BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN**



**TIỂU LUẬN**

**CƠ SỞ TOÁN CHO HỌC MÁY**

**Tên tiểu luận: Chương 5 – Mô hình hoá các phân bố**

.....

Giảng viên HD: **TS. Nguyễn Văn Hậu**

Học viên thực hiện: **Phạm Thị Huê**

Lớp: **H01222**

*Hưng Yên, 6/2023*

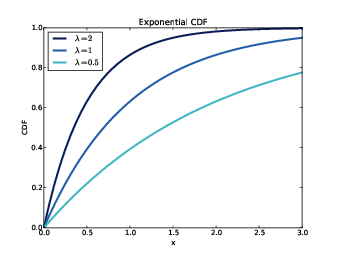
# Mô hình hoá các phân bố

Những phân bố mà chúng ta đã gặp cho đến giờ được gọi là **phân bố kinh nghiệm** vì chúng được dựa trên những quan sát kinh nghiệm, vốn là các mẫu với kích cỡ hữu hạn.

Một phân bố khác là **phân bố lý thuyết**, vốn được đặc trưng bởi một hàm phân bố lũy tích (CDF) là hàm toán học. Các phân bố lý thuyết có thể dùng để mô hình hoá những phân bố thực nghiệm. Ở đây, **mô hình** được hiểu là một cách giản hoá để bỏ qua những chi tiết không cần thiết. Chương này sẽ trình bày các phân bố lý thuyết thông dụng và dùng chúng để mô hình hoá dữ liệu từ nhiều nguồn khác nhau.

Mã lệnh của chương này có trong file analytic.py. Để biết cách tải về và làm việc với mã lệnh này, hãy xem Mục [0.2](https://greenteapress.com/thinkstats2/html/thinkstats2001.html#code).

## 5.1  Phân bố luỹ thừa



|  |
| --- |
| Hình 5.1: Các hàm CDF của những phân bố luỹ thừa với thông số khác nhau. |

Tôi sẽ bắt đầu với **phân bố lũy thừa** vì nó khá đơn giản. Hàm CDF của một phân bố lũy thừa là:

|  |
| --- |
| CDF(x) = 1 − e−λx |

Thông số λ quyết định hình dạng của phân bố. Hình [5.1](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#analytic_expo_cdf) cho thấy CDF này trông ra sao với λ = 0.5, 1, và 2.

Trong thực tế, các phân bố lũy thừa thường được bắt gặp khi ta quan sát một chuỗi các hiện tượng và đo khoảng thời gian giữa hai hiện tượng kế tiếp, mà chúng ta gọi là **thời gian giữa hai sự kiện**. Nếu các sự kiện có vẻ như xảy ra được bất kì lúc nào thì phân bố giữa khoảng thời gian liên tiếp này sẽ có xu hướng tuân theo một phân bố lũy thừa.

Hãy xét một ví dụ về khoảng thời gian giữa hai đứa trẻ chào đời liên tiếp. Vào ngày 18/12/1997, có 44 trẻ được sinh ra ở một bệnh viện tại Brisbane, Úc.[1](https://quangchien.wordpress.com/2012/01/10/ch4-thinkstats/#fn1) Thời điểm chào đời của cả 44 đứa bé được đăng trên một tờ báo địa phương; toàn bộ số liệu có ở một file mang tên babyboom.dat, trong bộ mã lệnh ThinkStats2.

df = ReadBabyBoom()

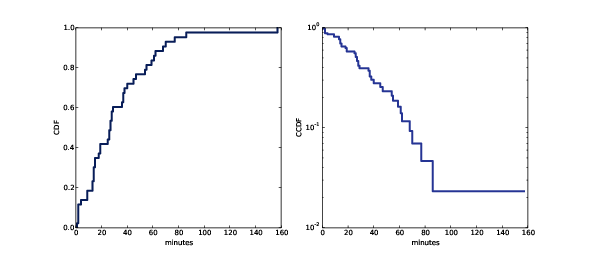
diffs = df.minutes.diff()

cdf = thinkstats2.Cdf(diffs, label='actual'

thinkplot.Cdf(cdf)

thinkplot.Show(xlabel='minutes', ylabel='CDF')

ReadBabyBoom đọc vào file dữ liệu và trả lại một DataFrame với các cột time, sex, weight\_g, và minutes, trong đó minutes là thời điểm chào đời được quy đổi thành số phút kể từ lúc nửa đêm.



|  |
| --- |
| Hình 5.2: CDF của các thời gian giữa hai sự kiện (trái) và CCDF trên hệ trục với trục tung theo thang loga (phải). |

diffs là khác biệt khoảng thời gian giữa các ca sinh, còn cdf là phân bố của khoảng thời gian giữa. Hình [5.2](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#analytic_interarrival_cdf) (trái) cho thấy hàm CDF. Dường như nó có hình dạng tựa như một phân bố lũy thừa, nhưng làm sao ta cho thấy được điều này?

