Джонсон. Одномерные непрерывные распределения

В рамках подготовки заметки о генерировании случайных чисел в Excel обратил внимание на фундаментальный обзор Джонсона с соавторами. Первое издание книги вышло еще в 1970 г., а второе, переведенное на русский язык – в 1994. Это серьезный математический труд, но интересовавшие меня вопросы вполне доступны для понимания)) В книге подробно излагаются свойства большого числа семейств распределений. Часть I: нормальное, логнормальное, Коши, Вейбулла, χ^2-, гамма-, обратное гауссовское, Парето, экспоненциальное. Часть II: логистическое, Лапласа, бета-, равномерное, экстремальных значений, F-, t-, нецентральное χ^2-, нецентральное F-, нецентральное t-, распределение коэффициента корреляции, времени жизни. Издание снабжено обширной библиографией, таблицами и графиками, необходимыми для активной работы с соответствующими семействами распределений. Я представляю отдельные фрагменты, связанные с моими интересами. Дополнения Excel набраны с отступом.

Н. Л. Джонсон, С. Коц, Н. Балакришнан. Одномерные непрерывные распределения (в 2-х частях). — М.: БИНОМ. Лаборатория знаний, 2017. — 703 с. + 603 с.

Купить цифровую книгу в ЛитРес

ГЛАВА 12. Непрерывные распределения (общие сведения)[1]

В случае непрерывных распределений полезно нормирование, т.е. использование случайной величины

$$(12.4)\ \frac{X - E[X]}{\sqrt{Var(X)}}$$

имеющей распределение с нулевым средним и единичным стандартным отклонением. В частности, форму распределения удобно описать заданием нормированных значений нескольких квантилей (т.е. значений случайной величины, в которых функция распределения принимает заданные значения). Следует различать нормированную и стандартную формы распределения. Последняя обычно удобна для получения формул, связанных с функцией плотности. Она может совпасть с нормированной формой, но это необязательно.

Кривая Лоренца

Кривая Лоренца для положительной случайной величины X является графиком отношения

[1] Главы нумеруются непрерывно. Первые 11 глав относятся к книге Одномерные дискретные распределения.

$$(12.7)\ L(F_X(x)) = \frac{E[X|X \le x]F_X(x)}{E[X]}$$

в зависимости от значений $F_X(x)$. Если случайная переменная X представляет годовой доход, то величина $L(p)$ есть доля общего дохода, полученного индивидуумами, имеющими по крайней мере 100p%-й доход. Видно, что $L(p) \le p$, $L(0) = 0$, $L(1) = 1$.

Типичная кривая Лоренца показана на рис. 12.1. Если доход всех индивидуумов равномерен, то $L(p) = p$. Площадь фигуры, ограниченной прямой $L(p) = p$ и кривой Лоренца, можно рассматривать как меру неравномерности дохода или, в более общем виде, как меру изменчивости распределения случайной величины X.

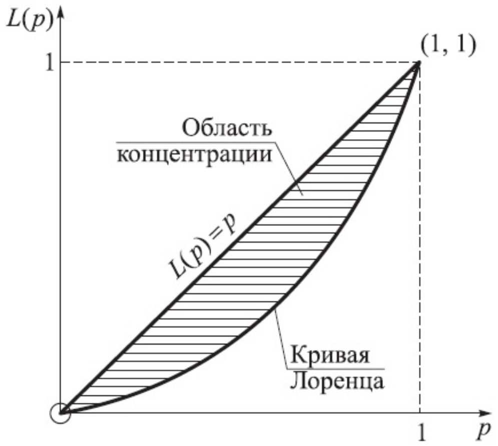


Рис. 12.1. Кривая Лоренца

Порядковые статистики

Порядковые статистики – это последовательность случайных величин, которая получается путем упорядочивания значений из выборки по их величине. Для понимания этого понятия полезно сначала вспомнить, что такое статистика и что такое выборка.

Статистика – это функция от выборки, которая используется для описания или вывода о генеральной совокупности. Примеры статистик: среднее значение, медиана, дисперсия. Выборка – набор наблюдений, полученных из генеральной совокупности.

Порядковые статистики являются функциями от выборки, которые представляют собой упорядоченные значения. Обычно они обозначаются как X_1, X_2, X_n, где n – размер выборки, а индекс указывает на порядковый номер в упорядоченной последовательности.

Примеры порядковых статистик:

Минимум выборки X_1: Это самое маленькое значение в выборке.

