.е. $h = 1$ и соответственно 22,67% случаев хорошего состояния здоровья; в 157 случаях (52,33% наблюдений) переменная z принимала значение, равное 0, те. $h=0&, что означало нормальное (обычное) здоровье и в 75 случаях (25% наблюдений) переменная z принимала значение “1”, т.е. $h = -1$, т.е. 25% случаев плохого состояния здоровья.

Как видно из результатов, представленных в табл. 2, оценка средней вероятности заболевания некоторой болезнью в данной выборке составляет 0,318, доверительный 95% интервал для оценок средних значений составляет [0,292; 0,343], медиана равна 0,247, а среднее квадратическое отклонение 0,226. Распределение ассиметрично и отличается от

нормального закона распределения. С учетом полученных значений вероятности p в данной выборке количество исходов заболевания составляет 66 случаев или 22%. Доверительный 95% интервал для оценок средней вероятности исхода заболевания равен [0,173; 0,267].

Следует отметить, что представленная демографическая выборка с различными указанными характеристиками может быть использована при планировании и оценке эффективности программ общественного здоровья [1], поскольку данный подход позволяет специалистам в области организации здравоохранения смоделировать различные процессы заболеваемости для той или иной популяции населения, оценить возможные социальные и экономические риски, определить потребность в необходимых ресурсах для проведения лечения и реализации профилактических программ, нацеленных на снижение заболеваемости [1, 3] (thesis01-897.html#thesis01-897-ref).

ЛИТЕРАТУРА

1. Hanzlíková A. a kol. Komunitné ošetrovateľstvo. Osveta, Martin, 2006. 280 s.
2. Kozlíková K, Trnka M. Úvod do spracovania a preyentovania dát v medicíne. EQUILIBRIA, s.r.o., Košice, 2018. 226 s.
3. Žiaková K. a kol. Ošetrovateľstvo. Teória a vedecký výskum. Osveta, Martin, 2009. 322 s.
4. Боровиков В. Statistica: искусство анализа данных на компьютере. Для профессионалов. – СПб.: Питер, 2001. – 656 с.