Một cách làm và vẽ đồ thị của **hàm bù CDF**, 1 −  CDF(x), theo trục tỉ lệ log-y. Với số liệu của một hàm lũy thừa, kết quả sẽ là một đường thẳng. Hãy xem rằng tại sao cách này phát huy tác dụng.

Nếu bạn vẽ hàm bù CDF (tức CCDF) của một bộ số liệu mà bạn cho rằng có phân bố lũy thừa, bạn sẽ trông đợi một hàm như:

|  |
| --- |
| y ≈ e−λx |

Lấy loga cả 2 vế ta được:

|  |
| --- |
| logy ≈ −λ x |

Vì vậy theo trục log-y hàm CCDF là một đường thẳng với độ dốc −λ. Sau đây là cách mà ta có thể vẽ được biểu đồ như vậy:

thinkplot.Cdf(cdf, complement=True)

thinkplot.Show(xlabel='minutes',

ylabel='CCDF',

yscale='log')

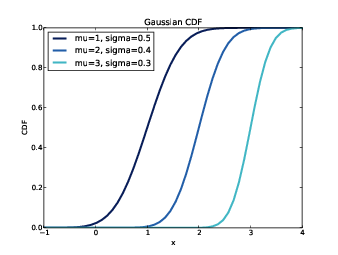
Với đối số complement=True, thinkplot.Cdf tính ra hàm bù CDF trước khi vẽ đồ thị. Và với yscale='log', thinkplot.Show đặt trục y theo thang loga.

Hình [5.2](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#analytic_interarrival_cdf) (phải) cho thấy kết quả. Đường này không hoàn toàn thẳng, nghĩa là phân bố lũy thừa chỉ là một cách xấp xỉ. Trong phần lớn trường hợp thì giả thiết nền tảng ở đây—việc sinh nở có khả năng xảy ra như nhau bất kể giờ trong ngày—là không hoàn toàn đúng. Dù sao, việc mô hình hoá bộ số liệu này bằng một phân bố luỹ thừa cũng là hợp lý. Với cách giản hoá đó, ta có thể tóm lược dạng phân bố này bằng một thông số duy nhất.

Thông số λ có thể được diễn giải như tốc độ thay đổi, nghĩa là số sự kiện xảy ra trung bình trong một đơn vị thời gian. Ở ví dụ này, 44 đứa trẻ chào đời trong khoảng 24 giờ, vậy tốc độ là λ = 0.0306 ca sinh mỗi phút. Giá trị trung bình của phân bố lũy thừa bằng 1/λ, nên thời gian trung bình giữa các ca sinh là 32.7 phút.

## 5.2  Phân bố chuẩn

**Phân bố chuẩn**, còn gọi là phân bố Gauss, là loại thường được dùng nhất vì nó mô tả rất nhiều hiện tượng, chí ít là gần đúng. Hóa ra còn một lý do giải thích được tính đa năng của phân bố này, mà ta sẽ xét đến trong Mục [14.4](https://greenteapress.com/thinkstats2/html/thinkstats2015.html#CLT).



|  |
| --- |
| Hình 5.3: CDF của phân bố chuẩn với một loạt các giá trị thông số. |

Phân bố chuẩn được đặc trưng bởi hai tham số: trị trung bình, µ, và độ lệch chuẩn σ. Phân bố chuẩn với µ=0 và σ=1 được gọi là **phân bố chuẩn hoá**. Hàm CDF của phân bố chuẩn không có dạng biểu thức chính xác nào, nhưng có những thuật toán giúp lập nên hàm này một cách khá hiệu quả. Một trong số đó được cung cấp bởi SciPy: scipy.stats.norm là một đối tượng biểu diễn cho phân bố chuẩn; nó cung cấp phương thức cdf, để tính ra hàm CDF chuẩn hoá:

>>> import scipy.stats

>>> scipy.stats.norm.cdf(0)

0.5

Kết quả này là đúng: số trung vị của phân bố chuẩn hoá thì bằng 0 (đúng bằng trị trung bình), và có một nửa số giá trị lọt xuống dưới mức trung vị, bởi vậy CDF(0) bằng 0.5.

norm.cdf nhận vào hai tham số không bắt buộc: loc, để quy định trị số trung bình, và scale, để quy định độ lệch chuẩn.