Максимум выборки X_n: Это самое большое значение в выборке.

Медиана выборки $X_{(n+1)/2}$: Это значение, которое делит упорядоченную выборку на две равные части.

Первый квартиль $X_{(n+1)/4}$ и третий квартиль $X_{3(n+1)/4}$: Это значения, которые делят упорядоченную выборку на четыре равные части.

n-ый квантиль X_{p*n}: Это значение, которое делит упорядоченную выборку так, что доля p наблюдений меньше этого значения.

ГЛАВА 13. Нормальное распределение

Датчики случайных чисел

В последнее время были построены многие алгоритмы, порождающие псевдо-случайные числа из нормального распределения. Конечно же для того, чтобы порождать псевдослучайные нормальные числа, можно использовать любой датчик равномерно распределенных псевдослучайных чисел в сочетании с функцией, обратной функции распределения (или ее эффективной аппроксимацией). Но были разработаны и другие, более простые, эффективные и быстрые методы; некоторые из них мы опишем здесь.

Метод Бокса–Мюллера

Исходя из двух независимых стандартных нормальных случайных величин X_1 и X_2 Бокс и Мюллер рассмотрели преобразование

$$(13.122)\ Y_1 = e^{\left\{-\frac{1}{2}(X_1^2+X_2^2)\right\}}, \qquad Y_2 = \frac{1}{2\pi} arctg\left(\frac{X_1}{X_2}\right)$$

и показали, что случайные величины Y_1 и Y_2 независимы и равномерно распределены в интервале (0,1). С другой стороны, если мы рассмотрим полярное преобразование

$$(13.123)\ X_1 = r sin\theta,\ \ X_2 = r cos\theta$$

то легко проверить, что плотность совместного распределения случайных величин r и θ имеет вид:[2]

$$(13.124)\ p(r,\theta) = \frac{\theta}{2\pi} e^{-\frac{r^2}{2}},\ \ 0 \le r < \infty,\ \ 0 \le \theta \le 2\pi$$

Случайные величины r и θ статистически независимы. Далее, случайные величины

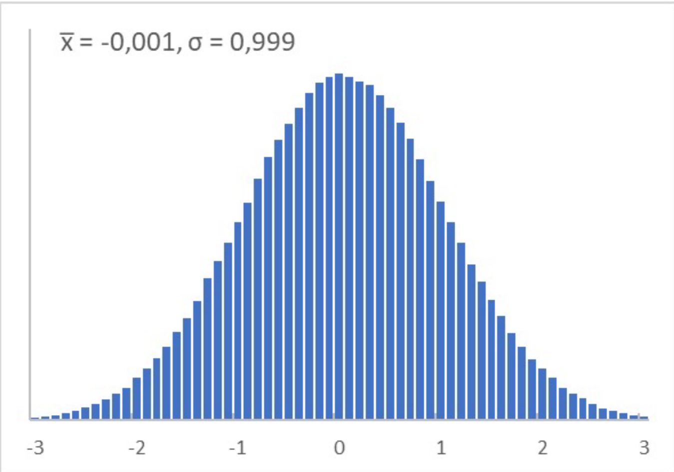
$$(13.125)\ U_1 = e^{-\frac{r^2}{2}}\ \text{ и }\ U_2 = \frac{\theta}{2\pi}$$

независимы и равномерно распределены на (0,1). Обратив преобразование, мы получаем, что величины

$$(13.126)\ X_1 = \sqrt{-2lnU_1}\cos(2\pi U_2)\,,\ \ X_2 = \sqrt{-2lnU_1}\sin(2\pi U_2)$$

представляют собой пару псевдослучайных стандартных нормальных наблюдений.

Я использовал первую формулу из (13.126) и сгенерировал 1M случайных значений. Получилось идеальное стандартное нормальное распределение (см. Excel файл 01. Метод Бокса–Мюллера):



ГЛАВА 14. Логнормальное распределение

Случайную переменную X с логнормальным распределением можно задать соотношением

$$(14.1)\ U = \gamma + \delta \ln(X - \theta)$$

где U – стандартная нормальная случайная величина, а γ, δ и θ – параметры. Из равенства (14.1) следует, что плотность распределения вероятностей случайной величины X имеет вид

$$(14.2)\ p_X(x) = \delta\left[(x-\theta)\sqrt{2\pi}\right]^{-1} \exp\left[-\frac{1}{2}\{\gamma + \delta \ln(X-\theta)\}^2\right], \qquad x > \theta$$

Можно перейти к другим обозначениям, заменив параметры γ и δ на математическое ожидание ζ и стандартное отклонение σ случайной величины $Z = ln(X - \theta)$. Эти два набора параметров связаны соотношениями

$$(14.2a)\ \zeta = -\frac{\gamma}{\delta}, \qquad \sigma = \delta^{-1}$$

[2] В книге приведена иная формула, мне представляется, что она содержит ошибку, и я исправил.