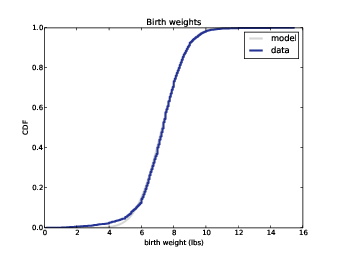
thinkstats2 giúp cho hàm này dễ dùng hơn một chút bằng cách cung cấp EvalNormalCdf, vốn nhận vào các tham số mu và sigma rồi tính ra CDF tại x:

def EvalNormalCdf(x, mu=0, sigma=1):

return scipy.stats.norm.cdf(x, loc=mu, scale=sigma)

Hình [5.3](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#analytic_gaussian_cdf) cho thấy CDF của một phân bố chuẩn với một loạt các giá trị tham số khác nhau. Hàm sigmoid của đường cong này là một đặc trưng dễ nhận biết của phân bố chuẩn.

Trong chương trước ta đã xét đến phân bố của cân nặng trẻ sơ sinh từ NSFG. Hình [5.4](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#analytic_birthwgt_model) cho thấy CDF kinh nghiệm của cân nặng tất cả trẻ được sinh ra và CDF của một phân bố chuẩn có cùng trị trung bình và độ lệch chuẩn.



|  |
| --- |
| Hình 5.4: CDF của cân nặng trẻ sơ sinh theo mô hình phân bố chuẩn. |

Phân bố chuẩn là một mô hình hợp lý cho bộ số liệu này, bởi vậy nếu ta có thể tóm gọn dạng phân bố với hai tham số µ = 7.28 và σ = 1.24, thì sai số thu được (hiệu số giữa mô hình và số liệu) là nhỏ.

Phía dưới số phần trăm thứ 10, đã có sự khác biệt giữa số liệu và mô hình; số liệu cho thấy nhiều trẻ nhẹ cân hơn ta mong đợi từ phân bố chuẩn. Nếu ta cần nghiaên cứu những ca sinh sớm thì rất cần phải mô phỏng đúng phần này của phân bố, vì vậy có lẽ sẽ không tốt nếu dùng mô hình phân bố chuẩn.

## 5.3  Đồ thị phân bố chuẩn

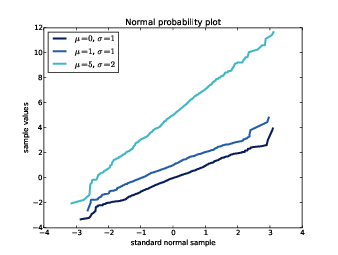
Đối với phân bố lũy thừa và một số loại phân bố khác, có những phép biến đổi đơn giản dùng được để kiểm xem liệu một phân bố liên tục có phải là mô hình tốt cho bộ số liệu hay không.

Đối với phân bố chuẩn thì không có phép biến đổi nào như vậy cả, nhưng có một cách làm thay thế là **đồ thị xác suất chuẩn**. Có hai cách phát sinh ra đồ thị xác suất chuẩn: cách khó và cách dễ. Nếu bạn quan tâm đến cách khó thì có thể đọc qua nó tại <https://en.wikipedia.org/wiki/Normal_probability_plot>. Sau đây là cách dễ:

1. Sắp xếp các giá trị trong mẫu.
2. Từ phân bố chuẩn hoá (μ = 0 và σ = 1), phát sinh một mẫu ngẫu nhiên có cùng kích cỡ với mẫu hiện có, rồi sắp xếp mẫu này.
3. Vẽ đồ thị các giá trị được sắp xếp từ mẫu so với các giá trị ngẫu nhiên.

Nếu dạng phân bố của một mẫu là xấp xỉ phân bố chuẩn athì kết quả sẽ là một đường thẳng cắt trục tung tại mu và có độ dốc bằng sigma. thinkstats2 cung cấp NormalProbability, vốn nhận vào một mẫu và trả lại hai mảng NumPy:

xs, ys = thinkstats2.NormalProbability(sample)



|  |
| --- |
| Hình 5.5: Đồ thị phân bố xác suất cho các mẫu ngẫu nhiên rút từ các phân bố chuẩn. |

ys chứa các giá trị đã sắp xếp từ mẫu sample; xs chứa các giá trị ngẫu nhiên từ phân bố chuẩn hóa.

Để kiểm tra NormalProbability tôi đã phát sinh vài mẫu giả vốn được rút từ chính các phân bố chuẩn với những tham số khác nhau. Hình [5.5](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#analytic_normal_prob_example) cho thấy kết quả. Các đường gần như là thẳng, với các giá trị ở hai đoạn đầu thì sai lệch nhiều hơn so với những giá trị ở gần mức trung bình.

Bây giờ ta hãy thử với số liệu thật. Sau đây là mã lệnh phát sinh một đồ thị xác suất chuẩn cho số liệu cân nặng trẻ sơ sinh từ mục trước. Nó vẽ ra một đường màu xám biểu thị cho mô hình và một đường xanh lam biểu thị cho số liệu.

def MakeNormalPlot(weights):

mean = weights.mean()

std = weights.std()

xs = [-4, 4]

fxs, fys = thinkstats2.FitLine(xs, inter=mean, slope=std)

thinkplot.Plot(fxs, fys, color='gray', label='model')

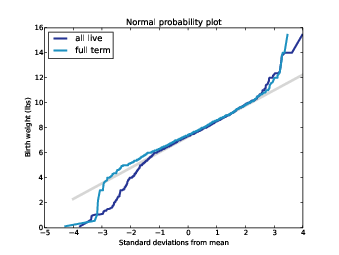
xs, ys = thinkstats2.NormalProbability(weights)

thinkplot.Plot(xs, ys, label='birth weights')

weights là một dãy Series của pandas chứa trị số cân nặng; mean là std trị trung bình và độ lệch chuẩn.

FitLine nhận vào một dãy các giá trị xs, một điểm giao cắt, và một độ dốc; nó trả lại xs và ys để biểu diễn một đường với các tham số đã cho, và ước lượng các giá trị trong xs.

NormalProbability trả lại xs và ys vốn chứa các trị số từ phân bố chuẩn hóa và những trị số từ weights. Nếu dạng phân bố của cân nặng là phân bố chuẩn, thì số liệu sẽ trùng khớp với mô hình.



|  |
| --- |
| Hình 5.6: Đồ thị xác suất chuẩn của cân nặng trẻ sơ sinh (đường xanh đậm: tất cả ca sinh thành công, đường xanh nhạt: những ca sinh đủ tháng). |

Hình [5.6](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#analytic_birthwgt_normal) cho thấy kết quả của tất cả những ca sinh thành công, cũng như các ca sinh đủ tháng (thai kỳ dài hơn 36 tuần). Cả hai đường đều khớp với mô hình ở gần trị trung bình và chệch khỏi mô hình ở hai đoạn đầu. Những bé nặng nhất thì nặng hơn so với cân nặng được mô hình ước tính, còn các bé nhẹ thì có số cân nặng nhỏ hơn so với mô hình ước tính.

Khi chỉ chọn các ca sinh đủ tháng thì ta đã loại bỏ một số cân nhẹ nhất rồi, điều này làm giảm mức độ chệch ở phía đầu thấp của đường phân bố.

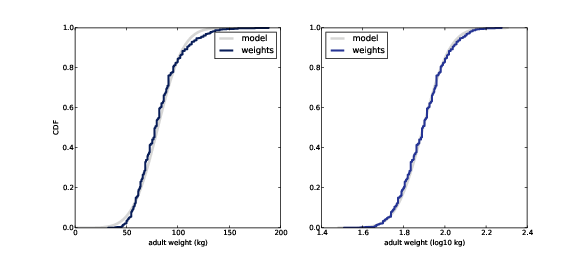
Đồ thị này cho thấy rằng mô hình chuẩn đã mô tả dạng phân bố khá tốt trong khoảng vài độ lệch chuẩn về hai phía trị trung bình, nhưng không tốt lắm ở hai đầu. Liệu rằng mô hình có đủ tốt với các mục đích thực tế không thì còn tùy vào mục đích cụ thể.

## 5.4  Phân bố loga chuẩn

Nếu logarit của một bộ các giá trị hợp thành phân bố chuẩn, thì bản thân các giá trị này sẽ có phân bố **loga chuẩn**. CDF của phân bố loga chuẩn cũng giống như CDF của phân bố chuẩn khi thay log x cho x.

|  |
| --- |
| CDFloga chuẩn(x) = CDFchuẩn(logx) |

Các tham số của phân bố loga chuẩn thường được viết là μ và σ. Song cần nhớ rằng những tham số này không phải là trị trung bình và độ lệch chuẩn; trị trung bình của phân bố loga chuẩn là exp(μ + σ2/2) và độ lệch chuẩn thì lằng nhằng hơn (xem <http://wikipedia.org/wiki/Log-normal_distribution>).