так что равенство (14.1) можно переписать в виде

$$(14.1') \; U = \frac{\ln(X-\theta)-\zeta}{\sigma}$$

а плотность (14.2) принимает форму

$$(14.2') \; p_X(x) = [(x-\theta)\sqrt{2\pi}\sigma]^{-1}\exp\left[-\frac{1}{2}\frac{\{\ln(X-\theta)-\zeta\}^2}{\sigma^2}\right], \qquad x > \theta$$

В большинстве приложений «известно», что параметр θ равен нулю (так что $Pr[X\leq 0] = 0$, или X есть «положительная случайная величина»). Этот важный случай получил название *двухпараметрического логнормального распределения* (с параметрами γ и δ или ζ и σ). При этом формула (14.1) приобретает вид

$$(14.3) \; U = \gamma + \delta \ln X$$

а равенство (14.1') выглядит так:

$$(14.4) \; U = \frac{\ln X - \zeta}{\sigma}$$

Для случайной величины X, подчиненной логнормальному распределению, r-й момент относительно нуля имеет вид

$$(14.7) \; \mu_r' = E[X^r] = E[expr(\zeta + U\sigma)] = \exp\left(r\zeta + \frac{1}{2}r^2\sigma^2\right)$$

Математическое ожидание случайной величины X равно

$$(14.8a) \; \mu_1' = E[X] = \exp\left(\zeta + \frac{1}{2}\sigma^2\right)$$

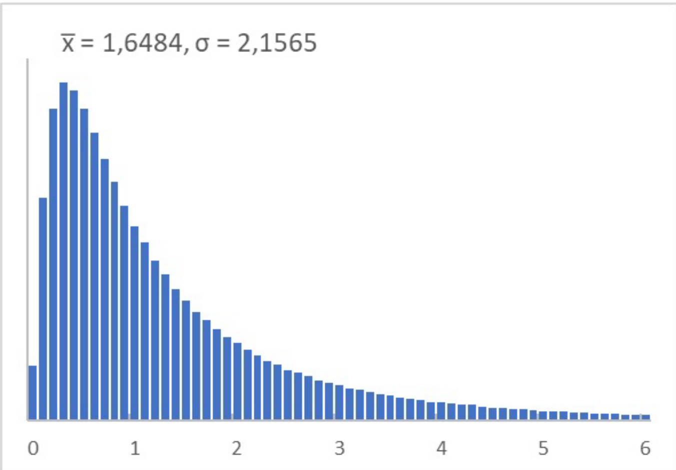
а ее дисперсия составляет

$$(14.8b) \; \mu_2' = D[X] = [\exp(\sigma^2) - 1]\exp(2\zeta + \sigma^2)$$

В Excel я воспользовался функцией ЛОГНОРМ.ОБР(p;μ;σ) и с помощью формулы…

=ЛОГНОРМ.ОБР(СЛМАССИВ(n;;0;1;ЛОЖЬ);μ;σ), n – число случайных значений, μ = 0 и σ = 1

…получил 1М случайных логнормальных значений. Разобрав их по карманам с шагом 0,1, получил (см. Excel файл 02. Логнормальное распределение):

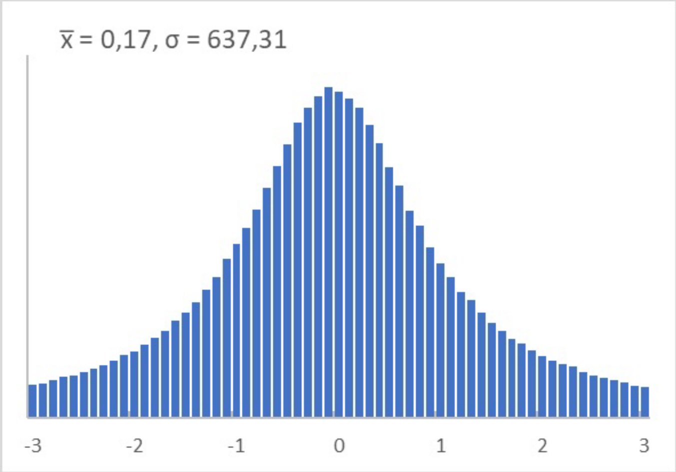


ГЛАВА 16. Распределение Коши

Удобный метод получения выборок из распределения Коши основан на обращении функции распределения $F(x)$. Если переменная U имеет равномерное распределение в интервале (0, 1), то случайная величина $tg(\pi(U - 1/2)$ подчинена стандартному распределению Коши.