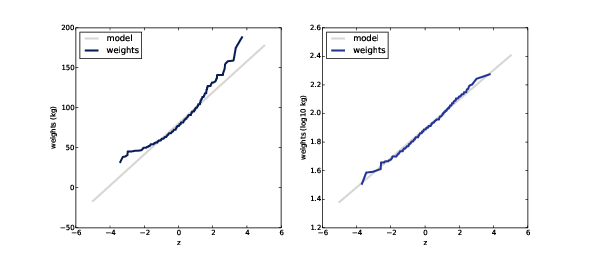
|  |
| --- |
| Hình 5.7: CDF của cân nặng người trưởng thành trên thang tuyến tính (trái) và thang loga (phải). |

Nếu một mẫu xấp xỉ phân bố loga chuẩn và bạn vẽ đồ thị CDF của nó theo thang log-x, thì nó sẽ có hình dáng đặc trưng của một phân bố chuẩn. Để kiểm tra xem mẫu này khớp với mô hình loga chuẩn đến mức nào, bạn có thể vẽ đồ thị phân bố chuẩn sử dụng các trị số loga của các số trong mẫu.

Lấy ví dụ, chẳng hạn ta xem phân bố cân nặng của người lớn; nó có dạng xấp xỉ loga chuẩn.[2](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#note2)

“National Center for Chronic Disease Prevention and Health Promotion” đã tiến hành cuộc điều tra thường niên như một phần của “Behavioral Risk Factor Surveillance System” (BRFSS).[3](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#note3) Vào năm 2008, họ đã phỏng vấn 414.509 người và lấy các thông tin về cấu trúc dân số, sức khỏe và rủi ro liên quan đến sức khỏe. Trong số dữ liệu thu thập được có cân nặng tính theo ki-lô của 398.484 người.

Kho dữ liệu của cuốn sách này có chứa CDBRFS08.ASC.gz, một file dạng ASCII có bề rộng cố định, bao gồm số liệu từ BRFSS, và file brfss.py, mã lệnh để đọc và phân tích số liệu.



|  |
| --- |
| Hình 5.8: Đồ thị phân bố chuẩn của cân nặng người trưởng thành trên thang tuyến tính (trái) và thang loga (phải). |

Hình [5.7](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#brfss_weight) (trái) cho thấy dạng phân bố của cân nặng người trưởng thành trên thang tuyến tính với mô hình chuẩn. Hình [5.7](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#brfss_weight) (phải) cho thấy cũng phân bố đó nhưng trên thang loga với mô hình loga chuẩn. Mô hình loga chuẩn thì khớp hơn, nhưng cách biểu diễn số liệu như thế này không cho thấy rõ sự khác biệt.

Hình [5.8](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#brfss_weight_normal) cho thấy đồ thị xác suất chuẩn cho cân nặng người trưởng thành, w, cùng trị số logarit tương ứng, log10 w. Bây giờ có thể thấy rõ là số liệu đã chệch hẳn khỏi mô hình chuẩn. Mặt khác, mô hình loga chuẩn lại khớp với số liệu.

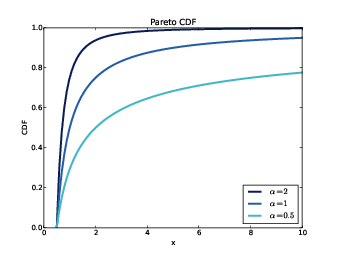
## 5.5  Phân bố Pareto

**Phân bố Pareto** được đặt tên theo nhà kinh tế học Vilfredo Pareto, người đã dùng nó để mô tả sự phân bố mức giàu nghèo (xem <http://wikipedia.org/wiki/Pareto_distribution>). Từ đó, nó được dùng để mô tả những hiện tượng trong khoa học tự nhiên và xã hội bao gồm quy mô các thành phố, kích cỡ các hạt cát hay thiên thạch, quy mô cháy rừng và động đất.

CDF của phân bố Pareto là:

\mathrm{CDF}(x) = 1 - \left( \frac{x}{x_m} \right)^{-\alpha} 

Các tham số xm và α quyết định vị trí và hình dạng của phân bố. xm là giá trị khả dĩ nhỏ nhất. Hình [5.9](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#analytic_pareto_cdf) cho thấy CDF của các phân bố Pareto với tham số xm = 0,5 và những giá trị khác nhau của α.



|  |
| --- |
| Hình 5.9: CDF của các phân bố Pareto với những tham số khác nhau. |

Có một cách kiểm tra đơn giản bằng mắt thường để phát hiện rằng liệu một phân bố kinh nghiệm có khớp với một phân bố Pareto hay không: trên thang log-log, hàm CCDF sẽ trông như một đường thẳng. Hãy xem tại sao có điều đó.

Nếu bạn vẽ đồ thị CCDF của một mẫu tuân theo phân bố Pareto lên một thang tuyến tính thì bạn sẽ lường trước một hàm số có dạng:

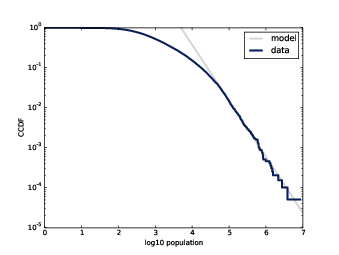
y ≈ (x / xm)−α

Lấy loga cả hai vế ta được:

|  |
| --- |
| logy ≈ −α (logx − logxm) |

Vì vậy nếu bạn vẽ log y theo log x, nó sẽ có hình dạng như một đường thẳng với độ dốc  −α và giao điểm  −α log xm với trục tung.

Lấy ví dụ, ta hãy xét kích cỡ các thành thị. Cục điều tra Hoa Kỳ (U.S. Census Bureau) công bố số dân của tất cả thành thị toàn nước Mỹ.



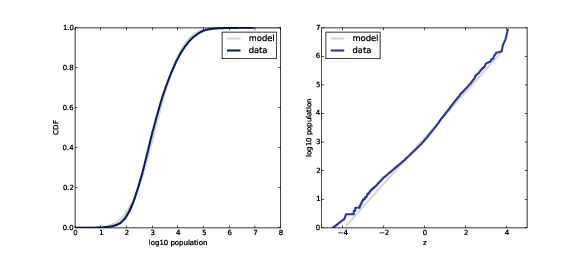
|  |
| --- |
| Hình 5.10: CCDF của dân số các thành thị trên thang log-log. |

Tôi đã tải số liệu của họ về từ <http://www.census.gov/popest/data/cities/totals/2012/SUB-EST2012-3.html>; nó nằm trong kho mã lệnh đi kèm cuốn sách này, ở file có tên PEP\_2012\_PEPANNRES\_with\_ann.csv. Kho mã lệnh cũng chứa populations.py, vốn đọc vào file rồi vẽ các đồ thị phân bố của dân số.

Hình [5.10](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#populations_pareto) cho thấy CCDF của các dân số trên thang log-log. Những thành thị trong nhóm 1%, dưới mức 10−2, thì giảm dần theo một đường thẳng. Vì vậy có thể kết luận, như một số nhà nghiên cứu đã nhận định, rằng đuôi của dạng phân bố này khớp với một mô hình Pareto.

Mặt khác, một phân bố loga chuẩn cũng mô hình hóa dữ liệu khá tốt. Hình [5.11](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#populations_normal) cho thấy CDF của dân số và một mô hình loga chuẩn (bên trái), và một đồ thị phân bố chuẩn (bên phải). Cả hai đồ thị đều cho thấy sự phù hợp giữa số liệu và mô hình.

Không mô hình nào trong số hai mô hình này là hoàn thiện. Mô hình Pareto chỉ áp dụng với nhóm 1% những thành phố lớn nhất, nhưng lại khớp tốt hơn ở khúc đó của phân bố. Còn mô hình loga chuẩn thì khớp tốt hơn cho phần 99% còn lại. Việc mô hình nào phù hợp hơn thì phụ thuộc vào phần nào của phân bố cần được quan tâm.



|  |
| --- |
| Hình 5.11: CDF của dân số thành thị trên thang log-x (trái), và đồ thị phân bố chuẩn của các dân số sau khi biến đổi loga (phải). |

## 5.6  Phát sinh các số ngẫu nhiên

Các CDF lý thuyết có thể được dùng để phát sinh ra số ngẫu nhiên với một hàm phân bố cho trước, p = CDF(x). Nếu có một cách làm hiệu quả để tính ra hàm CDF nghịch đảo, thì ta sẽ có thể phát sinh ra những giá trị ngẫu nhiên với dạng phân bố thích hợp bằng cách chọn một phân bố đều từ 0 đến 1, rồi chọn x = ICDF(p).

Chẳng hạn, CDF của phân bố lũy thừa là

|  |
| --- |
| p = 1 − e−λx |

Giải theo x ta được:

|  |
| --- |
| x = −log(1 − p) / λ |

Vì vậy trong Python, ta có thể viết

def expovariate(lam):

p = random.random()

x = -math.log(1-p) / lam

return x

expovariate nhận vào lam rồi trả lại một giá trị ngẫu nhiên chọn từ phân bố lũy thừa với tham số lam.

Có hai lưu ý với đoạn mã lệnh trên: Tôi gọi tham số là lam vì lambda trùng với một từ khóa của Python. Ngoài ra, vì log0 vô định, nên ta phải cẩn thận một chút. Phương thức random.random có thể trả lại giá trị 0 nhưng không thể trả lại 1, vì vậy 1 − p có thể bằng 1 nhưng không bằng 0, vì vậy log(1-p) sẽ luôn được xác định.