В Excel я воспользовался формулой (см. Excel файл 03. Распределение Коши)

=TAN(ПИ()*(СЛМАССИВ(n;;0;1;ЛОЖЬ)-0,5)), n – число случайных значений



ГЛАВА 21. Распределение Вейбулла

Плотность распределения вейбулловской случайной переменной X:

$$(21.4)\ p_X(x) = \frac{c}{\alpha}\left(\frac{x-\xi_0}{\alpha}\right)^{c-1} e^{-\{(x-\xi_0)/\alpha\}^c}, \qquad c>0, \alpha>0, x>\xi_0$$

Если положить $\xi_0 = 0$ и $\alpha = 1$, получим плотность *стандартного* распределения Вейбулла

$$(21.10)\ p_X(x) = cx^{c-1}e^{-x^c}, \qquad c>0, x>0$$

и соответствующую функцию распределения

$$(21.11)\ F_X(x) = 1 - e^{-x^c}, \qquad c>0, x>0$$

Математическое ожидание

$$(21.13)\ E[X] = \Gamma\left(\frac{1}{c}+1\right)$$

Дисперсия

$$(21.13a)\ Var[X] = \Gamma\left(\frac{2}{c}+1\right) - \left\{\Gamma\left(\frac{1}{c}+1\right)\right\}^2$$

Так как функция распределения F случайной величины, подчиненной трехпараметрическому распределению Вейбулла, записывается в аналитической форме (21.4), то соответствующие псевдослучайные наблюдения легко моделировать посредством обращения этой функции распределения. А именно, положим

$$(21.130)\ U = F(X;\xi_0,\alpha,c) = 1 - e^{-\{(x-\xi_0)/\alpha\}^c}$$

и, обратив это преобразование, получим

$$(21.131)\ X = \xi_0 + \alpha\{-\ln(1-U)\}^{1/c}$$

где U случайная величина равномерно распределенная в интервале (0; 1).

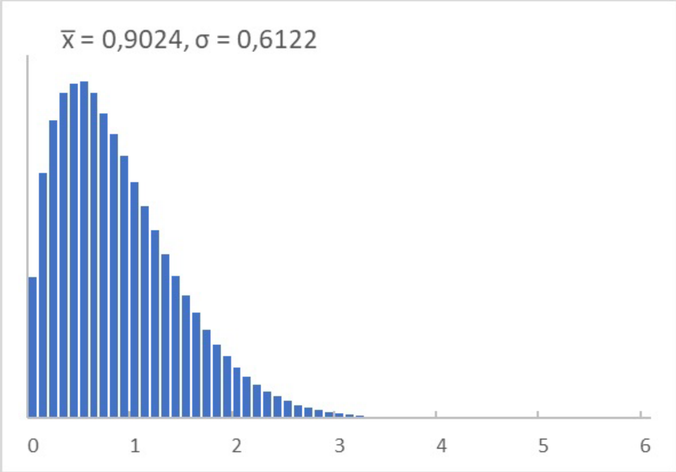
Другой способ моделирования наблюдения из распределения Вейбулла состоит в использовании любого эффективного датчика экспоненциально распределенных случайных чисел. В силу того что случайная величина $(X - \xi_0)/\alpha$ имеет стандартное экспоненциальное распределение, требуемое наблюдение X из распределения Вейбулла можно получить посредством преобразования

$$(21.133)\ X = \xi_0 + \alpha Z^{1/c}$$

где через Z обозначено уже смоделированное псевдослучайное наблюдение со стандартным экспоненциальным распределением.

В Excel я воспользовался формулой (см. Excel файл 04. Распределение Вейбулла)

= α *-LN(1-СЛМАССИВ(n;;0;1;ЛОЖЬ))^(1/c), n – число случайных значений


x̄ = 0,9024, σ = 0,6122
0
1
2
3
4
5
6