## 5.7  Tại sao cần có mô hình?

Ở đầu chương này, tôi đã nói rằng nhiều hiện tượng thực tế có thể được mô hình hóa bởi phân bố liên tục. Bạn có thể hỏi, “Vậy thì sao?”

Cũng như tất cả mô hình, các phân bố liên tục đều trừu tượng, theo nghĩa chúng lược bỏ tất cả những chi tiết nào được coi là thừa. Chẳng hạn, một phân bố được quan sát có thể chứa những sai số đo đạc hay nhiễu đặc thù của mẫu đó; các mô hình lý thuyết thì làm trơn tất cả những biến động này.

Mô hình lý thuyết cũng là một hình thức nén dữ liệu. Khi một mô hình khớp phù hợp với bộ số liệu thì một tập hợp ít các tham số có thể tóm tắt được cả một lượng số liệu rất lớn.

Đôi khi thật ngạc nhiên khi số liệu từ một hiện tượng tự nhiên lại khớp một phân bố lý thuyết, nhưng các quan sát này có thể dẫn tới nhìn nhận sâu sắc về hệ vật lý. Đôi khi chúng ta có thể giải thích tại sao một phân bố quan sát được lại có dạng riêng nào đó. Chẳng hạn, phân bố Pareto thường là kết quả của các quá trình phát sinh với phản hồi tích cực (thường gọi là quá trình gắn kết theo ý thích: hãy xem <http://wikipedia.org/wiki/Preferential_attachment>.).

Các phân bố liên tục rất thích hợp cho việc phân tích toán học, như ta sẽ được thấy ở Chương [14](https://greenteapress.com/thinkstats2/html/thinkstats2015.html#analysis).

Nhưng điều quan trọng cần nhớ rằng tất cả mô hình đều không hoàn hảo. Số liệu trên thực tế không bao giờ khớp hoàn toàn với phân bố lý thuyết. Đôi khi người ta nói như thể số liệu được phát sinh từ mô hình; chẳng hạn có thể nói rằng phân bố của chiều cao người có dạng chuẩn, hay phân bố thu nhập có dạng loga chuẩn. Theo nghĩa đen thì điều này không thể đúng; luôn có sự khác biệt giữa thực tế và mô hình toán.

Các mô hình có ích nếu chúng nắm bắt được các khía cạnh có liên quan của thực tế và bỏ qua những chi tiết không cần thiết. Như thế nào là “có liên quan” hay “không cần thiết” thì lại phụ thuộc vào mục đích bạn dùng mô hình để làm gì.

## 5.8  Bài tập

Với các bài tập sau đây, bạn có thể bắt đầu từ file chap05ex.ipynb. Lời giải của tôi có ở chap05soln.ipynb.

**Bài 1** Trong bộ số liệu BRFSS (xem Mục [5.4](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#lognormal)), dạng phân bố của chiều cao thì xấp xỉ phân bố chuẩn với các tham số µ = 178 cm và σ = 7.7 cm cho nam giới, và µ = 163 cm and σ = 7.3 cm cho phụ nữ.

Để gia nhập nhóm Blue Man Group, bạn phải là nam giới và cao từ 5’10” đến 6’1”. Có bao nhiêu phần trăm số nam giới Hoa Kỳ thuộc khoảng này? Gợi ý: hãy dùng *scipy.stats.norm.cdf*.

**Bài 2** Để hình dung được phân bố Pareto, hãy tưởng tượng thế giới sẽ ra sao nếu như cân nặng của con người tuân theo phân bố Pareto. Chọn các tham số xm = 1 m và α = 1.7, ta thu được một phân bố với chiều cao tối thiểu hợp lý là 1 m, và số trung vị 1.5 m.

Hãy vẽ dạng phân bố này. Giá trị trung bình chiều cao trong thế giới Pareto bằng bao nhiêu? Có bao nhiêu phần trăm dân số với chiều cao dưới trị trung bình? Nếu có 7 tỉ dân trong thế giới Pareto, thì theo bạn có bao nhiêu người sẽ cao hơn 1 km? Người cao nhất trong thế giới Pareto này sẽ cao bao nhiêu?

**Bài 3**

Phân bố Weibull là một dạng tổng quát của phân bố lũy thừa, xuất hiện trong phân tích sự cố CDF của nó là

|  |
| --- |
| CDF(x) = 1 − e−(x/ λ)k |

Bạn có thể tìm được một phép biến đổi nào khiến cho phân bố Weibull trở nên giống đường thẳng không? Khi đó độ dốc và tung độ giao điểm sẽ biểu thị điều gì?

Hãy dùng *random.weibullvariate* để phát sinh một mẫu từ phân bố Weibull rồi dùng nó để thử nghiệm phép biến đổi của bạn.

**Bài 4** Với các giá trị n nhỏ, ta không mong đợi rằng phân bố thực nghiệm sẽ khớp đúng với phân bố lý thuyết. Một cách để đánh giá mức độ khớp là phát sinh ra một mẫu từ một phân bố lý thuyết và xem nó khớp với số liệu đến đâu.

Chẳng hạn, ở mục [5.1](https://greenteapress.com/thinkstats2/html/thinkstats2006.html#exponential) ta đã vẽ đồ thị phân bố khoảng thời gian giữa các ca sinh và thấy rằng dạng phân bố này xấp xỉ phân bố lũy thừa. Nhưng phân bố này mới chỉ dựa trên 44 điểm số liệu. Để xem rằng liệu số liệu có bắt nguồn từ phân bố lũy thừa hay không, hãy phát sinh ra 44 giá trị từ một phân bố lũy thừa có cùng trị trung bình như số liệu, tức là khoảng 33 phút giữa các ca sinh.

Hãy vẽ đồ thị phân bố các giá trị ngẫu nhiên và so sánh nó với phân bố thực tế. Bạn có thể dùng *random.expovariate* để phát sinh ra các giá trị.

**Bài 5** Trong kho mã lệnh của cuốn sách này, bạn sẽ tìm thấy một tập hợp các file dữ liệu có tên *mystery0.dat*, *mystery1.dat*, v.v. Mỗi file có chứa một dãy các số ngẫu nhiên được phát sinh từ một dạng phân bố lý thuyết.

Bạn cũng sẽ thấy *test\_models.py*, một file lệnh để đọc dữ liệu từ một file và vẽ CDF theo một loạt các phép biến đổi. Có thể chạy file này như sau:

$ python test\_models.py mystery0.dat

Dựa trên các đồ thị này, bạn có thể suy diễn được dạng phân bố nào đã dùng để phát sinh ra từng file. Nếu vướng mắc, bạn có thể ngó vào file *mystery.py*, trong đó chứa mã lệnh để phát sinh các file.

**Bài 6**

Đôi khi, người ta dùng các dạng phân bố loga chuẩn và Pareto để mô hình hóa các phân bố về của cải và thu nhập. Để xem dạng nào tốt hơn, hãy nhìn vào vài số liệu sau.

Chương trình Current Population Survey (CPS) được thực hiện bởi Cơ quan thống kê lao động (Bureau of Labor Statistics) và Cục thống kê (Census Bureau) để tìm hiểu thu nhập và các biến số liên quan. Số liệu đã thu thập năm 2013 có ở <http://www.census.gov/hhes/www/cpstables/032013/hhinc/toc.htm>. Tôi đã tải về *hinc06.xls*, vốn là một bảng tính Excel với thông tin về thu nhập hộ gia đình, rồi chuyển nó về dạng *hinc06.csv*, một file CSV mà bạn sẽ thấy ở kho mã lệnh cuốn sách này. Bạn cũng sẽ thấy *hinc.py*, mã lệnh đọc file dữ liệu nêu trên.

Hãy xuất ra phân bố của thu nhập từ bộ số liệu này. Có dạng phân bố lý thuyết trong chương này là mô hình phù hợp cho số liệu trên không? Một lời giải bài tập này có ở file *hinc\_soln.py*.

## 5.9  Thuật ngữ

* **phân bố thực nghiệm (empirical distribution)**: Phân bố của các giá trị trong một mẫu.
* **phân bố lý thuyết (analytic distribution)**: Phân bố mà hàm luỹ tích (CDF) của nó là một hàm giải tích.
* **mô hình (model)**: Một cách giản hóa có ích. Các phân bố lý thuyết thường là mô hình tốt cho những phân bố thực nghiệm phức tạp hơn.
* **khoảng thời gian giữa (interarrival time)**: Khoảng thời gian trôi qua giữa hai sự kiện.
* **hàm phân bố luỹ tích bù (complementary CDF)**: Một hàm ánh xạ từ giá trị x đến tỉ lệ các giá trị vượt quá x, nghĩa là 1 − CDF(x).
* **phân bố chuẩn hoá (standard normal distribution)**: Dạng phân bố chuẩn có trị trung bình bằng 0 và độ lệch chuẩn bằng 1.
* **đồ thị xác suất chuẩn (normal probability plot)**: Đồ thị biểu diễn các giá trị trong một mẫu so với các giá trị ngẫu nhiên rút ra từ một phân bố chuẩn hoá.